Exploiting new technologies and new data sources – the opportunities and challenges associated with scanner data

David Fenwick, International Expert

Group of Experts on Consumer Price Indices – Geneva 2014

Introduction

The retail purchase information that is acquired by electronic devices that read coded tickets (bar codes) on products in shops can generate in virtually real-time, actual sales volumes and prices of items sold. The potential for using this scanner data in the compilation of official statistics is very appealing. It can be cost effective compared with purpose-designed surveys and it can add to the coherence of official statistics by drawing on the same source of data for a range of official statistics such as the consumer price index, average prices, purchasing power parities and, with more caution, retail sales. The price to be paid is the violation of some measurement protocols and conventions relating to official statistics. Issues also arise in connection with the inclusion of some business expenditure. Resilience of delivery of scanner data has been raised as a more general concern. But should these difficulties be barriers to the use of scanner data? The paper reviews the relative advantages and disadvantages of the use of scanner data for official statistics, particularly in computing a consumer price index, and investigates its potential as a source of data, focussing on some of the most pertinent measurement issues including adherence to compilation conventions.

Key words: scanner data, official statistics, measurement protocols, coherence, resilience, cost.

Background

Scanner data is based on electronic point of sale (EPOS) data recorded by bar-code readers at the time and point of purchase. As more shops move over to bar-code readers the potential benefits to compilers of consumer price indices increases. Scanner data provides the potential to deliver up-to-date and accurate information on:

- The number of sales over a chosen period of individual products (usually) uniquely identified by the barcode number.
- The total value of those sales and by implication the average transaction "value". This is normally assumed to equate with average "price" but there are some conceptual differences;
- A listing of the individual characteristics of the individual products concerned.
- An analysis of the above by a range of characteristics relating to the outlet. This can vary according to the scanner data supplier.
- Geographical analysis.
- Analysis by household type (some sources only).

In principle, the sampling error associated with sample surveys used with traditional price collection, particularly at the level of product variety, is avoided. In reality, the market coverage of scanner data varies between different shop types and commodity groups and the

amount and detail of data actually available can vary depending on the commercial source and on the individual product. In addition, definitions may not be compatible with index compilation. For example, scanner data will normally include purchases by business and purchases by non-resident households. The former are out-of-scope of a CPI and latter are out-of-scope if the national[1] concept is followed. Also the average transaction "price" recorded by scanner data, in reality the average revenue per unit of sale, does not take into account the specific needs of index compilers to measure the "actual transaction price" according to a strict set of pre-determined rules that disallow certain discounts such as those relating to damaged stock. Also "returned" goods are often recorded as a "negative" price and items given free when purchasing another product are recorded as being sold at a zero price.

Notwithstanding this, scanner data is being used in the compilation of consumer price indices: as a direct source of price data in its own right (including the computation of Paasche and Fisher Price Indices); for the estimation of appropriate quality adjustments when collectors are forced to select new items with different characteristics than the original; for the estimation of CPI weights; for probability sampling. In addition, scanner data has been used as a diagnostic tool e.g. for the identification of deficiencies in the item sample. But such studies can often highlight the difficulties faced. For example, a benchmarking exercise conducted some years ago involving a comparison of the prices for six products obtained from the then sampling procedures used for the UK CPI, when compared to unit values from scanner data, highlight difficulties in using each dataset. In particular, the then existing sampling techniques ran the risk of not having a representative sample, whereas scanner data had the opposite problem of including price quotes for items not in scope of the CPI, such as damaged goods or end-of-line sales[2]. But this didn't rule out the use of scanner data for local probability sampling as the sampling procedures using scanner data were considered to be more rigorous than asking price collectors to choose a "representative" item in each of the outlets visited[3]. Thus, the choice of data sources will depend on a comparison of the relative merits of each in generating reliable and representative information relating to the target sample.

The potential gains from utilising scanner data can be significant. But at what cost, not only in terms of the resilience of the data set for CPI compilation and the capacity of the national statistics institute to handle and quality assure exceedingly large files of data on sales and "prices", but also in terms of breaking the traditional measurement conventions associated with a CPI. The answer to the latter is dependent, in part, on the target measure. For instance, in the UK there are two main measure of inflation. The Retail Prices Index (RPI) and the Consumer Price Index (CPI). The RPI is an average measure of the change in prices of goods

[1] The "national" concept where non-residents are excluded from the CPI weights but the consumption of nationals abroad - for example while on holiday - is included.

[2] Price Collection and Quality Assurance of Item Sampling in the Retail Prices Index: How can Scanner Data Help? David Fenwick & Adrian Ball, *UK Office for National Statistics* & Peter Morgan and Professor Mick Silver, *Cardiff University*

[3] Static Samples in a Dynamic Universe: the potential use of scanner data and hedonic regression in the compilation of Consumer Price Indices. David Fenwick and Adrian Ball

and services bought for the purposes of consumption by the vast majority of households in the UK. The reference population is all private households with the exception of a) pensioner households that derive at least three-quarters of their total income from state pensions and benefits and b) "high income households" whose total household income lies within the top four per cent of all households. The reference expenditure items are the goods and services bought by the reference population for consumption. Prices used in the calculation of the index should reflect the cash prices typically paid by the reference population for these goods and services. The index is compiled mainly on an acquisition basis, in other words on the total value of goods and services acquired during a given period regardless of whether they are wholly paid for in that period. The main exception is owner-occupied housing where a user cost approach is adopted. In addition, The RPI excludes residents of communal establishments and foreign visitors to the UK[4]. In contrast, the CPI covers all of households and includes the expenditure of tourists and people in communal establishments[5] but does not include owner-occupier housing costs.

This paper poses the question – what is the price to be paid from using scanner data in terms of departing from the target measure?

Returning to the original question about the use of scanner data, we can ask what is the price to be paid in terms of departing from the target measure? This paper does not give the answers, only National Statistics Offices who have been engaged in the use of scanner data for their CPIs and who have comparative data can provide the answers. However, this paper does give an indication of the potential to lead to a departure from the target measure when using scanner data either to generate expenditure weights or to provide "transaction" prices. Such a departure is particularly problematic when it results in bias in the index rather than statistical "noise".

To reiterate and elaborate on the previous paragraphs, the main differences between scanner data and a traditional data set used for the compilation of a CPI are.

- The scope of a CPI is all transactions conducted in retail outlets by private households for private domestic consumption. Scanner data covers EPOS sales only (& therefore

[4] The weights for the RPI are derived from a number of sources but mainly from ONS's Living Costs and Food Survey. The expenditure of certain private households are excluded. Households whose income is within the top 4 per cent of all households and pensioner households which derive at least three quarters of their total income from state pensions and benefits are excluded on the grounds that the spending of these groups are significantly different from the majority. These restrictions are designed to make the RPI more representative of the 'typical household'. Pensioners who are mainly dependent on state benefits are represented in separate 'pensioner indices'. In addition, the RPI also excludes residents of communal establishments and foreign visitors to the UK.

[5] The CPI weights are based on the monetary expenditure of all private households in the UK, foreign visitors to the UK and residents of communal establishments such as nursing homes, retirement homes and university halls of residence. The weights are mainly derived from the Household Final Monetary Consumption Expenditure component of the National Accounts. The CPIH – a new measure of UK consumer price inflation - includes Owner Occupiers' Housing costs using the rental equivalence approach.

excludes sales where bar coding is not used) and can also exclude "own" brands. It does not distinguish between commercial customers and others;

- A CPI measures individual transaction prices according to index conventions. It therefore may exclude conditional discounts (for example, where a "club" card is required), two-for-one offers, personal discounts offered on a one-off basis by shop managers and discounts on discontinued or damaged stock. Scanner data measures average revenue generated after discounts given by whatever method, it will include discontinued or shop-soiled stock and will attribute discounts to the scanner code rather than to the transaction (for example, free video tapes given away with a recorder will be shown as a reduction in average revenue for video tapes);
- A CPI relates to prices charged in a fixed sample of outlets. It is not, therefore, designed to take account of people obtaining improved value for money by shopping in those outlets with the lowest prices. Scanner data, on the other hand, relates to current transactions in all outlets and therefore may include outlet substitution, depending on the particular source of scanner data.

Table 1 examines how scanner data complies with the regulations and guidelines relating to Eurostat's Harmonised Index of Consumer Prices. The numerical impact of breaking the measurement conventions will, of course, vary between countries depending on the structure of retailing. But we can observe a number of instances where scanner data does not follow HICP regulations and guidelines: for example.

- Expenditure coverage is Household final monetary consumption expenditure (HFMCE): expenditure on final consumption goods and services incurred by individual households on their own behalf. This should include institutional households and should exclude business expenditure. Among durables included in final monetary consumption expenditure is the transfer of ownership of some durables from an enterprise to a household – e.g. buying a second-hand car – and the dealer's margin for second-hand cars purchased via garages.
    - Scanner data will include some business expenditure &, possibly, expenditure incurred by Governments or non-profit institutions on goods or services provided to households free or at prices that are not economically significant such as social transfers in kind. But supplementary data collection will be required for some parts of the basket, such as second-hand goods and many services which are not included in scanner data. Thus, there is both a surplus and a deficit in coverage and some measurement conventions will be broken if scanner data is used.
- The prices should be the purchase prices paid by households to purchase individual goods and services in monetary transactions i.e. the actual purchase price including any taxes on the products less any discounts, subsidies and most kinds of rebates. Discounted prices shall be recorded as long as these are non-discriminatory, apply to an individual good or service, are known to the consumer in advance and are given at the time of purchase or close enough after that. Discounts on "damaged, shop soiled or defective goods and purchases for the purpose of sales" should be disregarded or treated according to the rules applied in the context of specification changes. Discounts on "goods close to the expiry date" should be disregarded or treated according to the rules applied in the context of specification changes.
    - Elementary price indices constructed from average "revenue" per transaction as per scanner data will include "returned goods" as a sale with a negative

price and free "offers" e.g. free DVDs with DVD player as sales with a zero price. Discriminatory discounts are included. Scanner data does not separately identify "damaged" goods or goods close to the expiry date..

But in other respects the use of scanner data as a single data source is helpful: for example.

- The HICP follows the "domestic concept" i.e. takes into account the (non-business) expenditures within the geographical boundaries of each Member State, whether made by residents or non-residents.
    - Scanner data includes sales both of the resident population and of tourists from abroad (but it will include business expenditure): most particularly, the types of goods and services purchased by tourists and, in principle at least, purchases in the type of outlets they use.

Despite the potential difficulties associated with scanner data a number of National Statistical Offices use this data to compile parts of their CPI. For instance, both the Netherlands and Sweden currently use scanner data in the compilation of parts of their CPI, whilst other countries, including Denmark and Norway are investigating the possibility. Where scanner data is being used it is generally obtained directly from the Head Offices of supermarket chains rather than through an intermediary such as AC Nielsen. The motivation for this is threefold: cost (obtaining data directly from the supermarket chain can be cheaper); the potential to obtain custom designed data sets (although not meeting all CPI conventions); the facility to address queries directly to the retailer instead of through a third party.

The non-adherence of scanner data to CPI conventions can be resolved, at least in part, by.

- Adjusting CPI protocols provided it this does not jeopardize the objective of measuring inflation.
- Making modifications to the scanner data set to make it conceptually more in line with the measurement objective.

The scope for enhancing scanner data to better meet the needs of index construction is limited in the majority of cases, as it is unlikely that retailers will be willing to go to the cost of modifying their systems if there is no business incentive for them to do so. If a retailer did agree to a change, National Statistical Offices would be asked to bear the costs. The latter could be significant. Thus, in reality CPI compilers are generally left with the first option – reviewing the protocols for index construction. Two instances arise.

- Conventions that will have no significant impact on measured inflation. The underlying challenges here are that.
    - The numerical impact may vary between different countries and between different parts of the basket. The former will impact on comparability and the latter will have a differential impact on CPI sub-indices and could be detrimental, for example, to the analysis of inflation.
    - Measuring the impact and being certain that this will not vary over time.

  It may be argued that the use of "average revenue" rather than transaction price comes into this category. But this needs to be properly tested. The associated decision is

whether to change the recognised conventions (and target measure) or accept that the CPI as computed departs from the target measure but not in a significant way.

- Conventions that can have a significant impact on measured inflation, especially those that may result in bias. Putting aside the issue of identifying which conventions have a significant impact and how that impact can vary, these conventions are far more problematic with no easy solution - they may limit the use of scanner data.
  The departure from Household Final Monetary Consumption Expenditure may be an instance where, at least for some countries, the impact may be significant, depending on the structure of the retail sector, and could more particularly lead to bias.

Additional market information held by the retailer may help to identify problematic areas. For example, supermarket chains may have special accounts for "business" customers that might give an indication of the significance of sales to businesses, both as a whole and by product, and this could be used to estimate the potential impact of their inclusion on measured inflation. For instance, sales of alcohol to businesses could lead to a long-term upward bias in the CPI if excise duty persistently increases above the rate of inflation.

Summary observations on the current position regarding the use of scanner data

The perceived advantages of scanner data include.

- It is a census rather than a survey and therefore will eliminate estimation error.
- It can, in theory, be cheaper as it involves no additional data collection.
- It reduces the burden on data suppliers.
- It is available across a number of countries and therefore increases the scope for harmonisation and international solutions to index methodology.
- It can add to the coherence of statistics if used for multiple purposes such as for computing CPIs, purchasing power parities and average prices, and when used to compute retail expenditure and expenditure weights as well as collecting "prices".

The above advantages need to be weighed against the following disadvantages associated with scanner data.

- It does not comply with all the current methodological conventions for price indices.
- It is not subjected to independent audit. This could be particularly problematical for National Statistical Offices, particularly given legal obligations in some countries in connection with the uses of consumer price indices for indexation. An error in the index might lead to claims for compensation.
- The logistic and quality control challenges involved in handling such large amounts of data.

Notwithstanding the reservations expressed above, scanner data has been used.

- To directly construct some elementary aggregate price indices of a CPI.
- To compute superlative price indices and Paasche price indices.
- To improve the sampling of products priced by the application of probability or quota sampling to control representativity when price collectors are asked to select the most

representative product variety in the shop being visited. Such controls provide a mechanism for ensuring better representation of different brands of hi-tech goods.

- To compute (weighted and un-weighted) hedonic regressions for explicit quality adjustment.

In some instances this has lead to significant and proven improvements in index construction, e.g. in addressing deterioration in the "representative" basket of hi-tech high turnover goods. In others instances the main motivation has been cost saving. In the latter case it is unclear what the cost has been in terms of methodological compromises and of a departure from the target index leading to a miss-measurement of inflation and possible index bias.

Whilst this paper does not subscribe to or oppose the view that scanner data can replace price survey data and expenditure data from a Household Income and Expenditure Survey it does highlight a number of issues which, in the view of the author, should be but have not been fully addressed by the international community of CPI compilers.

It is idealistic to presume that scanner data can be used for all expenditure categories in a CPI basket.

Consideration should also be given to other data sources, such as electronic price lists, which have some of the same characteristics and advantages of scanner data and in some aspects follow more closely the needs and conventions of a CPI, without necessarily having all the challenges associated with scanner data.

TABLE 1: Scanner data and compliance with HICP regulations & guidelines

HICP		Scanner data			Comments
HICP regulation/ guidelines	**Content**	**Fully compliant**	**Partially compliant**	**Not compliant**	
Coverage of the HICP					
Regulation 1749/96[6]	Household final monetary consumption expenditure (HFMCE): expenditure on final consumption goods and services incurred by individual households on their own behalf. Should include institutional households. Should exclude business expenditure. Among durables included in final monetary consumption expenditure is the transfer of ownership of some durables from an enterprise to a household – e.g. buying a second-hand car.			X	Will include some business expenditure & expenditure incurred by Governments or non-profit institutions on goods or services provided to households free or at prices that are not economically significant such as social transfers in kind. Supplemented data collection will be required for the purchases of second-hand goods.
Geographic coverage – Domestic concept					
Regulation 1688/98	Follows the "domestic concept": takes into account the (non-business) expenditures within the geographical boundaries of each Member State, whether made by residents or non-residents.	X			Does not differentiate sales by residence/nationality of customer – all included (but will included business expenditure – see above)[7].
Price concept					
Regulation 1687/98	The price should be the purchase prices paid by households to purchase individual goods and services in monetary transactions i.e. the actual purchase price including any taxes on the products less any discounts, subsidies and most kinds of rebates. Discounted prices shall be recorded as long as these are non-discriminatory, apply to an individual good or service, are known to the consumer in advance and are given at the time of purchase or close enough after that.		X		Elementary price indices constructed from average "revenue" per transaction. Will include "returned goods" as a sale with a negative price and free "offers" e.g. free DVDs with DVD player as sales with a zero price. Discriminatory discounts are included.
	Discounts on "damaged, shop soiled or defective goods and purchases for the purpose of sales" should be disregarded or treated according to the rules applied in the context of specification changes.			X	Scanner data does not separately identify such goods.
	Discounts on "goods close to the expiry date" should be disregarded or treated according to the rules applied in the context of specification changes.			X	Scanner data does not separately identify such goods.
Outlet coverage					
	The selection of outlets shall be designed to represent the purchasing patterns of the reference population. All the outlets from which the reference population makes purchases are in the scope of the HICP, and should be in the sampling frame when drawing the sample of outlets.		?		This depends on the scanner data which is accessed e.g. sales by small independent shops might be under-represented as will most "services". It is likely that the available scanner data will need to be supplemented by other sources.
Weights					
Regulation 1749/96	Weights will relate to final monetary consumption expenditure.			X	See earlier comments. Coverage will be broader than HFMCE.

[6] Member States shall provide information about expenditure included in household final monetary consumption expenditure but excluded from actual HICP coverage.

[7] Does not address issues relating to scanner data and the possible inclusion of internet purchases.

Regulation 2494/95.	The weightings of the HICP shall be updated with a frequency sufficient to meet the comparability requirement laid down in Article 4. This paragraph shall not require family budget surveys to be carried out more frequently than once every five years.....	X			Scanner data has the potential to be used both for constructing (un-weighted) elementary indices & for computing weights. Scanner data has the potential for updating the weights of the relevant elementary indices on a regular basis, e.g. annually, with a minimum time-lag.
Regulation 2454/97	Each month Member States shall produce HICPs using weightings which reflect consumers' expenditure patterns in a weighting reference period ending no more than seven years before the preceding December. Where reliable evidence ..... indicates a weighting change that would reflect the change in the HICP by more than 0.1% point on average over one year against the previous year, Member States shall adjust the weightings of the HICP appropriately**.**	X			See comments above.
	The weight reference period (the period to which the weights relate) should generally relate to 12 consecutive months (insurance apart).	X			Scanner data can be compiled for whatever period is required as it usually has a date marker.
	Price updating where very weight is multiplied by a factor which relates to the change of the price index of the respective aggregate between the two periods and then the respective weights are re-scaled.	Does not apply			Not necessary for scanner data as no significant time-lag in getting data.
	The weighting structure should follow the aggregation structure of the HICP, which is based on COICOP.			X	Recoding to COICOP will be necessary using some form of mapping. The latter will introduce some approximations.
	Stratum weights i.e. by region and shop type. In theory the level of detail will be determined by two factors that will have influenced the initial design stage: the extent to which marked differences in shopping patterns and prices justify this additional stratification; the extent of the available expenditure information.		?		This depends on the particularly scanner data-set. Also see earlier comment about supplementary data relating to outlets not covered by scanner data.
	Seasonal products: variable weights & determining the "out-of-season" period.	X			Scanner data can usually be made available on a regular basis e.g. weekly, monthly & therefore provides a rich source of information on purchasing patterns over time.
	Quality of weights: under-recording of expenditure on alcohol and tobacco.	X			All legitimate sales will be scanned. No potential for under-reporting.
Functional form of HICP					
Regulation 2494/95	The HICP is a Laspeyres-type: a close approximation to measuring current prices weighted by base quantities divided by base prices weighted by base quantities.			X	Indices compiled from scanner data are usually "unit value" indices where there is not a one-to-one relationship between the numerator & denominator of the price-relative i.e. not matched pairs.
Regulation 1749/96	HICPs shall be compiled using either of the ratio of arithmetic means or the ratio of geometric means		?		Such ratios can be calculated <u>but</u> the populations covered in the numerator & denominator will differ because of the lack of individual price quotes.
Minimum standards for sampling					
Regulation 1749/96	HICP should be constructed from the target samples which, for each category of COICOP/HICP and taking		?		This will depend on the number of elementary

	into account the weight of the category, have sufficient elementary aggregates to represent the diversity of items within the category and sufficient prices within each elementary aggregate to take account of the variation of price movement in the population shall be deemed reliable and comparable.				aggregates required to meet the regulation e.g. not all scanner data has a marker for "outlet-type".
Regulation 1687/98.	All COICOP categories that have a weight of 0.1 percent or more have to be explicitly covered in the HICP calculation.	X			Scanner data facilitates this for those parts of the basket where scanner data is available.
	Replacement sampling where one particular product-offer is replaced by another product-offer due to the disappearance or reduced importance of the initial product offer. The treatment of a new product-offer as a replacement for an existing one can be undertaken if the weight of the initial product-offer suitably reflects the sale of the new product-offer and if appropriate quality adjustment methods can be made to link the new price to the old price.		?		Scanner data does not compute price relatives for individual product offers but will generally identify disappearing goods and newly significant ones assuming the product codes are not re-cycled. Scanner data does not always give enough product detail to ascertain quality differences. Supplementary data will need to be collected e.g. from the internet.
Quality adjustment					
Regulation 1749/96	HICPs for which appropriate quality adjustments are made shall be deemed to be comparable. Where quality changes occur, Member States shall construct price indices by making appropriate quality adjustments based on explicit or implicit estimates of the value of the quality change.		?		See above comments
Regulation 2602/2000	Inducements and specification changes. The market value of an inducement temporarily offered to consumers to persuade them to purchase a particular product, may be deducted if known. That market value shall be added back at the time the offer is withdrawn.			X	It is very unlikely that these will be transparent in the scanner data as markers or different product codes are not usually given.
Newly significant goods and services (estimated consumers expenditure of at least one part per 1000)					
Regulation 1749/96	The inclusion of the price changes of a newly significant good or service shall be accomplished within 12 months of their identification either by adjusting the weights of or within the relevant category of COICOP classification or by assigning part of the weight specifically to the newly significant good or service.	X			Again, scanner data facilitates this.
Quality control for Sampling					
Regulation 1749/96	Member States shall… establish & maintain a clear statement of the target sample & shall maintain checks of price observations & price estimates. They shall provide Eurostat with such information to evaluate and ensure compliance.	?			Note: the opportunity to check individual price quotes in scanner data is limited because the data set comprises of "average" prices & therefore hides the detail. Also there is the danger of data-overload.
Timing of prices entering into the HICP					
Regulation 2601/2000	Prices for goods shall be entered into the HICP for the month in which they are observed. Prices for services shall be entered into the HICP for the month in which the consumption of the service at the observed prices can commence.		?		Scanner data will not always indicate when the consumption of a service can commence e.g. an air ticket, theatre ticket. The use of scanner data will have to be restricted to goods.
Treatment of missing price					
Regulation 1749/96	Where the target sample requires monthly observation, but observation fails due to non-availability of an item or for any other reason, estimated prices may be used for the first or second month but replacement prices shall be used from the third month.			X	Scanner data does not use individual price relatives for a unique product offer.

Frequency and price collection periods					
	Continuous price collection for volatile items	X			Scanner data can normally facilitate this.
Data Sources and Basic Information					
Regulation 322/97 (Statistics)	Price collection, compilation and calculation of the national and the overall HICP has to be in line with the subsidiary principle. Member States are responsible for price collection and the computation of their national HICPs.	Not applicable but relevant			The responsibility of Eurostat is to compile the average HICPs for the Community as a whole and for those countries within monetary union. Member States are not obliged to use scanner data or to hand over responsibility for compilation to Eurostat.
Regulation 2494/95	HICPs shall be considered to be comparable if they reflect only differences in price changes or consumption patterns between countries. HICPs which differ on account of differences in the concepts, methods or practices used in their definition and compilation, shall not be considered Comparable.			?	Comparability is unlikely to be fully achieved unless all Member States compile their HICPs using similar scanner data.