

CROWD-SOURCED PRICE DATA COLLECTION THROUGH MOBILE PHONES

Nada Hamadeh¹, Marko Rissanen², and Mizuki Yamanaka³

¹Senior Statistician, Development Data Group, World Bank, Email: nhamadeh@worldbank.org

²Consultant, Development Data Group, World Bank, Email: mrisanen@worldbank.org

³Consultant, Development Data Group, World Bank Email: Email: myamanaka@worldbank.org

Abstract

The World Bank Pilot Study for Crowd-Sourced Price Data Collection through Mobile Phones combines the need for high-frequency data, the recent developments in the information and communication technologies (ICT) sector, and power of crowd in an innovative way. Crowd-sourcing is the act of taking a job traditionally performed by designated agents, such as national price collectors, and outsourcing it to an undefined generally larger group of people in the form of an open call. Crowd-sourced data is data collected and reported by the user community using information and communication technologies. The objective of the pilot was to study the feasibility of crowd-sourced price data collection. The pilot was carried out by a World Bank contracted private company, JANA. The pilot method was implemented by non-professional price collectors (NPCs) as the “crowd,” using Personal Computers (PCs) and mobile phones as modern ICT means for collecting data. The NPCs were recruited using various social channels, such as Facebook and Twitter. A multilingual web microsite, PC and mobile phone accessible, was specifically developed for the pilot to facilitate all survey phases. In the pilot phase, price data was collected for 30 tightly specified food commodity items on a monthly basis for a period of approximately six months in eight pilot countries: Brazil, Bangladesh, India, Indonesia, Kenya, Nigeria, Pakistan and Philippines. After data verification and validation, NPCs received airtime rewards using JANA’s platform that has been integrated into the backend billing systems of over 235 mobile operator partners worldwide.

Keywords: Food prices, Crowdsourcing, Information and Communication Technologies

Acknowledgements: The authors wish to acknowledge Grant James Cameron and Michel Mouyelo-Katoula from the World Bank for their review and valuable inputs to this paper; Nathan Eagle and Kenvor Cothey from Jana for their work and contribution to this pilot study; and Morgan Brannon for her invaluable help with the preparation and editing of this paper.

1. Introduction

During the past few years, interest in high-frequency price data has grown steadily. This is partly due to recent major economic events, including the food crisis and the energy price surge, which have highlighted the need for timely and detailed data to inform decision making in modern globally interlinked economies. Thus, the need for detailed timely high-frequency data that is openly available to all users has increased. However, standard survey methods are struggling to meet this demand, due to the high cost of collecting detailed sub-national data, the time lag usually associated with publishing the results, and the limitations to publishing detailed data. This has led to a search for alternative ways to collect and disseminate price data. At the same time, the rapid development of the ICT sector is transforming the way people communicate and interact across the globe. Between 2000 and 2012, the number of mobile subscriptions in use worldwide has grown from fewer than one billion to over six billion, of which nearly five billion are in developing countries. As a result, 75% of the world's inhabitants now have access to a mobile phone. As mobile devices are becoming cheaper and more powerful, and network coverage and bandwidth is expanding, it is expected that the “mobile revolution” is currently only at the start of its growth curve.

The World Bank Pilot Study for Crowd-Sourced Price Data Collection through Mobile Phones combines the need for high-frequency data, the recent developments in the ICT sector, and power of crowd in an innovative way. Its objective was to study the feasibility of crowd-sourced price data collection. The pilot was carried out by a World Bank contracted private company, JANA. The findings of the pilot are presented in the following sections related to project design (section 2), data analysis (section 3) and challenges, successes, and lessons learned (section 4).

2. Project Design

2.1. Country coverage

The pilot originally covered seven countries: Brazil, China, Egypt, India, Kenya, Nigeria and Uganda. This country list was modified during the course of the project for various reasons: China was dropped from the list owing to the unclear legal landscape of operating the project; and activities in Egypt and Uganda were stopped given concerns around the ability to recruit and pay participants. Indonesia, Philippines and Pakistan were added to the pilot.

Table 1: Country Coverage

Country	Number of Survey Locations	Number of Supermarkets	Language	Status	Months covered
Bangladesh	47	1,282	English	implemented	Jan'12 – May'12
Brazil	13	41	Portuguese	implemented	Jan'12 – May'12
China	-	-	n/a	not implemented	
Egypt	-	-	English	stopped	Aug'11 – Sep'11
India	33	405	English	implemented	Aug'11 – May'12
Indonesia	41	237	Indonesian	implemented	Sep'11 – May'12
Kenya	41	1,447	English	implemented	Sep'11 – May'12
Nigeria	35	1,023	English	implemented	Sep'11 – May'12
Pakistan	49	516	English	implemented	Dec'11 – May'12
Philippines	38	212	English	implemented	Dec'11 – May'12
Uganda	-	-	English	stopped	Oct'11 – Nov'11
Total	297	5,163			

During the survey, prices were collected from around 5,000 supermarkets in close to 300 survey locations, spread across the pilot countries. Survey locations covered cities of varying size for the entire survey period. Table 1 below provides the country coverage for the pilot study.

In summary, the survey process gathered data successfully for at least six months in India, Nigeria and Kenya. By working to understand the dynamics of all the covered marketplaces, discovering and characterizing challenges and building strategies to combat those challenges, the project was also able to collect data continuously for at least four months in Bangladesh, Indonesia, Philippines, Pakistan and, to a lesser degree, in Brazil.

2.2. Recruitment, Retainment and Compensation of Non-Professional Price Collectors

The recruitment of Non-Professional Price Collectors (NPC), training and price collection survey was performed using a custom built website, developed in three languages (English, Portuguese, and Indonesian) to target focused geographies. Figure 1 depicts the project workflow and its various steps.



Figure 1: Project Workflow

In order to drive traffic to the website and recruit NPCs, the following tactics have been employed: (i) Internet and mobile advertising using social media networks including Facebook, Twitter, Buzzcity, and Orkut; (ii) Email campaigns; (iii) NPC referral by active NPCs; and (iv) Direct recruitment using online job boards. Table 2 below shows the number of active NPCs contributing in each country per month. In total, 4,751 NPCs submitted price data within the pilot project.

Table 2: Active Non-Professional Price Collectors

Country	Aug '11	Sep '11	Oct '11	Nov '11	Dec '11	Jan '12	Feb '12	Mar '12	Apr '12	May '12	Total
Bangladesh	n/a	n/a	n/a	n/a	n/a	21	240	671	1,333	162	2,428
Brazil	0	0	0	0	0	1	3	9	17	25	55
Indonesia	0	6	0	0	0	10	43	52	60	44	215
India	1	45	62	71	86	77	112	113	118	42	726
Kenya	0	5	24	73	142	84	151	364	576	493	1,912
Nigeria	0	3	3	53	65	70	221	249	274	53	991
Philippines	n/a	n/a	n/a	n/a	5	19	40	45	59	8	176
Pakistan	n/a	n/a	n/a	n/a	2	70	119	185	199	79	654
Total	3	60	95	212	300	331	689	1,017	1,303	744	4,751

The incentive structure was a crucial component with which participant behavior can be influenced. The original structure of value of airtime minutes paid was: \$1.00 as an initial bonus for providing supermarket information; \$0.50 for each supermarket identified up to a limit of three supermarkets; and \$5.00 for each verified item batch submission, consisting of price observations for each available item, containing thus a maximum of 30 price observations. Subsequently, a referral bonus of \$1.00 was granted to existing NPC for every new referral that enlisted as an NPC and submitted price data. Weekly payments were limited to \$20.00 per NPC in order to limit excessive payments in the event of fraudulent behavior. After reviewing recruitment, data submission rates, and following the change in the NPC workflow, the incentive structure was further modified. The incentive for item price

batch submission was increased for those countries where there was little activity: \$7.00 in Indonesia and \$15.00 for Brazil.

2.3. Data Collected

Data collected is broken into the three categories: (1) NPC data collected through the prequalification survey including Date of birth, Location, Languages spoken, Level of education, Mobile phone information, Level of internet access, and other contact details; (2) Supermarket data collected in the supermarket identification phase, including Name of the supermarket, Location information, and Photograph of the supermarket (optional); and (3) Price data collected during the price collection phase, for approximately 30 supermarket commodities. Figure 2 below shows an example of an item specification. The basic set of information collected covered: Price of the item identified; Weight or volume of the item (for standardization purposes); Additional comments (e.g. whether the item is in season or not); and Photograph of the item (optional).


[11] Long grain rice - Non-Parboiled		
Quantity	0.5 - 1.2 Kilogram	
Brand	Well known	
Type	Long grain, white rice (milled rice)	
Packaging	Pre-packed	
Quality	High grade	
Preparation	Uncooked, non-parboiled	
Other features	Not enriched, not aromatic (fragrant), not sticky	
Exclude	Premium rice e.g. Basmati rice, Jasmine rice	

Figure 2: Example of an Item Specification

To enhance the analysis process, the price collection website stored additional metadata alongside this primary NPC-submitted data, such as country specific currency, timing information, and fine grained (GPS) location information where it was available from the device used.

3. Data Analysis

3.1. Data Verification

Verification of the submitted data is a semi-automated process, involving analysis and manual inspection. Each NPC was scored based on various behaviors and the quality of the item batches that they submitted. High-scoring NPCs were automatically verified, negatively scoring NPCs were blacklisted, and those in a neutral standpoint were manually verified. Behaviors of NPCs such as frequency of collection, number of batches submitted, length of time taken to submit a batch, and the interval between item batches were of particular interest. Potential indicators of fraud included registering multiple times to get around limits as well as patterns of referrals. Additionally, the number of entries in a batch and the number of comments indicated the quality of data input. The verification measures were continually reviewed and revised as more data became available and patterns of behavior and common failure modes became apparent. Figure 3 below shows the item submissions accepted and rejected for the full collection period.

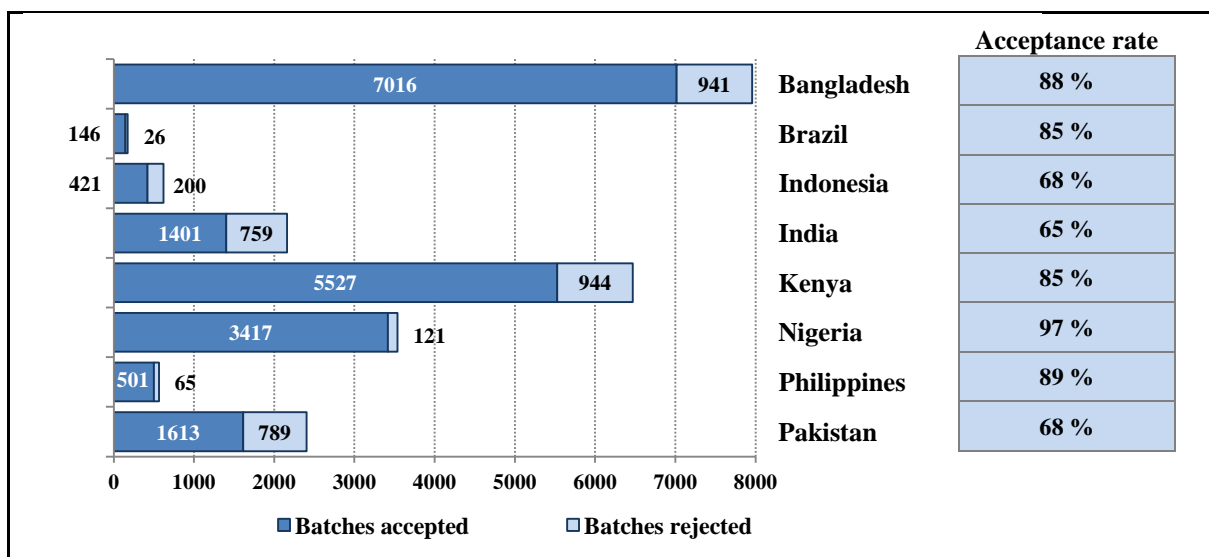


Figure 3: Submitted and Accepted Item Batches

3.2. Statistical Validation

Once the data is verified, price observations for each item at country level are validated using statistical measures such as t-ratio, Min/Max ratio and Coefficient of Variation¹ (CV), to flag potential outliers. Observations failing to pass the set thresholds of the indices were excluded from the datasets.

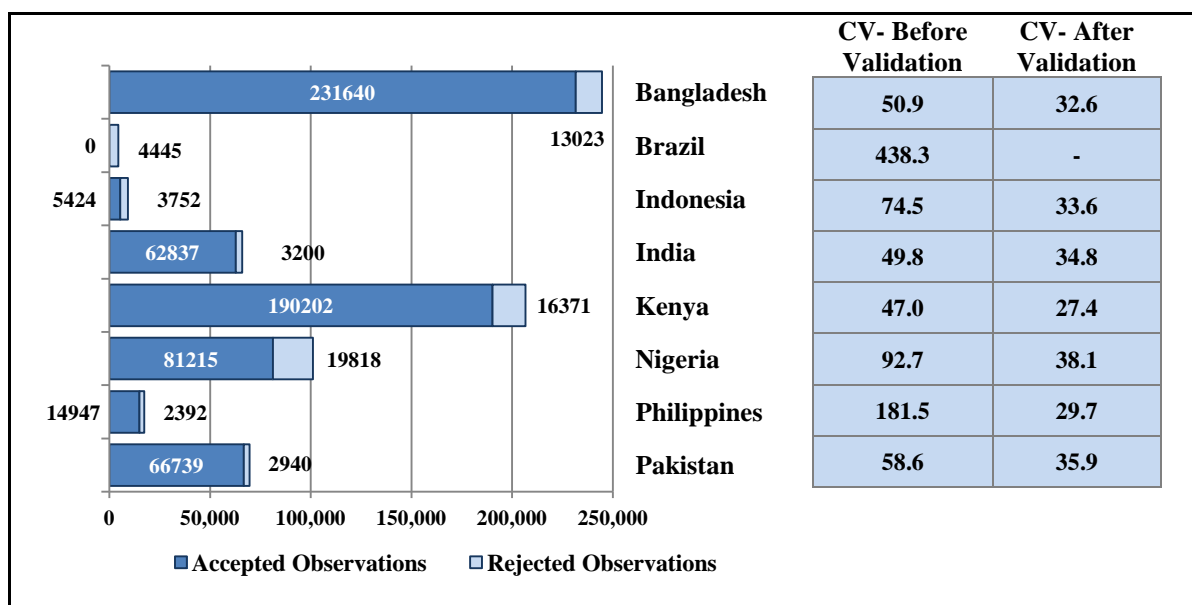


Figure 4: Number of Observations and Coefficient of Variation before and after Statistical Validation

As a result of the statistical validation, the CVs for all countries decreased substantially, leaving Nigeria and Indonesia with the highest price variation, and Kenya and Philippines with lowest price variation, as shown in Figure 4 above. Kenya and Nigeria witnessed the highest number of deleted observations, whereas Philippines and Pakistan had the least number of deleted observations. Brazil

¹ Coefficient of Variation = (Standard deviation / Average)*100

exhibited an extremely high initial CV with a low number of observations, thus the resulting average prices were unreliable. As a result, all remaining price observations for Brazil from data verification were deleted during the statistical validation

At the item level, in general more observations were excluded for items with high initial CVs by using the combination of above-mentioned statistical indices. After the statistical validation, CVs generally went down to 30-40 for most items. On average, ten percent of price observations were deleted during the statistical validation. Most problematic items were Tuna Steaks and Cornflakes, while the least difficult items were Coca Cola / Pepsi Cola, Chicken Eggs and White Sugar.

Table 3: Percentage of Deleted Observations for all Countries for the Whole Survey Period

Item	Deleted observations	Item	Deleted observations	Item	Deleted observations
Tuna steaks	21%	White rice	12%	Banana	9%
Cornflakes	17%	Shrimps	12%	White bread	8%
Wheat flour	15%	Instant coffee	11%	Brown Potatoes	8%
Cassava	15%	Dried white beans	11%	Soybean oil	8%
Flour White	14%	Orange	11%	Sunflower oil	8%
Canned sardines	14%	Maize	11%	Milk	8%
Long grain rice	14%	Onion	10%	Chicken legs	8%
Minced beef	13%	Apple	10%	White sugar	6%
Tea	13%	Palm oil	9%	Chicken eggs	6%
Cooking salt	13%	Carrots	9%	Coca Cola/ Pepsi Cola	6%

3.3. Assessment of Results

This section assesses the results from three aspects: temporally, sub-nationally, and spatially across countries.

3.3.1. Temporal analysis

Figure 5 below plots average prices in US\$ for seven selected items in India that have different trends over the period September 2011 to April 2012. Average prices were stable over the collection period for most of the countries, excluding most notably tuna steaks. Price fluctuations may be explained by seasonality, or by the difficulty in interpreting the item definitions.

Figure 5 also makes a comparison to India's monthly Food CPI (2000=100) for urban workers. Overall, the CPI and the individual items can have similar price change patterns, as can be seen for many of the pilot items.

This pilot does not aim to replace a country's CPI – that consists of thousands of items representative of the national food consumption behaviors – but instead it provides flash estimates on national and sub-national average price changes for key international food items, whose individual price changes can be larger or smaller than the whole national Food CPI basket. Thus, the data collected in this project complements the knowledge on price fluctuations within the countries, and provides insights to price changes across countries.

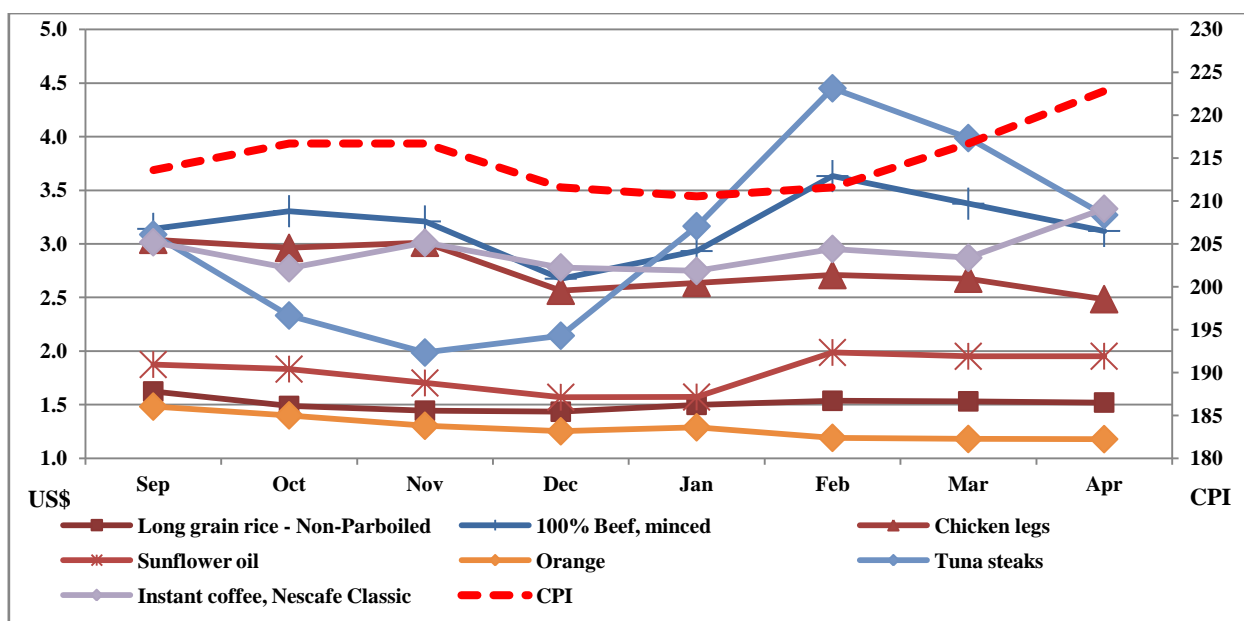


Figure 5: Monthly Average Prices for India, US\$

3.3.2. Sub-national spatial analysis

On the sub-national level, Table 4 below shows the average prices for five large Indian cities, Mumbai, Hyderabad, Pune, Nashik and Jabalpur, as well as average prices for India in total and for the remaining sample, for the period January-March 2012. As presented in Table 1 “Country coverage” above, prices were collected in 33 survey locations across India.

Table 4: Sub-National Average Prices for India, Q1 2012

Location	India	Mumbai	Hyderabad	Pune	Nashik	Jabalpur	Rest
Item	Prices in US\$, Q1 2012						
Rice	1.53	1.41	1.65	1.54	1.59	1.45	1.41
White rice	0.81	0.69	0.92	0.83	0.80	0.74	0.75
Cornflakes	2.02	2.03	2.44	1.68	1.66	1.91	1.89
Wheat flour	0.77	0.70	0.81	0.83	0.75	0.79	0.74
Maize Flour	1.67	1.70	1.72	1.59	1.46	1.71	1.73
White bread	0.99	0.90	1.01	1.11	1.00	0.94	0.99
Minced beef	3.42	3.77	3.36	3.26	3.09	3.29	3.53
Chicken legs	2.68	2.51	3.02	2.44	2.53	2.70	2.51
Tuna steaks	3.49	2.95	3.26	3.85	3.83	3.33	3.29
Shrimps	4.55	3.91	5.18	4.12	4.21	5.24	4.21
Canned sardines	0.94	0.84	1.00	0.93	0.81	1.26	0.88
Milk	0.92	1.02	1.04	0.87	0.82	0.91	0.80
Chicken eggs	0.48	0.48	0.51	0.50	0.48	0.44	0.47
Sunflower oil	1.89	2.00	1.71	1.89	1.78	2.10	2.08
Palm oil	1.50	1.59	1.49	1.42	1.30	1.59	1.61
Soybean oil	2.03	2.28	1.91	1.78	1.67	2.23	2.30
Banana	1.00	0.94	1.02	1.21	1.27	0.82	0.83

Location	India	Mumbai	Hyderabad	Pune	Nashik	Jabalpur	Rest
Item	Prices in US\$, Q1 2012						
Orange	1.21	1.03	1.20	1.37	1.55	1.13	1.05
Apple	2.18	2.34	2.25	1.91	1.93	2.27	2.24
Carrots	0.80	0.84	0.75	0.85	0.79	0.86	0.79
Onion	0.48	0.49	0.56	0.38	0.37	0.52	0.46
Maize	0.91	1.06	0.90	0.89	0.79	0.89	0.94
Potatoes	0.51	0.51	0.59	0.42	0.39	0.58	0.52
Cassava	0.75	0.56	0.86	0.81	0.81	0.72	0.62
Dried beans	0.55	0.62	0.49	0.60	0.58	0.51	0.58
White sugar	0.80	0.77	0.82	0.83	0.80	0.80	0.76
Cooking salt	0.50	0.43	0.59	0.49	0.49	0.48	0.46
Instant coffee	2.89	3.23	2.70	2.68	2.67	2.85	3.17
Tea	0.62	0.71	0.69	0.56	0.47	0.66	0.61
Coca Cola/ Pepsi Cola	1.40	1.41	1.28	1.63	1.63	1.36	1.34
Totals for all items							
Overall Price Ratio	1.00	0.98	1.04	0.99	0.95	1.01	0.97
Total # of obs.	27694	3662	8611	2656	4244	2019	6502
Av. # of obs. per item	923	122	287	89	141	67	217

3.3.3. Cross-country spatial analysis

The Price Level Index (PLI)² for categories of items reveals the price differentials between each country and another country or a group of countries. For example, in Table 5, in comparison to the geometric mean of the seven countries, prices for meat and fish in Bangladesh are relatively higher, while its PLIs for all the other categories are relatively lower.

Overall Food PLI for Nigeria is the highest among all the countries. Kenya and Philippines have slightly higher PLIs compared to the geometric mean, whereas Indonesia's price level is closest to the geometric mean. India and Pakistan have the lowest price levels.

Table 5: PLI for Categories of Items for Q1 2012 (Geometric mean of countries = 100)

Categories	Bangladesh	India	Indonesia	Kenya	Nigeria	Pakistan	Philippines	Geometric mean
Bread and cereals	76	79	94	96	244	74	103	100
Meat & Fish	114	80	95	126	118	92	85	100
Milk, cheese, and eggs & Oils and fat	74	75	96	111	148	92	124	100
Fruits & Vegetables	65	77	97	115	206	60	143	100
Sugar, jam, honey, chocolate, and confectionery	68	72	107	91	197	67	158	100
Non-alcoholic beverages	89	80	103	117	143	82	99	100
Overall Food PLI	79	77	98	109	171	77	116	100

The relationship of the price levels of the countries is quite similar to officially released statistics. The table below shows the comparison between results from this pilot survey and Food PLI released by the Asian Development Bank for 2009. Although the number of countries and the basket of items included in the analysis differ, the relationship between the selected countries is still comparable and the results from both are in line, as shown in Figure 6.

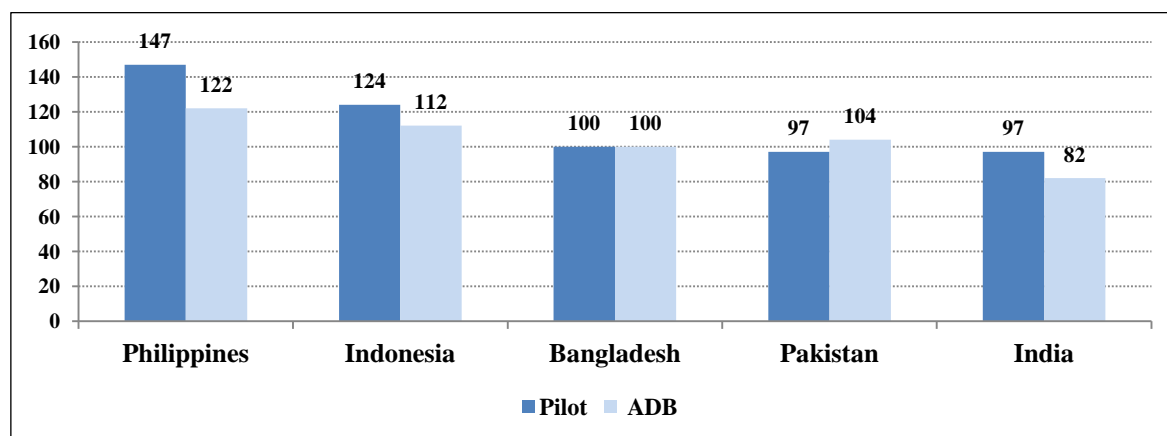


Figure 6: Overall Food PLI comparison Between the Pilot and the Asian Development Bank Comparison³

² PLI is defined as the ratio of the Purchasing Power Parities (PPPs), which is price relatives between countries, to the exchange rate and is expressed as a percentage. A country PLI that is greater (less) than 100 indicates that when the national average prices are converted at exchange rates, the resulting country prices tend to be higher (lower) on average than prices in the base country of the group.

³ For the pilot, the PLI is for all 30 food items, based on the data for the 7 pilot countries. For the ADB study, the PLI is for Food and Non-Alcoholic Beverages, based on the data for the 21 participating countries in the Asia comparison. Both figures are recalculated to make Bangladesh as the base (Bangladesh PLI = 100).

4. Challenges, Successes, and Lessons Learned

As with any pilot project, reporting the benefits and challenges is an important element for improving the project design and operation. The data collection survey process was challenging, particularly in certain countries. The main challenges encountered included:

(1) *Legal, Cultural, and Language Challenges:* Local laws pertaining to the use of the Internet, mobile phones, data collection and its intent, as well as photography impacted the project, especially in China. Cultural and language barriers needed to be addressed since they can significantly impact the success of the data collection process. For example, the website had to be translated into Portuguese and Indonesian to facilitate the project in Brazil and Indonesia respectively. Furthermore, levels of engagement differ greatly across countries. To address this concern, market-specific projects and localized project websites along with individual recruitment and incentive plans were entertained.

(2) *Building Trust:* Trust is difficult to build, yet is imperative between the NPCs and the project. The project had to instill a sense of trust in the NPCs to ensure utmost data quality while balancing with fraudulent identification measures.

(3) *Incentive Level and Structure:* The incentive level and structure affected the recruitment and retainment of NPCs. Recruitment was challenging in countries such as Brazil where the level of incentive was initially not high enough to attract engagement. Further research is necessary to determine the level of incentive required to completely engage NPCs at a productive standpoint. Another issue faced was a lag in payment due to an unreliable payment network with which the project partnered in some countries such as Egypt.

(4) *Survey Coverage:* The purpose of this pilot was to check what kind of penetration and survey coverage can be achieved. No predetermined survey framework was established. In the future, a survey framework needs to be established at the beginning of the project and specific number of observations per survey location need to be targeted. Also, this kind of collection method may not work well for social surveys that depend heavily on response rates for certain social classes, for example.

(5) *Managing and Analyzing Big Data:* Working with a huge dataset of microdata that results from such a project can be challenging and it requires database and validation systems to be put in place, in order to ensure robustness of the data validation process. Since the data is collected using crowd-sourcing, engaging the crowd can also be used in the validation process. “Crowd-validation” would entail that both data collectors and data users would participate in ensuring quality of the collected data.

Despite the challenges faced, the project was successful. The pilot provided insight to the potential use of alternative ways to sample and collect price data across countries. During the survey, prices were collected from around 5,000 supermarkets in close to 300 survey locations, spread across the pilot countries, by over 7,000 NPCs. The developed platform proved that once efficient verification and validation mechanisms are in place, data users across the globe can have access to the price data within days after the data collection.

The collected data showed that crowd-sourced price collection is feasible and thus the developed platform can have a wide variety of applications at the World Bank, international and regional agencies, national statistical offices and non-governmental organizations. The data produced by the method is cost-efficient and timely, enables high item and geographical coverage, is comparable

across countries and is open to users. In addition to the various uses of the data collected by combining crowd-sourcing and modern ICT means, this collection method benefits NPCs in low and middle income countries, through compensation in the form of free airtime and the feedback that they receive from the released data.

References

World Bank (2012). Little Data Book on Information and Communication Technology 2012.

Asian Development Bank (2012). 2009 Purchasing Power Parity Updates.

<http://www.jana.com/>