Seventh International Conference on Agricultural Statistics

ICASVII

2016 Rome, 26 | 28 October

Food and Agriculture Organization of the United Nations

Istat

Assessing the Impact of Agricultural Research: Data Requirements and Quality of Current Statistics in Europe

Alessandra Coli
University of Pisa, Department of Economics and Management – Pisa, Italy
alessandra.coli1@unipi.it

Fabio Bartolini
University of Pisa, Department of Agriculture, Food and Environment – Pisa, Italy
fabio.bartolini@unipi.it

Alessandro Magrini
University of Florence, Department of Statistics, Informatics, Applications – Florence, Italy
University of Pisa, Department of Agriculture, Food and Environment – Pisa, Italy
magrini@disia.unifi.it

Barbara Pacini
University of Pisa, Department of Political Sciences – Pisa, Italy
barbara.pacini@unipi.it

Linda Porciani
ISTAT, Tuscany – Florence, Italy
porciani@istat.it

ABSTRACT[1]

Assessing the impact of agricultural research on sustainability targets often implies to face two main issues: the complexity of the causal path, and the lack of appropriate data. In this paper, we discuss which data would be necessary to measure short- and long-term impacts in Europe, and we suggest a set of indicators to evaluate their quality, considering both metadata (qualitative indicators) and collected data (quantitative indicators). An application is shown for a selection of 20 variables, covering 15 EU countries, over the 1980-2015 period. According to our results, qualitative and quantitative indicators often provide contrasting indications. This result suggests that both qualitative and quantitative measures are necessary to assess the overall quality of available statistics.

Keywords: Data quality indicators; Data quality dimensions; Impact of agricultural research.

1. Introduction

Agriculture is an important target of EU and national policies. In particular, there is an increasing demand of knowledge on the effects of agricultural research on the EU sustainability targets. The achievement of such knowledge depends on two main factors: the complexity of the causal path, and the lack of appropriate data. Agricultural activities produce effects through a large number of pathways, from short-term impacts on agriculture production to long-term impacts on people’s sustainable well-being. Ideally, a unified analytical approach would jointly consider impacts across all the relevant sustainability dimensions at a local, national and over-national level. Methodologies in the literature range from disaggregate to aggregate analysis, from the assessment of economic rate of return to the assessment of multi-dimensional impacts. However, the extent of available statistical methods often contrasts with a general lack of appropriate data. This paper provides an insight into the quality of data (Eurostat data) when the purpose of research is measuring the effects of agricultural research on multiple targets in Europe. The paper is organized as follows. In Section 2, we concisely focus on the selection of variables necessary to investigate the short- and long-term effects of agricultural research in Europe. In Section 3, we suggest some synthetic indicators for measuring the quality of data. In Section 4, we compute such indicators for a selection of variables. Conclusions are provided in Section 5.

2. Data: themes of interest

Following Bartolini et al. (2014), we delineate the impact pathway from agricultural research expenditure to multiple sustainability dimensions through five interconnected levels: context/external drivers, investment, research activity, outcome and impact.

The context/external drivers level accounts for countries’ specific characteristics, which may act as a confounder of the relationships among the other levels. Context variables include macroeconomic variables (e.g. gross domestic product) as well as the disposal of agricultural resources (e.g. land and labour). External drivers take into account policies, regulations and laws, as

[1] Acknowledgements. The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 609448. The contents of this publication are the sole responsibility of the implementing partner of the project IMPRESA project (EU-FP7 project ,http://www.impresa-project.eu) and can in no way be taken to reflect the views of the European Union.

well as technology innovations from economic sectors other than agriculture (for instance, chemical and mechanic patent applications as pointed out by Thirtle et al., 2008).

The investment level includes the variables describing how agricultural research is funded within each country (e.g., general government and business enterprise expenditure).

Outputs of research activity represent the first and most immediate results of research investments. Campbell et al. (2013) consider human resources (e.g., number of high qualified researchers), quality of research (e.g., number of EU funded projects), innovation (e.g., number of patent applications), research infrastructures, industrial specialization and publications.

The outcome level includes the immediate impact of research activity on farm production. Productivity of the agricultural sector is the representative variable of this level.

The impact level contains variables non-immediately affected by research investment and encompasses multiple dimensions. This level includes for instance, changes in farmers' economic conditions and wellbeing, changes in environmental conditions (pollution emissions; biodiversity; soil and water quality) and changes in social conditions (health, education, food security, poverty, migration, etc.).

Output, outcome and impact variables identify the possible targets of European agricultural research. Applying textual mining techniques on the abstracts of EU funded research projects in agriculture[2], Bartolini *et al.* (2016) analyse changes in the share of budget among different research targets from 1994 to 2009. Authors find that, during late 90s, economic competitiveness and reduction of environmental pressure were the highest priorities of research projects, while, since 2000, the larger share of budget was finalized to support projects having an expected impact on the health of European consumers and citizens.

Assessing the impact of agricultural research requires recovering adequate data for each level of the impact pathway. Eurostat, FAO, OECD, ILO, the World bank and other international institutions disseminate data on most of them. However, due to the heterogeneity of the issues covered, availability and quality of data may vary significantly across countries and time. Along with well- established and harmonized statistics (e.g., labour or national accounts statistics), we find poor quality data. At first glance, statistics seem adequate but sometimes they conceal missing values, short time series or breaks in the series. In our view, it would be helpful if statistics were disseminated along with synthetic quality scores, in order to make immediately clear their actual usability. In the next sections, we suggest some quality indicators and present results for a selection of 20 variables representative of the impact pathway from agricultural research expenditure to multiple sustainability dimensions.

3. Quality indicators

Assessing the impact of agricultural research in European countries requires managing both time series and cross-section data sets. On the one hand, long time series for investment and research variables are required since their effects on target variables occur at different time lags. On the other hand, complete and comparable cross section statistics are needed to allow international comparisons. Thus, comparability over time and among countries represent the most important quality requirements.

Several institutions disseminate time series and cross-section data sets on the themes of interest detailed in Section 2, so that the identification of the best data source for each variable is a necessary first step. In this paper, we focus on Eurostat statistics only, as Eurostat is the primary data source for European countries, and disseminates the best metadata on data quality through single reports for each statistic (Euro-SDMX Metadata Structure files; ESMS files henceforth).

[2] 4th FP (1994-1998), 5th FP (1999-2002) and 6th FP (2002-2006) projects. Only projects with main topic 'Agriculture and food' or subtopic 'Agriculture' within the 'Biotechnology' topic were selected.

However, basing on Eurostat available metadata it is not immediate to detect the overall quality level of each variable, nor to understand for which analysis each variable could be fruitfully used (time series or cross sectional analysis, or both).

3.1 Qualitative indicators based on Eurostat metadata

A detailed report on data quality (ESMS file) is available for all the statistics in the Eurostat database. ESMS reports contain very useful information but their length (no less than 5 pages) and their level of detail may discourage users.

We summarize ESMS reports into four variables. The first variable considers the typology of data sources used to collect/produce data, assuming that the level of comparability and accuracy decrease going from Censuses to National Accounts, Surveys, Administrative data sources and Mixed data sources (such as inventories derived from various data sources). The second variable takes into account the ‘Institutional mandate’ section of ESMS files, which specifies if statistics are produced/collected on behalf of EU regulations and if they are disseminates on a mandatory, gentlemen’s agreement or voluntary basis. In this case, we assume that data quality improves if the collection, production and transmission of data are regulated. The third and fourth variables assign a quality level (low, medium, good and high) on the temporal and the geographical comparability, respectively. The quality level is derived directly from the assessments given in the ‘coherence and comparability’ section of ESMS documents.

3.2 Quantitative indicators based on data evidence

We develop several quality indicators on the basis of the evidence stemming from collected statistics. We considered two features of quality: missing data and outlier data. First, we focus on missing values, providing measures of their incidence both in time and space (i.e. across countries). Then, we consider the incidence of contiguous values in each time series. Finally, we focus on the detection of outliers, once all the time series are made stationary. The value of each indicator varies from 0 (minimum quality) to 1 (maximum quality).

Notation is the following. The set of countries is denoted as $j = 1, \dots, J$, and $X_{i,j,t}$ denotes the i-th variable ($i = 1, \dots, I$) in the j-th country at time slice t ($t = 1, 2, \dots, T$).

Missing data incidence

Let $o_{i,j,t}$ be a dummy variable such that $o_{i,j,t} = 1$ if the value of $X_{i,j,t}$ is available (not missing), otherwse $o_{i,j,t} = 0$. We define three indicators measuring the incidence of missing data.

- **Spatial Availability Index**. Proportion of available data for a certain variable in a certain country:

$$SAI_{ij} = \frac{1}{T}\sum_{t=1}^{T} o_{i,j,t}$$

- **Temporal Availability Index**. Proportion of available data for a certain variable at a certain time slice:

$$TAI_{it} = \frac{1}{J}\sum_{j=1}^{J} o_{i,j,t}$$

- **Contiguity Index**. Contiguity of available data for a certain variable in a certain country, computed as the proportion of available data adjacent to an available datum.

$$CI_{ij} = \frac{1}{T-1}\sum_{t=1}^{T-1} o_{i,j,t}\, o_{i,j,t+1}$$

Outlier data incidence

The distribution of a time series may change through time, that is it may contain an unit root or may not be stationary. If this is the case, the detection of outlier data does not make sense. For each variable i and for each country j, denote the order of integration as d_{ij}, that is the minimum number of differences required to obtain a significant result of the Dickey-Fuller test (rejection of the unit root hypothesis). Consider the *Skewness-adjusted Outlyingness* (Brys et al., 2005), a robust measure of outlyingness for skewned distributions:

$$\zeta_{ijt} = \begin{cases} \dfrac{\tilde{X}_{i,j,t} - M_{ij}}{R_{ij} - M_{ij}} & \tilde{X}_{i,j,t} \geq M_{ij} \\ \dfrac{M_{ij} - \tilde{X}_{i,j,t}}{M_{ij} - L_{ij}} & otherwise \end{cases}$$

with:

$$L_{ij} = \begin{cases} Q_{ij} - 1.5e^{-4MC_{ij}}(QQ_{ij} - Q_{ij}) & MC_{ij} \geq 0 \\ Q_{ij} - 1.5e^{-3MC_{ij}}(QQ_{ij} - Q_{ij}) & otherwise \end{cases}$$

$$R_{ij} = \begin{cases} QQ_{ij} + 1.5e^{-3MC_{ij}}(QQ_{ij} - Q_{ij}) & MC_{ij} \geq 0 \\ QQ_{ij} + 1.5e^{-4MC_{ij}}(QQ_{ij} - Q_{ij}) & otherwise \end{cases}$$

where $\tilde{X}_{i,j,t}$ represents $X_{i,j,t}$ after applying d_{ij} differences, whereas M_{ij}, Q_{ij}, QQ_{ij} and MC_{ij} are the median, the first quartile, the third quartile and the medcouple (an adjusted measure of skewness: Brys et al., 2004) of the i-th variable in the j-th country after applying d_{ij} differences, respectively. According to such outlyingness measure, $\tilde{X}_{i,j,t}$ is an outlier if $\zeta_{ijt} < L_{ij}$ or $L_{ij} > R_{ij}$. If this is the case, let $u_{i,j,t} = 0$, otherwise $u_{i,j,t} = 1$. We define the **Non-outlier Index** as the proportion of non-outlier data for a certain variable in a certain country:

$$OI_{ij} = \frac{1}{T}\sum_{t=1}^{T} u_{i,j,t}$$

4. Results

We selected a total of 20 variables representative of each level of the research impact pathway from agricultural research expenditure to multiple sustainability dimension, excepting the Output level, as we were not able to find Eurostat statistics on total factor productivity for Agriculture. Actually, according to Schreyer (OECD, 2015), in Europe only Statistics Denmark, Statistics

Finland, Statistics Sweden and ONS deliver estimates of total factor productivity for the A and B sectors of NACE classification.

We downloaded data and metadata from Eurostat website (http://ec.europa.eu/eurostat/data/database) in May 2016. We considered 15 EU countries (AT, BE, DE, DK, EL, ES, FI, FR, IE, IT, LU, NL, PT, SE and UK) in the period 1980-2015. For each of the selected variables, we computed quantitative (columns 4, 5 and 6) and qualitative (columns 8, 9, 10 and 11) indicators as defined in Section 3.2. Also, we derived an overall quantitative indicator I1 (column 7) and an overall qualitative indicator I2 (column 12). Indicator I1 is obtained as follows: quartiles of each quantitative indicator are computed, then values from 1 to 4 are assigned to each quantitative indicator for each variable depending on the nearest quartile (1 for the first quartile, 2 for the second, and so on), and finally such values are averaged for each variable. Indicator I2 is subjectively derived from the values taken by qualitative indicators. Results are shown in Table 1.

Figure 1 compares the I1 and I2 values after the variables under analysis are clustered into four groups: blue dots correspond to Economic variables, grey dots to Research variables, yellow dots to Social variables and orange dots to Environmental variables. We see that indicators I1 and I2 do not provide unanimous indication on data quality (Pearson correlation coefficient equal to -0.09): Research variables are characterized by low values of I2 and high values of I1, Economic and Social variables show a balance between the values taken by the two overall indicators, Environmental variables exhibit a heterogeneous combination.

Figure 1. Comparison of overall qualitative and quantitative indicators for a selection of variables.

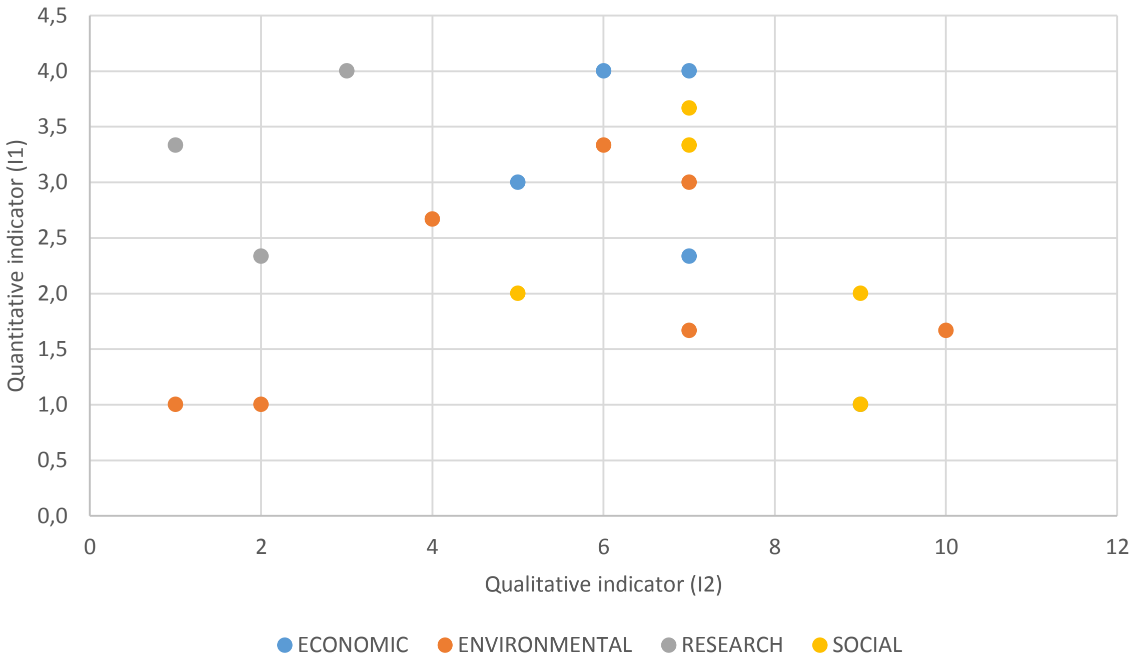

Table 1. Quality indicators based on collected data and quality reports from the Eurostat database on the 20 selected variables.

Description	Level of the impact pathway	Indicators computed on collected data				Indicators based on Eurostat ESMS files				
		Non-missing values	Contiguous over time	Non-outliers	I1: Overall quantitative	Primary data source	Institutional mandate	Comparability over time	Comparability across countries	I2: Overall qualitative
HA of arable land	ECONOMIC	0.86	0.83	0.99	4.0	Mixed sources (Economic Accounts for Agriculture)	EU Regulations (2004) sets harmonized methodology. Provision is on gentlemen's agreement	good	high	7
Business enterprise research expenditure for Agriculture	RESEARCH	0.31	0.29	0.98	2.3	Survey/Administrative	From 2004, collection is mandatory and regulated	Medium	Low	2
Greenhouse gas emissions from agriculture	ENVIRONMENTAL	0.64	0.63	0.98	3.0	Mixed sources (data produced by EEA)	Regulation in 2013. It is not clear if provision of data is mandatory	good	good	7
Energy for primary production	ENVIRONMENTAL	0.67	0.66	0.99	3.3	Censuses/Surveys (agricultural structure surveys)	Several EU Regulations. Provision is mandatory	medium	good	6
Share of energy from renewable sources in gross final energy consumption	ENVIRONMENTAL	0.28	0.26	0.98	1.7	Survey/Administrative	EC Regulations on methodologies. Data collection is voluntary	Good	good	7
Net entrepreneurial income of Agriculture – 2005=100	ECONOMIC	0.85	0.84	0.99	3.0	Mixed (Economic Accounts for Agriculture)	EU Regulations (2004) sets harmonized methodology. Provision is on gentlemen's agreement	Good	Medium	5
consumption estimate of Nitrogene	ENVIRONMENTAL	0.38	0.37	0.98	2.7	Survey/Administrative	Eurostat has no legal act in place requiring these data. The data are requested by gentlemen's agreement.	good	low	4
Government research expenditure for Agriculture	RESEARCH	0.9	0.89	0.99	4.0	Administrative	From 2004, collection is mandatory and regulated	Medium	Medium	3
Green house gas emission per capita	ENVIRONMENTAL	0.36	0.34	0.95	1.7	Mixed sources	Regulated and mandatory since 2004	high	high	10
Gross value added of Agriculture	ECONOMIC	0.88	0.88	0.99	4.0	Mixed (Economic Accounts for Agriculture)	EU Regulations (2004) sets the legal basis for a harmonized methodology. Provision is on gentlemen's agreement	Good	Good	6
Quota of persons with a good health status in rural areas	SOCIAL	0.27	0.25	0.93	1.0	Surveys/Administrative	Up to 2004(ECHP survey), data collaction was based on a gentleman's agreement. From2005 (Eusilc instrument) collection becomes mandatory.	High	High	9
Mean familiar income in rural areas	ECONOMIC	0.29	0.26	0.94	1.0	Surveys/Administrative	Up to 2004(ECHP survey), data collaction was based on a gentleman's agreement. From2005 (Eusilc instrument) collection becomes mandatory.	High	High	9
Annual work units salaried	SOCIAL	0.88	0.88	0.97	3.3	Mixed (Economic Accounts for Agriculture)	EU Regulations (2004) sets harmonized methodology. Provision is on gentlemen's agreement	good	high	7
Annual work units	SOCIAL	0.91	0.9	0.98	3.7	Mixed (Economic Accounts for Agriculture)	EU Regulations (2004) sets harmonized methodology. Provision is on gentlemen's agreement	good	high	7
Fully converted crop area (ha)	ENVIRONMENTAL	0.3	0.26	0.96	1.0	Administrative data	Gentlemen's agreement up to 2007. Then, transmission based on EU regulations	low	low	2
Number of agricultural patent applications: Agriculture, foresty, fishing	RESEARCH	0.84	0.81	0.99	3.3	Administrative	No official legal acts. Provision is voluntary	Low	Low	1
People at risk of poverty or social exclusion in thinly-populated area	SOCIAL	0.31	0.29	0.97	2.0	Surveys/Administrative	Up to 2004(ECHP survey), data collaction was based on a gentleman's agreement.	High	High	9
Utilised agricoltural area (1000 ha)	ECONOMIC	0.84	0.81	0.97	2.3	Mixed (Economic Accounts for Agriculture)	EU Regulations (2004) sets harmonized methodology. Provision is on gentlemen's agreement	good	high	7
Unemployment rate in rural areas	SOCIAL	0.61	0.6	0.97	2.0	Survey (EU-LFS)	Regulated and mandatory collection since 1998	medium	medium	5
Water used in Agriculture, forestry and fishing	ENVIRONMENTAL	0.01	0.01	-	1.0	Survey/Administrative	Data collection is voluntary	low	low	1

5. Concluding remarks

The evaluation of the short- and long-term impacts of agricultural research in Europe is an important theme for EU decision-making. To investigate this phenomenon, it is necessary to dispose of high quality time and cross section data for a large numbers of variables. At a first sight, official statistics supply a plenty of information on the themes of interests (levels of the impact pathway from agricultural research expenditure to multiple sustainability dimensions). However, quality deficiencies due to missing values, outliers, short time series and breaks in the series may considerably affect the reliability of statistical analysis.

In this paper, we propose some quality indicators, which, if provided, would help users to understand the actual usability of data. We compute such indicators on a subset of variables representative of each levels of the research impact pathway from agricultural research expenditure to multiple sustainability dimensions. These measures take into account qualitative information on data quality published by Eurostat (indicator I1) as well as quantitative quality measures computed on data (indicator I2). By comparing the values of quality indicators I1 and I2, we find contrasting indications: quality measures stemming from metadata seem not to be correlated to quality measures based on data. This result suggests that both qualitative and quantitative measures are necessary to assess the overall quality of available statistics.

REFERENCES

Bartolini, F., G. Brunori, A. Coli, C. Landi, and B. Pacini (2014). La Valutazione dell'Impatto della Spesa per Ricerca e Sviluppo in Agricoltura sulla Sostenibilità: Un'Analisi delle Principali Problematiche Metodologiche ed Empiriche. *Agriregionieuropa*, 38: 65-68.

Bartolini F., G. Brunori, A. Coli, A. Magrini, and B. Pacini. (2016). Assessment of Multiple Effects of Research at country level, Deliverable 4.2 Impresa project (EU 7th Framework Programme).

Brys, G., M. Hubert, and A. Struyf (2004). A robust measure of skewness. *Journal of Computational and Graphical Statistics*, 13:996–1017, 2004.

Brys G., M. Hubert, and P. J. Rousseeuw (2005). A Robustification of Independent Component Analysis. *Journal of Chemometrics*, 19, 1–12.

Campbell, D., Caruso, J. and Archambault, E. (2013). Cross-cutting Analysis, European Commission report.

OECD (2015). Expert Workshop: Measuring Environmentally Adjusted Total Factor Productivity for Agriculture. http://www.oecd.org/tad/events/environmentally-adjusted-total-factor-productivity-in-agriculture.htm

Thirtle, C., Piesse, J., and Schimmelpfennig, D. (2008). Modelling the Length and Shape of the R&D Lag: An Application to UK Agricultural Productivity. *Agricultural Economics*, 39, 73–85.