**Imputation of Producer Prices**

**Methodological improvements**

**Michele Vollaro**

**ESS – ECO Team**

The aim of this note is to propose a set of techniques to be used for imputing missing data in the FAO producer prices domain. It builds on economic theories and empirical evidences developed for testing the Law of One Price (LOP), the degree of integration of spatially separated markets and the transmission of price shocks across markets.

The proposed methodology for the imputation of missing data follows two implementation strategies: the first is the use of a FGLS panel technique for large series of missing prices, conditional on the choice of cointegrated reference series (the countries forming the panel); the second is the use of standard time-series techniques for a maximum of three consecutive missing data.

The note proceeds with an introduction about the Producer Price domain of the FAO and the presentation of the theoretical background about the LOP and the concept of market integration applied to the agricultural prices. The next sections present the proposed imputation techniques and the schedule for the next steps.

*The FAO domain of Producer Prices*

The process underlying the publication of data on Producer Prices is based on a sequence of steps characterized by the presence of specific risks, affecting both the quality and the quantity of the price data from the collection to the release stage. This calls for a constant and thorough check of each step, in order to avoid the transmission of qualitative and/or quantitative errors along the process.

The first step is the collection of producer prices through a questionnaire at country level. Prices are referred to primary crops, live animals and livestock primary products, and those received by farmers *at the point of initial sale* (prices paid *at the farm-gate*). Due to differences in data collection infrastructure and capacity, some countries do in fact collect *wholesale* or *local market* prices. While these may be acceptable proxies for farm-gate prices where marketing chains are short, they tend to be poor proxies in more complex production chains, where transaction costs and processors’ concentration largely affect the prices through transport fees and margins, which constitute a significant share of final prices. Some countries report even retail prices, which may be quite unrelated to producer prices.

The weakness of the producer prices domain often stems from its incomplete geographic and item coverage, particularly on livestock products. Moreover, produce price series reported in FAOSTAT were broken in 1991, so they are all relatively short. The reasons of such breaks are diverse and referable to two main aspects: the first was the absence of a systematic and mechanized data input system at FAO level with the consequence of losing important information about the collected series, such as the decimal places of the data, while the second was the wide wander behaviour of producer prices during the ’80 and ’90 characterized by hyperinflation, frequent exchange rates and changes in monetary policy regimes in several countries, leading to the loss of decimal points after the FAOSTAT standardization procedure to the new currencies. Another low-quality factor is the uncertainty of the price concept adopted by the countries and the point of transaction used to record prices.

Another factor affecting the overall quality of producer prices is a frequent lack of cross-country comparability over time. It occurs in particular when countries supplying agency changes or when countries respond to the annual questionnaire after an interval and time series display a price change that is not consistent with other price indicators or when item coverage differs from the previous one. The first action undertaken by the FAO team working on Producer Prices is to request further information to the country. When possible, missing terms of the series are interpolated. In extreme cases, new series are rejected, or old series are masked. In other cases, a country note informs users of the break in the series.

Currently, the dataset covers approximately 180 out of 185 questionnaire recipient countries, 30 of which replied once or twice only. It follows that official data represents less than 60 percent of the total price records.

Currently, missing data are imputed by applying simple techniques among the ones proposed by the FAO Statistical Standard on Imputation suggested by the FAO Office of the Chief Statistician. The suggested imputations techniques belong to both parametric and non-parametric techniques. Among the parametric ones there are the *regression imputation*, the *ratio* and *mean imputation*, while the non-parametric techniques include the *hot-deck* and *cold-deck* imputation and the *predictive mean matching*. Special cases of techniques belong to time-series imputation methods, including the *carry forward* and *carry backward imputations* derived from the ratio and *mean imputation* techniques. More specific imputation techniques, based on econometric models and advanced statistical inferential methods, are also suggested and allowed to be employed on a case-by-case basis, subject to an assessment and a validation process[[1]](#footnote-1).

However, so far, **imputed data are not published** because the employed imputation methodology, based on simple regression methods (trend), is not suitable for the large amount of consecutive missing data, resulting in low accuracy of imputed prices.

A contribution to the quality and quantity improvements in the overall production prices domain might be pursued, on one side, by ameliorating the early stages of price data collection; and on the other side, by adopting updated and reliable imputation techniques, based on inferential statistical methods and econometric models, defined specifically for the purpose of imputing price data.

*Agricultural prices: theoretical background*

The formation of agricultural production prices depends on endogenous elements, pertaining to farmers’ production choices, and exogenous factors, that characterize the social, economic and natural environment in which production takes place. It follows that changes in both endogenous and exogenous conditions lead production prices to change over time. Endogenous adjustments in agriculture are usually slow, as they stem from structural and technological improvement of the farms as well as from the adaptation to external conditions. Exogenous conditions, instead, might be less predictable and faster, hence inducing abnormal volatility and behaving as shocks. Shocks affect prices and thus markets equilibria in the short-run, but may also bear long-run consequences, as they may produce further adjustments back in endogenous factors, causing structural breaks in the price series. In each case, the determination of prices provides for a signal of *temporary* price efficiency.

Price efficiency is defined as *the* *set of* *prices leading to optimal (efficient) allocation of resources*, where the *optimum* corresponds to the *allocation of resources that maximizes individual’s utilities of output, conditional on the available, in terms of time and form, stock of resources* (Fackler and Goodwin, 2001). Price efficiency is the *conditio sine qua non* for the realization of *market* *efficiency*. This is defined as *the capacity of markets to minimize costs when they match supply and demand* (Listorti and Esposti, 2012). These costs are called transaction costs and refer to the set of costs and outlays required for transferring products between two different places. Transaction costs include transportation costs, which often constitute the largest share. *Spatial arbitrage* is the condition verified when the difference between prices for homogenous goods in two distinct locations is smaller or equal to transaction costs. Under this condition, price movements are horizontally coordinated (co-movements), and the markets are efficient. When price differentials are higher than transaction costs, the opportunity arises for obtaining arbitrage profits, that it, to obtain profits by moving homogeneous goods from one location to another. Under these conditions, the horizontal coordination of prices decreases and markets are considered inefficient (Fackler and Goodwin, 2001).

The horizontal price co-movement is a consequence of the law of one price (LOP) (Listorti and Esposti, 2012): *in markets linked by trade and arbitrage, homogeneous goods will have a unique price, when expressed in the same currency, net of transaction costs* (Fackler and Goodwin, 2001). Divergences from the theoretical conditions of the LOP have given rise to a vast academic literature, especially in terms of testing quantitatively the underlying conditions and contributions of different elements to affect the transmission of price shocks across markets linked through trade.

Another condition which is associated to the co-movements of prices is *market integration*. This is meant to be a measure of the degree of connections between markets, involving also *price transmission*. Indeed, according to Listorti and Esposti (2012), market integration refers *rather to the tradability of products between spatially distinct markets, irrespective of the presence or absence of spatial market equilibrium and efficiency*. This definition is backed by Fackler and Goodwin (2001). These authors state that *integration* does not necessarily correspond to correlation of prices, as price co-movements might depend upon other factors, independent from trade, such as, for example, other mechanisms involving information flows without trading flows (Listorti and Esposti, 2012). Indeed, Barrett et al. (2002) clearly state that *market integration* might be a synonymous of *market tradability* or *market contestability* (not necessarily actual trade) and it might manifest either through physical trade of goods, through price transmission or both[[2]](#footnote-2).

For the purpose of imputing producer prices, the concepts of *market integration* and *price transmission* are extremely important, because they provide for the conceptual background upon which it is possible to develop an imputation strategy based upon the over-time tracking of price co-movements between countries characterized by (degrees of) markets integration for homogenous goods.

From a wider perspective, producer prices might be affected also by signals and shocks coming from supply-chains to which they belong. This condition is known as vertical price transmission, and it is affected by factors that are different from those affecting spatial transmission.

Spatial price transmission – which is also termed “horizontal” in this context – can be symmetric, if prices movements are bilaterally transmitted between markets; or asymmetric, if transmission occurs only (or mostly) in one direction. This may happen, for instance, when two areas belong to the same trading network, but do not trade directly. In this case price shocks are transmitted through the trading linkages (Fackler and Goodwin, 2001). The same happens in vertical transmission, but in such a case, asymmetries are shaped by differences in the market power and non-competitive behaviours along the supply-chain.

There are several reasons why markets may be more or less integrated, and hence price shocks may be fully or partially transmitted. Conforti (2004) presented a systematic grouping of potential external factors that might affect both spatial and vertical price transmission: transport and transaction costs; market power; increasing return to scale in production; product homogeneity and differentiation; exchange rates and border and domestic policies. Beyond these, other factors like physical distance between markets, difference in market concentration of different locations, differences in the quality of market infrastructures, including transportation and communications, difference in physical infrastructures, including roads, harbours, ports, might affect price transmission as well.

In general, the efficiency of markets and price formation are affected by policies, productivity levels, climate anomalies and, in general, by differences in the local institutional and economic context pertinent to the market in question.

In this context, the framework presented so far serves the purpose of setting up an analytical framework for improving imputation techniques for the FAOSTAT producer prices domain. Qualifying and quantifying price transmissions sheds light on the behaviour of price shocks across markets, thus allowing to predict the transmission of shocks. Consolidated econometric techniques grounded on an economic theoretical background allow imputing missing information and providing solid evidence that can substitute for missing official information collected by countries.

*Imputation methodologies*

The proven capability of inferential and econometric techniques, developed under the analytical framework of market integration, in quantifying the transmission of prices across markets might be suitable for imputing missing price data, conditional upon the verifiable presupposition that the market with missing data is cointegrated with other markets chosen as reference. In terms of the producer prices domain of FAOSTAT, the assumption would be that the producer price of *item* *i* in country *j* responds to shocks the same way as the price of the same item in one or more other countries. That price may be a country group, sub-regional or regional price and it needs to be derived from cointegrated series. Two price series are cointegrated whenever the series of the estimation residual is stationary, implying the existence of a stable long-run relationship between the prices in two spatially separated markets.

More in details, according to Fackler and Goodwin (2001), cointegration tests evaluate the following simplified regression: , representing the equilibrium parity condition implied by the spatial arbitrage assumption, in which *p1t* and *p2t* are the prices for the same commodity in spatially separated markets over time. If the series of both prices are non-stationary, the estimated standard errors of the parameters are not consistent, but the estimated coefficients do. However, if the residual *ut* is stationary, then the price series are cointegrated, meaning that they are linked by a stable long-run relationship despite their short-run wandering behaviour. This test enables to select the countries for which the price series for specific commodities are cointegrated.

Accordingly, a first step in the proposed methodology is the application of the cointegration analysis in order to identify the group of reference countries for imputing the missing price series in other countries. A further step would be to test the observed price series for the existence of Granger lead/lag relationship among the selected countries, such that the direction and the speed of the shock transmission can be determined.

Once the country group has been selected, missing prices are imputed by hinging upon the aforesaid exogenous factors, given their empirically proven ability to affect the degree of price transmission. Within such conceptual framework, the imputation of prices of item *i* for country *j* can be operated as an out-of-sample prediction of a regression model, based on a *static spatial equilibrium* hypothesis. This implies that changes in the deviation from the regional price over time are explained by the variability of exogenous shifters.

Set as the missing price data of item *i* in country *j*; set as the aggregated mean price data for item *i*; set as the deviation of from , then . Using the observed price data of item *i* in the selected region, the *static spatial equilibrium* hypothesis will be tested by employing the following panel model: , where is the vector of both country-and group-level exogenous determinants of producer prices. By substituting into , the imputed price is obtained as .

A flexible analytical method is required to test the *static spatial equilibrium* hypothesis. This simple but very general hypothesis is tested by regressing the exogenous factors against the aggregated mean price and by controlling for the presence of correlation among the regressors (exogenous factors), in terms of both cross-sectional correlation (across countries) and country-specific serial correlation (along the years), and the presence of heteroskedasticity. To this end, the proposed econometric technique to employ is a panel model, applied through the generalized least squares (GSL) estimator, which allows, beyond the mentioned corrections, the inclusion of an autoregressive component of order one in the error term in order to account for potential biases due to country-level omitted variables.

The general form of the estimation model is based on pooled panel specification:, in which the conditional variance of the error term is the unknown non-singular matrix , estimated through the feasible GLS estimator. The model specification has the objective of estimating the effects of the exogenous shifters on the item price *i*, through the most efficient estimator of the panel models, conditional on having a balanced panel composed by the number of years greater than the number of countries, , such to ensure the correction of biases caused by the presence of serial- and cross- correlations.

This imputation approach, therefore, would be indicated to be employed in those cases in which data are missing over a long period or at least for a period greater than three years, for which other imputation methodologies based upon time-series analysis might be effective.

The external shifter might be identified among macroeconomic indicators conveying information on levels of development, and/or among sectoral information on agriculture, or wide-ranging phenomena such as climate indicators. Based also on Conforti (2004), examples of exogenous shifters include: per capita GDP, household shares of food consumption in income (Engels’ ratios), agricultural value added, agricultural capital stock, agricultural total factor productivity (TFP), number of farms, market concentration of processors; indicators of public investments, such as infrastructures for ICT and for transportation, research and extension services in the agricultural sector, import/export balance, exchange rates, border and domestic policy measures and more.

Where price series are characterized by few missing data points (up to three consecutive years), the employment of time-series analytical approaches would be more straightforward and effective, given the opportunity to take advantage of the short-run “path-dependency” properties of prices. Therefore, simpler time series models based upon autoregressive specification can be employed. Among these, the choice might be VAR, ARIMA or more complex ones, based on the Holt-Winters method, that manages seasonality feature of monthly data.

The basic model for forecasting (imputing) a price at time *t+1*, namely *πt+1*, is to make it dependent upon its lagged values, namely , and lagged values of other exogenous factors, namely , proven that they affect *π*. Based on Woolridge (2010), such a model can simply be specified as .

The general formulation of estimation model is , with in which *It-1* is a vector containing information of *π* and *x* for *t-1* and previous periods. This model is generally known as a Vector AutoRegressive (VAR) model. Once the parameters of the VAR have been estimated, the forecasted (imputed) price at *t+1*is computed by plugging the observed values of *π* and *x* at time *t*. The consistency of the estimated parameters is strictly dependent upon the correct specification of the lags to include in the model. This can be pursued by testing the model against the information criteria for different specifications, that is for different number of lags. Whenever the price series is tested to be non-stationary by the means of *unit root* tests, the proposed model is applied to the *difference* version of the variables. The degree of *difference* will define the order of *integration* *d* of the variables:

A more complex but flexible time-series model is the AutoRegressive-Integrated-MovingAverage (ARIMA) model, in which the autoregressive – AR(*p*) integrated I(*d*) - component is completed by the moving average – MA(*q*) - process of the price variable. By referring to Vander Donckt and Chan (2019), the practical application of the ARIMA models for imputation purposes can be extended to include exogenous factors and defined as ARIMAX, where *x* represents the set of exogenous factors. The general formulation of an ARIMAX(*p*,*d*,*q*) can be expressed in the following formulation:

, in which *π* is the producer price, *α* and *β* are the parameters to be estimated, *d* is the other of integration, *p* is the order of the AR process and *q* is the order of the invertible MA process and the AR and MA processes do not share common factors.

For all the other cases in which the price data are missing only in the last (current) year, beyond the described ARIMAX procedure, a *one-step-ahead forecast* approach might be followed. The basic concept for such procedure is to choose the forecasted (imputed) value for which the *loss* associated to the forecast error is minimized (Woolridge, 2002). One method belonging to the aforesaid approach is the *exponential smoothing*¸ specified according to the following equation:

, which, for and , can be reduced to the simplest equation . is the martingale of *π*, *α* is the smoothing parameter to be chosen, defined to be and *ft* is the forecasted value of *πt+1* for . It implies that the forecast of *πt+1* of time *t*, namely *ft*, is a weighted average of and its forecasted value *ft-1*.

A practical *exponential smoothing* method for imputing missing prices, including seasonality and trend features (for monthly price data), is the Holt-Winters (HW) method (Cipra et al., 1995). Beyond the smoothing parameter *α*, the HW method includes two more parameters accounting for trend and seasonality, specified according to the following general formulation proposed in Tularam and Saeed (2016):

, with ;

, with ;

, with .

Given the choice of three smoothing parameters, the HW method is also called *exponential triple smoothing* method, in which is the smoothing parameter, and are the estimated trend parameters, is the smoothing parameter for the trend estimates, is the estimated seasonality parameter, is the smoothing parameter for the seasonality estimates and *δ* is the number of periods in the seasonal cycle.

*Steps ahead*

The first step for the implementation of this methodology is a preliminary *peer review* of the proposal, with expected inputs for the steps to be taken for integrating it in the Statistical Working System (SWS). This is expected to happen **by the end of July 2020**.

In the meanwhile, preliminary estimation procedures will be run to test the cointegration, the FGLS and the time-series methods. During this step, different aspects of the imputation methodology will be scrutinized: i) the choice of cointegrated countries; ii) the selection of the exogenous shifters; and iii) the effectiveness of the estimation and of the forecast procedures.

The next step is to integrate the imputation stages in a unique routine, led by a mechanized procedure able, for the FGLS, to operate the choice of cointegrated countries (based on completed series); to run the estimation models by selecting iteratively the exogenous factors that maximize the model’s performance; to compute the forecasts based on the estimated parameters; and, for the time-series, to run the estimation models (VAR, ARIMAX and HW for monthly series) by selecting iteratively the exogenous factors that maximize the model’s performance and to compute the forecasts based on the estimated parameters.

A further option will that of employing first the time-series model in order to improve the availability of completed series upon which to operate the cointegration-FGLS imputation procedure.

This is expected to happen **by the end of September 2020**.

When the routines will be finetuned, they will be transferred to the SWS according to the best integration method to choose among a stand-alone package, a plug-in or an embedded procedure. The expectation is, however, to have inputs about the best integration method already during the peer review process. This is expected to happen **in October 2020**.

1. For a complete presentation of the imputation technique, see <https://home.fao.org/fileadmin/user_upload/scp/Standards_for_quality_compliance/,DanaInfo=intranet.fao.org+SSS_Imputation__Endorsed_Nov_2019_.pdf> [↑](#footnote-ref-1)
2. Barrett et al. (2002) specify that for contestable markets, two markets can be integrated *absent trade* if arbitrageurs face zero marginal returns, leaving them indifferent about trading. [↑](#footnote-ref-2)