A Nonlinear Regression Approach for Spatial Price Comparisons*

Ludwig von Auer
Universität Trier

Sebastian Weinand
Deutsche Bundesbank

May 6, 2022

PRELIMINARY VERSION

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 multilateral method is still inefficient and inference is invalid. Therefore, a nonlinear generalization of the CPD method is developed. This NLCPD method is admissible on all levels of aggregation and allows for a comprehensive spatial price comparison in a single coherent step. Its root mean squared error is smaller than that of the CPD method and, in contrast to the latter, it allows for inference. The relative performance of the CPD and NLCPD methods is compared in a comprehensive simulation study. Afterwards, the NLCPD method is applied to regional price information derived from Germany's consumer price index micro data relating to May 2019.

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

JEL Classification: C43 · E31

* 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 want 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. The opinions expressed in this paper are those of the authors and do not necessarily reflect the views of the Deutsche Bundesbank, the Eurosystem, or their staff.

1 Introduction

Important areas of economic theory and economic policy are concerned with regional indicators of economic activity. Well known examples 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 the *regional price levels*, while the primary task of national statistical offices is the tracking of *intertemporal price level changes*. The latter requires a very broad sample of different products. Thus, for a basic heading like pasta, say, in different regions prices of different types of pasta are recorded. By contrast, spatial price comparisons would benefit from a more selective sample where 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 is on offer and has been applied in case studies of countries from all over the world (e.g., Majumder and Ray, 2020, pp. 111-113; 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 the 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 for such cases the country-product-dummy (CPD) method. This regression approach also allows for statistical inference. However, the CPD method implicitly assumes that the included products exhibit the same regional price dispersion. Products that are assigned to the same basic heading (e.g., pasta products) are expected to satisfy this implicit assumption. Whether this optimism 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 becomes the CPD method's assumption of a uniform price dispersion.

Accordingly, the CPD method is primarily used for the computation of the regional price levels of a given basic heading. The aggregation of the regional price levels of the various basic headings into the overall regional price levels is usually conducted by an alternative method. Therefore, the final result involves a mix of different methods.

The previous considerations raise several fundamental questions. What are the statis-

tical 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?

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 where 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, the present paper introduces the *NLCPD method*, a non-linear 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 heterogeneous 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 the use of the NLCPD method. Section 6 applies this method to a large dataset 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, services, housing) 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 1/3.

	A	B	C	D
goods	3.0	3.0	3.0	3.0
housing	3.5	5.5	6.5	10.5
services	7.0	8.5	11.5	14.5

Table 1: Prices of Goods, Services, and Housing in Four Regions.

The general price levels of the four regions can be calculated by some 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 significanctly underrated drawback. The CPD regression implicitly assumes that the products included have the same price dispersion. The prices in Tab. 1 violate this assumption. Unfortunately, a violation makes the CPD regression inefficient and inference becomes invalid (as formally shown in Appendix A.3). Even worse, the CPD regression produces biased estimates of the regional price levels (as formally shown in Appendix A.2), barring two cases that are rarely satisfied in real world measurement problems.

The inefficiency of the CPD method is quite obvious. 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

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

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:

$$\widehat{P}^{\mathrm{A}} = 0.74\,, \;\; \widehat{P}^{\mathrm{B}} = 0.92\,, \;\; \widehat{P}^{\mathrm{C}} = 1.08\,, \;\; \widehat{P}^{\mathrm{D}} = 1.36\,. \tag{2}$$

The price levels are normalized such that $\widehat{P}^{\mathrm{A}} \cdot \widehat{P}^{\mathrm{B}} \cdot \widehat{P}^{\mathrm{C}} \cdot \widehat{P}^{\mathrm{D}} = 1$.

A graphical illustration of the CPD regression is provided in the upper left panel of 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 with $\ln P^r = \ln p_i^r$. 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 specif-

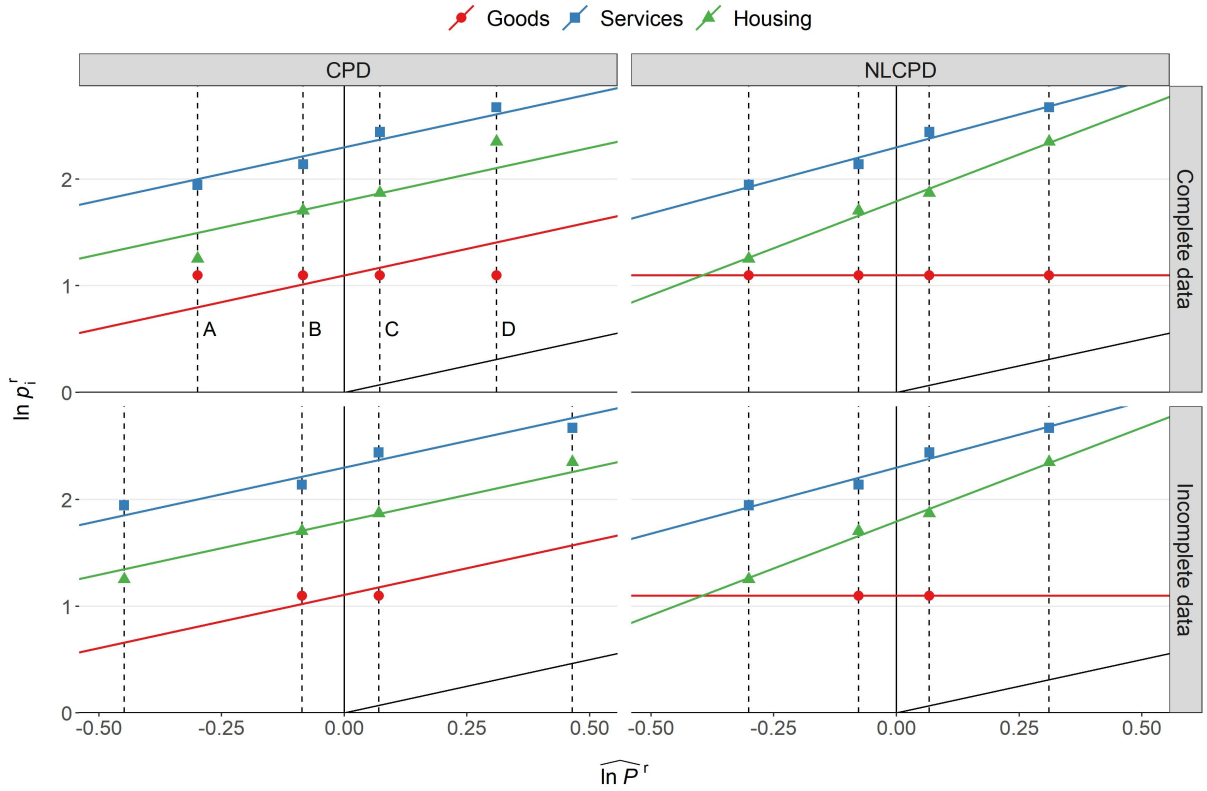

Figure 1: CPD and NLCPD regressions for the price data of Tab. 1, either with complete price data (top panels) or with missing prices for "goods" (bottom panels).

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

The CPD regression assumes that the errors, u_i^r, are independently distributed. However, the upper left panel of Fig. 1 reveals that the product-specific price dispersion causes the residuals to be both, autocorrelated and heteroskedastic. The autocorrelation 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.

Autocorrelation 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.3. Even worse, Appendix A.2 shows that the CPD regression is biased. The only exceptions are scenarios with complete price data or scenarios where the data gaps arise completely at random.

The cause of the bias is illustrated in the left panels of Fig. 1. The two outer vertical dashed lines in the upper left diagram 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 the product "goods" is observed in regions B and C, but not in regions A and D. Thus, the red circles corresponding to the latter two regions must be deleted. As a consequence, in region A a large negative disturbance vanishes. To reduce the sum of squared residuals of region A's remaining two price observations, the vertical dashed line of region A moves to the left. This effect is depicted in the lower left panel of Fig. 1. More generally, when a product with a low price dispersion is missing in the cheapest region, the estimated price level of that region always decreases below the level with complete data, that is, downward bias arises. Similarly, the missing observation in region D causes the dashed vertical line of that region to move to the right, that is, the estimated price level of region D is upward biased. Again, this can be seen in the lower left panel of Fig. 1. The corresponding price level estimates are

$$\widehat{P}^{\mathrm{A}} = 0.64\,,\ \ \widehat{P}^{\mathrm{B}} = 0.92\,,\ \ \widehat{P}^{\mathrm{C}} = 1.07\,,\ \ \widehat{P}^{\mathrm{D}} = 1.59\,.$$

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), the opposite direction of bias would arise.

All of these problems can be addressed by a simple generalization of the CPD model (1). The following section explains the basic concept, while the formal exposition is deferred to Section 4 and Appendix A.1.

3 Solution

To begin with, we consider the case of complete data with product-specific price dispersion. Then, 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 1, 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:

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

We denote this relationship as the nonlinear country-product-dummy (NLCPD) regression. The unknown values of the parameters δ_i determine the slopes of the solid lines.

In a CPD regression, all slope parameters, δ_i, are assumed to be equal to 1. The value $\delta_i = 1$ says that, in the absence of any disturbances ($u_i^r = 0$ for all i and r), each price ratio p_i^r/p_i^s ($i = 1,2,3$ and $r,s = A,B,C,D$) coincides with the ratio of the regional price levels P^r/P^s. In other words, all products exhibit the same regional price dispersion.

Economic models (e.g., Tabuchi, 2001, p. 105) as well as empirical evidence (e.g., Weinand and Auer, 2020, p. 430; Rokicki and Hewings, 2019, p. 94; Aten, 2017, p. 132-134) show that in the context of heterogeneous products the slope parameters, δ_i, should be allowed to deviate from 1. Products with $\delta_i > 1$ exhibit a stronger regional dispersion than the average of all products ("overdispersion"), while products with $\delta_i < 1$ exhibit a smaller dispersion ("underdispersion"). Products with prices that are invariant with respect to the regional price levels have a slope parameter, δ_i, close to 0. In our illustrative example, this case of underdispersion applies to the product "goods". The product "housing" seems to exhibit a strong overdispersion ($\delta_3 > 1$), while for the product "services" the slope parameter appears to be in the neighbourhood of 1 ($\delta_2 \approx 1$).

On average, the price ratios p_i^r/p_i^s must reflect the ratios of the regional price levels P^r/P^s. In Section 4 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 1, 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); further details are provided in Section 4.

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 $\widehat{\delta}_1 = 0$, $\widehat{\delta}_2 = 1.76$, and $\widehat{\delta}_3 = 1.24$. The estimated price levels are

$$\widehat{P}^{\mathrm{A}} = 0.74\,, \;\; \widehat{P}^{\mathrm{B}} = 0.93\,, \;\; \widehat{P}^{\mathrm{C}} = 1.07\,, \;\; \widehat{P}^{\mathrm{D}} = 1.36\,.$$

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 in regions A and D the prices of "goods" are missing. 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 0.056 (CPD regression) to 0.003 (NLCPD regression). Furthermore, in contrast to the CPD regression, the NLCPD method provides meaningful estimates of the standard errors of all estimated parameters, including regional price levels (formally shown in Appendix A.4). Thus, inference can be conducted.

4 Method

Multilateral price comparisons involve more than two regions. Therefore, any direct comparison between two regions should give the same price levels as an indirect comparison of these two regions via a third region. 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.

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 put into a proper regression model. In Section 4.2, the first-order conditions of this regression model as well as general formulas of the estimators are derived. The latter are 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 provides the 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 standard deviation of the error term is constant across products: $u_i^r \sim N\left(\mu = 0,\ \sigma_i = \sigma\right)$. Thus, in the CPD regression's minimization of squared residuals all observations are equally weighted. When expenditure shares or other indicators of the products' importance are available, it is recommended to use weighted least squares (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). This recommendation is usually implemented by the following assumption on the error: $u_i^r \sim N\left(\mu = 0,\ \sigma_i = \sigma/\sqrt{w_i}\ \right)$, where w_i are expenditure shares which add up to unity ($\sum_{i \in \mathcal{N}} w_i = 1$) and are uniform across regions. In the following, we apply weighted least squares.

In the CPD model (4) as well as in 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 is 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 by 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 dispersion of the regional price levels, $\ln P^s$, could be arbitrarily scaled up or down by the parameter λ.

The appropriate restriction can be easily derived from considering some pair of regions denoted by r and s. Both, the CPD method and the NLCPD method postulate that, in the absence of any disturbances, the logarithm of the price ratio p_i^s / p_i^r of a given product i

should be proportional to the logarithm of the ratio of the regional price levels of these two regions, P^r/P^s, that is,

$$\ln \frac{p_i^r}{p_i^s} = \ln \left(\frac{P^r}{P^s} \right) \delta_i \, . \tag{8}$$

This relationship should hold for all pairs of regions. The CPD method adds the assumption that δ_i is unity for all products $i \in \mathcal{N}$. In the NLCPD method no such restriction is imposed. Multiplying both sides of Eq. (8) by the expenditure weight w_i and summing over all products gives

$$\sum_{i \in \mathcal{N}} w_i \ln \frac{p_i^r}{p_i^s} = \ln \left(\frac{P^r}{P^s} \right) \sum_{i \in \mathcal{N}} w_i \delta_i \, . \tag{9}$$

Since the expenditure weights are assumed to be uniform across regions, the left-hand side of Eq. (9) is identical to the log-change index formulas of Törnqvist, Walsh-Vartia, and Sato-Vartia (e.g., Auer and Shumskikh, 2022, p. 6). To ensure that the NLCPD regression is consistent with these index formulas, the right-hand side of Eq. (9) must simplify to $\ln(P^r/P^s)$. Thus, the appropriate restriction on the δ_i-values is

$$\sum_{i \in \mathcal{N}} w_i \delta_i = 1 \, . \tag{10}$$

Note that the CPD model (4) satisfies this restriction by assumption ($\delta_i = 1$ for all $i \in \mathcal{N}$). By contrast, the NLCPD model (5) provides estimates for δ_i which have to satisfy restriction (10).

A very intuitive justification for restriction (10) arises when the term $\delta_i \ln P^r$ is interpreted as region r's product-specific logarithmic price level (not to be confused with the observed logarithmic price $\ln p_i^r$). Consequently, the weighted average of region r's product-specific logarithmic price levels should yield region r's overall price level:

$$\sum_{i \in \mathcal{N}} w_i \delta_i \ln P^r = \ln P^r \, .$$

To make this postulate a valid identity, restriction (10) is required.

This restriction 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), also one of the $\ln P^r$-values must be residually derived. Again, any product i and any region r can be chosen for this purpose. If product $i = 1$ and region $r = 1$ are selected, the NLCPD regression model (5) becomes

$$\ln p_i^r = \sum_{j \in \mathcal{N}} G_j \ln \pi_j + \left(\frac{G_1}{w_1} + \sum_{j \in \mathcal{N} \setminus \{1\}} \widetilde{G}_j \delta_j \right) \sum_{s \in \mathcal{R} \setminus \{1\}} \widetilde{D}^s \ln P^s + u_i^r \, , \tag{11}$$

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 by $\delta_1 = \left(1 - \sum_{j \in \mathcal{N} \setminus \{1\}} w_j \delta_j \right) / w_1$ 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 (11) 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 following, we derive the NLCPD method's weighted least squares estimators $\widehat{\ln \pi_i}$, $\widehat{\delta}_i$ and $\widehat{\ln P^r}$, and 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

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

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 where product i is priced. The set's number of products is denoted by R_i.

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 non-linear least squares approach we apply normalization (6) as well as restriction (10). As a consequence, one $\widehat{\delta}_i$-value as well as one $\widehat{\ln P^r}$-value cannot be used in the minimization, but is residually derived. The first-order conditions are

$$\frac{\partial S_{\hat{u}_i^r \hat{u}_i^r}}{\partial \widehat{\ln \pi_i}} = \sum_{r \in \mathcal{R}_i} w_i 2 \left(\ln p_i^r - \widehat{\delta}_i \widehat{\ln P^r} - \widehat{\ln \pi_i} \right) (-1) = 0 \tag{12a}$$

$$\frac{\partial S_{\hat{u}_i^r \hat{u}_i^r}}{\partial \widehat{\delta}_i} = \sum_{r \in \mathcal{R}_i} w_i 2 \left(\ln p_i^r - \widehat{\delta}_i \widehat{\ln P^r} - \widehat{\ln \pi_i} \right) \left(-\widehat{\ln P^r} \right) = 0 \tag{12b}$$

$$\frac{\partial S_{\hat{u}_i^r \hat{u}_i^r}}{\partial \widehat{\ln P^r}} = \sum_{i \in \mathcal{N}_r} w_i 2 \left(\ln p_i^r - \widehat{\delta}_i \widehat{\ln P^r} - \widehat{\ln \pi_i} \right) \left(-\widehat{\delta}_i \right) = 0 \ . \tag{12c}$$

Condition (12a) gives

$$\widehat{\ln \pi_i} = \frac{1}{R_i} \sum_{r \in \mathcal{R}_i} \left(\ln p_i^r - \widehat{\delta}_i \widehat{\ln P^r} \right) \ . \tag{13}$$

As the product weights are identical across regions, each region receives the same weight in the summation. Inserting the restriction $\widehat{\delta}_i = 1$ $(i \in \mathcal{N})$ in the NLCPD estimator (13), gives the corresponding CPD estimator:

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

where $\widehat{\ln P^{r\prime}}$ is the CPD estimator of the regional price levels as defined in Eq. (17), below.

For a product i that is priced in all regions ($R_i = R$), the estimator (13) simplifies to

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

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

Condition (12b) yields

$$\hat{\delta}_i = \frac{\sum\limits_{r\in\mathcal{R}_i} \widehat{\ln P^r}\left(\ln p_i^r - \widehat{\ln \pi_i}\right)}{\sum\limits_{r\in\mathcal{R}_i}\left(\widehat{\ln P^r}\right)^2} \ . \tag{15}$$

The numerator is the covariation (across regions) of the logarithmic regional price level, $\widehat{\ln P^r}$, 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 (15) 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, $\hat{\delta}_i$. If product i has a uniform price in all R regions, then $\widehat{\ln \pi_i} = \ln p_i^r$ and, therefore, the estimator (15) gives $\hat{\delta}_i = 0$. However, if this product has been observed only in $R_i < R$ regions, $\widehat{\ln \pi_i}$ and $\ln p_i^r$ can differ and $\hat{\delta}_i \neq 0$ can arise.

It can be shown that, for complete price data, formula (15) can be written in the form

$$\hat{\delta}_i = \frac{\sum\limits_{r\in\mathcal{R}} \widehat{\ln P^r} \ln p_i^r}{\sum\limits_{r\in\mathcal{R}} \widehat{\ln P^r} \cdot \sum\limits_{j\in\mathcal{N}} w_j \ln p_j^r} \ .$$

If the numerator is larger than the denominator, the prices of product i exhibit a stronger positive correlation with the regional price levels than the weighted average of the prices of all products. Therefore, we get $\hat{\delta}_i > 1$.

Condition (12c) can be rewritten as

$$\widehat{\ln P^r} = \frac{\sum\limits_{i\in\mathcal{N}_r} w_i \hat{\delta}_i \left(\ln p_i^r - \widehat{\ln \pi_i}\right)}{\sum\limits_{i\in\mathcal{N}_r} w_i \left(\hat{\delta}_i\right)^2} \ . \tag{16}$$

The numerator is the covariation (across products) of $\left(\ln p_i^r - \widehat{\ln \pi_i}\right)$ and the spread parameter $\hat{\delta}_i$. The denominator is the variation (across products) of $\hat{\delta}_i$. 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 (16) simplifies to the corresponding CPD estimator:

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

When all products i are priced in region r, we get $\sum_{i \in \mathcal{N}_r} w_i = 1$ and the resulting estimator (17) simplifies to the well known CPD formula (e.g., Rao, 2005, p. 577; Rao and Hajargasht, 2016, p. 417):

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

The nonlinear least squares formulas (13), (15) and (16) 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.[1] 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

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

(10) 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 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. (15) 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 (10). 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 Standard errors

In nonlinear regression models, approximations of the standard errors can be computed from the Jacobian matrix evaluated at final parameter estimates. This computation is documented in Appendix A.4. When the data set is complete, the approximated standard error of the NLCPD estimator $\widehat{\ln \pi_i}$ is

$$\widehat{se}\left(\widehat{\ln \pi_i}\right) = \widehat{\sigma} \sqrt{\frac{1}{R w_i}}\,, \tag{18}$$

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 by 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.3 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

$$\widehat{se}\left(\widehat{\delta}_i\right) = \widehat{\sigma}\sqrt{\frac{1}{\sum\limits_{r\in\mathcal{R}}\left(\widehat{\ln P^r}\right)^2}\left(\frac{1-w_i}{w_i}+\left(\widehat{\delta}_i-1\right)^2\right)}\,. \tag{19}$$

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 moves away from one.

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

$$\widehat{se}\left(\widehat{\ln P^r}\right) = \widehat{\sigma}\sqrt{\frac{1}{\sum\limits_{i\in\mathcal{N}} w_i\left(\widehat{\delta}_i\right)^2}\left(\frac{R-1}{R}+\left(\sum_{i\in\mathcal{N}} w_i\left(\widehat{\delta}_i\right)^2-1\right)\frac{\left(\widehat{\ln P^r}\right)^2}{\sum\limits_{s\in\mathcal{R}}\left(\widehat{\ln P^s}\right)^2}\right)}\,. \tag{20}$$

From restriction (10) and Jensen's (1906) inequality we know that

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

Thus, the root term in Eq. (20) 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, $\widehat{se'}\left(\widehat{\ln P^{r\prime}}\right) = \widehat{\sigma}'\sqrt{(R-1)/R}$, is biased because $\widehat{\sigma}'$ is biased.

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. 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.

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, for each product i true values of the parameters $\ln \pi_i$ and δ_i exists.

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 from $\ln \widetilde{P}^r$ the average preliminary price level of all regions 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%.

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 weights do not vary across regions. In the present context, they represent expenditure shares.

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 product prices to be greater zero while its positive skewness makes very expensive products occur less frequently.

The regional price dispersion of the products, δ_i, is expected to vary but should be 1 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.

In our simulation, we differentiate between three scenarios. The first two scenarios serve rather as a reference while the third scenario probably is the most realistic one.

Scenario 1: In the first scenario, we assume that the price data are complete, that is, there is exactly one price per product and region. This gives $NR = 300$ observations and the share of missing prices is equal to 0.

Scenario 2: In the second scenario, we assume that every third price is missing. This gives a total of 200 observations. The missing prices are chosen completely

at random. All other parameters are the same as in the first scenario.

Scenario 3: In the third scenario, we keep the setting of the second scenario but introduce the missing prices in a systematic manner: the larger 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 by formula (15).

We repeat these three steps $L = 2000$ times (with iterations $l = 1, 2, \ldots, L$) and obtain for each region r a set of 2000 $\widehat{\ln P^r}$-values for the CPD method and 2000 $\widehat{\ln P^r}$-values for the NLCPD method. Afterwards, we compare the performance of the two methods. To this end, we compute from the NLCPD results of the L iterations 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:

$$\text{Bias}\left(\widehat{\ln P}\right) = \frac{1}{R}\sum_{r\in\mathcal{R}}\left|\text{Bias}\left(\widehat{\ln P^r}\right)\right| = \frac{1}{R}\sum_{r\in\mathcal{R}}\left|\frac{1}{L}\sum_{l=1}^{L}\left(\widehat{\ln P_l^r} - \ln P_l^r\right)\right| \tag{21a}$$

$$\text{RMSE}\left(\widehat{\ln P}\right) = \frac{1}{R}\sum_{r\in\mathcal{R}}\text{RMSE}\left(\widehat{\ln P^r}\right) = \frac{1}{R}\sum_{r\in\mathcal{R}}\sqrt{\frac{1}{L}\sum_{l=1}^{L}\left(\widehat{\ln P_l^r} - \ln P_l^r\right)^2}\,, \tag{21b}$$

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 the simulation Scenarios 1 and 2, 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 3, $\ln P^r$-estimates of the CPD method are expected to be biased. 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 a higher accuracy. Consequently, the RMSE should be lower for the NLCPD method in all three simulation scenarios.

5.2 Discussion of results

The simulation results for the mean absolute bias and the mean RMSE of the $\ln P^r$-estimates can be found in Tab. 2.[2] 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 (see Scenarios 1 and 2 in Tab. 2). 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 (see Scenario 3 in Tab. 2).

	Scenario 1		Scenario 2		Scenario 3	
	CPD	NLCPD	CPD	NLCPD	CPD	NLCPD
Bias	0.0002	0.0001	0.0003	0.0002	0.0133	0.0002
RMSE	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, $\widehat{\ln P^{r\prime}}$.

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 all three simulation scenarios 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.

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 depicts the bias and RMSE for each region r on the horizontal axes. The regions 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 are unbiased as long as the data are complete or missing completely on random (Scenarios 1 and 2), but that the CPD method is biased when the data gaps are systematic (see the red dots in Scenario 3). More specifically, the more a region's true price level deviates from the average price level of all regions, the larger the bias. As predicted in Section 3, 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 the simulation Scenario 3 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

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

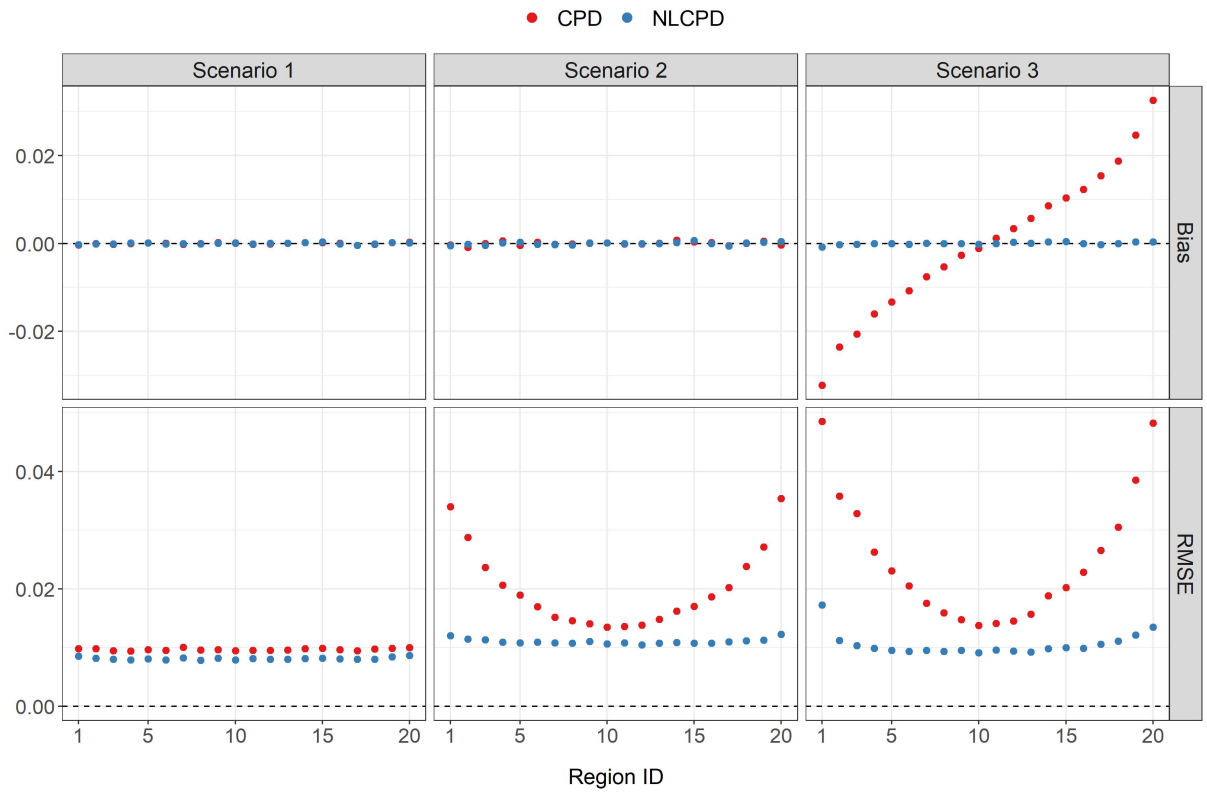

Figure 2: Bias and RMSE of the NLCPD estimates, $\widehat{\ln P^r}$, and the CPD estimates, $\widehat{\ln P^{r\prime}}$, for the three simulation scenarios.

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 3, the blue dots remain close to the horizontal baseline.

The NLCPD method outperforms the CPD method also with respect to the RMSE. 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 1), the advantage of the NLCPD method does not depend on a region's true price level. However, when data gaps occur (Scenarios 2 and 3), 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.

6 Empirical application

In the following, we apply the NLCPD method to regional price levels above the basic heading level, compiled from German official 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.[3]

6.1 Price data and aggregation approach

We have the privilege to work with micro price data of the German consumer price index (CPI) of May 2019. These data were provided to us by the Research Data Center of the Federal Statistical Office and Statistical Offices of the Länder. In total, the data contain more than 400,000 price observations for goods, services and rents which were collected in the 401 regions of Germany (294 counties and 107 cities). 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).

Due to methodological reasons, 70 basic headings with centrally collected prices cannot be exploited in a regional analysis.[4] They represent a combined expenditure weight of 13.44%. 36 other basic headings with a combined weight of 1.45% were too fragmentary

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 (in %, base year 2015). Usable and unusable weights add up to 100%. Source: Research Data Centre of the Federal Statistical Office and Statistical Offices of the Länder, CPI, May 2019; authors' own computations.

[3] The price index numbers of the German regions are available upon request.

[4] 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 when entering the micro data set.

to convey useful information for the interregional price comparison.[5] This leaves us with 677 basic headings the price information of which 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 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, the CPD method is applied to each basic heading. However, it is worthwhile to mention some improvements and modifications that we implement.

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 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.

The prices of fuels collected by the Federal Statistical Office represent four different basic headings. We replace them by two basic headings computed from a full sample, which was collected by the German Market Transparency Unit for Fuels in May 2019.[6]

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 are proportionally assigned to these 618 basic headings. This set

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

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

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 do 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 price level estimates of the CPD and NLCPD methods are highly correlated (Pearson correlation: 0.97). However, 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 of the number of data gaps and the NLCPD's estimates $\widehat{\delta}_i$ is -0.13.

The distributions of the estimated logarithmic regional price levels, $\widehat{\ln P^r}$ and $\widehat{\ln P^{r\prime}}$, are depicted in the left panel of Fig. 3. By definition, the average logarithmic price level of both methods is zero. The median price level is negative, indicating a positive skewness of the price level estimates. This effect is more pronounced for the CPD method than for the NLCPD method.

Also the estimates of $\ln \pi_i$ of the two methods are highly correlated (Pearson correlation: 0.96). The right panel of Fig. 3 shows that the lower bound of the range is similar, while the

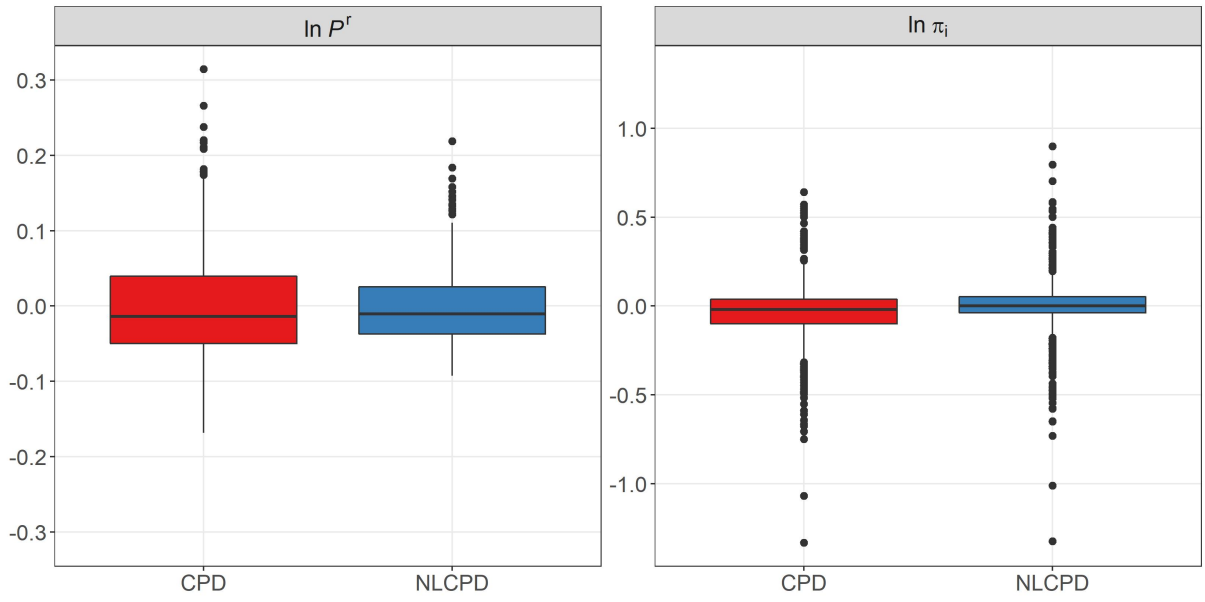

Figure 3: CPD and NLCPD estimates of the regional price levels, $\ln P^r$, and the general values of the products, $\ln \pi_i$.

upper bound differs. 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.

In order to transform the estimates of $\ln P^r$ into a regional price index they are expressed in relation to their population weighted average. For the NLCPD method the transformation is

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

where $\ln P^{\mathrm{Ger}} = \sum_{r=1}^{401} g^r \, \widehat{\ln P^r}$ and g^r is the population share of region r. The same transformation is applied to the CPD price level estimates.

Summary statistics of the price index numbers are reported in Tab. 4. When the NLCPD method is applied, the price level of the cheapest region is by $89.2/100 = 10.8\%$ below the population weighted average. The most expensive region exceeds that average by 21.8%. The spread between the most expensive and the cheapest region is $121.8/89.2 = 36.5\%$. These numbers are more pronounced for the CPD method, resulting in a regional price spread of 61.9%. As can be seen from Tab. 4, 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.

	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 the population weighted mean (= 100).

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. It's 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 neighbouring Munich, is the most expensive region in Germany.

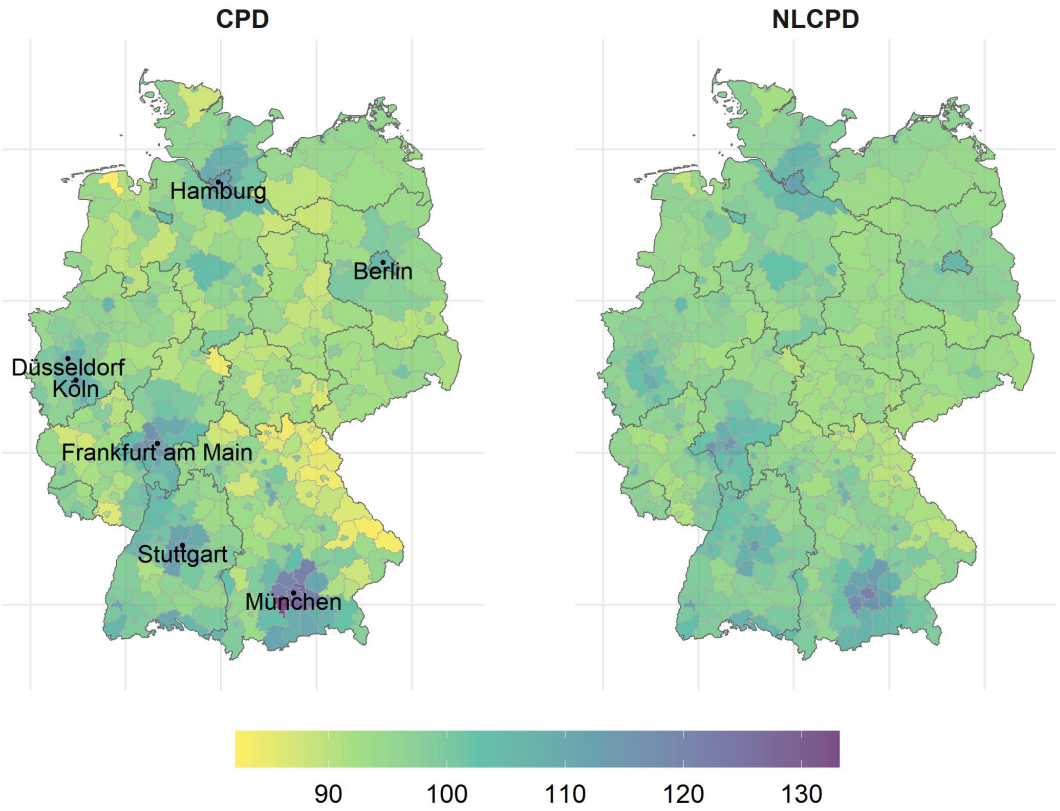

Figure 4: Price index numbers by CPD and NLCPD methods, each in relation to its population weighted average (= 100).

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 mandatory for the CPD method that the regional price dispersion of the various products is 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 amends the CPD method by parameters that capture the product-specific price dispersion. Their estimates indicate whether the CPD assumption of uniform price dispersion would have been reasonable. For all of the NLCPD estimators, the formulas for their estimated standard deviations have been derived. In a simulation, the deficiencies of the CPD method and the superiority of the NLCPD method has been shown. Finally, in a price level comparison of the 401 German regions, the practical applicability of the NLCPD method has been 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 may 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 coincides with the dispersion of the overall regional price levels (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 and GEKS-Jevons index number methods 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 both methods still coincide when all regions have the same distribution of expenditure shares (as in Tab. 1). 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 apply also to the other one. This is a relevant question, because not only the CPD method but also the GEKS-Jevons method 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 and Weinand and Auer, 2020, pp. 416-418). However, a careful analysis of this question must be left for future research.

A Mathematical derivations

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

A.1 The NLCPD Model and Special Cases

Model (11) can be written as

$$\breve{\boldsymbol{y}} = \breve{\boldsymbol{G}}\boldsymbol{\pi} + \left(\frac{G_1}{w_1} + \widetilde{\boldsymbol{G}}\boldsymbol{\delta}\right) \odot \left(\breve{\boldsymbol{D}}\boldsymbol{p}\right) + \breve{\boldsymbol{u}} \;, \tag{A.1}$$

where $\breve{\boldsymbol{D}} = \left(\widetilde{D}^2 \;\; \ldots \;\; \widetilde{D}^R\right)$, $\breve{\boldsymbol{G}} = (G_1 \;\; \ldots \;\; G_N)$ 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 error terms, u_i^r, respectively. The parameters are $\boldsymbol{\pi} = (\ln \pi_1 \;\; \ldots \;\; \ln \pi_N)^\intercal$, $\boldsymbol{p} = \left(\ln P^2 \;\; \ldots \;\; \ln P^R\right)^\intercal$, and $\boldsymbol{\delta} = (\delta_2 \;\; \ldots \;\; \delta_N)^\intercal$, 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 $\left(\breve{\boldsymbol{D}}\boldsymbol{p}\right)$. 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 $\breve{\boldsymbol{H}} = \begin{pmatrix} H^2 & \dots & H^R \end{pmatrix}$ 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.1) in the following linear form:

$$\breve{\boldsymbol{y}} = \breve{\boldsymbol{G}}\boldsymbol{\pi} + \breve{\boldsymbol{H}}\boldsymbol{p} + \breve{\boldsymbol{u}}\,. \tag{A.2}$$

It is assumed that the variance of the errors, u_i^r, is

$$\sigma_i^2 = \sigma^2 / \sqrt{w_i}\ . \tag{A.3}$$

Thus, a weighted least squares approach should be applied. To this end, we define for each product i a diagonal $(R_i \times R_i)$-matrix of weights, $\boldsymbol{W}_i = \operatorname{diag}\left(\sqrt{w_i} \ \dots \ \sqrt{w_i}\right)$, and combine them in the diagonal $(B \times B)$-matrix

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

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

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

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

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

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 estimation of model (A.2) with the weighting matrix $\boldsymbol{W}$ is equivalent to ordinary least squares estimation 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 $\boldsymbol{u} = \boldsymbol{W}\breve{\boldsymbol{u}}$. All entries of $\boldsymbol{u}$, $\sqrt{w_i}u_i^r$, are identically and independently distributed with $\sqrt{w_i}u_i^r \sim N\left(0, \sigma^2\right)$.

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

$$\widehat{\boldsymbol{\pi}} = \begin{pmatrix} \widehat{\ln \pi_1} & \ldots & \widehat{\ln \pi_N} \end{pmatrix}^\intercal \qquad \text{and} \qquad \widehat{\boldsymbol{p}} = \begin{pmatrix} \widehat{\ln P^2} & \ldots & \widehat{\ln P^R} \end{pmatrix}^\intercal$$

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

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

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

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

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,

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

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

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

Furthermore,

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

When the data set is complete, some additonal results can be derived. When product i is priced in all regions, we get $\boldsymbol{C}_i = \mathbf{1}_{(R-1)\times(R-1)}$. Then,

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

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

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

Furthermore,

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

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,

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

Concerning the two components of $(\boldsymbol{D}^\intercal \boldsymbol{M} \boldsymbol{D})^{-1}$ 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

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

because $\breve{\boldsymbol{D}}_i \mathbf{1}_{R\times R} = \mathbf{0}_{(R-1)\times(R-1)}$. Thus,

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

Furthermore,

$$\boldsymbol{D}^\intercal \boldsymbol{M} \boldsymbol{H} = \boldsymbol{I}_{R-1} + \mathbf{1}_{(R-1)\times(R-1)} \tag{A.25}$$

$$= \boldsymbol{D}^\intercal \boldsymbol{D}\,, \tag{A.26}$$

where we exploited the restriction $\sum_{i\in\mathcal{N}} w_i \delta_i = 1$. (A.24) and (A.26) imply that

$$\boldsymbol{D}^\intercal \boldsymbol{M} = \boldsymbol{D}^\intercal \qquad \text{and} \qquad \boldsymbol{M}\boldsymbol{D} = \boldsymbol{M}\boldsymbol{H} = \boldsymbol{D}\,. \tag{A.27}$$

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

$$\boldsymbol{G}^\intercal \boldsymbol{D} = \mathbf{0}_{N\times NR} \qquad \text{and} \qquad \boldsymbol{D}^\intercal \boldsymbol{G} = \mathbf{0}_{NR\times N}\,. \tag{A.28}$$

As a consequence,

$$\boldsymbol{G}^\intercal \boldsymbol{K} = \boldsymbol{G}^\intercal - \boldsymbol{G}^\intercal \boldsymbol{D} \left(\boldsymbol{D}^\intercal \boldsymbol{D}\right)^{-1} \boldsymbol{D}^\intercal = \boldsymbol{G}^\intercal\,. \tag{A.29}$$

Analogously, we have

$$\boldsymbol{G}^\intercal \boldsymbol{H} = \mathbf{0}_{N\times NR} \qquad \text{and} \qquad \boldsymbol{H}^\intercal \boldsymbol{G} = \mathbf{0}_{NR\times N} \tag{A.30}$$

and

$$\boldsymbol{G}^\intercal \boldsymbol{L} = \boldsymbol{G}^\intercal\,. \tag{A.31}$$

A.2 Bias of the CPD Estimators

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

$$\boldsymbol{y} = \boldsymbol{G}\boldsymbol{\pi} + \boldsymbol{D}\boldsymbol{p} + \boldsymbol{u}'\,. \tag{A.32}$$

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.32) is given by

$$E\left(\boldsymbol{u}'\right) = (\boldsymbol{H} - \boldsymbol{D})\,\boldsymbol{p}\,. \tag{A.33}$$

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

$$\widehat{\boldsymbol{p}} = \boldsymbol{p} + \left(\boldsymbol{H}^\intercal \boldsymbol{M} \boldsymbol{H}\right)^{-1} \boldsymbol{H}^\intercal \boldsymbol{M} \boldsymbol{u} \tag{A.34}$$

and

$$\widehat{\boldsymbol{\pi}} = \boldsymbol{\pi} + \left(\boldsymbol{G}^\intercal \boldsymbol{L} \boldsymbol{G}\right)^{-1} \boldsymbol{G}^\intercal \boldsymbol{L} \boldsymbol{u}\,. \tag{A.35}$$

Since $E(\boldsymbol{u}) = \mathbf{0}_{B\times 1}$, we get $E(\widehat{\boldsymbol{p}}) = \boldsymbol{p}$ and $E(\widehat{\boldsymbol{\pi}}) = \boldsymbol{\pi}$. Thus, the estimators (A.34) and

(A.35) would be unbiased.

The estimators of the CPD model (A.32) can be written in the form

$$\widehat{\boldsymbol{p}}' = \boldsymbol{p} + (\boldsymbol{D}^\intercal \boldsymbol{M} \boldsymbol{D})^{-1} \boldsymbol{D}^\intercal \boldsymbol{M} \boldsymbol{u}'$$

and

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

Inserting (A.33) and taking expectations yields

$$E\left(\widehat{\boldsymbol{p}}'\right) = (\boldsymbol{D}^\intercal \boldsymbol{M} \boldsymbol{D})^{-1} \boldsymbol{D}^\intercal \boldsymbol{M} \boldsymbol{H} \boldsymbol{p} \tag{A.36}$$

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:

$$(\boldsymbol{D}^\intercal \boldsymbol{M} \boldsymbol{D})^{-1} \boldsymbol{D}^\intercal \boldsymbol{M} \boldsymbol{H} = \boldsymbol{I}_{R-1} \,. \tag{A.37}$$

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.29) and (A.30) 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,

$$\boldsymbol{D}^\intercal \boldsymbol{M} \boldsymbol{D} = \boldsymbol{I}_{R-1} + \left(1 - \frac{w_j R}{R-1} \right) \boldsymbol{1}_{(R-1)\times(R-1)} \,. \tag{A.38}$$

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,

$$\boldsymbol{D}^{\intercal}\boldsymbol{M}\boldsymbol{H} = \boldsymbol{I}_{R-1} + \left(1 - \frac{w_j \delta_j R}{R-1}\right) \mathbf{1}_{(R-1)\times(R-1)} \,. \tag{A.39}$$

Inserting (A.38) and (A.39) in (A.36) yields

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

Rule (A.18) implies that

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

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 $E\left(\widehat{\boldsymbol{p}}'\right)$ we get

$$E\left(\widehat{\ln P^{r\prime}}\right) = \ln P^r + \frac{w_j\,(1-\delta_j)}{(1-w_j)\,(R-1)} \sum_{s \in \mathcal{R}\setminus\{1\}} \ln P^s \qquad \text{for } r = 2, ..., R \,. \tag{A.40}$$

For $\delta_j = 1$, the quotient in (A.40) is equal to 0 and we get $E\left(\widehat{\boldsymbol{p}}'\right) = \boldsymbol{p}$. 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 upward biased. 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 downward biased. 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.3 Inference in the CPD Method

When at least one δ_i-value is different from 1 and observations are (non-randomly) missing, the weighted CPD estimator $\widehat{\boldsymbol{p}}'$ is biased and inference is invalid, anyway. Therefore, one can restrict the following analysis to the case of complete data.

Exploiting (A.11), (A.27), and (A.29), 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

$$\widehat{\boldsymbol{u}}' = \boldsymbol{N}\boldsymbol{y}\,, \tag{A.41}$$

with

$$\boldsymbol{N} = \boldsymbol{I}_{NR} - \boldsymbol{G}\left(\boldsymbol{G}^\intercal\boldsymbol{G}\right)^{-1}\boldsymbol{G}^\intercal - \boldsymbol{D}\left(\boldsymbol{D}^\intercal\boldsymbol{D}\right)^{-1}\boldsymbol{D}^\intercal\,. \tag{A.42}$$

Relationships (A.28) and (A.30) imply that

$$\boldsymbol{N}\boldsymbol{D} = \boldsymbol{N}\boldsymbol{G} = \boldsymbol{0}_{NR\times NR}\,. \tag{A.43}$$

Furthermore, we know from (A.27) that $\boldsymbol{D}^\intercal\boldsymbol{H} = \boldsymbol{D}^\intercal\boldsymbol{M}\boldsymbol{H} = \boldsymbol{D}^\intercal\boldsymbol{D}$. Thus,

$$\boldsymbol{N}\boldsymbol{H} = \boldsymbol{H} - \boldsymbol{D}\,. \tag{A.44}$$

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,

$$E\left(\boldsymbol{u}'\boldsymbol{u}'^\intercal\right) = \boldsymbol{N}\boldsymbol{H}\boldsymbol{p}\boldsymbol{p}^\intercal\boldsymbol{H}^\intercal\boldsymbol{N} + \sigma^2\boldsymbol{I}_{NR}\,, \tag{A.45}$$

where σ^2 is the variance used in (A.3).

Since (A.27) to (A.31) apply, the variance-covariance matrix of the CPD estimators $\widehat{\boldsymbol{p}}'$ is

$$V\left(\widehat{\boldsymbol{p}}'\right) = \left(\boldsymbol{D}^\intercal\boldsymbol{D}\right)^{-1}\boldsymbol{D}^\intercal E\left(\boldsymbol{u}'\boldsymbol{u}'^\intercal\right)\boldsymbol{D}\left(\boldsymbol{D}^\intercal\boldsymbol{D}\right)^{-1}\,. \tag{A.46}$$

Inserting expression (A.45) in (A.46) and using (A.43) yields

$$V\left(\widehat{\boldsymbol{p}}'\right) = \sigma^2\left(\boldsymbol{D}^\intercal\boldsymbol{D}\right)^{-1}\,, \tag{A.47}$$

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

Using (A.29), the variance-covariance matrix of the CPD estimators $\widehat{\boldsymbol{\pi}}'$ is

$$V\left(\widehat{\boldsymbol{\pi}}'\right) = \left(\boldsymbol{G}^\intercal\boldsymbol{G}\right)^{-1}\boldsymbol{G}^\intercal E\left(\boldsymbol{u}'\boldsymbol{u}'^\intercal\right)\boldsymbol{G}\left(\boldsymbol{G}^\intercal\boldsymbol{G}\right)^{-1}\,. \tag{A.48}$$

Inserting expression (A.45) in (A.48) and using (A.43) yields

$$V\left(\widehat{\boldsymbol{\pi}}'\right) = \sigma^2 \left(\boldsymbol{G}^{\intercal}\boldsymbol{G}\right)^{-1} . \tag{A.49}$$

Substituting in (A.47) and (A.49) the variance σ^2 by its CPD estimator,

$$(\widehat{\sigma}')^2 = \frac{\widehat{\boldsymbol{u}}'^{\intercal}\widehat{\boldsymbol{u}}'}{NR-N-R+1} , \tag{A.50}$$

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

$$E\left(\widehat{\boldsymbol{u}}'^{\intercal}\widehat{\boldsymbol{u}}'\right) = \sigma^2 \left(NR-N-R+1\right) . \tag{A.51}$$

Substituting in (A.41) the vector $\boldsymbol{y}$ by $(\boldsymbol{G\pi}+\boldsymbol{Dp}+\boldsymbol{u}')$ 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.45) yields

$$E\left(\widehat{\boldsymbol{u}}'^{\intercal}\widehat{\boldsymbol{u}}'\right) = \operatorname{tr}\left(\boldsymbol{NHpp}^{\intercal}\boldsymbol{H}^{\intercal}\boldsymbol{N}\right) + \operatorname{tr}\left(\sigma^2\boldsymbol{N}\right) . \tag{A.52}$$

Note that

$$\operatorname{tr}\left(\sigma^2\boldsymbol{N}\right) = \sigma^2\left(NR-N-R+1\right) , \tag{A.53}$$

where we exploited the results (A.15) and (A.21). Thus, unbiasedness requires that in (A.52)

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

$$
\begin{aligned}
&(\boldsymbol{H}-\boldsymbol{D})\,\boldsymbol{p}\boldsymbol{p}^{\intercal}\,(\boldsymbol{H}^{\intercal}-\boldsymbol{D}^{\intercal}) \\
&= \left(\sum_{r\in\mathcal{R}\setminus\{1\}} (\ln P^r)^2\right)
\begin{bmatrix}
(\delta_1{-}1)\,(\delta_1{-}1)\,\sqrt{w_1 w_1}\breve{\boldsymbol{D}}_1\breve{\boldsymbol{D}}_1^{\intercal} & \dots & (\delta_1{-}1)\,(\delta_N{-}1)\,\sqrt{w_1 w_N}\breve{\boldsymbol{D}}_1\breve{\boldsymbol{D}}_N^{\intercal} \\
\vdots & \ddots & \vdots \\
(\delta_N{-}1)\,(\delta_1{-}1)\,\sqrt{w_N w_1}\breve{\boldsymbol{D}}_N\breve{\boldsymbol{D}}_1^{\intercal} & \dots & (\delta_N{-}1)\,(\delta_N{-}1)\,\sqrt{w_N w_N}\breve{\boldsymbol{D}}_N\breve{\boldsymbol{D}}_N^{\intercal}
\end{bmatrix},
\end{aligned}
$$

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

$$
\breve{\boldsymbol{D}}_i\breve{\boldsymbol{D}}_j^{\intercal} =
\begin{bmatrix}
R-1 & -1 & -1 & \cdots & -1 \\
-1 & 1 & 0 & \cdots & 0 \\
-1 & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
-1 & 0 & 0 & \cdots & 1
\end{bmatrix}
\qquad \forall i,j\in\mathcal{N}\,.
$$

Thus, $\mathrm{tr}\left(\breve{\boldsymbol{D}}_i\breve{\boldsymbol{D}}_i^{\intercal}\right) = (R-1)+(R-1) = 2\,(R-1)$ and

$$
\mathrm{tr}\left((\boldsymbol{H}-\boldsymbol{D})\,\boldsymbol{p}\boldsymbol{p}^{\intercal}\,(\boldsymbol{H}-\boldsymbol{D})^{\intercal}\right) = 2\,(R-1)\left(\sum_{r\in\mathcal{R}\setminus\{1\}} (\ln P^r)^2\right)\sum_{i\in\mathcal{N}} (\delta_i-1)^2\,w_i\,. \tag{A.54}
$$

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 $i\in\mathcal{N}$.

A.4 Inference in the NLCPD Model

In the following, it is assumed that the data set is complete. The NLCPD regression model was given in (A.1). The estimated model is

$$
\widehat{\breve{\boldsymbol{y}}} = \breve{\boldsymbol{G}}\widehat{\boldsymbol{\pi}} + \left(\frac{G_1}{w_1} + \widetilde{\boldsymbol{G}}\widehat{\boldsymbol{\delta}}\right)\odot\left(\breve{\boldsymbol{D}}\widehat{\boldsymbol{p}}\right)\,. \tag{A.55}
$$

One can write this model in the following more compact form:

$$
\widehat{\breve{\boldsymbol{y}}} = \breve{\boldsymbol{G}}\widehat{\boldsymbol{\pi}} + \widehat{\boldsymbol{H}}\widehat{\boldsymbol{p}}\,,
$$

with

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

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

The Jacobian matrix, $\boldsymbol{J}$, can be computed from the estimated regression model (A.55). This $(NR) \times (2N + R - 2)$-matrix has three submatrices. The first one is formed by the columns related to the derivatives with respect to $\widehat{\ln \pi_i}$. This submatrix is equal to $\breve{\boldsymbol{G}}$. The second submatrix is formed by the columns related to the derivatives with respect to $\widehat{\delta}_i$. Defining the diagonal $(R \times R)$-matrix

$$\widehat{\boldsymbol{P}} = \operatorname{diag}\left(-\widehat{\boldsymbol{p}}\, \boldsymbol{1}_{(R-1)\times 1} \quad \widehat{\boldsymbol{p}}^{\intercal}\right) ,$$

they can be written as $\left(\boldsymbol{I}_N \otimes \widehat{\boldsymbol{P}}\right)\widetilde{\boldsymbol{G}}$, where $\otimes$ denotes the Kronecker product. Note that $-\widehat{\boldsymbol{p}}\, \boldsymbol{1}_{(R-1)\times 1}$ is the residually computed value $\widehat{\ln P^1}$ and that

$$\left(\boldsymbol{I}_N \otimes \widehat{\boldsymbol{P}}\right)\widetilde{\boldsymbol{G}} = \begin{bmatrix} \widehat{\boldsymbol{P}}\widetilde{\boldsymbol{G}}_1 \\ \widehat{\boldsymbol{P}}\widetilde{\boldsymbol{G}}_2 \\ \vdots \\ \widehat{\boldsymbol{P}}\widetilde{\boldsymbol{G}}_N \end{bmatrix} = \begin{bmatrix} \widehat{\boldsymbol{P}}\widetilde{\boldsymbol{G}}_1 \\ \widehat{\boldsymbol{P}}\breve{\boldsymbol{G}}_2 \\ \vdots \\ \widehat{\boldsymbol{P}}\breve{\boldsymbol{G}}_N \end{bmatrix} ,$$

with

$$\widetilde{\boldsymbol{G}}_1 = \begin{bmatrix} -w_2/w_1 & -w_3/w_1 & \cdots & -w_N/w_1 \\ -w_2/w_1 & -w_3/w_1 & \cdots & -w_N/w_1 \\ \vdots & \vdots & \ddots & \vdots \\ -w_2/w_1 & -w_3/w_1 & \cdots & -w_N/w_1 \end{bmatrix} .$$

The third submatrix of $\boldsymbol{J}$ is formed by the columns related to the derivatives with respect to $\widehat{\ln P^r}$. This submatrix is equal to $\widehat{\boldsymbol{H}}$. Putting the three submatrices together, the Jacobian matrix can be written in the following compact form:

$$\boldsymbol{J} = \begin{bmatrix} \breve{\boldsymbol{G}} & \left(\boldsymbol{I}_N \otimes \widehat{\boldsymbol{P}}\right)\widetilde{\boldsymbol{G}} & \widehat{\boldsymbol{H}} \end{bmatrix} . \tag{A.56}$$

Using the diagonal $(NR) \times (NR)$-matrix $\boldsymbol{W}^{\intercal}\boldsymbol{W}$, we get the following quadratic form:

$$\boldsymbol{J}^{\intercal}\boldsymbol{W}^{\intercal}\boldsymbol{W}\boldsymbol{J} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{0}_{N\times(N-1)} & \boldsymbol{0}_{N\times(R-1)} \\ \boldsymbol{0}_{(N-1)\times N} & \boldsymbol{A}_{22} & \boldsymbol{A}_{23} \\ \boldsymbol{0}_{(R-1)\times N} & \boldsymbol{A}_{32} & \boldsymbol{A}_{33} \end{bmatrix} , \tag{A.57}$$

with

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

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

The inverse of the quadratic form (A.57) is denoted by

$$(\boldsymbol{J}^{\intercal}\boldsymbol{W}^{\intercal}\boldsymbol{W}\boldsymbol{J})^{-1} = \begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} & \boldsymbol{B}_{13} \\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22} & \boldsymbol{B}_{23} \\ \boldsymbol{B}_{31} & \boldsymbol{B}_{32} & \boldsymbol{B}_{33} \end{bmatrix}, \tag{A.58}$$

with

$$\boldsymbol{B}_{11} = \boldsymbol{A}_{11}^{-1} \tag{A.59}$$

$$\boldsymbol{B}_{22} = \left(\boldsymbol{A}_{22} - \boldsymbol{A}_{23}\boldsymbol{A}_{33}^{-1}\boldsymbol{A}_{32}\right)^{-1} \tag{A.60}$$

$$\boldsymbol{B}_{33} = \left(\boldsymbol{A}_{33} - \boldsymbol{A}_{32}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{23}\right)^{-1}\ . \tag{A.61}$$

Multiplying the diagonal elements of (A.58) by the estimated model variance,

$$\widehat{\sigma}^2 = \frac{S_{\hat{u}_i^r \hat{u}_i^r}}{RN - 2N - R + 2}\ , \tag{A.62}$$

and taking the square root of each of these products, gives the estimated standard errors of the NLCPD's estimated parameters. For that purpose, we derive the precise definitions of the matrices $\boldsymbol{B}_{11}$, $\boldsymbol{B}_{22}$, and $\boldsymbol{B}_{33}$.

Obviously, $\boldsymbol{B}_{11} = (1/R)\,\mathrm{diag}\,(\boldsymbol{w})^{-1}$. 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

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

Next, we insert the definitions of $\boldsymbol{A}_{22}^{-1}$ and $\boldsymbol{A}_{33}^{-1}$ into (A.60) and (A.61) and, finally, obtain

$$\boldsymbol{B}_{22} = \boldsymbol{A}_{33}^{-1} + \frac{1}{\sum\limits_{r\in\mathcal{R}} \left(\widehat{\ln P^r}\right)^2} \boldsymbol{V}$$

$$\boldsymbol{B}_{33} = \boldsymbol{A}_{22}^{-1} - \frac{1}{\sum\limits_{r\in\mathcal{R}} \left(\widehat{\ln P^r}\right)^2} \frac{1 - \sum\limits_{i\in\mathcal{N}} w_i \left(\widehat{\delta}_i\right)^2}{\sum\limits_{i\in\mathcal{N}} w_i \left(\widehat{\delta}_i\right)^2} \boldsymbol{Z} \,,$$

where

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

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

Multiplying the diagonal elements of $\boldsymbol{B}_{11}$, $\boldsymbol{B}_{22}$, and $\boldsymbol{B}_{33}$ by $\widehat{\sigma}$ gives the formulas (18), (20), and (19).

B Simulation results

Table 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. (21), but averaged over products instead of regions, e.g.:

$$\operatorname{Bias}\left(\widehat{\ln \pi}\right) = \frac{1}{N} \sum_{i\in\mathcal{N}} \operatorname{Bias}\left(\widehat{\ln \pi}_i\right) = \frac{1}{N} \sum_{i\in\mathcal{N}} \frac{1}{L} \sum_{l=1}^{L} \left(\widehat{\ln \pi}_{i,l} - \ln \pi_{i,l}\right)$$

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

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 (see third line of Tab. 5).

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

Auer, L. v. and Shumskikh, A. (2022). *Retrospective Computations of Price Index Numbers: Theory and Application*. Research Papers in Economics 01/22, Universität Trier.

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

Diewert, W. E. (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. (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.