A global database of public agricultural R&D investment: 1960-2022. Supplementary Information

Michiel van Dijk[1,2,*] Keith Fuglie[3] Paul W. Heisey[4] Haiyan Deng[5]

april 16, 2025

Contents

[1] Wageningen Economic Research, the Hague, the Netherlands
[2] International Institute for Applied Systems Analysis, Laxenbourg, Austria
[3] Economic Research Service, U.S. Department of Agriculture, Washington, DC, USA
[4] Agricultural economist, retired in 2019 after 21 years with the U.S. Department of Agriculture Economic Research Service and 13 years with the International Maize and Wheat Improvement Center (CIMMYT)
[5] School of Economics, Beijing Institute of Technology, 100081 Beijing, China

[*] Correspondence: Michiel van Dijk <michiel.van.dijk@wur.nl>

A Concepts to define and measure public agricultural R&D investment

Unless otherwise indicated, the definitions are directly taken from Chapter 2 and Annex 2 in OECD (2015):

- *Agricultural research intensity (ARI)* is the ratio of (public) agricultural R&D expenditure to agricultural GDP but sometimes also other denominators, such as agricultural output, number of farms and rural population, are used (@ Pardey, Roseboom, and Anderson 1991).
- *extramural R&D* is "any R&D performed outside of the statistical unit about which information is being reported".
- *Fields of research and development (FORD)* classification is used to classify R&D units and resources by fields of inquiry, namely, broad knowledge domains based primarily on the content of the R&D subject matter. For the purpose of this study FORD 4 - Agricultural and veterinary sciences is relevant. It consist of the following subcategories: 4.1 Agriculture, forestry and fisheries; 4.2 Animal and dairy science; 4.3 Veterinary science; 4.4 Agricultural biotechnology; and 4.5 Other agricultural sciences (OECD 2015, Table 2.2).
- *Full-time equivalent (FTE) of R&D personnel* is defined as "the ratio of working hours actually spent on R&D during a specific reference period (usually a calendar year) divided by the total number of hours conventionally worked in the same period by an individual or by a group".
- *Government budget allocations for R&D (GBARD)* "encompass all spending allocations met from sources of government revenue foreseen within the budget, such as taxation. Spending allocations by extra-budgetary government entities are only within the scope to the extent that their funds are allocated through the budgetary process".
- *Headcount (HC)* of R&D personnel is defined as "the total number of individuals contributing to intramural R&D, at the level of a statistical unit or at an aggregate level, during a specific reference period (usually a calendar year)".
- *Intramural R&D expenditures* "are all current expenditures plus gross fixed capital expenditures for R&D performed within a statistical unit during a specific reference period, whatever the source of funds".
- *Public R&D* is the sum of R&D activities in the government and higher education sectors. It excludes business enterprise and private non-profit sectors (based on Chapter 1 in OECD (2015), see individual chapters for definitions and description of the government and higher education sectors).
- *Research and experimental development (R&D)* "comprise creative and systematic work undertaken in order to increase the stock of knowledge – including knowledge of humankind, culture and society – and to devise new applications of available knowledge".
- *Researchers* are "professionals engaged in the conception or creation of new knowledge. They conduct research and improve or develop concepts, theories, models, techniques instrumentation, software or operational methods".
- *R&D personnel* "include highly trained researchers, specialists with high levels of technical experience and training, and other supporting staff who contribute directly to carrying out R&D projects and activities. R&D personnel are identified according to their R&D function: Researchers, Technicians, and Other supporting staff".
- *Socio-economic objectives (SEO) classification* is used to distribute GBARD. The criteria for classification should be the purpose of the R&D program or project, i.e. its primary objective. The allocation of R&D budgets to socio-economic objectives should be at the level that most accurately reflects the funder's objective(s). For the purpose of this study SEO 8 - Agriculture is relevant (OECD 2015, 337). This SEO covers all R&D aimed at the promotion of agriculture, forestry, fisheries and foodstuff production, or furthering knowledge on chemical fertilizers, biocides, biological pest control and the mechanization of agriculture, as well as concerning the impact of agricultural and forestry activities on the environment.

This also covers R&D aimed at improving food productivity and technology. It does not include R&D on the reduction of pollution (SEO 2); on the development of rural areas; on the construction and planning of buildings; on the improvement of rural rest and recreation amenities and agricultural water supply (SEO 4); on energy measures (SEO 5); or on the food industry (SEO 6). According to these definitions, R&D expenditures on food processing, in particular those on food manufacturing processes and the development of new food products are part of SEO 8 - Agriculture and not included in SEO 6 - Industrial production and technology. However, in practice, it will be difficult to separate and identify the research expenditures on food processing, which are at the intersection of SEO 6 and 8 (Pardey and Roseboom 1989; Heisey and Fuglie 2018).

- *Types of costs of R&D* "include individual current and capital cost categories for intramural R&D. Types of current costs include labor costs for internal R&D personnel and other current costs (for external R&D personnel, purchases of services, purchases of materials, and other costs not elsewhere classified. Types of capital costs include land and buildings, machinery and equipment, capitalized computer software and other intellectual property products".

B Assessment of imputation approaches

Dealing with missing data is a key problem when constructing a database time-series information on number of public agricultural researchers (HR) and related R&D expenditures (RD). Existing studies predominantly used the agricultural research intensity (ARI), which is defined as the share of RD in agricultural GDP, and linear interpolation to impute missing data Craig, Pardey, and Roseboom (1991). To better handle gaps in the data, we systematically compared the performance of a larger number of potential 'predictors' and imputation techniques.

We considered several additional variables that are expected to be correlated with public agricultural R&D investment, such as the government budget allocations for R&D (GBARD), which reflect the planned expenditures on public agricultural R&D. Another relevant indicator is the number of public agricultural researchers as labor costs constitute an large part of total RD. Similarly, data on research spending might also support the imputation of missing research personnel data. Finally, we also tested the use of more advanced approaches than linear interpolation to impute missing values (James Honaker and King 2010; J. Honaker, King, and Blackwell 2011).

We ran a number simulations to systematically compare the predictive power of these variables and under which conditions. More specifically, we conducted three experiments, each covering a particular missing data problem: (a) imputation of missing data for the most recent years (2 versions) and (b) imputation of missing data within time series. Imputation, when data is missing for distant years is similar to problem (a) and therefore not separately evaluated. Moreover, data on predictors that are needed for the imputation are not always available for the most distant time-periods (i.e. 1960s-1970s).

We used the same design for each experiment:

- Step 1: We selected the countries with the 'best' data, which we defined as the longest consecutive times series with no missing data points (see details for each experiment below). Depending on the experiment and selected predictor, the composition and size of the sample might differ.
- Step 2: We (randomly) removed streaks of observations (both for HR and RD) and replaced them by imputed values. We combined forecasting and imputation techniques, when data was missing for recent years, and imputation and interpolation approaches for data missing within the time-series. In practice the period for which data is missing can vary from a few to more than 30 years. To simulate this, we removed streaks of observations with different length between year t and 2022. In case of the within time-series imputation, where both the start and end year matter, we randomly deleted and imputed streaks within the time series, and repeated this ten times for each streak.
- Step 3: We evaluated the accuracy of the imputation using the mean absolute scaled error (MASE). The advantage of MASE over other metrics is that it is unit-free and therefore can be used to compare number of researchers and R&D expenditure imputations, which each have a different scale (Hyndman and Athanasopoulos 2021).

We started by exploring the correlations between HR and RD with potential predictors, including agricultural GDP (AGGDP), agricultural output (AGOUTPUT) and GBARD. Table S1 shows the results of a simple regression where HR/RD is the dependent variable and HR/RD, AGGDP, AGOUTPUT and GBARD are the independent variables. We included country fixed effects to control for time-constant unobserved heterogeneity in our panel data set. The table shows that there is a strong correlation between HR and RD and with AGGDP, AGOUTPUT and GBARD, confirming that all selected indicators potentially can support the imputation of missing information.

	HR-RD	HR-GBARD	RD-GBARD	HR-AGGDP	RD-AGGDP	HR-AGOUTPUT	RD-AGOUTPUT
Constant	33.75	527.39***	556.57***	−329.16	−74.85	−919.38*	−202.41*
	(317.89)	(150.80)	(90.52)	(411.05)	(83.95)	(397.62)	(80.76)
RD	3.07***						
	(0.05)						
AGGDP				338.74***	72.85***		
				(51.94)	(11.21)		
AGOUTPUT						1659.99***	358.05***
						(87.90)	(17.27)
GBARD		−0.83***	0.44***				
		(0.17)	(0.03)				
Country fixed effects	Yes	Yes	yes	yes	yes	yes	yes
Number of countries	161	38	42	162	162	162	162
R^2	0.92	0.97	0.97	0.86	0.81	0.87	0.82
Adj. R^2	0.91	0.97	0.97	0.85	0.80	0.86	0.81
Num. obs.	4792	848	1137	5021	5528	5021	5528

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table S1: Regression of the number of public agricultural researchers (HR) and public agricultural R&D expenditures (RD) on agricultural GDP (AGGDP), agricultural output (AGOUTPUT) and agricultural R&D government budget allocations (GBARD).

In the first experiment, we used the growth rate of HR/RD, AGGDP, AGOUTPUT and GBARD to impute missing observations of HR/RD. We selected countries for which complete information was available for HR, RD, AGGDP, AGOUTPUT and GBARD for a long time-period. Taking into account trade-offs between coverage of years and countries, we decided to take a cut-off of at least 21 years of data availability (i.e. a streak of max 20 years), resulting in a data set with a sample of 23 countries. All of these are countries covered by the EUROSTAT-OECD database, which is the only source of consistent HR, RD and GBARD data. AGGDP and AGOUTPUT were taken from the database with macro-economic data (see main text).

Apart from the growth rate of HR/RD, AGGDP, AGOUTPUT AND GBARD, we also used two common time-series forecasting approaches: simple exponential smoothing (SES) and Damped Trend (DT). SES is the preferred method for stationary data with no clear trend, giving more weight to recent observations while DT works best for data with a (long-run) trend (Hyndman and Athanasopoulos 2021). As these approaches require at least 10 observations to create reasonable forecasts, we only applied time-series forecasting techniques to impute data for streaks of 10 years or less. Finally, we added a NAIVE imputation, which carries the last available observation forward.

Figure S1 compares the result of the various imputation approaches between HR and RD. We also included an ensemble imputation (ENS), representing the mean of different methods, which is expected to result in more accurate outcomes (Wang et al. 2023). The figure shows that using GBARD provides the least accurate results (i.e. lowest MAPE), when data is missing for 5-20 years. For shorter streaks, AGGDP gives the worst results. Overall, the ensemble approach excluding GBARD (ENS-GB) shows the highest accuracy across all streaks.

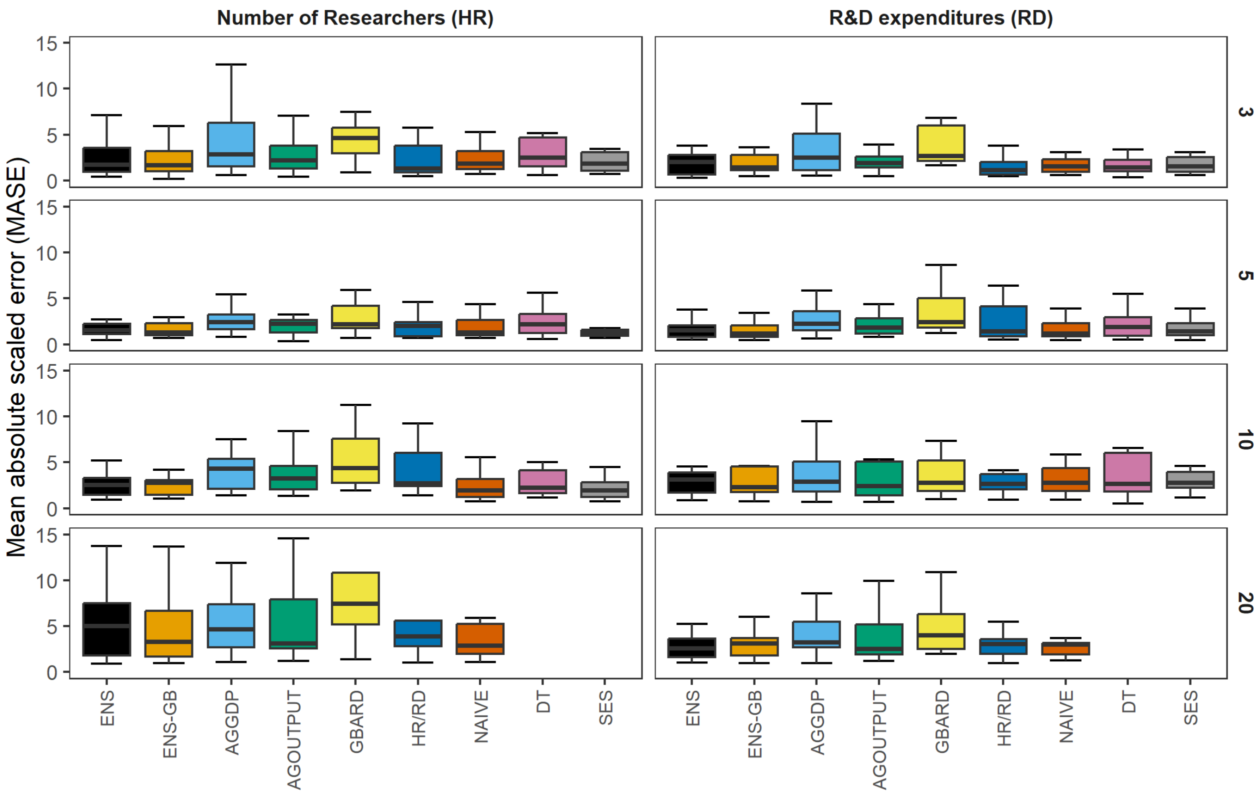

Figure S1: Assessment of different approaches (including GBARD) to impute missing RD and HR information at the end of the time-series. Numbers on the right y-axis indicate the length of the streak between year t and 2022 for which observations were removed and imputed. ENS: Ensemble imputation of AGGDP, AGOUTPUT, GBARD, HR/RD, NAIVE, and DT and SES (if available); ENS-GB: Ensemble imputation excluding GBARD; AGGDP: Imputation using AGGDP growth rate; AGOUTPUT: Imputation using AGOUTPUT growth rate; GBARD: Imputation using GBARD growth rate; NAIVE: Imputation using last available observation; DT: Imputation using DT forecasting and SES: Imputation using SES forecasting. Total number countries: 23.

In the second experiment, we focused on the performance of imputation techniques when data was missing for long streaks at the end of the time-series. This mostly occurs for small (island) countries that were covered by Pardey and Roseboom (1989) but for which no recent data was available. As the availability of GBARD data is a problem, we dropped this variable from the analysis and selected all countries that had complete information for HR, RD, AGGDP and AGOUTPUT covering at least a period of 31 consecutive years. This resulted in a data set with 27 low-, middle and high-income countries from all over the world.

The results, presented in Figure S2 are comparable to that of the first experiment. The ensemble imputation outperforms the results of the AGGDP, AGOUTPUT, HR/RD and NAIVE imputations for streaks of 10-30 years and is comparable or better than DT and SES approaches for shorter streaks. On the basis of this and the previous experiment, we decided to use the ensemble imputation, excluding GBARD, as a our preferred approach to handle missing information when data was missing for the most recent years. Imputations using DT and SES forecasting were only included if 15 original observations were available.

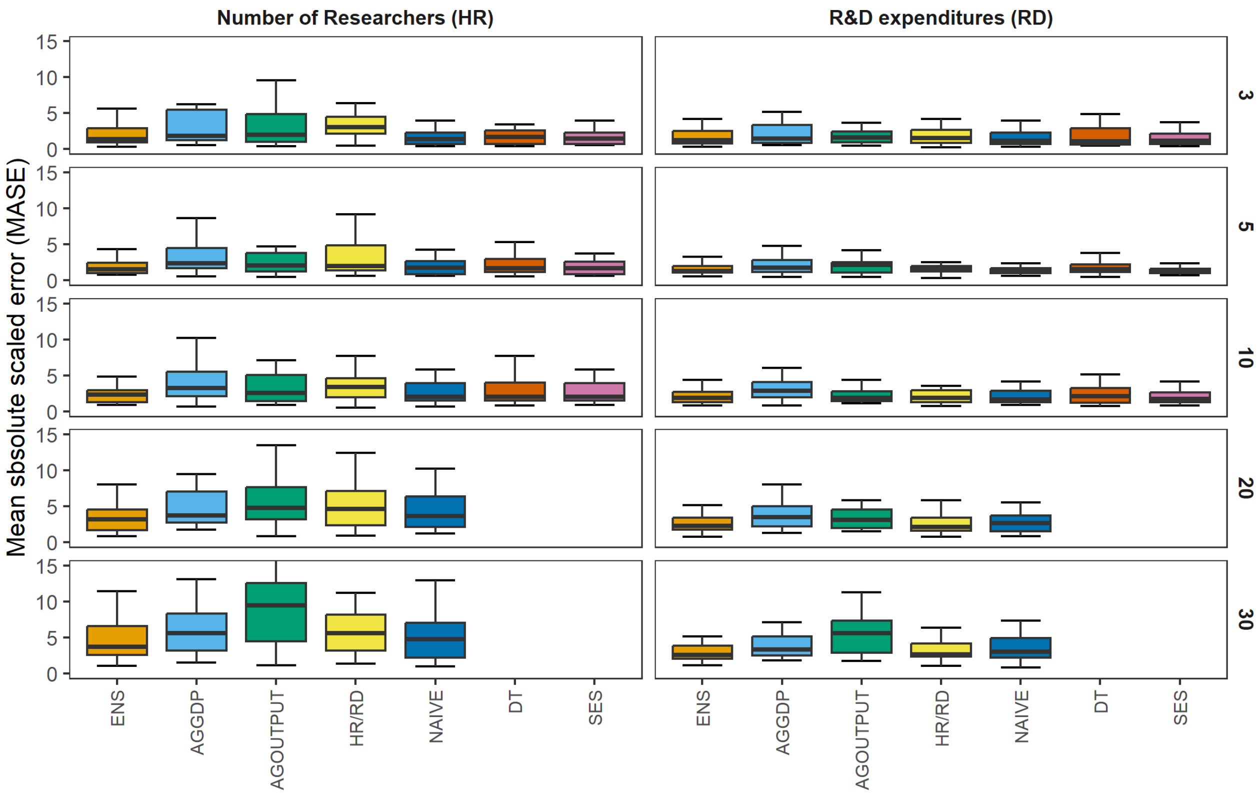

Figure S2: Assessment of different approaches (excluding GBARD) to impute missing RD and HR information when data is missing for recent years. Numbers on the right y-axis indicates the length of the streak for which observations were removed and imputed. ENS: Ensemble imputation of AGGDP, AGOUTPUT, HR/RD, NAIVE, and DT and SES (if available); AGGDP: Imputation using AGGDP growth rate; AGOUTPUT: Imputation using AGOUTPUT growth rate; NAIVE: Imputation using last available observation; DT: Imputation using DT forecasting and SES: Imputation using SES forecasting. Total number countries: 27.

In the final experiment, we investigated best-approaches to impute missing data within the time-series. This is a common problem as we are combing data on public agricultural R&D investment from different sources, often without overlap. In contrast to the previous two experiments, we now had to deal with start- and end-year information, which requires different techniques. The most simple approach is to apply linear imputation (LI) between the two years for which data is available. Another approach, often used in the literature (Pardey, Alston, and Chan-Kang 2013), is to linearly interpolate the ARI and then impute RD using information on agricultural GDP for years with missing data. Both approaches assume RD has a linear growth rate, which might contradict with the actual planning of government expenditures and budgets that often assume growth rates to set annual targets. For this reason, we also evaluated an exponential interpolation variant (EI) and its ARI equivalent (EI AGGDP). In line with the two previous experiments, we also added two comparable approaches that used the growth rate of AGOUTPUT instead of AGGDP as a reference (LI AGOUTPUT and EI AGOUTPUT). Finally, we tested the performance of Amelia (AMELIA), which uses a multiple imputation method based on bootstrapping, combined with data on potentially correlated variables, to estimate missing values (James Honaker and King 2010; J. Honaker, King, and Blackwell 2011). It was specifically developed to handle missing data in time-series-cross-sectional data. We included log AGGDP (in 2017 PPP$), log GDP per capita (in 2017 PPP$), log population, AGOUTPUT (2017 = 100), the share of rural population and the share of AGGDP in GDP as explanatory variables in Amelia. To run this experiment we restricted the test data set to at least 32 years of consecutive data (i.e. a maximum streak length of 30), resulting in a sample of 96 countries. Similar to the previous experiments, we also added the ensemble imputation (EI ENS and LI ENS).

Perhaps surprisingly, We found that the explanatory power of AMELIA was relatively low in comparison to the other approaches (Figure S3). Overall, the simple EI and LI approaches resulted in the most accurate imputations, with minimal difference between the two. The ensemble approach is comparable to the EI and LI approaches but with slightly worse performance for longer streaks in case of RD although for shorter streaks accuracy seems slightly higher. The AGOUTPUT and AGGDP variants have lower accuracy. As the implementation of AMELIA, AGGDP and AGOUTPUT requires additional information, which is not always available, and frequently long periods of missing data need to be imputed, we decided to adopt the simple EI approach to deal with missing data within the time-series.

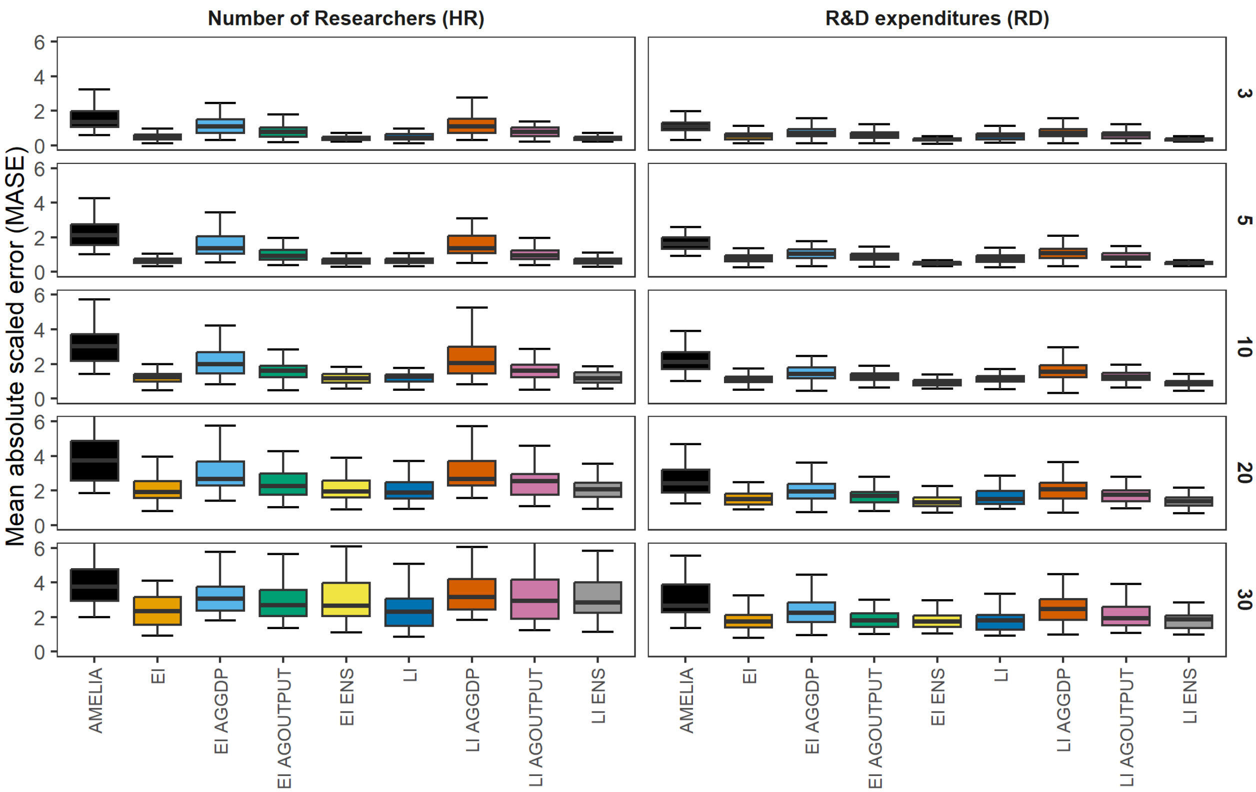

Figure S3: Assessment of different approaches to impute RD and HR missing information within the time-series. Numbers on the right y-axis indicates the length of the streak within the time-series for which observations were removed and imputed. For each streak data was randomly removed and imputed ten times. AMELIA: Imputation using the Amelia algorithm; EI: imputation using exponential growth rate; EI AGGDP: Imputation using exponential ARI growth rate; EI AGOUTPUT: similar as EI AGGDP but using AGOUTPUT, EI ENS: Ensemble imputation of EI, EI AGGDP, EI AGOUTPUT and AMELIA. The remainder are the same approaches but using linear instead of exponential growth rates. Total number countries: 96.

References

Craig, Barbara J, Philip G. Pardey, and Johannes Roseboom. 1991. "Internationally Comparable Growth, Development, and Research Measures." In *Agricultural Research Policy: International Quantitative Perspectives*, edited by Philip G Pardey, Johannes Roseboom, and Jock R Anderson, 131–72. Cambridge: Cambridge University Press.

Fuglie, Keith. 2018. "R&D Capital, R&D Spillovers, and Productivity Growth in World Agriculture." *Applied Economic Perspectives and Policy* 40 (3): 421–44. https://doi.org/10.1093/aepp/ppx045.

Heisey, Paul W., and Keith O. Fuglie. 2018. "Research Investment and Policy Reform in High-Income Countries." U.S. Department of Agriculture, Economic Research Service.

Honaker, James, and Gary King. 2010. "What to Do about Missing Values in Time-Series Cross-Section Data." *American Journal of Political Science* 54 (2): 561–81. https://doi.org/10.1111/j.1540-5907.2010.00447.x.

Honaker, J, G King, and M Blackwell. 2011. "Amelia II: A program for missing data." *Journal of Statistical*. http://artax.karlin.mff.cuni.cz/%7B~%7Dhans/src/doc/r-cran-amelia/amelia.pdf.

Hyndman, Rob J., and George Athanasopoulos. 2021. "Forecasting: Principles and Practice." OTexts. https://otexts.com/fpp3/.

OECD. 2015. *Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development*. The Measurement of Scientific, Technological and Innovation Activities. Paris: OECD. https://doi.org/10.1787/9789264239012-en.

Pardey, Philip G., Julian M. Alston, and Connie Chan-Kang. 2013. "Public agricultural R&D over the past half century: An emerging new world order." *Agricultural Economics (United Kingdom)* 44 (SUPPL1): 103–13. https://doi.org/10.1111/agec.12055.

Pardey, Philip G., and Johannes Roseboom. 1989. *ISNAR sgricultural research indicator series: A global data base on national agricultural research systems*. Cambridge: Cambridge University Press.

Pardey, Philip G., Johannes Roseboom, and Jock R. Anderson. 1991. "Regional perspectives on National Agricultural Research." In *Agricultural Research Policy: International Quantitative Perspectives*, edited by Philip G. Pardey, Johannes Roseboom, and Jock R. Anderson, 197–264. Cambridge: Cambridge University Press.

Wang, Xiaoqian, Rob J. Hyndman, Feng Li, and Yanfei Kang. 2023. "Forecast combinations: An over 50-year review." *International Journal of Forecasting* 39 (4): 1518–47. https://doi.org/10.1016/j.ijforecast.2022.11.005.