A Python framework for automated analysis of EU LFS microdata

Pavlos Sermpezis[1], Dimitrios Psarologos[2], Kostas Gourzis[2], and Stelios Gialis[2]

[1] Data & Web Science Lab, Aristotle University of Thessaloniki, Greece
[2] Labour Geography Research Lab, University of the Aegean, Greece

Introduction & Motivation

- **The "microdata":** Questionnaires of the Eurostat Labour Force Survey (LFS)
  - The LFS is conducted annually
  - The survey for each country and year contains several 1000s of questionnaires. For each questionnaire there are more than 150 fields/answers.
  - The data for each country and year can be more than 1GB!
  - The publication of the raw data (i.e., answers) of the questionnaires is strictly prohibited

- **Our key research question**
  - *How can we streamline the analysis of EU microdata, making it accessible to authorized users, while maintaining the flexibility and required power for large-scale data processing?*

Goals & Contributions

- **Goals**
  - Develop a system for the quick & automated analysis of LFS for multiple years & countries
  - Provide non-expert users with easy & intuitive tools to perform complex analyses
  - Facilitate the integration of advanced data science & machine learning & visualization techniques

- **Contributions**
  - Python framework for analysis of LFS data
    (framework = collection of methods, tools, libraries, etc.)
  - Web application with user-friendly interface for analysis (“in a few clicks”) & presentation of data

Web application: Overview

- Web application, i.e., it can be opened as a tab in a Web/mobile browser
- No need for knowledge of statistical software (e.g., SPSS, Stata)
- Complex requests with a few clicks on buttons (selections, options, filters, etc.)

  e.g., *“what is the employment per sector per region in country X at year Y among the age groups 15-19 years old of only female population with education level larger than Z?”*

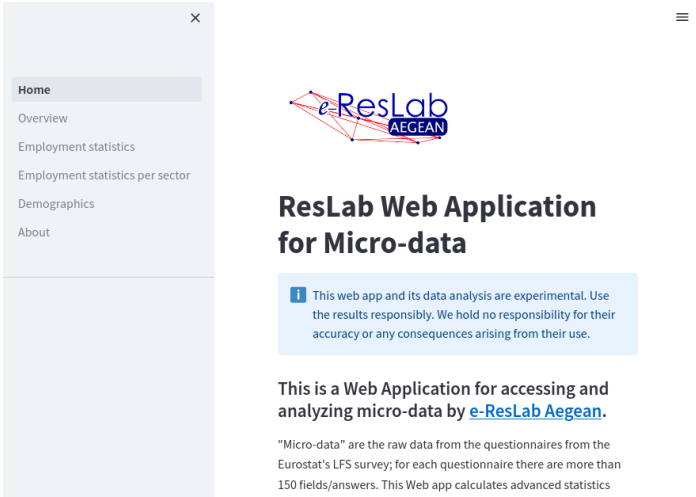
Web application: Overview

- Home page
  - Information about the Web application
  - Login form
  - Registration form
- Pages with analyses/statistics
  - Several subpages that calculate and present statistics, based on the selections of the user
  - Each subpage is focused on *different* analysis (of the *same* data)

Access to data is not public! You need to contact us & register

Home Page

- Home page
  - Information about the Web application
  - Login form
  - Registration form



"A Python framework for automated analysis of EU LFS microdata", ***Labour Geography Research Lab (LGRL)***, https://lgrl.aegean.gr/

Overview analysis Page

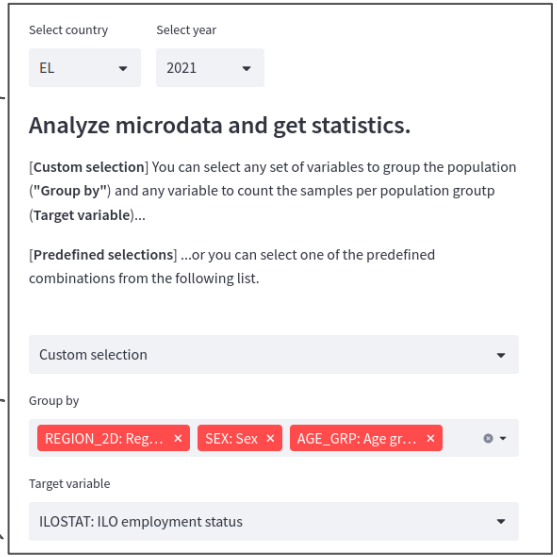
- “Overview analysis” page: Basic statistical analysis of the microdata
  - Select a country & a year
  - From the 150+ variable (geographical region, sex, age, employment status, education level, etc.)
    - Select any variable to analyze (*target variable*)
    - Group populations by any set of variables (*group variables*)
  - Calculate statistics for the target variable
  - Download results
    - in csv or excel format
    - for many countries & years at the same time
- Ease of use
  - Analysis per country,year → with 4 clicks, in ~2seconds
  - Analysis for 10 countries * 10 years → with 6 clicks, in ~3min

Overview analysis Page

Select county & year

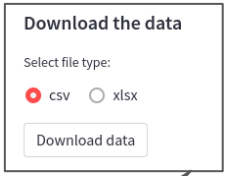
Select variables to group the data (out of 150+ options!)

Select the variable for which to calculate statistics



The Web app shows the results in a table!

REGION_2D	REGION NAME	SEX	AGE_GRP	ILOSTAT	Nb. of samples	Population (inferred
EL30	Attiki	Female	0-4 years of age	Outside the labour forc	100	89,941
EL30	Attiki	Female	10-14 years of age	Outside the labour forc	129	90,012
EL30	Attiki	Female	15-19 years of age	Employed	1	1,241
EL30	Attiki	Female	15-19 years of age	Outside the labour forc	125	98,458
EL30	Attiki	Female	15-19 years of age	Unemployed	2	1,155
EL30	Attiki	Female	20-24 years of age	Employed	36	27,698
EL30	Attiki	Female	20-24 years of age	Outside the labour forc	73	55,214



… and you can download the data!

… or download the data for more than one countries and/or years!

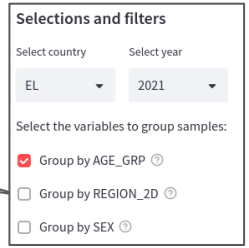
"A Python framework for automated analysis of EU LFS microdata", ***Labour Geography Research Lab (LGRL)***, https://lgrl.aegean.gr/

Employment statistics Page

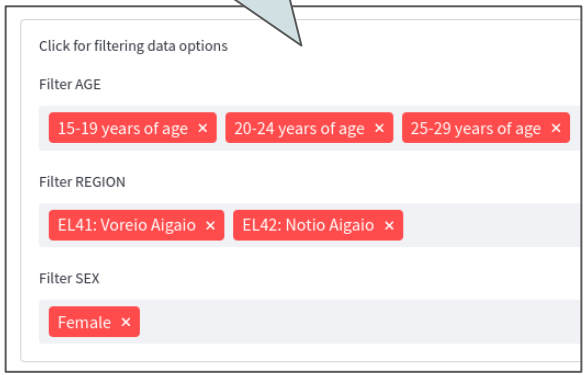
- “Employment statistics” page:
  - Similar functionality to the “Overview analysis” page, but …
    - Focused on the employment status of the population (*target variable*)
    - Group the population based on geographical region, sex, and age
    - Calculate statistics for their employment status
    - Advanced filtering options
    - Calculation of statistics
    - Calculation of statistical errors (for population)
    - Download results (csv or excel format)

Employment statistics Page

Select county & year and (optional) group variables



(Optionally) Filter the population of interest



Statistics are calculated in real-time & presented in a table

Statistical errors are also calculated!

	AGE_GRP	ILOSTAT	Nb. of samples	Population (inferred)	Perc. (samples)	Perc. (population)	Max error (samples)
0	0-4 years of age	Outside the labour force	4,981	344,579	0.025	0.0326	0.0007
23	5-9 years of age	Outside the labour force	7,289	497,932	0.0366	0.0471	0.0008
1	10-14 years of age	Outside the labour force	9,719	646,081	0.0488	0.0611	0.001
2	15-19 years of age	Employed	193	14,050	0.001	0.0013	0.0001
3	15-19 years of age	Outside the labour force	8,050	536,191	0.0404	0.0507	0.0009
4	15-19 years of age	Unemployed	157	8,628	0.0008	0.0008	0.0001
5	20-24 years of age	Employed	1,861	129,466	0.0093	0.0123	0.0004
6	20-24 years of age	Outside the labour force	3,737	283,673	0.0188	0.0268	0.0006
7	20-24 years of age	Unemployed	1,094	68,605	0.0055	0.0065	0.0003
10	25-29 years of age	Unemployed	1,508	123,292	0.0076	0.0117	0.0004

Total nb. samples
199,304

Total population (inferred)
10,567,546

Max error
0.1%

Quick stats about the results

+ Download options

"A Python framework for automated analysis of EU LFS microdata", ***Labour Geography Research Lab (LGRL)***, https://lgrl.aegean.gr/

Employment statistics per sector Page

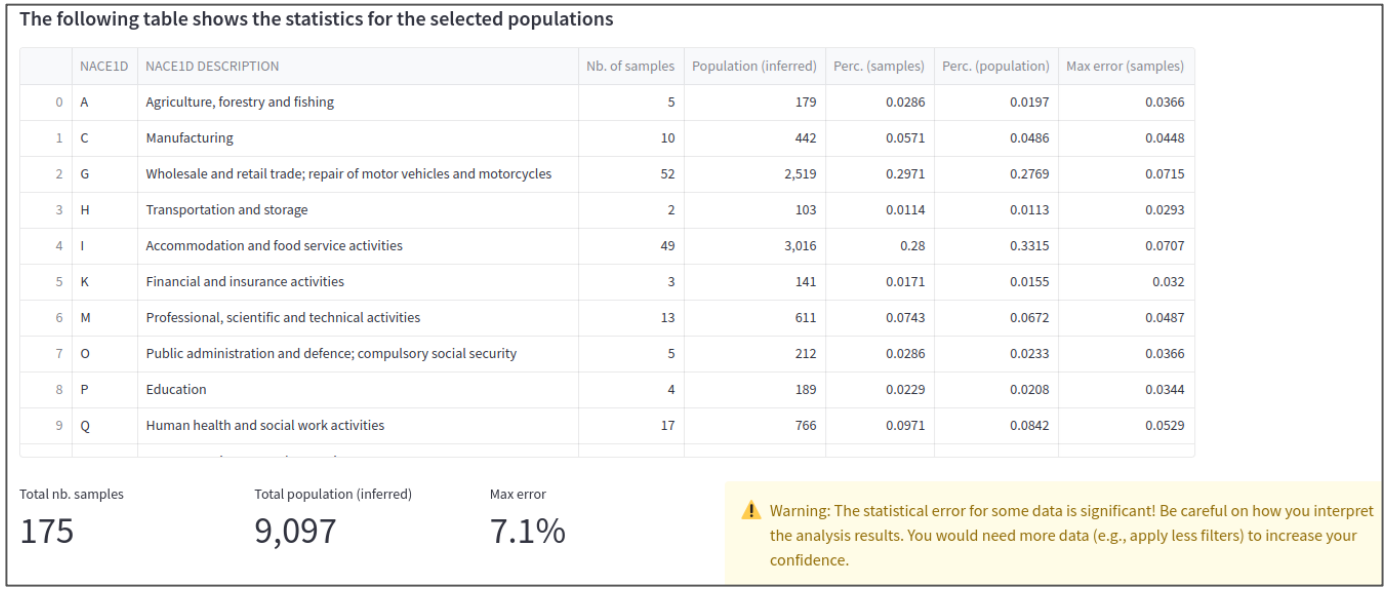
- “Employment statistics per sector” page:
  - Similar functionality to the “Employment statistics” page, but …
    - Focused on the employment sectors of the employed population
    - Advanced filtering options per employment sector

Employment statistics per sector Page

Selection & Group & Filter options

Calculation of statistics & presentation of data in table

+ Download options



The following table shows the statistics for the selected populations

	NACE1D	NACE1D DESCRIPTION	Nb. of samples	Population (inferred)	Perc. (samples)	Perc. (population)	Max error (samples)
0	A	Agriculture, forestry and fishing	5	179	0.0286	0.0197	0.0366
1	C	Manufacturing	10	442	0.0571	0.0486	0.0448
2	G	Wholesale and retail trade; repair of motor vehicles and motorcycles	52	2,519	0.2971	0.2769	0.0715
3	H	Transportation and storage	2	103	0.0114	0.0113	0.0293
4	I	Accommodation and food service activities	49	3,016	0.28	0.3315	0.0707
5	K	Financial and insurance activities	3	141	0.0171	0.0155	0.032
6	M	Professional, scientific and technical activities	13	611	0.0743	0.0672	0.0487
7	O	Public administration and defence; compulsory social security	5	212	0.0286	0.0233	0.0366
8	P	Education	4	189	0.0229	0.0208	0.0344
9	Q	Human health and social work activities	17	766	0.0971	0.0842	0.0529

Total nb. samples
175

Total population (inferred)
9,097

Max error
7.1%

⚠ Warning: The statistical error for some data is significant! Be careful on how you interpret the analysis results. You would need more data (e.g., apply less filters) to increase your confidence.

Warning if the confidence is small (i.e., high statistical error)

Demographics Page

- Basic statistical analysis for demographics
  - Simplified version of analysis
  - Calculate population per
    - geographical region
    - age
    - sex
  - Download results

Select country: ES

Select year: 2021

Select demographics: AGE_GRP: Age gr… × | REGION_2D: Reg… × | SEX: Sex ×

	REGION_2D	REGION NAME	SEX	AGE_GRP	Nb. of samples	Population (inferred)
0	ES11	Galicia	Female	0-4 years of age	153	39,409
1	ES11	Galicia	Female	10-14 years of age	258	58,235
2	ES11	Galicia	Female	100 years of age and older	6	1,071
3	ES11	Galicia	Female	15-19 years of age	292	63,533
4	ES11	Galicia	Female	20-24 years of age	241	56,369
5	ES11	Galicia	Female	25-29 years of age	174	53,162
6	ES11	Galicia	Female	30-34 years of age	176	52,593

Methodology - Python framework: Overview

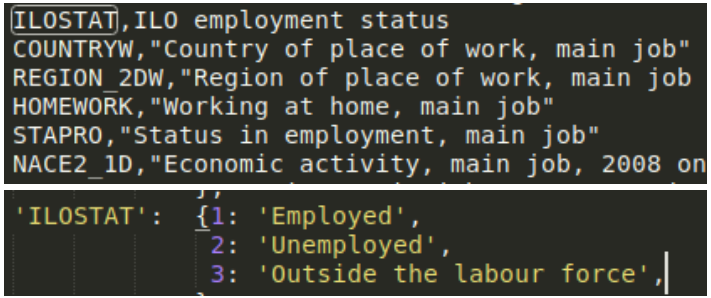
- Data analysis
- Data formatting
- Python libraries & technologies

Methodology - Data analysis

- Select country, year, target and group variables (user choices)
- Load all responses in the survey and group them based on the selections
  - *E.g. if the user has selected to group data based on sex (Male, Female) and selected to calculate the statistics for the employment status (Employed, Unemployed, Inactive), then we group the responses in 6 groups: Male/Employed, Male/Unemployed, Male/Inactive, Female/Employed, Female/Unemployed, Female/Inactive*
- For each group (e.g., Male/Employed)
  - (i) count the number of responses
  - (ii) calculate the inferred population (using the "COEFF" variable)
  - (iii) calculate the percentage of responses/population of each group over total responses/population
- Calculate statistical errors
  - Calculations for the percentages are susceptible to statistical errors (e.g., few responses)
  - Use the Wilson method to calculate for the percentage ***P*** its confidence interval ***[P-m, P+M]***

Methodology - Data formatting

- Changes in NUTS codes throughout the years
  - Re-codings (e.g., EL11 → EL51)
  - Merging of regions (e.g., DE41 and DE42 → DE40)
  - Splitting of regions (IE05 and IE06 ← IE02)
- Changes of variables names in LFS data (different years)
  - E.g. REGION → REGION_2D
  - Need to update code when updating data
- Coding of variable names & values
  - Need to define them in the code, i.e., hard-coded (for user-friendliness)
- NUTS names
  - Need to get names from external data (for user-friendliness)



Methodology - Python framework

- pandas, numpy
  - Libraries for data analysis, handling, calculations, etc.
- requests, json, excel, etc. libraries
  - Loading, saving, generating data
- streamlit
  - Library for easy creation of the web application

Summarizing …

- Goals & Contributions
  - Python framework & Web application
  - Easy, quick, automated analysis of LFS microdata
  - Use case on employment data

- Future work
  - Extra functionalities based on python libraries: Visualizations, ML/AI functions, analytics, etc.
  - Focus on other use cases → **tell us your ideas!**

Thank you!

Find out more about LGRL: https://lgrl.aegean.gr/

Contact us at: labourgeolab@gmail.com

The Project “Towards inclusive urban societies: Addressing labour and housing precarity based on advanced Geographical Information knowledge” is carried out within the framework of the National Recovery and Resilience Plan Greece 2.0 and is funded by the European Union - NextGenerationEU (Implementation Body: Hellenic Foundation for Research & Innovation).