

Curriculum Vitae (CV) Data Analytics & Intelligence

EUROPASS-2021-0081-NP-DSI-PHT-ASIA-CV-
DataAnalytics&Intelligence-008-20

Final Report

27/04/2022

Table of Contents

1. Abstract	7
2. Introduction	8
2.1 Background	8
2.2 Objectives	8
2.3 Summary of Outputs	9
2.4 Structure of the Report	10
3. Methodology.....	11
3.1 Overview.....	11
3.2 Data Sources	11
3.2.1 Europass User CVs	11
3.2.2 Statistical Frameworks and Classification Systems.....	11
3.2.3 External Sources of Statistics.....	12
3.2 Statistics Employed	13
3.2.1 Data Cleansing and Standardization	13
3.2.2 Weighting	14
3.2.3 Time-Series Analysis	14
3.2.4 Correlation and Regression Analysis	15
3.2.5 Association Rules	15
3.2.6 Revealed Comparative Advantage (RCA).....	16
4. Understanding the Europass CV dataset	17
4.1 Overview.....	17
4.2 Working with Naturally-Occurring CV Data.....	18
4.3 Technological Diffusion	19
4.4 User Composition Compared to the Labour Force.....	19
5. Occupations Analysis, Trends and Correlations	23
5.1 Disappearing and Newly-Emerging Occupations.....	23
5.1.1 Trends in Job Sectors	23
5.1.2 Highest Degree of Deviation in Trend per Country	24
5.1.3 Recruitments over Terminations in Time	25
5.1.4 Mean Net Hire Ratio of ISCO 2 Groups per Country.....	27
5.1.5 Next Steps.....	28
5.2 Changing Skill Requirements for Job Positions.....	29
5.2.1 Skills by Occupation and Birth Year	29
5.2.2 Next Steps.....	30
5.3 Usual Career Paths	30

5.3.1 Frequency of Group Change.....	31
5.3.2 Most Common Transitions to Different Groups	31
5.3.3 Work Experience by Age.....	32
5.3.4 Deviation of Observed and Expected Work Experience	34
5.3.5 Next Steps.....	35
5.4 Skills-to-Occupations Associations in the ESCO Model Compared to the Collected CV Data	36
5.4.1 Associations between Skills and Occupations	36
6. Cross Checking with Official Statistics	38
6.1 Employment (LFS)	38
6.1.1 Employment by Occupations per Age Group and Gender (EA19, 2019)	38
6.1.2 Employment by Occupation per Country (2019).....	40
6.1.3 Trends in Employment by Occupation (EA19).....	41
6.1.4 Gender Ratio by Occupation (EA19).....	42
6.2 Unemployment (LFS)	43
6.2.1 Unemployment by most Recent Occupation per Country (2019)	43
6.2.2 Education Level of the Long-Term Unemployed by Age Group per Country (2019)	44
6.3 Job Vacancies (Cedefop)	45
6.3.1 Supply and Demand (2019)	45
6.4 Job Tenure (LFS)	48
6.4.1 Job Tenure per Age Group and Gender (EA19, 2019)	48
6.4.2 Job Tenure per Occupation (EA19, 2019).....	49
7. Skills Analysis	50
7.1 Distribution of skills.....	50
7.1.1 Overall	50
7.1.2 Age Group.....	51
7.1.3 Gender.....	53
7.1.2 Country	54
7.2 Occupations and Skills	55
7.2.1 Skills by Occupation	55
7.2.2 Importance of Skills to Occupations	56
7.2.3 Skillscape	57
7.3 Skills of NEETs	58
7.3.1 NEETs in the Europass Dataset	58
7.3.2 Distribution of Skills among NEETs	59
7.3.3 Most Deviated Skills	60
8. Executive Summary	62
8.1 Discussion	62

8.2 Key Findings	64
8.3 Perspectives for Future Research.....	66
<i>Annex A: Methodological Notes</i>	67
Weighting	67
Time Series Analysis.....	67
Correlation Analysis	68
Association Rules	68
Revealed Comparative Advantage (RCA).....	68
<i>Annex B: Concordances</i>	69
<i>Annex C: Library for Classifying Multilingual Text of Qualification Titles - educationR</i>	70
EQF Level	70
Dataset	70
Feature Extraction	70
Classifier	71
Results	71
ISCED Fields of Education and Training (ISCED-F 2013)	72
Corpus of Documents.....	72
Pre-processing.....	73
Matching.....	73
Results	73
<i>Annex D: Exploratory Data Tool</i>	74
Explore Data	74
Text Analysis.....	78
Pairwise Analysis	80
Association Rules	83
Codebook	84
<i>Bibliography</i>	85

Table of Tables

Table 1. Completion Rate of CV Fields by Pillar.....	17
Table 2. Chi-squared Test and Residuals for Users in Age Group 15-24 in EA-19 Countries.....	21
Table 3. Chi-Squared Test and Residuals for Users in Age Group 25-49 in EA-19 Countries.....	21
Table 4. Overall Statistics for Italian model.....	71
Table 5. Statistics by Class for Italian model.	72
Table 6. Overall Statistics for Italian model.....	73
Table 7. Statistics by Class for Italian model.	73

Table of Figures

Figure 1. Data processing pipeline schematic	13
Figure 2. Age distribution reported in Europass CVs between 15 and 49 years old compared to the equivalent Eurostat demographics indicator.	14
Figure 3. Distribution of Work Experiences in the Europass CV Dataset with Respect to ISCO 1.	20
Figure 4. Age Distribution of Europass Users	22
Figure 5. Distribution of Country of Residence among Europass Users	22
Figure 6. Example of ISCO 2 occupations with an observed relative increase in recruitments in time. (EA-19).....	24
Figure 7. Example of ISCO 2 occupations with an observed relative decrease in recruitments in time. (EA-19).....	24
Figure 8. Example of the ISCO 2 occupations of Greece that display the highest deviation in yearly odds ratio change compared to the EA-19 average.	25
Figure 9. Evolution of net hire ratio for ISCO 1 occupations. (EA-19)	26
Figure 10. Evolution of net hire ratio across countries.	27
Figure 11. ISCO 2 occupations displaying the highest mean net hire ratio. (EA-19)	28
Figure 12. ISCO 2 occupations displaying the lowest mean net hire ratio. (EA-19)	28
Figure 13. Example of keywords more often included in younger users' skills with latest job Professionals.	29
Figure 14. Example of keywords more often included in older users' skills with latest job Professionals.	30
Figure 15. Example of ESCO skills more often entered by younger users with latest job Professionals.	30
Figure 16. Example of ESCO skills more often entered by older users with latest job Professionals.	30
Figure 17. ISCO 1 occupations with the lowest probability of group change.....	31
Figure 18. Probabilities of group change for ISCO 2 occupations of Professionals.	31
Figure 19. Transition between ISCO 1 occupations (excluding self-transition).	32
Figure 20. Relationship between age and years of work experience with ISCO 1 occupations placed based on their mean values.	33
Figure 21. Difference between expected and measured mean work experience by ISCO 1 groups. The expected work experience is calculated by using a linear model between age and work experience.	34
Figure 22. Relationship between age and years of work experience with ISCO 3 occupations placed based on their mean values	36
Figure 23. Associations between occupations and skills identified through market based analysis for Service and sales workers	37

Figure 24. Example of the comparison of distributions for EA-19 and ages 25-49, with both weighted and unweighted measurements	39
Figure 25. Example of the comparison of distributions for Greece and ages 15-49, with both weighted and unweighted measurements	40
Figure 26. Comparisons of trends for age group 25-49.....	41
Figure 27. Example of the evolution of gender ratio of occupations in time.....	42
Figure 28. Distribution of the latest ISCO 1 occupation for unemployed users	43
Figure 29. Distribution of education level for unemployed individuals in EA-19	45
Figure 30. Example of the distribution of the ISCO 1 for EU	46
Figure 31. Distribution of job tenure for age group 15-24	48
Figure 32. Distribution of job tenure for Managers	49
Figure 33. Distribution of ESCO skills found on Europass CVs.....	51
Figure 34. Age group breakdown of the most common ESCO skills mentioned in Europass CVs.....	52
Figure 35. ESCO skills displaying the highest deviation between the two genders.....	53
Figure 36. Skills distribution for Italy.....	54
Figure 37. Skills distribution for Portugal	54
Figure 38. Skills distribution for Romania.....	54
Figure 39. Top skill groups for waiters and bartenders.....	55
Figure 40. Top skill groups for administrative and specialised secretaries	55
Figure 41. Top skill groups for software and applications developers and analysts	56
Figure 42. Top skill groups by RCA for Waiters and bartenders.....	56
Figure 43. Top skill groups by RCA for Administrative and specialised secretaries.....	56
Figure 44. Top skill groups by RCA for Software and applications developers and analysts.....	57
Figure 45. Skillscape	58
Figure 46. Breakdown of the education and employment status of Europass users	59
Figure 47. Comparison of the inclusion of the most common ESCO skills between users in employment/education and NEET users.....	60
Figure 48. ESCO skills displaying the highest deviation between users in employment/education and NEET users.....	61
Figure 49. EQF level classification schematic	70
Figure 50. Cross-validation curves for the glmnet model for English. Cross-validation curves appear as the red dots, with upper and lower standard deviation shown as error bars.	71
Figure 51. ISCED field matching schematic	72

1. Abstract

In this study, we explore economic statistics assessing labour force characteristics using a dataset of 10 million anonymised CV entries contained in the Europass CV editor application's backup database. This data is naturally occurring and includes information on a variety of labour force aspects (e.g., region, age, work experience history, education, and more) that formal surveys do not typically collect in such extensive detail. The calculated statistics intend to expose the dataset's biases and limitations and investigate to which extent the type and quality of information encoded in various dimensions can provide a plausible and granular picture of the labour force. In addition to the analysis, we build algorithms and packages that assist labour market research in numerous ways, such as mapping free text of job titles, qualification titles, and skills, into standardised classifications, like the ESCO classification and the European Qualifications Framework. The necessary data cleansing, transformations, and classifications were performed using a data pipeline based on the R statistical computing language and the analysis is presented in reproducible form, whereby interlinked raw data, analysis, code and results can be inspected and independently reproduced by researchers.

2. Introduction

2.1 Background

Data collection from online resources and its subsequent analysis are growing research fields with a promise for deepening and broadening our understanding of a variety of socioeconomic research activities. Utilisation of innovative data sources and analytical methodologies is also becoming increasingly popular in labour market research. The availability of online tools for job and candidate search, job matching, and CV creation has revolutionised labour market interactions in many ways, especially with regard to job search. (Mang, 2012) As Internet access and use have spread across virtually all socioeconomic categories and countries, these online resources can help researchers gain a better understanding of micro-level issues, including employer skill demand, occupational change trends, wages, and working conditions.

Cedefop was responsible for the implementation, maintenance and development of the Europass web resources (web portal and the **online CV editor**) during the period 2005-2020. Europass is an initiative of the European Commission (EC) with the objective to increase transparency of qualifications and promote mobility of European citizens, established by Decision No 2241/2004/EC. The EC provided Cedefop with a mandate for the technical development of the Europass web resources until the new Europass portal was launched (1st July 2020). The Europass online CV editor has also maintained a database of anonymised statistical data on the CVs being created. The fields stored include: Education (qualification title, name of school, dates), Work experiences (job title, dates) and Skills (across four categories). The aim of this contract is to further explore this data in order to extract, through data analysis and machine learning, structured information on the Occupations, Skills and Qualifications from the free text entries inside the CVs using standard EU and international classification systems such as ESCO, ISCO, O*NET, NOC, EQF and ISCED-F.

2.2 Objectives

An exploratory data analysis of Europass CV data was performed within the scope of the Europass Survey, an opt-in survey conducted between June and September of 2019, which accumulated approximately 400K CVs. The present study serves as a continuation of that initial analysis on a larger scale, using the Europass CV editor database as its primary data source which includes anonymised data of over 10M users of the Europass CV editor.

The objectives of the assignment, as documented in Cedefop's Negotiated Invitation to Tender (NP/DSI/PHT-ASIA/CV_DataAnalytics&Intelligence/008/20) are presented below:

Task	Objective
Task 1	Produce a report with insights on: a) disappearing and newly-emerging occupations, b) changing skill requirements for jobs positions, c) usual career paths, d) skills-to-occupations association in the ESCO model compared to the collected CV data.
Task 2	Produce associations and cross-references of the CV data with other data sources and official statistics (notably, Eurostat's Labour Force Survey, Cedefop's Job vacancies data5, and others which the contractor may propose).
Task 3	Produce a report on different skill groups defined in the new ESCO (v1.0.5, May 2020) hierarchy of skills, like soft, digital and hard skills.
Task 4	Add concordances of the CV data with the O*NET and the Canadian NOC, comparing the results to those of ESCO.

Task	Objective
Task 5	Implement a library that can classify free multilingual text of Qualification titles (appearing in the CV field “Education & Training title”), to standardised taxonomies of EQF and ISCED FoET.
Task 6	Implement the visualisation and presentation of these results in an interactive user-friendly web interface.
Task 7	Publish the resulting code in a public, open-source code repository;
Task 8	Assist Cedefop in preparing a printed publication (document) to include analysis and interpretation of the results.

2.3 Summary of Outputs

The following table presents a summary of the study outputs.

Id	Output	Task(s)
O.01	Pipeline index – Internal documentation of the data processing pipeline. URL: https://epas-dsense.eworx.gr/results/ewa_index.html	Tasks 1 - 6
O.02	Codebooks – A set of standardised codebooks was produced for the needs of the study, as presented below: <ul style="list-style-type: none"> Demographics – provides insights on the demographic distribution of Europass CVs using 18 features (e.g., gender, nationality, work years). Occupations - describes aggregate job-related statistics based on the Europass CV database using 14 features (e.g., job esco, preceding work years). Skills/Competences – provides insights on the skills and competences appearing in the Europass CVs using 14 features (e.g., skill type, age group). Career Paths – provides insights on career paths identified in the Europass CV database using 16 features (e.g., recruitment year, termination year). URL: https://epas-dsense.eworx.gr/results/shiny/app/exploratory_data_tool/#!/codebook	Tasks 1 - 6
O.03	Occupations Analysis, Trends and Correlations report – Using naturally occurring labour market data aggregated from over 10M Europass CVs, the report presents insights on: <ul style="list-style-type: none"> disappearing and newly-emerging occupations; changing skill requirements for jobs positions; usual career paths; and skills-to-occupations association in the ESCO model compared to the collected CV data. URL: https://epas-dsense.eworx.gr/results/i-cv/regression_correlation.html See also: 5. <i>Occupations Analysis, Trends and Correlations</i>	Task 1
O.04	Cross checking with official statistics – Presents associations and cross-references of the aggregated Europass CV data with other data sources and official statistics (Eurostat’s European Union Labour Force Survey and Cedefop’s online job vacancy data). URL: https://epas-dsense.eworx.gr/results/i-cv/official_stats.html See also: 4. <i>Understanding the Europass CV dataset</i> , 6. <i>Cross Checking with Official Statistics</i>	Task 2
O.05	Europass CV Skills Analysis – Presents an analysis of the skill groups defined in the new ESCO (v1.0.5 and 1.0.8) hierarchy, including an analysis of skills by occupations and the distribution of skills of NEETs. URL: https://epas-dsense.eworx.gr/results/i-cv/skills_hierarchy.html See also: 7. <i>Skills Analysis</i>	Task 3
O.06	iscoCrosswalks library – Methods and classes for crosswalks between the ESCO, O*NET and Canadian NOC classifications of occupations URL: https://github.com/eworx-org/iscoCrosswalks See also: <i>Annex B: Concordances</i>	Task 4

0.07	Exploratory Data Tool – An online tool that allows researchers to explore the study’s datasets by applying custom queries, visualizing and exporting the data. URL: https://epas-dsense.eworx.gr/results/shiny/app/exploratory_data_tool/#!/explore_data <i>See also: Annex D: Exploratory Data Tool</i>	Task 6
0.08	educationR library – A new R library for the classification of multilingual text used in qualification titles was published in CRAN and Github (complementing the existing labour library). URL: https://github.com/eworx-org/educationR <i>See also: Annex C: Library for Classifying Multilingual Text of Qualification Titles - educationR</i>	Task 5; Task 7
0.09	2021-0081-NP-DSI-PHT-ASIA-CV_DataAnalytics&Intelligence-008-20 - Final report – The present final report, presents concrete insights supporting Cedefop in the production of its own publication.	Task 8

2.4 Structure of the Report

This Final Report presents a summary of the work performed under the Service Contract and presents its main findings. More specifically,

- **Chapter 3** presents the methodology applied including an overview of the data sources and statistics employed;
- **Chapter 4** documents the main characteristics of the Europass CV dataset and the biases identified as part of its study;
- **Chapters 5, 6 and 7** present an overview of the analysis of the main areas of concern of the study, i.e., occupations, employment and skills; and
- **Chapter 8** acts as a summary of the key findings of the study as well as a short discussion.

The main chapters of the report are accompanied by four annexes that detail methodological notes related to the analysis, provide documentation related to the software libraries produced (iscoCrosswalks and educationR) and offer a usage guide for the data exploration tool.

3. Methodology

3.1 Overview

An initial attempt for exploratory data analysis (Tukey 1977) of Europass CV data was made in the context of the Europass users' survey, an opt-in survey conducted between June and September of 2019 which accumulated approximately 400K CVs. Cedefop's analysis produced a suite of results, including a **data processing pipeline** based on machine learning methods, an **Insights Report** providing descriptive statistics, and an **interactive application** allowing researchers to explore the standardized dataset. The present analysis serves as a continuation of that attempt on a larger scale. Using the backup Europass database, which includes anonymised data for 10 million users of the Europass CV editor between Q1 2017 and Q2 2020, we aim to expose a number of different statistical techniques, compare with similar measurements present in other official statistical sources, and explore how data of this nature can be harnessed in the context of labour market research. We emphasize the dataset's benefits and drawbacks, as well as potential techniques for researchers and analysts to cope with methodological challenges arising from this type of data, by applying various methods in the context of specific tasks. Additionally, we propose approaches and situations under which this data could be used for research and policymaking, by presenting major components of our study alongside a comparative analysis with the Labour Force Survey and Online Job Vacancies.

For the purpose of this study, we develop a data processing pipeline for data cleansing, standardisation, classification and transformation that leads to standardised aggregated data, along with their codebooks. The main dataset encodes the CV's characteristics (e.g., gender, birth year, country, latest job) and the work experiences reported (e.g., recruitment and termination year). The data cleansing and standardization process is based on the ESCO taxonomy model, as well as custom groups (e.g., age groups) in comparison with other indicators of official statistics. The algorithm used to standardise multilingual free-text of work experiences and skills, is based on document vectorisation, and modified nearest-neighbours classifiers using the ESCO/ISCO hierarchy (see also the published classifier on CRAN, [labourR](#)). Emphasis is given on the usage of open-source technologies for reproducibility by using R and Docker to build the aforementioned pipeline, generate reports, and interactive data applications and develop open-source packages.

3.2 Data Sources

3.2.1 Europass User CVs

CVs created by visitors of the Europass CV editor application serve as the main source of data for this study. We document the characteristics of this data and its use on the present study in *Chapter 4. Understanding the Europass CV dataset*.

3.2.2 Statistical Frameworks and Classification Systems

To enable comparisons between measurements as well as aggregations with respect to each category (e.g., country or occupation) we make use of multiple statistical frameworks and classification systems. These include the ISO-369-1 codes for languages, the slightly modified ISO-3166-1 codes used by the EC for countries (Publications Office, n.d.), and more. Of utmost important to our study are the statistical frameworks related to classification of fields related to labour market research.

The **ESCO classification** is the multilingual classification of European Skills, Competences, Qualifications and Occupations. It identifies and categorizes skills, competences, qualifications and occupations in the context of the EU labour market. It documents relationships between the different concepts, such as the necessity of skills for occupations and also establishes a hierarchy above the

defined occupations based on the International Standard Classification of Occupations (**ISCO**). Namely, there are four levels of hierarchy, ISCO 1 through ISCO 4, in addition to ESCO occupations. Starting with version 1.0.5, skills are also organized hierarchically, allowing different levels of aggregation. As of version 1.0.8, which is the version used in this study, 2,942 occupations and 13,685 skills / competences and knowledge concepts are included in the ESCO classification, with labels and descriptions translated across 27 languages. We use ESCO to classify occupations mentioned as CVs' work experiences, as well as their skills.

The **European Qualifications Framework (EQF)** is a common European reference framework aiming to make national qualifications from countries in the European Union easier to understand and more comparable. EQF covers qualifications at all levels and sub-systems of education and training in Europe and defines eight reference levels tied to specific learning outcomes, ranging from basic (EQF level 1) to advanced (EQF level 8). We implement a machine learning classifier for standardizing qualification titles with respect to level of education based on EQF.

The **International Standard Classification of Education (ISCED)** is a framework for putting together and analysing cross-nationally comparable statistics on education. It is developed by UNESCO and acts as the reference classification for organizing education programs and related qualifications by levels and fields of education. The **ISCED Fields of Education and Training classification (ISCED-F 2013)** was used in this study. Specifically, we accommodate free text matching to the ISCED-F hierarchy, which defines 11 wide fields (2 digits), 29 narrow fields (3 digits), and 80 detailed fields (4 digits).

3.2.3 External Sources of Statistics

The **European Union Labour Force Survey (EU-LFS)** is a large-scale household sample survey regarding the labour participation of people aged 15 and over. It is conducted across all Member States of the European Union, as well as 4 candidate countries and 3 EFTA countries. The national statistical institutes are responsible for data collection on each country, and data processing is carried out by Eurostat. The centralized guidelines for the surveys, conducted by national institutes and the adherence to common standardised classifications allow for the harmonization of data at the European level. Data availability is in the form of indicators specific to the topic covered. Specific indicators from EU-LFS have been selected for comparisons with equivalent measurements in the Europass CV data. Specifically:

- Employment by sex, age, professional status and occupation [lfsa_egais];
- Previous occupations of the unemployed, by sex [lfsq_ugpis];
- Long-term unemployment (12 months and more) by sex, age, educational attainment level and NUTS 2 regions [lfst_r_lfu2ltu];
- Employment by sex, age and job tenure [lfsa_egad]; and
- Job tenure by sex, age, professional status and occupation [lfsa_qoe_4a2];

Cedefop's **Skills in Online Job Advertisements (OVATE)** project brings intelligence/insights on job vacancies based on online sources of job postings. It covers 28 European countries and is a result of analysis of more than 100 million online job as between July 2018 and December 2020. It is meant to serve a complementary role to other sources of official statistics. The online job vacancy data used in this study cover the distribution of online vacancies (mapped to ISCO 1) by country.

3.2 Statistics Employed

3.2.1 Data Cleansing and Standardization

The data pipeline developed for the Europass Survey CV analysis served as the basis for cleansing and standardising data derived from the backup database. A data flow of strict order is specified, and data is transformed from a semi-structured dataset into statistical information. To complement the data flow architecture, we used a reproducible approach based on literate programming (Xie, 2015) (Boettiger, 2015). Each step includes different data cleansing, data wrangling, and information retrieval methods and processes (see **Figure 1**). The tidy data standard (Wickham, 2014) was utilised to make data exploration and analysis easier, and the visualisation pipeline was built using Wilkinson's grammar of graphics (Wilkinson, 2010). Parts of the analysis were influenced by Cedefop's research project on Online job vacancies and skills analysis (Cedefop, 2019).

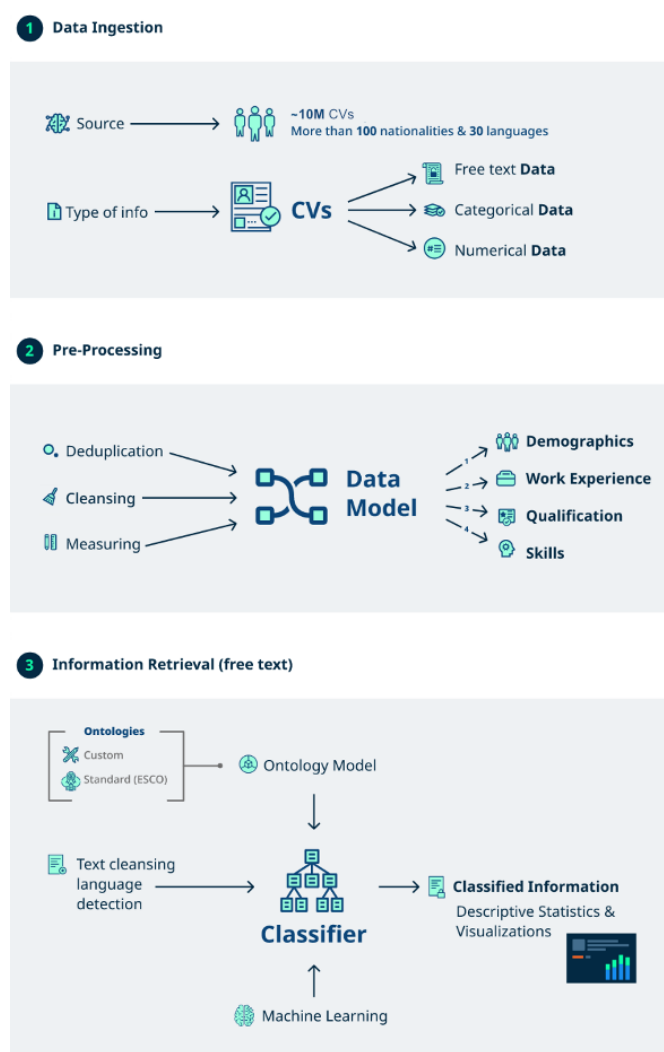


Figure 1. Data processing pipeline schematic

Unlike Europass Survey, where input data was in JSON format, processing data from the backup database required us to facilitate communication with a Microsoft SQL Server. Additionally, the volume of data was increased by a factor of 20 ×, which necessitated changes that ensured scalability through batch processing. The machine learning classifiers matching unstructured free text into the ESCO classification were updated according to version 1.0.8 of ESCO, and the one processing skills was modified to make use of the newly-introduced skills hierarchy. Several new branches were added in the data pipeline, according to the processing needs of each task.

3.2.2 Weighting

Weighting is a statistical technique in which data is adjusted to become more in line with the population being studied. It is typically employed by surveys, but can also be applied to data pulled from databases as an attempt to correct data gaps. It is an important analytical step, as it ensures the target population is fairly and equally represented in the results. Weighting was employed in our study to correct some of the selection bias with respect to age, since Europass CVs are more commonly created by people of younger ages (see **Figure 2**. Age distribution reported in Europass CVs between 15 and 49 years old compared to the equivalent Eurostat demographics indicator.). Additionally, weighting is used to create a European average that is not dominated by a small number of countries that are overrepresented in the dataset (e.g., Italy and Portugal). The European average used for reporting within this report is based on the 19 countries of the Euro area (EA-19), which are the most well represented within the dataset. Weighting is performed with a target population based on Eurostat's demographics dataset. The weights derived were restricted between lower and upper bounds set to 0.35 and 3 respectively.

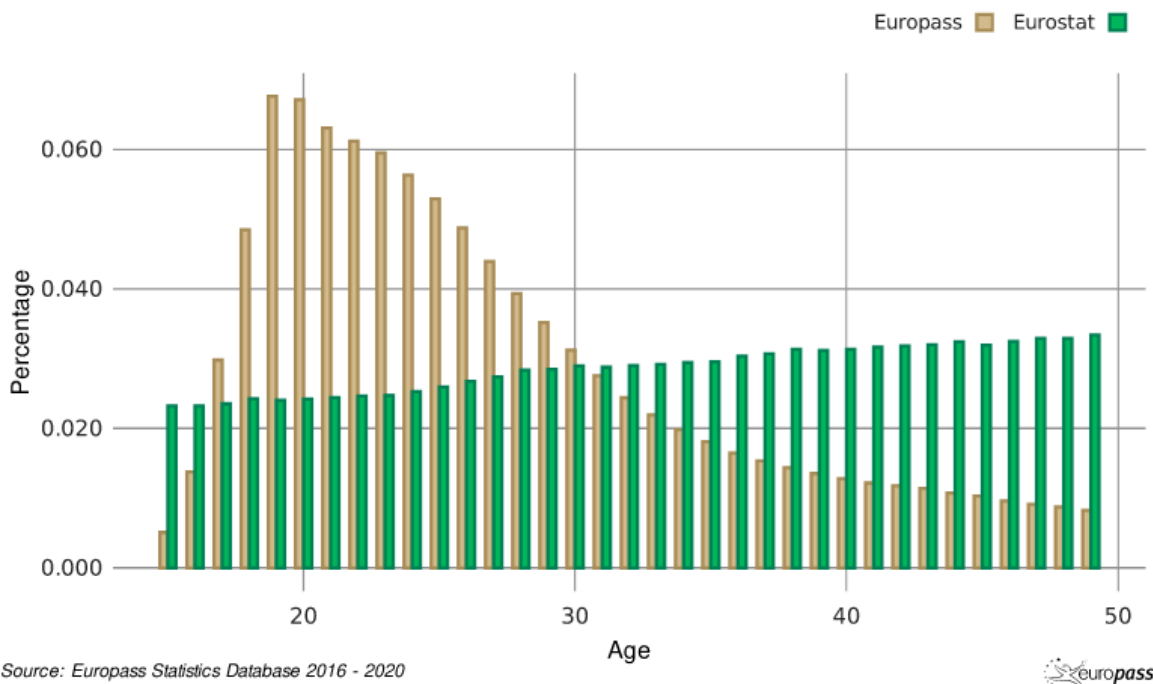


Figure 2. Age distribution reported in Europass CVs between 15 and 49 years old compared to the equivalent Eurostat demographics indicator.

Throughout the analysis, calculations were performed both on weighted and unweighted versions of the data. While bias remains encoded within the dataset, the weighting itself, as well as comparisons of results before and after weighting as well as with external indicators provide some quantitative feedback on how biases with respect to age and country of residence affect measurements. Unless otherwise noted, results presented in this report are weighted. Comparisons between weighted and unweighted results are presented in certain chapters (e.g., Chapter 6. *Cross Checking with Official Statistics*). Note that weighting was only performed on our primary source of CVs derived from the Europass backup database, and not on CVs from the Europass Survey.

3.2.3 Time-Series Analysis

Several time sequences can be defined based on fields in Europass CVs. For example, CVs include the date of recruitment and termination which is attached to a work experience along with the job title, dates related to attainment of education, and the year of birth. Through these variables, a number

of different time series emerge, allowing us to encode patterns and trends on the dataset. To gain a quantitative insight on trends that emerge with respect to variables such as recruitments per occupation, we perform regression analysis to these time series (Chambers, 1992). See *Annex A: Methodological* for more details.

Note that the web traffic of the Europass CV editor application throughout the period of data collection have been taken into account. Any attempt to measure the change of raw counts in time will result in observing the effect of web traffic. For that reason, all analysis is focused on the measurement of relative change between categories (e.g., the year-to-year percent increase/decrease of an occupation on the dataset compared to other occupations) or ratios between two different types of counts (e.g., recruitments over terminations as a function of time).

3.2.4 Correlation and Regression Analysis

The term correlation, often known as correlation analysis, is used to describe the link or association between two (or more) quantitative variables. The assumption of a linear relationship between the quantitative variables underpins this study (James, Witten, Hastie, & Tibshirani, 2013) (Chan, 2018). It assesses the "strength" or "extent" of an association between the variables, as well as its direction. A correlation coefficient, with values ranging from -1 to +1, is the end result of a correlation analysis.

A correlation coefficient of +1 indicates that the two variables are perfectly related in a positive and linear manner, while a correlation coefficient of -1 indicates that the two variables are perfectly related in a negative and linear manner, and a correlation coefficient of zero indicates that the two variables being studied have no linear relationship. As an empirical rule, a coefficient greater than 0.8 indicates a fairly strong positive relationship. Furthermore, by modelling the linear relationship, linear regression provides more insight into correlation by estimating the slope. The later, measures the rate of change of one variable with respect to the other, or else the steepness and the direction of the line. These tools are useful for comparing measures in the dataset to official indicators, as well as modelling the relationship between quantitative fields and answering queries like how age and job experience are related.

3.2.5 Association Rules

Association rules or otherwise referred to as Market Basket analysis, is a popular and well researched topic (Piatetsky-Shapiro, 1991) (Agrawal, Imieliński, & Swami, 1993) for knowledge discovery in databases. Each CV can be considered as a "transaction" that the CV owner performed with the itemset being the fields that the CV owner added in the CV (basket), such as job experiences, skills, demographic characteristics, qualifications etc.¹ These records can provide valuable insights and by applying this method we can retrieve existing CV patterns and also suggest new relations between categorical variables of interest for labour market intelligence.

Given the standardised dataset that resulted from the machine learning classifiers, association rules can be mined for the "itemset" of the ESCO classification model for skills and occupations. For example, when some ESCO skills are observed together with particular occupations in CVs more often than one would expect from their individual rates of occurrence, it can be said that this co-occurrence is an interesting pattern. These patterns can be evaluated by comparing them with the actual associations provided by the ESCO classification. The analysis can be continued by adding in the

¹ Even if the term "transaction" seems strange in this context, all data points that co-occur are considered to represent a transaction. Another example transaction could be the set of web pages that a user visits during a session (Chapman, 2015).

itemset any categorical field extracted from the database (e.g., age groups, qualifications etc.). See *Annex A: Methodological* for more details.

3.2.6 Revealed Comparative Advantage (RCA)

In order to identify what skills are truly relevant to each occupation, we measure how over- or underexpressed each particular skill group is for each occupation. For this purpose, we make use of the the revealed comparative advantage (*RCA*) index. See *Annex A: Methodological* for more details.

4. Understanding the Europass CV dataset

4.1 Overview

The Europass CV editor is an online application that allows visitors to create, store and share their curriculum vitae. It is part of the Europass initiative, which aims at increasing transparency of qualification and mobility of citizens in Europe. Cedefop developed and maintained Europass and all of its infrastructure until the second quarter of 2020, including the backup database of the Europass CV editor, which consists of anonymized data for over 10 million who created their CV between 2017 and the second quarter of 2020. CVs included on the backup database encompass demographic data such as the CV owners' gender and country of residence, information about their work experiences and qualifications, as well as some of their reported skills (see **Table 1**). In conformity with EU data protection provisions, information that can be used to identify CV owners (e.g., free-text entries of skills) has not been preserved within the database throughout this period.

The Europass backup database serves as the primary source of data throughout this study, and following a data cleansing and standardisation process, the information contained was transformed into aggregated datasets. Subsequently, an exhaustive exploratory data analysis within the scope of understanding the Europass CV data was performed, utilising a number of different disaggregations, statistical models and methods to examine the dataset. We extrapolated quantitative feedback on biases, evaluated the effectiveness of standard statistical methods, and identified representative breakdowns so as to gain a granular picture on the labour force supply.

A secondary source of data came from the Europass survey, a voluntary survey conducted between June and September of 2019. Almost 400,000 Europass visitors participated throughout its duration. The collected CVs were processed and transformed into anonymised aggregated datasets. CVs from the user survey included free-text entries (e.g., for skills) that were not preserved in the Europass backup database, resulting in smaller datasets which, however, included some additional fields of information. Once again, in accordance with EU data protection provisions, all personal data was removed 6 months after the completion of the survey. Unless otherwise stated, results from the present analysis refer to the primary source of the Europass backup database, but sections presenting an analysis of skills make use of the Europass survey data.

Table 1. Completion Rate of CV Fields by Pillar

Pillar	CV Field	Completion Rate
Demographics	CV Language	100%
Demographics	Creation Date	100%
Demographics	Country	96%
Demographics	Birth Year	50%
Demographics	Gender	44%
Work Experiences	Recruitment Year	97%
Work Experiences	Termination Year	95%
Work Experiences	Job Title (Label)	75%
Work Experiences	Job Title (ESCO classification)	22%
Work Experiences	Employer	8%
Qualifications	Qualification Title (Label)	91%
Qualifications	Enrolment Year	90%
Qualifications	Organisation Country	79%
Qualifications	Graduation Year	79%
Qualifications	EQF Level	12%
Qualifications	Educational Field (ISCED-F)	3%

4.2 Working with Naturally-Occurring CV Data

Following the cleansing and standardization process of the CVs stored in the Europass CV editor application's database, what remains is a dataset that encodes a large and rich amount of information. This section aims to document some of this dataset's characteristics and caveats that need to be considered before using it to generate labour market intelligence, as well as when interpreting any intelligence derived from it. As the dataset is not a result of randomized sampling, its major flaw is the bias with respect to the coverage and representativeness of the labour force that exists at the aggregate level. CVs generated online do not fully represent supply in the labour force, and Europass in particular is not equally used across the different breakdowns of the labour force. As such, the composition of the labour force characteristics encoded in CVs generated online is likely to be different from reality. These problems of bias and coverage exist for all online sources of information on the labour market from the standpoint of supply as well as demand, and have also been observed in the literature of online job vacancies. (Pouliakas, 2021)

The examined CV data needs to be understood as naturally-occurring, that is to say, data that is not elicited or generated as part of a research project. (Potter, 2002) Moreover, the data is specifically generated through users' interaction with an online tool and its quality is thus subject to user behaviour. Compared to a formal survey, information gathered through the Europass CV editor is likely to have high non-response or incomplete response bias. This observation can be attributed to a number of different reasons, including incomplete CVs left for editing later and never revisited, users leaving in the middle of the process due to not finding the editor compelling enough, and more. Additionally, the Europass CV editor in particular stores multiple revisions of the same CV, making it difficult to disambiguate between complete CVs and noise. To reduce some of the noise generated by incomplete CVs, a deduplication process was performed, and only the latest CV associated with each unique reported email address was considered on the analysis.

Since most fields on the Europass CV editor are optional, completion rate for specific information may vary by user and by pillar, even after deduplication. Given that most CVs are created with the intent of applying for a specific job, a user may choose to not disclose certain information with their potential employer (e.g., their age and sex) and may not consider every field as relevant for their application. Additionally, certain fields may require familiarity with standards (e.g., the European Qualifications Framework) that many users may not be informed about and may leave blank or fill erroneously (see **Table 1**). Some of the missing data (e.g., EQF level and ESCO occupations) in the Europass dataset was imputed in the cleansing process based on related information shared on other fields. Given the high dimensionality of the dataset, missing data is common and its treatment depends on the level of aggregation and the type of query being made.

One more source of bias that is inherent to studies that rely on self-reporting and is especially prevalent on a source of self-created CVs, is response or recall bias. The accuracy and completeness of users' recollection of their past work experiences cannot be guaranteed, and details are expected to be omitted, especially for CVs with extensive work experience. Moreover, pre-existing beliefs may impact the types of work experiences shared (e.g., past work experienced that are not relevant to a user's current career trajectory may not be included) and given that most CVs are completed by job seekers, they will likely elect to present themselves more positively to their potential employers. This can result in an increase or decrease in the strength of observed associations and can potentially lead to imprecise assumptions. One instance where recall bias can become an issue is when measuring labour market trends across a long period of time based on users' career history. Selecting an appropriate recall period is key to minimizing recall bias (Althubaiti, 2016) and in the case of the

Europass CV dataset, the shorter and more recent the period explored, the more reliable the results are expected to be.

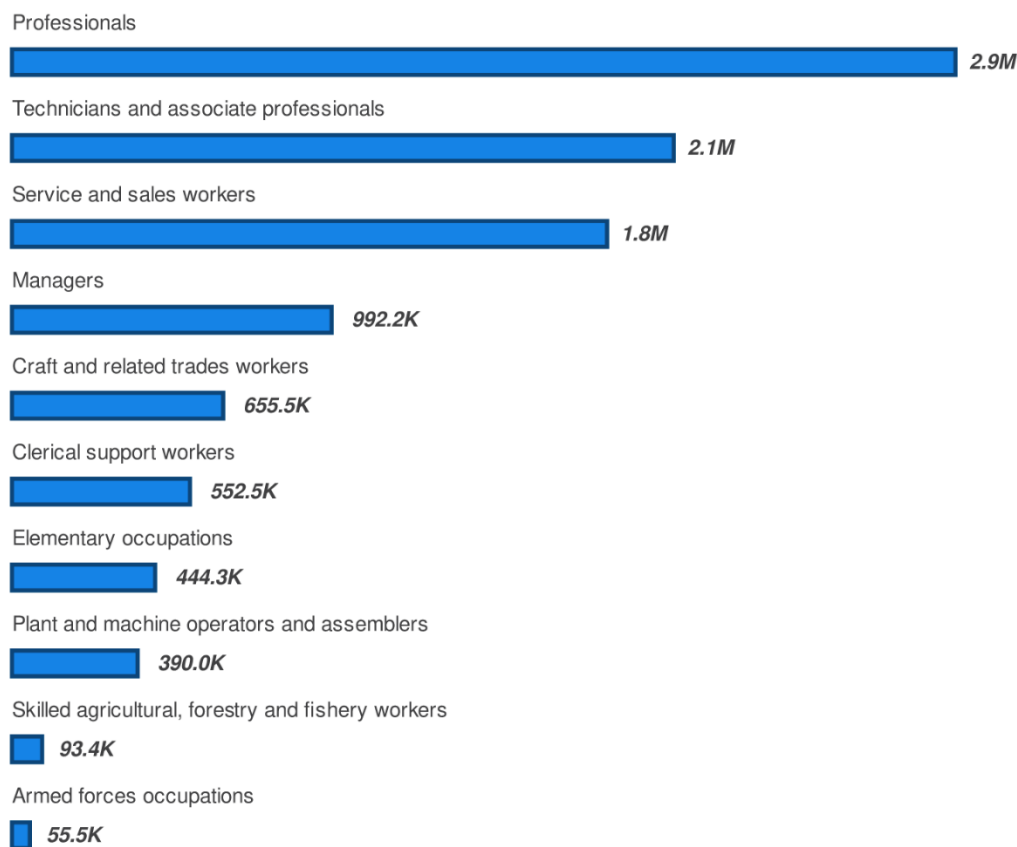
4.3 Technological Diffusion

Given the period of data collection between 2017 and 2020, a source of time-varying bias on the dataset is technological adoption/diffusion. A constant drift from “offline” resources (e.g., offline text editors) to free and/or low-cost online tools (Europass, LinkedIn, Indeed, etc.) takes place as a result of the wide adoption of the Internet / World Wide Web both from the point of view of job vacancies, and on the supply side of the labour force. Services developed by large organizations for CV submissions are expected to have better integration with popular online editors. For example, online forms for a vacancy of a large multinational company may be automatically filled upon the submission of a CV that is generated by a popular online editor, leading the user base of that service to grow.

From the standpoint of supply, (Barnichon, 2010) demonstrates that this shift in vacancy coverage closely resembles the S-shape typical of technology diffusion in the United States, as well as the similarly S-shaped fraction of internet users in that country. Similar results are expected for online CVs, leading to increasing coverage in time. This outcome is an additional source of time dependent bias, due to the rapid technological diffusion among the different breakdowns. For example, cooks in 2020 may use online CV editors more frequently than cooks in 2017, showing a false relative increase of cooks within this time span. To some extent, this effect can be smoothed out by comparing ratios, relative changes, or by introducing time dependent reweighting.

4.4 User Composition Compared to the Labour Force

The extent to which the composition of users’ characteristics is biased relative to the composition of the labour force, is hard to quantify, even for well represented categories in the dataset. The creation of a CV is most often linked to an individual’s intention to apply for a job; but it is expected that many vacancies in the job market are filled without submission of CV, for example, in the case of applications in smaller and family businesses or local stores where the interview process may be limited to an on-site conversation. Additionally, a considerable portion of the labour force will use offline tools, or even create a handwritten CV. Meanwhile, some job vacancies may lead to custom CV editors developed by their respective organizations, in which case a CV created on Europass will not be of use. Conversely, the promotion of the service is disproportional across different subsectors of the market, which may lead to overrepresentation of some and underrepresentation of others. More importantly, due to the features the Europass CV editor provides, it is anticipated to be used more frequently by users professing specific occupation types.



Source: Europass Statistics Database 2016 - 2020

 europass

Figure 3. Distribution of Work Experiences in the Europass CV Dataset with Respect to ISCO 1.

Due to the above circumstances as well as the fact that adoption of internet usage varies across different breakdowns, it is likely that middle and high skilled segments of the labour force use the Europass CV editor more. To this representativeness bias adds the fact that middle and higher skilled vacancies demand higher quality CVs, motivating candidates to use online editors that better meet the requirements. Following the cleansing and standardization process, we find that with respect to ISCO 1 digit classification, Professionals and Technicians and associate professionals represent the most commonly reported work experiences in the Europass dataset across the board (see **Figure 1**). Comparing the pool of users that were employed at the time of CV creation with the respective labour force statistics as presented on the Labour Force Survey (Eurostat, 2020) on the respective time frame, we find that the observed frequency distribution of occupations with respect to ISCO 1 is significantly different from the distribution in the labour force. The overrepresentation of Managers and Professionals make the greatest contribution to this discrepancy as evidenced by the analysis of the chi-square test residuals (Sharpe, 2015), while Craft and related trades workers, Clerical support workers, and Elementary occupations are especially underrepresented (see **Table 2**). These imbalances are even more pronounced in age group 15-24 (see **Table 3**). It should be noted that this finding is a differential representativeness and is expected to change in time, as online tools become more accessible and widely used, leading even lower-skilled positions to increase their CV quality requirements.

Table 2. Chi-squared Test and Residuals for Users in Age Group 15-24 in EA-19 Countries

$\chi^2 = 29138.18, df = 8, p < 2.2 \times 10^{-16}$			
Occupation (ISCO 1)	Observed Count	Expected Count	Standardized Residual
1 Managers	1769	124.15	148.13
2 Professionals	4340	1595.00	71.95
3 Technicians and associate professionals	3824	3150.86	13.19
4 Clerical support workers	961	1939.00	-23.5
5 Service and sales workers	4486	5182.74	-11.44
6 Skilled agricultural, forestry and fishery workers	227	361.48	-7.14
7 Craft and related trades workers	1071	2624.83	-32.78
8 Plant and machine operators and assemblers	651	1067.41	-13.14
9 Elementary occupations	886	2169.52	-29.36

Table 3. Chi-Squared Test and Residuals for Users in Age Group 25-49 in EA-19 Countries

$\chi^2 = 5841.72, df = 8, p < 2.2 \times 10^{-16}$			
Occupation (ISCO 1)	Observed Count	Expected Count	Standardized Residual
1 Managers	3444	1781.80	40.46
2 Professionals	10316	7206.16	41.28
3 Technicians and associate professionals	7342	6267.93	15.03
4 Clerical support workers	1642	3500.18	-33.17
5 Service and sales workers	5684	5576.42	1.58
6 Skilled agricultural, forestry and fishery workers	370	642.35	-10.85
7 Craft and related trades workers	1926	3758.41	-31.7
8 Plant and machine operators and assemblers	1406	2304.86	-19.39
9 Elementary occupations	1767	2858.88	-21.34

Another fact that adds up to the representation bias is that internet resources such as the Europass CV editor are more likely to be used by younger ages. Moreover, younger people are naturally more likely to be job seekers. Those two biases are entangled and lead to an overrepresentation of younger people in the dataset. Specifically, we find a mean age of 29, with 72% of users being under 32 years old (see **Figure 2**). Restricting the analysis to a narrow range of ages will likely elicit better results. Weighting with respect to age may smoothen out this bias, but it can also add bias related to unemployment in young people, leading them to build CVs more frequently than other age groups. This opposing effect also changes in time since internet usage is increasingly adopted across all age groups. Hence, bias with respect to age is also expected to decrease in time due to the wide adoption of online resources, but it obscures the picture of any time-series analysis of the data collected in the current study.

Finally, the Europass CV editor application is available internationally, but finds more adoption in countries on the European Union. Despite this fact, it is not evenly adopted across all countries. We find that Italy and Portugal are the top countries within the data, while countries like Denmark and Poland have an especially small sample size given their population (see **Figure 3**). Deriving statistics for countries with a larger sample size is generally more feasible with this type of dataset. Weighting can also be used to produce a general European indicator. We derive a restricted weighting scheme

based on countries in EA-19. Note that some country-specific bias is still encoded within that (see 3.2.2 Weighting).

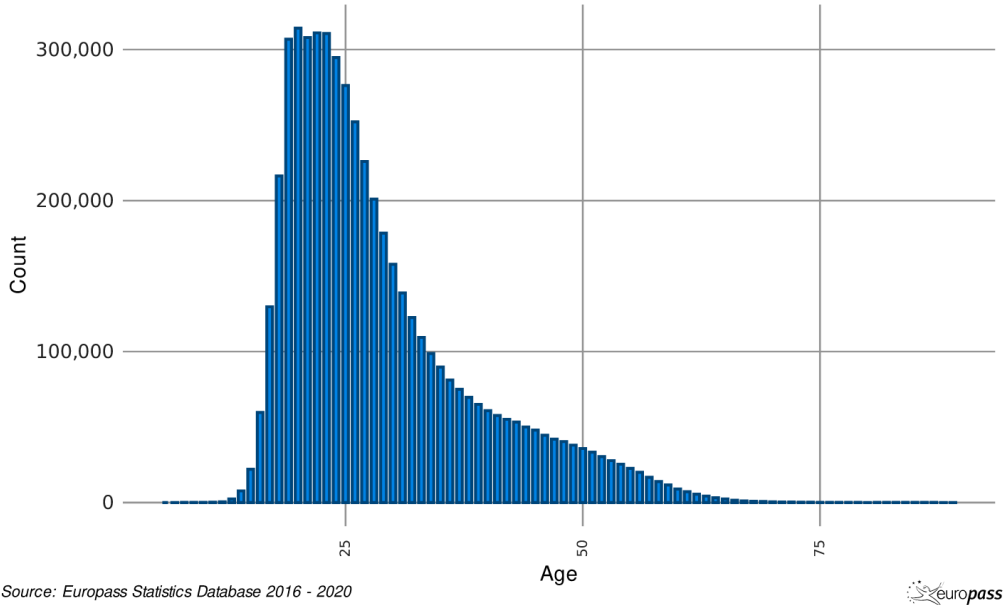
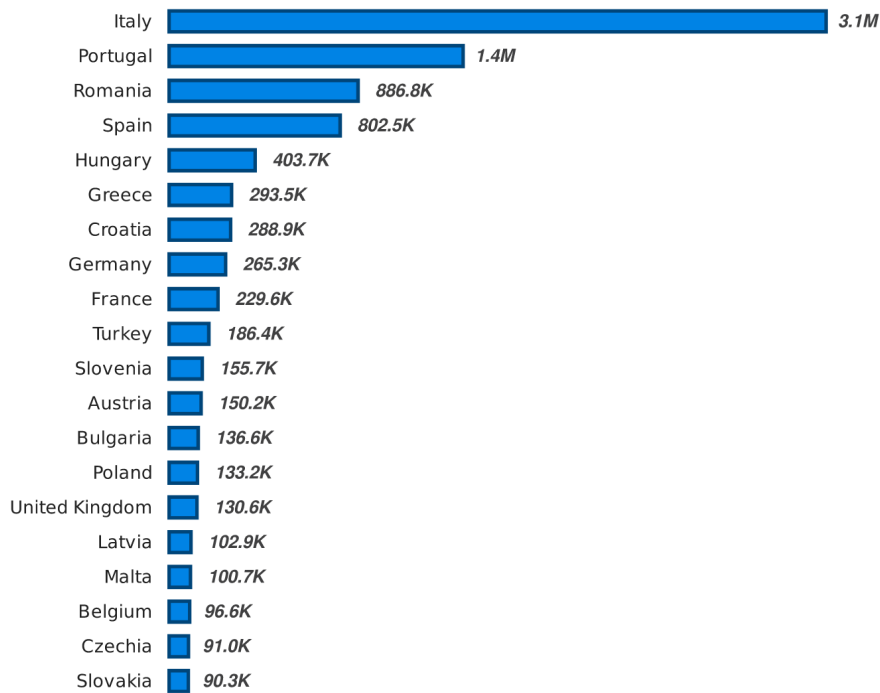


Figure 4. Age Distribution of Europass Users



Source: Europass Statistics Database 2016 - 2020



Figure 5. Distribution of Country of Residence among Europass Users

5. Occupations Analysis, Trends and Correlations

The Europass CV editor backup database includes CVs created by users between 2017 and the second quarter of 2020. Each CV may incorporate one or more work experiences a CV owner has had throughout their career. Metadata fields such as the years of recruitment and termination are attached to every work experience, allowing us to introduce a time component in the analysis. Following our data cleansing process, work experience is matched to the ESCO classification. We employed several different statistical transformations and techniques to measure trends and analyse patterns in the participation of the different subgroups of users in the labour market. We also used our secondary source of Europass Survey CVs to explore correlations and associations between occupations and skills included in user CVs (note that this exercise is further expanded in *Chapter 7. Skills Analysis*). Reporting of descriptive statistics in the following sections is undertaken with respect to ISCO 1, 2, or 3. Unless otherwise noted, users between ages 15 and 49 were considered, and the European averages reported are based on the EA-19 following a weighting procedure (see 3.2.2 *Weighting*).

5.1 Disappearing and Newly-Emerging Occupations

In order to identify growing and disappearing occupations within a defined period covered by the dataset, we calculated the frequency distribution of recruitments per occupation by year and performed time-series analysis. We fit a linear model whose slope is an indicator of the relative increase/decrease of an occupation in time. Through this process, we identify occupations with increased presence further back in time compared to more recent years and vice versa. Moreover, we show how the relationship between recruitments and terminations changes in time and how it differs for different occupations. This approach of comparing ratios helps to smooth out some of the biases encoded within the dataset given a longer time-span. Throughout this exercise, recruitments and terminations between 2000 and 2019 were considered based on Europass users' reported work experience histories.

5.1.1 Trends in Job Sectors

We performed time series analysis on recruitments per occupation by year.

- This approach results in the calculation of the odds ratio increase/decrease in recruitments per occupation for one unit in time.
- An odds ratio increase suggests that on average, there is an increase in recruitments for a given occupation relative to other occupations, while an odds ratio decrease suggests the opposite.
- As such, it is possible to identify which occupations trend upwards, which trend downwards, as well as those that show no major change.
- The exercise was repeated for EA-19 as well as individual countries, and for both weighted and unweighted data.

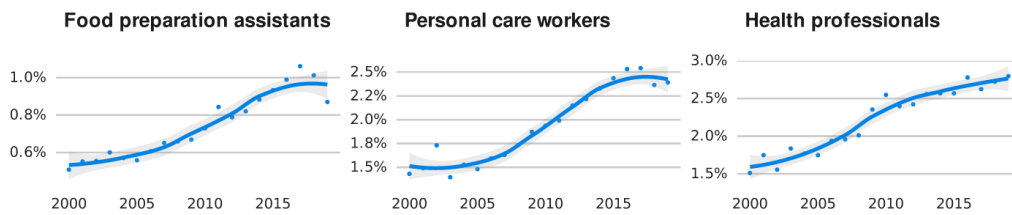


Figure 6. Example of ISCO 2 occupations with an observed relative increase in recruitments in time. (EA-19)

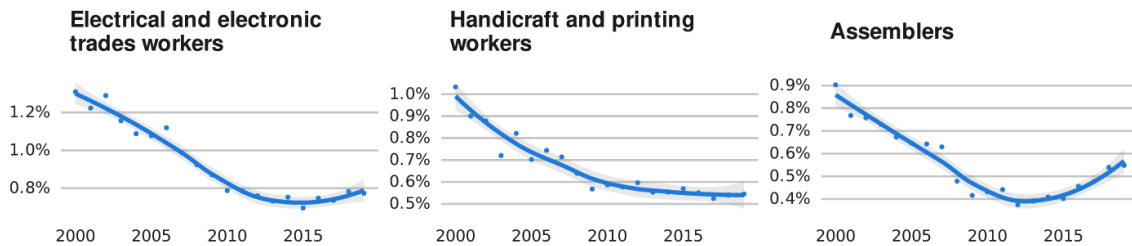


Figure 7. Example of ISCO 2 occupations with an observed relative decrease in recruitments in time. (EA-19)

- The presence of ISCO 2 occupations such as *Food preparation assistants*, *Personal care workers*, *Health professionals*, and *Personal service workers* in the dataset increases with time with respect to reported recruitments.
- A relative decrease is observed for ISCO 2 occupations like *Electrical and electronic trades workers*, *Assemblers*, *Building and related trades workers*, and *General and keyboard clerks*.

Discussion: Recruitments for occupations commonly professed at the start of a person’s career, such as *Personal service workers* and *Food preparation assistants* display an increase in more recent years. One likely explanation for this observation may be the fact that as a person’s career advances and they move to different jobs, they are less likely to include their early work experiences on their CV. This assumption relates to recall bias which is inherent to data of this nature. Looking at the reported career history, we can derive an idea of the labour market in years prior to the period of data collection, but it will be an incomplete picture.

5.1.2 Highest Degree of Deviation in Trend per Country

We compared the measurement of odds ratio percent change per occupation in EA-19 with the equivalent country-specific ones.

- By doing so it helps identify the occupations of each country whose recruitments display the most significant deviation in year-to-year increase/decrease in the dataset compared to the average.
- Greater deviation from the average suggests stronger patterns of increase/decrease in recruitments for a specific occupation in that country.

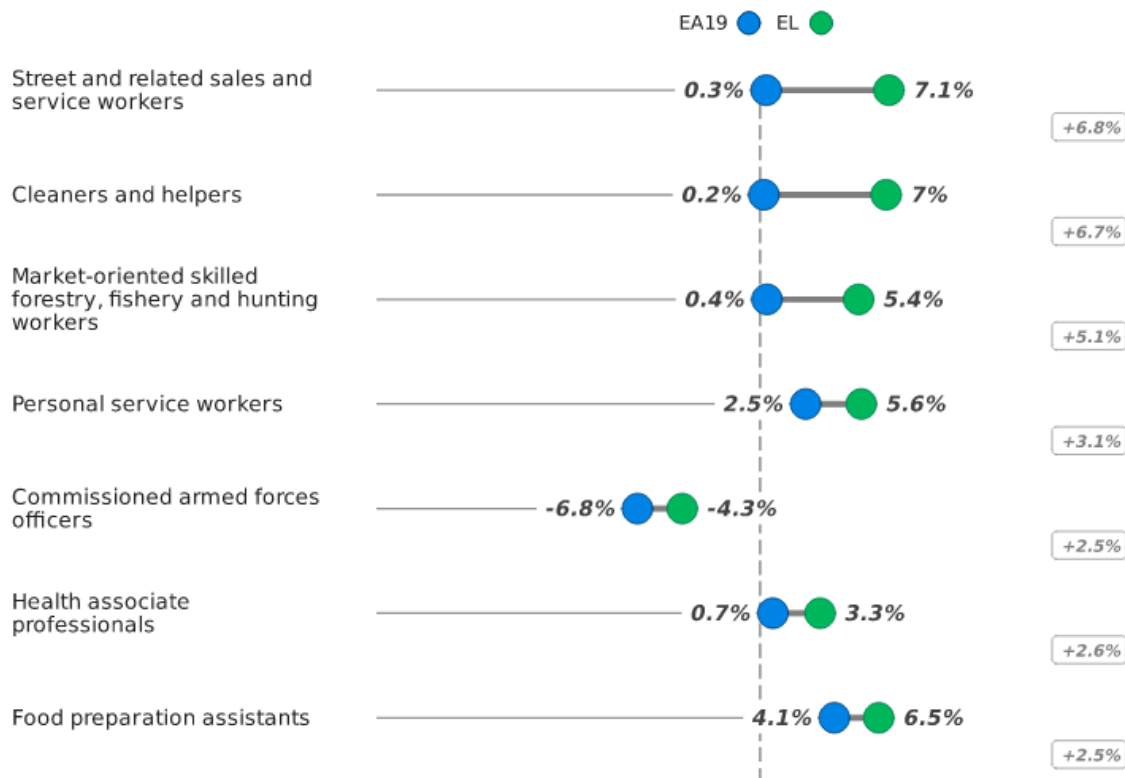


Figure 8. Example of the ISCO 2 occupations of Greece that display the highest deviation in yearly odds ratio change compared to the EA-19 average.

- Countries that rely heavily on tourism, such as Greece, Croatia, Portugal and Spain, display more positive trends in jobs related to tourism, such as ISCO 2 occupations *Personal service workers* and *Food preparation assistants*.
- On the other hand, ISCO 2 occupations related to the heavy industry, such as *Assemblers* and *Labourers in mining, construction, and transport* have increasing presence in Germany.

Discussion: Throughout a period of relative stability, such as the one studied in this analysis, it is anticipated that the core industries of each country will exhibit growth as opposed to decline. It is thus not surprising that jobs in strong industries of their respective countries display positive trends.

5.1.3 Recruitments over Terminations in Time

Another way of measuring the growth or decline of occupations is by exploring how the ratio between recruitments and terminations of employment changes in time. CVs of Europass users consist of work experiences that include a start date and, unless it is an ongoing job position, an end date, making this type of measurement per occupation possible.

- We record this relationship between recruitments and termination in the so-called net hire ratio:

$$\text{Net Hire Ratio} = \text{Total New Recruitments} / \text{Total Terminations}$$

- We study the evolution of the net hire ratio with respect to occupations and countries.
- Note that by using the term *Net Hire Ratio*, we refer to the ratio as it emerges from Europass CVs, and not in equivalent metrics that may emerge within the scope of a corporation or the real labour market of a country.

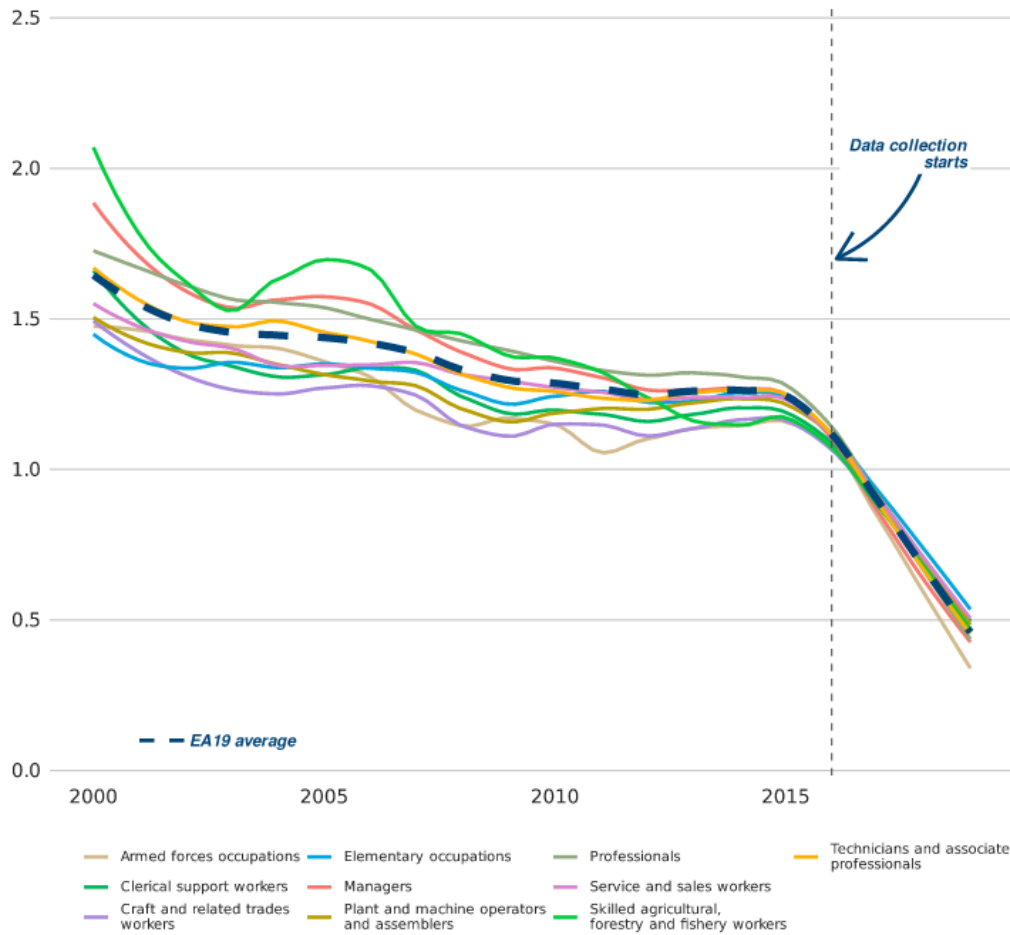


Figure 9. Evolution of net hire ratio for ISCO 1 occupations. (EA-19)

- For the period before data collection commenced, more recruitments than terminations are generally reported in Europass CVs, thus bringing the net hire ratio’s value over 1.0.
- Commencing with the year data collection begins, net hire ratio starts to dip below 1.0.

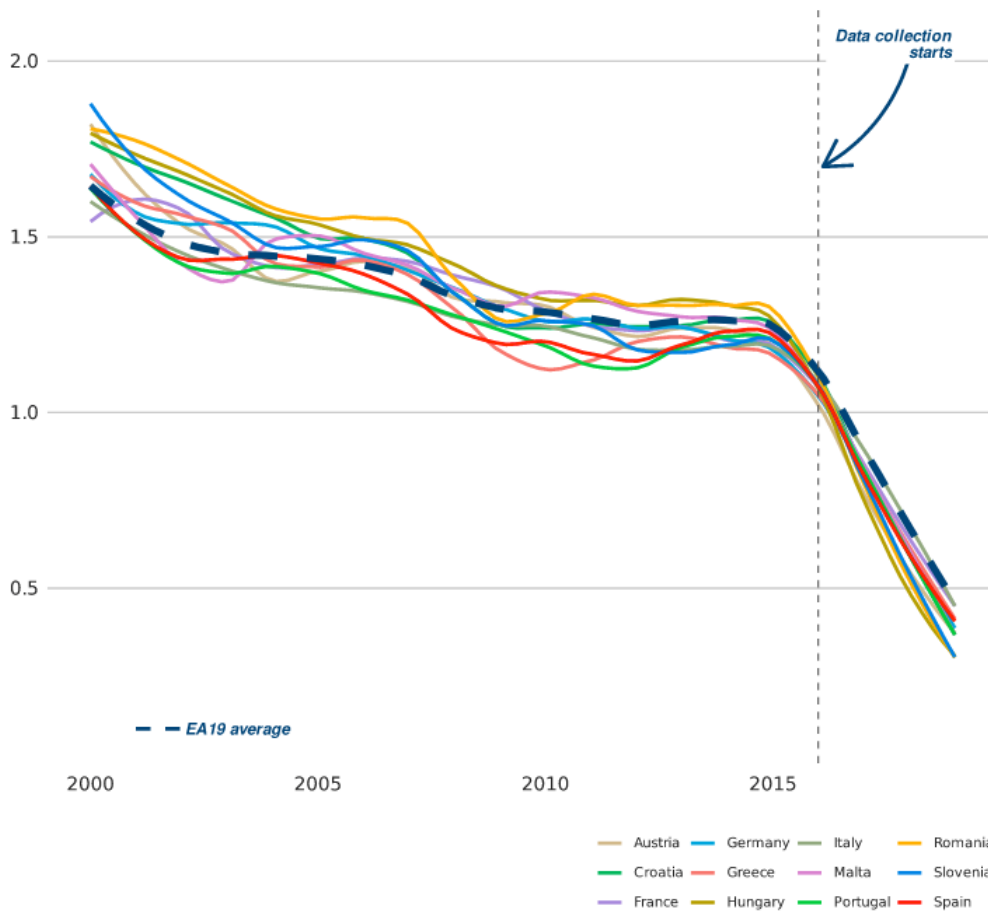


Figure 10. Evolution of net hire ratio across countries.

- Comparing each distinct line with the EA-19 average, phenomena local to specific countries may be observable in the countries graph.
- Notable is the sudden decline of the net hire ratio in countries of southern Europe, such as Greece, Spain and Portugal between the late 2000’s and the early 2010’s.

Discussion: As with previous results (see Chapter 5.1.2 Highest Degree of Deviation in Trend per Country), it is anticipated that throughout a period of relative stability, the labour market will experience growth, and thus more recruitments than terminations will be reported, as reflected between 2000 and 2016.

On the other hand, the sudden dip after CV collection begins may be explained by the fact that the application is more likely to be used by unemployed people searching for a job at the time of CV creation than employed ones who have recently started a new job. This effect is likely inherent to labour market analysis through naturally-occurring data of this nature, strongly indicating that this dataset is closer to the side of the supply of people actively seeking new employment.

It is also observed that the net hire ratio tends to be higher for years further away from the beginning of data collection. This effect may once again be related to the nature of the dataset, which tends to include people searching for new jobs and who are thus more likely to report more work experiences completed closer to the time of CV creation.

5.1.4 Mean Net Hire Ratio of ISCO 2 Groups per Country

In addition to studying the evolution of net hire ratio in time, we can also measure the mean net hire ratio of each occupation for a set period of time.

- This approach gives us a numerical indicator of the relationship between recruitments and terminations across the different occupations of each country.
- As more recruitments than terminations are noted overall, comparisons between these numerical indicators can reveal insight for which specific ISCO occupations exhibit relatively more growth.
- We measure mean net hire ratio in the period between 2000 and 2016, as previous analysis has shown that years subsequent to data collection are affected by the biases inherent to CV creation.



Figure 11. ISCO 2 occupations displaying the highest mean net hire ratio. (EA-19)



Figure 12. ISCO 2 occupations displaying the lowest mean net hire ratio. (EA-19)

- *Legal, social and cultural professionals, Health professionals, and Teaching professionals* are the ISCO 2 occupations displaying consistent growth across most European countries for the period studied.
- *Building and related trades workers, Assemblers, and Electrical and electronic trades workers* are among the ISCO 2 occupations that display the least amount of growth, meaning that compared to other ISCO 2 occupations, their number of recruitments is closer to their number of terminations.

Discussion: Generally, professions requiring more specialisation or more investment in education have the highest ratio between recruitments and terminations. This finding may suggest a rise in demand for highly specialised professionals, with new job positions opening faster than old ones are closing.

Many of the occupations that display a comparatively smaller ratio of recruitments and terminations are related to manual labour. One possible explanation for this finding is the increasing automation of manual work, and thus less job positions related to those fields open.

5.1.5 Next Steps

In general, in addition to the actual labour market trends that may be present, the measurements reported may also encode differences in the patterns of behaviour of individuals working across

different ISCO groups when it comes to CV creation. Continued research on those differences can shed light to the actual trends in the real labour market. Other questions that can be explored is how the recruitment / termination ratio is related to the actual net hire ratio in the real market.

Furthermore, calculation of the same trends for breakdowns with estimated high coverage (e.g., “*Professionals 25-35 in Italy*”) may lead to generalizable results. Exploration of such queries will be possible through the interactive data tool.

5.2 Changing Skill Requirements for Job Positions

Along with work experiences, Europass CVs include a number of skills. These skills are included as free text, which is matched to a classification in the ESCO model through our data cleansing process. As with work experiences, skills are aggregated along with specific associated demographic fields, such as the CV owner’s year of birth. One way to measure how the skill requirements of each job are changing, is to compare skills reported by younger users with skills of older users who report the same job position. As such, we calculate the birth year (or equivalently, age) distribution of each skill per ISCO occupation and perform regression analysis. In this case, the slope of the fitted statistical model is an indicator of the relative increase/decrease of the inclusion of a skill as the year of birth increases. As skill-related free text is not preserved in the Europass backup database, we make use of the secondary source of user CVs from the Europass Survey for this exercise. Note skills matching in this section is undertaken based on a version of the ESCO classification prior to the introduction of the skills hierarchy (which is utilized on Chapter 7. Skills Analysis).

5.2.1 Skills by Occupation and Birth Year

The time series analysis performed provided a measure of how the frequency of skill inclusion changes with birth year.

- We are able to identify skills more frequently included by older and others by younger people by ISCO occupation.
- The process was repeated for both the matched ESCO skills and keywords included in the free text individually for each ISCO 1 occupation, based on the users’ latest job.
- Ages between 15 and 49 were considered, as sample size of lower and higher ages was too low.

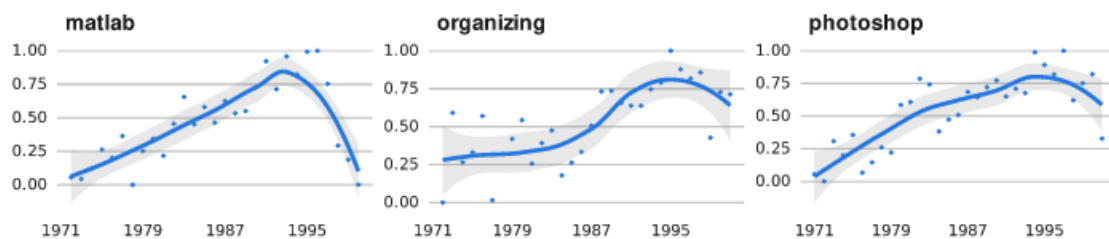


Figure 13. Example of keywords more often included in younger users’ skills with latest job Professionals.

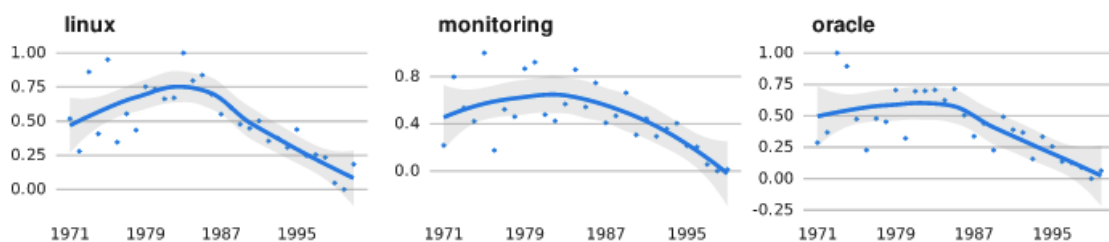


Figure 14. Example of keywords more often included in older users' skills with latest job Professionals.

- Keywords such as “photoshop”, “python” and “matlab” display a positive trend for ISCO 1 group Professionals with respect to birth year, meaning that they are more commonly reported by younger people compared to older ones, who mention keywords like “oracle”, “application”, and “security”.

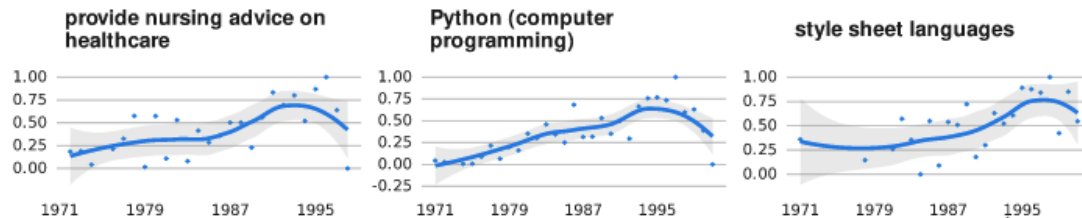


Figure 15. Example of ESCO skills more often entered by younger users with latest job Professionals.

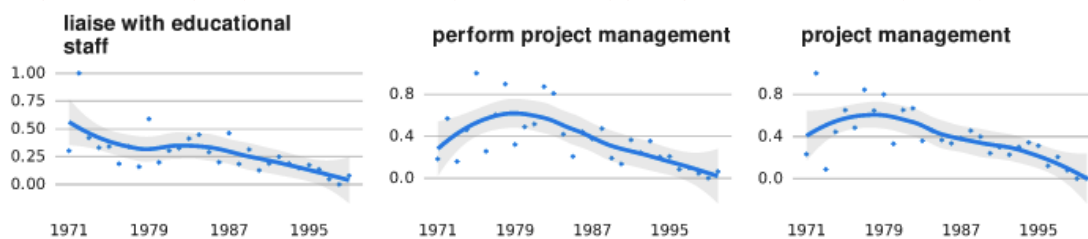


Figure 16. Example of ESCO skills more often entered by older users with latest job Professionals.

- Managerial ESCO skills (e.g., *project management*, *draft corporate emails*) are more frequently observed among older people, while ESCO skills related to childcare and students (e.g., *assist children with homework*, *communicate with youth*) are more frequently observed among younger people.

Discussion: Some low-resolution patterns related to the use of media editing software in younger ages and managerial skills in older ones can be noted for CVs where the most common occupations are reported (e.g., *Professionals* and *Technicians and associate professionals*). However, time series analysis cannot expose adequate insight in this case, as on top of the multilingualism problem, the overwhelmingly large number of ESCO skills, and the noisy nature of free-text skills data, the Europass survey provides a relatively small volume of data.

5.2.2 Next Steps

One element that may assist future work on this task is aggregation of the ESCO skills into meaningful groups. Starting with the version 1.0.5, ESCO has introduced a hierarchy that organizes its 13,000+ skills into respective groups, similar to how ISCO acts as a hierarchy for the almost 3000 ESCO occupations. This hierarchy has not been utilised in our current free-text matching algorithm. Adapting our approach in further research, can lead to improvement of text-mining itself (thanks to the richer corpus), but also better aggregation of skills (e.g., a total number of around 400 skills, instead of 13,000). Our next report focuses on skill analysis by using this hierarchy.

5.3 Usual Career Paths

Career paths of Europass users were examined using two main approaches. The first approach performed a series of statistical transformations and aggregations to gain insight on the frequencies of transitions between ISCO occupations. More specifically, we exploited the sequence of work experiences reported in CVs and aggregate over subsequent job positions to eventually form "from-to" pairs. The second approach investigated how recruitments are correlated with age and work experience. We quantified this relationship in order to understand how the different ISCO occupations accumulate work experience as a person's career advances.

5.3.1 Frequency of Group Change

Using career histories, we are able to explore sequences of work experiences as they appear on CVs.

- When changing jobs, CV owners may start a different occupation than their previous one.
- The probability of remaining in the same occupation differs from one ISCO group to another.
- By measuring the frequencies of transition between ISCO occupations, we calculate the probability of ISCO group change.

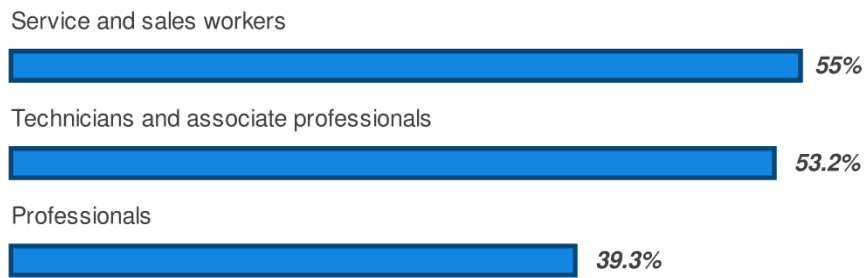


Figure 17. ISCO 1 occupations with the lowest probability of group change.



Figure 18. Probabilities of group change for ISCO 2 occupations of Professionals.

- In general, CV owners most commonly start a new job that belongs to the same ISCO group as their previous one.
- *Skilled agricultural, forestry and fishery workers* and *Clerical support workers* are the most likely to move to a different ISCO 1 group, while *Professionals* and *Technicians and associate professionals* are the least.
- ISCO 2 occupations *Health professionals*, *Information and communications technology professionals*, and *Teaching professionals* are particularly consistent in staying in their respective fields.

5.3.2 Most Common Transitions to Different Groups

Sequences of work experiences seen among CVs suggest that certain occupations are related and some patterns of transition between different ISCO occupations are more common than others.

- To quantify these relationships, we calculate the probability that a CV owner moves from an ISCO occupation to another.
- Transitions to the same group are excluded, because as documented, the most common transitions are, in general, between ISCO occupations of the same group.

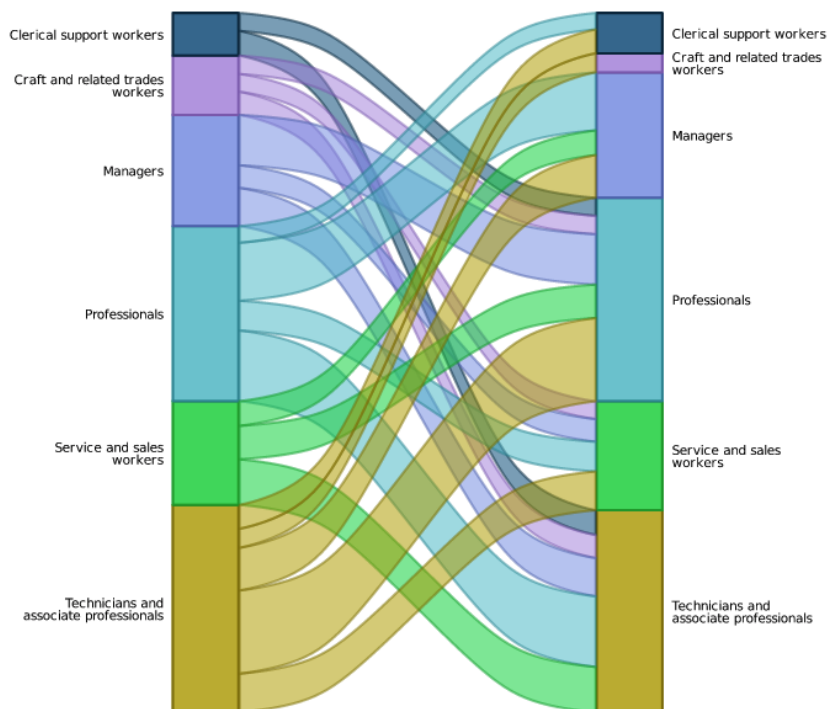


Figure 19. Transition between ISCO 1 occupations (excluding self-transition).

- When starting a job that belongs to a different ISCO 1 group than their previous one, CV owners most commonly become *Professionals* or *Technicians and associate professionals*.
- *Professionals* themselves are most likely to transition into *Managers* or *Technicians and associate professionals*.
- Transitions between related ISCO 2 groups (e.g., *Business and administration professionals* and *Administrative and commercial managers*, or *Legal, social and cultural professionals* and *Teaching professionals*) are common.

Discussion: Since transitions are generally made between related jobs, the exploration of common transitions may be more meaningful for lower levels of the ISCO hierarchy. The level of aggregation that ISCO 1 and 2 offer may not be able to expose some of the more interesting patterns that can be made evident through more specific queries. Caveats of the process of free-text matching also need to be kept in mind when exploring these transitions, as it is possible that occupations with similar corpus and definitions on the ESCO/ISCO taxonomy may be harder to disambiguate.

5.3.3 Work Experience by Age

As a person's career advances, they accumulate work experience through their participation in the labour force, making age naturally correlated with years of work experience.

- To quantify this relationship in our dataset, we have measured every CV owners's cumulative work experience and age at the time of recruitment on each occupation they reported.
- We have then proceeded to perform linear regression analysis on age vs. years of work experience and compared the fitted line with the mean age and mean work experience of each individual ISCO 1, 2 and 3 group.

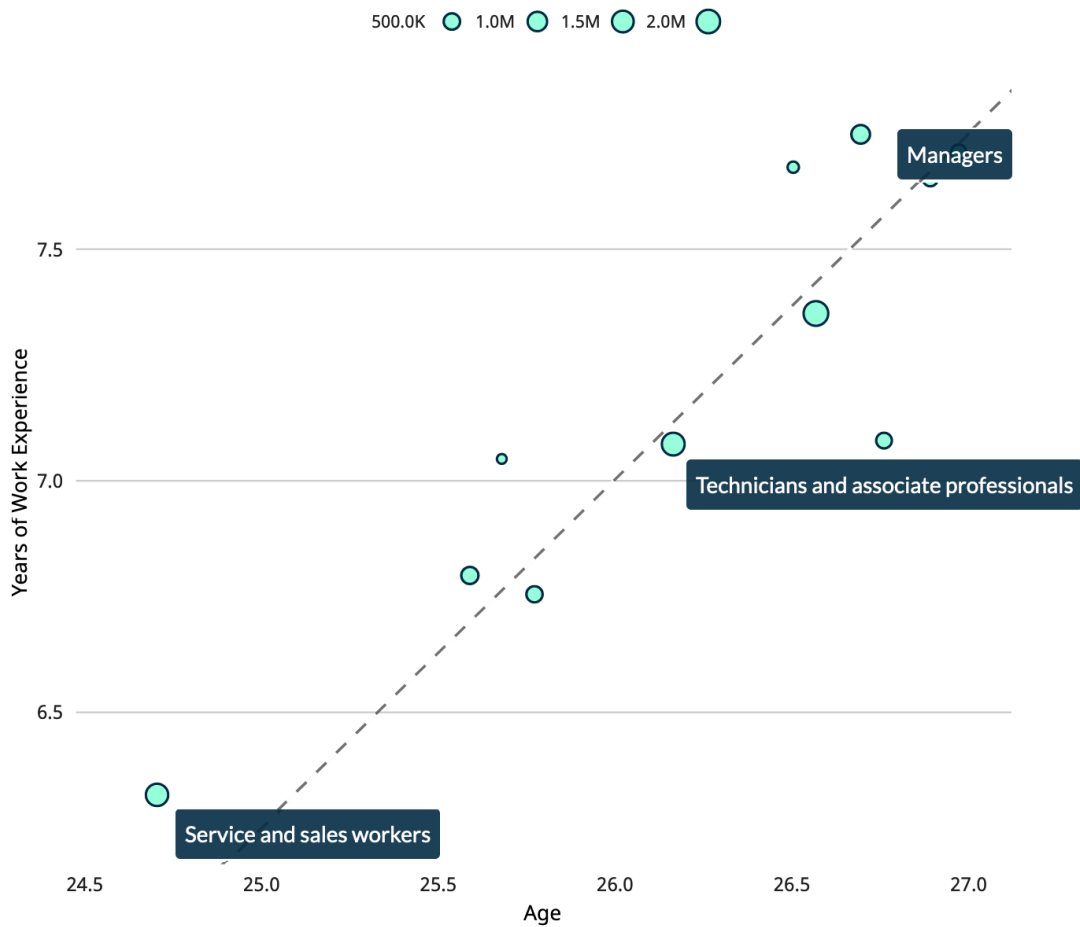


Figure 20. Relationship between age and years of work experience with ISCO 1 occupations placed based on their mean values.

The linear relationship between years of work experience and age as calculated by linear regression on all data points can be seen in the dashed line. Each of the green circles represents an ISCO 1 occupation and is placed based on the mean age and mean years of work experience at the time of recruitment of the users reporting it (regardless of what their previous jobs were). The occupations placed leftmost in the x-axis are for ISCO 1 occupations more commonly professed by younger individuals (e.g., Service and sales workers), while those rightmost are more commonly professed by relatively older ones (e.g., Managers). The occupations placed on the y-axis are for ISCO 1 occupations whose users had relatively more years of work experience at the time of recruitment (e.g., Managers), and those that lie on the bottom are entry-level jobs (e.g., Service and sales workers).

Looking at different ISCO hierarchical levels the relationship between age and work experience is strongly linear. When we look at recruitment type as a relation between age and work experience, we discover that young people with less work experience are more likely to be hired as waiters or serve in the army. As we move across the fitted line to the right, we gradually encounter vocations that require maturity, such as managers. Above the line are more likely skill-based occupations that allow entry into the labour market at a younger age and/or occupations that require more experience to be recruited like Managers. Those below the line, on the other hand, are more likely to be knowledge-based and/or that require less experience for recruitment, such as school teachers, life science experts, and programmers.

Discussion: The deviation between the dashed line and each occupation is a measure of the work experience of an individual who is recruited for a particular ISCO 1, compared to the expected value. For example, Managers appear above the dashed line, which means that individuals that start to work as Managers, have more years of work experience than the average newly-recruited individuals. The opposite is true for Elementary occupations that have less years of work experience than expected.

This approach can be used to measure how easily users from different ISCO groups accumulate work experience (e.g., with respect to unemployment), which groups tend to have an early entry in the job market, and which ones may have a later one. Deviations between the expected and observed value are explored in the next section.

5.3.4 Deviation of Observed and Expected Work Experience

The expected years of work experience by age appear on the dashed line in **Figure 20** derived through the regression analysis.

- The relationship between age and years of work experience is different for each ISCO group.
- Using the residuals of the regression line we fit on our data, we derive the mean residual of our data from the regression line by ISCO group.

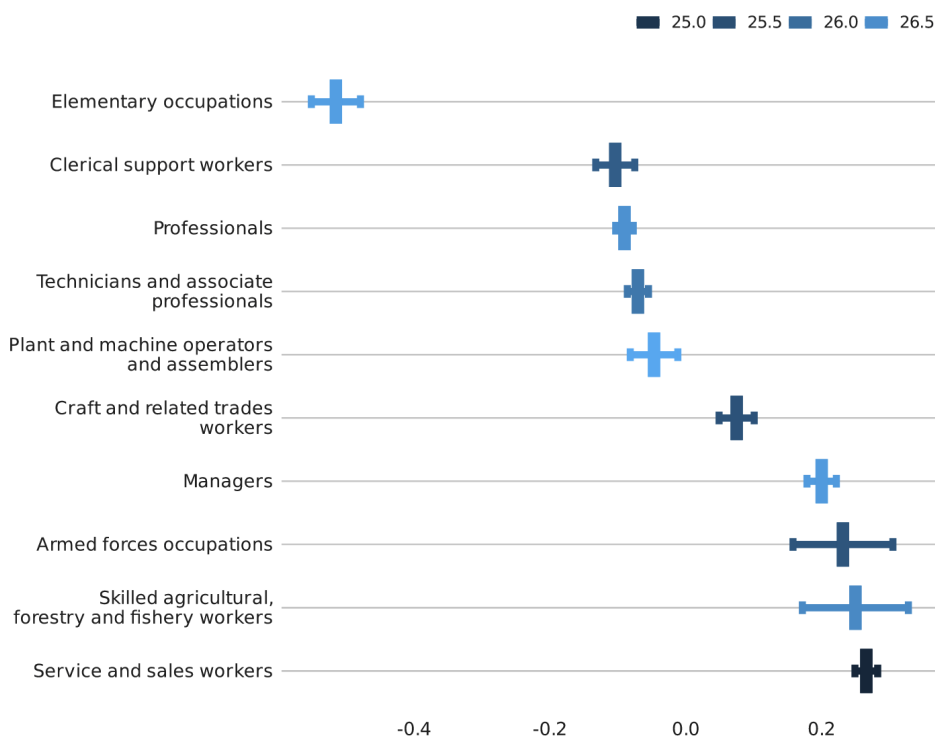


Figure 21. Difference between expected and measured mean work experience by ISCO 1 groups. The expected work experience is calculated by using a linear model between age and work experience.

- Individuals professing jobs that belong in the ISCO 1 groups *Professionals* and *Technicians and associate professionals* have roughly as many years of work experience as expected based on the main sequence.
- The observed work experience for *Service and sales workers* is higher than the expected value, while the opposite is true for *Elementary occupations*.

Discussion: Professions that require a lot of training and education before entering the job market include individuals that have less work experience than anticipated, likely due to late entry to the labour force.

Administrative and managerial jobs show more experience than anticipated, indicating a requirement of work experience accumulation before individuals are able to move to these occupations.

The observed work experience for entry-level jobs such as *Service and sales workers* is higher than the expected value, while the opposite is true for *Elementary occupations*.

5.3.5 Next Steps

Exploring more detailed, meaningful breakdowns of subgroups in the dataset can help derive more interesting insight about career paths. For example, focusing on a specific age group and/or individuals with a specific latest ISCO 1 job that is well represented in the data (e.g., *Professionals*) may allow to explore the question of job transitions with respect to ISCO 3, and can therefore capture transition patterns in higher resolution.

Additionally, the comparison of work experience and age can be expanded upon by employing clustering. Three main clusters of ISCO 3 groups can be noticed based on the relationship between age and years of work experience:

1. The main cluster, composed mainly of ISCO 3 occupations that are classified as *Professionals* and *Technicians and associate professionals* on ISCO 1, and have an average mean age and work experience;
2. A dense cluster composed mainly of occupations such as *Waiters and bartenders* and *Shop salespersons*; and
3. A sparser cluster composed of occupations generally professed by older individuals, such as *Managing directors and chief executives* and *Medical doctors*, but also *Heavy truck and bus drivers* and *Refuse workers*.

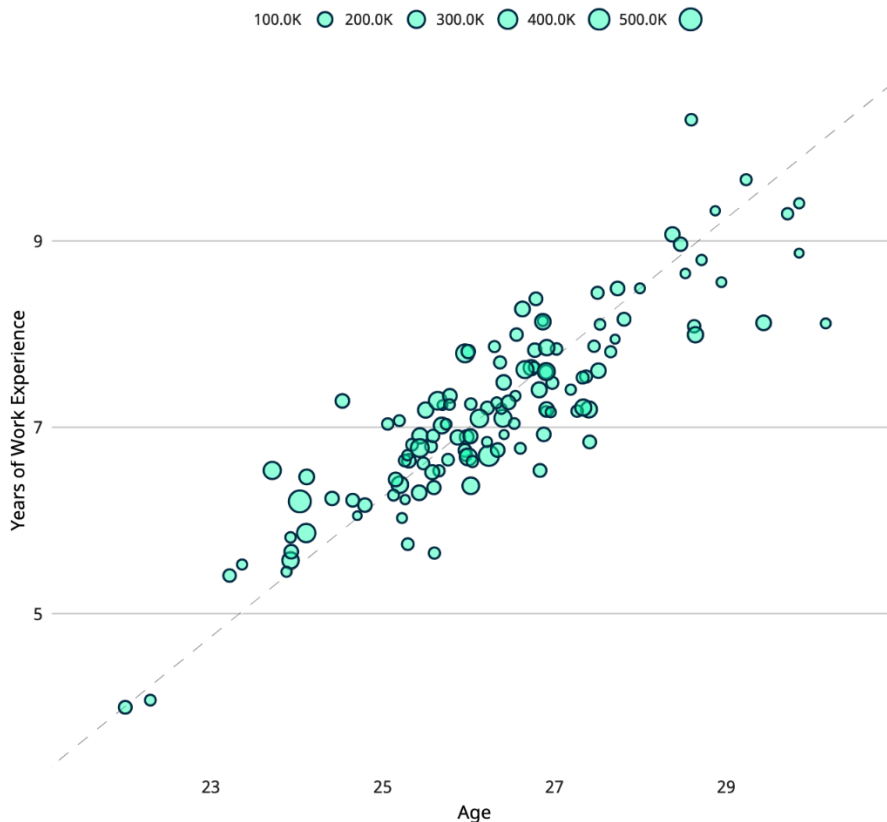


Figure 22. Relationship between age and years of work experience with ISCO 3 occupations placed based on their mean values

The characteristics of each of these clusters can be analysed separately with respect to age, gender, trends, career paths, and other features.

5.4 Skills-to-Occupations Associations in the ESCO Model Compared to the Collected CV Data

The skill necessity for an occupation is defined within the ESCO model, with skills being marked as either essential or optional for an occupation. As part of our data cleansing process, occupation and skill free-text entries entered by Europass users have been matched to an equivalent classification on the ESCO model. We apply association rules mining to CVs aiming to reveal the naturally-occurring associations between occupations and skills without considering the relationships pre-defined by the ESCO model.

5.4.1 Associations between Skills and Occupations

We aggregated our data so that they formed "baskets" on which market basket analysis can be applied. The "basket items" are defined as the latest occupation (mapped to ISCO 3) and each of the skills that were reported at the time of recruitment.

- We have extrapolated skill necessity for an ESCO occupation to the ISCO 3 level, with each skill being marked as essential or optional for at least one ESCO occupation of the respective ISCO 3 group, or undetermined otherwise.
- Through association rules mining, the relationship between an occupation and a skill has been quantified on a metric called "lift" (see annex). A high lift suggests a strong relationship between a skill and an occupation.

- By ordering occupation and skill pairs by lift, we identify the strongest associations between skills and occupations as they occur naturally in the CVs and whether or not they are already encoded in the ESCO model.

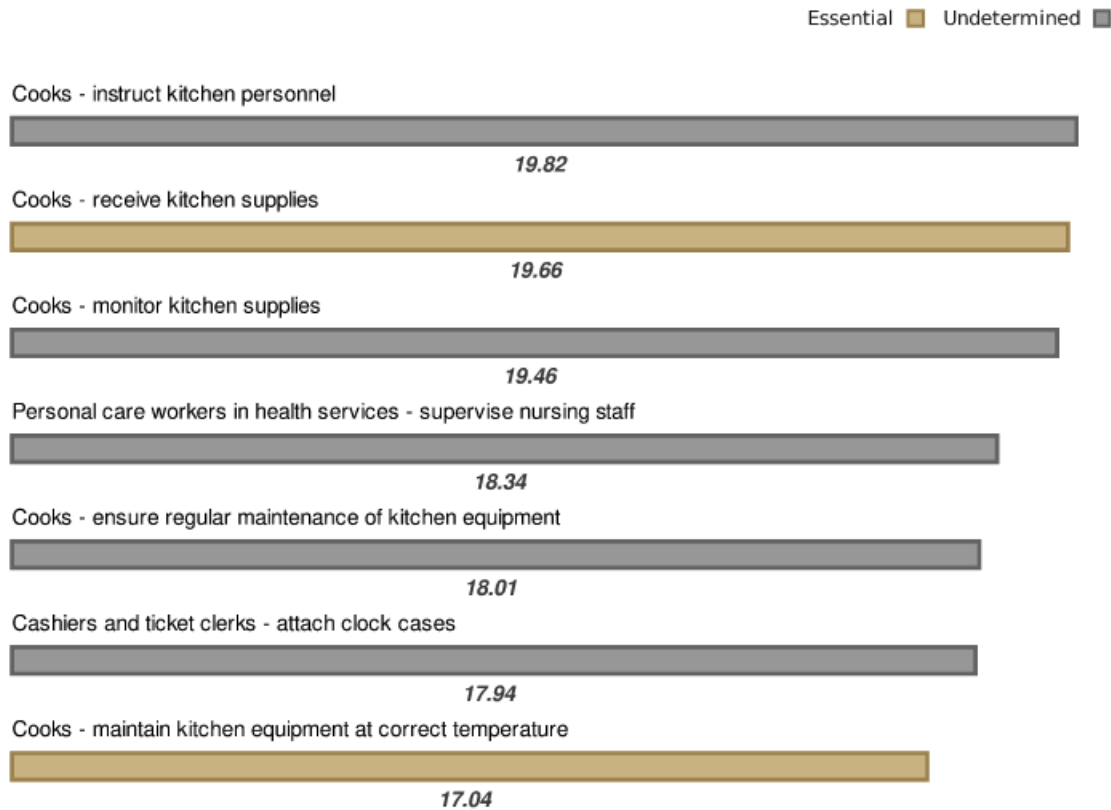


Figure 23. Associations between occupations and skills identified through market basket analysis for Service and sales workers

Measuring the inclusion of skills per occupation reveals new meaningful associations that are not encoded in the ESCO model. Examples of ISCO 3 and ESCO skill pairs include *Cooks - instruct kitchen personnel*, *Nursing and midwifery professionals - communicate with nursing staff*, and *Domestic, hotel and office cleaners and helpers - conduct cleaning tasks*.

Discussion: Identification of associated occupations and skills on the Europass CV data is possible without relying on the skill necessity defined by ESCO's model, through market basket analysis. Analysis of naturally-occurring data from Europass and other similar sources can inform and assist researchers to improve the ESCO model by documenting the most common associations.

6. Cross Checking with Official Statistics

We have transformed our dataset so that it can be compared with specific indicators in the European Union Labour Force Survey (EU-LFS), as well as statistics from Cedefop's OVATE. Our main goal with this exercise was not to identify or measure gaps in the labour market, but to test the representativeness and completeness of the information encoded in Europass. More specifically, we attempted to identify meaningful breakdowns of subgroups within the Europass userbase (e.g., "*Professionals in the age group 25-49*") for which insight can be more reliably derived given the dataset's biases.

6.1 Employment (LFS)

6.1.1 Employment by Occupations per Age Group and Gender (EA19, 2019)

The distribution of occupations observed in the Europass database for users reporting employment at the time of the creation of their CV (i.e., including no end date on their most recent work experience) was compared with the distribution of occupations of the general population reported by the Labour Force Survey.

- Specifically, the indicator reporting on *employment by sex, age, professional status and occupation* [lfsa_egais] was utilised. Statistics for people indicated as employed persons in their activity and employment status were considered. This indicator reports on quarterly data, so the mean of the four quarters was used to derive an estimate for the year.
- 2019 was selected as the year of focus as it provided the latest complete year available in the Europass database (which has data up to Q2 2020). Comparisons for Euro area (EA-19) were made with respect to two broad age groups (15-24 and 25-49) and gender.

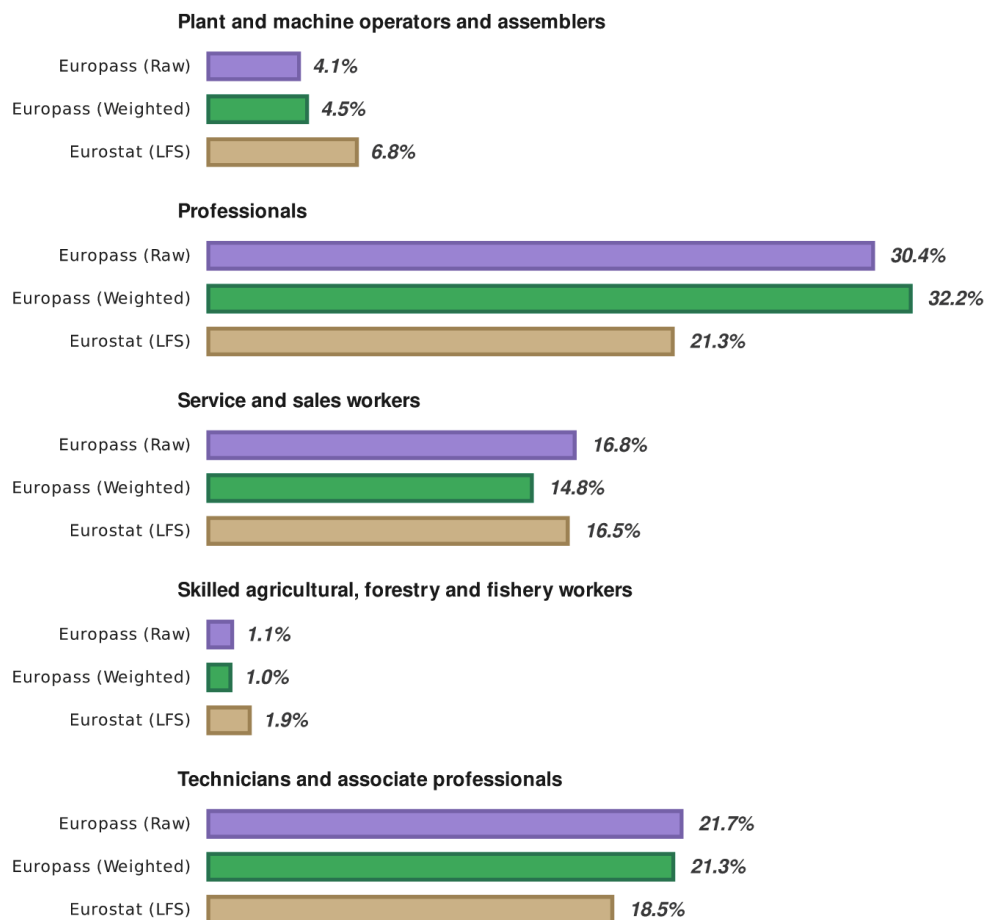


Figure 24. Example of the comparison of distributions for EA-19 and ages 25-49, with both weighted and unweighted measurements

- *Managers and Professionals* are significantly overrepresented in the Europass data. The former is reported almost twice as much among CVs in the age group Y25-49 compared to what is seen in the general population, while the latter is seen over 50%.
- Conversely, *Clerical support workers, Craft and related trades workers and Skilled agricultural, forestry and fishery workers* are significantly underrepresented, with CVs reporting these occupations roughly at a half rate than they appear in the Labour Force Survey.
- With a few exceptions, these imbalances are even more pronounced among individuals in the age group 15-24, where *Elementary occupations* also tends to be reported three times less than in the general population.
- Most patterns of differentiation between males and females observed in the Labour Force Market indicator are generally reflected in the Europass data. For example, *Plant and machine operators and Assemblers* are more likely to be male in both sources, while *Service and sales workers* are more likely to be female.

Discussion: A hypothesis for the encountered statistical deviation lies in the fact that Europass is an online CV creation tool, and is therefore not evenly used by people across different occupations. People working as managers are more likely to use the application than workers in craft and related trades, for example.

6.1.2 Employment by Occupation per Country (2019)

The previous exercise was repeated for each distinct country in EA-19.

- Male and female individuals aged between 15 and 49 from countries whose sample exceeds 800 CVs were included in the comparison.

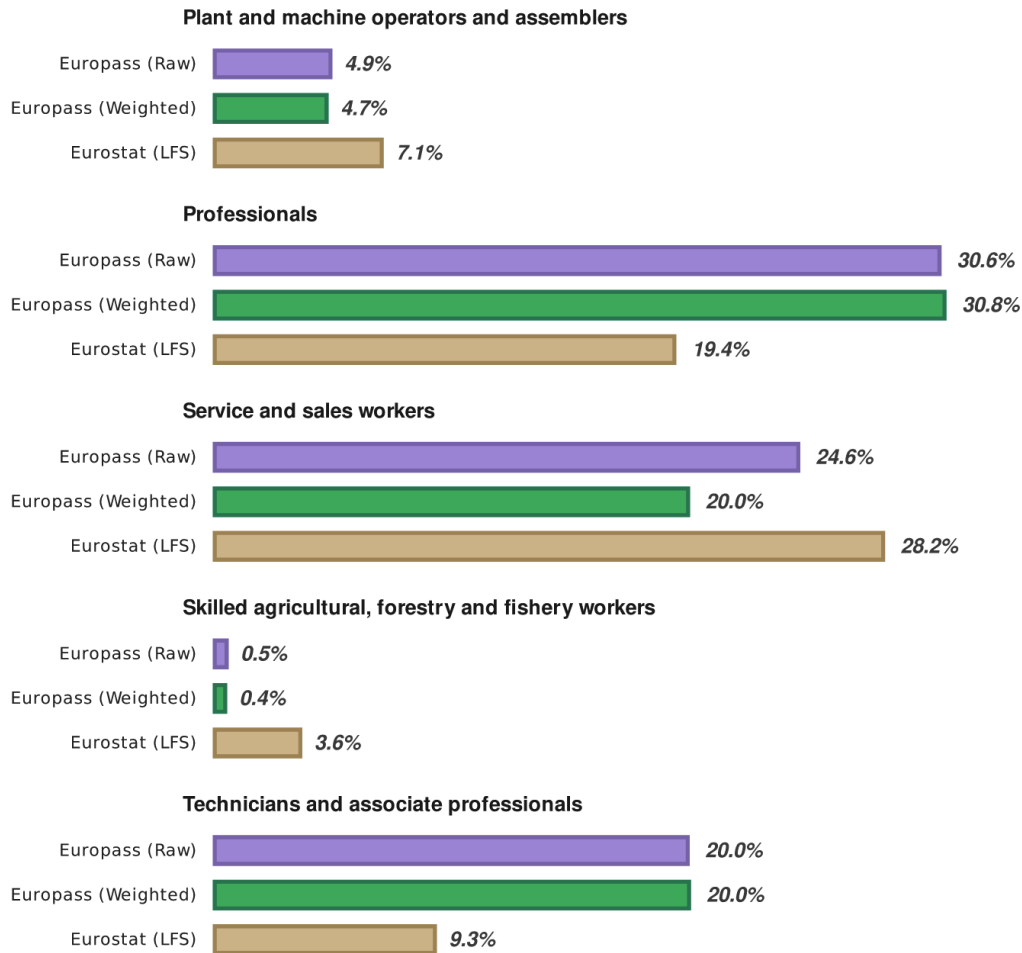


Figure 25. Example of the comparison of distributions for Greece and ages 15-49, with both weighted and unweighted measurements

- The tendency for *Managers* and *Professionals* to be overrepresented in the Europass database is observed among users of most countries. Additionally, *Technicians and associate professionals* are strongly overrepresented among individuals from certain countries such as Portugal, Spain and Greece, where this occupation appears twice as often as its respective statistic reported by the Labour Force Survey. A notable exception is that of Germany, where they are underrepresented.
- Likewise, *Clerical support workers*, *Craft and related trades workers* and *Skilled agricultural, forestry and fishery workers* are underrepresented across the board, with the third skill displaying a very small percentage even among countries with a robust primary sector like Greece, where 8.5% of the population undertakes occupation, but only 0.4% of CV's report it. *Elementary occupations* are also underreported across most countries, especially Spain and France.

Discussion: Patterns specific to certain countries noted by the Labour Force Survey are often also reflected in the Europass database. For example, countries in southern Europe such as Italy, Portugal and Greece have more users working as *Service and sales workers* than the EA-19 average according to both LFS and the measurements on the Europass database. This observation does not apply in

every case however, as the Europass userbase’s characteristics and behaviour may differ from country to country (e.g., in terms of age group and education background), and due to biases introduced by the multilingual nature of the dataset.

6.1.3 Trends in Employment by Occupation (EA19)

The Europass database includes data between Q1 2017 and Q2 2020. Using regression analysis, trends in employment with respect to occupations (ISCO 1) were measured for this period for both Europass data, and the respective Labour Force Survey indicator [lfsa_egais].

- Specifically, an estimate of the trend has been derived as the odds ratio change for one unit of time (1 year) with respect to age group for users in the Euro area (EA-19) as defined by Eurostat. For more specific details on the regression analysis, see Annex.
- We compared the trend derived from statistics reported by the Labour Force Survey with that from statistics measured in the Europass database.

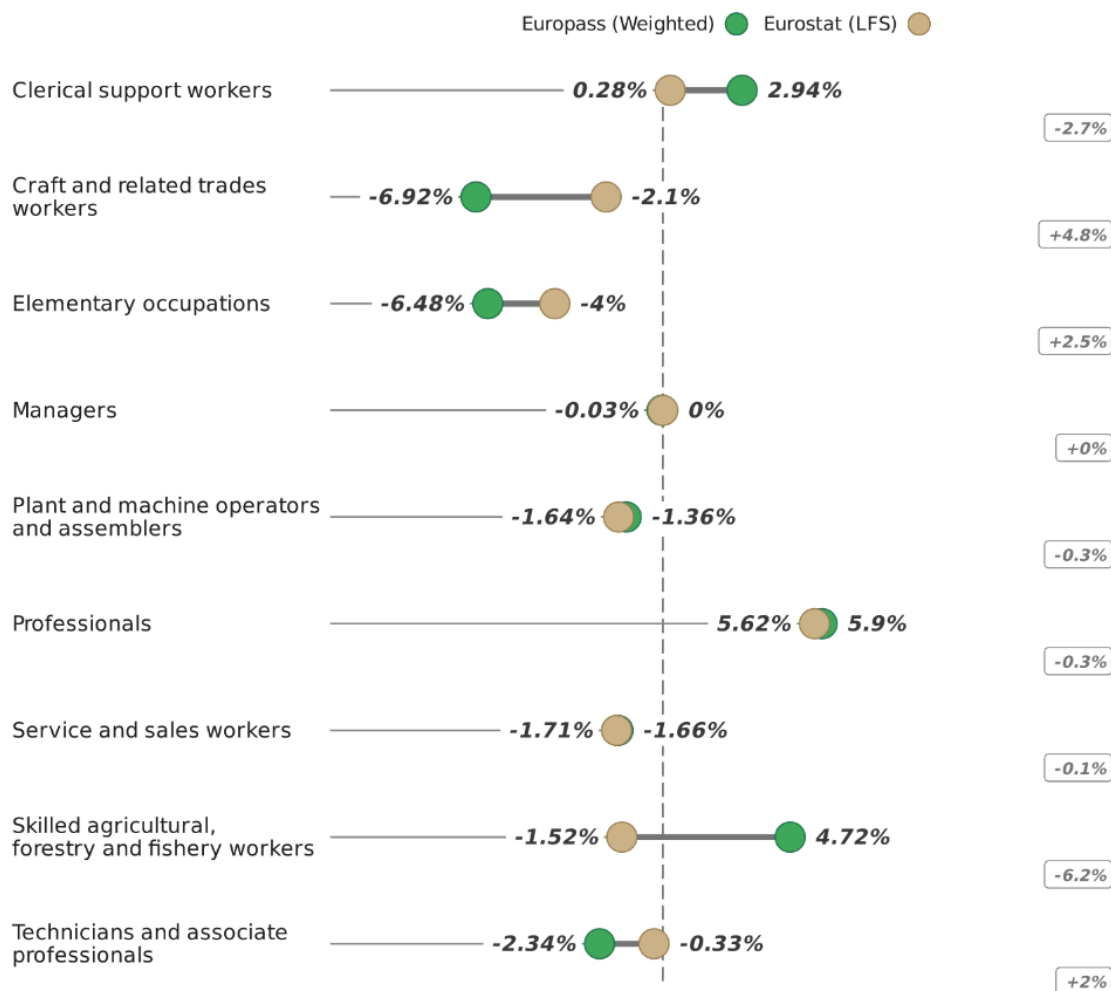


Figure 26. Comparisons of trends for age group 25-49

- For age group 25-49, the year-on-year ratio increase/decrease of each ISCO 1 occupation aligns relatively well. The measurements have a less than 1% difference for *Managers*, *Professionals*, *Plants and machine operators and assemblers*, and *Service and sales workers* and indicate relatively small yearly changes. For most other occupations, the trends measured have a greater difference in magnitude, but still align in direction in most cases.

- An exception is *Skilled agricultural, forestry and fishery workers*, whose trends are highly deviant between Europass and official statistics. This deviation is among the occupations that are relatively rare in the Europass database. The younger population displays different trends between Europass and LFS.

Discussion: Trends observed in the Europass database for the age group 25-49 align relatively well with the trends inferred by statistics in the Labour Force Survey. This alignment indicates that a dataset like Europass may be able to answer questions of this nature for this age group, provided that data collection is performed continuously over time.

The same is not true for younger population. This finding is likely because there is an inherent instability in the career path of younger individuals, which is also noted in the comparison of the distributions of occupations. Moreover, the sample size is relatively small for individuals aged between 15 and 20 especially in 2017. It is also possible that early adopters of the Europass application might be biased towards specific occupations, and this imbalance may even out in later years as the sample size increases.

6.1.4 Gender Ratio by Occupation (EA19)

Information about the gender breakdown of each occupation is available on both the Europass database and the LFS indicator [lfsa_egais]. We define the gender ratio of an occupation as the ratio of males to females for that occupation.

- We compared the evolution of the gender ratio in time between the two sources.
- In addition, we applied regression analysis to measurements for the Euro zone (EA-19) in order to quantify how the proportion of males in each occupation has changed in the period between 2017 and 2020 according to each source.

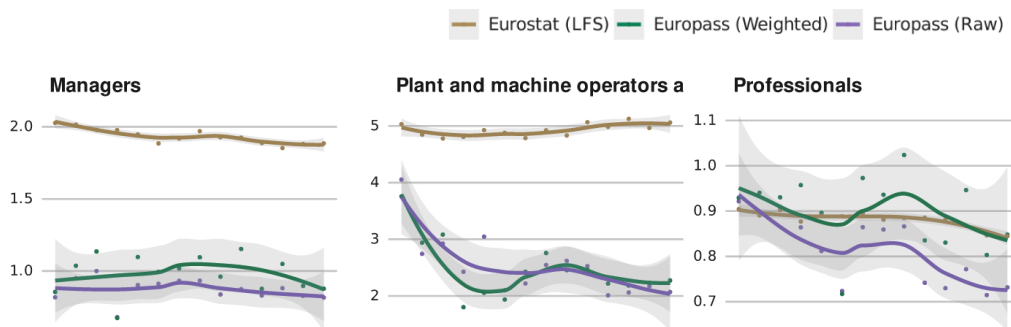


Figure 27. Example of the evolution of gender ratio of occupations in time.

- Barring some exceptions, the gender ratio reported by the Labour Force Survey generally reflects the gender ratio inferred by the analysis on the Europass database. There are over three times more males working as *Skilled agricultural, forestry and fishery workers* than there are females, while females working as *Service and sales workers* are almost twice as many as males. *Plant and machine operators and assemblers* as well as *Craft and related trades workers* are overwhelmingly male according to both sources, although the ratio difference is not as big in the Europass data as it is in the Labour Force Survey.
- Two major exceptions to this observation are *Managers* and *Clerical support workers*.

Discussion: *Managers* appear to be evenly distributed between males and females on Europass, even though the Labour Force Survey reports twice as many males than females. This finding may be partly explained by the fact that Europass data, even after weighting, are biased towards younger people (due to weight threshold) and it can be observed that the gender ratio is trending down year-on-year.

Trends inferred by both Labour Force Survey data and the Europass database display very small differences on the gender proportion of each occupation year-on-year. These trends do not align well between the two sources, and in most cases the Europass trends are biased in favour of male growth in the proportion of each occupation. A possible explanation for this is that there may have been more female early adopters for the Europass application, a difference that has evened out in later years.

6.2 Unemployment (LFS)

6.2.1 Unemployment by most Recent Occupation per Country (2019)

The distribution of the most recent occupation of unemployed individuals with at least one work experience was compared with the respective indicator on the Labour Force Survey. Specifically, the indicator reporting on *previous occupations of the unemployed, by sex* [lfsq_ugpis] was used, with the quarterly data included, to derive a yearly mean, with 2019 selected as the year of focus (2019 being the latest complete year available in the Europass database).

- We assumed that for CVs that did not specify an ongoing work experience at the time of creation, their owner is unemployed.
- No distinctions are made with regards to age group on the LFS indicator, so the measurements were compared with the entire age range of CV owners.

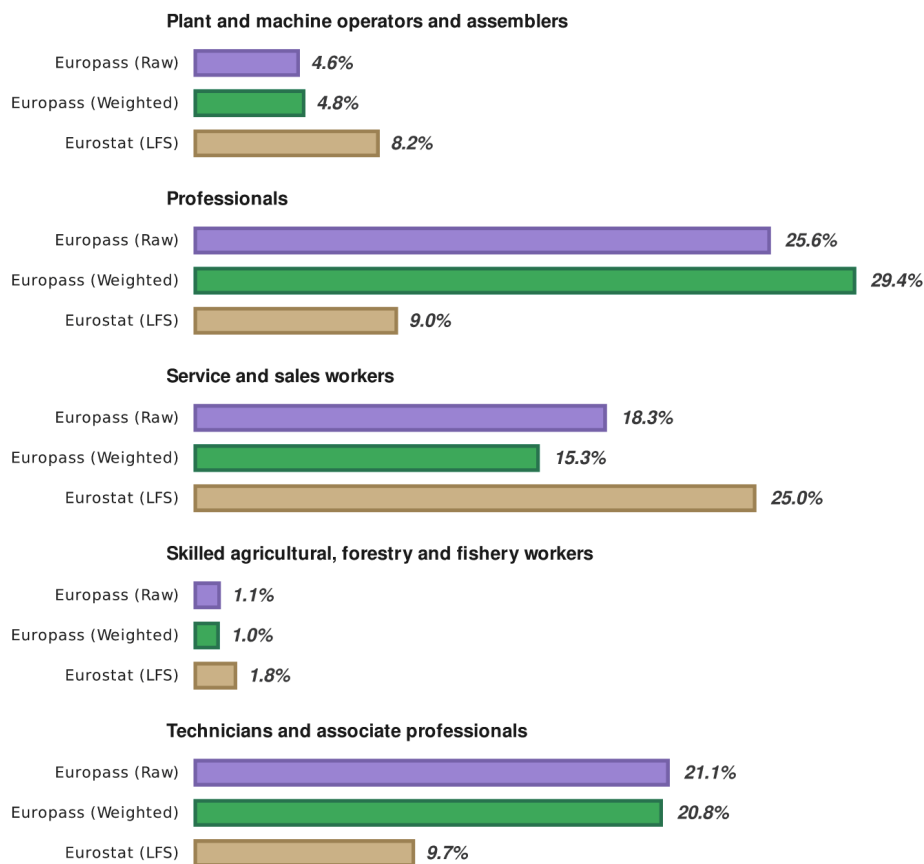


Figure 28. Distribution of the latest ISCO 1 occupation for unemployed users

- The distribution of the most recent occupations of the unemployed display major differences between the two sources. *Managers, Professionals and Technicians and associate professionals* are even more overrepresented in this case than they were in the case of employed individuals, with *Managers* appearing 5 times more often.

- The most underrepresented group in the Europass database are those in *Elementary occupations*, who appear 5 times less often than in the general population. *Service and sales workers* and *Clerical support workers* are also majorly underrepresented.
- These patterns are consistently observed in various degrees across most countries. A few exceptions to this finding are Portugal and Belgium, where the percentage of *Managers* is closer between the two sources. It should be noted that the Labour Force Survey does not report statistics for all occupations in every country. In those cases, the distribution is based on the occupations included.

Discussion: Biases towards the reported occupations once again has to do with Europass's nature as an online CV creator which is not evenly used by people of different occupations. Moreover, it may also have to do with the amount of time workers from different occupations stay unemployed. *Managers* and *Professionals* may jump from job to job relatively fast compared to people within *Elementary occupations*, meaning that the LFS is less likely to measure them as unemployed. Finally, the Europass data are biased towards the younger population, which may also play a role in the difference between the two distributions.

Overall, the distribution of occupations is similar between employed and unemployed individuals. We calculated the Pearson correlation coefficient (r) as a measure of the linear relationship between two distributions. The absolute value of r can be between 0 and 1, with 1 signifying perfect correlation.

Gender	Source	r
Total	Europass (Weighted)	0.9957607
Male	Europass (Weighted)	0.9945066
Female	Europass (Weighted)	0.9978480
Total	Europass (Raw)	0.9948595
Male	Europass (Raw)	0.9910250
Female	Europass (Raw)	0.9965291

The r coefficient is close to 1 for every gender and source. This finding suggests a strong correlation between the distributions of employed and unemployed individuals. It can be said that the distribution of occupations in Europass does not measure explicitly employed or unemployed population, but rather the distribution of occupations of people in the job search market.

6.2.2 Education Level of the Long-Term Unemployed by Age Group per Country (2019)

The relationship between long-term unemployment (i.e., unemployment for 12 months and more) and education is explored in the Labour Force Survey indicator *long-term unemployment (12 months and more) by sex, age, educational attainment level and NUTS 2 regions* [lfst_r_lfu2ltu].

- We made the equivalent calculation for CVs where the most recent work experience ended more than 1 year before the creation of the CV, as well as those who have reported no work experiences at all.

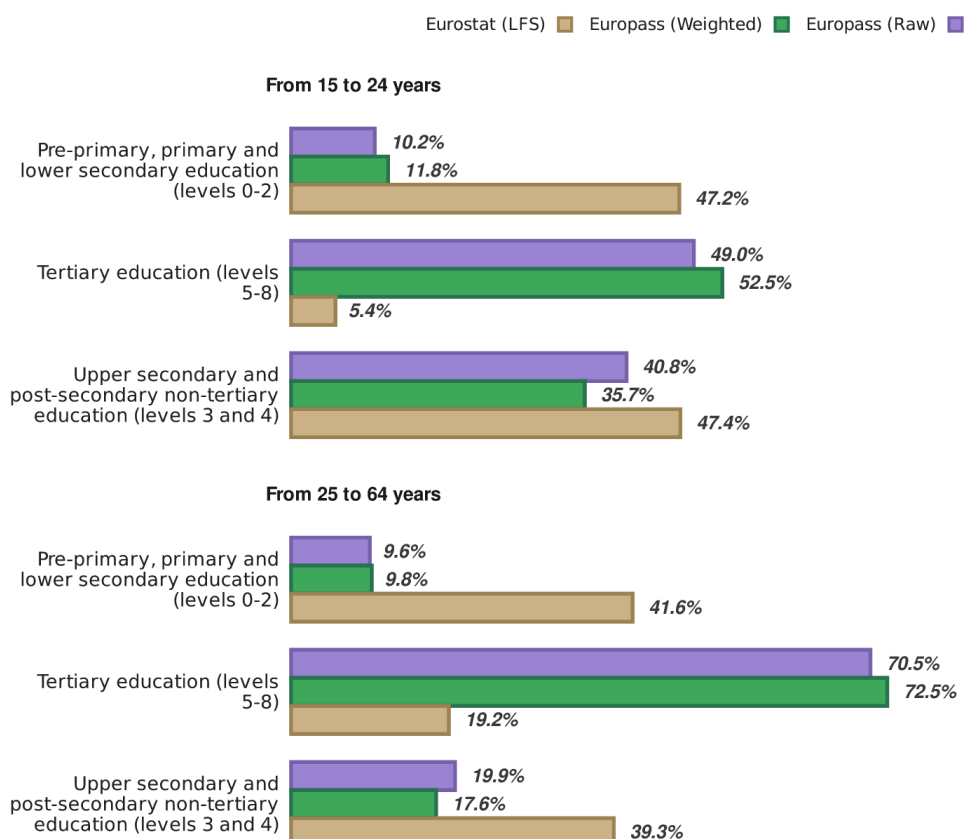


Figure 29. Distribution of education level for unemployed individuals in EA-19

- Compared to the general population as reflected on the Labour Force Survey, individuals report on average, a higher level of education. Specifically, unemployed individuals over 25 years old are more than 4 times likely to have received an education equivalent to 5-8 on ISCED.
- This imbalance becomes even more prominent for younger users on the age group 15-24.

Discussion: A likely explanation for this is the fact that Europass is generally used by people who work, or aim to work on occupations that necessitate at least some tertiary education. This explanation is indeed the case when observing the distribution of occupations, where *Managers* and especially *Professionals* are overrepresented compared to what is seen in the general population, while *Elementary occupations* that do not require special qualifications tend to be underrepresented.

Moreover, the dataset includes many young individuals who have only recently completed university education, and are thus unemployed at the time of CV creation. Individuals in those categories are likely to use an online application to create their CV before they apply for their first job, so sources such as Europass will usually be composed of people with higher-than-average education. Use of Europass also assumes technological literacy, which is higher among people who have achieved some post-secondary education.

6.3 Job Vacancies (Cedefop)

6.3.1 Supply and Demand (2019)

Job vacancy data from the OJA project was utilised with the purpose of measuring how the demand for jobs online compared with supply as inferred from the Europass database.

- Specifically, the "Job Applied For" field in Europass CVs of users between 15 and 49 was used to gauge supply.

- The distribution of supply and demand of occupations (ISCO 1) in 2019 was compared for countries in the Euro area (EA-19).
- The EU average for job vacancy data refers to EU27, while for Europass it refers to the EA-19.

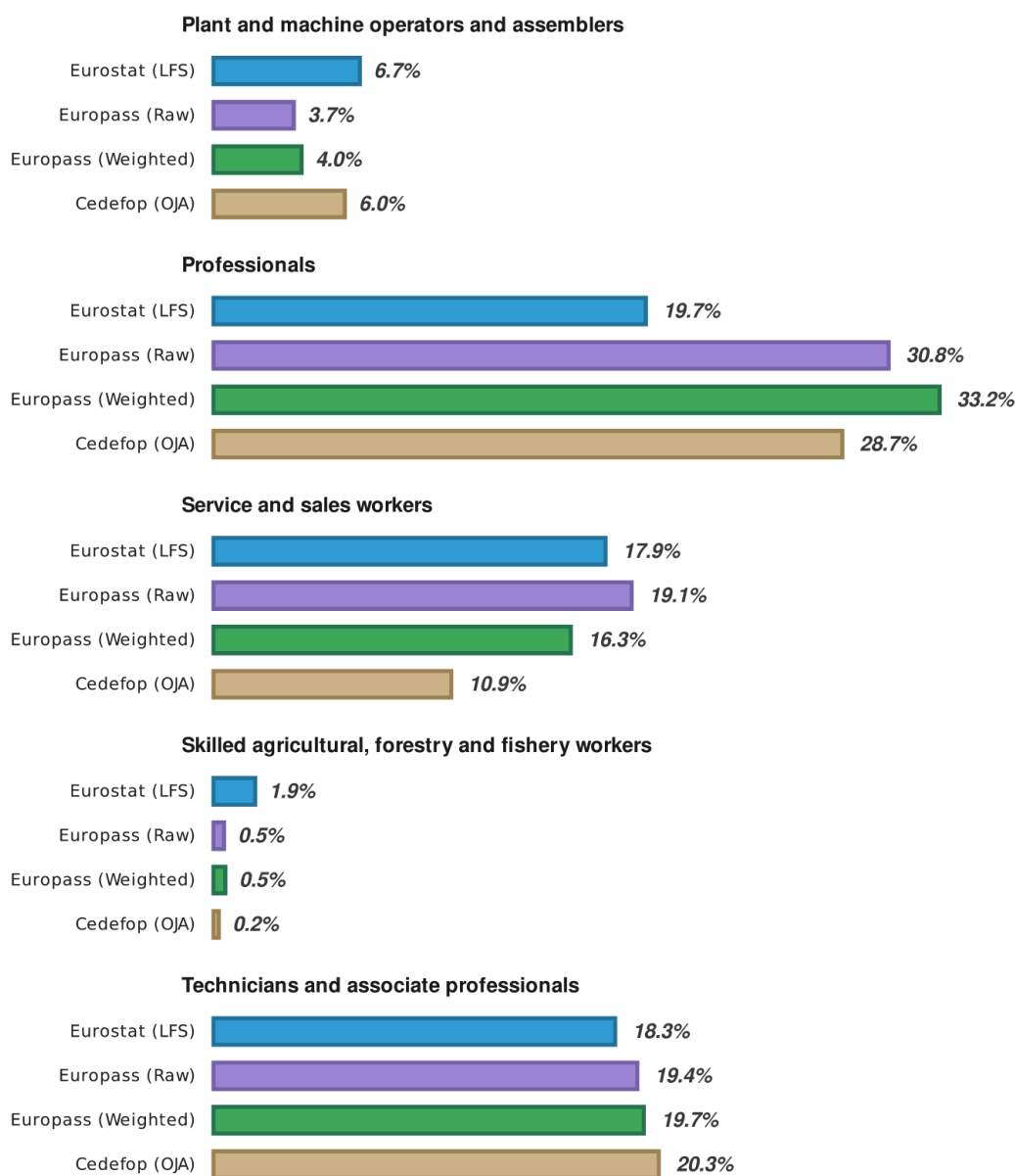


Figure 30. Example of the distribution of the ISCO 1 for EU

- It is observed that the distribution of occupations in Europass CVs resembles its equivalent in job vacancies, especially when compared to the general distribution of occupations professed by the employed population seen in the Labour Force Survey.
- The largest gap is noted for *Clerical Support Workers* and *Plant and machine operators and assemblers*, where the supply of jobs online is bigger than the demand in the CVs. The opposite can be noted for *Service and sales workers*, with many CVs reporting interest in these jobs, but not equally as many online vacancies.

Discussion: The gaps noted may indicate actual gaps in the labour force, or a difference in the behaviour of the population working across different occupations when it comes to online job

applications. Some discrepancies may also be a result of certain sectors, such as those occupying *Service and sales workers*, not announcing vacancies online as often as others.

As the resemblance between online supply and demand is noted qualitatively, we attempted to measure the linear relationship of the occupations distribution on Europass with that of online job vacancies through the Pearson correlation coefficient (r).

Country	Source	r	t	d.f	p-value
EU	Europass (Weighted)	0.9661474	9.908055	7	0.0000227
Italy	Europass (Weighted)	0.9176151	6.108085	7	0.0004872
Portugal	Europass (Weighted)	0.9298582	6.686741	7	0.0002808
Spain	Europass (Weighted)	0.9716155	10.866556	7	0.0000123
Greece	Europass (Weighted)	0.9678767	10.184984	7	0.0000190
France	Europass (Weighted)	0.8228066	3.830476	7	0.0064538
Slovenia	Europass (Weighted)	0.4395991	1.294898	7	0.2364269
Germany	Europass (Weighted)	0.9400677	7.294056	7	0.0001636
Austria	Europass (Weighted)	0.9394135	7.250745	7	0.0001698
Malta	Europass (Weighted)	0.6123692	2.049373	7	0.0796040
EU	Europass (Raw)	0.9387850	7.209772	7	0.0001760
Italy	Europass (Raw)	0.8700171	4.668861	7	0.0022906
Portugal	Europass (Raw)	0.9341104	6.923060	7	0.0002266
Spain	Europass (Raw)	0.9740810	11.393377	7	0.0000090
Greece	Europass (Raw)	0.9433405	7.521541	7	0.0001348
France	Europass (Raw)	0.8665648	4.594025	7	0.0025018
Slovenia	Europass (Raw)	0.4758030	1.431248	7	0.1954515
Germany	Europass (Raw)	0.9508351	8.122998	7	0.0000827
Austria	Europass (Raw)	0.9334679	6.885957	7	0.0002343
Malta	Europass (Raw)	0.5462385	1.725358	7	0.1281134

For the EU average, r coefficient is close to 1, suggesting a high correlation between supply and demand. This case is true for most countries as well, with the highest correlation being noted for Greece and Spain. Slovenia and Malta are the only cases where the correlation is under 0.5. As data from both sources have been processed through separate pipelines, this indicates that there is some correlation in labour data gathered from online sources, regardless of whether they measure supply or demand. It is also possible that the two datasets explored are subject to some similar biases.

6.4 Job Tenure (LFS)

6.4.1 Job Tenure per Age Group and Gender (EA19, 2019)

Job tenure most commonly refers to the length of time workers have been at their current job or with their current employer. The Labour Force Survey provides measurements of job tenure as part of the indicator reporting on *employment by sex, age and job tenure* [lfsa_egad].

- For the purpose of this comparison, job tenure was only measured for Europass CV owners employed at the time of their CV creation. Career histories were not used to measure tenure in their past jobs, as the statistic typically pertains to continuing spells of employment rather than completed ones.
- In this context, job tenure is measured with respect to job position instead of employer.
- The distribution of a specific set of job tenure lengths reported in the LFS indicator were compared with our measurements from the Europass database for countries in the Euro area (EA-19) as defined by Eurostat in 2019.

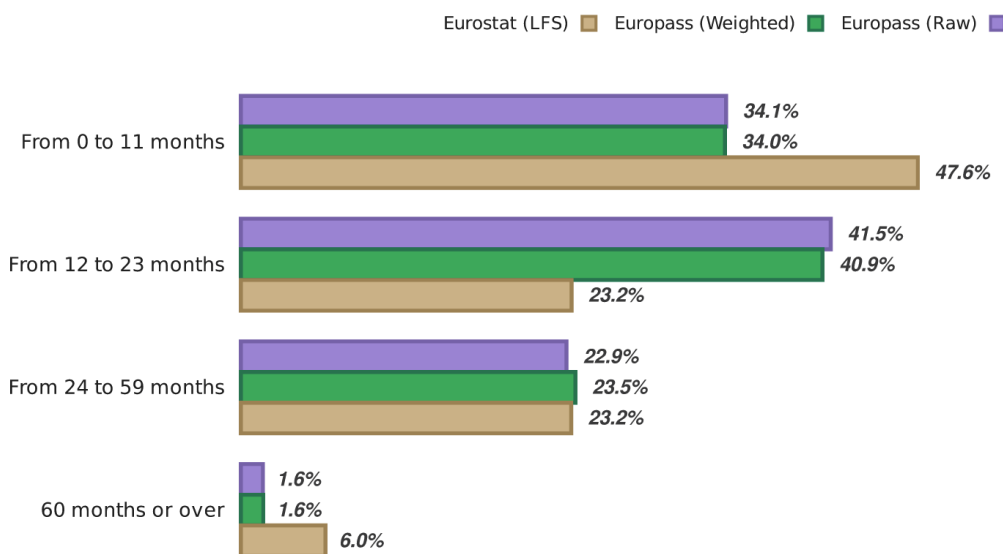


Figure 31. Distribution of job tenure for age group 15-24

- The distribution of job tenure observed in the Europass database greatly differs from the one reported by the Labour Force Survey indicator. This finding holds especially true for individuals in the age group 25-49, who appear to change jobs far more frequently than what Labour Force Survey indicates, where job tenure is greater than 5 years, a figure over 2 times less frequent than on the general population.
- The distribution is closer for younger individuals, who report job tenure between 1 and 2 years, roughly 20% more commonly than the general population.
- Differences between the two genders are not significant according to either source.

Discussion: Job tenure duration inferred from CVs in the Europass database is more evenly distributed between the four periods defined. This disparity can be attributed to the fact that Europass is an online CV creator, which is pointedly aimed to people looking for a new job. It is thus less likely for an individual who stayed with a particular employer or job title for a long time to use the application than someone who changes jobs more frequently.

Certain dataset biases might have a major effect on the measurements. As job tenure tends to vary for different kinds of jobs, it is hard to draw conclusions about the labour force in general, as parts of it are not well represented on the Europass dataset. For example, self-employed and/or individuals who run a business are not very likely to create a CV on Europass. This effect will likely arise with any attempt to measure job tenure via collected CVs. Moreover, the bias towards younger ages is expected to decrease the average job tenure.

Lastly, it should be noted that due to the fact that job tenure is measured with respect to the sequence of job positions and not employers, this may also decrease the average job tenure.

6.4.2 Job Tenure per Occupation (EA19, 2019)

Using the same assumptions as above, the distribution of job tenure of employed individuals at the time of their CV creation has been calculated for each occupation (ISCO 1). The LFS indicator *job tenure by sex, age, professional status and occupation* [lfsa_qoe_4a2] was used for comparison. In this case, the population with professional status “EMP” and age greater or equal to 25 years old was considered.

- The distribution of job tenure for each occupation in LFS was compared with equivalent measurements in Europass for countries in the Euro area (EA-19) in 2019.

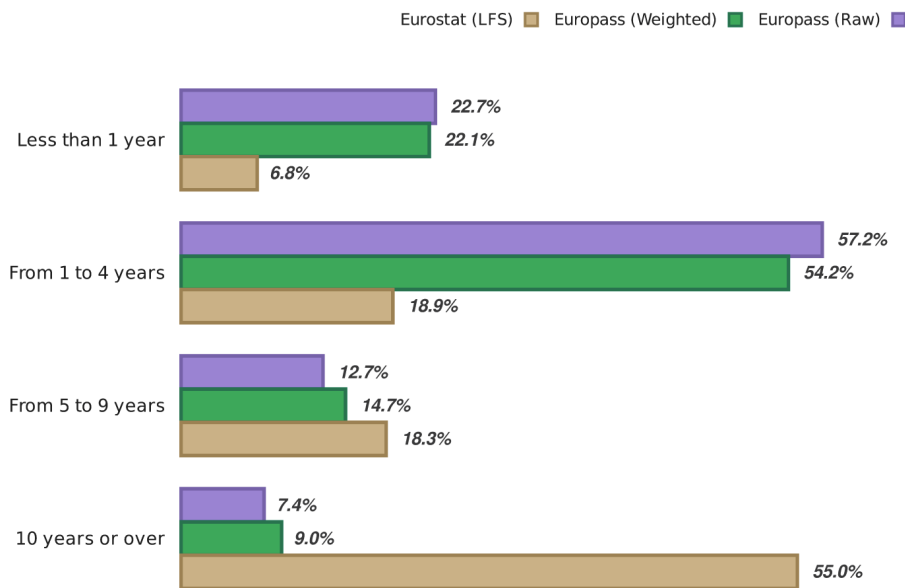


Figure 32. Distribution of job tenure for Managers

- The majority of employed individuals who created a Europass CV have worked on their current job between 1 and 4 years when they created their CV, with only small differences noticed between different ISCO occupations.
- This finding is in contradictory to the insight in Labour Force Survey, where more variance is noted in job tenure for each ISCO occupation. Biases and caveats of the dataset play a strong role in this disparity as documented previously.

7. Skills Analysis

In this section, we present our analysis on skills reported on Europass CVs. Given that information related to skills is not preserved in the Europass backup database, this section makes use of data from the Europass survey that was conducted between June and September of 2019. Specifically, the Europass CV editor allows for skills reporting across four categories:

- 1) Communication;
- 2) Organisational;
- 3) Job Related; and
- 4) Computer.

Every CV is thus tied to four free-text entries, which are then matched to the ESCO classification during the cleansing and standardisation process. A single free-text entry may be matched with one or more ESCO skills depending on its length.

Skills within the ESCO classification are organized in four broad categories:

- A – attitudes and values;
- K – knowledge;
- L – language skills and knowledge; and
- S – skills.

As with the occupation hierarchy defined by the combination of ESCO occupations and ISCO, the ESCO classification also define a skills hierarchy. The broad categories serve as the highest level of the hierarchy, and each one of those (with the exception of *L*) includes three additional levels of hierarchy, as well as the ESCO skills themselves, which serve as the leaf nodes, for a total of five levels. Unless otherwise noted, our reporting is done with respect to 3-digit ESCO skill groups, or in other words, the immediate parent nodes of ESCO skills. Linguistic skills are excluded.

7.1 Distribution of skills

7.1.1 Overall

We present initially the distribution of skills for all four skill category fields across all CVs.

- The distribution of non-linguistic skills observed overall is reflected in the following graph, with the number shown representing the total number of CV skill entries matched for each ESCO skill group.

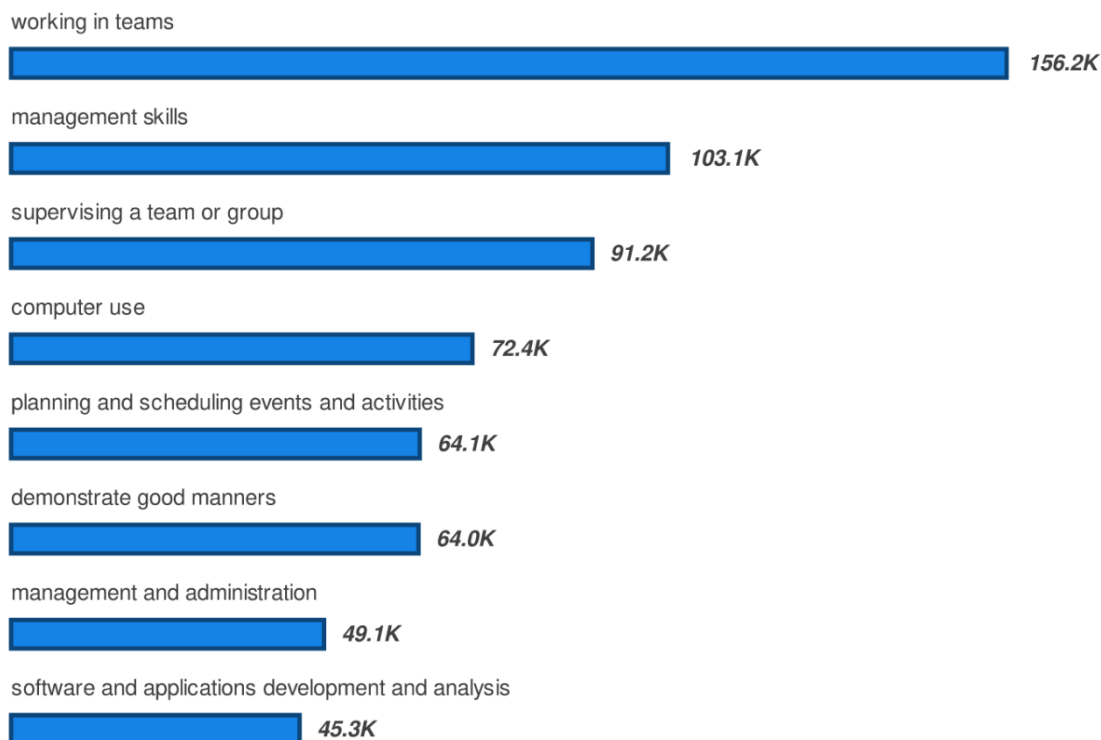


Figure 33. Distribution of ESCO skills found on Europass CVs

- The most commonly reported skills belong in ubiquitous skill groups such as *working in teams*, *management skills*, *supervising a team or group*, and *demonstrate good manners*.
- They are observed in every CV skill category, but especially in the Communication and Organisational skill categories.
- Some of the more concrete professional skills that are commonly observed in the Job related skills category, belong in groups such as *software and applications development and analysis*, *assure quality of processes and products*, and *law*.

7.1.2 Age Group

The most frequently mentioned skills across the entire user base are very similar regardless of age, meaning that directly comparing the distribution of skills of each age group does not elicit great results. Instead, quantitative feedback is provided on the age group breakdown of the 30 most common skill groups by comparing the ratio each age group represents across every one of them.

- Approximately 59% of users participating in the Europass Survey have included a birth year on their CV and are aged between 15 and 64 years old.
- Three broad age groups of 15-24, 25-49, and 50-64 have been defined and the reported skills were aggregated by age group.

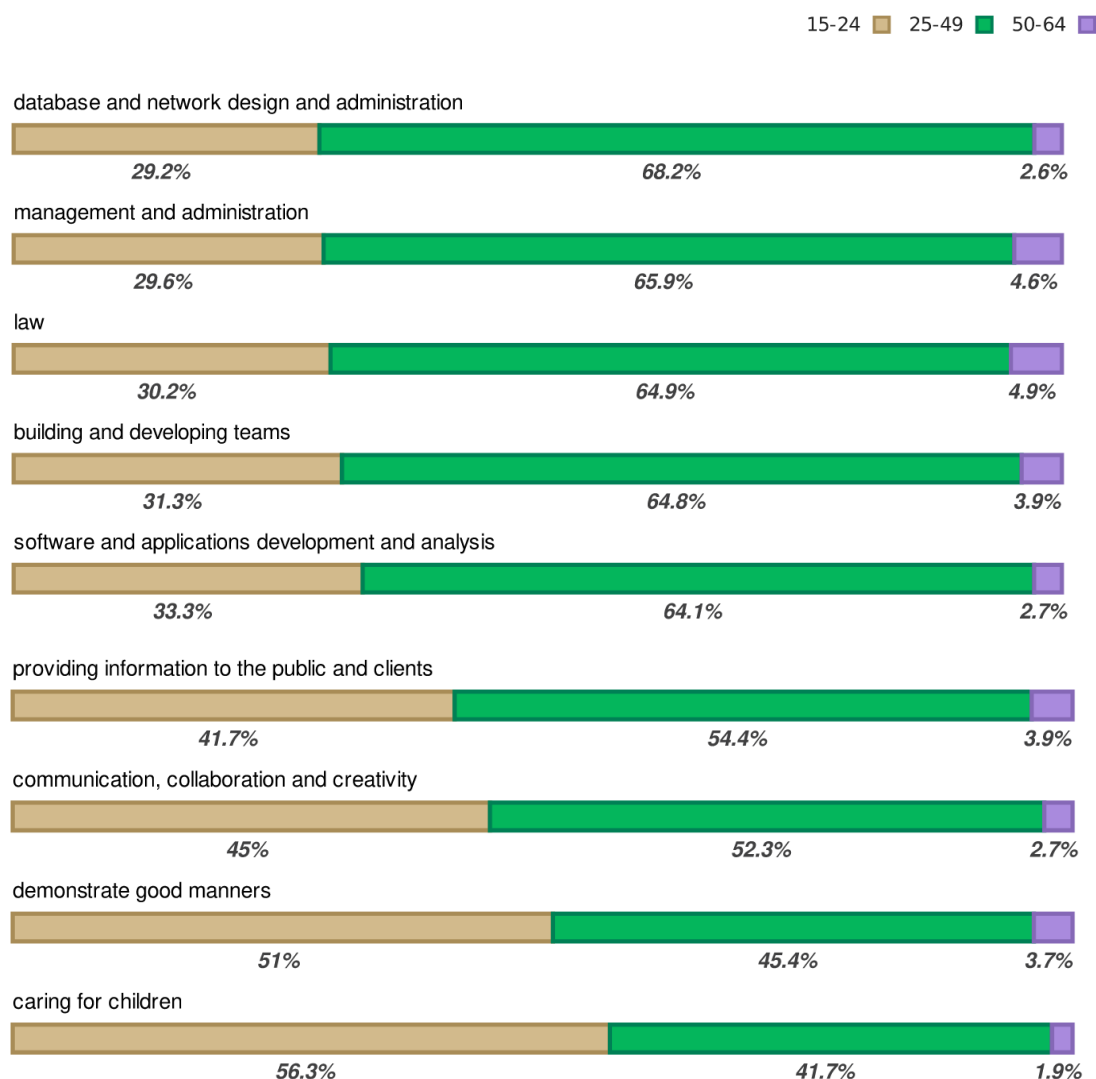


Figure 34. Age group breakdown of the most common ESCO skills mentioned in Europass CVs

- Skills that belong in skill groups that require more explicit professional experience, such as *management and administration*, *law*, and *software and application development and analysis* are more frequently referenced in the CVs in the age group 25-49.
- On the other hand, skills mentioned more frequently in the age group 15-24 tend to be more ubiquitous and generic in nature, such as *demonstrate good manners* and *caring for children*.
- With regard to Organisational and Communication skills, the older age groups tend to include concrete skills and knowledge on subject matters related to *law*, *building and developing teams*, and *management and administration*.
- Conversely, the younger age group more commonly includes skills acquired through experiences in *sports*, *language acquisition*, and soft skills like *demonstrating good manners* and *caring for children*.
- Younger individuals tend to focus on Computer skills related to *using digital tools for processing sounds and images* and *using word processing*, while older individuals report *database and network design and administration*, *setting up computer systems*, and *software applications development and analysis skill groups*.

7.1.3 Gender

As with age groups, the most common skills overall, are mostly the same for the two genders. Instead, we present the skills that display the biggest ratio difference in their gender breakdown.

- Approximately 53% of Europass CV owners have disclosed their gender as male or female on their CV.
- The gender breakdown of skill groups has been calculated, and skills have been sorted by gender ratio.

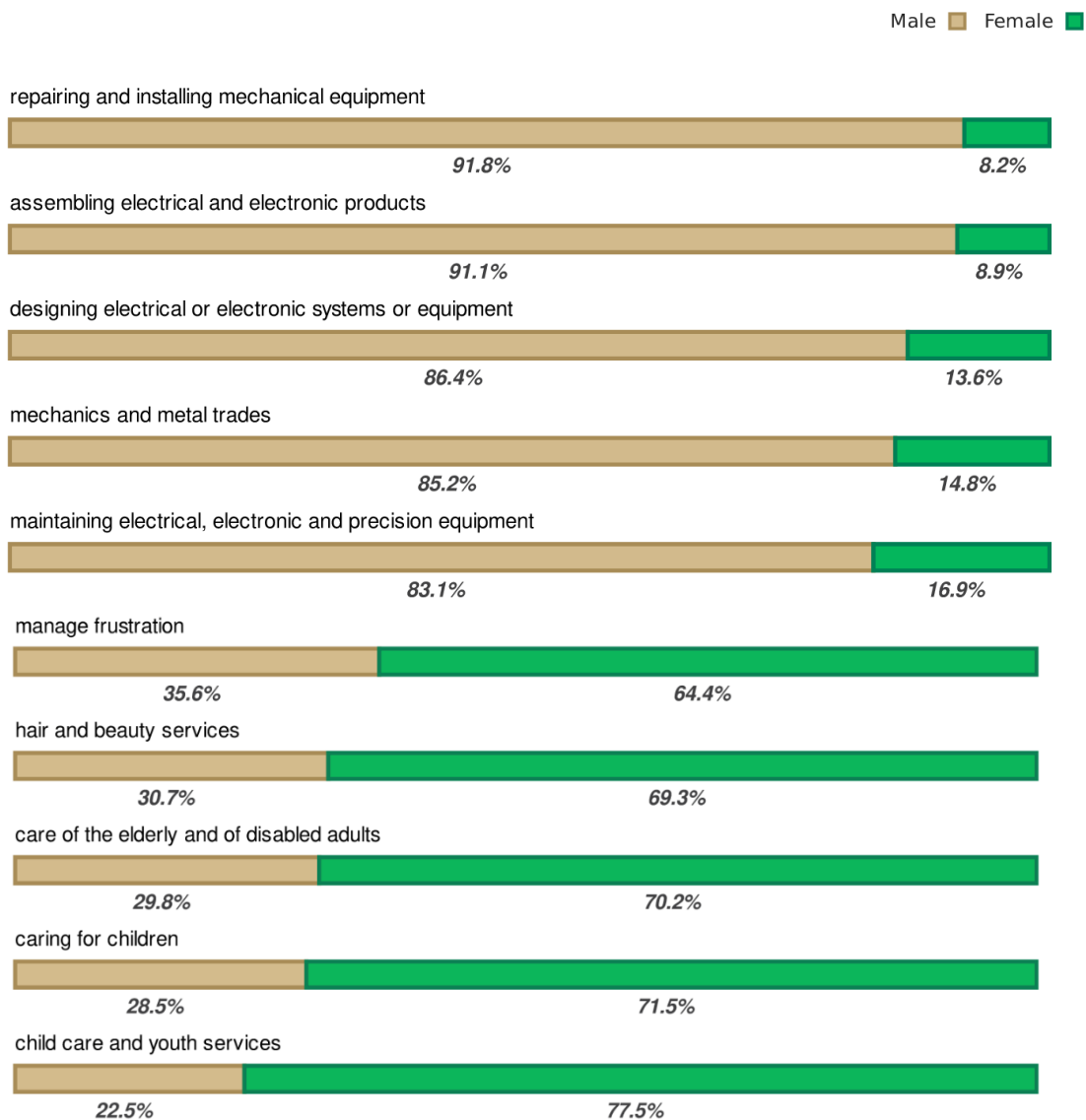


Figure 35. ESCO skills displaying the highest deviation between the two genders

- Male users tend to report technical skills more frequently across the board, such as those that belong in skill groups *repairing and installing mechanical equipment*, *assembling electrical and electronic products*, and *designing electrical or electronic systems or equipment*.
- Female users tend to report more skills where human interaction is required, such as *child care and youth services*, *caring for children*, and *care for the elderly and of disabled adults*.

7.1.2 Country

- Approximately 96% of Europass CV owners reported their country of residence at the time of CV creation.
- The top 5 skills groups of Italy, Portugal and Romania are displayed below.

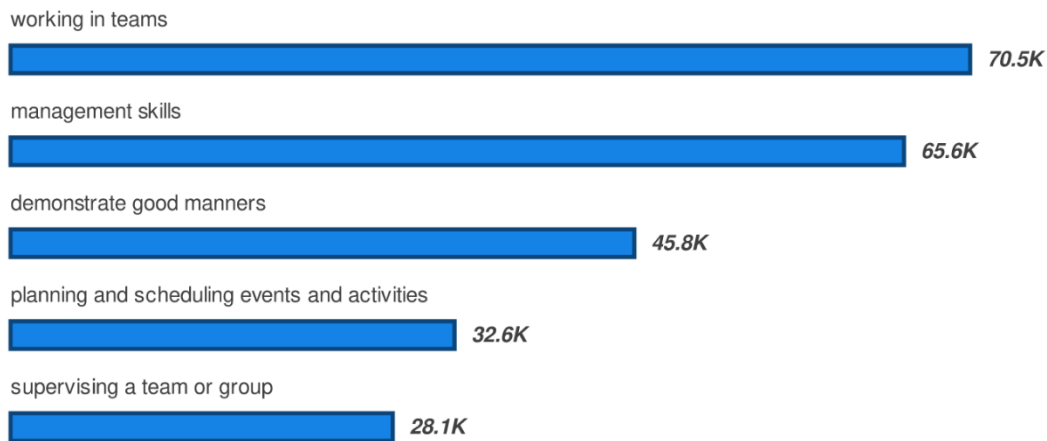


Figure 36. Skills distribution for Italy

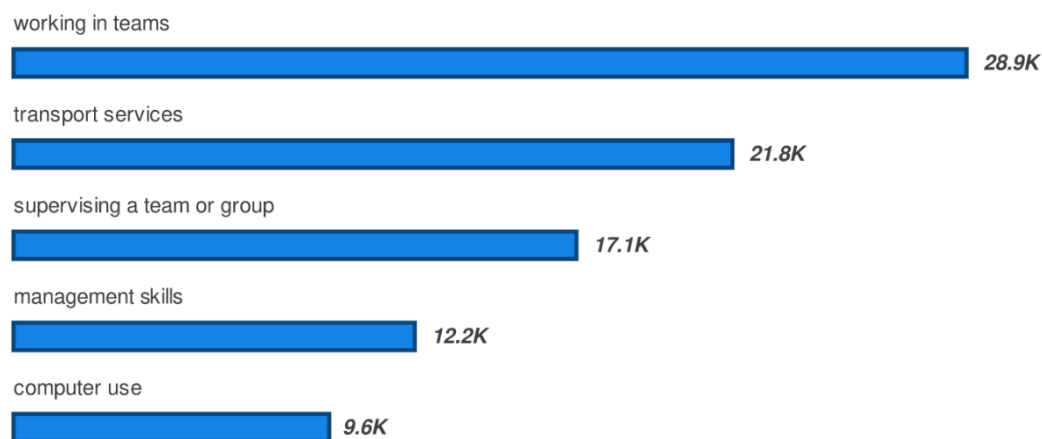


Figure 37. Skills distribution for Portugal

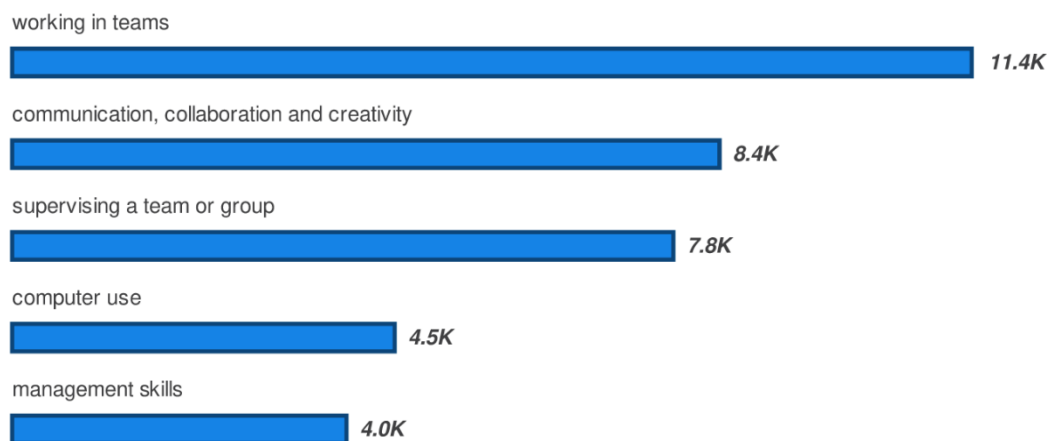


Figure 38. Skills distribution for Romania

- The most common skill groups observed overall, such as *working in teams*, *management skills*, *computer use*, and *supervising a team or group* appear relatively consistently among the top skill groups of all countries.

7.2 Occupations and Skills

7.2.1 Skills by Occupation

Given that Europass CVs contain information both for skills and for occupations, we attempted to understand how skill inclusion patterns differ between individuals from different professional backgrounds.

- We display the top 3-digit skill groups for the top ISCO 3 occupations by considering the chronologically last reported work experience of each CV owner as their latest job and all of their reported skills.

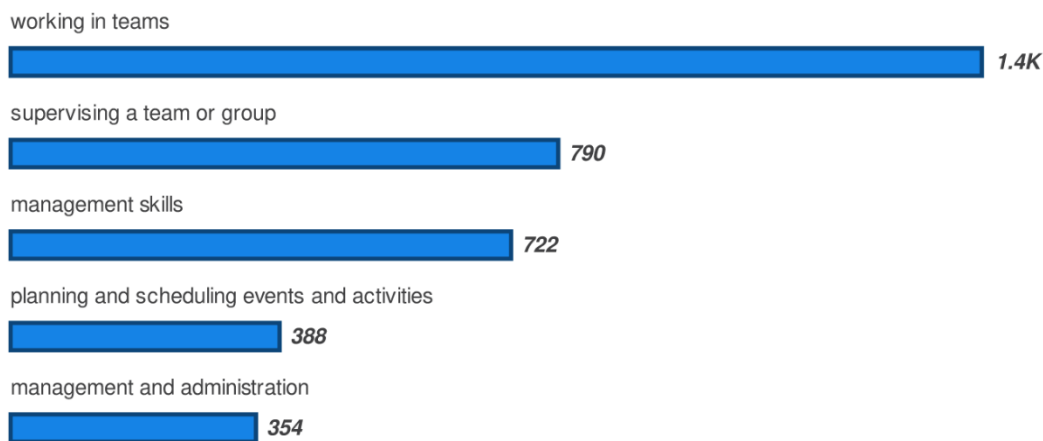


Figure 39. Top skill groups for waiters and bartenders

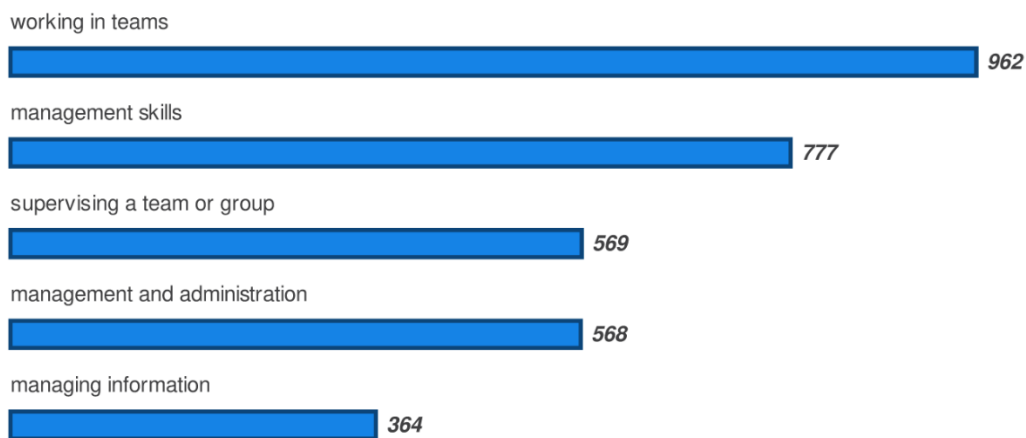


Figure 40. Top skill groups for administrative and specialised secretaries

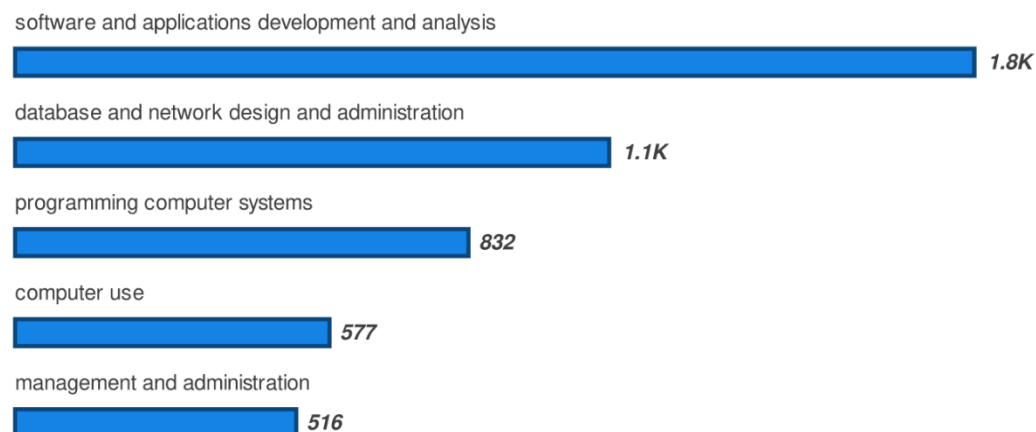


Figure 41. Top skill groups for software and applications developers and analysts

- We note that regardless of occupation, the most frequently mentioned skills generally belong in skill groups such as *working in teams* and *management skills*.
- Skills that are very specific to occupations, such as *law* for *Legal, social and cultural professionals* and *software and applications development and analysis* for *Software and applications developers and analysts* do occur, but in general the most common skills are ubiquitous.

7.2.2 Importance of Skills to Occupations

Given that certain skill groups are almost equally represented across the majority of occupations, in order to identify what skills are truly relevant to each occupation, we measured how over- or under-expressed each skill is for each occupation.

- We make use of the revealed comparative advantage (RCA) index, which we calculated for every pair of occupation and skill.
- An occupation-skill pair with $RCA > 1$ implies a comparative advantage, that is to say, an increased likelihood of inclusion of a particular skill in CVs of people in a particular occupation.

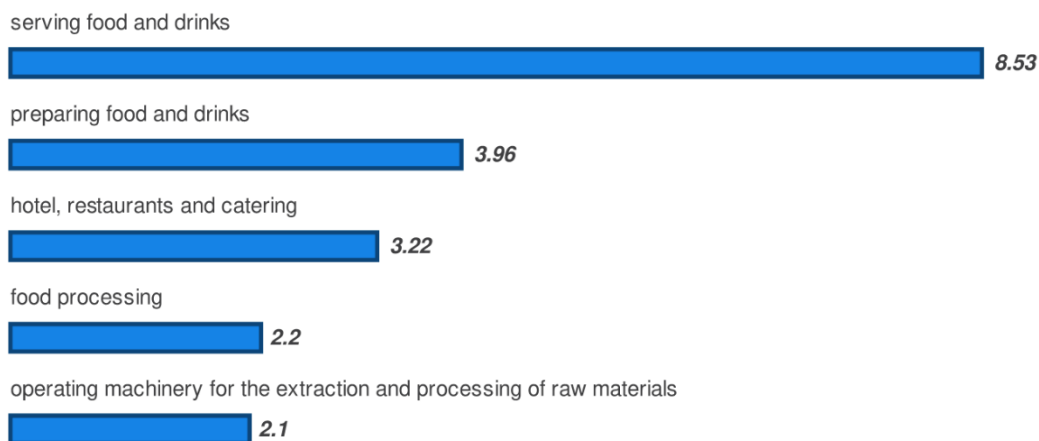


Figure 42. Top skill groups by RCA for Waiters and bartenders

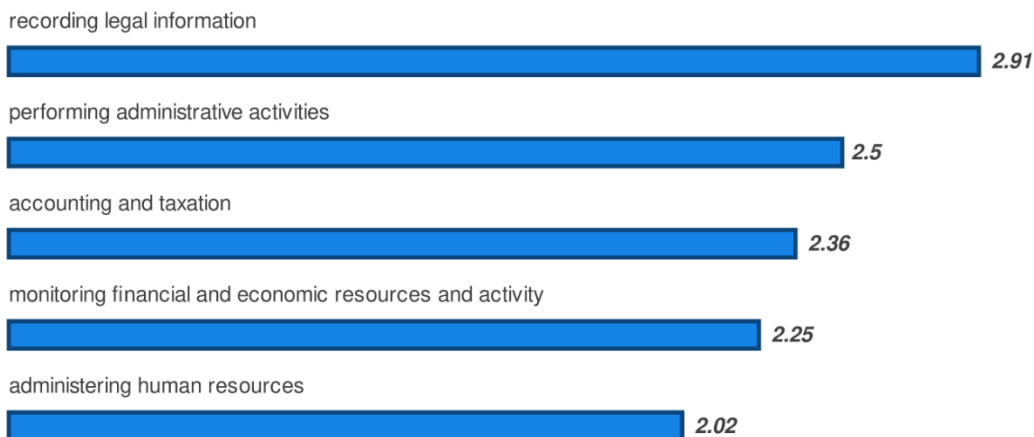


Figure 43. Top skill groups by RCA for Administrative and specialised secretaries

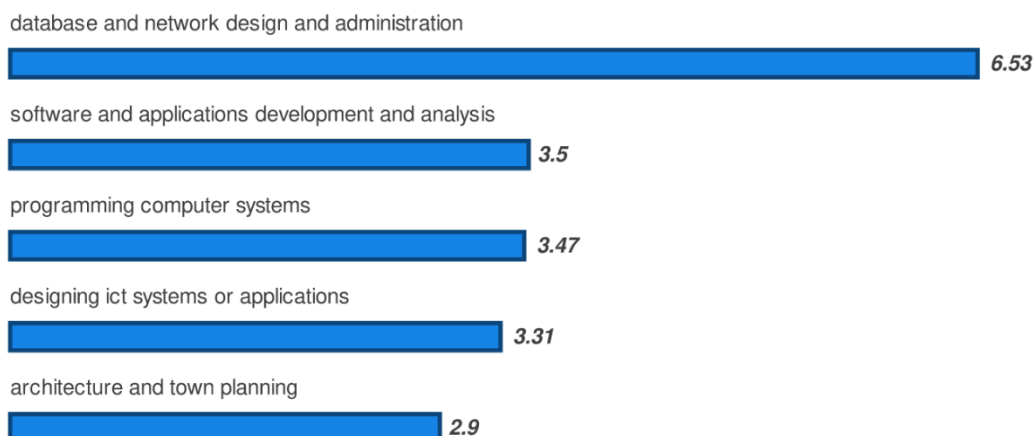


Figure 44. Top skill groups by RCA for Software and applications developers and analysts

- The RCA index reveals the skills expected for each job, with *-serving food and drinks* having over 8 times a comparative advantage for *Waiters and bartenders* and *database and network design and administration* being over 6 times more overexpressed among *Software and applications developers and analysts*.
- Skills that were shown to be ubiquitous, such as *working in teams* have received a low RCA score across the board and are therefore not characteristic of any occupation in particular.

Discussion: We find that using the RCA index is highly successful in mining associations in labour market data. Utilization of the RCA index, like with lift and confidence in association rules mining, can assist researchers developing the ESCO model to enhance it with new relationships between occupations and skills.

7.2.3 Skillscape

Using the RCA index, occupations can be distinguished based on their “effective use” of skills, through which we can also define skill complementarity as a measure of the frequency of a pair of skills being used effectively by the same occupation.

- Each node in the following graph represents a 3-digit ESCO skill group and they are linked according to skill complementarity.
- The colour of each node represents the ISCO 1 occupation for which the equivalent skill has the biggest comparative advantage.
- The Fruchterman-Reingold layout algorithm based on skill complementarity is utilized to create the following Skillscape.

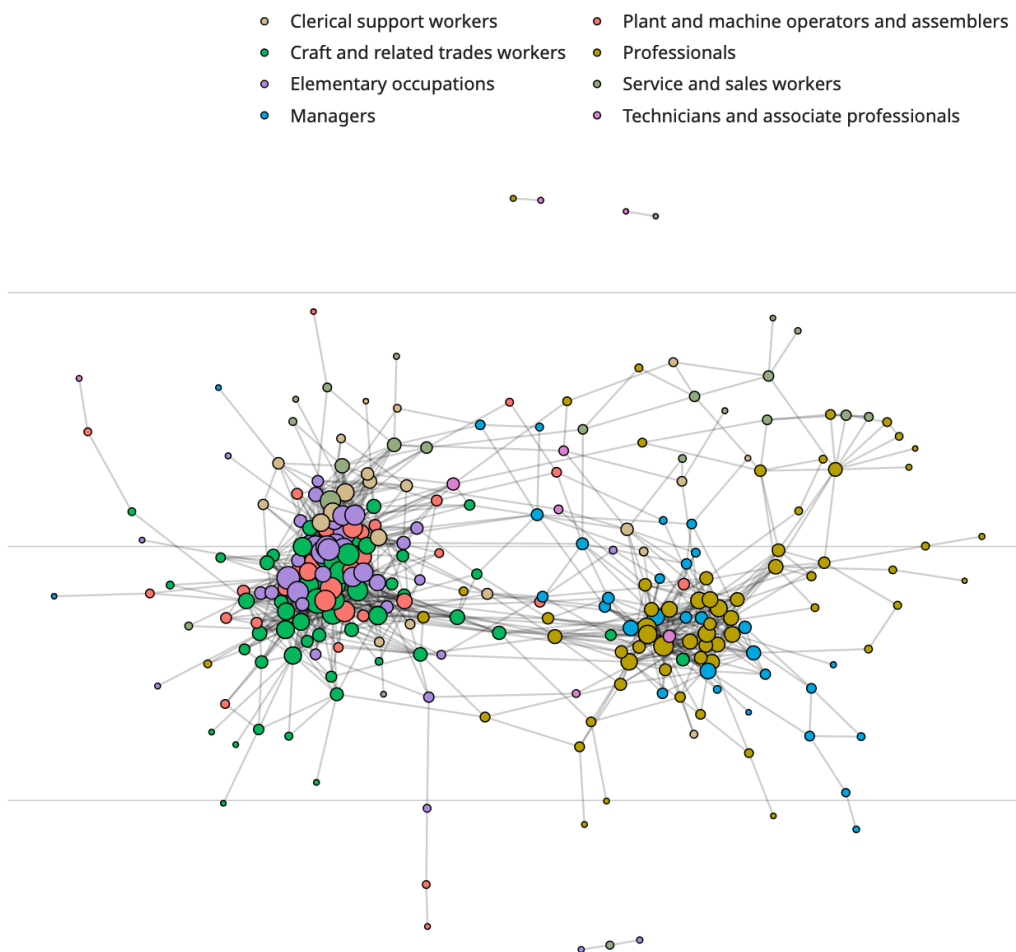


Figure 45. Skillscape

- Two main clusters are observed: one composed of skills primarily included on CVs of people whose latest job belongs on ISCO 1 occupations *Professionals* and *Managers*, and another of skills from occupations belonging on ISCO 1 occupations *Elementary occupations*, *Craft and related workers*, *Clerical support workers*, and *Service and sales workers*.

Discussion: This exercise is a reproduction of results observed in the literature. Skillscape (Alabdulkareem, et al., 2018) can be seen as an indicator of the polarization of workplace skills.

7.3 Skills of NEETs

7.3.1 NEETs in the Europass Dataset

Europass CVs may include information on the demographic characteristics, work experiences, qualifications, and skills of individuals. Using the temporal fields of work experiences and qualifications, it is possible to determine the current situation with regards to education, employment and training for individuals.

- CV owners have been determined as employed or in education if they have included an ongoing work experience or qualification respectively. If status on either area is not included, they are classified based only on the one disclosed. If neither employment nor education status is included, they are marked as “Unknown”. Category “NEET” is composed respectively of CV owners that shared both education and employment status, with NEETs between 15 and 29 marked separately from the rest.

- The graphs below compare the demographic characteristics of CV owners based on their status at the time of CV creation.

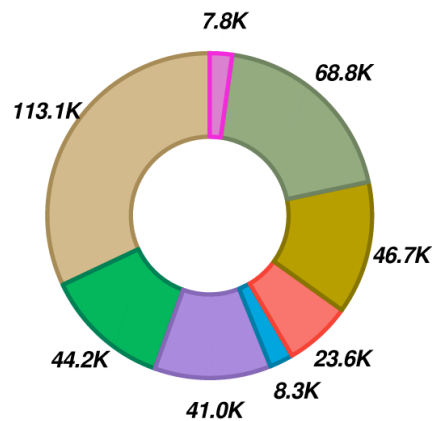
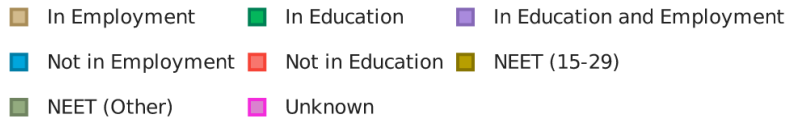


Figure 46. Breakdown of the education and employment status of Europass users

- Approximately 33% of CV owners are classified as NEET, of which 71% are between 15 and 29. Fifty-six percent (56%) are in employment, education, or both, while not enough information is provided to determine the status of the 11% of CV owners. About half of them who are in education are also employed.
- Noting the roughly 5% over-representation of males in the dataset, both male and female users are roughly equally likely to be classified as NEET. Active employment is slightly more likely for male individuals, while active education is slightly more like for female ones.
- The age breakdown of NEETs is slightly skewed towards younger ages compared to individuals who are active in education, employment, or training.
- Europass CV owners classified as NEET are over 40% in Portugal and Ireland. The lowest percentages are seen in Estonia and Finland, where it is under 25%.

7.3.2 Distribution of Skills among NEETs

A total of 123.5K CVs participating in the survey where the owners' reported age is between 15 and 29 have included information about their education and employment status. Of those, approximately 46.7K are *not in education, employment, or training (NEET)*, while the rest are in education, employment, or training (*EET*).

- The top 10 3-digit skill groups of CV owners classified as NEET overall and on each skill type are displayed in the graph below. The number shown is the percentage of NEET CVs that included each skill. The equivalent number for users in education, employment, or training is also included for comparison.

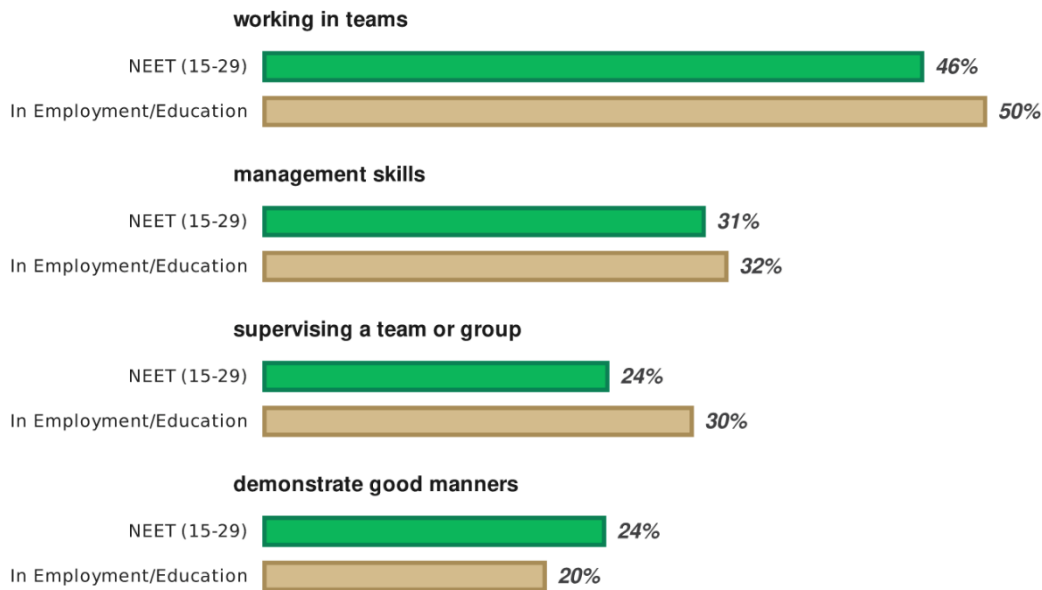


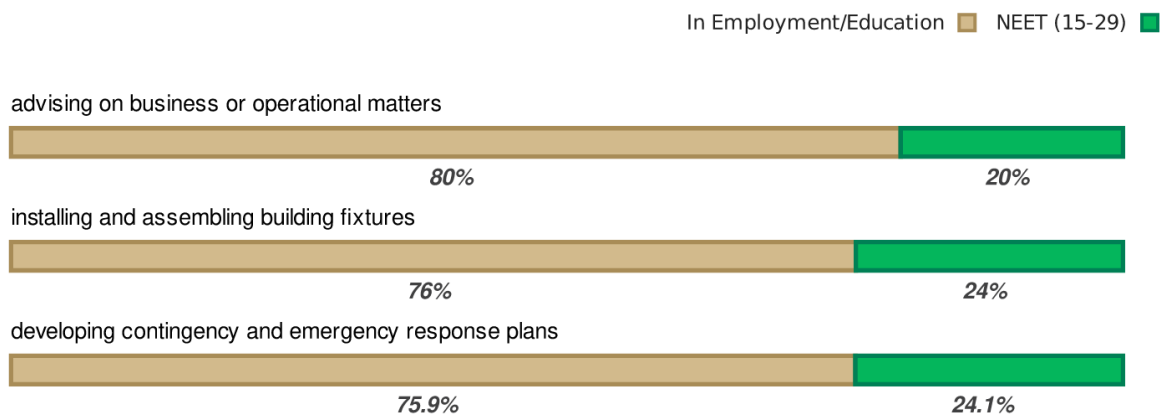
Figure 47. Comparison of the inclusion of the most common ESCO skills between users in employment/education and NEET users

- The most common skills seen in CVs created by NEETs are roughly the same as those seen in individuals active in education or employment. However, the average frequency of inclusion of skills is lower for NEETs.
- The frequency is especially lower for skills that suggest experience in the work force, such as *supervising a team or group*, and specialization in one particular area, such as *software and applications development and analysis*.
- On the other hand, inclusion of more generic skill groups such as *demonstrate good manners* and *providing information to public and clients* is slightly more frequently included.

7.3.3 Most Deviated Skills

It is possible to measure the difference in the relative frequency of inclusion of skills between CV owners that have been classified as NEET and those that are active in education, employment, or training. To do this, we calculated the ratio of the two categories (*NEET or not*) for each skill.

- The graph below reports the break down by status in education, employment or training of the 20 3-digit skill groups with at least 200 observations displaying the biggest relative difference overall and individually for each skill type. The number reported is the percentage that each category represents on each skill group respectively.



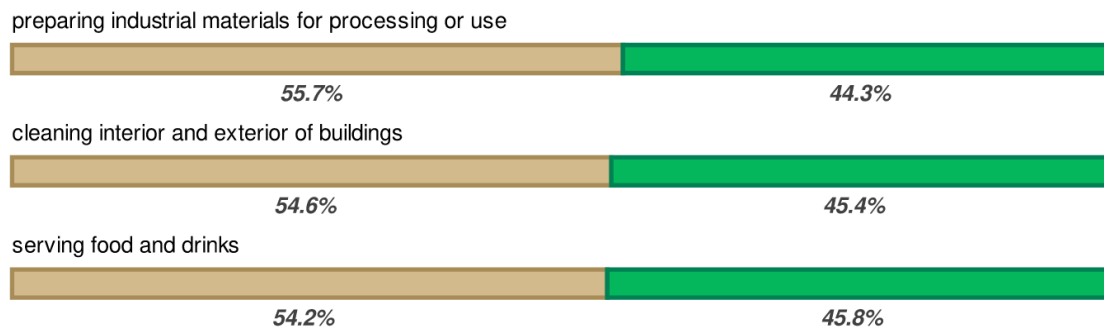


Figure 48. ESCO skills displaying the highest deviation between users in employment/education and NEET users

- Overall, CV owners currently active in education, employment, or training tend to report skills requiring more professional or academic specialization. For example, *advising or business or operational matters, providing financial advice and dental studies* are much more often reported by these individuals.
- On the other hand, CV owners determined as NEET have a higher frequency reporting more generic skills or skills related to manual labour. For example, *serving food and drinks, storing goods and materials, and demonstrating willingness to learn* are relatively more common among these individuals.

8. Executive Summary

8.1 Discussion

The use of new online data sources together with creative research methodologies to analyse such data is on the rise. This development impacts the ability to undertake labour market research both on the supply (i.e., the workforce) and the demand (i.e., available jobs) side (Edelman, 2012) (Sapleton, 2013) (Taylor, 2014). This study analyses online CV data in a variety of ways while also addressing methodological concerns with web-based/online sources. The degree of representativeness and the possibility to generalise the findings are two critical issues with the Europass CV dataset. We outlined a variety of statistical strategies for coping with bias effects, such as the platform's ever-increasing traffic, presented tools for accurate estimation of various population parameters, and harvest patterns from the Europass database. The samples in question were not chosen randomly; they were generated by the Europass online CV editor from 2017 up to the second quarter of 2020, which is similar to how online job ads are generated. Both the Europass CV dataset and datasets from the online job ads side, contain information from the "Online Labour Market" ecosystem and are based on naturally-occurring data.

Taking into account that every data collection exercise has its limitations, naturally-occurring data in the "Big Data" era can play an important supplementary, to official statistics, role. (Kitchin, 2015) Additionally, due to its vast size, this type of data can be considered as an accurate representation of particular economic environments of the labour market. With the trade-off of randomisation as it is employed by surveys it offers an appealing side of "realism". In economics, representativeness of the environment, rather than representative of the sampled population, in many cases is the most crucial variable in determining generalizability. (List, 2007)

Interestingly, in many studies analysing printed or online job advertisements, representativeness issues are not widely discussed while the findings and conclusions are not always generalised in a scientifically appropriate manner (Kureková, Beblavý, & Thum-Thysen, 2015). In our approach, we have tried to gain insights regarding the representativeness of our dataset by comparing our findings with the data of the EU Labour Force Survey (EU-LFS). This comparison reveals significant differences between the results of randomised surveys on the labour market and the Europass CV data. However, noting that the environments of data collection drastically differ, it can be argued that the Labour Force Survey data collection environment is not indicative of the actual market of people actively looking for a job. The latter is difficult to capture in a randomised manner, and Europass data offers an alternative corpus, which is especially comprehensive for younger populations i.e., the most representative group of people looking for a job.

Rather than dismissing research efforts based on online CV data and other types of web-based data due to representative flaws, we propose that a discussion is held on how these flaws can be compensated, and for which types of research questions and fields may be of less concern. (Daas, 2015) Calculating relative frequencies and comparing between groups are useful methods that we have employed as they also partially "regulate" the dataset; for example, variation in traffic that will be ruled out if we measure relative changes of ontologies across time.

Another interesting topic is **regression analysis between age and work experience**, that provides insights about career path and the process of job maturing of the average individual. Furthermore, time series analysis of count data can uncover trends in Occupations and Skills across representative breakdowns.

Furthermore, the Europass CV dataset proved to be particularly successful in mining associations between skills and occupations. By applying probabilistic estimations, relationships can be mined in a data-driven way, confirming known associations as well as helping establish previously undocumented ones. Additionally, association rules mining can easily be applied by considering the analogy of a CV to a shopping cart, where the user "buys" items from a given taxonomy (e.g., age group, ESCO skills, ISCO groups, etc.). Furthermore, by applying a similar method to association rules mining based on the adjusted RCA index, we observe that intuitive patterns between skills emerge and reproduce results met in the labour market research literature.

We highlight that adjusting for representativeness is a particularly difficult task and that can be attributed to the fact that the actual characteristics and structure of the population that is actively making their CV is practically unknown. The main caveat is the difficulty of actually identifying this part of the job market in its natural environment, and collecting a realistic dataset. For this reason, adjustment methods to our dataset, such as weighting, cannot fully be employed as to the best of our knowledge, no dataset accurately captures this segment of the labour force. Therefore, calibration methods should be used in a conservative way. Nevertheless, simple weighting is useful so to obtain a better understanding on the sensitivity of the results with respect to the distribution of candidate breakdowns, such as for example age groups. Thus, comparing adjusted with raw data can provide some useful feedback on the uncertainty of the results, although more sophisticated techniques like Monte Carlo simulations can give a much more realistic way of describing sensitivity of the results to particular variable distributions.

Based on the exploratory and expository data analysis of this overview and the statistical approaches we applied, we would also like to offer more general recommendations with respect to the usage of online CV data for future research.

Firstly, similar to job vacancy analysis, **the country-level representativeness and reliability of the data source must be assessed**. Focusing the analysis at country level, dominant market share of occupations in countries can be viewed as a good source of information that can lead to credible and transferable research findings.

Secondly, **representativeness and reliability must be evaluated in relation to a certain study topic**. Certain characteristics of coverage and sampling flaws can be addressed by using a data segment or sub-sample that can be regarded as representative. These biases are likely to be less pronounced in professions or labour market niches that are heavily exposed to the Internet (for example, IT-related professions). Unlike survey data, more detail is provided.

Thirdly, because online CV data and other sources of online vacancy data are strongly connected, **online CV data could be combined with other sources of online vacancy data depending on the research question**. We note that job sectors are likely to be exposed online similarly from the side of demand as well as supply. Moreover, online vacancies are expected to cause a driving effect on online CVs to be "tailed" to the jobs advertised. This leads towards an equilibrium between online job vacancies and online CVs created. Evidence of the later are found to the dataset of this study, since the distribution of estimated ISCO codes is highly correlated to that of online job advertisements, despite the well-known gap between supply and demand of the labour force.

Finally, **advanced statistical methods based on missing data, such as model-based approaches to the imputation of data not missing at random, could be employed to remove biases resulting from the structure of online CV data**. Also, it would be useful to link this dataset to the dataset generated by the new Europass CV editor. We also underline the fact that the Europass CV editor hosted by Cedefop stopped working a couple months after the COVID pandemic. By comparing these datasets,

insights could be derived for the characteristics of the users before and after the pandemic. This is a challenging task since a different UI/UX of the new web application might affect the user behaviour and introduces different sets of biases. However, large effects resulted from the pandemic and the digitization of the job market might still be observable and quantifiable.

The good news for the arguments on naturally-occurring data from online sources is that Internet-based applications from both the labour market supply and demand (search engines, online CV builders, online job advertisements, and so on) are likely to become the dominant ecosystem for large segments of the labour market used for job matching, which will considerably increase the percentage of workers and firms that participate in it (Askitas & Zimmermann, 2015). In comparison to traditional employment channels and processes, the online job market can offer a greater variety of options as well as increasingly sophisticated tools for assessing the suitability of a job or a job prospect.

Understanding the vast nature of the internet job market has far-reaching implications for attempts to revive underemployed populations and eliminate long-term unemployment. Career counselling services, the development of online CV preparation tools, second-chance education and training, and the integration of disadvantage job seekers into the labour market can all benefit from improved awareness of what employers' demand and what is available. Other sectors of public policy could benefit from the use of web-based labour market data. The desire to seek for new data sources and analytical methodologies, as well as to increase the reliability of the results, is motivated by not only intellectual, but also potential practical benefits.

8.2 Key Findings

Europass CV data provide significant labour market intelligence for highly-specialised young job seekers

The composition of Europass CV owners does not equally represent the different groups of the labour force (see 4. *Understanding the Europass CV dataset*). We find that people under the age of 32 create online CVs much more often than older individuals. Work experiences more typically reported on online CVs tend to require more specialization, which makes *Professionals* and *Managers* more common than on the general population. University education is most frequently observed. Additionally, the main motivation behind creation of CVs is job search. Making general statements about the labour market as a whole proves to be challenging, but sources like Europass can be used to get specific insight on some of the more well-represented subgroups through custom queries.

The study was able to reproduce some of the metrics observed in official statistics (EU-LFS)

Employment trends over the short period of data collection as they are calculated using the Europass CV data generally align with their counterparts from official sources of statistics for population over the age of 25 with respect to ISCO 1. Small year-to-year fluctuations noted on most groups of occupations are observable in Europass data, meaning that online CVs and similar sources of data can complement official sources on the identification of such trends and assist survey designers and policy-makers (see 6.1.3 *Trends in Employment by Occupation (EA19)*). Measurements over longer timespans using the career histories as they appear on CVs is more challenging due to recall bias (see 5.1.1 *Trends in Job Sectors*). However, through careful data calibration, focus on indicators of relative change instead of absolute values, and use of custom metrics (e.g., ratios) that help to smooth out some of the bias, patterns may also emerge over longer timespans via exploration of career histories (see 5.1.3 *Recruitments over Terminations in Time*).

A strong correlation between occupations in Europass CVs and online job vacancies was observed

The distribution of occupations observed in online vacancies resembles the equivalent distribution observed in Europass CVs (see Chapter 6.3 *Job Vacancies (Cedefop)*). This observation is true for the distribution of the most recent occupations in CVs of users that were employed at the time of CV creation, as well as users that were unemployed (see Chapter 6.2.1 *Unemployment by most Recent Occupation per Country (2019)*). This means that making measurements related to the demand side of the labour market online using CV data can not only enhance the already existing sources that measure demand, but also help to identify gaps between supply and demand either on the EU member state level, or across EU as a whole. Research of this gap and its evolution in the long-term can inform policies that help correct it, for example through new education and training programmes, or adjustments in migration and mobility policies.

The study demonstrates that the relationship between the ease of accumulation of work experience and the different occupations can be quantified through the analysis of online CV data

Using career histories reported on CVs, we find that accumulated work experience varies across the many occupations. This finding suggests significant variation in the career stability and ease of accumulation of experience of different groups of users, as well as numerous patterns with regard to their labour market entry (see Chapter 5.3.3 *Work Experience by Age*). Analysis of sources like Europass, which provide not only the current employment status of its users, but also a window to their past labour market participation, unlocks unique opportunities for identifying the more disadvantaged occupation groups and unravelling their demographic characteristics. This information is critical to taking measures against frictional and structural unemployment, which are two of the main causes of difficulty in work experience accumulation. Additionally, it can inform vocational planning and associated education and training policies.

The study shows that Online CVs research can extend the ESCO classification by revealing additional associations between occupations and job-related skills

The identification of skills that are specific to certain occupations is possible through analysis of online CVs. We have mined associations between occupations and skills by applying traditional market basket analysis and have been able to confirm relationships known to exist within the ESCO classification, as well as establish others that were not documented in the model (see 5.4 *Skills-to-Occupations Associations in the ESCO Model Compared to the Collected CV Data*). Additionally, filtering of ubiquitous skills has been shown to be possible using indices known from economics, helping to establish which skills are over-expressed (and thus more important) to which occupations (see Chapter 7.2.2 *Importance of Skills to Occupations*). We have demonstrated an application of this type of data transformation by reproducing the polarization of workplace skills as it is known in the literature using the Europass CV data (see 7.2.3 *Skillscape*).

Mapping unstructured text to statistical frameworks and classification systems through open-source libraries greatly enhances labour market research capabilities

Going from unstructured, free text to a restricted number of well-defined classes (e.g., occupations) is a major requirement when dealing with sources of online CVs. To this end, we developed software that takes user-input free text and matches it to the ESCO classification, as well as to the European Qualifications Framework and the ISCED Fields of Education and Training. We proceed to publish this software as open-source statistical packages that may help other researchers working with sources similar to Europass. `labourR`² has already been published on CRAN³, with `educationR`⁴ and

² <https://github.com/eworx-org/labourR>

³ <https://cran.r-project.org/web/packages/labourR/>

⁴ <https://github.com/eworx-org/educationR>

iscoCrosswalks⁵ currently available on GitHub and on the early phase of the CRAN submission process.

8.3 Perspectives for Future Research

Evaluation of skills mismatch

Qualifications are a strong indicator of the skills and individual possesses. Through mapping education and training to a given taxonomy of skills and qualifications, significant information on the skills present on the labour market might be gathered. A classification system of skills linked to a qualification system (e.g., link ESCO skills to ISCED) could create great opportunities for further insights in the attributes of the labour force, and the skills supply.

Too many people work in jobs that do not match their skills and qualifications (horizontal/vertical skills mismatches, underqualification/overqualification, etc.). Inferring skills from qualifications and training using an indirect approach can provide estimates for the supply of skills. On the other hand, occupations reported in a CV provide also an indirect estimate of skills, and can be considered as an approximation of demand. Comparing skills inferred from qualifications and occupational descriptions can be used to make econometric models to assess mismatches between labour supply and demand.

Research on language skills and mobility

The EU workforce is ageing and shrinking, leading to skills shortages in some cases. To compensate this, it is necessary to increase labour market participation and productivity. One important aspect is to increase the productivity and the efficiency of skilled workers, reduce brain drain, while facilitating mobility of citizens. On this ground, analysis of possible effects and correlations of language skills to working and/or studying abroad, could provide quantitative evidence of the effect of multilingualism and education on mobility of citizens.

Development of more open-source software to assist labour market research

The utility of software libraries in labour market research has been highlighted through exercises performed in this study, as well as feedback on the already published labourR library. Further development, maintenance and support of open-source software is critical for the future of labour market research, as not only does it accelerate the ability to answer specific in-depth questions without the need for separate research entities to solve the same problem again and again, but it also establishes the presence of labour market researchers in the broader software community for statistical research (e.g., on CRAN), enabling continuous feedback and improvements. Libraries missing from the community include additional text mining / machine learning libraries like labourR and education (e.g., skillsR), as well as bindings of relevant APIs (e.g., ESCO) to popular programming languages like R and Python.

⁵ <https://github.com/eworx-org/iscoCrosswalks/>

Annex A: Methodological Notes

Weighting

The weighting procedure employed is iterative proportional fitting (IPFP), which is an algorithm used in many different fields such as economics and social sciences. Through IPFP, the given distribution is updated with respect to given target marginal distributions. The R implementation of IPFP in the package *mipfp* was utilized. More specifically, we have focused on adjusting age so that it is less skewed towards younger population, and country of residence, so that we can derive a less biased European average. We have opted to use the official demographic statistics of Europe as published by Eurostat for our target marginal distributions.

The weighting procedure is applied initially for each country individually with respect to age, with the goal of adjusting the age distribution of each country. Subsequently, a separate weighting scheme has been derived for each potential European average. Specifically, we have adjusted the distribution of countries and ages in order to derive EA-19 (Euro area), EU-27 (UK excluded) and EU-28 (UK included) averages. Following the weighting procedure, weights are applied to the occupation dataset. More concretely, the distribution of occupations with respect to the variables measured is adjusted by multiplying the frequency of each disaggregation with its respective weight based on the country and age group observed.

To avoid adding bias through weighting, we have restricted the derived weights to be between 0.35 and 3. As such, biases with respect to country and age have not been eliminated entirely, so it is important to remember that metrics applying to the EA-19 are still subject to biases inherent to the dataset.

Time Series Analysis

In order to quantify trends, we have applied regression analysis to our data by fitting generalized linear models (GLM) to each breakdown of interest. Specifically, we have utilized the R package *glm* which estimates the coefficients β through maximum likelihood estimation (MLE). GLMs can be considered a generalization of ordinary linear regression that does not make the assumption that each observation necessarily comes from a normal distribution, $y_i \sim N(\mu_i, \sigma^2)$. Equivalently, this means that the error distribution of the response variable does not have to be the normal distribution.

A binomial distribution was assumed for our data, with logit serving as the link function,

$$estimate \times year + intercept = \ln\left(\frac{y}{n-y}\right)$$

And mean function,

$$\frac{y}{n} = \frac{1}{1 + e^{-(estimate \times year + intercept)}}$$

The process was repeated for both weighted and unweighted data. The results presented for the context of this report refer to trend estimations made for the weighted data, a sample of which is displayed below for EA-19.

The estimate statistic provides the percentage of the odds ratio increase for one unit of time (1 year),

$$odds \propto e^{estimate \times year}$$

where *odds* is defined as the ratio of the number of events that produce that outcome to the number that do not.

Correlation Analysis

Although scatterplots provide a lot of visual information, when there are a lot of variables, assessing the relationship between each pair using a single number might be useful. The Pearson product moment correlation coefficient is a measure of the relationship between two variables that may be estimated using the following formulae,

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma(x - \bar{x})^2 \Sigma(y - \bar{y})^2}}$$

where x and y are the samples, and the line above indicates the mean value. A positive covariance exists when x and y tend to go in the same direction across observations—to be both higher or lower than their respective means. There is no (linear) relationship between x and y if r is 0. While the variables have negative correlation coefficient, they tend to move in opposite directions in relation to their means: when x is lower, y tends to be higher.

Association Rules

Association rules mining through market basket analysis defines three specific metrics to quantify the relationships between items on each basket. Considering items A and B , the “support” of the rule $T(A \Rightarrow B)$ is the fraction of observations in the union of the two items. It can be viewed as an estimate of the probability of simultaneously observing both item sets in a randomly selected market basket.

$$Support(A \Rightarrow B) = T(A \Rightarrow B) = \frac{|(A \cap B) \subseteq T|}{|T|}$$

The “confidence” of the rule $C(A \Rightarrow B)$ is its support divided by the support of item A .

$$Confidence(A \Rightarrow B) = C(A \Rightarrow B) = \frac{T(A \Rightarrow B)}{T(A)}$$

Finally, the “lift” of the rule $L(A \Rightarrow B)$ is defined as the confidence divided by the support of item B , or consequently,

$$Lift(A \Rightarrow B) = L(A \Rightarrow B) = \frac{T(A \Rightarrow B)}{T(A)T(B)}$$

“Lift” essentially shows how much more likely it is for item A and B to co-occur on a randomly selected market basket when compared to the random chance. We have set thresholds $Support > 0.0001$, $Confidence > 0.03$ and $Lift < 20$, to identify the meaningful ISCO 3 and ESCO skill associations that display the highest lift.

Revealed Comparative Advantage (RCA)

In the context of skill - occupation pairs RCA is defined as:

$$RCA(o, s) = \frac{cv(o, s) / \sum_{s' \in \text{skill groups}} cv(o, s')}{\sum_{o' \in \text{ISCO3}} cv(o, s') / \sum_{s' \in \text{skill groups}, o' \in \text{ISCO3}} cv(o', s')}$$

Where $cv(o, s)$ is the number of CVs including a particular skill and occupation.

Annex B: Concordances

Eurostat, CEDEFOP, and other EU agencies have been looking into using the ESCO classification to link job–worker characteristics like skill importance and level ratings to European jobs. O*NET is a reliable source of occupational data for the European labour market because of its strong theoretical and empirical basis and the similarities between the European and US economic systems.

Drawing on O*NET necessitates mapping from one classification to another since the O*NET database classifies jobs differently than the EU. This is accomplished via a “concordance” (also known as a “crosswalk” or “correspondence table”). We propose the first publicly accessible software for methodologically transparent concordances, by constructing an R-package that we hope will greatly reduce the expenses and time required to perform an approximate mapping between the two categorization systems. We highlight various considerations, particularly in the mapping between different hierarchical levels between the two classes, in addition to presenting these concordances in an open, transparent manner.

ESCO began evaluating several options to tying skills to professions in the EU. The introduction of the Skills and Competencies Taxonomy, knowledge descriptors, which provides consistent nomenclature and enhances communication of occupational information in the EU, was the first stage in that process. One obvious source to consider as we evaluate the various options to utilize the ESCO Taxonomy is the US O*NET system. Indeed, to assess the skills and competencies of different segments of the EU labour market, various governmental and non-governmental organizations in the EU have built concordances between the US System of Occupational Classification (SOC) and ESCO.

The problem is that there is no clear one-to-one mapping from the EU ESCO to the US SOC utilized in O*NET for most jobs. As a result, many concordances have concentrated solely on occupations with clear one-to-one correspondences. Furthermore, these concordances are frequently proprietary, with little information or explanation about how they were created, making them inaccessible to others and limiting any critique of the approach adopted.

Understanding the variations in how jobs are organized in the US and the EU is necessary to realize the necessity for a more granular concordance. While the SOC and the ESCO have similar goals, there are significant distinctions in the organization and information included in occupational profiles. The US SOC has 867 6-digit classifications for statistics reporting (e.g., number of people employed or average wage). The EU ESCO, on the other hand, employs the ISCO 4-digit categorization, which has 436 unit groups, and adds an extra layer of information of approximately 3000 occupational distinctions. Both systems group occupations according to a four-level hierarchy, with ESCO adding an extra degree of detail.

We build an algorithm that conducts approximate matching between the ISCO and SOC classifications using concordances provided by the Institute for Structural Research and Faculty of Economics, University of Warsaw. The crosswalks offer a complete step-by-step mapping of O*NET data to ISCO-88 and ISCO-08 coding using an expanded version of SOC-00 and SOC-10 coding (used for most European data, including EU LFS). We propose a mapping method based on the afore-mentioned research that converts measurements to the smallest possible unit of the target taxonomy, and then performs an aggregation/estimate to the appropriate degree of detail.

Annex C: Library for Classifying Multilingual Text of Qualification Titles - educationR

EQF Level

Information on a person’s education level may be sourced from qualification titles sourced from a CV, online resources such as a LinkedIn profile, and more, and is most often in the form of free text. In various instances, ranging from the processing of applications in educational programmes, to recruitment, to labour market research, it is often required to standardize this information into a particular reference classification. The educationR package provides functionality for classifying free text into the European Qualifications Framework (EQF). The EQF was developed as a translation tool to make national qualifications from countries in the European Union easier to understand and more comparable. It defines eight reference levels tied to specific learning outcomes, ranging from basic (Level 1) to advanced (Level 8).

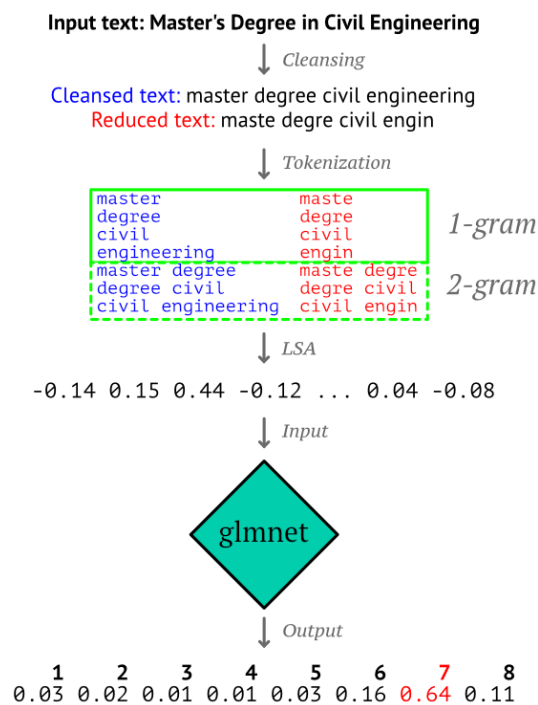


Figure 49. EQF level classification schematic

Dataset

The Europass backup database includes over 2.5M examples of qualification titles with their respective EQF level attached to them. We treat this as our labelled data which we use to derive our train/test split. Due to the dataset’s reliance on self-reporting and the nature of the process through which they are collected, where users’ familiarity with the European Qualifications Framework is presupposed, this labelled data may be considered noisy. Through manual labelling of a sample of this data, we find an accuracy of just over 70%. Given the volume of the data, we anticipate some of the noise to be eliminated through the pre-processing, but testing measurements should be treated with caution.

Feature Extraction

Every example in our data is composed of a document (i.e., the unstructured, free text used as a qualification title) and a label (1 through 8) representing its EQF level. In order to use this data to

train a predictive model, a pre-processing step where documents are embedded into a vector space is required. We use Latent Semantic Analysis (LSA), which combines the classical bag-of-words model used in text mining with Singular Value Decomposition (SVD). Prior to the application of LSA, a cleansing and transformation step is applied to simplify text, remove common words, as well as engineer new features (through substrings) to improve the accuracy of the predictive model. To help reduce noise, the vocabulary is pruned so that only terms that appear on at least 0.5% (and in no more than 50%) of all documents are considered. The number of dimensions used for LSA is 25.

Classifier

Multinomial logistic regression is used to train a classifier that takes as input a document embedded on the previously defined vector space, and outputs a predicted EQF level between 1 and 8. Specifically, the glmnet packaged is used, which offers implementation of fast algorithms for estimation of generalized linear models with ℓ_1 (lasso) and ℓ_2 (ridge regression) penalties, as well as mixtures of them (elastic net) using cyclical coordinate descent, computed along a regularization path. More specifically, glmnet was applied with 4-fold cross-validation and a separate model was produced for each language.

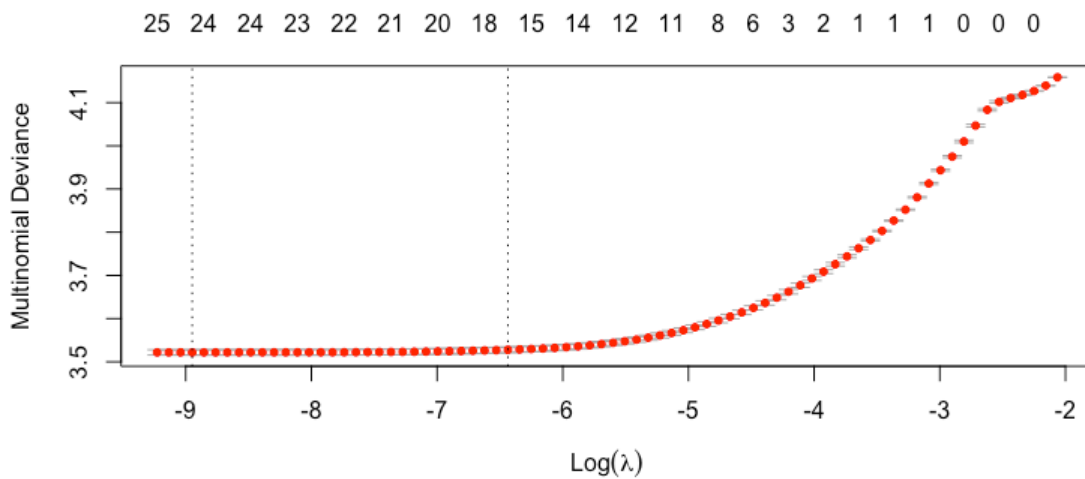


Figure 50. Cross-validation curves for the glmnet model for English. Cross-validation curves appear as the red dots, with upper and lower standard deviation shown as error bars.

Results

Table 4. Overall Statistics for Italian model

Accuracy	0.5338
95% CI	(0.5306, 0.537)
No Information Rate (NIR)	0.2027
p-value (Accuracy > NIR)	< 2.2e-16
Kappa	0.45

Table 5. Statistics by Class for Italian model.

	Class: 1	Class: 2	Class: 3	Class: 4	Class: 5	Class: 6	Class: 7	Class: 8
Sensitivity	0.30866	0.15175	0.55664	0.7576	0.25026	0.7758	0.74652	0.25527
Specificity	0.95133	0.96333	0.9415	0.8029	0.9385	0.9382	0.94011	0.9749
Pos Pred Value	0.2103	0.27291	0.58471	0.4942	0.47611	0.701	0.64848	0.45571
Neg Pred Value	0.97039	0.92604	0.93486	0.9287	0.84859	0.9573	0.96162	0.94083
Prevalence	0.0403	0.08315	0.1289	0.2027	0.18257	0.1574	0.12893	0.07607
Detection Rate	0.01244	0.01262	0.07175	0.1536	0.04569	0.1221	0.09625	0.01942
Detection Prevalence	0.05915	0.04624	0.12271	0.3107	0.09596	0.1742	0.14842	0.04261
Balanced Accuracy	0.62999	0.55754	0.74907	0.7802	0.59438	0.857	0.84332	0.61508

ISCED Fields of Education and Training (ISCED-F 2013)

An individual's participation on education programmes is often described on their CV, their LinkedIn profile, and other similar sources in the form of free text. For the purpose of statistical analysis (e.g., in labour market research) and other applications, it is often required to know which academic discipline or field of study corresponds to these qualification titles. The educationR package provides functionality for matching free text with the International Standard Classification of Education's Fields of Education and Training 2013 (ISCED-F 2013). ISCED-F 2013 is a classification of fields of education maintained by the United Nations Educational, Scientific and Cultural Organization (UNESCO) to increase the international comparability of education statistics. It contains a hierarchy of 11 broad fields (2 digits), 29 narrow fields (3 digits) and about 80 detailed fields (4 digits).

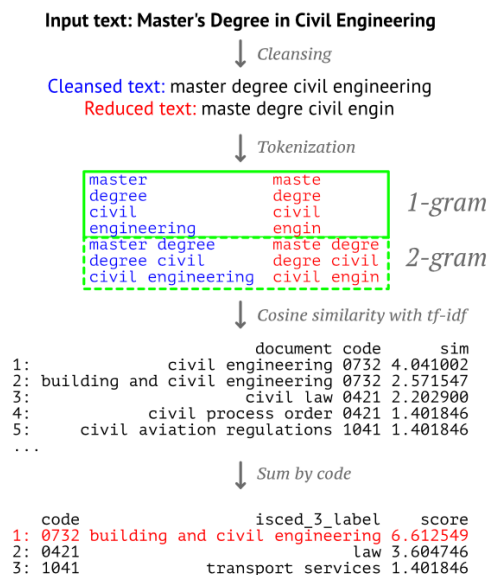


Figure 51. ISCED field matching schematic

Corpus of Documents

Matching free text to ISCED-F 2013 fields is performed through document similarity comparison. A corpus of documents is attached to every field, based primarily on the labels provided by UNESCO to all fields along the ISCED-F hierarchy. In addition to that, the ESCO classification is used to enhance this corpus. Specifically, ESCO utilizes ISCED-F 2013 as the Knowledge branch of its Skills & Competences hierarchy. Translations of the fields' labels are provided in over 27 languages, and

associated skills and competences are attached as leaf nodes to every ISCED-F detailed field (4 digit). As a result, over 2800 relevant ESCO skills are attached to each of the 80 detailed ISCED-F fields, with the corpus including not only labels of the fields (e.g., "travel, tourism and leisure") but also a lot of text related to them (e.g., "agritourism", "sightseeing information", "travel booking policites").

Pre-processing

Every document is composed of a text label and the 4 digit code corresponding to the ISCED-F field it refers to. Documents are initially pre-processed using a cleansing and transformation step which simplifies text, removes common words and engineers new features (through substrings). Consequently, a tf-idf (term frequency-inverse document frequency) statistics matrix is produced, reflecting how important a word is to each document in the corpus. Use of a transformed tf-idf weighted matrix instead of the more traditional bag-of-words matrix is useful given the nature of the problem being tackled.

Matching

The tf-idf matrix, along with the vectorizer and the fitted tf-idf model produced during the pre-processing step, are used for matching new documents with ISCED-F 2013. Free-text input is compared with the documents in the ISCED-F / ESCO skills corpus using cosine similarity with tf-idf. A similarity score between input and every document in the corpus is produced, and the ISCED-F code attached to the document that produced the highest score is given as output.

Results

A small number of labelled data exist in the Europass backup database. Note that labelled data are on an older version of ISCED Fields of Education and Training (2011) and we are only provided with codes to the 9 broad fields (2 digits) defined on this version. We transform results to this older version to produce some indicative metrics on the classifier's accuracy.

Table 6. Overall Statistics for Italian model

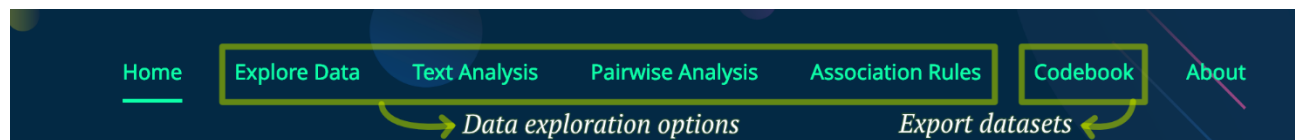
Accuracy	0.5038
95% CI	(0.5006, 0.507)
No Information Rate (NIR)	0.2835
p-value (Accuracy > NIR)	< 2.2e-16
Kappa	0.3857

Table 7. Statistics by Class for Italian model.

	Class: 0	Class: 1	Class: 2	Class: 3	Class: 4	Class: 5	Class: 6	Class: 7
Sensitivity	0.0229817	0.34339	0.52105	0.701	0.5288	0.43734	0.20112	0.38948
Specificity	0.9909666	0.96648	0.90062	0.7384	0.9097	0.96575	0.995703	0.94812
Pos Pred Value	0.0445714	0.29977	0.40508	0.5146	0.5903	0.74713	0.541763	0.37507
Neg Pred Value	0.9822421	0.9724	0.9354	0.8619	0.887	0.88122	0.980135	0.95104
Prevalence	0.0180068	0.04011	0.11494	0.2835	0.1975	0.18789	0.024639	0.07402
Detection Rate	0.0004138	0.01377	0.05989	0.1987	0.1044	0.08217	0.004955	0.02883
Detection Prevalence	0.0092846	0.04595	0.14784	0.3861	0.1769	0.10998	0.009147	0.07687
Balanced Accuracy	0.5069741	0.65493	0.71084	0.7197	0.7193	0.70155	0.598411	0.6688

Annex D: Exploratory Data Tool

To facilitate the dissemination and sharing of the several datasets produced throughout the analysis, a web application was developed using the Shiny framework in R. This tool provides users with the option to browse, filter and export data, as well as produce visualizations through custom queries.



Four main exploration options are provided, based on the type of data being dealt with:

- **Explore Data**
- **Text Analysis**
- **Pairwise Analysis**
- **Association Rules**

Additionally, the **Codebook** option offers the ability to export the raw datasets, as well as read brief documentation related to the information encoded in them.

Explore Data

The main datasets produced can be explored on “**Explore Data**”. Specifically, datasets available belong in six broad categories:

Category	Description	Value type
Survey	Produced in the context of the Europass Survey in the pilot phase of the analysis.	Aggregated/count
EQF Awareness	Produced for the “EQF Awareness” report in the pilot phase of the analysis.	Aggregated/count
EWA DB	Produced after analysis of data in the Europass backup database.	Aggregated/count
Occupations Analysis	Regression analysis of occupation data in the Europass backup database.	Indicator-like
Occupations Analysis (Weighted)	Regression analysis of occupation data in the Europass backup database, post-weighting.	Indicator-like
Skills Analysis (v1.0.8)	Produced after analysis of Europass Survey data, through utilization of the ESCO classification v1.0.8.	Both

Data exploration is performed in three main steps:

1. Prepare data ▼

Select data set:*

Demographics (Backup Database) ▼

Select variables:*

Narrow Age Group × Country × Latest Job ISCO 2 ×

EU countries only

Display missing values

Apply
× Reset variables

1. Prepare data:

a) A dataset must initially be initially selected. A short loading process will proceed, as the dataset’s metadata are gathered. **Note:** *In cases of error caused by heavy load on the application, clicking “Reset variables” will reinitialize the dataset.*

b) The user has to select the variables applicable to the query they want to make. Data will be aggregated by these variables, meaning that the less variables selected, the less rows (with higher *Count* values) will result. **Note:** *For datasets with indicator-like values, all variables must be selected.*

c) Checkboxes should be adjusted based on whether data from countries outside the EU should be included, and whether rows with NA values (e.g., age group not included on CV, or latest job unknown) should also be presented.

d) Clicking the **Apply** button will make the requested dataset appear on the **Table** tab. Server-side pagination is provided, along with sorting, text search, and export options for the presented results.

Table
Plot

Copy
Excel
PDF
Print

Show 20 entries

Search:

Narrow Age Group	Country	Latest Job ISCO 2	Count
21-25	Italy	Personal service workers	33799
36+	Italy	Business and administration associate professionals	25161
36+	Italy	Personal service workers	24284
26-30	Italy	Personal service workers	21401
Up to 20	Italy	Personal service workers	20329
36+	Portugal	Business and administration associate professionals	19102
21-25	Italy	Business and administration associate professionals	19099
36+	Italy	Legal, social, cultural and related associate professionals	18556

75 | Page

2. Filter data ▼

Narrow Age Group

Country

Latest Job ISCO 2

[✕ Reset filters](#)

[Filter](#)
[Download](#)

2. Filter data:

- a) The prepared data can optionally be filtered based on the variables selected. Depending on the application’s load, a load process of metadata that takes to 30 seconds may precede ability to filter data. *Note: In case of error caused by heavy load on the application, in the case of very in-depth queries (e.g., when there are over 100,000 rows in the prepared data), clicking the “Reset filters” or refreshing the browser’s tab will reinitialize prepared data.*
- b) Clicking the **Filter** button will make the requested filtered data appear on the **Table** tab.
- c) Clicking the **Download** button will provide a CSV file of the requested filtered data, which can be imported on third-party statistical software.

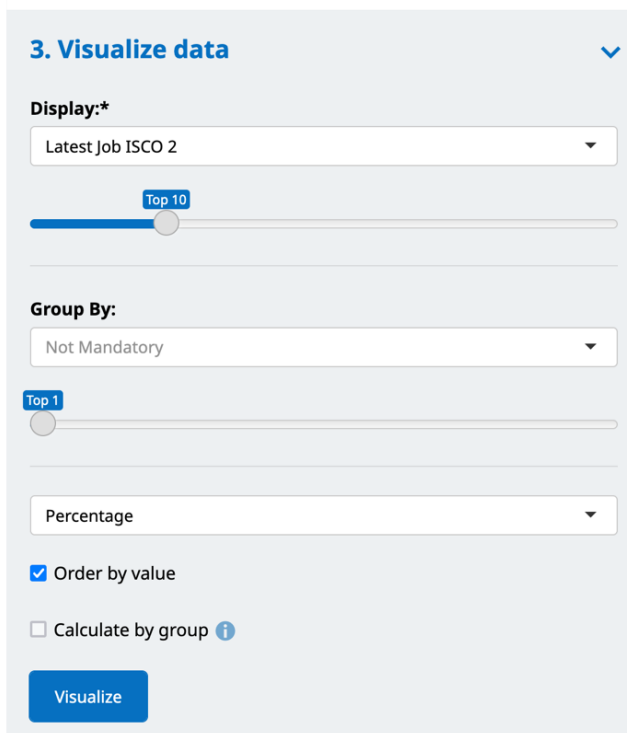
Table
Plot

Copy
Excel
PDF
Print

Show **20** entries

[Search:](#)

Narrow Age Group	Country	Latest Job ISCO 2	Count
26-30	Portugal	Business and administration associate professionals	10222
26-30	Portugal	Personal service workers	9392
26-30	Portugal	Health professionals	6538
26-30	Portugal	Hospitality, retail and other services managers	6282
26-30	Portugal	Science and engineering professionals	5670
26-30	Portugal	Teaching professionals	5469
26-30	Portugal	Business and administration professionals	5271



3. Visualize data:

a) Data can be aggregated by one (**Display**) or two (**Display** and **Group By**) variables and visualized through custom visualizations. *Note: As with step 1b, indicator-like datasets should not be aggregated. To proceed with visualization of indicator-like datasets, variables not selected as Display and Group By must be filtered on step 2.*

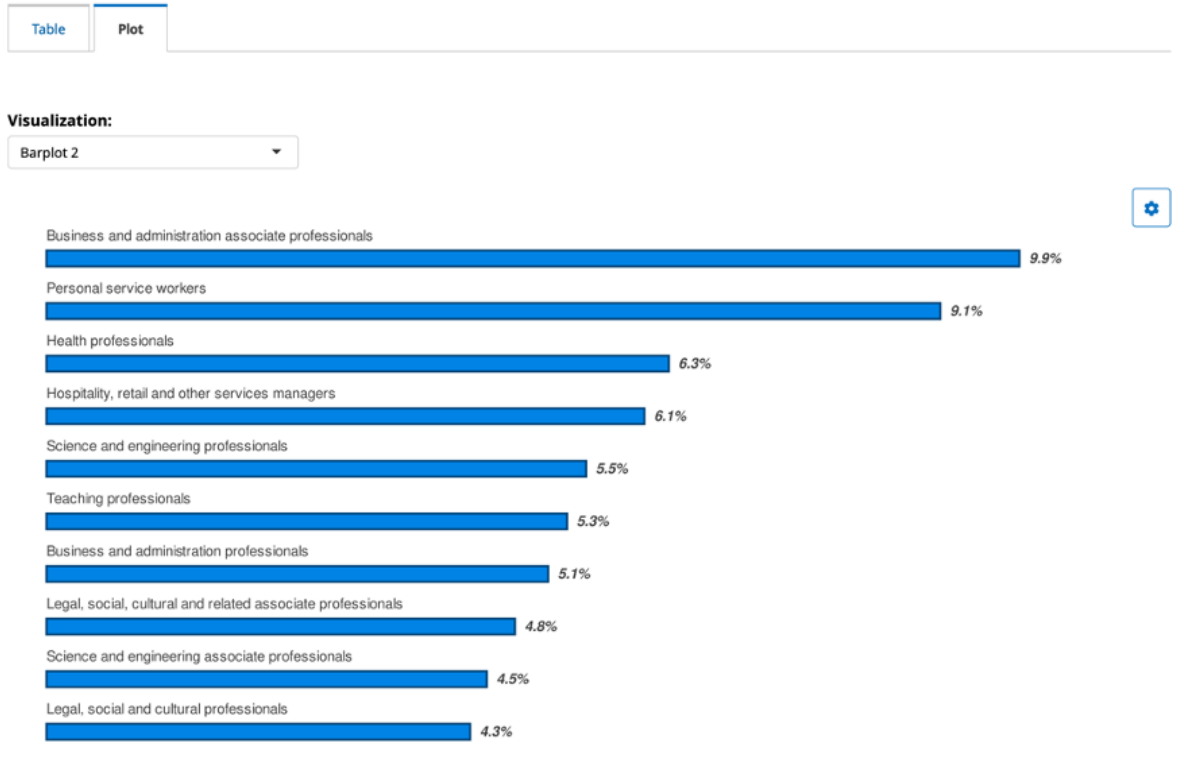
b) Selection of a **Display** variable is mandatory. This is the main variable whose values will be presented on the visualization. The number of values presented is based on user selection (e.g., top 10 values by count – filtering can accommodate visualization of non-top values if the research question requires a more specific subset of the data). If only **Display** is selected, basic visualizations such as *Barplot* and *Donut* will be enabled.

c) A second **Group By** variable can be selected optionally. If **Group By** is selected, more advanced, two-dimensional visualizations will be enabled. The values of the **Group By** variable serve as the grouping values of the visualizations (e.g., facets, or colours explained on the legend). Once again, the number of groups is based on user selection. *Note: If the Display variable is a temporal variable (e.g., recruitment year) and a Group By variable is selected, the Time series visualization is enabled.*

d) Using the drop-down box, the user can select to present visualizations either with respect to raw count data, as a percentage, or on a normalized scale. Through the checkboxes, they can optionally be ordered by the values of the Display variable. If a **Group By** variable is provided, the query parameters may also be calculated by group.

e) Clicking the **Visualize** button will make the visualization appear on the **Plot** tab.

f) Visualization type can be selected from the **Visualization** drop-down in the **Plot** tab. Clicking on the top right cog button also allows users to export the visualization.



Text Analysis

Text-based datasets can be explored on “Text Analysis”. Note that due to the fact that most free-text is not preserved in the Europass backup database, datasets available there derive from the Europass Survey data.

Data exploration is once again performed in three broad steps which are similar to the ones in **Explore Data**.

1. Select free text ▼

Select free text data set:*
Skills by Occupations ▼

Select language:*
English ▼

Select keyword or phrase:*
Phrase ▼

Select variable:
Job ISCO 2 ×

Keep missing values

Apply ✕ Reset variables

1. Select free text: A free text dataset is selected. One language can be explored at a time, and either specific words or two-word phrases can be explored. Variables can optionally be selected in order to drill-down on specific queries.

2. Filter free text ▼

Term

Job ISCO 2

Health professionals × Science and engineering professionals ×

[× Reset filters](#)

[Filter](#) [Download](#)

2. Filter text data: It is once again optionally possible to filter any one of the variables selected. If terms related to a specific job are being explored, the user can for example filter the job variable (e.g., Health professionals). If specific terms (e.g., terms related to entrepreneurship, as in EntreComp) are being explored, they can be selected in the term variable field.

3. Visualize free text ▼

Display:*

Term ▼

Top 8

Group By:

Job ISCO 2 ▼

Top 2

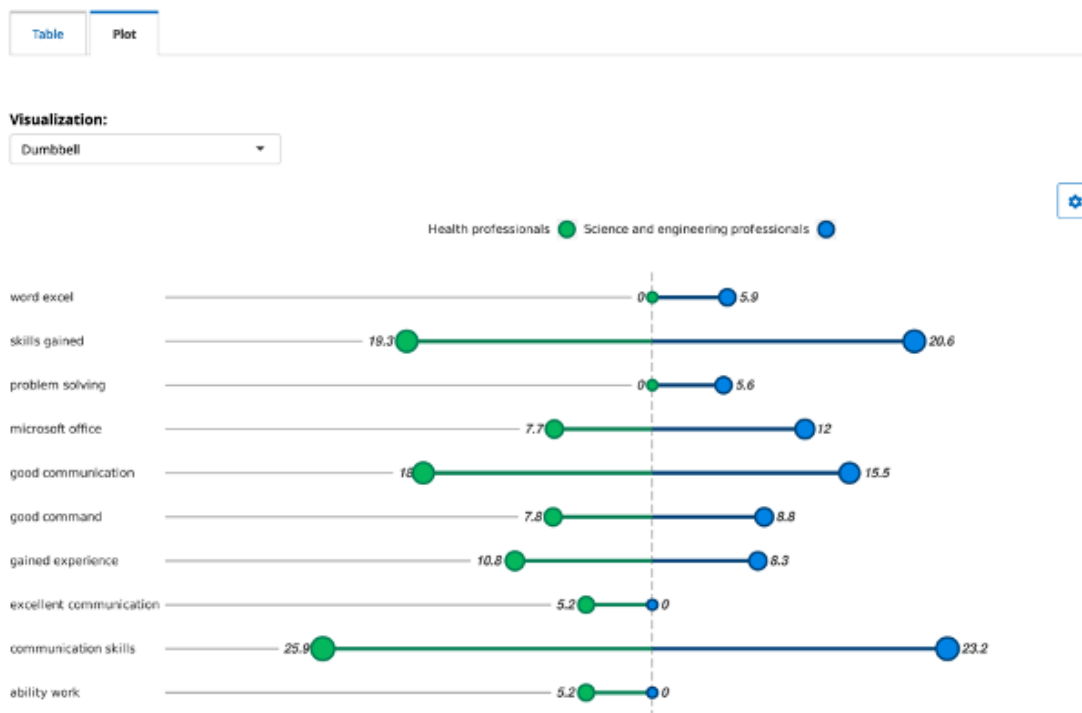
Normalize to 100 ▼

Order by value

Calculate by group ⓘ

[Visualize](#)

3. Visualize free text: Visualization of data is the same as in Explore Data. *Note: When comparing text inclusion between different subgroups (e.g., different ISCO jobs), normalizing to 100 and calculating by group can be especially useful.*



Pairwise Analysis

Graph-like data are exposed by encoding transitions from a state “A” to a state “B”. “**Pairwise Analysis**” facilitates exploration of this data through interactive network visualisations generated through custom queries. For example, in career paths, transitions between jobs can be visualised and clustered. The following datasets are provided:

- **Career paths** and **Academic paths** generated from the Europass Survey analysis.
- **Career paths** generated from analysis of the Europass backup database.
- **Skills/Competences** free-text data generated from the Europass Surevy analysis.
- The **Skillscape** dataset presented earlier, offered in higher resolution. *Note: Unlike the aforementioned datasets, which include aggregated count values, Skillscape includes indicator-like values and is subject to the restrictions documented in Explore Data. When selecting the Skillscape dataset, all variables must be included in Step 1 and all variables (except for the From Skill and To Skill) must be filtered in Step 2.*

As with the previous tools, visualizations can be produced in three main steps:

1. Select pairwise data ▼

Select data set:*
Career (Backup Database) ▼

Select from:*
From Job ISCO 2 ×

Select to:*
To Job ISCO 2 ×

Select variables:
Recruitment Year × Broad Age Group ×

Display missing values

[Apply](#) [× Reset variables](#)

1. Select pairwise data: The variables encoding the state transitions are selected in “Select from” and “Select to”. With the exception of the Skillscape dataset, variables in “Select variables” are optional.

2. Filter pairwise data ▼

From Job ISCO 2
All

To Job ISCO 2
All

Recruitment Year
2015 × 2016 × 2017 × 2018 × 2019 × 2020 ×

Broad Age Group
25-49 ×

[× Reset filters](#)

[Filter](#) [Download](#)

2. Filter pairwise data: If the research task requires a specific query, the selected variables can be filtered.

3. Visualize pairwise data

From:*
From Job ISCO 2

Top 15

To:*
To Job ISCO 2

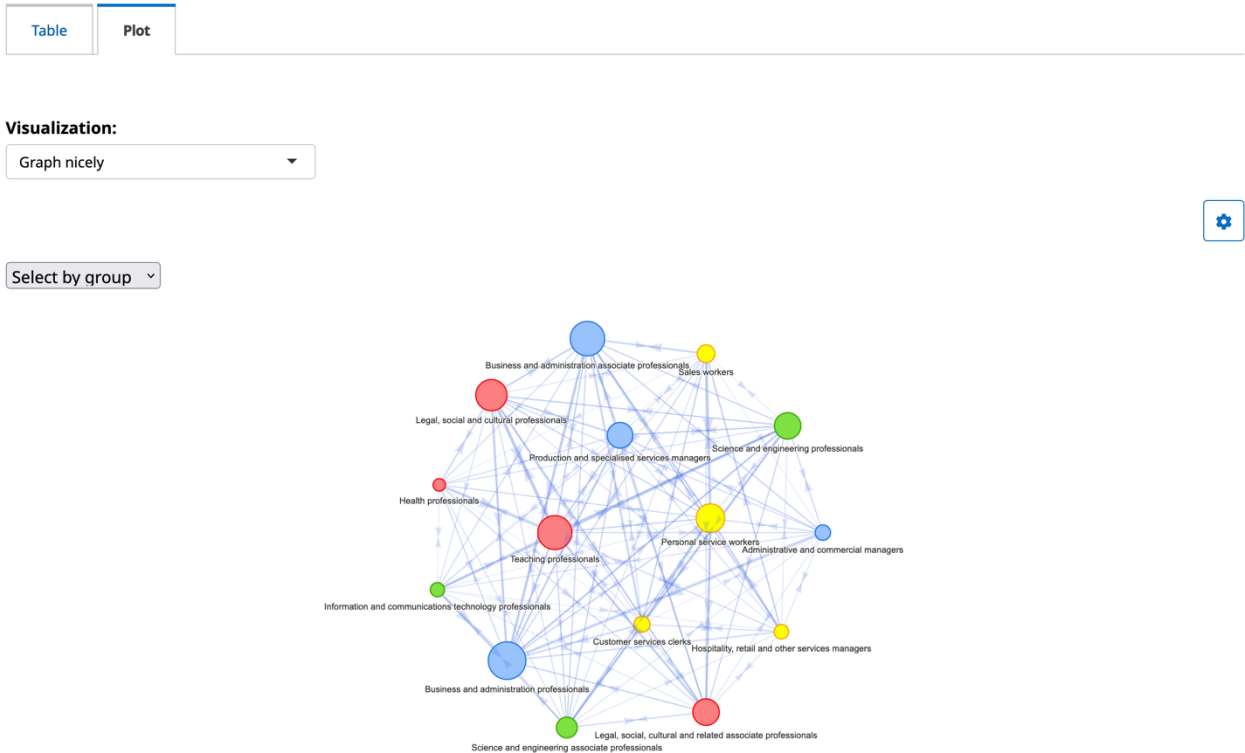
Top 15

Allow transitions to same level

Clustering *i*

Visualize

3. Visualize pairwise data: User may choose to allow transitions from one state to the same state (e.g., in career paths a person moves from a job classified in Professionals to a new job also classified in Professionals). Additionally, it is possible to enable **Clustering** of the results based on transitions.



Codebook

Data can be downloaded at any step of the exploration using the **Download** button in 2. Filter data, or using one of the options in the **Table** tab.



In addition to that, more information about each dataset and the ability to export the raw datasets is present in the **Codebook** tab. Codebooks provide brief documentation and description of each variable.

Raw data

These datasets provide anonymized aggregated data as they derived from the Europass Survey data analysis. Values are kept as they were recorded in the Europass XML schema (*example*). Free-text entries are mapped to ESCO and ISCO classification for occupations, skills and qualifications using text mining and machine learning.

Skills data (ESCO v1.0.8)

Dataset Overview

- **Name:** *skills_aggregate_v108*
- **Number of rows:** 11,101,502
- **Number of features:** 13
- **Description:** *This dataset describes aggregate skill-related statistics based on responses collected in the Europass Survey. The values of some of its features derive directly from the Europass CV creation form fields, while others are a result of data analysis upon the given responses. Certain features are simple statistics, while others are mappings of free text to standardized taxonomies made using fuzzy matching techniques. (ESCO v1.0.8)*
- **Sources:** *Europass*
- [Download](#)

Feature	Name	Description	Form ¹	Stat ²	Text ³
skill_hierarchy	ESCO Skill Hierarchy	Level on the ESCO Skill hierarchy.	•		📘
skill_type	Skill Category	Broad category of a classified skill response.	•		📘
skill_value	Skill Title	ESCO classification for skill free-text or drop down response.	•	•	📘
locale	CV Language	Language used to write CV.	•		📘
birth_year	Birth Year	Year of birth.	•		📘
gender	Gender	Female, male or missing value.	•		📘
country	Country	Country of residence.	•		📘
is_student	Student Status	Estimation of student status based on enrollment and graduation year.		•	📘

Bibliography

- Pouliakas, K. (2021). Artificial intelligence and job automation: an EU analysis using online job vacancy data. *Publications Office of the European Union*(Cedefop working paper; No 6), 10.1177/14614456020040040901. 3-4.
- Potter, J. (2002). Two kinds of natural. *Discourse Studies*(4(4)), 539–542. doi:10.1177/14614456020040040901.
- Alhubaiti, A. (2016). Information bias in health research: definition, pitfalls, and adjustment methods. *Journal of Multidisciplinary Healthcare*, 213. doi:10.2147/jmdh.s104807 .
- Barnichon, R. (2010). Building a composite Help-Wanted Index. *Economics Letters*(109(3)), 175–178. doi:10.1016/j.econlet.2010.08.029.
- Eurostat. (2020). EU - Labour Force Survey microdata 1983-2019, release 2020, version 1 (Version 1) [Data set]. *Eurostat*, 10.2907/LFS1983-2020V.1 .
- Sharpe, D. (2015). Chi-Square Test is Statistically Significant: Now What? *Practical Assessment, Research, and Evaluation*, 20(8), 2.
- Daas, P. J. (2015). Big Data as a Source for Official Statistics. *Journal of Official Statistics*, 31(2), 249–262. 10.1515/jos-2015-0016.
- Kitchin, R. (2015). Big Data and Official Statistics: Opportunities, Challenges and Risks. *SSRN Electronic Journal*, 10.2139/ssrn.2595075.
- Xie, Y. (2015). *Dynamic Documents with R and Knitr, Second Edition (Chapman & Hall/CRC the R Series)*. Routledge.
- Boettiger, C. (2015). An introduction to Docker for reproducible research. *ACM SIGOPS Operating Systems Review*, 49(1), 71–79. <https://doi.org/10.1145/2723872.2723882>.
- Wickham, H. (2014). Tidy Data. *Journal of Statistical Software*, 59(10), DOI: 10.18637/jss.v059.i10.
- Wilkinson, L. (2010). The grammar of graphics. *WIREs Computational Statistics*, 2(6), 673–677. doi: 10.1002/wics.118.
- Cedefop. (2019). Online job vacancies and skills analysis: a Cedefop pan European approach. *Publications Office*, DOI: 10.2801/097022 .
- Mang, C. (2012). Online job search and matching quality. *Ifo Working Paper*(147).
- Publications Office. (n.d.). *Interinstitutional Style Guide – 7.1. Countries – 7.1.1. Designations and abbreviations to use*. Retrieved from Publications Office: <http://publications.europa.eu/code/en/en-370100.htm>
- Chambers, J. M. (1992). Linear models. Chapter 4. In J. M. Chambers, & T. Hastie, *Statistical Models in S*.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). New York: Springer.
- Chan, B. K. (2018). Data analysis using R programming. In *Biostatistics for Human Genetic Epidemiology* (pp. 47-122). Springer.
- Piatetsky-Shapiro, G. (1991). Discovery, analysis, and presentation of strong rules. In *Knowledge discovery in databases* (pp. 229-238).

- Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *1993 ACM SIGMOD international conference on Management of data*, (pp. 207-216).
- Chapman, C. (2015). R for marketing research and analytics. In C. Chapman, & E. M. Feit. New York, NY: Springer.
- Edelman, B. (2012). Using Internet Data for Economic Research. *The Journal of Economic Perspectives*, 26, DOI: 10.2307/41495310.
- Sapleton, N. (2013). *Advancing Research Methods with New Technologies*, 1st edn. IGI Global.
- Taylor, L. S. (2014). Emerging practices and perspectives on Big Data analysis in economics: Bigger and better or more of the same? *Big Data & Society*, 1(2), 205395171453687. DOI: 10.1177/2053951714536877.
- List, J. A. (2007). Field experiments: a bridge between lab and naturally occurring data. *The BE Journal of Economic Analysis & Policy*, 5(2).
- Kureková, L. M., Beblavý, M., & Thum-Thysen, A. (2015). Using online vacancies and web surveys to analyse the labour market: a methodological inquiry. *IZA Journal of Labor Economics*, 4(1), 1-20.
- Askitas, N., & Zimmermann, K. (2015). The internet as a data source for advancement in social sciences. *Int J of Manpow*, 36(1), 2-12.
- Alabdulkareem, A., Frank, M. R., Sun, L., AlShebli, B., Hidalgo, C., & Rahwan, I. (2018). Unpacking the polarization of workplace skills. *Science Advances*, 4(7), DOI: 10.1126/sciadv.aao6030.

==== End of Report====