Hyperextended-Real-Valued Indexes of Absolute Dissimilarity*

Keir G. Armstrong

Department of Economics
Carleton University
Ottawa, ON K1S 5B6

karmstro@ccs.carleton.ca

September 17, 2004

* The author wishes to thank Prasada Rao, Alan Heston, Bettina Aten and Kam Yu for helpful comments on an earlier version of this paper, and the Social Science Research Council of Canada for financial support.

Abstract

This paper extends Diewert's (2002) study of similarity indexes to deal with the practical problem of the values of his preferred indexes of absolute dissimilarity going to infinity as any of their (per-capita) quantity arguments goes to zero; i.e., the inability of these indexes to distinguish between two quantity vectors that, between them, contain two or more zero components. Appropriately modified versions of these indexes are then used to generate alternative sets of weights for the purpose of implementing Hill's (1999) minimum spanning tree approach to multilateral international comparisons on the basis of each of two cross-sectional data sets produced by the United Nations International Comparison Project.

JEL Classification Numbers: C1, C43, C81.

Key Words: international comparisons; index numbers; similarity measures; axiomatic approach.

1. Introduction

Robert Hill (1999a,b) introduced a new method for making multilateral international comparisons of relative purchasing power that "provides a path by which to chain over space that involves the minimum number of binary comparisons to link all countries in a comparison" (Heston et al., 2001, p. 9). The idea behind this approach is that $n \geq 2$ countries can be compared in the same way as n time periods—viz. by chaining with respect to a superlative bilateral index such as the Fisher ideal—if an appropriate path can be found among the former that is analogous to the natural (linear) ordering of the latter. The notion that two non-adjacent time periods are best compared by chaining appropriate binary comparisons along the (sub-)path that connects them derives from the fact that adjacent time periods tend to be more similar in terms of the structure of their associated commodity baskets. Thus chaining provides a way to smooth the transition between two relatively dissimilar commodity baskets thereby improving the soundness of the corresponding comparison. The justification for this claim stems from the fact that the Paasche and Laspeyres indexes, which provide lower and upper bounds on the associated "true" cost-of-living index, will be close to one another if the underlying commodity baskets are similar. Consequently, such similarity "will lead to a very close approximation to the cost-of-living index" (Diewert, 1983, pp. 186–87).

Hill's (1999a,b) method constructs the minimum-spanning tree (MST) of a weighted connected graph with vertices corresponding to a bloc of countries and weights given by the Paasche-Laspeyres spreads among these countries. This choice of weights is supposed to reflect the pairwise degrees of similarity among the associated commodity baskets; i.e., the bigger is the difference between the Paasche and Laspeyres (price or quantity) indexes for a pair of countries, the more dissimilar are the associated commodity baskets. According to Diewert (2002, p. 2),

> [t]he problem with this measure of dissimilarity in the price [or quantity] structures of the two countries is that we could have [equality between the Paasche and Laspeyres numbers] (so that the Hill measure would register a maximal degree of similarity) but [the price or quantity vector of one country] could be very different [from that of the other]. Thus there is a need for a more systematic study of similarity (or dissimilarity) measures in order to pick the "best" one that could be used as an input into Hill's (1999a,b) spanning tree algorithm for linking countries.

Diewert (2002) then proceeds to provide just such a study by taking an axiomatic approach to both absolute indexes of quantity dissimilarity and relative indexes of price dissimilarity. The former regards the *per-capita* quantity vectors of two countries as being dissimilar if they are unequal, whereas the latter regards the price vectors of two countries as being dissimilar if one is not a positive scalar multiple of the other; i.e., if *relative* prices are not the same in both countries. The bottom line with respect to the analysis of absolute indexes of quantity dissimilarity is that two specific functional forms, the asymptotically linear and the asymptotically quadratic (defined below), give rise to the "preferred measures of absolute quantity dissimilarity [because] [t]hese indexes satisfy all of the important axioms" (Diewert, 2002, p. 22).

From a practical perspective, the weak point of this analysis is that the values of the preferred indexes of absolute dissimilarity go to infinity as any of their (per-capita) quantity arguments goes to zero. Consequently, neither of these indexes can distinguish between two quantity vectors that, between them, contain two or more zero components. This problem can be attributed to the fact that both indexes compute component-level degrees of dissimilarity between two quantity vectors in ratio terms. A possible alternative measure in absolute difference terms can only be invariant to changes in the units of measurement of the quantities at the cost of not being an increasing function of any given quantity when its counterpart is zero; i.e., given country i's per-capita quantity vector $x_i := (x_{i1}, \ldots, x_{im})^\top \in \Re_+^m$ and country j's per-capita quantity vector $x_j := (x_{j1}, \ldots, x_{jm})^\top \in \Re_+^m$, the ℓth component measure

$$d(x_{i\ell}, x_{j\ell}) = \frac{|x_{i\ell} - x_{j\ell}|}{x_{i\ell} + x_{j\ell}}, \ \ell \in \{1, \ldots, m\}, \tag{1}$$

of the dissimilarity index

$$D(x_i, x_j) = \tfrac{1}{m} \sum_{\ell=1}^{m} d(x_{i\ell}, x_{j\ell}) \tag{2}$$

is equal to one for every $x_{j\ell} > 0$ when $x_{i\ell} = 0$. Note that, as a non-convex function of $x_{j\ell}$ for $x_{i\ell} > 0$ since

$$\frac{\partial^2 d(x_{i\ell}, x_{j\ell})}{\partial x_{j\ell}^2} = \frac{\partial}{\partial x_{j\ell}} \left[\frac{2x_{i\ell}}{(x_{i\ell} + x_{j\ell})^2} \right] = \frac{-4x_{i\ell}}{(x_{i\ell} + x_{j\ell})^3} < 0,$$

d defined by (1) would be ruled out by Diewert (2002, p. 7). Besides, since this *d* goes to infinity as $(x_{i\ell}, x_{j\ell})$ goes to (0, 0), the associated *D* is unable to distinguish between quantity vectors that have two or more *corresponding* zero components.

The present paper solves the zero-component problem by first defining a hyperextended-real range for *d* and *D* in terms of transfinite ordinal numbers. The mathematical foundations of this definition are developed in Section 2. Section 3 re-casts Diewert's (2002) study of absolute dissimilarity measures in terms of hyperextended-real analysis and shows that his (suitably re-defined) preferred indexes continue "to satisfy all of the important axioms." Empirical illustrations of the MST method with weights given by these preferred indexes are the focus of Section 4. The sensitivity of the method with respect to the degree to which zero-components are allowed to influence the results is also shown. Section 5 offers some concluding remarks.

2. To Infinity and Beyond [1]

The starting point for the solution to the zero-component problem is the definition of the hyperextended real number system. A brief review of the fundamentals of transfinite set theory is provided in advance of this definition.

A set *X* is an *ordered set* with respect to an *ordering relation* $\prec$ if (i) for every pair of distinct elements *x* and *x′* in *X*, either $x \prec x'$ or $x' \prec x$, and (ii) for every trio of distinct elements *x*, *x′* and *x″* in *X*, if $x \prec x'$ and $x' \prec x''$, then $x \prec x''$; i.e., $\prec$ is transitive (Kamke, 1950, pp. 52–3). An ordered set *X* with an ordering relation $\prec_X$ is said to be *similar* to an ordered set *Y* with an ordering relation $\prec_Y$ if *X* can be mapped on *Y* so that if $x \in X$ corresponds to $y \in Y$ and $x' \in X$ corresponds to $y' \in Y$, then $x \prec_X x'$ implies $y \prec_Y y'$ (p. 55). An *order type* μ refers to an arbitrary representative *X* of a class of mutually similar ordered sets. The order type of the set of natural numbers, ordered according to increasing magnitude, is denoted by ω (p. 57). An ordered set *X* is *well-ordered* if it and all of its nonempty subsets have a first element with respect to the associated ordering relation $\prec$ (p. 79). An *ordinal number* is an order type that is represented by well-ordered sets (p. 80).

[1] Quote from Lasseter (1995) suggested by an anonymous seminar participant.

The first transfinite ordinal number is the order type of the set of all (finite) ordinal numbers preceding it; i.e., the order type of the set of all natural numbers $\Im := \{0, 1, 2, \ldots\}$, which is ω. This ordinal number is essentially different from those preceding it because, being a limit number, it has no immediate predecessor. However, as with every ordinal number, ω has an immediate successor: $\omega + 1$. Then comes $\omega + 2$, etc., leading to the sequence of ordinal numbers

$$0, 1, 2, \ldots, \omega, \omega + 1, \omega + 2, \ldots .$$

Since this sequence has order type $\omega + \omega =: \omega \cdot 2$, the ordinal number following it is $\omega \cdot 2$. The successor to this number is $\omega \cdot 2 + 1$, which is followed by $\omega \cdot 2 + 2, \ldots, \omega \cdot 3$; etc. Thus the beginning of the sequence of ordinal numbers is

$$0, 1, 2, \ldots, \omega, \omega + 1, \omega + 2, \ldots, \omega \cdot 2, \omega \cdot 2 + 1, \ldots, \omega \cdot z, \omega \cdot z + 1, \ldots (z \in \Im).$$

Since this sequence has order type $\omega \cdot \omega =: \omega^2$, the ordinal number following it is ω^2. This number is followed by $\omega^2 + 1, \omega^2 + 2, \ldots, \omega^2 + \omega, \ldots$—in general, all ordinal numbers of the form $\omega^2 + \omega \cdot z_1 + z_0$, where z_0 and z_1 are in $\Im$. Since the sequence of these ordinal numbers has order type $\omega^2 \cdot 2$, this is the ordinal number that follows it. Continuing in this manner yields all ordinal numbers that can be written in "polynomial" form:

$$\omega^k \cdot z_k + \omega^{k-1} \cdot z_{k-1} + \ldots + \omega \cdot z_1 + z_0 ,$$

where k and $z_k, z_{k-1}, \ldots, z_1, z_0$ are in $\Im$, and $\omega^t \cdot 1 \equiv \omega^t$ and $\omega^t \cdot 0 \equiv 0$ for any $t \in \Im$ (Kamke, 1950, pp. 98–9).

Define the term *hyperextended* (*non-negative*) *real number* to mean a number that can be expressed as the sum of a (finite or transfinite) ordinal number and a real number that is greater than or equal to zero and strictly less than one. Define the set of all such numbers, the *hyperextended real number system*,[2] denoted by $\Re_*$, as the sum of the set of all ordinal numbers, denoted by $\Im_*$, and the half-open unit interval $[0, 1) \subset \Re$; i.e.,

$$\Re_* := \{\mu + \lambda : \mu \in \Im_*, \lambda \in [0, 1)\} .$$

[2] The *extended real number system* normally refers to the set $\Re \cup \{\omega, -\omega\}$. See, for example, Rudin (1976, pp. 11–12)

3. Hyperextended Real Analysis of Absolute Dissimilarity Indexes

The focus of this section is the translation of Diewert's (2002) axiomatic framework for absolute dissimilarity indexes into one in which the domain of these indexes is $\Re_+^{2m}$ instead of $\Re_{++}^{2m}$ and the range is $\Re_*$ instead of $\Re_+$. To this end, let x_i and x_j be non-negative per-capita quantity vectors, and let $D(x_i, x_j)$ be the image of an absolute dissimilarity index $D : \Re_+^{2m} \to \Re_*$. Analogous to Diewert's (2002, §5) axioms B1–B8, desirable properties for D include

(i) non-negativity: $D(x_i, x_j) \geq 0$;

(ii) identity: $D(x_i, x_j) = 0$ if and only if $x_i = x_j$;

(iii) symmetry: $D(x_i, x_j) = D(x_j, x_i)$;

(iv) commensurability: $D(x_i, x_j) = D(\hat{a}x_i, \hat{a}x_j)$ for all $a := (a_1, \ldots, a_m)^\top \in \Re_{++}^m$, where $\hat{a}$ is the $m \times m$ diagonal matrix with $\hat{a}_{\ell\ell} = a_\ell$;

(v) monotonicity: $D(x_i, x_j)$ is increasing in $x_{j\ell}$ for each $\ell \in \{1, \ldots, m\} =: M$ if $x_j \geq x_i >> 0$;

(vi) continuity: D is a continuous function on $\Re_{++}^{2m}$;

(vii) ordering invariance: $D(x_i, x_j) = D(\tilde{I}_m x_i, \tilde{I}_m x_j)$ for any permutation of the columns of the $m \times m$ identity matrix $\tilde{I}_m$; and

(viii) additive separability: $D(x_i, x_j) = \sum_\ell d(x_{i\ell}, x_{j\ell})$ for some function $d : \Re_+^2 \to \Re_*$.

An additional desirable property for D is

(ix) triangle inequality: $D(x_i, x_j) \leq D(x_i, x_k) + D(x_k, x_j)$, where $x_k \in \Re_+^m$.

Any D that satisfies (ix) in addition to (i), (ii) and (iii) is a metric on $\Re_+^m$. As such, it possesses the most important properties of ordinary distance and is therefore a relatively intuitive measure of absolute dissimilarity.

Let $M_k := \{\ell : x_{k\ell} > 0\}$ be the set of commodities for which the associated country-k quantities are positive, and let $\overline{M}_k := \{\ell : x_{k\ell} = 0\}$ be the set of commodities for which the associated country-k quantities are zero.[3] The difference between the cardinality of the union and the cardinality of the intersection of the latter set for $k = i$ and

[3] Clearly, $M_k \cup \overline{M}_k \equiv M$.

$k = j$ is the number of commodities with associated zero-positive (or positive-zero) quantity pairs with respect to countries i and j. Each of these pairs is assigned the (transfinite ordinal) value ω, which corresponds to the ratio of the positive quantity to the zero quantity. Accordingly, the *asymptotically linear* and *asymptotically quadratic* indexes of absolute dissimilarity are defined, respectively, for $t = 1$ and $t = 2$, as

$$D_t(x_i, x_j) = \omega \cdot \left(\left| \overline{M}_i \cup \overline{M}_j \right| - \left| \overline{M}_i \cap \overline{M}_j \right| \right) + \frac{1}{m} \sum_{\ell \in M_i \cap M_j} \left[\left(\frac{x_{i\ell}}{x_{j\ell}} - 1 \right)^t + \left(\frac{x_{j\ell}}{x_{i\ell}} - 1 \right)^t \right] . \quad (3)$$

Note that the upper bound on $D_t(x_i, x_j)$ of $\omega \cdot m$ is achieved whenever the two per-capita baskets have no commodities in common but together cover all m of them, and the lower bound of 0 is achieved whenever the two baskets are identical.

THEOREM. $D_t(x_i, x_j)$ satisfies properties (i)–(viii) but not (ix). Hence $D_t(x_i, x_j)$ is not a metric on $\Re_+^m$.

The proof of this theorem can be found in the appendix.

Because the domain of D is $\Re_+^{2m}$, property (vi) is clearly a *weak* continuity requirement. Since

$$\lim_{x_i \to 0} D_t(x_i, x_j) = \omega \cdot m = D_t(0, x_j)$$

for any $x_j \in \Re_{++}^m$, D_t is a continuous function on $\Re_+^m \times \Re_{++}^m \supset \Re_{++}^{2m}$, and since

$$\lim_{x_j \to 0} D_t(x_i, x_j) = \omega \cdot m = D_t(x_i, 0)$$

for any $x_i \in \Re_{++}^m$, D_t is a continuous function on $\Re_{++}^m \times \Re_+^m \supset \Re_{++}^{2m}$ as well. D_t is *not* a continuous function on $\Re_+^{2m} \supset \Re_{+/++}^m \times \Re_{++/+}^m$, however, due to the existence of a discontinuity at each pair of baskets with one or more corresponding zero components; e.g., (0, 0). To see that this is so, let $x_j := \hat{a} x_i \in \Re_{++}^m$ for some $a := (a_1, \ldots, a_m)^\mathsf{T} \neq (1, \ldots, 1)^\mathsf{T}$ and take the limit of $D_t(x_i, x_j)$ as x_i approaches zero:

$$\lim_{x_i \to 0} D_t(x_i, \hat{a} x_i) = \frac{1}{m} \sum_{\ell=1}^{m} \left[\left(a_\ell^{-1} - 1 \right)^t + \left(a_\ell - 1 \right)^t \right] \neq 0 = D_t(0, 0) .$$

That a discontinuity exists with respect to two baskets with corresponding zero quantities is not too surprising given that such pairs represent, in effect, comparisons at a lower

dimensionality than those involving two baskets with fewer or no corresponding zero quantities.

The indexes defined by (3) assume implicitly that the degrees of dissimilarity between corresponding positive components of x_i and x_j are equally important by giving them equal weight.[4] This sort of assumption is not justifiable in many applications, which often call for weights that reflect the economic importance of the relevant commodities. In the present context, this can be done most appropriately by using the expenditure shares in the two countries. Thus, following Diewert (2002, p. 19), the *weighted asymptotically linear* and *weighted asymptotically quadratic* indexes of absolute dissimilarity are defined, respectively, for $t = 1$ and $t = 2$, as

$$D_t(x_i, x_j) = \omega \cdot \left(\left| \overline{M}_i \cup \overline{M}_j \right| - \left| \overline{M}_i \cap \overline{M}_j \right| \right) + \sum_{\ell \in M_i \cap M_j} s_\ell^{ij} \left[\left(\frac{x_{i\ell}}{x_{j\ell}} - 1 \right)^t + \left(\frac{x_{j\ell}}{x_{i\ell}} - 1 \right)^t \right], \quad (4)$$

where

$$s_\ell^{ij} := \frac{1}{2} \left(\frac{p_{i\ell} x_{i\ell}}{p_i \cdot x_i} + \frac{p_{j\ell} x_{j\ell}}{p_j \cdot x_j} \right) \quad (5)$$

is the mean expenditure share of countries i and j on commodity ℓ. Note that (4) with the weights s_ℓ^{ij} replaced by $1/m$ is the same as (3).

The indexes defined by (4) can be given a statistical interpretation as follows: Define the absolute dissimilarity of the ℓth quantity ratio between countries i and j as

$$d_\ell^{ij} = \begin{cases} \left(\frac{x_{i\ell}}{x_{j\ell}} - 1 \right)^t + \left(\frac{x_{j\ell}}{x_{i\ell}} - 1 \right)^t & : \quad x_{i\ell} > 0 \wedge x_{j\ell} > 0 \\ 0 & : \quad x_{i\ell} = 0 \wedge x_{j\ell} = 0 \\ \omega & : \quad x_{i\ell} > x_{j\ell} = 0 \vee x_{j\ell} > x_{i\ell} = 0 \end{cases}, \quad \ell = 1, \ldots, m. \quad (6)$$

Now define D^{ij} to be the discrete random variable that takes on the values $\{d_\ell^{ij}\}_{\ell=1}^m$ with probabilities $\{s_\ell^{ij}\}_{\ell=1}^m$. Note that $s_\ell^{ij} \in [0, 1]$ and $\sum_\ell s_\ell^{ij} = 1$. Thus the expected value of D^{ij} is

[4] The weights given to corresponding non-positive components of x_i and x_j are a non-issue since the associated measures of dissimilarity are either zero or infinity.

$$
\begin{aligned}
E(D^{ij}) &= \sum_{\ell=1}^{m} s_\ell^{ij} d_\ell^{ij} \\
&= \sum_{\ell \in \left(\overline{M}_i \cup \overline{M}_j\right) \setminus \left(\overline{M}_i \cap \overline{M}_j\right)} s_\ell^{ij} \omega + \sum_{\ell \in M_i \cap M_j} s_\ell^{ij} \left[\left(\frac{x_{i\ell}}{x_{j\ell}} - 1 \right)^t + \left(\frac{x_{j\ell}}{x_{i\ell}} - 1 \right)^t \right] \\
&= D_t(x_i, x_j) ,
\end{aligned}
$$

where the last equality follows by (4) since $s_\ell^{ij} \omega = \omega$. Consequently, as pointed out by Diewert (2002, p. 20), $D_t(x_i, x_j)$ can be interpreted as

> the expected value of the absolute dissimilarities of the quantity ratios between the two countries, where the m discrete quantity dissimilarities, [d_ℓ^{ij} defined by (6)], are weighted according to Theil's (1967, p. 138) probability weights, [s_ℓ^{ij} defined by (5)] for $\ell = 1, \ldots, m$.

4. Empirical Illustrations and Sensitivity Analysis

Consider a bloc of $n \geq 2$ countries indexed by the set $N := \{1, \ldots, n\}$ with positive country-specific price vectors $p_1, \ldots, p_n$ and non-negative per-capita quantity vectors $x_1, \ldots, x_n$, each corresponding to a common set of well-defined types of goods and services M. Hill's (1999a,b) MST method would measure the purchasing power parity (PPP) between any pair of countries in such a bloc as the product of the Fisher price indexes along a pre-determined path connecting these countries (via zero, one or more of the other countries). The pre-determined path within the bloc is supposed to correspond to the minimum total dissimilarity among the commodity baskets of the constituent countries subject to the constraint that there is a unique connection between each pair. More precisely, the MST PPP index for country $i \in N$ relative to country $j \in N$ is defined as the chain of Fisher price indexes across the minimum spanning tree T of the weighted connected graph G of order n with vertices N and weights (D^{ij}); i.e.,

$$
\rho_{MST}^{ij} = \rho_F^{ih} \rho_F^{hk} \cdots \rho_F^{tj} ,
$$

where $h, k, \ldots, t \in N$, $ih, hk, \ldots, tj$ are edges in T, and $\rho_F^{ij} := \left(\rho_L^{ij} \rho_P^{ij}\right)^{\frac{1}{2}}$, $\rho_L^{ij} := p_i^{\mathsf{T}} x_j / p_j^{\mathsf{T}} x_j$ and $\rho_P^{ij} := p_i^{\mathsf{T}} x_i / p_j^{\mathsf{T}} x_i$ are, respectively, the Fisher, Laspeyres and Paasche price indexes for country i relative to country j.

The weights used by Hill (1999a,b) to construct T are the Paasche-Laspeyres spreads $\left(D_{PLS}^{ij}\right)$, where

$$D_{PLS}^{ij} := \left|\ln\left(\rho_L^{ij} / \rho_P^{ij}\right)\right| \ .^{5}$$

An implication of the results of the preceding section is that a better choice would be $\left(D_t^{ij}\right)$, where D_t^{ij} is the weighted asymptotically linear (if $t = 1$) or weighted asymptotically quadratic (if $t = 2$) index of absolute dissimilarity given by (4). An empirical issue that arises from this choice relates to the precise treatment of zero-positive quantity pairs. In cases where the corresponding positive (mean) expenditure share is sufficiently small, it would be undesirable to treat the positive quantity as such because doing so would allow a relatively unimportant commodity to have a disproportionately large impact ($+\omega$) on the overall dissimilarity measure. A sensitivity analysis with respect to the magnitude of such a zero-quantity share cutoff is incorporated in the empirical illustrations below.

The bases for these illustrations are the 1980 and 1985 cross-sectional data sets produced by the United Nations International Comparison Project (ICP) and made freely available at the Center for International Comparisons Web site (pwt.econ.upenn.edu). The category PPPs and per-capita expenditures (in national currency units) of the major aggregate called "Private Final Consumption Expenditure"[6] for the forty-two countries that are common to both data sets were extracted and used to construct suitable price and per-capita quantity vectors[7] for the two years.[8] These price and quantity vectors were then used to construct five different sets of dissimilarity measures for each year: three based on the weighted asymptotically linear index with imposed zero-quantity share cutoffs of 0.005, zero and one; one based on the weighted asymptotically quadratic index with an imposed zero-quantity share cutoff of 0.005; and one equal to $\left(D_{PLS}^{ij}\right)$. Kruskal's (1956)

---

[5] Note that $\rho_L^{ij} / \rho_P^{ij} = \phi_L^{ij} / \phi_P^{ij}$, where $\phi_L^{ij} := p_j^{\mathrm{T}} x_i / p_j^{\mathrm{T}} x_j$ and $\phi_P^{ij} := p_i^{\mathrm{T}} x_i / p_i^{\mathrm{T}} x_j$ are, respectively, the Laspeyres and Paasche (per-capita) quantity indexes for country i relative to country j.

[6] Excluding the category "Net Purchases Abroad" to avoid the possibility of negative quantities.

[7] The former by dividing each category PPP by the corresponding U.S. value, and the latter by dividing each category per-capita expenditure multiplied by 1,000 by the corresponding element of the former. U.S.-dollar exchange rates were used in lieu of prices that could not be constructed due to missing category PPPs.

[8] The dimensionality of these vectors (m) is 107 for 1980 and 112 for 1985.

algorithm was then applied to G with each set of dissimilarity weights in turn to generate ten different MSTs.

Kruskal's (1956) algorithm begins by choosing an edge in G of minimum weight; i.e., $\arg\min_{ij}\{D^{ij} : i<j,\, i \in N\backslash\{n\}, j \in N\backslash\{1\}\} \in T$. Successive edges with progressively higher weights are examined and chosen if and only if doing so induces an acyclic subgraph of G. The algorithm stops once $n - 1$ edges have been selected.

The five MSTs for 1980 are depicted in Figure 1 and panel (a) of Figure 3, and the five MSTs for 1985 are depicted in Figure 2 and panel (b) of Figure 3. The vertices of each T therein are labelled with the appropriate ISO 3166(-1) A2 (two-letter Internet) country codes.[9] Panel (a) of each of Figures 1 and 2 shows the weighted asymptotically linear T with a zero-quantity share cutoff of 0.005, which means that any zero-positive quantity pairs with corresponding positive expenditure shares of half a percent or less were treated as if they were zero-zero quantity pairs. In Figure 1(a), the black lines indicate the edges of the relevant T. In Figure 2(a), the black lines indicate the edges that are also in Figure 1(a), and the grey lines indicate the edges that are not. The preponderance of grey lines therein (28 out of 41) illustrates what "Hill and others have noted, [namely that] the spanning tree will not necessarily be stable over time" (Heston et al., 2001, p. 10).

Panel (b) of each of Figures 1 and 2 shows the weighted asymptotically linear T with a zero-quantity share cutoff of zero, which means that all zero-positive quantity pairs were treated as such. Panel (c) shows the weighted asymptotically linear T with a zero-quantity share cutoff of one, which means that all zero-positive quantity pairs were treated as if they were zero-zero quantity pairs. Panel (d) shows the weighted asymptotically quadratic T with a zero-quantity share cutoff of 0.005. Figure 3 shows the MSTs that result from using the Paasche-Laspeyres spreads as weights.

[9] AT = Austria, BE = Belgium, BW = Botswana, CA = Canada, CI = Côte d'Ivoire, CM = Cameroon, DE = West Germany, DK = Denmark, ES = Spain, ET = Ethiopia, FI = Finland, FR = France, FY = Former Yugoslavia, GR = Greece, HK = Hong Kong, HU = Hungary, IE = Ireland, IN = India, IT = Italy, JP = Japan, KE = Kenya, KR = South Korea, LK = Sri Lanka, LU = Luxembourg, MA = Morocco, MG = Madagascar, ML = Mali, MW = Malawi, NG = Nigeria, NL = Netherlands, NO = Norway, PH = Philippines, PK = Pakistan, PL = Poland, PT = Portugal, SN = Senegal, TN = Tunisia, TZ = Tanzania, UK = United Kingdom, US = United States, ZM = Zambia, and ZW = Zimbabwe.

The black lines in each of panels (b), (c) and (d) of Figures 1 and 2 indicate the edges that are also in panel (a) of the same figure, and the grey lines indicate the edges that are not. Similarly, the black lines in Figures 3(a) and 3(b) indicate the edges that are also in Figures 1(a) and 2(a), respectively, and the grey lines indicate the edges that are not. Thus the proportions of grey relative to black lines in these figures illustrates the sensitivity of the MST method to the various choices of dissimilarity weights. With respect to the weighted asymptotically linear indexes, iterative application of Kruskal's (1956) algorithm to G with different zero-quantity share cutoffs yielded the range of such variation over which the associated T is affected. These ranges of variation are specified by the cutoff values stated in the captions for panels (b) and (c) of Figures 1 and 2.

Despite the fact that the details of the structure of T vary considerably with respect to the choice of dissimilarity weights, certain general characteristics do not. In particular, there is a strong tendency for the countries that have a high data-quality rating in Summers and Heston's (1984, pp. 259–60; 1991, pp. 363–6) estimation to form a sub-tree within T, and for the countries that have a low data-quality rating to do so as well. Figures 1(b), 2(a) and 2(d) are the best examples of this tendency since, in each case, all of the "A"- and "B"-rated countries form a sub-tree that is connected to another sub-tree comprised of all the "C"- and "D"-rated countries (via the GR-NG edge in the former case, and the GR-MA edge in the latter two).[10] Thus Heston et al.'s (2001, p. 8) notion that the MST "approach may help overcome some of the problems of quality control that have been difficult in spatial comparisons" is grounded to some extent in empirical fact.

5. Concluding Remarks

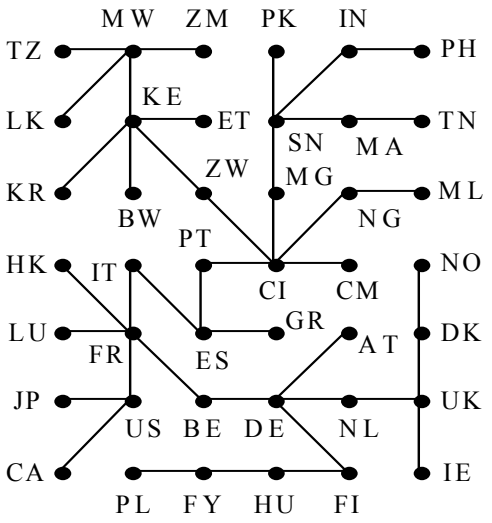
Heston et al. (2001, p. 9) suggested that "[o]ne could modify [the] EKS [method] to recognize the likely systematic differences in data quality or item qualities across countries." Since, as shown above, there is a tendency for countries of similar data quality to form clusters under the MST approach based on an appropriate index of absolute dissimilarity, it would seem that this method is a good candidate for effecting such a

[10] Note that Summers and Heston (1984; 1991) give India, Pakistan, the Philippines and Sri Lanka a "B" rating for 1980 and a "C" rating for 1985.

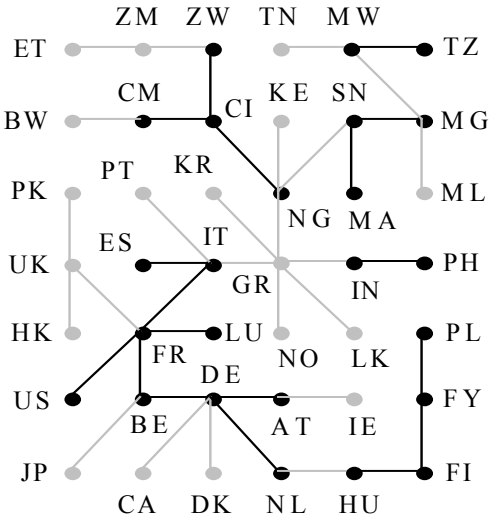
modification. For example, the PPPs among the countries in the low-quality cluster(s) could be determined by chaining with respect to the Fisher ideal index along the paths prescribed by the MST, whereas the PPPs among the countries in the high-quality cluster(s) could be determined using the EKS method thereby overruling the binary links within the associated sub-tree(s). The MST would also prescribe the binary link(s) between the high-quality cluster(s) and the low-quality cluster(s) thereby facilitating the calculation of the relevant cross-cluster PPPs. The resulting set of PPPs for the bloc as a whole would therefore be typed as an MST-EKS hybrid.

By providing a solution to the zero-component problem with respect to Diewert's (2002) preferred indexes of absolute dissimilarity, the present paper enables the calculation of these indexes on the basis of real-world data sets, within which the presence of zero quantities is not uncommon. By extension, the MST method using such index numbers as weights is also enabled. There is, however, a proviso to this assertion implicit in the empirical illustrations of the preceding section. The basis for these illustrations is a sub-aggregate of GDP rather than GDP itself because the latter includes possibly negative quantities associated with net foreign expenditures—quantities that do not fit within the absolute-dissimilarity-index framework. The problem here is that the preferred indexes therein compute component-level degrees of dissimilarity between two quantity vectors in ratio terms so that corresponding components with opposite signs maintain the same (negative) ratio as their absolute values are increased proportionately. Thus larger differences between such components do not register as larger measures of dissimilarity. The solution to this problem awaits further research.

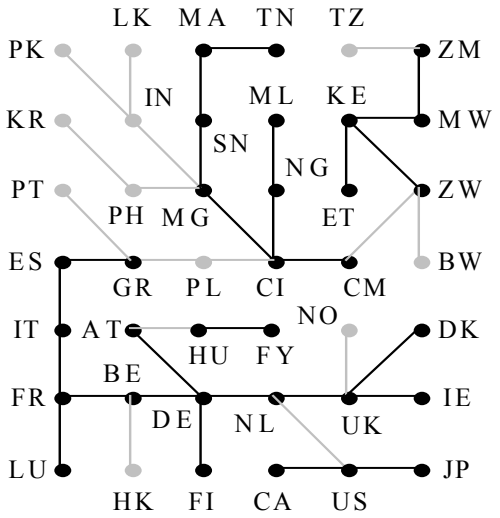
Figure 1.—42-Country Minimum Spanning Trees for 1980



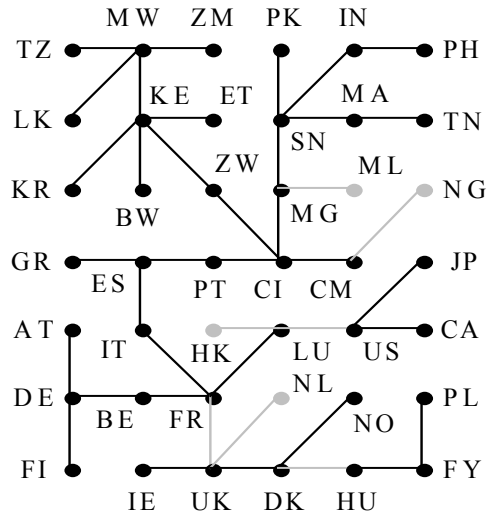
(a) Weighted asymptotically linear with a zero-quantity share cutoff of 0.005



(b) Weighted asymptotically linear with a zero-quantity share cutoff of 0.0000104 or less

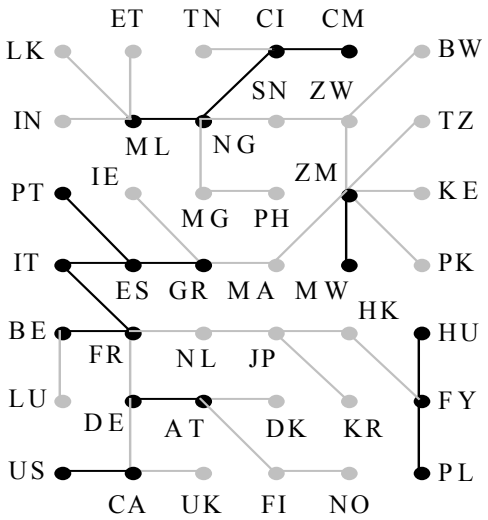


(c) Weighted asymptotically linear with a zero-quantity share cutoff of 0.0724 or more

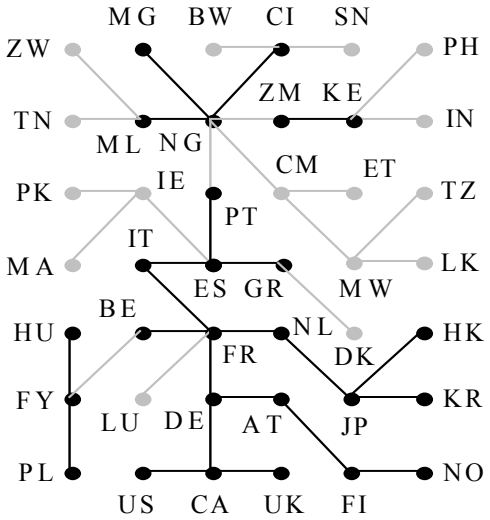


(d) Weighted asymptotically quadratic with a zero-quantity share cutoff of 0.005

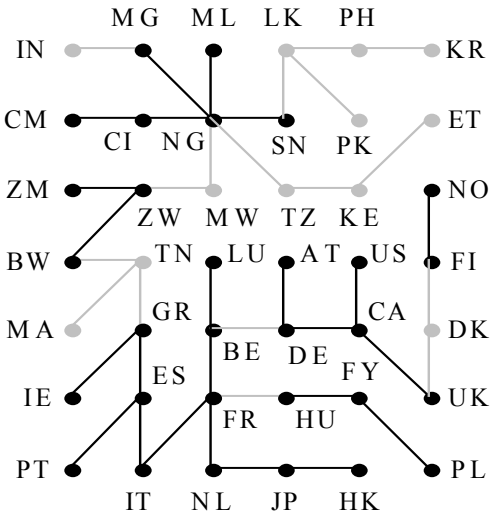
Figure 2.—42-Country Minimum Spanning Trees for 1985



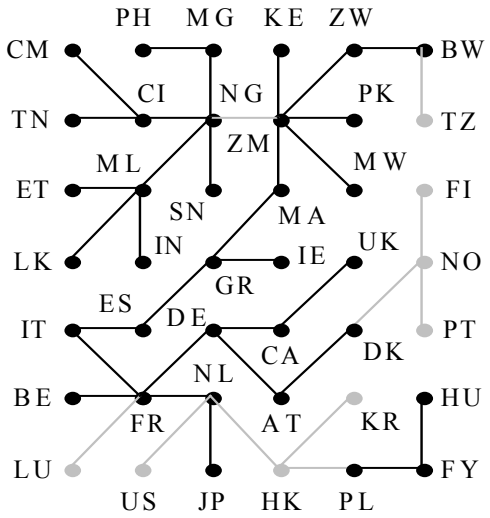
(a) Weighted asymptotically linear with a zero-quantity share cutoff of 0.005



(b) Weighted asymptotically linear with a zero-quantity share cutoff of 0.0000068 or less

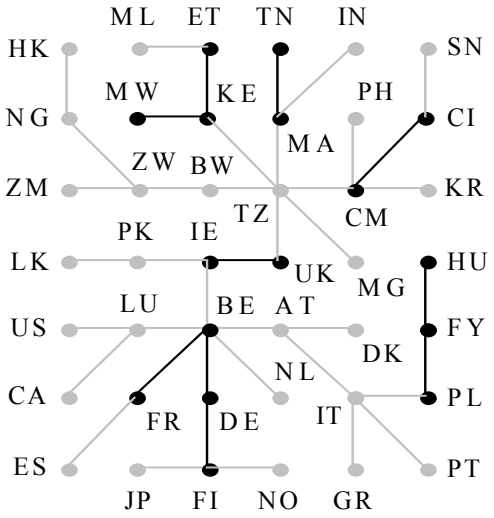


(c) Weighted asymptotically linear with a zero-quantity share cutoff of 0.0785 or more

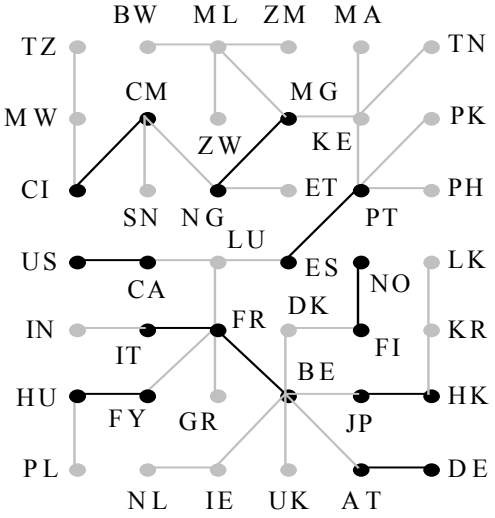


(d) Weighted asymptotically quadratic with a zero-quantity share cutoff of 0.005

Figure 3.—42-Country Minimum Spanning Trees Using Paasche-Laspeyres Spreads



(a) For 1980



(b) For 1985

Appendix

Proof of Theorem. (i) Satisfied since the cardinality of the union of two sets cannot be smaller than the cardinality of the intersection and since

$$\left(x_{i\ell} - x_{j\ell}\right)^2 \geq 0 \Leftrightarrow x_{i\ell}^2 - 2x_{i\ell}x_{j\ell} + x_{j\ell}^2 \geq 0 \Leftrightarrow \frac{x_{i\ell}^2 + x_{j\ell}^2}{x_{i\ell}x_{j\ell}} \geq 2 \Leftrightarrow \frac{x_{i\ell}}{x_{j\ell}} + \frac{x_{j\ell}}{x_{i\ell}} \geq 2 .$$

(ii) Satisfied since $\overline{M}_i \cup \overline{M}_j = \overline{M}_i \cap \overline{M}_j = \overline{M}_i$ and $x_{i\ell} / x_{j\ell} = 1$ for all $\ell \in M_i$.

(iii) Satisfied since the union, intersection and addition operators are commutative.

(iv) Satisfied since $a_\ell x_{i\ell} / a_\ell x_{j\ell} = x_{i\ell} / x_{j\ell}$. (v) Satisfied since

$$\frac{\partial D_t(x_i, x_j)}{\partial x_{j\ell}} = \frac{1}{m}\left[-t\left(\frac{x_{i\ell}}{x_{j\ell}} - 1\right)^{t-1} \frac{x_{i\ell}}{x_{j\ell}^2} + \frac{t}{x_{i\ell}}\left(\frac{x_{j\ell}}{x_{i\ell}} - 1\right)^{t-1} \right]$$

$$= \frac{t}{m}\left[\frac{1}{x_{i\ell}}\left(\frac{x_{j\ell} - x_{i\ell}}{x_{i\ell}}\right)^{t-1} - \frac{x_{i\ell}}{x_{j\ell}^2}\left(\frac{x_{i\ell} - x_{j\ell}}{x_{j\ell}}\right)^{t-1} \right]$$

$$= \frac{t}{m}\left[\frac{\left(x_{j\ell}-x_{i\ell}\right)^{t-1}}{x_{i\ell}^{t}} - \frac{x_{i\ell}\left(x_{i\ell}-x_{j\ell}\right)^{t-1}}{x_{j\ell}^{t+1}}\right]$$

$$= \frac{t}{m}\left[\frac{x_{j\ell}^{t+1}\left(x_{j\ell}-x_{i\ell}\right)^{t-1} - (-1)^{t-1}x_{i\ell}^{t+1}\left(x_{j\ell}-x_{i\ell}\right)^{t-1}}{x_{i\ell}^{t}x_{j\ell}^{t+1}}\right]$$

$$= \frac{t}{m}\left(x_{j\ell}-x_{i\ell}\right)\left[\frac{x_{j\ell}^{t^2-t+1} + x_{i\ell}^{t^2-t+1}}{x_{i\ell}^{t}x_{j\ell}^{t+1}}\right]$$

$$> 0\,,$$

where the last equality follows since $t \in \{1, 2\}$. (vi) Satisfied since $(x_{i\ell} / x_{j\ell} - 1)^t \in \Re$ is a continuous function for all $(x_{i\ell}, x_{j\ell}) \in \Re_{++}^2$ and $t \in \{1, 2\}$. (vii) Satisfied since the addition operator is commutative. (viii) Satisfied for

$$d_t(x_{i\ell}, x_{j\ell}) = \begin{cases} \frac{1}{m}\left[\left(\frac{x_{i\ell}}{x_{j\ell}}-1\right)^t + \left(\frac{x_{j\ell}}{x_{i\ell}}-1\right)^t\right] & : \quad x_{i\ell} > 0 \wedge x_{j\ell} > 0 \\ 0 & : \quad x_{i\ell} = 0 \wedge x_{j\ell} = 0 \\ \omega & : \quad x_{i\ell} > x_{j\ell} = 0 \vee x_{j\ell} > x_{i\ell} = 0 \end{cases}.$$

(ix) Not satisfied if $\overline{M}_i = \overline{M}_j = \overline{M}_k$ and $\min\{x_{i\ell}, x_{j\ell}\} < x_{k\ell} < \max\{x_{i\ell}, x_{j\ell}\}$ for all $\ell \in M_k$ since

$$D_1(x_i, x_k) + D_1(x_k, x_j) = \frac{1}{m}\sum_{\ell \in M_k}\left[\left(\frac{x_{i\ell}}{x_{k\ell}}-1\right)+\left(\frac{x_{k\ell}}{x_{i\ell}}-1\right)+\left(\frac{x_{k\ell}}{x_{j\ell}}-1\right)+\left(\frac{x_{k\ell}}{x_{j\ell}}-1\right)\right]$$

$$= \frac{1}{m}\sum_{\ell \in M_k}\left[\frac{x_{i\ell}}{x_{k\ell}} + \frac{x_{j\ell}}{x_{k\ell}} + \frac{x_{k\ell}}{x_{i\ell}} + \frac{x_{k\ell}}{x_{j\ell}} - 4\right]$$

$$= \frac{1}{m}\sum_{\ell \in M_k}\left[\frac{x_{j\ell}}{x_{k\ell}}\left(\frac{x_{i\ell}}{x_{j\ell}}+1\right) + \frac{x_{k\ell}}{x_{j\ell}}\left(\frac{x_{j\ell}}{x_{i\ell}}+1\right) - 4\right]$$

$$< \frac{1}{m}\sum_{\ell \in M_k}\left[\left(\frac{x_{i\ell}}{x_{j\ell}}+1\right)+\left(\frac{x_{j\ell}}{x_{i\ell}}+1\right)-4\right]$$

$$= \frac{1}{m}\sum_{\ell \in M_k}\left[\left(\frac{x_{i\ell}}{x_{j\ell}}-1\right)+\left(\frac{x_{j\ell}}{x_{i\ell}}-1\right)\right]$$

$$= D_1(x_i, x_j)$$

and

$$
\begin{aligned}
&D_2(x_i, x_k) + D_2(x_k, x_j) \\
&\quad = \tfrac{1}{m} \sum_{\ell \in M_k} \left[\left(\tfrac{x_{i\ell}}{x_{k\ell}} - 1\right)^2 + \left(\tfrac{x_{k\ell}}{x_{i\ell}} - 1\right)^2 + \left(\tfrac{x_{k\ell}}{x_{j\ell}} - 1\right)^2 + \left(\tfrac{x_{j\ell}}{x_{k\ell}} - 1\right)^2 \right] \\
&\quad = \tfrac{1}{m} \sum_{\ell \in M_k} \left[\left(\tfrac{x_{i\ell}}{x_{k\ell}}\right)^2 + \left(\tfrac{x_{j\ell}}{x_{k\ell}}\right)^2 + \left(\tfrac{x_{k\ell}}{x_{i\ell}}\right)^2 + \left(\tfrac{x_{k\ell}}{x_{j\ell}}\right)^2 - 2\left(\tfrac{x_{i\ell}}{x_{k\ell}} + \tfrac{x_{j\ell}}{x_{k\ell}} + \tfrac{x_{k\ell}}{x_{i\ell}} + \tfrac{x_{k\ell}}{x_{j\ell}}\right) + 4 \right] \\
&\quad = \tfrac{1}{m} \sum_{\ell \in M_k} \left\{ \left(\tfrac{x_{j\ell}}{x_{k\ell}}\right)^2 \left[\left(\tfrac{x_{i\ell}}{x_{j\ell}}\right)^2 + 1 \right] + \left(\tfrac{x_{k\ell}}{x_{j\ell}}\right)^2 \left[\left(\tfrac{x_{j\ell}}{x_{i\ell}}\right)^2 + 1 \right] - 2\left[\tfrac{x_{j\ell}}{x_{k\ell}} \left(\tfrac{x_{i\ell}}{x_{j\ell}} + 1\right) + \tfrac{x_{k\ell}}{x_{j\ell}} \left(\tfrac{x_{j\ell}}{x_{i\ell}} + 1\right) \right] + 4 \right\} \\
&\quad < \tfrac{1}{m} \sum_{\ell \in M_k} \left\{ \left[\left(\tfrac{x_{i\ell}}{x_{j\ell}}\right)^2 + 1 \right] + \left[\left(\tfrac{x_{j\ell}}{x_{i\ell}}\right)^2 + 1 \right] - 2\left[\left(\tfrac{x_{i\ell}}{x_{j\ell}} + 1\right) + \left(\tfrac{x_{j\ell}}{x_{i\ell}} + 1\right) \right] + 4 \right\} \\
&\quad = \tfrac{1}{m} \sum_{\ell \in M_k} \left\{ \left(\tfrac{x_{i\ell}}{x_{j\ell}}\right)^2 - 2\tfrac{x_{i\ell}}{x_{j\ell}} + 1 + \left(\tfrac{x_{j\ell}}{x_{i\ell}}\right)^2 - 2\tfrac{x_{j\ell}}{x_{i\ell}} + 1 \right\} \\
&\quad = \tfrac{1}{m} \sum_{\ell \in M_k} \left\{ \left(\tfrac{x_{i\ell}}{x_{j\ell}} - 1\right)^2 + \left(\tfrac{x_{j\ell}}{x_{i\ell}} - 1\right)^2 \right\} \\
&\quad = D_2(x_i, x_j) .
\end{aligned}
$$

The preceding inequalities follow from the fact that, for all $(x, y, z) \in \Re^3_{++}$,

$$
\begin{aligned}
&\frac{y}{z}\left(\frac{x}{y} + 1\right) + \frac{z}{y}\left(\frac{y}{x} + 1\right) < \left(\frac{x}{y} + 1\right) + \left(\frac{y}{x} + 1\right) \\
&\Leftrightarrow \frac{x}{z} + \frac{y}{z} + \frac{z}{x} + \frac{z}{y} < \frac{x}{y} + \frac{y}{x} + 2 \\
&\Leftrightarrow \frac{z - y}{x} + \frac{z - x}{y} + \frac{x + y}{z} < 2 \\
&\Leftrightarrow yz(z - y) + xz(z - x) + xy(x + y) < 2xyz \\
&\Leftrightarrow yz(z - y) + xz(z - x) < xy(z - x + z - y) \\
&\Leftrightarrow yz(z - y) + xz(z - x) < xy(z - x) + xy(z - y) \\
&\Leftrightarrow y(z - x)(z - y) + x(z - x)(z - y) < 0 \\
&\Leftrightarrow (x + y)(z - x)(z - y) < 0 \\
&\Leftrightarrow \min\{x, y\} < z < \max\{x, y\} \ . \blacksquare
\end{aligned}
$$

References

Diewert, W. Erwin. "The Theory of the Cost-of-Living Index and the Measurement of Welfare Change." In *Price Level Measurement*. Ed. W. Erwin Diewert and Claude Montmarquette. Minister of Supply and Services Canada, 1983, pp. 163–233.

Diewert, W. Erwin. "Similarity and Dissimilarity Indexes: An Axiomatic Approach." Discussion Paper No. 02-10. Department of Economics, University of British Columbia. 2002.

Heston, Alan, Robert Summers, and Bettina Aten. "Price Structures, the Quality Factor, and Chaining." TS. 13 April 2001.

Hill, Robert J. "Comparing Price Levels across Countries using Minimum-Spanning Trees" *Review of Economics and Statistics*, 81 (1999a), 135–142.

Hill, Robert J. "International Comparisons using Spanning Trees." In *International and Interarea Comparisons of Income, Output and Prices*. Ed. Alan Heston and Richard E. Lipsey. University of Chicago Press, 1999b, pp. 109–120.

Kamke, Erich. *Theory of Sets*. Trans. Frederick Bagemihl. New York: Dover Publications, 1950.

Kruskal, Joseph B. "On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem." *Proceedings of the American Mathematical Society*, 7, No. 1 (1956), pp. 48–50.

Lasseter, John, dir. *Toy Story*. With Tom Hanks and Tim Allen. Walt Disney Productions and Pixar Animation Studios, 1995.

Rudin, Walter. *Principles of Mathematical Analysis*. 3rd ed. [U.S.A.]: McGraw-Hill, 1976.

Summers, Robert, and Alan Heston. "Improved International Comparisons of Real Product and its Composition: 1950–1980." *Review of Income and Wealth*, 30 (1984), No. 2, 207–262.

Summers, Robert, and Alan Heston. "The Penn World Table (Mark 5): An Expanded Set of International Comparisons, 1950–1988." *Quarterly Journal of Economics*, 106 (1991), No. 2, 327–368.

Theil, Henri. *Economics and Information Theory*. Amsterdam: North-Holland, 1967.