We Just Ran Twenty-Three Million Queries of the World Bank's Website

Sarah Dykstra, Benjamin Dykstra, and Justin Sandefur

Abstract

Much of the data underlying global poverty and inequality estimates is not in the public domain, but can be accessed in small pieces using the World Bank's PovcalNet online tool. To overcome these limitations and reproduce this database in a format more useful to researchers, we ran approximately 23 million queries of the World Bank's web site, accessing only information that was already in the public domain. This web scraping exercise produced 10,000 points on the cumulative distribution of income or consumption from each of 942 surveys spanning 127 countries over the period 1977 to 2012. This short note describes our methodology, briefly discusses some of the relevant intellectual property issues, and illustrates the kind of calculations that are facilitated by this data set, including growth incidence curves and poverty rates using alternative PPP indices. The full data can be downloaded at www. cgdev.org/povcalnet.

Center for Global Development
www.cgdev.org

Working Paper 362
April 2014

We Just Ran Twenty-Three Million Queries of the World Bank's Website*

Sarah Dykstra
Center for Global Development

Benjamin Dykstra

Justin Sandefur
Center for Global Development

1 Introduction

The World Bank's PovcalNet database is perhaps the most important data resource for the study of poverty and inequality in the developing world on a comparable basis across countries and over time.[1] The database attempts to harmonize all available survey data on household income and consumption in the developing world, and serves as the basis for the World Bank's global estimates of "dollar-a-day" poverty (Chen and Ravallion, 2010) and Gini coefficients of inequality. Unfortunately, the PovcalNet web site offers the public only limited access to the underlying data sets, which are restricted to internal World Bank researchers. This prevents independent researchers from checking the robustness of the World Bank's parametric approximations, or from performing additional analysis of global income distributions beyond calculating headcount poverty rates.[2]

To overcome these limitations, we applied a simple, "brute force" method to compile the full distributions of income or consumption available through PovcalNet into a form that can be easily downloaded in comma-separated values (.csv) format and used by researchers in Stata or other statistical software packages. Note that this did *not* involve accessing any data that wasn't already public. Instead, we wrote a Python script to scrape large amounts of already public data from the World Bank's web site. The script queried the PovcalNet API just as a human user might. It entered a monetary value (a poverty line) and got back a headcount poverty rate. It then repeated this process roughly 23 million times over multiple weeks, tracing out 10,000 points from the cumulative distribution function (CDF) of income or consumption in each available survey year, multiplied by 942 surveys spanning 127 countries over the period 1977 to 2012.[3] The result is a dataset of distributions, *not* the original micro survey data.

This short paper documents the data set produced by this web-scraping exercise, which can be freely downloaded from the Center for Global Development's page on Harvard's IQSS Dataverse Network.[4] The data is entirely derived from PovcalNet, which is a product of the World Bank's research department, based on work by Shaohua Chen, Martin Ravallion,

---

[1] http://iresearch.worldbank.org/PovcalNet/index.htm

[2] For instance, using the application programming interface (API) on the PovcalNet web site it would be extremely difficult to create simple graphs of the distribution of income in a given country or time period, calculate alternative measures of poverty or inequality, or apply alternative purchasing power-parity deflators to a large number of countries.

[3] No valid PPP is available for 10 of the 952 surveys. PovcalNet only provides summary measures and estimates for the underlying Lorenz curve for these surveys.

[4] www.cgdev.org/povcalnet

Prem Sangraula, and Qinghua Zhao, and should be credited as such, though they bear no responsibility for any errors here.

Most of our 23 million web queries were entirely unnecessary. For most (but, crucially, not all) countries, PovcalNet estimates the income or consumption distribution using a parametric functional form based on just three parameters plus the mean, in a process documented in Datt (1998). It is important to note that the wealth of data implied by measuring the income or consumption distribution at 10,000 points is entirely illusory for surveys where a Povcal uses a parametric fit.[5] Povcal's underlying parameter estimates are publicly available on its web site, but again must be accessed one survey at a time in a pop-up window. We run a separate Python script to extract the four key parameters for each relevant survey year and provide them as a single download in .CSV format for researchers.

Several million of our web queries were not in vain, however. For 298 of the 952 surveys in the database (these are not specified in any of the documentation we found, but includes most of Latin America), Povcal does not use a parametric fit of the Lorenz curve, but reports estimates based on the micro data or other methods. There is often far more information available about the distribution for these countries lurking within PovcalNet for users with the patience to query it. Given that, and because the underlying methodology for these countries is currently not well documented, our simple-minded method of web scraping appears to be the only way to get a full picture of the database.

The rest of the paper is organized as follows. The following section describes some of the intellectual property issues related to international poverty data. Sections 3 documents our methodology for extracting the data from PovcalNet and 4 describes the features of the final combined data set. Finally, Section 5 illustrates calculations made possible with the data, including comparison of the full distribution of income across two countries (Nigeria and South Africa) over time, calculation of growth incidence curves, and poverty measurement using market exchange rates rather than purchasing-power parity dollars.

2 Intellectual property issues

Publishing data scraped from the World Bank's PovcalNet web site is, we believe, fully consistent with the letter and spirit of World Bank policy. As noted by Chen (2013), who is

---

[5]Confession: When we started this project, we didn't understand this. If we had, we likely could have saved some time! However, we take some consolation to hear that more than one staff member in the World Bank's research department confessed that they too were confused about what Povcal contains until the conversations provoked by this note.

one of the site's creators, "PovcalNet is a prime example of how the World Bank can serve as a knowledge bank, and it demonstrates that open research and open data can increase our institutions credibility." Going even further, another recent World Bank blog post argued that micro data from all publicly-funded surveys should be in the public domain and available to taxpayers (Demombynes, 2013).

From a legalistic perspective, our publication of the data is protected by two overlapping World Bank policies. First, as of April 2012, the World Bank adopted a new open access policy (World Bank, 2012b) making all research and data published by the Bank available under a Creative Commons Attribution (CC-BY) copyright license. The CC-BY license "allows anyone to distribute, reuse, and build upon the Bank's published work, even commercially, as long as the Bank is given credit for the original creation." At the launch of the policy, World Bank Managing Director Caroline Anstey said, "there is now unlimited potential for intermediaries to reuse and repurpose our content" (World Bank, 2012a).[6]

Second, the World Bank also has a separate Open Data Terms of Use which, like CC-BY, constitute an attribution license.[7] This applies to all entries in the World Bank data catalog. Povcal is linked to a catalog entry entitled "Poverty and Equity Database."[8] To the best of our knowledge, nothing we have done here goes beyond these terms of use specified by the World Bank.

A more political concern raised by World Bank staff in response to an earlier draft of this paper is that releasing the information we have collated will render countries less willing to cooperate with PovcalNet in the future. In response to this concern we have edited the text to highlight throughout, and we repeat here, that we are only accessing information that was already in the public domain. We have not accessed the original micro survey data underlying the PovcalNet estimates, or compromised the confidentiality of countries' survey data in any way.

3 Methodology

PovcalNet is an online tool that allows users to make country level poverty estimates using survey data collected by various national governments and international organizations, and

---

[6]Whether Povcal data is covered by the CC-BY license comes down to a question of whether it is a "formal publication" of the World Bank as defined in the policy. Formal publications are intended to include papers, reports, and accompanying data sets, as described here: http://documents.worldbank.org/curated/en/2012/04/16200740/world-bank-open-access-policy-formal-publications.

[7]http://data.worldbank.org/summary-terms-of-use

[8]http://data.worldbank.org/data-catalog/poverty-and-equity-database

collated by the World Bank. One output that the tool provides is the 'headcount,' the percentage of a country's population "living in households with consumption or income per person below [a user-defined] poverty line." The welfare measurement used in the surveys varies between income and consumption, but for simplicity we will refer only to income.

The PovcalNet data files we present are closely related to the World Income Distribution (WYD) data, also maintained by the World Bank and described in Milanovic (2002) and Milanovic (2012). There are three main differences: (i) the PovcalNet files described here include a handful of additional surveys from developing countries which have become available since the last update of the WYD; (ii) unlike the WYD database, the PovcalNet data does not incorporate surveys from developed economies as reported by the Luxembourg Income Survey[9]; and (iii) while WYD reports only ten to twenty points on the income distribution for most countries, we provide 10,000 points from each distribution in the PovcalNet data base. This last point sounds like a strength of PovcalNet vis-a-vis WYD, but in many cases that is entirely artificial for reasons already alluded to, and to which we now return.

For many, but not all surveys, the poverty numbers reported by PovcalNet and recorded here are based on parametric estimates of the underlying Lorenz curve. The parameterized Lorenz curve, $L = L(p, \pi)$ reports the consumption share of poorest p percent of the population, based on a vector of estimated parameters π. PovcalNet tests the fit of two alternative parameterizations, a general quadratic function and a beta function, and reports headcount poverty rates using the model that delivers the better fit. The general strengths and weaknesses of each parametric form, as well as the mechanisms of estimating the π vector, are outlined in Datt (1998). Purely for illustrative purposes, the left hand side of Figure 1 shows a zoomed-in slice of the CDF from two consumption surveys (China 2009 and Mexico 2010) where PovcalNet has clearly smoothed the data using a parametric form.

For other surveys, particularly in Latin America, PovcalNet does not use this methodology. The PovcalNet methodology page states that "when the parametric models are unlikely to work well (notably when using incomes with a bunching up at zero) PovcalNet estimates directly from the micro data." The right hand side of Figure 1 shows two examples: Brazil in 2009 and a different survey from Mexico in 2010. Both of these are income, as opposed to consumption surveys, and in both cases it is clear that the CDF has not been smoothed. The CDF line jumps around even at small fractions of a percentile. The four cases shown in Figure 1 illustrate our best guess, which is that greater detail is available for Latin America,

---

[9]http://www.lisdatacenter.org/our-data/lis-database/

and particularly for the income surveys which dominate that region – probably because this data is processed independently by an Argentine research center, the Centro de Estudios Distributivos, Laborales y Sociales (CEDLAS) at the Universidad Nactional de La Plata.

Because the PovcalNet documentation provides little guidance as to how many of the CDFs are calculated, or when and why one method is chosen over another, we opt here to simply query the database as provided by PovcalNet and report the numbers 'as is'.[10]

When a user inputs a poverty line, Z measured in 2005 purchasing-power parity dollars per person per month, PovcalNet returns the sum of the relative frequencies of all income levels below that line, X. By a simple repetitive process of querying many Z values at small increments, it is possible to trace out the estimated CDF of income for a given survey.[11]

We collect data in a series of passes, each covering the complete CDF. The first pass provides a rough starting point and every subsequent pass reveals the survey data in greater detail, with the end goal of identifying the lower and upper bound income levels for all values of headcount. The first pass queries headcount values at $0.10 income intervals up to $100, then at $1 intervals up to $1,000, at $10 intervals up to $10,000, and finally at $1000 intervals up to $100,000. At $100,000, the headcount values for all countries is 100%. Subsequent passes are conducted using a binary search algorithm which bisects the intervals between previous queries and conducts additional queries if the headcount is unknown at that income level (ie: if the headcount differs between income levels above and below the bisected value). The process terminates after the algorithm has explored all income distributions and identified the lowest and highest income level for headcount at $0.01 intervals, the maximum resolution provided by PovcalNet.

As noted in the introduction, this degree of precision is at times illusory in another important respect. In many cases PovcalNet stores only grouped data derived from the

---

[10] In principle, and in retrospect, we could have produced the data set presented here much more quickly in two steps. First, a preliminary scraping exercise would have determined the methodology used to estimate the Lorenz curve for each survey. For those using a general quadratic or beta function, it would be easy, in principle, to reproduce all PovcalNet data for a given survey year using only the estimated functional form parameters. Intense web scraping would have been required only for surveys that were found not to use an explicit functional form. Clearer documentation on the PovcalNet website would assist future researchers in this task.

[11] For a large share of the surveys in PovcalNet, headcounts are computed based on a parametric representation of the Lorenz curve estimated from grouped data. In these cases, the same data set could also have been produced by applying the estimated parameters of the Beta or Generalized Quadratic function. We opted instead for the "brute force" method, as it could be applied consistently across all surveys and ensures our results match precisely what the PovcalNet web site reports.

original micro survey data, and relies on a parametric representation of the Lorenz curve to return headcount values.[12] In these cases, the same data set could also have been produced by applying the estimated parameters of the Beta or Generalized Quadratic function. We opted instead for the "brute force" method of repeatedly querying the web site, as it could be applied consistently across all surveys and ensures our results match precisely what the PovcalNet website reports.

PovcalNet was initially developed to allow users to duplicate the World Bank's official poverty lines of $1, $1.25 and $2 a day. As highlighted on Povcal's website, the methods used to estimate headcounts were not designed to provide reliable estimates at the upper and lower tails of the distribution.

As to the mechanics of the process, for each query a Python script opens a web browser, navigates to PovcalNet, enters the query parameters and records headcount values. The script interacts with the website just like a human user, clicking on links with a virtual mouse and entering information with a virtual keyboard. The script then stores the results in a local Mongo database. A second script is used to export data from the database to produce a CDF for each country year.

Queries are executed at a throttled rate of one every three seconds. Each query can contain one or more countries. For each country, an income value and one or more years is selected. Our queries typically requested 20 distinct headcount values. While our target was to collect 10,000 data points for each distribution at increments of 0.01% from 0% to 100% (9.4 million data points in total), trial and error was required to find the monetary values corresponding to each increment on the CDF. In the end we requested 7.4 million data points through approximately 23 million queries. Due to occasional interruptions and re-starts (mostly attributable to the web browser we used, Firefox), the queries ran 24-hours a day on a dedicated computer for approximately 9 weeks.

Returning to the question of whether or not all these queries were really necessary, we examine how much new information was gleaned from this web scraping exercise. Table 1 provides a breakdown of PovcalNet survey coverage by the methodology used. As already noted, when PovcalNet uses a specified parametric fit (such as a Beta or General Quadratic form of the Lorenz curve), web scraping is very inefficient. Users could recreate an unlimited number of estimated points on the CDF using only the three estimated Lorenz parameters

---

[12]See Minoiu and Reddy (2009) for a critical examination of the approach to modeling Lorenz curves used in PovcalNet.

plus the mean. (This 'simple' process would still require thousands of clicks and manual cutting a pasting. For researchers' convenience, we queried all of these estimated Lorenz parameters, and provide them in a separate CSV file for download.[13]) As it turns out, PovcalNet does not use a specified functional form for the Lorenz curve in 298 out of the 952 cases reported on the web. For this 31% of the PovcalNet database, our method generates potentially new information that could not be accessed through the parametric estimates alone, and may justify our somewhat inefficient web-scraping approach.

4 Data set description

PovcalNet draws on data from 952 nationally representative household surveys conducted at irregular intervals between 1977 and 2012 to calculate poverty measures. Approximately a third of the surveys measure household income, primarily in Latin America, while the remaining surveys measure consumption. The database is comprised of both raw survey data and grouped tabulations, on which Povcal fits a Lorenz curve to provide headcount estimates. Additional information on the data underlying PovcalNet and the methodology used to calculate estimates can be found on the World Bank's website.[14]

Our database contains the returned results of 23 million Povcal queries across 942 country years. We construct a CDF for each country year which lists the percentage of people (X) who live on less than or equal to some income or consumption level per month (Z). The CDF is precise up to the following limits.

We estimate X for each Z at increments of $0.01 in 2005 PPP dollars, starting at one penny and ending when X reaches 100%. Povcal returns poverty headcount numbers (the X's) to two decimal points (e.g., 31.07% of people live at or below $$Z$). When a range of Z values correspond to a single X, we include the minimum Z value for that range in our data set. This implies a maximum feasible data set of 100% x 100 = 10,000 observations for each country year.

---

[13]PovcalNet generates a detailed report for each survey year when a headcount is requested. In most instances, this report contains the parameters for the General Quadratic and Beta Lorenz curves describing the survey data. The report also indicates which curve has the best fit and whether either is invalid. This script requests each report using the Povcal web interface. It then parses the report text, searching for the curve parameters and descriptive sentences. If they are present, the script stores them in a delimiter-separated values file. If they are not found, the script stores the country and year of the survey in a separate log file.

[14]http://iresearch.worldbank.org/PovcalNet

In practice, our data set will report fewer than 10,000 unique monetary values for each country year, for two reasons. First, underlying survey data sources are typically limited to several thousand observations. Our CDF cannot produce more granularity than the underlying data. Second, our decision to limit precision to one penny increments will aggregate households living on, say, $1.001/day and $1.003. If many households fall in these very narrow ranges, our CDF will produce jumps. For instance, hypothetically, we may find that 30.00% of the population lives at or below $1.00, while 30.02% of the population lives below $1.01. In this case, the initial data set we output would contain no data point for the CDF at 30.01%. In the data set presented here, we expand each country year to 10,000 observations and fill in missing Z values by carrying forward the previous existing Z.

In total, the combined data set for all 942 surveys contains $942 \times 10,000 = 9,420,000$ observations. Of these 7,184,166 contain unique Z values within a given survey, and the remainder contain duplicate Z values, which can be thought of as vertical stretches in a CDF graph.

5 Illustrating uses of the data

In this section we illustrate graphically two types of calculations that are facilitated by the publication of the PovcalNet data in a unified CSV or Stata file. Each of these calculations would require several hundred, and in some cases many thousands of separate queries from the PovcalNet website to perform, rendering them effectively impossible.

5.1 Growth incidence curves

One benefit of the data set presented here is that it allows researchers to compare entire distributions over time and across countries. As an illustrative example, consider the cases of Nigeria and South Africa. Moss (2013) notes that the upcoming rebasing of Nigeria's GDP suggests a major symbolic milestone is perhaps only five years away, when Nigeria surpasses South Africa as the largest economy in Sub-Saharan Africa. It is worth noting, however, how poorly the per capita incomes of the vast majority of Nigerians compare to their South African counterparts, and how little Nigeria's economic growth has done to raise those incomes.

Figure 2 shows the cumulative distribution functions for Nigeria and South Africa at two relatively comparable points in time (March 1996 and September 2009 in Nigeria's case, and January 1995 and August 2008 in South Africa's case). Two points stand out. First, Nigerians are much poorer than South Africans, which we already knew. Still, a detailed comparison is striking. In the late 2000s, the *median* South African had an income that would have placed them at roughly the 95th percentile of Nigerians, while there was essentially nobody in South Africa as poor as the median Nigerian. Second, over this thirteen-year period in which Nigeria became poised to surpass South Africa as the largest economy in the region, real incomes for the bottom two-thirds of Nigerians declined.

It is possible to zoom in even further to see the relative winners and losers from economic growth in each country at various points in time. Figure 3 shows growth incidence curves for Nigeria and South Africa for each of the available survey rounds.[15] No clear patterns emerge. Nigeria saw quite regressive growth (or rather, absolute declines in income) from 1985 to 1992, then a more progressive shift in the distribution despite continued slow growth in the mid-1990s. From 1996 to 2003 growth was slow and fairly uniform, becoming more regressive from 2003 to 2009 with absolute declines for most of the distribution. For South Africa, growth from 1993 to 1995 was markedly progressive, while the rest of the 1990s were slightly regressive. The first half of the 2000s saw gains at the top and bottom of the distribution but less growth in the middle (a U-shaped curve), while 2005-2008 was essentially the opposite (an inverted U).

After circulation of the first draft of this paper, the World Bank edited the Povcal web site to note that "densities near the bottom and top tails of the distribution could be quite unreliable, and no attempt has been made by the Banks staff to validate the tool for such purposes." It is notable that several of the growth incidence curves in Figure 3 show implausible patterns at the upper or lower tails. This is possibly an artifact of the parametric fit used for the Lorenz curve, or may reflect the paucity of survey data at these extremes. The desire to distinguish those two hypotheses is a good illustration of a situation where independent researchers might reasonably desire access to the original, anonymized micro data underlying PovcalNet.[16]

---

[15]The 2011 survey round for South Africa is excluded, as it shows growth rates that suggest non-comparability with earlier surveys, e.g., growth in excess of 30% in one year at various points in the distribution.

[16]Note that research on incomes in the top 1% or 0.01% typically eschews household survey data in favor of tax records, given the limited coverage of sample surveys for any small slice of the distribution. See Alvaredo, Atkinson, Piketty, and Saez (2014) for a full list of available top incomes data sets.

5.2 PPP calculations

This project of creating a usable Stata version of the PovcalNet database was originally motivated by an attempt to investigate the implications of using various alternatives to the standard purchasing power parity (PPP) deflators underlying poverty calculations. This issue is pursued in a separate paper which shows that replacing the official CPI series used to calculate changes in PPP over time with independent information from the same survey data used for poverty calculations leads to non-trivial revisions in poverty estimates, usually in the direction of less rapid poverty reduction (Sandefur, 2013).

The general point here, i.e., that the choice of PPP index has enormous implications for measured poverty rates is not news. It can be seen quite starkly in Figure 4. The horizontal axis shows dollar-a-day poverty using 2005 PPP dollars, while the vertical axis shows the poverty headcount measuring using the same line, but denominated in real 2005 U.S. dollars converted at market exchange rates. The two measures line up closely at the extremes of zero and one-hundred percent. But in between the discrepancies are typically very, very large. A country in which 40% of the population lives below PPP$1.25 per day will typically report poverty relative to the same line measured at market exchange rates of roughly 80%.

We focus on the extreme case of making *no* adjustment whatsoever for international differences in purchasing power solely for illustrative purposes. Deaton (2010) and Deaton and Dupriez (2011) demonstrate the sensitivity of international poverty estimates to *a priori* sensible methodological choices in defining PPPs. Nevertheless, we found it cumbersome and labor intensive to explore this issue using the Povcal web site. Our goal here is simply to put data in the public domain that would allow independent analysts to easily explore the implications of using other methodologies.

6 Conclusion

The main point of this short paper is to present the data underlying the World Bank's global poverty estimates in a form that allows non-World Bank researchers some of the same liberties afforded to World Bank researchers. Our aim has been to open up this data set from the confines of the World Bank API, and allowing users to manipulate the database in its entirety using other statistical programs such as Stata or R. All the information in the data set we have described is derived directly from the World Bank's PovcalNet website.

The full data can be downloaded at www.cgdev.org/povcalnet. It contains information from 942 surveys spanning 127 countries. Each survey is represented by 10,000 data points, representing the cumulative distribution function of per capita household income or consumption up to the second decimal place.

Stepping back, we draw a few simple recommendations stemming from this exercise, all aimed at increasing the replicability of global poverty estimates.

First, publish the underlying code. We found it impossible, for instance, to replicate the World Bank's regional aggregates based on the country data from PovcalNet, due to lack of a sufficiently detailed documentation for the process. We also still don't fully understand how the numbers Povcal reports for Latin America are derived, or why the World Bank has chosen to release so much more detail for these countries than for others.

Second, embrace open data formats. The World Bank could dramatically improve the usability of the Povcal data for research – an incredibly valuable resource, harmonized across countries at considerable expense – by providing the data in a truly open data, machine readable format. What we have done here is a partial substitute for that step, which will become outdated over time as new surveys emerge. We hope to see the World Bank make our efforts obsolete.

Third, the World Bank should release not just estimated and smoothed distributions, but full or partial extracts of the underlying data. For many countries, the Bank could post the entire unit record micro data, properly anonymized. For countries which object to the release of the unit record data, the Bank could release only grouped data sufficient to generate the parametric estimates of the Lorenz curve at the core of PovcalNet. In either case, the goal should be to post sufficient data so that independent researchers can recreate the estimates on the Povcal web site from scratch.

The Povcal project has contributed an enormous public good by harmonizing data on the global distribution of income and consumption. The value of this public good will only increase if it is readily accessible in the public domain.

References

ALVAREDO, F., T. ATKINSON, T. PIKETTY, AND E. SAEZ (2014): "The World Top Incomes Database," http://topincomes.g-mond.parisschoolofeconomics.eu, Online; accessed 2014-04-18.

CHEN, S. (2013): "DIY: Measuring Global, Regional Poverty Using PovcalNet," http://bit.ly/1gHSOyl, World Bank, "Open Data" blog post. Online; accessed 2014-04-18.

CHEN, S., AND M. RAVALLION (2010): "The Developing World Is Poorer Than We Thought, but No Less Successful in the Fight Against Poverty," The Quarterly Journal of Economics, 125(4), 1577–1625.

DATT, G. (1998): "Computational Tools for Poverty Measurement and Analysis," IFPRI FCND Discussion Paper No. 50.

DEATON, A. (2010): "Price Indexes, Inequality, and the Measurement of World Poverty," American Economic Review, 100(1), 5–34.

DEATON, A., AND O. DUPRIEZ (2011): "Purchasing Power Parity Exchange Rates for the Global Poor," American Economic Journal: Applied Economics, 3(2), 137–66.

DEMOMBYNES, G. (2013): "Opening Up Microdata Access in Africa," http://blogs.worldbank.org/impactevaluations/opening-up-microdata-access-in-africa, World Bank, "Development Impact" blog post. Online; accessed 2013-08-29.

MILANOVIC, B. (2002): "True World Income Distribution, 1988 and 1993: First Calculation Based on Household Surveys Alone," Economic Journal, 112(476), 51–92.

——— (2012): "Global inequality recalculated and updated: the effect of new PPP estimates on global inequality and 2005 estimates," Journal of Economic Inequality, 10(1), 1–18.

MINOIU, C., AND S. G. REDDY (2009): "Estimating poverty and inequality from grouped data: How well do parametric methods perform?," Journal of Income Distribution, 18(2), 160–178.

MOSS, T. (2013): "BRICN? When Will Nigeria Pass South Africa?," http://www.cgdev.org/blog/bricn-when-will-nigeria-pass-south-africa, Center for Global Development: Views from the Center blog post. Online; accessed 2013-08-26.

SANDEFUR, J. (2013): "Africa Rising? Using Micro Surveys to Correct Macro Time Series," Center for Global Development, mimeo.

WORLD BANK (2012a): "World Bank Announces Open Access Policy for Research and Knowledge, Launches Open Knowledge Repository," Press Release No:2012/379/EXTOP.

——— (2012b): "World Bank Open Access Policy for Formal Publications," Administrative Manual, AMS 14.40.

Appendix: Installing and using Python to download data from PovcalNet

Our Python script uses Firefox and MongoDB to download and store data from PovcalNet. Before running the script, the following programs and packages must be installed:

1. Firefox web browser

2. Python 3.3

3. Selenium webdriver

4. Pymongo

5. MongoDB 2.4.5

The Harvester.py script issues queries to PovcalNet according to the strategy laid out in Section 3 and stores query results in a MongoDB database, povcal. This script can be run from the Windows command line using the following steps:

1. Save the Harvester.py script in the Scripts folder in the directory where Python is installed.

2. Start the Mongod host process for the database. Open the command line (shortcut: enter `<cmd>` on the Windows start menu search bar) and enter the following commands:

   - `cd <Mongodb installation dir>` (for example: `cd C:\mongodb\bin`)
   - `./bin/mongod.exe`

3. Start the Harvester.py script. Open a second command line window, navigate to directory where the Python script is saved and call on the script using the following commands:

   - `cd <Python installation dir>` (for example: `cd C:\Python`)
   - `python harvester.py`

A Firefox window will automatically open and the program will start issuing queries to the Povcalnet website. If the script is stopped or errors at any point, repeat step 3 and the program will pick up where it left off.

Once Harvester.py has started running, data can be downloaded from MongoDB using the Unloader.py script. This script queries the povcal Mongo database and outputs a dataset which contains one observation for each headcount value in the dataset at the lower-bound poverty line value for that headcount (z).

Start the Unloader.py script. Open a third command line window, navigate to the directory where the Python script is save and call on the script using the following commands:

- `cd <Python installation dir>` (for example: `cd C:\Python\Scripts`)

- `python unloader.py`

Figure 1: Detail near the median of CDFs
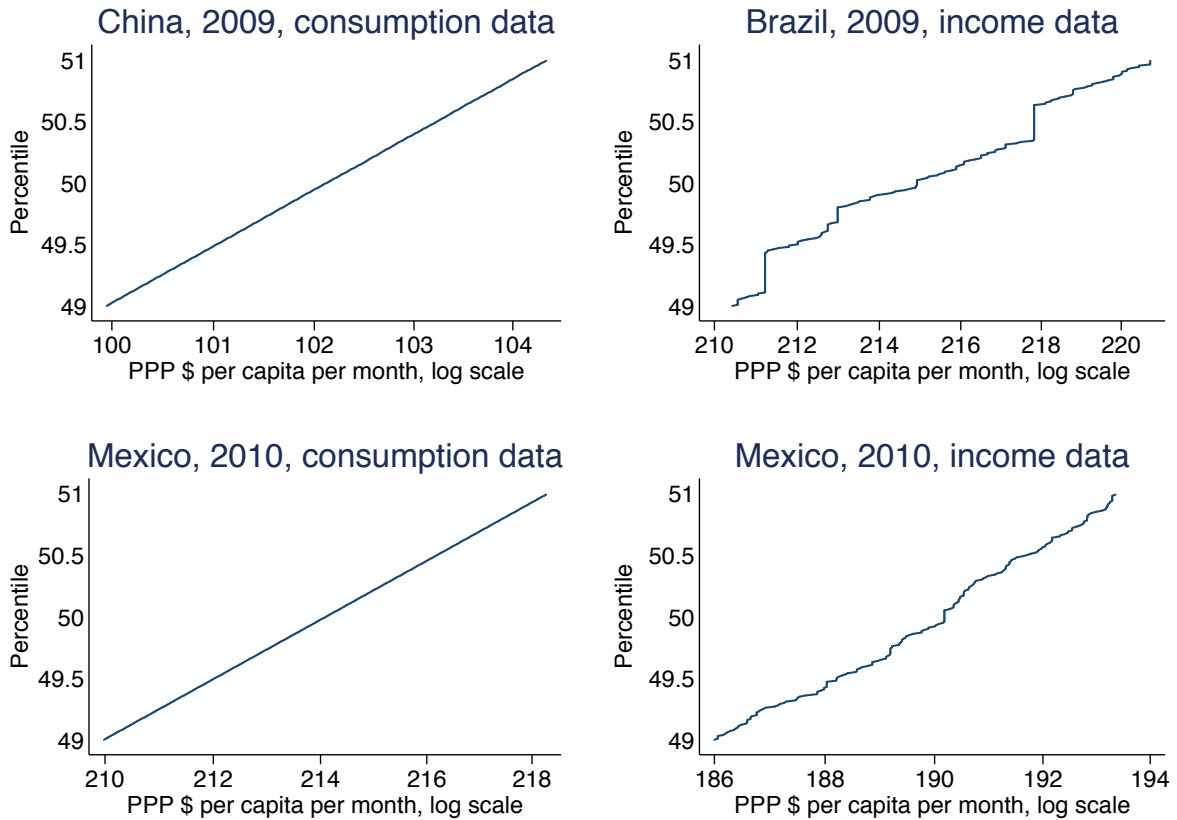
Each figure presents a small slice of a separate cumulative distribution function. Combined, the figures illustrate the greater amount of information available in Povcal for the income surveys from Latin American countries relative to other regions where Povcal uses smoothed parametric functions of consumption data.

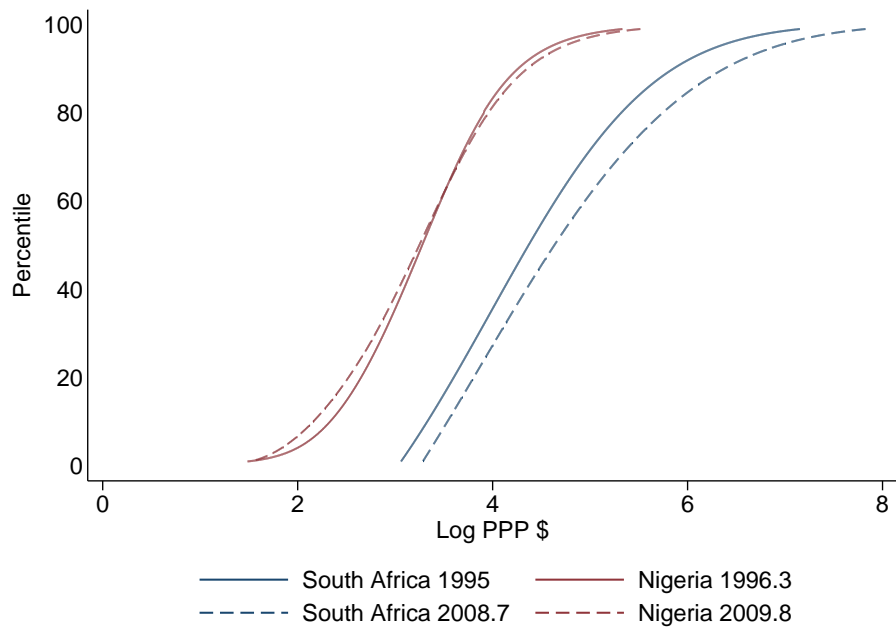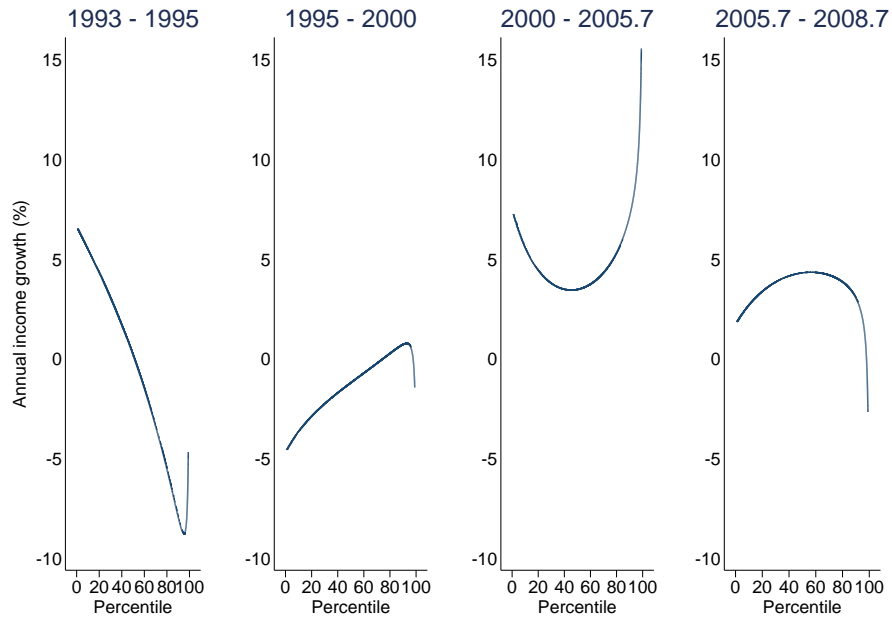Figure 2: Income distributions for Nigeria & South Africa

Figure 3: Growth incidence curves
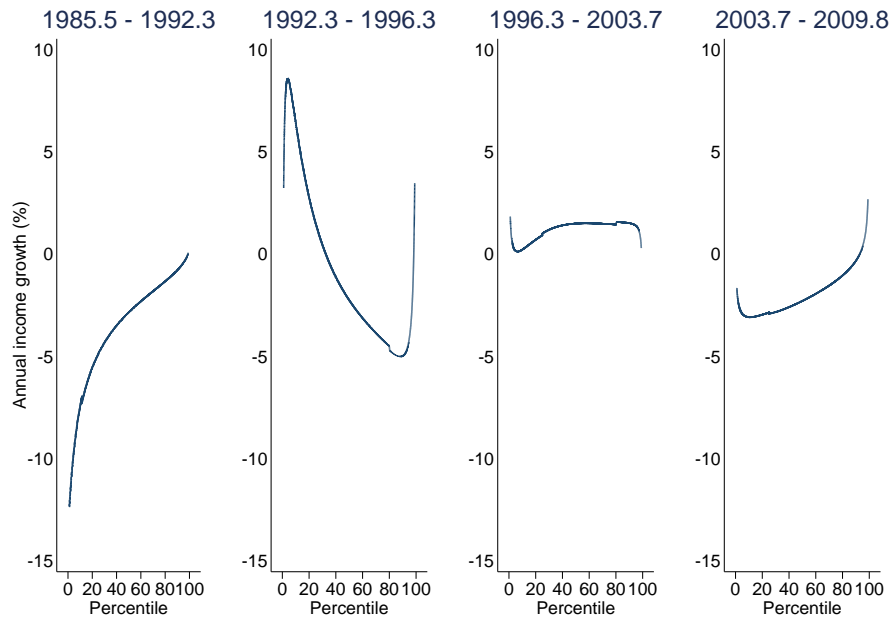
(a) South Africa

(b) Nigeria

Figure 4: Does PPP adjustment matter for poverty rates?

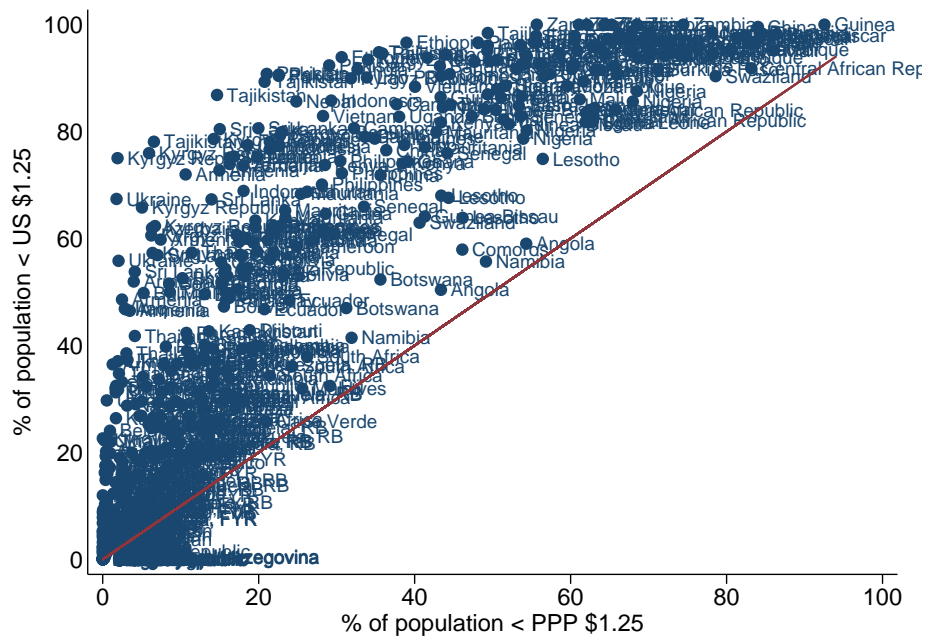Table 1: Number of Surveys by Estimation Method in PovcalNet, 1977-2012

	Explicit parametric form		Other[1]	Total
	Beta	GQ		
By region				
Europe & Central Asia	254	15	2	271
Latin America & Caribbean	48	10	258	316
Middle East & North Africa	37	0	0	37
Sub-Saharan Africa	58	82	0	140
South Asia	27	17	6	50
East Asia & Pacific	61	45	32	138
By type				
Consumption	407	149	38	594
Income	78	20	260	358
By decade				
1970-1979	2	1	1	4
1980-1989	64	29	21	114
1990-1999	140	67	107	314
2000-2009	258	58	157	473
2010-2012	21	14	12	47
Total	485	169	298	952

[1]Includes estimates based on microdata, as well as grouped data and aggregated Lorenz curves.