1. Cluster formation, data validation and outlier detection in the International Comparison Program: An application to the Africa region

Statistics Department, African Development Bank[1]

Abstract
The International Comparisons Program (ICP), an undertaking coordinated by the World Bank, compares the purchasing power of currencies and real output of almost all countries in the world. ICP is divided into six regions, each of which is required to provide prices and expenditures for a common list of basic headings that sum to GDP. A potential benefit of participation in ICP is that it can help countries improve their price collection and national accounts, because a fundamental part of ICP is validation of the data they supply. This validation process, to the extent that it focuses on differences across countries, can detect errors in the data that countries might otherwise have difficulty finding. A weakness of the current ICP data validation methodology, however, is that it does not pay enough attention to expenditure data. This paper proposes some new methods that are well suited to detecting anomalies in expenditure data. In addition to potentially improving the quality of the overall ICP-Africa comparison, these methods could help participating countries improve the quality of their national accounts, and thus, contribute to the capacity-building process.

Key words: *cluster analysis, purchasing power parities, national accounts.*

Résumé
Le Programme de comparaison internationale (PCI), une initiative coordonnée par la Banque mondiale, compare le pouvoir d'achat des monnaies et la production réelle de presque tous les pays dans le monde. Le PCI est divisé en six régions, dont chacune est tenue de fournir des prix et des dépenses pour une liste commune de positions élémentaires qui constituent le PIB. Un avantage potentiel de la participation au PCI est qu'il peut aider les pays à améliorer leur collecte des prix et des comptes nationaux, car une partie fondamentale du PCI est la validation des données qu'ils fournissent. Ce processus de validation, dans la mesure où il met l'accent sur les différences entre les pays, peut détecter des erreurs dans les données que les pays pourraient autrement avoir de difficultés à trouver. Cependant une faiblesse de la méthode actuelle de validation des données du PCI, est qu'elle ne prête pas suffisamment d'attention aux données de dépenses. Cet article propose quelques nouvelles méthodes qui sont bien adaptés à la détection d'anomalies dans les données des dépenses. En plus de potentielle-

1 A longer version of this paper was prepared by Prof. Robert Hill, University of Graz, Austria and AfDB Consultant as part of the review of the 2005 results of the ICP-Africa.

ment améliorer la qualité de la comparaison de l'ensemble du PCI-Afrique, ces méthodes pourraient aider les pays participants à améliorer la qualité de leurs comptes nationaux, et donc, contribuer au processus de renforcement des capacités.

Mots clés : *analyse en grappes, anomalies, comptes nationaux*

1. INTRODUCTION

The International Comparisons Program (ICP), an undertaking coordinated by the World Bank, compares the purchasing power of currencies and real output of almost all countries in the world. Its results underpin the Penn World Table (probably the most widely used data set in economics). The most recent comparison was made in 2005; the next is scheduled for 2011.

ICP is divided up into six regions, each of which is required to provide prices and expenditures for a list of basic headings (129 in ICP 2005) that sum to GDP. This list is common to all regions. A basic heading is the lowest level of aggregation for which expenditure weights are available. The basic heading prices are constructed from region-specific product lists. Each country supplies prices for a sample of products within each basic heading. In ICP 2005, these were aggregated to obtain the basic heading prices using the country-product-dummy (CPD) method (or a variant thereof) in all regions except the OECD-Eurostat region, which used the Jevons-S method (see for example Hill and Hill 2009 or Diewert 2010 for an explanation of these methods). The expenditure weights, for the most part, were obtained from the national accounts.

A potential benefit of participation in ICP is that it can help countries improve their price collection and national accounts, because a fundamental part of ICP is validation of the data supplied by the countries. This validation process, to the extent that it focuses on differences across countries, can detect errors in the data that countries might otherwise have difficulty finding. Correction of these errors should improve the quality of the macroeconomic statistics in these countries.

The African Development Bank (AfDB) manages the ICP comparison in the Africa region. Its role in ICP, however, extends to serving as a capacity-building platform for price statistics and national accounts in participating countries. Innovations in data validation methodology, therefore, serve the dual AfDB purposes of potentially improving the quality of the ICP-

Africa comparisons and simultaneously strengthening the capacity-building platform.

The ICP data validation tool set consists primarily of the Dikhanov Table and Quaranta Table. The Quaranta Table is used only to validate the price quotes within each basic heading. The Dikhanov Table is more versatile and can be used to validate both the price quotes within each basic heading and the prices at higher levels of aggregation, including at the basic heading level. However, neither method is used to validate the expenditure weights at the basic heading level, which is an important weakness in ICP's existing data validation methodology.

This study shows how the ICP data validation process can be improved using dissimilarity measures and cluster analysis methods. This approach can be used equally well to validate the price or the expenditure data. Based on this approach, by far, the greatest anomalies in the ICP 2005 data for Africa are in the expenditure weights. As noted above, identification of these anomalies could enable participating countries in the Africa region to improve their national accounts, thereby contributing to the capacity-building platform as well as improving the quality of the overall ICP-Africa comparison.

The starting point for this study was the observation that in Africa, and some other regions, identification of anomalous data points is made more difficult by the diversity of countries in the region. In an attempt to ameliorate this problem, ICP-Africa is divided into subregions. However, these subregions are determined primarily by geographical location rather than economic similarity. An alternative is to use cluster analysis methods to identify more natural economic groupings of countries that can then be more easily compared.

This study considers some ways in which this can be done. The approach begins by defining a dissimilarity measure that can be applied to either the basic heading prices or quantities (which are derived implicitly from the expenditure data) or to a combination of the two. This dissimilarity measure is then used to make bilateral comparisons between all possible pairings of countries in a region. The outcome is a K by K dissimilarity matrix (where K is the number of countries in the region).

This dissimilarity matrix is a useful data validation tool, in that unusually large elements within it signal problems with the underlying data. In fact, the dissimilarity matrices are at least as useful as the resulting dendrograms and clusters for identifying anomalies in the data.

Returning to the problem of cluster formation, the dissimilarity matrix is used as the input into a cluster analysis algorithm that generates a dendrogram as its output. A dendrogram is a tree-like structure that begins with each country as its own cluster, and then merges in a sequential and hierarchical manner until the countries are all in one large cluster. By choosing appropriate thresholds within the dendrogram, it is possible to obtain a list of country clusters that are exhaustive and non-overlapping.

A number of alternative dissimilarity measure formulas are considered in this report. Dissimilarity measures defined on the price relatives and quantity relatives, and the dendrograms derived from them, may be particularly useful for detecting anomalies in the data. Dissimilarity measures defined on per capita GDP, however, may be more useful for constructing stable clusters of economically similar countries that can be used to refine the search for data anomalies.

The new methods proposed in this study complement the existing ICP data validation framework. These methods could prove particularly useful for detecting anomalies in the expenditure data. Applying these methods to the ICP 2005 basic heading data for the Africa region, this study finds that the expenditure data contain far more outliers than the price data. More generally, it is important, therefore, that more attention is paid to validation of the expenditure data in the Africa region (and other regions) in ICP 2011.

2. DISSIMILARITY MEASURES

2.1 Description

Dissimilarity measures are useful for data validation and detecting outliers. They also provide the input for the cluster analysis methods described in section 3. The theory of dissimilarity measures in a price and quantity index context has been developed by Diewert (2001, 2009). A dissimilarity measure, denoted here by d_{jk}, is inherently bilateral. In this context, it is used to compare the dissimilarity between the price and/or quantity vectors of a pair of countries j and k. Four axioms that d_{jk} could be required to satisfy are the following:

A1: $d_{jk} = d_{jk}$;
A2: $d_{jk} \geq 0$;
A3: $d_{jk} = 0$ when $p_{kn} = \lambda p_{jn}$ for all n;
A4: $d_{jk} = 0$ when $q_{kn} = \mu q_{jn}$ for all n;

where n in A3 and A4 indexes the basic headings over which the comparison is being made.

If there are K countries in the comparison, a $K \times K$ dissimilarity matrix can be calculated with element d_{jk} in the jth row and kth column. A1 implies that the dissimilarity matrix is symmetric. A2 implies that each element of the matrix is nonnegative. A3 implies that an element $d_{jk} = 0$ if the Hicks (1946) aggregation condition (i.e., $p_{kn} = \lambda p_{jn}$ for all n) is satisfied. When the Hicks condition is satisfied, all bilateral price index formulas (worth considering) give the same answer. The corresponding quantity indexes can then be derived implicitly via the factor reversal test. Hence, in this case, there is no index number problem (in the sense that the answer obtained does not depend on the choice of bilateral index number formula). Conversely, A4 implies that an element $d_{jk} = 0$ if the Leontief (1936) aggregation condition (i.e., $q_{kn} = \mu q_{jn}$ for all n) is satisfied. When the Leontief condition is satisfied, all bilateral quantity index formulas (worth considering) yield the same answer. Given that corresponding prices indexes can then be derived implicitly via the factor reversal test, again, in this case, there is no index number problem.

More generally, a dissimilarity measure in this context can be interpreted as measuring the confidence we have in our estimated bilateral price and quantity indexes, with a lower dissimilarity score implying greater confidence. Maximum confidence is achieved when $d_{jk} = 0$. A direct implication of each of A3 and A4 is that the dissimilarity matrix has zeroes on its lead diagonal. It is unlikely in practice that zero terms will be observed anywhere else in the matrix.

The first dissimilarity measure considered here is defined as follows:

$$d_{jk}^{PLS} = [\max(P_{jk}^{P}; P_{jk}^{L}) / \min(P_{jk}^{P}; P_{jk}^{L})] - 1$$

where P_{jk}^{P} and P_{jk}^{L} are Paasche and Laspeyres price indexes.[2] These and their corresponding quantity indexes are defined as follows:

$$\text{Paasche: } P_{jk}^{P} = (\Sigma_{n=1}^{N} p_{kn}q_{kn})/ (\Sigma_{n=1}^{N} p_{jn}q_{kn}),$$
$$Q_{jk}^{P} = (\Sigma_{n=1}^{N} p_{kn}q_{kn})/ (\Sigma_{n=1}^{N} p_{kn}q_{jn}); \quad (1)$$

$$\text{Laspeyres: } P_{jk}^{L} = (\Sigma_{n=1}^{N} p_{kn}q_{jn})/ (\Sigma_{n=1}^{N} p_{jn}q_{jn}),$$
$$Q_{jk}^{P} = (\Sigma_{n=1}^{N} p_{jn}q_{kn})/ (\Sigma_{n=1}^{N} p_{jn}q_{jn}). \quad (2)$$

2 The superscript PLS in d_{jk}^{PLS} stands for Paasche-Laspeyres spread.

d_{jk}^{PLS} could equally well be defined as a ratio of Paasche and Laspeyres quantity indexes,

since, by construction, $P_{jk}^{P}/P_{jk}^{L} = Q_{jk}^{P}/Q_{jk}^{L}$.

It can be verified that d_{jk}^{PLS} satisfies all four axioms. One weakness of d_{jk}^{PLS}, however, is that $d_{jk}^{PLS} = 0$ is necessary but not sufficient for there to be no index number problem. This is because d_{jk}^{PLS} may attain the value of zero even when neither A3 or A4 is satisfied.

Our second and third dissimilarity measures are defined as follows:

$$d_{jk}^{P} = \Sigma_{n=1}^{N}\{[(s_{jn} + s_{kn})/2][p_{kn}/(P_{jk}Fp_{jn}) + (P_{jk}Fp_{jn}/p_{kn}) - 2]\};$$

$$d_{jk}^{Q} = \Sigma_{n=1}^{N}\{[(s_{jn} + s_{kn})/2][q_{kn}/(Q_{jk}Fq_{jn}) + (Q_{jk}Fq_{jn}/q_{kn}) - 2]\}.$$

The terms s_{jn} and s_{kn} denote expenditure shares defined as follows:

$$s_{jn} = (p_{jn}q_{jn})/(\Sigma_{m=1}^{N} p_{jm}q_{jm}); \; s_{kn} = (p_{kn}q_{kn})/(\Sigma_{m=1}^{N} p_{km}q_{km}).$$

By construction, $\Sigma_{n=1}^{N} s_{jn} = \Sigma_{n=1}^{N} s_{kn} = 1$. Also, P_{jk}^{F} and Q_{jk}^{F} denote Fisher price and quantity indexes, respectively. These indexes are defined as follows:

$$\text{Fisher: } P_{jk}^{F} = (P_{jk}^{P} \times P_{jk}^{L})1/2, \; Q_{jk}^{F} = (Q_{jk}^{P} \times Q_{jk}^{L})^{1/2}. \qquad (3)$$

The dissimilarity measures d_{jk}^{P} and d_{jk}^{Q} are considered by Diewert (2001, 2009). He refers to them as weighted asymptotically linear indexes of relative dissimilarity. Diewert also considers a number of other relative dissimilarity formulas. Focusing on the price index case, these formulas differ in the influence exerted on the overall measure by each of the n = 1,…, N elements

$[p_{kn}/(P_{jk}^{F} p_{jn}) + (P_{jk}^{F} p_{jn})/p_{kn}]$. d_{jk}^{P} and d_{jk}^{Q} are *linear* indexes in the sense that influence is a linear function of the size of each element. Diewert also considers measures that are concave or convex functions of the size of each

element. Convex functions may be overly sensitive to outlier elements, while concave functions may not be sensitive enough to outliers.

d_{jk}^{P} satisfies only axioms A1, A2 and A3, while d_{jk}^{Q} satisfies only A1, A2 and A4. A weakness of both $d_{jk}^{P} = 0$ and $d_{jk}^{Q} = 0$ is that each is sufficient but not necessary for there to be no index number problem. More specifically, when the Leontief condition (i.e., $q_{kn} = \mu q_{jn}$ for all n) is satisfied, in general, $d_{jk}^{P} > 0$. Conversely, when the Hicks condition (i.e., $p_{kn} = \lambda p_{jn}$ for all n) is satisfied, in general, $d_{jk}^{Q} > 0$.

Consider an additional axiom A5 defined as follows:

A5: $d_{jk} = 0$ if and only if either $p_{kn} = \lambda p_{jn}$ for all n or $q_{kn} = \mu q_{jn}$ for all n.

All three dissimilarity measures considered thus far violate A5. Suppose now we combine d_{jk}^{P} and d_{jk}^{Q} as follows:

$$d_{jk}^{PQ} = (d_{jk}^{P} \times d_{jk}^{Q})^{1/2}.$$

It can be verified that the dissimilarity measure, d_{jk}^{PQ} satisfies all five axioms. That is, it attains the value of zero if and only if either the Hicks or Leontief aggregation conditions are satisfied (and hence there is no index number problem).

Despite the attractive axiomatic properties of d_{jk}^{PQ}, from a data validation perspective, the dissimilarity measures d_{jk}^{P} and d_{jk}^{Q} are probably more useful. This is because d_{jk}^{P} is probably best for detecting anomalies in the price data, and dQ jk, for detecting anomalies in the quantity (expenditure) data. Given that the price and quantity data are obtained from largely independent sources (the former from ICP surveys and the latter from the national accounts) the holistic approaches provided by d_{jk}^{PLS} and d_{jk}^{PQ} may not be as effective at detecting errors and anomalies in either data set. However, they are useful for detecting inconsistencies between the price and quantity data.

An alternative criterion for cluster formation is differences in per capita GDP. A dissimilarity measure derived from this criterion that has the same fundamental structure as those considered above is the following:

$$d_{jk}^{GDP} = y_k/y_j + y_j/y_k - 2; \quad (4)$$

where y_j and y_k denote the per capita GDP of countries j and k expressed in units of a common currency. In the empirical analysis that follows, this method is implemented using per capita incomes calculated using both the GEKS and Iklé methods.[3] Alternatively, per capita GDP can be replaced by the price level, defined for country j as the ratio of its purchasing power parity P_j to its corresponding market exchange rate MER_j - as follows:

$$d_{jk}^{PLev} = (P_k/MER_k)/(P_j/MER_j) + (P_j/MER_j)/(P_k/MER_k) - 2. \quad (5)$$

These dissimilarity measures are simpler than those derived from price or quantity relatives in that they are derived from only a single input from each country (i.e., per capita GDP or price level) that, moreover, is always positive. Hence, no expenditure share weights are needed, or adjustments to allow for negative or zero elements. We turn to this latter issue next.

2.2 The problem of zero or negative quantities

Two problems can arise in the construction of the price relative and quantity relative dissimilarity measures. First, the expenditure estimates for a few basic headings in some countries are zero. This may be because the commodities in that basic heading are not available, as is the case for 1102111 Spirits, 1102121 Wine, and 1102131 Beer in some countries. Alternatively, it may be because expenditure in that heading is not measured in a country, as is the case for 111220 Prostitution and 111261 Financial Services Indirectly Measured (FISIM). In ICP 2005, the quantity data are derived implicitly by dividing expenditure by price for each basic heading in each country. A zero expenditure, therefore, implies a zero quantity. In the presence of zero quantities, neither d_{jk}^{Q} nor d_{jk}^{PQ} is defined.

3 The GEKS method was used by all regions except Africa to compute the aggregate level results in ICP 2005. Africa uses Iklé. Descriptions of these methods can be found, for example, in Dikhanov (1997), Hill (1997), and Diewert (2011).

The second problem is that in some cases, expenditures (and hence, quantities) are negative. This can occur for the balancing item basic headings (i.e., 111300 Net purchases abroad, 130225 Receipts from sales, 130425 Receipts from sales, 140115 Receipts from sales, 160000 Change in inventories and valuables, and 180000 Balance of exports and imports). Negative quantity relatives (i.e., $q_{kn}/(Q_{jk}^{F}\, q_{jn}) < 0$) are not allowed in the d_{jk}^{Q} formula. This means that when both $Q_{jk}^{F}\, q_{jn}$ and q_{kn} are negative, there is no problem, because in this case, the quantity relative will still be positive. Likewise, negative expenditure shares are not allowed in either the d_{jk}^{P} or d_{jk}^{Q} formulas.

To avoid negative expenditure shares, the shares used in the d_{jk}^{P} and d_{jk}^{Q} are modified as follows:

$$s_{jn}^{*} = |s_{jn}|/(\Sigma_{m=1}^{N} |s_{jm}|)\ ;\ s_{kn}^{*} = |s_{kn}|/(\Sigma_{m=1}^{N} |s_{km}|).$$

The modified d_{jk}^{P} formula takes the following form:

$$d_{jk}^{P} = \Sigma_{n=1}^{N} \{[(s_{jn}^{*} + s_{kn}^{*})/2][p_{kn}/(P_{jk}^{F}\, p_{jn}) + (P_{jk}^{F}\, p_{jn})/p_{kn} - 2]\}. \qquad (6)$$

Further modifications to the d_{jk}^{Q} formula are required to allow for the possibility of zero quantities and negative quantity relatives.

$$d_{jk}^{Q} = \Sigma_{n=1}^{N} [(s_{jn}^{*} + s_{kn}^{*})/2]\, d_{jk,n}^{Q}, \qquad (7)$$

where

$$d_{jk,n}^{Q} = \min[20,\ q_{kn}/(Q_{jk}^{F}\, q_{jn}) + (Q_{jk}^{F}\, q_{jn})/q_{kn} - 2],\ \text{for } q_{kn}/(Q_{jk}^{F}\, q_{jn}) > 0;$$

$$d_{jk,n}^{Q} = 20\ \Sigma_{n=1}^{N} [(s_{jn}^{*} + s_{kn}^{*})/2],\ \text{for } q_{kn}/(Q_{jk}^{F}\, q_{jn}) < 0;\ q_{jn} = 0;\ \text{or } q_{kn} = 0.$$

Here an upper limit of 20 is imposed on each element $d_{jk,n}^{Q}$ in (7). This ensures that $d_{jk,n}^{Q}$ is defined when either q_{jn} or q_{kn} equals zero, and that it is well behaved when $q_{kn}/(Q_{jk}^{F}\, q_{jn})$ is negative. The inclusion of an upper limit on $d_{jk,n}^{Q}$ also ensures that the dendrograms derived from the quantity dissimilarity matrices are not oversensitive to outliers among the quantity relatives. In a price context, none of the elements $[p_{kn}/(P_{jk}^{F}\, p_{jn}) + (P_{jk}^{F}\, p_{jn})/$

$p_{kn} - 2]$ exceeds 20 in the ICP 2005 data for Africa. Hence there is no need to impose such an upper limit in (6).

3. PRICE AND QUANTITY RELATIVES AND DETECTION OF OUTLIERS

Price and quantity relatives are useful diagnostic tools for detecting outliers (and perhaps errors) in the basic heading data. Here, an upper price relative UPR_{jn} for basic heading n in country j is defined as the geometric mean of the price relatives for heading n between country j and each other country in the comparison, with the country j price in the denominator. A lower price relative LPR_{jn} for basic heading n in country j is simply the reciprocal of UPR_{jn}. That is, now the country j price is in the numerator. The upper quantity relative UQR_{jn} and lower quantity relative LQR_{jn} for basic heading n in country j is defined in an analogous way. Again, for quantity relatives, the complication arises that either q_{jn} or q_{kn} may equal zero, or the ratio qkn/qjn may be negative. In such cases, this particular element is set to 1 in the UQR_{jn} and LQR_{jn} formulas. This ensures that UQR_{jn} and LQR_{jn} are both defined, while at the same time not drawing unnecessary attention to these observations. A quantity of zero is clearly problematic and should be checked anyway. Negative quantity relatives are only observed for balancing items. These should also always be carefully scrutinized in each round of ICP.

$$\text{Lower price relative: } UPR_{jn} = \prod_{k=1}^{K} [p_{kn}/(P_{jk}^{F} p_{jn})]^{1/K};$$

$$\text{Upper price relative: } LPR_{jn} = \prod_{k=1}^{K} [(P_{jk}^{F} p_{jn})/p_{kn}]^{1/K};$$

$$\text{Lower quantity relative: } UQR_{jn} = \prod_{k=1}^{K} [q_{kn}^{*}/(Q_{jk}^{F} q_{jn}^{*})]^{1/K};$$

$$\text{Upper quantity relative: } LQR_{jn} = \prod_{k=1}^{K} [(Q_{jk}^{F} q_{jn}^{*})/q_{kn}^{*}]^{1/K};$$

where $q_{kn}^{*}/q_{jn}^{*} = q_{kn}/q_{jn}$ when $q_{kn}/q_{jn} > 0$;

$q_{kn}^{*}/q_{jn}^{*} = Q_{jk}^{F}$ otherwise.

Also, P_{jk}^{F} and Q_{jk}^{F} again denote Fisher price and quantity indexes, respectively, as defined in (3).

An unusually large upper price relative UPR_{jn} suggests that pjn may be too high, while an unusually large lower price relative LPR_{jn} suggests that pjn may be too low. For example, suppose $UPR_{jn} = 10$. This implies that the price of heading n is, on average, 10 times higher in country j than in the other countries in the comparison.

Such a large difference is probably attributable to an error in the calculation of p_{jn}. Similarly, an unusually large upper quantity relative UQR_{jn} suggests that q_{jn} may be too high, while an unusually large lower quantity relative LQR_{jn} suggests that q_{jn} may be too low. Inclusion of bilateral quantity indexes in UQR_{jn} and LQR_{jn} implies that they adjust for differences in the economic size of countries. Hence, when $UQR_{jn} = 10$, this means that the per capita quantity consumed of heading n in country j is, on average, 10 times larger than in the other countries in the comparison (thus, again strongly suggesting an error in the calculation of q_{jn}).

In the empirical analysis that follows, this methodology is applied to the ICP 2005 data for Africa. The country-basic headings jn with the largest upper and lower price and quantity relatives are identified. A number of these clearly warrant closer scrutiny.

4. DENDROGRAMS AND CLUSTERS

The literature on cluster analysis methods is extensive (for example, Everitt, Lander, Leese and Stahl 2011). Much of this literature assumes that the observations over which the clusters are to be formed can be represented as points in an N dimensional space (which is often Euclidean). Dissimilarity matrices and dendrograms can be constructed for such data sets if desired. However, this is just one of a number of approaches that can be used to construct clusters on such data sets.

Our starting point, by contrast, is a dissimilarity matrix, without any corresponding representation in an N dimensional space. This narrows the range of options somewhat. The nature of the data lends itself to an agglomerative approach to cluster formation using dendrograms. Agglomerative methods start with each observation (in this case, a country) as an individual cluster. At each step, the closest pair of clusters is merged until all the observations have been merged into a single cluster. This requires a criterion for measuring cluster proximity. Two widely used criteria are the nearest neighbor and group average. The former, at each step, seeks out the pair of observations in different clusters with the smallest dissimilarity measure, and then merges

them. The latter compares all possible pairs of observations in two clusters and then calculates the average of their dissimilarity measures. This average distance AD_{JK} between a pair of clusters J and K, is calculated as follows:

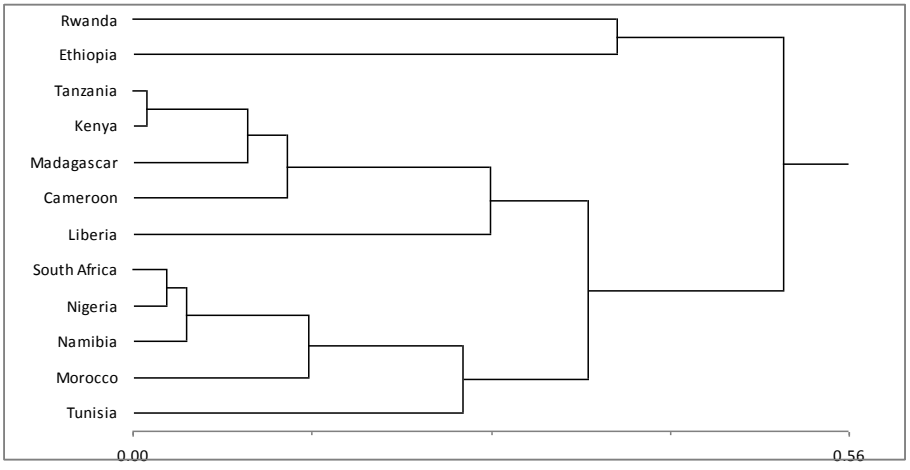
$$AD_{JK} = (1/N_J N_K) \sum_{j-1}^{N_J} \sum_{k-1}^{N_K} d_{jk};$$

where N_J and N_K denote the number of observations in clusters J and K, and j and k index the observations in each cluster. These averages are then compared across all possible pairs of clusters. The pair with the smallest average is then merged. In our context, the group average criterion is more appropriate, since the objective is to identify groups of countries with collectively similar relative price and/or quantity structures.

The resulting pattern of cluster formation can be described by a dendrogram. A dendrogram is a tree-like diagram which shows the order in which the clusters were merged. The lengths of the branches in the dendrogram measure the distance between clusters when they were linked. A simple example consisting of 12 countries from the Africa region in ICP 2005 is provided below in Figure 1. The dissimilarity matrix in this example is computed over the full set of basic headings using the Paasche-Laspeyres spread dissimilarity measure d_{jk}^{PLS}, and the dendrogram is calculated using the group average criterion.

Figure 1. Clustering strategy and dendrogram for 12-country example

Clustering strategy			
Cluster	**1st item**	**2nd item**	**Distance**
1	Tanzania	Kenya	0.011
2	South Africa	Nigeria	0.026
3	Cluster 2	Namibia	0.042
4	Cluster 1	Madagascar	0.089
5	Cluster 4	Cameroon	0.120
6	Cluster 3	Morocco	0.136
7	Cluster 6	Tunisia	0.257
8	Cluster 5	Liberia	0.277
9	Cluster 8	Cluster 7	0.354
10	Rwanda	Ethiopia	0.376
11	Cluster 10	Cluster 9	0.506



The clustering strategy table in Figure 1 gives the order in the which the clusters are merged, starting from an initial situation in which each country is its own cluster. The table also provides the average distance between elements of a cluster. A list of clusters can be derived from a dendrogram by specifying a critical threshold for the average distance between elements of a cluster. Once this threshold is reached, the algorithm terminates. For example, if the threshold in Figure 1 is set to 0.01, there are 12 clusters (i.e., each country is its own cluster). This is because even if we try to form a cluster between the two countries with the smallest dissimilarity measure—in this case, Tanzania and Kenya—the average distance of 0.011 will surpass the allowed critical threshold of 0.01.

By contrast, if the threshold is set to 0.51, there is only one cluster containing all 12 countries. Clearly these two extreme cases defeat the whole objective of the exercise. For the results to be of any use, it is necessary to obtain at least 2 clusters and fewer than 12. Suppose the average distance threshold is set to 0.25; now we obtain 6 clusters. The constituent countries in each cluster are as follows:

Cluster 1: Rwanda;
Cluster 2: Ethiopia;
Cluster 3: Tanzania, Kenya, Madagascar, Cameroon;
Cluster 4: Liberia;
Cluster 5: South Africa, Nigeria, Namibia, Morocco;
Cluster 6: Tunisia.

If the threshold is raised to 0.3, the number of clusters falls to 4, with the constituent countries being as follows:

Cluster 1: Rwanda;
Cluster 2: Ethiopia;
Cluster 3: Tanzania, Kenya, Madagascar, Cameroon, Liberia;
Cluster 4: South Africa, Nigeria, Namibia, Morocco, Tunisia.

An important implication of this example, therefore, is that a dendrogram itself does not indicate how many clusters there are. The process of cluster formation involves a subjective element, the choice of the critical threshold, which should be decided only after the dendrogram has been constructed.

In an ICP context, a rigid threshold should probably not be applied across the whole dendrogram, because such an approach is too inflexible. Rather, the critical threshold should be allowed to vary in different parts of the dendrogram, subject to the constraint that the resulting set of clusters is non-overlapping and exhaustive (i.e., each country is included in only one cluster).

To see why, again consider Figure 1. The ultimate objective is to obtain a list of clusters that are useful in an ICP context. A problem with applying a rigid threshold to the whole dendrogram is that it may generate too many singleton clusters. Singleton clusters are not helpful if the goal is to identify groups of countries that may usefully be compared as part of the data validation process. Visual inspection of Figure 1 suggests that in an ICP context it might be best to group the countries into 3 clusters:

Cluster 1: Rwanda, Ethiopia;
Cluster 2: Tanzania, Kenya, Madagascar, Cameroon, Liberia;
Cluster 3: South Africa, Nigeria, Namibia, Morocco, Tunisia.

It is not possible to obtain this arrangement of clusters using a single average distance threshold. It can, however, be obtained if, starting from the top of the dendrogram, the critical threshold for cluster 1 is, say, 0.4, and thereafter, the critical threshold is, say, 0.3.

In the empirical examples that follow, dendrograms are calculated using the full sample of 48 countries participating in Africa in ICP 2005. In some cases, suggested groupings of countries derived from these dendrograms using a flexible group average distance threshold are also provided.

5. AN APPLICATION TO ICP 2005 DATA FOR THE AFRICA REGION

5.1 The ICP 2005 data set for the Africa region

A total of 48 countries from the Africa region participated in ICP 2005. The ICP 2005 global comparison was made over 129 basic headings. A basic heading is the lowest level of aggregation at which expenditure data are available. A basic heading consists of a group of similar products defined within a general product classification. In the overall global comparison, food and non-alcoholic beverages account for 29 headings, alcoholic beverages and tobacco for 4 headings, clothing and footwear for 5 headings, etc. (Blades 2007). The basic heading price indexes, which in ICP 2005 in the Africa region were calculated using the Country-Product-Dummy (CPD) method, together with their corresponding expenditure data, provide the building blocks from which the overall comparison is constructed.

The prices of all headings in ICP 2005 are expressed relative to those in the United States. That is, for all headings, the US price is normalized to 1. The price data have no gaps; that is, a price is available for all 129 headings for every country). Furthermore, all the prices are strictly positive.

The expenditure data for each basic heading in each country are expressed in units of domestic currency. In the Africa region, the expenditure level is often set to zero for 14 of the 129 basic headings. Typically, this is because of a failure of measurement rather than because the products in that heading are not actually available. Also, sometimes expenditure is negative. In all but one case, this is because the basic heading in question is a balancing item (e.g., 111300 Net purchases abroad, 130225 Receipts from sales, 130425 Receipts from sales, 140115 Receipts from sales, 160000 Change in inventories and valuables, and 180000 Balance of exports and imports). Expenditure on balancing items can legitimately be negative. The exception is 111261 FISIM, which is negative in Niger even though FISIM is not a balancing item.

As explained above, the presence of basic headings with zero or negative expenditures complicates computation of the dissimilarity measures used in the empirical analysis that follows. The inputs to the dissimilarity measures are prices and quantities. The quantities are obtained by dividing expenditure by price. A zero expenditure implies a zero quantity, while a negative expenditure implies a negative quantity and a negative expenditure share.

As explained in section 2.2, the dissimilarity measure formula must be modified to accommodate such cases.

5.2 Outliers identified by the upper and lower price and quantity relatives

The 25 largest outliers among the country-basic headings jn in terms of the upper and lower price relatives UPR_{jn} and LPR_{jn} are shown in Table 1. The largest UPR_{jn} of 6.27 is for the basic heading "Education" for Mauritius. This implies that the price of "Education" is, on average, 6.27 times greater in Mauritius than in the other 47 countries in the comparison. Such a large difference is probably attributable to an error in the calculation of the price of "Education" in Mauritius. Certainly this price, and indeed, all the prices identified in Table 1, should be checked. A good rule of thumb might be that all country-basic heading prices pjn with either upper or lower price relatives UPR_{jn} and LPR_{jn} that exceed 2 require checking.

The lower price relatives LPR_{jn} are generally larger than the upper price relatives UPR_{jn}. This may be because erroneously low prices are less conspicuous than erroneously high prices, and hence, the former are less likely to be detected during data validation.

Two countries stand out in Table 1. Djibouti appears four times in the list of the 25 largest upper price relatives UPR_{jn}. Even more significantly, three of the five largest lower price relatives LPR_{jn} relate to "compensation of employees" in Chad, and need checking.

Table 1. Extreme Upper and Lower Price Relatives

25 Most Extreme Upper Price Relatives		
Mauritius	111000 Education	6.27
Sudan	110621 Medical Services	5.93
Botswana	110724 Other services in respect of personal transport equipment	5.37
Zimbabwe	110613 Therapeutical appliances and equipment	4.77
Gabon	110915 Repair of audio-visual, photographic and information processing equipment	4.62
Guinea	110623 Paramedical services	4.36
Comoros	1102121 Wine	3.97
Ethiopia	111000 Education	3.95
Djibouti	110622 Dental services	3.86
Egypt	111220 Prostitution	3.81
Zambia	110623 Paramedical services	3.80
Malawi	110723 Maintenance and repair of personal transport equipment	3.74
Djibouti	111211 Hairdressing salons and personal grooming establishments	3.73
Rwanda	111231 Jewellery, clocks and watches	3.71
Djibouti	111220 Prostitution	3.70
Benin	110613 Therapeutical appliances and equipment	3.66
Cape Verde	110915 Repair of audio-visual, photographic and information processing equipment	3.65
Angola	110622 Dental services	3.61
Liberia	140111 Compensation of employees	3.56
Niger	110960 Package holidays	3.50
Djibouti	110621 Medical Services	3.48
Namibia	110512 Carpets and other floor coverings	3.48
Burundi	1101183 Confectionery, chocolate and ice cream	3.36
Burundi	110512 Carpets and other floor coverings	3.36
Congo, Rep.	110613 Therapeutical appliances and equipment	3.35
25 Most Extreme Lower Price Relatives		
Chad	140111 Compensation of employees	12.29
Chad	130421 Compensation of employees	12.06

Lesotho	110623 Paramedical services	10.50
Ghana	110410 Actual and imputed rentals for housing	7.98
Chad	130221 Compensation of employees	7.48
Zimbabwe	1105621 Domestic services	6.31
Mauritius	110810 Postal services	5.39
Burundi	110810 Postal services	5.16
Swaziland	110623 Paramedical services	4.96
Gambia, The	110622 Dental services	4.95
Ethiopia	1105621 Domestic services	4.56
Central African Republic	110623 Paramedical services	4.56
Tunisia	110440 Water supply and miscellaneous services relating to the dwelling	4.55
Guinea	1105621 Domestic services	4.30
Swaziland	110621 Medical Services	4.26
Zambia	110440 Water supply and miscellaneous services relating to the dwelling	4.24
Equatorial Guinea	110623 Paramedical services	4.16
Liberia	110941 Recreational and sporting services	4.14
Morocco	110452 Gas	4.05
Madagascar	111000 Education	4.01
Central African Republic	110941 Recreational and sporting services	3.99
Comoros	110613 Therapeutical appliances and equipment	3.92
Zimbabwe	130421 Compensation of employees	3.89
Mauritius	110820 Telephone and telefax equipment	3.80
Egypt	110736 Other purchased transport services	3.79

The 25 largest outliers among the country-basic headings jn in terms of the upper and lower quantity relatives UQR_{jn} and LQR_{jn} are shown in Table 2. The most striking aspect of Table 2 is how much larger the quantity relatives are than their corresponding price relatives in Table 1. The upper quantity relatives are larger than the upper price relatives by a factor of about 10, and the lower quantity relatives are larger than the lower price

relatives by a factor of about 200. These large quantity relatives cannot be attributed to differences in the economic size of countries, since as noted above, UQR_{jn} and LQR_{jn} include bilateral quantity indexes that adjust for such differences. A value of LQR_{jn} of 1,000 (the ten largest lower quantity relatives in Table 2 exceed this value) implies that the per capita quantity consumed of this heading n is, on average, a 1,000 times lower in country j than in the other countries in the comparison. Except for balancing items (see below), or perhaps exceptionally, for alcoholic beverages such as "Beer," such results seem implausible.

Table 2. Extreme Upper and Lower Quantity Relatives

25 Most Extreme Upper Quantity Relatives		
Lesotho	110623 Paramedical services	89.28
Malawi	110915 Repair of audio-visual, photographic and information processing equipment	69.76
Zambia	110935 Veterinary and other services for pets	54.12
Zambia	110915 Repair of audio-visual, photographic and information processing equipment	50.10
Egypt	1101143 Cheese	44.37
Chad	1101143 Cheese	43.29
Zambia	110941 Recreational and sporting services	36.01
Niger	110933 Gardens and pets	35.27
Cape Verde	110452 Gas	33.41
Kenya	110935 Veterinary and other services for pets	33.20
Malawi	110724 Other services in respect of personal transport equipment	32.08
Ethiopia	1101173 Frozen or preserved vegetables	31.05
South Africa	110921 Major durables for outdoor and indoor recreation	30.08
Zambia	110442 Miscellaneous services relating to the dwelling	29.45
Zambia	110931 Other recreational items and equipment	29.07
Guinea-Bissau	1101151 Butter and margarine	28.75
South Africa	1101182 Jams, marmalades and honey	28.74
Liberia	110533 Repair of household appliances	28.65
Morocco	110452 Gas	23.88
Guinea-Bissau	110915 Repair of audio-visual, photographic and information processing equipment	23.71

São Tomé and Principe	1101143 Cheese	23.22
Kenya	110915 Repair of audio-visual, photographic and information processing equipment	23.21
Tunisia	111120 Accommodation services	22.77
Rwanda	1102131 Beer	22.43
Ethiopia	110513 Repair of furniture, furnishings and floor coverings	22.40
25 Most Extreme Lower Quantity Relatives		
Ghana	160000 Change in inventories and valuables	76241.77
Congo, Dem. Rep.	110935 Veterinary and other services for pets	2297.18
Senegal	110532 Small electric household appliances	2087.73
Botswana	110921 Major durables for outdoor and indoor recreation	1564.47
Egypt	1102121 Wine	1431.78
Morocco	1101122 Pork	1339.61
Mozambique	111250 Insurance	1288.77
Tanzania	111250 Insurance	1176.71
Ethiopia	1101115 Pasta products	1057.09
Senegal	110531 Major household appliances whether electric or not	1026.43
Sudan	1102131 Beer	863.41
Chad	110611 Pharmaceutical products	801.17
Chad	110453 Other fuels	794.61
Malawi	111270 Other services n.e.c.	725.50
Egypt	1102131 Beer	635.80
Angola	1103111 Clothing materials and accessories	567.49
Congo, Dem. Rep.	110810 Postal services	507.39
Ethiopia	110712 Motor cycles	411.57
Angola	1101183 Confectionery, chocolate and ice cream	402.82
Angola	110512 Carpets and other floor coverings	294.51
Gambia	110551 Major tools and equipment	258.67
Botswana	111262 Other financial services n.e.c	242.11
Niger	1101122 Pork	233.77
Chad	110451 Electricity	208.66
Sudan	1101132 Preserved fish and seafood	197.40

As with the price data, the fact that the lower quantity relatives LQR_{jn} are larger than the upper quantity relatives UQR_{jn} is probably attributable to erroneously low quantities being less conspicuous than erroneously high quantities. From Table 2 it is clear that it is especially important that data validation focus on implausibly low quantities. Ironically, the largest of the lower quantity relatives in Table 2—"Change in inventories and valuables" for Ghana—is not necessarily an error, since balancing items can legitimately equal zero. However, this entry is the only balancing item heading in the list of the 25 largest lower quantity relatives in Table 2. It is highly likely that the other 24 quantities listed in the lower half of Table 2 (with the possible exception of "Beer") are too low by a factor of at least 100.

For the lower quantity relatives, countries that stand out are Angola and Chad (each of which appears three times in the list of the top 25). For the upper quantity relatives, the country that stands out is Zambia (which appears five times in the list of the top 25).

In summary, Table 2 reveals that the outliers in the quantity data are far more pronounced than in the price data. It is, therefore, important that more attention be focused on the validation of the quantity (i.e., expenditure) data in ICP 2011, particularly, low quantities.

5.3 Dendrograms and clusters for Africa in ICP 2005

Dendrograms are constructed here based on the price relative, quantity relative and Paasche-Laspeyres spread dissimilarity measures discussed in sections 2.1 and 2.2. These dendrograms are calculated over different combinations of basic headings. The combinations considered are as follows:

All headings (GDP)
Group 1 headings (Final consumption expenditure by households)
Group 2 headings (Individual consumption expenditure by government)
Group 3 headings (Collective consumption expenditure by government)
Group 4 headings (Expenditure on gross fixed capital formation)
Group 5 headings (Changes in inventories and acquisitions, less disposals of valuables)
Group 6 headings (Balance of exports and imports)

An immediate consideration when computing d_{jk}^{P}, d_{jk}^{Q} and d_{jk}^{PQ} dissimilarity measures for subcomponents of GDP is whether the Fisher price and quantity indexes, P_{jk}^{F} and Q_{jk}^{F}, within these formulas should be defined on GDP or on the same subcomponents of GDP as the dissimilarity measure itself. It

is probably preferable to use price and quantity indexes defined on GDP regardless of which headings the dissimilarity measure is computed over.

In particular, in the limiting scenario where there is only one heading, as is the case in Groups 5 and 6 in ICP 2005, the dissimilarity measures d_{jk}^{P}, d_{jk}^{Q} and d_{jk}^{PQ} all collapse to zero. This is because when P_{jk}^{F} and Q_{jk}^{F} are calculated over just this one heading, by construction, P_{jk}^{F} = pkn/pjn and $Q_{jk}^{F} = q_{kn}/q_{jn}$. The same happens to d_{jk}^{PLS}, because, likewise, $P_{jk}^{P} = P_{jk}^{L} = p_{kn}/p_{jn}$ in this case. This problem does not arise, however, in Groups 5 and 6 when the price and quantity indexes P_{jk}^{F} and Q_{jk}^{F} in d_{jk}^{P} and d_{jk}^{Q} are always defined on GDP.

However, another problem arises for the dissimilarity measure d_{jk}^{Q} for Groups 5 and 6. As a result of these headings both being balancing items, many quantities are negative. This limits the usefulness of computing d_{jk}^{Q} (and likewise, d_{jk}^{PQ}). Hence, the only dendrograms considered here for Groups 5 and 6 are calculated using the dissimilarity measure djkP. For the headings in GDP and Groups 1 to 4, however, dendrograms are presented using all four dissimilarity measures (i.e., d_{jk}^{P}, d_{jk}^{Q}, d_{jk}^{PQ} and d_{jk}^{PLS}). The results (dendrograms and resulting clusters) for the headings in GDP are shown in Figure 2a. Dendrograms have also been calculated for Groups 1, 2, 3, 4, 5, and 6. To save space, these dendrograms are not presented here.

Country clusters obtained from price relative dissimilarity dendrogram (GDP) in Figure 2a

Cluster 1
Zimbabwe
Chad
Cluster 2
Rwanda
Burundi
Madagascar
Zambia
Ethiopia
Egypt, Arab Rep.
Cluster 3
Mali
Benin
Niger
Burkina Faso
Cameroon
Uganda
Tanzania
Togo
Kenya
Central African Republic
Mozambique
Côte d'Ivoire
Sudan
Lesotho
Botswana
Angola
Senegal
Djibouti

Cluster 4
Sierra Leone
Nigeria
Mauritania
Guinea-Bissau
Ghana
Gambia, The
Guinea
Malawi
Cluster 5
Gabon
Comoros
Congo, Dem. Rep.
Equatorial Guinea
Congo, Rep.
São Tomé and Principe
Cluster 6
Liberia
Cluster 7
Tunisia
Mauritius
South Africa
Morocco
Cape Verde
Swaziland
Namibia

Figure 2a. Dendogram for Price Dissimilarity Matrix (GDP)

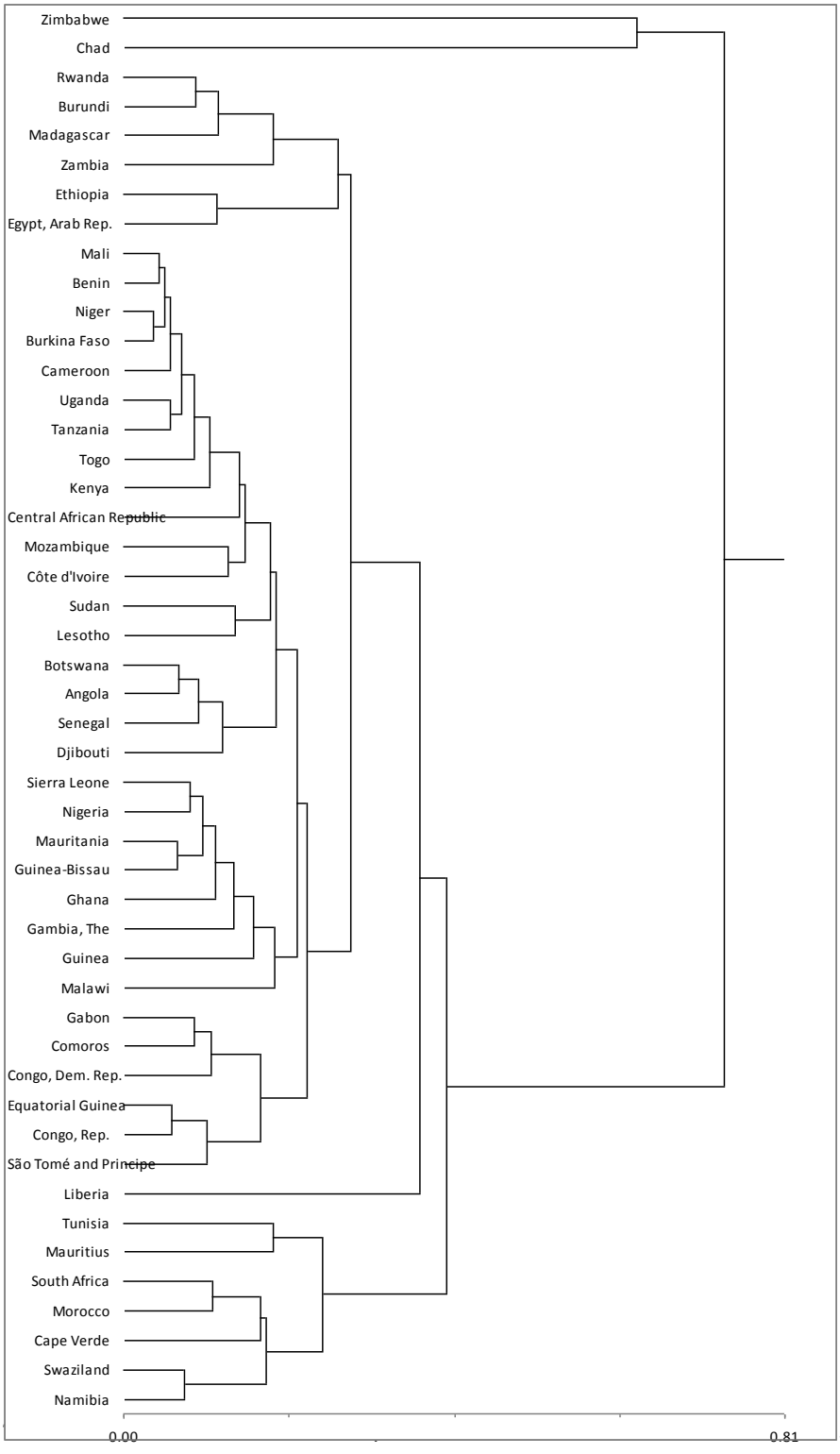
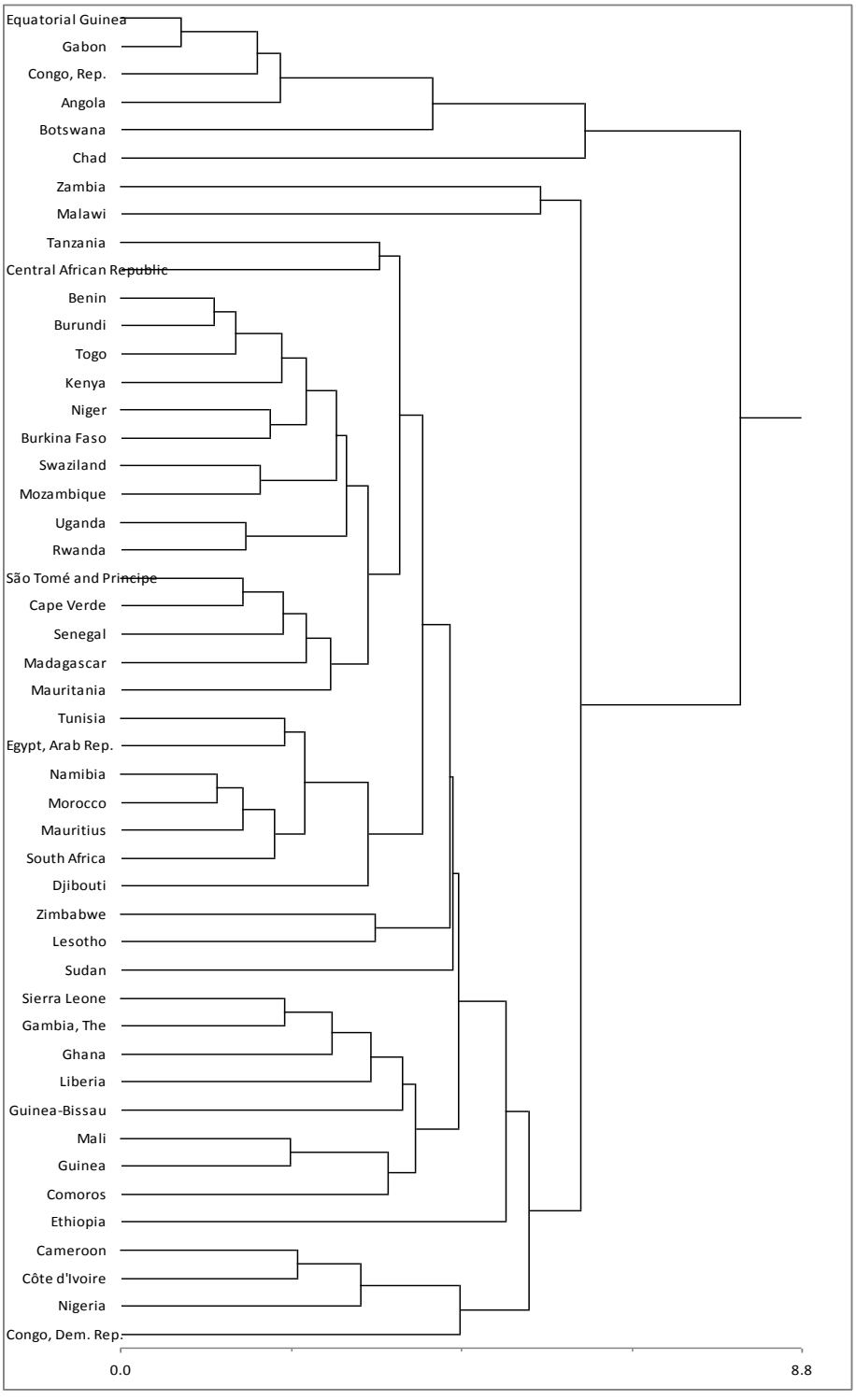


Figure 2(b) Dendrogram for quantity dissimilarity matrix (GDP)

Country clusters obtained from quantity relative dissimilarity dendrogram (GDP) in Figure 2b
Cluster 1
Equatorial Guinea
Gabon
Congo, Rep.
Angola
Botswana
Chad
Cluster 2
Zambia
Malawi
Cluster 3
Tanzania
Central African Republic
Cluster 4
Benin
Burundi
Togo
Kenya
Niger
Burkina Faso
Swaziland
Mozambique
Uganda
Rwanda
Cluster 5
São Tomé and Principe
Cape Verde
Senegal
Madagascar
Mauritania
Cluster 6
Tunisia
Egypt, Arab Rep.
Namibia
Morocco
Mauritius
South Africa
Djibouti
Cluster 7
Zimbabwe
Lesotho
Cluster 8
Sudan
Cluster 9
Sierra Leone
Gambia, The
Ghana
Liberia
Guinea-Bissau
Mali
Guinea
Comoros
Cluster 10
Ethiopia
Cluster 11
Cameroon
Côte d'Ivoire
Nigeria
Congo, Dem. Rep.

Figure 2(b) Dendogram for quantity dissimilarity matrix (GDP)



Country clusters obtained from geometric mean of price relatives and quantity relatives dissimilarity dendrogram (GDP) in Figure 2c

Cluster 1
Liberia
Côte d'Ivoire
Cluster 2
Niger
Burkina Faso
Mali
Togo
Benin
Cameroon
Tanzania
Madagascar
Burundi
Uganda
Rwanda
Guinea
Central African Republic
Cluster 3
Senegal
Mauritania
São Tomé and Principe
Guinea-Bissau
Sierra Leone
Ghana
Gambia, The
Cluster 4
Mozambique
Kenya
Lesotho
Sudan
Cluster 5
Congo, Dem. Rep.
Comoros

Cluster 6
Djibouti
Cluster 7
Ethiopia
Egypt, Arab Rep.
Cluster 8
Tunisia
Morocco
Mauritius
Swaziland
Namibia
South Africa
Cape Verde
Cluster 9
Zambia
Cluster 10
Equatorial Guinea
Gabon
Congo, Rep.
Nigeria
Botswana
Angola
Cluster 11
Malawi
Cluster 12
Zimbabwe
Cluster 13
Chad

Figure 2(c) Dendogram for price-quantity dissimilarity matrix (GDP)

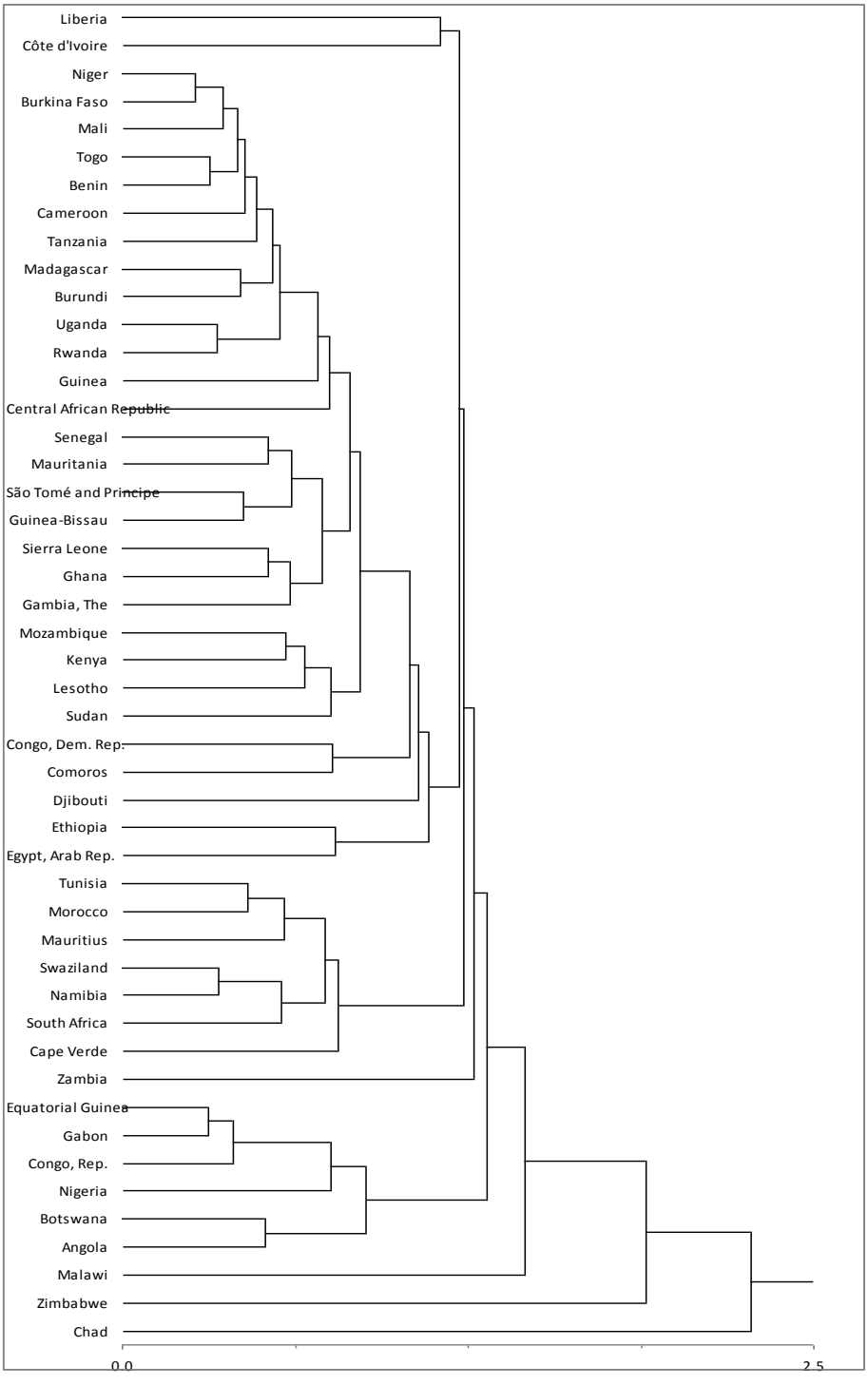
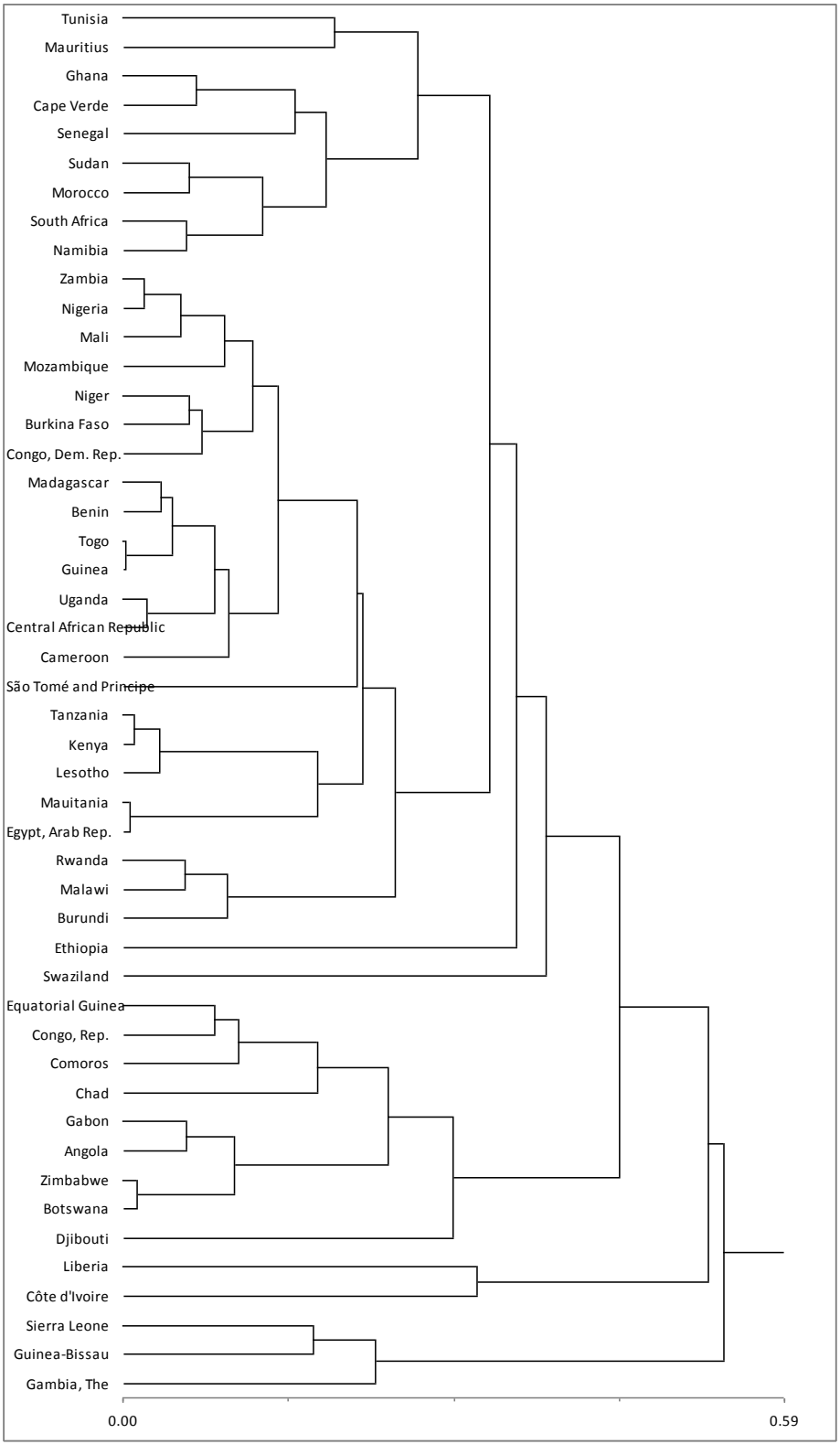


Figure 2d. Dendrogram for Paasche-Laspeyres spread matrix (GDP)

Country clusters obtained from Paasche-Laspeyres spread dissimilarity dendrogram (GDP) in Figure 2d
Cluster 1
Tunisia
Mauritius
Ghana
Cape Verde
Senegal
Sudan
Morocco
South Africa
Namibia
Cluster 2
Zambia
Nigeria
Mali
Mozambique
Niger
Burkina Faso
Congo, Dem. Rep.
Madagascar
Benin
Togo
Guinea
Uganda
Central African Republic
Cameroon
São Tomé and Principe
Cluster 3
Tanzania
Kenya
Lesotho
Mauritania
Egypt, Arab Rep.
Cluster 4
Rwanda
Malawi
Burundi
Cluster 5
Ethiopia
Cluster 6
Swaziland
Cluster 7
Equatorial Guinea
Congo, Rep.
Comoros
Chad
Gabon
Angola
Zimbabwe
Botswana
Djibouti
Cluster 8
Liberia
Côte d'Ivoire
Cluster 9
Sierra Leone
Guinea-Bissau
Gambia, The

Figure 2d. Dendogram for Paasche-Laspeyres Spread Matrix (GDP)



A comparison of dendrograms across the price relative, quantity relative and Paasche-Laspeyres spread dissimilarity measures and the various groups reveals that they are not robust to changes in the dissimilarity measure formula or data.

Also presented in Figure 2 is a possible configuration of clusters derived from each dendrogram. As noted previously, the choice of clusters for a given dendrogram has a subjective element. Here, configurations have been chosen consisting of 6 to 13 clusters, depending on how the dendrogram is structured. The clusters can be made broader or narrower to suit the needs of users. For example, on the dendrogram for d_{jk}^{P} in Figure 2a, some possible configurations of clusters are listed below:

Three clusters:

Cluster 1: Zimbabwe, Chad;
Cluster 2: Rwanda, Burundi, Madagascar, Zambia, Ethiopia, Egypt Arab Rep., Mali, Benin, Niger, Burkina Faso, Cameroon, Uganda, Tanzania, Togo, Kenya, Central African Republic, Mozambique, Côte d'Ivoire, Sudan, Lesotho, Botswana, Angola, Senegal, Djibouti, Sierra Leone, Nigeria, Mauritania, Guinea-Bissau, Ghana, The Gambia, Guinea, Malawi, Gabon, Comoros, Congo Dem. Rep., Equatorial Guinea, Congo Rep., São Tomé and Principe, Liberia;
Cluster 3: Tunisia, Mauritius, South Africa, Morocco, Cape Verde, Swaziland, Namibia.

Five clusters:

Cluster 1: Zimbabwe, Chad;
Cluster 2: Rwanda, Burundi, Madagascar, Zambia, Ethiopia, Egypt Arab Rep.;
Cluster 3: Mali, Benin, Niger, Burkina Faso, Cameroon, Uganda, Tanzania, Togo, Kenya, Central African Republic, Mozambique, Côte d'Ivoire, Sudan, Lesotho, Botswana, Angola, Senegal, Djibouti, Sierra Leone, Nigeria, Mauritania, Guinea-Bissau, Ghana, The Gambia, Guinea, Malawi, Gabon, Comoros, Congo Dem. Rep., Equatorial Guinea, Congo Rep., São Tomé and Principe;
Cluster 4: Liberia;
Cluster 5: Tunisia, Mauritius, South Africa, Morocco, Cape Verde, Swaziland, Namibia.

Six clusters:

Cluster 1: Zimbabwe, Chad;
Cluster 2: Rwanda, Burundi, Madagascar, Zambia, Ethiopia, Egypt Arab Rep.;
Cluster 3: Mali, Benin, Niger, Burkina Faso, Cameroon, Uganda, Tanzania, Togo, Kenya, Central African Republic, Mozambique, Côte d'Ivoire, Sudan, Lesotho, Botswana, Angola, Senegal, Djibouti, Sierra Leone, Nigeria, Mauritania, Guinea-Bissau, Ghana, The Gambia, Guinea, Malawi;
Cluster 4: Gabon, Comoros, Congo Dem. Rep., Equatorial Guinea, Congo Rep., São Tomé and Principe;
Cluster 5: Liberia;
Cluster 6: Tunisia, Mauritius, South Africa, Morocco, Cape Verde, Swaziland, Namibia.

Nine clusters:

Cluster 1: Zimbabwe, Chad;
Cluster 2: Rwanda, Burundi, Madagascar, Zambia;
Cluster 3: Ethiopia, Egypt Arab Rep.;
Cluster 4: Mali, Benin, Niger, Burkina Faso, Cameroon, Uganda, Tanzania, Togo, Kenya, Central African Republic, Mozambique, Côte d'Ivoire, Sudan, Lesotho;
Cluster 5: Botswana, Angola, Senegal, Djibouti;
Cluster 6: Sierra Leone, Nigeria, Mauritania, Guinea-Bissau, Ghana, The Gambia, Guinea, Malawi;
Cluster 7: Gabon, Comoros, Congo Dem. Rep., Equatorial Guinea, Congo Rep., São Tomé and Principe;
Cluster 8: Liberia;
Cluster 9: Tunisia, Mauritius, South Africa, Morocco, Cape Verde, Swaziland, Namibia.

Thirteen clusters:

Cluster 1: Zimbabwe, Chad;
Cluster 2: Rwanda, Burundi, Madagascar, Zambia;
Cluster 3: Ethiopia, Egypt Arab Rep.;
Cluster 4: Mali, Benin, Niger, Burkina Faso, Cameroon, Uganda, Tanzania, Togo, Kenya, Central African Republic;
Cluster 5: Mozambique, Côte d'Ivoire;
Cluster 6: Sudan, Lesotho;

Cluster 7: Botswana, Angola, Senegal, Djibouti;
Cluster 8: Sierra Leone, Nigeria, Mauritania, Guinea-Bissau, Ghana, Gambia The, Guinea, Malawi;
Cluster 9: Gabon, Comoros, Congo Dem. Rep.;
Cluster 10: Equatorial Guinea, Congo Rep., São Tomé and Principe;
Cluster 11: Liberia;
Cluster 12: Tunisia, Mauritius;
Cluster 13: South Africa, Morocco, Cape Verde, Swaziland, Namibia.

Some implications for data validation of the dendrograms and clusters identified in Figure 2 are discussed in the next section.

For dissimilarity measures derived from differences in per capita GDP, dendrograms are constructed for the dissimilarity matrix computed using the dissimilarity measure in (4), with per capita GDP computed using GEKS and Iklé. Also considered here is the dendrogram derived from the dissimilarity measure in (5) that measures differences in the price level across countries. These three dendrograms and associated clusters are shown in Figure 3.[4]

4 The price-level dendrogram in Figure 9(c) is calculated from the official ICP 2005 results, which, as noted above, used the Iklé method in the Africa region. Zimbabwe is excluded because no market exchange rate is available for 2005.

Figure 3a. Dendrogram for per Capita GDP calculated using GEKS

Country clusters obtained from GEKS per capita GDP dissimilarity dendrogram in Figure 3a
Cluster 1
Liberia
Burundi
Congo, Dem. Rep.
Cluster 2
Gambia, The
Central African Republic
Malawi
Niger
Sierra Leone
Rwanda
Mozambique
Guinea-Bissau
Ethiopia
Zimbabwe
Cluster 3
São Tomé and Principe
Lesotho
Chad
Senegal
Mauritania
Côte d'Ivoire
Benin
Kenya
Sudan
Cameroon
Nigeria
Djibouti

Cluster 4
Togo
Guinea
Uganda
Madagascar
Tanzania
Mali
Comoros
Zambia
Ghana
Burkina Faso
Cluster 5
Swaziland
Namibia
Egypt, Arab Rep.
Tunisia
Cape Verde
Angola
Congo, Rep.
Morocco
Cluster 6
Mauritius
Equatorial Guinea
Gabon
Botswana
South Africa

Figure 3a. Dendogram for Per Capita GDP Calculated using GEKS

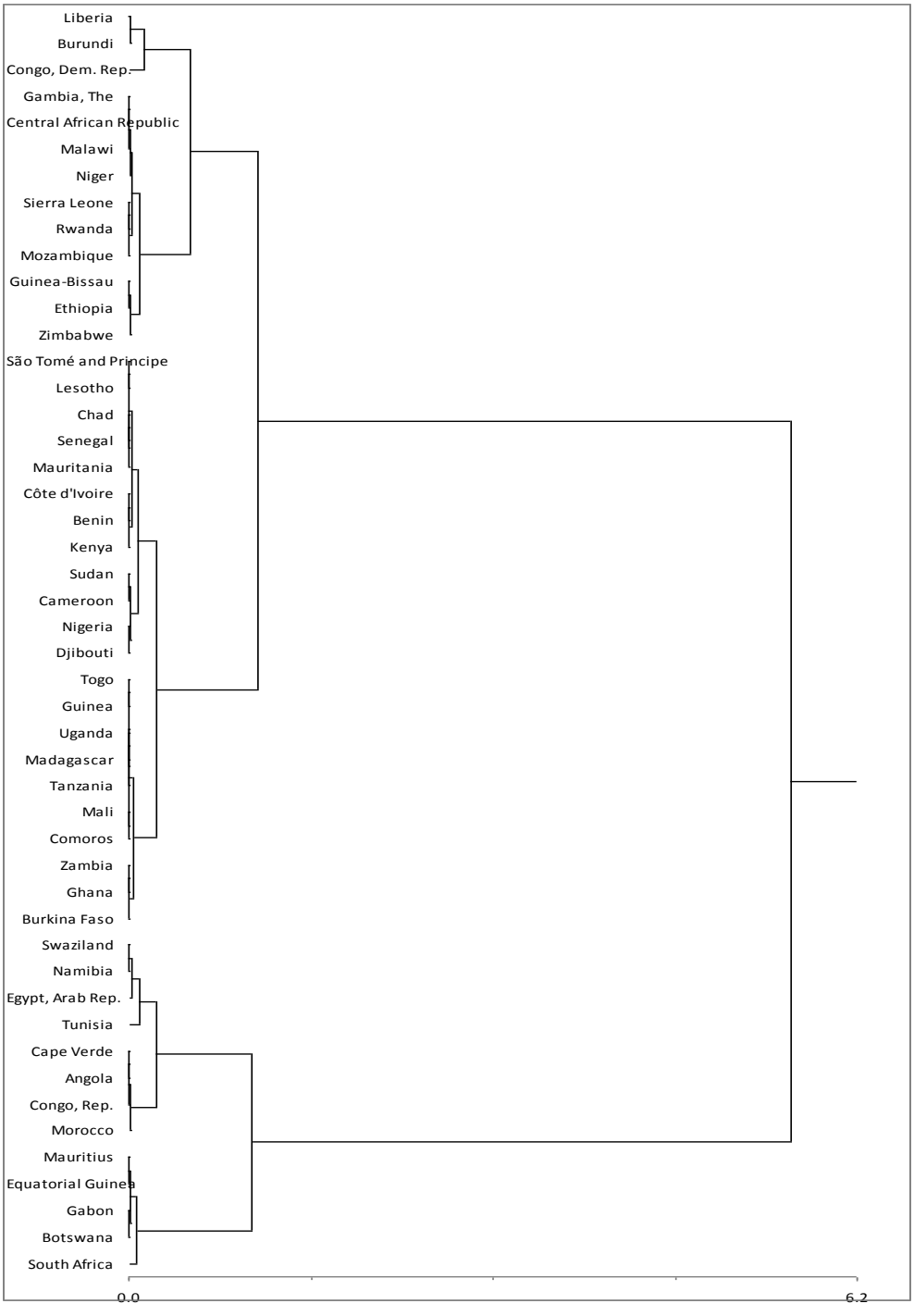


Figure 3b. Dendrogram for per capita GDP calculated using Ikle

Country clusters obtained from Ikle per capita GDP dissimilarity dendrogram in Figure 3b
Cluster 1
Zimbabwe
Guinea-Bissau
Niger
Ethiopia
Burundi
Cluster 2
Malawi
Central African Republic
Mozambique
Gambia, The
Sierra Leone
Rwanda
Cluster 3
Togo
Guinea
Tanzania
Mali
Uganda
Madagascar
Comoros
Zambia
Burkina Faso
Ghana
Cluster 4
Senegal
Mauritania
Chad
Côte d'Ivoire
Lesotho
Benin
Kenya
São Tomé and Principe
Djibouti
Cameroon
Nigeria
Sudan
Cluster 5
Congo, Dem. Rep.
Liberia
Cluster 6
Swaziland
Namibia
Egypt, Arab Rep.
Tunisia
Morocco
Angola
Congo, Rep.
Cape Verde
Cluster 7
South Africa
Mauritius
Equatorial Guinea
Botswana
Gabon

Figure 3b. Dendogram for Per Capita GDP Calculated using Ikle

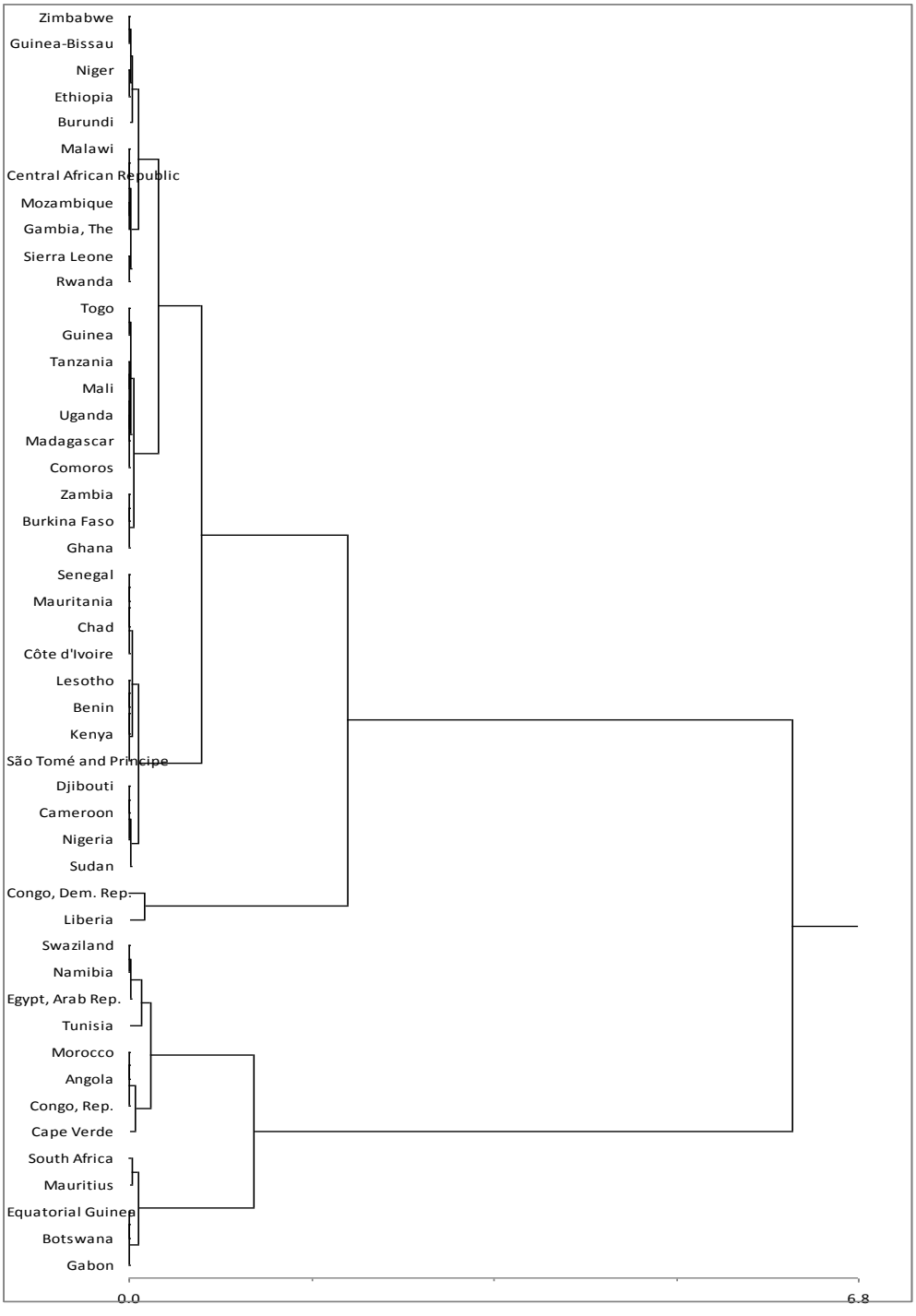
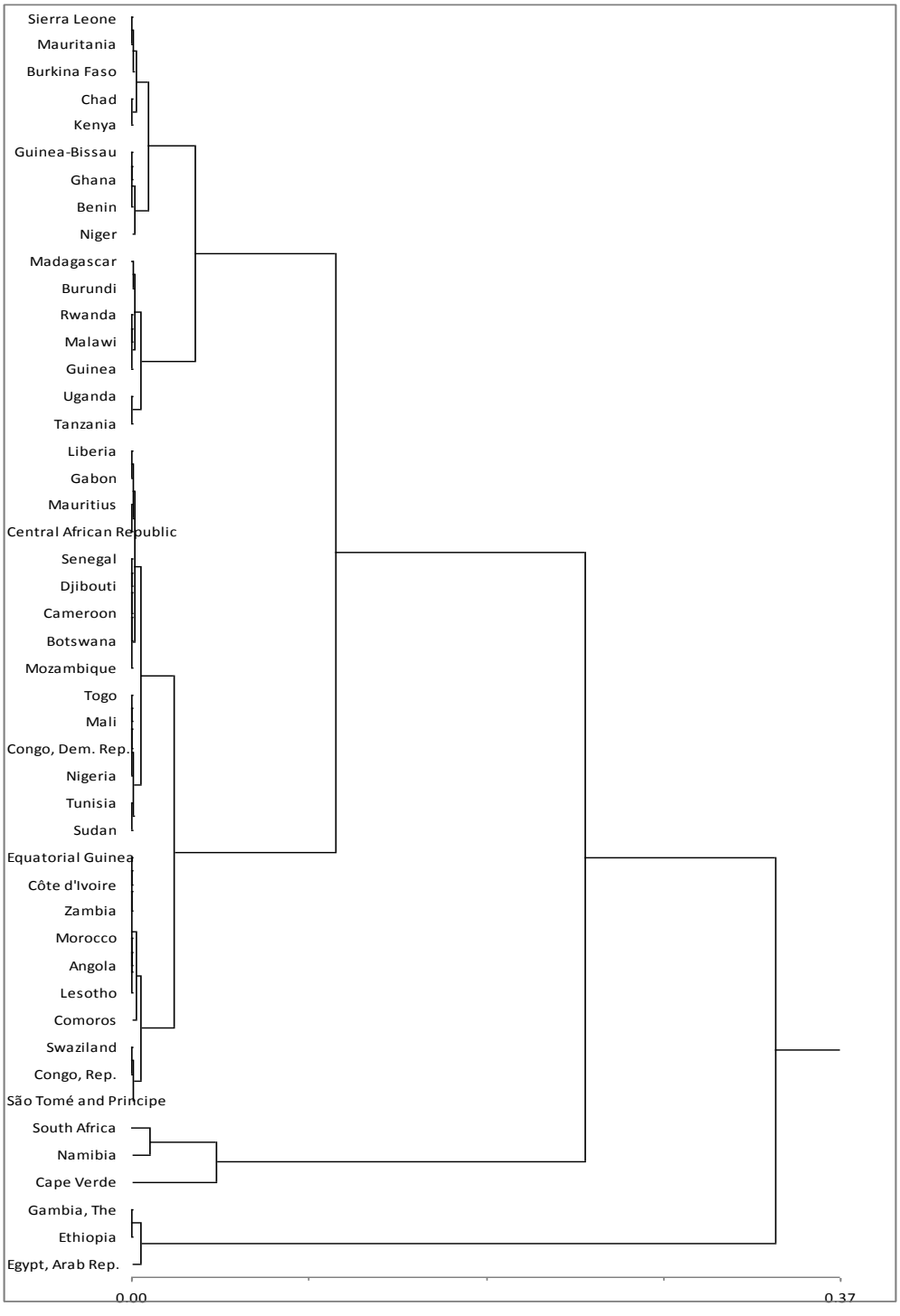


Figure 3c. Dendrogram for country price levels

Country clusters obtained from price level dissimilarity dendrogram in Figure 3c
Cluster 1
Sierra Leone
Mauritania
Burkina Faso
Chad
Kenya
Cluster 2
Guinea-Bissau
Ghana
Benin
Niger
Cluster 3
Madagascar
Burundi
Rwanda
Malawi
Guinea
Uganda
Tanzania
Cluster 4
Liberia
Gabon
Mauritius
Central African Republic
Senegal
Djibouti
Cameroon
Botswana
Mozambique

Cluster 5
Togo
Mali
Congo, Dem. Rep.
Nigeria
Tunisia
Sudan
Cluster 6
Equatorial Guinea
Côte d'Ivoire
Zambia
Morocco
Angola
Lesotho
Comoros
Swaziland
Congo, Rep.
São Tomé and Principe
Cluster 7
South Africa
Namibia
Cape Verde
Cluster 8
Gambia, The
Ethiopia
Egypt, Arab Rep.

Figure 3c. Dendogram for Country Price Levels



When applied to the dissimilarity measure in (4) derived from per capita GDP, the group-average dendrogram begins by ranking the countries by per capita GDP and then divides them into clusters that preserve this order. In

other words, if countries A and B are in a cluster, a country C with a per capita income between that of countries A and B must be in the same cluster.

A comparison of the GEKS and Iklé per capita GDP dendrograms and clusters reveals considerable similarity (the price level dendrogram and clusters, by contrast, are quite different). In particular, cluster 3 under GEKS is identical to cluster 4 under Iklé; cluster 4-GEKS is identical to cluster 3- Iklé; cluster 5-GEKS is identical to cluster 6- Iklé; and cluster 6-GEKS is identical to cluster 7- Iklé. Differences arise only for the lowest per capita GDP countries. Given that the differences between the GEKS and Iklé per capita GDP estimates are nontrivial and alter the ranking of countries substantially (particularly at the lower end), this suggests that dendrograms constructed on per capita GDP attain a certain degree of stability. By implication, the per capita GDP clusters identified using ICP 2005 should still be broadly applicable to ICP 2011 data.

The actual subregions in ICP-Africa are shown in Table 3. These subregions are largely geo-political and do not necessarily make sense from an economic perspective. There is not much overlap between the official ICP-Africa subregions in Table 3 and the clusters in Figures 3a and 3b. The use of the existing ICP-Africa subregions, therefore, may make it harder to validate the ICP data. It would be preferable if these geo-political subregions were replaced in ICP 2011 by clusters derived from a measure of economic similarity such as those in Figures 3a and 3b.

Table 3. Sub-Regional Organization of ICP-Africa

AFRISTAT Group	COMESA Group	ECOWAS Group
Benin	Burundi	Gambia
Burkina Faso	Djibouti	Ghana
Cameroun	Egypt	Liberia
Cap Vert	Eritrea	Nigeria
Central African Republic	Ethiopia	Sierra Leone
Chad	Kenya	
Comoros	Madagascar	
Congo	Rwanda	
Cote d'Ivoire	Sudan	

Congo, Dem. Rep.	Uganda	
Equatorial Guinea		
Gabon		
Guinea Bissau		
Guinea Conakry		
Mali		
Mauritania		
Niger		
São Tomé and Principe		
Senegal		
Togo		
SADC Group	**UMA Group**	**Other African countries**
Angola	Libya	Somalia
Botswana	Morocco	
Lesotho	Tunisia	
Malawi	Algeria	
Mauritius	Mauritania	
Mozambique		
Namibia		
Seychelles		
South Africa		
Swaziland		
Tanzania		
Zambia		
Zimbabwe		

5.4 How cluster analysis can improve data validation

The price and quantity relatives discussed in section 5.2 provide a way of detecting anomalies in the individual basic heading price and quantity data of each country. The dendrograms, by contrast, provide an indication of the reliability of the complete set of headings over which the dendrogram has been constructed. Focusing first on the price data and the dissimilarity measure d_{jk}^{P}, it can be seen from the dendrogram in Figure 2a that Zimbabwe and Chad are clear outliers. This could be either because their price data are of poor quality or because the relative price structures in these two countries are very different from those in the rest of Africa. From a data validation

perspective, the key point is that the ICP regional office for Africa should carefully check the price data of Zimbabwe and Chad for errors.

Some progress can be made in identifying the sources of discrepancies in the price data of Zimbabwe and Chad by looking at the dendrograms generated by d_{jk}^{P} for each of the six constituent groups of GDP. In the dendrogram (not included here) covering only the price data for the headings in Group 1 (i.e. final consumption expenditure by households), Zimbabwe and Chad are no longer outliers. This indicates that the problems lie elsewhere. The dendrogram (also not included) covering the price data for headings in Group 2 (i.e., individual consumption expenditure by government) is revealing in that it now identifies a cluster of four countries with price structures that are very different from those of the other 44 countries in the comparison. The other two countries in this cluster, in addition to Zimbabwe and Chad, are Congo Dem. Rep. and Equatorial Guinea.

The dendrogram (not included) focusing on the price data for headings in Group 3 (i.e., collective consumption expenditure by government) shows that Chad is a huge outlier. This strongly suggests a problem with Chad's price data for these headings.

The dendrogram (not included) focusing on the price data for headings in Group 4 (i.e., expenditure on gross fixed capital formation) is also striking in that Zimbabwe is the main outlier. Malawi, too, is somewhat of an outlier. This dendrogram strongly suggests a problem with Zimbabwe's price data for these headings.

The dendrogram (not included) focusing on the price data in Group 5 (i.e., changes in inventories and acquisitions, less disposals of valuables) does not contain any strong outliers. Finally, Zimbabwe is again a strong outlier in the dendrogram (not included) focusing on the price data in Group 6 (i.e., balance of exports and imports).

In the quantity data at the level of GDP and the dissimilarity measure dQ jk, no clear outliers appear in the dendrogram in Figure 2b. In the dendrogram (not included) focusing on the quantity data for the headings in Group 1 (i.e., final consumption expenditure by households), Malawi, Chad, Ethiopia, Zambia, Zimbabwe and Djibouti are all outliers, although none are extreme outliers.

The dendrogram (not included) focusing on the quantity data for headings in Group 2 (i.e., individual consumption expenditure by government) identifies

Tanzania and Liberia as outliers. In the dendrogram (not included) focusing on the quantity data for headings in Group 3 (i.e., collective consumption expenditure by government), Tanzania and Zambia are the main outliers. In the dendrogram (not included) focusing on the quantity data for headings in Group 4 (i.e., expenditure on gross fixed capital formation), Congo. Dem. Rep., Congo Rep., and Chad are the major outliers. The relevant headings for these countries should probably be checked.

It is also useful to compare the magnitude of the final group average link in the dendrograms (i.e., the link that joins all the countries into a single cluster). The value of the group average at which this link occurs in the dendrograms calculated using the d_{jk}^{P} metric defined over GDP, Group 1, 2, 3, 4, 5 and 6 is, respectively, 0.74, 0.51, 3.23, 4.25, 1.17, 0.07, 1.95. From this it can be deduced that the biggest problems with the price data occur in Groups 2 and 3 (i.e., individual and collective consumption expenditure by government), since this is where the group average is highest when all countries are included in a single cluster. Similarly, the value of the group average in the dendrograms calculated using the d_{jk}^{Q} metric defined over GDP, Group 1, 2, 3, 4, 5 and 6 is, respectively, 8.01, 8.01, 13.4, 5.9, and 6.6, suggesting that the biggest problems in the quantity data arise in Group 2 (i.e., individual consumption expenditure by government). The much larger group averages at the final link in the quantity data dendrograms, compared with the price data dendrograms, are striking. This suggests much greater variability in the quantity relatives across countries than in the price relatives.

The dissimilarity measures d_{jk}^{PQ} and d_{jk}^{PLS}, while potentially useful for detecting outliers, do not indicate whether the problems lie with the price or quantity data. Their findings can also differ substantially. For example, Figure 2c, which constructs a dendrogram using the dissimilarity measure d_{jk}^{PQ} over GDP, identifies Chad and Zimbabwe as outliers, but Figure 2d, which constructs a dendrogram using the dissimilarity measure d_{jk}^{PLS} over GDP, does not. Zimbabwe and Chad, however, do emerge as strong outliers according to both d_{jk}^{PQ} and d_{jk}^{PLS} for Group 2 (i.e., individual consumption expenditure by government).

An alternative use for dendrograms in data validation is to identify clusters of countries with reasonably similar price relative or quantity relative structures. It may then be easier to detect data errors if prices and quantities are compared across countries belonging to the same cluster. For example, in Figure 2b, cluster 4 consists of Benin, Burundi, Togo, Kenya, Niger, Burkina Faso, Swaziland, Mozambique, Uganda, and Rwanda.

Upper and lower quantity relatives for the 10 countries in this cluster are compared in Table 4. With regard to the upper quantity relatives, Kenya (with eight entries) and Niger (with six entries) in the top half of Table 4 stand out as countries whose quantity (i.e., expenditure) data should be checked. No country stands out in the lower half of Table 4 with regard to the lower quantity relatives. Nevertheless, as in Table 2, the lower quantity relatives are larger than the upper quantity relatives.

Table 4. Extreme Upper and Lower Quantity Relatives for Cluster 4 in Figure 2b

25 Most Extreme Upper Quantity Relatives		
Niger	110933 Gardens and pets	27,07
Kenya	110533 Repair of household appliances	24,28
Rwanda	1101173 Frozen or preserved vegetables	18,14
Kenya	110512 Carpets and other floor coverings	17,96
Kenya	110915 Repair of audio-visual, photographic and information processing equipment	17,28
Kenya	110921 Major durables for outdoor and indoor recreation	16,03
Togo	110513 Repair of furniture, furnishings and floor coverings	14,24
Kenya	110935 Veterinary and other services for pets	14,05
Togo	1102111 Spirits	13,46
Kenya	111250 Insurance	12,66
Swaziland	110621 Medical Services	12,50
Mozambique	1101173 Frozen or preserved vegetables	12,06
Mozambique	1101143 Cheese	11,28
Niger	110921 Major durables for outdoor and indoor recreation	11,27
Niger	111232 Other personal effects	10,53
Niger	110513 Repair of furniture, furnishings and floor coverings	10,29
Swaziland	110520 Household textiles	9,53
Kenya	110513 Repair of furniture, furnishings and floor coverings	9,31
Niger	1101115 Pasta products	9,16
Niger	1103141 Cleaning and repair of clothing	8,97
Burkina Faso	1101122 Pork	8,84

Kenya	110612 Other medical products	8,77
Togo	1101115 Pasta products	8,56
Mozambique	110960 Package holidays	8,48
Burkina Faso	110533 Repair of household appliances	8,28
25 Most Extreme Lower Quantity Relatives		
Mozambique	111250 Insurance	1315,42
Niger	1101122 Pork	663,52
Niger	110735 Combined passenger transport	149,30
Uganda	1101142 Preserved milk and milk products	55,13
Swaziland	110533 Repair of household appliances	51,34
Rwanda	1103141 Cleaning and repair of clothing	51,03
Swaziland	110733 Passenger transport by air	46,70
Togo	110452 Gas	41,52
Niger	1102131 Beer	34,58
Benin	140115 Receipts from sales	25,48
Niger	1101173 Frozen or preserved vegetables	24,41
Rwanda	1102111 Spirits	23,36
Mozambique	110943 Games of chance	22,87
Uganda	1101173 Frozen or preserved vegetables	19,89
Mozambique	110915 Repair of audio-visual, photographic and information processing equipment	17,67
Swaziland	110921 Major durables for outdoor and indoor recreation	16,61
Uganda	1101115 Pasta products	15,77
Togo	1101151 Butter and margarine	15,66
Togo	1101143 Cheese	15,64
Kenya	1101115 Pasta products	14,71
Uganda	111232 Other personal effects	14,55
Burkina Faso	110820 Telephone and telefax equipment	14,13
Rwanda	110551 Major tools and equipment	13,33

Rwanda	111120 Accommodation services	12,81
Uganda	1101162 Frozen, preserved or processed fruits	12,79

More generally, applying such an approach to a subset of countries belonging to the same cluster may reveal anomalies in the data that might otherwise be missed if the approach was applied only to the full set of 48 countries.

In this context, it would be useful if appropriate clusters for ICP 2011 could be identified based on the ICP 2005 data. Data validation then could begin before all countries have submitted their data. As well, if the clusters are formed based on unvalidated ICP 2011 data, errors in the data could cause some countries to be placed in the wrong clusters.

A drawback, however, to using the ICP 2005 data to construct country clusters that are then used to validate the ICP 2011 data is that most of the dissimilarity measures considered here do not generate robust dendrograms and clusters. The exception is the dissimilarity measure in (4) that measures differences in per capita GDP, which seems to be reasonably robust to changes in the data. Hence, it is perhaps worth considering for the ICP 2011 data validation process.

Finally, the dendrogram approach to cluster formation often generates one or more singleton clusters (although this does not happen in either the per capita GDP-GEKS or per capita GDP-Iklé dendrograms in Figures 3a and 3b). For example, both Sudan and Ethiopia end up in singleton clusters in the suggested configuration arising out the dendrogram in Figure 2b. Therefore, the clustering approach does not provide guidance about which other countries should be used as points of reference when validating the quantity data of these two countries. A solution is to refer back to the dissimilarity matrix to find the countries closest to Sudan and Ethiopia in their dissimilarity measures. The two countries closest to Sudan, based on djkQ applied at the level of GDP, are Burundi and Benin; the two countries closest to Ethiopia are Mali and Burundi. Hence Sudan's quantity data can be validated in comparison to the quantity data of Burundi and Benin, and Ethiopia's quantity data can be validated in comparison to the quantity data of Mali and Burundi. This process, however, should not be reversed. Burundi's quantity data should be validated using the countries in its cluster, and the same is true for Benin and Mali. With slight modifications, this general approach can be applied to clusters that contain only two countries if it is felt that these clusters are too small.

6. CONCLUSION

A few main themes emerge from this study. First, anomalies in the data can be identified from the dissimilarity matrices themselves, before the application of cluster analysis methods. Second, dissimilarity matrices can be constructed in a number of ways, each of which may help identify different types of outliers. Third, the dendrograms and resulting clusters in general are not very robust to the way the dissimilarity measure is defined or the sample of data over which they are calculated. Fourth, cluster analysis may prove useful in data validation in at least two ways. The construction of dendrograms over the full set of countries helps to highlight those with anomalous data. In addition, the division of countries into clusters before the application of data validation methods may increase the power of these methods. This might lead to the detection of anomalies that would otherwise have been masked by the inherent heterogeneity of the countries in the region. However, the latter ideally requires a dissimilarity measure that generates clusters that are reasonably stable from ICP 2005 to ICP 2011. A possible candidate in this regard is the dissimilarity measure derived from the difference in per capita GDP. Fifth, it seems that not enough attention was devoted to validation of the basic heading expenditure data in ICP 2005. This may be partly because neither of the two main data validation tools that were used (the Quaranta and Dikhanov Tables) is designed for this purpose. Dissimilarity measures and the dendrograms derived from them may prove useful in ICP 2011. Finally, it is recommended that the main empirical calculations in this report be repeated on the ICP 2011 data for Africa as part of data validation. The dendrograms in Figures 3a and 3b could also be used as a guide when forming subregions in Africa, thereby allowing data validation at a more refined level.

REFERENCES

Blades, D. (2007), "GDP and Main Expenditure Aggregates," Chapter 3, *ICP 2003-2006 Handbook*. Washington D.C.: The World Bank.

Diewert, W.E. (2001), "Similarity and Dissimilarity Indexes: An Axiomatic Approach," Discussion Paper No. 02-10, Revised March 2006. Vancouver, British Columbia: Department of Economics, University of British Columbia.

Diewert, W.E. (2009), "Similarity Indexes and Criteria for Spatial Linking," in Rao D.D. (ed.), *Purchasing Power Parities of Currencies: Recent Advances in Methods and Applications*. Cheltenham United Kingdom: Edward Elgar.

Diewert, W. Erwin (2010), "New Methodological Developments for the International Comparison Program," *Review of Income and Wealth* 56, Special Issue 1, S11-S31.

Diewert, W.E. (2011), "Methods of Aggregation above the Basic Heading Level within Regions," Chapter 7 in Rao D.S. and F. Vogel (eds.), *Measuring the Size of the World Economy: A Framework, Methodology and Results from the International Comparison Program (ICP),* forthcoming.

Dikhanov, Y. (1997), "Sensitivity of PPP-Based Income Estimates to Choice of Aggregation Procedures." Washington D.C. Development Data Group, International Economics Department, The World Bank.

Everitt, B.S., S. Lander, M. Leese and D. Stahl (2011), *Cluster Analysis*, Fifth Edition, Wiley Series in Probability and Statistics. Chichester: John Wiley & Sons.

Heston, A. (1994), "A Brief Review of Some Problems in Using National Accounts Data in Level of Output Comparisons and Growth Studies," *Journal of Development Economics*, vol. 44, no. 1, pp. 29-52.

Hicks, J.(1946), *Value and Capital*, Second Edition. Oxford: Clarendon Press.

Hill, R.J. (1997), "A Taxonomy of Multilateral Methods for Making International Comparisons of Prices and Quantities," *Review of Income and Wealth*, vol. 43, no. 1, pp. 49-69.

Hill, R.J. and T. P. Hill (2009), "Recent Developments in the International Comparison of Prices and Real Output," *Macroeconomic Dynamics*, vol. 13, supplement 2, pp. 194-217.

Leontief, W. (1936), "Composite Commodities and the Problem of Index Numbers," *Econometrica*, vol. 4, no. 1, pp. 39-59.