

Experiencing the Semper Validation Software: Genuine African solutions for data validation within the International Comparison Program

Mathieu B. Djayeola and Roland Rittenau¹

Summary

Since the first African participation in 1970, this is the first time an African institution has taken the lead to manage the International Comparison Program and provide support to more than forty countries in the region. This support covered the design of price survey instruments for price data collection and office editing tools for data validation at the national and regional levels. The development of the Semper Validation Software has provided user-friendly tools to assist countries in their price survey data validation activities. Beyond the technicalities of computer software, the Semper could be appreciated as a communication vehicle, a code, and a language between the national ICP teams and the regional coordinator. The main purpose is to question the quality of the data provided for international validation. An answer is provided within minutes with graphical tables on which action is required from the user.

Key words

International comparison - price survey - data validation

Résumé

Depuis la première participation africaine en 1970, c'est la première fois qu'une institution africaine est à la tête du Programme de comparaison internationale et fournit un appui à plus de quarante pays de la région. Cet appui a porté sur la conception d'instruments d'enquêtes sur les prix destinés à la collecte de données et sur des outils d'édition en vue de la validation des données sur le plan national et régional. La mise au point du logiciel Semper se situe dans le cadre d'un processus visant à fournir des outils conviviaux destinés à aider les pays dans leurs activités de validation des données issues des enquêtes sur les prix. Outre sa technicité en tant que logiciel informatique, le Semper peut être aussi considéré comme un moyen de communication, un code et un langage entre les équipes nationales du PCI et le coordon-

1: This paper was prepared by Mathieu B. Djayeola (b.djayeola@afdb.org), Statistician Economist and Roland Rittenau (r.rittenau@afdb.org), Principal Statistician, both members of the Regional Coordination Team of ICP-Africa, African Development Bank, Tunis, Tunisia.

nateur régional. Il sert surtout à tester la qualité des données en vue de leur validation internationale. L'utilisateur est assuré d'une réponse en quelques minutes dans un tableau à thèmes graphiques sur lequel il devra intervenir.

Mots clés

Comparaison internationale – enquête sur les prix – validation des données.

1. Background

The International comparison Program (ICP) is a global statistical initiative established to produce internationally comparable price level, expenditure values, and Purchasing Power Parity (PPP) estimates, with the objective of facilitating cross-country comparisons of economic aggregates and price levels for GDP and its sub-aggregates. The genesis of the ICP was an early recognition that measures of economic aggregates based on exchange rates do not reflect differences in price levels between countries, and as such they are unstable for policy decisions which in principle relate to volumes only, free of price and exchange rate distortions.

The prices collected for the ICP can be used for other comparative purposes such as poverty-specific price comparison². Whatever the purpose, it is important that countries collect prices for products that are comparable both within the country and between countries. More generally, the quality of international price comparisons, like the quality of national consumer price indices (CPI), is highly dependent on the quality of the basic input data. The validation process described in this paper is designed to ensure that countries provide the ICP regional coordinator with good quality price data for a selection of comparable products.

This paper is concerned with the intra-country validation of prices collected for the ICP – that is, the validation of its price data that a country carries out before reporting them to the ICP regional coordinator. The process of inter-country validation of prices subsequently carried out by the ICP regional coordinator, and its supporting software are described elsewhere³.

2: Based on a poverty-specific sub-set of products and a specific weighting pattern with the aim to recalculate international poverty lines into local currencies in a much more appropriate way than it is currently done.

3: See for example, Annex IV Quaranta editing procedure, Eurostat-OECD Methodological manual on purchasing power parities, Luxembourg, 2005

2. Terms of Reference for the Development of a Validation Software

The development of Semper stems from the need to verify, clean and validate price data in the context of CPIs and ICP. In most of the countries, this validation process is carried out manually or using non integrated and specialized tools. Therefore, there is no assurance of consistency in methods within a country or over time. This is compounded when comparing validation practices between countries.

The product list was the main ICP survey instrument provided to the countries. It was developed through an extensive participatory process that involved the regional member countries and the African Development Bank. To check the quality of prices collected for these products, the regional coordination team of the ICP-Africa entrusted Mathieu B. Djayeola, Statistician Economist, to develop a specific validation tool⁴ according to the following terms of reference:

1. Verify the consistency between reference and observed unit of measurement types⁵, e.g.: litre is incompatible with kilogrammes and vice-versa;
2. Check the relationship between observed quantity and unit of measurement, like entering 800 kilogrammes instead of 800 grammes;
3. Process and flag prices reported in different currency units, e.g.: South African Rand and its cents;
4. Generate average prices for each product;
5. Compute price dispersion indicators such as minimum, maximum and standard deviation;
6. Identify outliers and potential mistakes;
7. Analyze price escalation over time to control trends;
8. Identify wrongly reported or inputted price quotations: e.g. misplaced decimal point, confusing use of sub-multiple of currency unit, etc.

The Semper software was named after the Latin dictum, *Semper aliquid novi Africa affert*⁶ to underline the novelty as well as the African peculiarity of the application. This Microsoft Excel Visual Basic procedure facilitates data validation at the country level. It aims at verifying the consistency of units of measure, quantities and prices, as well as whether price variations are within acceptable limits.

The quest for data quality and the increased awareness for product comparability have positive impact on national statistical data gathering systems, especially the CPI. As the ICP surveys cover a twelve-month period in Africa, a time series analysis

4: The software was developed under the overall supervision of Michel Mouyelo-Katoula, ICP-Africa Regional Coordinator, and the technical guidance of Roland Rittenau, Principal Statistician.

5: Unit of measurement types refer to weight, capacity, linear scales, units, etc.

6: "There is always something new out of Africa": Author unknown, recorded by Pliny the Elder.

tool was essential. Functionalities of the Semper Time Line are specifically relevant for CPI-type data validation.

What is the Semper validation software? What are the targeted tasks to be accomplished by the software? How does it operate? What are its specific features? What are the advantages and limits of this application? This paper elaborates on one year of experience with ICP data collection and validation in forty-eight African countries.

3. What is the Semper Validation Software?

The Semper validation software builds on the concept of price quotation, which, in the context of the International Comparison Program, includes the following minimum information:

1. To which group of products, or in ICP terminology, to which basic heading does the product belong? According to international classifications, a bottle of orange drink purchased in a city shop is different from the same bottle purchased for immediate drinking in a restaurant. From a national accounts point of view, these bottles of orange drink do not belong to the same classification category as they do not serve the same consumption purpose and the service element is different. For instance, in the context of the ICP, the bottle bought in a shop is a beverage pertaining to the basic heading “Soft drinks and concentrates” while the bottle bought in a restaurant relates to the basic heading “Catering in hotels and restaurants”.
2. The identification of the product: A code is necessary for automatic processing of information related to products. The name of the product is needed for understanding what a specific product code refers to. This is also a classification issue but may also be relevant when making economic analysis of data collected in the field. All characteristics and modalities of the products are important identification factors. For example, “Tinned Pineapple” is fully defined by the following product specification: [Basic heading: Preserved and processed fruits; Product code: 025.07; Reference quantity & unit of measurement: 850 Milliliters; Product presentation: Tin; 800-900 Milliliters; Type: Pineapple; Form: Chunks; Juice: Syrup (water + sugar)].
3. The observed unit of measurement: It is necessary to ensure that the observed unit of measurement is identical or at least convertible into the reference unit of measurement. Example, if the reference unit of measurement for “Tinned Pineapple” is milliliter, the actually observed unit can be any multiple of a milliliter, whereas units of weight (such as gram, pound, etc.) will not be accepted.

4. The observed quantity: the quantity observed must be reported to estimate the unit price (i.e. related to the reference quantity) known as the “recalculated price” in the application.
5. The observed price.
6. The place of observation: the outlet, the location, the country where a price is collected are important price determining factors. The price of mineral water in a tourist area will not be the same as in a neighborhood shop of a residential area.
7. The time of data collection may also be critical for some products. Seasonal product prices vary over the months and prices of perishable goods may vary according to the time of observation in the day. For example, prices of fresh seafood may go down in the evening.
8. Comments and remarks can provide additional information on a particular price quotation, in terms of survey circumstances or deviation from the product specification.

The Semper validation software is an integrated application designed for office editing of field data collected in the context of the ICP-Africa. Its three-stage procedure is as follows:

1. Check survey constants: Conformity of reported product codes, name, quantity, unit of measurement, and other characteristics to the reference specification in the product list. Any product not included in the ICP-Africa product list is immediately rejected by the Semper.
2. Establish the dual mathematical relationship between observed and reference quantity and unit of measurement and effect necessary conversions to rescale observed prices.
3. Analyze rescaled product prices to identify outliers and potential errors.

4. How Does it Operate?

The Semper Validation Software analyses “*country data*” in the context of the ICP-Africa. It needs clean and standardized files, structured and verified database to run properly. The application first checks these requirements before processing. Like a factory, the Semper works with an Input-Process-Output system (Figure 1).

The Input – What the Semper requires.

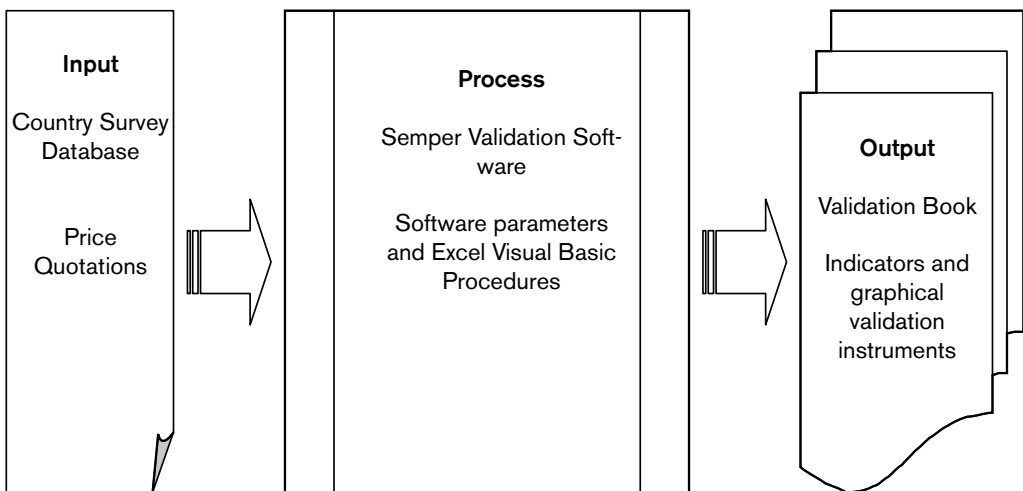
The Semper validation tool generates its input from the ICP-Africa Data Entry Sheets. The Data Entry Sheets are specific devices for data entry with a high degree of se-

curity, protection of cells and scroll functions to guide users during the data entry process, guarantee a homogenous structure of the data sets and avoid typing mistakes as far as possible. These sheets are typically structured by price collection center in a particular country and are relatively heavy in size, due to the big number of control functions. They are merged into one single input file referred to as “*Country Data*” through an automatic Excel macro⁷.

The Process – What does the Semper do?

The Semper program file contains five worksheets and a built in Visual Basic module. The first sheet (the PS or Product Specification Sheet) contains general information on the product specification and individual data validation parameters such as the reference quantities, the reference unit of measurement and the acceptable outlier percentage. The second sheet (the UoM or Unit of Measurement sheet) contains information on compatible unit of measurement and subsequent conversion factors, the fourth sheet (the User Information Sheet) contains basic information on the way the price database is built and how the outlier indicators are computed, and the fifth sheet (the Comments Sheet) is reserved for the user’s comments and remarks. The third sheet (the Program Info Sheet) is the core part of the Semper. It contains the Start button to launch the application and the processing log.

Figure 1: Structure of the Semper software



7: This macro named “*MergeCountryDataMacro*” is necessary for fast and automatic preparation of a national price database for validation.

The Output: What do you get out of the Semper validation software?

The Output (or “Validation Book”) is a specific Excel workbook containing three sheets: (1) the “*Computing Area*” Sheet, (2) the “*Summary*” Sheet to store temporarily processed information, and (3) The “*Miscellaneous*” Sheet, meant to serve as user notebook (Table 1).

Using converted prices (to the reference quantity) for a particular product, Semper computes selected indicators for survey data analysis. These are:

- (1) Averages prices
- (2) Count of price quotations
- (3) Minimum price
- (4) Maximum price
- (5) Standard deviation of prices
- (6) Count of representative quotations
- (7) Count of poverty relevant quotations
- (8) Count of poor location quotations

Once all these indicators are processed, the Semper starts the individual data validation by generating, simultaneously, an outlier indicator for individual quotations and minimum - maximum indicator for a group of quotations related to each product. Then, it highlights the isolated quotations with a range of colors: no color when the price is close to average, yellow when the price is within an acceptable distance from the average, pink to indicate a likely outlier and red to highlight obvious outliers and errors (Table 2).

Users can save the output generated by the software in different names or specific locations for future consideration. Any attempt to corrupt the structure of the file or inappropriate database configuration can run into errors and stop the program.

In the following example (Table 1) we see information at the level of the overall price report (Grand_Aggregation): this specific country has reported 4732 price quotations. The overall average price is statistically irrelevant, but shows that no technical problem occurred, as all individual data could be treated.

One level below, the product level is “00104_Aggregation”, referring to product 00104. In this example three products were observed: 00104, 00105 and 00107, all of them belonging to the first basic heading 001 – Rice. At the level of the product,

the table shows the number of price quotations (Price_Count) and the average price. For instance, for product 00104, this country reported 6 quotations with an average price of 553 currency units. All this information is called “recalculated”, because the observed prices are recalculated to the reference quantity. The outlier indicator for product 00104 is red, as it is below - 30 meaning that the minimum price is much less than half of the maximum price quotation.

At the level of individual price quotations, index numbers identify each quotation easily in the country reports. This is important when correcting errors. Outlier indicators for individual quotations are highlighted when they are too far from the average price for the selected product.

Table 1: A screenshot of the validation book (computing area)

Index	Product Name [02]	Outlier_Indicator	Requested Outlet type [03]	Product Code [01]	Recalculated Price_Count	Recalculated Price_Average	...
				Grand_Aggregation	4732	1292	
		-32		00104_Aggregation	6	553	
01225	Long-grained rice	8	Supermarket	00104	1	632	
01277	Long-grained rice	-4	Supermarket	00104	1	340	
03873	Long-grained rice	-16	Supermarket	00104	1	898	
04097	Long-grained rice	6	Supermarket	00104	1	452	
04140	Long-grained rice	12.5	Supermarket	00104	1	582	
04370	Long-grained rice	3	Supermarket	00104	1	416	
		30		00105_Aggregation	5	1740	
00687	Long-grained rice	9	Neighbourhood shop, Market	00105	1	1960	
00688	Long-grained rice	13	Neighbourhood shop, Market	00105	1	1666	
01721	Long-grained rice	9.5	Neighbourhood shop, Market	00105	1	1540	
02119	Long-grained rice	7.5	Neighbourhood shop, Market	00105	1	1470	
03661	Long-grained rice	6	Neighbourhood shop, Market	00105	1	2065	
		-30		00107_Aggregation	32	70	
00124	Medium-grained rice	8	Neighbourhood shop	00107	1	60	
00280	Medium-grained rice	8.5	Neighbourhood shop	00107	1	80	
00480	Medium-grained rice	1	Neighbourhood shop	00107	1	90	
00642	Medium-grained rice	11.5	Neighbourhood shop	00107	1	65	
00982	Medium-grained rice	13.5	Neighbourhood shop	00107	1	68	
...

For more detailed explanation, a user-guide is available.

5. Special Features

Unlike most of common Microsoft Visual Basic applications, the Semper builds on both internal and external procedures of the mother software – Microsoft Excel. Internal procedures are usual Excel actions such as formula, copy and paste tasks, as well as sub grouping procedures. Beyond the regular Visual Basic, Semper interacts with the operating system to gather information on language specific parameters: e.g. where does Excel store sub-group information when computing subtotal? In the original English Excel version, the sub-group identification is on the left of the string, while the identification is on the right in some French versions. By querying the computer system files, the Semper locates these information items and takes them into account while running.

As the application intends to collect all necessary parameters on its own, Semper runs specific procedures that fit the size of the price database. This re-scaling procedure aims at allowing the application to self-extract information on the number of price quotations, the number of products surveyed and the number of quotations per product.

Strict compliance to the common ICP data validation standards was the core objective of the Semper. The price database input to the Semper (“*Country Data*”) is derived from the products list and the data entry form, which were originally compiled in strict compliance with the World Bank requirements for data transmission. The Semper has an extension for time series analysis called Time Line. An interesting point from a CPI perspective is that this component is capable of validating data over time.

6. Advantages

Semper generates a central matrix (located in the *Validation Book*) which can be used for further graphical data analysis: navigation, filtering and grouping. There are three levels of analysis in the Validation Book. The lower the level, the finer the analysis:

- (1) The overall price report level,
- (2) The product level, and
- (3) The price quotation level,

It is easy to navigate through the Validation Book: the user can concentrate on one product, select a range of products or select price quotations within a product to be analyzed together. Semper also allows for the selection of price quotations on the basis of their outlier indicator value, the level of representativity and/or the relevance of the product to poverty analysis.

Fatal errors are clearly flagged with two specific codes as outlier indicator: the software creates a specific “outlier code” when either of the key input to the computation of the recalculated price (price, quantity, or unit of measurement) is missing [Code 9696]; and when, despite the availability of all these variables, the recalculation of prices is not possible [Code 9898]. This case occurs if the observed unit of measurement is not compatible with the requested unit of measurement (Figure 3).

The outlier indicator is set to positive values for “non problematic” price observations and negative values for “problematic” price observations. It is calculated following a formula, which is determined by the setting of an acceptable degree of price variation to the average for a particular product. These settings are variable and can be modified by the user. It is also possible to set different tolerance limits for different type of products to reflect the heterogeneity of the markets in the validation procedures. The most problematic items are attributed a higher number in absolute terms. This way, Excel filters will highlight these items and direct the user to them before showing less problematic price observations and those for which the Semper did not identify any anomaly.

The color-coding of outliers has proved to be very practical and is highly rated by most users. However, in few cases, there is the psychological danger of avoiding colors by simply deleting highlighted quotations. This is an inappropriate data validation practice and must be avoided. It has been noticed that some of these observations depict the market reality of the country so they have to be maintained in the data set.

When processing the survey data, Semper preserves the original copy of the “*Country Data*” file to allow exclusively human intervention in the validation process. The software does not take any step towards the correction or deletion of any “problem-

atic” price quotation. It does not directly point out any record as unsuitable and does not generate any outside table summarizing the information in the price database.

Table 2: Overview of color codes in the Semper

Absolute deviation of price quotations	Mini/Max ratio of product prices	Color coding
0-30% to average	< 0.5 – Ration is less than 1 to 2	None
30-40% to average	10% less than a 1 to 2 ratio	Yellow
40-50% to average	20% less than a 1 to 2 ratio	Pink
More than 50% to average	30% less than a 1 to 2 ratio	Red
Errors cases (“9696” or “9898”)	#N/A, #DIV/0! Error types	Red

Instead, the Semper is restricted to the role of a data analysis tool, providing easy-to-read indicators to help the user in the identification of problematic items and unclear price quotations.

The underlying philosophy is that the validation and the cleaning of the price data base cannot be automated. Any automatic deletion of outliers could lead to biased results. Outliers need to be verified and only deleted or corrected if found to be wrong. Reality is sometimes surprising and different from expectations. In other words, outliers can reflect market conditions.

Therefore, the task of a software shall not be the automatic cleaning of a price report, but a highlighting of potentially problematic cases. Flagged observations have to be verified and corrected in case of errors or confirmed if reflecting reality.

7. Discussing Recurrent Issues

One specification of the Semper lies in its nature as an intra-country validation tool, consequently the possibilities of analysis in the case of just one price quotation for a certain product description are limited. In such cases no meaningful analysis on the basis of average prices and variation can be done. However, cases with just one price-quotation per product description can be analyzed in the framework of time series or through cross-country analysis of price data. Tools for analyzing these cases are included in other software and data analysis procedures such as the Quaranta tables and the Time Line analysis, an extension of the Semper validation software.

It should be highlighted that the core Semper software makes the strong assumption that each monthly price database is self-determined. It is assumed that apart from the base parameter such as the conversion factors, the outlier tolerance ratio limit, and the minimum-maximum price ratio limit, all other data analysis features are included in the database. Obviously, there is no way to validate isolated price quotations.

Other features of the software are of a conceptual nature: (1) the exclusive use of graphical instruments to highlight problematic items, instead of words and letters (2) the software runs from the Random Access Memory (RAM), and therefore requires a minimum capacity of 256 Kilobyte, (3) the ignorance of the time dimension within the country by the core Semper application is fixed by adding a Time Line component.

The first feature stems from a technical choice: for instance, it would have been possible to exclude “problematic” items from the computation of recalculated prices and outlier indicators. Instead, the Semper made the option of contaminating all quotations of a product if one of them is a fatal error. Unless all compatibility problems are resolved, the product level indicator will not be processed, and till the last product is analyzed, the overall group indicators will not be computed. One can argue about this strategy of contamination as it puts emphasis on the errors until the user solves all highlighted errors.

Secondly, the Semper is designed to run using exclusively the RAM. It does not manage internal resources of the computer. The memory is heavily used and information is stored in the RAM each time the application runs within one working session. As the Semper is not a resident installed application, it runs directly from the RAM, where the temporary information, generated when the application is running, is stored. Unfortunately, freeing the RAM will mean removing the Semper from it. There is a design rule in programming software stated as follows: “you can not cut a branch on which you are installed”. This rule limits the performance of the Semper and after the application processes approximately 50,000 price quotations, it is recommended to start a new session by restarting the computer, depending on which Operating System is used. However, in standard country-validation procedures, the number of 50,000 price quotations for one survey is rarely ever reached, but can be exceeded in case of treating a huge survey several times a day.

Concerning the third feature, some room for future improvement is with the consideration of the time dimension within countries. Especially for bigger and culturally di-

verse countries, it makes a difference, if price variations occur within a certain region or between regions. The same way, in a country with high inflation, variation can occur from one month to the other. In the core Semper application, all price-quotations are treated on the national level for calculating the averages and outlier indicators. During the validation process, country experts may even now follow the regional origin of each quotation also in the Semper output; this is an important aspect when judging about plausibility of variation.

8. Practical Country Experiences

The Semper is now used in more than 40 countries on a monthly basis for one year and after some “teething problems” (mainly referring to surrounding software versions and language settings) having been fixed, it runs without any technical difficulty all over the continent. Originally, the Semper was meant to have the following main characteristics:

- Using standard software with easy or no installation procedure;
- Performing the required tasks in a tailor-made way to satisfy ICP-Africa needs;
- Working also on older models of computers;
- Being operational in different language settings;
- Being modifiable for future developments and modified uses.

The philosophy of the software creators was not to do data validation automatically, as this is considered to be a genuine human task, needing individual interpretation of specific data situations. The aim of the software is to support and facilitate this human task by highlighting dubious situations, but never to “correct” applying a standard rule.

Monthly data collection started in most African countries for the main household consumption component in June–July 2005. Three countries have started earlier in January and few launched the ICP surveys in April–May 2005. Participating countries have submitted Semper validated data from the start of data collection up to June 2006. The editing and validation of field data was organized in five sub-regions managed by sub-regional organizations namely (1) Afristat, comprising Francophone West and Central African countries; (2) COMESA, composed of Eastern African countries; (3) ECOWAS handling activities in Anglophone West African countries; (4) MAGHREB for Northern African countries directly managed from the African Development Bank, and (5) SADC for southern African countries.

Periodic sub-regional workshops have been organized in all sub-regions, with the objective of giving practical training on how to use the Semper and sharing experience on the quality of the data collected in the participating countries. The international validation was mainly based on the analysis of Quaranta tables generated by other software specifically designed for the ICP.

The Semper provided necessary tables to assess (1) the coverage in terms of the number of products covered out of the 853 items of the regional products list of the main household consumption component for urban and rural areas, (2) the number of quotations per products, (3) the average, minimum and maximum prices for reference quantities of each product specification, (4) obvious errors in terms of unit of measures, product specifications, data entry errors etc., and (5) variation in view of identifying outliers and reducing them.

9. Conclusion

The Semper Validation Software is an integrated application designed to facilitate office editing of price data collected during field surveys in the context of the International Comparison Program for Africa (ICP-Africa). The development of the application was inspired by the Input – Process – Output approach used in the industrial community.

Since its release in July 2005, the application has helped to carry out intra-country data validation in more than 40 countries participating in the ICP for the Africa region. After twelve month of experience, it is important to stop over and share about the experience reported by the countries.

As the software highlights potentially problematic price observations in various colors (red being strongest), some countries showed initially the tendency to eliminate the respective observations just because of their marking. This was subject to discussions in several workshops and has now been overcome by most countries. Countries report consistently that the Semper is a very appropriate tool to separate errors from pure price variation.

The agreement is that good quality data do reflect the reality. If reality sees pure

price variations due to transport problems, climate, culture and other influences, a high quality data set will still have some cases highlighted.

10. References

African Development Bank. (2005). Training of Field Supervisors and Price Data Collectors, A Trainer's Guide, ADB, Tunis, Tunisia, 1-77.

Adam, A. (2005). Quality Assurance Guidelines, ADB, Tunis, Tunisia, 1-18.

Astin, J. (2004). ICP 2004 Operational Manual, What National Coordinators Need to Know, Canterbury, UK, 48 - 50.

Eurostat-OECD (2005). PPP Methodological Manual - Annex IV Quaranta Editing Procedure, Luxembourg, Luxembourg, 1-6.

Kokil, B. et al. (2005). Data Validation with the Semper Excel based Software, ADB, Tunis, Tunisia, 1 - 5.

Rittenau, R. (2005). The Quality of Data in ICP-Africa, ADB, Tunis, Tunisia, 1-6.

Sergeev, S. (2003). Description of the VBA Program for the Computation of the EKS-PPP at the Basic Heading Level and the "Quaranta" Tables, Vienna, Austria, 1 - 27.

The World Bank. (2004). Price Collection ToolPack, Price Collection Module User Guide, Washington, USA, 1 - 62.