STATISTICAL EDUCATION AND OFFICIAL STATISTICS – TRAINING FUTURE DATA SCIENTISTS

Maria Eduarda Silva[1] and Pedro Campos[1,2]
[1]*University of Porto, Faculty of Economics, R. Dr. Roberto Frias, Porto, Portugal*
[2]*Statistics Portugal, Av. António José de Almeida, 1, Lisboa, Portugal*
pcampos@fep.up.pt

EMOS (The European Master in Official Statistics) was set up to strengthen the collaboration within academia and producers of official statistics and help develop professionals able to work with European official data at different levels in the fast-changing production system of the 21st century. In this paper we address the need for training in Official Statistics, particularly in current times, where new skill sets and competencies are necessary. In particular, the needs for new data sources currently used by national statistical systems require the development of new methodologies. For that purpose, we do a matching between National Statistical Offices (NSO) needs and the offer from universities.

1. INTRODUCTION

EMOS (The European Master in Official Statistics) is a joint project of universities and data producers in Europe. It was set up to strengthen the collaboration within academia and producers of official statistics and help develop professionals able to work with European official data at different levels in the fast-changing production system of the 21st century. After several years, and two calls for interest, the network comprises now 30 programmes in some 20 countries.

Academia is aware of the relevance of official statistics at a global level. Furthermore, machine learning and algorithms for Big Data are more and more important for innovation in official statistics. At University of Porto Faculty of Economics (FEP.UP) it is acknowledged that future professionals need to understand the relevance of official statistics, use different kinds of data sources, and methodologies, and need to be able to present data in an effective way. Therefore, recently, University of Porto joined the EMOS network, through the creation of a new track in the Master in Data Analytics.

In this paper we address the need for training in Official Statistics, particularly in current times, where new sets of skills and competencies are necessary. In particular, the need for new data sources currently used by national statistical systems require the development of new methodologies. For that purpose, we do a matching between National Statistical Offices (NSO) needs and the offer from universities, which are based on the learning outcomes. These include programming skills, new production models, and various statistical and data mining techniques, specifically targeted for the extraction of actionable knowledge from large volumes of existing data. Learning outcomes are usually defined by Eurostat (2021), and are aligned with the goals of the European Statistical System (ESS):

1. System of Official Statistics
2. Production model and methods (quality issues, sources, architectures)
3. Specific Themes (economic, Finance, Population, Environment, Energy...)
4. Statistical Methods (sampling, index numbers, ...)
5. Dissemination

In: R Helenius, E Falck (Eds.), *Statistics Education in the Era of Data Science*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
Aug-Sept 2021, Online conference. ©2021 ISI/IASE iase-web.org/Conference_Proceedings.php

2. A MATCHING BETWEEN NATIONAL STATSTICAL OFFICES' DEMAND AND UNIVERSITIES' OFFER

It is important to establish a relationship between the needs of NSO and the offer of universities in terms of training future data scientists. We identified the most important areas in which training is needed: Administrative sources, Big Data and Statistical Confidentiality.

2.1. Administrative Data

Administrative data is data reported to administrative authorities by individual persons or legal entities for legal compliance or to access government services. It can be also data recording decisions made by administrative authorities and data generated by administrative authorities to support planning, implementation, monitoring and evaluation of administrative programmes (United Nations Statistical Division, 2021)

The use of administrative data has several advantages for the production of official statistics. Administrative sources are an alternative to primary data collection that allows for the reduction of costs, and decreases the response burden. In addition, administrative data improves data coverage and availability [14]. Often, the results of sample surveys pertaining to population sub-groups cannot be presented due to unacceptably large sampling error. So, we need to improve the quality and efficiency across the entire statistical production system, using administrative data sources. Data from administrative sources needs to be adapted and processed to make it suitable for statistical compilation: transform administrative entities into statistical units and transform administrative variables into statistical variables.

2.2. Big Data

Statistical organizations regard Big Data as Data that is difficult to collect, store or process within the conventional systems of statistical organizations. Either, their volume, velocity, structure or variety requires the adoption of new statistical software processing techniques and/or IT infrastructure to enable cost-effective insights to be made. According to the HLG project on Big Data in Official Statistics (2014), there is a 'data-oriented' approach nowadays, where statistical organizations ask how they can make use of new sources such as:

- Energy Consumption statistics and trends
- Credit card and Consumer Loyalty Information
- Web Search Information
- Satellite and ground sensor data
- Mobile device location data

The problem with Big Data in official statistics is that it is difficult to collect, store or process within the conventional systems of statistical organizations (Hackl, 2014). However, there are a number of scenarios in which Big Data could be used in statistical organizations (HLG project on Big Data in Official Statistics, 2014):

Scenario 1: use as auxiliary information to improve all or part of an existing survey
Scenario 2: supplementing/replacing all or part of an existing survey with Big Data
Scenario 3: producing a predefined statistical output either with or without supplementation of survey data
Scenario 4: producing a statistical output guided by findings from the data (HLG project on Big Data in Official Statistics, 2014).

In order to assess the main potential of Big Data, the Generic Statistical Business Process Model (GSBPM v5.0) may be used.

2.3. Statistical Confidentiality

Statistical confidentiality is a fundamental principle of official statistics enshrined in the Treaty and in the European statistics Code of Practice. With the use of new potential sources for the production of Official Statistics, people are ever more concerned about the protection of individual data under the GPDR (General Data Protection Regulation). Harmonisation of principles and guidelines as regards protection of confidential data is the obligation of Eurostat and national statistical authorities in the European Statistical System.

Table 1 shows a cross tabulation between the required EMOS learning outcomes, and the ones we have identified above as the currently most important official statistical needs (Pratesi and Campos, 2021).

EMOS learning outcomes	Administrative Data	Big data/Smart Statistics	Privacy and Confidentiality
1. System of Official Statistics	√	√	
2. Production model and methods (quality issues, sources, architectures)	√	√	
3. Specific Themes (economic, Finance, Population, Environment, Energy...)	√	√	
4. Statistical Methods (sampling, index numbers, ...)	√	√	√
5. Dissemination	√	√	

Table 1 - Matching between the required EMOS learning outcomes, and the current important topics addressed in Official Statistics

Source: Pratesi and Campos (2021)

3. Discussion and conclusions

The business informatisation, the internet popularisation and the emergence of the big data phenomenon lead to non-conventional data sources which may contribute to official statistics. However, to make this contribution effective the next generation of statisticians must be able to master methods to analyse data characterized by Volume, Velocity, Variety. Moreover, the new statisticians must be familiar with methods that allow to check for Veracity and to extract Value from the data. In addition, it is important to know how to deal with new sources of data, namely administrative data, and the corresponding issues of metadata and quality. In a world with more and more data sources, it is increasingly important to ensure the protection of individual data. To this end, it is urgent to teach methods of Statistical Disclosure Control. These requirements point to the need of extending the skills of competencies of the statisticians into the realm of Data Science. This is precisely the aim of the EMOS track of the Maser in Data Analytics.

Eurostat (2021), Learning Outcomes of the EMOS programmes, accessed in July 2021, available at: https://ec.europa.eu/eurostat/cros/content/learning-outcomes-emos-programmes_en

Hackl, Peter. 'Big Data: What Can Official Statistics Expect?', Statistical Journal of the International Association of Official Statistics 1 Jan. 2016 : 43 – 52.

HLG project on Big Data in Official Statistics, (2014), "How big is Big Data? Exploring the role of Big Data in Official Statistics", draft version available at: https://statswiki.unece.org/pages/viewpage.action?pageId=99484307&preview=/99484307/99451129/Virtual%20Sprint%20Big%20Data%20paper.docx

Pratesi, M., Campos, P. (2021), EMOS reloaded: unlock the future of education in official statistics with a new partnership with Universities, Statistical Journal of the International Association of Official Statistics (to appear)

United Nations Statistical Division, (2021), Use of administrative data for official statistics: The Global Perspective, UN Statistics Division/DESA