
Efficiency and Diversity of R&D in Knowledge-Intensive Services (2005–2023)

How **efficiently** selected European countries convert R&D spending into researcher human capital within the knowledge-intensive services sector?

Does increasing **female participation** in researcher roles correlates with chosen metrics of development?



git



GitHub

BY DOMINIKA DRAŻYK
OCTOBER, 2025

Introduction

The **knowledge-intensive services sector** ('G-N' by NACE) has become a major engine of innovation in Europe.

About companies embracing gender diversity within the R&D:

- (...) more innovative, better at solving complex problems (*Phillips, 2014*);
- (...) tend to achieve higher productivity and innovation performance (*Hoogendoorn et al., 2019*);
- (...) consistently outperform rivals (*Carucci, 2024*).

PRACTICAL BUSINESS QUESTION:

Does **investing in a more gender-diverse R&D workforce** help or hinder the efficient use of R&D budgets in G-N sector?

The **knowledge-intensive services sector** ('G-N' by NACE) has become a major engine of innovation in Europe.

About companies embracing gender diversity within the R&D:

- (...) more innovative, better at solving complex problems (*Phillips, 2014*);
- (...) tend to achieve higher productivity and innovation performance (*Hoogendoorn et al., 2019*);
- (...) consistently outperform rivals (*Carucci, 2024*).

PRACTICAL BUSINESS QUESTION:

Does **investing** in a more gender-diverse R&D workforce help or hinder the efficient use of R&D budgets in G-N sector?

Business enterprise R&D expenditure in high-tech sectors by NACE Rev. 2 **{htec_sti_exp2}**

Business enterprise R&D personnel in high-tech sectors by NACE Rev. 2 **{htec_sti_pers2}**

R&D personnel and researchers in business enterprise sector by NACE Rev. 2 activity and sex **{rd_p_bempoccr2}**



publicly available

regularly updated

structured, but require pre-processing

Approach

O1: Automated web-scraping of datasets and metadata

O2: Cleaning and reshaping datasets

O3: Merging datasets

Python {BeautifulSoup, selenium, pandas, pyjstat}

Jupyter Notebook as environment

Git & GitHub for version control

Example – web scraping

```
print("• Extracting expenditure dataset metadata")
url_exp2_meta = 'https://ec.europa.eu/eurostat/databrowser/view/htec_sti
print(f"  Source: {url_exp2_meta}")

chrome_options = Options()
driver_exp2 = webdriver.Chrome(options = chrome_options)
driver_exp2.get(url_exp2_meta)
print("  • Webpage opened")
time.sleep(20)
r = driver_exp2.page_source
print("  • Page source extracted")
soup_exp2 = bs(r, "html.parser")
driver_exp2.close()
print("  • Browser closed")

print("  • Parsing metadata fields")
body = soup_exp2.find('body')
marker = body.find('span', string = "last update")
tag = marker.find_next("b", class_ = "infobox-text-data")
exp2_date = tag.get_text(strip = True)
print(f"    - Last updated: {exp2_date}")
```

SOLUTION

*I use selenium library to automatically open and scrap a website.
Then I use BeautifulSoup library to access its HTML code
and extract metadata of the source.*

O1: Automated web-scraping of datasets and metadata

O2: Cleaning and reshaping datasets

O1: Merging datasets

Example – reshaping dataset

```
print("• Transforming personnel data to wide format")
data_pers2_wide = data_pers2.pivot(index = ['nace_r2', 'geo', 'time', 'prof_pos'],
                                   columns = ['unit'],
                                   values = 'value').reset_index()
data_pers2_wide = data_pers2_wide.pivot(index = ['nace_r2', 'geo', 'time'],
                                       columns = ['prof_pos'],
                                       values = ['FTE', 'HC']).reset_index()
data_pers2_wide.columns = ["_".join([str(c) for c in col if c != ""])
                           for col in data_pers2_wide.columns.to_flat_index()]
data_pers2_wide = data_pers2_wide.rename(columns=lambda x: f"pers2_{x}"
                                       if x not in ['nace_r2', 'geo', 'time'] else x)
print(f" ✓ Personnel data: {data_pers2_wide.shape[0]:,} rows
      × {data_pers2_wide.shape[1]} columns")
print(f" Sample: {data_pers2_wide.shape}")
```

SOLUTION

I use pandas library to pivot tables into preferred wide format, rename newly created columns and display the characteristics of the transformed DataFrame.

O1: Automated web-scraping of datasets and metadata

O2: Cleaning and reshaping datasets

O1: Merging datasets

Intermediate results



eu_efta_countries.csv

a list of EU and EFTA countries for filtering data;



scraper_data.csv

*the main dataset merged from the external data sources
{htec_sti_exp2, htec_sti_pers2, rd_p_bempoccr2};*



scraper_metadata.csv

*the metadata from external data sources including
dataset IDs, source, title and the date of their last update;*



scraper_code.ipynb / scraper_code.py

*the Python codes that can be re-used in the future
in case of an Eurostat database update.*

Further Approach

I have generated three new data sources and an reusable code:



eu_efta_countries.csv

a list of EU and EFTA countries for filtering the main dataset;



scraper_data.csv

the main dataset merged from the three external data sources {htec_sti_exp2, htec_sti_pers2, rd_p_bempoccr2};



scraper_metadata.csv

the metadata about three external data sources including dataset IDs, source, title and the date of their last update;



scraper_code.ipynb / scraper_code.py

the Python codes that can be used in the future to re-scrap the data in case of an Eurostat database update.

O1: Dataset preprocessing

O1.1 Loading datasets and metadata

O1.2 Filtering and renaming dataset variables

O1.3 Missing data review and country selection

O2: Analysing chosen efficiency metrics

O2.1 Calculating Annual Spending Efficiency across countries

O2.2 Calculating Annual Labor Intensity across countries

O3: Analysing participation of female researchers

O3.1 Calculating Share of Female Researchers in all Researchers

O3.2 Spending Efficiency vs. Female Researchers Share

O3.3 Calculating Growth Rates of chosen metrics

Python {pandas, NumPy, scipy, matplotlib, seaborn}

Jupyter Notebook as environment

Git & GitHub for version control

Example – filtering dataset

```
def filter_countries(row):
    global set_of_countries
    if row['geo'] in set_of_countries:
        return 'in'
    else:
        return 'out'

print("• Filtering by EU + EFTA countries")
global set_of_countries
set_of_countries = euefta['geo'].values
df['geo_euefta'] = df.apply(filter_countries, axis = 1)
df = df[df['geo_euefta'] == 'in']
print(f"• Countries included: {len(df.geo.unique())} EU + EFTA countries")
df = pd.merge(df, euefta, on = ['geo'], how = 'left')
```

SOLUTION

As the analysis focuses only on EU and EFTA countries, I wrote a simple function that effectively allows me to filter the main dataset by a list of EU and EFTA countries, previously scrapped from the Eurostat website.

Further Approach

O1: Dataset preprocessing

O1.1 Loading datasets and metadata

O1.2 Filtering and renaming dataset variables

O1.3 Missing data review and country selection

O2: Analysing chosen efficiency metrics

O2.1 Calculating Annual Spending Efficiency across countries

O2.2 Calculating Annual Labor Intensity across countries

O3: Analysing participation of female researchers

O3.1 Calculating Share of Female Researchers in all Researchers

O3.2 Spending Efficiency vs. Female Researchers Share

O3.3 Calculating Growth Rates of chosen metrics

Missing data review

Key variables related to knowledge-intensive services sector

GDE Euro: R&D expenditure (millions €)

FTE All: number of full-time positions

FTE Researcher: number of full-time researcher positions

FTE Researcher Fem: number of full-time researcher positions held by women

Missing Data by Column:

- GDE Euro: 26% missing
- FTE All: 27% missing
- FTE Researcher: 29% missing
- FTE Researcher Fem: 35% missing

INSIGHT

*The number of female researcher full-time equivalents variable (**FTE Researcher Fem**) includes significant proportion of missing data (~**35%**) compared to the rest of chosen metrics.*

Further Approach

O1: Dataset preprocessing

O1.1 Loading datasets and metadata

O1.2 Filtering and renaming dataset variables

O1.3 Missing data review and country selection

O2: Analysing chosen efficiency metrics

O2.1 Calculating Annual Spending Efficiency across countries

O2.2 Calculating Annual Labor Intensity across countries

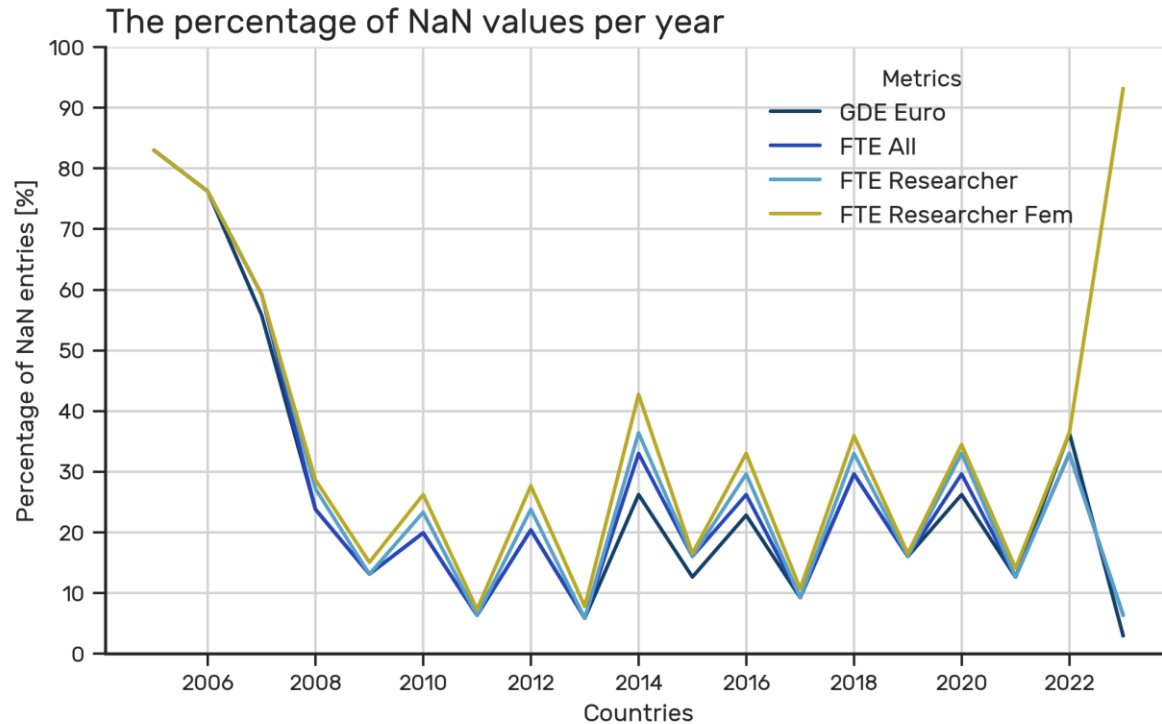
O3: Analysing participation of female researchers

O3.1 Calculating Share of Female Researchers in all Researchers

O3.2 Spending Efficiency vs. Female Researchers Share

O3.3 Calculating Growth Rates of chosen metrics

Missing data review



INSIGHT

Data entry issues:

- pre-2008 period with missing data across all metrics;
- post-2022 period with gaps in FTE Researcher Fem.

These patterns inform the selection of the **2009-2021** analysis period for optimal data coverage.

Further Approach

O1: Dataset preprocessing

O1.1 Loading datasets and metadata

O1.2 Filtering and renaming dataset variables

O1.3 Missing data review and country selection

O2: Analysing chosen efficiency metrics

O2.1 Calculating Annual Spending Efficiency across countries

O2.2 Calculating Annual Labor Intensity across countries

O3: Analysing participation of female researchers

O3.1 Calculating Share of Female Researchers in all Researchers

O3.2 Spending Efficiency vs. Female Researchers Share

O3.3 Calculating Growth Rates of chosen metrics

Missing data review

```
df_nans = df.groupby(['Country', 'geo'])[['GDE Euro', 'FTE All', 'FTE Researcher', 'FTE Researcher Fem']].apply(
    lambda x: (x.isna().sum() * 100 / len(x)).sort_values(by = 'Country')
)
df_nans = df_nans[(df_nans['FTE Researcher'] <= 20)
    & (df_nans['FTE Researcher Fem'] <= 20)
    & (df_nans['FTE All'] <= 20)
    & (df_nans['GDE Euro'] <= 20)]
df_nans.reset_index(inplace = True)
```

	Country	geo	GDE Euro	FTE All	FTE Researcher	FTE Researcher Fem
0	Bulgaria	BG	5.263	5.263	5.263	10.526
1	Croatia	HR	0.000	0.000	0.000	5.263
2	Czechia	CZ	0.000	0.000	0.000	0.000
3	Estonia	EE	15.789	10.526	10.526	15.789
4	Hungary	HU	0.000	0.000	0.000	5.263
5	Italy	IT	10.526	10.526	10.526	15.789
6	Poland	PL	10.526	5.263	5.263	8.271
7	Portugal	PT	10.526	10.526	10.526	10.526
8	Slovakia	SK	0.000	0.000	0.000	5.263

INSIGHT

*I used pandas library to filter countries that demonstrate **data completeness rates of at least 80%** across all analytical metrics.*

*Only those **nine countries** listed above fulfill that requirement and those will be chosen for further analysis.*

Further Approach

O1: Dataset preprocessing

O1.1 Loading datasets and metadata

O1.2 Filtering and renaming dataset variables

O1.3 Missing data review and country selection

O2: Analysing chosen efficiency metrics

O2.1 Calculating Annual Spending Efficiency across countries

O2.2 Calculating Annual Labor Intensity across countries

O3: Analysing participation of female researchers

O3.1 Calculating Share of Female Researchers in all Researchers

O3.2 Spending Efficiency vs. Female Researchers Share

O3.3 Calculating Growth Rates of chosen metrics

Example - metrics calculation

```
print("• Calculating Annual Spending Efficiency")
efficiency_calc = df.groupby(['Country', 'Year'])[['GDE Euro', 'FTE Researcher']].apply(
    lambda x: pd.Series({
        'SpendEff': x['GDE Euro'].sum() / x['FTE Researcher'].sum()
        if not (x['GDE Euro'].isna().any() or x['FTE Researcher'].isna().any())
        else float('nan')})).reset_index()
df = df.merge(efficiency_calc, on=['Country', 'Year'], how = 'left')
nan_pct = df['SpendEff'].isnull().sum()*100/len(df['SpendEff'])
print(f"• SpendEff: {nan_pct:.0f}% values converted to NaN")
```

Calculating Annual Spending Efficiency
SpendEff: 6% values converted to NaN

```
g = sns.relplot(kind = 'line', x = 'Year', y = 'SpendEff', data = df, hue = 'Country', errorbar = None)
sns.move_legend(g, "upper right", bbox_to_anchor = (0.975, 0.95), ncol = 1)
g.set(title = 'Annual Spending Efficiency per a Researcher (FTE)',
      xlabel = "Calendar year", ylabel = "Spending Efficiency [MIO € / 1 Researcher FTE]")
g.fig.set_size_inches(10,4)
```

SOLUTION

I used pandas library to calculate annual spending efficiency, defined as R&D expenditure per researcher full-time equivalent.

Then I visualised the calculated metrics across years and chosen set of countries using matplotlib and seaborn libraries.

Further Approach

O1: Dataset preprocessing

O1.1 Loading datasets and metadata

O1.2 Filtering and renaming dataset variables

O1.3 Missing data review and country selection

O2: Analysing chosen efficiency metrics

O2.1 Calculating Annual Spending Efficiency across countries

O2.2 Calculating Annual Labor Intensity across countries

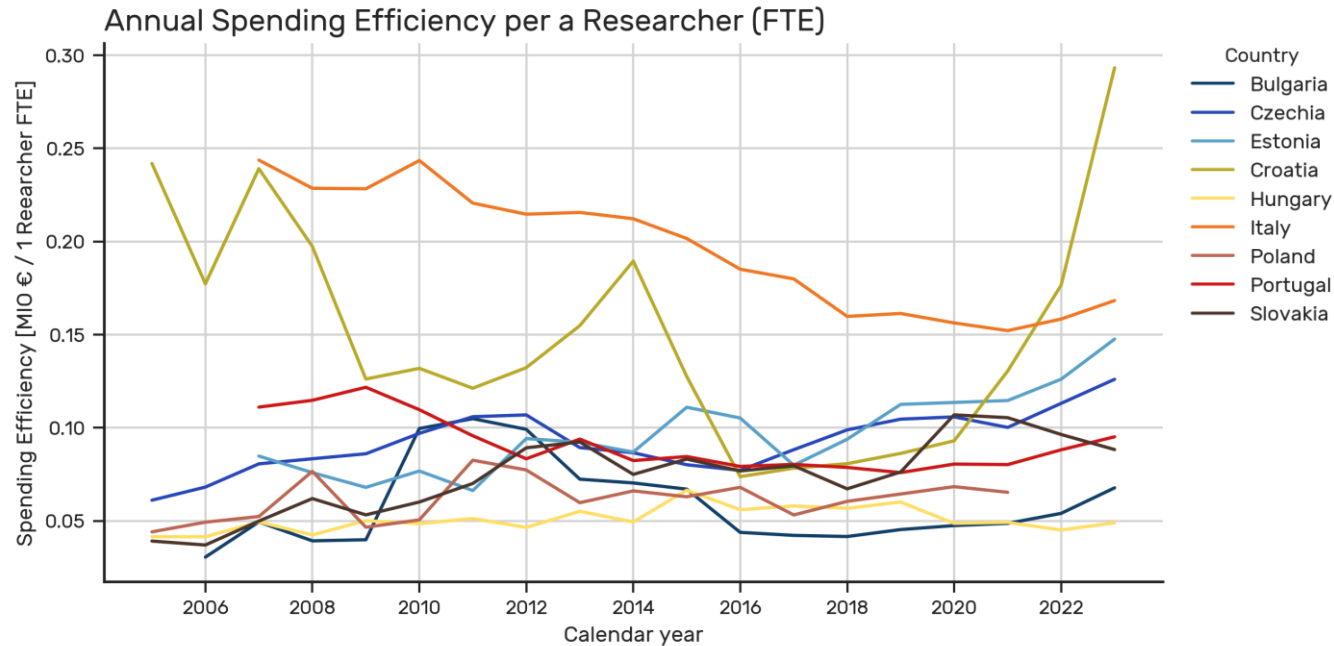
O3: Analysing participation of female researchers

O3.1 Calculating Share of Female Researchers in all Researchers

O3.2 Spending Efficiency vs. Female Researchers Share

O3.3 Calculating Growth Rates of chosen metrics

Visualization



INSIGHT

A rising curve present for most countries, except Portugal and Italy, showing a decrease. This could be due to austerity periods following the 2008 crisis (Wearden, 2011).

A decrease could indicate labor scaling, budget cuts, or both.

Croatia presented a volatile trend (probably due its late EU accession).

Poland presents a generally low but stable spending efficiency.

Further Approach

O1: Dataset preprocessing

O1.1 Loading datasets and metadata

O1.2 Filtering and renaming dataset variables

O1.3 Missing data review and country selection

O2: Analysing chosen efficiency metrics

O2.1 Calculating Annual Spending Efficiency across countries

O2.2 Calculating Annual Labor Intensity across countries

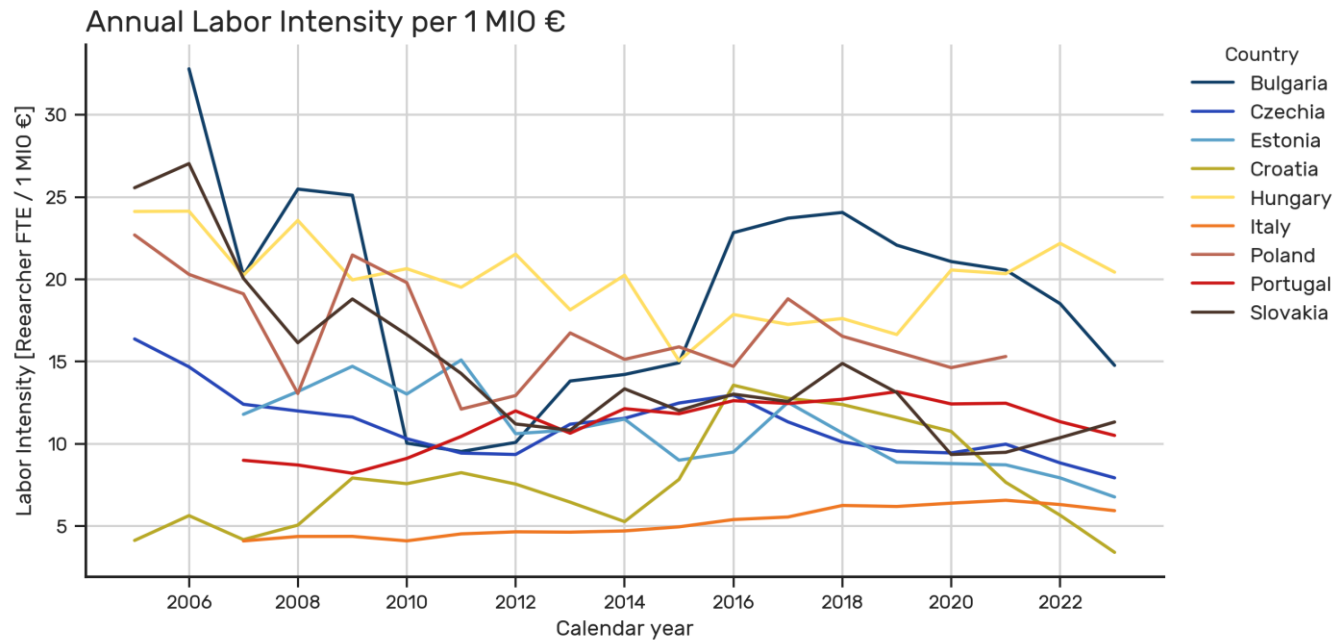
O3: Analysing participation of female researchers

O3.1 Calculating Share of Female Researchers in all Researchers

O3.2 Spending Efficiency vs. Female Researchers Share

O3.3 Calculating Growth Rates of chosen metrics

Visualization



INSIGHT

Labor Intensity: researcher FTEs per unit of expenditure.

A rise was observed only for Italy and Portugal, which could partially explain their decrease in spending efficiency. All remaining countries increased spending faster than the number of researcher FTEs.

Poland presented a generally high labor intensity, with a volatile pattern between 2007 and 2011, that could be caused by its late EU accession.

Further Approach

O1: Dataset preprocessing

O1.1 Loading datasets and metadata

O1.2 Filtering and renaming dataset variables

O1.3 Missing data review and country selection

O2: Analysing chosen efficiency metrics

O2.1 Calculating Annual Spending Efficiency across countries

O2.2 Calculating Annual Labor Intensity across countries

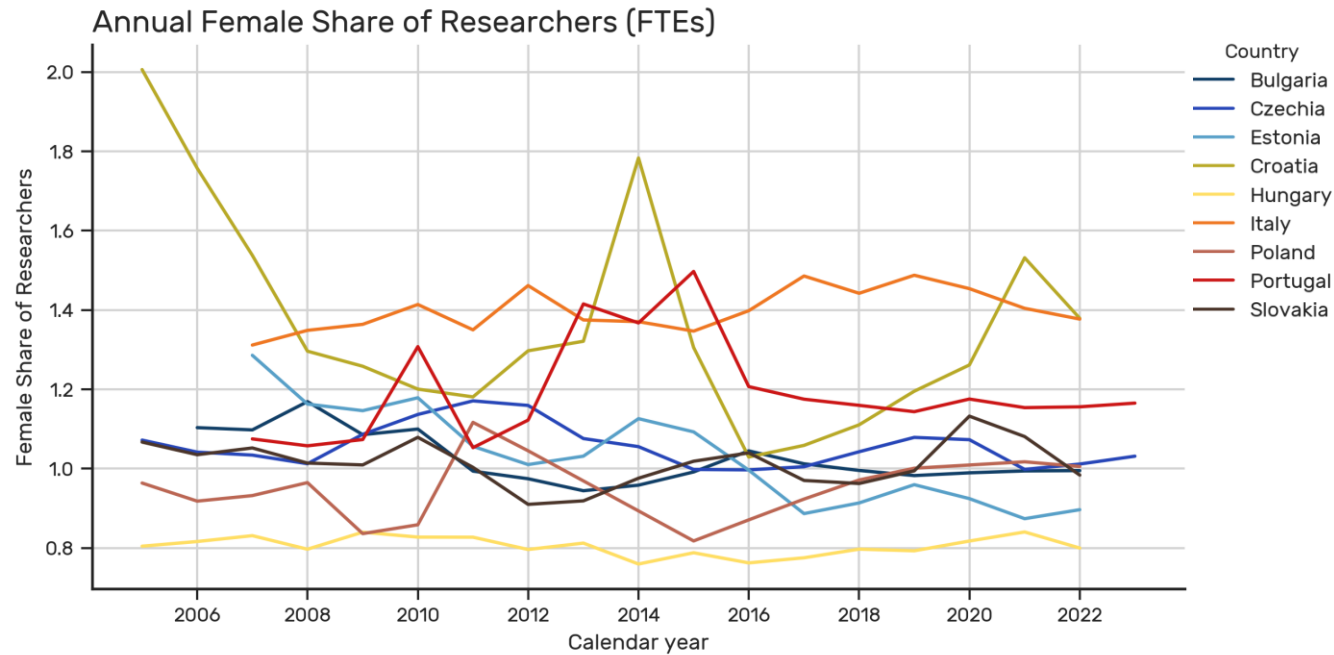
O3: Analysing participation of female researchers

O3.1 Calculating Share of Female Researchers in all Researchers

O3.2 Spending Efficiency vs. Female Researchers Share

O3.3 Calculating Growth Rates of chosen metrics

Visualization



INSIGHT

Female Share: the proportion of female researchers in all researcher FTEs.

A small rise for Poland, Italy and Portugal. Only for Poland was it combined with an increase in spending efficiency and a decrease in labor intensity.

Both Portugal and Croatia recorded a rapid increase in female researcher FTEs between 2013 and 2015 that could be caused by EU-funded gender-inclusion projects (e.g., Horizon 2020).

Further Approach

O1: Dataset preprocessing

O1.1 Loading datasets and metadata

O1.2 Filtering and renaming dataset variables

O1.3 Missing data review and country selection

O2: Analysing chosen efficiency metrics

O2.1 Calculating Annual Spending Efficiency across countries

O2.2 Calculating Annual Labor Intensity across countries

O3: Analysing participation of female researchers

O3.1 Calculating Share of Female Researchers in all Researchers

O3.2 Spending Efficiency vs. Female Researchers Share

O3.3 Calculating Growth Rates of chosen metrics

Example – correlation analysis

```
for c in new_df['Country'].unique():
    x = new_df[new_df['Country'] == c]['FemShare']
    y = new_df[new_df['Country'] == c]['SpendEff']

    sh_x = shapiro(x)
    sh_y = shapiro(y)

    ax = plt.gca()
    ax.legend(title = 'Country', bbox_to_anchor = (1.35, 0.99))
    m, b = np.polyfit(x, y, 1)
    X_plot = np.linspace(ax.get_xlim()[0]+0.05, ax.get_xlim()[1]-0.05, 100)

    if (sh_x[1] <= 0.05) or (sh_y[1] <= 0.05):
        stat, p = sp.stats.spearmanr(a = x, b = y)
        print(f" - {c}: Normality violation - Spearman's r = {stat:.2f}, p = {p:.2E}")
        if p <= 0.05:
            plt.plot(X_plot, m*X_plot + b, '-')
        else:
            plt.plot(X_plot, m*X_plot + b, ':')
    else:
        stat, p = sp.stats.pearsonr(x = x, y = y)
        print(f" - {c}: Normal distribution - Pearson's r = {stat:.2f}, p = {p:.2E}")
        if p <= 0.05:
            plt.plot(X_plot, m*X_plot + b, '-')
        else:
            plt.plot(X_plot, m*X_plot + b, ':')
```

SOLUTION

I used NumPy, scipy, matplotlib and seaborn libraries to perform and visualise the correlation test between two chosen metrics.

Before performing a correlation (Pearson's or Spearman's) on each country, I have checked the test assumptions related to the normality of data distribution (Shapiro-Wilk test).

Further Approach

O1: Dataset preprocessing

O1.1 Loading datasets and metadata

O1.2 Filtering and renaming dataset variables

O1.3 Missing data review and country selection

O2: Analysing chosen efficiency metrics

O2.1 Calculating Annual Spending Efficiency across countries

O2.2 Calculating Annual Labor Intensity across countries

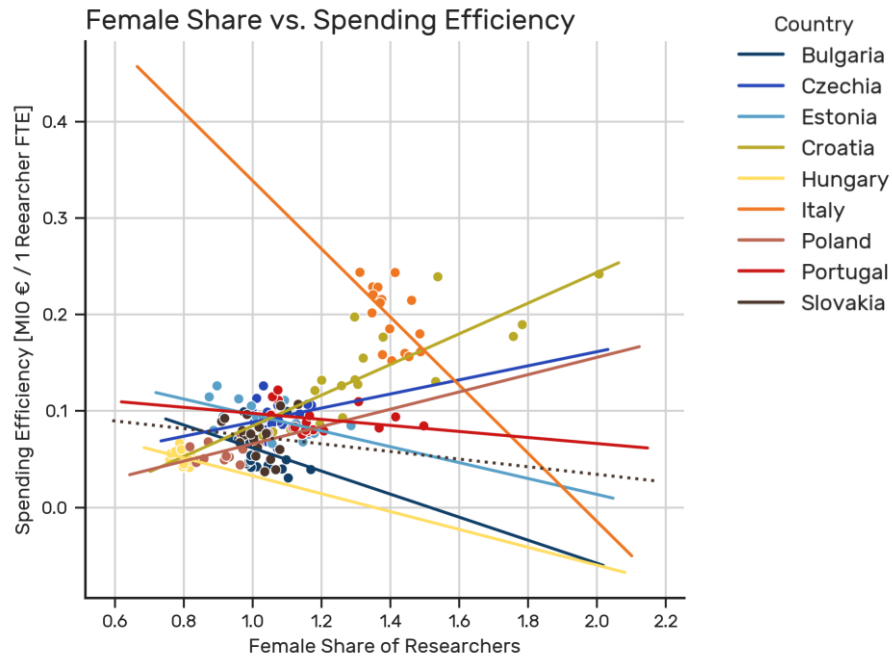
O3: Analysing participation of female researchers

O3.1 Calculating Share of Female Researchers in all Researchers

O3.2 Spending Efficiency vs. Female Researchers Share

O3.3 Calculating Growth Rates of chosen metrics

Visualization



Bulgaria: Normality violation - Spearman's $r = -0.55$, $p = 6.38E-11$
Czechia: Normality violation - Spearman's $r = 0.31$, $p = 2.75E-04$
Estonia: Normality violation - Spearman's $r = -0.62$, $p = 1.03E-12$
Croatia: Normality violation - Spearman's $r = 0.87$, $p = 6.89E-40$
Hungary: Normality violation - Spearman's $r = -0.33$, $p = 1.51E-04$
Italy: Normality violation - Spearman's $r = -0.56$, $p = 8.86E-11$
Poland: Normality violation - Spearman's $r = 0.56$, $p = 7.61E-09$
Portugal: Normality violation - Spearman's $r = -0.33$, $p = 2.97E-04$
Slovakia: Normality violation - Spearman's $r = -0.16$, $p = 7.36E-02$

INSIGHT

A positive correlation: Croatia, Czechia and Poland.

A negative correlation: Italy, Portugal, Estonia, Hungary and Bulgaria.

Further Approach

O1: Dataset preprocessing

O1.1 Loading datasets and metadata

O1.2 Filtering and renaming dataset variables

O1.3 Missing data review and country selection

O2: Analysing chosen efficiency metrics

O2.1 Calculating Annual Spending Efficiency across countries

O2.2 Calculating Annual Labor Intensity across countries

O3: Analysing participation of female researchers

O3.1 Calculating Share of Female Researchers in all Researchers

O3.2 Spending Efficiency vs. Female Researchers Share

O3.3 Calculating Growth Rates of chosen metrics

Example – CAGR calculation

```
print("• Calculating Researcher FTE CAGR")
res_cagr_calc = df.query("(Year >= '2009') and (Year <= '2021')").sort_values(
    ['Country', 'Year']).groupby(
    ['Country'])[['FTE Researcher']].apply(
    lambda x: pd.Series({
        'Res CAGR 2009_2021': ((x.iloc[-1,].item() / x.iloc[0,].item())**(1/13) - 1)
        if not (x.iloc[0,].isna().any())
        else float('nan'))}).reset_index()
nan_pct = res_cagr_calc['Res CAGR 2009_2021'].isnull().sum()
            *100/len(res_cagr_calc['Res CAGR 2009_2021'])
print(f" - Res CAGR: {nan_pct:.0f}% values converted to NaN")
df = df.merge(res_cagr_calc, on=['Country'], how = 'left')
```

```
g = sns.catplot(kind = 'bar', x = 'Country', y = 'CAGR value',
                hue = 'CAGR types', data = df_cagr)
g.set(title = 'Compound Annual Growth Rates between 2009 and 2021',
      xlabel = "Country", ylabel = "CAGRs")
sns.move_legend(g, "upper right", bbox_to_anchor = (0.95, 0.95), ncol = 1)
g.fig.set_size_inches(10,4)
```

SOLUTION

I used pandas library to calculate compound annual growth rates of chosen metrics in a previously established 2009-2021 period.

After reshaping the dataset, I used matplotlib and seaborn to plot the comparison of all CAGR metrics between countries.

Further Approach

O1: Dataset preprocessing

O1.1 Loading datasets and metadata

O1.2 Filtering and renaming dataset variables

O1.3 Missing data review and country selection

O2: Analysing chosen efficiency metrics

O2.1 Calculating Annual Spending Efficiency across countries

O2.2 Calculating Annual Labor Intensity across countries

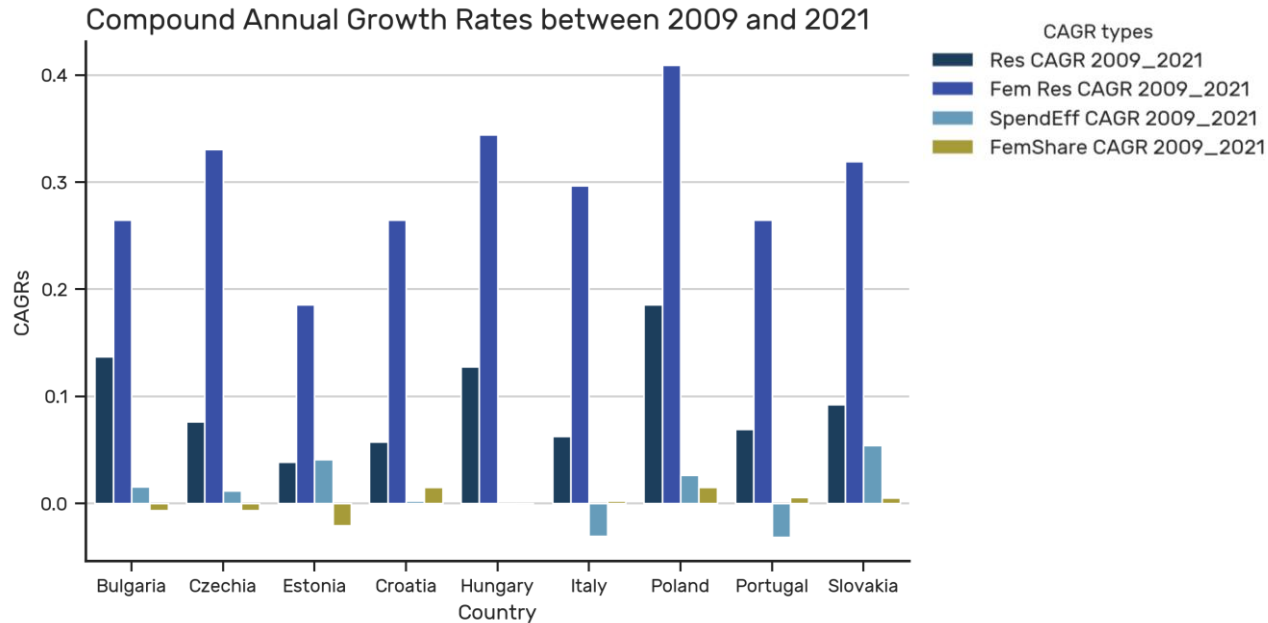
O3: Analysing participation of female researchers

O3.1 Calculating Share of Female Researchers in all Researchers

O3.2 Spending Efficiency vs. Female Researchers Share

O3.3 Calculating Growth Rates of chosen metrics

Visualization



INSIGHT

Positive growth rates for spending efficiency and female share observed only for Poland and Slovakia.

Croatia presented a stagnant spending efficiency.

Bulgaria, Czechia and Estonia: increasing growth rates in spending efficiency with decreased female share.

Despite a growing workforce, Italy and Portugal recorded a decrease in spending efficiency.

Hungary presented a stagnant spending efficiency and share of female researcher positions.

Further Approach

O1: Dataset preprocessing

O1.1 Loading datasets and metadata

O1.2 Filtering and renaming dataset variables

O1.3 Missing data review and country selection

O2: Analysing chosen efficiency metrics

O2.1 Calculating Annual Spending Efficiency across countries

O2.2 Calculating Annual Labor Intensity across countries

O3: Analysing participation of female researchers

O3.1 Calculating Share of Female Researchers in all Researchers

O3.2 Spending Efficiency vs. Female Researchers Share

O3.3 Calculating Growth Rates of chosen metrics

Source metadata

```
print("Data Sources:")
print("• Current analysis was prepared based on the following information sources:")

for i, row in mdf.iterrows():
    print(f" {i+1}. {mdf.dataset_id[i]} dataset provided by: {mdf.dataset_source[i]}")
    print(f"      Last updated: {mdf.dataset_last_updated[i]}")
print()
```

Data Sources:

- Current analysis was prepared based on the following information sources:
 1. htec_sti_exp2 dataset provided by: Eurostat
Last updated: 29/09/2025 23:00
 2. htec_sti_pers2 dataset provided by: Eurostat
Last updated: 29/09/2025 23:00
 3. rd_p_bempoccr2 dataset provided by: Eurostat
Last updated: 02/05/2025 11:00

SOLUTION

I display the metadata related to web-scraped datasets used during the analysis to inform future data interpretations

Further Approach

O1: Dataset preprocessing

O1.1 Loading datasets and metadata

O1.2 Filtering and renaming dataset variables

O1.3 Missing data review and country selection

O2: Analysing chosen efficiency metrics

O2.1 Calculating Annual Spending Efficiency across countries

O2.2 Calculating Annual Labor Intensity across countries

O3: Analysing participation of female researchers

O3.1 Calculating Share of Female Researchers in all Researchers

O3.2 Spending Efficiency vs. Female Researchers Share

O3.3 Calculating Growth Rates of chosen metrics

Summary – analysis limitations

- Data Coverage:
missing data exists before 2008 and after 2022, with optimal metric comparability limited to the 2009-2021 period.
- Country selection:
only 9 countries met data quality criteria (at least 80% completeness).
- Sector heterogeneity:
NACE G–N covers multiple service subsectors that can influence spending and hiring patterns.
- Correlation:
does not imply causation, therefore observed links between female share and efficiency may reflect confounders.

Important note:

Offered insights are not based on a structured academic knowledge about EU policy or geo-political structure of discussed countries.

My aim was to simply demonstrate my ability to understand complex datasets, data characteristics and measures that lay outside my primary academic expertise.

Summary – key findings

GENERAL TRENDS

- **2005 – 2023:** spending efficiency generally increased and female researcher representation gradually improved.
- **2009 – 2021:** number of female researchers grew faster than overall employment.

COUNTRY-WISE OBSERVATIONS

- **Poland:** stable but relatively low spending efficiency, paired with high but stable labor intensity.
- **Croatia, Czechia, Poland:** female inclusion coincided with more effective use of R&D resources.
- **Italy, Portugal, Estonia, Hungary, Bulgaria:** female inclusion coincided with less effective use of R&D resources.
- **Poland, Slovakia:** simultaneous increases in spending efficiency, female share, and workforce size.
- **Croatia, Italy, Portugal:** stagnant or declining efficiency but strong female inclusion.
- **Bulgaria, Czechia, Estonia:** improved efficiency but declines in female share.

Summary – Categorization

GENERAL TRENDS


- **2005 – 2023:** spending efficiency generally increased and female researcher representation gradually improved.
- **2009 – 2021:** number of female researchers grew faster than overall employment.

COUNTRY-WISE OBSERVATIONS


- **Poland:** stable but relatively low spending efficiency, paired with high but stable labor intensity.
- **Croatia, Czechia, Poland:** female inclusion coincided with more effective use of R&D resources.
- **Italy, Portugal, Estonia, Hungary, Bulgaria:** female inclusion coincided with less effective use of R&D resources.
- **Poland, Slovakia:** simultaneous increases in spending efficiency, female share, and workforce size.
- **Croatia, Italy, Portugal:** stagnant or declining efficiency but strong female inclusion.
- **Bulgaria, Czechia, Estonia:** improved efficiency but declines in female share.




Balanced growth (Poland, Slovakia)

 **Business Insight:** diversity initiatives combined with targeted funding and organizational change could result in a stable growth.

Initial investment (Italy, Portugal, Croatia)

 **Business Insight:** rising female participation can cause declining spending efficiency (initial integration costs) but may also bring long-term efficiency benefits.

Efficiency prioritization (Bulgaria, Czechia, Estonia)

 **Business Insight:** Without corresponding diversity policies, improved spending efficiency may have favored male-dominated recruitment or sub-sectors.

Final results

References



eu_efta_countries.csv

list of EU and EFTA countries for filtering data;



scraper_data.csv

*main dataset merged from the external data sources
{htec_sti_exp2, htec_sti_pers2, rd_p_bempoccr2};*



scraper_metadata.csv

*metadata from external data sources including
dataset IDs, source, title and the date of their last update;*



scraper_code.ipynb / scraper_code.py

*Python codes that can be re-used in the future
in case of an Eurostat database update.*



analysis_data.csv

main preprocessed dataset ready for data analysis;



cagr_analysis_data.csv

CAGR metrics dataset ready for data analysis;



analysis_code.ipynb / analysis_code.py

*the Python codes that can be re-used in the future
in case of an Eurostat database update.*

Carucci, R. (2024)

One More Time: Why Diversity Leads to Better Team Performance.
Forbes

Hoogendoorn, S., Oosterbeek, H., & van Praag, M. (2019)

When Gender Diversity Makes Firms More Productive.
Harvard Business Review

Niederle, M., Segal, C., & Vesterlund, L. (2008)

How Costly is Diversity? Affirmative Action in Light
of Gender Differences in Competitiveness.
NBER Working Paper

Phillips, K. (2014)

How Diversity Makes Us Smarter.
Scientific American

Wearden, G. (2011)

EU debt crisis: Italy hit with rating downgrade.
The Guardian