Chapter 5 Price Collection and Validation

A. Introduction

5.1. This chapter discusses the principles and procedures followed in collecting prices, the practicalities of price collection, and the processes of quality assuring, validating and editing the prices that have been obtained.

This chapter focuses mainly on traditional methods of price collection where price collectors visit outlets and record prices on paper forms, handheld computers or tablets, or where prices are collected centrally by postal enquiry, email, fax or telephone. However, other methods of price collection are covered in this manual, most particularly web-scraping is discussed in an appendix to this chapter, and chapter 10 examines scanner data.

5.2. It should be noted that this chapter covers operational issues relating to the collection of prices as well as principles concerning pricing conventions such as the treatment of discounts. It begins with the latter and then moves on to the practical application of the pricing rules and the organization and management of the price collection operation. Some operational issues will depend on local circumstances in terms of the structure and nature of retailing. These issues include geography, where large land masses and dispersed populations can add to the challenges of price collection; the presence of retail chains which can make things less complicated with set price lists and significant market share and highly computerised accounting systems that generate sales and price information; and, in some cases, scanner data. Advice on operational issues cannot be overly prescriptive as the challenges can vary between countries.

B. Background

5.3. Several factors will determine the choice of which price collection methods a national statistical office (NSO) uses. Consideration needs to be given as to how best to collect prices in terms of efficiency, accuracy and representation of consumers’ purchasing patterns. For example, local price collection is costly but can have the advantage of covering a wide range of locations and item selections, particularly for food, alcohol, tobacco and durable goods such as clothing, furniture and electrical products. On the other-hand, central collection, either at head office or in regional offices, can be cheaper and may be usefully deployed for products where there are national pricing policies, such as for rail fares, or where prices cannot be observed directly in retail outlets, such as can be the case for many professional services. In terms of representing consumers’ purchasing patterns, the price collection method also needs to reflect methods of shopping. For instance, goods purchased through mail order and catalogues or via the internet need to be properly reflected in the sample. Their exclusion could lead to an imbalance in the sample of prices covered in the CPI as the prices paid and the items purchased may be different from those in physical outlets.

5.4. Bearing the above in mind, price collection will require different practical solutions in different countries according to local circumstances. Thus, the most appropriate sampling and survey methods, and the best data sources for price collection will depend on the structure of retailing in terms of the characteristics of outlets, their geographical spread and the range of goods and services available to purchase, plus the shopping behaviour of the households.
covered by the index. The compiler should always be guided by the principles and objectives of the price index being compiled, as addressed in earlier chapters.

5.5. It should be noted that price collection is becoming increasingly multi-modal with prices being web-scrapped and even obtained from interrogating scanner data, as well as by being collected by hand from outlets, from the internet, and by telephone enquiry. Issues of coherence can arise when integrating price data from different sources. The consequences are not trivial. The internet prices charged by online retailers can differ from the prices charged by traditional outlets. Similarly, the internet prices obtained from a retail chain's online website and the associated sales volume can differ from the corresponding in-store prices. The prices collected from each source will need to be weighted according to the respective values of sales via the different channels to ensure the index properly represents the purchases of the target population. Moreover, products representative of in-store sales may not be representative of internet sales and vice versa. Constructing separate elementary aggregates for each data source, where the latter represent different outlets or outlet types, and aggregating the elementary aggregates using explicit weights relating to the respective turnovers can help to ensure a balanced sample. By way of example, prices from a sample of bakeries (using price collectors) and prices for bread from supermarkets (using scanner data) plus prices obtained via web-scraping have to be aggregated in a correct and coherent fashion taking into account the weights of the different channels. Ideally, indices per data source should be weighted and prices from different sources should not be mixed; however, in practice this will depend upon the availability of data to develop weights.

5.6. The activity of price collection needs to be carefully planned and a number of organisational options can be followed. As mentioned in the introduction, the focus of this chapter is on traditional forms of price collection where price surveys are undertaken in outlets, using paper forms, handheld computers, or tablets, and supplemented by prices obtained centrally from postal, email or telephone enquiry. However, the general principles described in the paragraphs that follow apply to all types of price collection.

5.7. A separate section details the collections of prices over the internet. Additionally, as mentioned in the introduction, the use of scanner data is discussed in chapter 10. Chapter 10 considers the conceptual as well as practical issues associated with making use of these data.

5.8. The relative advantages of central and local price collection will depend on individual country circumstances. A further option involves the outsourcing of certain elements of price collection. In reviewing these options, this chapter covers the relationship between the field and head office and the handling of the flow of information.

5.9. The following section briefly considers the organisational options, prior to the chapter considering the principles of price collection and the practical aspects.

C. Organization options

1. Local collection

5.10. Local price collection involves collectors visiting individual outlets to collect prices for a variety of goods and services. This is the predominant method of price collection in most countries although more countries are now migrating to other methods. The range and
number of outlets visited, and the types of goods and services priced will vary between countries.

5.11. Although the precise method of local price collection will vary, each price collector will usually be responsible for collecting prices from a certain location or from certain types of outlets. Collectors will visit the same outlets in each collection period to attempt to price the same items. Through this type of arrangement, price collectors can build up effective relationships with retailers and gain specialist knowledge.

5.12. There are several important criteria relating to the conduct of the collection. These criteria include:

a. Collectors should be appropriately dressed and polite—they are representing the NSO.

b. They should carry identification to confirm their role and status.

c. They should make themselves known to the retailer or store manager when they arrive, and before they begin collecting prices.

d. They should comply with any request from the shopkeeper whenever possible. For instance, if the store is very busy and the shopkeeper asks the collector to return later in the day and this does not invalidate the sample.

e. The collection should be carried out as quickly as possible, causing minimal disruption to store business.

5.13. Collectors should also follow rules of common sense in preparing for a collection. These will include making sure that they have: spare pens; the appropriate forms; a clipboard; a local map; spare batteries (if the collection is digitalized); money for parking; and wet weather clothing, if needed. A mobile telephone will also be useful.

2. Central collection

5.14. Central shop-collected prices are prices obtained from the central offices of major retailing chains with national pricing policies. Branches of these chains may be excluded from the local collection if data can be collected more effectively centrally. Data suppliers may provide information on paper forms, or by entering price data on spreadsheets and forwarding them to the NSO by email, fax or on external media such as a flash memory stick. Mail order catalogues can also be treated as central shops. These prices are then combined with those for the same items from the local collection. Prices can also be collected centrally online or by using web-scraping or scanner data (see paragraphs 5.200 to 5.232 and, for scanner data, chapter 10).

5.15. Price data for services or fees may be collected centrally from organizations such as trade associations, national or local governmental departments and so on. This is often the case for tariffs, for example for gas and electricity or for telephone and internet charges where data from administrative sources is required on pricing structures and purchasing patterns (see chapter 11)\(^1\). Whenever possible, centrally collected prices are obtained from

\(^1\) It can also be noted that when a CPI basket of goods and services is updated, item weights are updated by drawing on data from a variety of sources including administrative data as well as data from Household Budget Surveys (HBS), detailed national accounts household final consumption expenditure data and market research
The principles of price collection

5.16. The procedures for price collection should follow the general concepts underlying the CPI and should reflect the purchases of the target population.

1. Defining the price

5.17. If the aim of the CPI is to measure monetary inflation, prices are defined according to what consumers pay in the market place, i.e. the actual transaction prices of goods and services purchased by the consumer, including indirect taxes. However, in practice, the advertised price is usually taken as this is the price that can be observed. But the use of this general definition does not eliminate practical difficulties stemming from the existence of, for example, discounts being offered when stock is being sold off because the outlet is about to close or when end-of-line or damaged stock or special stock brought in at sale-time is being offered at a discounted price - in these cases special procedures apply. Another complication that can arise is bargaining, where a price might not be displayed or where goods are purchased on the black market. Procedures to deal with these situations are discussed below in paragraphs 5.34 – 5.40.

Non-transaction prices

...
5.18. It should be noted that there are a small number of exceptions that deviate from the stated aim of measuring actual transaction prices. The most notable example is the treatment of owner-occupier housing costs, where alternative conceptual approaches require different measurement methodologies. The choice of conceptual treatment depends on the main purpose of the index. More details are given in chapter 11. For some services the transaction price may be represented by a tariff or daily or hourly rate. Tariffs are also considered in more detail in chapter 11.

**Catalogue and list prices (not mail order)\(^2\)**

5.19. It can be argued that the use of catalogue or list prices, which supposedly give the prices at which goods and services are sold in outlets, is contrary to the principle of recording transaction prices, as the catalogue or list price may only be the recommended price and not the actual price sold at. Even where it is supposed to be the actual price charged retailers may not always comply. However, in many cases the list price is identical to the transaction price. In practice, list prices can be a cost-effective method of obtaining prices. They can be used, but their reliability should be confirmed periodically.

**Price reductions and related issues**

5.20. As noted above, transaction prices may differ from the advertised prices if, for example, at time of purchase a discount is offered. In practice, however, discriminatory discounts, which are available only to a restricted group of households (as opposed to non-discriminatory discounts that are available to all), are generally excluded on principle. For example, money-off coupons and loyalty rewards for previous expenditure are normally ignored and the non-discounted price is recorded. Also, it may be difficult to obtain the price paid if this is subject to individual bargaining. It may therefore not come as a surprise that, while the application of the general rule above may appear simple, there are a number of cases requiring special treatment either because of conceptual issues or because of practical difficulties. These extend beyond different methods of giving discounts. The following practices are followed by a number of countries and are recommended to deal with different types of price reductions. Price collectors should make extensive notes on the situations confronted so that decisions on how to treat specific cases can be reviewed and confirmed at head office.

a. *Discounted prices*: these should only be taken if generally available to everyone with no conditions attached; otherwise the non-discounted or unsubsidized price should be recorded. The general practice is to ignore money-off coupons and loyalty rewards. A judgement needs to be made, however, relating to the interpretation of “generally available”. For instance, reduced prices for payment by direct debit may be collected depending on the extent to which consumers have access to, and use, such a service. A judgement is required in the latter case on the threshold to be set for access, above which action is taken for inclusion in the index. Alternatively, for tariffs, different payment methods may all be priced individually (for example, separate data collection for

\(^{2}\) The pricing of goods and services ordered by mail can be conducted by reference to mail order catalogues. The index will need to include posting and packaging costs.
electricity payments by cash, direct debit and pre-payment) and the prices weighted together to form a single price index for that item.

b. **Price discrimination**: discounts systematically available only to a restricted group of households should be disregarded because they are discriminatory, unless they are significant and are available either to the vast majority of the population or to identifiable subgroups who qualify for such discounts on the basis of demographic or other characteristics not requiring action by the individuals concerned at the time of purchase. If included, they should be treated as stratification or coverage issues in item sampling. Again, some judgement is required. Examples of allowable price discrimination may include lower prices offered to senior citizens (for example, discounted travel or haircuts) and discounts for people who receive state benefits. Another example of a case where prices are not universally available to all and where judgement is required is when a nominal or token membership fee is required by the retail outlet. In these cases, the take-up of such membership, which is widely available to all, needs to be considered in terms of thresholds and general spending patterns of the consumers and the conditions placed on membership which may make the latter restrictive (for example, minimum levels of purchase). See later section for a discussion on price bargaining (paragraphs 5.34 to 5.40).

c. **Sale or special offer prices**: these should be recorded if they are temporary reductions on goods that are likely to be available again at normal prices or are stock-clearing sales (such as January sales or summer sales). However, before designating a price as a “sale” price, special care should be taken to ascertain that there is a genuine sale with price reductions on normal stock. On occasion, stock is continually sold below the recommended retail price or advertised as a special offer even though these prices are available all year. In such cases, prices should not be considered as sale prices, but should still be collected. Special purchases of end-of-season, damaged, or defective goods should not normally be priced, as they are likely not to be the same quality as, or comparable with, goods previously priced and are unlikely to be available in future. If the special offer is limited to the first customers, the item should not be priced, as the offer is not available to everyone. Introductory special offers may be included if they are available to all. However, given the need to price the same “basket” each month, such offers will not be chosen as representative items unless they have sufficient volume of sales to be considered representative and are introduced at the time of an update of the “basket”, or when a replacement item needs to be chosen. Discounts on goods close to expiry dates should be disregarded or treated as specification or quality changes.

d. **Bonus offers, extras and free gifts**: prices for items temporarily bearing extra quantities (for example, 30% extra free) should be adjusted to take account of the increased quantity unless it can be determined that the extra quantities involved are not be wanted by most consumers, will not have influenced the decision to purchase and will not be consumed. Similarly, “2 for the price of 1” offers should normally be included. The reasoning behind this is that if the CPI is tracking the price of a specific product, say a 330 ml can of a diet soda, and the offer is two cans for the price of one, the shopper will always take the two representing a 50% discount on a single can. However, an offer of three cans for the price of two would be disregarded as this is not the same item. Money-off coupons for future purchases are usually disregarded as these may not be used or wanted. Free items with other purchases (such as a free gift with every product purchased) are generally
disregarded. For example, free gifts such as plastic toys in cereal boxes should be ignored because they are not included in the list for price observations; it is the price to be paid to get the cereal in the box that is relevant. Also, such free gifts can be difficult to value. Similarly, receiving a free toothbrush when buying a bottle of mouthwash would be disregarded because it would be difficult to value the free gift and the consumer may only want the mouthwash and not a toothbrush. The treatment of bonus offers, extras and free gifts is subject to interpretation on a case by case basis. Collectors should be aware that temporary “special offer” weight changes (X% extra free) could become a permanent weight change (for example, cans of alcoholic drinks changing size from 440 ml to 500 ml) and should feed the information back to head office as they become aware of it. In this way, CPI staff can issue new or amended guidance to price collectors about item specifications.

e. **Stamps**: sometimes purchasers are given special stamps, which can be accumulated and subsequently exchanged for goods and services. If a discount is available as an alternative to such stamps, then the discounted price should be recorded. Otherwise, the stamps should be disregarded (see (a)).

f. **Trade-ins**: in general, the price reduction obtained by trading in an old item (for example, a car) compared with the nominal full price should be ignored. This treatment follows convention, as the transaction essentially relates to a second-hand good and only the service charge levied by the outlet in buying and selling the good comes under the scope of the index. In practice, however, the situation is not so clear-cut. For instance, a garage may well give a discount which is greater than the retail value of the traded-in car and, therefore, in effect gives a genuine discount on the new car. In many cases, discounts from trade-ins are very difficult to evaluate. The trade-in value may be negotiable in each case, and the full nominal price – which is used as the benchmark against which the discount is measured – may not be known. It is therefore generally best to report the list price or asking price.

g. **Sales taxes**: when an indirect tax is not included in the price of individual items in a shop but is instead added on when the customer pays for the item, great care must be taken to record the price including tax. To make sure of this, with items for which the price is normally quoted pre-tax, and in areas where a general sales tax is added to the bill, the price collection forms should require the collector to indicate whether or not the price recorded does include the tax, as a price check, so that it can be added where necessary.

h. **Tips for services**: if a compulsory service charge is included, for example on a restaurant bill, only the compulsory amount should be included in the price, but not any additional discretionary tips. For services which are free in principle, but which in practice can rarely be obtained without what amounts to a tip, or where tipping at a standard rate is the common practice, such tips may be added to the specified price. But there is no agreed convention.

i. **Regular rebates or refunds** should only be considered when attributable to the purchase of an individual identifiable product and granted within a time period from the actual purchase such that they are expected to have a significant influence on the quantities buyers purchase. For example, money-back deposits on bottles should be deducted from the price if these are a sufficient incentive for returning the bottle, while money-back
offers on lawn mowers after a five-year period should be disregarded. In all cases, a consistent decision for each item must be applied over time. Decisions about the treatment of rebates are not easy to recommend, as many decisions are made on an individual basis. They may reflect income rather than expenditure changes and may require different treatment for, say, national accounts use.

j. Irregular rebates or refunds: as with regular rebates or refunds, these should only be taken into account when they apply to the purchase of an individual product and are granted within a time period such that they are expected to have a significant influence on the quantities purchasers are willing to buy. Loyalty rebates or coupons associated with previous expenditure at the outlet, to be used for similar or other purchases, should be disregarded, as they are conceptually out of scope because they are discriminatory relying on previous purchases. One-off rebates should be disregarded as they do not relate to the specific time period of the consumption and are unlikely to affect levels of consumption. They are viewed more as a source of additional income.

k. Credit card and other payment arrangements involving interest, service charges or extra charges: charges incurred as a consequence of failing to pay within a specified period of time from the purchase should be disregarded. For example, zero interest as well as positive interest loans granted to finance a purchase should be disregarded when determining the price. These charges come under financial services.

l. Cash back: Whether to include a cash back offer depends on the circumstances. Some large retailers offer credit card accounts for their customers. In this case, a cash back offer could be considered a form of discount. Each time shoppers use the retailer’s credit card, they could earn a percentage of the total amount paid in the form of cash back. For example, if the card pays 2 percent cashback and the customer spends £100 in the retail outlet, they will earn £2. Most cash back cards credit the amount earned by the customer onto their statement, thereby reducing their credit card bill. Like loyalty cards, data collectors should collect additional information to determine what proportion of shoppers use the retailer’s credit card to judge whether the “discount” provided by the cash back offer is widely available or should be considered discriminatory. If deemed widely available, the CPI price should reflect the net purchase price (price less 2 percent cash back). If discriminatory, these offers should be excluded. In other cases, cash back offers are provided by the bank who issues a credit card. The cash back offer provides an incentive for consumers to use a particular credit card for purchases but does not represent a discounted price offered by the retailer and should be excluded. Cash back offers can also be tied to loyalty cards. A percentage of the total amount spent accrues and can be used as a discount on future purchases. These types of cash back offers are excluded because they apply to future purchases and the price paid is not discounted in any way.

m. Dual pricing for cash and for credit or debit card purchases: some outlets may sell goods at different prices depending on whether the item is paid for in cash or a bank debit or credit card is used. The primary objective is to achieve a representative sample and continuity. The price collector should determine the proportions of people who pay by the different methods and this should then be used to judge what prices need to be collected to ensure that the prices collected for the CPI are representative. If the price collector
collects a price based on cash payment during a particular month, then a price based on cash payment must be obtained in each of the following months. If a price is collected based on the swiping of a bank or credit card during a particular month, then the price collector should get the price based on swiping a bank debit or credit card in the following months. Widely available reductions for cash payments may be included, but care should be taken to ensure representative purchases are measured, consistent treatment from one period to the next and adherence to the fixed basket concept. Thus, where relevant, item descriptions should include method of payment and the proportions that pay by different methods should be kept constant between reviews of the CPI basket.

n. **Foreign currency denominated prices.** Where the sale of goods in a foreign currency is common enough practice to mean that its exclusion from the CPI basket would undermine how representative the index is of the prices paid by the average consumer, the price collector should collect prices of items quoted in foreign currency from outlets selling goods in foreign currency. The price should then be converted into local currency using an average foreign currency rate obtained by the head office from the central bank or appropriate dealers in the foreign exchange market. The price in local currency (obtained after conversion) is the one which should be used in the CPI.

o. **Product with a local and foreign currency price.** In some countries there are some outlets which sell products in local currency and foreign currency, i.e. a purchaser is given an option to pay for an item either in local currency or foreign currency. The local currency price is used when purchases are made in local currency. However, if purchases are made in foreign currency, the foreign currency price is converted to local currency as in the previous example. In the event that more items are sold in local currency or vice versa, the enumerator has to ask the shop owner which currency is being used more frequently for making purchases. The currency which is used more frequently in purchasing the item is the one which is used for pricing purposes, unless purchases are frequently made in both currencies, in which case the price in both currencies is priced and a weighted average used after conversion to the local currency.

5.21. To re-iterate, the appropriate practice to be followed will be determined by individual circumstances, which might vary between different countries. Price collectors should make extensive notes on the situations confronted so that decisions on how to treat specific cases can be quality assured and reviewed at head office.

2. **Unavoidable costs that are not part of the advertised cost**

5.22. Some goods are sold in a manner whereby the consumer has no choice but to pay an extra cost, in addition to the advertised price of the product itself, to obtain the use of the product. In cases where most customers will treat and pay this cost as part of the purchase price, the extra cost should be added to the advertised price to determine the price for CPI purposes.

5.23. A common example of this situation is the sale of large household appliances and furniture. The majority of consumers would not be able to provide their own transport to carry these products home. As a result, many stores arrange delivery to the customer’s home for an additional cost. It can be argued that most of the customers must pay for the delivery service, and this additional charge should be include this charge as part of the purchase price.
for CPI purposes. The concept still applies when the transport is provided by another business, if this is part of the transaction. In this case, the cost of transport would be priced separately but added to the cost of the appliance when the index is being compiled. However, if the customer arranges for the transportation under a separate contract then the item should be priced separately under "transport services".

5.24. A similar situation can occur in marketplaces when live poultry is bought for meat. If consumers regularly buy live fowls and then go to other stalls to have the fowls killed and plucked, then the two purchases can be combined to calculate the cost of buying poultry meat.

3. Price bargaining

5.25. Bargaining relates to a situation where prices are individually negotiated between sellers and purchasers and are not pre-determined. The process of negotiation is a characteristic of, for example, some market-places where almost everything to be purchased must be negotiated to arrive at an agreed price, including a wide range of life’s daily necessities that can account for a large part of household consumption. The system of bargaining is characterized by its great flexibility in the setting of prices. Final transaction prices and quantities will vary from one transaction to another and cannot be determined until the purchase has been made. Similarly, there will be variations between transactions in the quality of the goods being purchased which will impact on the price that is paid. For example, the bargained price for fresh fruit is likely to be lower the longer the fruit is on display. Clearly, these special conditions require special methods to determine purchasers’ prices for inclusion in the CPI.

5.26. It can be argued from the viewpoint of the system of national accounts that bargaining is a form of price discrimination. A purchaser is not free to choose the purchase price because the seller can charge different categories of purchasers, different prices for identical goods and services sold under exactly the same circumstances. It follows that “identical” products sold at different prices should be recognized as having the same quality, and their prices must be averaged to obtain a single price relative to calculate price indexes. In practice, the variation in transaction price can rarely be associated with identifiable price-related categories of customers. Rather, purchasers may inadvertently buy at a higher price than may be found elsewhere or could have been finally negotiated. Notwithstanding this, collectors of prices should guard against the presumption that price differences do not relate to quality (or quantity) differences.

5.27. Where prices are determined by bargaining, standard price survey methods, which consist of collecting prices directly from sellers, can generate erratic price indexes that do not reflect actual price movements in a market. The prices that are collected should be ‘bargained prices’ and will depend on the ability, willingness and power of price collectors to bargain in the same way as genuine purchasers and to achieve the same success in obtaining a lower price to get the actual prices paid by members of the public. In addition, prices can vary during the course of a day as well as from one day to the next, adding an extra dimension to the concept of representativeness. A number of survey methods and price collection techniques have been developed to overcome the difficulties inherent in measuring prices that have been bargained.
a. **Survey by purchase of products, simulating a shopper.** The principle is that price collection should be carried out in conditions that simulate as closely as possible situations in which real transactions actually take place. Price collectors behave like regular purchasers by actually purchasing items to be priced and spreading their purchases over the day to ensure representativeness. In each case, the field manager will need to carry out regular checks on quantities and prices obtained by collectors. The following approaches may be taken:

i. Price collectors buy items to determine the relevant price through bargaining. They should be trained to behave as normal purchasers and strive to get the lowest possible price from selected outlets and sellers. Given the high turnover of sellers, the sample of sellers should be partially renewed on a regular basis to ensure that it remains representative and should be chained in as appropriate.

ii. Price collectors buy items and, in addition, are given an incentive to get the best price. For example, a price ceiling may be set, and the collector may receive a proportion of the difference between the ceiling and the bargained price. This incentive system guards against potential difficulties caused by the collector not getting the lowest price because, unlike an ordinary customer, he or she is not concerned with maximizing value for money and is not constrained by income although there might be a limit on the money available for purchasing.

b. **Survey of purchasers.** The prices purchasers have paid are collected throughout the day immediately after the purchaser leaves the outlet or market stall, together with a record of the quantity and quality of the product purchased. The extent of the haggling or bargaining should be determined (for example, opening and closing prices) together with an indication of the relevant parameters determining the price. A form of incentive payment for survey participation may be needed where there is reluctance among purchasers to submit to such time-consuming questions. Determining the quantity and quality of the items purchased can be difficult (see later paragraph).

For the survey by purchase of products and the survey of purchasers, all CPI basket items are subject to bargaining should be covered. The number of prices collected needs to be sufficient both to cover all relevant items and to provide a reliable guide to average price. This may be difficult to determine beforehand, although previous price collections should provide some guidance. It is suggested that price collectors engaging in a survey of purchasers are given a form on which to record the number of quotations per stall or shop, as indicated by the various respondents. This can be used to check the number of quotes obtained against the target number set by head office.

c. **Survey of trends in wholesale prices.** A limited parallel collection of wholesale prices can be a useful supplement for problematic items where the information obtained from the above survey techniques is only partially successful, for example where there is a deficit in the number of observations obtained. Ideally, prices should be obtained from the wholesalers where the relevant retailers get their goods. All factors should be observed which might result in increases in the corresponding retail prices, such as changes in taxes on retail activities, licence fees and the rental for the market stall. On the assumption that these factors remain constant over time, the evolution of wholesale prices may be used as a proxy for the retail prices of relevant items. The price of an item for the current period
would be estimated by multiplying the price of the previous period by the corresponding evolution in wholesale price.

5.28. Determination of the prices paid by a purchaser can be problematic where the final price is for a bundle of items, for example where a stall holder gives the purchaser extra quantities as a bonus for buying a number of goods. If the bonus comprises several categories of items, including the item on which a transaction price was being directly negotiated, then the purchase has to be split into as many sub-transactions as item categories. In these cases, a common-sense approach is needed regarding treatment in the CPI. For example, in a case where some shoppers will be living on a subsistence income and therefore will consume all food purchases even where the food is past its best date. In such cases, the extra quantities involved will be wanted and will be consumed and should be included when calculating the price paid. Purchasers will have actively bartered an overall price for the total basket of purchases, including any “free” goods thrown in.

5.29. The method for determining the price paid by the purchaser is illustrated in the following example: a purchaser wants to buy 5 kg of carrots and is offered a bonus consisting of 500 grams of carrots, 100 grams of lettuce and 200 grams of onions.

5.30. Three transactions can be identified, involving: 5.5 kg of carrots; 100 grams of lettuce; and 200 grams of onions. The bonus has to be valued at prices at which the seller would have sold the items and the purchaser would have bought them. The assumption made is that prices, in local currency units (LCU), would have been determined through bargaining on the same conditions as the price of the item that initiated the purchase (carrots). In this example the opening value of 5 kg of carrots is LCU 15,000 and the closing value LCU 12,000, whereas the opening values of other food products included in the bonus are LCU 990 for a bunch of 264 grams of lettuce and LCU 4,620 for a pile of onions of 4.4 kg. The actual closing price of carrots will be determined as shown in Table 5.1. The actual purchaser’s price of carrots is found to be LCU 2.0967 per gram or LCU 2,096.7 per kilo.

5.31. If the price collector does not know the closing price at which lettuce and onions would have been sold by the seller of the carrots, then it can be estimated. This is done by collecting opening values and standard quantities from a sample of sellers in the same market or at different outlets in the same area. The average opening price of an item is equal to the sum of opening values of the item divided by the sum of relevant standard quantities. For each bonus item (lettuce and onions), the resulting average opening price will be divided by the bargaining ratio calculated on the item needed (carrots) to estimate a closing price for that bonus item. The value of each bonus item is obtained by multiplying the closing price by the quantity offered. If the packet of bonus items contains an item of the same quality as the requested item, that bonus item will be valued on the basis of the closing value of the requested item.

Table 5.1 Method for determining the actual price paid by the purchaser when bargaining takes place
### Requested items

<table>
<thead>
<tr>
<th></th>
<th>Carrots</th>
<th>Carrots</th>
<th>Lettuce</th>
<th>Onions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Opening value of standard quantities (local currency units)</td>
<td>15,000</td>
<td>15,000</td>
<td>990</td>
<td>4,620</td>
</tr>
<tr>
<td>2. Standard quantities (grams)</td>
<td>5,000</td>
<td>5,000</td>
<td>264</td>
<td>4,400</td>
</tr>
<tr>
<td>3. Opening unit price of standard quantities (local currency units per gram) ((1) ÷ (2))</td>
<td>3</td>
<td>3</td>
<td>3.75</td>
<td>1.05</td>
</tr>
<tr>
<td>4. Opening unit price of quantities and bonus quantities (local currency units per gram) (as per (3) above)</td>
<td>3</td>
<td>3</td>
<td>3.75</td>
<td>1.05</td>
</tr>
<tr>
<td>5. Quantities and bonus quantities received (grams)</td>
<td>5,000</td>
<td>500</td>
<td>100</td>
<td>200</td>
</tr>
<tr>
<td>6. Opening value of quantities and bonus quantities (local currency units) ((5) at unit price at (4))</td>
<td>15,000</td>
<td>1,500</td>
<td>375</td>
<td>210</td>
</tr>
<tr>
<td>7. Closing value of items received (local currency units) (assumes bargaining ratio for lettuce and onions same as that for carrots i.e. 12000÷15000 = 1.25)</td>
<td>12,000</td>
<td>1,200</td>
<td>300</td>
<td>168</td>
</tr>
<tr>
<td>8. New price (local currency units per gram) ((7)/(5))</td>
<td>2.4</td>
<td>2.4</td>
<td>3</td>
<td>0.8</td>
</tr>
<tr>
<td>9. Bargaining ratio ((4) ÷ (8) e.g. old price for carrots/new price for carrots i.e. 15000÷12000 = 1.24)</td>
<td>1.25</td>
<td>1.25</td>
<td>1.25</td>
<td>1.25</td>
</tr>
<tr>
<td>10. Payment (local currency units)</td>
<td>12,000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. Estimated closing value of bonus (local currency units) (1200÷300+168=1668)</td>
<td>1,668</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. Actual value of requested item (all carrots) (local currency units) (12000-1668=10332)</td>
<td>10,332</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13. Quantity received of requested item (grams)</td>
<td>5,500</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14. Actual purchaser’s unit price of requested item (local currency units per gram) ((12000-300-168) ÷5500) = 2.0967)</td>
<td>2.0967</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15. Improved bargaining ratio ((4)/(14)) i.e. 3÷2.0967=1.52</td>
<td>1.52</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 4. The principle of the fixed basket

5.32. The important principle underlying price collection is the need to compare prices on a like-to-like basis from one period to the next. This has two consequences.
Where there is a choice of variety of product to be priced initially, an important consideration should be whether that variety will be available to price over a reasonably long period. This is in addition to being typical of what is sold to customers. Note that tight specifications are of no use if the described item cannot be found in the outlets.

A record should be kept of additional information needed to ensure the unique identification of the variety priced so that:

i. The same variety continues to be priced in the case of subsequent price collection being carried out by a different person.

ii. The identification and adjustment for any quality change can be made when the variety disappears and is replaced by a different one.

5. Item specifications

The recording of item specifications is particularly relevant for traditional forms of price collection where products are sampled and where price collection involves price collectors visiting retail outlets to collect prices or where prices are collected by postal enquiry, fax, email or telephone most particularly from central sources.

There are no firm rules, especially regarding the use of loose or tight item specifications: each NSO may choose their preferred methods. However, there are a number of considerations in deciding on item specifications. These include:

a. Loose specifications leave more discretion to the price collector, so the reliability and training of collectors are factors to consider when deciding whether to use loose or tight specifications.

b. Particular care should be taken to ensure that the specification is more detailed for heterogeneous items where there is scope for significant difference between one variety and another, and for items which by their very nature are highly-specified. Cars and hi-tech goods fall into the latter category.

c. Loose specifications not only allow the index to more broadly reflect regional differences in tastes and preferences, but also account for differences due to socio-economic factors as well.

d. Tight specifications allow for the calculation of meaningful average prices.

i. Average prices are useful in identifying outliers and in assessing the accuracy of the reported prices.

ii. Subject to satisfactory sample design, average prices allow comparisons of price levels, including between, regions or urban and rural areas. Tight specifications facilitate the use of CPI prices in the computation of purchasing power parities.

e. However, specifications should not be so detailed that the item cannot be found in an adequate number of outlets.

Responsibility for deciding whether to use loose or tight specifications and for specifying the items to be priced should normally rest with head office. In general, loose specifications can be useful for food, beverages, clothing, and personal items. Tight specifications tend to be useful for electronic and other items with high rates of turnover.
Specifications should be reviewed on a regular basis in order to determine whether they continue to be relevant. A revision of specifications could be implied by:

a. A large number of missing price quotations.

b. A large number of substitutions.

c. A wide variation in the distribution of collected price levels.

d. An observed distortion in the achieved sample, e.g. where an item is stocked only by one retail chain indicating that the item is unlikely to be fully representative of what households buy or where there is a reduction in the number of brands available possibly indicating a fall in demand.

5.36. Practical difficulties stem from the existence of a temporal dimension in price collection. Issues such as the timing of entering prices into the index, the treatment of missing prices, seasonality and the frequency of price collection are, in essence, temporal issues which need to be addressed to maintain the quality of the index. Aspects relating to the frequency and timing of price collection are addressed in the following sections.

6. Frequency

5.37. Decisions about the frequency of price collection are governed by several factors. The most important are: the volatility of prices; the characteristics of the market from which the prices are collected; the known regularity of price changes; and the frequency and method of calculation of the CPI. The general principle is that each item should be priced as often as is necessary to ensure that the index reflects a reliable and meaningful measure of price change. Thus, the desired frequency of price collection will vary by product, depending on how frequently the prices to be observed change. For instance, the prices of some products (such as fees for government services and utilities) might need to be collected only once a year if it is known that prices are reviewed annually at a regular point in time. However, it is advisable for the price collector to occasionally check that the assumption of regular and predictable reviews still holds. In contrast, the prices of products with more volatile prices, such as fresh food, might have to be collected more frequently than the frequency of index calculation and publication. Also, price collection will need to reflect user needs in terms of whether the target index is a point-in-time or period-of-time index.

7. Period-of-time or point in time price collection

5.38. NSOs should strive to calculate an index based on prices covering the whole period (e.g. a month). Price collection should be organized so that prices are collected from different outlets throughout the month. This ensure that the prices used to calculate the index more broadly reflect the average price for the reference month. While resource constraints could restrict price collection to a specific period during the month, every effort should be made to maximize the price collection period.

5.39. If the CPI is used for deflating income, expenditure or sales, the index should relate to the time period of these money flows. For economic analysis, where the index will be used in conjunction with other economic statistics, most of which relate to a period rather than to a point in time, it seems logical that the CPI should do the same.
5.40. Under the traditional approaches to price measurement, where price collectors visit retail outlets and record prices, spreading price collection over a period of time, will result in a more even workload. This avoids some of the operational problems and inefficiencies associated with point-in-time price collection but will involve collecting sufficient price quotations to obtain a reliable average price level for each product over the period.

5.41. From an operational perspective the uneven workloads associated with point-in-time price collection when using traditional price collection techniques, can be inefficient and have a negative impact on price collector performance at peak times. For instance, a sizeable price collection team will be required for a short length of time in each collection period. This has implications for the recruitment and deployment of price collectors and also for management of the fieldwork. Communication and planning between collectors and their supervisors to manage events like absence due to illness must be prompt and effective. Staff at head office will similarly be confronted with a heavy workload of data checking and editing of price data over a short period of time.

5.42. With the point-in-time approach, major price setters, notably the government, can influence the index according to whether their price changes take effect on a day just before or just after the day for which their price information is obtained, or on the day of collection. Since prices are often collected centrally from such price setters, it should be possible to obtain information from them about both the amount and timing of price changes at the end of each month, so that in applying the period-of-time approach, an average price for the whole month can be calculated. For example, if electricity charges are made quarterly and prices increase part way through the three-month period, individual customers’ payments could include zero, one, two or three months at the higher rate depending on when the price change is implemented.

5.43. The point-in-time or period is chosen to represent a pre-specified reference period, normally the calendar month in which the price observations take place. Whether collection is continuous or point-in-time, the interval between successive price observations must be held constant, e.g. by visiting an outlet during the same fixed time period each month or quarter.

5.44. The sampling variance will differ according to whether a period or point-in-time index is compiled and the frequency of collection. In considering the timing and frequency of price collection, consideration also should be given more generally to the trade-off between statistical accuracy and cost, particularly when prices are collected by visits to local outlets as this can be a costly activity. The budget for price collection can limit the available options.

5.45. The timing of the publication of the resulting price indexes can act as a constraint on the price collection schedule and vice versa. For example, there may be legal constraints on the timing of the publication of indexes such as a requirement to publish at a set time each month. In such cases, prices must be collected to a schedule that allows quality assurance, processing and aggregation procedures to be completed before the deadline. The quality management of the price collection process is covered in paragraphs 5.95 to 5.134.

8. Timing of price collection
5.47. The interval between price observations should be uniform for each outlet. Since the length of the month varies, this uniformity must be defined carefully, for instance, not by date but by a formula such as "second Monday in the month".

5.48. Regular timing is important, particularly when inflation is rapid. Where there is a specific collection day, it is important that the most volatile prices are collected on that specific day rather than on the days around it. Particular items, where prices can be more volatile, include fresh fruit and vegetables, fresh meat and fish, and petrol. In the case of food products sold in market-places, the time of day as well as the day of the week is important. Prices can be higher in the mornings when produce is fresh and lower in the evenings. Thus, the precise timing of price collection is particularly important for fresh produce. Prices of fresh fruit and vegetables and of fresh meat and fish may be cheaper in the evenings especially if there is limited refrigeration.

5.49. When inflation is low and stable there will be little difference between the annual inflation rate rates based on point-in-time and period-of-time collections. For example, there is likely to be very little difference between the annual rate of change in the index from Monday, January 9, 2017 to Monday, January 8, 2018 and the corresponding annual rate of change between the complete months of January 2017 and January 2018. This will not be the case if inflation is high or the inflation rate changes significantly during the year. The difference between January 1 and February 1 and average January to average February inflation rates may be different, particularly if so-called “sale” periods fall on regular dates each year or are limited by laws or local ordinance, as they are in some countries. For certain products with high index weights, where price changes are sudden and tend to affect the whole market on about the same day, the choice between point-in-time or period-of-time price collection is especially influential. Examples are petrol, electricity and telecommunication prices. Here it can be argued that the case for taking an average price for the period is stronger. Where this is done, the calculation of the average monthly price should relate to the periodicity of the collection, taking into account the appropriate pricing periods. For example, if prices are increased a third of the way through the period then two-thirds of the average price over the month should reflect the higher pricing. In these cases, different areas or locations should be scheduled for price collection at different times of the month according to a regular pattern to be repeated each month. This not only makes the use of the collectors’ time more efficient, but also has the advantage of providing a spread of collection dates for many representative items. It is also important that individual price observations are carried out at the same time each month so that the index does not move because of a change in the length of interval between collection dates. A further important consideration, of particular consequence in some countries, are cases where prices can vary by day of the week (for example, depending when market day is) or time of the day according to various marketing special offers to attract more customers at less busy times or to reflect the freshness of the goods.

5.50. Preferably, days of the week and times of the month should be chosen considering when purchases are concentrated and where prices and goods in stock are known to be representative of the month as a whole. In some countries, results of the HBS suggest that most households do the shopping on the day of the market or souk. However, retailers may be
Before they are busy, so a balance needs to be struck between the ideal timing for collection and the impact on response rates. It should be noted that an entirely fixed interval is impossible because of the varying length of a month and the timing of public and religious holidays. One solution is to take sequences of four and five weeks, so maintaining a relatively stable monthly or quarterly period where price collection takes place on a fixed day or days each month; another is to follow a rule such as collecting prices on the regular market day or on a Wednesday through to Friday of the first full week in the month.

5.51. Price collection days (and sometimes the time of collection) need to be set in advance and it is important for public perception regarding the integrity of the index that the NSO explains the procedures used for setting collection dates and the underlying objectivity of the method. Any data suppliers who supply prices directly to head office staff need to know the collection date at least a short time in advance to be able to prepare and supply the necessary price returns. To enhance transparency, it is recommended practice to include the price collection period in the CPI metadata.

5.52. Price collection days need to be set after considering a variety of factors that affect prices and shopping patterns. Holidays and weekends should be avoided except for items with large sales during these times, such as petrol, leisure services and entertainment (for example, restaurant meals and tourist attractions). Some countries have limited shop openings on some days or half-day closing on other days, which can limit the number of prices that can be collected or bias the location sample to certain types of outlets or service providers.

9. Taking account of hyperinflation

5.53. Special arrangements may need to be put into place where there is hyperinflation. Hyperinflation is an economic phenomenon that occurs when inflation increases very rapidly. Although there is not a precise number that defines hyperinflation, economists typically consider monthly inflation rates of above 50 percent as hyperinflation episodes. In these circumstances it becomes even more important that the prices of individual items in individual outlets are collected at precisely the same time each month, otherwise misleading figures may result. Consideration should be given to the more frequent collection of prices and, correspondingly, a more frequent compilation of the index. For example, where prices are normally collected on a quarterly basis, it may be sensible to collect them more frequently. If this is not feasible, it may be appropriate to adjust prices proportionately by some relevant indicator, such as a subset of the index, to provide an approximation to a monthly index. If this is done, however, great care needs to be taken in choosing the appropriate comparator, particularly as relativities between prices i.e. comparative prices can change dramatically in periods of hyperinflation. The same considerations apply to collecting prices online.

5.54. In some circumstances, rapid or frequent price changes may be associated with certain items only and action should be taken accordingly. For example, food prices may rise disproportionately because of a bad harvest and it may be sensible to increase the frequency

---

3 There is no rule of thumb on what constitutes an acceptable response rate as this will depend on the sample design that has been adopted and the structure of the retail sector (most particularly, the variability of product and item varieties available and their prices and the range of outlets in which they are stocked).
of the index for food items only, possibly publishing a separate index. Alternatively, a simpler way of dealing with this situation may be to monitor a small number of relevant prices on a regular basis without producing a full price index. Such sub-indexes can be published separately or used to adjust the index as mentioned above, or to provide background briefing for analytical purposes. These items may be chosen according to their importance for the household budget and whether they are particularly susceptible to big price increases.

E. The practical aspects of price collection

5.55. As mentioned in the introduction, the most appropriate sampling and survey procedures will vary depending on the use of the CPI and on national circumstances. This also applies to the practical aspects of price collection. The following sections look at aspects of planning and organisation and provide indicative guidance on the processes and procedures that contribute to successful price collection in the field.

1. Practical procedures for local price collection: planning and organization

5.56. For the sake of simplicity, the discussion that follows is based on period-of-time pricing rather than point-in-time pricing (see earlier references), but the concepts discussed will generally apply to both collection methods.

5.57. The procedures governing the collection of prices need to cover not only the requirements of obtaining usable prices from the outlets, but also the practical problems faced in managing travel to and from the various localities, the transfer of data, and the validation of these data back in the office. The overall operation can only be achieved by cooperation between the price collectors, their supervisors and, of course, the retailers selected for the price survey.

5.58. An overview of local price collection is given in Appendix A at the end of this chapter. It is in the form of a flow chart and shows the different situations that confront the price collector and the decisions or referrals to head office for further action.

Planning the collection schedule

5.59. The collection schedule should include sufficient time for a collector to travel around all required locations within a reasonable number of working hours in a day. Allowance will also need to be made for the price collector to perform all the necessary checking of prices and to answer queries from the supervisor or head office, and where necessary re-visiting the outlet concerned. The collection schedule also has to allow for the transfer of data and forms between the data collectors, regional offices (if applicable), and head office. If paper forms are being posted or hand-delivered, time must be allowed to ensure that all information reaches its destination by the required deadline. If electronic transmission of data and forms is used, time needs to be allowed to ensure that all the data has arrived in the correct format and to address potential problems in data connection leading to files being corrupted. Also, plans for recovery from technical breakdowns need to be in place.

5.60. Public holidays can occur on days that would otherwise be price collection days. In general, prices due to be collected on the holiday should be collected as close as practicable to that day. This usually means adjusting the regular price collection schedule to the
immediately preceding or following few days. Any adjustments to the regular schedule should be made in such a way that prices reflect the normal pattern of buyer-seller behaviour. Consideration also needs to be given to the work schedule for index compilation and publication.

**Dealing with queries: enquiry management system**

5.61. Price collection queries need to be dealt with in a timely and efficient manner both because of the tight schedules associated with the compilation of a CPI and because it is very difficult to correct errors in pricing retrospectively given the dynamic nature of retailing where prices and stock can change very quickly without warning. The progress in dealing with queries needs to be monitored and the system for monitoring must be simple, in order to be effective and flexible to the needs of the CPI production cycle and to any problems that may arise. An enquiry management system should be able to monitor progress and provide an effective audit trail. Information recorded is likely to include: date price collection form received and who from; date due; collection date/trip/period (if applicable); current progress (with date); dates when queries were raised and responses sent; and date completed. It is particularly important that decisions are recorded and signed-off to provide price collectors with confirmation for their own purposes of verification and for audit trails for quality assurance purposes at head office. Audit trails are an effective way of ensuring that processes have been carried out correctly and for reviewing the effectiveness of those processes. The information gathered should also feed into a quality management system for the computation of the CPI (see chapter 13).

5.62. A simple system might use emails to send the query. As these will be seen and read they satisfy the notification requirement. A template form requiring simple ticks and dates to be entered could be created in a spreadsheet and a new copy printed for every collector each collection period.

5.63. An enquiry management system can be either paper or electronic. The two methods should be matched to the resources and infrastructure available. For instance, messages to and from price collectors to their supervisor might be on paper but messages between a regional office and head office could be by email.

5.64. The raising of queries about collected prices is a two-way process. Head office may want to question some prices collected and request a check while price collectors may ask for further guidance on situations arising in the field such as the selection of a suitable replacement for a disappearing good.

2. **Practical collection procedures: collecting prices in the field**

**Price collection, price collectors and questionnaire design**

5.65. It should be re-emphasised that price collection procedures should adhere to the definition of the CPI. It is important that the items priced are representative of what people actually spend their money on. Also, prices need to be collected at different types of shops and outlets, from supermarkets to corner shops and outdoor markets and souks in urban and rural areas, throughout the country. In this way the index properly represents the various places where people go shopping. The items chosen within each shop should be representative of that shop and should be listed in the price collection questionnaire. It is also important to try to keep a consistent routine of price collection from month to month: ideally,
individual prices should be collected on the same day each month and, particularly in the case of fresh produce, at the same time of day even when the pricing schedule for the basket as a whole is spread over a number of days or over the full month.

5.66. Price collection is a complex process that needs to be effectively managed by head office. The proper training of price collectors in the concepts and practices of price collection is an essential element in achieving a representative and error-free sample of prices. Price collectors should also be given help and practical support when undertaking price collection. The collection of prices should be supported by good documentation. Also, the scheduling of price collection and the drawing up of recovery plans to deal with, for example, price collector illness or other factors outside the control of the price collection team, should be in place so that the price collection operation goes smoothly. The value of the effective management of the price collection process cannot be over-emphasised and should be given the necessary attention to reflect its importance. More detailed guidance on best practice is given in the paragraphs that follow.

**Design of price collection questionnaire**

5.67. Good design of the questionnaire form (or its electronic equivalent) is essential for the successful collection of prices. Not only is it important that the price collectors are given the appropriate direction and find it easy to use, but the format and layout should easily facilitate the extraction of data (price, item description, comments, and so on) by head office for effective quality assurance.

5.68. Whether recording price information obtained from personal visits on paper form or via handheld computer or tablet, the price collector needs to be provided with all the information required to successfully collect prices and, correspondingly, head office will want from price collectors all the information needed to quality assure the prices collected. The price collection form should cover: collection date and collector’s name; name or location of the particular outlet; product name and specification of the actual item to be priced; price entry; collector comments about the product or price movement or changes in the representative item being priced such as package size.

5.69. The optional inclusion in the form of the price collected in the previous period raises concerns that the price collector may be inclined to automatically record the previous price or be too influenced by it when identifying the item in the outlet to be priced. However, this information is sometimes included to assist the collector in ensuring that the correct item and price are being recorded and, where paper forms are being used, to identify any unusual price movements that need to be investigated. On the former, price collectors are encouraged to add additional information to the "standard" description to facilitate the unique identification of the product to be priced without a reliance on price. Such information might include: brand, make, size, model name and number, reference number, distinctive features and location in outlet (for example, “bottom shelf at rear of shop”).

5.70. Items should appear on the form in the same order in which prices are collected: a correctly ordered list will reduce time-consuming swapping and searching between pages. Even with a handheld computer or tablet, searching and pressing of navigation buttons will be less time-consuming if the item list is ordered to match the shop layout.
5.71. For fruit and vegetables, weight and quantity should be part of the item description, but it is also essential to record the amount actually priced so that a unit price can be calculated. Weight and quantity are important as it is often the practice that the unit price is lower if larger quantities are purchased. If the price collector prices a bunch of bananas and the item description is one kilo of bananas, then the price collector should choose a bunch weighing approximately one kilo, weigh the bunch and calculate the price per kilo.

5.72. At the time an index basket is updated, the questionnaire will need to list all items included in the old as well as the new basket. Where an updating of the outlet sample is planned to coincide with the introduction of a new basket, it will be necessary in the chain link month for the price collection to cover both the old sample of locations, outlets and items and the new sample of locations, outlets and items.

5.73. An example of a typical price collection form is given in Appendix B. This example is of a form used by the collector for recording prices when visiting an outlet and could be either in paper form or electronic. It is also possible to ask the shopkeepers concerned to complete the form themselves and to send it to the NSO although there are clear risks in the form not being completed correctly. Such a form may therefore serve for reporting as well as for collecting. If the form has space for recording prices over a whole sequence of months, the collector may keep the form and transcribe the prices from it each month onto a separate report form, which is sent to the NSO. Where the form used for collection is also used for reporting, there are two main options: either the form has space for recording prices over a whole sequence of months, and the form is shuttled backwards and forwards monthly between the collector and head office; or new forms for collection and reporting are printed out each month. In the latter case it can be useful if the form contains the prices recorded in the previous month alongside the spaces for recording the current month’s prices although as already mentioned this might encourage the price collector or data provider to give over-reliance on the previous recorded price. It should be noted that the transfer of the prices to another form or system, especially when done manually, may lead to transcription errors and is best avoided. Note that, in the example given in Appendix B, the previous month's price is shown (see paragraphs 5.204 to 5.213 on Computer Assisted Data Collection for a more detailed discussion of the advantages and disadvantages of giving price history). For simplicity, the example assumes all prices entered on to the form will include sale tax.

5.74. Completion of the price collection forms, whether by paper or electronically, should comply with the following guidelines.

a. All prices should be entered in the collection sheet in full even if there is no price change. This is good practice as it helps to ensure that the correct item is being priced and the price has been recorded correctly and that the price collector has not relied on information given by the shopkeeper without carrying out a check. For instance, a price may not have changed but the package size might have reduced, a practice sometimes used by retailers to make price increases less transparent, but to the shopkeeper there has been no price change. A price should never be recorded as, for instance, ‘no change’. All other entry fields on the price collection sheet should have something entered if only to indicate that they were not accidentally missed, for instance, “NA (not available)”.

b. If a price is not available a reason must always be supplied. The information given will be useful to the collector’s supervisor and the index compilers as well as to the price
The price collector may need to ask the shopkeeper for an explanation and may also need to choose a replacement. For example, the latter will be the case if the shopkeeper indicates that the item is no longer to be stocked.

5.75. No matter which method of local price collection is used, it is essential to provide some procedures to allow the tracking of activities and formal sign-off, confirming that processes have been completed and necessary actions undertaken on data checking and the transfer of information to head office. Audit trails are essential for work and quality management.

5.76. As recommended earlier, price collectors should be required to provide full descriptions of the items being priced. This enables checks to be put in place to ensure that collectors are properly following instructions, particularly on the selection of items to be priced. It also ensures that any changes, including changes in the quality of the items, are being properly identified, with enough detailed information to enable decisions to be taken on quality adjustment. Price collectors should be given a checklist or set of codes to record relevant information on changes relating to outlets, items or prices. The information needs to be systematically collected. For instance, codes to help with quality adjustment need to reflect those characteristics that most influence price. Prior research, for example, based on the hedonic method (see chapter 6), can help to pre-determine these.

5.77. Codes for managing the sample of outlets can be helpful and may include:

a. **Closed down**: outlet permanently shut or closed down.

b. **Temporarily unavailable**: outlet temporarily closed, but likely to be open next month.

c. **Refusal**: owner or staff refuses to cooperate.

d. **Change of details**: change of ownership or name or change of purpose.

**Continuity and the use of price collection codes**

5.78. Continuity is one of the most important principles of price collection. As the index measures pure price changes, it is vital that the same item is priced every month to establish a true picture of price change. So, if, for example, a jar of a supermarket’s own brand strawberry jam has been selected, that particular brand and flavour should continue to be collected. If it is out of stock, another brand and flavour should not be used without further investigation to establish whether this is a temporary situation or likely to be permanent. In the latter case, and if another flavour of the same brand, size and quality is available, then this item should in normal circumstances be chosen as a “comparable” item and the item description suitably amended. If a different brand, size or quality product is available then this should be selected as a “new” item, but only where no comparable items are available. The same principles apply to other items such as clothes and fresh fruit and vegetables. With clothes, it is important that colour, fabric, country of origin, logos and size are specified to ensure that the same item is priced each month. For fresh fruit and vegetables, useful

---

4 The replacement should be either a) as similar as possible to the previous one or b) the most popular “similar” product in the shop or c) the “similar” product that most likely will be available for future pricing. Unlike approach (a) which leaves the sample “static” with the risk that it will be increasingly out-of-date and difficult to collect prices for, approaches (b) and (c) have the advantage of introducing an element of sample replenishment.
attributes to record may be “country of origin”, “class” and variety. For electrical equipment, it may be the specifications and features given in the manufacturer’s catalogue that are important.

5.79. It is important that a detailed description of the items being priced is recorded to assist the price collector and head office in choosing or confirming the suitability of a replacement for an item that has been withdrawn and in identifying changes in quality. The focus should be on recording price-determining characteristics. Should the regular price collector be unable to carry out the normal collection, full and accurate descriptions will also enable a relief collector to carry out the collection without any doubt as to the correct items to be priced.

5.80. Most of the time, the item will be exactly as collected the previous month and all that will be recorded is a current price. However, should there be a change or uncertainty in the item, then it will be necessary for price collectors to use their own judgement and to inform head office, bearing in mind that head office staff should be responsible for making the final decision. A pre-coded specification will be less time consuming and will provide better guidance to the price collector on what information should be reported. The codes should cover situations faced regularly when collecting prices. The codes can be numeric or alphanumeric and each one should indicate the action to be taken by the price collector or supervisor and will be associated with corresponding procedures to be followed in the calculation of the index by head office. The indicator codes might include:

a. **Comparable (C)**: The original item is no longer stocked but a similar alternative has been collected that does not differ in terms of major price-determining attributes. The price is likely to be in a similar range although this may not always be the case.

b. **Non-comparable (N)**: The item has been replaced by something new that is not really comparable but is equally representative of that product group. If possible, the collector should try and find out the price of the “new” item in the previous or base period.

c. **Sale or special offer (S)**: A price decrease because of a genuine sale or special offer, with a sale or discount sticker present. This does not include damaged or out of date stock or clearance goods. The latter should never be included. A price reduction where there is no notice of a sale or special offer is not a “sale”; the item should still be priced, but without the S indicator code.

d. **Recovery (R)**: A return to the normal selling price after a sale or special offer. This need not be a return to the price prior to the sale or special offer. A referral back to the previous price collected and consultation with the shop keeper will normally be required.

e. **Temporarily out of stock (T)**: Guidance will need to be given to the price collector concerning the meaning of “temporarily” (in terms of expected duration, which may vary for different items). It may be advisable to replace items immediately (for example, fashion clothing, if it is unlikely that the identical item will come back into stock). Typically, "T" indicators should not be used for more than two consecutive months, and in the third month, a replacement should be chosen. In food outlets, it is very unusual for items to go permanently out of stock. The collector should always try and check future availability with the retailer.
f. **Missing (M):** Used where the outlet does not stock an item or no longer intends to stock an item and there is no appropriate alternative replacement. In these circumstances, it is recommended that the item is checked at subsequent collections to ensure that a suitable replacement item has not come into stock.

g. **Weight (W):** A permanent weight or quantity change to the product.

h. **Query (Q):** Such a code may be used to supply extra retail information to head office (for example, “10% extra free”, “2 for the price of 1”, or a strange price difference that is not covered by one of the other indicators, such as a bumper issue of a magazine at an increased price. Arrangements need to be in place for head office to respond to these comments and to treat the price quotes accordingly.

5.81. Even if the retailer says there have been no price changes since the previous month, the price collector should confirm prices anyway. This will require some diplomacy, but it is important because it is easy for the shopkeeper to overlook a small number of price increases, forget when the last increase occurred or even deliberately mislead the price collectors. Also checks need to be made to ensure that there have been no changes in the price determining characteristics of the item such as pack size.

5.82. As a general rule, a price should be recorded only if the exact product being priced is on display and immediately available for sale. No price should be recorded if a product is temporarily out of stock, but the price collector should record when new stocks are expected to arrive. For certain large items such as furniture and cars, however, where the item must normally be ordered, the price should be recorded as long as the retailer confirms that it is available for delivery within the normal time period.

**Unit prices**

5.83. Some food items, such as meat, fish and cheese, can be sold in variable weights and it is necessary to collect prices per unit of weight. This should be taken from the package labelling or calculated directly by the collector. Roughly the same package size and type should be used each month, as the unit price might be lower for larger pack sizes or differ between package types. Other items, such as eggs, are often sold in specified quantities. For these, it is essential that collectors record prices for the specified quantity, as total and unit prices usually depend on the number bought. If X eggs are to be priced and the price for the number is not quoted directly, then the price of one egg can be obtained and multiplied by X to get the required price. Care does need to be taken, however, to ensure that unit price does not decrease with quantity: significant changes in weight or quantity are to be avoided. Other examples are where herbs, such as mint, or vegetables, such as cabbage leaves, are sold in bunches of variable size with no label giving the weight. In this instance, a number of bunches should be weighed and priced to obtain an average price per kilogram. It will be necessary to purchase bundles and weigh them back at head office when there are no reliable scales available at the shop or market stall.

5.84. Certain food items, such as fruit or vegetables, are more difficult to price as some outlets might price items per number purchased while others might price by weight. For example, peppers may be priced by weight or by unit no matter what the size. Garlic may be priced per bulb, clove or by weight. Various types of berries may be priced by weight or by container, which may differ in size or how full it is. In these instances, care needs to be taken...
with the product descriptions. The item description given on the price collection form should refer to the pack size, weight or quantity being priced. Collectors need to be aware of the importance of collecting the same thing from one month to another, so that genuine price changes are recorded and not quantity or quality price changes.

3. Practical collection procedures: quality management in the field

5.85. Adequate field procedures are required as a precursor to ensuring that the quality of the price index is not compromised through errors in price collection. Price collection needs to be carefully planned and managed, and effective instructions and training given to price collectors. In traditional price collection, most prices are likely to be collected through price collectors visiting individual outlets. Checks need to be carried to ensure the accuracy of the data. This section gives guidance on field procedures relating to local price collection and provides an overview of quality management. It focuses on data validation in the field. Chapter 13 takes a broader look at the organisation and management issues associated with central price collection and with the complete process of producing a CPI.

Data validation in the field

5.86. Data validation should be carried out throughout the entire process and steps involved with the compilation of a CPI, from the collection of individual prices to their aggregation into indexes.

5.87. The questionnaires (and the data collection software where handheld computers or tablets are being used) facilitate quality assurance of the prices collected at or close to the time of collection and to record the results of the checks made in the field as part of an audit trail.

5.88. Collected prices can be compared to the prices of the same product sold in the same outlet with its previously collected price, and large movements can be checked for accuracy. Preferably, head office should provide guidance on what is viewed as acceptable ranges for price movements based on previous price movements. If it is the case, the retailer should be asked to explain why a price has changed significantly or remained unchanged for a very long period.

5.89. Primarily, it is the responsibility of price collectors to ensure that the prices and relevant characteristics of items have been correctly recorded, and that all the items in the sample have been priced. For instance, possible relevant differences in package size and weight should be immediately verified and recorded. The cooperation of the retailer or other data provider is essential during this phase.

5.90. In support of the key responsibilities of price collectors in providing accurate prices for input into the CPI, field supervisors and independent auditors should be deployed, where resources allow, to check the validity of the prices and related information recorded by the price collectors and to provide assistance and help when required as part of a collaborative effort. The required level of the validity checks may vary depending on the nature of retailing and the data collection procedures. For example, the real time data editing and price collection reports facilitated by using handheld computers and tablets reduces the chance of errors in price collection.
In the following sections the focus is on price collection and the validity of the prices and the editing of data. Chapter 13 addresses quality management and reporting in the context of CPI production and publication.

**Data validation: field supervisors**

Checks to ensure that data are complete and correct should be carried out as early as possible in the collection and compilation processes. A return to the shop to re-input prices becomes increasingly less feasible as time goes on, and there is a greater risk that the prices in the shops will have changed since the initial collection. The use of handheld computers or tablets by price collectors facilitates much more detailed checking at the time of the initial collection of prices in the shop than the equivalent paper system (see paragraphs 5.204 to 5.213 on electronic reporting).

Field supervisors have two main roles: one-to-one training of price collectors during joint field trips to correct any deviations from the procedures and tasks laid down in the price collection documentation; and reviewing the work done by price collectors in previous days in order to verify the quality of the work for which he or she was trained and to facilitate the correction of errors.

In an ideal system, field supervisors will be employed to regularly check that price collectors are adhering to the price collection schedule and are undertaking the required checks at the appropriate time. The supervisor should check that price collectors are completing price collection sheets correctly. A sample of collection sheets from each collector should be checked where it is not practical for supervisors to check every sheet. Checks may be made, for example, on whether the price collector has attempted to collect all prices from all outlets, that explanations have been given where prices were not obtained, and adequate descriptions entered where replacements for disappearing goods have been priced. The supervisor may also be required to check the accuracy with which data are transferred from paper to the computer. This is an essential task associated with the quality assurance process and needs to be allocated to somebody other than the person who initially inputs the data so that an independent check can be performed.

Supervisors should also be encouraged to visit shops and check the individual prices collected by price collectors. This can be organised either on a random basis or chosen on the basis of indicative information, such as extreme price variations, which suggests that a particular price collector may be experiencing difficulties. A typical audit report will include the percentage error rate and a breakdown of whether the errors are likely to have a high impact (wrong price, wrong item, item available but listed as temporarily unavailable) or a low impact (incomplete product description or inadequate reason given for price change). The audit report should be followed up by a formal request to the price collector asking for corrective action and confirmation that all necessary follow-up actions have been carried out.

Field supervisors should check for consistency and credibility in price movements recorded by all the collectors under their control. If, for instance, one location is reporting different price movements from the other locations within the collection region, some explanation or a follow-up price collection will be required to check the accuracy of the prices collected. Preferably, this should be done once data have been entered into the computer system and have been checked for transcription errors. Tabulations of price changes grouped by product or elementary aggregate should be provided to enable the
supervisor to conduct these checks efficiently. This will enable the supervisor to quickly identify extreme or inconsistent movements. Extreme or inconsistent movements may indicate either errors in collection or unexpected behaviour in the market. These checks should be conducted regularly during the price collection period.

**Quality checks in head office: data entry queries and the role of head office**

5.97. Once the price collection has been completed and the prices have been submitted to head office, a series of further validation checks can be run. In deciding on what checks should be carried out, account should be taken of the validation checks carried out in the field. For example, the use of handheld computers or tablets will increase the potential for validation at time of price collection and reduce the need for detailed scrutiny at head office. In addition, it would clearly not be productive or cost-effective to repeat tests already carried out.

5.98. The range of tests carried out on individual price quotes can include:

a. *Price change:* The price entered is compared with the price for the same product in the same outlet in the previous month and triggers a query where the price difference is outside pre-set percentage limits. These limits will vary, depending on the item or group of items, and may be determined by looking at historical evidence of price variation for the product or item concerned. If there is no valid price for the previous month because, for example, the item was out of stock, the check can be made against the price two months or three months before.

b. *Maximum/minimum prices:* A query is raised if the price entered exceeds a maximum or is below a minimum price for the item of which the particular variety is representative. The range may be derived from the validated maximum and minimum values observed for that item in the previous month expanded by a standard scaling factor. This factor may vary between items, again based on previous experience. Where necessary and possible, the maximum/minimum price should take account of any significant differences in average prices between, for example, regions.

5.99. If a handheld computer or tablet is used (see paragraphs 5.204 to 5.213), both tests can be implemented easily to take place at the time of collection; otherwise they will need to be conducted in head office as soon as possible after collection and prior to the computation of the index. A failure in either test should not result in the collector being unable to price the item but should prompt the collector to check and correct or confirm the entry, and prompt for an explanatory comment.

5.100. Queries raised may be either dealt with at head office or sent to the price collector for resolution. For example, scrutiny of a form might show that a big price difference has arisen because the item priced was a new product replacing another that has been discontinued. In this case, there may be no need to raise a query with the price collector, unless there is evidence to suggest that labelling the item “new product” is incorrect.

5.101. Where a price collection error is discovered and the collection of the correct price is too late to resolve in the process of computing a non-revisable CPI, head office will need to reject the price and exclude that item from that month’s index and the base month or treat it as a missing price and impute a price using the price movements of similar products. Where a CPI is revisable, it can be re-computed and the corrected figure published the following
month. In some countries the CPI is first published as a provisional figure to facilitate the late take-on of data including the situation just described.

5.102. Collectors should be encouraged to give feedback to head office on their experiences of price collecting. Collectors are a valuable source of information and often give good early feedback on changes in the market-place. Collectors can often warn of size or product changes before head office is able to obtain this information from other sources such as trade magazines. Collectors’ feedback can be used to support observed price movements and to provide supplementary briefing material. It can also form the basis of a newsletter for collectors.

5.103. The whole periodic routine of collecting prices in the field needs to be carefully planned and monitored, with arrangements in place to reflect local conditions. Circumstances vary, so it is not appropriate to be too prescriptive. It is, however, important to ensure that price collectors send in information when it is due. If they do not do so, it is necessary to find out the reason and to take appropriate action. It is also important to check that the information that has been submitted is accurate and complete.

Quality checks of local price collection: the role of auditors.

5.104. One way of monitoring the work being carried out by price collectors and addressing any issues is to employ auditors to occasionally accompany collectors during the field collection, or to carry out a retrospective check on the data that have been collected. The function of the auditor is to check the validity of the prices collected and to initiate corrective action which may extend beyond correcting an individual price quote to reviewing and updating instructions to price collectors and to general re-training. The function can cover more than one geographical area, but it does not normally extend to the supervisor's role of managing price collectors and the price collection process. However, sometimes the function of the auditor is combined with that of the field supervisor. The observations and comments of auditors are an essential part of quality management.

5.105. The range of tasks that an auditor carries out will vary from one NSO to another. Monitoring the standard of price collection will always be the main focus of the auditor. There are a number of other areas, however, in which auditors can be called upon to contribute. Auditors may be required to help with the sampling of locations and items and check that proposed collection locations contain an adequate range of outlets. They can also advise on economic conditions in these locations and on any dangerous areas. Auditors can carry out product reviews. For example, if a particular item seems to be causing difficulty for price collectors, auditors can speak to collectors and retailers to determine the reasons for these difficulties. Auditors can also advise on changes to basket composition. They can ensure that products suggested by head office are available across the country and can suggest item descriptions and weight bands. Furthermore, auditors can provide reports on price collection in existing locations. For example, head office may raise a query about a particular outlet in a particular location; auditors can visit this outlet to find the answer to the question or to persuade a retailer to continue with the survey.

5.106. The main purpose of audits is to ensure that each collector is following the procedures laid down for price collection so that the risk of errors is reduced. However, it is also important to recognise other benefits ensue which are of greater strategic importance in terms of continuous quality improvement. These are:
a. Raising awareness of quality.

b. The identification of the scope for introducing improvements to quality including rectifying weaknesses in procedures and documentation and also in price collection skills.

**Quality checks of local price collection: back-checking**

5.107. Another approach to monitoring the standard of price collection is to carry out a back-check, i.e. a retrospective check of a proportion of the prices recorded during the collection.

5.108. Back-checks can be used to:

a. Assess the standard of competence of individual price collectors.

b. Audit the overall standard of price collection.

c. Identify general training needs or the specific needs of an individual.

d. Highlight any key issues including, for example, problems with documentation or instructions issued by head office.

e. Identify areas where collection is problematical; for example, all collectors may have problems in certain types of outlets, prompting the need for more detailed instructions from head office.

5.109. Back-checking should be done by an expert independent of the process (preferably employed by the NSO), such as an auditor. Back-checking is carried out by visiting the selected outlet and re-collecting the prices and other relevant information, such as attribute or description codes. This activity should be carried out close to the original collection period to avoid problems of price changes occurring in the interim. It is important that back-checkers seek permission from the shopkeeper beforehand and follow the general criteria of conduct for local collection.

5.110. For a back-check to be a useful exercise it is important that performance criteria are determined to which all back-check results can be compared. These criteria should set, for example, the acceptable number of pricing errors per number of items checked. Well-defined criteria will enable easy identification of a poorly performing collector or location following a back-check.

5.111. A back-check may include a range of tests to identify the following:

a. Price difference – if the price is different, the auditor should check with shopkeepers to see if there has been a price change since the original collection took place.

b. Insufficient item description – each item should be uniquely defined so that another collector can step in to cover the absence of the usual collector, for example in the event of illness.

c. Wrong item priced – such as incorrect size or brand being chosen.

d. Items wrongly recorded as missing or temporarily out of stock.

5.112. A report should be sent to head office for scrutiny once the back-check has been completed.

5.113. Head office will then need to take appropriate action, which may include, for example, retraining the price collector or sending out supplementary instruction.
Quality checks conducted centrally by head office

5.114. Four kinds of regular checking are necessary at head office:

a. To ensure that the price collectors’ reports are sent in when they are due. If this is not done, it is necessary to find out the reason and to take appropriate action to obtain the reports.

b. To confirm that the reports contain what they are supposed to contain, i.e. that fields which must be filled in have not been left blank, that numeric fields contain numbers and that non-numeric fields do not.

c. Review and edit each return. Substitutions may have to be made centrally, or those made by the collectors may have to be approved. Unusual (or simply large) price changes may need to be queried. Items priced in multiple units or varying weights may have to be converted to price per standard unit. Missing prices must be dealt with according to standard rules relating to the cause.

d. Errors, introduced when keying the numbers into the computer or transcribing them onto worksheets, must be found and corrected, and preferably avoided in the first place by eliminating the need to transcribe.

5.115. It should be noted that the way the data are organized in worksheets or in the computer may differ from the way they are arranged on receipt. For instance, the latter data may arrive at head office organized for price collection purposes by collector, outlet and item but will be entered on to a spreadsheet designed to reflect the computational needs of CPI compilation. The data, in original format, should be recorded so that reference to it can be made if processing discloses any problems with the data. This facilitates operational management when dealing with queries. Furthermore, even if the same set of codes are used both in price collection and in the processing of collected prices, other codes may have to be used for information which comes in from the collectors in non-coded form.

5.116. How the checking is organized will vary from country to country. In some cases, local or regional supervisors will do some of it; in other cases, it will be more appropriate for it all to be done centrally. Some of these tasks can be done by computer, others manually. Therefore, no general recommendation can be made about the sequence of the work or about its division into different parts.

5.117. Procedures should be in place to check that all documents, messages or files are returned from the field so that price collectors can be contacted about missing returns. Initial checks should then be carried out to ensure that data are complete and correct. For instance, checks should be run to ensure that unexpected duplicate prices (i.e. for the same item, in the same shop, in the same location) are not taken on, and that the location, outlet and item identifier codes which accompany each price exist and are valid. If any prices fail these checks, a query should be raised with the price collector for clarification. Since some of the checking may require reference back to the price collectors (or to their supervisors or respondents when direct mail questionnaires are used), the timetable for producing the index must allow for this communication to take place.

5.118. In deciding on what checks should be done in head office, account should be taken of the validation checks carried out in the field. The use of handheld computers or tablets will increase the potential for validation at the time of price collection and reduce the need for
detailed scrutiny at head office. It would clearly not be productive or cost-effective to repeat all the tests already carried out locally, except as a secondary audit or random check that those checks have been completed.

Data reports

5.119. Reports can help the analyst pick out prices for which the level or change stands out as different from that reported for similar varieties elsewhere, or simply where a change may need to be queried because it lies outside certain specified limits. Thus, a comparative analysis of other prices collected for an item can be undertaken. A computer printout can list all prices which either fall well outside the range of prices obtained for that representative item, or for which the percentage change from last time falls outside a specified range, and a similar list can be compiled identifying outliers based on the recent price behaviour of the same item in the same outlet. The limits used will vary from item to item and can be amended in the light of experience. The analyst can then work through the printout, first ascertaining whether there has been a keying-in error, and then examining whether any explanation furnished by the collector adequately explains the divergent price behaviour or whether a query should be sent back to the supervisor or collector. Again, the timetable should allow for this, and anomalous observations should be discarded where an acceptable explanation or correction cannot be obtained in time.

5.120. Other reports may be produced regularly covering several periods (e.g. three monthly) to detect accumulated patterns, thus enabling broader problems to be detected. For example:

a. One collector’s reports might show many more “outlet closed” remarks than those of other collectors, perhaps indicating either a motivational or training need on the part of that collector, or a change in retail trade patterns in a particular area.

b. Variety substitution for a particular representative item might become more numerous than before, suggesting a possible need for revision of the specification or the choice of another representative item.

c. Where tight specifications list a number of brands and models of which one is to be chosen, but a large number of collected prices are for items not specified in the original list, this will suggest that the specified brands and models are no longer appropriate and that a review of the list is required.

d. The dispersion of price changes for a specific representative item might be much larger than it used to be, raising the question of whether it has been appropriately specified.

5.121. Routine computer-generated reports should enable those in charge of the index to detect the existence of such problems. Two types of reports are particularly useful: index dispersion reports, and price quote reports.

a. Index dispersion report. This is a list of items indicating the current index for each item, the number of valid quotes for each item, and the number of price relatives and their values. The ratios of current to previous valid prices can be compared and queries generated, where these fall out of acceptable ranges based on previous price behaviours and taking into account any special circumstances such as the introduction of discounted prices in seasonal sales. The index dispersion reports can be used to identify quotes with price relatives that fall outside the range of the main bulk of quotes. These suspect quotes can then be investigated and appropriate action taken.
b. **Price quote report.** This consists of a range of information on an item that the index dispersion report has highlighted as warranting further investigation. Information listed may include current price, recent previous prices and base price, together with locations and types of shop. The report can be used to identify the quotes that require further investigation and also to investigate rejected prices.

5.122. It should be noted that while price collectors should examine every price they collect, it is not considered necessary nor feasible for collection supervisors and index compilers to subject every collected price to the same level of scrutiny. It is recommended that, to improve cost-effectiveness, some form of significance rating should be applied to determine how much time and effort should be expended on examining and, where necessary, editing individual prices.

a. In general, all things being equal, prices from elementary aggregates with relatively small price samples should receive more attention from the index compiler. This is because, if the weights of the elementary aggregates are broadly equal, each individual price movement within these elementary aggregates can have a much more significant influence on the index calculation than any individual price movement from within an elementary aggregate with a large number of price quotes.

b. Price samples from elementary aggregates with high expenditure weights should also be examined critically as a higher expenditure weight will cause all price movements within the sample to have a greater influence on the CPI.

c. The highest risk is associated with elementary aggregates with relatively large weights but few price quotes and with complex index construction methods. This situation is associated with utilities and other services which account for relatively large expenditures and where there may be only one or a handful of suppliers and prices are based on complex tariffs.

5.123. Some of the techniques discussed in the following section work best when applied to large quantities of data and have the advantage of being automated with the intervals prescribed for closer examination being generated by the prices data. Abnormal individual prices such as sale prices, or price movements such as sale recovery prices, may be excluded from manual and automated procedures for the detection of outliers, particularly in the calculation of upper and lower bounds, as they are not representative of the general trend in prices. Nevertheless, such prices should be checked for accuracy, for instance by reference to previous price history. Automated checking can be applied to seasonal sale prices and prices for seasonal items.

5.124. Automated checking essentially performs the same basic filtering purpose for the identification of outliers as the manual techniques described above. It is sometimes referred to as statistical checking compared with the manual techniques which are sometimes referred to as non-statistical checking.
F. Data validation and editing

1. Data validation and editing by head office: automated (statistical) checking and the use of algorithms

5.125. The main conceptual difference between automated (statistical) checking and manual checking is that the automated technique calculates the limits for acceptable movement based on the data collected. These techniques have the benefit of automatically updating the acceptable limits in line with any overall change in price volatility, as observed when new price data are received, and the limits are re-calculated. These techniques require a large amount of data to provide reliable results and so are best suited to data handlers and index compilers in regional offices and head office, where prices data from several collection centres will have been collated and stored, rather than at the local level.

5.126. Automated (statistical) checking compares each price change with changes in the other items from a given price sample. The chosen price sample is usually the sample to which the item being checked belongs, but the sample for testing may be a combination of price samples for similar products. It can also be updated as more prices are received from the field. For each of the methods described below the price ratios may measure the price change over any time period: for instance, the change from the previous period or the change from the same period in the previous year.

The use of median and quartile values

5.127. One method of setting the limits to determine whether a movement is a possible error is based on the median and quartile values of the price ratios from the sample. The acceptable limits are set as a predefined multiple of the range between the median and the quartiles. Any observation with a price change outside this range is identified as a possible error. The major benefit of a method like this is that it is not affected by any single outlier value. A numerical example is provided in Appendix C.

5.128. The basic approach to estimating sensible upper and lower limits of acceptable price movement relies on the assumption that the observed price changes are normally distributed. Under this assumption, the distance between each of the first and third quartiles ($R_{Q1}$ and $R_{Q3}$) and the median ($R_M$) will be the same: call this distance ‘$D_M$’. Operating under this assumption, the proportion of price changes that are likely to lie outside specified upper ($L_U$) and lower ($L_L$) limits can be estimated from a normal distribution table. The limits can be defined as:

$$L_U = R_M + C \times D_M; \text{ and}$$

$$L_L = R_M - C \times D_M$$

where $C$ is a user defined value.

5.129. It will be seen later (paragraphs 5.145 to 5.154) that a variation of this basic approach is recommended to allow for the skewed distribution of price changes which can be observed in practice.

5.130. If $C$ is defined as equal to one, then approximately 50 per cent of the observations will lie between the upper and lower limits. Using the standardized normal distribution, this is
equivalent to setting the limits at plus or minus 0.7 times the standard deviation (σ) from the median. The following table provides approximate multiples of σ for selected values of C and the associated percentage of the observations that will be flagged as possible errors and outliers.

Table 5.2 Selected values of C and the proportion of observations flagged

<table>
<thead>
<tr>
<th>C</th>
<th>σ multiplier</th>
<th>Expected proportion of observations flagged</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.68</td>
<td>50%</td>
</tr>
<tr>
<td>2</td>
<td>1.37</td>
<td>17%</td>
</tr>
<tr>
<td>3</td>
<td>2.07</td>
<td>4%</td>
</tr>
<tr>
<td>4</td>
<td>2.75</td>
<td>0.7%</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>0.14%</td>
</tr>
</tbody>
</table>

5.131. In practice, there are serious shortcomings with this method as described.

5.132. In normal circumstances, the majority of observations for many products will not show any price movement. Therefore, the values of the quartiles are likely to be very close to the median value. As a result of this, using small values for C is likely to cause the majority of price movements to be flagged as possible errors and outliers. To demonstrate this effect, 16 additional observations indicating no price movements have been added to the sample of 30 observations used in the first example. These are shown in example 2 in Appendix C. A price sample with at least one-third of the observations showing no movements would not be unusual for many categories of items. If C is set to 2, then 60 per cent of the actual price movements would be flagged as possible errors, compared with 30 per cent in the unadjusted sample.

5.133. The index compiler should experiment with different values of C for different product groups or outlet types to determine appropriate values for local use. It is recommended that a relatively low value of C should be used. C need not be an integer.

5.134. It should be noted at this point that the distribution of prices and price movements is rarely normal, rather in most cases a skewed distribution exists. Thus, the underlying assumption of a normal distribution is not valid, and the use of symmetrical upper and lower limits will result in a skewed distribution of prices flagged up as possible errors or outliers. This is operationally inefficient and the examination of differing proportions of “low” and of “high” prices and price movements could lead to bias.

A modified use of median and quartile values

5.135. To use the above method in practice, three modifications are recommended.

5.136. Based on the simple price ratios, the distances from the median represented by price decreases are not as large as the distances represented by price increases. As an example, consider a case where a product is on special offer at half price. This is represented by a price decrease of 50 per cent. However, to return to the original price requires a 100 per cent increase. To make the calculation of the distance from the centre the same for extreme
changes for price decreases as for the price increases, the price ratios should be transformed. The transformed distance, $S_i$, for the $i^{th}$ price observation can be calculated as:

$$S_i = 1 - \frac{R_{M}}{R_i}$$

if $0 < R_i < R_M$

$$S_i = \frac{R_i}{R_M} - 1$$

if $R_i \geq R_M$

5.137. The observations with a price ratio lower than $R_M$ have now been transformed into the negative of the increase required to return the price ratio to the value of $R_M$. Any observations with a price ratio equal to $R_M$ will have a transformed price movement of zero. Observations with a price ratio greater than $R_M$ have been transformed to show changes as though they had increased from $R_M$. The procedure is then carried out on the set of $S_i$.

5.138. In situations where the quartiles ($R_{Q1}$ and $R_{Q3}$) are quite close in value to the median ($R_M$), many small price movements are likely to be identified as possible errors or outliers. To reduce this problem, items with no price movements may be removed from the calculations. If the acceptance interval is still very narrow, some minimum distance should be set. A starting value is 5 per cent for monthly changes but it is up to the CPI compiler to choose, based on personal judgement.

5.139. The third modification is intended to overcome the problem of using small samples. When using a small sample, the impact of one observation on the distances between the quartiles and the median might be considered to be too significant. In practice, the sample sizes for many elementary aggregates will be small. To improve the usefulness of this method, the samples from several similar elementary aggregates can be combined. In this regard, elementary aggregates can be considered similar if their prices are believed to exhibit similar behaviour.

5.140. The Hidiroglou and Berthelot method can easily be extended according to the described in the following paragraphs. The variable $s_i$ is independent of the price levels. To address the issue that the level of the prices can influence the acceptance interval, $s$ may be transformed into a new variable, $E$:

$$E_i = s_i \cdot \left( \max \left\{ p_{t-1}^i, p_t^i \right\} \right)^U, \quad 0 \leq U \leq 1$$

5.141. $E$ is calculated as $s$ multiplied by the largest of the prices in period $t$ or $t-1$, raised to the power $U$. The variable $U$ determines to what degree the price level influence the acceptance interval. The larger $U$, the larger the influence of the price level. If $U = 0$ the price level plays no role. This facility is helpful, if one wishes to pay more attention to a price increase from 1000 to 1100 than one from 10 to 11.

5.142. A further transformation can be made to ensure a minimum acceptance interval to avoid too many price changes are identified as possible errors for elementary aggregates with
only small price changes. To do this, for each elementary aggregate the median, \( E_M \), and first and third quartile, \( E_{Q1} \) and \( E_{Q3} \), of the \( E_i \)’s are found, and the following values calculated:

\[
\begin{align*}
    d_{Q1} &= \text{Max} \{ E_M - E_{Q1}, |AE_M| \} \\
    d_{Q3} &= \text{Max} \{ E_{Q3} - E_M, |AE_M| \}
\end{align*}
\]

5.143. \( A \) is a constant that enters \( |AE_M| \) to ensure a minimum acceptance interval. A low value of \( A \) raises the probability that \((E_M - E_{Q1})\) or \((E_{Q3} - E_M)\) determines \( d_{Q1} \) or \( d_{Q3} \), and vice versa. If, for instance, \( A \) is set to 0.05, \( |AE_M| \) will be quite small so that \((E_M - E_{Q1})\) or \((E_{Q3} - E_M)\) are likely to determine \( d_{Q1} \) or \( d_{Q3} \), even if the dispersion of the \( E_i \)’s are relatively small. If, on the other hand, the dispersion of the \( E_i \)’s becomes very small, \( |AE_M| \) determines \( d_{Q1} \) and \( d_{Q3} \). Hence, \( A \) can be used to avoid that too many price changes are identified as possible errors for elementary aggregates with only small price changes. The acceptance interval is finally defined as:

\[
\text{Acceptance interval} = \{ E_M - C \cdot d_{Q1}; E_M + C \cdot d_{Q3} \},
\]

where \((E_M - C \cdot d_{Q1})\) is the lower bound and \((E_M + C \cdot d_{Q3})\) is the upper bound of the interval. \( C \) is an extra variable that may be introduced. The larger \( C \), the larger the acceptance interval, and the fewer extremes and potential errors will be identified.

5.144. For a more detailed explanation of this approach the reader is directed to Hidiroglou and Berthelot\(^5\) (1986).

**The Tukey Algorithm**

5.145. The Tukey algorithm overcomes the problem of having many instances where there has been no price change, i.e. where many price relatives are equal to one indicating no price movement. The first step is to sort the sample of price relatives. The highest and lowest 5 per cent are flagged for examination as possible errors or outliers and removed from further calculation. All observations with no price movement are also removed from the sample before further calculations are done. The next step is to calculate the arithmetic mean \((AM)\) of the remaining observations (referred to as the Tukey sample). This value is then used as the dividing value to separate the observations into two smaller samples: an upper and a lower set of price ratios. The arithmetic mean of each of these two samples is then calculated \((AM_U, AM_L)\). The upper and lower Tukey limits \((T_U, T_L)\) are then calculated for the Tukey set as:

\[
\begin{align*}
    T_U &= AM + 2.5(AM_U - AM) \\
    T_L &= AM - 2.5(AM - AM_L)
\end{align*}
\]

5.146. All observations that are greater than \( T_U \) or less than \( T_L \) are flagged as possible errors or outliers.

\(^5\) “Statistical editing and imputation for periodic business surveys”, Survey Methodology, Volume 12, No 1, pp 73-83.
5.147. As this method excludes all observations with no price movement, the calculated limits are unlikely to be close to the mean. Therefore, there will be no need to impose a minimum difference as in the first method. However, the problem of requiring a reasonably large number of observations in the sample remains. Again, it may be necessary to combine the samples of similar elementary aggregates. Example 3 in Appendix D shows that 5 observations would have been flagged by this method in comparison to 18 observations by the previous method based on the modified use of median and quartile values.

5.148. Statistical methods of filtering have an advantage over simple filtering in that the limits are set by the data and they can be re-calculated over time. The disadvantage is that filtering cannot be done until sufficient quantities of data have been collected, unless the index compiler uses approximations from past experience. The processes can be repeated as additional prices are received. Compilers should aim to set filters so that most of the records flagged as potential errors do turn out to be errors (or outliers requiring explanation). The aim of all these methods of filtering is to indicate which records require examination, not to flag records for automatic deletion from the sample. Each price movement should be checked for credibility and that it is representative. Only if the movement is an error or unrepresentative should modification of the movement be considered. There should not be a presumption that an outlier is “wrong until proven right”. Put another way, outliers should not be treated as incorrect prices.

2. Visual data validation

5.149. In data validation, it can be more practical to use visualization. In visualization it is easy to use plot charts which are readily available in tools like Microsoft Excel or can be programmed to the IT system. From plot charts (see figure 5.1 with arrows highlighting outliers) it is easy to spot the outliers and focus validation to these. Visualization can be more convenient than focusing validation to the change of prices, especially when there are extreme price changes such as during sales periods or in the case of fresh fruit and vegetables. If the plot chart is programmed into the IT system, from the outlier it is possible to make a direct link to the observation.

Figure 5.1 Price changes in plot chart during sales season
3. Dealing with outliers

5.150. Outliers are extreme or large price movements compared to the average price movements which have been verified as correct. As already indicated above, the detection of price observations which are outliers may be conducted through an examination not only of price levels but also of price movements. In this case, outliers can be defined as price movements that are exceptionally large compared with the majority of movements. The movements will have been verified as being based on correctly collected and recorded data but may not be representative of the behaviour of the section of the market that they are meant to represent. This leads to the concern that a different sample would have produced a greatly different and more representative average price movement.

5.151. It is recommended that a strategy is followed which uses resources efficiently on validation and checking of input data and checking is done by impact/output and the focus is on finding the most important errors/outliers. The general rule should be to include verified prices; to exclude prices until proven to be correct or to modify them should be the exception. The aim should be to reflect the reality.

5.152. The tests for outliers are the same as those for identifying potential errors, as described in the section on data validation and editing (paragraphs 5.135 to 5.175). Outliers can be determined by comparing the price movement against defined allowable limits. These may be either pre-determined numerically or pre-defined based on statistical tests.

5.153. If, by exception, outliers are to be modified, they are usually modified to lie on the pre-defined boundaries of acceptable movement or to be imputed by the movement of a suitable sample of prices. Imputation by the average price change of the product group to which the item belongs gives a similar result to its exclusion (the same result within the elementary aggregate), but such imputations can have operational advantages insofar as they employ protocols already in the calculation system for the imputation of missing prices. But automatic adjustment should be deployed with extra caution and should generally be avoided.
It certainly shouldn't be used to reduce volatility in an index, for example, at the elementary aggregate level. The index compiler should consider each case on its individual merits, following agreed guidelines and using all relevant information in coming to a decision. Prices should be modified or discarded only if there is sufficient justification. It should be noted that the CPI protocols followed by the NSO may forbid the modification or exclusion of outliers.

5.154. It is worth re-emphasising that price collectors and their supervisors are responsible for providing as much information as possible about the reasons for extreme price movements or levels and why they accepted the price quote as valid. In addition to checking for simple accuracy, supervisors can also be instructed to compare the price movements for equivalent products obtained by all the collectors they supervise.

4. No price change

5.155. As mentioned earlier, if the price observations are collected in a way that prompts the respondent with the previously reported price, the respondent may report the same price as a matter of convenience. This can happen even though the price may have changed, or even when the particular product being surveyed is no longer available or has changed its price-determining characteristics. As prices for many items do not change frequently, this kind of error is unlikely to be spotted by normal checks. Often the situation comes to light when the contact at the responding outlet changes and the new contact has difficulty in finding something that corresponds to the price previously reported. It is advisable, therefore, to keep a record of the last time a respondent reported a price change. When that time has become suspiciously long, the analyst should verify with the respondent that the price observation is still valid. What constitutes too long will vary from product to product and the level of overall price inflation, but, in general, any price that has remained constant for more than a year is suspect.

5. Missing prices

5.156. Treatment of missing prices is dealt with in more detail in chapter 6. This short section discusses ways of minimising the occurrence of missing observations.

5.157. It is important to maintain the relevance of the sample of items priced. As part of the longer-term maintenance of price samples, items and locations for which prices are missing can be examined for common patterns. For instance, if many retailers are missing the same item, there may be some supply problem of which no single retailer is aware. This may be an indication that an item will have to be replaced. If the number of regularly missing items is growing, then the sample itself might need to be reviewed. If a particular shop is recorded as having a relatively large number of missing prices it may no longer be an appropriate outlet for the particular items assigned to it, or the varieties of products whose prices are collected in the shop may need to be reviewed.

5.158. Where prices are collected by means of a questionnaire sent to the shopkeeper, individual shopkeepers often follow a regular pattern, some return their price surveys promptly, others take some time. Staff should be encouraged to become familiar with these patterns. If the system for recording the return of these surveys also records the expected return date, then unexpected non-returns can be flagged even though the final deadline for return of survey forms has not passed. These shopkeepers can be contacted in advance of the final deadline to ensure that the survey form has not been forgotten. Early contact can reduce
the number of prices still missing by the deadline. Shopkeepers that provide prices for heavily weighted items can also be monitored and contacted earlier rather than later.

6. Credibility checking

5.159. Credibility checking tests the reasonableness of the input data and the results obtained. Credibility checking of the results should take place after the checking of the numerical accuracy of the data at or shortly after price collection as described above. These early checks are the responsibility of the price collectors and their supervisors but also involve outlier detection at head office. These early checks should discover all straightforward errors like incorrect coding, such as wrongly attributing a price as a sale price, and the incorrect recording of prices.

5.160. Dealing with other potential errors is not so straightforward. Results that fail a data check, such as exceeding the pre-defined movement limits described earlier, may be judged by the index compiler to be valid as a result of referring to other information such as market intelligence. Other potential errors might only be resolved after checking with the shopkeeper, if time allows.

5.161. If it is still possible with individual price quotations to re-survey the price or obtain a satisfactory explanation from the shopkeeper, then the query can be sent back to the price collector and the data can be flagged as being verified and then subsequently corrected if found to be an error. Even if it is not possible to check with the shopkeeper before the computation deadline, the shopkeeper could be questioned during the next regular visit, as the answer may assist the NSO’s understanding of market behaviour for the particular product or retail sector. When a satisfactory explanation is not available, the CPI procedures should provide guidelines to aid the compiler in making a decision. For instance, the compiler could omit the price, allowing the processing system to impute a price, or modify the price to keep the price change within a pre-defined limit, but this is best avoided and should be the exception to the rule. If prices are modified without verification from the shopkeeper, it is recommended that price collectors are informed of potential problems during the next collection.

5.162. NSOs can minimise problems caused by unusual prices and price movements by training price collectors to recognise these situations and to check prices when first observed and to collect relevant explanatory information during the initial price collecting visit. Avoiding return visits or calls keeps costs down and also reduces the burden placed on shopkeepers.

7. Checking by impact or data output checking

5.163. Filtering by impact, or output editing, is based on calculating the impact that an individual price change has on an index to which it contributes. This index can be an elementary aggregate index, the total index, or some other aggregate index. The impact that a price change has on an index is its percentage change times its effective weight. However, the exact calculation of the impact will depend on which formula has been applied for the elementary indices. It is possible to set a maximum value for this impact, so that all price changes that cause an impact greater than this can be flagged for review. The impact of a price change on a higher-level index will also depend on the weight of the elementary index in the aggregate.
5.164. At the lowest level, the appearance and disappearance of products in the sample cause the effective weight of an individual price to change substantially. The effective weight is also affected if a price observation is used as an imputation for other missing observations. The evaluation of effective weights in each period is possible, though complicated. As an aid to highlighting potential errors, the nominal weights, as a percentage of their sum, will usually provide a reasonable approximation. If the impact of 12-month changes is required to highlight potential errors, approximations are the only feasible filters to use, as the effective weights will vary over the period.

5.165. One advantage of identifying potential errors in this way is that it focuses on the results. Another advantage is that this form of filtering also helps the analyst to describe the contributions to change in the price indices. In fact, much of this kind of analysis is often done after the indexes have been calculated, as the analyst often wishes to highlight in the statistical press release those indices that have contributed the most to overall index changes. Sometimes the analysis results in finding, for instance, that some retail sectors have a relatively high contribution to the overall price change – a contribution that is considered unrealistic and counter-intuitive. The change may be traced back to an error, but it may be late in the production cycle and jeopardize the scheduled release date. There is thus a case for identifying such unusual contributions early on as part of the data editing procedures rather than for analytical purposes. The disadvantage of this method is that in practice the final calculation of an elementary index’s change may be rejected only after the CPI has been computed. It may be necessary to over-ride the calculated index, though this should be only a stopgap measure until the index sample is redesigned.

G. Price collector training

5.166. As indicated by passing references, the training of local price collectors and good work instructions are a vital element in ensuring the quality of the prices data and of the CPI. Collectors need to be properly trained, require adequate instructions and easy access to guidance, particularly so given the fact that:

a. Price collection is not a trivial matter.
b. Snap judgements often need to be made.
c. Collectors often work remotely and on their own
d. Instant communication is not always possible.
e. Collectors work in a dynamic environment.
f. It is difficult to rectify errors after the event.

5.167. Documents are needed to explain what is to be done, when it should be done, how it should be done and why it should be done. Reviewing the documentation also provides an opportunity to review the procedures.

5.168. The issue of good documentation as part of an integrated quality management system is addressed in chapter 13. The current chapter deals specifically with the documentation needs of price collectors and the issue of training.
5.169. The training of price collectors should enable them to successfully perform all essential activities and deal with all potential situations including not just recording the price of an item but, for example:

a. Persuading new outlets to become price providers.
b. Understand and recognise occasions when prices provided are unacceptable.
c. Recording relevant information to describe quality change in a product.
d. Recognising unusual price movements when checking their collected prices.

1. Price collector induction training

5.170. Induction training should be given to all price collectors so that they gain the necessary skills before collecting prices for the CPI. It can also be a motivational tool.

5.171. A typical training schedule might consist of a one-day training course at head office (which may include some refresher training for experienced collectors) covering:

a. Background to the CPI and the NSO.
b. Use of the CPI and importance of accurately recording prices.
c. The general principles of index compilation and price collection.
d. How local price collection fits into the process.
e. Retailer recruitment, getting permission to enter shops etc.
f. Practical price collection issues. For example: product identification /descriptions; pricing (item descriptions, definition of price/sale price, rules relating to seasonal items, quantity conversions, quality adjustment - when is an item or product equivalent?).
g. The timetable and administrative arrangements.

5.172. Practical examples and practice collections should be an integral part of the learning process. For example, there should be an opportunity for:

a. Discussions about “equivalent” replacements using photographs and item descriptions.
b. Practice collections in office.
c. Supervised practice collection in the field.

5.173. Tests and evaluation of individual performance should be an integral part of the training. This could be achieved through:

a. Written tests at the end of induction day.
b. Evaluation by supervisors of practice collections in the field.
c. Feedback to new collectors, including additional training needs.
d. Evaluation by collectors of training provided (essential for ensuring training is relevant and effective).

5.174. The evaluation of individual collector performance is essential. Collectors must pass the required standards, against a checklist of tasks, before being allowed to conduct a real collection.
2. **Follow-up training and refresher training**

5.175. The longer-term element of training is just as important as induction training to the integrity of price collection, particularly with the evolution of the retail sector and CPI methodology and when CPI baskets are updated. One way of facilitating this is for the price collector’s supervisor to:

a. Accompany the new price collector on a live price collection.

b. Conduct a back-check of the prices collected to identify any problems.

c. Produce an evaluation report which will provide a basis for further training of the price collector. The evaluation report can include a scorecard against a checklist of required actions.

5.176. Where resources allow NSOs should consider conducting regular accompanied checks and back-checks of all price collectors in addition to special checks of poorly performing collectors. The information gathered can be used to compile scorecards for individual price collectors, supervisors and groups of price collectors.

5.177. A good training program will also help to motivate price collectors.

5.178. Regular refresher training workshops should be considered, especially where evidence from the field indicates a pressing need or where price collection procedures and conventions change, or the CPI basket or sample of retail outlets has been updated. These present an opportunity to: raise awareness of the importance of getting correct prices; provide formal training on revised guidelines; resolve recurring or recent problems; provide price collectors with an opportunity to assist each other in managing problem situations encountered in the field, such as dealing with reluctant respondents.

3. **Training of supervisors and head office staff**

5.179. Supervisors must be at least as well informed as the price collectors. As the supervisor will normally be the first point of contact when a difficult situation is encountered during price collection, they also need a good understanding of the methodology and theory behind the selection of the product sample. Supervisors are also part of the management team. Their training should cover such matters as:

a. Team management.

b. Performance appraisal (where this is normal office practice).

c. Project management.

5.180. Head office staff should also be provided with a basic training in price collection. The benefits are three-fold:

a. It gives head office staff a better understanding of collectors’ needs.

b. It helps editing (staff at head office will know what to look out for).

c. It supports Disaster recovery (head office staff will be able to undertake price collection in an emergency). For more information on disaster recovery, please see 5.197 to 5.199.
H. Documentation: work instructions

5.181. Accessible, relevant and up-to-date work instructions are essential both for price collectors and their supervisors. Documentation should cover all aspects of the job and should in large part reflect what has been covered in training.

5.182. Price collectors should be provided with work instructions on:
   a. How to approach shopkeepers.
   b. How to ask questions to ensure that the required information is obtained.
   c. Appropriate personal behaviour and dress codes.
   d. Procedures for recording and passing on collected prices and other relevant information.
   e. Data checking.
   f. Creating collection schedules.
   g. Recognising when recorded prices appear to be incorrect.

5.183. The work instructions for supervisors of price collectors should be in the form of a supplement to the price collectors’ work instructions and should cover:
   a. Checking the quality of the price collectors’ work.
   b. Checking the accuracy and completeness of the prices collected.
   c. Official recording of resource use (for example, cars and bicycles for transport and funds for buying goods in markets).
   d. Official procedures for maintenance of resources (for example, testing the accuracy of scales).
   e. Creating complementary collection timetables for all collectors within the supervisor’s area of responsibility.

5.184. Most of the documentation should be prepared by head office, with input as appropriate from regional offices and fieldwork supervisors. Centrally prepared documentation will help ensure consistent practices in the field including between regions. It should be readily available to all collectors and supervisors. The documents could be available in paper or electronic form, whichever suits the organisation, but whichever medium is chosen it needs to be accessible to the relevant staff.

5.185. All documentation should be kept up-to-date. Effective documentation control systems should be in place. With paper-based documentation this could mean keeping the instructions in a loose-leaf folder and issuing individual updates. The amendment pages should include version number and date printed and should be kept to a reasonable number for ease of reference. Editorial access should be restricted and should be password protected. A judgement will need to be made on when a re-print of a chapter or the complete working instructions is justified. Documentation is an essential part of a quality management system.

5.186. An example of a documentation control template is given at Appendix D.
I. Disaster recovery

5.187. The prices data should be stored onto a database with hardware and software that is robust and supported to minimise the business continuity risks associated with running the existing system but even with a resilient production system, contingency planning and operational continuity in the field and head office are essential.

5.188. In a world of rapidly changing statistical needs, it is imperative that a statistical system is able to respond quickly and well to changing demands and that it has the resilience needed to ensure continuity in the production of statistics. This is not possible if systems are old, inflexible and extensively tailored to past requirements. Building a modern statistical infrastructure of methods, tools to implement them and a technical environment to support the statistical processes, is a significant component of achieving quality CPI compilation and computation. Similarly, there needs to be contingency plans when the unexpected happens, for example, when there is a systems failure, or the price collection team is depleted by a serious bout of illness. Disaster recovery plans address these risks by, for example, saving regular and frequent back-up copies of the prices database on a secondary computer and by having the capacity to collect prices when a significant numbers of price collectors are indisposed. Two strategies have been followed to accommodate a short-term unexpected deficit in the numbers of price collectors.

a. The collection of prices from a sub-sample of shops chosen to be representative of the full sample, thereby not needing so many price collectors. The price evolution from one period to the subsequent period can then be calculated from the sub-sample using matched pairs of price observations.

b. The training of head office staff in price collection, e.g. as part of their initial induction training, so that they can provide cover. Head office can also be given responsibility for price collection on a routine basis in a shopping area close to head office, with individual head office staff allocated the task of price collection on a rotating basis. An added benefit of this is that head office staff become more familiar with the issues confronted by price collectors.

5.189. In cases where no fieldwork can be performed, for example where there is an all-out strike and no prices can be collected by visiting shops, an indicative figure may be possible using data from other sources, such as retailers websites, but only if it can be established that like-for-like comparisons are being made and that the fixed basket principle is being adhered to.

J. Other methods of price collection

5.190. This chapter has so far focussed in large part on traditional methods of price collection where price collectors visit outlets and record prices on paper forms. It now considers other methods of price collection. As noted at the beginning of the chapter, scanner data is given a separate consideration in chapter 10.

1. Electronic reporting

5.191. Electronic reporting for centrally collected prices and use of handheld computers and tablets for local price collection can introduce greater efficiency into price collection and
processing, as well as providing more scope for effective quality assurance of prices and 
auditing, but both are dependent on the introduction of effective quality control procedures. 
Electronic reporting through electronic point of sale, commonly referred to as scanner data, is 
another option.

5.192. Centrally collected data can be collected electronically in a number of ways. Once 
initial contact has been made with data suppliers, a mutually convenient electronic data 
collecting procedure can be initiated. Options include:

a. Emailing data collection spreadsheets between the NSO and the retailer.
b. Emailing of price lists at agreed times by retailers.
c. Touch-tone dialling facilities for data to be supplied in an agreed format.
d. The use of the internet (supplemented, if necessary, by telephone calls to clarify 
definitions and availability and whether the prices displayed on the internet are the same 
as those displayed in the corresponding shops). See also paragraphs 5.214 to 5.229 on 
collecting prices online and web-scraping and chapter 10 on scanner data.

5.193. The use of electronic websites/portals, where respondents can report prices online. 
This is becoming increasingly common and can also be excellent channels for 
communicating with the respondents.

2. Collection by telephone

5.194. The prices for certain items, particularly services such as electricians’ and plumbers’ 
charges and the cost of home security, may be obtained by telephoning the business or 
organisation concerned. This is the case where it is relatively easy to avoid ambiguities in 
price as the outlets provide standard items or services. However, even if prices are obtained 
by telephone, the retailer should be visited occasionally. This helps to maintain cooperation 
through personal contact and to ensure that there are no misunderstandings over the prices. 
This will be more important for some retailers than others. For example, the price of hiring a 
van may be less certain than the cost of an eye exam.

3. Computer Assisted Data Collection (CADC): the use of mobile telephones, 
handheld computers and tablets

5.195. A number of NSOs have successfully used mobile telephones, handheld computers or 
tables for local price collection. These technologies are now available at competitive prices 
and the necessary infrastructures are generally in place to make the use of CADC an 
attractive option.

5.196. A major perceived advantage of the use of a CADC system is that it leads to 
improvements in the quality of CPI data, particularly as increased quality control at the point 
of data entry helps to identify anomalies and ensure that prices are correct. CADC has the 
potential to significantly improve the quality of the final CPI in the following ways:

a. Price history. The price collection program might allow for a more comprehensive price 
history to be available to the price collector, rather than the one previous price included 
on paper forms. The availability of such data leads to less judgmental editing at the point
of data collection and helps ensure the comparability of items, particularly where prices for a particular item are variable. It has been argued that the machines should be programmed to reveal the price history only after a price quote has been entered so that collectors are not overly influenced by price when locating the item to be priced or when choosing a replacement item, but others argue that an early sight of the price history is useful information for the price collector as it assists them in their work.

b. **Quality checks in the field.** The price collection program can include several automatic validity checks which can be used to identify where the price entered varies by +/- a certain percentage from the previous month’s price and the average price for that item over a number of months and to flag up where data were not entered in all required fields (price, weight, indicator code). These checks provide a useful marker for when a price needs to be double-checked: in a paper-based system such checks are carried out in the central office after the data has been collected so audits can only be carried out after the collection period when prices may have changed.

c. **Transcription.** There is a major risk of errors when transcribing paper forms. This is not a risk when using CADC, where data can be transferred electronically to the central office.

5.197. In addition to these advantages, CADC technology allows improvements in quality through enhanced operational procedures (see paragraphs 5.204 to 5.213).

5.198. The use of CADC also significantly reduces the time taken to make data available electronically at the central office and between data collection and finalization. This can be achieved through:

a. **Transcription:** Data collected on paper must be transcribed onto a desktop computer for computation. This process is time-consuming and resource intensive. When data is collected most particularly on handheld computers or tablets, the data can be directly transferred electronically to the servers at the central office potentially in real time.

b. **Transmission from regions:** Electronic transmission will allow price collectors or regional offices to directly transmit an electronic data file to the central office thus negating the need for postal or courier services or hand deliver forms. This significantly increases the speed of data transmission to head office and reduces the cost of doing so. In addition, head offices are able to look up the latest returns of price data from all regions immediately on receipt and at the same time thereby identifying early on any issues.

c. **Quality checks in advance:** As the functionality is available to run certain quality checks in the field that would normally be run in the office after data has been transcribed, the time taken for quality checking centrally can be reduced, or, alternatively, extra supplementary checks can be carried out.

5.199. These improvements to the speed of the processing system can facilitate an earlier publication or provide opportunities to spend more time on analysis and interpretation, the production of press releases and associated briefing or the collection of more prices.

5.200. A CADC system enables certain checks that improve how efficiently the CPI is managed. These include:

a. A check that all prices have been collected before the collector leaves the outlet: an electronic data collection form can easily check whether all prices have been collected
and flag when they have not. This mitigates the risk of the price collector inadvertently forgetting to price an item.

b. A check on when prices were inputted. Electronic data collection can automatically record a date/time when the prices were entered in the machine. This is useful for validation purposes.

c. Indicator Codes. CADC provides the opportunity for additional features to be included in the data collection form. One such feature would be indicator codes, these codes (represented by a single letter) can be used to show when a price collected is for an item on sale, a replacement item, a missing item, a discontinued item etc. This is a simple tool to enhance the ease of validation and the management of the item list (see 5.90).

d. Having price histories more readily at hand can:

i. Make briefing of price collectors prior to fieldwork more effective, e.g. by a better appreciation of when an “outlier” is a legitimate price change and vice-versa.

ii. Add to the quality assurance processes through assisting with analysis when the index has been compiled and the briefing is being put together. These advantages are particularly relevant when there can be significant regional variations in price levels and trends.

5.201. Clearly there is a short-term cost associated with the introduction and implementation of CADC for CPI price collection. Costs include:

a. The purchase of equipment.

b. Upgrading ‘back-office’ systems to enable interaction with the handheld computers, mobile telephones or tablets.

c. The development of appropriate software for local price collection building on the experience of others. Costs depend on the functionality and sophistication of the program. Some NSOs have developed software for CADC that they may be ready to share.

d. Training of field staff and NSO staff on using the new systems, including pilot price collection.

5.202. There will also be longer-term costs associated with maintenance of the system and training of new staff but the additional expenditure on the latter is likely not to be significant as new staff must be trained whatever system of price collection is used.

5.203. When planning to use handheld computers, mobile telephones or tablets for local price collection, the decision needs to be made on whether the software is designed specifically for a certain hardware or not. If it is hardware dependent, the life cycle of the hardware is usually dictated by the life cycle of the software. In data transmission, the issue of confidentiality needs to be addressed as well as how to secure the transmission action in practice in terms of reliability: the speed and reliability of any network which is used should be tested from all price collection locations.

5.204. There are several advantages when using CADC, but as with all data collection methods it has its risks and limitations. With CADC the first risks are with the devices: the battery might not last the whole day of price collection, especially if price collection is done in extreme conditions, and the data might be lost if the device breaks down during price
collection period. Also, the devices have life cycles and if the software is coded to operate on a certain device this may need to be re-written for a replacement device. This causes costs when the devices come to the end of their life cycle. There are also risks in data transfer especially if the connections are poor: what happens if during downloading the data from CADC to the database the connection fails? Some of these risks can be avoided using different solutions, like programming the software independent from the device choice.

4. Collecting prices online and web-scraping

5.205. A distinction should be made between collecting the prices of goods and services purchased online, to reflect the increasing importance of the internet as a channel for making purchases, and collecting prices online as part of a strategy to reduce costs by making price collection more efficient through reducing the resource-demanding process of manual price collection. As mentioned in the introduction to this chapter, prices for online sales can differ from the prices charged at physical outlets - even for the same retailer - and the profile of goods and services purchased online can be different to other purchases. The prices in the CPI must be representative and accurate. The motivation in adopting different data sources and extraction techniques is important. The motivation for collecting prices online can be twofold - efficiency and to ensure online shopping is properly representative. The distinction is important.

5.206. Retailers with an online presence - either as sellers of goods and services online or as retailers who do not sell online but use the internet to list prices - should be treated like any other retailer and be contacted first by head office and be invited to participate in the price survey even though this will not involve a physical interface.

Online collection of prices

5.207. This section relates to collecting prices online from publicly accessible websites, those prices of goods and services also sold in the corresponding physical outlets. It is a way of increasing the efficiency of price collection for a traditional sample of retail outlets and fixed basket of goods and services. It does not relate to web-scraping, the collection of prices for online purchases or the use of scanner data.

5.208. Instead of the traditional way of sending a price collector to a retail outlet, extracting prices online directly from websites will significantly reduce the price collection costs. Similarly, the response burden on data providers will reduce close to zero when extracting prices online directly from websites to replace postal or online questionnaires. Collecting prices online is relatively straightforward although care needs to be taken and checks put in place, especially so when automated technical solutions are adopted. The prices obtained online must represent the transaction price in the physical outlet. Checks need to be made that the price advertised online is the same as the price advertised on the retailer's website and in the physical outlet and that no overheads associated with buying online, such as delivery charges, are included. Also, online data needs to contain sufficient information on

---

6 Charges relating to a delivery service arranged by the customer himself either for specific items or for bulk delivery with other items purchased are legitimate for inclusion in a CPI but should be recorded under a separate heading - transport services - as indicated in COICOP.
characteristics to detect changes in quality. When using online data sources checks also need to be made that product code numbers, if used to identify the good or service, haven’t changed between price collection periods and that the codes are unique.

5.209. Collecting prices online - which can be relatively easy and cheap - will not always be a suitable substitute for price collectors, however enticing it is.

**Collection of prices for online purchases**

5.210. Goods and services purchased via the internet need to be properly reflected in the sample of prices used to compile the CPI.

5.211. Elementary product groups should be stratified to reflect products purchased online. Sale information for weighting and the drawing of samples of internet purchases can be taken from HBSs, which should record information about outlet-type (including internet purchases), and from information supplied by online retailers and market research companies.

5.212. Prices associated with online purchases can be extracted from retailer's websites in the same way as the online collection from websites of prices charged by physical outlets. The sample of items to be priced should be representative of all online purchases and will be different from online collection of prices charged by physical outlets. The prices recorded should represent the full cost of purchase. Online purchases may include standard extras such as delivery charges. Unavoidable charges that are directly connected to the purchase of the priced product and which are not separately invoiced should be included in the price for the purpose of CPI compilation. If the charge is separately invoiced or relates to the purchase of a number of items, then the treatment is less clear cut. One option is to include these charges under transport services but issues relating to the classification of individual consumption by purpose (COICOP) arise (for more information on COICOP, see chapter 2). Another option is to follow the HICP approach where such unavoidable charges that are not part of the basic advertised price may be treated as an inseparable bundle of a good and a service and can be treated as one product (for additional information, see the section on internet purchases in chapter 11).

**Web-scraping**

5.213. Web-scraping is the process of automated collection of data from the internet through a set of computer software techniques for extracting information from websites (Webpages) or using an Application Programming Interface (API), which is a set of routines, protocols, and tools for building software applications. Web-scraping identifies and retrieves relevant data and downloads and organises it in a suitable format for computing a CPI.

5.214. It is important to be aware that there are technical measures applied in some websites to avoid web-scraping activities. These measures are based mainly by blocking the scraper IP address access to the site or by blocking the response to the scraper http browser agent identification. The action is triggered after identifying an ‘abnormal’ behaviour (by analysing the activity log) or by filtering access from some agents (through the robots.txt server configuration file)\(^7\).

---

\(^7\) Web Robots (also known as Web Wanderers, Crawlers, or Spiders), are programs that traverse the Web automatically. Search engines use them to index the web content. Website owners use the /robots.txt file to give
5.215. Anti-web-scraping devices by retailers reinforces the need to gain the cooperation of online retailers prior to data scraping to avoid being blocked from data collection. Retailers should always be clearly informed about the nature, extension, and frequency of web-scraping actions prior to any web-scraping taking place. NSOs should inform and ask permission from retailers and agree to the most suitable web-scraping technique with the retailer’s management. There may also be legal constraints to web-scraping.

5.216. When undertaking web-scraping, the same considerations apply as with online collection, most particularly whether the aim is to download from publicly accessible websites, the prices of goods and services sold in physical outlets to increase the efficiency of price collection or whether to download prices paid for online purchases.

5.217. For web-scraping to replace traditional price collection, it needs to be demonstrated that prices, both online and in traditional outlets, are the same. While true for some retailers and some products, it is almost certainly not true for all retailers and all products. In such cases, the price collection from a website should be seen as a different outlet type, which should be sampled along with traditional outlets and then combined with prices collected via the traditional collection methods. It should be noted that to integrate these prices into a CPI it is necessary to evaluate if there are some extra fees associated with the purchase and not included in the prices listed on the website (see collection of prices for online purchases).

5.218. As mentioned in the introduction to this chapter, integration of prices from different data sources needs to take into account differences in sampling regimes (for instance, the relatively bigger samples facilitated by web-scraping) and the relative values of sales. It is one of the reasons why online purchases are often treated as separate elementary aggregates with separate weights.

5.219. Appendix F provides more details on web scraping.

5. Calculation of average price from different data sources in the elementary aggregate

5.220. Another aspect to be taken in consideration when combining traditional data sources with web-scraped data is the different collection frequencies. Traditional price collection deals with price ‘snapshots’ scheduled in such a way that the price series for a product in a shop respects both frequency and equidistance in terms of time. In contrast, one of the perceived advantages of web-scraped data is that it can collect a price daily during a certain period, extract the average value and take that as the ‘snapshot’ price for that week. A problem arises when trying to integrate data collected in ‘continuous daily’ frequency into the regular monthly CPI. When using web-scraped data, care must be taken not to apply the raw collected data directly on the calculation of the average monthly price at the elementary aggregate level. Since the number and nature of observations (probably much more than one per month per outlet/period for price obtained by web-scraping) are different, they should be transformed into compatible data. This can be accomplished simply by calculating an ‘outlet monthly price’ for the product-offer. This average price will then have the same importance
as other prices generated by the traditional snapshot approach. The consequences of using directly the online prices will lead to a distorted estimate of inflation, since average prices will be calculated using a disproportionate number of prices that came from online collection, regardless of any relative weight information. Appendix E gives a brief example of these calculations and how a disregard of the differences in the frequency of price collection can lead to an under-estimate of inflation. However, the use of an empirical country-based data set is needed for a more accurate evaluation.

5.221. There are a variety of tools to aid the scraping activity. For example, many programs are available in common programming languages (C, Python, JavaScript, etc.), for standalone applications or, most commonly, as add-ins to the browser. In most cases, web-scraping tools are primarily designed to fulfil web application testing and verification. For that reason, most of the tools used are implemented as browser add-ins (or plug-ins). This is not the ideal situation in terms of IT architecture. See the references at the end of this chapter for further reading.

5.222. When combining prices using traditional price collection techniques with prices obtained from web-scraping, allowance needs to be made in the computation of average product prices for the different collection techniques being deployed, especially the frequency of price collection where web-scraping is sometimes carried out at a greater frequency than traditional price collection. Again, the reader is referred to Appendix E.

**Summary of recommendations**

- Price collection methods and organization decisions will depend upon country specific circumstances and available resources.

- Collect prices should reflect actual transaction prices including any tax and reflecting any discounts, sales, or promotions.

- Items should be priced as often as necessary to ensure that the index reflects a reliable and meaningful measure of price change.

- NSOs should strive to calculate an index based on prices covering the whole period (e.g. month or quarter).

- The interval between price observations should be uniform for each outlet.

- Price collection period should be made available in advance to all users and stakeholders.

- Proper training of price collectors is essential. Detailed documentation on data collection procedures and processes should be drafted and made available to data collectors and CPI staff.

- Data validation techniques are necessary to ensure the accuracy and reliability of the collected prices.
• Outlier detection methods should be defined and implemented. All questionable prices should be verified and corrected as necessary.

• Quality assurance procedures should be defined and implemented to ensure the accuracy of collected prices.
Appendix A Price collection procedures

Is the outlet open?
- Yes
  - Is it permanently shut?
    - Yes
      - Select another outlet according to instructions. Report it.
    - No
      - Is the outlet willing to cooperate?
        - Yes
          - Report the facts to HO
        - No
          - Is the product type to be priced currently available to purchase?
            - Yes
              - Report the non-availability and the reason for it.
            - No
              - Is there a large difference from the price recorded last month?
                - Yes
                  - File a deviation for damaged or sold items.
                - No
                  - Is there a difference between the special offer, black market price, replacement offer, or replacement item?
                    - Yes
                      - Disregard the price.
                    - No
                      - Is the outlet likely to be permanently shut next month?
                        - Yes
                          - Is the variety expected to become permanently unavailable next month?
                            - Yes
                              - Report the facts to HO
                            - No
                              - Price next item.
                        - No
                          - State the reason, e.g. sale, special offer, black market price, replacement offer, replacement item.
                      - Is the outlet authorized to find a replacement?
                        - Yes
                          - Is the replacement the same or is it a substitute?
                            - Yes
                              - Report the replacement.
                            - No
                              - Select another variety likely to remain available. Record a description of it in sufficient detail to cover quality differences and to enable exact identification.
                        - No
                          - Estimate and report the amount of price difference reflecting the value of the quality difference, revealed by the descriptions of the original variety and its replacement.

- No
  - Try again later if the instructions say so, otherwise report the facts.
Appendix B Example of a price collection form

### Consumer Price Index: Pricing Form - September 2008

<table>
<thead>
<tr>
<th>Collection Period:</th>
<th>WEEK 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unique Number:</td>
<td>8370435</td>
</tr>
<tr>
<td>Brand Name:</td>
<td>Woolworths</td>
</tr>
<tr>
<td>Product Name:</td>
<td>Sandwich White Bread</td>
</tr>
<tr>
<td>Quantity:</td>
<td>1 Loaf</td>
</tr>
<tr>
<td>Size:</td>
<td>700</td>
</tr>
<tr>
<td>Origin:</td>
<td>South Africa</td>
</tr>
<tr>
<td>Code:</td>
<td>K043</td>
</tr>
<tr>
<td>Feature:</td>
<td>Product presentation</td>
</tr>
<tr>
<td>Description:</td>
<td>Plastic bag</td>
</tr>
<tr>
<td>Variety:</td>
<td>Sliced</td>
</tr>
</tbody>
</table>

#### Price

<table>
<thead>
<tr>
<th>2008/08</th>
<th>2008/09</th>
</tr>
</thead>
<tbody>
<tr>
<td>R 7.95</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type of price:</th>
<th>REGULAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantity:</td>
<td>1 Loaf</td>
</tr>
<tr>
<td>Size:</td>
<td>700</td>
</tr>
<tr>
<td>Unit of size:</td>
<td>Gram</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LAST REGULAR PRICE</th>
<th>QUANTITY</th>
<th>SIZE</th>
<th>UNIT</th>
<th>COLLECTION PERIOD</th>
</tr>
</thead>
<tbody>
<tr>
<td>R 7.95</td>
<td>1 Loaf</td>
<td>700</td>
<td>Gram</td>
<td>2008/08</td>
</tr>
</tbody>
</table>

### Field Message:

Office Use Only
Unit Status Code:
Appendix C Automated data checking

*Worked Example 1*

Worked Example 1 demonstrates the use of median and quartile values. Table C.1, column 1, shows the price ratios for the illustrative sample.

<table>
<thead>
<tr>
<th>Price Ratio</th>
<th>$S_i$</th>
<th>Flagged using PR Above 1.08723 or below 0.87628</th>
<th>Flagged using $S_i$ Above 0.10894 or below -0.10894</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8138</td>
<td>-0.20638</td>
<td>Extreme</td>
<td>Extreme</td>
</tr>
<tr>
<td>0.8525</td>
<td>-0.15161</td>
<td>Extreme</td>
<td>Extreme</td>
</tr>
<tr>
<td>0.8760</td>
<td>-0.12072</td>
<td>Extreme</td>
<td>Extreme</td>
</tr>
<tr>
<td>0.8990</td>
<td>-0.09205</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>0.9086</td>
<td>-0.08051</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>0.9135</td>
<td>-0.07471</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>0.9339</td>
<td>-0.05124</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>0.9414</td>
<td>-0.04286</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>0.9553</td>
<td>-0.02769</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>0.9608</td>
<td>-0.02180</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>0.9658</td>
<td>-0.01651</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>0.9668</td>
<td>-0.01546</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>0.9702</td>
<td>-0.01190</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>0.9724</td>
<td>-0.00962</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>0.9817</td>
<td>-0.00005</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>0.9818</td>
<td>0.00005</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>0.9843</td>
<td>0.00260</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>0.9869</td>
<td>0.00525</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>1.0034</td>
<td>0.02205</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>1.0050</td>
<td>0.02368</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>1.0061</td>
<td>0.02480</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>1.0301</td>
<td>0.04925</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>1.0571</td>
<td>0.07675</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>1.0824</td>
<td>0.10252</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>1.0909</td>
<td>0.11118</td>
<td>Extreme</td>
<td>Extreme</td>
</tr>
<tr>
<td>1.0931</td>
<td>0.11342</td>
<td>Extreme</td>
<td>Extreme</td>
</tr>
<tr>
<td>1.1300</td>
<td>0.15101</td>
<td>Extreme</td>
<td>Extreme</td>
</tr>
<tr>
<td>1.1550</td>
<td>0.17647</td>
<td>Extreme</td>
<td>Extreme</td>
</tr>
<tr>
<td>1.2296</td>
<td>0.25246</td>
<td>Extreme</td>
<td>Extreme</td>
</tr>
<tr>
<td>1.2304</td>
<td>0.25327</td>
<td>Extreme</td>
<td>Extreme</td>
</tr>
</tbody>
</table>
\[ S_i = 1 - \frac{R_M}{R_i} \]

if \( 0 < R_i < R_M \)

\[ S_i = \frac{R_i}{R_M} - 1 \]

if \( R_i \geq R_M \)

The first and third quartiles (\( R_{Q1} \) and \( R_{Q3} \)) and the median (\( R_M \)) can be obtained using the quartile function in Microsoft Excel. The average distance of the quartiles from the median (\( D_M \)) is defined as:

\[ D_M = \frac{(R_{Q3} - R_{Q1})}{2} \]

The upper and lower limits are then calculated as:

\[ L_U = R_M + C \times D_M; \text{ and} \]

\[ L_L = R_M - C \times D_M \]

where the multiplier \( C \) is a user defined value and has been set equal to 2 to limit the number of observations flagged up as potential errors.

The resulting upper and lower limits are shown in the “Series PR” column in Table C.2.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Series PR</th>
<th>Series S_i</th>
</tr>
</thead>
<tbody>
<tr>
<td>( R_{Q1} )</td>
<td>0.94488</td>
<td>-0.03907</td>
</tr>
<tr>
<td>( R_M )</td>
<td>0.98175</td>
<td>0.00000</td>
</tr>
<tr>
<td>( R_{Q3} )</td>
<td>1.05035</td>
<td>0.06988</td>
</tr>
<tr>
<td>( D_M )</td>
<td>0.05274</td>
<td>0.05447</td>
</tr>
<tr>
<td>( C )</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>( L_L )</td>
<td>0.87628</td>
<td>-0.10894</td>
</tr>
<tr>
<td>( L_U )</td>
<td>1.08723</td>
<td>0.10894</td>
</tr>
</tbody>
</table>

As described in the main part of the chapter, the price ratio series can be transformed to provide more equal weighting between negative and positive price movements. The transformations are repeated here as:

\[ S_i = 1 - \frac{R_M}{R_i} \]

if \( 0 < R_i < R_M \)

\[ S_i = \frac{R_i}{R_M} - 1 \]

if \( R_i \geq R_M \)
The transformed observations are shown in the column labelled ‘Series Si’ in Table C.1. The quartiles, median and calculated limits for the transformed series are shown in the third column in Table C.2. The increased value for $D_M$ for the transformed sample shows that the transformation has increased the distances for the price decreases while leaving the distances for positive movements the same.

The last two columns in Table C.3 show, respectively, the observations that would be flagged for further examination (indicated by the word ‘extreme’) for the original price ratios and the transformed price movements.

**Worked Example 2**

Worked Example 2 demonstrates the same statistical filtering method than Worked Example 1, but with an additional 16 price ratios added to the sample. All the new price ratios show zero change. The same calculations are done but on a sample of 46 instead of 30 observations. Table C.3 shows the sample of price ratios and the transformed price movements, as well as the observations flagged for further observations. Table C.4 shows the parameters and calculated limits.

<table>
<thead>
<tr>
<th>Price Ratio</th>
<th>Si</th>
<th>Flagged using PR Above 1.03490 or below 0.96510</th>
<th>Flagged using Si Above 0.03598 or below -0.03598</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8138</td>
<td>-0.2288</td>
<td>Extreme</td>
<td>Extreme</td>
</tr>
<tr>
<td>0.8525</td>
<td>-0.1730</td>
<td>Extreme</td>
<td>Extreme</td>
</tr>
<tr>
<td>0.8760</td>
<td>-0.1416</td>
<td>Extreme</td>
<td>Extreme</td>
</tr>
<tr>
<td>0.8990</td>
<td>-0.1123</td>
<td>Extreme</td>
<td>Extreme</td>
</tr>
<tr>
<td>0.9086</td>
<td>-0.1006</td>
<td>Extreme</td>
<td>Extreme</td>
</tr>
<tr>
<td>0.9135</td>
<td>-0.0947</td>
<td>Extreme</td>
<td>Extreme</td>
</tr>
<tr>
<td>0.9339</td>
<td>-0.0708</td>
<td>Extreme</td>
<td>Extreme</td>
</tr>
<tr>
<td>0.9414</td>
<td>-0.0622</td>
<td>Extreme</td>
<td>Extreme</td>
</tr>
<tr>
<td>0.9553</td>
<td>-0.0468</td>
<td>Extreme</td>
<td>Extreme</td>
</tr>
<tr>
<td>0.9608</td>
<td>-0.0408</td>
<td>Extreme</td>
<td>Extreme</td>
</tr>
<tr>
<td>0.9658</td>
<td>-0.0354</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>0.9668</td>
<td>-0.0343</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>0.9702</td>
<td>-0.0307</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>0.9724</td>
<td>-0.0284</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>0.9817</td>
<td>-0.0186</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>0.9818</td>
<td>-0.0185</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>0.9843</td>
<td>-0.0160</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>0.9869</td>
<td>-0.0133</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>1.0000</td>
<td>0.0000</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>1.0000</td>
<td>0.0000</td>
<td>OK</td>
<td>OK</td>
</tr>
</tbody>
</table>
A comparison of the results from the two examples demonstrates the effect of having a significant number of observations with no price movement. The distance from the median (DM) is reduced and the number of observations flagged for further examination is significantly increased.

**Worked Example 3**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Series PR</th>
<th>Series Si</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_{Q_1}$</td>
<td>0.96765</td>
<td>-0.03343</td>
</tr>
<tr>
<td>$R_M$</td>
<td>1.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>$R_{Q_3}$</td>
<td>1.00255</td>
<td>0.00255</td>
</tr>
<tr>
<td>$D_M$</td>
<td>0.01745</td>
<td>0.01799</td>
</tr>
<tr>
<td>$C$</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>$L_L$</td>
<td>0.96510</td>
<td>-0.03598</td>
</tr>
<tr>
<td>$L_U$</td>
<td>1.03490</td>
<td>0.03598</td>
</tr>
</tbody>
</table>
Worked Example 3 demonstrates the alternative statistical filtering method, the Tukey algorithm. The enlarged sample from Worked Example 2 is used here to demonstrate the benefit of this method when the sample has a large proportion of price ratios indicating no movement. Table C.5 presents the intermediate data stages in addition to the basic sample and the indicator flagging extreme observations as possible errors.

Table C.5 Price relatives showing movement from previous period - Example 3

<table>
<thead>
<tr>
<th>Price Relative</th>
<th>Less 5% tails and Less zero movement</th>
<th>T lower set</th>
<th>T upper set</th>
<th>Flagged as extreme</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8138</td>
<td></td>
<td></td>
<td></td>
<td>Extreme</td>
</tr>
<tr>
<td>0.8525</td>
<td></td>
<td></td>
<td></td>
<td>Extreme</td>
</tr>
<tr>
<td>0.8760</td>
<td>0.8760</td>
<td>0.8760</td>
<td></td>
<td>Extreme</td>
</tr>
<tr>
<td>0.8990</td>
<td>0.8990</td>
<td>0.9086</td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>0.9086</td>
<td>0.9086</td>
<td>0.9086</td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>0.9135</td>
<td>0.9135</td>
<td>0.9135</td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>0.9339</td>
<td>0.9339</td>
<td>0.9339</td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>0.9414</td>
<td>0.9414</td>
<td>0.9414</td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>0.9553</td>
<td>0.9553</td>
<td>0.9553</td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>0.9608</td>
<td>0.9608</td>
<td>0.9608</td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>0.9658</td>
<td>0.9658</td>
<td>0.9658</td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>0.9668</td>
<td>0.9668</td>
<td>0.9668</td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>0.9702</td>
<td>0.9702</td>
<td>0.9702</td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>0.9724</td>
<td>0.9724</td>
<td>0.9724</td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>0.9817</td>
<td>0.9817</td>
<td>0.9817</td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>0.9818</td>
<td>0.9818</td>
<td>0.9818</td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>0.9843</td>
<td>0.9843</td>
<td>0.9843</td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>0.9869</td>
<td>0.9869</td>
<td>0.9869</td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td>OK</td>
</tr>
</tbody>
</table>
The price ratios for the sample are shown in the first column. The first step was to remove the highest and lowest 5 per cent of price ratios. Five per cent of this sample equals 1.5 observations. This was rounded up to 2 observations so the two highest and the two lowest price ratios were removed. Observations with zero price movement were also removed. The remaining observations are shown in the second column. The arithmetic mean (AM) of the remaining set of observations is calculated. This value, along with other parameter calculations is shown in Table C.6. The arithmetic means of the lower and upper sets of data are then calculated (labelled $AM_L$ and $AM_U$ respectively). The lower and upper data sets have been presented in columns 3 and 4, respectively, of Table C5 purely for explanatory purposes. The Tukey lower and upper limits are then calculated as:

$$T_L = AM - 2.5(AM - AM_L)$$

$$T_U = AM + 2.5(AM_U - AM)$$

The results are shown in Table C.6.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>0.99429</td>
</tr>
<tr>
<td>AM_L</td>
<td>0.94990</td>
</tr>
<tr>
<td>AM_U</td>
<td>1.06531</td>
</tr>
<tr>
<td>T_L</td>
<td>0.88332</td>
</tr>
<tr>
<td>T_U</td>
<td>1.17184</td>
</tr>
</tbody>
</table>

Using this method, five observations would be selected for further examination - many fewer than the 18 selected in example 2.
Appendix D Documentation Control Template

The documentation control template is an essential element of documentation, production and dissemination and control. Documentation control contributes to better quality management as access to non-authors is restricted to "read only". It provides checks, background (including explanations for changes) and an audit trail. Two further benefits accrue when combined with an electronic system.

- More efficient production of documentation as it helps with initial compilation and updates and reduces the need to print and circulate paper copies.
- Better informed staff because they have immediate electronic access to the latest documentation, including desk instructions, with search facility by subject and author.

<table>
<thead>
<tr>
<th>DATE ISSUED</th>
<th>DOCUMENTATION</th>
<th>Ref</th>
<th>DETAILS OF CHANGE</th>
<th>REASON FOR CHANGE</th>
<th>NAME OF ISSUER</th>
</tr>
</thead>
<tbody>
<tr>
<td>XX/XX/XX</td>
<td>Calculating CPI food item weights (non-seasonal)</td>
<td>2.1</td>
<td>Change in process with effect from...</td>
<td>CPI Technical Board has agreed that in the future weights should be taken from National Accounts.</td>
<td>Mr Smith, Consumer Price Statistician</td>
</tr>
<tr>
<td>XX/XX/XX</td>
<td>Calculating and updating price index for telecommunications services</td>
<td>2.5</td>
<td>Change in process with effect from...</td>
<td>Methodology changes in pricing structures for mobile telephones – new methodology agreed by CPI Technical Board. Reflects changing market.</td>
<td>Mr Smith, Consumer Price Statistician</td>
</tr>
<tr>
<td>XX/XX/XX</td>
<td>Desk instructions for checking and editing of prices</td>
<td>3.1</td>
<td>Additional checks to be carried out based on month-on-month price change.</td>
<td>Last audit indicated current checks inadequate resulting in incorrect prices entering the CPI.</td>
<td>Ms Brown, CPI Operations Manager</td>
</tr>
</tbody>
</table>
Appendix E The calculation of average product price when combining prices from different price collection methods and for different price collection frequencies

1. When combining the traditional with the web-scraping price collection methods some care must be taken regarding average product price computation. There are two options: web-scraping can follow the same calendar collection that is used in traditional collection. In this option, the advantages of web-scraping are disregarded. The second option will be to collect prices more frequently with the web-scraping techniques. This will increase the number of observations used in the computation, potentially providing more reliability to the estimation. In the following simplified example relating to just one item, the price for a specific product A is collected in a sampled local shop in the framework of traditional price collection twice along the considered time frame of 10 days. The price for the same product is collected on a daily basis using web-scraping. This data is collected for two months and the price at day 1 and at day 2 is the same for both outlet/collection types. However, the price averages and the indexes will differ according to the way these are computed at the elementary aggregate level.

2. When calculating the average price (geometric mean) there are two ways of computing it at the elementary aggregate level. The assumption is made that there are no expenditure weights (and each observation at a given point in time has the same weight) and prices are collected to be representative of all sales.

   a. Method 1 – compute the geometric mean in two steps: first an average price is calculated by kind of outlet/collection methodology; secondly a geometric average is calculated with the two average prices per outlet.

   b. Method 2 – compute the geometric mean in a single stage using all prices.

3. Taking into consideration the different collection frequencies, method 2 will undervalue price change (noting that in this example prices are falling) since it takes into account a huge number of observations taken day by day where the price change is not significant and in a way it is providing more ‘weight’ to the internet shop with a price series that is more stable than to the local shop where the price being observed each 10 days is recording a higher change. Due to these differences and the circumstance described, the method 1 is recommended since it will provide the same ‘weight’ to both kinds of outlets/collection methods. In summary, the prices can only be averaged in a single step when there are an equal number of price observations generated by each price collection for each item over a given period. Averaging prices in a single step when the frequency of price collection varies between different price collection methodologies leads to distorted results.
<table>
<thead>
<tr>
<th>product A</th>
<th>Month 1 Days</th>
<th>Geometric mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 3 4 5 6 7 8 9 10</td>
<td></td>
</tr>
<tr>
<td>local shop - traditional collection</td>
<td>500</td>
<td></td>
</tr>
<tr>
<td>Internet website - web scraping</td>
<td>500 500 500 497 497 490 489 483 480 480</td>
<td></td>
</tr>
<tr>
<td>Geometric mean: Method 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geometric mean: Method 2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>product A</th>
<th>Month 2 Days</th>
<th>Geometric mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 3 4 5 6 7 8 9 10</td>
<td></td>
</tr>
<tr>
<td>local shop - traditional collection</td>
<td>450</td>
<td></td>
</tr>
<tr>
<td>Internet website - web scraping</td>
<td>450 452 452 448 445 445 445 440 435 430</td>
<td></td>
</tr>
<tr>
<td>Geometric mean: Method 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geometric mean: Method 2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Indexes

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Local shop - traditional collection</td>
<td>89.79</td>
</tr>
<tr>
<td>Internet website - web scraping</td>
<td>90.25</td>
</tr>
<tr>
<td>geomean 1</td>
<td>90.02</td>
</tr>
<tr>
<td>geomean 2</td>
<td>90.17</td>
</tr>
</tbody>
</table>
Appendix F: Web scraping

Introduction

The collection of prices is an integral component of the production of a CPI. Several data collection modes are available and used by countries. These include personal visits, online, telephone, administrative data, and transactions data. More recently, with the growth of online retailing, pricing information may be obtained directly from websites. Advances in technology and automated scraping software have enabled large scale data collection from the internet. This is referred to as web scraping.

Web scraping extracts large amounts of data from websites. Prices can be collected as frequently as needed for all products using purpose-built programs that scan the websites of retailers, find the relevant information, and store the information in a time series. The process can be run automatically and as frequently as needed (daily/weekly), providing high frequency information.

This annex provides insights on web scraping experiences from NSOs who have implemented web scraping to compile their CPI. A general description of processes and approaches to web scraping are described in detail.

Benefits of web scraping

Web scraping enables many more products to be priced, and for these products to be priced more often than would be possible using traditional data collection methods. Web scraping provides an opportunity to significantly enhance the sample of products and prices collected by expanding product coverage. Other benefits of web scraping include:

- Automated data collection at a reduced collection cost;
- Enhanced price representativity;
- Increased price collection frequency (i.e., daily vs once per month) allows for a more representative price for the period to be obtained. Secondary effects of using an average period price include reduced volatility;
- Faster identification of new and disappearing products;
- Reduced respondent burden; and
- Rich source of metadata (i.e., product characteristics) that can be extracted, stored and potentially used for explicit quality adjustment (e.g., hedonics). Such metadata may also complement other data sources such as transactions data.

Limitations of web scraping

The lack of expenditure information from web scraping means that products/product groupings cannot be weighted by economic importance in a timely manner. A lack of expenditure information also limits the types of index formulas that can be used to compile the index. Other potential limitations of web scraping include:
Web scraping is limited to retail outlets that have an online presence (i.e., potential for under coverage);

Web scraping requires regular IT maintenance. Website changes may cause the web scraping program to fail. Businesses can block IP addresses if they detect the web scraping activity and wish to prevent it;

If performed within the NSO, web scraping requires analysts with intermediate programming knowledge to deal with the regular IT maintenance; and

Web scraping distinct prices for different geographic regions may be difficult if the website detects the physical location of the computers IP address. This may require assumptions that retailers use national pricing for regional price indexes.

The Process of web scraping

For those NSOs using web scraped data for research and/or production purposes, the process of performing web scraping has focused on two main methods:

- Web scraping performed within the NSO using statistical software (see Van Loon and Roels, 2018); and
- Web scraping services procured from a third party/private company (see Bentley and Krsinich, 2017).

The choice to perform web scraping within the NSO versus contracting with an external vendor will depend on the local context in which the NSO operates (i.e., budgets, NSO programming capacity, maintenance costs, etc.). With respect to those NSOs performing web scraping internally using statistical software, the process of web scraping information from the internet can be broken down into three main steps:

- Confirm the website allows scraping;
- Scraping the website; and
- Cleaning the collected data.

There are two main ways to determine whether a website is eligible for web scraping. First, staff setting up web scrapers should check the terms and conditions section of a website for “conditions of use”. Here, websites will often specify if web scraping is allowed or prohibited. Additionally, a “robots.txt” file can be located within the root directory of a website. These text files contain detailed information and may outline the conditions for web scraping possibly including who is allowed to web scrape, which information is available for scraping, and anything for which scraping is forbidden.

Once it is determined that a website is available for scraping, a scraper is set up for the website. The scraper locates the website’s category structure and identifies all the relevant categories to be scraped. The programmer defines the parts of the structure to be included and excluded. For example, a website may list all the individual product categories, then additional categories such as “new products” or “all products” which duplicate the products in the individual categories. These additional categories can be excluded by the programmer.
The scraper then proceeds to download all products and prices from the internet. An attempt is made to show as many products on each page as possible, by experimenting with URL options on the website prior to setting up the scraper. There are two options available for pulling in the information. First, data may be collected as text by copying and pasting from the website. In this case, the process of cleaning the text is carried out following the scraping. Alternatively, products and prices designated according to a particular tag or class name can be pulled in via the HTML tags which can provide a more targeted approach to extracting and cleaning the data. Another advantage of this approach is that product identifiers can occasionally be hidden in the HTML so pulling them in using the tags allows these to be added to the product description (as opposed to relying purely on the text description). However, the use of HTML tags is not easy for every website.

For information collected as text, the raw data need to be cleaned post-scraping so that only the set of products and prices remain. A pattern in the data needs to be uncovered and coded to separate the products and prices from the “noise” (including removing all the information prior to and after the list of products). For products which are on sale, multiple prices may be listed. In these instances, the scraper records the sale prices as the current price of the product.

In terms of the available information to construct price indexes, Table F.1 below provides a summary of a typical metadata scraped by NSOs. In summary, the data frame will typically include:

- Date: specific day of the scrape (date);
- Retailer: name of the retailer (text);
- Category: retailer’s website classifications (text);
- Product ID: text description of product (text); and
- Price: specific price of product (numeric).

<table>
<thead>
<tr>
<th>Date</th>
<th>Retailer</th>
<th>Category</th>
<th>Product ID</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>10-Jul-16</td>
<td>Retailer ABC</td>
<td>Women’s Shirts</td>
<td>Brand XYZ - Short Sleeve Polo Shirt</td>
<td>$45.00</td>
</tr>
<tr>
<td>10-Jul-16</td>
<td>Retailer ABC</td>
<td>Women’s Shirts</td>
<td>Brand XYZ - S/S Regular Shirt</td>
<td>$55.00</td>
</tr>
<tr>
<td>10-Jul-16</td>
<td>Retailer ABC</td>
<td>Women’s Shirts</td>
<td>Brand XY - Short Sleeve Regular T-Shirt</td>
<td>$15.00</td>
</tr>
<tr>
<td>10-Jul-16</td>
<td>Retailer ABC</td>
<td>Women’s Shirts</td>
<td>Brand XYZ - Long Sleeve Regular Shirt</td>
<td>$65.00</td>
</tr>
<tr>
<td>10-Jul-16</td>
<td>Retailer ABC</td>
<td>Women’s Shirts</td>
<td>Brand XYZ - Short Sleeve Regular Shirt</td>
<td>$35.00</td>
</tr>
</tbody>
</table>
Practical considerations

The basic information required to compile a price index includes prices, expenditure information (or reasonable assumptions on substitution if no expenditure information is available) and classifications (both for products and product groupings). While web scraped data (such as the example in Table F.1) may appear reasonably consistent with these requirements, it is important to ensure the conceptual requirements of the CPI (e.g., quality adjustment) are maintained. According to Bentley and Krsinich (2017, p.6) “Big data (transaction, online and administrative data) is ‘found data’ in the sense that measuring CPI inflation is a secondary use of the data – the data was not created with this use in mind”.

This sub-section describes some of the main challenges identified by NSOs when using web scraped data, including:

- Classifying web scraped data;
- Options to define individual products; and
- Index aggregation options.

Classification of web scraped data

The classification of web scraped data involves similar considerations as described in Chapter 10 for the use of scanner data. Web scraped data typically has some basic product text and category description that are required to be mapped to a NSOs hierarchical classifications (e.g., COICOP). Approaches considered by NSOs to solve these classification problems include:

- Text string searches: check for the presence or absence of keywords in the description string for classification.
- Category mapping: some of the datasets (or parts thereof) contain retailer categories for each product; if one of these categories sits within a classification, the category can be mapped to the classification.
- Supervised learning algorithms: provide training data (e.g., using one or multiple methods above) to statistical learning algorithm that identifies patterns between text and training decision for automatic classification.

Options to define individual products

An essential part of price measurement is accounting for quality change and the introduction of new products. The CPI measures the price inflation in a basket of goods and services priced at constant quality. If the quality of a product changes over time, then prices are adjusted so that the index movements reflect pure price change. This has important implications for web scraping.

The appearance and disappearance of products from a CPI sample has the potential to bias
the index unless any corresponding changes in the quality of the sample are dealt with appropriately. This poses a problem for the calculation of indexes incorporating all (or most) web scraped prices due to the large number of prices these datasets contain, the high rate of product attrition, and the tendency for products to have unusual price movements near the start and end of their life cycles.

A natural approach for dealing with this problem is to estimate the price change between two periods using the products available at both time points only, thus excluding the prices of new and disappearing products. Clearly, the matched model method (as described in Chapter 6) involves discarding information about new and disappearing products. However, it gains strength from the comprehensive coverage from the census of products represented in the full web scraped dataset. This can be considered as an application of the overlap method, as it is almost certain that sales of new and disappearing products would overlap with the sales of other products that consumers would use as substitutes.

An assumption behind the overlap method is that price differences between products are reflective of quality differences. This assumption seems reasonable in a competitive marketplace and in normal circumstances; however, however, disappearing products are sometimes sold at discounted prices to clear remaining stock (end of lifecycle), and if not linked to a product of comparable quality, may produce a “relaunch” problem and could potentially create a downward bias in the index (Chessa, 2016). This problem has been identified in various price index studies on different types of products including high technology goods (Silver and Heravi, 2005), clothing (Chessa, 2017) and personal care products (Chessa, 2016).

To overcome this problem, several practical strategies have been proposed, primarily focused on extracting characteristics information (e.g., brand, shirt type) from text strings to form broader product definitions. Techniques proposed include:

- Use of a broader product category (e.g., women’s shirts in Table F.1).
- Text/regular expression functions: used to extract characteristics from semi-structured text data (e.g., text left of the hyphen in Table F.1 extracts the characteristic “brand”).
- Approximate (fuzzy) matching functions: used to approximately match text strings using a penalty function.
- Supervised learning algorithms (see Abe and Shinozaki, 2018): provide training data to statistical learning algorithm that identifies patterns between text and training decision for product classification.
- Unsupervised learning algorithms (see Metcalfe et al, 2016): use characteristics (e.g., text string, price) and algorithms to automatically define “clusters” of products.

Options to aggregate prices

Web scraped data do not contain quantity or expenditure information. This would naturally restrict the choice of index formula to an unweighted price index formula (e.g., Jevons), which is the current choice for various NSOs in the CPI production.

The use of web scraped data in compiling price indexes continues to be the subject of
extensive research. Some researchers (Chessa and Griffioen, 2016) have experimented with approximating expenditure weights based on data observed from the websites. Using brand and product type definitions and the number of products as a proxy for quantity data, results demonstrate that the web scraped data approximate a benchmark index (scanner data) using the Geary-Khamis (GK) method. Additional published (e.g., Metcalfe et al 2016) experiment with several bilateral and multilateral price index methods. Research findings showed a substantial amount of drift across most bilateral and multilateral indexes with granular product definitions for clothing products. To minimize this drift problem, they propose the Clustering Large Datasets into Price Indexes (CLIP) as an alternative method. The CLIP calculates a fixed based Jevons with broader product definitions (derived via an unsupervised learning algorithm).

Country case study – web scraping

As part of a broader project to modernise the CPI, one NSO\(^8\) began collecting prices using web scrapers beginning May 2016. Web scrapers are currently programmed and maintained by the NSO staff using Microsoft excel (Visual Basic for Applications), collecting approximately 500,000 prices per week across more than 50 retailers.

From the second quarter 2017, the NSO began using web scraped data in the calculation of the CPI. Varieties are selected using the same methods as other forms of price collection (field collection, online collection, or other methods). An average price is calculated for each item over a given period.

Care is taken to select representative and stable products. When transitioning a respondent from using traditional data collectors to web scraped prices, an attempt is made to link the current field collected product to the identical product on the website. A determination regarding quality adjustment is made for each product to ensure that the new web scraped product’s base period price is correct.

For each web scraped product in the sample, the price for a given period (month or quarter) is an arithmetic average of the prices which fall within the specified period. If a product disappears, a replacement product is selected from among the other products within the relevant category from that respondent. A quality adjustment is performed to ensure that only pure price change is shown. The combination of category name, brand/product description and price history of both the old and new product is sufficient to enable an accurate quality adjustment to be applied.

The NSO has adopted a phased approach to implementing new retailers and price index methods using web scraped data. With respect to respondents, an assessment is made for each respondent to be transitioned to web scraped data, taking into account the quality of web scraped data over a period of time, the correlation between online prices and field-collected prices in each city, and the potential for collection efficiencies and sample improvements. With respect to price index methods, development work continues within the NSO on text mining techniques (to form broader product definitions, especially for clothing).

---

\(^8\) See Australian Bureau of Statistics, 2015
and price index methods (both bilateral and multilateral index methods) that maximize and automate the use of web scraped data.