**Annual producer prices missing value imputation procedures**

The Autoregressive Integrated Moving Average Exogenous Variable (ARIMAX) based approach is used to estimate the missing observations on the annual producer prices.

Among different crops and livestock, we consider countries for which a limited number of data points is missing (countries with observations greater or equal to 20) to implement time series estimation. In this regard, we first fit a group of ARIMA models with additional exogenous variables to the producer price data. Thereafter, based on Bayesian Information Criteria (BIC), we select the best in-sample specification which is subsequently used to estimate the missing values.

The ARIMAX models extend ARIMA models through the inclusion of exogenous variables that have linear effect on the stationary response series . The general form of the ARIMAX model estimated separately for each country is:

Where AR polynomial of : , and the MA polynomial of order : is the logarithmic producer price at period t, is to the order of integration. The parameter represents the constant term, and is a vector of observed exogenous regressors with coefficients .

To estimate the model in (1), we use the STATA built-in *arima* function. After having written the model in state-space form, maximum likelihood parameter estimates are obtained based on the Kalman filter. See Becketti (2013) for further detailed explanation on STATA’s *arima* command.

The steps of the implementation of the ARIMAX approach are summarized as follows:

1. As for each crop or livestock, input country-specific data series, take the natural log transformation of all series to reduce variability in skewed data, then check that all logged series are stationary and transform them appropriately (first differences, etc.) if necessary.
2. For each country under crop , the endogenous variable is the logarithmic producer price, which contains missing values. Within a loop, fit a group of models and single out the one with the best in-sample fit based on minimization of the BIC criterion. The list of candidate models is defined along two dimensions:
3. The autoregressive, integration orders, and moving average of the ARIMA model, with . Due to the shortage of observations (we only have 28 years of time coverage for all the series), we divide the specifications into three to avoid “*flat log likelihood encountered, cannot find uphill direction*” error during the best model selection process in STATA. This means the best model selection process will be carried out three times for each country under the following specifications:
4. The following set of exogenous regressors (before transformation): GDP, GDP per capita, gross fixed capital formation of total economy, value added of agriculture, forestry and fishing[[1]](#footnote-1), gross fixed capital formation of agriculture, forestry and fishing[[2]](#footnote-2), agriculture trade openness index, agriculture exports-to-value added ratio[[3]](#footnote-3), and yields for each crop[[4]](#footnote-4). When one of the above country-specific time series data is incomplete, we exclude this regressor from the exogenous regressors group in the *arima* model selection process.
5. After selecting the country-specific preferred model, construct one-step ahead forecasts to estimate missing data points on the logarithmic producer price.
6. Store results, minimum BIC value and best model specification (with parameter estimates).
7. Find the optimal model from the three best models under the above different specifications and store the final estimated results.
8. Compare the minimum BIC value, store the lowest BIC value and corresponding model specifications.
9. Calculate the Root Mean Square Error (RMSE) for estimated producer price (reverse the estimated the logarithmic producer price) for the three best model and store the lowest RMSE and corresponding model specifications.
10. Compare the model specifications from the lowest BIC value and the lowest RMSE, if they are the same, we store it as the final model specification, if not, we graph the forecasted producer prices from both model specifications as well as the observed value to allow for graphical inspection so that final model specifications can be chosen.

# Bibliography

Beketti, S. (2013). *Introduction to Time Series Using Stata .* College Station, TX: Stat Press.

1. The first four time series can be obtained directly from the Macro Indicators domain on FAOSTAT. [↑](#footnote-ref-1)
2. This time series can be obtained directly from Capital Stock domain on FAOSTAT. [↑](#footnote-ref-2)
3. The main sources for trade related time series are the FAO Trade Database and the UNCTAD Merchandise trade matrix. [↑](#footnote-ref-3)
4. Yields data can be obtained directly from Crops, Crops Processed and Livestock Primary and Livestock Processed domains on FAOSTAT. [↑](#footnote-ref-4)