Empowering and interacting with statistical produsers: a practical example with Eurostat data as a service

Jacopo Grazzini, Jean-Marc Museux and Martina Hahn[1] ({first}.{last}@ec.europa.eu)

Keywords: *data analytics, open algorithm, open data, reproducible computational workflow, interactive computing platform, statistical literacy, produser*

1. Context: running transparent data analytics in a post-truth society

Since policy advice is becoming increasingly supported by data resources [1], public organisations are leveraging the use of these resources to inform decisions. In this context, supporting *data analytics*, which aims eventually at extracting valuable information from data to use in intelligent ways by means of advanced statistical and computational techniques [2], offers big opportunities – and challenges – for enhanced insight and decision-making [3]. *Data analytics* draws not only on existing and new sources of ever-growing data, it builds also on new methodologies and emerging Information and Communication Technologies tools, and advances thanks to innovative initiatives. In particular, with increased availability of *open data*, new developments in open technologies and recent breakthroughs in *data science*[2], it is believed that *data analytics* can help improve current governance processes by enabling policy-driven data-informed evidence-based decision-making and potentially reduce the bias, costs and risks of policy decisions [4][5]. Eventually, it seems very logical, and appealing, to deploy complement the statistical offices' toolbox with data analytics tools. Still, the development and deployment of such tools will need sound judgment as an abundance of data and computing power does not automatically guarantee good decision-making.

Nowadays political decisions are expected to be accompanied by the access to the data analysed, the detailed information about the sources, the underlying assumptions (models and methods) and also the tools (software) used to support the decision[3] [4]. At the time where the citizens' demands for more transparency in the EU institutions are growing, it also underpins the movement towards not only more open, transparent and defensible [5], but also more participative decision-making systems. In the context of a "post-truth" society, *data analytics* presents substantial promises for E-government, openness and transparency, and the interaction between governments and *produsers*, *e.g.* statisticians, scientists and also citizens [1]. Additionally, it is also necessary to create a framework to provide *produsers* with the ability to both perform the analysis and repeat it with different hypotheses, parameters, or data, hence translating questions that are asked into a series of well-understood computational methods [3][6].

While the importance of openness and transparency in statistical processes [6][7], and how these can be supported through *open algorithms* and *open data* [4][5], has been

[1] Methodology and innovation in official statistics Unit, Eurostat (DG ESTAT) – European Commission.

[2] Without discussing the conceptual difference(s) – whether or not they diverge by scope, purpose or focus – between *data analytics*, *data science*, *data mining*, and *knowledge discovery* (or even…*statistics* [2]), we simply admit here they all combine logical reasoning, mathematics, statistics, programming, and computer science, and often involve methods at the intersection of signal processing, pattern recognition, machine learning, artificial intelligence, simulation and optimisation. They also generally imply the extensive use of data and require the capacity to collect, prepare, link and visualise the data.

[3] Ultimately, the purpose is to build public trust in the data, the models and the tools.

already emphasized, this contribution aims at showcasing an approach similar to [8] where algorithms and data are delivered as interactive, reusable and reproducible computing services. This will eventually provide *produsers* with the necessary tools to perform, for themselves, *data analytics* on Eurostat data in a straightforward manner.

2. OPENING AND SHARING DATA AND ALGORITHMS MAY NOT BE ENOUGH

It is expected that a greater openness in designing production processes will result in a better grip on the complete transparency, as well as the overall quality, of the statistical processes involved in decision-making [4]. However, the sole dissemination of comprehensive models and methods in manuscripts, guidelines, wikis, *etc*… should not be the end of the road towards more transparency in official statistics. Beyond increasing efficiency of statistical processes and timeliness of products, *open algorithms*, together with *open data*, can, in this aspect, enable to track the totality of the decision-making process as well as its progress [5]. Still, the requirement of openness and transparency is not automatically equivalent to publicising or accessing data and algorithm. Even when code and data are shared, variability in computing environments, operating systems or software versions used during the original analysis make it difficult to reproduce results.

In addition, the trend towards algorithmic decision-making and automated data processing techniques – the role of models being overwhelmed by the data and the algorithms to analyse them – raises methodological and empirical difficulties which can lead to important biases or issues that may be hard to identify [9]. Algorithms can indeed underpin fully automated or semi-automated decision-making. For instance, decision-making criteria can be programmed-in through relatively simple instructions, or can be learned by the algorithm through machine learning approaches and analytics techniques [3][9]. Yet, a general implication of the basic functioning of machine learning – *e.g.*, through the design and training of a model on massive datasets – is that opening the algorithm may not be enough to understand and explain the decision. More meaningful solutions will be needed to ensure accountability and fairness and a clearer understanding of effective openness and transparency [9]. Following [8], we further highlight the importance of capturing and sharing data, algorithms and software as well as the computational components needed to "*generate the same results from the same inputs*".

3. EXPOSING DATA AND ALGORITHMS ON INTERACTIVE COMPUTING PLATFORMS

Prior to designing truly "*explainable statistics*" [9], and beyond just opening data and algorithms [5], it is already possible to provide the public with further insights into the workings of decision-making systems – may they be design-based, model-based, model-assisted or algorithm-based – by exposing fully reproducible (and reusable) computational statistics workflows[4]. Such workflows will not only enable *produsers* to fully reproduce experiments [6], but also allow them to "*judge for themselves if they agree with the analytical choices, possibly identify innocent mistakes and try other routes*" [7]. Today's technological solutions – *e.g.* flexible Application Programming Interface, lightweight virtualised container platforms, versatile interactive notebooks and code source control – make the development and deployment of reproducible statistical workflows easy by supporting an approach where data and algorithms are delivered as portable, scalable, standardised and encapsulated interactive computing *services*.

[4] In regard, see for instance M.Upson's blog on reproducible analytical pipelines.

First, the provision of computing notebooks offers the prospect of actively engaging ("empowering*"*) the public in the creation, implementation, testing and validation of statistical products[5]. Notebooks are both human-readable documents containing the analysis description and the results, as well as executable documents which can be run to perform *data analytics*. They offer a convenient tool for rerunning or tweaking previous data analyses. Notebooks foster statistical literacy by alternating, in practice, live code (in various programming languages) with narrative and explanatory text, in the literate programming paradigm [10]. Indeed, *(prod)users*, can benefit from storytelling since the code is exposed together with the narrative following its structure, in the order demanded by the logic and flow of thoughts. This further supports the more generic concept of literate computing, in which the interactive exploratory computing is captured along with its motivations and results [8]. Last, *(prod)users* can share their analysis to iterate on it.

Then, to further improve reproducibility, it is possible to use virtualised containers that allow computational processes to be run the exact same way in any environment [8]. Since it is common to use one or more software libraries within a given statistical process, using these libraries creates a dependency. Besides, the underlying infrastructure used during development and deployment, *e.g.*, operating system, installed tools and libraries, *etc…*, create further constraints. Containers can reduce these constraints, while maintaining the dependencies, by wrapping software into a minimal virtual machine with a predefined computing environment that includes everything the processes need to run.

4. A PRACTICAL USE CASE: ACCESSING EUROSTAT GISCO DATA AND SERVICES

Marek is working on a project to identify local events over Europe: indeed, he is trying to create a map of local events where cities are classified by regional NUTS2 code. Though he could do this manually using the legal acts, *e.g.* based on the online NUTS code *vs.* name classification, Marek would rather automate the process. He is aware of Eurostat geospatial web-services, *e.g.*, geocoding tools, NUTS identifier, provided by GISCO, and knows some scripting. Thus, he believes it is possible to automatically retrieve the NUTS code, at any level, of a given geolocation based on its name or geographical coordinates.

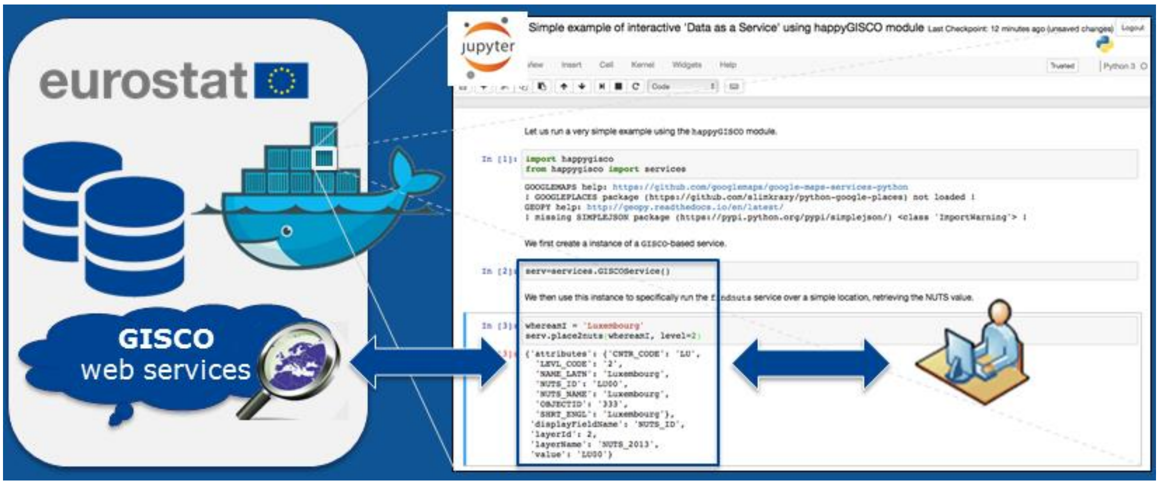
To solve his problem, Marek will be able to use the *happyGISCO* module provided in a virtualised container and embedded in a computing interface on top of GISCO web-services, like in this example. In practice, *happyGISCO* is built upon the *Jupyter Notebook Data Science Stack* which combines *docker* virtualised container and *Jupyter* notebook to provide with a portable environment with interactive computing tools and ("*agnostically"*) and support different programming languages, namely *R* and *Python* (see *Figure* next page).

5. CONCLUSION: CURRENT STATUS AND FUTURE PLANS

We propose herein a new form of participatory knowledge production that can contribute to new models of interaction between authorities and *produsers* to engage in *data analytics* for decision-making. In doing so, this approach foster statistical literacy and can provide with a way to understand "*more epidemiology on how people collect, manipulate, analyse, communicate and consume data"* [7]. We also emphasize the need for computational statistics to exceed the limits of the statistical offices so as to empower *produsers*, since perfectly configured and ready-to-use computing environments can be

[5] Of course, one could argue that notebooks may not be the right tool for codifying production pipelines and producing official statistics. However they offer many other advantages, not discussed herein.

distributed with any newly published official statistics. In the future, we aim at adopting a similar design for accessing all Eurostat datasets from its online open database.



Simple example of interactive 'Data as a Service' using happyGISCO module Last Checkpoint: 12 minutes ago (unsaved changes)

Let us run a very simple example using the happyGISCO module.

```
In [1]: import happygisco
        from happygisco import services

        GOOGLEMAPS help: https://github.com/googlemaps/google-maps-services-python
        ! GOOGLEPLACES package (https://github.com/slimkrazy/python-google-places) not loaded !
        GEOPY help: http://geopy.readthedocs.io/en/latest/
        ! missing SIMPLEJSON package (https://pypi.python.org/pypi/simplejson/) <class 'ImportWarning'> !
```

We first create a instance of a GISCO-based service.

```
In [2]: serv=services.GISCOService()
```

We then use this instance to specifically run the findnuts service over a simple location, retrieving the NUTS value.

```
In [3]: whereamI = 'Luxembourg'
        serv.place2nuts(whereamI, level=2)

   [3]: {'attributes': {'CNTR_CODE': 'LU',
          'LEVL_CODE': '2',
          'NAME_LATN': 'Luxembourg',
          'NUTS_ID': 'LU00',
          'NUTS_NAME': 'Luxembourg',
          'OBJECTID': '333',
          'SHRT_ENGL': 'Luxembourg'},
         'displayFieldName': 'NUTS_ID',
         'layerId': 2,
         'layerName': 'NUTS_2013',
         'value': 'LU00'}
```

Figure – A produser can run a container with an interactive Jupyter notebook that is already configured with the ready-to-use Python happyGISCO module and can query Eurostat GISCO web-services. *happyGISCO module is available here (documentation and examples). Ultimately, the container/notebook should be distributed as a web-service (instead of a downloadable container to install locally); in this perspective, docker will help tackle the problem of providing data as a service in a scalable format.*

References

[1] Höchtl, J. *et al.* (2016): Big data in the policy cycle: Policy decision making in the digital era, *Journal of Organizational Computing and Electronic Commerce*, 26(1–2):147–169, doi:10.1080/10919392.2015.1125187.

[2] Donoho D. (2017): 50 Years of Data Science, *Journal of Computational and Graphical Statistics*, 26(4):745–766, doi:10.1080/10618600.2017.1384734.

[3] H. Hassani *et al.* (2014): Data mining and official statistics: The past, the present and the future, *Big Data*, doi:10.1089/big.2013.0038.

[4] J. Grazzini and F. Pantisano (2015): Guidelines for scientific evidence provision for policy support based on Big Data and open technologies, *Publications Office of the European Union*, doi:10.2788/329540.

[5] J. Grazzini *et al.* (2018): *"Show me your code, and then I will trust your figures"*: towards software-agnostic open algorithms in statistical production, in Proc. *Quality* conference.

[6] V. Stodden (2014): The reproducible research movement in statistics, *Statistical Journal of the IAOS*, doi: 10.3233/SJI-140818.

[7] J. Leek *et al.* (2017): Five ways to fix statistics, *Nature*, 551:557–559, doi:10.1038/d41586-017-07522-z.

[8] B.K. Beaulieu-Jones and C.S. Greene (2017): Reproducibility of computational workflows is automated using continuous analysis, *Nature Biotechnology*, 35: 342–346, doi: 10.1038/nbt.3780.

[9] Guidotti, R. *et al.* (2018), A survey of methods for explaining black box models, arXiv:1802.01933.

[10] D. Knuth (1984): Literate Programming, *The Computer Journal*, 27(2): 97–111, doi:10.1093/comjnl/27.2.97.