DATA QUALITY IN THE AGRI-ECONOMIC RESEARCH IN THE REPUBLIC OF MACEDONIA

Kotevska A.*, Martinovska Stojcheska A., Dimitievski D., Janeska Stamenkovska I.

Ss. Cyril and Methodius University in Skopje, Faculty of Agricultural Sciences and Food-Skopje, Republic of Macedonia

*Corresponding author: ana.kotevska@fznh.ukim.edu.mk

Abstract

The transition process in the South-East European countries opened new research topics and encouraged use of new methodologies in agricultural economic studies. The process of transformation of the agri-food sector towards market economy principles required introduction of complex micro and macroeconomic models to evaluate the alternative solutions to the ongoing challenges. The models' relevance and their usefulness depend on the availability and quality of input data. Data unavailability and big differences among databases make data sets unreliable and require adjustments to make them suitable for use. As an EU candidate country since 2005, the Republic of Macedonia conducts institutional and systemic harmonization, including the national statistics. The aim of this paper is to analyse the role of data and their quality in the agricultural economics research in Republic of Macedonia, through a review of the available literature and experiences of the stakeholders (based on users' survey) in the last decade. The analysis uses the attribute-based approach (Wang, Reddy and Kon, 1995), defining data quality as multi-dimensional and hierarchical concept. The paper underlines the role of data availability and their quality for the needs of evidence-based policy decision making; and the need for more active involvement of the agricultural economics science and practice in the creation, harmonization, quality assurance and sharing of data necessary for research.

Keywords: agricultural economics, attribute-based approach, data quality.

Introduction

The transition process in the South-East European countries, along with the economic reforms, opened new research topics and encouraged use of new approaches in agri-economic studies. The process of transformation of the agri-food sector towards market economy required contemporary approach in monitoring the development of the sector and providing projections to support the policy decision process. During the early 2000s, countries that obtained EU candidate status applied models to compare the results among each other and with some of the EU countries. At the same time, these experiences show data

unavailability and extreme differences between databases, which make them unreliable and require adjustments to make data suitable for use (Macours and Swinnen, 1997). The availability and quality of input data strongly affect the usefulness of complex micro and macroeconomic models. To obtain more realistic projections, they often utilize mathematical or econometric methods for which longer data series are needed. To compare with other countries, data need to be harmonized by using the same definition and methodology of data collection and processing. Otherwise, as Wang et al. (1995, p.349) point out "inaccurate, out-of-date, or incomplete data can have significant impacts both socially and economically".

Data are considered to be reflections of reality, whereas data quality (DQ) is their "fitness for use" (Tayi and Ballou, 1998), and cannot be assessed unrelated of data consumers (Strong et al., 1997). Data are being collected from multiple data sources and stored in different database formats. The usefulness and usability transform data into information, a valuable organizational, policy or research resource.

As an EU candidate country since 2005, the Republic of Macedonia passes through a process of institutional and systemic harmonization, including the national statistics. The changes in the country to large extent affected data providers' attitude for sharing their data. Having all this in mind, the aim of this paper is to analyze the quality of data used in agri-economic research in Macedonia in the last decade (2007-2017). The paper is based on the available literature review and experts' experiences. It analyses data quality from three groups of sources: primary data (from own survey), secondary data from State Statistical Office (SSO) and secondary data from other sources. The paper uses the attribute-based approach (Wang et al., 1995) defining data quality (DQ) as a multi-dimensional and hierarchical concept, analyzing it from data user perspective.

After the introductory part, the next section describes the material and method used. Following the presentation of results, the discussion is organized by source group (primary data, statistical data and other data sources), and the conclusion is given in the end.

Material and methods

This research employed both qualitative and quantitative data collection and analysis techniques. The paper is mainly based on expert experiences collected through an online survey. The researchers included in the survey are selected based on the list of authors that have published contributions regarding Macedonian agriculture in the main agricultural economics related publications in the country during the last decade (2007-2017). In addition, we contacted representatives from the analytical units in the Ministry of Agriculture, Forestry and Water Economy (MAFWE) and the Agency for Financial Support of Agriculture and Rural Development (AFSARD), as well as few non-governmental organizations working in agricultural research. In total, 41 individuals were contacted, out of which 25

responded (response rate of 61%). The survey was conducted online in September 2017.

The survey emphasizes users' perception of quality of data from three groups of sources: primary data (own surveys), secondary data from SSO and secondary data from other sources. Each DQ dimension was assessed on a 10-point scale (1 being the worst grade, and 10 being the best grade).

According to the attribute-based approach (Wang et al., 1995), data quality is a multi-dimensional concept based on data characteristics defined in four categories: intrinsic, accessibility, contextual and representational DQ. The intrinsic DQ is defined with data accuracy, objectivity, believability and reputation; accessibility aspect includes access, but also access security; contextual DQ is determined with data relevancy, value-added, timeliness, completeness, and amount of data; whereas representational DQ with data interpretability, ease of understanding, conciseness and consistency (Strong et al., 1997). This paper focuses on the most important DQ dimensions, such as: accuracy, objectivity, believability, accessibility, understandability, conciseness and consistency (as the identified by Ballou et al., 1985 in Wang et al., 1995).

Results

In this part, we present the assessment of the selected DQ dimensions and the most frequent sources of primary and secondary data. In addition, we present the main data sources in FASF papers published in the last decade, and in the researches that have been implemented continuously during several years.

Primary data

Table 1 presents the results from the online survey relevant to the primary data. Respondents evaluated primary data as quite used (mean score 6.92.), but with a lower level of accessibility (4.32) and cost acceptability (4.52).

Table 1. Assessment of primary data

Level of perception of primary data related issues	Mean score*
- Research based on primary data	6.92
- Easy access to primary data	4.32
- Cost acceptability of collecting primary data	4.52
- Readiness to share own databases with other researchers	7.60

Source: Own survey; Note: *Scale 1-10: 1-not at all; 10-fully.

Individual farmers are identified as the main source of primary data (72%), whereas the agricultural enterprises and governmental institutions are used by 24% of all respondents, respectively (Table 2).

From the pool of responses, the most frequently data collection technique is face-to-face survey or interview, personally from the researcher (36%) or through other surveyors (31%). The respondents' explanations for selection of this technique are

multiple: it allows direct contact with farmers hence enabling quality control and high level of accuracy. As common challenges in the process the respondents identify the lack of trust, lack of finances, difficulty getting full and accurate answers, as well as difficulties in sampling in terms of size, randomness or region. Other data collection techniques are less represented: online surveys (10%), focus groups (10%) and observation (7%), whereas data collection by post or telephone and experiments are rarely used.

Table 2. Most frequent source of primary data

Most frequent source of primary data	Respondents	Percent frequency*
- Data from individual producers	18	60%
- Data from agricultural enterprises	6	20%
- Data from governmental institutions	6	20%

Source: Own survey; Note: *Multiple answers possible.

Secondary (statistical and other) data

Due to the complexity of the agricultural sector, the agricultural economic research often requires use of data from various sources (Table 3). Although SSO is the major source of data (46%), it is often complemented with data from other institutions, such as MAFWE, AFSARD, others' reports and studies, as well as other not listed sources, such as National Extension Agency (NEA), Ministry of Finance, Central Register, etc, or data from tailor-made surveys as part of own research projects.

Table 3. Most frequent source of secondary data

Source	Respondents	Percent frequency*
State Statisticial Office	23	43%
MAFWE	9	17%
AFSARD	7	13%
Others' reports and studies	10	19%
Other	4	8%

Source: Own survey; n= 25; *Multiple answers possible

Table 4 describes the perception of use and the DQ dimensions of secondary data, separately the official statistical data from SSO and other data sources.

Data from SSO were perceived as the most used (mean score of 7.29). The analysed DQ dimensions were generally well assessed, especially the understandability, accessibility and consistency dimensions (7.63, 7.50 and 7.17, respectively). Timeliness (6.54) and completeness (6.04), as a measure of contextual DQ, and accuracy, objectivity, believability, as measures of intrinsic

DQ, are also well assessed (6.04); whereas comparability with other sources was estimated to be lowest (4.46).
Data from other sources were generally less used (6.17) than the official statistical and primary data. Accordingly, their DQ dimensions are evaluated lower. Accessibility (4.21) and representational DQ, measured through consistency (4.75) and understandability (5.75), are the least well assessed. The contextual DQ, measured with completeness (5.29) and timeliness (5.58), and the intrinsic DQ, i.e. accuracy, objectivity and believability (6.13), are assessed better, but still lower when compared to the assessment of data from SSO.

Table 4. Assessment of secondary data (mean score)

	SSO	Other
Level of use data from SSO/Other	7.29	6.17
Easy access	7.50	4.21
Accuracy, objectivity, believability	6.04	6.13
Timeliness	6.54	5.58
Completeness (sufficient coverage)	6.04	5.29
Consistency (used same format, comparable with previous data)	7.17	4.75
Understanding (clearly defined and easy to understand)	7.63	5.75
Comparability (availability of same/similar data to be checked)	4.46	-

Source: Own survey; Note: Scale 1-10 (1-not at all; 10-fully), n=25

Table 5 summarizes part of the survey that focused on the occurrence of continuous research. Half of the respondents have conducted the same research continuously in few consecutive years. Most of those researches encompassed period up to five years, but there are researches which maintained longer time series. Primary data were the most frequent dominant data source in those researches, but also SSO, NEA, AFSARD and Customs office were used as well. Almost all of these researches have built a database from these researches.

Table 5. Research implemented continuously in several years

Total number	13	52%
Period		
- up to 5 years	5	38%
- 5-10 years	2	15%
- more than 10 years	3	23%
Used data in these research		
- Primary data	6	46%
- SSO	4	31%
- Other secondary data	3	23%
Built database from these research	11	85%

Source: Own survey

In addition to the survey, we reviewed 204 scientific and applicative agricultural economics related papers published by the FASF staff in the period 2007 to 2017 (Table 6). Out of them, around one-third address theoretical issues, and the rest use quantitative data to fulfil the aims of the respective researches. Looking deeper at the latter, most papers use field surveys collecting primary data directly from farmers (46%); about 31% use data from the SSO as a main source, 17% use raw or processed data from other sources (such as Central Register, Customs, MAFWE, NEA Farm Monitoring System, studies, etc.), and the remaining 6% of the papers use a combination of various primary and secondary data.

Table 6. Data sources in FASF staff papers 2007-2017

Dominant source of data in paper	**No.**	**Share**	
Theoretical or review paper, no quantitative data	60	-	29.4%
Primary data, own field research	66	45.8%	32.4%
State statistical Office (SSO)	44	30.6%	21.6%
Other secondary data	28	19.4%	13.7%
Combination	6	4.2%	2.9%
Subtotal: Papers with quantitative data	144	100%	
Total: All papers	204		100%

Source: Own survey

Discussion

Before we discuss the results from this research, we briefly discuss the general concerns of the DQ dimensions, and then we present the observed qualities and concerns of the DQ dimensions of the primary data, statistical data and other data sources in the Republic of Macedonia, separately.

Multiple sources of the same data, such as primary data, statistical data and other data sources, cause concerns regarding the intrinsic DQ (accuracy, believability, objectivity, and reputation). If the accuracy of different sources cannot be confirmed, conflicting data develop into a believability problem. For instance, the SSO and MAFWE both publish different data on the same issue. In addition, subjectivity in the data production process appears as a concern about data objectivity; thus, interpreted data is considered to be of lower quality than raw data. This is often case with secondary data from other sources (studies and reports). Over time, accumulated believability or objectivity problems develop in a poor reputation of the data producer. All this is viewed as having little added value to the data users and results in reduced use of this data (Strong et al., 1997). We have not identified such case, but it is a risk worth paying attention.

Accessibility DQ is often related to technical accessibility (available connections, granted access permission, and installed access methods). Although data users are aware of the confidential nature of some data and realize the importance of access security for individual records, still the data operators and permissions are

perceived as barriers to accessibility because they could not provide data access without approval (such examples are the disaggregated data from FADN or on budgetary payments.

The broad accessibility of data users exceeds technical accessibility. It also includes the ease to use the data (the ease of access and the ease to understand data). Any access barrier that hardens the use (such as the use of different definition, measures or representation of a similar data, or the need of excessive access time because of the data volume) is perceived as accessibility problem. Therefore, the representational DQ dimensions appears to be an underlying cause of some accessibility DQ problems (*ibid)*. Other representational concerns that become a barrier to data accessibility are data for which specific expertise is required to interpret; data that is difficult to be analyzed across time due to changing sample or format (such as the agricultural policy measures), or a large amount of data that require time to to access, leading to timeliness problem.

The contextual DQ is different to measure, because data users evaluate it relative to their tasks. The common contextual DQ problems that occur is the incomplete data, inadequately defined or measured data, and data that could not be appropriately aggregated (*ibid*).

Primary data

Considering the farmers' mentality in Balkan countries and the reluctance to provide data (due to general distrust), the most viable way to conduct a survey and collect micro-economic data is by direct (face-to-face) interview through people who they know and who they trust, most often extension agents (Kotevska and Martinovska Stojcheska, 2015). Although it is more time-consuming and relatively more expensive, it ensures the necessary number of the respondents and lowers the occurrence of missing values (*ibid*).

This approach to some extent influences the intrinsic dimension of DQ. The process of full randomization of the respondents is affected by the lack of full and accurate list of farmers (with contact details), but also from external factors. For instance, due to an unstable political situation, the generally low level of trust in the country were even stronger and farmers were very reluctant to communicate with interviewers they did not know (Huber et al., 2016).

Intrinsic DQ can also be affected by the language used in the survey, since translation in some cases cannot fully grasp the spirit of the local languages, which affects the respondents' understanding and presumably pushes them to claim a more neutral position (Kotevska and Martinovska Stojcheska, 2015). This is a very important issue to be considered when translating questionnaires from other languages (such as English), when translating to the other official languages in the country, but also when adopting the vocabulary understandable to both the interviewer and the respondents.

As already explained in other context, the accessibility of primary data is limited, the most evident reason being farmers' inaccessibility due to their lack of trust, but

also due to high collection costs to cover larger sample and follow the principles of randomization.

Regarding the contextual and representational DQ, primary data collection as a source of data mostly depends on the researcher and the methodological approach. The data use is limited to the collectors; they are not publicly shared, but often their utilization is maximized through several analyses from different research perspectives. Surprisingly, the respondents in the online survey stated favourably to share the databases with other researchers (mean score of 7.6, Table 1).

The organization and presentation of the primary collected data depends on the few individuals responsible for them. Still, an advantage is the timeliness, and relevance, but it can be a problem to obtain longer data series with certain level of consistency. FASF's Institute of Agricultural Economics has a positive experience of providing long data series with high level of consistency. Such reliable and continuous series of cost of production budgets were produced in the period 1967-1985, for a dozen of more important products. Efforts to maintain such budgets were revisited on several occasions (Martinovska Stojcheska et al., 2010). A system of cost of production budgets that would be continuously maintained is not acknowledged enough by policy makers, although it would be highly beneficial to them, but also to researchers and to other stakeholders, such as banks, advisors etc.

Statistical data

The State Statistical Office (SSO) has an established methodology and infrastructure for data collection. However, during the process of transition, changes in the institutions lead to changes in the methodology as well, which reflected the length of comparable time series (since 1995). This caused some difficulties in the studies conducted in the past, but now, 20 years later, there is a consistency in the data formats.

In the process towards the EU accession, the methodology, data organization and data presentation gradually harmonized with that of Eurostat. Several joint projects of the Western Balkan countries initiated the comparison of the agricultural sector and the relevant policies among the countries and at some level with the EU (Volk, 2010; Volk et al., 2014). These projects observed the effect of the EU approximation process on the national statistics development. In regards to the harmonization of the agricultural statistics with the EU, Macedonia had highest progress comparative to the other countries in the region (*ibid*).

The accessibility dimension of the SSO improved over time, through the increased number of publications and the MAKSTAT database available online, though there is possibility to expand the data scope and time series length.

To increase data application a further improvement of the remaining DQ dimensions is needed. For example, since recently, regional statistics are available separately for urban and rural areas. These data are quite broad and interesting; however, the definition of rural areas used in the methodology does not correspond to the definition in the Law of Agriculture and Rural Development (OG 49/2010),

according to which their development is being planned and monitored. This example shows the importance of relevancy, as one of the contextual DQ dimensions, and data interpretability, as one of the representational DQ dimensions, for the policy design and decision process.

Other data sources

The researcher is challenged to best evaluate the intrinsic DQ, if multiple sources of the same data are available. The intrinsic DQ, especially the accuracy dimension, can be increased by appropriate control techniques. For example, the accuracy of the FADN data is influenced from data collection and data entry. It is recognized the importance of DQ controls in all stages: from farmer statements via data collectors and data processors to data transmitters to ensure its accuracy and objectivity (Martinovska Stojcheska et al., 2010).

The accessibility dimension of the other data sources is often limited by many administrative barriers introduced by the corresponding institution. For example, there is an administrative barrier to use disaggregated FADN data for research purposes. MAFWE (as the Liaison Office) provides only aggregated FADN data (for 600 farms covered in the survey on an annual bases), which could not always be of use for some research purposes where disaggregated data are needed (Dimitrievski et al., 2017).

The accessibility of data from some other secondary sources can be limited due to extremely high costs for their use. For instance, data on agricultural enterprises from Central register are expensive to acquire, even for research purposes.

A representational DQ problem is observed in the MAFWE registers that are different in format and data entries, thus hindering their further aggregation and use in different policy analyses. The inconsistent definitions, measures and data representations are mainly caused by autonomous design decisions in each division. The need for a common data warehouse with common data definitions and representations for cross-divisional data use has been identified recently and an actual project in MAFWE is trying to solve it.

Expert opinion can be also used as another source of data. It is often used to resolve problems in the presence of gaps in the current data or future uncertainty (Kotevska, 2013), that makes it a valuable input in the research or decision making process. Still, one can pose questions suspecting their accuracy and believability. The attempts to lower the level of subjectivity in building assumptions are often difficult, especially when a smaller group of experts are consulted and when their scientific views are weighted by the individual analyst.

Conclusions

This paper emphasizes the role of data quality in the agri-economic studies in Republic of Macedonia. It deepens our understanding of researchers' experience with DQ dimensions and their common DQ concerns. It depicts the most frequent obstacles or advantages in obtaining relevant data from primary and secondary

sources, such as data collection approaches, randomization, data collection challenges, the level of harmonization with other national and international statistics, comparability among years and with other data sources, availability of specific data formats, as well as institutional cooperation in obtaining data. The paper underlines the need for more active involvement of the agri-economic science and practice in the creation, harmonization, quality monitoring and sharing of data necessary for research and analysis. The results of this research may be used as an empirical basis for improving the DQ management in any institutional data collection, storage, and presentation.

References

Huber, M., Kotevska, A., Martinovska-Stojcheska, A. and Solovyeva, A. 2016. Evaluating an information campaign about rural development policies in (FYR) Macedonia. Working papers SES, N. 469, V.2016. Université de Fribourg.

Kotevska, A., Dimitrievski, D., and Erjavec, E. 2013. Partial equilibrium models as a tools for policy analysis in agriculture: An empirical evidence of Macedonia. Proceedings of Seminar: Agriculture and Rural Development - Challenges of Transition and Integration Processes. Belgrade- Zemun, September 2013, pg. 52-65.

Kotevska, A., and Martinovska Stojceska, A. (ed). 2015. Impact of socio-economic characteristics of rural population on success of the rural development policy, Skopje: Association of Agricultural Economists of the Republic of Macedonia

Macours, K. and Swinnen, J.F.M. 2000. "Causes of Output Decline in Economic Transition: The Case of Central and Eastern European Agriculture". Journal of Comparative Economics, 28(1), pp. 172-206.

Martinovska Stojcheska, A., Dimitrievski, D., and Georgiev, N. 2010. The Farm Accountancy Data Network in the Republic of Macedonia with emphasis on data quality. In: Peševski, M., Gjoševski, D., Georgiev, N. (Eds.) Agriculture in the process of adjustment to the common agricultural policy, University Ss Cyril and Methodius, Faculty of Agricultural Sciences and Food, Skopje, 137-150.

Dimitrievski, D., Martinovska Stojcheska, A., Janeska Stamenkovska, I., and Kotevska, A. 2017. Farm Productivity in the Western Balkan Countries: Technical efficiency analysis of Macedonian vegetable farms (study report). EC JRC IPTS Project: Analysis of agricultural and rural development policies in Western Balkan countries. p.50.

OG (Official Gazette of the Republic of Macedonia), 49/2010. Law on Agriculture and rural development.

Strong, D. M., Lee, Y. W., and Wang, R. Y. 1997. Data quality in context. Communications of the ACM, 40(5), 103-110.

Tayi, G. K., and Ballou, D. P. 1998. Examining data quality. Communications of the ACM, 41(2), 54-57.

Volk, T. (ed). 2010. Agriculture in the Western Balkan Countries. Halle (Salle): IAMO, 57.

Volk, T., Erjavec, E., and Mortensen, K. (Eds.). 2014. Agricultural policy and European integration in Southeastern Europe Budapest: Food and Agriculture Organization of the United Nations. (121–134).
Wang, R. Y., Reddy, M. P., and Kon, H. B. 1995.Toward quality data: An attribute-based approach. Decision support systems, 13(3-4), 349-372.