

ESS Coding Labs

Eurostat - the statistical office of the European Union - and National Statistical Institutes in Greece, Italy, Spain and Romania seek to put together master level students to work in virtual coding projects as 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.

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

The projects focus on learning rather than performing. Therefore, accent is put on documenting all approaches and always make the effort to challenge the ideas of the team. To stimulate creativity and innovation, students will never be blamed for failing or for getting bad results, but will be encouraged to question the evidence and propose new ideas. Thus, do not be afraid to apply to the labs even if you do not match all the requirements. Your interest and motivation are the most important things to apply. If you are willing to invest more time to train and learn, with the support of the tutor, your contribution will be appreciated.




Students involved in the project will receive a certificate of attendance. Depending on the local university rules, the project work may be recognised officially as part of the study program e.g. by serving as a basis for writing a Master thesis or other ways, subject to a case-by-case agreement with university supervisor. Successful candidates will publish their results and reference their work on the statistics coded page github domain and potentially in paper(s) of internationally recognised events.

Please apply by **23 July 2021** to express your interest in the topics explained below through this [online form](#). Selection will be done by project leaders, possibly with a short remote interview. All applicants will be informed of the result.

For reference please see [2020 coding labs](#). Should you need any further information or wish to send additional attachments, please contact estat-emos@ec.europa.eu

| Topic | Skills |
|---------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Graph analysis techniques to process international trade relations by ISTAT Italy | Intermediate to advanced <ul style="list-style-type: none"> – Intermediate to advanced skills in scientific programming with Python. – Prior knowledge of interactive computing notebooks and dashboard technologies would be an advantage. – Prior knowledge of interactive computing notebooks and dashboard technologies, e.g. Jupyter and Voila (https://github.com/voila-dashboards/voila). – Prior knowledge of data processing machine learning and statistics. |



| | | |
|-----------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|
| <p>Time series clustering for Population and Migration Official Statistics</p> <p>By Eurostat F2</p> | <p>Beginner to intermediate</p> <ul style="list-style-type: none"> – Programming skills in R (Python is also a possibility), including: <ul style="list-style-type: none"> ○ Knowledge of some basic data manipulation libraries, eg tidyverse, dplyr or data.table ○ Knowledge of visualisation packages, eg ggplot2 or plotly ○ Knowledge of R markdown notebooks – Good notions of mathematics, including for example <ul style="list-style-type: none"> ○ Topology ○ Time series analysis ○ General Statistics ○ Optimization – Interest in the topic of migrations and demography – Creative spirit |  |
| <p>Data analytics and Official Statistics</p> <p>By ELSTAT Greece</p> | <p>Intermediate</p> <p>Students with a programming background in languages such as R, Python, C, Java etc. would be ideal candidates. individuals who have undertaken courses around statistics and math or know algorithmic logic would also be eligible:</p> <ul style="list-style-type: none"> – Programming skills: Desired (mandatory) – Math skills: Desired (optional) – Statistics: Desired (optional) – Analytical Thinking: Desired (optional) |  |
| <p>Synthetic Mobile Network Operator data processing for population statistics</p> <p>By INE Spain & INSSE Romania</p> | <p>Intermediate to advanced</p> <p>a) Scientific programming skills:</p> <ul style="list-style-type: none"> – Intermediate (to advanced) R programming: – Understanding of XML markup language and XSD schema language – Working with git version control system – Proficiency with RStudio IDE and/or Jupyter notebooks – Basic understanding of GIS – C++ could be an advantage <p>b) Statistics and probabilities:</p> <ul style="list-style-type: none"> – Basic knowledge of probability theory: random variables, probability distribution functions, confidence intervals |  |



1. *Graph analysis techniques to process international trade relations*



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

Graphs are a very useful tools for modeling and analyzing relationships among entities in a wide range of contexts, e.g., disease detection, genetics, healthcare, and banking. Graphs can be a key instrument to analyze international trade relations, allowing to gain insights and make better data driven decisions. As an example, in the context of international trade relations, graphs easily allow to analyze the effects of the COVID-19 pandemic on trade relations.

The applicants will analyze international trade relations using graph analysis techniques applied to COMEXT data <http://epp.eurostat.ec.europa.eu/newxtweb/>.

The applicants will further develop machine learning algorithms to analyze the relationship of international trade.

The applicants will code the graph analysis and machine learning algorithms in Python (using for example Jupyter notebooks and NetworkX package) on COMEXT data. Then the implemented algorithms will be exposed via rest APIs (implemented using Python flask) on cloud environment using docker containers.

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

- You will learn about COMEX Eurostat's reference database for detailed statistics on international trade in goods.
- You will manage and transform international trade data using python language.
- You will implement graph analysis algorithms in python language.
- You will implement machine learning algorithms in deep learning framework TensorFlow.
- You will implement REST API's to integrate your graphs in real world applications.
- You will improve your analytics skills from simple exploration of datasets to complex visualization of graphs.

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

- Intermediate to advanced skills in scientific programming with Python.
- Prior knowledge of interactive computing notebooks and dashboard technologies would be an advantage.
- Prior knowledge of interactive computing notebooks and dashboard technologies, e.g. Jupyter and Voila (<https://github.com/voila-dashboards/voila>).
- Prior knowledge of data processing machine learning and statistics.

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

If selected, you will be working on a 3-month project under direct tutoring from ISTAT staff.

The main tutors for this project are **Fabrizio De Fausti**, **Mauro Bruno** and **Francesco Amato**.



The project work will be conducted remotely and there is no need for you to travel physically to Istat's premises. Interactions with the tutors will be in the form of informal videoconferences, emails, and collaborative platforms (e.g. github) 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: | September – November (flexible) |
| Working method: | Remote interaction videoconferences, emails and collaborative platforms. Working language is English. |
| Coding expertise: | Intermediate to advanced |
| Contact ISTAT Italy: | Fabrizio De Fausti, email: defausti@istat.it Francesco Amato, email: framato@istat.it Mauro Bruno, email: mbruno@istat.it |
| Application by: | 23 July 2021 |
| Send application to: | https://ec.europa.eu/eusurvey/runner/Codinglab |

References:



2. Time series clustering for Population and Migration Official Statistics



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

Migration has become an increasingly important phenomenon for European societies.

Patterns of migration flows may change greatly over time, with the size and composition of the migrant population reflecting both current and historical patterns of migration flows.

This projects aims to exploring data on stocks and flows of migrants in the EU.

During the first part, which will serve as an introduction to the topic, students will be in charge of **producing the most original, insightful and well-designed visualization** using [Eurostat's data on Asylum and Managed Migration statistics](#).

The second part will be an extension of the work on visualizations, but focusing on the purpose of data validation, which is at the heart of Official Statistics. Therefore, students will be in charge of **producing visualizations that can improve the quality of the validation in the field of population statistics**.

Finally, the third and major part of the project will be focused on time series clustering. **Time Series Clustering** is an unsupervised data mining technique for organizing data points into groups based on their similarity. The objective is to maximize data similarity within clusters and minimize it across clusters. Students will start with a bibliographic work to benchmark state-of-the art techniques on clustering of time series and they will then apply those techniques to Eurostat's data on Asylum and Managed Migration Statistics. The goal would be to identify common patterns in terms of flows of migrants, for example considering the asylum requests, the issue of residence permits or the returns of irregular migrants.

2.2. What you will learn – Outcomes and benefits

- You will learn about Official Statistics on Population and Migrations
- You will interact with Official Statistics through Eurostat API by using dedicated client packages: learn to use the API, query, extract, load and transform data from Eurostat database.
- You will explore modern solutions for data visualisations
- You will learn state-of-the-art techniques on time series clustering, and how to implement them with R
- You will learn how to work on an innovation subject in Official Statistics, which includes
 - Bibliographic research
 - Testing ideas
 - Leading a team
 - Documenting your work

Depending on the interests of the students and of the success of the project, the additional outcomes may include :

- Organisation of a lunchtime presentation to present the work in Eurostat
- Use of the results in a publication by Eurostat F2
- Publication and reference to the work on Eurostat github domain



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

- Programming skills in R (Python is also a possibility), including:
 - Knowledge of some basic data manipulation libraries, eg [tidyverse](#), [dplyr](#) or [data.table](#)
 - Knowledge of visualisation packages, eg [ggplot2](#) or [plotly](#)
 - Knowledge of R markdown notebooks
- Good notions of mathematics, including for example
 - Topology
 - Time series analysis
 - General Statistics
 - Optimization
- Interest in the topic of migrations and demography
- Creative spirit

2.4. How you will work – Organisation of the project

You will be working on a 3/4 months project under direct tutoring by Jules Zaccardi, statistician at Eurostat.

The team will work remotely. There will be 1 weekly meeting in visio conference to coordinate the work of the team with the tutor. You will have one bilateral meeting with the tutors for the kick-off and then occasionally if specific support is needed. The group of students will handle part of the work autonomously.

About once a month, there will be a meeting with the other labs to share ideas and progress between each coding lab. It will be an opportunity to summarize your project and get new ideas from your peers.

Summary – Additional information

| | |
|-----------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Duration/workload: | 3 months, 5-10h/week minimum + meetings |
| Period: | End of August – Beginning of November (flexible) |
| Working method: | Language EN Weekly team meetings in Visio conference Brainstorming sessions using tools such as Miro Continuous discussion over slack and emails Bilateral discussions from time to time when needed |
| Coding expertise: | Intermediate or advanced |
| Contact Eurostat: | Jules ZACCARDI, ESTAT-AMM-STATISTICS@ec.europa.eu |
| Application by: | 23 July 2021 |
| Send application to: | https://ec.europa.eu/eusurvey/runner/Codinglab |

References:

- <https://ec.europa.eu/eurostat/documents/3217494/12278353/KS-06-20-184-EN-N.pdf/337ecde0-665e-7162-ee96-be56b6e1186e?t=1611320765858>
- Aghabozorgi, Saeed, Ali Seyed Shirkhorshidi, and Teh Ying Wah. "Time-series clustering—a decade review." *Information Systems* 53 (2015): 16-38.



3. Data analytics and Official Statistics



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

The courses aim to introduce students, in an applied way, to data science. The student during the courses will learn to apply the basic techniques - strategies used in data mining. This includes programming skills, visualization, statistics and math. At the end the student will be able to set successfully the basic business-research questions that can be answered for a specific data set.

3.2. What you will learn – Outcomes and benefits

The main points of the courses are briefly mentioned below, divided into sections.

Section 1

Familiarity with the R programming language and the Rstudio IDE: data types, data structures (vectors, matrices, databases, lists, factors), conditionals, loops, functions, libraries, scripts, suggested data science vocabulary in R.

<https://www.rstudio.com>

Section 2

Familiarity with data wrangling in R: application of SQL queries through the libraries of the tidyverse package (aggregation, selection, filtering, ordering, joins). Exercises and tasks on data sets.

<https://www.tidyverse.org/packages/>

Section3

- Introduction to visualization through the ggplot2 library. Create descriptive visualizations by data type. Use of GUI libraries to produce graphs faster.
- Introduction to code version management and online collaboration on software projects. Use of Git and Github.

Section4

Introduction to literate programming. Use of markdown to create dynamic reports and export them to formats such as html, .pdf, .docx.

Libraries: knitr, slidify

Section5

Introduction to unsupervised learning: principal component analysis, multiple correspondence analysis, Centroids-based Clustering (Partitioning methods), Connectivity-based Clustering (Hierarchical clustering).

Library: Factominer

<http://factominer.free.fr>

Section6

-Introduction to supervised learning: creating models for binary classification. Logistic regression, decision trees, random forests, support vector machines, neural networks, Bayesian networks.

Library: caret

<https://topepo.github.io/caret/>



Section7

- Resampling techniques (bootstrap, cross-validation)
- Bagging and Boosting techniques
- Use of parallel programming for model training in R.
- Comparison of model performance

Library: caret, caret ensemble

Section8

- Introduction to the construction of interactive app for the presentation of data science projects.
- Description of basic concepts of HTML, CSS, Javascript.
- Introduction and construction of a forecasting application and production of dynamic report.

Library: shiny

<https://shiny.rstudio.com>

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

In general, it is important for the learner to have prior knowledge of a programming language (or even to know algorithmic logic) although people with very good computer skills could also successfully attend the course. Students with a programming background in languages such as R, Python, C, Java etc. would be ideal candidates. Also, individuals who have undertaken courses around statistics and math would benefit the lectures.

- Programming skills: Desired (mandatory)
- Math skills: Desired (optional)
- Statistics: Desired (optional)
- Analytical Thinking: Desired (optional)

3.4. How you will work – Organisation of the project

The lectures will be carried out with the method of synchronous remote learning using a platform (eg zoom, webex, etc.). An asynchronous e-learning platform (eg. moodle) will also be used to organize the classroom and add announcements, material, exercises and assignments.

Summary – Additional information

| | |
|--------------------------------|-------------------------------------------------------------------------------------------------------------|
| Duration/workload: | 8 weeks |
| Period: | 10/2021-12/2021 |
| Working method: | Synchronous-asynchronous |
| Coding expertise: | Intermediate |
| Contact ELSTAT, Greece: | smos@statistics.gr |
| Application by: | 1 October 2021 |
| Send application to: | https://ec.europa.eu/eusurvey/runner/Codinglab |



4. *Synthetic Mobile Network Operator data processing for population statistics*



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

Mobile Network Operator (MNO) data has proved to be an outstanding data source for the production of statistics in general, and for Official Statistics, in particular. There are several statistics domains that can benefit from MNO data: population statistics, tourism, migration, transportation etc. This project starts from the end-to-end methodological framework developed during the ESSnet Big Data II project which comprises the following modules:

- Geolocation - this module focuses on the computation of location probabilities for each device across a reference grid used for the statistical analysis.
- Deduplication - this module focuses on the computation of multiplicity probabilities for each device, i.e. probabilities of a given device to be carried by an individual jointly with one or several other devices
- Statistical filtering - this module focuses on the algorithmic identification of mobile devices of individuals of the target population such as domestic tourists, commuters, inbound tourists, etc.
- Aggregation - this module focuses on the computation of probability distributions for the number of individuals detected by the network (i.e. with mobile devices) across different territorial units.
- Inference - this module focuses on the computation of probability distributions for the number of individuals of the target population across different territorial units.

Except the statistical filtering, all modules are implemented in specific R packages, freely available. The objective of the project is twofold:

- to run the entire process using synthetic data, generated by a simulation software specially designed for this purpose and build visualisations of the outputs of each module
- to improve the computation of the estimated target population by improving the already existing methods, starting from the computation of the location probabilities of each device and going up to the final estimates of the target population.

The first step of the project will be to build and run different simulations using different network configurations and mobility patterns of the population. Then, for each output of the simulation software some visualisations should be build: the signal strength / signal dominance, the location probabilities, the trajectory of each person, the estimated number of mobile devices detected by the network, the estimated target population, the real target population, and the dynamics of the population. All visualisation functions could be assembled in an R package.

The module computing the location probabilities is one of utmost importance for the entire process. Currently its implementation is based on Hidden Markov Models. You will explore different ways of optimizations of this implementation and possibly develop new methods.

You will analyse and propose different optimizations for the other modules too, showing their advantages.



4.2. What you will learn – Outcomes and benefits

- You will learn about the structure of MNO data, the specificity of these data sets, how are they generated by the interaction between mobile devices and the network, the information that they carry, and how these data can be used for official statistics.
- You will use R (optionally with C++) to develop your own approach by improving the already existing analyses.
- You will improve your analytics skills.
- You will improve your data visualization and interpretation skills.

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

c) Scientific programming skills:

- Intermediate (to advanced) R programming:
 - data types, objects, arrays and matrices, lists and data frames, reading/writing data from/to files, control statements, user defined functions, graphical procedures
 - advanced graphics with ggplot2 package, advanced data manipulation with data.table package, the tidyverse suite of R packages, building web apps with shiny R package.
- Understanding of XML markup language and XSD schema language
- Working with git version control system
- Proficiency with RStudio IDE and/or Jupyter notebooks
- Basic understanding of GIS
- C++ could be an advantage

d) Statistics and probabilities:

- Basic knowledge of probability theory: random variables, probability distribution functions, confidence intervals

4.4. How you will work – Organisation of the project

You will be working on a 2 and ½ months project under direct tutoring from experts from INE Spain and NSI Romania. The main tutors for this project are David SALGADO, PhD., and Bogdan OANCEA, PhD.

All the activities of this project will be conducted remotely. Interactions with the tutors will be in the form of online video conferences/calls, via emails and via other collaborative platforms (github, slack).

The timing of the project will be flexible.

The software tools developed during this project will be uploaded on a github repository.



Summary – Additional information

| | |
|-----------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------|
| Duration/workload: | 2 and ½ months, 10-16 hours a week |
| Period: | 1.09.2021 – 15.11.2021. The period is flexible. |
| Working method: | Remote teams interacting on github and through online meeting calls, the working language is English. |
| Coding expertise: | Intermediate: R, XML, XSD, C++ (optional) |
| Contact: | david.salgado.fernandez@ine.es bogdan.oancea@insse.ro |
| Application by: | 23 July 2021 |
| Send application to: | https://ec.europa.eu/eusurvey/runner/Codinglab |

References:

Salgado, D., Sanguiao, L., Oancea, B. et al. An end-to-end statistical process with mobile network data for official statistics. *EPJ Data Sci.* **10**, 20 (2021).

<https://doi.org/10.1140/epjds/s13688-021-00275-w>

Salgado, D., Sanguiao, L., Oancea, B., Barragan, S., Necula, M., Towards a modular end-to-end statistical production process with mobile network data, *SPANISH JOURNAL OF STATISTICS* Vol.2 No. 1 2020, Pages 41–77 doi: <https://doi.org/10.37830/SJS.2020.1.04>

Salgado, D., Sanguiao, L., Barragan, S., Oancea, B., Necula, M.(2020) , A proposed production framework with mobile network data, Technical report, Statistics Spain (INE) and Statistics Romania (INS)

https://ec.europa.eu/eurostat/cros/system/files/wpi_deliverable_i3_a_proposed_production_framework_with_mobile_network_data_2020_11_26_final.pdf

Oancea, B., Barragan, S., Sanguiao, L., Salgado, D. (2020), Some IT tools for the production of official statistics with mobile network data, Technical report, Statistics Romania (INS) and Statistics Spain (INE),

https://ec.europa.eu/eurostat/cros/sites/default/files/WPI_Deliverable_I4_Some_IT_tools_for_the_production_of_official_statistics_with_mobile_network_data_2020_11_05_Final_corrected.pdf

Oancea, B., Necula, M., Sanguiao, L., Salgado, D., Barragán, S. (2019), A simulator for network event data, Technical report, Statistics Romania (INS) and Statistics Spain (INE),

https://ec.europa.eu/eurostat/cros/sites/default/files/WPI_Deliverable_I2_Data_Simulator_-_A_simulator_for_network_event_data.pdf

Sanguiao L, Barragán S, Salgado D (2020) destim: an R package for mobile devices position estimation. R package version 0.1.0. <https://github.com/Luis-Sanguiao/destim>

Oancea B, Barragán S, Salgado D (2020) deduplication: an R package for deduplicating mobile device counts into population individual counts. R package version 0.1.0.

<https://github.com/bogdanoancea/deduplication>

Oancea B, Barragán S, Salgado D (2020) aggregation: an R package to produce probability distributions of aggregate number of mobile devices. R package version 0.1.0.

<https://github.org/bogdanoancea/aggregation>

Oancea B, Barragán S, Salgado D (2020) inference: R package for computing the probability distribution of the number of individuals in the target population. R package version 0.1.0. <https://github.com/bogdanoancea/inference>