DEUTSCHE
BUNDESBANK
EUROSYSTEM

Discussion Paper

Deutsche Bundesbank
No 45/2022

A nonlinear generalization of the Country-Product-Dummy method

Ludwig von Auer
(Universität Trier)

Sebastian Weinand
(Deutsche Bundesbank)

Deutsche Bundesbank, Wilhelm-Epstein-Straße 14, 60431 Frankfurt am Main,
Postfach 10 06 02, 60006 Frankfurt am Main

Tel +49 69 9566-0

Please address all orders in writing to: Deutsche Bundesbank,
Press and Public Relations Division, at the above address or via fax +49 69 9566-3077

Internet http://www.bundesbank.de

ISBN 978–3–95729–924–6
ISSN 2749–2958

Non-technical summary

Research Question

The Country-Product-Dummy (CPD) method is a linear regression approach widely used for comparisons of regional price levels. The CPD method implicitly assumes that all products of the consumption basket exhibit a uniform regional price dispersion. However, the prices of most consumption goods are relatively constant across regions, say, while the cost of housing varies considerably more. Such empirical observations raise several fundamental questions. What are the statistical consequences when the CPD method is applied, even though the price dispersion is 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 the same price dispersion? Are there alternative estimation methods that remain unbiased even when price dispersion is product-specific?

Contribution

As a solution to these problems, the present paper introduces the NLCPD method, a non-linear generalization of the CPD method. Both index number methods estimate the regional price levels and the general values of the individual products. However, only the NLCPD method also provides estimates of the price dispersion of the various products. These estimates indicate whether the assumption of a uniform price dispersion would be justified.

Results

The present paper shows that the CPD method's statistical inference is invalid when there is product-specific price dispersion. Even worse, the estimates of the regional price levels are biased, unless the set of price data is complete, or the data gaps occur completely at random. By contrast, the regional price levels estimated by the NLCPD method remain unbiased even when the price data exhibit product-specific price dispersion and systematic data gaps exist. But also in cases where the data set is complete or the data gaps are completely at random, the NLCPD method outperforms the CPD method. Finally, the NLCPD method is applied to regional price information derived from Germany's consumer price index micro data of May 2019, resulting in a price index for the 401 regions of Germany.

Nichttechnische Zusammenfassung

Fragestellung

Die Country-Product-Dummy (CPD)-Methode ist ein linearer Regressionsansatz, der häufig für regionale Preisniveauvergleiche genutzt wird. Die CPD-Methode unterstellt allen Produkten des Warenkorbs eine einheitliche regionale Preisstreuung. Allerdings unterscheiden sich zum Beispiel die Preise vieler Konsumgüter regional kaum voneinander, während Mieten deutlich stärker schwanken. Solche empirischen Beobachtungen werfen mehrere grundlegende Fragen auf. Welche statistischen Konsequenzen ergeben sich, wenn die CPD-Methode angewendet wird, obwohl die Preisstreuung produktspezifisch ist? Werden die regionalen Preisniveaus weiterhin unverzerrt geschätzt? Ist Inferenz nach wie vor zulässig? Falls nicht, gibt es eine Möglichkeit zu prüfen, ob eine Auswahl an Produkten die gleiche Preisstreuung aufweist? Existieren alternative Schätzmethoden, die auch bei einer produktspezifischen Preisstreuung unverzerrt bleiben?

Beitrag

Als Lösung dieser Probleme stellt das vorliegende Papier die NLCPD-Methode vor, eine nichtlineare Verallgemeinerung der CPD Methode. Beide Indexmethoden schätzen die regionalen Preisniveaus und die überregionalen Durchschnittspreise der einzelnen Produkte. Einzig die NLCPD-Methode liefert jedoch auch Schätzwerte für die Preisstreuung der Produkte. Diese Schätzer geben an, ob die Annahme einer einheitlichen Preisstreuung gerechtfertigt ist.

Ergebnisse

Das vorliegende Papier zeigt, dass statistische Inferenz der CPD-Methode im Falle produktspezifischer Preisstreuung nicht zulässig ist. Zudem werden die Preisniveaus verzerrt geschätzt, außer wenn die vorliegenden Preisdaten vollständig sind oder Lücken zufällig auftreten. Im Gegensatz dazu schätzt die NLCPD-Methode die Preisniveaus selbst dann unverzerrt, wenn die Preisdaten eine produktspezifische Preisstreuung und systematische Datenlücken aufweisen. Aber auch wenn der Datensatz vollständig ist oder Datenlücken zufällig auftreten, schneidet die NLCPD-Methode besser als die CPD-Methode ab. Schließlich wird die NLCPD-Methode auf regionale Preisinformationen angewendet, welche von Mikrodaten des deutschen Verbraucherpreisindex im Mai 2019 abgeleitet wurden. Dies liefert einen Preisindex für die 401 Regionen Deutschlands.

Deutsche Bundesbank Discussion Paper No 45/2022

A Nonlinear Generalization of the Country-Product-Dummy Method*

Ludwig von Auer
Universität Trier

Sebastian Weinand
Deutsche Bundesbank

November 16, 2022

Abstract

The present paper shows that product-specific regional price dispersion usually causes the Country-Product-Dummy (CPD) method to be biased. In cases where it is not, this index number method is still inefficient and inference is invalid. In view of this, a nonlinear generalization of the CPD method has been developed. This NLCPD method can be employed at all levels of aggregation and allows for inference. A comprehensive simulation reveals that the NLCPD method's root mean squared error is smaller than that of the CPD method, even in cases where the latter is unbiased. Finally, this paper applies the NLCPD method to regional price information derived from Germany's consumer price index micro data. Price levels of the 401 German districts are computed.

Keywords: multilateral price index · regional price levels · CPD method · measurement bias

JEL classification: C43 · E31

* Correspondence to vonauer@uni-trier.de or sebastian.weinand@bundesbank.de. 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 seminar participants at the Deutsche Bundesbank as well as the "Statistische Woche 2022" in Münster. The paper has greatly benefited from the generous advice we received from Gholamreza Hajargasht. The opinions expressed in this paper are those of the authors and do not necessarily reflect the views of the Deutsche Bundesbank or the Eurosystem.

1 Introduction

Important areas of economic theory and economic policy utilize regional indicators of 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 are not hard to find. The production of regional real indicators requires reliable information on *regional price levels*, while national 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, pp. 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 case studies of countries from all over the world (surveyed by Majumder and Ray, 2020, pp. 111-113 and Weinand and Auer, 2020, pp. 416-419). The choice between the various methods also depends on the available data set. Some studies cover only parts of a country. Others cover the complete country, but the regions 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. Summers (1973) proposes the Country-Product-Dummy (CPD) method for such cases. This linear regression approach also allows for statistical inference. However, the CPD method implicitly assumes that the variance of the logarithmic prices across regions is identical for all included products. Put more simply, the products' (regional) price dispersion is uniform. Whether this assumption is justified is an empirical question. The higher the level of aggregation and the more heterogeneous the included products (e.g., pasta versus shoes), the less plausible the CPD method's assumption of a uniform price dispersion.

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). The aggregation of these regional price levels into the overall regional price levels is usually conducted by employing an alternative method. The final result therefore 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 price dispersion 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 price dispersion? Are there alternative estimation methods that remain unbiased even when price dispersion is not uniform, but product-specific?

The present paper answers all of these questions. When there is product-specific price dispersion, the CPD method's statistical inference is invalid. Even worse, the estimates of the regional price levels are biased unless the set of price data is complete (a situation in which the CPD method is rarely used) or the data gaps occur completely at random (a situation that is difficult to achieve in real-world price data samples).

As a solution to these problems, this paper introduces the *NLCPD method*, a nonlinear generalization of the CPD method. Both of these multilateral index number methods compute the regional price levels and the general values of the individual products. However, only the NLCPD method also provides estimates of the price dispersion of the various products. These estimates indicate whether the assumption of a uniform price dispersion 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 price dispersion 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 price dispersion.

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 use of the NLCPD method. Section 6 applies this method to a large data set of regional prices. Section 7 concludes.

2 Problem

In subnational price comparisons, the prices of manufactured goods are found to be rather uniform across the regions, while the cost of housing varies considerably (e.g., Weinand and Auer, 2020, pp. 430-431 for Germany; Aten, 2017, pp. 130-131 for the United States). The prices of services take an intermediate position. Tab. 1 shows the same features. It lists the prices of three products ($i =$ goods, housing, services) in four different regions ($r =$ A, B, C, D). For simplicity, it is assumed that within each region the expenditure share of each

product is the same.

The general price levels of the four regions can be calculated using some manner of multilateral measurement approach. A well-established approach is the CPD method introduced by Summers (1973). He emphasizes that his regression approach allows for statistical inference, which differentiates it from many other approaches to index number theory. However, this regression approach also has a significantly understated drawback. The CPD regression implicitly assumes that the products included have the same price dispersion. The prices in Tab. 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 the following, we demonstrate that, with product-specific price dispersion, the CPD regression produces biased estimates of the regional price levels (as formally shown in 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 (as formally shown in Appendix A.4.2).

Let p_i^r denote the price of product i in region r. The CPD regression assumes that each price can be explained by the linear relationship

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

where P^r is the price level of region r, π_i is the general value of product i, and $u_i^r \sim N\left(0, \sigma^2\right)$ is an error term (see Summers, 1973). To estimate the values of $\ln P^r$ and $\ln \pi_i$, the CPD model (1) is transformed into a regression equation with a set of dummy variables that represent the regions and the products. In the example related to Tab. 1, the CPD regression yields estimates of the logarithmic price levels, $\widehat{\ln P^r}$, of the four regions. Taking anti-logs gives the following regional price levels:

$$\widehat{P}^{\mathrm{A}} = 0.74\,, \;\; \widehat{P}^{\mathrm{B}} = 0.92\,, \;\; \widehat{P}^{\mathrm{C}} = 1.10\,, \;\; \widehat{P}^{\mathrm{D}} = 1.34\,. \tag{2}$$

The price levels are normalized such that $\widehat{P}^{\mathrm{A}} \cdot \widehat{P}^{\mathrm{B}} \cdot \widehat{P}^{\mathrm{C}} \cdot \widehat{P}^{\mathrm{D}} = 1$. As a consequence, the logarithmic prices of product i observed in regions A to D, $\ln p_i^r$, fluctuate around this product's estimated logarithmic general value, $\widehat{\ln \pi_i}$.

A graphical illustration of the CPD regression is provided in the upper left panel of

	A	B	C	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

Table 1: Prices of goods, housing, and services in four regions.

Fig. 1. It shows on the vertical axis the observed values of the dependent variable, $\ln p_i^r$, and on the horizontal axis the unknown regional logarithmic price levels, $\ln P^r$. The black diagonal indicates all points in which $\ln P^r = \ln p_i^r$. 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 determined by the CPD regression. More specifically, the intersection of each line with the horizontal axis is the estimated value $\widehat{\ln P^r}$. Thus, the four intersection points indicated in the upper left panel of Fig. 1 are the logarithms of the price levels listed in (2). To each product i, a solid straight line is depicted that runs parallel to the diagonal. The intersection of this solid line with the vertical axis is the estimated value of $\ln \pi_i$.

Changing the estimated value of $\ln \pi_i$ causes a parallel vertical shift of the solid line relating to product i. Changing the estimated value of $\ln P^r$ 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, $\widehat{u}_i^r$. 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) vertical distances between the observations and their respective solid lines is minimized. The upper

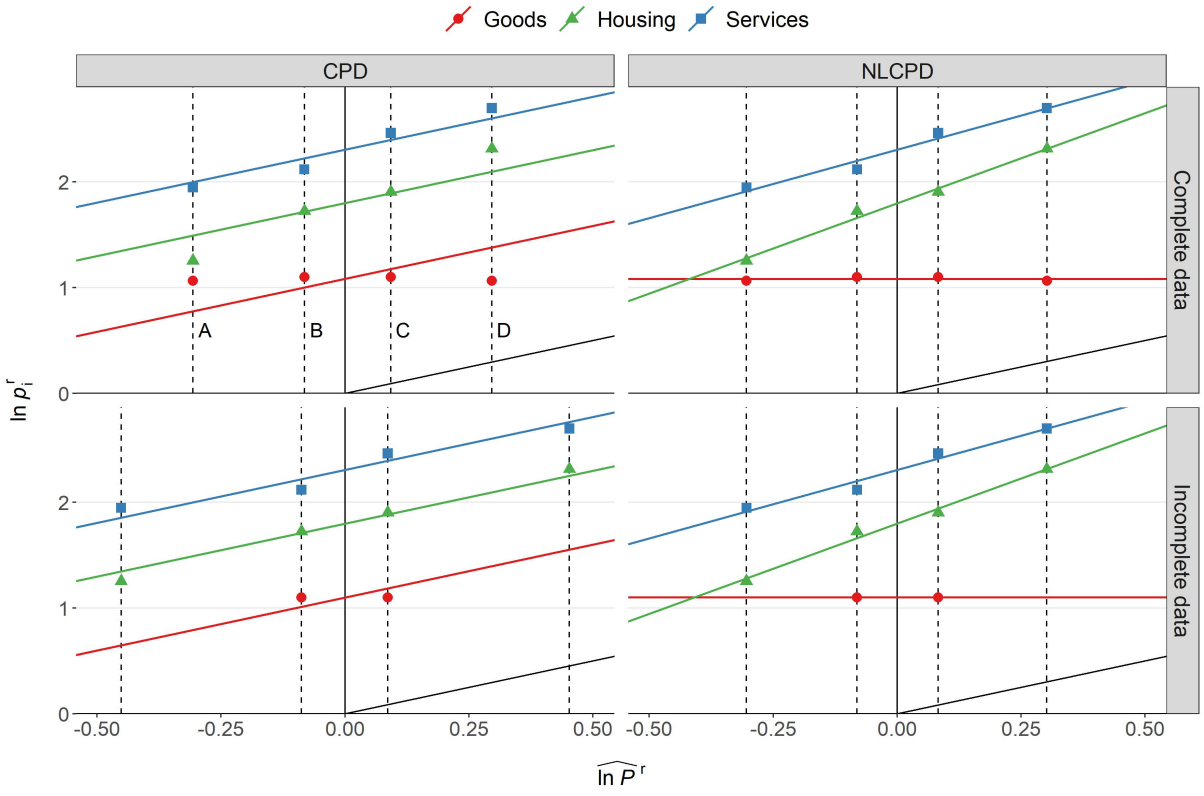

Figure 1: CPD and NLCPD regressions for the price data of Tab. 1, either with complete price data (top panels) or with missing prices for "goods" (bottom panels).

left panel of Fig. 1 depicts the solution to this minimization problem.

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 error term is zero: $\mathrm{E}(u_i^r|\mathbf{x}_i^r) = 0$. However, the left panels of Fig. 1 illustrate that missing prices usually lead to $\mathrm{E}(u_i^r|\mathbf{x}_i^r) \neq 0$. 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 depicted in the lower left panel of Fig. 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 Fig. 1 vanishes, that is, $\mathrm{E}(u_i^{\mathrm{A}}|\mathbf{x}_i^{\mathrm{A}}) < 0$. 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 Fig. 1). More generally, when a product with a low price dispersion 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 $\mathrm{E}(u_i^{\mathrm{D}}|\mathbf{x}_i^{\mathrm{D}}) > 0$. 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 Fig. 1). The corresponding price level estimates are

$$\widehat{P}^{\mathrm{A}} = 0.64\,,\ \ \widehat{P}^{\mathrm{B}} = 0.92\,,\ \ \widehat{P}^{\mathrm{C}} = 1.09\,,\ \ \widehat{P}^{\mathrm{D}} = 1.57\,.$$

Compared to the situation with complete price data, the price level of region A falls by 14% while the price level of region D increases 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 dispersion) instead of "goods" (the product with the lowest price dispersion), bias in the opposite direction would arise.

If no prices were missing, the CPD regression would be unbiased (upper left panel of Fig. 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 Fig. 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 $\widehat{u}_1^r$ (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 $\widehat{u}_3^r$ (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 dispersion 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 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. Recall, however, that this remedy requires either complete price data (as in the upper left panel of Fig. 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 nonlinear generalization of the CPD model (1) that provides for each product an estimate of its price dispersion. The following section explains the basic concept, while the formal exposition is deferred to Section 4 and Appendix A.

3 Solution

To begin with, we consider the case of complete data with product-specific price dispersion. In this case, the CPD regression is unbiased, but inefficient and inference is invalid. The three solid lines in the upper left panel of Fig. 1 have a slope of one, that is, they are parallel to the diagonal. 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:

$$\ln p_i^r = \ln \pi_i + \delta_i \ln P^r + u_i^r \; . \tag{3}$$

We denote this relationship as the NLCPD regression model.

The unknown values of the parameters δ_i determine the slopes of the solid lines. Products with $\delta_i > 1$ exhibit a stronger regional price dispersion than the average of all products (in Tab. 1, the product "housing"), while products with $\delta_i < 1$ exhibit a smaller price dispersion. Products with prices that are all but invariant with respect to the regional price levels have a slope parameter, δ_i, close to 0 (in Tab. 1, the product "goods").

CPD regressions assume that all included products exhibit a uniform price dispersion. This assumption is formalized by the restriction $\delta_i = 1$ (for all i). Thus, in the absence of any disturbances ($u_i^r = 0$ for all i and r), each price ratio p_i^r/p_i^s would coincide with

the ratio of the regional price levels P^r/P^s. 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, pp. 132-134) show that price dispersion is usually product-specific. The same is true for the illustrative price data listed in Tab. 1. Therefore, the CPD model (1) is misspecified.

The NLCPD model (3) accounts for product-specific price dispersion. Its specification is such that the price ratios depend not only on the price level $\ln P^r$ but also on δ_i and, therefore, are product-specific: $p_i^r/p_i^s = \delta_i(P^r/P^s)$ (for all i, r, and s). On average, the price ratios must reflect the ratios of the regional price levels P^r/P^s. In Section 4.2, it is shown that this intuitive condition leads to the following restriction: $\sum_{i=1}^{3} \delta_i/3 = 1$. Since the CPD model (1) implicitly assumes that all δ_i-values are equal to one, that model automatically satisfies this restriction. For the NLCPD model (3), it is a restriction that must be appropriately implemented in the estimation procedure. The estimation of the NLCPD model (3) uses exactly the same set of dummy variables as the estimation of the CPD model (1).

For the price data listed in Tab. 1, the fitting of the NLCPD regression lines to the data is depicted in the upper right panel of Fig. 1. The estimates of the slopes of the regression lines are $\widehat{\delta}_1 = 0.00$, $\widehat{\delta}_2 = 1.71$, and $\widehat{\delta}_3 = 1.29$. The estimated price levels are

$$\widehat{P}^{\mathrm{A}} = 0.74\,,\ \ \widehat{P}^{\mathrm{B}} = 0.92\,,\ \ \widehat{P}^{\mathrm{C}} = 1.09\,,\ \ \widehat{P}^{\mathrm{D}} = 1.35\,.$$

They are very similar to those obtained from the CPD regression when no prices are missing.

The lower right panel of Fig. 1 depicts the case where 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 price levels $\widehat{P}^{\mathrm{A}}$ to $\widehat{P}^{\mathrm{D}}$. 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 Fig. 1), the sum of squared residuals divided by the degrees of freedom falls from 0.055 (CPD regression) to 0.004 (NLCPD regression). Furthermore, in contrast to the CPD regression, the NLCPD method provides meaningful estimates of the standard errors of all estimated parameters (formally shown in Appendix A.4.1). Thus, the statistical significance of the coefficients $\widehat{\ln P^r}$, $\widehat{\ln \pi_i}$, and $\widehat{\delta}_i$ can be examined. When in the NLCPD regression at least one coefficient $\widehat{\delta}_i$ significantly deviates from one, the CPD model would have been misspecified (unless the data are complete or missing completely at random).

4 Method

The NLCPD model (3) is a generalization of the linear CPD model (1). The model function is nonlinear in its parameters. Consequently, parameter estimates must be derived by nonlinear regression. In Section 4.1, it is shown how the NLCPD model can be placed into a proper regression model. In Section 4.2, the NLCPD estimators are derived and compared to the formulas known for the CPD method. Since nonlinear regressions involve iterative search procedures, parameter start values are typically required. In Section 4.3, three strategies for the derivation of such start values are presented. Section 4.4 discusses the issue of weighting and provides formulas of the estimators' standard errors.

4.1 Regression model

Let $\mathcal{R} = \{r : r = 1, 2, \ldots, R\}$ denote the set of regions and $\mathcal{N} = \{i : i = 1, 2, \ldots, N\}$ the set of products included in the price comparison. To transform the CPD and NLCPD models in (1) and (3) into proper regression models, two sets of dummy variables are required. For each region $s \in \mathcal{R}$, a dummy variable D^s is defined such that $D^s = 1$ when $r = s$, and $D^s = 0$ otherwise. Similarly, for each product $j \in \mathcal{N}$, a dummy variable G_j is defined such that $G_j = 1$ when $i = j$, and $G_j = 0$ otherwise. With these dummy variables, the CPD model (1) can be written in the form

$$\ln p_i^r = \sum_{j \in \mathcal{N}} G_j \ln \pi_j + \sum_{s \in \mathcal{R}} D^s \ln P^s + u_i^r \tag{4}$$

and the NLCPD model (3) in the form

$$\ln p_i^r = \sum_{j \in \mathcal{N}} G_j \ln \pi_j + \sum_{j \in \mathcal{N}} G_j \delta_j \sum_{s \in \mathcal{R}} D^s \ln P^s + u_i^r \ . \tag{5}$$

Summers (1973) assumes that the variance of the error term is constant across products and regions: $u_i^r \sim N\left(0,\ \sigma^2\right)$. Thus, the CPD regression model (4) can be estimated using ordinary least squares. 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, pp. 745-746; Selvanathan and Rao, 1992, pp. 338-339; Diewert, 2005, pp. 562-563; Rao, 2005, pp. 574-575; Hajargasht and Rao, 2010, p. S39).

In both the CPD model (4) and the NLCPD model (5), perfect multicollinearity would arise. To avoid this problem, one of the π_j-values or $\ln P^s$-values can be set equal to 0. Alternatively, the normalization

$$\sum_{s \in \mathcal{R}} \ln P^s = 0 \tag{6}$$

can be applied and one of the $\ln P^s$-parameters derived as a residual from (6) instead of

being estimated. Any of the $\ln P^s$-parameters can be used for this purpose. If region $s = 1$ is chosen, the CPD model (4) becomes

$$\ln p_i^r = \sum_{j \in \mathcal{N}} G_j \ln \pi_j + \sum_{s \in \mathcal{R} \setminus \{1\}} \widetilde{D}^s \ln P^s + u_i^r \ , \tag{7}$$

where $\widetilde{D}^s = \left(D^s - D^1\right)$ and the parameter $\ln P^1$ is residually calculated using the expression $\ln P^1 = -\sum_{s \in \mathcal{R} \setminus \{1\}} \ln P^s$.

The NLCPD regression requires an additional condition. Since $\delta_j \ln P^s = (\delta_j \lambda) \ln P^s / \lambda$, the estimation of the parameters in model (5) requires a restriction on the δ_i-values. Otherwise, the regional price levels, $\ln P^s$, could be arbitrarily scaled up or down by the parameter λ. We recommend the restriction

$$\sum_{i \in \mathcal{N}} w_i \delta_i = 1 \ , \tag{8}$$

where w_i is the average expenditure share spent on product i. The justification for this restriction is deferred to Section 4.2. Note that the CPD model (4) satisfies restriction (8) by assumption ($\delta_i = 1$ for all $i \in \mathcal{N}$). By contrast, the NLCPD model (5) provides estimates for δ_i that have to satisfy restriction (8).

Restriction (8) implies that one of the δ_i-values must not be estimated but is to be derived as a residual. As in the CPD model (4), one of the $\ln P^r$-values must also be residually derived. Again, any product i and any region r can be chosen for these purposes. If product $i = 1$ and region $r = 1$ are selected, the NLCPD regression model (5) becomes

$$\ln p_i^r = \sum_{j \in \mathcal{N}} G_j \ln \pi_j + \left(\frac{G_1}{w_1} + \sum_{j \in \mathcal{N} \setminus \{1\}} \widetilde{G}_j \delta_j \right) \sum_{s \in \mathcal{R} \setminus \{1\}} \widetilde{D}^s \ln P^s + u_i^r \ , \tag{9}$$

where $\widetilde{D}^s = \left(D^s - D^1\right)$ and $\widetilde{G}_j = (G_j - (w_j / w_1) G_1)$. The parameters δ_1 and $\ln P^1$ are defined as $\delta_1 = \left(1 - \sum_{j \in \mathcal{N} \setminus \{1\}} w_j \delta_j\right) / w_1$ and $\ln P^1 = -\sum_{s \in \mathcal{R} \setminus \{1\}} \ln P^s$, respectively.

Note that the only difference between the NLCPD regression model (9) and the CPD regression model (7) is the factor in brackets. For observations of product $i = 1$, this factor simplifies to the above definition of δ_1, and for all other observations, the factor simplifies to the parameter δ_i.

4.2 Estimator

In the previous section, it was asserted that Eq. (8) is the appropriate restriction on the δ_i-values. In the following, we justify this assertion. To this end, we apply the NLCPD method to the bilateral case ($R = 2$). In such a context, the NLCPD estimators should turn into an attractive bilateral price index.

As illustrated in the upper right panel of Fig. 1, the NLCPD method fits straight lines through the observed logarithmic prices. In the bilateral case, the fit is perfect. All straight lines run through the observed logarithmic prices, that is, all residuals become 0:

$$\ln p_i^r - (\widehat{\delta}_i \widehat{\ln P^r} + \widehat{\ln \pi}_i) = 0 \tag{10a}$$

$$\ln p_i^s - (\widehat{\delta}_i \widehat{\ln P^s} + \widehat{\ln \pi}_i) = 0 \,, \tag{10b}$$

where r and s denote the two regions.

Using the normalization $\widehat{\ln P^s} = 0$, we get from Eq. (10b) the following simple estimator:

$$\widehat{\ln \pi}_i = \ln p_i^s \,. \tag{11}$$

Inserting this result into Eq. (10a), solving for $\widehat{\delta}_i$, multiplying the resulting expression by some weight w_i, and summing over all i, yields

$$\sum_{i \in \mathcal{N}} w_i \widehat{\delta}_i \widehat{\ln P^r} = \sum_{i \in \mathcal{N}} w_i \left(\ln \frac{p_i^r}{p_i^s} \right) \,. \tag{12}$$

The right hand side becomes the Törnqvist index when the weights w_i are defined as

$$w_i = \frac{1}{2} \left(w_i^r + w_i^s \right) \,, \tag{13}$$

where w_i^r and w_i^s denote the expenditure shares of product i in regions r and s, respectively (Diewert, 1995, pp. 11-12, Diewert, 2005, pp. 564-565).[1] The generalization of definition (13) to the case $R > 2$ is

$$w_i = \frac{1}{R} \sum_{r \in \mathcal{R}} w_i^r \,. \tag{14}$$

The Törnqvist index on the right hand side of Eq. (12) expresses the logarithmic price level of region r relative to the logarithmic price level of region s, that is, to $\ln P^s = 0$. Thus, the left hand side of Eq. (12) must simplify to $\widehat{\ln P^r}$. This requires that $\sum_{i \in \mathcal{N}} w_i \widehat{\delta}_i = 1$, which is restriction (8). Therefore, a case has been made for restriction (8), where the weights w_i represent average expenditure shares as defined in Eq. (14).

Substituting in Eq. (10a) the variables $\widehat{\ln P^r}$ and $\widehat{\ln \pi}_i$ with the terms $\sum_{i \in \mathcal{N}} w_i \ln(p_i^r / p_i^s)$ and $\ln p_i^s$, respectively, and solving the resulting expression for $\widehat{\delta}_i$ yields the intuitive estimator

$$\widehat{\delta}_i = \frac{\ln(p_i^r / p_i^s)}{\sum_{j \in \mathcal{N}} w_j \ln(p_j^r / p_j^s)} \,. \tag{15}$$

---

[1] If in Eq. (13), instead of an arithmetic mean of the expenditure shares w_i^r and w_i^s, a geometric mean were used, the right hand side of Eq. (12) would become the Walsh-Vartia index. With a logarithmic mean, the Sato-Vartia index would result.

The estimators (11), (12), and (15) apply to bilateral regional price comparisons for normalization $\widehat{\ln P^s} = 0$. In the following, we consider the multilateral case. In such 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 price levels.

In the following, we derive the NLCPD method's weighted least squares estimators $\widehat{\ln \pi_i}$, $\widehat{\delta}_i$, and $\widehat{\ln P^r}$ as well as the CPD method's estimators, $\widehat{\ln \pi'_i}$ and $\widehat{\ln P^{r\prime}}$, as special cases. The residuals $\hat{u}_i^r$ of the NLCPD regression model (3) are defined by $\hat{u}_i^r = \ln p_i^r - \widehat{\delta}_i \widehat{\ln P^r} - \widehat{\ln \pi_i}$. Accordingly, the weighted sum of squared residuals, $S_{\hat{u}_i^r \hat{u}_i^r}$, can be written as

$$S_{\hat{u}_i^r \hat{u}_i^r} = \sum_{r \in \mathcal{R}} \sum_{i \in \mathcal{N}_r} w_i \left(\ln p_i^r - \widehat{\delta}_i \widehat{\ln P^r} - \widehat{\ln \pi_i} \right)^2 = \sum_{i \in \mathcal{N}} \sum_{r \in \mathcal{R}_i} w_i \left(\ln p_i^r - \widehat{\delta}_i \widehat{\ln P^r} - \widehat{\ln \pi_i} \right)^2 \ , \quad (16)$$

where $\mathcal{N}_r$ defines 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 R_i. A discussion of the choice of weights, w_i, is deferred to Section 4.4.

The formulas of $\widehat{\ln \pi_i}$, $\widehat{\delta}_i$, and $\widehat{\ln P^r}$ can be derived by minimizing $S_{\hat{u}_i^r \hat{u}_i^r}$. In this nonlinear least squares approach, we apply normalization (6) as well as restriction (8). As a consequence, one $\widehat{\delta}_i$-value and one $\widehat{\ln P^r}$-value cannot be used in the minimization. They are residually derived. The first-order conditions are

$$\frac{\partial S_{\hat{u}_i^r \hat{u}_i^r}}{\partial \widehat{\ln \pi_i}} = \sum_{r \in \mathcal{R}_i} w_i 2 \left(\ln p_i^r - \widehat{\delta}_i \widehat{\ln P^r} - \widehat{\ln \pi_i} \right) (-1) = 0 \quad (17a)$$

$$\frac{\partial S_{\hat{u}_i^r \hat{u}_i^r}}{\partial \widehat{\delta}_i} = \sum_{r \in \mathcal{R}_i} w_i 2 \left(\ln p_i^r - \widehat{\delta}_i \widehat{\ln P^r} - \widehat{\ln \pi_i} \right) \left(-\widehat{\ln P^r} \right) = 0 \quad (17b)$$

$$\frac{\partial S_{\hat{u}_i^r \hat{u}_i^r}}{\partial \widehat{\ln P^r}} = \sum_{i \in \mathcal{N}_r} w_i 2 \left(\ln p_i^r - \widehat{\delta}_i \widehat{\ln P^r} - \widehat{\ln \pi_i} \right) \left(-\widehat{\delta}_i \right) = 0 \ . \quad (17c)$$

Condition (17a) gives

$$\widehat{\ln \pi_i} = \frac{1}{R_i} \sum_{r \in \mathcal{R}_i} \left(\ln p_i^r - \widehat{\delta}_i \widehat{\ln P^r} \right) \ . \quad (18)$$

As the product weights are identical across regions, each region receives the same weight in the summation. Inserting the restriction $\widehat{\delta}_i = 1$ $(i \in \mathcal{N})$ into the NLCPD estimator (18), gives the corresponding CPD estimator:

$$\widehat{\ln \pi'_i} = \frac{1}{R_i} \sum_{r \in \mathcal{R}_i} \left(\ln p_i^r - \widehat{\ln P^{r\prime}} \right) \ , \quad (19)$$

where $\widehat{\ln P^{r\prime}}$ is the CPD estimator of the regional price levels as defined in Eq. (22), below.

For a product i that is priced in all regions ($R_i = R$), the NLCPD estimator (18) simplifies to

$$\widehat{\ln \pi_i} = \frac{1}{R}\sum_{r\in\mathcal{R}} \ln p_i^r - \widehat{\delta}_i \frac{1}{R} \overbrace{\sum_{r\in\mathcal{R}} \widehat{\ln P^r}}^{=0} = \frac{1}{R}\sum_{r\in\mathcal{R}} \ln p_i^r \,,$$

which coincides with the CPD estimator for such a product (e.g., Diewert, 2004, p. 7).

Condition (17b) yields

$$\widehat{\delta}_i = \frac{\sum\limits_{r\in\mathcal{R}_i} \widehat{\ln P^r}\left(\ln p_i^r - \widehat{\ln \pi_i}\right)}{\sum\limits_{r\in\mathcal{R}_i} \left(\widehat{\ln P^r}\right)^2} \,. \tag{20}$$

The numerator is the covariation (across regions) of the logarithmic regional price levels, $\widehat{\ln P^r}$, and $\left(\ln p_i^r - \widehat{\ln \pi_i}\right)$. The denominator is the variation (across regions) of the logarithmic regional price levels. Therefore, the estimator (20) can be viewed as the ordinary least square estimator of the slope parameter of a simple linear model where $\left(\ln p_i^r - \widehat{\ln \pi_i}\right)$ is regressed on $\widehat{\ln P^r}$. The covariation represented by the numerator is usually positive. The larger this covariation, the larger the estimated price dispersion, $\widehat{\delta}_i$. If product i has a uniform price, then $\widehat{\ln \pi_i} = \ln p_i^r$ and, therefore, the estimator (20) gives $\widehat{\delta}_i = 0$.

Condition (17c) can be rewritten as

$$\widehat{\ln P^r} = \frac{\sum\limits_{i\in\mathcal{N}_r} w_i \widehat{\delta}_i \left(\ln p_i^r - \widehat{\ln \pi_i}\right)}{\sum\limits_{i\in\mathcal{N}_r} w_i \left(\widehat{\delta}_i\right)^2} \,. \tag{21}$$

The numerator is the covariation (across products) of $\left(\ln p_i^r - \widehat{\ln \pi_i}\right)$ and the spread parameter $\widehat{\delta}_i$. 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 price dispersion, $\widehat{\delta}_i$. In expensive regions $\left(\widehat{\ln P^r} > 0\right)$, prices above the general level dominate.

Setting $\widehat{\delta}_i = 1$ for all products $i \in \mathcal{N}_r$, the estimator (21) simplifies to the corresponding CPD estimator:

$$\widehat{\ln P^{r\prime}} = \frac{\sum\limits_{i\in\mathcal{N}_r} w_i \left(\ln p_i^r - \widehat{\ln \pi_i'}\right)}{\sum\limits_{i\in\mathcal{N}_r} w_i} \,. \tag{22}$$

When all products i are priced in region r, we get $\sum_{i\in\mathcal{N}_r} w_i = 1$, and the resulting estimator

(22) simplifies to the well-known CPD formula (e.g., Rao, 2005, p. 577; Rao and Hajargasht, 2016, p. 417):

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

For the derivation of the nonlinear least squares formulas (18), (20), and (21), the actual definition of the weights, w_i, was irrelevant. Note, however, that the weights are assumed to be uniform across regions (w_i instead of w_i^r) 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 weights that vary across regions. These estimators are derived in Appendix A.1. To express restriction (8) in terms of region-specific weights, w_i^r, one merely has to substitute w_i with the right hand side of Eq. (14).

The nonlinear least squares formulas (18), (20), and (21) do not provide explicit solutions for the parameters $\widehat{\ln \pi_i}$, $\widehat{\ln P^r}$, and $\widehat{\delta}_i$. Instead, an iterative optimization routine is necessary.[2] Such routines require appropriate start values for the model parameters.

4.3 Parameter start values

The choice of appropriate start values is important for two reasons. First, it is more likely that the optimization algorithm successfully converges in the allowed number of iterations. Second, singularities can prevent any optimization if initial parameter start values are not set adequately. Strategies for deriving start values are usually data and model-driven (e.g. Gallant, 1975, p. 76). In the following, we provide three simple strategies for the derivation of parameter start values in the NLCPD regression.

In strategy S1, parameter start values are derived from the calculation of simple price averages across products and regions. Defining the weighted logarithmic average price in region r as $\ln \bar{p}^r = \sum_{i \in \mathcal{N}_r} w_i \ln p_i^r$, the start values $\overline{\ln P^r}$ and $\overline{\ln \pi_i}$ can be computed from

$$\overline{\ln P^r} = \ln \bar{p}^r - \frac{1}{R_i} \sum_{s \in \mathcal{R}_i} \ln \bar{p}^s \qquad \text{and} \qquad \overline{\ln \pi_i} = \frac{1}{R_i} \sum_{r \in \mathcal{R}_i} \ln p_i^r \ .$$

The start values for δ_i are set equal to one for all $i \in N$. This assumption satisfies restriction (8) and is also the assumption underlying the CPD regression model. The calculations are easy to implement and computationally efficient.

In the event of incomplete price data, however, start values for $\ln P^r$ and $\ln \pi_i$ derived

[2] Common methods are Gauss-Newton, Levenberg-Marquardt, (L-)BFGS, Nelder-Mead, and gradient descent. A comprehensive overview can be found in Kelley (1999). Our implementation of the NLCPD method relies on a modification of the Levenberg-Marquardt algorithm (see Elzhov *et al.*, 2016; Moré, 1978).

by strategy S1 might be a poor guess. Using the CPD method's estimates of $\ln P^r$ and $\ln \pi_i$, is a more appealing approach, irrespective of any data gaps. This is strategy S2. Again, the start values for δ_i are set equal to one. When the price data are complete, this strategy provides the same set of start values as strategy S1.

If it is known that some δ_i-values deviate from one (e.g., for products with uniform prices across regions), setting $\delta_i = 1$ is inappropriate. Therefore, strategy S3 is identical to strategy S2, but computes the start values of δ_i from Eq. (20), where the CPD estimates of $\ln P^r$ and $\ln \pi_i$ provide the values of $\widehat{\ln P^r}$ and $\widehat{\ln \pi_i}$, respectively. The resulting $\widehat{\delta}_i$-values do not necessarily satisfy restriction (8). Therefore, to obtain the proper start values, they are divided by $\sum_{i \in \mathcal{N}} w_i \widehat{\delta}_i$.

When the price data are complete, the choice between the three strategies hardly matters. Strategy S3 takes exactly one iteration less than the other two strategies because start values for $\widehat{\delta}_i$ are directly derived from the first-order condition. With incomplete price data, the start values of the three strategies differ. Our simulations indicate that strategy S3 outperforms strategies S2 and S1. The number of iterations until convergence is slightly smaller, the percentage of successful completions is marginally higher, and the sum of squared residuals achieved at convergence is slightly lower.

4.4 Weighting and standard errors

The weights in Eq. (16) are the products' average expenditure shares, w_i. 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, pp. 562-563; Rao, 2005, p. 575) and/or it can reflect the 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 error term u_i^r. Thus, the economic and the econometric motivation for weighting may lead to different sets of weights. This complicates the CPD and NLCPD estimators of the regional price levels.

In the NLCPD model (9), the error 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 error term, u_i^r, is given by $w_i \sigma^2$ 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 error 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 (see Appendix A.4.1). In nonlinear regression models, approximations of these standard errors can be computed from the Jacobian matrix evaluated at final parameter estimates. When the data set is complete and the weighted error term $\sqrt{w_i}u_i^r$ is homoskedastic, the approximated standard error of the NLCPD estimator $\widehat{\ln \pi_i}$ is

$$\widehat{se}\left(\widehat{\ln \pi_i}\right) = \widehat{\sigma}\sqrt{\frac{1}{Rw_i}}\,, \tag{23}$$

with

$$\widehat{\sigma} = \sqrt{\frac{S_{\hat{u}_j^r \hat{u}_j^r}}{NR-R-2N+2}}\,.$$

To obtain the corresponding estimator of the CPD method, $\widehat{se}'\left(\widehat{\ln \pi_i'}\right)$, the estimator $\widehat{\sigma}$ must be replaced with the estimator $\widehat{\sigma}' = \sqrt{S_{\hat{u}_j^{r\prime} \hat{u}_j^{r\prime}}/(NR-R-N+1)}$, with $\hat{u}_j^{r\prime}$ denoting the residuals of the CPD regression. In Appendix A.4.2 it is shown that the estimator $\widehat{\sigma}'$ and, therefore, the estimator $\widehat{se}'\left(\widehat{\ln \pi_j'}\right)$ are biased.

The approximated standard error of the estimator of δ_i is

$$\widehat{se}\left(\widehat{\delta}_i\right) = \widehat{\sigma}\sqrt{\frac{1}{\sum\limits_{r\in\mathcal{R}}\left(\widehat{\ln P^r}\right)^2}\left(\frac{1-w_i}{w_i} + \left(\widehat{\delta}_i - 1\right)^2\right)}\,. \tag{24}$$

This standard error falls as the product weight, w_i, increases, the fluctuation of the estimated logarithmic price levels, $\widehat{\ln P^r}$, increases, and the $\widehat{\delta}_i$-value approaches one.

For the NLCPD estimator of the regional price levels, $\widehat{\ln P^r}$, the following standard error is derived:

$$\widehat{se}\left(\widehat{\ln P^r}\right) = \widehat{\sigma}\sqrt{\frac{1}{\sum\limits_{i\in\mathcal{N}} w_i\left(\widehat{\delta}_i\right)^2}\left(\frac{R-1}{R} + \left(\sum_{i\in\mathcal{N}} w_i\left(\widehat{\delta}_i\right)^2 - 1\right)\frac{\left(\widehat{\ln P^r}\right)^2}{\sum\limits_{s\in\mathcal{R}}\left(\widehat{\ln P^s}\right)^2}\right)}\,. \tag{25}$$

Restriction (8) and Jensen's (1906) inequality yield

$$\sum_{i\in\mathcal{N}} w_i\left(\widehat{\delta}_i\right)^2 \geq \left(\sum_{i\in\mathcal{N}} w_i\widehat{\delta}_i\right)^2 = 1^2 = 1\,.$$

Thus, the root term in Eq. (25) is always positive. In addition, one can show that it is smaller or equal to $\sqrt{(R-1)/R}$. The root term increases with the number of regions, R, and the estimated logarithmic price level, $\widehat{\ln P^r}$. If for all products $i \in \mathcal{N}$ the estimated

price dispersion were $\widehat{\delta}_i = 1$, the formula would simplify to $\widehat{se}\left(\widehat{\ln P^r}\right) = \widehat{\sigma}\sqrt{(R-1)/R}$. Note that the CPD formula, $\widehat{se}'\left(\widehat{\ln P^{r\prime}}\right) = \widehat{\sigma}'\sqrt{(R-1)/R}$, would be biased because $\widehat{\sigma}'$ is biased.

By default, most statistical software would use formulas (23) to (25) to compute standard errors in a weighted NLCPD regression. In other words, such software would implicitly follow the position of Clements and Izan (1987), who argue that the weight w_i correctly addresses product i's economic importance and that the weighted error term $\sqrt{w}_i u_i^r$ is homoskedastic because the weight w_i is negatively related to the variance of the error 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 error term, u_i^r, is homoskedastic and the weighted error term, $\sqrt{w}_i u_i^r$, is heteroskedastic. Rao (2004, pp. 17-18) and Hajargasht and Rao (2010, pp. 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.

Fig. 1 revealed that product-specific price dispersion results in a heteroskedastic error term, u_i^r. With this type of price dispersion, the NLCPD regression model is preferable. If its unweighted error term, u_i^r, is homoskedastic, the weighted error 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, pp. S45). If both, the unweighted error term, u_i^r, and the weighted error term, $\sqrt{w}_i u_i^r$, are heteroskedastic, an even more general formula is required (see Appendix A.4.1).

5 Simulation

Imposing the restriction $\delta_i = 1$ for all products i in the NLCPD model (3) yields the CPD model (1). However, the restriction is quite unrealistic as regional price level dispersions can be expected to vary across basic headings and sometimes even within basic headings.[3] 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. The simulation setting is described in Section 5.1, while the results are provided in Section 5.2.

[3] *Basic heading* is the official terminology for groups of products with an identical consumption purpose.

5.1 Setting

In the simulation, we consider $N = 15$ products or basic headings available in $R = 20$ regions. The data generating process (DGP) in Eq. (3) assumes that each region r has a true but unknown price level $\ln P^r$. Similarly, true values of the parameters $\ln \pi_i$ and δ_i exist for each product i. Four different scenarios are considered. They differ with respect to the number and structure of missing observations and to the variance of the δ_i-values.

The true regional price levels, $\ln P^r$, are generated in two steps. First, preliminary price levels, $\ln \widetilde{P}^r$, are independently sampled from a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 0.1$, that is, $\ln \widetilde{P}^r \sim N\,(\mu = 0,\ \sigma = 0.1)$. Second, the price levels are normalized. Subtracting the average preliminary price level of all regions from $\ln \widetilde{P}^r$ yields the true regional price levels, $\ln P^r$:

$$\ln P^r = \ln \widetilde{P}^r - \frac{1}{R} \sum_{s \in \mathcal{R}} \ln \widetilde{P}^s \ .$$

By definition, their mean is zero. The resulting average price level spread between the most expensive region and the cheapest region is almost 50%.

In the simulation, the weights w_i represent the products' expenditure shares. For each product i, preliminary weights, $\widetilde{w}_i$, are sampled from a uniform distribution with $\widetilde{w}_i \sim U\,(\min = 1,\ \max = 100)$. The normalized weights are $w_i = \widetilde{w}_i / \sum_{j \in \mathcal{N}} \widetilde{w}_j$.

The products' general values, $\ln \pi_i$, are drawn from a log-normal distribution with $\ln \pi_i \sim LN\,(\mu = 0,\ \sigma = 0.5)$. The log-normal distribution ensures that product prices are greater than zero while its positive skewness makes very expensive products occur less frequently.

It is only in the first scenario that the regional price dispersion of the products conforms with the CPD assumption: $\delta_i = 1$ for all N products. In the other three scenarios, the δ_i-values are product-specific but equal to one on average. The preliminary values of $\widetilde{\delta}_i$ are sampled from a normal distribution with $\widetilde{\delta}_i \sim N\left(\mu = 1,\ \sigma = \sqrt{0.5}\right)$. The normalized values are $\delta_i = \widetilde{\delta}_i / (\sum_{j \in \mathcal{N}} w_j \widetilde{\delta}_j)$. Thus, $\sum_{i \in \mathcal{N}} w_i \delta_i = 1$.

The error term u_i^r is sampled from a normal distribution with a product-specific standard deviation: $u_i^r \sim N\left(\mu = 0,\ \sigma_i = \sigma / \sqrt{w_i}\ \right)$ with $\sigma = 1/100$ being the "global" standard deviation of the error term. This setting ensures that weighted variants (or weighted least squares) of the CPD and NLCPD methods are the appropriate choice of estimation.

The first of the four scenarios represents the most artificial scenario, 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.

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 set to 1. Note that this yields 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 the artificial price data by inserting the sampled values of $\ln P^r$, $\ln \pi_i$, δ_i, w_i, and u_i^r into the DGP defined in Eq. (3). 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 regional price dispersion. Third, we apply both the (weighted) CPD method and the (weighted) NLCPD method to the price data generated during the first step. For the starting values of the NLCPD method, we apply strategy S3. That is, 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 (20).

We repeat these three steps $L = 2{,}000$ times (with iterations $l = 1, 2, \ldots, L$) and obtain for each region r a set of 2,000 $\widehat{\ln P^{r\prime}}$-values for the CPD method and 2,000 $\widehat{\ln P^r}$-values for the NLCPD method. Afterwards, we compare the performance of the two 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, $|\text{Bias}\left(\widehat{\ln P^r}\right)|$, and also the root mean squared error, $\text{RMSE}\left(\widehat{\ln P^r}\right)$. Then, we take the average of these numbers across all regions:

$$\text{Bias}\left(\widehat{\ln P}\right) = \frac{1}{R}\sum_{r\in\mathcal{R}}\left|\text{Bias}\left(\widehat{\ln P^r}\right)\right| = \frac{1}{R}\sum_{r\in\mathcal{R}}\left|\frac{1}{L}\sum_{l=1}^{L}\left(\widehat{\ln P_l^r} - \ln P_l^r\right)\right| \tag{26a}$$

$$\text{RMSE}\left(\widehat{\ln P}\right) = \frac{1}{R}\sum_{r\in\mathcal{R}}\text{RMSE}\left(\widehat{\ln P^r}\right) = \frac{1}{R}\sum_{r\in\mathcal{R}}\sqrt{\frac{1}{L}\sum_{l=1}^{L}\left(\widehat{\ln P_l^r} - \ln P_l^r\right)^2}\,, \tag{26b}$$

where $\widehat{\ln P_l^r}$ denotes the *estimated* parameter of region r's price level obtained in iteration l by the NLCPD method, while $\ln P_l^r$ is the corresponding *true* parameter. For the CPD method, $\text{Bias}\left(\widehat{\ln P'}\right)$ and $\text{RMSE}\left(\widehat{\ln P'}\right)$ are derived in the same way.

For Scenarios 1 to 3, we expect both 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). Although the degrees of freedom in the NLCPD 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.

5.2 Discussion of results

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

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 to 4, the computed mean RMSE of $\ln P^r$-estimates is lower for the NLCPD method than for the CPD method (see bottom line of Tab. 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. 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 Fig. 2. Its structure is similar to Tab. 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 Fig. 2 reveals that in all regions both the CPD and the NLCPD method

	Scenario 1		**Scenario 2**		**Scenario 3**		**Scenario 4**	
	CPD	NLCPD	CPD	NLCPD	CPD	NLCPD	CPD	NLCPD
Bias	0.0002	0.0002	0.0002	0.0001	0.0003	0.0002	0.0133	0.0002
RMSE	0.0097	0.0097	0.0097	0.0081	0.0201	0.0110	0.0250	0.0105

Table 2: Mean absolute bias and mean RMSE of the NLCPD estimates, $\widehat{\ln P^r}$, and the CPD estimates, $\widehat{\ln P^{r\prime}}$.

[4] In Appendix B, mean absolute bias and mean RMSE are also reported for the estimates of $\ln \pi_i$ and δ_i, respectively.

are unbiased as long as the data are complete or missing completely at random (Scenarios 1 to 3), but that the CPD method is biased when the data gaps are systematic (see the red 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. Recall that in Scenario 4, the number of data gaps is negatively correlated with the product's true regional price dispersion, δ_i. Switching to a positive correlation, one would observe the 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. Also in Scenario 4, the blue dots remain close to the horizontal baseline.

The NLCPD method outperforms the CPD method with respect to the RMSE, too. This is shown in the bottom row of Fig. 2. The blue dots are closer to the base line. 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 become more pronounced. The u-shape of the red dots implies that the largest inaccuracies arise for the cheapest and the most expensive regions.

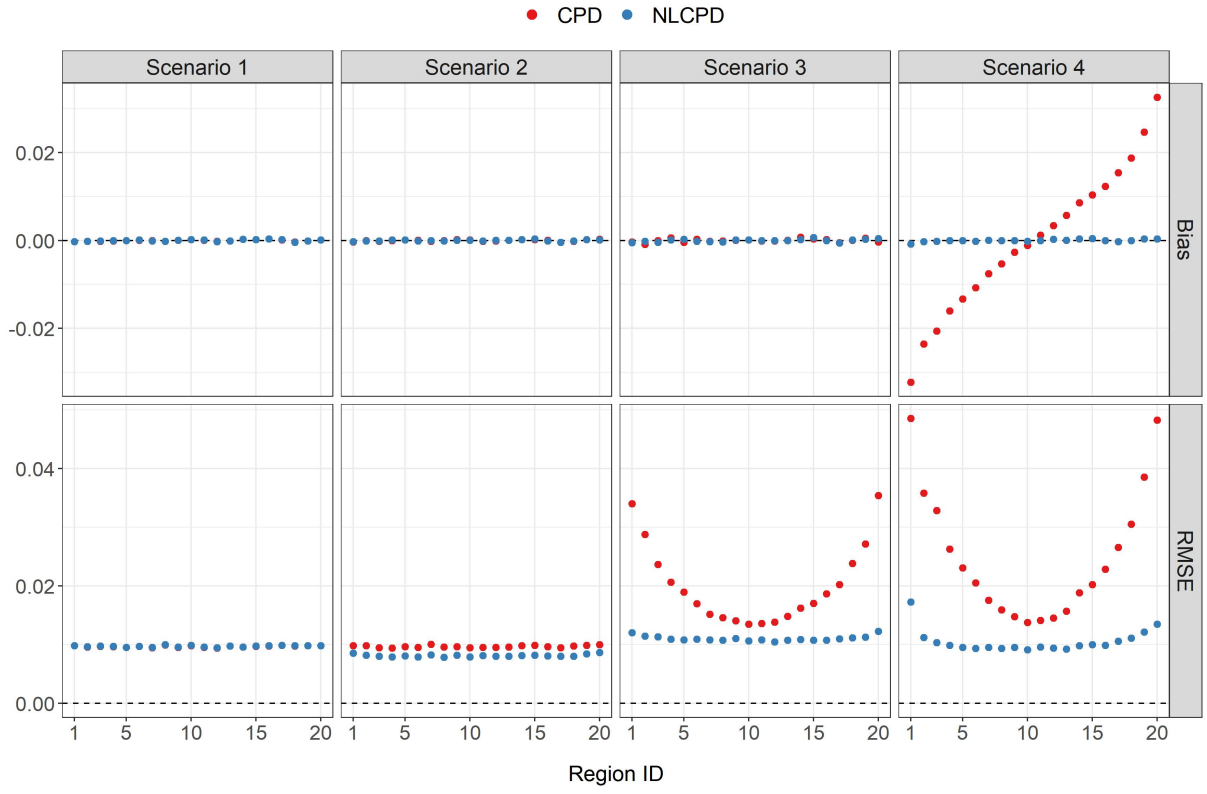

Figure 2: Bias and RMSE of the NLCPD estimates, $\widehat{\ln P^r}$, and the CPD estimates, $\widehat{\ln P^{r\prime}}$, for the three simulation scenarios.

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 the degree of price dispersion can be expected to vary between basic headings (e.g., rents versus manufactured goods), while the CPD method assumes a uniform degree of price dispersion. Therefore, we also compare the results of the NLCPD method to those we would obtain from the CPD method. The estimated price levels are transformed into a regional price index for Germany.[5]

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 that 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 Tab. 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.[6] They represent a combined expenditure weight of 13.44%. 36 other basic headings with a combined weight of 1.45% were too fragmentary to convey useful information for the interregional price comparison.[7] This leaves us with 677 basic headings for which the price information can be included in the regional price comparison. As can be seen from Tab. 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 by the overall price index.

For each of the remaining 677 basic headings, we assume that the price dispersion of the items within a basic heading is identical. Thus, the set of regional price levels of a given

[5] The price index numbers for the German districts are available upon request.

[6] 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.

[7] For example, the priced items of the basic heading "gloves" were not identical and, therefore, not comparable.

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). However, it is worth mentioning some of the improvements on and modifications to the data preparation and aggregation in Weinand and Auer (2020) that we have implemented.

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, pp. 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.

ID	**Division**	**#BH**	**Expenditure weight**	
			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 maint.	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

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%. Source: Research Data Center of the Federal Statistical Office and Statistical Offices of the Länder, CPI, May 2019; authors' own computations.

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, which was collected by the German Market Transparency Unit for Fuels in May 2019.[8]

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 data base for the NLCPD as well as the CPD estimation. Both 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 provide estimates of the overall logarithmic price levels of the 401 German regions, $\widehat{\ln P^r}$ and $\widehat{\ln P^{r\prime}}$ $(r = 1, \ldots, 401)$, and of the basic headings' general values, $\widehat{\ln \pi_i}$ and $\widehat{\ln \pi_i'}$ $(i = 1, \ldots, 618)$, respectively. Only the NLCPD method additionally provides estimates of the basic headings' price dispersion, $\widehat{\delta}_i$ $(i = 1, \ldots, 618)$.

Except for very few outliers, the NLCPD method's estimates $\widehat{\delta}_i$ appear highly plausible. For the two basic headings with constant regional price levels, the NLCPD method yields an estimated price dispersion of $\widehat{\delta}_i = 0$. For rents (existing tenancies) and for new lease rents we get $\widehat{\delta}_i = 3.23$ and $\widehat{\delta}_i = 4.82$, respectively. On average, the $\widehat{\delta}_i$-values of goods are the smallest ones. The $\widehat{\delta}_i$-values of rents are among the largest ones, while most of the $\widehat{\delta}_i$-values of services take a middle position. The results clearly confirm that the regional price dispersion varies between the basic headings. 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).[9] They are depicted in Fig. 3, where the blue dots represent the seven cities with the highest number of inhabitants in Germany. The estimated logarithmic price levels obtained from the NLCPD method range between -0.09 and 0.22, while those of the CPD method 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 price dispersion results in an upward biased estimate of the spread of the estimated regional price levels. In the present case, the Spearman correlation

[8] The data were downloaded from `https://creativecommons.tankerkoenig.de/` where historical fuel prices are provided on a daily basis.

[9] 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.

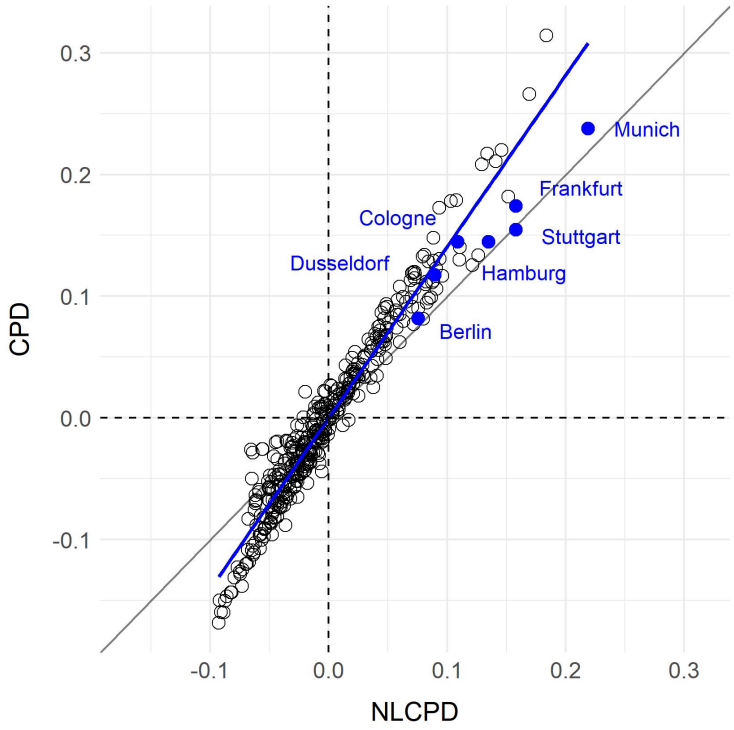

Figure 3: Estimates of the regional logarithmic price levels, $\ln P^r$, by NLCPD method (horizontal axis) and CPD method (vertical axis). Ordinary least squares regression as solid blue line.

of the number of data gaps and the NLCPD's estimates $\widehat{\delta}_i$ is -0.13. Consequently, the CPD method produces biased price level estimates.

In order to transform the logarithmic price level estimates $\widehat{\ln P^r}$ and $\widehat{\ln P^{r\prime}}$ into a regional price index, they are expressed in relation to their respective population-weighted averages. For the NLCPD method, the transformation is

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

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

As can be seen from Tab. 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

	Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**	**Sd.**
CPD	82.2	92.5	95.9	97.5	101.2	133.1	7.6
NLCPD	89.2	94.3	96.9	98.0	100.4	121.8	5.2

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

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%. For both methods, the unweighted mean is below the population-weighted mean, indicating that a region's price level tends to increase with its population.

The spatial pattern of the price index numbers of the 401 German regions is depicted in Fig. 4. 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, the same ranking of the seven cities arises and Starnberg, a region neighboring Munich, is the most expensive region in Germany.

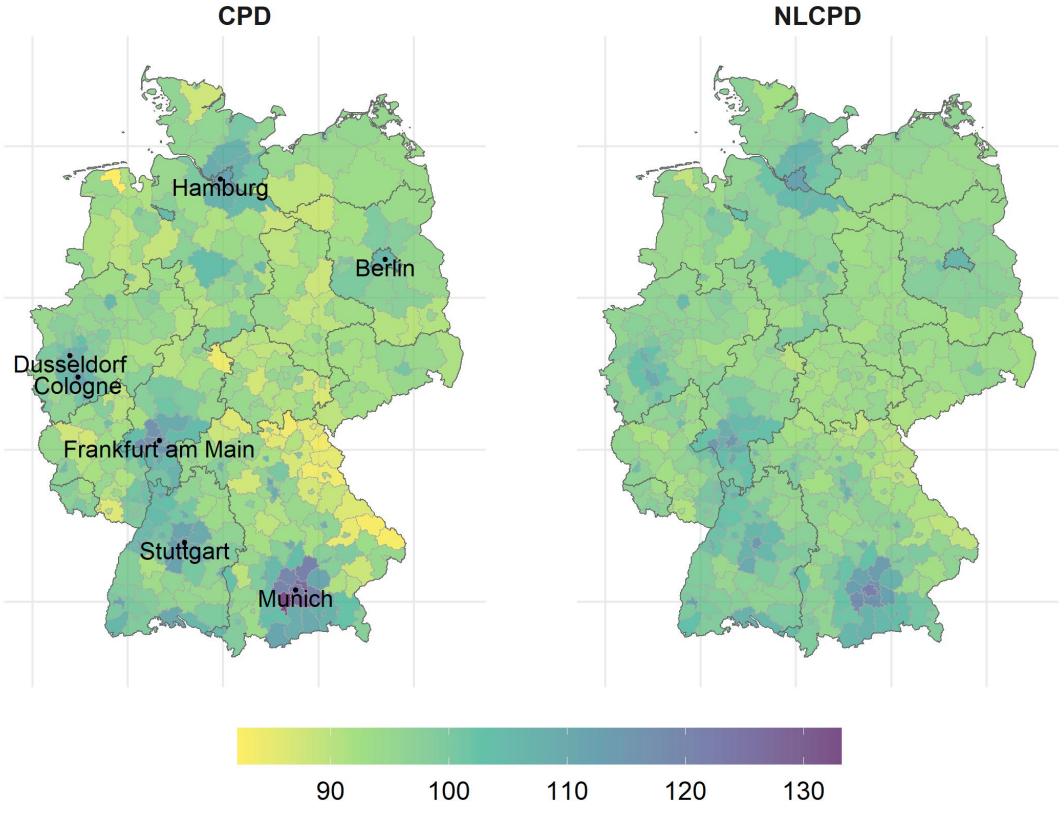

Figure 4: Price index numbers by CPD and NLCPD methods, each in relation to its population-weighted average (= 100).

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 has shown that it is necessary for the CPD method for the regional price dispersion of the various products to be uniform. If it is not, the estimates of the standard errors are biased. Even worse, when the data gaps are not completely at random, the estimates of the regional price levels are systematically biased.

As a solution, this paper introduced the NLCPD method, a nonlinear generalization of the CPD method. The NLCPD method expands on the CPD method to include parameters that capture the product-specific price dispersion. Their estimates indicate whether the CPD assumption of uniform price dispersion would have been reasonable. In a simulation, the deficiencies of the CPD method and the superiority of the NLCPD method were shown. Finally, in a price level comparison of the 401 German regions, the practical applicability of the NLCPD method was demonstrated.

The only drawback of the NLCPD method as compared to the CPD method is its nonlinear specification. As a consequence, iterative estimation procedures are required. When the variation in the regional price levels is small and a product has only very few observations, the iterative estimation of its price dispersion 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 price dispersion, one can impose the restriction that the product's price dispersion is unity and, thus, coincides with the dispersion of the overall regional price levels. Recall that the CPD method imposes this restriction on *all* products. Such a restricted NLCPD regression would still outperform the CPD regression.

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, pp. 35-37) show that the weighted CPD method and the GEKS-Törnqvist index coincide when the weights in the CPD method are the products' expenditure shares and these shares are uniform across regions. Even when data gaps are present, there is a close relationship between the two methods (Weinand, 2022). Consequently, one could argue that any issue of one approach is likely to also apply to the other one. This is a relevant concern because it is not only the CPD method but also the GEKS approach that is used in the International Comparison Program (World Bank, 2020, p. 82) and in various national studies (surveyed in Majumder and Ray, 2020, pp. 105-109; Weinand and Auer, 2020, pp. 416-418). However, a careful analysis of this question has to be left for future research.

The CPD method's drawbacks in the presence of product-specific price dispersion 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 (4) 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 intertemporal dispersion of prices over time is uniform across products. Consequently, the TPD method has the same statistical issues as the CPD method. Ongoing research is examining whether the nonlinear TPD (NLTPD) method can solve these problems.

A Mathematical derivations

In the following, we provide the mathematical derivations underlying the paper. In particular, this includes results on bias and inference for the CPD method as well as the formulas of the NLCPD method's standard errors.

A.1 NLCPD estimators when weights are region-specific

Suppose that we have region-specific weights w_i^r with $\sum_{i\in\mathcal{N}_r} w_i^r = 1$ for each region r. Then, the minimization problem is

$$\min_{\widehat{\ln P^r},\widehat{\delta}_i,\widehat{\ln \pi_i}} S_{\hat{u}_i^r \hat{u}_i^r} ,$$

with

$$S_{\hat{u}_i^r \hat{u}_i^r} = \sum_{r\in\mathcal{R}} \sum_{i\in\mathcal{N}_r} w_i^r \left(\ln p_i^r - \widehat{\delta}_i \widehat{\ln P^r} - \widehat{\ln \pi}_i\right)^2 .$$

The first-order conditions are

$$\frac{\partial S_{\hat{u}_i^r \hat{u}_i^r}}{\partial \widehat{\ln P^r}} = \sum_{i\in\mathcal{N}_r} w_i^r 2\left(\ln p_i^r - \widehat{\delta}_i \widehat{\ln P^r} - \widehat{\ln \pi}_i\right)\left(-\widehat{\delta}_i\right) = 0 \tag{A.1a}$$

$$\frac{\partial S_{\hat{u}_i^r \hat{u}_i^r}}{\partial \widehat{\delta}_i} = \sum_{r\in\mathcal{R}_i} w_i^r 2\left(\ln p_i^r - \widehat{\delta}_i \widehat{\ln P^r} - \widehat{\ln \pi}_i\right)\left(-\widehat{\ln P^r}\right) = 0 \tag{A.1b}$$

$$\frac{\partial S_{\hat{u}_i^r \hat{u}_i^r}}{\partial \widehat{\ln \pi_i}} = \sum_{r\in\mathcal{R}_i} w_i^r 2\left(\ln p_i^r - \widehat{\delta}_i \widehat{\ln P^r} - \widehat{\ln \pi}_i\right)(-1) = 0 . \tag{A.1c}$$

Rearranging condition (A.1a) gives

$$\widehat{\ln P^r} = \frac{\sum_{i\in\mathcal{N}_r} w_i^r \widehat{\delta}_i \left(\ln p_i^r - \widehat{\ln \pi}_i\right)}{\sum_{i\in\mathcal{N}_r} w_i^r \left(\widehat{\delta}_i\right)^2} .$$

Condition (A.1b) can be rewritten as

$$\widehat{\delta}_i = \frac{\sum_{r\in\mathcal{R}_i} w_i^r \widehat{\ln P^r}\left(\ln p_i^r - \widehat{\ln \pi}_i\right)}{\sum_{r\in\mathcal{R}_i} w_i^r \left(\widehat{\ln P^r}\right)^2} \,.$$

Condition (A.1c) yields

$$\widehat{\ln \pi}_i = \sum_{r\in\mathcal{R}_i} \frac{w_i^r}{\sum_{s\in\mathcal{R}_i} w_i^s}\left(\ln p_i^r - \widehat{\delta}_i \widehat{\ln P^r}\right) \,.$$

A.2 NLCPD model and special cases

Model (9) can be written as

$$\breve{\boldsymbol{y}} = \breve{\boldsymbol{G}}\boldsymbol{\pi} + \left(\frac{G_1}{w_1} + \widetilde{\boldsymbol{G}}\boldsymbol{\delta}\right) \odot \left(\breve{\boldsymbol{D}}\boldsymbol{p}\right) + \breve{\boldsymbol{u}} \,, \tag{A.2}$$

where $\breve{\boldsymbol{D}} = \left(\widetilde{D}^2 \;\; \ldots \;\; \widetilde{D}^R\right)$, $\breve{\boldsymbol{G}} = \left(G_1 \;\; \ldots \;\; G_N\right)$ and $\widetilde{\boldsymbol{G}} = \left(\widetilde{G}_2 \;\; \ldots \;\; \widetilde{G}_N\right)$. The vectors $\breve{\boldsymbol{y}}$ and $\breve{\boldsymbol{u}}$ contain the logarithmic prices, $\ln p_i^r$, and the errors, u_i^r, respectively. The parameters are $\boldsymbol{\pi} = \left(\ln \pi_1 \;\; \ldots \;\; \ln \pi_N\right)^\intercal$, $\boldsymbol{p} = \left(\ln P^2 \;\; \ldots \;\; \ln P^R\right)^\intercal$, and $\boldsymbol{\delta} = \left(\delta_2 \;\; \ldots \;\; \delta_N\right)^\intercal$, where the symbol $^\intercal$ denotes the transpose. The operator $\odot$ denotes the Hadamard product, that is, the elementwise multiplication of the column vectors $\left(G_1/w_1 + \widetilde{\boldsymbol{G}}\boldsymbol{\delta}\right)$ and $\left(\breve{\boldsymbol{D}}\boldsymbol{p}\right)$. The observations are sorted by product and then region. When the price data are complete, the number of price observations, B, is equal to NR.

If all δ_i-values were known (but possibly different from unity), we could define the matrix $\breve{\boldsymbol{H}} = \left(H^2 \;\; \ldots \;\; H^R\right)$ with $H^s = \left(G_1/w_1 + \sum_{j\in N\setminus\{1\}} \widetilde{G}_j \delta_j\right) \tilde{D}^s$ and we could write the NLCPD model (A.2) in the following linear form:

$$\breve{\boldsymbol{y}} = \breve{\boldsymbol{G}}\boldsymbol{\pi} + \breve{\boldsymbol{H}}\boldsymbol{p} + \breve{\boldsymbol{u}} \,. \tag{A.3}$$

To apply a weighted least squares approach, we define for each product i a diagonal $(R_i \times R_i)$-matrix of weights, $\boldsymbol{W}_i = \operatorname{diag}\left(\sqrt{w_i} \;\; \ldots \;\; \sqrt{w_i}\right)$, and combine them in the diagonal $(B \times B)$-matrix

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{W}_1 & \boldsymbol{0}_{R_1\times R_2} & \ldots & \boldsymbol{0}_{R_1\times R_N} \\ \boldsymbol{0}_{R_2\times R_1} & \boldsymbol{W}_2 & \ldots & \boldsymbol{0}_{R_2\times R_N} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0}_{R_N\times R_1} & \boldsymbol{0}_{R_N\times R_2} & \ldots & \boldsymbol{W}_N \end{bmatrix} \,,$$

where $\boldsymbol{0}_{R_i\times R_j}$ is a $(R_i \times R_j)$-matrix, all of whose entries are zero.

Furthermore, we define the three matrices $\boldsymbol{G}$, $\boldsymbol{D}$, and $\boldsymbol{H}$. The matrix $\boldsymbol{G}$ is defined by

$$\boldsymbol{G} = \boldsymbol{W}\breve{\boldsymbol{G}} = \begin{bmatrix} \sqrt{w_1}\breve{\boldsymbol{G}}_1 \\ \sqrt{w_2}\breve{\boldsymbol{G}}_2 \\ \vdots \\ \sqrt{w_N}\breve{\boldsymbol{G}}_N \end{bmatrix},$$

with the $(R_i \times N)$-matrices

$$\breve{\boldsymbol{G}}_1 = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 \end{bmatrix}, \quad \cdots \quad , \breve{\boldsymbol{G}}_N = \begin{bmatrix} 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

The matrices $\boldsymbol{D}$ and $\boldsymbol{H}$ are given by

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{D}_1 \\ \boldsymbol{D}_2 \\ \vdots \\ \boldsymbol{D}_N \end{bmatrix} = \begin{bmatrix} \sqrt{w_1}\,\breve{\boldsymbol{D}}_1 \\ \sqrt{w_2}\,\breve{\boldsymbol{D}}_2 \\ \vdots \\ \sqrt{w_N}\,\breve{\boldsymbol{D}}_N \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{H} = \begin{bmatrix} \boldsymbol{H}_1 \\ \boldsymbol{H}_2 \\ \vdots \\ \boldsymbol{H}_N \end{bmatrix} = \begin{bmatrix} \sqrt{w_1}\,\breve{\boldsymbol{H}}_1 \\ \sqrt{w_2}\,\breve{\boldsymbol{H}}_2 \\ \vdots \\ \sqrt{w_N}\,\breve{\boldsymbol{H}}_N \end{bmatrix}, \tag{A.4}$$

where

$$\breve{\boldsymbol{D}}_i = \begin{bmatrix} -1 & -1 & \cdots & -1 \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \qquad \text{and} \qquad \breve{\boldsymbol{H}}_i = \begin{bmatrix} -\delta_i & -\delta_i & \cdots & -\delta_i \\ \delta_i & 0 & \cdots & 0 \\ 0 & \delta_i & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \delta_i \end{bmatrix} = \delta_i \breve{\boldsymbol{D}}_i\,, \tag{A.5}$$

when product i is priced in all regions. When product i is missing in some region r, line r of $\breve{\boldsymbol{D}}_i$ and $\breve{\boldsymbol{H}}_i$ must be deleted. In any case, $\breve{\boldsymbol{D}}_i$ and $\breve{\boldsymbol{H}}_i$ are $(R_i \times (R-1))$-matrices.

Weighted least squares of model (A.3) with the weighting matrix $\boldsymbol{W}$ is equivalent to ordinary least squares of the model

$$\boldsymbol{y} = \boldsymbol{G}\boldsymbol{\pi} + \boldsymbol{H}\boldsymbol{p} + \boldsymbol{u}\,, \tag{A.6}$$

where $\boldsymbol{y} = \boldsymbol{W}\breve{\boldsymbol{y}}$ and $\boldsymbol{u} = \boldsymbol{W}\breve{\boldsymbol{u}}$.

The least squares estimators of model (A.6) are

$$\begin{aligned}\begin{bmatrix}\widehat{\boldsymbol{\pi}}\\ \widehat{\boldsymbol{p}}\end{bmatrix} &= \begin{bmatrix}\boldsymbol{G}^\intercal\boldsymbol{G} & \boldsymbol{G}^\intercal\boldsymbol{H}\\ \boldsymbol{H}^\intercal\boldsymbol{G} & \boldsymbol{H}^\intercal\boldsymbol{H}\end{bmatrix}^{-1}\begin{bmatrix}\boldsymbol{G}^\intercal\boldsymbol{y}\\ \boldsymbol{H}^\intercal\boldsymbol{y}\end{bmatrix}\\ &= \begin{bmatrix}(\boldsymbol{G}^\intercal\boldsymbol{L}\boldsymbol{G})^{-1}\boldsymbol{G}^\intercal\boldsymbol{L}\boldsymbol{y}\\ (\boldsymbol{H}^\intercal\boldsymbol{M}\boldsymbol{H})^{-1}\boldsymbol{H}^\intercal\boldsymbol{M}\boldsymbol{y}\end{bmatrix}, \end{aligned} \tag{A.7}$$

with $\widehat{\boldsymbol{\pi}} = \left(\widehat{\ln\pi_1} \;\; \dots \;\; \widehat{\ln\pi_N}\right)^\intercal$, $\widehat{\boldsymbol{p}} = \left(\widehat{\ln P^2} \;\; \dots \;\; \widehat{\ln P^R}\right)^\intercal$, and

$$\boldsymbol{L} = \boldsymbol{I}_B - \boldsymbol{H}\left(\boldsymbol{H}^\intercal\boldsymbol{H}\right)^{-1}\boldsymbol{H}^\intercal \tag{A.8}$$

$$\boldsymbol{M} = \boldsymbol{I}_B - \boldsymbol{G}\left(\boldsymbol{G}^\intercal\boldsymbol{G}\right)^{-1}\boldsymbol{G}^\intercal\,, \tag{A.9}$$

where $\boldsymbol{I}_B$ is the identity matrix with dimensions $B \times B$.

When all δ_i ($i \in \mathcal{N}$) are equal to unity, we get $\boldsymbol{H} = \boldsymbol{D}$, and model (A.6) becomes the CPD model:

$$\boldsymbol{y} = \boldsymbol{G}\boldsymbol{\pi} + \boldsymbol{D}\boldsymbol{p} + \boldsymbol{u}\,. \tag{A.10}$$

The corresponding estimators are

$$\begin{bmatrix}\widehat{\boldsymbol{\pi}}'\\ \widehat{\boldsymbol{p}}'\end{bmatrix} = \begin{bmatrix}(\boldsymbol{G}^\intercal\boldsymbol{K}\boldsymbol{G})^{-1}\boldsymbol{G}^\intercal\boldsymbol{K}\boldsymbol{y}\\ (\boldsymbol{D}^\intercal\boldsymbol{M}\boldsymbol{D})^{-1}\boldsymbol{D}^\intercal\boldsymbol{M}\boldsymbol{y}\end{bmatrix}, \tag{A.11}$$

with

$$\boldsymbol{K} = \boldsymbol{I}_B - \boldsymbol{D}\left(\boldsymbol{D}^\intercal\boldsymbol{D}\right)^{-1}\boldsymbol{D}^\intercal\,. \tag{A.12}$$

For the following derivations, some useful results are established. It can be shown that

$$\boldsymbol{G}\left(\boldsymbol{G}^\intercal\boldsymbol{G}\right)^{-1}\boldsymbol{G}^\intercal = \begin{bmatrix}\boldsymbol{G}_{11} & \boldsymbol{0}_{R_1\times R_2} & \cdots & \boldsymbol{0}_{R_1\times R_N}\\ \boldsymbol{0}_{R_2\times R_1} & \boldsymbol{G}_{22} & \cdots & \boldsymbol{0}_{R_2\times R_N}\\ \vdots & \vdots & \ddots & \vdots\\ \boldsymbol{0}_{R_N\times R_1} & \boldsymbol{0}_{R_N\times R_2} & \cdots & \boldsymbol{G}_{NN}\end{bmatrix}, \tag{A.13}$$

with the $(R_i \times R_i)$-submatrices

$$\boldsymbol{G}_{ii} = \begin{bmatrix}1/R_i & 1/R_i & \cdots & 1/R_i\\ 1/R_i & 1/R_i & \cdots & 1/R_i\\ \vdots & \vdots & \ddots & \vdots\\ 1/R_i & 1/R_i & \cdots & 1/R_i\end{bmatrix} = \frac{1}{R_i}\boldsymbol{1}_{R_i\times R_i}\,. \tag{A.14}$$

Thus,

$$\operatorname{tr}\left(\boldsymbol{G}\left(\boldsymbol{G}^{\intercal}\boldsymbol{G}\right)^{-1}\boldsymbol{G}^{\intercal}\right)=\sum_{i\in\mathcal{N}}R_i\frac{1}{R_i}=N\,. \tag{A.15}$$

For the matrix $\boldsymbol{D}$, we get the following result:

$$\boldsymbol{D}^{\intercal}\boldsymbol{D}=\breve{\boldsymbol{D}}^{\intercal}\boldsymbol{W}^{\intercal}\boldsymbol{W}\breve{\boldsymbol{D}}=\sum_{i\in\mathcal{N}}w_i\breve{\boldsymbol{D}}_i^{\intercal}\breve{\boldsymbol{D}}_i=\sum_{i\in\mathcal{N}}w_i\left(\boldsymbol{I}_{R-1}+\boldsymbol{C}_i\right)\,,$$

where the $(R-1)\times(R-1)$-matrix $\boldsymbol{C}_i$ is defined by

$$\boldsymbol{C}_i=\begin{bmatrix}c^1 & c^1 & \dots & c^1\\ c^1 & c^2 & \dots & c^1\\ \vdots & \vdots & \ddots & \vdots\\ c^1 & c^1 & \dots & c^{R-1}\end{bmatrix}\,,$$

with $c^r=1$ when product i is observed in region r and $c^r=0$ otherwise. Note that

$$\boldsymbol{D}^{\intercal}\boldsymbol{M}\boldsymbol{D}=\sum_{i\in\mathcal{N}}w_i\left(\breve{\boldsymbol{D}}_i^{\intercal}\breve{\boldsymbol{D}}_i-\frac{1}{R_i}\breve{\boldsymbol{D}}_i^{\intercal}\boldsymbol{1}_{R_i\times R_i}\breve{\boldsymbol{D}}_i\right)\,. \tag{A.16}$$

For the matrix $\boldsymbol{H}$ we get

$$\boldsymbol{H}^{\intercal}\boldsymbol{H}=\sum_{i\in\mathcal{N}}\left(\delta_i\right)^2w_i\breve{\boldsymbol{D}}_i^{\intercal}\breve{\boldsymbol{D}}_i=\sum_{i\in\mathcal{N}}\left(\delta_i\right)^2w_i\left(\boldsymbol{I}_{R-1}+\boldsymbol{C}_i\right)\,.$$

Furthermore,

$$\boldsymbol{D}^{\intercal}\boldsymbol{M}\boldsymbol{H}=\sum_{i\in\mathcal{N}}\delta_iw_i\left(\breve{\boldsymbol{D}}_i^{\intercal}\breve{\boldsymbol{D}}_i-\frac{1}{R_i}\breve{\boldsymbol{D}}_i^{\intercal}\boldsymbol{1}_{R_i\times R_i}\breve{\boldsymbol{D}}_i\right)\,.$$

When the data set is complete, some additonal results can be derived. First, we get $\boldsymbol{C}_i=\boldsymbol{1}_{(R-1)\times(R-1)}$. Thus, $\boldsymbol{D}^{\intercal}\boldsymbol{D}=\boldsymbol{I}_{R-1}+\boldsymbol{1}_{(R-1)\times(R-1)}$ and

$$\left(\boldsymbol{D}^{\intercal}\boldsymbol{D}\right)^{-1}=\boldsymbol{I}_{R-1}-\frac{1}{R}\boldsymbol{1}_{(R-1)\times(R-1)}\,, \tag{A.17}$$

where we exploited the rule that the inverse of some matrix $[\boldsymbol{I}_Z+k\boldsymbol{1}_{Z\times Z}]$, with k being some constant, is

$$\left[\boldsymbol{I}_Z+k\boldsymbol{1}_{Z\times Z}\right]^{-1}=\boldsymbol{I}_Z-\frac{k}{Zk+1}\boldsymbol{1}_{Z\times Z}\,. \tag{A.18}$$

Furthermore,

$$\boldsymbol{D}\left(\boldsymbol{D}^{\intercal}\boldsymbol{D}\right)^{-1}\boldsymbol{D}^{\intercal} = \begin{bmatrix} \boldsymbol{D}_{11} & \boldsymbol{D}_{12} & \dots & \boldsymbol{D}_{1N} \\ \boldsymbol{D}_{21} & \boldsymbol{D}_{22} & \dots & \boldsymbol{D}_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{D}_{N1} & \boldsymbol{D}_{N2} & \dots & \boldsymbol{D}_{NN} \end{bmatrix}, \tag{A.19}$$

with the $(R \times R)$-matrices

$$\boldsymbol{D}_{ij} = \sqrt{w_i w_j}\left(\boldsymbol{I}_R - \frac{1}{R}\mathbf{1}_{R\times R}\right). \tag{A.20}$$

Thus,

$$\operatorname{tr}\left(\boldsymbol{D}\left(\boldsymbol{D}^{\intercal}\boldsymbol{D}\right)^{-1}\boldsymbol{D}^{\intercal}\right) = R-1\,. \tag{A.21}$$

Concerning the two components of $\boldsymbol{D}^{\intercal}\boldsymbol{M}\boldsymbol{D}$ as specified in (A.16) we get

$$\breve{\boldsymbol{D}}_i^{\intercal}\breve{\boldsymbol{D}}_i = \boldsymbol{I}_{R-1} + \mathbf{1}_{(R-1)\times(R-1)} \tag{A.22}$$

and

$$\frac{1}{R_i}\breve{\boldsymbol{D}}_i^{\intercal}\mathbf{1}_{R\times R}\breve{\boldsymbol{D}}_i = \mathbf{0}_{(R-1)\times(R-1)}\,. \tag{A.23}$$

Thus, $\boldsymbol{D}^{\intercal}\boldsymbol{M}\boldsymbol{D} = \boldsymbol{I}_{R-1} + \mathbf{1}_{(R-1)\times(R-1)} = \boldsymbol{D}^{\intercal}\boldsymbol{D}$ and, therefore,

$$\left(\boldsymbol{D}^{\intercal}\boldsymbol{M}\boldsymbol{D}\right)^{-1} = \left(\boldsymbol{D}^{\intercal}\boldsymbol{D}\right)^{-1}. \tag{A.24}$$

Furthermore,

$$\begin{aligned} \boldsymbol{D}^{\intercal}\boldsymbol{M}\boldsymbol{H} &= \boldsymbol{I}_{R-1} + \mathbf{1}_{(R-1)\times(R-1)} && \text{(A.25)} \\ &= \boldsymbol{D}^{\intercal}\boldsymbol{D}\,, && \text{(A.26)} \end{aligned}$$

where we exploited the restriction $\sum_{i\in\mathcal{N}} w_i\delta_i = 1$. It can be easily checked that

$$\boldsymbol{D}^{\intercal}\boldsymbol{M} = \boldsymbol{D}^{\intercal}\,. \tag{A.27}$$

Thus,

$$\boldsymbol{D}^{\intercal}\boldsymbol{H} = \boldsymbol{D}^{\intercal}\boldsymbol{D}\,. \tag{A.28}$$

Since $\breve{\boldsymbol{G}}_i^{\intercal}\breve{\boldsymbol{D}}_i = \mathbf{0}_{N\times(R-1)}$ and $\boldsymbol{G}^{\intercal}\boldsymbol{D} = \sum_{i\in\mathcal{N}} w_i\breve{\boldsymbol{G}}_i^{\intercal}\breve{\boldsymbol{D}}_i$, we get

$$\boldsymbol{G}^{\intercal}\boldsymbol{D} = \mathbf{0}_{N\times(R-1)} \qquad \text{and} \qquad \boldsymbol{D}^{\intercal}\boldsymbol{G} = \mathbf{0}_{(R-1)\times N}\,. \tag{A.29}$$

As a consequence,

$$\boldsymbol{G}^{\intercal}\boldsymbol{K} = \boldsymbol{G}^{\intercal} - \boldsymbol{G}^{\intercal}\boldsymbol{D}\left(\boldsymbol{D}^{\intercal}\boldsymbol{D}\right)^{-1}\boldsymbol{D}^{\intercal} = \boldsymbol{G}^{\intercal}\,. \tag{A.30}$$

Analogously, we have

$$\boldsymbol{G}^{\intercal}\boldsymbol{H} = \boldsymbol{0}_{N\times(R-1)} \qquad \text{and} \qquad \boldsymbol{H}^{\intercal}\boldsymbol{G} = \boldsymbol{0}_{(R-1)\times N} \tag{A.31}$$

and

$$\boldsymbol{G}^{\intercal}\boldsymbol{L} = \boldsymbol{G}^{\intercal}\,. \tag{A.32}$$

A.3 Bias of the CPD estimators

We use $\boldsymbol{u}'$ to denote the vector of errors arising in the estimation of the CPD model (A.10) when model (A.6) is the correct model:

$$\boldsymbol{y} = \boldsymbol{G}\boldsymbol{\pi} + \boldsymbol{D}\boldsymbol{p} + \boldsymbol{u}'\,. \tag{A.33}$$

Therefore, $\boldsymbol{H}\boldsymbol{p}+\boldsymbol{u} = \boldsymbol{D}\boldsymbol{p}+\boldsymbol{u}'$ and the expected value of the error term of the CPD model (A.33) is given by

$$E\left(\boldsymbol{u}'\right) = (\boldsymbol{H}-\boldsymbol{D})\,\boldsymbol{p}\,. \tag{A.34}$$

The estimators of the NLCPD model (A.6) were stated in (A.7). They can be written in the form

$$\widehat{\boldsymbol{p}} = \boldsymbol{p} + (\boldsymbol{H}^{\intercal}\boldsymbol{M}\boldsymbol{H})^{-1}\,\boldsymbol{H}^{\intercal}\boldsymbol{M}\boldsymbol{u} \tag{A.35}$$

and

$$\widehat{\boldsymbol{\pi}} = \boldsymbol{\pi} + (\boldsymbol{G}^{\intercal}\boldsymbol{L}\boldsymbol{G})^{-1}\,\boldsymbol{G}^{\intercal}\boldsymbol{L}\boldsymbol{u}\,. \tag{A.36}$$

Since $E(\boldsymbol{u}) = \boldsymbol{0}_{B\times 1}$, we get $E(\widehat{\boldsymbol{p}}) = \boldsymbol{p}$ and $E(\widehat{\boldsymbol{\pi}}) = \boldsymbol{\pi}$. Thus, the estimators (A.35) and (A.36) would be unbiased.

The estimators of the CPD model (A.33) can be written in the form

$$\widehat{\boldsymbol{p}}' = \boldsymbol{p} + (\boldsymbol{D}^{\intercal}\boldsymbol{M}\boldsymbol{D})^{-1}\,\boldsymbol{D}^{\intercal}\boldsymbol{M}\boldsymbol{u}' \tag{A.37}$$

and

$$\widehat{\boldsymbol{\pi}}' = \boldsymbol{\pi} + (\boldsymbol{G}^{\intercal}\boldsymbol{K}\boldsymbol{G})^{-1}\,\boldsymbol{G}^{\intercal}\boldsymbol{K}\boldsymbol{u}'\,. \tag{A.38}$$

Inserting (A.34) and taking expectations yields

$$E\left(\widehat{\boldsymbol{p}}'\right) = (\boldsymbol{D}^{\intercal}\boldsymbol{M}\boldsymbol{D})^{-1}\,\boldsymbol{D}^{\intercal}\boldsymbol{M}\boldsymbol{H}\boldsymbol{p} \tag{A.39}$$

and

$$E\left(\widehat{\boldsymbol{\pi}}'\right) = \boldsymbol{\pi} + (\boldsymbol{G}^{\intercal}\boldsymbol{K}\boldsymbol{G})^{-1}\,\boldsymbol{G}^{\intercal}\boldsymbol{K}\boldsymbol{H}\boldsymbol{p}\,.$$

For a complete data set, we get the following result:

$$(\boldsymbol{D}^{\intercal}\boldsymbol{M}\boldsymbol{D})^{-1}\,\boldsymbol{D}^{\intercal}\boldsymbol{M}\boldsymbol{H} = \boldsymbol{I}_{R-1}\,. \tag{A.40}$$

Thus, the CPD estimators $\widehat{\boldsymbol{p}}'$ remain unbiased, provided the data set is complete. The same is true for the CPD estimators $\widehat{\boldsymbol{\pi}}'$, because (A.30) and (A.31) imply that $\boldsymbol{G}^{\intercal}\boldsymbol{K}\boldsymbol{H} = \boldsymbol{G}^{\intercal}\boldsymbol{H} = \boldsymbol{0}_{N\times(R-1)}$.

With data gaps, however, the matrix $(\boldsymbol{D}^{\intercal}\boldsymbol{M}\boldsymbol{D})^{-1}\,\boldsymbol{D}^{\intercal}\boldsymbol{M}\boldsymbol{H}$ does not simplify to the identity matrix. Suppose that the only data gap is product j in region 1 ($B = NR-1$ and $R_j = R-1$). Then, instead of (A.22) and (A.23), we get for product j

$$w_j\left(\breve{\boldsymbol{D}}_j^{\intercal}\breve{\boldsymbol{D}}_j - \frac{1}{R_j}\breve{\boldsymbol{D}}_j^{\intercal}\boldsymbol{1}_{R_j\times R_j}\breve{\boldsymbol{D}}_j\right) = w_j\left(\boldsymbol{I}_{R-1} - \frac{1}{R-1}\boldsymbol{1}_{(R-1)\times(R-1)}\right)\,.$$

However, for the other products, $i \neq j$, relationships (A.22) and (A.23) remain valid. Thus,

$$\boldsymbol{D}^{\intercal}\boldsymbol{M}\boldsymbol{D} = \boldsymbol{I}_{R-1} + \left(1 - \frac{w_j R}{R-1}\right)\boldsymbol{1}_{(R-1)\times(R-1)}\,. \tag{A.41}$$

Also, relationship (A.25) no longer applies. For product j we have

$$w_j\delta_j\left(\boldsymbol{I}_{R-1} + \boldsymbol{1}_{(R-1)\times(R-1)}\right) = w_j\delta_j\left(\boldsymbol{I}_{(R-1)} - \frac{1}{R-1}\boldsymbol{1}_{(R-1)\times(R-1)}\right)\,,$$

while for the other products, $i \neq j$, relationships (A.22) and (A.23) remain valid. Thus,

$$\boldsymbol{D}^{\intercal}\boldsymbol{M}\boldsymbol{H} = \boldsymbol{I}_{R-1} + \left(1 - \frac{w_j\delta_j R}{R-1}\right)\boldsymbol{1}_{(R-1)\times(R-1)}\,. \tag{A.42}$$

Inserting (A.41) and (A.42) in (A.39) yields

$$E\left(\widehat{\boldsymbol{p}}'\right) = \left(\boldsymbol{I}_{R-1} + \left(1 - \frac{w_j R}{R-1}\right)\boldsymbol{1}_{(R-1)\times(R-1)}\right)^{-1}\left(\boldsymbol{I}_{R-1} + \left(1 - \frac{w_j\delta_j R}{R-1}\right)\boldsymbol{1}_{(R-1)\times(R-1)}\right)\boldsymbol{p}\,.$$

Rule (A.18) implies that

$$\left(\boldsymbol{I}_{R-1} + \left(1 - \frac{w_j R}{R-1}\right)\boldsymbol{1}_{(R-1)\times(R-1)}\right)^{-1} = \boldsymbol{I}_{R-1} + \frac{1 - R(1-w_j)}{(R-1)\,R(1-w_j)}\boldsymbol{1}_{(R-1)\times(R-1)}\,.$$

Furthermore,

$$1 - \frac{w_j\delta_j R}{R-1} = \frac{R-1-w_j\delta_j R}{R-1} = \frac{(R-1-w_j\delta_j R)\,R(1-w_j)}{(R-1)\,R(1-w_j)}\,.$$

Thus,

$$E\left(\widehat{\boldsymbol{p}}'\right) = \left(\boldsymbol{I}_{R-1} + \frac{w_j\,(1-\delta_j)}{(1-w_j)\,(R-1)}\mathbf{1}_{(R-1)\times(R-1)}\right)\boldsymbol{p}$$

and for each entry $E\left(\widehat{\ln P^{r\prime}}\right)$ of the vector $E\,(\widehat{\boldsymbol{p}}')$ we get

$$E\left(\widehat{\ln P^{r\prime}}\right) = \ln P^r + \frac{w_j\,(1-\delta_j)}{(1-w_j)\,(R-1)} \sum_{s\in\mathcal{R}\setminus\{1\}} \ln P^s \qquad \text{for } r = 2, ..., R\,. \tag{A.43}$$

For $\delta_j = 1$, the quotient in (A.43) is equal to 0 and we get $E\,(\widehat{\boldsymbol{p}}') = \boldsymbol{p}$. For $\delta_j < 1$, the quotient becomes positive. If region 1 (the region where product j is missing) is cheaper than average, the average of the logarithmic price levels of the other regions is positive: $\sum_{s\in\mathcal{R}\setminus\{1\}} \ln P^s > 0$. Thus, the estimated logarithmic price levels $\widehat{\ln P^{r\prime}}$ $(r = 2, ..., R)$ are biased upward. Since $\widehat{\ln P^{1\prime}} = -\sum_{s\in\mathcal{R}\setminus\{1\}} \widehat{\ln P^{s\prime}}$, the estimated logarithmic price level of region 1 is biased downward. If region 1 were more expensive than the average of all regions, the opposite bias would arise. For $\delta_j > 1$, the directions of bias are exactly opposite to those arising with $\delta_j < 1$.

A.4 Inference in the NLCPD and CPD models

A.4.1 NLCPD model

The NLCPD estimators that minimize the sum of squared residuals defined in Eq. (16) can be combined in the vector $\widehat{\boldsymbol{\beta}} = \left(\widehat{\boldsymbol{\pi}}^{\mathsf{T}}\ \widehat{\boldsymbol{\delta}}^{\mathsf{T}}\ \widehat{\boldsymbol{p}}^{\mathsf{T}}\right)^{\mathsf{T}}$, where $\widehat{\boldsymbol{\pi}} = \left(\widehat{\ln\pi}_1\ \ldots\ \widehat{\ln\pi}_N\right)^{\mathsf{T}}$, $\widehat{\boldsymbol{\delta}} = \left(\widehat{\delta}_2\ \ldots\ \widehat{\delta}_N\right)^{\mathsf{T}}$, and $\widehat{\boldsymbol{p}} = \left(\widehat{\ln P^2}\ \ldots\ \widehat{\ln P^R}\right)^{\mathsf{T}}$. The estimated NLCPD model can be written in the form

$$\widehat{\ln p_i^r} = \widehat{\delta}_i \widehat{\ln P^r} + \widehat{\ln\pi}_i\,, \tag{A.44}$$

where $\widehat{\ln p_i^r}$ is the estimated logarithmic price of product i in region r. Each of the B estimated logarithmic prices is a function of the NLCPD estimators $\widehat{\boldsymbol{\beta}}$: $\widehat{\ln p_i^r} = f_i^r(\widehat{\boldsymbol{\beta}})$. More specifically, these functions have the following form:

$$\begin{aligned}
\widehat{\ln p_1^1} &= f_1^1(\widehat{\boldsymbol{\beta}}) = -\frac{1-\sum_{i=2}^N w_i\widehat{\delta}_i}{w_1} \sum_{r=2}^R \widehat{\ln P^r} + \widehat{\ln\pi}_1 \\
&\vdots \\
\widehat{\ln p_N^R} &= f_N^R(\widehat{\boldsymbol{\beta}}) = \widehat{\delta}_N \widehat{\ln P^R} + \widehat{\ln\pi}_N\,.
\end{aligned}$$

Each of these functions can be differentiated with respect to each of its arguments in $\widehat{\boldsymbol{\beta}}$. Combining these derivatives in the Jacobian matrix $\boldsymbol{J}$ yields

$$\boldsymbol{J} = \begin{bmatrix} \frac{\partial f_1^1}{\partial \widehat{\ln \pi_1}} & \dots & \frac{\partial f_1^1}{\partial \widehat{\ln \pi_N}} & \frac{\partial f_1^1}{\partial \widehat{\delta}_2} & \dots & \frac{\partial f_1^1}{\partial \widehat{\delta}_N} & \frac{\partial f_1^1}{\partial \widehat{\ln P^2}} & \dots & \frac{\partial f_1^1}{\partial \widehat{\ln P^R}} \\ \frac{\partial f_1^2}{\partial \widehat{\ln \pi_1}} & \dots & \frac{\partial f_1^2}{\partial \widehat{\ln \pi_N}} & \frac{\partial f_1^2}{\partial \widehat{\delta}_2} & \dots & \frac{\partial f_1^2}{\partial \widehat{\delta}_N} & \frac{\partial f_1^2}{\partial \widehat{\ln P^2}} & \dots & \frac{\partial f_1^2}{\partial \widehat{\ln P^R}} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_N^R}{\partial \widehat{\ln \pi_1}} & \dots & \frac{\partial f_N^R}{\partial \widehat{\ln \pi_N}} & \frac{\partial f_N^R}{\partial \widehat{\delta}_2} & \dots & \frac{\partial f_N^R}{\partial \widehat{\delta}_N} & \frac{\partial f_N^R}{\partial \widehat{\ln P^2}} & \dots & \frac{\partial f_N^R}{\partial \widehat{\ln P^R}} \end{bmatrix}.$$

Approximated standard errors of the NLCPD estimators $\widehat{\boldsymbol{\beta}}$ can be derived from the estimated asymptotic variance matrix, $\widehat{\boldsymbol{V}}\left(\widehat{\boldsymbol{\beta}}\right)$. Following Cameron and Trivedi (2005, pp. 156-157), this matrix is

$$\begin{aligned} \widehat{\boldsymbol{V}}\left(\widehat{\boldsymbol{\beta}}\right) &= \left(\boldsymbol{J}^\intercal \boldsymbol{W}^\intercal \boldsymbol{W} \boldsymbol{J}\right)^{-1} \boldsymbol{J}^\intercal \boldsymbol{W}^\intercal \boldsymbol{W} \widehat{\boldsymbol{\Omega}} \boldsymbol{W}^\intercal \boldsymbol{W} \boldsymbol{J} \left(\boldsymbol{J}^\intercal \boldsymbol{W}^\intercal \boldsymbol{W} \boldsymbol{J}\right)^{-1} \\ &= \left(\boldsymbol{X}^\intercal \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\intercal \boldsymbol{W} \widehat{\boldsymbol{\Omega}} \boldsymbol{W}^\intercal \boldsymbol{X} \left(\boldsymbol{X}^\intercal \boldsymbol{X}\right)^{-1}, \end{aligned} \tag{A.45}$$

with $\boldsymbol{X} = \boldsymbol{W}\boldsymbol{J}$ and $\widehat{\boldsymbol{\Omega}} = \operatorname{diag}\left(\left(\widehat{u}_1^1\right)^2 \dots \left(\widehat{u}_N^R\right)^2\right)$, where $\widehat{u}_i^r = \ln p_i^r - \widehat{\delta}_i \widehat{\ln P^r} - \widehat{\ln \pi}_i$. If the unweighted error term, u_i^r, is homoskedastic, we get $\widehat{\boldsymbol{\Omega}} = \widehat{\sigma}^2 \boldsymbol{I}$ and Eq. (A.45) becomes

$$\widehat{\boldsymbol{V}}\left(\widehat{\boldsymbol{\beta}}\right) = \widehat{\sigma}^2 \left(\boldsymbol{X}^\intercal \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\intercal \boldsymbol{W} \boldsymbol{W}^\intercal \boldsymbol{X} \left(\boldsymbol{X}^\intercal \boldsymbol{X}\right)^{-1}.$$

If the weighted error term, $\sqrt{w_i} u_i^r$, is homoskedastic, then $\boldsymbol{W} \widehat{\boldsymbol{\Omega}} \boldsymbol{W}^\intercal = \widehat{\sigma}^2 \boldsymbol{I}$ and, therefore, Eq. (A.45) simplifies to

$$\widehat{\boldsymbol{V}}\left(\widehat{\boldsymbol{\beta}}\right) = \widehat{\sigma}^2 \left(\boldsymbol{X}^\intercal \boldsymbol{X}\right)^{-1}.$$

The latter case was considered in Section 4.4 when the standard errors of the NLCPD estimators $\widehat{\boldsymbol{\beta}}$ were discussed. For a complete data set, the estimated standard errors were given by formulas (23) to (25). In the following, these formulas are derived.

The $(NR) \times (2N + R - 2)$-matrix $\boldsymbol{X}$ has three submatrices. The first one is formed by the N columns related to the derivatives with respect to $\widehat{\ln \pi}_i$. This submatrix is equal to $\boldsymbol{G}$. The second submatrix is formed by the $(N-1)$ columns related to the derivatives with respect to $\widehat{\delta}_i$ $(i \in N \setminus \{1\})$. Defining the $(NR \times (N-1))$-matrices

$$\begin{aligned} \boldsymbol{Q} &= \boldsymbol{W} \left(\mathbf{1}_{N \times (N-1)} \otimes \left(-\mathbf{1}_{1 \times (R-1)} \widehat{\boldsymbol{p}} \quad \widehat{\boldsymbol{p}}^\intercal\right)^\intercal\right) \\ \boldsymbol{S} &= \begin{bmatrix} \mathbf{1}_{R \times (N-1)} \operatorname{diag}\left(-w_2/w_1 \;\; -w_3/w_1 \;\; \dots \;\; -w_N/w_1\right) \\ \boldsymbol{I}_{N-1} \otimes \mathbf{1}_{R \times 1} \end{bmatrix}, \end{aligned}$$

they can be written as $\boldsymbol{Q} \odot \boldsymbol{S}$. The third submatrix of $\boldsymbol{X}$ is formed by the $(R-1)$ columns

related to the derivatives with respect to $\widehat{\ln P^r}$. This submatrix is equal to

$$\widehat{\boldsymbol{H}} = \boldsymbol{W} \begin{bmatrix} \widehat{\boldsymbol{H}}_1 \\ \widehat{\boldsymbol{H}}_2 \\ \vdots \\ \widehat{\boldsymbol{H}}_N \end{bmatrix}, \quad \text{with} \quad \widehat{\boldsymbol{H}}_i = \begin{bmatrix} -\widehat{\delta}_i & -\widehat{\delta}_i & \cdots & -\widehat{\delta}_i \\ \widehat{\delta}_i & 0 & \cdots & 0 \\ 0 & \widehat{\delta}_i & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \widehat{\delta}_i \end{bmatrix} \qquad \text{for } i \in \mathcal{N} \,,$$

where $\widehat{\delta}_1 = \left(1 - \boldsymbol{w}^\intercal \widehat{\boldsymbol{\delta}}\right) \big/ w_1$ and $\boldsymbol{w} = (w_2 \, w_3 \, \ldots \, w_N)^\intercal$.

Putting the three submatrices together, the matrix $\boldsymbol{X}$ can be stated in the following compact form:

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{G} & \boldsymbol{Q} \odot \boldsymbol{S} & \widehat{\boldsymbol{H}} \end{bmatrix} . \tag{A.46}$$

Thus,

$$\boldsymbol{X}^\intercal \boldsymbol{X} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{0}_{N\times(N-1)} & \boldsymbol{0}_{N\times(R-1)} \\ \boldsymbol{0}_{(N-1)\times N} & \boldsymbol{A}_{22} & \boldsymbol{A}_{23} \\ \boldsymbol{0}_{(R-1)\times N} & \boldsymbol{A}_{32} & \boldsymbol{A}_{33} \end{bmatrix}, \tag{A.47}$$

with

$$\begin{aligned} \boldsymbol{A}_{11} &= R \, \mathrm{diag}(w_1 \, \boldsymbol{w}) \\ \boldsymbol{A}_{22} &= \left(\mathrm{diag}\,(\boldsymbol{w}) + \frac{\boldsymbol{w}^\intercal \boldsymbol{w}}{w_1} \right) \sum_{r \in \mathcal{R}} \left(\widehat{\ln P^r} \right)^2 \\ \boldsymbol{A}_{33} &= \left(\boldsymbol{I}_{(R-1)} + \boldsymbol{1}_{(R-1)\times(R-1)} \right) \sum_{i \in \mathcal{N}} w_i \left(\widehat{\delta}_i \right)^2 \\ \boldsymbol{A}_{23} &= (\boldsymbol{A}_{32})^\intercal = \mathrm{diag}\,(\boldsymbol{w}) \, \widetilde{\boldsymbol{d}} \widetilde{\boldsymbol{p}}^\intercal \,, \end{aligned}$$

where $\widetilde{\boldsymbol{d}} = \left(\widehat{\boldsymbol{\delta}} - \widehat{\delta}_1 \, \boldsymbol{1}_{(N-1)\times 1} \right)$ and $\widetilde{\boldsymbol{p}} = \left(\widehat{\boldsymbol{p}} - \widehat{\ln P^1} \, \boldsymbol{1}_{(R-1)\times 1} \right)$.

The inverse of the matrix (A.47) is denoted by

$$(\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} = \begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} & \boldsymbol{B}_{13} \\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22} & \boldsymbol{B}_{23} \\ \boldsymbol{B}_{31} & \boldsymbol{B}_{32} & \boldsymbol{B}_{33} \end{bmatrix}, \tag{A.48}$$

with

$$\boldsymbol{B}_{11} = \boldsymbol{A}_{11}^{-1}$$

$$\boldsymbol{B}_{22} = \left(\boldsymbol{A}_{22} - \boldsymbol{A}_{23} \boldsymbol{A}_{33}^{-1} \boldsymbol{A}_{32} \right)^{-1} \tag{A.49}$$

$$\boldsymbol{B}_{33} = \left(\boldsymbol{A}_{33} - \boldsymbol{A}_{32} \boldsymbol{A}_{22}^{-1} \boldsymbol{A}_{23} \right)^{-1} . \tag{A.50}$$

Obviously, $\boldsymbol{B}_{11} = (1/R)\,\mathrm{diag}\,(w_1\ \boldsymbol{w})^{-1}$. For the derivation of $\boldsymbol{B}_{22}$ and $\boldsymbol{B}_{33}$ we need the inverses of $\boldsymbol{A}_{22}$ and $\boldsymbol{A}_{33}$. For the latter, we can invoke rule (A.18) and get

$$\boldsymbol{A}_{33}^{-1} = \left(\boldsymbol{I}_{R-1} - \frac{1}{R}\,\mathbf{1}_{(R-1)\times(R-1)}\right)\left(\sum_{i\in\mathcal{N}} w_i\left(\widehat{\delta}_i\right)^2\right)^{-1} .$$

For the derivation of $\boldsymbol{A}_{22}^{-1}$ we make use of a generalization of rule (A.18) that is due to Miller (1981, pp.68-69) and obtain

$$\boldsymbol{A}_{22}^{-1} = \left(\mathrm{diag}(\boldsymbol{w})^{-1} - \mathbf{1}_{(N-1)\times(N-1)}\right)\frac{1}{\sum\limits_{r\in\mathcal{R}}\left(\widehat{\ln P^r}\right)^2} .$$

Next, we insert the definitions of $\boldsymbol{A}_{22}^{-1}$ and $\boldsymbol{A}_{33}^{-1}$ into (A.49) and (A.50) and, finally, obtain

$$\boldsymbol{B}_{22} = \boldsymbol{A}_{22}^{-1} + \frac{1}{\sum\limits_{r\in\mathcal{R}}\left(\widehat{\ln P^r}\right)^2}\,\boldsymbol{V}$$

$$\boldsymbol{B}_{33} = \boldsymbol{A}_{33}^{-1} - \frac{1}{\sum\limits_{r\in\mathcal{R}}\left(\widehat{\ln P^r}\right)^2}\,\frac{1-\sum\limits_{i\in\mathcal{N}} w_i\left(\widehat{\delta}_i\right)^2}{\sum\limits_{i\in\mathcal{N}} w_i\left(\widehat{\delta}_i\right)^2}\,\boldsymbol{Z} ,$$

where

$$\boldsymbol{V} = \widetilde{\boldsymbol{d}}\widetilde{\boldsymbol{d}}^{\intercal} + \left(\widehat{\delta}_1 - 1\right)\left(\widetilde{\boldsymbol{d}}\mathbf{1}_{1\times(N-1)} + \mathbf{1}_{(N-1)\times 1}\widetilde{\boldsymbol{d}}^{\intercal}\right) + \left(\widehat{\delta}_1 - 1\right)^2\,\mathbf{1}_{(N-1)\times(N-1)}$$

$$\boldsymbol{Z} = \widetilde{\boldsymbol{p}}\widetilde{\boldsymbol{p}}^{\intercal} + \widehat{\ln P^1}\,\left(\widetilde{\boldsymbol{p}}\mathbf{1}_{1\times(R-1)} + \mathbf{1}_{(R-1)\times 1}\widetilde{\boldsymbol{p}}^{\intercal}\right) + \left(\widehat{\ln P^1}\right)^2\,\mathbf{1}_{(R-1)\times(R-1)} .$$

Multiplying the diagonal elements of $\boldsymbol{B}_{11}$, $\boldsymbol{B}_{22}$, and $\boldsymbol{B}_{33}$ by the estimated model variance,

$$\widehat{\sigma}^2 = \frac{S_{\hat{u}_i^r\hat{u}_i^r}}{RN - 2N - R + 2} , \tag{A.51}$$

and taking the square root of each of these products, gives the formulas (23) to (25).

A.4.2 CPD model

When at least one δ_i-value is different from one and observations are (non-randomly) missing, the weighted CPD estimators $\widehat{\boldsymbol{p}}'$ are biased and, therefore, inference is invalid. Thus, the following analysis can be restricted to the case of complete data. In the following, the formulas for the standard errors of the weighted CPD estimators $\widehat{\boldsymbol{\pi}}'$ and $\widehat{\boldsymbol{p}}'$ are derived. For

comparability with the corresponding formulas of the NLCPD method, it is assumed that the weighted error term, $\sqrt{w_i}u_i^r$, is homoskedastic:

$$var(\sqrt{w_i}u_i^r) = \sigma^2 \,. \tag{A.52}$$

It is shown that the estimated standard errors of the CPD method are upward biased.

Exploiting (A.11), (A.27), and (A.30), the estimated CPD model can be written in the form

$$\widehat{\boldsymbol{y}}' = \left(\boldsymbol{G}\left(\boldsymbol{G}^\intercal\boldsymbol{G}\right)^{-1}\boldsymbol{G}^\intercal + \boldsymbol{D}\left(\boldsymbol{D}^\intercal\boldsymbol{D}\right)^{-1}\boldsymbol{D}^\intercal\right)\boldsymbol{y}\,.$$

Inserting this result in $\widehat{\boldsymbol{u}}' = \boldsymbol{y} - \widehat{\boldsymbol{y}}'$ yields

$$\widehat{\boldsymbol{u}}' = \boldsymbol{N}\boldsymbol{y}\,, \tag{A.53}$$

with

$$\boldsymbol{N} = \boldsymbol{I}_{NR} - \boldsymbol{G}\left(\boldsymbol{G}^\intercal\boldsymbol{G}\right)^{-1}\boldsymbol{G}^\intercal - \boldsymbol{D}\left(\boldsymbol{D}^\intercal\boldsymbol{D}\right)^{-1}\boldsymbol{D}^\intercal\,. \tag{A.54}$$

Relationships (A.28), (A.29), and (A.31) imply that $\boldsymbol{D}^\intercal\boldsymbol{H} = \boldsymbol{D}^\intercal\boldsymbol{D}$, $\boldsymbol{G}^\intercal\boldsymbol{D} = \boldsymbol{0}_{N\times(R-1)}$, $\boldsymbol{D}^\intercal\boldsymbol{G} = \boldsymbol{0}_{(R-1)\times N}$, and $\boldsymbol{G}^\intercal\boldsymbol{H} = \boldsymbol{0}_{N\times(R-1)}$, respectively. Thus,

$$\boldsymbol{N}\boldsymbol{D} = \boldsymbol{0}_{NR\times(R-1)}\,, \qquad \boldsymbol{N}\boldsymbol{G} = \boldsymbol{0}_{NR\times N}\,, \qquad \text{and} \qquad \boldsymbol{N}\boldsymbol{H} = \boldsymbol{H} - \boldsymbol{D}\,. \tag{A.55}$$

We know that $\boldsymbol{u}' = (\boldsymbol{H}-\boldsymbol{D})\boldsymbol{p} + \boldsymbol{u} = \boldsymbol{N}\boldsymbol{H}\boldsymbol{p} + \boldsymbol{u}$ and $E(\boldsymbol{u}) = \boldsymbol{0}_{NR\times 1}$. Thus,

$$E\left(\boldsymbol{u}'\boldsymbol{u}'^\intercal\right) = E\left((\boldsymbol{N}\boldsymbol{H}\boldsymbol{p}+\boldsymbol{u})\left(\boldsymbol{p}^\intercal\boldsymbol{H}^\intercal\boldsymbol{N}+\boldsymbol{u}^\intercal\right)\right) = \boldsymbol{N}\boldsymbol{H}\boldsymbol{p}\boldsymbol{p}^\intercal\boldsymbol{H}^\intercal\boldsymbol{N} + \sigma^2\boldsymbol{I}_{NR}\,, \tag{A.56}$$

where σ^2 is the variance used in (A.52).

Since (A.27) to (A.32) as well as (A.37) apply, the variance-covariance matrix of the CPD estimators $\widehat{\boldsymbol{p}}'$ is

$$V\left(\widehat{\boldsymbol{p}}'\right) = E\left[\left(\widehat{\boldsymbol{p}}'-\boldsymbol{p}\right)\left(\widehat{\boldsymbol{p}}'-\boldsymbol{p}\right)^\intercal\right] = \left(\boldsymbol{D}^\intercal\boldsymbol{D}\right)^{-1}\boldsymbol{D}^\intercal E\left(\boldsymbol{u}'\boldsymbol{u}'^\intercal\right)\boldsymbol{D}\left(\boldsymbol{D}^\intercal\boldsymbol{D}\right)^{-1}\,. \tag{A.57}$$

Inserting expression (A.56) in (A.57) and using (A.55) yields

$$V\left(\widehat{\boldsymbol{p}}'\right) = \sigma^2\left(\boldsymbol{D}^\intercal\boldsymbol{D}\right)^{-1}\,, \tag{A.58}$$

where the precise form of $\left(\boldsymbol{D}^\intercal\boldsymbol{D}\right)^{-1}$ was given in (A.17).

Using (A.30), the variance-covariance matrix of the CPD estimators $\widehat{\boldsymbol{\pi}}'$ is

$$V\left(\widehat{\boldsymbol{\pi}}'\right) = \left(\boldsymbol{G}^\intercal\boldsymbol{G}\right)^{-1}\boldsymbol{G}^\intercal E\left(\boldsymbol{u}'\boldsymbol{u}'^\intercal\right)\boldsymbol{G}\left(\boldsymbol{G}^\intercal\boldsymbol{G}\right)^{-1}\,. \tag{A.59}$$

Inserting expression (A.56) in (A.59) and using (A.55) yields

$$V\left(\widehat{\boldsymbol{\pi}}'\right) = \sigma^2 \left(\boldsymbol{G}^{\intercal}\boldsymbol{G}\right)^{-1} . \tag{A.60}$$

Substituting in (A.58) and (A.60) the variance σ^2 with its CPD estimator,

$$(\widehat{\sigma}')^2 = \frac{\widehat{\boldsymbol{u}}'^{\intercal}\widehat{\boldsymbol{u}}'}{NR-N-R+1} , \tag{A.61}$$

yields the estimated CPD variance-covariance matrices $\widehat{V\left(\widehat{\boldsymbol{p}}'\right)}$ and $\widehat{V\left(\widehat{\boldsymbol{\pi}}'\right)}$. The square roots of the diagonal elements of $\widehat{V\left(\widehat{\boldsymbol{p}}'\right)}$ and $\widehat{V\left(\widehat{\boldsymbol{\pi}}'\right)}$ are the estimators of the standard errors of the CPD estimators:

$$\widehat{se}'\left(\widehat{\ln P^{r\prime}}\right) = \widehat{\sigma}'\sqrt{(R-1)/R}$$
$$\widehat{se}'\left(\widehat{\ln \pi_i'}\right) = \widehat{\sigma}' \Big/ \sqrt{Rw_i} .$$

For these estimators to be unbiased, the estimate of σ^2 must be unbiased. Thus, we have to examine whether

$$E\left(\widehat{\boldsymbol{u}}'^{\intercal}\widehat{\boldsymbol{u}}'\right) = \sigma^2/\left(NR-N-R+1\right) . \tag{A.62}$$

Substituting in (A.53) the vector $\boldsymbol{y}$ with $(\boldsymbol{G\pi}+\boldsymbol{Dp}+\boldsymbol{u}')$ yields

$$\widehat{\boldsymbol{u}}' = \boldsymbol{NG\pi}+\boldsymbol{NDp}+\boldsymbol{Nu}' = \boldsymbol{Nu}' .$$

Therefore,

$$\widehat{\boldsymbol{u}}'^{\intercal}\widehat{\boldsymbol{u}}' = \boldsymbol{u}'^{\intercal}\boldsymbol{N}^{\intercal}\boldsymbol{Nu}' = \operatorname{tr}\left(\boldsymbol{u}'^{\intercal}\boldsymbol{Nu}'\right) = \operatorname{tr}\left(\boldsymbol{uu}'^{\intercal}\boldsymbol{N}\right) .$$

Taking expections gives

$$E\left(\widehat{\boldsymbol{u}}'^{\intercal}\widehat{\boldsymbol{u}}'\right) = E\left(\operatorname{tr}\left(\boldsymbol{u}'\boldsymbol{u}'^{\intercal}\boldsymbol{N}\right)\right) = \operatorname{tr}\left(E\left(\boldsymbol{u}'\boldsymbol{u}'^{\intercal}\right)\boldsymbol{N}\right) .$$

Inserting (A.56) yields

$$E\left(\widehat{\boldsymbol{u}}'^{\intercal}\widehat{\boldsymbol{u}}'\right) = \operatorname{tr}\left(\boldsymbol{NHpp}^{\intercal}\boldsymbol{H}^{\intercal}\boldsymbol{N}\right) + \operatorname{tr}\left(\sigma^2\boldsymbol{N}\right) . \tag{A.63}$$

Note that

$$\operatorname{tr}\left(\sigma^2\boldsymbol{N}\right) = \sigma^2\left(NR-N-R+1\right) , \tag{A.64}$$

where we exploited the results (A.15) and (A.21). Thus, unbiasedness requires that in (A.63)

we have $\operatorname{tr}(\boldsymbol{N}\boldsymbol{H}\boldsymbol{p}\boldsymbol{p}^\intercal\boldsymbol{N}\boldsymbol{H}^\intercal) = 0$ or, equivalently, $\operatorname{tr}(\boldsymbol{H}-\boldsymbol{D})\,\boldsymbol{p}\boldsymbol{p}^\intercal\,(\boldsymbol{H}^\intercal-\boldsymbol{D}^\intercal) = 0$. However,

$$
\begin{aligned}
&(\boldsymbol{H}-\boldsymbol{D})\,\boldsymbol{p}\boldsymbol{p}^\intercal\,(\boldsymbol{H}-\boldsymbol{D})^\intercal = \\
&\begin{bmatrix}
\sqrt{w_1}\sqrt{w_1}\,(\delta_1-1)\,(\delta_1-1) & \sqrt{w_1}\sqrt{w_2}\,(\delta_1-1)\,(\delta_2-1) & \dots & \sqrt{w_1}\sqrt{w_N}\,(\delta_1-1)\,(\delta_N-1) \\
\sqrt{w_2}\sqrt{w_1}\,(\delta_2-1)\,(\delta_1-1) & \sqrt{w_2}\sqrt{w_2}\,(\delta_2-1)\,(\delta_2-1) & \dots & \sqrt{w_2}\sqrt{w_N}\,(\delta_2-1)\,(\delta_N-1) \\
\vdots & \vdots & \ddots & \vdots \\
\sqrt{w_N}\sqrt{w_1}\,(\delta_N-1)\,(\delta_1-1) & \sqrt{w_N}\sqrt{w_2}\,(\delta_N-1)\,(\delta_2-1) & \dots & \sqrt{w_N}\sqrt{w_N}\,(\delta_N-1)\,(\delta_N-1)
\end{bmatrix} \\
&\otimes \begin{bmatrix}
\left(\sum_{r\in\mathcal{R}\setminus\{1\}} \ln P^r\right)^2 & -\left(\sum_{r\in\mathcal{R}\setminus\{1\}} \ln P^r\right)\ln P^2 & \dots & -\left(\sum_{r\in\mathcal{R}\setminus\{1\}} \ln P^r\right)\ln P^R \\
-\ln P^2\left(\sum_{r\in\mathcal{R}\setminus\{1\}} \ln P^r\right) & \ln P^2 \ln P^2 & \dots & \ln P^2 \ln P^R \\
\vdots & \vdots & \ddots & \vdots \\
-\ln P^R\left(\sum_{r\in\mathcal{R}\setminus\{1\}} \ln P^r\right) & \ln P^R \ln P^2 & \dots & \ln P^R \ln P^R
\end{bmatrix},
\end{aligned}
$$

where $\otimes$ denotes the Kronecker product. The trace of this matrix is

$$
\operatorname{tr}\left((\boldsymbol{H}-\boldsymbol{D})\,\boldsymbol{p}\boldsymbol{p}^\intercal\,(\boldsymbol{H}-\boldsymbol{D})^\intercal\right) = \sum_{i\in\mathcal{N}} w_i\,(\delta_i-1)^2 \left(\left(\sum_{r\in\mathcal{R}\setminus\{1\}} \ln P^r\right)^2 + \sum_{r\in\mathcal{R}\setminus\{1\}} (\ln P^r)^2\right). \quad \text{(A.65)}
$$

This expression is larger than zero and, therefore, the estimator $(\widehat{\sigma}')^2$ is larger than σ^2, except when $\delta_i = 1$ for all $i \in \mathcal{N}$.

B Simulation results

Tab. 5 provides error metrics for all parameters of the simulation setting described in Section 5.1. Mean absolute bias and mean RMSE of the estimates of $\ln P^r$ are replicated from Tab. 2. Mean absolute bias and mean RMSE of the estimates of $\ln\pi$ and δ are analogously defined to Eq. (26), but averaged over products instead of regions, e.g.:

$$
\begin{aligned}
\operatorname{Bias}\left(\widehat{\ln\pi}\right) &= \frac{1}{N}\sum_{i\in\mathcal{N}} \operatorname{Bias}\left(\widehat{\ln\pi}_i\right) = \frac{1}{N}\sum_{i\in\mathcal{N}} \frac{1}{L}\sum_{l=1}^{L}\left(\widehat{\ln\pi}_{i,l} - \ln\pi_{i,l}\right) \\
\operatorname{RMSE}\left(\widehat{\ln\pi}\right) &= \frac{1}{N}\sum_{i\in\mathcal{N}} \operatorname{RMSE}\left(\widehat{\ln\pi}_i\right) = \frac{1}{N}\sum_{i\in\mathcal{N}} \sqrt{\frac{1}{L}\sum_{l=1}^{L}\left(\widehat{\ln\pi}_{i,l} - \ln\pi_{i,l}\right)^2}.
\end{aligned}
$$

for the $\ln\pi_i$-parameters of the NLCPD method. For the CPD method, $\operatorname{Bias}\left(\widehat{\ln\pi'}\right)$ and $\operatorname{RMSE}\left(\widehat{\ln\pi'}\right)$ are defined in the same way.

The CPD method does not provide any estimates for δ_i but implicitly assumes that $\delta_i = 1$. Consequently, in the computation of $\operatorname{Bias}(\widehat{\delta}')$ and $\operatorname{RMSE}(\widehat{\delta}')$ we set $\widehat{\delta}'_{i,l} = 1$ for all products i and iterations l. Due to this exogenous restriction of the CPD model, the estimates $\widehat{\delta}'$ are found to be markedly biased except for Scenario 1 (see third line of Tab. 5).

		Scenario 1		Scenario 2		Scenario 3		Scenario 4	
		CPD	NLCPD	CPD	NLCPD	CPD	NLCPD	CPD	NLCPD
Bias	$\ln P^r$	0.0002	0.0002	0.0002	0.0001	0.0003	0.0002	0.0133	0.0002
	$\ln \pi_i$	0.0002	0.0002	0.0003	0.0003	0.0004	0.0004	0.0005	0.0004
	δ_i	0.0000	0.0021	0.5527	0.0023	0.5527	0.0035	0.5527	0.0035
RMSE	$\ln P^r$	0.0097	0.0097	0.0097	0.0081	0.0201	0.0110	0.0250	0.0105
	$\ln \pi_i$	0.0130	0.0130	0.0130	0.0130	0.0205	0.0167	0.0218	0.0178
	δ_i	0.0000	0.1398	0.6262	0.1397	0.6262	0.1850	0.6262	0.2157

Table 5: Mean absolute bias and mean RMSE of estimated parameters.

References

Aten, B. H. (2017). Regional Price Parities and Real Regional Income for the United States. *Social Indicators Research*, **131** (1), 123–143.

Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications*. New York: Cambridge University Press.

Clements, K. W. and Izan, H. Y. (1981). A Note on Estimating Divisia Index Numbers. *International Economic Review*, **22** (3), 745–747.

— and — (1987). The Measurement of Inflation: A Stochastic Approach. *Journal of Business & Economic Statistics*, **5** (3), 339–350.

—, Izan, I. H. Y. and Selvanathan, E. A. (2006). Stochastic Index Numbers: A Review. *International Statistical Review*, **74** (2), 235–270.

Crompton, P. (2000). Extending the Stochastic Approach to Index Numbers. *Applied Economics Letters*, **7** (6), 367–371.

de Haan, J., Hendriks, R. and Scholz, M. (2021). Price Measurement Using Scanner Data: Time-Product Dummy Versus Time Dummy Hedonic Indexes. *Review of Income and Wealth*, **67** (2), 394–417.

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

— (2004). *On the Stochastic Approach to Linking the Regions in the ICP*. Discussion Paper 04/16, The University of British Columbia, Vancouver.

— (2005). Weighted Country Product Dummy Variable Regressions and Index Number Formulae. *Review of Income and Wealth*, **51** (4), 561–570.

Egner, U. (2019). Verbraucherpreisstatistik auf neuer Basis 2015. In *Wirtschaft und Statistik*, no. 5 in 2019, Statistisches Bundesamt, pp. 86–106.

Elzhov, T. V., Mullen, K. M., Spiess, A.-N. and Bolker, B. (2016). *minpack.lm: R Interface to the Levenberg-Marquardt Nonlinear Least-Squares Algorithm Found in MINPACK, Plus Support for Bounds*. R package version 1.2-1.

Gallant, A. R. (1975). Nonlinear Regression. *The American Statistician*, **29** (2), 73–81.

Hajargasht, G. and Rao, D. S. P. (2010). Stochastic Approach to Index Numbers for Multilateral Price Comparisons and their Standard Errors. *Review of Income and Wealth*, **56** (s1), S32–S58.

Jensen, J. L. W. V. (1906). Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica*, **30** (0), 175–193.

Kelley, C. T. (1999). *Iterative Methods for Optimization*. Society for Industrial and Applied Mathematics.

Majumder, A. and Ray, R. (2020). National and Subnational Purchasing Power Parity: A Review. *Decision*, **47** (2), 103–124.

Miller, K. S. (1981). On the Inverse of the Sum of Matrices. *Mathematics Magazine*, **54** (2), 67–72.

Moré, J. J. (1978). The Levenberg-Marquardt Algorithm: Implementation and Theory. In *Lecture Notes in Mathematics*, vol. 630, Springer Berlin Heidelberg, pp. 105–116.

Rao, D. S. P. (2004). The Country-Product-Dummy Method: A Stochastic Approach to the Computation of Purchasing Power Parities in the ICP, SSHRC International Conference on Index Number Theory and the Measurement of Prices and Productivity, Vancouver, Canada.

— (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** (4), 571–580.

— and Banerjee, K. S. (1986). A Multilateral Index Number System Based on the Factorial Approach. *Statistische Hefte*, **27** (1), 297–313.

— and Hajargasht, G. (2016). Stochastic Approach to Computation of Purchasing Power Parities in the International Comparison Program (ICP). *Journal of Econometrics*, **191** (2), 414–425.

Rokicki, B. and Hewings, G. J. D. (2019). Regional Price Deflators in Poland: Evidence from NUTS-2 and NUTS-3 Regions. *Spatial Economic Analysis*, **14** (1), 88–105.

Selvanathan, E. A. and Rao, D. S. P. (1992). An Econometric Approach to the Construction of Generalized Theil-Tornqvist Indices for Multilateral Comparisons. *Journal of Econometrics*, **54** (1), 335–346.

Summers, R. (1973). International Price Comparisons Based upon Incomplete Data. *Review of Income and Wealth*, **19** (1), 1–16.

Tabuchi, T. (2001). On Interregional Price Differentials. *Japanese Economic Review*, **52** (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** (1), 117–143.

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

— and — (2020). Anatomy of Regional Price Differentials: Evidence from Micro-Price Data. *Spatial Economic Analysis*, **15** (4), 413–440.

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

World Bank (2020). *Purchasing Power Parities and the Real Size of World Economies: Results from the 2017 International Comparison Program*. Washington DC: World Bank.