Now-casting Food Consumer Price Indexes with Big Data: Public-Private Complementarities

Sangita Dubey and Pietro Gennari
Food and Agriculture Organization of the United Nations
Statistics Division, Viale delle Terme di Caracalla, 00153 Rome, Italy
Telephone (+39) 06-5705-5890; Sangita.Dubey@fao.org
Telephone (+39) 06-5705-53599; ESS-Director@fao.org

Abstract: Policy makers, particularly central banks, rely increasingly on big data for information, or "nowcasts", about the current state of the economy, where official statistics, such as GDP and unemployment rates, are available only with a significant lag. Official statistics, however, remain hesitant about adopting or using big data based on concerns about data quality, representativity, and legal issues.

This paper presents the uses of big data in the domain of food prices, from producing official statistics to nowcasts for food security early warning. In the context of private sector data production, it reviews some big food price sources, namely, from supermarket scanners, web-scraping, and crowd-sourcing, with an illustration using Brazilian food prices. It proposes comparative advantages and complementarities of private-public production, particularly in the food security context, concluding that while data quality issues can be addressed, organizational mandates and legislative requirements create more difficult hurdles in public-private partnerships in the official use of private food price statistics.

Acknowledgement: The authors thank Joseph Reisinger and Jonathan Cross of Premise for sharing their data, analysis and methodological notes, and Franck Cachia for his statistical and analytical support.

Keywords: big data, nowcasting, food price statistics, scanner data, web-scraping, crowd-sourcing, consumer price index, mobile applications, scanner data, food security.

Disclaimer: The views expressed are purely those of the authors and may not in any circumstances be regarded as stating an official position of the Food and Agriculture Organization of the United Nations.

1. INTRODUCTION

"Big data," characterized by its three "v's" of volume, velocity and variety, has opened the door to more timely, frequent, detailed and cost-effective data. This enables policy makers to obtain "nowcasts," or current period forecasts, of key economic phenomena, such as GDP growth, unemployment rates, retail sales, and consumer inflation, which better inform fiscal and monetary policy, and serve as early warnings of turning points in the economy (Armah, 2013; Askitas and Zimmerman, 2009; Banbura et al, 2010; Choi and Varian, 2009a and 2009b; Galbraith and Tkacs, 2013; Khan, 2012; McLaren and Shanbhogue, 2011; Wu and Brynjoflsson, 2009). Big data also helps address the limitation that official statistics used for many policy decisions are available only with a significant time lag and lacking the detail and disaggregation required.

The use of big data in nowcasting official statistics raises several important questions. Is there a role for big data in official statistics? As new private sector big data producers appear, what is their role vis-a-vis official statistics? How do these roles vary by type of data, producer, and use?

These issues and questions are being debated in both national and international statistics offices alike (Karlbeg and Skaliotis, 2013; Pierson, 2013; United Nations Global Pulse, 2012, Oct 2013, June 2013; UNSC, 2014). Opponents point out the very valid risks in using private sector big data to produce official statistics, including concerns arising from representativity, data quality, privacy, legal and institutional mandates, and ongoing production. Supporters, on the other hand, point out the benefits of low-cost, low-burden, timely and detailed information, and suggest that the risks raised can be addressed, much in the way that it has with respect to the use of administrative data.

The domain of food price statistics is particularly helpful in contributing to these debates, both because of its history in use of big data, and because of the importance of frequent, detailed and real time food price data in monitoring food security and providing

an early warning of food insecurity. For example, the food price crisis of 2007-08 saw world food prices increase significantly, resulting in an increase in food insecurity, social unrest and political and economic instability. A World Bank report attributes food riots to sharp food price increases (World Bank, 2014). More recently, the Ebola outbreak in West Africa has seen sharp increases in food prices and food insecurity, arising from a combination of factors including travel restrictions and disruptions in agricultural production and transportation.

Though the need for high-frequent real-time food price data is undisputed, official food price statistics are typically available monthly, and only at the end of the month or a week or two after, and rarely with the kind of detail needed for food security early warning, monitoring and policy response. Many policy departments, particularly in developing countries, have adopted big data approaches to "nowcast" food prices and monitor food security and inform their market information and early warning systems. At the same time, many national statistics offices (NSOs) have also started to adopt big data in compiling their official food price statistics, namely, the food and non-alcoholic beverage component of the consumer price index (food CPI), though this has rarely led to more timely statistics.

This paper examines the actual and potential role and impact of big data on food price statistics, particularly big data produced by the private sector. Towards that end, the paper is structured as follows. Section 2 provides a brief overview of the collection and production of the food CPI by national statistics offices (NSOs), which is necessary to evaluate private sector big food data. Section 3 discusses three approaches in using private sector big data to compute or nowcast official food price statistics: 1) retail point-of-sale scanner data; 2) web-scraped food prices; and 3) crowd-sourced mobile app data collection. Section 4 provides an example in the context of Brazil, comparing official Brazilian food CPI data with private sector data provided by the San Francisco based IT company, Premise. Section 5 concludes with some of the public and private sector

comparative advantages and complementarities in collecting and producing food CPIs and other food price statistics.

2. Food price statistics, official food CPIs and nowcasting

Food CPIs, a sub-component of the CPI, are the most traditional food price statistics compiled by NSOs, and much of its importance derives from the importance of the CPI as an official statistic used for both public and private sector decision making.

The importance of the CPI stems from its wide uses as a monetary policy target; to monitor inflation; to escalate social security and pension payments; to index tax thresholds; and to escalate both private and public sector wage contracts. Within NSOs, it is among their group of "mission critical" statistics that receive the highest attention to data quality and timeliness, as the CPI impacts a country's interest and exchange rates, its government revenues and expenditures, and private sector wage compensation bills. And the food CPI subcomponent, as mentioned earlier, is important in its own right in order to monitor price-related food insecurity.

Given its importance, great efforts are made to continually improve the quality and comparability of the CPI within and across countries, with well-established international guidelines and methodologies, and a vast volume of theoretical and applied research. Despite this, all countries deviate from the first best methodology in compiling CPIs, largely due to costs and the need for timely data, with implications for the quality of the food CPI subcomponent in monitoring food security. To examine this, the next subsection provides a brief overview of CPI data collection and compilation.

2.1. A brief primer on the CPI methodology - data collection to compilation

The food CPI follows the same statistical methodology as the CPI itself, with international guidelines provided in the CPI Manual: Theory and Practice (ILO, 2004). Most countries compile the CPI using a Laspeyres-type index, P_L, as follows:

$$P_L = \frac{\sum_{i=1}^{n} p_i^t q_i^0}{\sum_{i=1}^{n} p_i^0 q_i^0} = \sum_{i=1}^{n} \left(\frac{p_i^t}{p_i^0}\right) \frac{p_i^0 q_i^0}{\sum_{i=1}^{n} p_i^0 q_i^0} = \sum_{i=1}^{n} \left(\frac{p_i^t}{p_i^0}\right) s_i^0$$

Where n is the number of commodities, i represents commodity i, t is the time period, $s_i^0 = \frac{p_i^0 q_i^0}{\sum_{i=1}^{n} p_i^0 q_i^0}$ is the share of expenditure on commodity i in period 0, and $\left(\frac{p_i^t}{p_i^0}\right)$ is the price relative of commodity i between periods 0 and t (ILO, 2004). In practice, commodity i refers to a set of individual products, whose individual expenditure weights are unavailable. As a result, the "price relative" for this set of products is itself an elementary price index, generally estimated by the geometric mean of price relatives of individual products (a Jevons index).

A Laspeyres index has the property that quantity or expenditure weights are kept fixed for the base period, 0, to enable the index to measure only the pure price change between two periods. However, since prices are typically collected monthly, while the expenditure weights are typically computed annually, the index is not a pure Laspeyres index, but pertains to the more general category of Lowe-type indices. Furthermore, the expenditure weights or shares pertain to broad groupings of commodities as opposed to detailed commodity breakdowns. Both the weighting and the groupings take into account the high costs of data collection, the former of which is typically obtained through household expenditure surveys. Since consumers are known to substitute from higher to lower priced items, this index is known to have an upward bias, which increases the less frequently expenditure weights are updated. The opposite bias is found for Paasche-type indices, where expenditure weights refer to the current period.

The index methodology used differs from the ideal chained indices (which require monthly expenditure weights), largely due to the high costs of conducted regular expenditure surveys. Furthermore, the expenditure weights are designed to reflect a "representative" consumer, but the lack of coverage of expenditures in small cities and rural areas results in an index that more likely represents the "representative big city" consumer. Due to transportation, storage and post-harvest food loss, it is reasonable to expect that food price inflation is higher in urban areas relative to small cities and rural areas, particularly in developing countries, with the possible exception of regions and products facing import-dependency, where the opposite may be true.

For price collection, the NSO determines the regions covered, which are often only large urban areas; and for each commodity group, the sample of markets or outlets within the region and the sample of products within the outlet for which prices are collected. Commodity groupings are based on the Classification of Individual Consumption by Purpose (COICOP), an international classification system, or some variant thereof. To ensure only price change is measured, prices are collected for the same commodity for the same outlet, which means CPI data are longitudinal in nature.

The selection of the sample of products within each commodity grouping is based on the main commodities purchased within that group in a particular outlet. What is "main" is often based on the advice of the outlet, which may reflect judgment as opposed to statistical evidence. Furthermore, while every effort is made to price the same commodity over time, it is not always possible to ensure that quality remains fixed. Furthermore, CPIs also need to take into account the introduction of new items in the market, and emerging brands sold at higher prices.

While this description does not fully explain the methodology behind CPI data collection and compilation, it does show that the actual practice deviates from the ideal.

The deviation is driven largely by costs, response burden, and the need for timely and reliable CPI data to inform monetary policy.

In summary, while traditional food CPI data collection and compilation is guided by internationally accepted guidelines, its practice suffers from the following limitations, particularly with respect to its use in monitoring food security and warning of turning points. Even when provided monthly, the CPI and food CPI are available with a lag of several weeks, as they are published between end of the month or within several weeks after. Its focus on urban areas leads to lack of representativity of consumers in smaller cities and rural areas, who likely face lower levels of food price inflation, particularly in developing countries, apart from those areas dependent on food imports. The lack of frequent updating of expenditure weights creates an upward bias in measuring food price inflation, because it ignores the fact that consumers substitute from higher to lower priced items of a similar nature. Unless there is some ex-post re-weighting to get at a better representativity of a typical basket of food items purchased, it is likely that official price statistics will overestimate food price inflation. And finally, the food CPI lacks the level of food product detail and geographical detail necessary to pinpoint the products types and locations where price-related food insecurity is likely to occur, a data need particularly importing for food security monitoring and early warning.

3. The use of private sector "Big Data" in computing food CPIs

As mentioned earlier, the types of big data used for computing food CPIs include retail point-of-sale scanner data, data scraped from internet sites, and food price data collected using mobile applications on hand-held devices such as mobile phones. A fourth method, based on internet search queries, is not discussed in this paper. Unlike many other domains in official statistics, where there are still significant methodology and data quality based objections to the use of private sector big data, the use of private sector big

data to compile the CPI provides a counter example, as scanner data is being directly used in CPI compilation by several countries.

3.1 Scanner Data

The use of scanner data for computing CPIs, particularly food CPIs has been advocated by leading price index theorists, such as Erwin Diewert, Robert Feenstra, Denis Fixler, and Jack Triplett (Feenstra and Shapiro, 2003); discussed several times in the international meetings of the Ottawa Group on price indices; featured in the May 2014 ILO-led expert group meeting on the CPI; and advocated by Eurostat to its member countries in compiling the Harmonised Indices of Consumer Prices (HICP). At the national level, it has been tested and/or implemented in several countries, including the Netherlands (van der Grient and de Haan, 2010), Norway (Rodriguez and Haraldsen, 2006.), Switzerland (R. Müller et al, 2006), and the United Kingdom (James and Campbell, 2012).

Scanner data for data obtained at the retail point-of-sale when purchases are scanned by bar-code readers. For food prices, these are typically gathered at supermarket checkouts. The information collected includes the product purchased, its characteristics, the expenditure, and the time and sale location/outlet. Bar codes use detailed classification systems, such as the International European Article Number, formerly the European Article Number (EAN), or the Universal Product Code (UPC), both of which enable a mapping to the COICOP. Using this common COICOP classification is essential for cross-country comparisons of CPIs.

In countries where scanner data is used, advantages include improvements in data quality and representativity of the main items purchased, reduced costs of data collection and response burden, and the availability of near real time data (available with a two to three day delay) as well as a census of transactions for the retailers covered (James and

Campbell, 2012; Mueller, 2006; Siler and Heravi, 2001). These advantages increase when food sales are concentrated among a few retailers, such as Switzerland, where the two biggest retail chains account for 70% of sales, or the UK, where the top four account for over 75% and the top six for almost 90% of sales.

Two limitations do exist. The first is the fact that food sales in mom-and-pop stores are not captured. The second, and more considerable limitation, arises from the risk associated with reliance on the private sector for data collection and maintenance. IT glitches can delay the reporting of data by retailers, putting at risk either the quality of the CPI, or the requirement that the NSO publish the CPI according to a fixed, pre-announced schedule. Private firms can experience financial difficulties, reducing their ability to place effort in activities, such as data sharing, that do not contribute to sales. Where retail sales are concentrated, the impact of missing data is significant, and since the CPI is among the statistical indicators that impact financial markets, these risks are not trivial. While the first issue cannot be resolved using scanner data, a solution to the latter is establish formal and legally binding contracts with private sector providers of scanner data.

While this example is noteworthy in demonstrating how official statistics utilize private sector big food data, its application in developing countries may be limited due to low shares of food sales in supermarket chains that electronically scan bar codes.

3.2 Web-scraped price data, and the Billion Prices Project at MIT

The Billion Prices Project (BPP) at MIT is an example of the use of web-scraping of on-line prices to nowcast the CPI. This project, originally an academic initiative at MIT, collects online prices for millions of items sold by a large number of retailers to produce real-time national inflation indexes for 22 countries, as well as a global inflation index. The indices are sold through the BPP's private sector partner, PriceStats. To protect this bottom line, only data for the United States and Argentina are made publicly available for free, and that with a 10 day lag relative to access available by paying clients.

Some of the advantages of this approach include low costs and response burden, greater timeliness and frequency, more detailed commodity prices collected, coverage of a large number of countries, including developing countries, and the provision of (near) real-time inflation measures. Relative to scanner data, the ability to include developing countries in producing comparable inflation measures is an important advantage (Cavallo, 2013).

For purposes of now-casting, James Surowecki wrote in the New Yorker in 2011: "... after Lehman Brothers went under, in September, 2008, the project's data showed that businesses started cutting prices almost immediately, which suggested that demand had collapsed. The government's numbers, by contrast, didn't show this deflationary pressure until that November. This year, there's been a mild uptick in annual inflation, and again the BPP detected the new trend before the Consumer Price Index did. That kind of early heads-up could help governments make more timely decisions."

As a disadvantage, the indices deviates from internationally recommended practices in that it does not use expenditure weights or the Jevons index for elementary aggregated, and furthermore, collects prices in some countries for a very limited number of retailers and/or cities (e.g. 1 retailer in Argentina in 2012) (Cavallo, 2012). The bigger limitation, for purposes of this paper, is the absence of food price sub-indices, which in developing countries, may reflect the fact that most food prices may not be advertised on-line.

3.3 Crowd-sourced mobile app price data collection

A third big food data source comes from data collected using mobile apps in hand held devices, such as cell phones. For NSOs that equip their trained enumerators with this technology, it becomes analogous to computer-assisted personal interview (CAPI) applications, already widely used by NSOs in data collection, so little more need be said.

The low cost of this technology, however, has also led to its use by non-NSO government departments and private sector firms to collect and publish early warning/market information on food prices. The absence of international/statistical guidelines, however, limit the value of this data source in providing comparable cross-country data.

To expand the value of this tool beyond the usual CAPI applications, organizations have turned to crowd-sourcing as a way of collecting larger amounts of data, with crowdsourcing defined by Wikipedia as “the process of obtaining needed services, ideas, or content by soliciting contributions from a large group of people … it combines the efforts of numerous self-identified volunteers or part-time workers.” Organizations, both private and public, who use this approach to collect food price data have two options. They can use pure crowdsourcing, where the “crowd” determines what foods and markets and retailers to cover. In this case, the use of this data would not be appropriate for food CPI compilation, though it could well monitor food security issues. Alternatively, an organization can allocate its price data collection across the “crowd” to pinpoint specific markets, outlets and commodities, in which case, the approach may approximate the collection methodology most countries used to obtain food CPI data. This latter approach is taken by Premise, though as they correctly point out, this allocation of data requirements is typically not seen as crowdsourcing.

4. Public versus private food price statistics for Brazil

4.1 The IBGE and official Brazilian food price data

Brazil's national statistics office (NSO), the Instituto Brasileiro de Geografia e Estatistica (IBGE), created in 1937, is the main provider of official Brazilian statistics, including its CPI and food CPI sub-component. The IBGE follows the CPI Manual guidelines to produce monthly Laspeyres-type CPIs and food CPIs, and like most NSOs, adheres to the ten Fundamental Principles of Official Statistics, established by the United Nations' Statistical Commission in 1994 (http://unstats.un.org/unsd/methods/statorg/default.htm). These principles include: relevance, impartiality and equal access to statistics; professionalism and accountability in the use and reporting of methods and procedures for the collection, processing, storage and presentation of data; choice over the source of data based on quality, timeliness, costs and respondent burden; and international coordination and cooperation.

The IBGE produces several measures of the consumer price index, which vary based on locations and households covered. The Indice Nacional de Precos au Consumidor (IPCA), used for this analysis, covers ten key metropolitan areas and two municipalities: Belém, Fortaleza, Recife, Salvador, Belo Horizonte, Rio de Janeiro, São Paulo, Curitiba, Vitória and Porto Alegre, Brasília, and the municipalities of Goiânia and Campo Grande. The IPCA includes families dwelling in these areas with monthly income from any source, "ranging from 1 (one) to 40 (forty) minimum wages" (from the IBGE website). This contrasts with another of the IBGE's consumer price index measures, which covers the same geographic areas, but includes only families with monthly income ranging from 1(one) to 5 (five) minimum wages and whose head of household is paid a salary for their main activity. Data collection for the IPCA occurs from day 1 to day 30 of the reference month.

Food CPI subcomponents include meats, fruits, vegetables, fish, fats, beverages, herbs, cereals, processed meats and fish, poultry & eggs, dairy, bread, flours, roots, and sugars. The overall CPI and the food CPI are also published for each of the 12 urban areas covered.

Because the private sector data compared looks at only those foods purchased for food preparation at home, the CPI sub-component of IPCA used for this analysis is the "Alimentação no domicílio," or food-at-home, subcomponent of the food CPI. This is one of the two key groupings of the Brazilian food CPI, with the other referring to food prepared and/or eaten outside the home.

4.2 Premise and its food price data

Premise, a recently established San Francisco based IT firm, collects and compiles food price statistics for Brazil, as well as Argentina, China, India, and the United States, with plans to expand into Africa, starting with Nigeria and Ghana. Its data collection methods include "crowd-sourced" mobile data collection, combined with web-scraped prices. Unique in its approach is the fact that it attempts to follow internationally accepted practices in obtain price data and calculating an index to approximate the food CPI, using official statistics for expenditure weights, adjusted through local analysis of more current food expenditure habits.

For its data collection, Premise has established approaches fairly similar to that used by NSOs to obtain food prices. Both its online data and offline data rely on local experts to identify the key outlets (internet domains or retail outlets) and the main food items, similar to the judgmental sampling used by NSOs. These experts are also used to validate the data collected. For its online data, the Premise web-crawler collect and records pre-tax, pre-shipping prices. For its offline data, prices for food items of interest are "crowd-sourced" using mobile apps downloaded on smart phones, though in reality, the

allocation of work renders their approach similar to a Computer Assistance Personal Interview (CAPI), in which its data collection strategy makes the Premise mobile app a CAPI-type application.

For its offline sampling strategy, Premise uses multi-stage sampling to obtain prices for a minimum of 5 cities in each country for a set of key food products. Each city is divided into coverage areas, with elementary strata defined by store type (by size and chain) and food products. Field workers that form the "crowd" are assigned to a coverage area, and encouraged to collect prices from as many strata as possible. These field workers also submit metadata and photographs for the priced items. Their incentive for data collection and quality is the fee they receive per price quote that passes quality control.

In reality, the approach is not strictly crowd sourcing, as field workers, usually university students, are screened and recruited, obtain field training on both the mobile app as well as the data and metadata required, and are assigned items and locations on which to collect food prices, metadata and product photos. Furthermore, Premise also incents shop-keepers to allow this price data collection by providing them with reports on prices of goods in similar stores in the area.

Combined with its underlying sampling strategy, Premise's approach could be viewed as similar to traditional NSO price data collection strategies where interviewers are hired on contract, CATI applications are used for data capture and quality control, and well designed sampling strategies determine the market, outlet and product for which prices are collected.

In its data processing and index compilation, Premise normalizes product prices for size and quantity, classifies products into categories, performs outlier detection, ensures a minimum sample size in constructing Jevons-type price relatives for its food subcomponent, compiles its aggregate country level "Food Staples Index" (FSI) and

sub-component indices using a Laspeyres-type index, and applies official NSO expenditure weights in the FSI.

As the company and its data series are relatively new, dating back to 2013, Premise currently computes a 7-day and 30-day inflation measure from its FSI, and publishes indices for its subcomponents, such as processes meat, fruit, vegetables, oils and fats, and dairy and eggs. However, it also publishes price levels of key individual food commodities at city level for purposes of food security monitoring, such as potato and green pepper prices in Brazil, or wheat bread and vegetable prices in Buenos Aires.

Premise data and metadata is available to paying clients or can be requested for a trial period.

The FSI price index subcomponents include meat, fruit, vegetables, fish & seafood, oils & fats, beverages, herbs, processed grains, processed meats, dairy & eggs, flours, roots, sugars, grains and nuts, processed fruits and vegetables, sweets, and other snacks

4.3 A comparison of IBGE and Premise food consumer price indexes

So how does the Premise FSI compare with the food-at-home component of the IBGE's IPCA? This can be assessed against the typical dimensions of a statistical quality assurance framework: relevance, timeliness, accuracy, accessibility, interpretability and coherence. These multi-dimensional elements of quality assurance cover the notion of "fitness-for-use," and hence, are typically interpreted from the perspective of the user (Statistics Canada, 2002).

In terms of relevance, the uses and users of IBGE and Premise statistics potentially overlap, but are not identical. The IBGE food CPI is largely used by government to inform policy, though it is also used by the private sector for generic analysis of food price changes as well as food and agriculture researchers. Given the types of food subcomponents available (see Table 2), there is likely other users and uses of which the

authors are unaware. Premise data users include private sector firms, including international banks and hedge funds, interested in monitoring real-time price movements for purposes of corporate, lending and investment decisions, though Premise is also looking to expand users to include governments interested in food prices to monitor food security. The key advantage to IBGE data arises from its indices available for each of its 12 urban areas covered.

The key advantage to Premise arises in terms of food security monitoring from the fact that it provides daily indices in near real-time, as well as prices available for some individual products and their comparison locally and historically. The latter also suggest a potential use by consumers in identifying lower prices, though with a day or two lag, though it is unclear how Premise would obtain revenues from providing this data publicly, particularly given that it would likely lost revenues from clients who currently purchase it.

The key advantage from the IBGE arises from its longer time series, which enables more thorough analysis of historical trends. This also points to a risk created in using private sector food price indices: any risk to continuity in their data collection and index publication will reduce the relevance of their data, given the longitudinal nature of the CPI.

Timeliness leans in favor of Premise, both because the FSI is a daily index, and because it can make available a monthly index similar to the food CPI 10 days before month-end and up to 25 days before the official IBGE release. Indeed, the later analysis will show that the first 7-day average of Premise's daily FSI, available before mid-month, does a reasonably good job of predicting or now-casting the monthly food-at-home CPI.

Accessibility clearly favors the IBGE, which ensures impartial access to all, as per the Fundamental Principles of Official Statistics; while Premise provides access, as expected, mainly to paying clients.

In terms of accuracy, the methodology and strategy followed by Premise appears to be as rigorous as that of an NSO, but as a private firm, its data are not subject to the

international scrutiny or government auditing and quality assurance faced by an NSO. Both indices appear to cover similar food commodity groups, both the IBGE and Premise provide both an overall food-at-home price index and indices for key food subcomponents. If both the IBGE and Premise apply the same expenditure weights, it is expected that household coverage will also be similar. Premise updates these weights with recent and local analysis of current food expenditure patterns, though costs may render this less robust in comparison to official household expenditure surveys. IPCA, on the other hand, covers a broader geographic area than Premise, including 12 urban areas compared to 5 cities covered by Premise, and produces food CPIs for these areas, which may better inform local decision making.

Some degree of assessment of accuracy may be informed by a statistical comparison of the FSI and its subcomponents against the IBGE's food-at-home CPI/IPCA and its subcomponents. This assessment is limited to simple analysis - correlation analysis and simple linear regression forecasts/predictions - given the limited time series in Premise data. Please keep in mind that differences in Premise and IBGE indices can arise from differences in sample size and selection (and hence, in sampling errors), geographic coverage, and food product/item coverage. As a result, one cannot conclude that one set of indices are necessarily more accurate than the other based on index comparisons and analyses alone.

Table 1 provides the IBGE food-at-home CPI and select sub-component food price indices for the IPCA, with indices rebased to June 2013 to facilitate comparison. Table 2 and 3 provide the Premise FSI and select subcomponents: Table 2 provides the daily average index for the first 7 days, while Table 3 provides the daily average index for the first 30 days (except for February, which contains a 28 day average). Monthly indexes were also compiled, though are not presented, for the daily average of the first 15 days, and the the first 21 days (3 weeks), to evaluate prediction accuracy relative to lead times.

The index subcomponents were selected to enable comparison of as similar product groupings between the IBGE and Premise as possible. Month-over-month food price inflation was calculated for the four Premise series, to compare Premise data's predictive power relative to the monthly IBGE food price inflation series.

Chart 1 plots the four series based on the Premise daily FSI (7-day average, 15-day average, 21-day average, 30-day average) against the Brazilian IBGE food-at-home CPI, with all indices rebased to June 2013. All four Premise series track seem to track well the official Brazilian statistic, with similar trends except in July and August 2014, when Premise data shows increasing prices while official Brazilian data shows the reverse.

Chart 2, which shows month-over-month inflation for the five series confirms these results. Again, except for July and August 2014, official data and the Premise series all seem to have similar movements. Between September 2013 and January 2014, it almost appears that official data lag Premise series. This interpretation would not be valid, however, given that both data sets set out to measure the same phenomenon. Most problematic, however, is the July and August 2014 data, in which official statistics show a fall in food prices, while Premise data shows an increase. Though it may be tempting to conclude that the private sector data is faulty, the World Cup event in Brazil, and the news of its upwards pressure on inflation might lead one to speculate if, instead, it wasn't Premise that got it right?

To evaluate Premise data in terms of its ability to now-cast or predict official food-at-home price inflation, a simple linear regression model was constructed for each of the four Premise series (7-day average, 15-day average, 21-day average, 30-day average), with each regression using Premise data as the explanatory variable. Different regression models were constructed, for each series, to predict food-at-home price inflation from April to August 2014. April inflation was predicted using June 2013 to March 2014 monthly Premise indices; May inflation using June 2013-April 2014 data; June inflation

using June 2013-May 2014 data; July inflation using June 2013-June 2014 data; and August inflation using June 2013-July 2014 data.

To evaluate the predictive power of each of the series, a Mean Absolute Prediction Error (MAPE) was computing using the following formula:

$$MAPE = \sum_{t=1}^{n} abs(\frac{At-Ft}{At})$$

Where At is the actual value from IBGE data; Ft is the forecast based on a simple linear regression using the Premise FSI as the independent variable; and n=5 is the number of months forecasted. Table 4 provides the results of the predictions/now-casts, the MAPE, and the lead times for each of the four monthly Premise series.

None of the values of the MAPEs are particularly compelling, with values in the 95% to 97% range based on the forecasts for the five months from April to August 2014, inclusive. More importantly, the signs predicted for July and August are incorrect. Interestingly the 7-day series, produced about 25 days before the IBGE publishes the IPCA, has the same MAPE as the 15-day and 30-day series.

Keeping in mind questions about the July and August data, MAPEs were also calculated for only April through June of 2014, with values ranging from 15% for the 7-day series to 7% for the 30-day series. These predictions have a much more acceptable MAPE, the signs are all correct, and, as in the case of most now-casts, the addition of more information in the Premise series reduces the MAPE. The 15-day series is the most attractive in its trade-off between lead-time of 17 days before official data publication, and MAPE. Waiting to obtain the full month of day (the 30-day series), often published around the same day or a day or two ahead of the official series, only gains a marginal advantage in prediction power. This 15-day series is also compelling, in that many NSOs compile their CPIs and its subcomponents based on data collected during the first 15 days of the month.

This analysis suggests there is some predictive power in the use of Premise's big food price data. If indeed the official food CPI statistics were incorrect in measuring food-at-home price inflation in July and August of 2014, the analysis suggests that Premise data would not only provide a valuable and timely now-cast of food price inflation, but could also help validate official statistics.

5. Conclusion: Public-private sector comparative advantages and complementarities

The descriptions and analyses above lead back to the original questions, particularly in the context of food price statistics: Is there a role for big data in official statistics? And what is the role of private sector producers of big data vis-a-vis official statistics? How do these roles vary by type of data, producer, or use?

In the case of food CPIs, there is no question that there has been and is a role for big data in official statistics, as the use of scanner data in compiling official CPIs demonstrates. This role varies by country, as some NSOs use scanner data directly to compile their food CPIs, while other use it to validate their CPIs, and many do not use it at all.

Statistical organizations have also begun to adopt other big data tools described in this paper, such as web-scraping and mobile tools. Eurostat, for example, is developing a generic tool to collect web-scraped prices to improve its CPI, and Italy's Istat is experimenting with web-scraping and text-mining for its Survey on information and communication technology in enterprises. Eurostat, New Zealand and Slovenia obtain microdata on mobile phone call/text times and positions to enhance their population and migration statistics, which data-sharing legislation in place in Slovenia to obtain this data for free from its private sector providers (UNECE website, Big Data Home).

In adopting tools such as web-scrapers and mobile apps, NSOs need to consider technical, legislative and security issues, such as stability of the application, reliability of

mobile networks, security of confidential information transmitted, and the personal security of interviewers. This leads to the final two questions on the role of the private sector as big data producers, and the varying types of roles.

Again, in the case of scanner data, NSOs already obtain scanner data from private supermarket outlets, for which the key risks arising from IT glitches and delays in data transmission are managed with legal contracts. Furthermore, since the data is collected for a different purpose than an NSO's use, namely, to inform market research and wholesale food purchases and marketing campaigns, and NSOs publish aggregate data, most producers of scanner data do not compromise their business line by sharing this data with NSOs.

For some of the newer private producers of big food data, such as Premise and PriceStats, the considerations are quite different, and stem from their business model, in which the data itself is a key product. Such firms earn revenues primarily from the data, statistics and analysis they provide to paying clients, who receive this intelligence in advance of their competitors or the public at large. The comparative advantage of private sector production in using crowd-sourced and/or web scraped prices likely arises from the lower per unit costs of data collection incurred by specialist firms, as well as their flexibility in modifying and improving their data collection and production processes over time.

Such private firms normally focus on a narrow areas of statistics, for which they recruit and train staff that specialize in one subject matter domain and one set of IT platforms (contrasted with NSOs, where efficiency gains accrue from generic IT platform and knowledge across multiple subject matter areas). The value these private firms bring their clients is in producing (near) real-time, frequent and detailed data with restricted access. Their clients, in turn, benefit from this high frequency, real-time proprietary information which enables them to make decisions ahead of their competitors, or with more detailed product and geographic information than available from official statistics. Not surprisingly,

some of the key clients of Premise and PriceStats include hedge funds and other investment firms, who rely on this type of just-in-time detailed intelligence necessary to inform profitable business decisions.

This business model underlying firms like Premise and PriceStats circumscribes the type of public-private partnership possible. Since NSOs are required by law to publicly provide statistics on their country, economy and peoples, and the ten fundamental principles of official statistics require impartial access and transparency in the sharing of the underlying data collection and compilation methodology (which may be viewed as a trade secret in a private firm), the direct use of such private sector big data in compiling official statistics would likely run into legislative problems and political problems. Since the CPI has the ability to move financial markets, the knowledge that some firms have advanced access to even part of the official CPI would likely create, at a minimum, adverse public reaction. On the other hand, if an NSO could republish the CPI data purchased from a private firm, this would undermine the profitability of such firms who make their income from selling data. Finally, while most NSOs have some financial stability, given their legislative mandates and tax-funding, private firms lack this financial security. In short, if a private sector data producer goes bust, we say “bye-bye” to the data they produce. This is particularly problematic for the CPI and food CPI, which rely on a relatively long monthly series of longitudinal data.

This differences in mandates and financial security does suggest some alternative and complementarities. On the one extreme, NSOs can and have adopted the big data tools pioneered by private sector firms, including the development of web-scraping technologies and mobile apps. In at least one country, and NSO has bought out the private sector pioneer. This does lead to a separate set of questions regarding public sector crowding out of private sector firms.

In the middle of the spectrum, NSOs can use, sometimes at a purchase price, private sector big food data for validation or quality assurance of official statistics. This has been the case in some countries with respect to scanner data. The Premise-based now-casts suggests a similar role with respect to this company's data. Similarly, policy departments can and do use private sector big data to nowcast official statistics, such as GDP growth and employment statistics, with the analysis of this paper suggesting this may be extended to nowcasting official food CPIs. Furthermore, the complementary and timely nature of private sector big food data suggests a role for central banks, finance departments and ministries of agriculture to use this data source to monitor food security, and provide an early warning of key turning points. There is already precedence for such use, as many central banks and policy departments purchase private sector economic forecasts as inputs into their fiscal and monetary policy decisions and to nowcast key official statistics.

At the other extreme lies the purchase of private sector big food data for direct use in compiling official statistics, though differences in business and institutional models and mandates and legal obligations provide the key factors in determining what types of big data and private sector providers can serve this function.

6.

REFERENCES

[1] Michaela Agafitei and Sorina Vaju, *Addressing the Challenge of Producing European Comparable Data using Administrative Data*. Presented at the Seminar on Statistical Data Collection, 25-27 September 2013. Geneva: United Nations Economic Commission for Europe, 2013.

[2] Jose Ramon G. Albert, Big Data: Big Threat or Big Opportunity for Official Statistics? Published by Paris21, http://www.paris21.org/newsletter/fall2013/big-data-dr-jose-ramon-albert, 2013.

[3] Pedro Less Andrade et al., From Big Data to Big Social and Economic Opportunities: Which Policies Will Lead to Leveraging Data-Driven Innovation's Potential? in: The Global Information Technology Report 2014: Rewards and Risks of Big Data, 2014, pp.81-86.

[4] Nii Ayi Armah, Big Data Analysis: The Next Frontier, in: Bank of Canada Review, Summer 2013, pp 32-39.

[5] N. Askitas and K. F. Zimmermann, Google Econometrics and Unemployment Forecasting. Applied Economics Quarterly, 55 (2), 2009, pp 107–20.

[6] Marta Banbura et al, Nowcasting, European Central Bank Working Paper Series No 1275, December 2010.

[7] Benat Bilbao-Osorio et al, The Global Information Technology Report 2014: Rewards and Risks of Big Data, Geneva, 2014.

[8] Danah Boyd and Kate Crawford, Six Provocations for Big Data, in: A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society, September 2011.

[9] Alberto Cavallo, Online and official price indexes: Measuring Argentína's inflation, in Journal of Monetary Economics, 2012.

[10] Alberto Cavallo, Scraped Data and Sticky Prices, MIT Sloan Working Paper, May 2013.

[11] H. Choi and H. Varian, Predicting the Present with Google Trends. Google Inc, April 2009a.

[12] H. Choi and H. Varian, Predicting Initial Claims for Unemployment Benefits. Google Inc, July 2009b.

[13] P. Daas and M. van der Loo, Big Data (and Official Statistics), presented at the Meeting on the Management of Statistical Information Systems, Paris and Bangkok, 23–25 April 2013.

[14] Robert C. Feenstra and Matthew D. Shapiro, Eds, Scanner Data and Price Indexes, University of Chicago Press, 2003.

[15] John W. Galbraith and Greg Tkacz, Nowcasting GDP: Electronic Payments, Data Vintages and the Timing of Data Releases, CIRANO working paper, Montreal, 2013.

[16] International Labour Organization (ILO), Consumer Price Index Manual: Theory and Practice, 2004.

[17] International Telecommunication Union, Measuring the Information Society, Geneva, 2013.

[18] Adam Jacobs, The Pathologies of Big Data, in: Communications of the ACM, 52 (8), August 2009.

[19] Sara James and Richard Campbell, Obtaining Scanner Data Project, presented to the Workshop on Scanner Data for HICP, Stockholm, 8 June 2011.

[20] Martin Karlberg and Michail Skaliotis, Big Data for Official Statistics – Strategies and Some Initial European Applications, presented to The Conference of European Statisticians, Geneva, 25-27 September 2013.

[21] Irfan Khan, Nowcasting: Big data predicts the present, in: IT World, Oct 2012.

[22] Robert Kirkpatrick, Beyond Targeted Ads: Big Data for a Better World, presented at the O'Reilly Strata Conference, United Nations Global Pulse, Oct 2012.

[23] O. Lamont, Do 'Shortages' Cause Inflation? in: Reducing Inflation: Motivation and Strategy, C. D. Romer and D. H. Romer, eds., University of Chicago Press, Chicago, 1997, pp. 281–306.

[24] Clifford Lynch, Big Data: How do your data grow? in: Nature 455, 3 September 2008, pp. 28-29.

[25] James Manyika et al, Big Data: The next frontier for innovation, competition and productivity, McKinsey & Company, San Francisco, 2011.

[26] N. McLaren and R. Shanbhogue, Using Internet Search Data as Economic Indicators, in: Bank of England Quarterly Bulletin Q2, 2011, pp. 134-140.

[27] R. Müller et al, Recent Developments in the Swiss CPI: Scanner Data, Telecommunications and Health Price Collection, presented to the 9th meeting of the Ottawa Group Meeting on Prices, London, 2006, pp. 14-16.

[28] National Research Council of the National Academies, Improving Data to Analyze Food and Nutrition Policies, Washington, 2005.

[29] OECD. Exploring data-driven Innovation as a new source of growth: Mapping the policy issues raised by "Big Data," Paris, June 2013.

[30] Steve Pierson, Big Data: A Perspective from the BLS, in AMSTATNEWS, 1 January 2013.

[31] J. Rodriguez and F. Haraldsen, The Use of Scanner Data in the Norwegian CPI: The 'New' Index for Food and Non-Alcoholic Beverages, in: Economic Survey 4, 2006, pp. 21–28.

[32] Hillary Sanders, et al, The Relationship between Premise Price Data & Official Government Releases,

[33] Monica Scannapieco et al, Placing Big Data in Official Statistics: A Big Challenge? Presented to the New Techniques and Technologies for Statistics conference, United Nations Economic Commission for Africa, 2013.

[34] Mick Silver and Saeed Heravi, Scanner Data and the Measurement of Inflation, in: The Economic Journal 111, June 2001, pp. 383-404.

[35] Michail Skaliotis, The role of official statistics in a big data ecosystem: what will change? presented to the European Central Bank Workshop on Big Data for Forecasting and Statistics, Frankfurt, 2014.

[36] T. Suhoy, Query Indices and a 2008 Downturn: Israeli Data, Bank of Israel Discussion Paper No. 2009-06, 2009.

[37] James Surowiecki, A Billion Prices Now, in: The New Yorker, May 30, 2011.

[38] Statistics Canada, Statistics Canada's Quality Assurance Framework, 2002.

[39] Statistics Sweden, Issues in the use of scanner data, (undated document).

[40] Jack E. Triplett, Should the Cost-of-Living Index Provide the Conceptual Framework for a Consumer Price Index? Brookings Institution, 2000.

[41] United Nations Economic Commission for Europe (UNECE), Big Data Home, UNECE Statistics Wikis, 2014l

[42] United Nations Global Pulse, Big Data for Development: A Primer, June 2013.

[43] United Nations Global Pulse, Mobile Phone Network Data for Development, Oct 2013.

[44] United Nations Global Pulse, Big Data for Development: Challenges and Opportunities, May 2012.

[45] United Nations Statistics Division (UNSD), Fundamental Principles of Official Statistics, 1994-2013.

[46] United Nations Statistical Commission, Big data and modernization of statistical systems. Report of the Secretary-General presented at the Forty-fifth session, 4-7 March 2014.

[47] Heymerik van der Grient and Jan de Haan, The use of supermarket scanner data in the Dutch CPI, Statistics Netherlands, 2010.

[48] Joe Weisenthal, Is MIT's Billion Prices Project Warning of a large spike up in the CPI? in: Business Insider, 25 April 2011.

[49] Wikibon, A Comprehensive List of Big Data Statistics. Wikibon Blog, 1 Aug 2012. Available at http://wikibon.org/blog/big-data-statistics/.

[50] World Bank, Food Price Watch, May 2014.

[51] L. Wu and E. Brynjolfsson, The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales. Sloan School of Management, Massachusetts Institute of Technology, 2009.

Table 1: Official Brazilian food price statistics – the food CPI and select subcomponents

	Food CPI	Meats	Fruits	Vegetables	Fish	Fats	Drinks	Herbs
Jan-13	96.32	104.24	92.50	78.84	103.38	107.76	99.31	95.70
Feb-13	97.89	104.10	92.90	89.62	102.59	107.70	99.44	95.66
Mar-13	99.22	102.41	97.09	97.66	104.41	107.03	100.30	97.90
Apr-13	100.31	100.58	100.23	107.74	105.56	104.84	100.97	99.37
May-13	100.36	99.87	101.25	105.10	102.60	101.99	100.58	99.87
Jun-13	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Jul-13	99.27	100.08	97.40	86.96	99.85	98.25	100.46	99.90
Aug-13	98.93	100.23	97.07	77.40	100.26	97.07	100.90	100.34
Sep-13	98.90	101.11	99.88	69.11	100.40	96.89	101.25	99.10
Oct-13	99.96	104.32	101.88	70.87	101.44	96.37	102.33	97.36
Nov-13	100.37	105.28	102.46	72.81	104.19	96.55	102.85	95.92
Dec-13	101.16	107.73	107.09	74.97	106.78	96.87	103.24	96.90
Jan-14	102.07	111.04	110.77	74.34	113.01	97.45	104.52	97.76
Feb-14	102.30	111.04	113.89	75.38	112.47	97.37	105.06	97.67
Mar-14	104.78	113.54	116.12	91.93	115.91	99.26	105.37	98.03
Apr-14	106.38	115.61	116.78	97.99	119.26	101.85	106.05	98.96
May-14	106.81	116.08	114.22	98.73	118.81	102.98	106.67	100.93
June-14	106.17	116.55	110.19	89.77	115.91	103.15	107.44	101.88
July-14	105.63	116.69	109.54	77.46	113.58	102.26	108.11	103.07
Aug-14	104.99	117.19	107.40	71.64	112.89	99.40	108.22	103.45

Source: Instituto Brasileiro de Geografia e Estatistica. Indexes re-based to June 2013.

Table 2: Premise food price statistics – the Food Staples Index and select subcomponents, first 7 day average of daily indices

	Food Staples Index	Meat	Fruit	Vegetables	Fish & Seafood	Oils & Fats	Beverages	Herbs, spices & condiments
May-13	101.49	100.28	97.35	104.41	99.59	98.48	99.36	100.38
Jun-13	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Jul-13	97.85	98.19	97.20	93.53	98.75	99.17	100.25	100.18
Aug-13	97.69	100.26	96.29	93.00	98.76	97.38	101.69	98.64
Sep-13	99.69	102.65	100.83	92.11	100.13	95.92	103.11	99.10
Oct-13	100.34	104.76	102.31	89.42	100.71	94.88	105.22	99.38
Nov-13	101.68	106.64	102.07	90.00	101.03	95.78	104.58	98.96
Dec-13	102.96	108.45	105.46	91.54	100.77	95.41	105.51	98.18
Jan-14	103.62	109.16	106.65	94.36	104.19	95.83	105.76	98.72
Feb-14	102.54	107.73	104.96	92.07	102.11	95.43	105.30	98.58
Mar-14	103.91	108.61	106.11	96.11	104.70	96.32	105.11	98.49
Apr-14	105.75	109.73	105.68	98.85	106.06	97.53	106.82	99.94
May-14	105.96	109.75	103.99	99.37	108.20	97.76	107.24	101.79
Jun-14	104.87	108.65	100.35	95.54	107.53	97.57	107.17	102.55
Jul-14	105.51	109.34	100.29	92.93	107.32	97.76	107.16	102.95
Aug-14	107.17	108.77	101.28	92.75	107.12	97.83	108.06	104.15

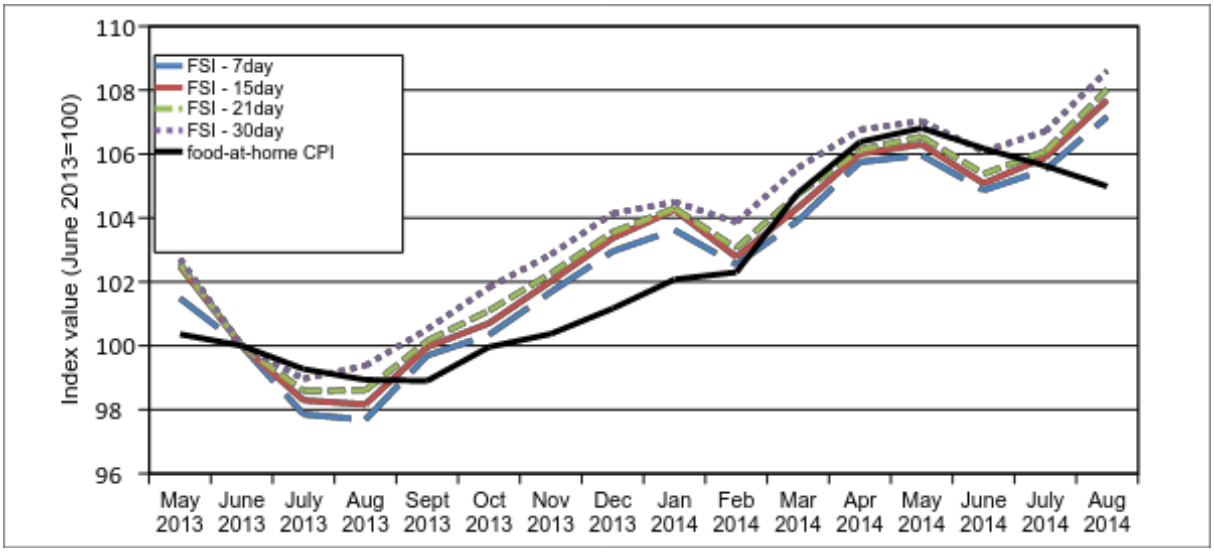
Source: Premise. Indexes re-based to June 2013.

Table 3: Premise food price statistics – the Food Staples Index and select subcomponents, 30-day average of daily indices

	Food Staples Index	Meat	Fruit	Vegetables	Fish & Seafood	Oils & Fats	Beverages	Herbs, spices & condiments
May-13	102.72	101.55	99.86	105.94	104.02	100.27	99.88	100.64
Jun-13	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Jul-13	98.96	99.78	97.80	96.32	99.53	98.78	100.57	100.13
Aug-13	99.38	101.85	99.37	95.90	99.77	97.53	101.30	98.97
Sep-13	100.52	104.40	102.70	92.98	100.70	96.38	103.47	99.38
Oct-13	101.83	107.00	102.68	92.63	101.29	95.64	104.76	98.92
Nov-13	102.86	108.18	104.17	93.03	101.05	96.60	104.28	98.51
Dec-13	104.14	109.94	106.78	94.92	102.44	96.29	105.38	98.19
Jan-14	104.50	110.50	107.09	96.25	105.03	96.48	105.16	98.58
Feb-14	103.87	109.40	107.03	95.40	103.75	96.15	104.65	98.46
Mar-14	105.57	110.02	106.95	100.78	105.82	97.62	105.43	99.25
Apr-14	106.76	110.89	106.38	101.06	107.97	98.46	106.68	100.38
May-14	107.03	110.70	104.20	101.53	108.96	98.51	106.85	102.15
Jun-14	106.08	109.68	101.14	98.08	107.87	98.57	106.87	102.87
Jul-14	106.71	109.91	101.01	94.26	108.60	98.42	106.25	102.99
Aug-14	108.60	111.83	102.90	96.71	108.37	98.27	107.91	104.17

Source: Premise. Indexes re-based to June 2013.

Chart 1: Brazil's food-at-home CPI versus the Premise FSI, Jan 2013-Aug 2014



Data Source: IBGE and Premise

Chart 2: Consumer Food Price Inflation, Brazil's official statistics versus the Premise Food Staples Index, Jan 2013-Aug 2014

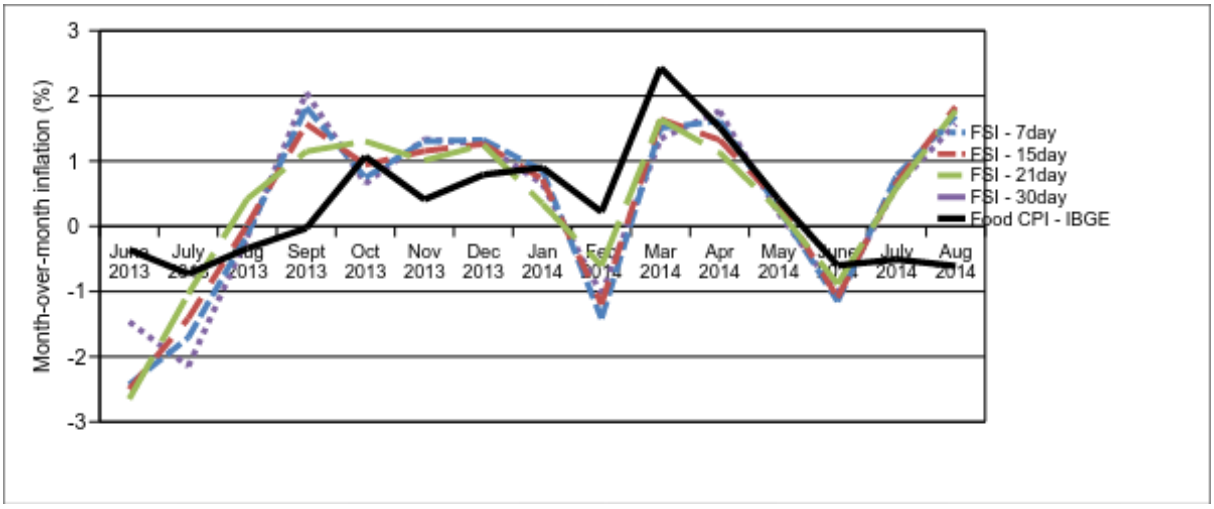


Table 4: Using Premise data to predict consumer food price inflation: Mean Absolute Prediction Errors (MAPE) and Lead Times

	Predicted food inflation using daily average Premise indices (Ft)				IBGE value
	7day	15day	21day	30day	(At)
April	0.81	0.60	0.55	0.50	1.52
May	0.09	0.11	0.17	0.12	0.41
June	-0.44	-0.47	-0.49	-0.43	-0.60
July	0.28	0.34	0.32	0.30	-0.51
August	0.69	0.69	0.83	0.88	-0.61
MAPE, April - Aug	0.97	0.97	0.95	0.97	
MAPE, April - June	0.15	0.08	0.08	0.07	
Lead Time	25 days	17 days	10 days	2 days	

Source data: Premise and IBGE.