

Statistical issues and activities in a changing environment: improvements in the commodity price input used for the ECB's analysis of HICP food prices

Andrew Kanutin^{1, 2, 3}

1. Introduction

The ECB regularly assesses commodity prices⁴ as part of its inflation analysis and forecasting framework. Given the ECB's primary mandate of price stability⁵ not only are the outturns examined but also the causes for deviations from the forecasts. In 2007 the ECB started to see sustained errors relating to the forecasting of the HICP food price sub-index. In the calculation of consumer price indices, Eurostat makes a distinction between processed and unprocessed food and provides separate HICP aggregates. The HICP unprocessed food index includes meat, fish, fruit and vegetables; while HICP processed food index covers the rest of the food sub-indices plus alcohol and tobacco. For the purpose of explaining domestic food prices, the commodities should be defined as close as possible to that of the HICP food sub index. Around one third of the forecast error could be attributed to the raw commodity price index used which was not reflecting the costs being borne by European Union (EU) food consumers as it reflected world market prices rather than EU prices which may be affected by the Common Agricultural Policy (CAP) and other regional affects. After examination of alternative sources and weighting schemes it became clear that rather than raw commodity prices it was necessary to examine prices one step further along the production chain i.e. farm gate or wholesale market prices in the EU.

2. Commodity price data used at the ECB

The ECB has used several different data sources and approaches to analyse the effect of commodity prices on inflation. The following two commodity price indices are compiled by the ECB and are still used for analytical purposes but did not prove to be optimal for projections of the HICP food sub-index.

¹ (ECB/DG-S/DIV MAC).

² European Central Bank. Kaiserstrasse 29, 60311 Frankfurt am Main, Germany. E-mail: andrew.kanutin@ecb.europa.eu.

³ The author would like to thank Daniela Schackis and Cyril Papadacci for their comments. All remaining errors are his. The views expressed are those of the author and do not necessarily reflect the view of the European Central Bank or the Eurosystem.

⁴ The term "commodity" means different things to different users. There is not a strictly established definition worldwide and very generally speaking any tangible good can be categorised as a commodity. Usually the commodity should be a comparatively homogeneous product that can typically be bought in bulk. Well-established physical commodities are actively traded at spot and derivative markets. However, the term commodity may not be restricted to raw materials only (oil, cotton, cocoa, or silver) but can also describe a manufactured product used to make other things, for example, microchips used in personal computers.

⁵ "Price stability is defined as a year-on-year increase in the Harmonised Index of Consumer Prices (HICP) for the euro area of below 2%." The Governing Council has also clarified that, in the pursuit of price stability, it aims to maintain inflation rates below, but close to, 2% over the medium term.

2.1 Non-energy commodity price index weighted using euro area imports (NECPI)

The NECPI was used for forecasting HICP food-price inflation when the analytical issues first started to be examined. It is a Laspeyres-type price index which has been produced by the ECB since 2002 as a monthly index with a monthly price collection frequency and steady improvement of the data quality. The NECPI is tailor-made for ECB needs, including the selection of commodities and the frequency of updating the weights. The index includes 18 food commodities⁶ weighted by import value.

In the first half of 2008, the NECPI commodity coverage was updated and new weights referring to 2004–2006 euro area imports were introduced. The old price series using the imports over the period 1999–2001 as weights were linked with the new price series using December 2002 as a linking month. Applying accumulated three-year weights aims at reducing one-off effects on the weights. The source for the weighting data is the European Commission external trade statistics, as available from Eurostat's COMEXT databank.

2.2 Non-energy commodity price index weighted using domestic demand (UWI)

The first approach to reduce the projection errors was to develop a food non-energy commodity price index using a more adapted weighting structure but with the same commodity coverage and commodity price data as the NECPI. The weights of the UWI are based on estimated euro area domestic demand, or “use”, taking into account information on imports, exports and the domestic production of each commodity (ignoring for simplicity – as well as lack of appropriate and comprehensive source data – inventories, which are assumed to be relatively stable over the observed period). In terms of its theoretical properties this use-weighted commodity price index was believed to be more appropriate for the assessment of price pressures stemming from global commodity price changes on the HICP processed food index than the NECPI. Furthermore, its composition is closer to the product coverage of the food components of both the NECPI and the HICP.

While this approach improved the projection accuracy there were still significant errors recorded in the overall HICP projections from this source. The series are still used for several purposes within the ECB.

3. The current approach for food-price inflation projections: European Commission's DG-Agriculture data-set

The European Commission's Directorate General for Agriculture (DG-AGRI) has collected data on wholesale market or / farm-gate⁷ prices for a range of agricultural products over several decades. The data are self-reported by Member States on a weekly basis and are presented as monthly averages. They are not subject to any significant verification by DG-AGRI who simply collate the data, calculate rudimentary EU averages and make them available via their web-site.⁸

⁶ Barley, maize, rice, wheat, soya beans, sunflower seeds, coconut oil, palm oil, sunflower seeds oil, beef, swine meat, cocoa, coffee, sugar, tea, tobacco, bananas, and oranges.

⁷ These are two different concepts. Farm gate prices exclude any transport costs to the wholesale market and any wholesaler margin. Wholesale prices include these two aspects. National practices differ regarding when in the supply chain the data are collected.

⁸ See http://ec.europa.eu/agriculture/markets/prices/monthly_en.xls

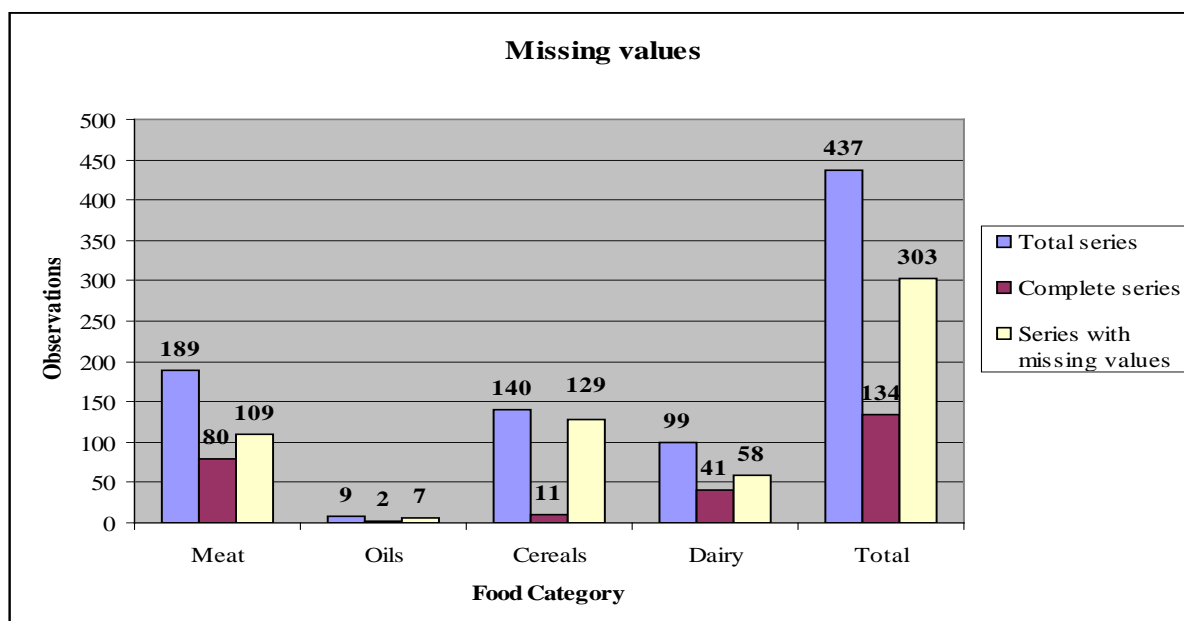
The advantage of the data set is that it reflects true input prices paid by EU food producers which then are passed through to consumers. The data-set contains almost 450 time series at a monthly frequency with a time horizon from 1997 onwards. There is significant heterogeneity in time ranges: some of the newer Member States' price series begin at later dates, but not always in line with their accession date to the European Union and other price series are not available after certain dates due to longer reporting lags. All of the data are split into four food categories (see Table 1).

Table 1
Food categories

Food categories			
Meat	BOV	CALVE	Veal
		BEEFF	Beef
		COWS1	Cows
		HEIFR	Beef Live
	POR	PORK1	Pork 20kg
		PORK2	Pork Regulated
	OVI	LAMB1	Lamb
		LAMB2	Lamb
	CHIC	CHIC1	Chicken 83%
		CHIC2	Chicken PAC
	CHIC0	Every type of chicken	
Cereals	CER	OAT01	Feed oats
		OAT02	Milling oats
		WHEA1	Common wheat
		WHEA2	Breadmaking common wheat
		WHEA3	Durum wheat
		MAIZE	Feed maize
		BARL1	Malting barley
		BARL2	Feed barley
		RYE01	Rye
		RYE02	Rye
Dairy	LAI	MILP1	Skimmed milk powder
		MILP2	Skimmed milk powder
		BUTTR	Butter
		CHEDA	Cheddar
		EDAM1	Edam
	OEV	EGGS1	Eggs L&M 63gr
	EGGS0	All types of eggs	
Oils and Fats	OIL	OILS1	Oil extra-virgin 0.5%
		OILS2	Oil extra-virgin 0.8%
		OILS3	Olive sauce

Some monthly series remain constant for several successive months, which may hide missing values in data sources. Others have true missing values i.e. holes in the time series (See Graph 1). In order to enhance the data-set the ECB examined how best to estimate some of the missing data to complete the data set.

Graph 1
Missing Values in DG-AGRI data set



Missing data are a common problem in large datasets and the problem is broadly discussed in the statistics and data analysis literature. However, no standardised method of estimating missing values in time series exists. The approach taken was to first classify the data gaps per series according to the following rules:

Case A	Time series has at least 30 observations before a missing value (gap) occurs and a gap is no longer than six consecutive missing values.
Case B	Time series does not have sufficient observations (<30) before the gap.
Case C	Time series has sufficient observations (≥ 30) before the gap but the gaps extends to seven or more consecutive missing values.

Only in case A did the ECB made an attempt made to fill the holes in the data. The other two cases have too much missing data and hence any attempt to fill gaps is hard to defend statistically.

The two most basic techniques used to fill gaps in data-sets are *mean substitution* (replacing all missing data in a variable by the mean of that variable) and the *repetition* of the previous value. While either method would be a simple and fast way to complete the dataset, they both have the significant disadvantage that the variability in the data set is artificially decreased in direct proportion to the number of missing data points, with an impact on dispersion and correlation measures.

For these reasons two further methods were investigated: a *correlation method* and the *Box-Jenkins' autoregressive and moving average models*.

3.1 Correlation method

In this method missing values are estimated through a correlation procedure that identifies which country series can be considered as an indicator of another price series with missing values. The basic idea is that for any missing data of a series and country, the most

comparable series of another country is sought on which then the estimation of the missing data is based.

Using FAOSTAT 2008 production data, four geographic macro areas of EU countries were created. Classification criteria for each group were based on the characteristics they share such as similar agricultural outputs (e.g. grapes and wheat in the Southern area and cow milk, whole, fresh and potatoes in Western area) and similar weather conditions (e.g. Mediterranean climate in Southern countries and maritime sub-arctic climate in Northern countries). The resultant groups are shown in Table 2 below.

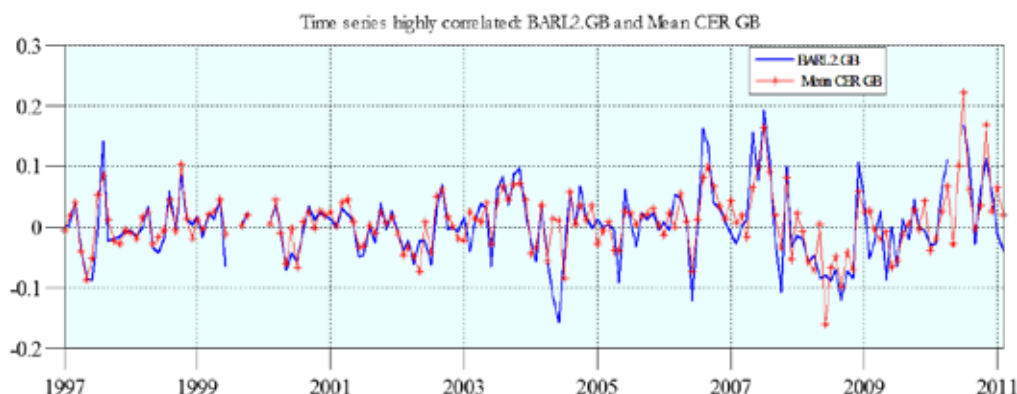
Table 2
Grouping of Member States for correlation

Country		Joined	Southern-Europe	Western-Europe	Northern-Europe	Eastern-Europe
Italy	IT	Founder	X			
Spain	ES	1986	X			
Austria	AT	1995		X		
Belgium	BE	Founder		X		
Bulgaria	BG	2007				X
Cyprus	CY	2004	X			
Czech Rep.	CZ	2004				X
Germany	DE	Founder		X		
Denmark	DK	1973			X	
Estonia	EE	2004			X	
Finland	FI	1995			X	
France	FR	Founder		X		
UK	GB	1973			X	
Greece	GR	1981	X			
Hungary	HU	2004				X
Ireland	IE	1973			X	
Lithuania	LT	2004			X	
Luxembourg	LU	Founder		X		
Latvia	LV	2004			X	
Malta	MT	2004	X			
Netherlands	NL	Founder		X		
Poland	PL	2004				X
Portugal	PT	1986	X			
Romania	RO	2007				X
Sweden	SE	2004			X	
Slovenia	SI	2004	X			
Slovakia	SK	2004				X
EU-27	V1					

For each national data series correlation coefficients were calculated using month-on-month percentage changes in the food price series. These are then compared to the countries in the same geographic area and the same food category. Furthermore, in order to reveal additional patterns and find connections, each series was compared to aggregate series of the average month-on-month percentage changes in prices. For example, the British barley price series was correlated not only with barley series of other Northern countries (cross-correlation) but also with the average of the total British cereals month-on-month percentage changes and with the average of barley month-on-month percentage changes for the Northern Area as a whole.

In the graph below it can be seen that while the barley price series is highly volatile, there is a strong correlation with the mean of all the remaining cereal time series for the UK (rho coefficient = 0.8039).

Graph 2
UK barley price correlations



Once this full set of correlations was undertaken, the most highly correlated price series was used to fill the gaps of those series with missing values by applying to the last available price the month-on-month percentage change of the highly correlated time series multiplied by the correspondent correlation coefficient.⁹ However, in order to ensure robustness of the results, the correlation coefficient must be greater than a threshold of 0.65¹⁰ for the method to be applied and the number of missing monthly observations that need to be filled should be six or fewer.

Applying this approach to the 169 “Case A” series¹¹ it is possible to complete 70 time series.

3.2 Modelling method

For the 99 “Case A” series that had a correlation coefficient lower than 0.65 a further attempt was made to fill the data gaps using modelling methods. Univariate ARIMA models were used to forecast missing values in the price series using the information contained in their own past values and in current and past values of an error term (see Annex A for more details). This model based approach yielded only a further seven series that would have missing data filled. Given the large overhead required to maintain the models and the low cost-benefit ratio it was decided to not pursue this approach further.

⁹ In order to be consistent with the Counter-seasonal estimation procedure to estimate missing observations in the HICP (Commission Regulation (EC) No 330/2009), the first missing value is filled adjusting the last available price by the average month-on-month percentage changes over previous 13 months and from the second month on, missing values are filled with the correlation method procedure.

¹⁰ While it can be argued that this threshold is arbitrary a sensitivity analysis was undertaken to examine the impact that the change has on the number of series that could be filled and 0.65 was the best compromise regarding quality and increasing the number of series that are completed. For example, by changing the threshold from 0.65 to 0.75 the percentage of series filled falls from 41.4 % to 24.9%.

¹¹ Case A price series are those series that have at least 30 observations before the missing value and that have no more than six consecutive gaps to be filled.

4. Weighting scheme

DG-AGRI publishes EU aggregates of the specific data collected. However, mixed methods are used to calculate these averages (for example sometimes simple averages while at other times weighted averages are used depending on the underlying data availability). Given that the intention of the ECB is to use a methodology that is closer to the HICP, a different approach is used to calculate aggregates. In order to compile the euro area aggregate the underlying countries are weighted according to their relative size of private consumption in the euro area, while for the non-euro area and EU aggregates the relative size of private consumption in the EU is utilised.

At the same time the ECB additionally calculates for each Member State, for the euro area (in both the current composition and moving composition forms) and the EU, special aggregates for four categories: Meat, Cereal, Dairy (cheese and eggs) and Oils & Fats.

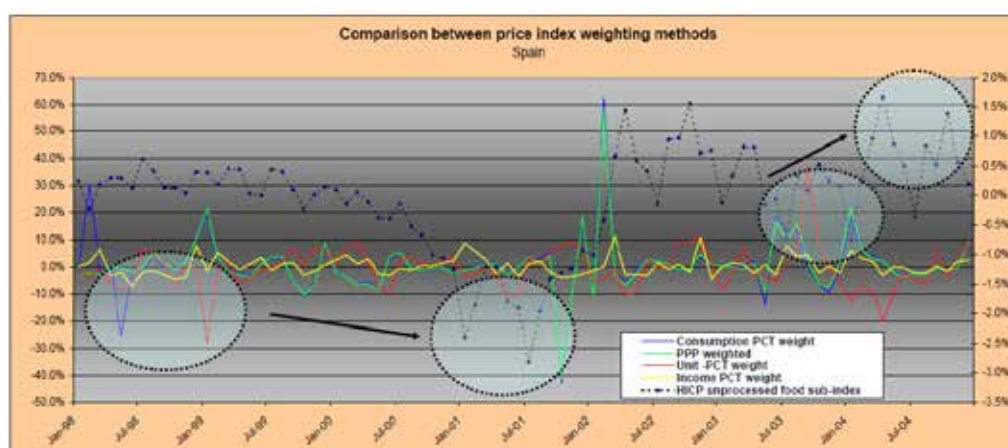
In order to calculate these four special aggregates, elementary series for each country and for a particular product group are calculated – these are un-weighted geometric averages of the price data for the cases where we have no weights (e.g. edam and cheddar cheese are combined). Confidential Purchasing Power Parities (PPP) consumption weights supplied to the ECB by Eurostat are then used to aggregate the series to get to country series for each of the four product groups. For simplicity, monthly rates of change with 2009 PPP weights are currently used to build up the aggregates. The euro area series are then created from these country aggregates and weighted using the HICP weights and index formula. The moving coverage series follow the same moving coverage country coverage as in the HICP, i.e. each reference month reflects the composition of the euro area at that moment, while the current composition series reflects the current euro area 17 members throughout the series. A similar approach is taken when computing the EU series.

As a quality check three alternative weighting methods were compared against the PPP weighted European series described above. These alternative weighting schemes are based on different items from the Economic Accounts for Agriculture (EAA):¹²

1. Units of production- these are obtained by dividing current values of producer prices by the corresponding physical quantities. Their value is given in 1000 tonnes (quantities), Euro per tonnes (unit value) or national currency per tonnes (unit values). Unit values are converted to euro using market exchange rates where appropriate.
2. Food consumption- i.e. gross human apparent consumption of main food items expressed in values (1000 tonnes).
3. Income – the price receivable by the producers from the purchaser for a unit of a goods or services produced as output plus any subsidy receivable on that unit as a consequence of its production or sale minus any tax payable on that unit as a consequence of its production or sale. The producer price is expressed in millions of euro (from 1.1.1999)/millions of ECU (up to 31.12.1998). Income values are converted to euro using market exchange rates where appropriate

¹² Eurostat provides data on national agricultural accounts.

Graph 3



Graph 3 compares the different food indexes with the HICP unprocessed food sub-index for Spain. It shows that for food the currently used PPP consumption index is most similar to the food index weighted using gross human apparent consumption.

In order to assess further the best weighting method, the predictive performance of changes in the food price index weighted with the current and the alternative methods was compared with the HICP unprocessed food sub-index. The indicator used to evaluate the accuracy of the food index against the benchmark is the Mean Absolute Percentage Error (MAPE). It computes the absolute percentage difference between the Food index and the HICP unprocessed food sub-index. Empirical analysis confirms that a time-lag exists between movements in the food commodity prices and the HICP index. For this reason, a series of lagged MAPEs are computed to properly evaluate the performance of food indexes which the PPP approach gives the lowest MAPE values and therefore it is confirmed as the most appropriate weighting for the use of forecasting changes in the HICP unprocessed food sub-index.

5. Enhancements related to quality control

The farm gate and wholesale price data-set as delivered by DG-AGRI is extremely volatile. Extreme values can be defined as observations numerically distant from the rest of the time series; they can be the result of either an error in recording or of the heavy-tailed distribution of the sample mainly due to extreme food prices. Raw material prices may be very volatile due to the fragility of both supply and demand factors. The challenge is to identify and remove only erroneous outlying observations that can bias estimates but to leave in the dataset all outliers that correctly show the extreme price movements. A very pragmatic approach was taken in the end for this data-set and outliers have not been removed with the sole exception of zero values, as it is assumed that no farm provides goods for free. Not removing further outliers in the absence of any additional information is believed to support best the main use of the data-set as input into the forecasting process.

6. Results

The farm gate and wholesale price dataset is only relatively recently being used in the ECB analysis and therefore it is difficult to give it a full and categorical endorsement. Nonetheless,

early indications show that it is performing significantly better than the previous datasets. One possible reason is that the previous data used international prices while the DG-AGRI data reflect the prices actually paid by EU food producers. This is particularly important as the EU prices will directly include the effect of the CAP. In times of buoyant commodity prices the effect will be minimal but, if world prices are below CAP thresholds then the effect is likely to be significant.

7. Future

While, as discussed in the previous section, the change to using the farm gate and wholesale price data set has seen improvements in the analysis and projection of commodity price feed through within the ECB further improvements could be envisaged. These include:

- Addition of additional commodities – one area which is not published in the current DG-AGRI dataset is the prices of fresh fruit and vegetables. These data are collected by DG-AGRI but up to now are believed to be of such a low quality and as having extreme volatility so as not to be of publishable quality. This is an area where further investigation could be undertaken in order to either use the available, but poor quality, DG-AGRI data or, alternatively, to see if an alternative data source could be found.
- Inclusion of non-indigenous commodities – Commodities which are not grown in the EU in sizeable quantities such as coffee or cocoa are not included in the DG-AGRI data as there is no corresponding farm-gate price. However, these commodities have a relatively high share in the Food sub-index in the HICP. One potential approach for these commodities would be to investigate ways to include these items into the current dataset.
- Publication – up to now the dataset is available only to ESCB internal users. However, the data are likely to be of wider interest and it could be considered if the EU aggregates could be released as “ECB experimental” data.
- While significant data cleaning and quality control is already undertaken further improvements are likely to be desirable.

8. Conclusions

This paper has presented the commodity price data that is being used at the ECB for price pass through analysis and projections. Historically, international price primary commodity indices were used for the task but since 2007, when these data started to give rise to sustained projection errors, alternative approaches have been explored. These have culminated in the current dataset which is based on raw information made available by the European Commission’s Directorate General for Agriculture. As a result the projection errors and analysis of pass through effects have been improved. However, there are more improvements that can be explored.

Annex A: Model-based procedure: detailed methodology

1. **Data preparation:** Time series were differenced just enough to become stationary and no patterns such as a trend or seasonality are left.
2. **Model identification:** Sample autocorrelations are compared with the theoretical ones for different orders of AR, MA and ARMA models (see box Chatfield, 2004).¹³

Summary on Auto regression and Moving Average models (ARMA)

To model time series dependence, we use univariate ARMA models assuming that the level of current observations depends on the level of lagged observations.

An autoregressive model (AR) is simply a linear regression of the current value of the series against one or more prior values of the series. An AR model of order p , denoted as AR(p), can be expressed as:

$$y_t = \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + u_t$$

where u_t is a white noise disturbance term.

Another common approach for modeling univariate time series models is the moving average (MA) model where it is assumed that the observations of a random variable at time t are not only affected by the shock at time t , but also by previous shocks. Under the assumptions that u_t ($t = 1, 2, 3, \dots$) behave as a white noise with $E(u_t) = 0$ and $\text{Var}(u_t) = \sigma_2$.

$$\text{Then } y_t = \mu + u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2} + \dots + \theta_q u_{t-q}$$

is a q th order moving average model, denoted MA(q).

The autoregressive model AR includes lagged terms on the time series itself, and that the moving average model MA includes lagged terms on the noise or residuals. If we combine both types of lagged terms, we obtain the autoregressive-moving-average, or ARMA models. The order of the ARMA model is included in parentheses as ARMA(p,q), where p is the autoregressive order and q the moving-average order.

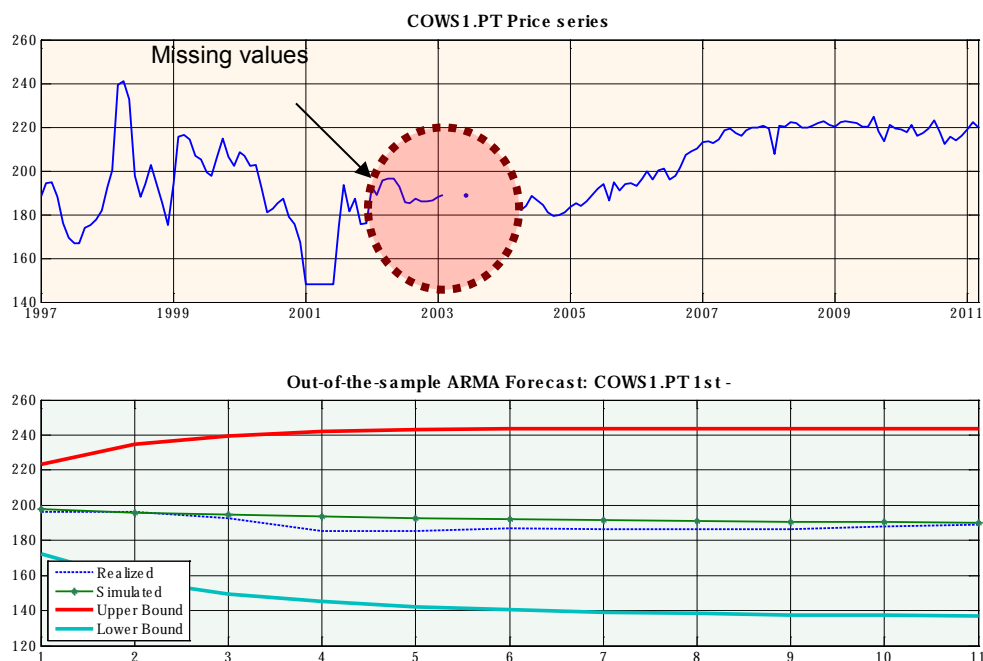
3. **Parameter estimation:** Polynomial AR, MA and ARMA models are estimated for time-series data.
4. **Model checking:** We determine if the model estimated in the previous step is adequate for the time series. If this is found to be inadequate, we need to go back to Step 2 and try to identify a better model. *Residual diagnostics* technique is used to check the model adequacy. This method implies checking residuals for

¹³ "The analysis of time series: an introduction", Chatfield 2004

independency, homoscedasticity¹⁴ and normal distribution and if these properties are present, it means that the model originally specified is adequate to capture the features of the data.

5. **Forecasting:** After that we have selected, estimated and checked the model, the last step is to forecast missing values of a series given its previous values and/or previous values of an error term. Mean absolute percentage error (MAPE)¹⁵ is used as a measure of accuracy in a fitted time series value. For differentiated time series, it could happen that we have zero values so in these cases instead of MAPE we measure the forecast accuracy with Symmetric Mean Absolute Percentage Error (SMAPE).¹⁶ Firstly, we proceed with in-sample forecasts that are those generated for the same set of data that was used to estimate the model's parameters, then in the second part of the procedure, if the model shows good fitted values (i.e. R square greater than 0.70), we proceed with the out-of-sample forecast to fill the missing values in the series.

The graph below shows an example of forecast for Portuguese cow prices. We can notice that the proposed method ARMA(1,1) shows good forecasting accuracy, with MAPE of 0.021.



¹⁴ A vector of random variables is homoscedastic if all random variables in the sequence or vector have the same finite variance.

¹⁵ MAPE formula:
$$M = \frac{1}{n} \sum_{t=1}^n \left| \frac{A_t - F_t}{A_t} \right|$$

¹⁶ SMAPE formula:
$$SMAPE = \frac{1}{n} \sum_{t=1}^n \frac{|A_t - F_t|}{A_t + F_t}$$