

Eurostat Coding Labs





[Eurostat](#) is the statistical office of the European Union providing statistics at European level that enable comparisons between countries and regions. Eurostat is committed to fostering education of future official statisticians and developing new skills in the rapidly changing data environment.

Eurostat seeks to put together graduate students to work remotely on the projects described below. These projects will provide with a unique **learning opportunity for state-of-the-art programming techniques and statistical methodological developments**. It will be the occasion to collaborate with Eurostat staff and get further acquainted with the way *Official Statistics* are produced and disseminated.

Students involved in the project will receive a certificate of attendance. Depending on the local university rules, the project work may also be recognized officially as part of the study program, it may grant ECTS credits, or it may serve as basis for writing a Master thesis (with support by a university advisor). If you are interested in exploring the possibility of such arrangements, get in touch with Eurostat at the email contact below.

The current call proposes four projects of about 2-3 months under direct tutoring from Eurostat staff. **The work will be conducted remotely** mostly through the collaborative development platforms or informal visioconference/calls and via emails

Deadline for free-form applications is **30 June 2020**. Selection will be done by Eurostat project leaders based on CVs and possibly a short remote interview. All applicants will be informed of the result.

Topic	Skills	
Statistics explained through literate programming	Beginner to intermediate <ul style="list-style-type: none"> – Beginner to intermediate programming skills in R, Python or Javascript. – Prior knowledge of data processing and visualisation tools and modelling techniques, e.g. time-series forecasting models (preferred) 	
Gallery of Eurostat (meta)data visualisations	Intermediate to advanced (preferred) <ul style="list-style-type: none"> – Programming skills in R, Python or Javascript. – Prior knowledge of visualisation tools¹, in particular Javascript d3 framework (preferred). – Prior knowledge of interactive computing notebooks and dashboard technologies, e.g. Jupyter, R Markdown and voila 	
Semantic annotation of real world data through the eye of Official Statistics	Advanced <ul style="list-style-type: none"> – Advanced programming skills in R and Python (preferred). – Prior knowledge of interactive computing notebooks and dashboard technologies, e.g. Jupyter, R Markdown and voila. – Understanding of machine learning, artificial intelligence and deep learning. – Interest in computer vision and natural language processing. – Interest in semantic analysis, ontology matching and metadata management. 	
Processing of Mobile Network Operator data for spatial statistics	Intermediate to advanced <ul style="list-style-type: none"> – Intermediate to advanced skills in scientific programming with R or Python or Julia. – Prior knowledge of interactive computing notebooks and dashboard technologies would be an advantage. – Basic knowledge of GIS software and C/C++ programming language would be an advantage 	

¹ <https://python-graph-gallery.com/> - <https://www.r-graph-gallery.com/> - <https://www.d3-graph-gallery.com/>

1. *Statistics explained* through literate programming

1.1. *What you will do – Description of the project and objectives*

Communicating and interacting with the public is essential to improve perception and awareness of *Official Statistics*. Ultimately, dissemination shall also become an integral part of the whole statistical production workflow. Inspired by the reproducibility movement in [Open Science](#), and driven by the opening and sharing of all assets – making not only the data, but also the methods, tools, and software open – we promote an innovative bottom-up approach to dissemination that enables collaboration and increases participative forms of statistical services' design and sharing.



This project aims at creating computational narratives of how data are collected and processed to produce simple descriptive statistics. In practice, the [Statistics Explained](#) pages published on Eurostat websites will be reimplemented (and possibly extended with more complex analyses) into computational documents. Interactive and portable notebooks (such as [Jupyter](#) and [R Markdown](#)) that combine – in the so-called [literate programming](#) paradigm – explanatory text with executable code and simple visualisations will be used. They enable to dynamically reproduce the on-the-fly results typically displayed on the *Statistics Explained* pages through calling [Eurostat Application Programming Interface](#) (API). First examples (and templates) of what the project aims for can be explored on the [statistics-coded](#) page of [Eurostat github domain](#). The students will chose the pages they want to illustrate from their domain of interest.

1.2. *What you will learn – Outcomes and benefits*

- You will learn about data (and metadata) in *Official Statistics*, the way they are formatted, disseminated and shared.
- You will interact with *Official Statistics* through Eurostat API by using dedicated client packages²: learn to use the API, query, extract, load and transform data from Eurostat database.
- You will use computational notebooks to reproduce and verify the narratives based on *Official Statistics*, and possibly develop your own by extending existing analyses.
- You will improve your analytics skills from simple exploration of datasets to complex modelling of indicators.
- You will learn best practices from *Open Science* in terms of replicability and reproducibility, including versioning and testing of your code.
- **If successful, you will publish your results and reference your work on the statistics-coded page of Eurostat github domain.**

²<https://github.com/eurostat/restatapi>
<https://github.com/eurostat/pyrostat>.

<https://github.com/eurostat/eurostat.js>

1.3. What you will need – Desired/required knowledge and skills

- Beginner to intermediate programming skills in [R](#) and [Python](#) or [Javascript](#). **This project does welcome students with beginner skills.**
- Prior knowledge of data processing and visualisation tools and modelling techniques, e.g. time-series forecasting models (preferred).

1.4. How you will work – Organisation of the project

If selected, you will be working on a 3-4 months project under direct tutoring from Eurostat staff. The team will actually be guided by Jacopo GRAZZINI, PhD. Computer Science, and Matyas MESZAROS, PhD. Applied Economics.

The project work will be conducted remotely and there is no need for you to travel physically to Eurostat's premises. You will work and communicate mostly through the collaborative development platform, e.g. github. Technical questions, e.g. regarding code/programming, will be addressed through the ticket issuing facility of the platform. Interactions with the tutors, typically in the form of informal visioconference/calls and via emails, will be arranged flexibly depending on the project needs. Note that **the timeline of this project is actually flexible.**

Summary – Additional information

Duration/workload:	3-4 months/5 hours a week – 3 weeks equivalent full time at least. Investment can be light since it will vary depending on the number of pages the student will want to reproduce.
Period:	July-October (flexible)
Working method:	Remote teams interacting on github and through visio/calls, working language is English
Coding expertise:	Beginner to intermediate
Contact:	Jacopo Grazzini, email: ESTAT-Methodology@ec.europa.eu
Deadline for application:	30.06.2020

References:

- Grazzini J., Gaffuri J. and Museux J.-M. (2019): [Delivering Official Statistics as Do-It-Yourself services to foster producers' engagement with Eurostat open data](#), doi:[10.5281/zenodo.3240272](https://doi.org/10.5281/zenodo.3240272).
- Project Jupyter *et al.* (2018): [Binder 2.0 - Reproducible, interactive, sharable environments for science at scale](#), doi:[10.25080/Majora-4af1f417-011](https://doi.org/10.25080/Majora-4af1f417-011).
- Grazzini J., Museux J.-M. and Hahn M. (2018): [Empowering and interacting with statistical producers: A practical example with Eurostat data as a service](#), doi:[10.5281/zenodo.3240557](https://doi.org/10.5281/zenodo.3240557).

2. Gallery of Eurostat (meta)data visualisations

2.1. What you will do – Description of the project and objectives

Nowadays, many advanced technologies to present data in fancy and visually attractive ways are available. Animated and interactive graphs, mapping tools, charting components, *etc...* are now common techniques, and many examples of attractive and well-designed data visualisation are to be found on the internet in particular³. Many of these concepts can be used to visualise *Official Statistics*. Indeed, we think people expect *Official Statistics* to follow these trends so that they can use the same tools and concepts to consult *Official Statistics*.



This project aims at promoting the adoption of modern visualisation tools to expose and explore Eurostat data and metadata, and showcase them in a gallery of original (or not so original) examples. The objective is to build various dynamic examples of on-the-fly visualisations through the call to [Eurostat Application Programming Interface](#) or the fetching of [Eurostat metadata](#). Existing examples (from the abovementioned galleries or from the press) will be carefully discussed and selected by the students together with the tutors so as to be adapted to Eurostat (meta)data. First examples of what the project aims for can be explored on [Eurostat github domain](#)⁴.

2.2. What you will learn – Outcomes and benefits

- You will learn about data (and metadata) in *Official Statistics*, the way they are formatted, disseminated and shared.
- You will interact with *Official Statistics* through Eurostat API and its metadata repositories: learn to use the API, query, extract, load, transform and visualise metadata and data from Eurostat database.
- You will use computational notebooks to reproduce and verify the narratives based on *Official Statistics*, and possibly develop your own by extending existing analyses.
- You will improve your analytics skills from simple exploration of datasets to complex visualisation of indicators.
- **If successful, you will publish your results and reference your work on the d3-examples page of Eurostat github domain.**

2.3. What you will need – Desired/required knowledge and skills

- Intermediate to advanced (preferred) programming skills in [R](#) and [Python](#) or [Javascript](#).
- Prior knowledge of visualisation tools⁵, in particular [Javascript d3](#) framework (preferred).
- Prior knowledge of interactive computing notebooks and dashboard technologies, e.g. [Jupyter](#), [R Markdown](#) and [voila](#).

³ <https://observablehq.com/@d3/gallery> - <http://christopheviau.com/d3list/> - <https://voila-gallery.org/>

⁴ <https://github.com/eurostat/d3.examples> - <https://github.com/eurostat/d3.sunburst>

⁵ <https://python-graph-gallery.com/> - <https://www.r-graph-gallery.com/> - <https://www.d3-graph-gallery.com/>

2.4. How you will work – Organisation of the project

If selected, you will be working on a 3-4 months project under direct tutoring from Eurostat staff. The team will actually be guided by Jacopo GRAZZINI, PhD. Computer Science, and Julien GAFFURI, PhD. Geographical Information Sciences.

The project work will be conducted remotely and there is no need for you to travel physically to Eurostat's premises. You will work and communicate mostly through the collaborative development platform, *e.g.* github. Technical questions, *e.g.* regarding code/programming, will be addressed through the ticket issuing facility of the platform. Interactions with the tutors, typically in the form of informal visioconference/calls and via emails, will be arranged flexibly depending on the project needs. Note that **the timeline of this project is actually flexible**.

Summary – Additional information

Duration/workload:	3-4 months/5 hours a week – 3 weeks equivalent full time at least. Investment can be light since it will vary depending on the number of visualisations the student will want to produce.
Period:	July-October (flexible)
Working method:	Remote teams interacting on github and through visio/calls, working language is English
Coding expertise:	Intermediate to advanced
Contact:	Jacopo Grazzini, email: ESTAT-Methodology@ec.europa.eu
Deadline for application:	30.06.2020

References:

- Grazzini J., Gaffuri J. and Museux J.-M. (2019): [Delivering Official Statistics as Do-It-Yourself services to foster producers' engagement with Eurostat open data](#), doi:[10.5281/zenodo.3240272](https://doi.org/10.5281/zenodo.3240272).
- ESS Visualisation Workshop (2016): [presentation material](#).
- [ONS guidance for data visualisation](#).

3. Semantic annotation of real world data through the eye of Official Statistics

3.1. What you will do – Description of the project and objectives

Although *Official Statistics* are heavily used by numerous actors (e.g. businesses, academia, data journalists and other organisations) and for various purposes (e.g. policymaking, urban planning, and research and development), their potential for value creation grows when they are combined with other data. However, the reuse of statistical data, including Eurostat open data, is hampered by semantic interoperability challenges, i.e. challenges related to the interpretation of the meaning of data and metadata coming from different sources, in different types and various formats.



This project aims at demonstrating the potential of semantic matching for the integration and the linking of *Official Statistics* with external datasets. The students will select the format of the external data sources, which may come as text⁶ or images⁷. They will then determine the semantic elements (e.g., through semantic segmentation for visual information or content classification for textual information) that hold just enough information to make these resources linkable with Official Statistics, and also identify the metadata to identify a particular representation of these resources in Eurostat database. In this aspect, the work will focus specifically on facilitating the matching of information from different sources (*“bridging the semantic gap”*) while it will not address issues of interoperability.

3.2. What you will learn – Outcomes and benefits

- You will learn about data from *Official Statistics*, the way they are formatted, disseminated and shared.
- You will comprehend statistical taxonomy, more specifically how metadata are handled.
- You will interact with *Official Statistics* through [Eurostat Application Programming Interface](#) and its metadata repositories.
- You will improve your analytics skills from simple data exploration to data processing.
- You will learn best practices in terms of replicability and reproducibility, including versioning and testing of your code.

- You will develop a methodological approach for semantic matching building on concepts derived from artificial intelligence, natural language processing, computer vision, etc...
- **If successful, you will publish your results/software and reference your work on [Eurostat github domain](#).**
- **If successful, the project may lead to a scientific publication co-authored by you and by your Eurostat tutors.**

⁶ for instance: https://www.wikidata.org/wiki/Wikidata:Main_Page.

⁷ for instance: <https://www.mapillary.com/dataset/places>.

3.3. *What you will need – Desired/required knowledge and skills*

- Advanced programming skills in [R](#) and [Python](#) (preferred).
- Prior knowledge of interactive computing notebooks and dashboard technologies, *e.g.* [Jupyter](#), [R Markdown](#) and [voila](#).
- Understanding of machine learning, artificial intelligence and deep learning.
- Interest in computer vision and natural language processing.
- Interest in semantic analysis, ontology matching and metadata management.

3.4. *How you will work – Organisation of the project*

If selected, you will be working on a 4-5 months project under direct tutoring by Eurostat staff. The team will actually be guided by Jacopo GRAZZINI, PhD. Computer Science, and Jean-Marc MUSEUX, PhD. Theoretical Physics.

The project work will be conducted remotely and there is no need for you to travel physically to Eurostat's premises. You will work and communicate mostly through the collaborative development platform, *e.g.* github. Technical questions, *e.g.* regarding code/programming, will be addressed through the ticket issuing facility of the platform. Interactions with the tutors, typically in the form of informal visioconference/calls and via emails, will be arranged flexibly depending on the project needs.

Summary – Additional information

Duration/workload:	4-5 months/5-10 hours a week – 1-month equivalent full time at least.
Period:	August-December (flexible)
Working method:	Remote teams interacting on github and through visio/calls, working language is English
Coding expertise:	Advanced
Contact:	Jacopo Grazzini, email: ESTAT-Methodology@ec.europa.eu
Deadline for application:	30.06.2020

4. Processing of Mobile Network Operator data for spatial statistics

4.1. What you will do – Description of the project and objectives

As mobile phones interact with the mobile network infrastructure, they reveal their approximate location to the mobile network, at least at the level of the radio cell. As most people nowadays carry a mobile phone, the mobile network can be used as a large-scale “sensor” to measure presence and mobility of the entire population of mobile subscribers across the whole country. Transforming Mobile Network Operator (MNO) data into aggregate statistics involves a number of technical and non-technical challenges, from the privacy and legal aspects connected to data access, to the development of a robust, flexible and transparent methodology to transform raw MNO data into reliable statistics.



In this project you will focus on the methodological aspects of MNO data processing to produce spatial statistics about human presence and mobility. Your work will contribute to advance the state-of-the-art in MNO data processing. Within the [methodological framework](#)⁸ under development by Eurostat you will focus on a particular component of a larger methodological chain, for which you will be implementing and comparing alternative solutions, unveiling their respective advantages, limitations and trade-offs and possibly develop new novel approaches. The project involves the generation of synthetic and semi-synthetic data.

The project is expected to result in a scientific publication and in the release of an open notebook.

4.2. What you will learn – Outcomes and benefits

- You will learn about MNO data, how they are generated and what spatial information they contain, what are the sources of errors and uncertainty affecting such data, how to analyse and correctly interpret them.
- You will use computational notebooks to reproduce and verify the narratives based on previous published papers, and possibly develop your own approach by extending/correcting existing analyses.
- You will improve your analytics skills, from simple exploration of datasets to more complex analysis, and your data visualization and interpretation skills.
- **If successful, you will publish your results in a scientific publication.**

⁸ F. Ricciato, G. Lanzieri, A. Wirthmann, G. Seynaeve, “Towards a methodological framework for estimating present population density from mobile network operator data” working paper. <https://europa.eu/!uk88NQ>

4.3. *What you will need – Desired/required knowledge and skills*

- Intermediate to advanced skills in **scientific programming** with [R](#), [Python](#) or [Julia](#).
- Prior knowledge of interactive computing notebooks and dashboard technologies would be an advantage.
- Basic knowledge of GIS software and C/C++ programming language would be an advantage.

4.4. *How you will work – Organisation of the project*

If selected, you will be working on a 3 months project under direct tutoring from Eurostat staff. The main tutor for this project is Fabio RICCIATO, PhD., and Jacopo GRAZZINI, PhD.

The project work will be conducted remotely and there is no need for you to travel physically to Eurostat’s premises. Interactions with the tutors will, typically in the form of informal visioconference/calls, via emails and via collaborative platforms (e.g. github), will be arranged flexibly depending on the project needs. Note that **the timeline of this project is flexible**.

Summary – Additional information

Duration/workload:	3 months with 10-20 hours a week (extendable).
Period:	July-October (flexible)
Working method:	Remote interaction visio/calls, emails and collaborative platforms. Working language is English.
Coding expertise:	Intermediate to advanced
Contact:	Fabio RICCIATO, email: ESTAT-Methodology@ec.europa.eu
Deadline for application:	30.06.2020

References:

[1] F. Ricciato, G. Lanzieri, A. Wirthmann, G. Seynaeve, “Towards a methodological framework for estimating present population density from mobile network operator data” working paper. <https://europa.eu/!uk88NQ>