[Github link](https://github.com/Carloselrecharlie/ML_assessment_repeat.git)

Wordcount:1296

Contents

[1. Introduction 1](#_Toc142256148)

[2. EDA 1](#_Toc142256149)

[2.1 Earnings 7](#_Toc142256150)

[3. Forecast 11](#_Toc142256151)

[4. Sentiment analysis 12](#_Toc142256152)

[5. References 12](#_Toc142256153)

[Appendix 13](#_Toc142256154)

# Introduction

The source data from Ireland (*EHQ10 - Public Sector Employment and earnings* 2023) is licensed under Creative Commons Attribution 4.0. The data from Spain was taken from the National Institute of Statistics following the instructions of the section “Re-use of the information contained on this website” from the legal notice of their website (INE, no date). And according to that notice this document is referenced as “Own compilation with data taken from the INE website: [www.ine.es](http://www.ine.es)”, being the source data updated up to the first quarter of 2023. Finally the data from the United Kingdom (ONS 2023) is published under the Open Government License (OGL).

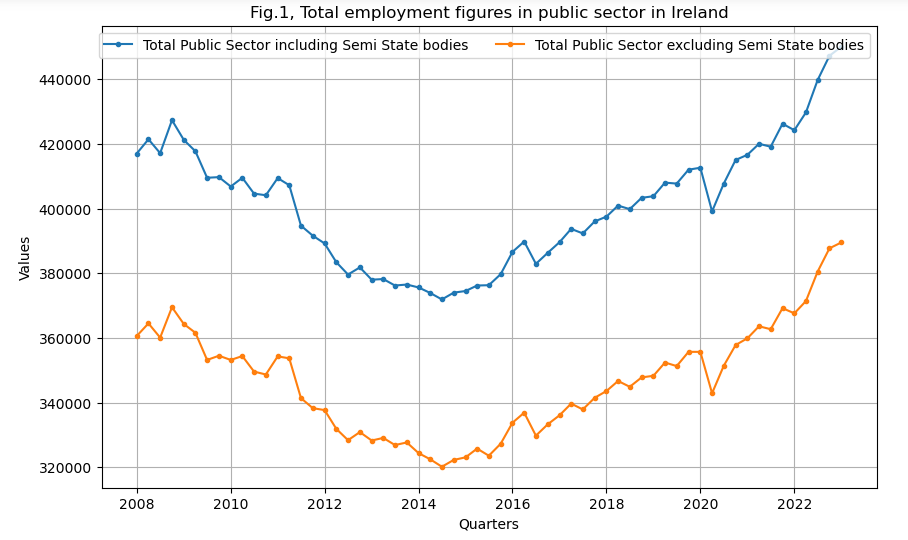
The data mining methodology chosen for this project is CRISP-DM, the main reason is because it includes business understanding stage which is necessary to clarify the expectations from the organization (IBM 2016), a state in this case, and states also are businesses, just like a big corporation with the same needs. This step really helps this project to point to a successful direction.

It also includes the implementation stage which focuses in taking advantage of what achieved through the previous work. At least from the design point of view, this methodology is the most complete

# EDA

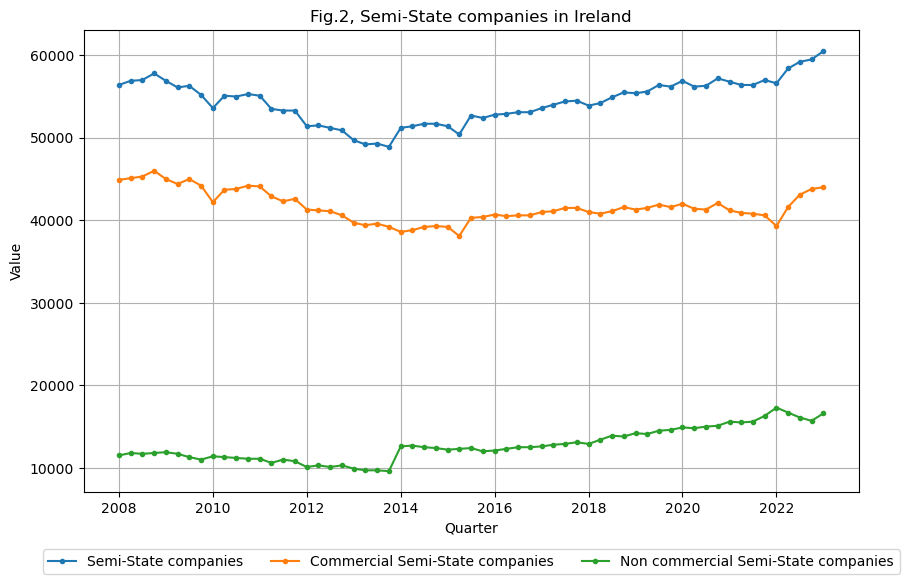
The charts of time series between 2008 and 2023 from total public employment in Ireland shows two facts essentially:

1. The trend changed from decreasing to increasing since the third quarter of 2014
2. The semi state bodies are a significant part of the total public employment, which plot would same a similar shape than the corresponding ones below, given these two also have very similar spikes.



Regarding semi-state companies, these are mainly represented by commercial ones with 3- or 4 more-times employment, with decreasing figures until 2015, where after a slight increase the third quarter of 2015, it remains more stable until 2022, where another positive spell is obvious.

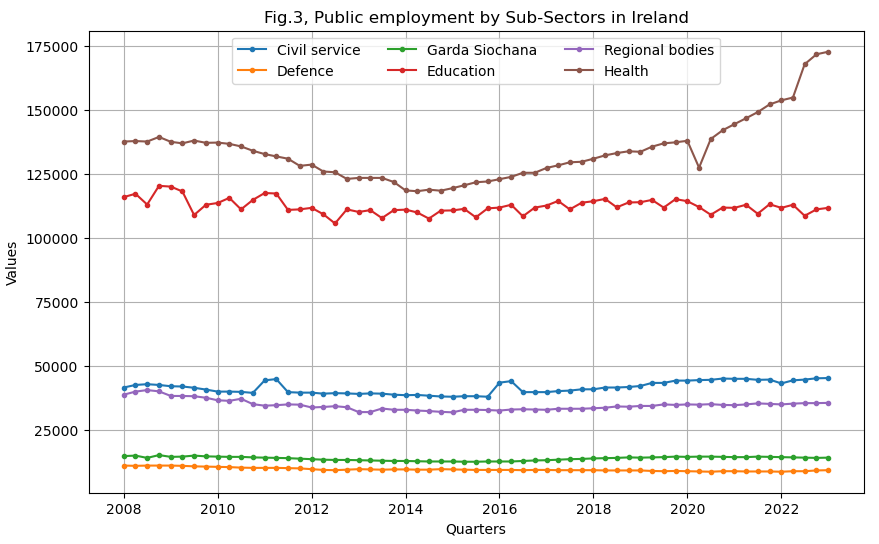
Similarly, the non-commercial ones were slightly decreasing until 2014 but the first quarter of this year and year 2022, there were positive spikes in employment.



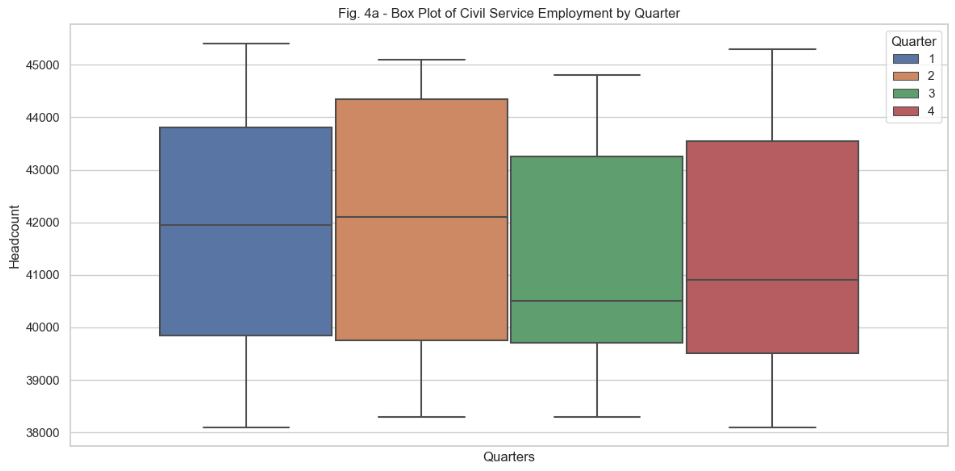
In general the public employment by subsectors in Ireland has not seemed to vary meaningfully within the last 15 years. The main character has been the health sector, for being the leading employer (ranging between 120 and 175 thousand of employees) and the group which has increased the most. In fact, between 2008 and 2014 the tendency was negative with a progressive decrement, however after a stable year 2015 the tendency switches to a slight employment increase. And this only changes again in the second quarter of 2020 with COVID pandemic. Anyhow the positive balance keeps growing the following quarter, especially the third quarter of 2022.

Education is the second most important public employment sub-sector (above 100 and below 125 thousand employees) and it clearly shows a seasonal pattern where it decreases every fourth quarter, corresponding to the summertime (holiday). Finally civil service and regional bodies have have similar figures (35-40 thousand) and it is worth mentioning that the civil service had two two two-quarters increments at the beginning of 2011 and 2016. Perhaps the first increment was related to the Public Service Reform Plan which took place that year.

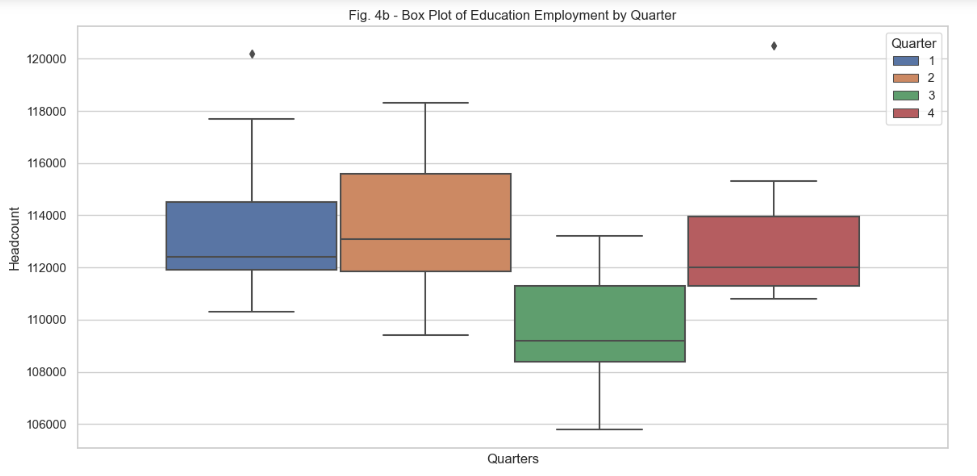
Finally Garda Siochana and Defense have the smallest contribution to public employment (11-13 thousand) showing the most stable figures (flattest plots).



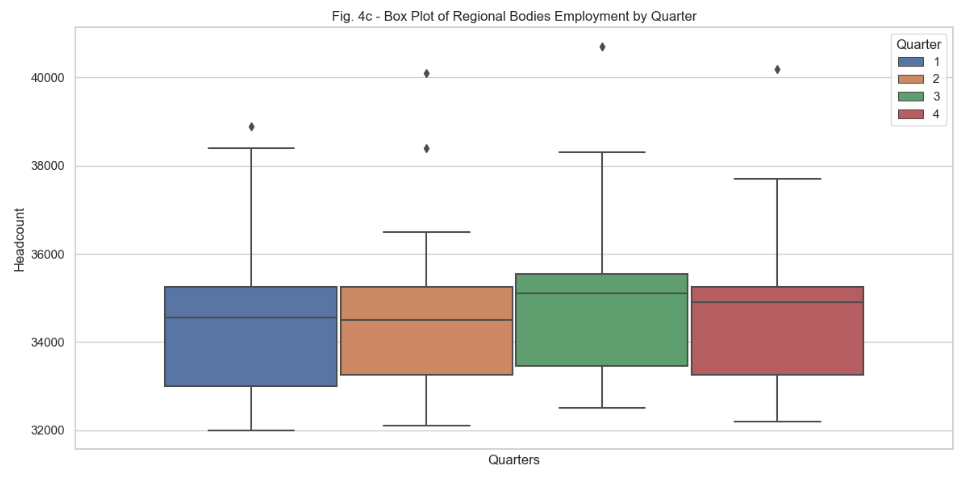
The quarterly distribution by sub-sectors is as follows: in Civil Service a lot of the data gather around the mean, and the 3r quarter stands out for having many observations between the second and third quartile:

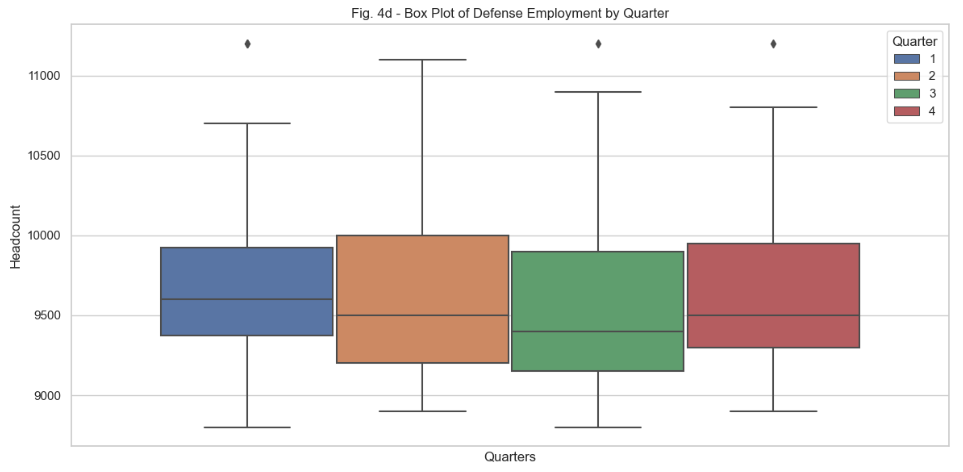


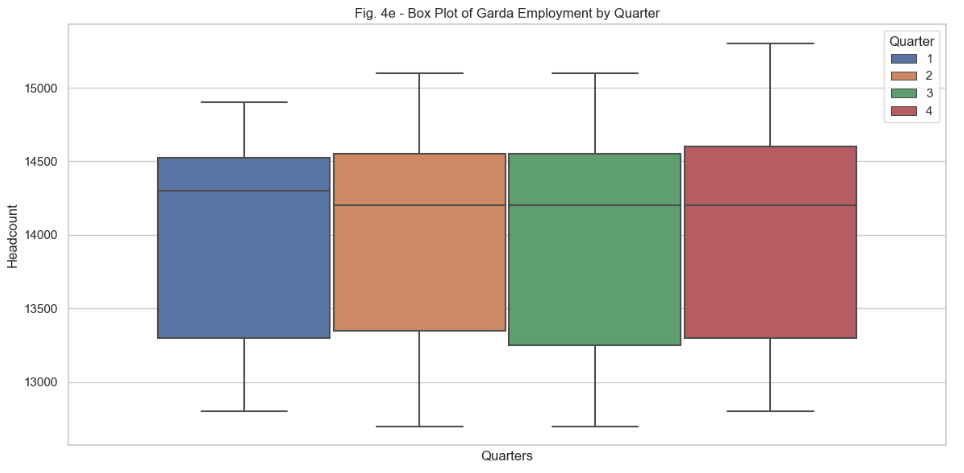
Again the evidence that the employment in Education decreases in the third quarter (summer).There is more variability between quarters:

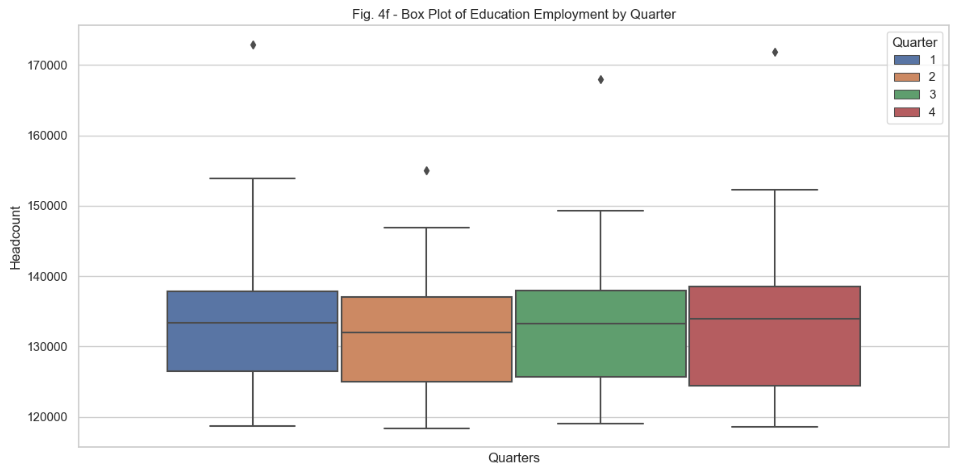


Regarding the three remaining sub-sectors it seems like they have a similarity, and it is the fact that there is little variance between the quarters, they have a similar distribution within each sub-sector:





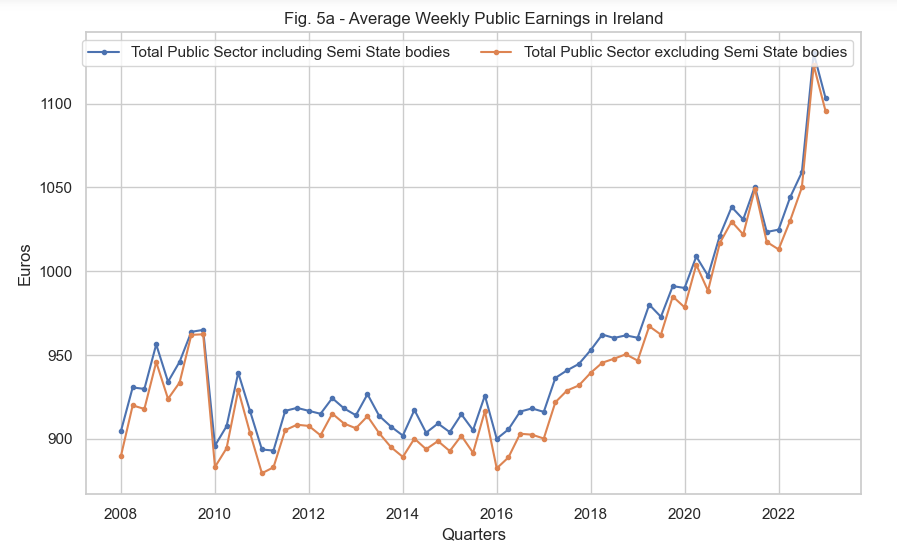


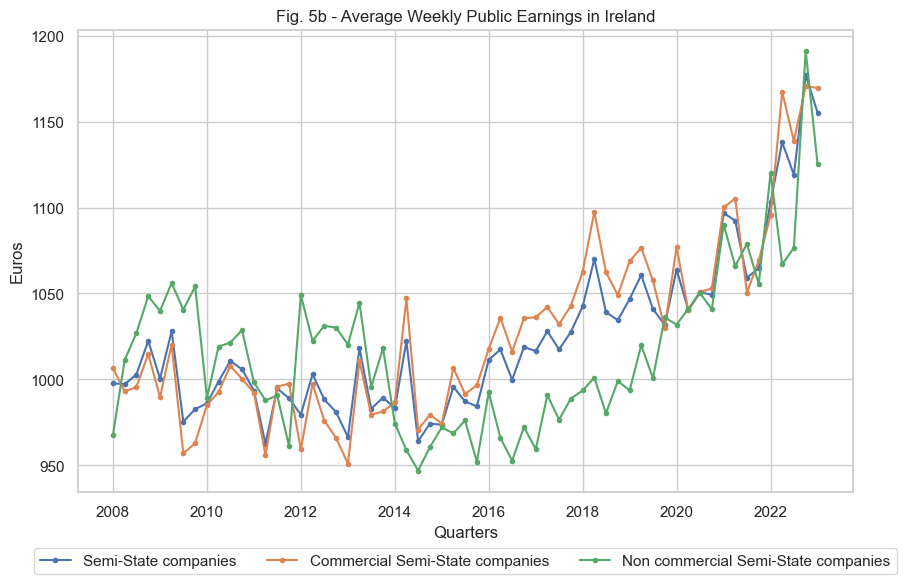


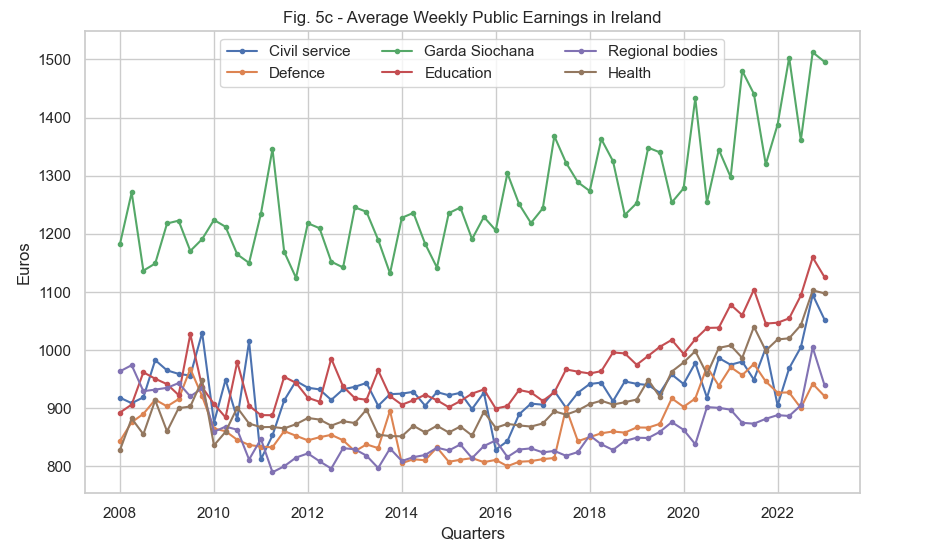
## 2.1 Earnings

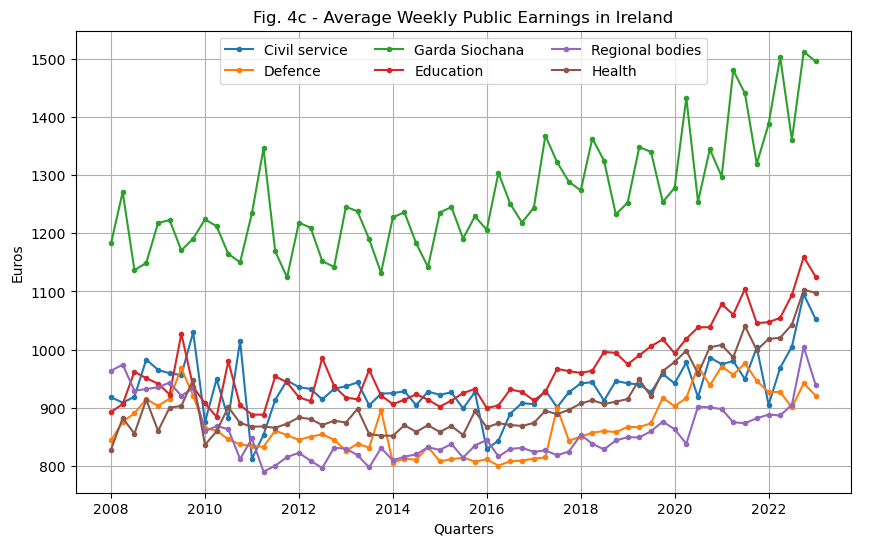
The public salaries in Ireland were mainly above 900 euros weekly until 2016, when there is a clear change in the pattern with a slow and progressive increment starts until the last observation, which is around 1,100 euros per week. The semi-state bodies follow a fairly similar pattern and did increase the average salaries between 2011 and 2019. Before and after these years they did not seem to add anything significant to the metric:

As far as semi-state companies are concerned, the non-commercial had higher salaries until 2013 but they fell behind the year after and from 2020 onwards all three types of semi-state companies showed relatively similar salaries:





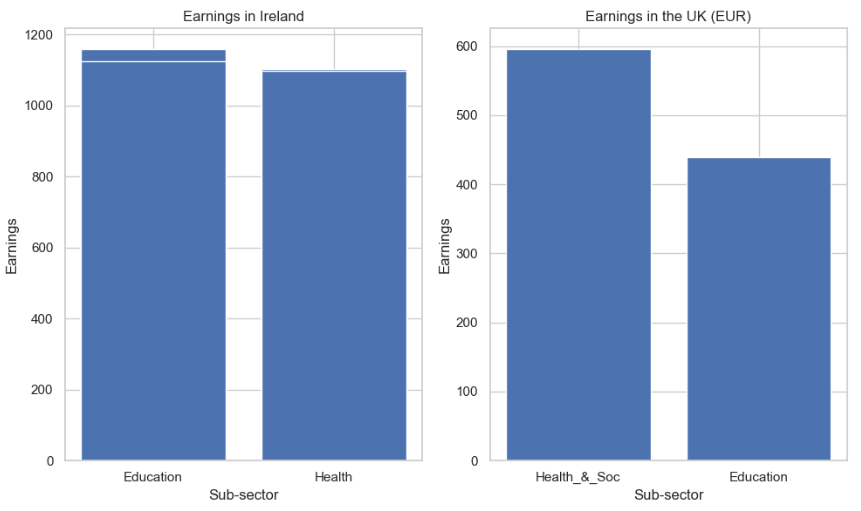


Regarding sub-sectors it is more than obvious that Garda Siochana has always had a salary greater than the other subsectors, however they all have increased similarly and seem to follow the same approximate trend keeping their ranges within the same boundaries, with Education following the Garda, then Health, Civil Service, Defense and finally the Regional Bodies. Perhaps Civil Service has more variability around 2010.

As far as the U.K. is concerned, the time frame covered in their data is wider but only the period matching the Irish one was considered. Related to employment the focus is on Total Public Sector, Local Government, Civil Service, HM forces, Police, Education and NHS. Since the Irish dataset includes both full time and part-time employees (as confirmed with CSO by email), only the HC (headcount) columns from UK were taken into account as opposed to FTE (full-time equivalent).

And about earnings only Public sector, Education and Health match between data sets and could be compared, however the data from the UK Public Sector covers a time frame different to the Irish one so finally just Education and Health can be compared. The process of exchange currency was done by converting British pounds to Euros, which was achieved with a function to assign the corresponding average exchange rate per each of the years considered (rates taken from [www.ofx.com](http://www.ofx.com)).

The comparison of average weekly earnings between Education and Health employees really stands out, note the difference of the scale between both Y axis, despite both being in Euros:



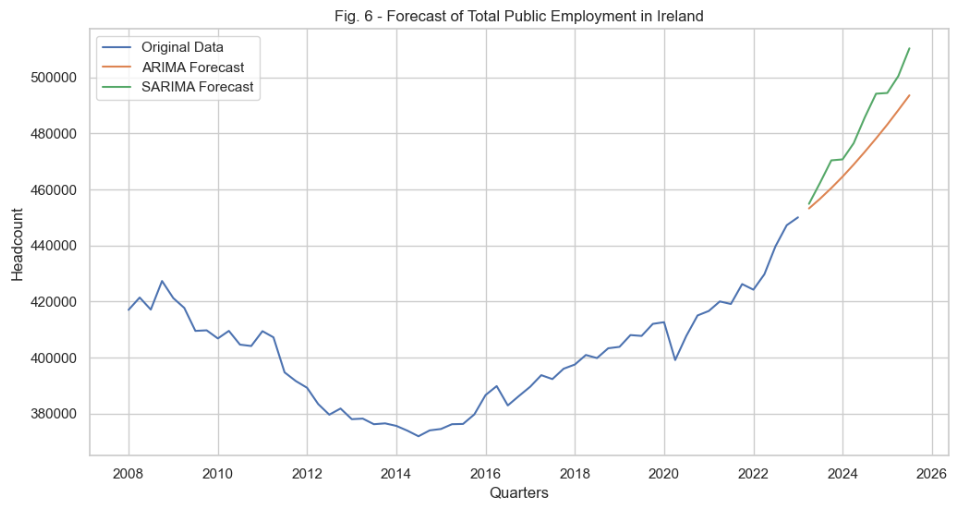
# Forecast

Since there is no dependent variable at first there cannot be supervised learning and considering the type of data it just makes sense to perform a time series analysis. Therefore ARIMA context was deemed suitable for this project and both ARIMA and SARIMA models were implemented as they are designed to identify and retain the autoregressive (AR) and moving average (MA) components of the data. They are also suitable for capturing trends and seasonal patterns in time series.

Considering this scenario there application of techniques like dimensionality reduction was not taken into account.

ParameterGrid was used to combine in all possible ways the order values for ARIMA and the seasonal order values for SARIMA. Then an iterative process ran through these combinations, fitted the models, and selected the combination with the lowest AIC (Akaike Information Criterion) as the most suitable set of parameters (Scikit-learn Developers, *3.2. tuning the hyper-parameters of an estimator*), since the lower AIC values, the better trade-off between model fit and complexity.

The following 10 values of Total Public Employment were forecast and both models show a clear increment in this sector, being SARIMA more optimistic:



There were issues with indexes when calculating scoring metrics but for the evaluation of performance MAE was selected to weigh the average magnitude of errors, MSE because penalizes larger errors and RMSE as it uses same units as data (headcount).

# Sentiment analysis

A Twitter developer account was setup for this purpose which provided access to a subset of Twitter API v2 endpoints and limited v1.1 endpoints however there were difficulties accessing the allowed endpoints and all the attempts failed. It did not seem an issue with the account or tokens/keys but with accessing the correct endpoints.

Researched online but there did not seem to be datasets on a related topic, reason why next step was to bulk download from [www.archive.org](http://www.archive.org) and filter by topic. The monthly tar files from January and February 2013 (>70 GB) had no matches for none of the following strings 'public employment','public sector employment','health employment','education employment','civil service employment', 'defence employment','regional bodies employment','garda employment'.

# References

*Archive team: The Twitter Stream Grab* (2013) *Archive Team: The Twitter Stream Grab : Free Web : Free Download, Borrow and Streaming : Internet Archive*. Available at: [https://archive.org/details/twitterstream?tab=collection&sort=-publicdate&and%5B%5D=year%3A"2013](https://archive.org/details/twitterstream?tab=collection&sort=-publicdate&and%5B%5D=year%3A%222013)" (Accessed: 26 July 2023).

Central Statistics Office (2023) *EHQ10 - Public Sector Employment and earnings*, *Data.Gov.IE*. Available at: https://data.gov.ie/dataset/ehq10-public-sector-employment-and-earnings?package\_type=dataset (Accessed: 08 July 2023).

*IBM (*2016*). IBM SPSS Modeler Crisp-DM Guide*. Available at: https://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/18.0/en/ModelerCRISPDM.pdf (Accessed: 12 July 2023).

*Legal notice* (no date) *INE*. Available at: https://www.ine.es/dyngs/AYU/en/index.htm?cid=125 (Accessed: 17 July 2023).

ONS (2023) *Public Sector Employment Time Series - Office for National Statistics*. Available at: https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/publicsectorpersonnel/datasets/publicsectoremploymenttimeseriesdataset (Accessed: 15 July 2023).

*ONS* (2023) *Average weekly earnings time series - Office for National Statistics*. Available at: https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/datasets/averageweeklyearnings (Accessed: 18 July 2023).

Scikit-learn Developers (no date) *3.2. tuning the hyper-parameters of an estimator*, *scikit*. Available at: https://scikit-learn.org/stable/modules/grid\_search.html (Accessed: 01 August 2023).

Twitter (no date) *Determining tweet types | docs | twitter developer platform*, *Twitter*. Available at: <https://developer.twitter.com/en/docs/tutorials/determining-tweet-types> (Accessed: 15 May 2023).

# Appendix

UK data: Public sector employment time series - Office for National Statistics -> pse.csv <https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/publicsectorpersonnel/datasets/publicsectoremploymenttimeseriesdataset>

UK data: Average weekly earnings time series -> emp.csv <https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/datasets/averageweeklyearnings>

GBP to EUR average annual exchange rates <https://www.ofx.com/en-ie/forex-news/historical-exchange-rates/yearly-average-rates/>

Twitter documentation:

<https://developer.twitter.com/en/docs/twitter-api/getting-started/getting-access-to-the-twitter-api>

<https://developer.twitter.com/en/docs/twitter-api/v1>