DEUTSCHE
BUNDESBANK
EUROSYSTEM

# Discussion Paper

Deutsche Bundesbank
No 12/2020

## Measuring spatial price differentials: A comparison of stochastic index number methods

Sebastian Weinand

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

Tel +49 69 9566-0

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

Internet http://www.bundesbank.de



ISBN 978–3–95729–682–5 (Printversion)
ISBN 978–3–95729–683–2 (Internetversion)

# Non-technical summary


### Research Question

Comparisons between the price levels of different regions depend to a large extent on the quality of the underlying price data. Below the basic heading level, these data often have large gaps. Therefore, stochastic index number methods are used for aggregation into regional price levels. Two of the best known approaches are the Country-Product-Dummy method (CPD method) and the GEKS method, which is named after its developers (Gini, Eltetö, Köves and Szulc). The two methods generate identical estimates for regional price levels if prices are fully available for all products in the regions under analysis. This is no longer the case once there are gaps in the price data. To date, the literature has given little attention to the reasons behind this and the effects on price level estimation.

### Contribution

The present study expands the existing theoretical basis of the two index number methods. First, it shows that both the CPD method and the GEKS method can be deduced from the same stochastic model. Second, the formula of the estimated regional price levels is derived for a specific case of incomplete price data. The impact of missing prices on price level estimation is analysed by means of several simulation studies.

### Results

Where price data are incomplete, the regional price levels of the CPD method and the GEKS method are estimated by correction terms which factor prices from other regions into the calculation. Differences in the estimated price levels of the two index number methods arise solely through differences in the weighting of these correction terms. Particularly in cases of highly fragmentary price data, the CPD method exhibits better statistical properties in terms of estimation efficiency. An application to micro price data from the official consumer price statistics confirms this result.

# Nichttechnische Zusammenfassung


### Fragestellung

Vergleiche zwischen den Preisniveaus unterschiedlicher Regionen hängen in hohem Maße von der Qualität der zugrundeliegenden Preisdaten ab. Unterhalb der Elementarebene weisen diese oftmals große Lücken auf. Für die Aggregation in regionale Preisniveaus werden deshalb stochastische Indexmethoden eingesetzt. Zu den wohl bekanntesten Vertretern zählen die Country-Product-Dummy-Methode (kurz: CPD-Methode) und die nach ihren Entwicklern (Gini, Eltetö, Köves und Szulc) benannte GEKS-Methode. Beide generieren identische Schätzwerte für die regionalen Preisniveaus, wenn in den betrachteten Regionen Preise vollständig für alle Produkte vorliegen. Sobald die Preisdaten Lücken aufweisen, ist dies nicht länger der Fall. In der Literatur sind die Gründe hierfür und die Auswirkungen auf die Schätzung der Preisniveaus bislang kaum untersucht.

### Beitrag

Die vorliegende Studie erweitert die existierenden theoretischen Grundlagen der beiden betrachteten Indexmethoden. Es wird zum einen gezeigt, dass sowohl die CPD-Methode als auch die GEKS-Methode aus demselben stochastischen Modell abgeleitet werden kann. Zum anderen wird für einen spezifischen Fall lückenhafter Preisdaten die Formel der geschätzten regionalen Preisniveaus ermittelt. Über verschiedene Simulationsansätze wird der Einfluss fehlender Preise auf die Schätzung der Preisniveaus untersucht.

### Ergebnisse

Die regionalen Preisniveaus der CPD-Methode und der GEKS-Methode werden bei lückenhaften Preisdaten über Korrekturterme geschätzt, welche die Preise anderer Regionen in die Berechnung einbeziehen. Unterschiede in den geschätzten Preisniveaus der beiden Indexmethoden entstehen einzig durch eine unterschiedliche Gewichtung dieser Korrekturterme. Insbesondere bei sehr lückenhaften Preisdaten weist die CPD-Methode bessere statistische Eigenschaften in Bezug auf die Effizienz der Schätzung auf. Eine Anwendung auf Mikropreisdaten der amtlichen Verbraucherpreisstatistik bestätigt dieses Ergebnis.

DEUTSCHE BUNDESBANK DISCUSSION PAPER NO 12/2020

# Measuring spatial price differentials: A comparison of stochastic index number methods

Sebastian Weinand*
Deutsche Bundesbank

**Abstract**

Spatial price comparisons rely to a high degree on the quality of the underlying price data that are collected within or across countries. Below the basic heading level, these price data often exhibit large gaps. Therefore, stochastic index number methods like the CPD method and the GEKS method are utilised for the aggregation of the price data into higher-level indices. Although the two index number methods produce differing price level estimates when prices are missing, the present paper demonstrates that both can be derived from exactly the same stochastic model. In addition, for a specific case of missing prices, it is shown that the formula underlying these price level estimates differs between the CPD method and the GEKS method only with respect to the weighting pattern applied. Lastly, the impact of missing prices on the efficiency of the price level estimates is analysed in two simulation studies. It can be shown that the CPD method slightly outperforms the GEKS method. Using price data of Germany's Consumer Price Index, it can be observed that more narrowly defined products lead to efficiency gains in the estimation.

**Keywords:** Spatial price comparisons, below basic heading, multilateral index number methods, CPD method, GEKS method, product definition.

**JEL classification:** C43, E31, R10.

* Contact address: Deutsche Bundesbank, Wilhelm-Epstein-Straße 14, 60431 Frankfurt am Main, Germany. E-mail: sebastian.weinand[@]bundesbank.de. This work was part of the research project *RegioPreis* at Trier University. I am indebted to the RDC of the Federal Statistical Office and Statistical Offices of the Federal States for granting me access to the Consumer Price Index micro data of May 2011. I received valuable support from the Statistical Office of Bavaria as well as from Timm Behrmann, Florian Burg, Bernhard Goldhammer and Karsten Sandhop. I would also like to thank Ludwig von Auer, Bert Balk, Jan Becker and Thomas Knetsch for their detailed and helpful comments on the paper. The views expressed in this paper are those of the author and do not necessarily reflect the views of the Deutsche Bundesbank or the Eurosystem.

# 1 Introduction

In general, index number theory is divided into three primary strands: the *test approach* (e.g. Balk, 1995), which relies on a framework of desirable properties for the valuation of index number methods; the *economic approach* (e.g. Diewert, 1995), which builds on microeconomic theory in the context of cost and utility functions; and the *stochastic approach* (e.g. Clements, Izan and Selvanathan, 2006), which embeds the index number theory into a statistical framework. The stochastic approach to index numbers has especially gained increasing attention in recent years. Within the International Comparison Program (ICP), for example, it is used for the compilation of Purchasing Power Parities (PPPs) in the participating countries (see Rao and Hajargasht, 2016, pp. 415-416 and World Bank, 2015, pp. 256-257 for the ICP round in 2011 and Diewert, 2010, p. S14 for the ICP round in 2005). More specifically, in the first stage of aggregation, the PPPs are calculated using the two probably best-known index number methods of the stochastic approach: the Country-Product-Dummy (CPD) method, originally developed by Summers (1973), and the GEKS method, named after its authors Gini (1924, 1931), Eltetö and Köves (1964) and Szulc (1964).

The CPD method is a simple case of a hedonic regression. It explains the price of some product by the product itself and the region where that price was observed. In the literature it is well-known that the GEKS method can be put into a regression approach as well (e.g. Rao and Timmer, 2003, pp. 498-500). Initially, however, it was designed as a technique to adjust a set of bilateral index numbers such that these satisfy internal consistency in a multilateral context. The CPD and GEKS methods might be complemented by the much less prominent Country-Dummy (CD) method that reaches back to Summers (1973) as well. Similarly, within a regression framework, it explains the regional price ratio of some product by the general price level of the regions. A comprehensive survey of these stochastic index number methods is provided, for example, by Balk (2008) and Auer (2012).

In the literature, the CPD and GEKS methods are typically considered independently of each other. This, in fact, makes sense because of the different rationale behind their model specifications. Balk (1981, p. 75), however, pointed out that both methods are closely related.[1] In our paper, we demonstrate that both the CPD and the GEKS methods can be derived from the same stochastic model originally introduced by Summers (1973, p. 5) and Selvanathan and Rao (1992, pp. 338-340). Moreover, we show that the CD method also traces back to this model. This deeper anchoring of the three multilateral index number methods into the stochastic approach is this paper's first contribution.

[1] In a temporal framework, Balk (1980, 1981) applied the CPD and GEKS methods to the case of seasonally unavailable products.

Below the basic heading level, a spatial price comparison relies exclusively on the prices that were collected in different regions for a number of products.[2] If the price of each product is observed in each region, then the available price matrix is said to be complete. For a complete price matrix, it is known that the CPD method and the GEKS method generate exactly the same estimates for the regional price levels (e.g. World Bank, 2013, pp. 115-116). Strictly speaking, this is true when the bilateral price index numbers underlying the GEKS method are calculated as a Jevons index (henceforth, we use the term GEKS-Jevons method). A complete price matrix, however, is rarely available. More often there are large gaps in the price data due to missing prices. In this case, the CPD and GEKS-Jevons methods no longer produce the same results, although it would be helpful to know how these differences evolve.

Ferrari, Gozzi and Riani (1996) consider a price matrix with two groups of regions. For the first group of regions the matrix is complete and for each region of the second group the same prices are missing. The authors show that in this case the CPD method and the GEKS-Jevons method generate different results, and that the differences are due to different weights in a correction term. We extend the work of Ferrari *et al.* (1996) by a more general scenario of missing prices where all regions exhibit gaps. We show that their results also remain valid in this new setting. Now, however, the price level estimates rely not on one but on two correction terms that are weighted differently between the CPD and the GEKS-Jevons methods. Moreover, it can be shown that the price levels differ between intragroup comparisons (the prices of two regions that belong to the same group of regions are compared) and intergroup comparisons. For intragroup comparisons, the CPD price levels correspond to the Jevons index of the two regions under consideration. These further insights into the calculation of price levels in the case of missing prices are the paper's second contribution.

Our theoretical derivations draw on a specific case of an incomplete price matrix. To evaluate the impact of missing prices on the price level estimates also in a more general setting, we conduct a Monte Carlo analysis. For that purpose, we build artificial price data, randomly introduce gaps into these data sets and apply the CPD and GEKS-Jevons methods (a similar approach was undertaken by Dikhanov, 2010). This enables us to evaluate the impact of missing prices on the estimation efficiency separately for both index number methods and, in addition, to analyse possible differences between them. Not surprisingly, it turns out that the estimation efficiency in general suffers from an increasing number of gaps. Moreover, the CPD method slightly outperforms the GEKS-Jevons method in terms of estimation efficiency under different tested scenarios. These findings are the paper's third contribution.

We also adopt our simulation strategy to more realistic price data. For that purpose, we

[2] We use the term region in place of countries, cities or any other geographical entity.

draw on a subset of the micro price data underlying the official Consumer Price Index (CPI) of Germany. On the basis of these data, we are able to confirm the findings of our first simulation study. Moreover, we use the product descriptions provided in the data to analyse how the estimation efficiency reacts to narrower product definitions. This issue has practical relevance for two reasons. First, with respect to our simulation results, it shows that one may increase the estimation efficiency with a narrower product definition. Second, it also reveals that this gain in the efficiency is closely related to the regional volatility of the prices. More specifically, with low regional price fluctuations, one could rely on relatively loose product definitions as narrower ones do not significantly enhance the estimation efficiency. This finding may have an important implication for the compilation of regional price indices in practice. A narrow product definition using CPI data usually entails a lot of data preprocessing (e.g. Weinand and Auer, 2019, pp. 9-11). Our results indicate that this extensive workload can be reduced when the regional volatility of prices within a basic heading is taken into account. This is the paper's fourth contribution.

The remainder of the paper is laid out as follows. Section 2 provides an overview of the stochastic approach to index numbers in the context of spatial price comparisons below the basic heading level. Section 3 discusses appropriate error term specifications in the light of empirical studies on spatial price comparisons. Section 4 presents the theoretical derivations for a specific case of incomplete price data and the results of our simulation studies while Section 5 concludes.

# 2 Stochastic approach to spatial price index numbers

Two central requirements for spatial price comparisons are transitivity and characteristicity of the price index numbers. They are defined in Section 2.1, along with some other basic concepts. In Section 2.2, we derive the CPD and CD methods from a stochastic model initially proposed by Summers (1973) and Selvanathan and Rao (1992) in the context of spatial price comparisons. Likewise, in Section 2.3, we derive the GEKS method from that model.

## 2.1 Basic concepts and definitions

Usually, the price levels of more than two regions are compared. A basic requirement of such multilateral price comparisons is called *transitivity*. It postulates that $P^{sr}$, the relative price level between the regions $r$ and $s$, should be equal to the product of the price levels $P^{st}$ and $P^{tr}$, where $t$ is some arbitrary third region that serves as a bridge (e.g. Rao and Banerjee, 1986, p. 304). Consequently, transitivity ensures the internal consistency of some multilat-

eral system of index numbers. A second postulate, initially advocated by Drechsler (1973), is denoted as *characteristicity*. It states that the price comparison between two regions $r$ and $s$ should be based exclusively on information relating to these two regions. Both requirements play a central role for price comparisons below and above the basic heading level.

Below the basic heading level, neither expenditure weights nor quantity information are available. In this case, elementary price indices are used for the aggregation of prices into higher-level indices. An elementary index number formula widely used among statistical offices is the Jevons (1865) index (e.g. OECD, 2018, pp. 8-9). For the regions $r$ and $s$, it is defined by:

$$\dot{P}_{\mathrm{J}}^{sr} = \prod_{i=1}^{N} \left(p_i^r / p_i^s\right)^{\frac{1}{N}} , \tag{1}$$

where $p_i^r$ is the price of product $i$ in region $r$ and $N$ the number of products.[3]

The Jevons index outperforms most other elementary index number formulas under the axiomatic approach to index numbers and is also (weakly) supported under the economic approach (e.g. Diewert, 1995, pp. 5-20). In particular, Hill and Hill (2009, p. 198) point out that the Jevons index numbers are transitive if prices are available for each product and region. Moreover, from (1), it is obvious that each index number on its own is characteristic. In practice, however, price information for individual products are frequently missing below the basic heading level. Equation (1) shows that the Jevons index is only applicable to regionally matched price observations. Thus, price comparisons between different pairs of regions (e.g. $\dot{P}^{st}$ versus $\dot{P}^{sr}$) might stem from varying sets of matched prices. Consequently, a multilateral system of bilateral Jevons index numbers would still be characteristic, but no longer transitive. Taking into account the trade-off between transitivity and characteristicity, the stochastic approach to index numbers offers alternatives to ensure transitivity even in the event that prices are missing.

Following the stochastic model advocated by Summers (1973, p. 5) and Selvanathan and Rao (1992, pp. 338-340), the price ratio of product $i$ for regions $r$ and $s$ is defined by the multiplicative relationship of two terms: the general price level of region $r$ relative to region $s$, $P^{sr}$, and a random component, $\epsilon_i^{rs}$. If transitivity is assumed, $P^{sr}$ can be written as $P^r/P^s$ (e.g. Rao and Banerjee, 1986, pp. 304-306). Hence, the logarithm of this multiplicative relationship can be expressed by

$$\ln\left(p_i^r / p_i^s\right) = \ln\left(P^r / P^s\right) + u_i^{sr} , \tag{2}$$

where $u_i^{sr} = \ln \epsilon_i^{sr}$ is assumed to be some normally distributed random variable with ex-

[3] In the following, we denote bilateral price index numbers by a dot, e.g. $\dot{P}^{sr}$, in order to indicate that the price index number is not necessarily transitive in a multilateral context.

pected value 0 and variance $\sigma^2$ for all products $i = 1,2,\ldots,N$ and regions $s,r = 1,2,\ldots,R$.[4] In Appendix A.1.2, it is shown that $u_i^{sr}$ is not mutually independent. Instead, it follows that $u_i^{sr} = u_i^{vr} - u_i^{vs}$. In the following, we show that the stochastic model in (2) serves as a starting point for the derivation of the CD, the CPD and the GEKS methods, respectively.

## 2.2 CPD method and Country-Dummy method

Taking the sum over all regions $s = 1,\ldots,R$ in Equation (2) and rearranging leads to

$$\ln p_i^r = \ln P^r + \frac{1}{R}\sum_{s=1}^{R} \ln\left(p_i^s / P^s\right) + \frac{1}{R}\sum_{s=1}^{R} u_i^{sr} \,. \tag{3}$$

Although the price $p_i^s$ on the right-hand side of the equation is initially known, the price level $P^s$ is not. Therefore, the arithmetic average of the logarithmic price to price level ratios, $\frac{1}{R}\sum_{s=1}^{R} \ln\left(p_i^s / P^s\right)$, is also unknown. We denote this term by $\ln \pi_i$. From an economic point of view, it represents the average deflated price of product $i$. This interpretation reveals similarities to the "international price" of the Geary-Khamis method.[5] In addition, we define $\frac{1}{R}\sum_{s=1}^{R} u_i^{sr} = u_i^r$. Consequently, (3) can be rewritten as

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

Equation (4) represents the logarithmic form of the CPD method's underlying model (Summers, 1973, p. 10). It explains the price of product $i$ in region $r$, $p_i^r$, by region $r$'s general price level $P^r$ and product $i$'s general value $\pi_i$. Because $u_i^r$ is a linear combination of the disturbances $u_i^{sr}$ in (2), it follows a normal distribution with expected value 0. The variance of the disturbances is assumed to be identical among the regions and products in the original form of the CPD method.

In order to transform (4) into a standard regression model, we introduce for each region $t$ ($t = 1,\ldots,R$) the dummy variable $region^t$ and for every product $j$ ($j = 1,\ldots,N$) the dummy variable $product_j$:

$$region^t = \begin{cases} 1 & \text{if } r = t \\ 0 & \text{if } r \neq t \end{cases} \quad \text{and} \quad product_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \,. \tag{5}$$

[4] In order to derive the CCD index (see Caves, Christensen and Diewert, 1982) under the stochastic approach, Selvanathan and Rao (1992, pp. 338-340) assume heteroskedastic disturbances. In the context of intertemporal price comparisons, Clements and Izan (1981, pp. 745-746) show that the Divisia index can be derived from (2) under plausible specifications of the error term.

[5] Geary (1958) and Khamis (1972) included country and product specific quantities in the definition of the international price. If the quantities are identical across countries, the definition simplifies to $\frac{1}{R}\sum_{s=1}^{R} p_i^s / P^s$.

Defining $\alpha^t = \ln\left(P^t/k\right)$ and $\beta_j = \ln\left(k \cdot \pi_j\right)$, with $k$ being some constant, we can express Equation (4) by

$$\ln p_i^r = \sum_{t=1}^{R} \alpha^t \, region^t + \sum_{j=1}^{N} \beta_j \, product_j + u_i^r \quad \forall \; r = 1,\ldots,R \text{ and } i = 1,\ldots,N\,. \tag{6}$$

Equation (6) can be viewed as a linear regression model, albeit one suffering from perfect multicollinearity. Furthermore, we are interested in estimates of the price levels $P^t$. Since $\alpha^t = \ln\left(P^t/k\right)$, we first need to specify $k$. Both problems can be simultaneously solved by specifying $k$ in terms of the parameter $\alpha^t$.

If we define region $t = 1$ (or some other region) as the base region that serves as a reference for the price levels of the other regions, that is, $k = P^1$, it follows that $\alpha^1 = \ln\left(P^1/P^1\right) = 0$. As a consequence, $\alpha^1 \, region^1 = 0$ for all observations. Therefore, the term $\alpha^1 \, region^1$ can be dropped from Equation (6):

$$\ln p_i^r = \sum_{t=2}^{R} \alpha^t \, region^t + \sum_{j=1}^{N} \beta_j \, product_j + u_i^r \quad \forall \; r = 1,\ldots,R \text{ and } i = 1,\ldots,N\,. \tag{7}$$

Perfect multicollinearity is removed. The parameters $\alpha^t$ are estimated using ordinary least squares (OLS). The corresponding estimator, $\widehat{\alpha}^t$, is defined as the logarithmic price level relative between region $t$ and the base region. By definition, these estimated price levels satisfy the requirement of transitivity (e.g. Rao and Banerjee, 1986, pp. 304-306). Alternatively, we could avoid perfect multicollinearity in (6) by setting $\sum_{t=1}^{R} \alpha^t = 0$. Consequently, $\widehat{\alpha}^t$ would express the logarithmic price level of region $t$ relative to the unweighted average price level of all regions. Diewert (2004, pp. 6-8) describes an elegant way of estimating the parameters $\alpha^t$ in this setting.

Alternatively to the approach outlined above, one can set in Equation (2) region $s$ as a fix reference for product $i$'s price ratios:

$$\ln\left(p_i^r/p_i^s\right) = \ln\left(P^r/P^s\right) + u_i^{sr} \quad \forall \; r \in R_s^* \text{ and } i = 1,\ldots,N \tag{8}$$

where $R_s^* = \{r \in \mathbb{N}^+ \mid r \leq R,\; r \neq s\}$. Equation (8) represents the Country-Dummy method. It assumes that any product-specific price ratio between two regions $r$ and $s$ can be explained by the overall price level relative of these regions. The disturbances $u_i^{sr}$ remain a normally distributed random variable with expected value 0 and variance $\sigma^2$. As pointed out by Summers (1973, p. 10), it follows that $\operatorname{cov}\left(u_i^{sr},\, u_i^{sv}\right) = \frac{1}{2}\sigma^2$ for regions $r \neq v$. If additionally $\operatorname{cov}\left(u_i^{sr},\, u_j^{sr}\right) = \operatorname{cov}\left(u_i^{sr},\, u_j^{sv}\right) = 0$ for products $i \neq j$ is assumed, the disturbances are autocorrelated blockwise (see Appendix A.1.2 for a formal proof).

In addition to the dummy variable $region^t$ in (5), we need to define a second dummy variable that refers to the price of region $s$ in the price ratio $\ln\left(p_i^r/p_i^s\right)$. For that purpose, we introduce for each region $t$ ($t = 1,\ldots,R$) the dummy variable

$$\widetilde{region^t} = \begin{cases} 1 & \text{if } s = t \\ 0 & \text{if } s \neq t \end{cases} . \tag{9}$$

The two dummy variables, $region^t$ and $\widetilde{region^t}$, are complemented by the additional parameter of region $t$'s logarithmic price level, $\alpha^t$. Defining $\alpha^t = \ln\left(P^t/P^1\right)$, the regression model of the CD method can be expressed by

$$\ln\left(p_i^r/p_i^s\right) = \sum_{t=2}^{R} \alpha^t \left(region^t - \widetilde{region^t}\right) + u_i^{sr} \quad \forall\ r \in R_s^* \text{ and } i = 1,\ldots,N\,. \tag{10}$$

Since $\alpha^1 = \ln\left(P^1/P^1\right) = 0$, it follows that $\alpha^1\left(region^1 - \widetilde{region^1}\right)$ is not included in (10). Due to the known autocorrelation structure, the remaining parameters $\alpha^2,\ldots,\alpha^R$ are estimated using generalised least squares (GLS). They express the logarithmic price level relative between region $t$ and the base region $t = 1$. Moreover, they are transitive in a multilateral context. In comparison to (7), it is worthwhile to note that the lower number of model parameters is exactly compensated by a lower number of observations. As a result, the degrees of freedom in models (7) and (10) are identical.

Subtracting the definition of product $i$'s price in region $r = 1$, $\ln p_i^1 = \ln P^1 + \ln \pi_i + u_i^1$, from Equation (4) and rearranging yields the CPD's regional price level with

$$\ln\left(P^r/P^1\right) = \ln\left(p_i^r/p_i^1\right) - \left(u_i^r - u_i^1\right) .$$

From (4), it is known that $u_i^r = \frac{1}{R}\sum_{s=1}^{R} u_i^{sr}$ and $u_i^1 = \frac{1}{R}\sum_{s=1}^{R} u_i^{s1}$. Consequently, $u_i^r - u_i^1$ can be written as $\frac{1}{R}\sum_{s=1}^{R}\left(u_i^{sr} - u_i^{s1}\right)$. In Appendix A.1.2, it is shown that $u_i^{sr} - u_i^{sv} = u_i^{vr}$. Similarly, $u_i^{sr} - u_i^{s1} = u_i^{1r}$ applies. As a result, the previous equation can be rewritten as (8), the CD method's price level, $\ln\left(P^r/P^1\right)$, relative to region $s = 1$. This suggests that the CPD and CD methods are equivalent approaches and, therefore, should give equal price level estimates.

In the case of a complete price matrix, the CPD and CD method's price level estimator, $\exp\left(\widehat{\alpha}^t\right)$, is defined as a geometric average of the price ratios between region $t$ and the base region (see, for example, Rao and Hajargasht, 2016, pp. 418-419, and Appendix A.1 for the derivation of that result).[6] Consequently, the estimated price levels coincide with the Jevons index in (1). Furthermore, it follows that the CPD estimator for product $j$'s general value, $\widehat{\beta}_j$,

[6] Kennedy (1981, p. 801) recommends calculating $P^t$ by $\exp\left(\widehat{\alpha}^t - 0.5\widehat{\mathrm{var}}\left(\widehat{\alpha}^t\right)\right)$ instead of $\exp\left(\widehat{\alpha}^t\right)$ in order to reduce the upward bias that would arise from the convex transformation $\exp\left(\widehat{\alpha}^t\right)$.

is defined by $\frac{1}{R}\sum_{t=1}^{R}\ln\left(p_j^t/\exp\left(\widehat{\alpha}^t\right)\right)$.[7] This expression is already known from (3). It reveals that the prices of product $j$ are deflated by the respective regional price levels.

## 2.3 GEKS method

Following Hill (2008, p. 3), the GEKS method is not a price index in the proper sense. Strictly speaking, it is a two-stage technique to convert a set of bilateral price index numbers into a multilateral system of transitive index numbers. The first stage encompasses the calculation of the bilateral index numbers, $\dot{P}^{sr}$, for each regional pair $r$ and $s$. These, however, may lack transitivity, with the result that they differ from the multilateral index numbers, $P^{sr}$. For that reason, the second stage incorporates an adjustment of the characteristic bilateral into transitive multilateral index numbers. Drechsler (1973, p. 28) points out that the GEKS method is designed with the aim of keeping this adjustment as small as possible with respect to the trade-off between characteristicity and transitivity. A mathematical formulation of this optimisation problem can be found in Hill and Timmer (2006, pp. 368-369) and Rao and Timmer (2003, pp. 497-500).

In the following, we demonstrate that the multilateral GEKS-Jevons method, like the CPD method, can be derived from the stochastic model defined in (2). Taking the sum over all products $i = 1,\ldots,N$ in (2) and rearranging yields

$$\frac{1}{N}\sum_{i=1}^{N}\ln\left(p_i^r/p_i^s\right) = \ln\left(P^r/P^s\right) + \frac{1}{N}\sum_{i=1}^{N}u_i^{sr}\,. \tag{11}$$

The term on the left-hand side of the equation is the logarithmic form of the Jevons index in (1).[8] Therefore, we denote it by $\ln\dot{P}_{\mathrm{J}}^{sr}$. In addition, we define $\frac{1}{N}\sum_{i=1}^{N}u_i^{sr} = u^{sr}$. Consequently, (11) can be rewritten as

$$\ln\dot{P}_{\mathrm{J}}^{sr} = \ln\left(P^r/P^s\right) + u^{sr}\,. \tag{12}$$

This model specification of the GEKS-Jevons method is widely documented (e.g. Hill, 2016, p. 408). It states that the bilateral Jevons index numbers, $\ln\dot{P}_{\mathrm{J}}^{sr}$, and the corresponding transitive index numbers, $\ln\left(P^r/P^s\right)$, differ only with respect to the disturbances $u^{sr}$. Since the disturbances $u^{sr}$ are a linear combination of $u_i^{sr}$ in (2), they follow a normal distribution with expected value 0. Their variance is assumed to be identical.

In order to transform (12) into a standard regression model, we introduce for each region $t$ ($t = 1,\ldots,R$) the two dummy variables $region^t$ and $\widetilde{region}^t$. Their definitions can be found

[7] If the restriction $\sum_{t=1}^{R}\alpha^t = 0$ applies, then $\widehat{\beta}_j = \frac{1}{R}\sum_{t=1}^{R}\ln p_j^t$ follows (see also Diewert, 2004, p. 7).

[8] Rao and Banerjee (1986, p. 306) underline the importance that the bilateral index numbers satisfy the country-reversal test. The Jevons index exhibits that property.

in (5) and (9). Defining $\alpha^t = \ln\left(P^t/P^1\right)$, Equation (12) can be written as

$$\ln \dot{P}_{\mathrm{J}}^{sr} = \sum_{t=2}^{R} \alpha^t \left( region^t - \widetilde{region^t} \right) + u^{sr} \quad \forall\ s = 1,\ldots,R-1 \text{ and } r = s+1,\ldots,R\,. \tag{13}$$

Regression model (13) draws on non-redundant bilateral price index numbers only. Because $\alpha^1 = \ln\left(P^1/P^1\right) = 0$, it follows that $\alpha^1 \left( region^1 - \widetilde{region^1} \right)$ is not included in (13). The remaining parameters $\alpha^2,\ldots,\alpha^R$ are estimated using OLS. In Appendix A.1, it is shown that the corresponding estimator, $\exp\left(\widehat{\alpha}^t\right)$, is defined by

$$\exp\left(\widehat{\alpha}^t\right) = \prod_{r=1}^{R} \left( \dot{P}_{\mathrm{J}}^{1r} \cdot \dot{P}_{\mathrm{J}}^{rt} \right)^{\frac{1}{R}}\,, \tag{14}$$

which is the typical presentation of the GEKS-Jevons method (e.g. ILO, IMF, OECD, UNECE, Eurostat and World Bank, 2004, p. 498).[9] Moreover, when we insert the definitions of $\dot{P}_{\mathrm{J}}^{1r}$ and $\dot{P}_{\mathrm{J}}^{rt}$ into (14), it simplifies to

$$\exp\left(\widehat{\alpha}^t\right) = \prod_{i=1}^{N} \left( p_i^t/p_i^1 \right)^{\frac{1}{N}} = \dot{P}_{\mathrm{J}}^{1t}$$

in the case of a complete price matrix. Thus, the estimated price levels of the GEKS-Jevons method, $\exp\left(\widehat{\alpha}^t\right)$, are defined as ordinary Jevons indices (e.g. Hill, 2016, p. 408). This is due to the fact that the bilateral Jevons index numbers are transitive when no product prices are missing. As a consequence, no adjustment of the bilateral index numbers is necessary.

# 3 Discussion on the error term specification

In the previous section, it was shown that the stochastic approach to index numbers provides point estimates for the price level of some region. The reliability of these estimates depends to a high degree on the quality of the collected price data. In particular, differences in the quality of products, missing prices across regions or selection bias may lead to distortions in the regional price levels and a loss in representativity. These "non-stochastic" sources of error are widely discussed in the literature (e.g. Balk and Kersten, 1986; Kokoski, Moulton and Zieschang, 1999; Silver, 2009).

As one of its main advantages over the economic and the test approach, the stochastic approach to index numbers provides measures of precision for the estimated price levels

[9] Clearly, other bilateral price index formulas instead of the Jevons index could be used in (14) as well. In its origin, the GEKS method was constructed based on binary Fisher indices. Caves *et al.* (1982, p. 78) propose the use of the Törnqvist index. Both indices require quantity information.

(e.g. standard errors, confidence intervals). Even if price level estimates are unbiased, the interpretation of these measures relies highly on the choice of model specification and the assumptions on the error term.[10] Model (2), for example, postulates that the price ratio of regions $r$ and $s$ for product $i$ solely deviates from the general relative price level by a random error term. This is a simple but plausible assumption. The assumption that the error term's variance is identical among the regions and products, however, is rather restrictive (e.g. Summers, 1973, p. 6). In particular, it is not only restrictive but might be false when it does not fit to the underlying empirical data. As a consequence, the measures of precision would be biased and, thus, meaningless. In the following, we address the importance of appropriate error term specifications in light of empirical studies on spatial price comparisons.

**Variance of the disturbances:** The error term in (2) is assumed to be homoskedastic. This assumption, however, might be inappropriate when the price ratios behave systematically different among the regions and/or products (e.g. due to pricing policies that differ among the products). For example, suppose that some basic heading consists of two products $i$ and $j$. Product $i$ is uniformly priced in the regions while product $j$ is not. Using (2), the error terms would be realised by

$$\exp\left(u_i^{sr}\right) = P^s/P^r \quad \text{and} \quad \exp\left(u_j^{sr}\right) = \left(P^s/P^r\right) \cdot \left(p_j^r/p_j^s\right),$$

suggesting a higher dispersion of product $j$'s error term. The assumption of homoskedastic rather than heteroskedastic disturbances would be difficult to defend in this case.[11]

Moreover, from a statistical viewpoint, some of the price ratios might be more reliable than others. Several researchers have stressed this issue by incorporating more plausible specifications on the error term into the CPD and GEKS methods.[12] Rao (2001, pp. 4-8) introduced a weighting concept into the GEKS method where the variance of the error term, $\text{var}(u^{sr})$, depends on individual weights, $w^{sr}$, for the underlying bilateral price index numbers. Consequently, with $\text{var}(u^{sr}) = \sigma^2/w^{sr}$, it is possible to discriminate between different pairs of regions. Within this framework, Rao and Timmer (2003, pp. 498-500) developed and tested various weighting schemes (e.g. weights based on the number of product matches or

[10] Hajargasht and Rao (2010, pp. S38-S44) show that the CPD model under different distributional assumptions on the error term leads to various multilateral index number methods (e.g. Iklé index, Rao system).

[11] In the context of intertemporal price comparisons, Crompton (2000) and Selvanathan (2003) recommend the use of White's (1980) heteroskedasticity-consistent covariance matrix. Following Crompton, "the exact nature of the error variance is of no concern, and can remain unidentified."

[12] We do not consider the GEKS method proposed by Eurostat-OECD (2012, pp. 243-244) at this point because it takes into account additional information on the representativity of some product rather than incorporating new specifications on the error term. For the same reason, we omit the CPRD method (the "R" stands for the additional representativity dummy variables in the CPD model; see World Bank, 2013, pp. 109-111) from our discussion.

the economic distance of regions) while Hill and Timmer (2006, pp. 370-371) derived standard errors as a weighting factor that "penalizes bilateral comparisons where the overlap of products is small". Similarly, Rao (2001, pp. 15-16), Rao (2005, pp. 574-575) and Diewert (2005) incorporated weights into the CPD method that reflect the importance of a single price observation. In the absence of expenditure share data within a basic heading, the ICP uses "importance weights" that distinguish between important (weight of 3) and unimportant (weight of 1) price observations (see World Bank, 2013, p. 110). In this scenario, the price levels are estimated by weighted rather than ordinary least squares.

**Covariance of the disturbances among regions:** The covariance of different regions (CPD method) or different pairs of regions (GEKS method) is usually assumed to be zero, meaning that the error terms are spatially uncorrelated. Empirical studies, however, have shown that spatial autocorrelation can be found in prices as well as price levels. Aten (1997, 1996) was the first to explore the spatial nature in price levels. Using household consumption data for 64 countries in 1985, Aten (1996, pp. 160-162) reported for 15 out of 23 product categories "significantly spatially autocorrelated residuals". Rao (2001, pp. 18-20) found for seven out of eight highly aggregated product categories, such as food and furniture, significant spatial autocorrelation. These findings have been confirmed in more recent studies that used the price data underlying the official CPI. Biggeri, Laureti and Polidoro (2017, pp. 109-111) computed sub-national price levels on the basis of official CPI data for seven basic headings that were collected in Italy in 2014. They reported that "an autocorrelation among disturbances was observed for all the BHs [basic headings] under analysis even if Moran's $I$ is quite low in some cases." Similarly, Weinand and Auer (2019, pp. 29-31) computed a regional price index with price data underlying the German CPI in 2016. They found positive spatial autocorrelation in the regional price levels which is mainly driven by housing and, to a much lesser degree, by services and goods.

The empirical studies show that spatial autocorrelation plays an important role in spatial price comparisons. More specifically, from a statistical viewpoint, ordinary least squares would no longer provide efficient estimates when the disturbances are spatially autocorrelated. Therefore, various concepts have been proposed to address this issue. Rao (2004, pp. 8-11) reformulated the original CPD model into a spatial error model (e.g. Anselin, 2003, p. 316). In this modified version, the disturbances are assumed to be spatially autocorrelated, with $\text{cov}\left(u_i^r,\, u_i^s\right) \neq 0$ for regions $r \neq s$. In contrast, Montero, Laureti, Mínguez and Fernández-Avilés (2019, pp. 8-10) propose a spatially-penalised version of the CPD method where a penalty for the differences in the price levels of neighbouring regions is included in the CPD model. For the GEKS method, Cuthbert (2003, pp. 77-78) recommended on the basis of an OECD data set the use of an "idealised" variance-covariance-matrix with constant variances and covariances that are defined by $\text{cov}\left(u^{sr},\, u^{st}\right) > 0$ and $\text{cov}\left(u^{sr},\, u^{ut}\right) = 0$

for different regions $r, s, t$ and $u$.

**Covariance of the disturbances among products:** A relatively new field in price statistics is the collection of online price data using web scraping techniques. Empirical studies show that many online retailers are characterised by uniform pricing policies, meaning that the prices on their website do not depend on the buyers' location (e.g. Cavallo, 2018, pp. 15-21). Thus, spatial autocorrelation might not play a dominant role in online price data. However, with web scraped price data, new issues might arise that affect the error term specification in multilateral index number methods. More specifically, online prices that are adjusted by algorithms in response to competitors' price changes for the same product or for a substitute might lead to correlated disturbances among the products. A survey of the European Commission (2017, pp. 175-177) on e-commerce strengthens this assumption. It revealed that "53% of the respondent retailers track the online prices of competitors [...]" while the "majority of those retailers that use software to track prices subsequently adjust their own prices to those of their competitors (78%)". Moreover, Chen, Mislove and Wilson (2016, pp. 1344-1346) found evidence of dynamic pricing in the online marketplaces of Amazon while Calvano, Calzolari, Denicolò and Pastorello (2018, pp. 24-25) and Klein (2018, pp. 10-17) studied experimentally Q-learning algorithms and showed, broadly speaking, that these are able to coordinate on price setting.

The issue of correlated product prices is likely to be of more concern for the CPD method rather than for the GEKS method. From (12), it is obvious that the latter does not rely on individual product prices. Consequently, possible correlations among the prices of products vanish in the calculation of the bilateral index numbers. In contrast, for the CPD method, it would imply that $\text{cov}\left(u_i^r,\, u_j^r\right) \neq 0$ for products $i \neq j$. Furthermore, taking into account the findings on regionally uniform pricing of online retailers by Cavallo (2018), this might be extended to $\text{cov}\left(u_i^r,\, u_j^s\right) \neq 0$ for products $i \neq j$ and regions $r \neq s$. Lastly, it is worthwhile noting that correlated product prices are likely to have more relevance in temporal price comparisons as the algorithmic adjustment of prices to those of competitors is more of a temporal rather than a spatial issue. Nevertheless, the considerations may give rise to future research in this field.

# 4 Price level estimates when prices are missing

In Section 2, it was shown that the CPD, CD and GEKS-Jevons methods yield identical price level estimates under suitable assumptions on the error terms and in the case of complete price data. Strictly speaking, the price levels are defined by Equation (1) as a Jevons index. It is well known that this equivalence no longer applies when prices are missing (e.g. World

Bank, 2013, p. 108), though there is little knowledge about how price level estimates change. One exception is the work of Ferrari *et al.* (1996). They consider the case where prices for exactly the same products are missing in some regions while prices are fully available in the remaining regions. It turns out that the CPD and GEKS-Jevons price level estimators are still defined as a geometric average of those price ratios that are commonly available in the regions to be compared, though multiplied with a correction term. The correction term is weighted differently in both methods.

## 4.1 Some insights from a specific case of missing prices

In the following, we expand the considered case in Ferrari *et al.* (1996). For that purpose, we randomly divide the regions into the nonempty and disjoint subsets $R_k$ and the products into the disjoint subsets $N_k$ (henceforth, we will refer to region and product groups).[13] We suppose that our price data consists of two groups of regions and products, respectively, that is, $k \in \{1,2\}$. Furthermore, we assume that the prices of products $i \in N_k$ are only available in regions $r \in R_k$. Thus, we have two complete, but non-connected blocks of prices (e.g. World Bank, 2013, p. 98). Using graph-theoretic concepts, Rao (2004, pp. 11-17) shows that the computation of price levels with the CPD method, however, requires a connected price data graph.[14] This is a remarkable result that also applies to the stochastic GEKS approach (e.g. Ferrari and Riani, 1998, pp. 102-105). Therefore, we introduce an additional, nonempty set of products, $N_0$, whose prices are fully available in all regions. As a consequence, our price data are said to be connected since all regions are linked either through direct or indirect comparisons of product prices.[15] In total, they encompass $\sum_{k=1}^{2} |R_k| = R$ regions and $\sum_{k=0}^{2} |N_k| = N$ products, where $|R_k|$ is the number of regions and $|N_k|$ the number of products in group $k$, respectively. The corresponding price incidence matrix is sketched in Table 3 of the Appendix.

We denote the price level estimator of region $t$ compared to some arbitrary base region $s$ by $\widehat{P}_m^{st}$. The subscript $m$ indicates if the price level stems from the CPD or the GEKS-Jevons

[13] The concept of product groups in this context is only a theoretical one and should not be mixed up with the classification of similar products into product groups as is the usual practice in official price statistics.

[14] Using graph-theoretic concepts, Hajargasht and Rao (2019) derive necessary and sufficient conditions for the existence and uniqueness of various index number methods.

[15] The general form of the price data is the same as in Hajargasht, Rao and Abbas (2019, pp. 105-106, panel *e* of Figure 1), where the authors derive the formula for the estimated variance of the CPD price levels.

method. If region $s \in R_1$, it is shown in Appendix A.2 that

$$\widehat{P}_m^{st} = \prod_{i \in N_0} \left( \frac{p_i^t}{p_i^s} \right)^{\frac{1}{|N_0|}} \cdot \left( \prod_{r \in R_1} \Lambda_1^{sr} \right)^{\frac{1}{\lambda_{m,1}}} \cdot \begin{cases} \left( \prod_{r \in R_1} \Lambda_1^{rt} \right)^{\frac{1}{\lambda_{m,1}}} & \text{if } s,t \in R_1 \\ \left( \prod_{r \in R_2} \Lambda_2^{rt} \right)^{\frac{1}{\lambda_{m,2}}} & \text{if } s \in R_1 \wedge t \in R_2 \end{cases} \tag{15}$$

where the correction term of product group $N_k$ for regions $r$ and $t$, $\Lambda_k^{rt}$, as well as its weighting factor, $\lambda_{m,k}$, are defined by

$$\Lambda_k^{rt} = \frac{\prod_{i \in N_0 \cup N_k} \left( p_i^t / p_i^r \right)^{\frac{1}{|N_0|+|N_k|}}}{\prod_{i \in N_0} \left( p_i^t / p_i^r \right)^{\frac{1}{|N_0|}}} \quad \text{and} \quad \lambda_{m,k} = \begin{cases} |R_k| & \text{if } m = \text{CPD} \\ R & \text{if } m = \text{GEKS} \end{cases}$$

for $k \in \{1,2\}$.[16] The correction term $\Lambda_1^{sr}$ for regions $s$ and $r$ is defined in the same way.

The basic structure of the correction term is obviously the same for the CPD and GEKS-Jevons methods. Consequently, the price levels $\widehat{P}_{\text{CPD}}^{st}$ and $\widehat{P}_{\text{GEKS}}^{st}$ differ only due to a different weighting of the two correction terms. Since $|R_k| < R$ for all $k$, the CPD method assigns greater weights to the correction terms than the GEKS-Jevons method. One could say that the CPD method's weights are somewhat plutocratic (Diewert, 1986, pp. 18-19) because they differ with respect to the regional group sizes. In contrast, the correction terms of the GEKS-Jevons method are weighted independently of the regional group sizes and, thus, more or less "democratic".

Equation (15) reveals that the calculation of $\widehat{P}_m^{st}$ distinguishes between a price comparison involving two regions of the same (intragroup comparisons) or of a different regional group (intergroup comparisons). For intragroup comparisons, it actually simplifies to

$$\widehat{P}_m^{st} = \prod_{i \in N_0} \left( \frac{p_i^t}{p_i^s} \right)^{\frac{1}{|N_0|}} \cdot \left( \Lambda_1^{st} \right)^{\frac{|R_1|}{\lambda_{m,1}}} \quad \text{if } s,t \in R_1 \, . \tag{16}$$

The price level solely relies on the prices of the two regions under consideration.[17] Therefore, it is fully characteristic. In contrast, the price levels of intergroup comparisons, $\widehat{P}_m^{st}$ (with $s \in R_1 \wedge t \in R_2$), are defined in (15) by a geometric average of those prices that are commonly available in regions $s$ and $t$ (first term), multiplied with two regional sequences of the

[16] It is worthwhile to note that we can replicate the results in Ferrari *et al.* (1996) by setting $|N_k| = 0$ for $k = 1$. Technically speaking, we drop products $i \in N_1$ from our price data but keep regions $r \in R_1$. Consequently, the prices in regions $r \in R_2$ are fully available for all products.

[17] In Appendix A.2, it is shown that we obtain the price level of regions $s, t \in R_2$ when we replace $\left(\Lambda_1^{st}\right)^{|R_1|/\lambda_{m,1}}$ with $\left(\Lambda_2^{st}\right)^{|R_2|/\lambda_{m,2}}$ in (16).

correction terms, $\Lambda_1^{sr}$ and $\Lambda_2^{rt}$ (second and third term). The latter two put the prices of regions $s$ and $t$ in relation to those of other regions $r$ in the same regional group. As a result, intragroup comparisons generate characteristic price levels while intergroup comparisons do not.

The GEKS-Jevons price levels are generated such that the overall quadratic deviation to the initial Jevons price levels, $\dot{P}_J^{st}$, is kept at a minimum (e.g. Rao and Timmer, 2003, p. 499; Laureti and Rao, 2018, p. 126). Therefore, it would be the natural choice to use the GEKS-Jevons method in every situation in order to approximate $\dot{P}_J^{st}$ on average as close as possible. Equation (15) shows that this might be appropriate when intergroup comparisons are of relevance, due to the smoother weighting of the correction terms, $\Lambda_1^{sr}$ and $\Lambda_2^{rt}$, by the GEKS-Jevons method.[18] For intragroup comparisons, however, this is not the case. With $\lambda_{\text{CPD},1} = |R_1|$, the CPD price level estimator, $\widehat{P}_{\text{CPD}}^{st}$, in (16) simplifies to

$$\widehat{P}_{\text{CPD}}^{st} = \prod_{i \in N_0 \cup N_1} \left(p_i^t / p_i^s\right)^{\frac{1}{|N_0|+|N_1|}} = \dot{P}_J^{st} \quad \text{if } s, t \in R_1$$

and, thus, corresponds to the bilateral Jevons price level of regions $s$ and $t$, $\dot{P}_J^{st}$. In contrast, the GEKS-Jevons price level estimator, $\widehat{P}_{\text{GEKS}}^{st}$, equals $\dot{P}_J^{st}$ only if $p_i^t = p_i^s$ for all products $i \in N_0 \cup N_1$. As a result, it seems that the CPD method assigns a higher priority on the "accuracy" of intragroup price levels while the GEKS-Jevons method treats intragroup and intergroup price levels as equally important. This leads to the question of how the CPD and GEKS-Jevons price level estimators behave in a generalised setting, namely when prices are randomly rather than group-wise missing.

## 4.2 A generalised setting: Simulation of artificial price data

In Section 2, it was shown that the CPD method as well as the GEKS-Jevons method can be derived from the same data generating process (DGP) in Equation (2). In the following, we exploit the DGP to create artificial price data that can be used for a deeper comparison between the CPD and the GEKS-Jevons price level estimators. Kackar and Harville (1984, p. 860) recommend including a relatively simple estimator into the comparison of the error metrics. In our case, the logarithm of the Jevons index, $\dot{\alpha}_J^t = \ln \dot{P}_J^{1t}$, would be a natural choice that serves as our baseline in the following.

We conduct a Monte Carlo analysis with $L = 2{,}000$ iterations ($l = 1, 2, \ldots, L$). We set the

[18] For intergroup comparisons, the bilateral Jevons price level is defined by $\dot{P}_J^{st} = \prod_{i \in N_0} \left(p_i^t / p_i^s\right)^{\frac{1}{|N_0|}}$. When $\prod_{r \in R_1} \Lambda_1^{sr} \gtrless 1$ and $\prod_{r \in R_2} \Lambda_2^{rt} \lessgtr 1$, it is technically possible that the CPD price levels approximate $\dot{P}_J^{st}$ closer. In Appendix B.1, however, it is shown that the GEKS-Jevons price level estimates are in most cases closer to $\dot{P}_J^{st}$ than the corresponding CPD price levels.

number of regions in each iteration to $R = 30$ in order to receive a constant amount of price level estimates. In each region, there are the same $N = 50$ products ($i = 1,\ldots,N$) available. The price levels, $P^t$ ($t = 2,\ldots,R$), are drawn independently for each region and iteration from a log-normal distribution with $P^t \sim LN\left(\mu = 0, \sigma^2 = 0.02\right)$. In addition, we exogenously fix the price level of the base region to one, i.e. $P^1 = 1$. This setting ensures a sufficient fluctuation around the base region's price level while the maximum price level spread between the most expensive and the cheapest region is roughly four. Furthermore, we assume that the error term is *iid* with $u_i^{sr} = \ln \epsilon_i^{sr} \sim N\left(\mu = 0, \sigma^2 = 0.04\right)$. As a result, we obtain $L = 2{,}000$ data sets with regional price ratios on the product level.

We apply the GEKS-Jevons method to each of these simulated data sets and obtain a set of transitive price levels, $\widehat{\alpha}^t_{\text{GEKS}}$. The CPD method, however, requires absolute prices rather than price ratios. Therefore, we additionally assume to know the price for each product in at least one region. This enables us to compute all of the remaining absolute prices, to transform these into a full price matrix and, consequently, to apply the CPD method as well. As a result, we receive the logarithmic price level estimators, $\widehat{\alpha}^t_{\text{CPD}}$ and $\widehat{\alpha}^t_{\text{GEKS}}$ ($t = 2,\ldots,R$), that were computed from exactly the same price data. Moreover, from the DGP, we also know the "true" logarithmic price levels, $\alpha^t = \ln\left(P^t/P^1\right)$, that were used within the simulation of each data set. Thus, we are able to evaluate the performance of the estimators in terms of bias and root mean squared error (RMSE):

$$\text{Bias}\left(\widehat{\alpha}^t_m\right) = \frac{1}{L} \cdot \sum_{l=1}^{L} \left(\widehat{\alpha}^t_{m,l} - \alpha^t_l\right) \quad \text{and} \quad \text{RMSE}\left(\widehat{\alpha}^t_m\right) = \sqrt{\frac{1}{L} \cdot \sum_{l=1}^{L} \left(\widehat{\alpha}^t_{m,l} - \alpha^t_l\right)^2},$$

where $L = 2{,}000$ is the number of simulation runs and $m \in \{\text{CPD, GEKS, Jevons}\}$.

We know from Section 2 that the CPD and GEKS-Jevons price level estimators, $\widehat{\alpha}^t_{\text{CPD}}$ and $\widehat{\alpha}^t_{\text{GEKS}}$, coincide when no prices are missing. Moreover, we know that they also coincide with the logarithm of the Jevons index, $\dot{\alpha}^t_{\text{J}}$. Consequently, regardless of how many simulations we would run, the estimated bias as well as the estimated RMSE is the same for these three estimators. However, when prices are missing, it is well known that the simple Jevons index no longer generates transitive price levels (e.g. ILO *et al.*, 2004, p. 498). In addition, it was shown in Section 4.1 that the transitive price levels produced by the CPD and GEKS-Jevons methods differ. Therefore, in order to evaluate the performance of the price level estimators under these circumstances, we incorporate gaps into our simulated price data by dropping prices for certain products and regions. The selection of the prices that are removed happens randomly, but is restricted to three conditions. First, no matter how many prices are removed, the price data must stay connected in order to ensure the feasibility of price level computations. Second, for each product, prices are always available in at least two different

| Gaps | Bias | | | RMSE | | |
|---|---|---|---|---|---|---|
| | CPD | GEKS | Jevons | CPD | GEKS | Jevons |
| 0% | 0.00025 | 0.00025 | 0.00025 | 0.01942 | 0.01942 | 0.01942 |
| 25% | -0.00009 | -0.00007 | -0.00022 | 0.02304 | 0.02306 | 0.02608 |
| 50% | 0.00005 | 0.00002 | -0.00058 | 0.02911 | 0.02948 | 0.04067 |
| 60% | 0.00043 | 0.00051 | -0.00011 | 0.03315 | 0.03409 | 0.05272 |
| 70% | -0.00011 | -0.00017 | 0.00016 | 0.03954 | 0.04163 | 0.07503 |
| 80% | -0.00048 | -0.00108 | -0.00601 | 0.05282 | 0.05726 | 0.10775 |

**Table 1:** Estimated bias and RMSE by percentage of missing prices for the CPD, GEKS-Jevons and Jevons price level estimators, respectively. Calculations on the basis of $L = 2{,}000$ simulated price data sets with $R = 30$ regions and $N = 50$ products.

regions. Third, the deletion of prices is path-dependent for a single price data set.[19]

Table 1 contains the simulation results in terms of bias and RMSE for the three estimators. It further illustrates how the two error metrics change when the gaps in our artificial price data gradually increase from 0% (no missing prices) to 80% (highly fragmented). As can be seen, the estimated bias is roughly zero and, thus, indicates that the estimators are unbiased. The estimated RMSE, on the other hand, increases for each estimator in reaction to an increased share of gaps in our price data. Not surprisingly, it is the highest for the simple Jevons index. The CPD and GEKS-Jevons estimators clearly outperform the Jevons index in terms of efficiency.[20]

The simulation setting leading to the results in Table 1 represents the case when all regions may exhibit gaps in the price data (including the base region $s$ and the comparison region $t$ of some estimated price level, $\widehat{\alpha}_m^t = \ln \widehat{P}_m^{st}$). Nevertheless, it neglects scenarios where either the base region, the comparison region or both of them provide full price information while the other regions do not. Therefore, we extended our simulation study by these scenarios. The overall simulation results can be found in Table 4 of the Appendix. A subset of them in terms of the RMSE is visualised in Figure 1. The first panel (from left to right) depicts the RMSE that arises when both, the base as well as the comparison region, provide full price information while all other regions in the data set may exhibit gaps. As can be seen, the RMSE of the CPD method and the simple Jevons index coincide.[21] Moreover, it does not change when prices are missing in other regions (represented by the horizontal line). This, in contrast, is not true for the RMSE of the GEKS-Jevons method. The second and third panels highlight the case when either the comparison or the base region is the only region

[19] Path dependency in this context means that prices that are already missing remain missing in the updated price data when we further increase the number of gaps. Specifically, it ensures that the impact of a gradual increase in the number of gaps can be properly evaluated.

[20] In Table 4 of the Appendix, it is shown that this is also true for other error metrics.

[21] This result traces back to identical price level estimates and, thus, indicates that the CPD price level estimator of two regions that provide full price information is defined as a Jevons index.

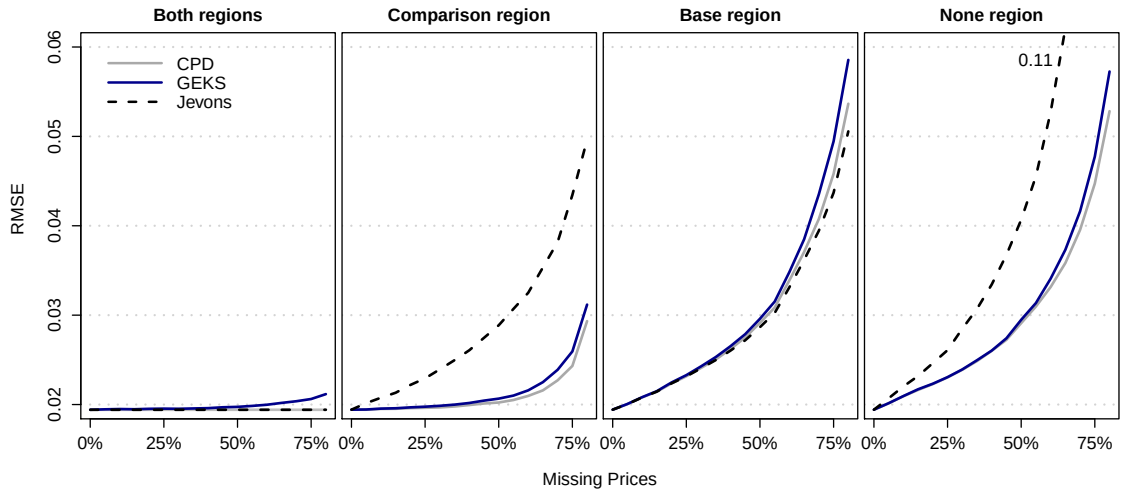


**Figure 1:** RMSE (vertical axis) by percentage of missing prices (in %, horizontal axis), index method and scenario (four panels). Calculations on the basis of $L = 2{,}000$ simulated price data sets with $R = 30$ regions and $N = 50$ products.

that provides full price information. Interestingly, both the CPD and GEKS-Jevons estimators perform far better when the prices are fully available in the comparison region rather than in the base region. Lastly, the fourth panel captures the RMSE values of Table 1 where all regions may exhibit gaps.

Overall, Figure 1 shows that the baseline RMSE of the Jevons index vastly increases as soon as prices are missing in at least one of the two regions under consideration (see dotted line in panels two to four). Moreover, the RMSE of the CPD price level estimator lies slightly below that of the GEKS-Jevons estimator in all four scenarios. The change associated with an increased share of missing prices, however, is fairly similar. This result is not surprising for two reasons. First, from Equation (15), we know that the CPD and GEKS-Jevons price level estimates differ only due to a different weighting of the correction terms. Now, even in this more general setting, the estimators $\widehat{\alpha}^t_{\text{CPD}}$ and $\widehat{\alpha}^t_{\text{GEKS}}$ are almost perfectly correlated.[22] Second, and perhaps more importantly, the deletion of prices within the simulation happened randomly, with the result that the gaps in our price data are uniformly distributed among the regions and products. In practice, however, this is a rather unrealistic situation as regions provide price information at varying frequencies. Similarly, specific products are less frequently available across regions than others. Therefore, we adapt our simulation study in the next section to a more realistic setting.

[22] Their correlation does not fall below 0.99, including in cases when 80% of the prices are missing. In contrast, the correlation of $\left(\widehat{\alpha}^t_{\text{CPD}}, \dot{\alpha}^t_{\text{J}}\right)$ and $\left(\widehat{\alpha}^t_{\text{GEKS}}, \dot{\alpha}^t_{\text{J}}\right)$ drops to nearly 0.84 in each case.

## 4.3 A more realistic setting with official micro price data

The official CPI in Germany is constructed as a stratified, non-random sample.[23] The prices of narrowly defined products are collected on a monthly basis in different regions, outlet types (e.g. supermarkets, discount stores) and basic headings (e.g. rice, milk). The actual collection of the price data is mainly carried out by the Statistical Offices of the Federal States (Statistische Landesämter) in selected regions of Germany. These data are supplemented by the Federal Statistical Office (Statistisches Bundesamt) which gathers the prices of products that are known to be regionally identical (e.g. books and cigarettes) or affected by particularly complex pricing policies (e.g. package holidays).

We have the privilege to work with a subset of these CPI data that was provided to us by the Research Data Centre (RDC) of the Federal Statistical Office and Statistical Offices of the Federal States. The price data were collected in $R = 19$ Bavarian regions in May 2011 (see left panel of Figure 2). They contain 23,642 consumer prices for goods, services and rents that are divided at the lowest classification level into 607 basic headings. The basic headings' expenditure weights add up to 71.79% of the German consumption basket.[24] Moreover, the data set contains information about the region where a price was collected. A unique identifier of the product for which that price was observed, however, is missing. Instead, semi-structured product descriptions are available (e.g. Behrmann, Deml and Linz, 2009, pp. 5-6; Zimmer, 2016, pp. 44-45). These include information about the product's amount (e.g. the weight or quantity), the respective unit of measurement (e.g. litre) and, subject to the basic heading, a number of supplementary characteristics like the brand or the packaging. In addition, special offer prices and the outlet type where the price was observed are indicated.

The collected rents for flats and single-family houses in the data set are accompanied by much richer "product descriptions" compared to those for goods and services. Therefore, one would typically draw on more sophisticated methods for the computation of regional price levels than the simple CPD and GEKS-Jevons methods (e.g. more complex hedonic regressions). However, since our simulation analysis concentrates on the latter two, we omit the rent data from that analysis. As a result, 21,783 price observations in $B = 601$ basic headings (expenditure weight: 51.05%) remain. For those basic headings, we rely on the product descriptions to define directly comparable products as precisely as possible. The choice of how narrowly we define such a product, however, is left to us and is thus more or less subjective. Therefore, we distinguish the following evaluation of the estimators' performances

[23] Rents are a subcategory of the CPI. In contrast to the prices for goods and services, however, they are collected from a stratified random sample since 2016 (Goldhammer, 2016, pp. 93-95).

[24] The prices collected by the Federal Statistical Office are not included in the data set. They add up, together with a small fraction of seasonal products, to the remaining expenditure weight.

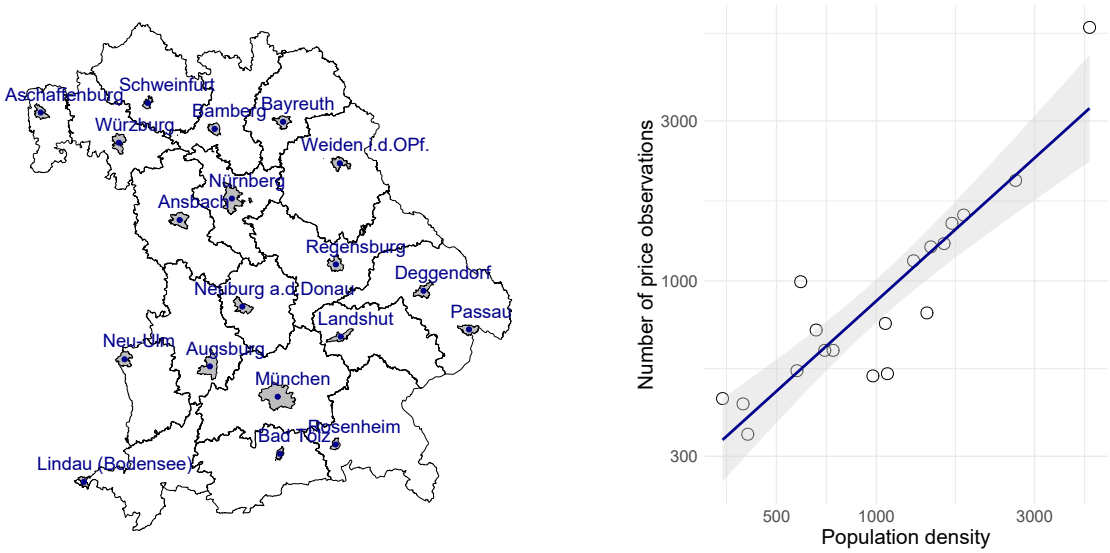


**Figure 2:** Bavarian regions where prices were collected in 2011 (left panel, grey shaded areas) and relationship between population density and number of collected prices across those regions (right panel, logarithmic scale). Source: RDC of the Federal Statistical Office and Statistical Offices of the Federal States, Consumer Price Index, May 2011, own calculations.

by the level of product definition. For price comparisons at the *product level* (a product is defined as narrowly as possible by all available characteristics), we identify 1,291 unique products that are priced in at least two different regions. In contrast, this number reduces to 652 at the *outlet level* (a product is defined only by the outlet type within a basic heading) and to 371 at the *basic heading level* (no product definition; all prices within a basic heading are assumed to be directly comparable). Weinand and Auer (2019, pp. 31-32) speak in this context of "simplified compilation procedures", since a definition at the outlet or at the basic heading level does not require any prior processing of the product descriptions.

As in the previous section, we perform a Monte Carlo analysis with $L = 2{,}000$ iterations ($l = 1, 2, \ldots, L$). This time, however, we do not randomly introduce gaps into our price data. Instead, we mimic the underlying structure of the CPI data set, i.e., we create artificial basic headings that adopt the observed basic headings' structure. In this way, we take into account that the number of collected prices varies by region (see right panel of Figure 2). More specifically, it is positively correlated with the population density. Those regions with a relatively low population density do not provide prices for each basic heading. As a result, most of the basic headings in the data set are incomplete.

Our simulation strategy is as follows. First, we randomly choose one of Eurostat's main HICP special aggregates.[25] Second, within the aggregate, we randomly select $N = 10$ spe-

[25] The likelihood of choosing either (1) *processed food, alcohol and tobacco*, (2) *unprocessed food*, (3) *energy*, (4) *non-energy industrial goods* or (5) *services* depends on the relative frequency of these aggregates in the underlying CPI data.

cific products from the original CPI data set without replacement. This setting ensures that unprocessed food, say, is not mixed up with services. Third, we add to these products the corresponding regions that originally collected prices. Consequently, we receive a new composition of products and regions. The regional distribution of available product prices, however, is adopted from the original price data. Lastly, we add artificial prices. For this purpose, we sample independently for each iteration the true regional price levels, $P^t$ $(t = 1,\ldots,R)$, and error terms, $u_i^{sr}$, in line with the DGP in (2):

$$P^t \sim LN\left(\mu = 0, \sigma^2 = 0.02\right) \quad \text{and} \quad u_i^{sr} = \ln \epsilon_i^{sr} \sim N\left(\mu = 0, \sigma^2 = 0.04\right) .$$

In this way, we generate $L$ = 2,000 data sets. On average, 66.7% of the prices are missing within these data sets. The distribution of available prices across the regions is highly correlated with that of the original CPI data $\left(\rho = 0.78\right)$. We apply the CPD, the GEKS-Jevons and the simple Jevons estimators to each of these data sets. Their performance in terms of bias and RMSE is documented in Table 2.

| **Product definition** | **Bias** | | | **RMSE** | | |
|---|---|---|---|---|---|---|
| | CPD | GEKS | Jevons | CPD | GEKS | Jevons |
| Product level | 0.00044 | 0.00115 | -0.00049 | 0.22680 | 0.23971 | 0.24453 |
| Outlet level | -0.00395 | -0.00239 | -0.00271 | 0.28745 | 0.29268 | 0.29479 |
| Basic head. level | -0.00387 | -0.00387 | -0.00387 | 0.29566 | 0.29566 | 0.29566 |

**Table 2:** Simulation results in terms of bias and RMSE for the CPD, GEKS-Jevons and Jevons price level estimators, respectively. Calculations on the basis of $L$ = 2,000 incomplete price data sets with $N$ = 10 products.

The simulation results show how the regional price level estimators perform on "real world data". Moreover, they demonstrate the relevance of the product definition level for the estimation efficiency. As can be seen, the estimated bias and RMSE are the same for the three estimators when there is no product differentiation within a basic heading (see line "Basic heading level"). Otherwise, with a product differentiation, the RMSEs differ. Strictly speaking, they slightly decrease for product definitions at the outlet level and considerably at the much narrower product level. In both cases, the RMSE is the lowest for the CPD method.

Lastly, it is worthwhile to note that the RMSE comparison between the different levels of product definition also depends on the regional volatility of prices, that is, how much the prices of some product fluctuate across the regions. Unsurprisingly, when we lowered the regional volatility of the prices in our simulation study, the RMSE values dropped to roughly 0.22, including for product definitions at the outlet and basic heading level.[26] As a conse-

[26] One could imagine a basic heading with identical product prices in all regions. Independent of the level of product definition, the regional price level estimates would be the same.

quence, in future work with official CPI data, one could rely on product definitions at the outlet level for those basic headings with low regional price fluctuations. In contrast, for those basic headings with high regional price fluctuations, the estimation efficiency clearly benefits from a narrow product definition. This mixed strategy would heavily reduce the costly data preprocessing reported by Weinand and Auer (2019, pp. 9-11).

# 5 Concluding remarks

The main goal of this paper was to expand the theoretical foundations of the stochastic approach to index numbers in light of spatial price comparisons. To this end, we examined the most prominent representatives of the stochastic approach: the CPD method and the GEKS method. In particular, we analysed the impact of missing prices below the basic heading on the estimation of regional price levels. For a specific case of missing prices, we derived the formula underlying the price level estimates and showed that differences between the CPD and GEKS-Jevons methods stem solely from the assignment of a different weighting pattern. Moreover, using simulation techniques, we studied the statistical properties of the CPD and GEKS-Jevons price level estimators in terms of bias and RMSE. Our results revealed lower RMSE values for the CPD method in four tested scenarios. For spatial price comparisons, it is worthwhile keeping in mind that the estimation efficiency improved especially in those cases where the comparison region provided complete prices.

Notwithstanding these differences, our results demonstrate that the regional price level estimates of the CPD and the GEKS-Jevons methods are closely related. Therefore, we do not want to speak generally in favour of one of the two methods. However, two thoughts are worth mentioning. First, from a practical point of view, statistical offices collect absolute prices rather than price ratios or price index numbers. These price data form the building blocks for CPI measurement purposes and would be a unique data source for the calculation of regional price levels as well. The application of standard regression techniques to these raw data (CPD method) therefore seems more straightforward than first converting prices into bilateral index numbers (GEKS method). In addition, the regression approach underlying the CPD method would allow for extensions in the sense of more careful quality adjustments, for example, by including additional product characteristics (e.g. Balk, 2008, p. 258). Second, from a statistical point of view, we showed in Section 4 that the estimation efficiency of the CPD method outperforms that of the GEKS-Jevons method, especially in the case of substantial gaps in the price data. This result strengthens the application of the CPD method below the basic heading level where data gaps are frequently an issue.

In our second simulation study, we used a subset of the price data underlying Germany's

CPI. These price data come with precise but relatively unstructured product descriptions. The importance of these product descriptions for spatial price comparisons is widely documented in the literature (e.g. ILO *et al.*, 2004, p. 73; World Bank, 2013, p. 590), as they enable price statisticians to identify directly comparable products as precisely as possible. Utilising the product descriptions would be a natural choice for statistical offices to compare only like with like across regions and thereby avoid any distortions in their estimates of regional price levels. However, one drawback that statistical offices would face is the extensive preprocessing of the product descriptions (e.g. Weinand and Auer, 2019, pp. 9-11). Our findings have practical relevance for two reasons as they address both issues and may therefore serve as guidance for statistical offices carrying out spatial price comparisons. First, our simulation results underline the importance of the product definition for the estimation efficiency. In particular, they show an improvement in the estimation efficiency owing to more narrowly defined products, though this is usually accompanied by more gaps in the price data. Second, our simulation results reveal that statistical offices may reduce the workload associated with preprocessing the product descriptions by following a mixed strategy that takes into account the regional price volatilities of the basic headings.[27] For those basic headings with a low regional price volatility, statistical offices could rely on looser product definitions, such as the outlet level, which do not require any data preprocessing.

We conclude with two points that are worth mentioning but beyond the focus of this paper. First of all, Hajargasht and Rao (2019) recently examined the theory on multilateral index numbers in light of graph theory. Although they do not explicitly mention the CPD and GEKS methods, their derivations might be relevant to our setting as well. Basically, not only does the percentage of missing prices directly influence the efficiency of the price level estimates, the manner in which the prices and thus the gaps within the collected data are distributed among the regions ("degree of connectedness") are also relevant. This consideration may give rise to future research. With respect to simulation analyses, greater attention in future work could be focused towards different patterns in the price data (e.g. spatially correlated prices). Moreover, the adoption of various extensions of multilateral index number methods such as the CPRD method proposed by Cuthbert and Cuthbert (1988, pp. 55-58) into the simulation framework would be interesting. As information on the representativity of some product is usually lacking in the price data gathered by statistical offices, we focused our analyses on the straightforward CPD and GEKS-Jevons methods.

[27] Expert judgement of price statisticians on specific basic headings could be included as well.

# A Derivation of price level estimators

## A.1 Complete price data

In the following, we consider a complete price matrix with exactly one price, $p_i^r$, per product ($i = 1,2,\ldots,N$) and region ($r = 1,2,\ldots,R$). In Sections A.1.1 to A.1.3, we derive the price level estimators of the CPD, CD and GEKS-Jevons methods.

### A.1.1 CPD method

The estimated regression model of (7) can be cast in general matrix notation. To that end, we define the vectors $\mathbf{y} = \left(\mathbf{y}_1 \ldots \mathbf{y}_N\right)'$ and $\widehat{\mathbf{u}} = \left(\widehat{\mathbf{u}}_1 \ldots \widehat{\mathbf{u}}_N\right)'$, with $\mathbf{y}_i = \left(\ln p_i^1 \ldots \ln p_i^R\right)'$ and $\widehat{\mathbf{u}}_i = \left(\widehat{u}_i^1 \ldots \widehat{u}_i^R\right)'$ for products $i = 1,\ldots,N$. The $(NR \times (R-1+N))$-matrix $\mathbf{X}$ comprises the $R-1$ dummy variables $region^t$ ($t = 2,\ldots,R$) and the $N$ dummy variables $product_j$ ($j = 1,\ldots,N$). With complete price data, there are $N \cdot R$ observations and $(N-1)\cdot(R-1)$ degrees of freedom available. The OLS estimator, $\hat{\boldsymbol{\beta}}_{\text{CPD}}$, can be written as

$$\hat{\boldsymbol{\beta}}_{\text{CPD}} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_{\text{CPD}}^1 \\ \hat{\boldsymbol{\beta}}_{\text{CPD}}^2 \end{pmatrix} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \tag{17}$$

where $\hat{\boldsymbol{\beta}}_{\text{CPD}}^1 = \left(\widehat{\alpha}^2 \ldots \widehat{\alpha}^R\right)'$ contains the estimated logarithmic regional price levels and $\hat{\boldsymbol{\beta}}_{\text{CPD}}^2 = \left(\widehat{\beta}_1 \ldots \widehat{\beta}_N\right)'$ the estimated logarithmic product prices.

The $((R-1+N)\times(R-1+N))$-matrix $\mathbf{X}'\mathbf{X}$ in (17) takes the following form:

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} N \cdot \mathbf{I}_{R-1} & \mathbf{J}_{(R-1)\times N} \\ \mathbf{J}_{N\times(R-1)} & R \cdot \mathbf{I}_N \end{pmatrix},$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{J}$ is a matrix of ones. The diagonal of the upper left submatrix indicates the number of products that are observed in the respective region ($= N$), whereas the lower right submatrix indicates the number of regions in which the product has been observed ($= R$). The upper right and the lower left submatrices convey the same information. They specify for each combination of products and regions if a price is available ($= 1$) or not ($= 0$). Since we have a complete data set, each element in the submatrices has the value one. Using computation rules on block matrices (e.g. Horn and Johnson, 2012,

p. 18), the inverse of $\mathbf{X}'\mathbf{X}$ is defined by

$$\left(\mathbf{X}'\mathbf{X}\right)^{-1} = \begin{pmatrix} \frac{1}{N}\cdot(\mathbf{I}_{R-1}+\mathbf{J}_{R-1}) & -\frac{1}{N}\cdot\mathbf{J}_{(R-1)\times N} \\ -\frac{1}{N}\cdot\mathbf{J}_{N\times(R-1)} & \frac{1}{R}\cdot(\mathbf{I}_N+\frac{R-1}{N}\cdot\mathbf{J}_N) \end{pmatrix}. \tag{18}$$

The same result can be found in Rao and Hajargasht (2016, p. 418).

The first block of the $((R-1+N)\times 1)$-vector $\mathbf{X}'\mathbf{y}$ contains for each region $(t=2,\ldots,R)$ the sum of its logarithmic product prices while the second block comprises for each product $(j=1,\ldots,N)$ the regional sum of logarithmic prices:

$$\mathbf{X}'\mathbf{y} = \left(\sum_{i=1}^{N}\ln p_i^2 \ldots \sum_{i=1}^{N}\ln p_i^R \quad \sum_{r=1}^{R}\ln p_1^r \ldots \sum_{r=1}^{R}\ln p_N^r\right)'. \tag{19}$$

Inserting (18) and (19) in (17) yields the OLS estimator $\hat{\boldsymbol{\beta}}_{\mathrm{CPD}}$. The estimated logarithmic price levels, $\widehat{\alpha}^2,\ldots,\widehat{\alpha}^R$, can be found in $\hat{\boldsymbol{\beta}}^1_{\mathrm{CPD}}$ by multiplying the top row of $\left(\mathbf{X}'\mathbf{X}\right)^{-1}$ with $\mathbf{X}'\mathbf{y}$:

$$\hat{\boldsymbol{\beta}}^1_{\mathrm{CPD}} = \left(\widehat{\alpha}^2\ldots\widehat{\alpha}^R\right)' = \begin{pmatrix} \frac{1}{N}\cdot\sum_{i=1}^{N}\ln\left(p_i^2/p_i^1\right) \\ \vdots \\ \frac{1}{N}\cdot\sum_{i=1}^{N}\ln\left(p_i^R/p_i^1\right) \end{pmatrix}.$$

Thus, taking anti-logs, $\exp\left(\hat{\boldsymbol{\beta}}^1_{\mathrm{CPD}}\right)$ yields the estimated regional price levels as a geometric average of the product prices in region $t$ (with $t=2,\ldots,R$) relative to those of the base region $(t=1)$. Furthermore, multiplying the bottom row of $\left(\mathbf{X}'\mathbf{X}\right)^{-1}$ with $\mathbf{X}'\mathbf{y}$ results in

$$\hat{\boldsymbol{\beta}}^2_{\mathrm{CPD}} = \left(\widehat{\beta}_1\ldots\widehat{\beta}_N\right)' = \begin{pmatrix} \frac{1}{R}\cdot\sum_{r=1}^{R}\ln p_1^r - \widehat{\alpha}^r \\ \vdots \\ \frac{1}{R}\cdot\sum_{r=1}^{R}\ln p_N^r - \widehat{\alpha}^r \end{pmatrix}.$$

Therefore, $\exp\left(\hat{\boldsymbol{\beta}}^2_{\mathrm{CPD}}\right)$ is defined as a geometric average of product $j$'s regional prices (with $j=1,\ldots,N$), deflated by the corresponding regional price level, $\exp\left(\widehat{\alpha}^r\right)$.

Using (18), the estimated variance-covariance-matrix is given by

$$\widehat{\mathbf{V}}\left(\hat{\boldsymbol{\beta}}^1_{\mathrm{CPD}}\right) = \widehat{\sigma}^2_{\mathrm{CPD}}\left(\mathbf{X}'\mathbf{X}\right)^{-1} = \widehat{\sigma}^2_{\mathrm{CPD}}\cdot\frac{1}{N}\cdot(\mathbf{I}_{R-1}+\mathbf{J}_{R-1}) \tag{20}$$

where the estimated variances of $\widehat{\alpha}^2,\ldots,\widehat{\alpha}^R$ can be found on the diagonal, with $\widehat{\mathrm{var}}\left(\widehat{\alpha}^t\right) = (2/N)\cdot\widehat{\sigma}^2_{\mathrm{CPD}}$ for regions $t=2,\ldots,R$.

### A.1.2 CD method

From (2), it is known that $u_i^{sr} = \ln\left(p_i^r / p_i^s\right) - \ln\left(P^r / P^s\right)$. Hence,

$$\begin{aligned} -u_i^{sr} &= \ln\left(p_i^s / p_i^r\right) - \ln\left(P^s / P^r\right) \\ &= u_i^{rs} \end{aligned} \tag{21}$$

for all regions $r$ and $s$. Furthermore, with some arbitrary third region $v$, it follows that

$$\begin{aligned} u_i^{vr} - u_i^{vs} &= \ln\left(p_i^r / p_i^v\right) - \ln\left(P^r / P^v\right) - \ln\left(p_i^s / p_i^v\right) + \ln\left(P^s / P^v\right) \\ &= \ln\left(p_i^r / p_i^s\right) - \ln\left(P^r / P^s\right) \\ &= u_i^{sr} \ . \end{aligned} \tag{22}$$

The variance of the disturbances is assumed to be identical among the products and regions. Therefore, $\mathbb{E}\left[\left(u_i^{sr}\right)^2\right] = \mathbb{E}\left[\left(u_i^{sv}\right)^2\right] = \mathbb{E}\left[\left(u_i^{vr}\right)^2\right] = \sigma^2$. Using (21), the relationship in (22) can be rewritten as $u_i^{vr} = u_i^{sr} - u_i^{sv}$. Taking the expected value of this expression yields

$$\begin{aligned} \mathbb{E}\left[\left(u_i^{vr}\right)^2\right] &= \mathbb{E}\left[\left(u_i^{sr} - u_i^{sv}\right)^2\right] \\ &= \mathbb{E}\left[\left(u_i^{sr}\right)^2\right] + \mathbb{E}\left[\left(u_i^{sv}\right)^2\right] - 2 \cdot \mathbb{E}\left[u_i^{sr} u_i^{sv}\right] \\ \sigma^2 &= \sigma^2 + \sigma^2 - 2 \cdot \operatorname{cov}\left(u_i^{sr},\ u_i^{sv}\right) \\ \operatorname{cov}\left(u_i^{sr},\ u_i^{sv}\right) &= \frac{1}{2}\sigma^2 \end{aligned} \tag{23}$$

for regions $r \neq v$. Furthermore, it is assumed that the disturbances of two different products $i$ and $j$ are uncorrelated:

$$\operatorname{cov}\left(u_i^{sr},\ u_j^{sr}\right) = \operatorname{cov}\left(u_i^{sr},\ u_j^{sv}\right) = 0 \ . \tag{24}$$

The estimated regression model of (10) can be put in general matrix notation. We set region $s = 1$ as the reference for the price ratios, that is, $\ln\left(p_i^r / p_i^1\right)$ for products $i = 1, \ldots, N$ and $r = 2, \ldots, R$. Accordingly, we define the vector of residuals $\hat{\mathbf{u}} = \left(\hat{\mathbf{u}}_1 \ldots \hat{\mathbf{u}}_N\right)'$ and the vector of price ratios $\mathbf{y} = \left(\mathbf{y}_1 \ldots \mathbf{y}_N\right)'$, with $\hat{\mathbf{u}}_i = \left(\hat{u}_i^{12} \ldots \hat{u}_i^{1R}\right)'$ and $\mathbf{y}_i = \left(\ln\left(p_i^2 / p_i^1\right) \ldots \ln\left(p_i^R / p_i^1\right)\right)'$. The $(N(R-1) \times (R-1))$-matrix $\mathbf{X}$ contains the $R-1$ dummy variables $\left(region^t - \widetilde{region^t}\right)$ for regions $t = 2, \ldots, R$. With complete price data, there are $N \cdot (R-1)$ observations and $(N-1) \cdot (R-1)$ degrees of freedom available.

From (23) and (24), it follows that the disturbances of the Country-Dummy method are autocorrelated blockwise (see also Summers, 1973, p. 10). The variance-covariance-matrix

of the disturbances, $\mathbf{V}(\mathbf{u})$, can be written as

$$\begin{aligned}\mathbf{V}(\mathbf{u}) &= \sigma^2 \ (\mathbf{I}_N \otimes \mathbf{V}_i) \\ &= \sigma^2 \, \Omega \, ,\end{aligned} \tag{25}$$

where $\mathbf{I}$ is a $(N \times N)$-identity matrix and $\mathbf{V}_i = \frac{1}{2} \cdot (\mathbf{I}_{R-1} + \mathbf{J}_{R-1})$ for products $i = 1, \ldots, N$. OLS would lead to inefficient estimates of the regression coefficients. However, since the form of the $(N(R-1) \times N(R-1))$-matrix $\Omega$ is known, GLS can be applied to obtain unbiased and efficient estimates.

The GLS estimator is defined by

$$\hat{\boldsymbol{\beta}}_{\text{CD}}^{\text{GLS}} = \left(\hat{\alpha}^2 \ldots \hat{\alpha}^R\right)' = \left(\mathbf{X}' \Omega^{-\mathbf{1}} \mathbf{X}\right)^{-1} \mathbf{X}' \Omega^{-\mathbf{1}} \mathbf{y} \, . \tag{26}$$

The inverse of $\Omega$ in (25) can be expressed by the Kronecker product $\Omega^{-1} = \mathbf{I}_N \otimes \mathbf{V}_i^{-1}$. Following Graybill (1983, pp. 190-193), it can be shown that $\mathbf{V}_i^{-1} = 2 \cdot \left(\mathbf{I}_{R-1} - \frac{1}{R} \cdot \mathbf{J}_{R-1}\right)$. As a result, the $((R-1) \times (R-1))$-matrix $\mathbf{X}' \Omega^{-\mathbf{1}} \mathbf{X}$ can be written as

$$\left(\mathbf{X}' \Omega^{-\mathbf{1}} \mathbf{X}\right) = N \cdot \underbrace{2 \cdot \left(\mathbf{I}_{R-1} - \frac{1}{R} \cdot \mathbf{J}_{R-1}\right)}_{=\mathbf{V}_i^{-1}} .$$

The inverse of $\left(\mathbf{X}' \Omega^{-\mathbf{1}} \mathbf{X}\right)$ is defined by

$$\left(\mathbf{X}' \Omega^{-\mathbf{1}} \mathbf{X}\right)^{-1} = \frac{1}{N} \cdot \underbrace{\frac{1}{2} \cdot (\mathbf{I}_{R-1} + \mathbf{J}_{R-1})}_{=\mathbf{V}_i} . \tag{27}$$

Furthermore, the $((R-1) \times 1)$-vector $\mathbf{X}' \Omega^{-\mathbf{1}} \mathbf{y}$ can be written as

$$\mathbf{X}' \Omega^{-1} \mathbf{y} = 2 \cdot \begin{pmatrix} \sum_{i=1}^{N} \ln\left(p_i^2 / p_i^1\right) - \frac{1}{R} \cdot \sum_{r=2}^{R} \sum_{i=1}^{N} \ln\left(p_i^r / p_i^1\right) \\ \vdots \\ \sum_{i=1}^{N} \ln\left(p_i^R / p_i^1\right) - \frac{1}{R} \cdot \sum_{r=2}^{R} \sum_{i=1}^{N} \ln\left(p_i^r / p_i^1\right) \end{pmatrix}. \tag{28}$$

Inserting (27) and (28) in (26) yields the definition of the GLS estimator:

$$\hat{\boldsymbol{\beta}}_{\text{CD}}^{\text{GLS}} = \begin{pmatrix} \frac{1}{N} \cdot \sum_{i=1}^{N} \ln\left(p_i^2 / p_i^1\right) \\ \vdots \\ \frac{1}{N} \cdot \sum_{i=1}^{N} \ln\left(p_i^R / p_i^1\right) \end{pmatrix} .$$

Thus, like the CPD method, the estimated regional price levels in $\exp\left(\hat{\boldsymbol{\beta}}_{\text{CD}}^{\text{GLS}}\right)$ are defined as a

geometric average of the product price ratios between regions $t = 2,\ldots,R$ and the base region ($t = 1$), respectively.

The estimated variance of the disturbances is

$$\widehat{\sigma}^2_{\text{CD}} = \frac{\hat{\mathbf{u}}'\Omega^{-\mathbf{1}}\hat{\mathbf{u}}}{(N-1)\cdot(R-1)}\,.$$

From $\widehat{u}_i^{1r} = \widehat{u}_i^r - \widehat{u}_i^1$, it follows that $\hat{\mathbf{u}}'\Omega^{-\mathbf{1}}\hat{\mathbf{u}} = 2\cdot\sum_{r=1}^{R}\sum_{i=1}^{N}\left(\widehat{u}_i^r\right)^2$. Thus, $\widehat{\sigma}^2_{\text{CD}} = 2\cdot\widehat{\sigma}^2_{\text{CPD}}$. Using (27), the estimated variance-covariance-matrix of the GLS estimator is given by

$$\begin{aligned}\widehat{\mathbf{V}}\left(\hat{\boldsymbol{\beta}}^{\text{GLS}}_{\text{CD}}\right) &= \widehat{\sigma}^2_{\text{CD}}\left(\mathbf{X}'\Omega^{-\mathbf{1}}\mathbf{X}\right)^{-1}\\ &= \widehat{\sigma}^2_{\text{CD}}\cdot\frac{1}{N}\cdot\frac{1}{2}\cdot(\mathbf{I}_{R-1}+\mathbf{J}_{R-1})\\ &= \widehat{\sigma}^2_{\text{CPD}}\cdot\frac{1}{N}\cdot(\mathbf{I}_{R-1}+\mathbf{J}_{R-1})\,,\end{aligned} \tag{29}$$

which coincides with $\widehat{\mathbf{V}}\left(\hat{\boldsymbol{\beta}}^{1}_{\text{CPD}}\right)$ in (20). The estimated variances of $\widehat{\alpha}^2,\ldots,\widehat{\alpha}^R$ can be found on the diagonal, with $\widehat{\text{var}}\left(\widehat{\alpha}^t\right) = (1/N)\cdot\widehat{\sigma}^2_{\text{CD}} = (2/N)\cdot\widehat{\sigma}^2_{\text{CPD}}$ for regions $t = 2,\ldots,R$.

### A.1.3 GEKS method

The estimated regression model of (13) is converted in standard matrix notation. The vector $\mathbf{y}$ encompasses the logarithmic bilateral Jevons price index numbers, $\ln \dot{P}_{\text{J}}^{sr}$, for regions $s = 1,\ldots,(R-1)$ and $r = (s+1),\ldots,R$. With complete prices, there are $R\cdot(R-1)/2$ non-redundant, bilateral price index numbers available (Hill and Timmer, 2006, p. 368). The matrix $\mathbf{X}$ contains the $R-1$ dummy variables $\left(region^t - \widetilde{region^t}\right)$ for regions $t = 2,\ldots,R$. Thus, the degrees of freedom are $(R-2)\cdot(R-1)/2$. The OLS estimator

$$\hat{\boldsymbol{\beta}}_{\text{GEKS}} = \left(\widehat{\alpha}^2\ldots\widehat{\alpha}^R\right)' = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \tag{30}$$

yields the estimated regional price levels $\widehat{\alpha}^2,\ldots,\widehat{\alpha}^R$.

The $((R-1)\times(R-1))$-matrix $\mathbf{X}'\mathbf{X}$ contains on the diagonal for each region $t = 2,\ldots,R$ the number of bilateral index numbers. It can be written as

$$\mathbf{X}'\mathbf{X} = (R\cdot\mathbf{I}_{R-1} - \mathbf{J}_{R-1})\,,$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{J}$ is a matrix of ones. The inverse of $\mathbf{X}'\mathbf{X}$ is defined by

$$\left(\mathbf{X}'\mathbf{X}\right)^{-1} = \frac{1}{R}\cdot(\mathbf{I}_{R-1}+\mathbf{J}_{R-1})\,. \tag{31}$$

The $((R-1)\times 1)$-vector $\mathbf{X}'\mathbf{y}$ contains for each region $t=2,\ldots,R$ the sum of its logarithmic bilateral index numbers $\dot{P}_{\mathrm{J}}^{rt}$:

$$\mathbf{X}'\mathbf{y}=\left(\sum_{r=1}^{R}\ln\dot{P}_{\mathrm{J}}^{r2}\ \ldots\ \sum_{r=1}^{R}\ln\dot{P}_{\mathrm{J}}^{rR}\right)'. \tag{32}$$

Inserting (31) and (32) in (30) yields the OLS estimator of the GEKS-Jevons method:

$$\hat{\boldsymbol{\beta}}_{\mathrm{GEKS}}=\frac{1}{R}\cdot\begin{pmatrix}\sum_{r=1}^{R}\ln\dot{P}_{\mathrm{J}}^{r2}+\left[\sum_{r=1}^{R}\ln\dot{P}_{\mathrm{J}}^{r2}+\ldots+\sum_{r=1}^{R}\ln\dot{P}_{\mathrm{J}}^{rR}\right]\\ \vdots\\ \sum_{r=1}^{R}\ln\dot{P}_{\mathrm{J}}^{rR}+\left[\sum_{r=1}^{R}\ln\dot{P}_{\mathrm{J}}^{r2}+\ldots+\sum_{r=1}^{R}\ln\dot{P}_{\mathrm{J}}^{rR}\right]\end{pmatrix}. \tag{33}$$

It is well known in literature that the Jevons index satisfies the country-reversal test, that is, $\ln\dot{P}_{\mathrm{J}}^{rs}=-\ln\dot{P}_{\mathrm{J}}^{sr}$. Thus, with complete prices, it can be shown that the term in squared brackets simplifies to $\sum_{r=1}^{R}\ln\dot{P}_{\mathrm{J}}^{1r}$. Accordingly, (33) can be rewritten as

$$\hat{\boldsymbol{\beta}}_{\mathrm{GEKS}}=\begin{pmatrix}\frac{1}{R}\cdot\sum_{r=1}^{R}\left(\ln\dot{P}_{\mathrm{J}}^{r2}+\ln\dot{P}_{\mathrm{J}}^{1r}\right)\\ \vdots\\ \frac{1}{R}\cdot\sum_{r=1}^{R}\left(\ln\dot{P}_{\mathrm{J}}^{rR}+\ln\dot{P}_{\mathrm{J}}^{1r}\right)\end{pmatrix}. \tag{34}$$

Inserting (1), the definition of the Jevons index, in (34) and taking anti-logs yields the estimated regional price levels as a geometric average of the product price ratios between regions $t=2,\ldots,R$ and the base region ($t=1$), respectively. Therefore, the GEKS-Jevons approach leads to identical estimates of the regional price levels as the CPD and CD methods in the event of complete prices.

Using (31) yields the estimated variance-covariance-matrix as

$$\widehat{\mathbf{V}}\left(\hat{\boldsymbol{\beta}}_{\mathrm{GEKS}}\right)=\widehat{\sigma}^2_{\mathrm{GEKS}}\left(\mathbf{X}'\mathbf{X}\right)^{-1}=\widehat{\sigma}^2_{\mathrm{GEKS}}\cdot\frac{1}{R}\cdot\left(\mathbf{I}_{R-1}+\mathbf{J}_{R-1}\right), \tag{35}$$

with $\widehat{\mathrm{var}}\left(\widehat{\alpha}^t\right)=(2/R)\cdot\widehat{\sigma}^2_{\mathrm{GEKS}}$ for regions $t=2,\ldots,R$. With complete prices, the bilateral index numbers are transitive and therefore coincide with the multilateral index numbers. As a result, $\widehat{\sigma}^2_{\mathrm{GEKS}}=0$ and $\widehat{\mathrm{var}}\left(\widehat{\alpha}^t\right)=0$.

## A.2 Missing prices

In the following, we consider a price matrix with $R$ regions and $N$ products. We randomly divide the regions $r=2,\ldots,R$ into the nonempty and disjoint sets $R_1$ and $R_2$. The base region $r=1$ belongs, by definition, to set $R_1$. Similarly, we randomly divide the products

$i = 1, 2, \ldots, N$ into the nonempty and disjoint sets $N_0$, $N_1$ and $N_2$. We assume that the prices of products $i \in N_1$ are available only in regions $r \in R_1$ while the prices of products $i \in N_2$ are available only in regions $r \in R_2$. In contrast, the prices of products $i \in N_0$ are fully available in all regions. The price incidence matrix of this scenario is shown in Table 3. For illustration purposes, its entries are ordered column-wise by the product group and row-wise by the region group.

$$M = \begin{array}{c} \xrightarrow{\text{products}} \\ \left(\begin{array}{ccc:ccc:ccc} 1 & \cdots & 1 & 1 & \cdots & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & \cdots & 1 & 1 & \cdots & 1 & 0 & \cdots & 0 \\ \hdashline 1 & \cdots & 1 & 0 & \cdots & 0 & 1 & \cdots & 1 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & \cdots & 1 & 0 & \cdots & 0 & 1 & \cdots & 1 \end{array}\right) \end{array} \Bigg\downarrow \text{regions}$$

**Table 3:** Price incidence matrix of blockwise missing prices, indicating whether a specific price is available in the data (= 1) or not (= 0).

### A.2.1 CPD method

The OLS estimator, $\hat{\boldsymbol{\beta}}_{\text{CPD}}$, is defined in (17). Using block matrix notation, the inverse of the $((R-1+N)\times(R-1+N))$-matrix $\mathbf{X}'\mathbf{X}$ can be written as

$$\left(\mathbf{X}'\mathbf{X}\right)^{-1} = \begin{pmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}' & \mathbf{C} \end{pmatrix}, \tag{36}$$

where the matrix $\mathbf{X}_1 = (\mathbf{X}_{11}\, \mathbf{X}_{12})$ contains the $R-1$ dummy variables $region^t$ and the matrix $\mathbf{X}_2 = (\mathbf{X}_{20}\, \mathbf{X}_{21}\, \mathbf{X}_{22})$ the $N$ dummy variables $product_j$.[28] Furthermore, the $((R-1+N)\times 1)$-vector $\mathbf{X}'\mathbf{y}$ can be expressed by

$$\mathbf{X}'\mathbf{y} = \begin{pmatrix} \mathbf{X}_1'\mathbf{y} & \mathbf{X}_2'\mathbf{y} \end{pmatrix}', \tag{37}$$

where the $((R-1)\times 1)$-vector $\mathbf{X}_1'\mathbf{y}$ contains for each region $t$ (except the base region $t = 1$) the sum of its logarithmic prices. In contrast, the $(N\times 1)$-vector $\mathbf{X}_2'\mathbf{y}$ encompasses for each product $j$ the sum of its logarithmic prices. Inserting (36) and (37) in (17) results in

$$\hat{\boldsymbol{\beta}}^1_{\text{CPD}} = \mathbf{A}\,\mathbf{X}_1'\mathbf{y} + \mathbf{B}\,\mathbf{X}_2'\mathbf{y}, \tag{38}$$

[28] The matrix $\mathbf{X}_{11}$ encompasses the dummy variables $region^t$ for regions $t \in R_1 \wedge t \neq 1$ while the matrix $\mathbf{X}_{12}$ holds the dummy variables $region^t$ for regions $t \in R_2$. Similarly, $\mathbf{X}_{20}$ contains the dummy variables $product_j$ for products $j \in N_0$, $\mathbf{X}_{21}$ for $j \in N_1$ and $\mathbf{X}_{22}$ for $j \in N_2$.

the estimator of logarithmic regional price levels.

In the scenario under consideration, it can be shown that the $((R-1)\times(R-1))$-submatrix $\mathbf{X}_1'\mathbf{X}_1$ and the $((R-1)\times N)$-submatrix $\mathbf{X}_1'\mathbf{X}_2$ can be written as

$$\mathbf{X}_1'\mathbf{X}_1 = \begin{pmatrix} a\cdot\mathbf{I}_{|R_1|-1} & \mathbf{0}_{(|R_1|-1)\times|R_2|} \\ \mathbf{0}_{|R_2|\times(|R_1|-1)} & b\cdot\mathbf{I}_{|R_2|} \end{pmatrix}$$

and

$$\mathbf{X}_1'\mathbf{X}_2 = \begin{pmatrix} \mathbf{J}_{(|R_1|-1)\times|N_0|} & \mathbf{J}_{(|R_1|-1)\times|N_1|} & \mathbf{0}_{(|R_1|-1)\times|N_2|} \\ \mathbf{J}_{|R_2|\times|N_0|} & \mathbf{0}_{(|R_2|)\times|N_1|} & \mathbf{J}_{|R_2|\times|N_2|} \end{pmatrix},$$

where $\mathbf{0}$ is the null matrix. The scalars $a = |N_0| + |N_1|$ and $b = |N_0| + |N_2|$ show the number of available price observations per region. Inserting the definitions of $\mathbf{X}_1'\mathbf{X}_1$ and $\mathbf{X}_1'\mathbf{X}_2$ into (36) yields the matrices

$$\mathbf{A} = a^{-1}\cdot\begin{pmatrix} \mathbf{I}+\mathbf{J} & \mathbf{J} \\ \mathbf{J} & (a/b)\cdot\mathbf{I}+(1+c)\cdot\mathbf{J} \end{pmatrix} \text{ and } \mathbf{B} = -a^{-1}\cdot\begin{pmatrix} \mathbf{J} & \mathbf{J} & \mathbf{J} \\ d\cdot\mathbf{J} & e\cdot\mathbf{J} & f\cdot\mathbf{J} \end{pmatrix}, \tag{39}$$

with $c = \frac{a\cdot|N_2|\cdot|R_1|+b\cdot|N_1|\cdot|R_2|}{b\cdot|N_0|\cdot|R_1|\cdot|R_2|}$, $d = 1+\frac{|N_1|}{|N_0|\cdot|R_1|}$, $e = 1-\frac{1}{|R_1|}$ and $f = 1+\frac{|R_1|\cdot|N_0|+R\cdot|N_1|}{|N_0|\cdot|R_2|\cdot|R_1|}$.[29] Inserting (39) in (38) results in the regional price level estimator, $\exp\left(\hat{\boldsymbol{\beta}}^1_{\text{CPD}}\right) = \widehat{P}^{1t}_{\text{CPD}}$:

$$\widehat{P}^{1t}_{\text{CPD}} = \prod_{i\in N_0}\left(\frac{p_i^t}{p_i^1}\right)^{\frac{1}{|N_0|}}\cdot\begin{cases} \Lambda_1^{1t} & \text{if } t\in R_1 \\ \left(\prod_{r\in R_1}\Lambda_1^{1r}\right)^{\frac{1}{|R_1|}}\cdot\left(\prod_{r\in R_2}\Lambda_2^{rt}\right)^{\frac{1}{|R_2|}} & \text{if } t\in R_2 \end{cases}, \tag{40}$$

where $\Lambda_1^{1r}$ and $\Lambda_2^{rt}$ are two correction terms that depend on the prices in region group $R_1$ and $R_2$. They are defined by

$$\Lambda_1^{1r} = \frac{\prod_{i\in N_0\cup N_1}\left(p_i^r/p_i^1\right)^{\frac{1}{|N_0|+|N_1|}}}{\prod_{i\in N_0}\left(p_i^r/p_i^1\right)^{\frac{1}{|N_0|}}} \quad\text{and}\quad \Lambda_2^{rt} = \frac{\prod_{i\in N_0\cup N_2}\left(p_i^t/p_i^r\right)^{\frac{1}{|N_0|+|N_2|}}}{\prod_{i\in N_0}\left(p_i^t/p_i^r\right)^{\frac{1}{|N_0|}}}. \tag{41}$$

The correction term $\Lambda_1^{1t}$ for regions 1 and $t$ is defined in the same way.

Due to transitivity of the price levels in (40), $\widehat{P}^{st}_{\text{CPD}}$ for regions $s,t\in R_2$ can be calculated

[29] The submatrices in $\mathbf{A}$ and $\mathbf{B}$ encompass the same dimensions as the submatrices in $\mathbf{X}_1'\mathbf{X}_1$ and $\mathbf{X}_1'\mathbf{X}_2$, respectively. A description of the matrix $\mathbf{C}$ is omitted at this point because it is not required for the derivation of the regional price level estimates.

by

$$\widehat{P}^{st}_{\text{CPD}} = \widehat{P}^{1t}_{\text{CPD}} / \widehat{P}^{1s}_{\text{CPD}} = \prod_{i \in N_0} \left( \frac{p_i^t}{p_i^s} \right)^{\frac{1}{|N_0|}} \cdot \Lambda_2^{st} \,,$$

where $\widehat{P}^{1t}_{\text{CPD}}$ ($t \in R_2$) and $\widehat{P}^{1s}_{\text{CPD}}$ ($s \in R_2$) are defined in (40), respectively.

Hajargasht *et al.* (2019, p. 106) provide the formula underlying the estimated variance of these price levels.

### A.2.2 CD method

The GLS estimator, $\hat{\boldsymbol{\beta}}^{\text{GLS}}_{\text{CD}}$, is defined in (26). The matrix $\mathbf{X} = (\mathbf{X}_1 \, \mathbf{X}_2)$ encompasses the $R-1$ dummy variables $\left( region^t - \widetilde{region^t} \right)$, where those of regions $t \in R_1 \wedge t \neq 1$ can be found in $\mathbf{X}_1$ and those of $t \in R_2$ in $\mathbf{X}_2$. As a result, the $((R-1) \times (R-1))$-matrix $\mathbf{X}' \boldsymbol{\Omega}^{-\mathbf{1}} \mathbf{X}$ can be written as:

$$\left( \mathbf{X}' \boldsymbol{\Omega}^{-\mathbf{1}} \mathbf{X} \right) = 2 \cdot \left( \begin{bmatrix} a \cdot \mathbf{I}_{|R_1|-1} - \frac{|N_1|}{|R_1|} \cdot \mathbf{J}_{|R_1|-1} & \mathbf{0}_{(|R_1|-1) \times |R_2|} \\ \mathbf{0}_{|R_2| \times (|R_1|-1)} & b \cdot \mathbf{I}_{|R_2|} - \frac{|N_2|}{|R_2|} \cdot \mathbf{J}_{|R_2|} \end{bmatrix} - \frac{|N_0|}{R} \cdot \mathbf{J}_{R-1} \right) ,$$

where the scalars $a$ and $b$ are defined on page 31. Following Miller (1981, p. 67), the inverse of that matrix is given by

$$\left( \mathbf{X}' \boldsymbol{\Omega}^{-\mathbf{1}} \mathbf{X} \right)^{-1} = \frac{1}{2} \cdot \mathbf{A} \,, \tag{42}$$

with $\mathbf{A}$ being defined in (39). The $((R-1) \times 1)$-vector $\mathbf{X}' \boldsymbol{\Omega}^{-\mathbf{1}} \mathbf{y}$ can be written as

$$\mathbf{X}' \boldsymbol{\Omega}^{-\mathbf{1}} \mathbf{y} = 2 \cdot \begin{pmatrix} \mathbf{X}_1' \boldsymbol{\Omega}^{-\mathbf{1}} \mathbf{y} \\ \mathbf{X}_2' \boldsymbol{\Omega}^{-\mathbf{1}} \mathbf{y} \end{pmatrix} , \tag{43}$$

where the element $x_1^t$ of the $((|R_1|-1) \times 1)$-vector $\mathbf{X}_1' \boldsymbol{\Omega}^{-\mathbf{1}} \mathbf{y}$ is defined by

$$x_1^t = \frac{1}{R} \cdot \sum_{r \in R_1 \cup R_2} \sum_{i \in N_0} \ln\left( p_i^t / p_i^r \right) + \underbrace{\sum_{i \in N_1} \ln\left( p_i^t / p_i^1 \right) - \frac{1}{|R_1|} \cdot \sum_{r \in R_1} \sum_{i \in N_1} \ln\left( p_i^r / p_i^1 \right)}_{= \frac{1}{|R_1|} \cdot \sum_{r \in R_1} \sum_{i \in N_1} \ln\left( p_i^t / p_i^r \right)}$$

for regions $t \in R_1 \wedge t \neq 1$, while the element $x_2^t$ of the $(|R_2| \times 1)$-vector $\mathbf{X}_2' \boldsymbol{\Omega}^{-\mathbf{1}} \mathbf{y}$ is

$$x_2^t = \frac{1}{R} \cdot \sum_{r \in R_1 \cup R_2} \sum_{i \in N_0} \ln\left( p_i^t / p_i^r \right) + \underbrace{\sum_{i \in N_2} \ln\left( p_i^t / p_i^s \right) - \frac{1}{|R_2|} \cdot \sum_{r \in R_2} \sum_{i \in N_2} \ln\left( p_i^r / p_i^s \right)}_{= \frac{1}{|R_2|} \cdot \sum_{r \in R_2} \sum_{i \in N_2} \ln\left( p_i^t / p_i^r \right)}$$

for regions $t \in R_2$. Because the initial region $s = 1$ would not allow the calculation of any price ratio for products $i \in N_2$, another region $s \in R_2$ is used instead. This is evident from the definition of $x_2^t$.

Inserting (42) and (43) in (26) finally yields $\exp\left(\hat{\boldsymbol{\beta}}_{\text{CD}}^{\text{GLS}}\right)$ as defined in (40). Thus, it follows that $\hat{\boldsymbol{\beta}}_{\text{CD}}^{\text{GLS}} = \hat{\boldsymbol{\beta}}_{\text{CPD}}^{1}$.

### A.2.3 GEKS method

The OLS estimator, $\hat{\boldsymbol{\beta}}_{\text{GEKS}}$, is defined in (34).[30] Some of the logarithmic bilateral index numbers, $\ln \dot{P}_{\text{J}}^{rt}$, change due to a decrease in commonly available price observations between regions $t$ and $r$ (e.g. Rao and Hajargasht, 2016, p. 415). It follows that the sum of logarithmic bilateral Jevons index numbers for region $t$ is defined by

$$\sum_{r=1}^{R} \ln \dot{P}_{\text{J}}^{rt} = \begin{cases} \frac{1}{|N_0|+|N_1|} \sum_{r \in R_1} \sum_{i \in N_0 \cup N_1} \ln\left(p_i^t / p_i^r\right) + \frac{1}{|N_0|} \sum_{r \in R_2} \sum_{i \in N_0} \ln\left(p_i^t / p_i^r\right) & \text{if } t \in R_1 \\ \frac{1}{|N_0|+|N_2|} \sum_{r \in R_2} \sum_{i \in N_0 \cup N_2} \ln\left(p_i^t / p_i^r\right) + \frac{1}{|N_0|} \sum_{r \in R_1} \sum_{i \in N_0} \ln\left(p_i^t / p_i^r\right) & \text{if } t \in R_2 \end{cases} . \tag{44}$$

Inserting (44) in (34) yields the price level estimator, $\exp\left(\hat{\boldsymbol{\beta}}_{\text{GEKS}}\right) = \widehat{P}_{\text{GEKS}}^{1t}$, as

$$\widehat{P}_{\text{GEKS}}^{1t} = \prod_{i \in N_0} \left(\frac{p_i^t}{p_i^1}\right)^{\frac{1}{|N_0|}} \cdot \begin{cases} \left(\Lambda_1^{1t}\right)^{\frac{|R_1|}{R}} & \text{if } t \in R_1 \\ \left(\prod_{r \in R_1} \Lambda_1^{1r}\right)^{\frac{1}{R}} \cdot \left(\prod_{r \in R_2} \Lambda_2^{rt}\right)^{\frac{1}{R}} & \text{if } t \in R_2 \end{cases} . \tag{45}$$

The two correction terms, $\Lambda_1^{1r}$ and $\Lambda_2^{rt}$, are defined in (41). Thus, the GEKS-Jevons price level estimator only differs from that of the CPD method in (40) in terms of a different weighting of the correction terms. Due to transitivity of the price levels in (45), $\widehat{P}_{\text{GEKS}}^{st}$ for regions $s, t \in R_2$ can be calculated by

$$\widehat{P}_{\text{GEKS}}^{st} = \widehat{P}_{\text{GEKS}}^{1t} / \widehat{P}_{\text{GEKS}}^{1s} = \prod_{i \in N_0} \left(\frac{p_i^t}{p_i^s}\right)^{\frac{1}{|N_0|}} \cdot \left(\Lambda_2^{st}\right)^{\frac{|R_2|}{R}} ,$$

where $\widehat{P}_{\text{GEKS}}^{1t}$ ($t \in R_2$) and $\widehat{P}_{\text{GEKS}}^{1s}$ ($s \in R_2$) are defined in (45), respectively.

The estimated variance-covariance-matrix is defined in (35). Because prices are missing, the bilateral price index numbers are no longer transitive. Thus, $\widehat{\sigma}_{\text{GEKS}}^2 > 0$ and, accordingly, $\widehat{\text{var}}\left(\widehat{\alpha}^t\right) = (2/R) \cdot \widehat{\sigma}_{\text{GEKS}}^2 > 0$ for regions $t = 2, \ldots, R$.

[30] Compared to Section A.1.3, the definitions of $\mathbf{X}$, $\mathbf{y}$ and, thus, the OLS estimator in (33) remain unchanged. Moreover, Equation (33) simplifies to (34) in the scenario of missing prices considered here.

# B Simulation results

## B.1 Intragroup and intergroup price level estimates

In Section 4.1, it is shown that the CPD price level is defined as a simple Jevons index for *intragroup comparisons*. This does not apply to the GEKS-Jevons method. By contrast, for *intergroup comparisons*, the GEKS-Jevons price level seems to approximate the Jevons index closer due to a smoother weighting of the correction terms. In order to strengthen this assumption, we draw on $L = 2{,}000$ ($l = 1, 2, \ldots, L$) simulated price data sets. The number of regions and products varies across the data sets. The prices are drawn independently for each region $r$ and product $i$ from a normal distribution with constant mean and variance, i.e. $p_i^r \sim N(10, 1)$. We estimate for each data set the regional price levels, $\widehat{P}^{st}_{\text{CPD}}$ and $\widehat{P}^{st}_{\text{GEKS}}$, and normalise them by the bilateral Jevons price level, $\dot{P}^{st}_{\text{J}}$, respectively.

Figure 3 depicts on the vertical axis the normalised price levels. The left panel shows that the normalised CPD price levels are always one for intragroup comparisons and, therefore, represented by a straight horizontal line. For intergroup comparisons, however, this is not the case. It can be seen from the right panel that the GEKS-Jevons price levels fluctuate in roughly 90% of cases more closely around $\dot{P}^{st}_{\text{J}}$ than the CPD price levels.

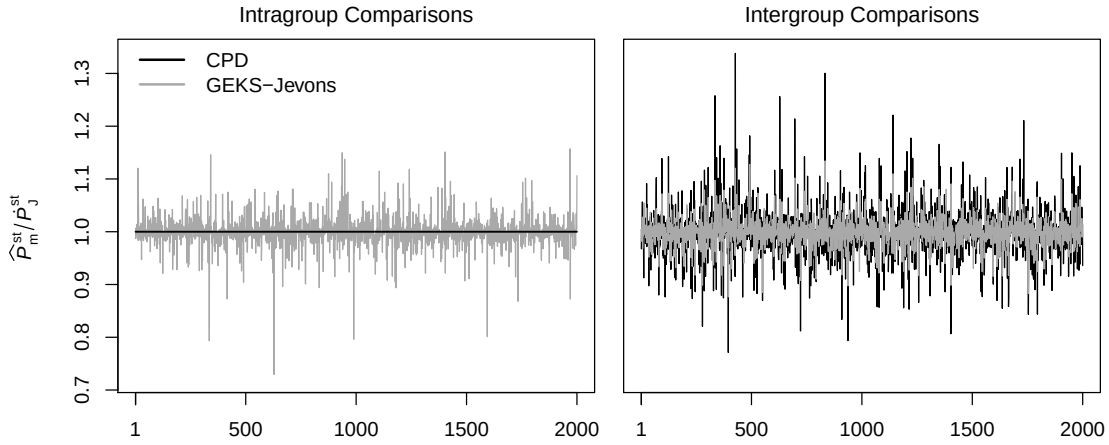


**Figure 3:** Estimated CPD and GEKS-Jevons price levels, $\widehat{P}^{st}_{m}$, for intragroup (left panel) and intergroup comparisons (right panel), normalised by the respective bilateral Jevons price level, $\dot{P}^{st}_{\text{J}}$. Calculations on the basis of $L = 2{,}000$ simulated price data sets.

## B.2 Error metrics of price level estimators

In the following, we showcase the simulation results of Section 4.2. Figure 4 depicts the change in the estimated root mean squared error (RMSE) when we vary the number of prod-

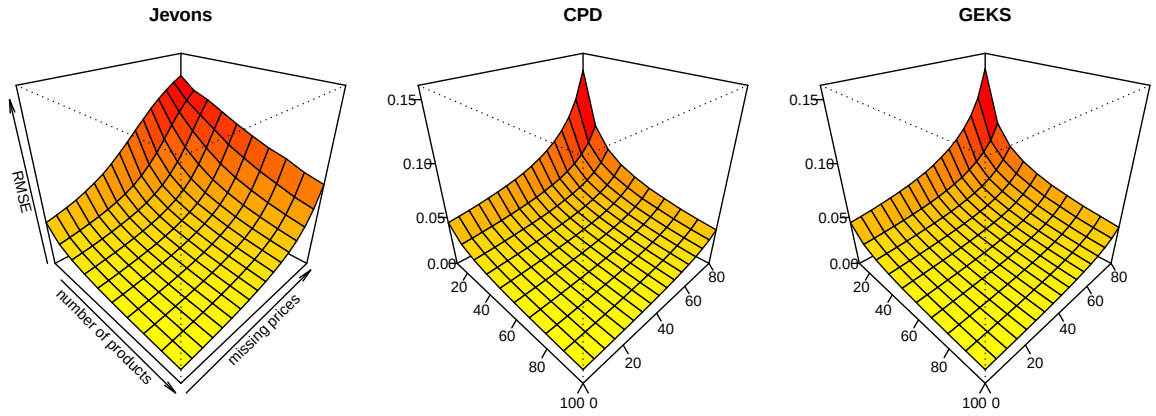


**Figure 4:** RMSE (vertical axis) by number of products (left horizontal axis) and percentage of missing prices (in %, right horizontal axis). Calculations on the basis of $L = 2{,}000$ simulated price data sets with $R = 30$ regions.

ucts in our artificial price data within a sequence from 10 to 100 but exogenously set the number of regions to $R = 30$. As can be seen, the number of products and the RMSE are inversely related. This result is not surprising as it mirrors a gain in the efficiency of the price level estimators that arises from an increased number of price observations per region.

In contrast, Figure 5 shows the simulation results in terms of the RMSE when we set the number of products to $N = 50$ but vary the number of regions within a range from 10 to 100. As can be seen, especially with a high percentage of missing prices, the RMSE of the CPD and the GEKS-Jevons estimators decreases when we raise the number of regions in our price data. The same does not apply to the simple Jevons index. This is due to its bilateral nature that considers only the prices of the two regions under consideration when calculating some price level. The CPD and GEKS-Jevons methods, by contrast, additionally use the price information of all other regions and, thus, are able to increase the efficiency of the respective price level estimator.

Summarising, Figures 4 and 5 reveal that the RMSE depends, among other drivers, on the number of regions and products present in some price data set. The general interpretation of introducing gaps into the price data, however, basically remains unaffected. Therefore, it seems acceptable to exogenously fix both the number of products and regions in the core simulation study.

We conduct our simulation study with $L = 2{,}000$ artificial price data sets for $R = 30$ regions and $N = 50$ products. Table 4 depicts the corresponding error metrics (bias, maximum-minimum error range, mean absolute error and RMSE) for the price level estimator, $\widehat{\alpha}_m^t = \widehat{\ln P_m^{st}}$ (with $m \in \{\text{Jevons}, \text{CPD}, \text{GEKS}\}$). Moreover, the error metrics are calculated separately for four different scenarios of missing prices:

- **Scenario I**: The base region $s$ as well as the comparison region $t$ provide full prices

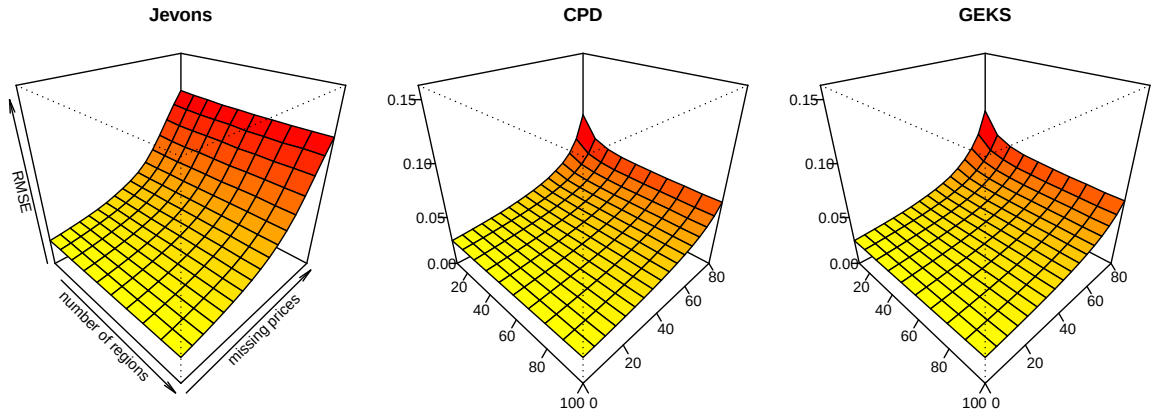


**Figure 5:** RMSE (vertical axis) by number of regions (left horizontal axis) and percentage of missing prices (in %, right horizontal axis). Calculations on the basis of $L = 2{,}000$ simulated price data sets with $N = 50$ products.

while all other regions $r$ (with $r \neq s, t$) may exhibit gaps.

- **Scenario II**: The comparison region $t$ is the only region that provides full prices.
- **Scenario III**: The base region $s$ is the only region that provides full prices.
- **Scenario IV**: Neither the base region $s$, the comparison region $t$ or some other region provide prices for all products.

Figure 6 illustrates the absolute deviations between the estimated and the true logarithmic price levels for each scenario.

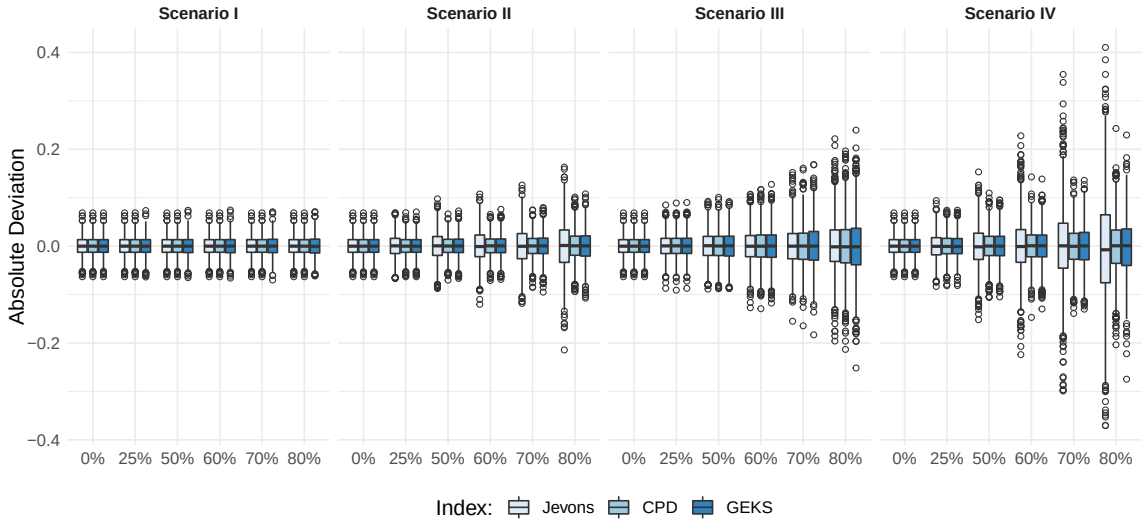


**Figure 6:** Boxplots of absolute deviations between estimated and true logarithmic price levels (vertical axis) by percentage of missing prices (in %, horizontal axis). Calculations on the basis of 2,000 simulated price data sets with $R = 30$ regions and $N = 50$ products.

| | Missing | Bias | | | MaxMin | | | MAE | | | RMSE | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | CPD | GEKS | Jevons | CPD | GEKS | Jevons | CPD | GEKS | Jevons | CPD | GEKS | Jevons |
| **Scenario I** | 0% | 0.00028 | 0.00028 | 0.00028 | 0.13190 | 0.13190 | 0.13190 | 0.01550 | 0.01550 | 0.01550 | 0.01941 | 0.01941 | 0.01941 |
| | 25% | 0.00028 | 0.00020 | 0.00028 | 0.13190 | 0.13715 | 0.13190 | 0.01550 | 0.01558 | 0.01550 | 0.01941 | 0.01954 | 0.01941 |
| | 50% | 0.00028 | 0.00013 | 0.00028 | 0.13190 | 0.13901 | 0.13190 | 0.01550 | 0.01573 | 0.01550 | 0.01941 | 0.01974 | 0.01941 |
| | 60% | 0.00028 | 0.00002 | 0.00028 | 0.13190 | 0.14052 | 0.13190 | 0.01550 | 0.01593 | 0.01550 | 0.01941 | 0.01998 | 0.01941 |
| | 70% | 0.00028 | 0.00014 | 0.00028 | 0.13190 | 0.14078 | 0.13190 | 0.01550 | 0.01619 | 0.01550 | 0.01941 | 0.02037 | 0.01941 |
| | 80% | 0.00028 | 0.00009 | 0.00028 | 0.13190 | 0.13273 | 0.13190 | 0.01550 | 0.01686 | 0.01550 | 0.01941 | 0.02117 | 0.01941 |
| **Scenario II** | 0% | 0.00026 | 0.00026 | 0.00026 | 0.13190 | 0.13190 | 0.13190 | 0.01551 | 0.01551 | 0.01551 | 0.01943 | 0.01943 | 0.01943 |
| | 25% | 0.00027 | 0.00016 | 0.00035 | 0.13240 | 0.13391 | 0.13619 | 0.01563 | 0.01572 | 0.01840 | 0.01964 | 0.01976 | 0.02293 |
| | 50% | 0.00049 | 0.00046 | 0.00003 | 0.13655 | 0.13947 | 0.18521 | 0.01609 | 0.01639 | 0.02299 | 0.02023 | 0.02066 | 0.02887 |
| | 60% | 0.00042 | 0.00043 | -0.00007 | 0.13661 | 0.14474 | 0.22764 | 0.01666 | 0.01711 | 0.02596 | 0.02096 | 0.02159 | 0.03250 |
| | 70% | 0.00029 | 0.00021 | -0.00038 | 0.15987 | 0.17454 | 0.24447 | 0.01809 | 0.01889 | 0.03064 | 0.02271 | 0.02389 | 0.03818 |
| | 80% | 0.00072 | 0.00017 | -0.00001 | 0.19507 | 0.21520 | 0.37746 | 0.02331 | 0.02483 | 0.03940 | 0.02933 | 0.03118 | 0.04963 |
| **Scenario III** | 0% | 0.00026 | 0.00026 | 0.00026 | 0.13190 | 0.13190 | 0.13190 | 0.01551 | 0.01551 | 0.01551 | 0.01942 | 0.01942 | 0.01942 |
| | 25% | 0.00068 | 0.00069 | 0.00068 | 0.17995 | 0.17724 | 0.17207 | 0.01852 | 0.01864 | 0.01854 | 0.02316 | 0.02329 | 0.02316 |
| | 50% | 0.00033 | 0.00030 | 0.00032 | 0.18864 | 0.17930 | 0.18118 | 0.02338 | 0.02377 | 0.02304 | 0.02914 | 0.02960 | 0.02866 |
| | 60% | -0.00009 | -0.00009 | 0.00004 | 0.24611 | 0.24512 | 0.23431 | 0.02688 | 0.02757 | 0.02627 | 0.03390 | 0.03483 | 0.03313 |
| | 70% | 0.00010 | 0.00023 | 0.00022 | 0.32552 | 0.35160 | 0.30728 | 0.03236 | 0.03468 | 0.03122 | 0.04076 | 0.04353 | 0.03945 |
| | 80% | -0.00026 | -0.00055 | -0.00011 | 0.40972 | 0.49109 | 0.41767 | 0.04200 | 0.04585 | 0.03944 | 0.05365 | 0.05855 | 0.05056 |
| **Scenario IV** | 0% | 0.00025 | 0.00025 | 0.00025 | 0.13190 | 0.13190 | 0.13190 | 0.01550 | 0.01550 | 0.01550 | 0.01942 | 0.01942 | 0.01942 |
| | 25% | -0.00009 | -0.00007 | -0.00022 | 0.15590 | 0.15548 | 0.17732 | 0.01839 | 0.01843 | 0.02087 | 0.02304 | 0.02306 | 0.02608 |
| | 50% | 0.00005 | 0.00002 | -0.00058 | 0.21551 | 0.20012 | 0.30521 | 0.02314 | 0.02349 | 0.03231 | 0.02911 | 0.02948 | 0.04067 |
| | 60% | 0.00043 | 0.00051 | -0.00011 | 0.29018 | 0.26839 | 0.45189 | 0.02641 | 0.02701 | 0.04099 | 0.03315 | 0.03409 | 0.05272 |
| | 70% | -0.00011 | -0.00017 | 0.00016 | 0.27549 | 0.26593 | 0.65357 | 0.03142 | 0.03312 | 0.05775 | 0.03954 | 0.04163 | 0.07503 |
| | 80% | -0.00048 | -0.00108 | -0.00601 | 0.44658 | 0.50425 | 0.78107 | 0.04174 | 0.04548 | 0.08442 | 0.05282 | 0.05726 | 0.10775 |

**Table 4:** Simulation results in terms of bias, maximum-minimum error range (MaxMin), mean absolute error (MAE) and root mean squared error (RMSE) for the CPD, GEKS-Jevons and Jevons price level estimators, respectively. Calculations on the basis of 2,000 simulated price data sets with $R = 30$ regions and $N = 50$ products.

# References


ANSELIN, L. (2003). Spatial Econometrics. In B. H. Baltagi (ed.), *A Companion to Theoretical Econometrics*, Blackwell Publishing Ltd, pp. 310–330.

ATEN, B. H. (1996). Evidence of Spatial Autocorrelation in International Prices. *Review of Income and Wealth*, **42** (2), 149–163.

— (1997). Does Space Matter? International Comparisons of the Prices of Tradables and Nontradables. *International Regional Science Review*, **20** (1-2), 35–52.

AUER, L. V. (2012). Räumliche Preisvergleiche: Aggregationskonzepte und Forschungsperspektiven. *AStA Wirtschafts- und Sozialstatistisches Archiv*, **6** (1), 27–56.

BALK, B. M. (1980). A Method for Constructing Price Indices for Seasonal Commodities. *Journal of the Royal Statistical Society. Series A (General)*, **143** (1), 68–75.

— (1981). A Simple Method for Constructing Price Indices for Seasonal Commodities. *Statistische Hefte*, **22** (1), 72–78.

— (1995). Axiomatic Price Index Theory: A Survey. *International Statistical Review*, **63** (1), 69–93.

— (2008). *Price and Quantity Index Numbers*. Cambridge University Press.

— and KERSTEN, H. M. P. (1986). On the Precision of Consumer Price Indices Caused by the Sampling Variability of Budget Surveys. *Journal of Economic and Social Measurement*, **14** (1), 19–35.

BEHRMANN, T., DEML, S. and LINZ, S. (2009). *Verwendung von Einzeldaten aus der Verbraucherpreisstatistik für regionale Preisvergleiche*. Research Note 36, RatSWD.

BIGGERI, L., LAURETI, T. and POLIDORO, F. (2017). Computing Sub-National PPPs with CPI Data: An Empirical Analysis on Italian Data Using Country Product Dummy Models. *Social Indicators Research*, **131** (1), 93–121.

CALVANO, E., CALZOLARI, G., DENICOLÒ, V. and PASTORELLO, S. (2018). *Artificial Intelligence, Algorithmic Pricing and Collusion*. Discussion Paper 13405, Centre for Economic Policy Research, London.

CAVALLO, A. (2018). *More Amazon Effects: Online Competition and Pricing Behaviors*. Working Paper 25138, National Bureau of Economic Research.

CAVES, D. W., CHRISTENSEN, L. R. and DIEWERT, W. E. (1982). Multilateral Comparisons of Output, Input, and Productivity Using Superlative Index Numbers. *The Economic Journal*, **92** (365), 73–86.

CHEN, L., MISLOVE, A. and WILSON, C. (2016). An Empirical Analysis of Algorithmic Pricing on Amazon Marketplace. In *Proceedings of the 25th International Conference on World Wide Web*, pp. 1339–1349.

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

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

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

CUTHBERT, J. R. (2003). On the Variance/Covariance Structure of the Log Fisher Index, and Implications for Aggregation Techniques. *Review of Income and Wealth*, **49** (1), 69–88.

— and CUTHBERT, M. (1988). *On Aggregation Methods of Purchasing Power Parities*. OECD Economics Department Working Paper 56, OECD Publishing, Paris.

DIEWERT, W. E. (1986). *Microeconomic Approaches to the Theory of International Comparisons*. Working Paper 53, National Bureau of Economic Research.

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

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

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

— (2010). New Methodological Developments for the International Comparison Program. *Review of Income and Wealth*, **56** (s1), S11–S31.

DIKHANOV, Y. (2010). Assessing Efficiency of Elementary Indices with Monte Carlo Simulations, African Development Bank Workshop, Washington DC.

DRECHSLER, L. (1973). Weighting of Index Numbers in Multilateral International Comparisons. *Review of Income and Wealth*, **19** (1), 17–34.

ELTETÖ, O. and KÖVES, P. (1964). On a Problem of Index Number Computation Relating to International Comparison. *Statisztikai Szemle*, **42**, 507–518.

EUROPEAN COMMISSION (2017). *Final Report on the E-commerce Sector Inquiry*. Brussels: European Commission.

EUROSTAT-OECD (2012). *Eurostat-OECD Methodological Manual on Purchasing Power Parities*. Methodologies and Working Papers, Luxembourg: Publications Office of the European Union.

FERRARI, G., GOZZI, G. and RIANI, M. (1996). Comparing CPD and GEKS Approaches at the Basic Headings Level. In Eurostat (ed.), *CPI & PPP: Improving the Quality of Price Indices*, pp. 323–337.

— and RIANI, M. (1998). On Purchasing Power Parities Calculation at the Basic Heading Level. *Statistica*, **58** (1), 91–108.

GEARY, R. C. (1958). A Note on the Comparison of Exchange Rates and Purchasing Power Between Countries. *Journal of the Royal Statistical Society. Series A (General)*, **121** (1), 97–99.

GINI, C. (1924). Quelques Considerations au Sujet de la Construction des Nombres Indices des Prix et des Questions Analogues. *Mentron*, **4** (1), 3–162.

— (1931). On the Circular Test of Index Numbers. *International Statistical Review*, **9** (2), 3–25.

GOLDHAMMER, B. (2016). Die neue Mietenstichprobe in der Verbraucherpreisstatistik. In *Wirtschaft und Statistik*, no. 5 in 2016, Statistisches Bundesamt, pp. 86–101.

GRAYBILL, F. A. (1983). *Matrices with Applications in Statistics*. Belmont: Wadsworth International Group, 2nd edn.

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.

— and — (2019). Multilateral Index Number Systems for International Price Comparisons: Properties, Existence and Uniqueness. *Journal of Mathematical Economics*, **83** (2019), 36–47.

—, — and ABBAS, V. (2019). Reliability of Basic Heading PPPs. *Economics Letters*, **180**, 102–107.

HILL, R. J. (2016). A Least Squares Approach to Imposing Within-Region Fixity in the International Comparisons Program. *Journal of Econometrics*, **191** (2), 407–413.

— and HILL, T. P. (2009). Recent Developments in the International Comparison of Prices and Real Output. *Macroeconomic Dynamics*, **13** (S2), 194–217.

— and TIMMER, M. P. (2006). Standard Errors as Weights in Multilateral Price Indexes. *Journal of Business & Economic Statistics*, **24** (3), 366–377.

HILL, T. P. (2008). Elementary Indices for Purchasing Power Parities. Joint UNECE/ILO Meeting on Consumer Price Indices, Geneva, http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2008/mtg1/zip.24.e.pdf, accessed: 2019-12-04.

HORN, R. A. and JOHNSON, C. R. (2012). *Matrix Analysis*. New York: Cambridge University Press, 2nd edn.

ILO, IMF, OECD, UNECE, EUROSTAT and WORLD BANK (2004). *Consumer Price Index Manual: Theory and Practice*. Geneva: International Labour Office.

JEVONS, W. S. (1865). On the Variation of Prices and the Value of the Currency since 1782. *Journal of the Statistical Society of London*, **28** (2), 294–320.

KACKAR, R. N. and HARVILLE, D. A. (1984). Approximations for Standard Errors of Estimators of Fixed and Random Effects in Mixed Linear Models. *Journal of the American Statistical Association*, **79** (388), 853–862.

KENNEDY, P. (1981). Estimation with Correctly Interpreted Dummy Variables in Semilogarithmic Equations. *American Economic Review*, **71** (4), 801.

KHAMIS, S. H. (1972). A New System of Index Numbers for National and International Purposes. *Journal of the Royal Statistical Society. Series A (General)*, **135** (1), 96–121.

KLEIN, T. (2018). *Assessing Autonomous Algorithmic Collusion: Q-Learning Under Short-Run Price Commitments*. Discussion Paper 2018/056/VII, Tinbergen Institute, Amsterdam and Rotterdam.

KOKOSKI, M. F., MOULTON, B. R. and ZIESCHANG, K. D. (1999). Interarea Price Comparisons for Heterogeneous Goods and Several Levels of Commodity Aggregation. In *International and Interarea Comparisons of Income, Output, and Prices*, University of Chicago Press, pp. 123–169.

LAURETI, T. and RAO, D. S. P. (2018). Measuring Spatial Price Level Differences within a Country: Current Status and Future Developments. *Estudios de Economía Aplicada*, **36** (1), 119–148.

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

MONTERO, J.-M., LAURETI, T., MÍNGUEZ, R. and FERNÁNDEZ-AVILÉS, G. (2019). A Stochastic Model with Penalized Coefficients for Spatial Price Comparisons: An Application to Regional Price Indexes in Italy. *Review of Income and Wealth*, forthcoming.

OECD (2018). OECD Calculation of Contributions to Overall Annual Inflation. http://www.oecd.org/sdd/prices-ppp/OECD-calculation-contributions-annual-inflation.pdf, accessed: 2019-12-04.

RAO, D. S. P. (2001). Weighted EKS and Generalised CPD Methods for Aggregation at Basic Heading Level and above Basic Heading Level, Joint World Bank - OECD Seminar on Purchasing Power Parities, Washington, DC.

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

— (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.

— and TIMMER, M. P. (2003). Purchasing Power Parities for Industry Comparisons using Weighted Elteto-Koves-Szulc (EKS) Methods. *Review of Income and Wealth*, **49** (4), 491–512.

SELVANATHAN, E. A. (2003). Extending the Stochastic Approach to Index Numbers: A Comment on Crompton. *Applied Economics Letters*, **10** (4), 213–215.

— 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.

SILVER, M. (2009). *The Hedonic Country Product Dummy Method and Quality Adjustments for Purchasing Power Parity Calculations*. Working Paper 09/271, International Monetary Fund.

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

SZULC, B. J. (1964). Indices for Multiregional Comparisons. *Przeglad Statystyczny*, **3**, 239–254.

WEINAND, S. and AUER, L. V. (2019). *Anatomy of Regional Price Differentials: Evidence from Micro Price Data*. Discussion Paper 2019/04, Deutsche Bundesbank.

WHITE, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, **48** (4), 817–838.

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 (2015). *Purchasing Power Parities and the Real Size of World Economies: A Comprehensive Report of the 2011 International Comparison Program*. Washington DC: World Bank.

ZIMMER, E. (2016). Die neue Mietenstichprobe im Verbraucherpreisindex. In *Zeitschrift für amtliche Statistik*, no. 2 in 2016, Statistik Berlin Brandenburg, pp. 42–47.