Error Assessment of Imputation for Agricultural Production Data in FAOSTAT

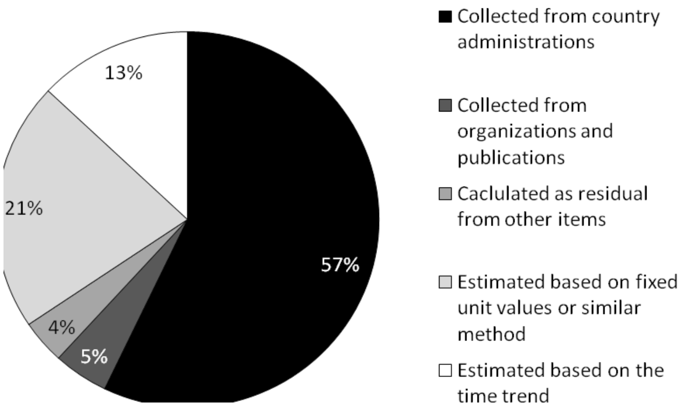
Hoffmeister, Onno
Food and Agriculture Organization of the United Nations (FAO)
Economics and Social Statistics Division (ESS)
Viale delle Terme di Caracalla, I-00153 Rome (Italy)
E-mail: onno.hoffmeister@fao.org

Khaira, Hansdeep
Food and Agriculture Organization of the United Nations (FAO)
Economics and Social Statistics Division (ESS)
Viale delle Terme di Caracalla, I-00153 Rome (Italy)
E-mail: hansdeep.khaira@fao.org

This study assesses the applicability and accuracy of new imputation methods which are currently being developed for FAOSTAT, the world's largest database on agriculture, nutrition, fisheries, forestry, food aid and land use. The time series of annual data on agricultural production, collected by the Production and Trade Team of ESS for FAOSTAT and processed further in the Food Balance Sheets, provide the main input for the estimation of food availability from which the FAO derives its estimation of undernourishment rates all over the world.

FAO undertakes best efforts to collect a large amount of the source data on agricultural production from questionnaires filled by country administrations. These are supplemented by information from official publications or from international organizations, especially in cases of non-response. As not all data can be obtained from the aforementioned sources at the required level of quality (see figure 1), statistical imputation of missing data constitutes an essential component in the data production process. The error assessment presented in this study is part of the FAO's efforts to enhance the methods applied for that imputation.

Figure 1: Sources of Agricultural Production Data for FAOSTAT (1990-2009)[1]



Note: [1]) Non-missing data points on primary crops production, area harvested for primary crops and meat production, disseminated in FAOSTAT. Source: FAO 2011.

Below, three different imputation methods (linear interpolation, trend smoothing, and benchmarking of growth rates on aggregates) are evaluated with regard to their applicability and accuracy. For this, they have been applied on a random sample of data points for which official figures are available, so the imputed values can be compared with their counterpart values obtained from official sources. In its present stage, the study focuses on the disseminated FAOSTAT data on production of primary crops, area harvested for primary crops and on meat production.

By comparing the accuracy of imputations in pooled time series, this paper has a similar objective as the studies by Hu and Salvucci (1998) and Cooper (2010) which evaluate different imputation methods applied to cross-section survey data, as well as by Lawrence *et al.* (1985), Marcellino (2005), Chen (2007) and Castillo and Useche (2010) which focus on the accuracy of imputations in time series. It touches on the bulk of papers on error assessment in time-series forecasting which followed the pioneering works of Reid (1969, 1975), Newbold and Granger (1974) and Makridakis and Hibon (1979). See Armstrong (2006) and De Gooijer and Hyndman (2006) for comprehensive overviews.

The next section (Section 1) describes the imputation methods applied. Section 2 describes the baseline data and the sampling; Section 3 outlines the methodology of the error assessment; Section 4 presents the results; and Section 5 concludes.

1. Imputation Methods

The imputations currently under development by the Production and Trade Team of ESS within FAO are aimed to represent as closely as possible the unobserved true value of a particular country in a given year. Thereby, the emphasis is on one of the various objectives often followed with imputation of missing data, notably on the objective of assigning values at the microlevel with maximum accuracy and thereby “allowing analyses to be conducted as if the data set were complete” (Kalton and Kasprzyk 1982, p. 22). Other objectives, such as preserving the characteristics of the distribution throughout the dataset or to derive reliable estimators for a target population based on a sample, are attributed lower attention in the scope of the present study. For that reason, it has been decided to focus primarily on *deterministic* imputations, based for example on regressions or calculation of means, without any random procedure involved, as these have been shown to produce in general more exact estimators for the single imputed data point than *probabilistic* imputations. *Probabilistic* imputations, in contrary, are commonly more efficient in preserving the distributional properties of a variable throughout the dataset as a whole (*ibd.*, pp. 24ff.).

In this early stage of research, the choice of imputation methods has been strongly guided by concerns about practicability and apprehensibility. Due to constraints on time and resources, the chosen approach needs to be easy to be implemented, interpreted and verified; computational efforts are intended to be low. Based on the considerations above, out of the variety of imputation approaches[1] we analyze the following three in the scope of this study:

a) Linear interpolation:

A linear trend is assumed to exist between the start- and endpoints of gaps in the time series. Let y_0, y_1, ..., y_{t-l} denote the data points with values obtained from official sources before the gap and y_{t+r}, y_{t+r+1}, ..., y_m denote the data points with official values after the gap. The imputed values are calculated as

$$\hat{y}_t = y_{t-l} + l\frac{y_{t+r} - y_{t-l}}{l+r}. \qquad (1)$$

b) Trend smoothing:

A regression is run on the model:

[1] For overviews see Dagum and Cholette (2006), Durrant (2005).

$$y_t = \alpha + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \beta_4 t^4 + \rho u_{t-1} + u_t\,, \tag{2}$$

where y_t is an official value recorded for year t and u_t is the residual in that year. This is a special version of trend smoothing models, initially proposed by Holt (1957), Winters (1960) and Brown (1963),[2] in which the smoothed trend is modeled by a fourth-order polynomial of time (t). It is equivalent to an MA-1 model, following a Box and Jenkins (1976) approach, with a deterministic polynomial trend.

The (MA) term ρu_{t-1} has been included in the model only when serial correlation could be considered to be significant. This has been assumed to be the case when the value of the Durbin-Watson statistics fell into the range between 1.5 and 2.5. ρ has then been estimated following the method of Prais and Winsten (1954). When the absolute amount of the estimator of ρ reached levels greater than 0.95, the model above has been considered to be inappropriate for the prediction of missing data points.

The terms $\beta_1 t$, $\beta_2 t^2$, $\beta_3 t^3$ and $\beta_4 t^4$ remain in the model only if the respective β-coefficient has been approved to be non-zero by a t-test at the 5% significance level. The final model has been derived using backward selection: starting from the complete model (2), step-by-step the least significant coefficient (identified by the highest p-value) has been set to zero until all remaining β coefficients turned out significant.

Once the parameters of model (2) have been estimated, the imputed values are calculated as

$$\hat{y}_t = \alpha + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \beta_4 t^4 + \rho \hat{u}_{t-1}\,, \tag{3}$$

Trend smoothing has not been applied when less than ten observations are available for the regression, when a dependent variable turned out to be collinear with the trend or an exponential transformation of the trend, when the residuals of the basic model (without MA-term) show an autoregression coefficient close to or greater than 1 or close to or smaller than -1, or when the approximation method, used in the Prais-Winsten approach for the determination of ρ, failed to achieve convergence.

<u>c) Benchmarking of growth rates on aggregates</u>

The average annual growth rate observed in the time span from t-l to t is calculated for a group of commodities or countries to which the missing data point belongs:

$$\bar{r} = \left(\left(\frac{\sum\limits_{c \neq c_0} y_{c,t-l+1}}{\sum\limits_{c \neq c_0} y_{c,t-l}} \right) \left(\frac{\sum\limits_{c \neq c_0} y_{c,t-l+2}}{\sum\limits_{c \neq c_0} y_{c,t-l+1}} \right) \cdots \left(\frac{\sum\limits_{c \neq c_0} y_{c,t}}{\sum\limits_{c \neq c_0} y_{c,t-1}} \right) \right)^{1/l} - 1\,, \tag{4}$$

where c is the identifier of the commodity or countries belonging to the same group as the missing data point, c_0. The imputed value is derived by applying this average growth rate to the last official value which can be found before the missing data point (in year t-l).

$$\hat{y}_{c_0,t} = y_{c_0,t-l}\left(1 + \bar{r}\right)^l\,, \tag{5}$$

In the case of a time-series gap before which the last official value is zero, the imputed value has been derived by back-casting the first official value behind the gap using the average growth rate of the benchmark group observed from year t to year t+r (analogously to formula 4).

The choice of the commodity and/or country group on which the growth rate is benchmarked has been done according to the following order of priorities:

1. Parent commodity group in the same country
2. Same commodity in the region (a group of countries, smaller than a continent)
3. Parent commodity group in the region
4. Same commodity in the continent
5. Parent commodity group in the continent

[2] See Gardener (1985; 2006).

6. Same commodity in the world's top 20 countries[3]
7. Parent commodity group in the world's top 20 countries

When the average growth rate could not be calculated at one of the aforementioned levels, because of missing data, the next higher level has been chosen.

2. Data

For the assessment of the accuracy of the imputation methods above random samples of official values have been drawn from the FAOSTAT data. The universe of those samples is represented by around 180 thousand data points with official values recorded in the FAOSTAT database for production of primary crops, area harvested for primary crops and meat production from 1990 to 2009. This baseline dataset has been divided up into three strata which correspond to the aforementioned three groups of data (primary crops; area harvested; meat production). From each stratum approximately thousand data points have been selected randomly without replacement.

Table 1: Sample Sizes Used for the Simulation of Missing Values in Different Surroundings

Surrounding (draw)		Area harvested for crops	Primary crops production	Meat production
Middle of a series gap,	gap length: 1 year	936 (78,550)	948 (94,142)	963 (9,808)
Middle of a series gap,	gap length: 3 years	919 (69,123)	933 (83,616)	970 (8,669)
Middle of a series gap,	gap length: 5 years	915 (61,267)	949 (74,668)	971 (7,754)
Right margin of a series gap,	gap length: 2 years	935 (75,228)	934 (90,888)	961 (9,425)
Right margin of a series gap,	gap length: 3 years	920 (72,812)	927 (88,329)	979 (9,139)
Right margin of a series gap,	gap length: 5 years	922 (68,594)	932 (83,996)	973 (8,646)
Truncation of a series	distance to series end: 1 year	938 (87,185)	949 (103,739)	967 (10,780)
Truncation of a series	distance to series end: 2 years	924 (83,584)	929 (100,144)	957 (10,348)
Truncation of a series	distance to series end: 3 years	927 (80,530)	947 (97,063)	978 (10,005)
Truncation of a series	distance to series end: 5 years	932 (75,993)	952 (92,339)	983 (9,459)

Note: Number of data points in the baseline dataset (universe) in parentheses.

The random draws have been repeated ten times to select only data points which can be used for the

[3] The 20 countries in the world for which the largest amounts of the given item and element have been recorded.

simulation of different situations in which missing values can occur. With regard to the situations we distinguish, firstly, between gaps in the time series and truncations of the series. In the case of a time-series gap, data is missing throughout a number of consecutive years, and official values are recorded before and behind the analyzed missing value. In the case of a truncation of the series, official data can be found before but not behind the missing value. Secondly, we distinguish between different lengths of the time-series gaps or, in the case of truncated series, different distances of the missing value to the time-series endpoint. Furthermore, in the case of a time-series gap, a distinction has been made whether the missing value is located in the middle or at the right margin of that gap. In each random draw, the data points which do not allow a simulation of the particular surrounding have been dropped from the universe. Table 1 shows the exact sample size and the size of the universe, differentiated by stratum, in each draw.

3. Assessment of Accuracy and Applicability

Once the samples of data points with official values have been drawn, the imputation methods described in Section 1 above have been applied to the data points in these samples. Imputed values could not be compiled for all data points in the samples. In the case of the aggregate-benchmarking approach this could be caused by insufficient data available for the calculation of growth rates or zero values before and after the time-series gap or truncations. In the trend-smoothing approach successful imputation could be jeopardized by statistical limitations of the applied regression model (see above, Section 1). The extent to which an imputation method leads valid results, i.e. its applicability, constitutes an important quality criterion. It has been measured by the proportion of data points with imputed values in the total sample of data points, in the following referred to as “coverage rate”.

For the successfully imputed data points comparison of the imputed values with their corresponding official figures allows us assessing the accuracy of the respective method applied. Accuracy has been measured in terms of the root mean squared error, defined as

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2}\ . \tag{6}$$

Before computation of that indicator the data recorded in FAOSTAT have been indexed to the time-series mean calculated over the time period from 1990 to 2009. This ensures that an error with an amount of one percent of the time-series mean is given equal weight throughout the entire sample, regardless of the unit of measurement, size of the country or economic importance of the commodity group to which the data point belongs.

To analyze the statistical significance of the results – which are based on sampling – confidence intervals have been compiled using bootstrapping with 100 repetitions.

4. Results

The graphs in the figures A1 to A3 in the Annex show the coverage rate and root mean squared error achieved with different imputation methods, for different types of variables, and for different situations in which missing values occur. The following findings can be derived from these graphs:

- The trend smoothing approach leads imputable values for fewer cases than the linear interpolation and the benchmarking approach. This is the case for all scenarios analyzed. In time-series gaps the coverage rate ranges between 65 and 80 percent (see figures A1 and A2); for missing values behind time-series endpoints this rate is even smaller, amounting to around a half (see figure A3).
- In contrast to trend smoothing, the coverage rate achieved with benchmarking on aggregates is consistently high, ranging between 90 to 98 percent for all situations analyzed. For gaps in the time series this approach is thereby only slightly less exhaustive than linear interpolation which leads by

definition a coverage rate of 100 percent.

- The accuracy achieved with the three methods does not differ much for imputations within time-series gaps. The differences are in many situations not statistically significant. The ordering of imputation methods depends on the analyzed variable. In imputations for meat production data, for instance, benchmarking of growth rates on aggregates leads consistently, and significantly, a greater root mean squared error than linear interpolation and trend smoothing. In the middle of time-series gaps, the trend smoothing approach tends to be less exact than the other two methods; this is not the case for imputations at the right margin of these gaps. The accuracy of the benchmarking approach tends to diminish as the size of the gap increases.
- For extrapolations behind the time-series endpoints, the benchmarking approach leads a significantly and remarkably higher accuracy than trend smoothing. While for benchmarking on growth rates the observed root mean squared error is always below 50 percent of the time series mean, for trend smoothing it takes values from 50 to almost 260 percent.
- Data on meat production seem on average better predictable than data on crop production and harvested area.

5. Conclusions

All in all, the findings above reveal that benchmarking on growth rates is much more often applicable than trend smoothing. For imputations behind the endpoints of time series (extrapolation), this approach leads also a greater accuracy. For imputations within time-series gaps, it is dependent on the analyzed variable and gap-length which approach produced the more exact results. Linear interpolations are always feasible within time-series gaps. Their accuracy is in no case significantly weaker than the accuracy of any of the other two approaches. However, linear interpolations are not an available option for imputations behind the endpoint of a time series.

The comparatively low applicability of trend smoothing, revealed by the results above, could be explained, firstly, by the limited amount of information available for underlying the trend regressions. The annual time series on which the regressions are based have a maximum length of 39 years, notably from 1961 to 2009. In many cases, however, they consist of less than 30 valid observations, so the degrees of freedom are often not sufficient to guarantee reliable prediction. Secondly, a substantial number of time series have been found to be non-stationary. At the current stage of research, in these cases no estimation has been carried out. In a later stage it is intended to also apply estimations in first differences or growth rates and thereby increase the coverage rate of trend smoothing.

That in the case of extrapolations the accuracy of trend smoothing is lower than of benchmarking growth rates on aggregates arises from the fact that the former approach makes use of any information only for past years. Benchmarking of growth rates, in contrary, is based on information from the current year, which is a valuable resource for the prediction of missing data, even if that information does not refer to the identical commodity or country, but to a group of familiar items. Data on agricultural production are known to show frequently deviations from the trend observed in the past, caused for instance by unpredictable changes in meteorological conditions, animal or plant diseases, floods, fires and other natural catastrophes.

For the Production and Trade Team within the ESS Division of FAO, the evaluation of accuracy of imputation methods represents work in progress. In the following time the study is intended to be further developed, especially by undertaking the following next steps:

- Refinement of the models applied for the trend estimation, especially with the aim to cope with time series that have a unit root;
- refinement of the benchmarking approach, especially by enhancing the method for selection of benchmark groups, for example by identifying ‘nearest neighbours’ based on statistical techniques;
- extension of the present analysis to other commodity groups.

REFERENCES

Armstrong, J.S., 2006, Findings from Evidence-based Forecasting: Methods for Reducing Forecast Error, *International Journal of Forecasting,* 22, 583-598.

Box, G.E.P. and G.. M. Jenkins, 1974, *Time Series Analysis. Forecasting and Control,* Holden-Day, San Francisco.

Brown, R.G., 1959, *Smoothing, Forecasting and Prediction of Discrete Time Series,* Prentice-Hall, Englewood Cliffs, NJ.

Castillo, M.J. and P.P. Useche, 2010, Missing Agricultural Price Data: An Application of Mixed Estimation, *Applied Economics Letters,* 17/4-6, 537-541.

Chen, B., 2007, *An Empirical Comparison of Methods for Temporal Distribution and Interpolation at the National Accounts,* Bureau of Economic Analysis, Washington, http://www.bea.gov/papers/pdf/chen_temp_aggregation_wp.pdf.

Cooper, D., 2010, Imputing Household Spending in the Panel Study of Income Dynamics: A Comparison of Approaches, *Federal Reserve Bank of Boston, Working Papers,* 12/12.

Dagum, E.B. and P.A. Cholette, 2006, *Benchmarking, Temporal Distribution, and Reconciliation Methods for Time Series,* Springer, New York.

De Gooijer, J.G. and R.J. Hyndman, 2006, 25 Years of Time Series Forecasting, *International Journal of Forecasting,* 22, 443-473.

Durrant, G.B., 2005, Imputation Methods for Handling Item-Nonresponse in the Social Sciences: A Methodological Overview, ESRC National Centre for Research Methods and Southampton Statistical Sciences Research Institute, *NCRM Methods Review Papers,* NCRM/002.

Food and Agriculture Organization of the United Nations (FAO), 2011, *FAOSTAT,* http://faostat.fao.org, data extraction from 6 April 2011.

Gardener, E.S. Jr., 1985, Exponential Smoothing: The State of the Art, *Journal of Forecasting,* 4, 1-28.

Gardener, E.S. Jr., 2006, Exponential Smoothing: The State of the Art – Part II, *International Journal of Forecasting,* 22, 637-666.

Holt, C.C., 1957, Forecasting Seasonals and Trends by Exponentially Weighted Moving Averages, *ONR Memorandum,* 52, PA, Pittsburgh; reprinted in: International Journal of Forecasting, 2004, 20, 5-10.

Hu, M. and M.S. Salvucci, 1998, Evaluation of Some Popular Imputation Methods, *American Statistical Association,* Proceedings from the Survey Research Method Section, 308-313.

Kalton, G. and D. Kasprzyk, 1982, Imputing for Missing Survey Responses, *American Statistical Society, Proceedings of the Survey Research Methods Section*, 22-31.

Lawrence, M.J. R.H. Edmundson and M.J. O"Connor, 1985, An Examination of the Accuracy of Judgemental Extrapolation of Time Series, *International Journal of Forecasting,* 1/1, 25-35.

Makridakis, S. and M. Hibon, 1979, Accuracy of Forecasting: An Empirical Investigation (with Discussion), *Journal of the Royal Statistical Society,* A 142, 97-145.

Marcellino, M., 2005, Pooling-based data interpolation and backdating, *C.E.P.R. Discussion Papers,* 5295.

Newbold , P. and C. W. J. Granger, 1974, Experience with Forecasting Univariate Time Series and the Combination of Forecasts, *Journal of the Royal Statistical Society,* Series A, 137/0, 131-146, 1974.

Prais, S. J. and C. B. Winsten, 1954. Trend Estimators and Serial Correlation, *Cowles Commission Discussion Paper,* No 383, Chicago.

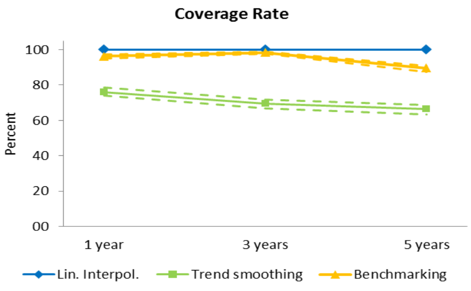
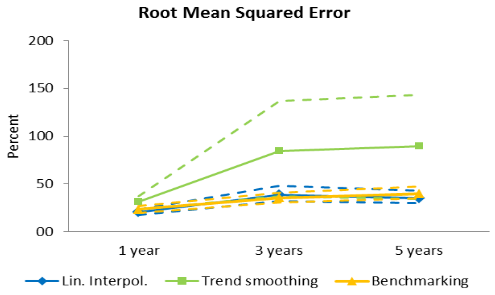
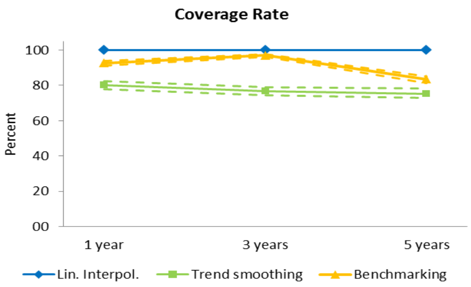
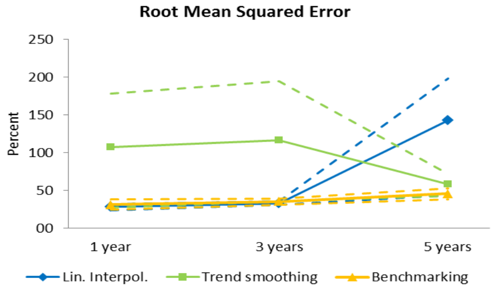
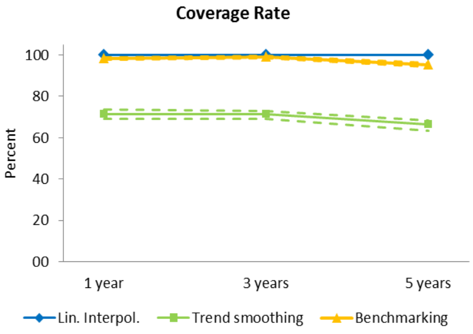
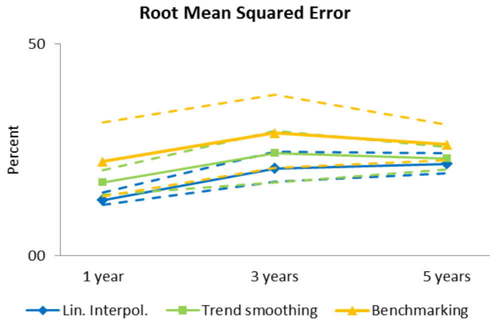
Reid, D.J., 1969, *A Comparative Study on Time Series Prediction Techniques on Economic Data,* PhD Thesis, Department of Mathematics, University of Nottingham.

Reid, D.J., 1975, A Review of Short Term Projection Techniques, in: H.D. Gordon (ed.), *Practical Aspects of Forecasting,* Operational Research Society, London, 8-25.

Winters, P.R., 1960, Forecasting Sales by Exponentially Weighted Moving Averages, *Management Science,* 6, 324-342.

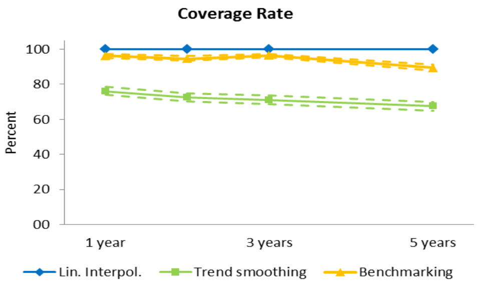
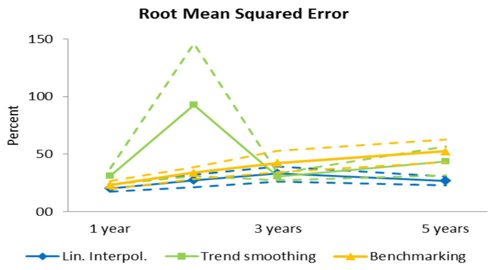
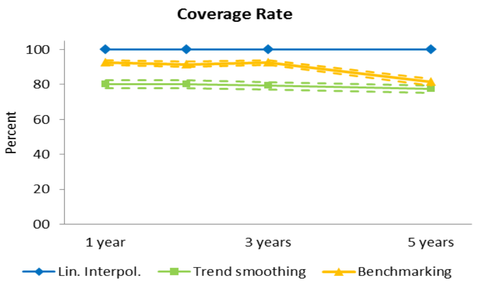
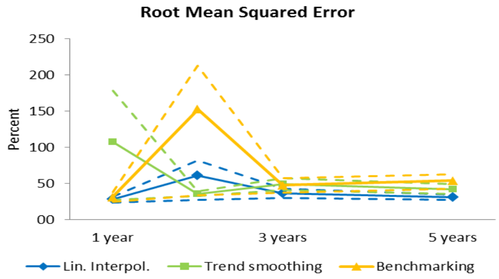
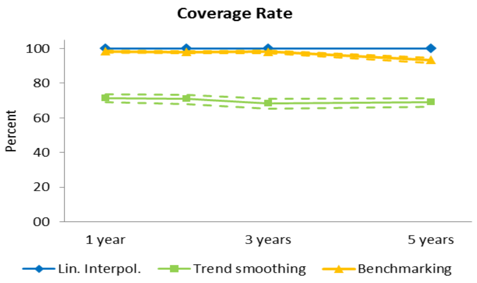
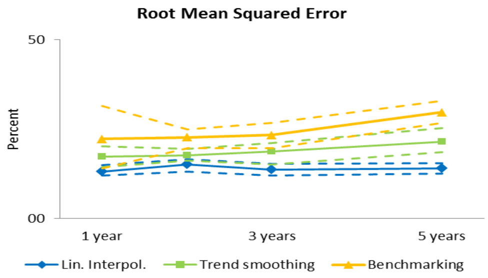
Annex: Charts

Figure A1: Applicability and accuracy of imputations – in the middle of a time-series gap

	Coverage rate[1]	Root mean squared error
Area harvested for crops		
Crops production		
Meat production		

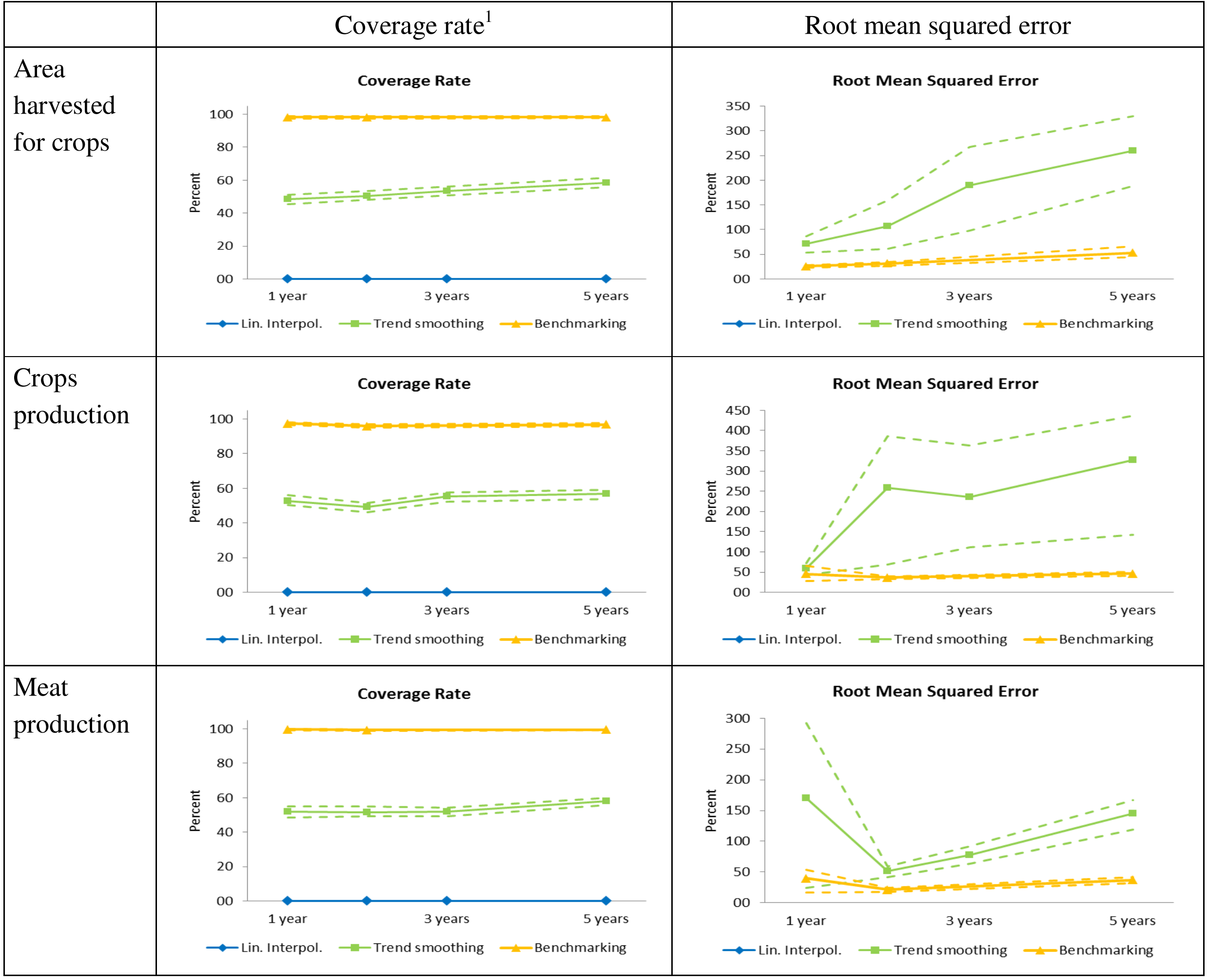
Notes: X-axis measures the length of the series gap. [1]) Proportion of imputed data points.

Figure A2: Applicability and accuracy of imputations – at the right margin of a time-series gap

	Coverage rate[1]	Root mean squared error
Area harvested for crops		
Crops production		
Meat production		

Notes: X-axis measures the length of the series gap. [1]) Proportion of imputed data points.

Figure A3: Applicability and accuracy of imputations – behind the time-series endpoints



Notes: X-axis measures the distance to the end point of the time series. [1]) Proportion of imputed data points.