On the Stochastic Approach to Linking the Regions in the ICP

by

W. Erwin Diewert

OCTOBER 2004

Discussion Paper No.: 04-16

DEPARTMENT OF ECONOMICS
THE UNIVERSITY OF BRITISH COLUMBIA
VANCOUVER, CANADA V6T 1Z1

http://www.econ.ubc.ca

On the Stochastic Approach to Linking the Regions in the ICP

Erwin Diewert,[1]
Discussion Paper No. 04-16,
Department of Economics,
The University of British Columbia,
Vancouver, Canada, V6T 1Z1.
email: diewert@econ.ubc.ca November 9, 2004.

Abstract

The paper looks at the problems involved in making international comparisons of prices at the first stage of aggregation, where detailed information on expenditure weights may not be available. However, even in this situation, it is argued that a target index concept should be defined that may require information on expenditure weights. With a target index defined, then various "practical" approximations to the ideal target index can be evaluated. The paper suggests a weighted generalization of Summer's Country Product Dummy (CPD) method for making international price comparisons. It is argued that this weighted CPD method is a natural generalization of Theil's stochastic approach to index number theory (where there are only two countries in the comparison with each country purchasing positive amounts of all commodities) to the case where there are many countries and not every commodity is transacted in each country.

Journal of Economic Literature Classification Numbers

C21, C43, O57.

Keywords

Purchasing power parities, index numbers, country product dummy method, weighted multilateral method, International Comparisons Program (ICP), Törnqvist price index, stochastic approaches to index number theory.

1. Introduction

This paper takes a systematic look at some of the problems that are involved in making international comparisons of prices. We focus on the problems that occur at the *first stage of aggregation*, where accurate information on expenditure weights that are associated with price information collected by individual countries at the basic heading

[1] My thanks to Bert Balk, Yonas Biru, Angus Deaton, Yuri Dikhanov, Peter Hill, Alice Nakamura, Sergey Sergeev and Prasada Rao for helpful comments. Earlier versions of this paper were presented at the Workshop: Estimating Production and Income Across Nations at UC Davis, Institute of Governmental Affairs organized by Robert Feenstra, April 13-14, 2004, and on May 25 and September 23, 2004 at meetings of the Technical Advisory Group for the International Comparisons Project at the World Bank, Washington, D.C. This research was partially supported by the World Bank and a SSHRC research grant. Neither institution is responsible for any views expressed in the paper.

level is not available.[2] Without information on weights, "normal" index number theory is not applicable.[3] Hence, the present paper will focus on statistical or stochastic approaches to the index number problem at this first stage of aggregation. However, in some sections of the paper, we assume that information on appropriate expenditure weights is available and under these circumstances, we consider *weighted stochastic approaches* for making international comparisons of prices. These weighted approaches can serve as *target concepts* for the more practical unweighted stochastic approaches.

The most promising statistical approach to the multilateral aggregation problem at the first stage of aggregation is the *Country Product Dummy* (CPD) method for making international comparisons of prices, proposed by Robert Summers (1973). In section 2, we will review the algebra of this method assuming that we are attempting to make an international comparison of prices between C countries over a reasonably homogeneous group of say N items. In this section, we also assume that no expenditure weights are available for the price comparisons and that exactly K outlets are sampled for each of the N items in each of the C countries. Thus there are CNK price quotes collected across all of the countries. This generalizes the usual CPD method in that we model price variability right down to the level of the individual prices that are collected by the countries involved in the comparison.[4]

In section 3, the assumption that an equal number of price quotes for each product is collected by each country in the comparison is dropped; i.e., we develop the algebra of the Country Product Dummy method without assuming equal sample sizes for each product in each country. We also allow for gaps in the data. This model is our basic unweighted CPD model.

Section 4 specializes the model presented in section 3: we consider the special case where price information on a particular product is collected in only one country. Intuitively, it seems reasonable that prices on a product that are collected in only one country should have no effect on the Purchasing Power Parities between the countries and this intuition turns out to be correct.

Section 5 is the key section in the paper. In this section, we assume (unrealistically) that each country is able to collect price and quantity information for every transaction that took place in the reference period for the class of products that are included in the basic heading category. With this information, it is possible to set up a weighted version of the CPD least squares regression problem that is consistent with normal index number theory, where prices are weighted according to their economic importance. Although the model developed in this section is not immediately "practical", it is important because it

[2] For additional material on this first stage aggregation problem, see Hill (2004) and Rao (2004).

[3] At higher stages of aggregation where information on prices and quantities (or expenditures) are available, many satisfactory multilateral aggregation methods exist; see Balk (1996) (2001) or Diewert (1999) for surveys of the various methods.

[4] The CPD models described by Hill (2004) and Rao (2004) work with models that have aggregated product prices over outlets within a country so that their models work with CN prices rather than our CNK prices.

gives statisticians a reasonable *target* index. Given this target index, various approximations to it can then be evaluated.

To show that the target ideal index that is defined in section 5 is indeed a reasonable index, in section 6, we specialize the general multilateral model presented in section 5 to the case of two countries. In this two country case, the PPP between the two countries can be worked out as an explicit index number formula. We find that the resulting bilateral index has very reasonable properties and it turns out to closely approximate a superlative bilateral index number formula.

Sections 7 and 8 are not required reading for the remainder of the paper but they do present some material that may be of some interest. It is well known that Theil (1967; 136-137) developed a simple stochastic approach to bilateral index number theory where the logarithm of his suggested index is simply the mean of a certain discrete distribution of the logarithms of the price relatives for the products in the two countries being compared. In sections 7 and 8, we present two discrete distribution interpretations for our basic model that was presented in section 5 and we argue that the section 5 PPP's are suitable generalizations of Theil's bilateral PPP to the multilateral situation. We note that the model presented in section 8 has a wider range of applicability; in particular, it could be applied to hedonic regression models where expenditure information on the various models is available.

Section 9 returns to the basic model presented in section 5. However, instead of assuming that *complete information* on expenditures associated with each collected price is available, it is assumed that only *approximate information* on expenditure weights is available. In particular, it is assumed that each collected price quote can be labeled as being *representative* or *unrepresentative*. We show how the fundamental weighted CPD model developed in section 5 can be adapted to this situation.

The models presented in sections 3 and 9 (the unweighted CPD and the approximately weighted CPD model respectively) are the two basic "practical" models that we suggest should be used in ICP 2004.

Sections 10, 11 and 12 present various numerical examples that illustrate the various CPD methods that were developed in previous sections.

In section 10, we consider a data set that was constructed by Hill (2004) where he postulated prices for 10 items and 4 countries. We illustrate the unweighted and weighted CPD PPP's suggested in sections 3 and 9 using this data set.

In section 11, we use the data listed in section 10 to illustrate some ideas on linking countries suggested by Robert Hill.[5] His basic idea is this: countries which are most similar in their price structures (i.e., their prices are closest to being proportional across items) should be linked first. This idea is a very good one at higher levels of aggregation,

[5] See Robert Hill (1995) (1999a) (1999b) (2001) (2004). The basic idea of spatially linking countries that have the most similar price and quantity structures dates back to Fisher (1922; 271-272).

where complete price and expenditure data are available, but it is not obvious that the same methodology can be applied at the elementary level, where complete data on expenditures associated with each price quote are missing. In section 11, we construct an example which indicates that the bilateral linking approach of Robert Hill is not appropriate when data are sparse. Thus, we feel that the multilateral models developed in sections 3 and 9 are more appropriate at this first stage of aggregation where complete price and quantity information on each product will not be available.

Cuthbert and Cuthbert (1988; 57) introduced an interesting generalization of the Country Product Dummy method that can be used if information on *representativity* of the prices is collected by the countries in the comparison project along with the prices themselves. Hill (2004) explains this method in some detail and he called the method the *extended CPD Method* or *CPDR Method*. In section 12, we examine this CPRD method. Our tentative conclusion is that the CPRD method is likely to be an improvement over the unweighted CPD method suggested in section 3 but it is not likely to be an improvement over the weighted CPD method suggested in section 9, which is our preferred method.

A complication that we have not dealt with up to now is that the current ICP project is proceeding in two stages. The world is divided up into 6 regions[6] r with C(r) countries in each region r for r = 1,…,6. Within each of the 6 regions, PPP's at the basic heading level will be constructed more or less independently for each region. In the second stage, the regions will be linked. In section 13, we consider some of the complications involved in modeling this situation. It turns out that our preferred model presented in section 9 can readily be adapted to deal with this somewhat complicated linking problem. In section 13, we explain that there are two variants of the theory that could be used. In one variant, the regions are linked using the fixed within region parities that have been estimated by the regions. In the other variant, the between region and within regions are estimated at the same time. Sections 14 and 15 present a numerical example that illustrates the two variants.

Up to this point, the paper is concerned with linking at the basic heading level. In section 16, we briefly consider the problem of linking the regions at the final stage of aggregation under the assumption that at least for some regions, the within region parities are to be respected.

Section 17 concludes.

2. The Country Product Dummy Method with Equal Item Sample Sizes Across Countries

The *Country Product Dummy* (CPD) method for making international comparisons of prices can be viewed as a very simple type of hedonic regression model that was

[6] The 6 regions are: (1) Africa with 50 participating countries; (2) Latin America with 10 countries; (3) Asia with 23 countries; (4) Commonwealth of Independent States (CIS) with 12 countries; (5) Western Asia with 13 countries and (6) the OECD and EU countries with 40 participating countries. Thus there are 147 participating countries in all.

proposed by Robert Summers (1973) where the only characteristic of the commodity is the commodity itself. The CPD method can also be viewed as an example of the stochastic approach[7] to index numbers. In this section, we will review the algebra of this method assuming that we are attempting to make an international comparison of prices between C countries over a reasonably homogeneous group of say N items.[8] In this section, we also assume that no expenditure weights are available for the price comparisons and that exactly K outlets are sampled for each of the N items in each of the C countries. Thus there are CNK price quotes collected across all of the countries. These assumptions are not very realistic but it is useful to present this model as an introduction to more complex models.[9]

It should be noted that aggregation of prices in the International Comparisons Program (ICP) of the World Bank takes place at two levels of aggregation:

- Aggregation at the basic heading level;
- Aggregation above the basic heading level.

Aggregation at the basic heading level generally proceeds without expenditure weights whereas aggregation above the basic heading level uses national expenditure weights for the class of transactions that are in the domain of definition for each basic heading category of transactions. This paper is concerned only with the aggregation problem for a particular basic heading category where expenditure weights for each outlet price are not generally available.[10]

Let p_{cnk} denote the price of item n in outlet k in country c for c = 1,…,C; n = 1,…,N; k = 1,…,K. Each item n must be measured in the same quantity units across countries but the

[7] See Theil (1967; 136-138), Balk (1980) and Selvanathan and Rao (1994) for examples of the stochastic approach to index number theory. A main advantage of the CPD method for comparing prices across countries over traditional index number methods is that we can obtain *standard errors* for the country price levels. This advantage of the stochastic approach to index number theory was stressed by Summers (1973) and more recently by Selvanathan and Rao (1994).

[8] Using the language of the International Comparison of Prices (ICP) project, we are making a comparison of prices at the basic heading level. In the current ICP project headed up by the World Bank, there are 155 basic headings.

[9] A special case of the present model can be obtained by setting K equal to 1 and the price p_{cn1} can be set equal to the geometric mean of all of the outlet prices collected for product n in country c. The geometric mean is chosen over other methods for aggregating the outlet prices because, in the absence of weights, it seems to have the best axiomatic properties; e.g., see Diewert (2004). (Note however, that when aggregating using geometric means, the micro prices should not approach zero). This is the "traditional" CPD model and it is discussed by Hill (2004) and Rao (2004) in some detail. The problem with this model is that it neglects of the variability of the outlet prices *within* a country c, product n, cell. The advantage of the traditional CPD model is that the associated algebra is much simpler and hence, much easier to understand.

[10] Thus the aggregation problem to be studied in the present paper is a generalization to the case of C countries or time periods from the case where C equals 2, which is the usual *elementary index number aggregation problem* studied in the time series context, for the case of comparing prices over two periods when no expenditure weight information is available. See Hill (2004) and Rao (2004) for additional discussion of the more complex interspatial aggregation problems at the first stage of aggregation when no weight information is available.

prices can be in local currency units. The basic statistical model that is assumed is the following one:

(1) $p_{cnk} = a_c b_n u_{cnk}$; $c = 1,\ldots,C;\ n = 1,\ldots,N;\ k = 1,\ldots,K$

where the a_c and b_n are unknown parameters to be estimated and the u_{cnk} are independently distributed error terms with means 1 and constant variances. The parameter a_c is to be interpreted as the *average level of prices* (over all items in this group of items) in country c relative to other countries and the parameter b_n is to be interpreted as the *average* (over all countries) *multiplicative premium* that item n is worth relative to an average item in this grouping of items. Thus the a_c are the basic heading country price levels that we want to determine while the b_n are item effects. The basic hypothesis is that the price of item n in country c is equal to a country price level a_c times an item commodity adjustment factor b_n times a random error that fluctuates around 1. Taking logarithms of both sides of (1) leads to the following model:

(2) $y_{cnk} = \square_c + \square_n + \square_{cnk}$; $c = 1,\ldots,C;\ n = 1,\ldots,N;\ k = 1,\ldots,K$

where $y_{cnk} \equiv \ln p_{cnk}$, $\square_c \equiv \ln a_c$, $\square_n \equiv \ln b_n$ and $\square_{cnk} \equiv \ln u_{cnk}$.

The model defined by (2) is obviously a linear regression model where the independent variables are dummy variables. The least squares estimators for the $\square_c$ and $\square_n$ can be obtained by solving the following minimization problem:

(3) $\min_{\square_c,\square_n} \{\square_{c=1}^{C} \square_{n=1}^{N} \square_{k=1}^{K} [y_{cnk} \square \square_c \square \square_n]^2\}$.

However, it can be seen that the solution for the minimization problem (3) cannot be unique: if $\square_c^*$ for c = 1,…,C and $\square_n^*$ for n = 1,…,N solve (3), then so does $\square_c^* + \square$ for c = 1,…,C and $\square_n^* \square \square$ for n = 1,…,N, for any arbitrary number $\square$. Thus it will be necessary to impose an additional restriction or normalization on the parameters $\square_c$ and $\square_n$ in order to obtain a unique solution to the least squares minimization problem (3). Two possible normalizations are (4) or (5) below:

(4) $\square_1 = 0$ or $a_1 = 1$;
(5) $\square_{c=1}^{C} \square_c = 0$ or $\square_{c=1}^{C} a_c = 1$.

The normalization (4) means that country 1 is chosen as the numeraire country and the parameter a_c for c = 2,…,C is the PPP (Purchasing Power Parity) of country c relative to country 1 for the class of commodity prices that are being compared across the C countries. On the other hand, the normalization (5) treats all countries in a symmetric manner: the geometric mean of the PPP's a_c is set equal to 1.[11] In this section, we will choose to work with the normalization (5).[12]

[11] Note that $\square_{c=1}^{C} a_c = 1$ is equivalent to $\square_{c=1}^{C} a_c^{1/C} = 1$.

[12] However, if we obtain a solution to the least squares minimization problem (3) subject to the normalization (5), say $\square_1^*, \square_2^*, \ldots, \square_C^*, \square_1^*, \square_2^*, \ldots, \square_N^*$, then the solution to (3) subject to the

Initially, we ignore the constraint (5) and we differentiate (3) with respect to α_c and β_n for c = 1,…,C and n = 1,…,N and set the resulting partial derivatives equal to 0. The resulting C + N equations simplify to the following equations:

$$(6)\ \Sigma_{n=1}^{N} \Sigma_{k=1}^{K} y_{cnk} = NK\, \alpha_c + K\, \Sigma_{n=1}^{N} \beta_n\, ; \qquad c = 1,\ldots,C;$$
$$(7)\ \Sigma_{c=1}^{C} \Sigma_{k=1}^{K} y_{cnk} = K\, \Sigma_{c=1}^{C} \alpha_c + CK\, \beta_n\, ; \qquad n = 1,\ldots,N.$$

If we tentatively set $\Sigma_{c=1}^{C} \alpha_c = 0$, then equations (7) imply the following least squares solutions for the β_n:

$$(8)\ \beta_n^{*} \equiv \Sigma_{c=1}^{C} \Sigma_{k=1}^{K} y_{cnk}/CK\, ; \qquad n = 1,\ldots,N.$$

Thus β_n^{*} is simply the arithmetic average of all of the log prices $y_{cnk} \equiv \ln p_{cnk}$ of item n over all countries and all outlets. Now substitute equations (8) into (6) and we obtain the following least squares solutions for the α_c:

$$\begin{aligned}(9)\ \alpha_c^{*} &\equiv \Sigma_{n=1}^{N} \Sigma_{k=1}^{K} y_{cnk}/NK - \Sigma_{n=1}^{N} \beta_n^{*}/N\, ; \qquad c = 1,\ldots,C \\ &= \Sigma_{n=1}^{N} \Sigma_{k=1}^{K} y_{cnk}/NK - \Sigma_{c=1}^{C} \Sigma_{n=1}^{N} \Sigma_{k=1}^{K} y_{cnk}/CNK.\end{aligned}$$

Thus each α_c^{*} is equal to the arithmetic average of the logarithms of all item prices in country c less the global arithmetic average of the logarithms of all item prices over all countries.

We need to check that the α_c^{*} defined by (9) satisfy the restrictions (5):

$$\begin{aligned}(10)\ \Sigma_{c=1}^{C} \alpha_c^{*} &= \Sigma_{c=1}^{C} \{\Sigma_{n=1}^{N} \Sigma_{k=1}^{K} y_{cnk}/NK - \Sigma_{d=1}^{C} \Sigma_{n=1}^{N} \Sigma_{k=1}^{K} y_{dnk}/CNK\} \\ &= \Sigma_{c=1}^{C} \Sigma_{n=1}^{N} \Sigma_{k=1}^{K} y_{cnk}/NK - C\, \Sigma_{n=1}^{N} \Sigma_{k=1}^{K} y_{dnk}/CNK \\ &= 0.\end{aligned}$$

Thus (8) and (9) give the unique solution to the least squares minimization problem (3) subject to the normalization (5). Note in particular that this solution can be calculated simply by calculating various averages of log prices without having to do any complicated matrix inversions.[13]

normalization (4) is $\alpha_1^{*} = 0$, $\alpha_2^{*} - \alpha_1^{*}$, ... , $\alpha_C^{*} - \alpha_1^{*}$, $\beta_1^{*} + \alpha_1^{*}$, $\beta_2^{*} + \alpha_1^{*}$, ... , $\beta_N^{*} + \alpha_1^{*}$. Rao (2004) works with the normalizations (4) for the special case of our model where K=1, whereas Hill (2004) introduces an additional parameter to represent the overall logarithmic mean of the prices and then imposes the extra two normalizations $\alpha_1 = 1$ and $\beta_1 = 1$. With these extra normalizations, the overall mean price parameter becomes the mean logarithmic price for product 1 in country 1. All three methods of normalization will lead to the same relative purchasing power parities but the resulting confidence intervals for the PPP's in the three models will be somewhat different. For computing confidence intervals, the normalization (5) is the most appropriate one for ICP purposes.

[13] This solution is well known in the analysis of variance literature; e.g., see Rao (1965; 209-211). For additional references to the statistics literature on this type of model, see Hill (2004).

It is of some interest to calculate the difference between any two of the log parities between say countries c and d:

$$
\begin{aligned}
(11)\ \square_c^* \,\square\, \square_d^* &= \square_{n=1}^{N} \square_{k=1}^{K} y_{cnk}/NK \,\square\, \square_{i=1}^{C} \square_{n=1}^{N} \square_{k=1}^{K} y_{ink}/CNK \\
&\quad \square \{\square_{n=1}^{N} \square_{k=1}^{K} y_{dnk}/NK \,\square\, \square_{i=1}^{C} \square_{n=1}^{N} \square_{k=1}^{K} y_{ink}/CNK\} \quad \text{using (9) twice} \\
&= \square_{n=1}^{N} \square_{k=1}^{K} y_{cnk}/NK \,\square\, \square_{n=1}^{N} \square_{k=1}^{K} y_{dnk}/NK.
\end{aligned}
$$

Using (11) and the definitions $y_{cnk} \equiv \ln p_{cnk}$, we can calculate the PPP parity between countries c and d as follows:

$$
\begin{aligned}
(12)\ a_c/a_d &= \exp[\square_c^* \,\square\, \square_d^*] \\
&= \square_{n=1}^{N} \square_{k=1}^{K} p_{cnk}^{1/NK} / \square_{n=1}^{N} \square_{k=1}^{K} p_{dnk}^{1/NK}.
\end{aligned}
$$

Thus the PPP between countries c and d can be calculated as the geometric mean of all of the country c prices divided by the geometric mean of all of the country d prices. Hence the PPP's are *transitive* in this equal sample size case so that $[a_c/a_d]\ [a_d/a_e] = [a_c/a_e]$ for any 3 countries, c, d and e.[14] Note also if we dropped some countries from the comparison, then as long as the sample of prices in the remaining countries was not altered, the PPP's in the remaining countries would remain invariant in the ratio form given by (12). This is a very useful property.

Once the least squares estimators $\square_n^*$ and $\square_c^*$ have been determined by (8) and (9) above, the sample residuals e_{cnk} can be calculated as follows:

$$(13)\ e_{cnk} \equiv y_{cnk} \,\square\, \square_c^* \,\square\, \square_n^* ; \qquad c = 1,\ldots,C;\ n = 1,\ldots,N;\ k = 1,\ldots,K.$$

Standard least squares regression theory[15] tells us that these residuals may be used in order to calculate the following unbiased estimator for the variance $\square^2$ of the true error terms $\square_{cnk}$:

$$(14)\ \square^{*2} \equiv \square_{c=1}^{C} \square_{n=1}^{N} \square_{k=1}^{K} e_{cnk}^{2}/[CNK \,\square\, (C \,\square\, 1 + N)].$$

Note that if all of the sample residuals e_{cnk} happen to equal 0, then the international sample of prices satisfy the following equations:

$$(15)\ p_{cnk} = a_c^* b_n^* ; \qquad c = 1,\ldots,C;\ n = 1,\ldots,N;\ k = 1,\ldots,K$$

where $a_c^* \equiv \exp[\square_c^*]$ for $c = 2,\ldots,C$ and $b_n^* \equiv \exp[\square_n^*]$ for $n = 1,\ldots,N$. Thus if all of the sample residuals e_{cnk} equal 0, then the item prices are *proportional* across the C countries in the comparison and a_c^* is the factor of proportionality for country c. In the general case where the sample residuals e_{cnk} are not all equal to 0, then $\square^{*2}$ defined by (14) can

[14] This result was obtained by Triplett and McDonald (1977) in the context of a hedonic regression model. For the case where K = 1, Ferrari, Gozzi and Riani (1996), Hill (2004) and Rao (2004) obtained this result.

[15] See for example Theil (1971; 114).

serve as a quantitative measure of *the lack of proportionality* of the international sample of prices or as a measure of the relative *dissimilarity* of the prices.[16]

In order to work out the distribution of the estimated log parities π_c^*, we need to calculate the means and variances of the π_c^*. Using equations (2), (5) and (9), it can be shown that the mean and variance of π_c^* are given by the following expressions:

(16) $E\, \pi_c^* = \pi_c$; $c = 1,\ldots,C$;
(17) $\mathrm{Var}\, \pi_c^* = [C - 1]\sigma^2 / CNK$; $c = 1,\ldots,C$.

If in addition to our previous assumptions, we assume that the ε_{cnk} are independently normally distributed with means 0 and variances σ^2, then it can be shown[17] that the following statistics have t distributions with $CNK - (C - 1 + N)$ degrees of freedom:

(18) $[\pi_c^* - \pi_c][CNK]^{1/2}/[C - 1]^{1/2}\, \sigma^*$; $c = 1,\ldots,C$;

where σ^* is the square root of the σ^{*2} defined by (14).

We turn now to the much more realistic case where the item sample sizes are not equal across countries and where some countries may not be able to find some of the N items in their countries.

3. The Unweighted Country Product Dummy Method with Unequal Sample Sizes

In real life applications of the CPD method for making international comparisons of prices, it is almost never the case that all items from the common list of N items can be priced in all countries in the comparison. In fact, it can happen that an item from the common list is only present in a single country. In this section, we show how the equal sample size model presented in the previous section can be modified to deal with these difficulties.

We need to introduce some additional notation. For country c and item n, let K(c,n) be the number of item n price quotes that are collected in country c. Define the total number of item n price quotes that are collected across all C countries as K(0,n); i.e.:

(19) $K(0,n) \equiv K(1,n) + K(2,n) + \ldots + K(C,n)$; $n = 1,\ldots,N$.

Define the total number of price quotes collected in country c over all items and outlets as K(c,0); i.e.:

[16] If we want to bound the dissimilarity measure between 0 (minimum dissimilarity) and 1 (maximum dissimilarity), then we could use the measure $\sigma^{*2}/[1 + \sigma^{*2}]$. Diewert (2002a) took an axiomatic approach to measures of relative price dissimilarity but considered only the case of two countries. For the case C = 2, Allen and Diewert (1981) suggested the sum of squared sample residuals (which is (14) times a constant) as a measure of nonproportionality of two price vectors.

[17] See for example Theil (1971; 131).

$$(20)\ K(c,0) \equiv K(c,1) + K(c,2) + \ldots + K(c,N)\ ; \qquad c = 1,\ldots,C.$$

For any c,n, it can happen that K(c,n) = 0, which means that no item n prices were collected in country c. However, we assume that row and column totals, K(0,n) and K(c,0), are all positive so that the price of item n is collected in at least one country and each country collects at least one item price. The total number of item prices collected over all countries is K and this total can be obtained by summing the K(0,n) over n or the K(c,0) over c; i.e., we have:

$$(21)\ K \equiv \Sigma_{c=1}^{C}\ \Sigma_{n=1}^{N}\ K(c,n) = \Sigma_{c=1}^{C}\ K(c,0) = \Sigma_{n=1}^{N}\ K(0,n).$$

The following linear regression model is a counterpart to the equal sample size model (2) presented in the previous section:

$$(22)\ y_{cnk} = \pi_c + \eta_n + \varepsilon_{cnk}\ ; \qquad c = 1,\ldots,C;\ n = 1,\ldots,N;\ k = 1,\ldots,K(c,n)$$

where $y_{cnk} \equiv \ln p_{cnk}$ as in section 2, the π_c and η_n are parameters to be estimated and the ε_{cnk} are independently distributed error terms with means 0 and variances σ^2. If for any c and n, K(c,n) = 0 so that there are no item n prices collected in country c, then the corresponding equations in (22) are dropped.

The least squares estimators for the π_c and η_n can be obtained by solving the following minimization problem:

$$(23)\ \min_{\pi_c,\eta_n} \{\Sigma_{c=1}^{C}\ \Sigma_{n=1}^{N}\ \Sigma_{k=1}^{K(c,n)}\ [y_{cnk} - \pi_c - \eta_n]^2\}.$$

As in the previous section, the parameters π_c and η_n cannot be uniquely identified so we will choose to set the purchasing power parity of country 1, $a_1 \equiv \exp[\pi_1]$, equal to 1, which implies the following normalization on the parameters appearing in (23):

$$(24)\ \pi_1 = 0.$$

After substituting (24) into (23), we can differentiate (23) with respect to $\pi_2, \pi_3, \ldots, \pi_C$ and set the resulting partial derivatives equal to 0. The resulting $C - 1$ equations simplify to the following equations:[18]

$$(25)\ K(c,0)\pi_c + \Sigma_{n=1}^{N}\ K(c,n)\eta_n = \Sigma_{n=1}^{N}\ \Sigma_{k=1}^{K(c,n)}\ y_{cnk}\ ; \qquad c = 2,3,\ldots,C.$$

Now differentiate (23) with respect to $\eta_1,\ldots,\eta_N$ and set the resulting partial derivatives equal to 0. The resulting N equations simplify to the following equations:[19]

$$(26)\ \Sigma_{c=2}^{C}\ K(c,n)\pi_c + K(0,n)\eta_n = \Sigma_{c=1}^{C}\ \Sigma_{k=1}^{K(c,n)}\ y_{cnk}\ ; \qquad n = 1,\ldots,N.$$

[18] If K(c,n) = 0, then the corresponding y_{cnk} terms on the right hand side of (25) are omitted.

[19] If K(c,n) = 0, then the corresponding y_{cnk} terms on the right hand side of (26) are omitted.

The system of estimating equations (22), with (24) imposed, can be written in matrix form as follows:

$$(27)\ y = X\begin{bmatrix}\pi\\ \beta\end{bmatrix} + \varepsilon$$

where ε is a column vector of the ε_{cnk}, y is a column vector of the y_{cnk}, $\pi \equiv [\pi_2,\ldots,\pi_C]^T$, $\beta \equiv [\beta_1,\ldots,\beta_N]^T$ and X is a K by C − 1 + N matrix of dummy variables whose elements are equal to 0 or 1. The vector of least squares estimators for the π and β vectors which occur in (27) is given by:

$$(28)\ \begin{bmatrix}\pi^*\\ \beta^*\end{bmatrix} = (X^TX)^{-1}X^Ty$$

where $\pi^* \equiv [\pi_2^*, \pi_3^*,\ldots, \pi_C^*]^T$ and $\beta^* \equiv [\beta_1^*, \beta_2^*,\ldots, \beta_N^*]^T$. It turns out that the C − 1 + N by C − 1 + N matrix X^TX can be obtained by reading off the coefficients of the π_c and β_n in equations (25) and (26). More explicitly, we have:

$$(29)\ X^TX = \begin{bmatrix}
K(2,0) & 0 & \ldots & 0 & K(2,1) & K(2,2) & \ldots & K(2,N)\\
0 & K(3,0) & \ldots & 0 & K(3,1) & K(3,2) & \ldots & K(3,N)\\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots\\
0 & 0 & \ldots & K(C,0) & K(C,1) & K(C,2) & \ldots & K(C,N)\\
K(2,1) & K(3,1) & \ldots & K(C,1) & K(0,1) & 0 & \ldots & 0\\
K(2,2) & K(3,2) & \ldots & K(C,2) & 0 & K(0,2) & \ldots & 0\\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots\\
K(2,N) & K(3,N) & \ldots & K(C,N) & 0 & 0 & \ldots & K(0,N)
\end{bmatrix}.$$

This matrix can readily be constructed using the item sample sizes K(c,n) and the sums of these sample sizes across countries by item, the K(0,n), and the sample sizes across items by country, the K(c,0).[20]

It turns out that the C − 1 + N column vector matrix X^Ty can be obtained by reading off the sums on the right hand sides of equations (25) and (26). More explicitly, we have:

[20] See Rao (2004) and Dikhanov (2004; 3) for an explicit formula for $(X^TX)^{-1}$ in the case where there is at most one product price for each country.

$$(30)\ X^T y = \begin{bmatrix} \sum_{n=1}^{N} \sum_{k=1}^{K(2,n)} y_{2nk} \\ \sum_{n=1}^{N} \sum_{k=1}^{K(3,n)} y_{3nk} \\ \ldots \\ \sum_{n=1}^{N} \sum_{k=1}^{K(C,n)} y_{Cnk} \\ \sum_{c=1}^{C} \sum_{k=1}^{K(c,1)} y_{c1k} \\ \sum_{c=1}^{C} \sum_{k=1}^{K)c,2)} y_{c2k} \\ \ldots \\ \sum_{c=1}^{C} \sum_{k=1}^{K(c,N)} y_{cNk} \end{bmatrix}.$$

Thus the first element on the right hand side of (30) is the sum of all of the log prices collected in country 2, the second element is the sum of all of the log prices collected in country 3, … , and the $C - 1$ element is the sum of all of the log prices collected in country C. The last N elements on the right hand side of (30) are: the sum of all the item 1 log prices collected over all countries, the sum of all the item 2 log prices collected over all countries, … , and the sum of all the item N log prices collected over all countries.[21]

Once X^TX and X^Ty have been calculated using (29) and (30), it is straightforward to calculate the vectors of least squares estimators for the α and β vectors using (28).[22]

Having calculated α^* and β^* using (28), we can now calculate the sample residuals e_{cnk} using the following equations:[23]

$$(31)\ e_{cnk} \equiv y_{cnk} - \alpha_c^* - \beta_n^* ; \qquad c = 1,\ldots,C;\ n = 1,\ldots,N;\ k = 1,\ldots,K(c,n).$$

Standard least squares regression theory[24] tells us that these residuals may be used in order to calculate the following unbiased estimator for the variance σ^2 of the true error terms ε_{cnk}:

$$(32)\ \sigma^{*2} \equiv \sum_{c=1}^{C} \sum_{n=1}^{N} \sum_{k=1}^{K(c,n)} e_{cnk}^2/[K - (C - 1 + N)]$$

[21] Some countries c participating in the International Comparison Project may be forbidden to release the individual log price product data, y_{cnk}, for various reasons. By examining (29) and (30), it can be seen that the central processing agency can calculate the elements of X^TX and X^Ty provided that each country c reports the following data to the center: K(c,1), K(c,2),..., K(c,N) (the number of prices collected in country c for each item n) and $\sum_{k=1}^{K(c,1)} y_{c1k}$, $\sum_{k=1}^{K(c,2)} y_{c2k}$, ..., $\sum_{k=1}^{K(c,N)} y_{cNk}$ (the sum of the outlet log prices for each item n in country c). Thus as long as each product cell has more than one outlet price collected for it, individual price data need not be forwarded to the central processing agency.

[22] Using the results in the Appendix, it can be seen that a sufficient condition for the existence of the inverse for X^TX is that there exists a country that collects at least one price quote for each of the N commodities in the basic heading category.

[23] Define $\alpha_1^* \equiv 0$.

[24] See for example Theil (1971; 114).

where $K = □_{c=1}^{C} □_{n=1}^{N} K(c,n)$ is the total number of price quotes collected across all countries.[25]

Define the first $C□1$ diagonal elements of $(X^TX)^{□1}$ as $□_2, □_3, \ldots , □_C$. Assuming that the true residuals $□_{cnk}$ are independently normally distributed with mean 0 and variance $□^2$, then it can be shown[26] that the following statistics have t distributions with $K \quad □(C □ 1 +N)$ degrees of freedom:

$$(33) \quad [□_c^* □ □_c] □_c^{1/2} / □^* ; \qquad c = 2,\ldots,C$$

where $□^{*2}$ is defined by (32).

Comparing our new distributional results (33) with our previous distributional results in the equal sample size case, (18), it can be seen that in the present unequal sample size case, the variances of the $□_c^*$ will *vary* as the $□_c$ vary (whereas this did not happen in the equal sample case). Note that the variance of $□_c^*$ will decrease if $□_c$ decreases. It can be shown that $□_c$ will decrease as K(c,0) increases, where K(c,0) is the total number of price quotes collected in country c.[27] This makes intuitive sense: the variances on the purchasing power parities will be *smaller* for the countries that have collected a *larger* number of log price quotes that enter into the overall regression.

As in the previous section, it seems reasonable to use the estimated variance of the regression, $□^{*2}$ defined by (32), as an overall measure of the degree of dissimilarity in the price structures of the countries in the comparison.

In the following section, we show that if an item is priced in only one country, then those item prices have no effect on the least squares log purchasing power parities, $□_2^*, □_3^*, \ldots, □_C^*$.

4. The Case where an Item is Priced in Only One Country

Consider the model presented in the previous section but suppose now that the prices for item or product n* are collected only in country c*.[28] In this section, we show that the

[25] Formula (31) requires the individual log price data y_{cnk} in order to calculate the individual log price residuals e_{cnk}. As before, it may not be possible for some countries c to give the central processing agency the individual log price data. In this case, note that for each c and n, $□_{k=1}^{K(c,n)} e_{cnk}^2 = □_{k=1}^{K(c,n)} [y_{cnk} □ □_c^* □ □_n^*]^2 = □_{k=1}^{K(c,n)} y_{cnk}^2 □ 2[□_{k=1}^{K(c,n)} y_{cnk}][□_c^* + □_n^*] + K(c,n)][□_c^* + □_n^*]^2$. Thus in order to calculate $□_{k=1}^{K(c,n)} e_{cnk}^2$, country c needs to report only the sum of the log prices for each product n, $□_{k=1}^{K(c,n)} y_{cnk}$, over all outlets k, the sum of the squares of the log prices for each product n, $□_{k=1}^{K(c,n)} y_{cnk}^2$, over all outlets k and the number of price quotes collected in each product n cell, K(c,n). Recall that this information set is also sufficient to calculate the $□_c^*$ and the $□_n^*$.

[26] See for example Theil (1971; 131).

[27] The diagonal element of X^TX that corresponds to $□_c$ is K(c ,0); see the northwest block of (29). It turns out that $\partial □_c(t)/\partial t = □[□_c]^{□2}$ where $t \equiv K(c,0)$. Thus $□_c$ decreases as K(c,0) increases. This last result can be derived using the matrix differentiation formula $dA^{□1}(t)/dt = □ A^{□1}(t) [dA(t)/dt]A^{□1}(t)$.

[28] We assume $c^* \neq 1$.

least squares purchasing power parities from the full sample of prices are identical to the least squares purchasing power parities that result from a least squares model that simply omits the prices of item n* from the sample of prices.

Since item n* is priced only in country c*, we have:

(34) $K(c^*,n) = 0$ for all $n \neq n^*$ and
(35) $K(c,n^*) = 0$ for all $c \neq c^*$.

Thus, using (34), equation $n = n^*$ in (26) becomes:

$$(36)\ K(c^*,n^*)\square_{c^*} + K(c^*,n^*)\square_{n^*} = \square_{k=1}^{K(c^*,n^*)} y_{c^*n^*k}.$$

Using (35), equations (25) for $c \neq c^*$ become:

$$(37)\ \square_{n=1,n\neq n^*}^{N} K(c,n)\square_c + \square_{n=1,n\neq n^*}^{N} K(c,n)\square_n = \square_{n=1,n\neq n^*}^{N} \square_{k=1}^{K(c,n)} y_{cnk}\ ;\quad c = 2,3,\ldots,C \text{ but } c \neq c^*.$$

Equation (25) for $c = c^*$ can be rewritten in the following form:

$$(38)\ [\square_{n=1,n\neq n^*}^{N} K(c^*,n) + K(c^*,n^*)]\square_c + [\square_{n=1,n\neq n^*}^{N} K(c^*,n) + K(c^*,n^*)]\square_n = \square_{n=1,n\neq n^*}^{N} \square_{k=1}^{K(c^*,n)} y_{c^*nk} + \square_{k=1}^{K(c^*,n^*)} y_{c^*n^*k}.$$

Now subtract (36) from (38) and we obtain the following equation:

$$(39)\ \square_{n=1,n\neq n^*}^{N} K(c^*,n)\square_c + \square_{n=1,n\neq n^*}^{N} K(c^*,n)\square_n = \square_{n=1,n\neq n^*}^{N} \square_{k=1}^{K(c^*,n)} y_{c^*nk}\ .$$

Thus if $\square_2^*, \square_3^*,\ldots, \square_C^*$ and $\square_1^*, \square_2^*,\ldots, \square_N^*$ satisfy the full sample equations (25) and (26) in the previous section, then we have shown that $\square_2^*, \square_3^*,\ldots, \square_C^*$ and the $\square_n^*$ for $n \neq n^*$ satisfy equations (37) and (39) and equations (26) for all n but excluding the equation for n*. But this latter set of C□1 plus N□1 equations are the precise counterparts to the full sample equations (25) and (26), except that all log prices pertaining to commodity n* are dropped from the comparison.

Thus we can obtain the least squares logarithms of the purchasing power parities, the $\square_2^*$, $\square_3^*,\ldots, \square_C^*$, in one of two equivalent ways if commodity n* is priced only in country c*: (a) we can solve the full sample set of least squares equations (25) and (26) in the previous section or (b) we can drop commodity n* from the comparison, solve equations (37), (39) and equations (26) (omitting the equation for n = n*) in order to obtain the same parities $\square_2^*, \square_3^*,\ldots, \square_C^*$.[29]

[29] Equation (36) may be used to calculate $\square_{n^*}$ if we take alternative path (b).

Thus if an item is priced in only one country, then these item prices do not enter into the least squares purchasing power parities between the C countries in the comparison.[30]

In the following section, we develop a weighted counterpart to the unweighted model presented in section 3.

5. The Weighted Country Product Dummy Model with Unequal Sample Sizes

The stochastic model of prices described by (22)-(24) in section 3 above assumed that each log price y_{cnk} was of equal importance in the least squares minimization problem (23). However, best practice index number theory typically involves weighting prices by their economic importance. Thus a particular log price may represent a transaction that has either a large or small expenditure associated with it and it does not seem "fair" that prices that represent large expenditures are given the same weight as those representing small expenditures. However, there are several ways that this economic importance could be measured. One could weight by either the *quantities transacted* in the two situations or by the *expenditures* pertaining to that component. However, since we are comparing prices across large and small countries, then using either of these two methods of weighting will give too much weight to the large countries. Hence, we will follow the example of Theil (1967; 136-137) and weight the importance of each log commodity price by its *share* in the country's national expenditures in the class of commodities under consideration in the comparison.[31] Note that this weighting scheme is a "democratic" weighting scheme, where each country's prices are given the same aggregate weight, as opposed to a "plutocratic" weighting scheme, which would give more weight to countries that had larger expenditures (in a common currency) on the class of expenditures under consideration.

Note that we are considering a rather idealized situation in this section, where instead of collecting a sample of prices for the commodity group under consideration, we envisage a situation where the national price collector *has detailed price and associated expenditure data on every transaction made in the country on the relevant class of commodities within the reference period*. With the advent of scanner data, this assumption is not completely unrealistic but even if the relevant data are not available, it is useful to work out what the "best" purchasing power parities would be if all relevant information were available. This ideal case can then provide some guidance for price collection strategies in non-ideal cases.

Thus in the present section, items 1 to N now are regarded as a comprehensive and complete list of all items in the relevant domain of definition for the basic heading price comparison project under consideration over all countries in the comparison. We also assume that within each country, we have a comprehensive listing of each transaction for

[30] However, these item n* prices would enter into the estimated variance $\square^{*2}$ defined by (32) if the full sample were used in the comparison.

[31] Other papers that pursue a weighted approach are Prasada Rao (1990), (1995) (2001) (2002) (2004), Heston, Summers and Aten (2001), Sergueev (2001) (2003) and Hill (2004). The approach in this section generalizes the approach taken by Rao down to the level of individual transactions.

each of the N items. Thus for item n in country c, we assume that there are K(c,n) transactions involving the item[32] and that the unit value price for the kth such transaction is p_{cnk} and the associated quantity transacted is q_{cnk} for k = 1,2,…,K(c,n). As in section 3, $y_{cnk} \equiv \ln p_{cnk}$ is the logarithm of the price p_{cnk}. Within each country c, we use the prices and quantities p_{cnk} and q_{cnk} in order to form the following transactions expenditure shares:

(40) $s_{cnk} \equiv p_{cnk}q_{cnk} / \square_{i=1}^{N} \square_{j=1}^{K(c,n)} p_{cij}q_{cij}$; n = 1,…,N ; k = 1,…,K(c,n).

For each country c, these expenditure shares sum up to 1:

(41) $\square_{n=1}^{N} \square_{k=1}^{K(c,n)} s_{cnk} = 1$; c = 1,…,C.

Our *weighted according to economic importance* counterpart to the unweighted least squares minimization problem (23) in section 3 is:

(42) $\min_{\square_c,\square_n} \{\square_{c=1}^{C} \square_{n=1}^{N} \square_{k=1}^{K(c,n)} s_{cnk} [y_{cnk} \square \square_c \square \square_n]^2\}$.

As in section 3, the parameters $\square_c$ and $\square_n$ cannot be uniquely identified so we will choose to set the purchasing power parity of country 1, $a_1 \equiv \exp[\square_1]$, equal to 1, which implies the following normalization on the parameters appearing in (43):

(43) $\square_1 = 0$.

In order to obtain a classical regression model that has a solution consistent with the least squares minimization problem (42) subject to the constraint (43), we need to multiply each y_{cnk} by the square root of the associated expenditure share s_{cnk} defined by (40); i.e., the counterparts to our linear regression equations (22) are now the following equations:

(44) $s_{cnk}^{1/2}\, y_{cnk} = s_{cnk}^{1/2}\, \square_c + s_{cnk}^{1/2}\, \square_n + \square_{cnk}$; c = 1,…,C; n = 1,…,N; k = 1,…,K(c,n)

where $y_{cnk} \equiv \ln p_{cnk}$ as in section 3, the $\square_c$ for c = 2,…,C and $\square_n$ for n = 1,….,N are parameters to be estimated ($\square_1$ is set equal to 0) and the $\square_{cnk}$ are independently distributed error terms with means 0 and variances $\square^2$. If for any c and n, K(c,n) = 0 so that there are no item n prices collected in country c, then the corresponding equations in (44) are dropped.

In order to rigorously justify the linear regression model (44), we need to assume that the variance of y_{cnk} is proportional to $\square^2/s_{cnk}$ for c = 1,…,C; n = 1,…,N; k = 1,…,K(c,n).[33]

[32] Of course, K(c,n) could be 0. Note that the K(c,n) that appears in this section (the number of transactions involving commodity n in country c) is very much bigger than the K(c,n) that was used in section 3, which denoted the number of price quotes collected for product n in country c.

[33] An alternative way for justifying the weighted model (44) is to argue that each logarithmic price ln p_{cnk} should be repeated according to its economic importance; i.e., if consumers are spending e_{cnk} dollars on commodity n in country c, then ln p_n^c should appear e_{cnk} times in the regression instead of only once. In order to standardize these weights across countries, we change the e_{cnk} weight to s_{cnk}.

This means that the smaller is the expenditure share s_{cnk}, the bigger will be the variance of y_{cnk}. This assumption may not be precisely justified from a statistical point of view but we feel that solving the weighted least squares problem (42) leads to very reasonable purchasing power parities from the viewpoint of classical index number theory, where weighting by economic importance is regarded as being extremely important. It is worth quoting Irving Fisher on the importance of weighting:

> "It has already been observed that the purpose of any index number is to strike a 'fair average' of the price movements—or movements of other groups of magnitudes. At first a *simple* average seemed fair, just because it treated all terms alike. And, in the absence of any knowledge of the relative importance of the various commodities included in the average, the simple average *is* fair. But it was early recognized that there are enormous differences in importance. Everyone knows that pork is more important than coffee and wheat than quinine. Thus the quest for fairness led to the introduction of weighting." Irving Fisher (1922; 43).

> "But on what principle shall we weight the terms? Arthur Young's guess and other guesses at weighting represent, consciously or unconsciously, the idea that relative *money values* of the various commodities should determine their weights. A value is, of course, the product of a price per unit, multiplied by the number of units taken. Such values afford the only common measure for comparing the streams of commodities produced, exchanged, or consumed, and afford almost the only basis of weighting which has ever been seriously proposed." Irving Fisher (1922; 45).

Thus we argue that since the statistical model defined by (44) corresponds to the weighted least squares minimization problem (42) that uses economic weighting, then the statistical model (44) will be theoretically consistent with economic weighting and be approximately correct from a statistical perspective so that it can be used to provide approximate standard errors for the estimated purchasing power parities.[34]

After substituting (43) into (42), we can differentiate (42) with respect to $\square_2, \square_3, \ldots, \square_C$ and set the resulting partial derivatives equal to 0. The resulting C □ 1 equations simplify to the following equations:[35]

$$(45)\ \square_{n=1}^{N} \square_{k=1}^{K(c,n)} s_{cnk}\, \square_c + \square_{n=1}^{N} \square_{k=1}^{K(c,n)} s_{cnk}\, \square_n = \square_{n=1}^{N} \square_{k=1}^{K(c,n)} s_{cnk}\, y_{cnk}\ ; \quad c = 2,3,\ldots,C.$$

Now differentiate (42) with respect to $\square_1,\ldots,\square_N$ and set the resulting partial derivatives equal to 0. The resulting N equations simplify to the following equations:[36]

[34] There is another way of proceeding and that is to solve the weighted least squares problem but instead of assuming the stochastic specification given by (44), assume that $y_{cnk} = \square_c + \square_n + \square_{cnk}$ where the $\square_{cnk}$ are independently distributed and have mean zero and variance $\square^2$. Thus we still solve equations (45) and (46) for the weighted least squares $\square_c$* and $\square_n$* but the resulting parameter estimates are no longer minimum variance unbiased for the new stochastic specification. However, the resulting estimates are still unbiased under the new stochastic specification and they are representative from the viewpoint of index number theory. Hill and Timmer (2004) and Deaton (2004) take this point of view.

[35] If K(c,n) = 0, then the corresponding y_{cnk} terms in equations (45) can be omitted. Alternatively, we can set the corresponding sum of expenditure shares, $\square_{k=1}^{K(c,n)} s_{cnk}$, equal to 0. We follow this latter convention in equations (47) and (48) which follow shortly.

$$(46)\ \Sigma_{c=2}^{C}\, \Sigma_{k=1}^{K(c,n)}\, s_{cnk}\, \alpha_c + \Sigma_{c=1}^{C}\, \Sigma_{k=1}^{K(c,n)}\, s_{cnk}\, \beta_n = \Sigma_{c=1}^{C}\, \Sigma_{k=1}^{K(c,n)}\, s_{cnk}\, y_{cnk}\ ; \qquad n = 1,\ldots,N.$$

As in section 3, we can write the system of estimating equations (44) in the matrix dummy variable form (27) and as before, the vector of least squares estimators for $\alpha \equiv [\alpha_2,\ldots,\alpha_C]^T$ and $\beta \equiv [\beta_1,\ldots,\beta_N]^T$ can be defined using (28). The new $C - 1 + N$ by $C - 1 + N$ matrix X^TX can be obtained by reading off the coefficients of the α_c and β_n in equations (45) and (46). More explicitly, we have:

$$(47)\ X^TX =$$

$$\begin{bmatrix}
\sum_{n=1}^{N}\sum_{k=1}^{K(2,n)} s_{2nk} & 0 & \cdots & 0 & \sum_{k=1}^{K(2,1)} s_{21k} & \cdots & \sum_{k=1}^{K(2,N)} s_{2Nk} \\
0 & \sum_{n=1}^{N}\sum_{k=1}^{K(3,n)} s_{3nk} & \cdots & 0 & \sum_{k=1}^{K(3,1)} s_{31k} & \cdots & \sum_{k=1}^{K(3,N)} s_{3Nk} \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & 0 & \cdots & \sum_{n=1}^{N}\sum_{k=1}^{K(C,n)} s_{Cnk} & \sum_{k=1}^{K(C,1)} s_{C1k} & \cdots & \sum_{k=1}^{K(C,N)} s_{CNk} \\
\sum_{k=1}^{K(2,1)} s_{21k} & \sum_{k=1}^{K(3,1)} s_{31k} & \cdots & \sum_{k=1}^{K(C,1)} s_{C1k} & \sum_{c=1}^{C}\sum_{k=1}^{K(c,1)} s_{c1k} & \cdots & 0 \\
\sum_{k=1}^{K(2,2)} s_{22k} & \sum_{k=1}^{K(3,2)} s_{32k} & \cdots & \sum_{k=1}^{K(C,2)} s_{C2k} & 0 & \cdots & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\sum_{k=1}^{K(2,N)} s_{2Nk} & \sum_{k=1}^{K(3,N)} s_{3Nk} & \cdots & \sum_{k=1}^{K(C,N)} s_{Ck} & 0 & \cdots & \sum_{c=1}^{C}\sum_{k=1}^{K(c,N)} s_{cN}
\end{bmatrix}.$$

It should be noted that $\Sigma_{n=1}^{N}\, \Sigma_{k=1}^{K(c,n)}\, s_{cnk} = 1$ for $c = 2,3,\ldots,C$ and so the $C-1$ by $C-1$ block in the northwest corner of the expression (47) for X^TX is equal to I_{C-1}, an identity matrix of rank $C-1$.

The above matrix can readily be constructed using the country shares on each transaction, the s_{cnk}. It turns out that the $C - 1 + N$ column vector matrix X^Ty can be obtained by

[36] If $K(c,n) = 0$, then the corresponding y_{cnk} terms in equations (46) can be omitted. Alternatively, we can set the corresponding sum of expenditure shares, $\Sigma_{k=1}^{K(c,n)}\, s_{cnk}$, equal to 0. We follow this latter convention in equations (47) and (48).

reading off the share weighted sums on the right hand sides of equations (45) and (46). More explicitly, we have:

$$(48)\ X^Ty = \begin{bmatrix} \sum_{n=1}^{N} \sum_{k=1}^{K(2,n)} s_{2nk} y_{2nk} \\ \sum_{n=1}^{N} \sum_{k=1}^{K(3,n)} s_{3nk} y_{3nk} \\ \ldots \\ \sum_{n=1}^{N} \sum_{k=1}^{K(C,n)} s_{Cnk} y_{Cnk} \\ \sum_{c=1}^{C} \sum_{k=1}^{K(c,1)} s_{c1k} y_{c1k} \\ \sum_{c=1}^{C} \sum_{k=1}^{K)c,2)} s_{c2k} y_{c2k} \\ \ldots \\ \sum_{c=1}^{C} \sum_{k=1}^{K(c,N)} s_{cNk} y_{cNk} \end{bmatrix}.$$

Thus the first element on the right hand side of (48) is the share weighted sum of all of the log prices collected in country 2, the second element is the share weighted sum of all of the log prices collected in country 3, … , and the C − 1 element is the share weighted sum of all of the log prices collected in country C. The last N elements on the right hand side of (48) are: the share weighted sum of all the item 1 log prices collected over all countries, the share weighted sum of all the item 2 log prices collected over all countries, … , and the share weighted sum of all the item N log prices collected over all countries.

Once X^TX and X^Ty have been calculated using (47) and (48), it is straightforward to calculate the vectors of least squares estimators for the π and β vectors using (28). In the Appendix, we develop a sufficient condition for the existence of the inverse of X^TX.[37]

Having calculated π^* and β^* using (28), we can now calculate the sample residuals e_{cnk} using the following equations:[38]

$$(49)\ e_{cnk} \equiv s_{cnk}^{1/2}\,[y_{cnk} - \pi_c^* - \beta_n^*]\ ; \qquad c = 1,\ldots,C;\ n = 1,\ldots,N;\ k = 1,\ldots,K(c,n).$$

Standard least squares regression theory[39] tells us that these residuals may be used in order to calculate the following unbiased estimator for the variance σ^2 of the error terms ε_{cnk} which appear in the transformed linear regression model (44):

$$(50)\ \sigma^{*2} \equiv \sum_{c=1}^{C} \sum_{n=1}^{N} \sum_{k=1}^{K(c,n)} e_{cnk}^2/[K - (C - 1 + N)]$$

where $K = \sum_{c=1}^{C} \sum_{n=1}^{N} K(c,n)$ is the total number of price quotes collected across all countries.

[37] The sufficient condition is the existence of a country that has a transaction for each of the N commodities in the basic heading category.

[38] Define $\pi_1^* \equiv 0$.

[39] See for example Theil (1971; 240).

Define the first C□1 diagonal elements of $(X^TX)^{\square 1}$ as $\square_2, \square_3, \dots , \square_C$. Assuming that the true (transformed) residuals $\square_{cnk}$ are independently normally distributed with mean 0 and variance $\square^2$, then it can be shown[40] that the following statistics have t distributions with K □ (C □ 1 +N) degrees of freedom:

$$(51)\ [\square_c^{*} \square \square_c]\, \square_c^{1/2} / \square^{*}\ ; \qquad c = 2,\dots,C$$

where $\square^{*2}$ is defined by (32).

As in previous sections, it seems reasonable to use the estimated variance of the regression, $\square^{*2}$ defined by (50), as an overall measure of the degree of dissimilarity in the price structures of the countries in the comparison.[41]

In the following section, we test the reasonableness of the share weighted least squares estimators for the country PPP’s by looking more closely at the case where there are only two countries in the comparison.

6. The Weighted Country Product Dummy Model for the Two Country Case

In order to gain some insight into the structure of the weighted country product dummy purchasing power parities defined in the previous section, in this section we consider the case where there are only two countries involved in the price comparison. In this case, equations (45) and (46) become the following equations:

$$(52)\ \square_{n=1}^{N} \square_{k=1}^{K(2,n)} s_{2nk}\, \square_2 + \square_{n=1}^{N} \square_{k=1}^{K(2,n)} s_{2nk}\, \square_n = \square_{n=1}^{N} \square_{k=1}^{K(2,n)} s_{cnk}\, y_{cnk}\ ;$$

$$\begin{aligned}(53)\ [\square_{k=1}^{K(1,n)} s_{1nk} + \square_{k=1}^{K(2,n)} s_{2nk}]\square_n &= \square_{k=1}^{K(1,n)} s_{1nk}\, y_{1nk} + \square_{k=1}^{K(2,n)} s_{2nk}\, y_{2nk} \square \square_{k=1}^{K(2,n)} s_{2nk}\, \square_2 \\ &= \square_{k=1}^{K(1,n)} s_{1nk}\, y_{1nk} + \square_{k=1}^{K(2,n)} s_{2nk}\, [y_{2nk} \square \square_2]. \end{aligned} \qquad n = 1,\dots,N$$

The N equations in (53) may be used to solve for the $\square_n$ in terms of $\square_2$.[42] If we substitute these equations into equation (52), we obtain a single equation in the single unknown, $\square_2$. The solution to this equation is:

$$(54)\ \square_2^{*} = W^{\square 1} \square_{n=1}^{N}\{\square_{i=1}^{K(1,n)} \square_{j=1}^{K(2,n)} [\square_{k=1}^{K(1,n)} s_{1nk} + \square_{k=1}^{K(2,n)} s_{2nk}]^{\square 1}\, s_{1ni} s_{2nj} \ln[p_{2nj}/p_{1ni}]\}$$

where the total weight factor W is defined as

$$(55)\ W \equiv \square_{n=1}^{N}\{\square_{i=1}^{K(1,n)} \square_{j=1}^{K(2,n)} [\square_{k=1}^{K(1,n)} s_{1nk} + \square_{k=1}^{K(2,n)} s_{2nk}]^{\square 1}\, s_{1ni} s_{2nj}\}.$$

[40] See for example Theil (1971; 240).

[41] For related ideas, see Hill an Timmer (2004).

[42] We need to assume that the share sums [$\square_{k=1}^{K(1,n)} s_{1nk} + \square_{k=1}^{K(2,n)} s_{2nk}$] are greater than 0 for n = 1,...,N. This means that for each item n, at least one country has a transaction involving that item.

Thus it can be seen that $\square_2^*$ is a somewhat complicated *weighted average of all possible log price relatives of an item n price in country 2 relative to an item n price in country 1* (these are the $\ln[p_{2nj}/p_{1ni}]$ over all possible positive prices for n in both countries) *over all items n*. Since the weights are symmetric in the data for the two countries, it can be seen that the bilateral index number formula defined by the exponential of (54) satisfies the time reversal test.[43] It can also be seen that the index number formula defined by (54) and (55) can deal with situations where say item n* has transactions in one country but not the other. In this situation, it can be seen that these item n* prices play no role in (54) since the corresponding item transaction shares will be zero in one of the two countries and thus the prices of item n* will be zeroed out in the formula (54).[44] Examination of formula (54) shows that greater weight will be placed on the prices that have relatively large expenditure shares in both countries.[45]

Finally, let us assume that there is at most one transaction for each item n in each of the two countries. In this case, formula (54) simplifies to the following formula:

$$(56)\ \square_2^* = W^{\square 1}\square_{n=1}^{N} [s_{1n1} + s_{2n1}]^{\square 1}\, s_{1n1}s_{2n1} \ln[p_{2n1}/p_{1n1}]$$

where

$$(57)\ W \equiv \square_{n=1}^{N} [s_{1n1} + s_{2n1}]^{\square 1}\, s_{1n1}s_{2n1}.$$

The above formula allows for one of the item expenditure shares in either country to be zero for each item but we require at least one item that is priced in both countries.

Finally, if there is exactly one transaction for each item in each country, then the shares s_{1n1} and s_{2n1} are positive for n = 1,…,N and formula (56) further simplifies to the following one:[46]

[43] Thus if we interchanged the data for countries 1 and 2, we find that the interchanged data PPP equals the reciprocal of the original data PPP; see Fisher (1922; 64) for a formal statement of the time reversal test in the context of bilateral index number theory. It should be noted that our present model covers a more general situation than traditional bilateral index number theory where there are N commodities, N positive prices and N positive expenditure shares in each country. Our present model is based on individual transactions rather than market totals so that the total number of transactions in each country can be quite different. Also our present model allows for some items to be present in one country (the corresponding transaction expenditure shares will be positive in that country) but not in the other (the corresponding transaction expenditure shares will be zero in that country).

[44] This property suggests that it may not be necessary to work out a separate penalty structure to penalize comparisons between countries where the degree of product matching varies between countries since the transactions weighted country product dummy method for making comparisons automatically adjusts for the lack of matching. For some related approaches on how to deal with different degrees of product matching, see Hill and Timmer (2004).

[45] This is also obvious from the nature of the weighted least squares problem (42).

[46] This result was obtained by Diewert (2002b; 4) who showed that this formula was a pseudo superlative formula; i.e., it approximates a superlative index number formula to the second order around an equal price and quantity vector for the two countries being compared; see Diewert (1978; 888) for material on superlative and pseudo superlative indexes.

$$(58)\ \square_2^* = W^{-1}\sum_{n=1}^{N} \{(1/2)[s_{1n1}]^{-1} + (1/2)[s_{2n1}]^{-1}\}^{-1} \ln[p_{2n1}/p_{1n1}]$$

where

$$(59)\ W \equiv \sum_{n=1}^{N} \{(1/2)[s_{1n1}]^{-1} + (1/2)[s_{2n1}]^{-1}\}^{-1}.$$

Note that $\{(1/2)[s_{1n1}]^{-1} + (1/2)[s_{2n1}]^{-1}\}^{-1}$ is the harmonic mean of the country 1 and 2 expenditure shares for item n.

While the bilateral index number formula defined by the exponential of (58) is not known to be superlative, it will approximate the superlative Törnqvist formula[47] reasonable closely.

Our overall conclusion at this point is that the transaction share weighted CPD purchasing power parities defined in the previous section provide a reasonable target set of parities that could be used in international comparisons of prices.

In the following section, we provide a descriptive statistics justification for the transaction share weighted CPD purchasing power parities defined in section 5.

7. A Descriptive Statistics Interpretation of the Transaction Weighted CPD Parities

Theil (1967; 136-137) proposed a stochastic approach to making comparisons of prices between two countries. He argued as follows. Assume that there is only one transaction for each item in each country so that the prices p_{1n1} and p_{2n1} are positive and the associated national expenditure shares s_{1n1} and s_{2n1} are also positive. Suppose we draw price relatives at random in such a way that each dollar of expenditure in country one has an equal chance of being selected. Then the probability that we will draw the nth price relative is equal to s_{1n1}, the country 1 expenditure share for commodity n. Then the overall mean (country 1 weighted) logarithmic price change for country 2 relative to country 1 is $\sum_{n=1}^{N} s_{1n1} \ln(p_{2n1}/p_{1n1})$. Now repeat the above mental experiment and draw price relatives at random in such a way that each dollar of expenditure in country 2 has an equal probability of being selected. This leads to the overall mean (country 2 weighted) logarithmic price change of $\sum_{n=1}^{N} s_{2n1} \ln(p_{2n1}/p_{1n1})$. Each of these measures of overall logarithmic price change seems equally valid so we could argue for taking a symmetric average of the two measures in order to obtain a final single measure of overall logarithmic price change[48]. Theil[49] argued that a nice symmetric index number formula

[47] This is the index P_T defined by (60) below.

[48] "The [asymmetric] price index (1.6) has certain merits. It is, for example, independent of the units in which we measure the quantities of the various commodities (tons, gallons, etc.). It has the disadvantage, however, of being one sided in the sense that it is based on the distribution of expenditure in the *a*th region. We could equally well apply our random selection procedure to the *b*th region, in which case, w_{ia} is replaced by w_{ib} in (1.5) and (1.6). We must conclude that (6) is an asymmetric index number, which is a disadvantage because the question asked is symmetric: If the price level of the *b*th region exceeds that of the *a*th by a factor 1.2, say, we should expect that the price level of the latter region exceed that of the former by a factor 1/1.2." Henri Theil (1967; 137).

can be obtained if we make the probability of selection for the nth price relative equal to the arithmetic average of the country 1 and 2 expenditure shares for commodity n. Using these probabilities of selection, Theil's final measure of overall logarithmic price change was

$$(60)\ \ln P_T(p^1,p^2,s^1,s^2) \equiv \Sigma_{n=1}^{N} (1/2)(s_{1n1} + s_{2n1}) \ln(p_{2n1}/p_{1n1})$$

where p^1 and p^2 are the N dimensional vectors of country 1 and 2 prices and s^1 and s^2 are the N dimensional vectors of country 1 and 2 expenditure shares.

Theil gave the following statistical interpretation of the right hand side of (60). Define the nth logarithmic price ratio r_n by:

$$(61)\quad r_n \equiv \ln(p_{2n1}/p_{1n1}) \quad \text{for } n = 1,\ldots,N.$$

Now define the discrete random variable, R say, as the random variable which can take on the values r_n with probabilities $\rho_n \equiv (1/2)[\, s_{1n1} + s_{2n1}]$ for $n = 1,\ldots,N$. Note that since each set of expenditure shares, s_{1n1} and s_{2n1}, sums to one, the probabilities ρ_n will also sum to one. It can be seen that the expected value of the discrete random variable R is

$$(62)\quad E[R] \equiv \Sigma_{n=1}^{N} \rho_n r_n = \Sigma_{n=1}^{N} (1/2)(s_{1n1} + s_{2n1}) \ln(p_{2n1}/p_{1n1}) = \ln P_T(p^0,p^1,s^0,s^1).$$

using (60) and (61). Thus the logarithm of the index P_T can be interpreted as *the expected value of the distribution of the logarithmic price ratios* in the domain of definition under consideration, where the N discrete price ratios in this domain of definition are weighted according to Theil's probability weights, $\rho_n \equiv (1/2)[\, s_{1n1} + s_{2n1}]$ for $n = 1,\ldots,N$. Taking antilogs of both sides of (60), we obtain the Törnqvist (1936) (1937) Theil price index, P_T.

Theil's stochastic approach is a nice one: the logarithm of the price index is simply the mean of a discrete probability distribution of the log price ratios and *it is not necessary to make assumptions about the exact distribution of error terms*. Our goal in this section is to obtain an analogue to Theil's approach in our much more complicated framework with many countries and many transactions in the same item instead of a single transaction in each item. Note that in the two country case, the situation is simplified because we can focus on the distribution of *relative* prices in the two countries. When we move to three or more countries, there are many relative price ratios and it becomes necessary to shift to an *absolute* price level framework with item price effects.

We use the same notation and framework as in section 5. We follow the example of Theil and define the transaction expenditure share s_{cnk} as the probability that $y_{cnk} \equiv \ln p_{cnk}$ occurs in country c for the kth transaction of item n in country c. However, this gives us

[49] "The price index number defined in (1.8) and (1.9) uses the n individual logarithmic price differences as the basic ingredients. They are combined linearly by means of a two stage random selection procedure: First, we give each region the same chance _ of being selected, and second, we give each dollar spent in the selected region the same chance ($1/m_a$ or $1/m_b$) of being drawn." Henri Theil (1967; 138).

a discrete probability distribution of log prices that pertains to prices in a given country c. We need to consider a global distribution of log prices across all countries. Thus we assign the global probability that the log price y_{cnk} occurs across all countries as:

(63) $\square_{cnk} \equiv s_{cnk}/C$; $\quad$ $c = 1,\ldots,C;\ n = 1,\ldots,N;\ k = 1,\ldots.K(c,n)$.

Since the country shares s_{cnk} sum to 1 over n and k for each c, it can be seen that the global probabilities $\square_{cnk}$ sum to 1; i.e., we have:

(64) $\square_{c=1}^{C}\ \square_{n=1}^{N}\ \square_{k=1}^{K(c,n)}\ \square_{cnk} = 1$.

Note that we are also following the example of Theil in that the log prices in each country have an *equal chance* of being drawn in the global distribution of log prices.

Define Y as a discrete random variable that takes on the values y_{cnk} with probability $\square_{cnk}$ for $c = 1,\ldots,C;\ n = 1,\ldots,N;\ k = 1,\ldots,K(c,n)$. Thus Y is a discrete random variable that summarizes the distribution of item prices by their relative transaction importance across the C countries in the comparison of prices.

In what follows, we will attempt to represent Y as the sum of 3 discrete random variables, A, B and E, each of which takes on values with the same probabilities as Y. Thus, we define A, B and E as discrete random variables that take on the values a_{cnk}, b_{cnk} and e_{cnk} respectively with the probabilities $\square_{cnk}$ for $c = 1,\ldots,C;\ n = 1,\ldots,N;\ k = 1,\ldots,K(c,n)$. However, we will impose some restrictions on the values a_{cnk}, b_{cnk} and e_{cnk} that the random variables A, B and E take on.

The realizations of the random variable A are restricted as follows:

(65) $a_{1nk} = \square_1 \equiv 0$ $\quad$ for $c = 1;\ n = 1,\ldots,N;\ k = 1,\ldots,K(1,n)$;
$a_{cnk} = \square_c$ $\quad$ for $c = 2,3,\ldots,C;\ n = 1,\ldots,N;\ k = 1,\ldots,K(c,n)$.

Thus a_{cnk} is equal to $\square_c$ with probability $\square_{cnk}/\square_{i=1}^{N}\ \square_{j=1}^{K(c,i)}\ \square_{cij}$ for $n = 1,\ldots,N$ and $k = 1,\ldots,K(c,n)$. Note that a_{cnk} depends only on c and equals the constant $\square_c$ for any commodity n that is priced in country c and for any outlet k that transacted commodity n in country c. Thus $\square_c$ can be interpreted as an average country price level effect for log prices that are collected in country c.

The realizations of the random variable B are restricted as follows:

(66) $b_{cnk} = \square_n$ for $c = 1,\ldots,C;\ k = 1,\ldots,K(c,n)$.

Thus b_{cnk} is equal to $\square_n$ with probability $\square_{cnk}/\square_{i=1}^{C}\ \square_{j=1}^{K(i,n)}\ \square_{inj}$ for $c = 1,\ldots,C$ and $k = 1,\ldots,K(c,n)$. Note that b_{cnk} depends only on n and equals the constant $\square_n$ for any commodity n that is priced in country c and for any outlet k that transacted commodity n in country c. Thus $\square_n$ can be interpreted as an average product price level effect for log prices that are collected in any country c for product n.

Thus the discrete random variables A and B have very simple structures: $a_{cnk} = \square_c$ for all n and k and $b_{cnk} = \square_n$ for all c and k.

The random variable E is unrestricted except that we impose the following C+N restrictions on the e_{cnk}:

$$(67)\ \square_{n=1}^{N} \square_{k=1}^{K(c,n)} \square_{cnk}\, e_{cnk} / \square_{i=1}^{N} \square_{j=1}^{K(c,i)} \square_{cij} = 0; \qquad c = 1,\ldots,C;$$
$$(68)\ \square_{c=1}^{C} \square_{k=1}^{K(c,n)} \square_{cnk}\, e_{cnk} / \square_{i=1}^{C} \square_{j=1}^{K(i,n)} \square_{inj} = 0; \qquad n = 1,\ldots,N.$$

Conditions (67) can be interpreted as follows. Conditional on a particular country c, the discrete probability distribution of E takes on the value e_{cnk} with probability $\square_{cnk}/\square_{i=1}^{N}\square_{j=1}^{K(c,i)} \square_{cij}$ for n = 1,…,N and k = 1,…,K(c,n). Equations (67) impose the restrictions that each of these C conditional probability distributions has mean 0. Similarly, conditions (68) can be interpreted as follows. Conditional on a particular item n, the discrete probability distribution of E takes on the value e_{cnk} with probability $\square_{cnk}/\square_{i=1}^{C} \square_{j=1}^{K(i,n)} \square_{inj}$ for c = 1,…,C and k = 1,…,K(c,n). Equations (68) impose the restrictions that each of these N conditional probability distributions has mean 0. In what follows, we interpret E as an "error" random variable and A and B as the systematic parts of the random variable Y.

We now set Y equal to the sum of A, B and E,

$$(69)\ Y = A + B + E,$$

and we ask whether it is possible to write Y in the additive form (69) where A, B and C satisfy the restrictions (65)-(68). Using these restrictions, we can use (69) to express the e_{cnk} in terms of the y_{cnk}, a_{cnk} and b_{cnk} as follows:

$$(70)\ e_{cnk} = y_{cnk} \square \square_c \square \square_n\, ; \qquad c = 1,\ldots,C;\ n = 1,\ldots,N;\ k = 1,\ldots,K(c,n)$$

where $\square_1 = 0$.

Using (70), it can be seen that equations (67) and (68) are equivalent to the following equations:

$$(71)\ \square_{n=1}^{N} \square_{k=1}^{K(c,n)} s_{cnk}\, e_{cnk} = 0; \qquad c = 1,\ldots,C;$$
$$(72)\ \square_{c=1}^{C} \square_{k=1}^{K(c,n)} s_{cnk}\, e_{cnk} = 0; \qquad n = 1,\ldots,N.$$

It can be seen that if the last $C\square 1$ equations in (71) are satisfied along with the N equations (72), then the first equation in (71) will also be satisfied.

Now substitute equations (70) into (71) and (72). If we omit the first equation in the resulting equations, it can be seen that we obtain precisely equations (45) and (46) in section 5. Thus the $\square_c$ and $\square_n$ that are required to obtain the additive decomposition of Y given by (69) along with the restrictions (65)-(68) can be obtained by setting the $\square_c$ and $\square_n$ equal to the corresponding weighted least squares estimators $\square_c^*$ and $\square_n^*$ that solve the transactions weighted least squares minimization problem (42) subject to the restriction (43).

Thus the C+N conditional distributions of Y have the same means as the sum of the corresponding conditional distributions of A and B where the conditional distributions of A and B are constant with respect to c and n respectively. This decomposition can be viewed as a suitable generalization of Theil's stochastic approach to index number theory.

Why is the above result important? First, we stress the main advantage of Theil's stochastic approach. The main advantage of his approach is that it is completely nonparametric; i.e., we do not have to have to make problematical assumptions as to what the "true" distribution of log price relatives is: the distribution is simply the empirical population distribution and we take the mean of this (weighted) log price distribution as our desired summary measure of this distribution of log price relatives. Now, as soon as we move to more than 2 time periods or countries, we encounter a new difficulty; i.e., there are many relative price ratios between countries (not just a single price ratio for each commodity). To deal with this difficulty, we move away from the relative price formulation to a price levels formulation but now we have a two dimensional classification to deal with: the country and product dimensions both influence price.[50] A counterpart to Theil's approach in the many country case then is to assume that *the country and product conditional means of log prices* (these are the realizations of the A and B random variables) *are the descriptive statistics of interest.*

In the following section, we give an alternative descriptive statistics interpretation for the weighted least squares estimators $\square_c^*$ and $\square_n^*$ that solve the transactions weighted least squares minimization problem (42) subject to the restriction (43). In order to obtain this alternative justification, we convert the usual linear regression model that is based on continuous random variables into a discrete random variable framework.

8. A Discrete Random Variable Approach to Linear Regression Models

Let $y_{cnk} \equiv \ln p_{cnk}$ be the log price for outlet k for product n in country c and let $\square_{cnk}$ defined by (63) in the previous section be the global probability that this price occurs across all countries in the comparison. As in the previous section, we assume that Y is a discrete random variable that takes on the value y_{cnk} with probability $\square_{cnk}$ for c = 1,…,C; n = 1,…,N and k = 1,…,K(c,n).

As in section 3 above, we assume that the realizations of the random variable Y (the y_{cnk}) have the following decomposition:

[50] In the two country case, by considering relative prices, we remove the commodity classification problem.

(73) $y_{cnk} = \square_c + \square_n + u_{cnk}$; $\qquad$ $c = 1,\ldots,C$; $n = 1,\ldots,N$; $k = 1,\ldots,K(c,n)$;
(74) $\square_1 = 0$.

The $\square_c + \square_n$ terms on the right hand side of (73) are regarded as constants (to be determined somehow) whereas the terms u_{cnk} are regarded as realizations of another discrete random variable U which takes on the values u_{cnk} with probabilities $\square_{cnk}$ for $c = 1,\ldots,C$; $n = 1,\ldots,N$ and $k = 1,\ldots,K(c,n)$. Note that Y and U have the same probabilities. The $\square_c + \square_n$ terms on the right hand side of (73) are regarded as the systematic components of the realizations of the random variable Y while U is regarded as the "nonsystematic" or "random" part of Y.

It will be convenient to rewrite the model defined by (73) and (74) in the form of a linear regression model. Thus we rewrite the realizations of Y, the y_{cnk}, in terms of a single index j (rather than in terms of the 3 indexes c, n and k) so that $\{y_{cnk} : c = 1,\ldots,C ; n = 1,\ldots,N ; k = 1,\ldots,K(c,n)\}$ becomes $\{y_j : j = 1,\ldots,J\}$ where $J = \square_{c=1}^{C}\square_{n=1}^{N} K(c,n)$ is the total number of price quotations collected over all countries in the comparison for the particular basic heading under consideration. The probabilities $\square_{cnk}$ are also reordered in terms of a single index j so that the jth log price, y_j, now has the probability $\square_j$ for $j = 1,\ldots,J$. The model defined by (73) and (74) can now be rewritten as follows:

(75) $y_j = \square_{k=1}^{C+N\square 1} x_{jk}\, \square_k + u_j$; $\qquad$ $j = 1,\ldots,J$

where $\square^T \equiv [\square_1,\ldots,\square_{C+N\square 1}] \equiv [\square_2,\ldots,\square_C; \square_1,\ldots,\square_N]$ is the combined vector of log purchasing power parities (the $\square_c$'s) and the vector of log product premiums (the $\square_n$'s) and the x_{jk} are known constants, equal to either 0 or 1. The J y_j observations in (75) can be stacked into a vector $y \equiv [y_1,\ldots,y_J]^T$ and we obtain the following *linear regression model* in matrix form:

(76) $y = X\square + u$

where $X \equiv [x_{jk}]$ is the J by $C+N\square 1$ matrix of dummy variables and $u \equiv [u_1,\ldots,u_J]^T$ is the vector of realizations of the random variable U. The vector of probabilities associated with the discrete random variables Y and U is $\square \equiv [\square_1,\ldots,\square_J]^T$ where $\square_j$ is the probability associated with y_j and u_j, the jth realizations of the random variables Y and U.

In the model defined by (76), it is assumed that we know y and X but that we do not know the vector of constants $\square$ and the "error" vector u at this stage. The matrix X is regarded as a fixed nonrandom matrix but y and u are regarded as the realizations of discrete probability distributions where both y_j and u_j have known probability $\square_j$ for $j = 1,\ldots,J$.

Our estimation problem is to determine the vector of fixed effects, $\square$. The principle that we use to determine the components of $\square$ is the following one: we require that, on average, the error vector u is orthogonal to each column of the X matrix. This translates into the following C+N−1 requirements:[51]

(77) $\sum_{j=1}^{J} \pi_j x_{jk} u_j = 0$; $\qquad\qquad$ k = 1,…,C+N−1.

Thus, on average, we require that the "error" vector u be perpendicular to each column of the X matrix.[52]

Using matrix notation, conditions (77) can be rewritten as follows:

(78) $X^T \Pi u = 0_{C+N-1}{}^T$

where Π is a diagonal J by J matrix with the elements of the vector π running down the main diagonal. Using (76), the vector u is equal to $y - X\beta$ and replacing u in (78) by $y - X\beta$ shows that conditions (77) are equivalent to the following matrix equation:

(79) $X^T \Pi [\, y - X\beta] = 0_{C+N-1}{}^T$ or

(80) $X^T \Pi y = X^T \Pi X\beta$.

Hence if $(X^T \Pi X)^{-1}$ exists, the vector β is uniquely determined by (80) as follows:

(81) $\beta = (X^T \Pi X)^{-1} X^T \Pi y$.

Thus if $(X^T \Pi X)^{-1}$ exists, then conditions (77) lead to a unique solution for the β vector defined by (81) and hence, the required log PPP's are determined by this discrete variable linear regression approach.

Recall the weighted CPD model that was defined by the weighted least squares minimization problem (42) subject to the normalization (43). Note that if we divide the country expenditure shares s_{cnk} by the constant 1/C, these s_{cnk} turn into the country probabilities $\rho_{cnk} \equiv s_{cnk}/C$ defined by equations (63) in the previous section. Dividing each s_{cnk} by C will not change the solution to the weighted least squares minimization problem (42) subject to the normalization (43) but it can be verified that the least squares minimization problem that modifies (42) by dividing each s_{cnk} by C leads precisely to equations (80) above. Hence, the β solution defined by (81) will also solve the weighted least squares minimization problem described in section 5. Thus the model developed in the present section can be used to justify the weighted least squares model that was used

[51] If the X matrix consisted of just a single column of ones, then conditions (77) would reduce to the single condition $\sum_{j=1}^{J} \pi_j u_j = 0$ and this condition is equivalent to requiring that the expectation of the discrete random variable U be 0; i.e., (77) reduces to EU = 0 in this case. For the case of a general X, conditions (77) can be interpreted as the conditions $EX^TU = 0_{C-1+N}$ where X is regarded as a fixed matrix.

[52] If a linear combination of the columns of the X matrix is equal to a column of ones, then it can be shown that each column of the X matrix is *uncorrelated* with U; i.e., we have for each k, $\sum_{j=1}^{J} \pi_j [x_{jk} - EX_{\bullet k}][u_j - EU] = 0$ where $EU \equiv \sum_{j=1}^{J} \pi_j u_j$ and $EX_{\bullet k} \equiv \sum_{j=1}^{J} \pi_j x_{jk}$.

in section 5. However, even though these two models have the same solution, their assumptions are different. In particular, there is a sampling theory associated with the model defined in section 5 so that confidence intervals for the log parities (the π_c) are developed there but do not exist for the present model. Thus the model developed in the present section is again a *descriptive statistics model* along the lines of Theil's stochastic model for the case of two countries. The models in the present section and the previous section essentially give us various conditional means of a discrete sampling distribution but there is no distribution theory associated with these conditional means.[53]

It is possible to follow Theil's (1967) example and define a measure of goodness of fit of the model. Thus define the following measure of *the degree of nonproportionality of prices*:

$$
\begin{aligned}
(82)\ T &\equiv \sum_{j=1}^{J} \omega_j \,[y_j - \sum_{k=1}^{C+N-1} x_{jk}\,\beta_k]^2 && \text{where the } \beta_k \text{ are defined by (81)}\\
&= [y - X\beta]^T \,\Omega\, [y - X\beta] \\
&= u^T \,\Omega\, u && \text{where u is defined in (76)}\\
&= (1/C) \sum_{c=1}^{C} \sum_{n=1}^{N} \sum_{k=1}^{K(c,n)} e_{cnk}{}^2
\end{aligned}
$$

where the errors e_{cnk} were defined by (49) in section 5.[54] It can be seen that T must be nonnegative. If T equals zero, then prices are exactly proportional across countries. On the other hand, the more positive T is, then the more nonproportional are the prices across the countries in the comparison and the more uncertain are the associated PPP's.

It should be noted that the discrete distribution of errors approach to linear regression analysis outlined in this section could be applied in many other contexts, including in particular, hedonic regression analysis where information on model sales is available.

9. The Use of Approximate Sampling Weights

The material in section 5 gives us a theoretical framework for a comprehensive stochastic approach to the determination of purchasing power parities. However, in real life, national price statisticians will not be able to collect prices for every item and every transaction for each item in the applicable transactions domain of definition for the price comparison project under consideration. Thus only a sample of items will be chosen in each country and only a sample of transaction prices for each of the chosen items will be

[53] Thus these descriptive models have the disadvantage that we cannot develop confidence intervals for the country log PPP's but they have the advantage that we do not have to make particular distributional assumptions about the error terms as is necessary when using the traditional regression approach. The major problem with the traditional approach is that it gives rise to endless discussions about what is the "right" assumption to make about the distribution of the error terms. Our present discrete approach largely avoids these somewhat fruitless discussions but of course, we still have to make the orthogonality assumptions (77).

[54] If a linear combination of the columns of X equal a vector of ones, we can convert T into an R^2 in the usual way; i.e., define $T_1 \equiv [y - 1_J y^*]^T \,\Omega\, [y - 1_J y^*]$ where 1_J is a vector of ones of dimension J and y^* is the mean of the y_j and define $R^2 \equiv 1 - (T/T_1)$. If $R^2 = 1$ (so that T = 0), then prices are exactly proportional across countries whereas if $R^2 = 0$, then prices are essentially randomly distributed across countries and products.

collected. If we have absolutely no information on the expenditures that can be associated with each sampled price quotation, then it would seem that there is nothing we can do except apply the unweighted CPD method explained in section 3. However, if the national price collectors can form some judgments about the relative expenditure importance of the various item prices that are collected, then more can be done.

At present, OECD and Eurostat price statisticians decide whether a price quote is "representative" or "unrepresentative".[55] Representative prices are given a heavier weight in the OECD and Eurostat index number computations than unrepresentative prices.

It is possible to apply the model outlined in section 5 in this sampling framework where we have representative prices and unrepresentative prices: all that is required is that the national price statistician form some rough judgment as to the relative size of expenditures that can be associated with the two types of price quote.[56] If the price p_{cnk} in country c is regarded as unrepresentative, then it is given the sampling weight $w_{cnk} = 1$. If the price p_{cnk} in country c is regarded as representative, then it is given the sampling weight w_{cnk} which is some multiple of 1, such as 10 if it is thought that 10 times the value is associated with a representative price quote versus an unrepresentative one. Thus instead of defining the transactions share s_{cnk} corresponding to the price p_{cnk} by (40), we use the approximate sampling weights w_{cnk} to form estimates for the s_{cnk} as follows:

$$(83)\ s_{cnk} \equiv w_{cnk} / \square_{i=1}^{N} \square_{j=1}^{K(c,n)} w_{cij}\ ; \qquad n = 1,\ldots,N\ ;\ k = 1,\ldots,K(c,n).$$

Now all of the algebra developed in developed in section 5 can be applied to this model where we use the approximate shares defined by (83) in place of the true expenditure shares.

Since some countries c participating in the ICP may not be able to submit to the central processing agency the *individual* log prices for product n and outlet k, y_{cnk}, for confidentiality reasons, it will be useful to look carefully at the X^TX matrix defined by (47) and the X^Ty vector defined by (48) and see what aggregated information from the countries is sufficient to send to the center for processing into the weighted least squares estimators for the log PPP's, the $\square_c$. It is evident that each country c needs to report the following data to the center: $\square_{k=1}^{K(c,1)} s_{c1k}\, y_{c1k}$, $\square_{k=1}^{K(c,2)} s_{c2k}\, y_{c2k}$, …, $\square_{k=1}^{K(c,N)} s_{cNk}\, y_{cNk}$

[55] Peter Hill (2004) explains the methods used by OECD and Eurostat price statisticians making price comparisons at the elementary level and calls their methods EKS methods, Variant 1 is essentially due to Eltetö and Köves (1964) and Szulc (1964). Variant 2 was developed by a group of Eurostat experts and variant 3 is due to Sergeev (2003). Sergeev (2002; 10) notes a problem with this method: some countries classify virtually all items as representative whereas other countries do not. However, it should be noted that these EKS methods are based on the countries reporting to the center only average prices over all outlets for the product under consideration and then that average price is graded as being "representative" or "not representative". The methodology being developed here grades the *individual* outlet price quotes as being "representative" or "not representative". If there is only one price quote per product (as in Hill's (2004) examples), then this methodological distinction vanishes.

[56] The N items in the model are now interpreted as only a sample of the items rather than the universe of items.

(the sum of the national share weighted outlet log prices for each item n in country c where the s_{cnk} are defined by (83)[57]) and $\square_{k=1}^{K(c,1)} s_{c1k}$, $\square_{k=1}^{K(c,2)} s_{c2k}$, ..., $\square_{k=1}^{K(c,N)} s_{cNk}$ (the sum of the national outlet shares for each item n in country c where the s_{cnk} are defined by (83)) Thus as long as each product cell has more than one outlet price collected for it, individual price data need not be forwarded to the central processing agency in order to calculate the log PPP's, the $\square_c$'s.

In order to find confidence intervals for the log PPP's, it is also necessary to calculate the sum of squared residuals for the weighted regression, $\square_{c=1}^{C} \square_{n=1}^{N} \square_{k=1}^{K(c,n)} e_{cnk}^2$; see (49) and (50) above. Using definitions (49), we have for each country c and product n, we have:

$$(84)\ \square_{k=1}^{K(c,n)} e_{cnk}^2 = \square_{k=1}^{K(c,n)} s_{cnk} [y_{cnk} \square \square_c^* \square \square_n^*]^2$$
$$= \square_{k=1}^{K(c,n)} s_{cnk}\, y_{cnk}^2 \square 2\, [\square_{k=1}^{K(c,n)} s_{cnk}\, y_{cnk}][\square_c^* + \square_n^*] + [\square_{k=1}^{K(c,n)} s_{cnk}][\square_c^* + \square_n^*]^2.$$

Thus the center can calculate the sum of squared residuals if the individual countries send the following information to the center (in addition to the information already noted above): $\square_{k=1}^{K(c,1)} s_{c1k}\, y_{c1k}^2$, $\square_{k=1}^{K(c,2)} s_{c2k}\, y_{c2k}^2$, ..., $\square_{k=1}^{K(c,N)} s_{cNk}\, y_{cNk}^2$ (the sum of the national share weighted squares of the outlet log prices for each item n in country c).

Since the center will not have the individual log prices y_{cnk} under the above conditions, it will be necessary for the individual countries to eliminate outliers from their country data before submitting the above information to the center.

10. An Example due to Peter Hill

In order to illustrate the methods suggested in sections 2 and 9 , we consider a data set that was used by Hill (2004) where he postulated prices for 10 items and 4 countries. The prices may be found in Table 1 below. Representative price quotes are marked with an "r" in the Table.

Table 1: Price Data for 10 Items for 4 Countries

Item	Country 1	Country 2	Country 3	Country 4
1	2 r	100	10 r	25 r
2	5 r	250	12 r	60
3	6 r	270	15 r	22 r
4	8 r	320	70	250
5	8 r	280	100	120 r
6	7	210 r	60	120
7	16	400 r	50 r	140 r
8	6	120 r	12 r	100

[57] Note that if country c happens to collect the same number of outlet prices in each of the N item cells, say K prices, then for each product n, $\square_{k=1}^{K(c,n)} s_{cnk}\ y_{cnk} = \square_{k=1}^{K} (1/NK)\ y_{cnk} = \square_{k=1}^{K} (1/NK) \ln p_{cnk} = (1/N)\square_{k=1}^{K} (1/K) \ln p_{cnk}$, which is 1/N times the logarithm of the geometric mean of the K outlet prices for product n.

9	2	30 r	20	10 r
10	10	100 r	50	100

Since there are an equal number of price quotes for each item (namely one), we can apply the *unweighted model* of section 2 and obtain the following (transitive) purchasing power parities:[58]

(85) $a_1 = 1, \quad a_2 = 28.507, \quad a_3 = 4.948, \quad a_4 = 11.241.$

If we use the model suggested in the previous section and assign the weight 1 for an unrepresentative price quote and the weight w for a representative price quote (so that expenditures associated with the representative quotes are thought to be w times as big as the expenditures on unrepresentative quotes), then we obtain the following Table of purchasing power parities for various values of w, the weighting factor for representative price quotes:

Table 2: Weighted CPD Country Parities for Various Values for Representativity

Weight w	Country 1	Country 2	Country 3	Country 4
w = 1	1	28.507	4.948	11.241
w = 2	1	27.075	4.468	10.551
w = 4	1	26.205	4.060	9.925
w = 10	1	25.996	3.668	9.273

Unfortunately, Table 2 shows that the choice of the weighting factor for representative price quotes versus unrepresentative quotes makes a substantial difference. Since traditional index number theory strongly suggests that weighted price indexes are preferable to unweighted ones, it seems preferable to use even rough weights in the ICP. But the practical question then is: *exactly how do we choose w*? Obviously, if we have expenditure weights to go along with the item price quotes, this question can be answered in an objective manner. But in the more usual case where approximate weights are not available, then it appears that the national price statisticians collecting the price data will have to use their judgment and give the central processing agency their best estimate for the magnitude of the weighting factor w.[59]

11. Bilateral Linking Strategies

It is possible to use the model presented in section 3 or the approximate version of it presented in section 9 to develop a strategy for linking the countries in a bilateral fashion.

[58] These are the same parities as were obtained by Hill (2004). Hill also shows that these parities are also equal to the Variant 1 EKS parities but they differ from the Variant 2 and 3 EKS parities which were 1, 27.27, 4.03, 9.60 for both variants in this particular case.

[59] My guess is, that under "normal" conditions, w = 3 or 10 will be "better" than w = 2, since in the time series index number context, it is known that best selling items will be purchased much more frequently than slowly selling items.

The basic idea is this: countries which are most similar in their price structures (i.e., their prices are closest to being proportional across items) should be linked first. This basic idea has been successfully exploited by Robert Hill at higher levels of aggregation,[60] where complete price and expenditure data are available, but it is not obvious that the same methodology can be applied at the elementary level, where complete data on expenditures associated with each price quote are missing. In order to make the method operational, all that is required is a suitable measure of relative price dissimilarity. In the present context, we suggest that the estimated variance of the regression, □^{*2} defined by (50) for the case where C = 2, as a measure of the degree of dissimilarity in the price structures of the 2 countries in the comparison. Alternatively, the dissimilarity measure T defined by (82) could be used.[61]

To illustrate this suggested method, consider the empirical example of Peter Hill discussed in the previous section. If we do not use weights, then the Robert Hill methodological approach applied to this example gives the same purchasing power parities as those given by (85) in the previous section, since all the bilateral parities are transitive in this particular example. However, when we use approximate sampling weights as was suggested in the previous section, the situation changes.

If we use the approximate weight 1 for unrepresentative price quotes and 4 for representative price quotes, we find that the bilateral regression variance estimates for □^{*2} are as follows: □^{*2} = 0.0094286 for the countries 1 and 2 regression with $PPP_{2/1}$ = 28.50666 ; □^{*2} = 0.016692 for the countries 1 and 3 regression with $PPP_{3/1}$ = 4.083539; □^{*2} = 0.018571 for the countries 1 and 4 regression with $PPP_{4/1}$ = 10.16165; □^{*2} = 0.026992 for the countries 2 and 3 regression with $PPP_{3/2}$ = 0.1513575; □^{*2} = 0.017054 for the countries 2 and 4 regression with $PPP_{4/2}$ = 0.3785348; □^{*2} = 0.020493 for the countries 3 and 4 regression with $PPP_{4/3}$ = 2.226068. The smallest variance bilateral regressions are the 1 and 2 regression, the 1 and 3 regression and the 2 and 4 regression and these three regressions lead to the following purchasing power parities:

(86) $a_1 = 1, \quad a_2 = 28.507, \quad a_3 = 4.084, \quad a_4 = 10.791.$

These parities turn out to be different than the multilateral parities in the third line of Table 2, which also used the weighting factor, w = 4.[62]

While the Robert Hill linking strategy *could* be applied at the basic heading level, it probably should *not* be applied at this level of aggregation due to the problem of *sparseness*; i.e., the Hill bilateral linking strategy is best suited to linking at higher levels of aggregation where price parities and expenditure weights are available for each basic heading category of expenditure. To illustrate this point, consider the following data set involving 3 countries and three products with one price quote per product except that

[60] See Robert Hill (1995) (1999a) (1999b) (2001) (2004).

[61] These two measures differ only by a constant and hence will give the same answer in the bilateral case.

[62] These parities were 1, 26.205, 4.060, 9.925 for countries 1,2,3 and 4 respectively.

product 1 is not available in country 3, product 2 is not available in country 1 and product 3 is not available in country 2:

Table 3: Prices for a Sparse Data Example

		Prices	
Product	Country 1	Country 2	Country 3
1	1	2	_
2	_	1	2
3	1	_	3

Thus for each pair of countries, we have a direct bilateral comparison of prices. Thus using a direct comparison of prices between countries 1 and 3, the PPP of country 3 relative to country 1 will be 3. However, for each pair of countries, we can also make a comparison of prices by traveling through the remaining country. Thus using the indirect comparison between the prices of countries 1 and 3, traveling through country 2, we find the indirect PPP of country 3 relative to country 1 will be 2 times 2 or 4. Hence using some sort of average of the direct and indirect parities between 3 and 1 should give us a PPP somewhere between 3 and 4. It can be seen that the Robert Hill bilateral spatial chaining approach breaks down under these conditions. Hence, it seems preferable to use either the *unweighted multilateral approach* outlined in section 3 or if weights are available, then use the *weighted multilateral approach* outlined in section 9 in circumstances where the matrix of price quotes is not complete.[63]

12. The Extended CPD Method or the CPDR Method

Cuthbert and Cuthbert (1988) introduced an interesting generalization of the Country Product Dummy method that can be used if information on representativity of the prices is collected by the countries in the comparison project along with the prices themselves. Hill (2004) termed the method the CPRD method and he justified the method as follows:

> "Thus, the price of a given product may be relatively high or low in a country depending on whether or not it is representative. These relationships are not consistent with the basic assumption underlying traditional CPD methods that the pattern of relative prices is the same in all countries. The CPD model should therefore be modified to take account of representativity when information about representativity is available. The influence of representativity on price is explicitly taken into account in the EKS 2 and 3 methods, but not in the CPD." Peter Hill (2004; 24).

The CPRD method generalizes the unweighted model (22) above as follows. Define $y_{cnkr} = \ln p_{cnkr}$ where p_{cnkr} is the logarithm of the kth outlet price collected in country c for product n and r is an index that denotes whether the collected price is representative (in which case r = 1) or unrepresentative (in which case r = 2). The basic (unweighted) statistical model that is assumed is the following one:

(87) $y_{cnkr} = \square_c + \square_n + \square_r + \square_{cnkr}$; $c = 1,\ldots,C;\ n = 1,\ldots,N;\ k = 1,\ldots,K(c,n)\ ;\ r = 1,2$

[63] Using the unweighted CPD approach outlined in section 3 gives rise to the following parities: 1, 1.817, 3.302.

where the $\square_c$ are the log country PPP's, the $\square_n$ are the log product price effects and the $\square_r$ are the two log representativity effects and the $\square_{cnkr}$ are independently distributed random variables with mean zero and constant variances. In order to identify the parameters, we impose the following normalizations:

(88) $\square_1 = 0$; $\square_1 = 0$.

Thus the present model is much the same as the model presented in section 3 except that we have an analysis of variance model that has 3 classifications instead of 2.

A potential problem with the new model can be explained as follows. Let $\square_c^*$, $\square_n^*$ and $\square_r^*$ denote the least squares estimators for the parameters in the linear regression model defined by (87) and (88). Define the least squares sample residual errors e_{cnkr} as follows:

(89) $e_{cnkr} \equiv y_{cnkr} \square [\square_c^* + \square_n^* + \square_r^*]$; $c = 1,\ldots,C; n = 1,\ldots,N; k = 1,\ldots,K(c,n) ; r = 1,2.$

Since each column in the X matrix that corresponds to the linear regression model defined by (87) and (88) is orthogonal to the vector of errors $e \equiv [e_1,\ldots,e_J]^T$ which is obtained by stacking the errors e_{cnkr} into a column vector, it can be seen that the sum of the errors that correspond to nonrepresentative observations where r = 2 is zero; i.e., we have:

(90) $\square_c\square_n\square_k\, e_{cnk2} = 0$.

This is an extra constraint that the errors in this CPRD type model satisfy compared to the plain vanilla CPD model defined earlier by (22). In most cases, this will not cause any great problems with the estimates of the PPP's; i.e., in most cases, the two models will give more or less the same PPP's. However, this is not always the case. Consider the following two country example, involving 3 products with one price quote per product:

Table 4: A Two Country Example Illustrating the CPRD Method

	Prices	
Product	Country 1	Country 2
1	1 r	2 r
2	2 r	4 r
3	3 r	7

Thus all prices are *representative* (denoted by r in the Table) in each country except for the price of item 3 in country 2. For the first two items, the price of the item for country 2 divided by the corresponding price for country 1 is 2 and for the third item, the price relative is 7/3, which is 2 1/7. Thus it seems clear that the PPP for country 2 relative to country 1 should be a number that is somewhat greater than 2. However, if we calculate the CPRD $PPP_{2/1}$, we find that $a_2 \equiv \exp(\square_2) = 2$; i.e., the CPRD $PPP_{2/1}$ for 2 relative to 1 turns out to be *exactly* 2 instead of a number greater than 2. The problem is that the number of nonrepresentative price quotes across the 2 countries is *sparse*; in fact, there is only one nonrepresentative price quote and the result (90) implies that the corresponding error will be zero. The extra parameter $\square_2$ is used to set this error equal to zero and this causes some bias in the PPP.[64] However, this example is rather extreme and is caused by the extreme sparseness of nonrepresentative price quotes. As long as nonrepresentative price quotes are not too sparse, the CPRD model can be regarded as a very useful extension of the basic unweighted CPD model.

We agree with Hill that the *unweighted* CPD parities will usually be biased. However, it is not clear to us the *weighted* CPD parities defined in section 9 above will necessarily be biased compared to Hill's CPRD parities. These weighted CPD parities are listed below in Table 5 for the data in Table 4 for various values of the weighting factor w, which gives the weight of a representative quote relative to an unrepresentative price quote.

Table 5: Weighted CPD Country Parities for Various Values for Representativity

Weight w	Country 1	Country 2
w = 1	1	2.105
w = 2	1	2.080
w = 4	1	2.056
w = 10	1	2.030

Thus as the weighting factor w increases, the relative weight of product 3 in country 2 (the unrepresentative price quote) in Table 4 decreases and the $PPP_{2/1}$ approaches 2 as could be expected.[65]

Our tentative conclusion is that it is not necessary to have the extra representativity parameter in the *weighted* CPD model but if the data are not too sparse, it is likely that the unweighted CPRD model will give better results than the 3 EKS models (since they do not utilize all of the data and hence they cannot be statistically efficient). It is also likely that the unweighted CPRD model will give better results than the unweighted CPD model. However, as soon as weights are available, we recommend the weighted CPD model defined in section 9 over the weighted CPRD model, since the weighted CPD model can be regarded as an approximation to the theoretically sound target index defined in section 5.

13. The Two Stage Linking Procedure for the Current Round of the ICP

A complication that we have not dealt with up to now is that the current ICP project is proceeding in two stages. The world is divided up into 6 regions r with C(r) countries in each region r for r = 1,…,6. Within each of the 6 regions, PPP's at the basic heading level will be constructed more or less independently for each region. In the second stage,

[64] By increasing arbitrarily the price of product 3 in country 2, we can increase the bias arbitrarily.

[65] Hill (2004) also considered a weighted version of his CPRD model along the lines of the model presented in section 9. However, for the data in Table 4, Hill's weighted $PPP_{2/1}$ remains equal to 2 for all possible positive weights w, which is clearly biased for the present model.

the regions will be linked. In this section, we consider some of the complications involved in modeling this situation.

We first consider a generalization of the unweighted CPD model presented in section 3. We need to generalize this above model to allow for a reorganization of the list of C countries into 6 regions and C(r) countries in each region. With these changes, the basic model becomes:

(91) $p_{rcnk} \approx a_r b_{rc} c_n$; $r = 1,\ldots,6$; $c = 1,\ldots.,C(r)$; $n = 1,\ldots,N$; $k = K(r,c,n)$[66] ;
(92) $a_1 = 1$;
(93) $b_{r1} = 1$; $r = 1,\ldots,6$.

The normalization (92) means that we have to choose a numeraire region. The normalizations (93) mean that within each region, we need to choose a numeraire country in order to identify all of the parameters uniquely. Thus the parameters a_r and b_{rc} replace our initial model parameters a_c. Note that the total number of parameters remains unchanged when we group all of the countries in the comparison into regions and countries within the regions.

Taking logarithms of both sides of (91) and then adding error terms ε_{rcnk} (with means 0) leads to the following regression model:

(94) $\ln p_{rcnk} = \ln a_r + \ln b_{rc} + \ln c_n + \varepsilon_{rcnk}$; $r = 1,\ldots,6$; $c = 1,\ldots.,C(r)$; $n = 1,\ldots,N$; $k = K(r,c,n)$;
$= \alpha_r + \beta_{rc} + \gamma_n + \varepsilon_{rcnk}$

where we impose the following normalizations on the parameters in order to uniquely identify them:

(95) $\alpha_1 = 0$;
(96) $\beta_{r1} = 0$; $r = 1,\ldots,6$

where $\alpha_r \equiv \ln a_r$, $\beta_{rc} \equiv \ln b_{rc}$, $\gamma_n \equiv \ln c_n$.

If all of the data collected for each regional comparison could be pooled and if there are product overlaps between the regions, then there will be 155 regressions of the form (94) to run, one for each basic heading category. In the above model, the interregional log parities (the α_r) are estimated along with the within region country log parities (the β_{rc}) and the product log price premiums (the γ_n). Call this the *first approach* to estimating the

[66] We have noted already that there are 6 regions in the ICP model and 147 countries in all. Within each region, there will be about 20 separate items that will be priced in each basic heading category of commodities across all countries in the ICP comparison project so if there were no overlap in the items selected across regions, then N would be equal to 120 (equal to 6 times 20). But of course, we require some overlap of items across regions so that we can identify the parameters a_r. For each country and for each item that is priced in that country, there will typically be multiple price quotes collected for each item, say 5, so that K(r,c,n) will typically equal 5.

regional parities for each basic heading.[67] It uses all of the available information in making comparisons between all of the countries.

However, the above one big regression approach (for each basic heading) is *not consistent* with approaches that use only the regional data to determine the within region parities, the $\square_{rc}$ parameters, holding r fixed. But a principle of the current ICP methodology is that regions should be allowed to determine their own parities, independently of other regions.[68] However, the regression model (94) can be modified to deal with this problem. If the regional log parities $\square_{rc}$ are known, then the term $\square_{rc}$ (which is equal to ln b_{rc}) can be subtracted from both sides of (94), leading to the following regression model:

(97) $\ln p_{rcnk}\square \ln b_{rc} = \ln a_r+\ln c_n + \square_{cnk}$; $r = 1,\ldots,6$; $c = 1,\ldots.,C(r)$; $n = 1,\ldots,N$; $k = K(r,c,n)$

or

(98) $\ln [p_{rcnk}/b_{rc}] = \square_r + \square_n + \square_{rcnk}$;

where the normalization (95) still holds. Thus if the within region parities are known, then prices in each region p_{rcnk} can be divided by the appropriate regional parity for that country in that region $\square_{rc}$, and these regionally adjusted prices can be used as inputs into the usual CPD model that has now only the regional log parities $\square_r$ and the commodity adjustment factors $\square_n$ as unknown parameters to be estimated.[69] Call the model defined by (95) and (98) the *second approach* to estimating the regional parities for each basic heading. This second approach respects the within region parities that have been constructed by the regional price administrators. It is possible that this second approach will be used in ICP 2005.

Of course, weighted versions of the two approaches can also be implemented.

Within each basic heading category, the regional coordinators have developed product lists, consisting of approximately 2 to 30 separate products. Unfortunately, it appears that there is *virtually no overlap* between the 6 regional product lists.[70] Thus the interregional CPD model at each basic heading level defined by (95) and (98) cannot be implemented using just the data that the regions collect to do the within region comparisons!

McCarthy (2004) outlines two broad strategies that could be used to deal with this lack of regional matching problem:

[67] This approach is broadly consistent with the approach favored by Yuri Dikhanov, who advocated using all the available information in making the ring comparisons.

[68] See Hill (2004).

[69] Thus we have saved 144 degrees of freedom in this model compared to our previous example where we had 625 observations and 249 parameters to estimate.

[70] Fred Vogel pointed this out to the author in a personal communication.

- At least one country in each region would be required to collect prices for the product list of at least one other region; the countries who agree to collect these extra price quotes are called *bridge countries* by McCarthy;
- A group of countries (called *core countries*), with at least one country in each of the 6 regions, would be chosen and they would work out a separate joint interregional product list for each basic heading category, which each core country would then price out.

For either strategy, the unweighted CPD model defined by (95) and (98) (or a weighted counterpart using the methodology outlined in section 9 above)[71] could be used to estimate the interregional PPP's for each basic heading category. Thus there would be 155 separate interregional CPD models that would have to be estimated by the center.

We concur with McCarthy's (2004) judgment that it would be far more expensive and time consuming to follow the core country strategy where separate international product lists would have to be drawn up and then priced by the core countries. It is not clear that there is enough time available to accomplish this design task either. Hence, we recommend that the bridge country strategy be followed in order to link the regions.

14. Linking at the Basic Heading Level: A Three Region Example using the Second Approach

We consider the problem of comparing the prices in three regions at one of the 155 basic heading levels. We assume that there are three ring countries in region A, countries A1, A2 and A3, where A1 is the numeraire ring country for region A, there are two ring countries in region B, B1 and B2, where B1 is the numeraire ring country for region B, and there are two ring countries in region C, C1 and C2, where C1 is the numeraire ring country for region C. For the particular basic heading category under consideration, there are 6 items on the ring list and the ring countries collect the following item prices listed in Table 6 in their own national currencies. Representative prices are denoted with an r.[72]

[71] Our preference is for the weighted CPD model over the unweighted version (which is unlikely to converge to the "truth" no matter how much information is collected).

[72] As explained earlier, price collectors in the regions are instructed to determine whether a particular product price that they have collected for the ICP project is *representative* or *nonrepresentative*. A representative price corresponds to a product that has a higher volume of sales in the country as compared to a nonrepresentative product. Since representative prices are more highly demanded by purchasers of the product, it is likely that prices for nonrepresentative prices are relatively higher in the local market as compared to representative prices for similar products.

Table 6: Item Prices in Domestic Currencies for the Ring Countries at a Specific Basic Heading Level

Item	Region A			Region B		Region C	
	Country A1	Country A2	Country A3	Country B1	Country B2	Country C1	Country C2
1	1 r	2 r	4 r	2 r	6 r	3 r	30 r
2	2 r	5	__	6	10 r	5 r	__
3	6 r	10 r	__	10 r	__	22	160 r
4	4 r	__	18	__	30	10 r	150
5	5 r	__	20 r	__	__	__	180
6	12 r	30	40 r	24 r	72 r	36 r	360 r

There are 7 countries and 6 items in the above table so the maximum number of price quotes is 42 but there are missing observations so the table has only 32 item prices. Of these item prices, 24 are representative and 8 are not representative. Note that all of the product prices collected by country A1 are representative whereas all other countries have at least one nonrepresentative price.

We now suppose that each region has determined an appropriate PPP between the ring countries in its region for this particular basic heading category of transactions.[73] These regional PPP's for this basic heading category are listed in Table 7 below.

Table 7: Basic Heading PPP's for Each Region for the Ring Countries

Country	A1	A2	A3	B1	B2	C1	C2
Basic Heading PPP	1	2	4	1	3	1	10

Thus in region A, it has been determined by the region A experts that 2 units of country A2's currency buys the same amount of final demand in the given basic heading category as 1 unit of country A1's currency and 4 units of country A3's currency buys the same amount of final demand in the given basic heading category as 1 unit of country A1's currency. Similarly, in region B, it has been determined by the region B experts that 3 units of country B2's currency buys the same amount of final demand in the given basic heading category as 1 unit of country B1's currency and in region C, it has been determined by the region C experts that 10 units of country C2's currency buys the same amount of final demand in the given basic heading category as 1 unit of country C1's currency.

In order to preserve these regional parities for the ring countries when linking the regions, it is necessary to divide the ring country prices by these within region parities and this essentially converts the country prices within a region into common regional prices. Thus looking at Table 1, it is necessary to divide the item prices listed there by the factor 2 for country A2 and by the factor 4 for country A3. Similarly, it is necessary to divide the country B2 item prices by the factor 3 in order to convert these prices into country B1

[73] Thus we are now considering an example of the *second approach* that was described in the previous section.

equivalents. Finally, it is necessary to divide the country C2 item prices by the factor 10 in order to convert these prices into country C1 equivalents. The resulting regional prices are listed in Table 8 below.

Table 8: Item Prices in Numeraire Country Currencies for the Ring Countries at a Specific Basic Heading Level

Item	Region A			Region B		Region C	
	Country A1	Country A2	Country A3	Country B1	Country B2	Country C1	Country C2
1	1 r	1 r	1 r	2 r	2 r	3 r	3 r
2	2 r	2.5	__	6	3 1/3 r	5 r	__
3	6 r	5 r	__	10 r	__	22	16 r
4	4 r	__	4.5	__	10	10 r	15
5	5 r	__	5 r	__	__	__	18
6	12 r	15	10 r	24 r	24 r	36 r	36 r

The country prices listed in Table 8 can be regarded as "outlet" prices within each region and hence the data in Table 8 can be used in the usual Country Product Dummy regressions (or other methods) in order to determine parities between the 3 regions for the particular basic heading category under consideration. Thus the 14 prices listed in the Country A1-A3 columns can be regarded as item prices in a common region A currency, the 8 prices in the Country B1-B2 columns can be regarded as item prices in a common region B currency and the 10 prices in the Country C1-C2 columns can be regarded as item prices in a common region C currency.

Applying the CPD method[74] and the CPRD method [75] to the data in Table 8 (treating the data as pertaining to 3 countries, which are regions in this case) leads to the following PPP's between the 3 regions:

Table 9: Basic Heading CPD and CPRD PPP's between the Three Regions

Method	Region A	Region B	Region C
CPD	1	2.047	3.006
CPDR	1	2.014	2.937

Thus the use of the representativity dummy variable causes the region B and C parities to fall with respect to region A compared to the ordinary unweighted CPD regression (about 3 and 7 percentage points respectively). An explanation for the change in the parities as

[74] The basic model that we use is (98) with the normalization (95) imposed. Thus we want to estimate the two parameters $\square_2$ and $\square_3$ (the log of the regional PPP for region 2 relative to region 1 and the log of the regional PPP for region 3 relative to region 1 respectively) and the 6 product premium parameters, $\square$,...,$\square_6$. In the Tables below, the exponentials of $\square_2$ and $\square_3$ are reported and we set the region B and C PPP's reported in Tables 9 and 10 equal to the exponentials of $\square_2$ and $\square_3$ respectively.

[75] See Hill (2004) for an exposition of this method. In brief, we add another dummy variable to the regression model defined by (98) and the normalization (95). This dummy variable takes on the value 0 if the price quote is representative and takes on the value 1 if the corresponding price quote is not representative.

we shift from CPD to CPRD can be made as follows. Note that the ratio of nonrepresentative observations relative to representative observations in the three regions are 3/11 = .27 for region A, 2/6 = .33 for region B and 3/7 = .43 for region C. Since nonrepresentative prices can be expected to be relatively higher than representative prices, we can expect the coefficient on the representativity variable to be positive (since the dummy variable is 0 for representative prices and 1 for nonrepresentative prices). Thus including the representativity dummy will tend to *lower* the regional PPP's for regions that have *higher* ratios of nonrepresentative prices relative to nonrepresentative prices[76] and this is what happens in our example.

We also calculate weighted versions of the CPD parities using the methodology suggested in sections 5 above. This weighted method ensures that each country (or region in this case) is given equal importance in the weighted least squares optimization problem that is the basis for the method.[77] The key parameter in this method is to decide on what relative expenditure weight to give representative versus nonrepresentative observations. Thus in Table 10 below, if w = 2, then representative price quotes get twice the weight in the weighted CPD least squares minimization problem that nonrepresentative price quotes get. This means that we believe it is likely that region expenditures that are associated with representative price quotes are twice as large as the expenditures that are associated with nonrepresentative price quotes. Hill (2004) also suggested that the same methodology could be applied to the CPRD method, leading to a weighted version of the CPRD method.[78] In Table 10 below, we table the weighted CPD parities for the regions for weighting factor w equal to 1, 2, 3 and 10 respectively and weighted CPRD parities for the regions for weighting factor w equal to 1, 2, 3 and 10 respectively.[79]

Table 10: Basic Heading Weighted CPD and CPRD PPP's between the Three Regions

Method	Region A	Region B	Region C

[76] The positive coefficient on the representativity dummy will tend to lead to lower coefficients for the country dummy variables that have the highest proportions of nonrepresentative product prices.

[77] Thus the unweighted CPD method in our present example gives more weight to region A's prices (14 observations) than regions C's prices (10 observations) and region B gets the smallest weight (8 observations). In each version of the weighted CPD method, each country's item prices get an equal weight in the least squares minimization problem.

[78] If the proportion of nonrepresentative to representative price quotes differs across the regions (as it does in this case), then I agree with Hill (2004) that it is sensible to include the representativity dummy variable in the CPD regression, leading to a CPRD regression. However, if we use the weighted CPD method, then choosing a w greater than 1 does much the same job as including the representativity dummy in the unweighted CPRD regression and hence including the representativity dummy in the weighted CPD is not necessary. Thus in our particular example, the regional parities for the unweighted CPRD method are 1, 2.014 and 2.937 (see Table 9 above), which are fairly close to the weighted CPD parities with w = 3 which are 1, 2.004 and 2.956 (see Table 10 below).

[79] These weighted CPD and CPRD parities for w = 1 are not equal to the unweighted CPD and CPRD parities in Table 9 because the number of price quotes in each region is not equal, and hence the unweighted CPD and CPRD parities give undue influence to the price quotes of region A. However, since the number of price quotes in each region is not all that different (14 for A, 8 for B and 10 for C), the differences between CPD and weighted CPD with w = 1 are not that great.

CPD w = 1	1	2.053	3.006
CPDR w = 1	1	2.009	2.933
CPD w = 2	1	2.022	2.971
CPDR w = 2	1	1.977	2.915
CPD w = 3	1	2.004	2.956
CPDR w = 3	1	1.965	2.911
CPD w = 10	1	1.966	2.928
CPDR w = 10	1	1.947	2.910

Comparing the weighted CPD parities for w = 1 in Table 10 with the unweighted CPD parities in Table 9, we see that the weighted parity for region B relative to A has increased slightly from 2.047 to 2.053, which is a negligible change, and the weighted and unweighted parities for region C relative to A have remained unchanged at 3.006. Comparing the CPD and CPRD parities for the same level of weighting w, it can be seen that the weighted CPRD parities are consistently below the corresponding weighted CPD parities for regions B and C. The reason for this difference is the same as explained above: regions B and C have the highest ratios of nonrepresentative price quotes compared to region A and this leads to the discrepancies. It can be seen that the weighted CPD and CPRD parities for regions B and C decrease as the weighting factor w increases. This is explained by the fact that region A has the highest percentage of representative price quotes and regions B and C the highest percentages of nonrepresentative price quotes, which tend to be relatively high. Thus as w increases, the influence of these relatively high nonrepresentative price quotes diminishes, and the parities for B and C are lowered as a result. Finally, note that the difference between the weighted CPD and weighted CPRD parities for B and C relative to A decreases as w increases. This is also understandable, since as w increases, the influence of the nonrepresentative prices diminishes and hence the difference between the weighted CPD and weighted CPDR regressions becomes less.[80]

All of the regressions give much the same answer. My own preference is for the weighted CPD regression with w = 3 but whether this is a good choice or not depends on the target index that is chosen and somewhat subjective judgments about the relative magnitudes of expenditures that are associated with representative versus unrepresentative price quotes.

The above computations illustrate the *second approach* to linking the regions that was discussed in the previous section; i.e., the within region PPP's for the basic heading are respected and treated as exogenous variables in the regression that links the regions. In the following section, we use the same data set to illustrate *first approach* to linking the regions, where the within region parities are not imposed but are estimated along with the interregional parities.

15. Linking at the Basic Heading Level: A Three Region Example using the First Approach

[80] In the limit, as w became very large, the objective functions for the two weighted least squares problems would approach each other.

We illustrate the first approach for linking the regions explained in section 13 using the data listed in Table 8. In this approach, the within region PPP's are estimated along with the interregional PPP's.

The basic regression model is defined by equations (94)-(96) above. When we specialize this model to cover the case corresponding to the data in Table 8, we find that there are three interregional parities in the resulting model that need to be estimated, a_1, a_2 and a_3, which we relabel as A, B and C. Within region 1, there are three countries and using the notation used in equations (94), these within region country parities are labeled as b_{11}, b_{12} and b_{13}, which we now relabel as A1, A2 and A3. Within region 2, there are 2 countries and in (94), these within region country parities are labeled as b_{21} and b_{22}, which we now relabel as B1 and B2. Within region 3, there are 2 countries and in (94), these within region country parities are labeled as b_{31} and b_{32}, which we now relabel as C1 and C2. The identifying restrictions (95) and (96) in the general model boil down to the following restrictions for the particular model represented by the data in Table 8:

(97) A = 1 ; A1 = 1 ; B1 = 1 ; C1 = 1.

Applying the CPD method and the CPRD method to the data in Table 8 (treating the data as pertaining to 3 countries, which are regions in this case) leads to the following PPP's between region and within region PPP's:

Table 11: Basic Heading CPD and CPRD PPP's between the Three Regions

Method	PPPA	PPPB	PPPC	PPPA1	PPPA2	PPPA3	PPPB1	PPPB2	PPPC1	PPPC2
CPD	1	2.184	2.934	1	2.206	3.857	1	2.742	1	10.876
CPDR	1	1.924	2.720	1	1.785	3.548	1	2.861	1	10.161

Thus the use of the representativity dummy variable causes the region B and C parities to fall with respect to region A compared to the ordinary unweighted CPD regression (about 16 and 21 percentage points respectively). In the previous section, when the CPD and CPRD methods were compared, we found a similar result in that the addition of the representativity dummy variable caused the parities to fall in regions with a high proportion of nonrepresentative quotes compared to regions with a high proportion of representative quotes. However, in the previous section, the addition of the representativity dummy caused the region B and C parities to fall with respect to region A compared to the ordinary unweighted CPD regression by only 3 and 7 percentage points compared to the 16 and 21 percentage points drops reported in the present section. The reason for this change is probably due to the fact that the model estimated in the present section has an additional four parameters that have to be estimated (the within region country parities) compared to the model estimated in the previous section. Since there are only 32 observations in all and a total of 13 parameters to be estimated in this first approach CPRD model, there may not be enough degrees of freedom to accurately estimate all 13 parameters. Thus the addition of the extra representativity dummy variable causes the CPRD regression to "fit the errors" to a certain extent.

In Table 12 below, we table the weighted CPD parities for the regions for weighting factor w equal to 1, 2, 3 and 10 respectively. In constructing the weights, we could use country weights, in which case, each country's weights would sum to one. However, since our focus is on obtaining the regional PPP's, we continue to use the regional weights (that sum to one for each region) rather than use country weights. Thus the weighting matrices used in Table 12 are the same as the weighting matrices used in Table 10.

Table 12: Weighted CPD PPP's between and within the Three Regions

w	PPPA	PPPB	PPPC	PPPA1	PPPA2	PPPA3	PPPB1	PPPB2	PPPC1	PPPC2
1	1	2.193	2.939	1	2.215	3.849	1	2.732	1	10.852
2	1	2.079	2.833	1	2.106	3.763	1	2.812	1	10.955
3	1	2.024	2.795	1	2.045	3.714	1	2.855	1	10.915
10	1	1.912	2.748	1	1.909	3.604	1	2.947	1	10.631

There is very little difference between the weighted CPD parities for the weighting factor w = 1 and the unweighted CPD parities reported in Table 11.[81] As the weighting factor increases, the influence of the unrepresentative price quotes diminishes and the parities of countries and regions that have relative large proportion of unrepresentative price quotes will fall. Thus as w increases, the PPP of region B relative to A falls and the PPP of region C relative to A falls, as in the previous section, since region A has the highest proportion of representative price quotes.

Based on the above example, it appears that either approach 1 or approach 2 could be used to link the regions.

16. Linking the Regions at the Final Stage of Aggregation

The above method for obtaining regional PPP's can be repeated for each basic heading category of transactions. Extending our example to the case of 6 regions, we would obtain the following table of basic heading interregional PPP's:

Table 13: Basic Heading Interregional PPP's

Basic Heading	PPP_A	PPP_B	PPP_C	PPP_D	PPP_E	PPP_F
1	1	$PPP_B(1)$	$PPP_C(1)$	$PPP_D(1)$	$PPP_E(1)$	$PPP_F(1)$
2	1	$PPP_B(2)$	$PPP_C(2)$	$PPP_D(2)$	$PPP_E(2)$	$PPP_F(2)$
…						
155	1	$PPP_B(155)$	$PPP_C(155)$	$PPP_D(155)$	$PPP_E(155)$	$PPP_F(155)$

At this stage, we can use the interregional PPP's in Table 13 along with the within region country PPP's by basic heading category to form a complete world matrix of PPP's by

[81] If the number of observations in each region is exactly equal, there will be no change. In our case, the differences in the number of quotes in each region is not large enough to make a difference.

basic heading and by country. Suppose that there are 147 countries in the ICP project. Then we would have a 155 by 147 matrix of country PPP factors and the country expenditures to go along with them. At this stage, any suitable multilateral method could be used to aggregate up these data into a set of 147 country PPP's. Call this *Approach 1*. However, the problem with this approach is that the multilateral method to be used would not necessarily respect the regional PPP's unless it was restricted in some manner.

Thus we consider *Approach 2*, which will link the regions, while respecting the within region overall PPP's that the regions deem best for their purposes. The first step is to match the 155 times 6 interregional PPP's in Table 13 with the corresponding regional expenditures. If there are C(A) countries in region A and for basic heading category heading n, if the within region PPP's for region A and country c in region A are P_{An}^{c} for n = 1,2,…,155 and c = 1,2,…,C(A)[82] and if the within region expenditure by country c in region A for expenditure category n is E_{An}^{c} in country n's currency unit, then region A's total expenditures for basic heading category n are:

$$(98)\ E_{An} \equiv \sum_{c=1}^{C(A)} E_{An}^{c}/P_{An}^{c}\ ; \qquad n = 1,2,\ldots,155.$$

Now make analogous definitions for the countries in regions B,C,…,F; i.e., define E_{Bn}^{c} and P_{Bn}^{c} ,…, E_{Fn}^{c} and P_{Fn}^{c} in a manner analogous to E_{Bn}^{c} and P_{Bn}^{c} [83]and then define the regional expenditures by basic heading in regional numeraire currencies by the following counterparts to (98):

$$(99)\ E_{Bn} \equiv \sum_{c=1}^{C(B)} E_{Bn}^{c}/P_{Bn}^{c}\ ; \qquad n = 1,2,\ldots,155.$$
$$(100)\ E_{Cn} \equiv \sum_{c=1}^{C(C)} E_{Cn}^{c}/P_{Cn}^{c}\ ; \qquad n = 1,2,\ldots,155.$$
$$(101)\ E_{Dn} \equiv \sum_{c=1}^{C(D)} E_{Dn}^{c}/P_{Dn}^{c}\ ; \qquad n = 1,2,\ldots,155.$$
$$(102)\ E_{En} \equiv \sum_{c=1}^{C(E)} E_{En}^{c}/P_{En}^{c}\ ; \qquad n = 1,2,\ldots,155.$$
$$(103)\ E_{Fn} \equiv \sum_{c=1}^{C(F)} E_{Fn}^{c}/P_{Fn}^{c}\ ; \qquad n = 1,2,\ldots,155.$$

Now we can form a matrix of regional expenditures in the regional numeraire currencies by basic heading category that will match up with the prices in Table 13, and this is Table 14 below.

Table 14: Basic Heading Expenditures by Region in Regional Numeraire Currencies

Basic Heading n	E_{An}	E_{Bn}	E_{Cn}	E_{Dn}	E_{En}	E_{Fn}
1	E_{A1}	E_{B1}	E_{C1}	E_{D1}	E_{E1}	E_{F1}
2	E_{A2}	E_{B2}	E_{C2}	E_{D2}	E_{E2}	E_{F2}
…						
155	E_{A155}	E_{B155}	E_{C155}	E_{D155}	E_{E155}	E_{F155}

[82] In our example, these region A parities for the 3 ring countries for basic heading category 1 say were 1, 2 and 4 but now we have to include the within region parities for the basic heading category under consideration for *all* countries in the region and *not j*ust the ring countries.

[83] We assume that the first country in each region is the numeraire ring country and it is also the numeraire country in each region.

The price and expenditure data by 155 expenditure categories and 6 regions in Tables 13 and 14 can now be regarded as a complete set of "country" data and any desired multilateral method can be used to form the PPP's between the 6 regions using this data set.[84] Of course, once these regional parities have been determined, the fixed within region parities can be used to obtain a complete set of PPP's for each country in the comparison project. The resulting set of country parities will respect the within region parities that have been determined by the regions. The overall procedure does not depend on the choice of numeraire countries, either within regions or between regions; i.e., the relative country parities will be the same no matter what the choices are for the numeraire countries.

17. Conclusion

The paper argues that the weighted CPD model developed in section 5 is a suitable theoretical target index that could be used at the first stage of aggregation. A practical approximation to this theoretical target index is developed in section 9 and this is our preferred method of aggregation at the basic heading level. One major unresolved issue with this method is that it is necessary to choose a relative weighting factor for the economic importance of representative price quotes versus unrepresentative quotes and we have not been able to provide definitive advice on the magnitude of this weighting factor.

Appendix: Some Properties of the Inverse Matrix in a Weighted CPD Regression Model

Recall (47) and (48) which defined the X^TX matrix and X^Ty vector that were used to define vectors of least squares estimators for the □ and □ vectors in the weighted CPD model using (28). In this Appendix, we find an easily checked sufficient condition for the existence of the inverse of X^TX and we also develop the monotonicity properties of the elements of $(X^TX)^{-1}$.

We need to express the elements of X^TX using matrix notation. In order to do this, we need to make a series of definitions. Thus define the *expenditure share (over all outlets) of country c on product n* as follows:[85]

(A1) $s_{cn} \equiv \sum_{k=1}^{K(c,n)} s_{cnk}$; $c = 1,\ldots,C;\ n = 1,\ldots,N.$

Define the *country c vector of expenditure shares* on the N products as

(A2) $s_c \equiv [s_{c1},\ldots s_{cN}]^T$; $c = 1,\ldots,C.$

Letting 1_N denote a column vector of ones of dimension N and 1_N^T its transpose, it can be seen that the following relations hold since the country expenditure shares sum to one:

[84] For surveys of possible multilateral methods, see Balk (1996) or Diewert (1999).

[85] If commodity n is not transacted in country c so that K(c,n) is 0, then define $s_{cn} \equiv 0$.

(A3) $1_N{}^T s_c = 1$; $c = 1,\ldots,C.$

Define the *world expenditure share sum vector* s as the sum of the country expenditure share vectors:[86]

(A4) $s \equiv \sum_{c=1}{}^C s_c$.

Finally, define the N by C−1 *matrix of country expenditure share vectors*, excluding the expenditure share vector of country 1, as follows:

(A5) $S \equiv [s_2,s_3,\ldots,s_N]$.

Substituting the above definitions into (47) and using equations (A3) yields the following expression for X^TX:

(A6) $$X^TX = \begin{bmatrix} I_{C-1} & S^T \\ S & \hat{s} \end{bmatrix}$$

where I_{C-1} is an identity matrix of size C−1 by C−1 and $\hat{s}$ is an N by N diagonal matrix with the elements of the world share sum vector s running down the main diagonal.

Assuming for the moment that the inverse of the N by N matrix $\hat{s} - SS^T$ exists, we can use elementary block row operations in order to obtain the following formula for the inverse of X^TX:

(A7) $$(X^TX)^{-1} = \begin{bmatrix} I_{C-1} + S^T(\hat{s} - SS^T)^{-1}S & -S^T(\hat{s} - SS^T)^{-1} \\ -(\hat{s} - SS^T)^{-1}S & (\hat{s} - SS^T)^{-1} \end{bmatrix}.$$

Thus if $(\hat{s} - SS^T)^{-1}$ exists, then (A7) can be used to construct $(X^TX)^{-1}$. We now find a set of conditions that are sufficient to imply the existence of the inverse of $\hat{s} - SS^T$. We first show that this matrix is positive semidefinite.[87] Using definition (A5), it can be seen that:

(A8) $$SS^T = [s_2,s_3,\ldots,s_C][s_2,s_3,\ldots,s_C]^T$$
$$= \sum_{c=2}{}^C s_c s_c{}^T .$$

Using (A4) and (A8), we have:

(A9) $$\hat{s} - SS^T = \sum_{c=1}{}^C \hat{s}_c - \sum_{c=2}{}^C s_c s_c{}^T$$
$$= \hat{s}_1 + \sum_{c=2}{}^C [\hat{s}_c - s_c s_c{}^T].$$

[86] Of course, s is not a share vector since it is a sum of C share vectors.

[87] It is easy to verify that it is a symmetric matrix.

Define the N by N matrices A_c as follows:

$$\text{(A10)}\quad A_c \equiv [\hat{s}_c - s_c s_c^T]\,; \qquad\qquad c = 2,3,\ldots,C.$$

We now show that each of the matrices A_c is positive semidefinite.[88] Thus for each N dimensional vector z, we need to show that

$$\begin{aligned}\text{(A11)}\quad 0 &\le z^T A_c z \\ &= z^T [\hat{s}_c - s_c s_c^T]\, z \\ &= z^T \hat{s}_c\, z - z^T s_c s_c^T z \\ &= z^T \hat{s}_c\, z - [s_c^T z]^2.\end{aligned}$$

Thus we need to show that for all z:

$$\text{(A12)}\quad [s_c^T z]^2 \le z^T \hat{s}_c\, z.$$

By the Cauchy Schwarz inequality, for any two N dimensional vectors, we have:

$$\text{(A13)}\quad [x^T y]^2 \le x^T x\, y^T y.$$

Since the country c share vector s_c is nonnegative, the matrix $\hat{s}_c$ is diagonal with the nonnegative elements of s_c running down the main diagonal. Thus the diagonal square root matrix $\hat{s}_c^{\,1/2}$, which has the nonnegative square roots of the diagonal elements of $\hat{s}_c$ running down the main diagonal, is well defined. Define

$$\text{(A14)}\quad x \equiv \hat{s}_c^{\,1/2}\, 1_N\,;\ y \equiv \hat{s}_c^{\,1/2}\, z\,.$$

Where 1_N is a vector of ones of dimension N. Upon substituting (A14) into (A13) and using (A3), we find that (A12) is true for $c = 2,3,\ldots,C$.

Using the positive semidefiniteness of the matrices A_c and using (A9), we see that $\hat{s} - SS^T$ is equal to the positive semidefinite matrix $\hat{s}_1$ plus the sum of $C - 1$ positive semidefinite matrices A_c for $c = 2,\ldots,C$ and hence is positive semidefinite.

Thus a simple condition that is sufficient to imply the existence of the inverse of $\hat{s} - SS^T$ is that $\hat{s}_1$ be positive definite or equivalently, that all of the N elements of s_1 be strictly positive. This simple sufficient condition can be written as follows:[89]

$$\text{(A15)}\quad s_1 \gg 0_N.$$

[88] It is obvious that each A_c is symmetric.

[89] Notation: 0_N is a vector of zeros of dimension N.

This condition means that the numeraire country has collected at least one price for each of the N commodities that are in the comparison.[90]

We now show that condition (A15) is also sufficient to imply that all of the elements of $[\hat{s} - SS^T]^{-1}$ are nonnegative. First, note that for c = 2,…,C, we have:

$$\text{(A16)}\ [\hat{s}_c - s_c s_c^T]1_N = s_c - s_c s_c^T 1_N$$
$$= s_c - s_c 1 \qquad \text{using (A3)}$$
$$= 0_N .$$

Using (A9) and (A16), it can be seen that:

$$\text{(A17)}\ [\hat{s}_c - SS^T]1_N = \hat{s}_1 1_N - 0_N = s_1 >> 0_N$$

where the strict inequality follows from assumption (A15). The matrix $\hat{s}_c - SS^T$ with positive elements on the main diagonal and nonnegative elements elsewhere is known in the economics literature as a Leontief matrix.[91] It turns out that condition (A17) is sufficient to imply that all elements of $[\hat{s}_c - SS^T]^{-1}$ are nonnegative; see Gale and Nikaido (1965; 86). Hence, recalling the formula for $(X^TX)^{-1}$ given by (A7), it can be seen that the northwest and southeast blocks of the inverse consist of nonnegative elements and the northeast and southwest blocks of the inverse consist of nonpositive elements.

Now regard the log PPP's for countries 2,…,C, $\pi_2,\ldots,\pi_C$, as functions of the prices of country 1, 2,…,C, which we denote by the price vectors $p^1, p^2, \ldots, p^C$ respectively. Using (48), (A7) and the nonnegativity properties of $(X^TX)^{-1}$ noted in the paragraph above, it can be seen that $\pi_2,\ldots,\pi_C$ are nonincreasing in the components of p^1; i.e., we have:

$$\text{(A18)}\ \nabla_{p^1}\pi_2(p^1,p^2,\ldots,p^C) \le 0_{N^1};\ \nabla_{p^1}\pi_3(p^1,p^2,\ldots,p^C) \le 0_{N^1};\ldots;\ \nabla_{p^1}\pi_C(p^1,p^2,\ldots,p^C) \le 0_{N^1}$$

where $\nabla_{p^1}\pi_c$ denotes the vector of first order partial derivatives of π_c with respect to the components of p^1 and N_1 is the total number of price quotes collected in country 1.[92] Thus if any price in country 1 increases, then the log PPP's of the other countries either remain unchanged or decrease relative to country 1, an intuitively plausible result.[93]

[90] Using the discussion below equation (3) in the main text, it can be shown that the inverse of $\hat{s} - SS^T$ will exist provided that any country c in the comparison has a strictly positive share vector s_c. While this condition is sufficient for the existence of the inverse, it is not necessary; i.e., weaker conditions will imply the existence of the inverse.

[91] See Hawkins and Simon (1949) and Gale and Nikaido (1965; 86).

[92] In this comparative statics exercise, as the components of p^1 increase, we hold constant all of the weighting shares.

[93] Melser (2005) has additional results on the monotonicity properties of CPD models for the case of two countries.

It is also straightforward to use (48), (A7) and the nonnegativity properties of $(X^TX)^{\square 1}$ to establish the following results:

(A19) $\square_2(p^1,\square p^2,p^3,\ldots,p^C) = \square_2(p^1,p^2,\ldots,p^C) + \ln \square$; for all $\square > 0$;
$\square_3(p^1,p^2,\square p^3,\ldots,p^C) = \square_2(p^1,p^2,\ldots,p^C) + \ln \square$; for all $\square > 0$;

…
$\square_C(p^1,p^2,p^3,\ldots,\square p^C) = \square_2(p^1,p^2,\ldots,p^C) + \ln \square$; for all $\square > 0$.

Define the PPP's for countries 2,…,C relative to the numeraire country 1 as the exponentials of the corresponding log parities:

(A20) $a_c(p^1,p^2,p^3,\ldots,p^C) \equiv \exp[\square_c(p^1,p^2,p^3,\ldots,p^C)]$; $c = 2,3,\ldots,C$.

Using (A19), it can be seen that each parity is linearly homogeneous in its own prices; i.e., we have:[94]

(A21) $a_2(p^1,\square p^2,p^3,\ldots,p^C) = \square a_2(p^1,p^2,\ldots,p^C)$; for all $\square > 0$;
$a_3(p^1,p^2,\square p^3,\ldots,p^C) = \square a_2(p^1,p^2,\ldots,p^C)$; for all $\square > 0$;

…
$a_C(p^1,p^2,p^3,\ldots,\square p^C) = \square a_2(p^1,p^2,\ldots,p^C)$; for all $\square > 0$.

More work remains to be done in developing the axiomatic properties of the weighted CPD parities.

References

Allen, R.C. and W.E. Diewert (1981), "Direct versus Implicit Superlative Index Number Formulae", *The Review of Economics and Statistics* 63, 430-435.

Balk, B.M. (1980), "A Method for Constructing Price Indices for Seasonal Commodities", *Journal of the Royal Statistical Society A* 143, 68-75.

Balk, B.M. (1996), "A Comparison of Ten Methods for Multilateral International Price and Volume Comparisons", *Journal of Official Statistics* 12, 199-222.

Balk, B.M. (2001), "Aggregation Methods in International Comparisons: What Have we Learned?", Research in Management Series, Report ERS-2001-41-MKT, Erasmus Research Institute of Management, Erasmus University Rotterdam.

Cuthbert, J. and M. Cuthbert (1988), "On Aggregation Methods of Purchasing Power Parities", Working Paper No. 56, November, Paris: OECD.

[94] Dikhanov (2004) has developed additional axiomatic properties for CPD models.

Deaton, A. (2004), "Interpreting Regressions", presentation at the Meeting of the Technical Advisory Group for the International Comparisons Project at the World Bank, Washington, D.C., September 22-24.

Diewert, W. E. (1978), "Superlative Index Numbers and Consistency in Aggregation", *Econometrica* 46, 883-900; reprinted as pp. 253-273 in *Essays in Index Number Theory, Volume 1*, W. E. Diewert and A. O. Nakamura (eds.), Amsterdam: North-Holland, 1993.

Diewert, W.E. (1999), "Axiomatic and Economic Approaches to International Comparisons", pp. 13-87 in *International and Interarea Comparisons of Income, Output and Prices*, A. Heston and R.E. Lipsey (eds.), Studies in Income and Wealth, Volume 61, Chicago: The University of Chicago Press.

Diewert, W.E. (2002a), "Similarity and Dissimilarity Indexes: An Axiomatic Approach", Department of Economics, Discussion Paper 02-10, University of British Columbia, Vancouver, B.C., Canada, V6T 1Z1.

Diewert, W.E. (2002b), "Weighted Country Product Dummy Variable Regressions and Index Number Formulae", Department of Economics, Discussion Paper 02-15, University of British Columbia, Vancouver, B.C., Canada, V6T 1Z1.

Diewert, W.E. (2004), "Elementary Indices", pp. 355-371, Chapter 20 in *Consumer Price Index Manual: Theory and Practice*, ILO/IMF/OECD/UNECE/Eurostat/The World Bank, Geneva: International Labour Office.

Dikhanov, Y. (2004), "Comments on 'The Ring Comparisons' by Peter Hill", unpublished note, The World Bank.

Dikhanov, Y. (2004), "Assessing Efficiency of Elementary Indices with Monte Carlo Simulations", paper presented at the Washington Meeting of the Technical Advisory Group for the International Comparisons Project at the World Bank, September 22-24, 2004.

Eltetö, O. and P. Köves (1964), "On a Problem of Index Number Computation Relating to International Comparisons", *Statisztikai Szemle* 42, 507-518.

Ferrari, G., G. Gozzi and M. Riani (1996), "Comparing CPD and GEKS Approaches at the Basic Heading Level", pp. 323-337 in *CPI and PPP: Improving the Quality of Price Indices*, Proceedings of the Firenze Conference: Luxemburg: Eurostat.

Fisher, I. (1922), *The Making of Index Numbers*, Houghton-Mifflin, Boston.

Gale, D. and H. Nikaido (1965), "The Jacobian Matrix and the Global Univalence of Mappings", *Mathematische Annalen* 159, 81-93.

Hawkins, D. an H.A. Simon (1949), “Note: Some Conditions of Macroeconomic Stability”, *Econometrica* 17, 245-248.

Heston, A., R. Summers and B. Aten (2001), “Some Issues in Using Chaining Methods for International Real Product and Purchasing Power Comparisons”, Paper presented at the Joint World Bank and OECD Seminar on Purchasing Power Parities, January 30-February 2, Washington D.C.

Hill, Peter, (2004), “The Estimation of PPPs for Basic Headings”, Chapter 10, *ICP Manual*, [Draft], August 10, Washington DC: The World Bank.

Hill, R.J. (1995), Purchasing Power Methods of Making International Comparisons, Ph. D. dissertation, Vancouver: The University of British Columbia.

Hill, R.J. (1999a), “Comparing Price Levels across Countries Using Minimum Spanning Trees”, *The Review of Economics and Statistics* 81, 135-142.

Hill, R.J. (1999b), “International Comparisons using Spanning Trees”, pp. 109-120 in *International and Interarea Comparisons of Income, Output and Prices*, A. Heston and R.E. Lipsey (eds.), Studies in Income and Wealth Volume 61, NBER, Chicago: The University of Chicago Press.

Hill, R.J. (2001), “Measuring Inflation and Growth Using Spanning Trees”, *International Economic Review* 42, 167-185.

Hill, R.J. (2004), “Constructing Price Indexes Across Space and Time: The Case of the European Union”, *American Economic Review*, forthcoming.

Hill, R.J. and M. Timmer (2004), “Standard Errors as Weights in Multilateral Price Indexes”, revised version of a paper presented at the Workshop: Estimating Production and Income Across Nations at UC Davis, Institute of Governmental Affairs organized by Robert Feenstra, April 13-14, 2004, revised September 22.

McCarthy, P. (2004), “Alternatives for Linking Regions in the ICP”, paper presented at the Washington Meeting of the Technical Advisory Group for the International Comparisons Project at the World Bank, September 22-24, 2004.

Melser, D. (2005), “The Hedonic Regression Time-Dummy Method and the Monotonicity Axioms”, forthcoming in the *Journal of Business and Economic Statistics*.

Rao, C.R. (1965), *Linear Statistical Inference and Its Applications*, New York: John Wiley & Sons.

Rao, D.S. Prasada (1990), "A System of Log-Change Index Numbers for Multilateral Comparisons", pp. 127-139 in *Comparisons of Prices and Real Products in Latin America*, J. Salazar-Carillo and D.S. Prasada Rao (eds.), New York: Elsevier Science Publishers.

Rao, D.S. Prasada (1995), "On the Equivalence of the Generalized Country-Product-Dummy (CPD) Method and the Rao System for Multilateral Comparisons", Working Paper No. 5, Centre for International Comparisons, University of Pennsylvania, Philadelphia.

Rao, D.S. Prasada (2001), "Weighted EKS and Generalized CPD Methods for Aggregation at the Basic Heading Level and Above Basic Heading Level", Paper presented at the Joint World Bank and OECD Seminar on Purchasing Power Parities, January 30-February 2, Washington D.C.

Rao, D.S. Prasada (2002), "On the Equivalence of Weighted Country Product Dummy (CPD) Method and the Rao System for Multilateral Price Comparisons", School of Economics, University of New England, Armidale, Australia, March.

Rao, D.S. Prasada (2004), "The Country-Product-Dummy Method: A Stochastic Approach to the Computation of Purchasing Power parities in the ICP", paper presented at the SSHRC Conference on Index Numbers and Productivity Measurement, June 30-July 3, 2004, Vancouver, Canada.

Szulc, B. (1964), "Indices for Multiregional Comparisons", *Przeglad Statystyczny* 3, 239-254.

Selvanathan, E. A. and D. S. Prasada Rao (1994), *Index Numbers: A Stochastic Approach*, Ann Arbor: The University of Michigan Press.

Sergueev, S. (2001), "Measures of the Similarity of the Country's Price Structures and their Practical Application", Conference on the European Comparison Program, U. N. Statistical Commission. Economic Commission for Europe, Geneva, November 12-14, 2001.

Sergeev, S. (2002), "Calculation of Equi-Characteristic PPPs at the Basic Heading Level (Modification of the Method of 'Asterisks')", Paper presented at the Meeting of the Working Party on Purchasing Power Parities, Eurostat, Luxembourg, June 12-13, 2002.

Sergeev, S. (2003), "Recent Methodological Issues: Equi-Representativity and Some Modifications of the EKS Method at the Basic Heading Level", Working Paper No. 8, Statistical Commission and Economic Commission for Europe, Joint Consultation on the European Comparison Programme, Geneva, March 31-April 2, 2003.

Summers, R. (1973), “International Comparisons with Incomplete Data”, *Review of Income and Wealth* 29:1, 1-16.

Theil, H. (1967), *Economics and Information Theory*, Amsterdam: North-Holland.

Theil, H. ((1971), *Principles of Econometrics*, New York: John Wiley and Sons.

Törnqvist, L. (1936), “The Bank of Finland's Consumption Price Index”, *Bank of Finland Monthly Bulletin* 10, 1-8.

Törnqvist, L. and E. Törnqvist (1937), Vilket är förhällandet mellan finska markens och svenska kronans köpkraft?”, *Ekonomiska Samfundets Tidskrift* 39, 1-39 reprinted as pp. 121-160 in *Collected Scientific Papers of Leo Törnqvist*, Helsinki: The Research Institute of the Finnish Economy, 1981.

Triplett, J. E. and R. J. McDonald (1977), “Assessing the Quality Error in Output Measures: The Case of Refrigerators”, *The Review of Income and Wealth* 23:2, 137-156.