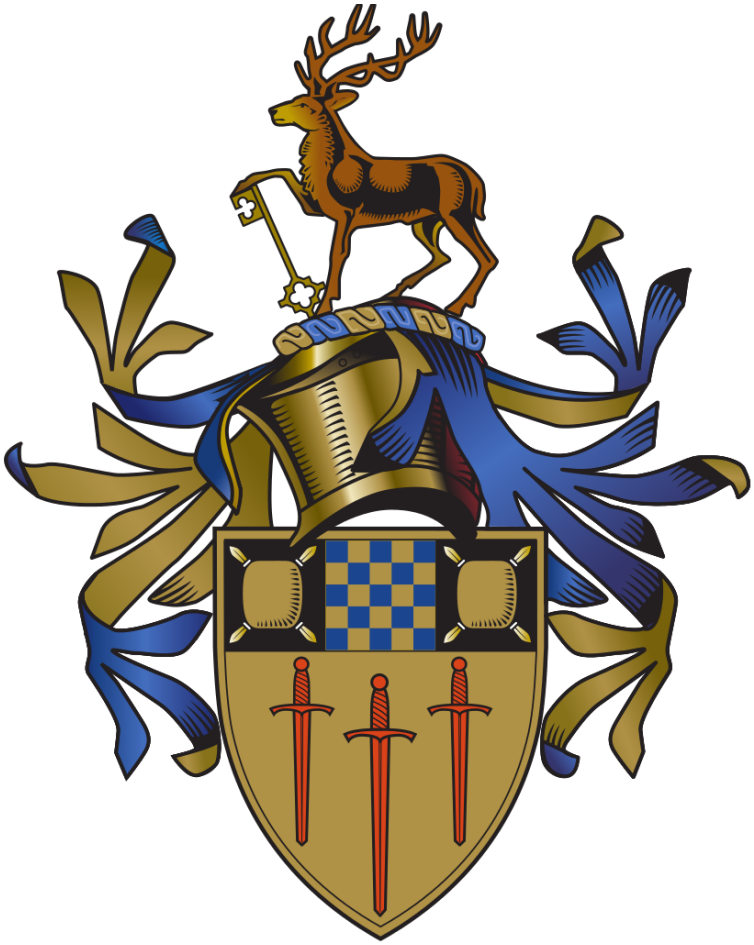
Shrinkage and Selection in "Big Data": Nowcasting the United Kingdom's Labour Market with Google Trends



Arthur Grainger

University of Surrey

MSc Economics (Econometrics and Big Data)

September 2025

Abstract

This paper examines the potential of Google Trends (GT) data to nowcast the United Kingdom's unemployment rate from January 2004 to March 2025. We apply a two-stage procedure where a rolling-window approach is combined with shrinkage estimations (Lasso and Elastic Net) for variable selection, followed by OLS estimation to identify the most predictive variables and track how their importance evolves over time. Our in-sample analysis reveals a distinction between stable predictors and unstable 'event-driven' ones whose importance shifts across economic regimes. Out-of-sample evaluation shows that linear GT-augmented models provide only marginal and statistically insignificant improvements in nowcasting accuracy, with value confined to periods of high economic volatility. However, a Random Forest model using the top eight selected variables achieves statistically significant improvement over the AR(3) benchmark, suggesting the linear results reflect functional form misspecification rather than GT data limitations. The predictive signal decays quickly, leaving simple autoregressive models most effective for 12-month forecast horizons.

Contents

Word Count: 9,574

[1]To ensure reproducibility, all Python scripts, datasets, and package dependencies used in this research are available in the Replication Package.

1 Introduction

The unemployment rate for the United Kingdom (UK) is released with a publication lag of approximately 6-8 weeks. Consequently, policymakers may be slow to react to emerging labour market trends, while investors and businesses are left with an incomplete picture for decision making. This delay necessitates the use of more timely, high-frequency data to form an accurate picture of the current state of the labour market. The formal term for this predictive exercise is *nowcasting*. Popularised in economics by Giannone et al. (2008), it is broadly defined as the prediction of the present, the very near future, and the very recent past.

To bridge this information gap, data from Google Trends (GT) offers a compelling real-time source of information, particularly as the internet has become a primary resource for job searching (Kuhn, and Mansour, 2022). For example, Faberman and Kudlyak (2019) find that approximately 38% of the employed and 74% of the unemployed reported using the internet for their job search. Given this substantial reliance on internet-based job searching, the choice of search engine becomes crucial for capturing representative search behaviour patterns. Google holds 93.4% of the search engine market in the UK (StatCounter, 2025), making it the natural choice for analysing online job-search activity.

We wish to investigate two core questions. First, which specific GT variables are selected the most and when, focusing on coefficient performance during significant economic regimes such as the Great Recession and the Covid-19 pandemic. Second, assess whether GT data provides incremental predictive power for nowcasting UK unemployment and when these improvements are most prevalent. As these investigations involve a large number of predictors and a limited sample size, the scale of the data necessitates a methodology capable of addressing these selection challenges.

A major challenge in the literature is selecting the most relevant search terms to predict a target variable (Smith, 2016). This paper addresses this issue by adopting a "big data" approach that uses a wide range of GT search terms, topics, and websites. This method allows for comprehensive coverage of unemployment-related search behaviour, capturing not only direct job search activity but also benefit-seeking, regional labour market dynamics, and sectoral employment patterns. To systematically manage this high-dimensional data, we employ a rolling-window approach combined with shrinkage estimations (Lasso and Elastic Net) to algorithmically identify not only which variables are most predictive, but also how their importance evolves over time.

Our methodology applies a post-selection Ordinary Least Squares (OLS) framework to

a high-dimensional monthly dataset spanning from January 2004 to March 2025. This process first uses shrinkage models (Lasso and Elastic Net) for automated variable selection within a constrained 48-month rolling window. These algorithms are well designed to handle a challenging "$p = n$" scenario where the number of predictors ($p = 48$) equals the number of observations. A final OLS model is then estimated using only this selected subset of predictors. A key advantage of this data-driven approach is that it circumvents the need for subjective, *a priori* variable selection, allowing the algorithm to automatically identify the most informative predictors.

Before modelling, all data series are transformed to ensure stationarity, and predictors are seasonally adjusted. The two-stage post-shrinkage OLS procedure serves two goals: (1) for the out-of-sample evaluation, it generates forecasts which are compared against a benchmark model using RMSE and the Diebold-Mariano test; (2) for the in-sample analysis, the framework is used to investigate predictor importance by tracking the selection frequency and coefficient evolution of the variables. This reveals not only which predictors are most consistently selected but also how their predictive power changes across different economic regimes.

This paper contributes to the literature in three main ways. First, this paper investigates using a shrinkage framework to the variable selection problem, algorithmically identify the most salient variables from a large set of UK search queries. This approach is then combined with a rolling-window analysis, enabling us to move beyond a static view of predictor importance and instead investigate how the set of key variables evolves over time.

Second, our analysis demonstrates that the predictive power of GT data is not static. Its primary value is concentrated in the very short term and is most pronounced during periods of high economic volatility, such as the Great Recession and the Covid-19 pandemic. During stable periods, its predictive advantage is minimal. Furthermore, we confirm that GT data acts as a complement, not a substitute, to traditional economic indicators by providing timely signals before official data is released. A robustness check using a Random Forest model further highlights the potential for non-linear relationships. It achieved a statistically significant improvement in nowcasting accuracy over the autoregressive benchmark, whereas our linear models did not.

Finally, this paper makes an analytical contribution by distinguishing between two types of predictors: more stable indicators and unstable 'event-driven' indicators. This distinction offers a compelling explanation for an observed asymmetry in our models' performance. While the GT-augmented models perform well during the onset

of a crisis, they often struggle in the immediate aftermath. We hypothesise that this occurs because the rolling-window architecture retains crisis-relevant 'event-driven' predictors, which may then become less effective or even misleading as the economy transitions into a new more stable regime.

The remainder of this paper is structured as follows. Section 2 reviews the relevant literature, tracing the evolution of Google Econometrics, discussing common approaches for incorporating GT data, and establishing the applicability of shrinkage methods. Section 3 details the methodology, outlining the data sources, pre-processing steps, and the modelling framework. We then present our empirical results, first analysing the in-sample variable selection and coefficient evolution in Section 4, before evaluating the out-of-sample nowcasting performance in Section 5. Finally, Section 6 discusses the implications of these findings and offers concluding remarks.

2 Literature Review

2.1 The Evolution of Google Econometrics

The foundational work of Ettredge et al. (2005) provided the initial proof-of-concept for using internet search data as a leading indicator of unemployment. They employ a simple linear regression on employment-related search terms (such as "jobs", "job search", "resume", "employment", and "job listings") over a modest 77-week sample from September 2001 to March 2003. Building on this preliminary foundation, the seminal paper by Google's economists Choi and Varian (2012) – coupled with the 2008 release of Google Insights for Search, later rebranded as Google Trends (Smith, 2016) – sparked broader interest in exploring how internet search data can enhance standard time series models when nowcasting. This interest was further amplified by the urgent need for more timely economic indicators during the 2008 financial crisis (D'Amuri, 2009), establishing the methodological framework that would define the field.

Following Choi and Varian's breakthrough, the field expanded rapidly, with researchers applying increasingly sophisticated methodologies to a range of economic variables. The evolution has moved from simple augmented regressions to complex machine learning approaches, while the data sources have expanded from individual search terms to comprehensive topic categories and real-time indicators explored in economic monitoring systems like the OECD Weekly Tracker (Woloszko, 2020). This methodological sophistication proved prescient during the Covid-19 pandemic, when the need for real-time economic monitoring became even more critical as tra-

ditional statistics struggled to capture the unprecedented speed and scale of market disruptions (Bank of England, 2020).

Standard time series modelling in the literature tends to include simple autoregressive models (AR) or more sophisticated autoregressive integrated moving average (ARIMA) models. These are then augmented with manually selected internet-search data (Askitas, and Zimmermann, 2009; Choi, and Varian, 2012; D'Amuri, and Marcucci, 2017; Katris, 2019; McLaren, and Shanbhogue, 2011; Pavlicek, and Kristoufek, 2015; Vosen, and Schmidt, 2011). Many of these papers then compare out-of-sample forecasts of their economic variables, evaluating the models with and without the internet search variables, using metrics such as root mean square errors (RMSE), Diebold and Mariano (1995) tests of equal forecast accuracy between specifications, and R^2 statistics for in-sample fit.

A representative example of this approach comes from McLaren and Shanbhogue (2011) in the Bank of England's Quarterly Bulletin, who demonstrate a simple application to the UK's labour market. This paper explores how well GT-augmented models nowcast the change in the UK's unemployment rate, among other variables, compared to models augmented by more traditional leading indicators (claimant count and GfK consumer confidence survey).

More advanced statistical techniques and machine learning methods for macroeconomic nowcasting and forecasting have gained considerable traction in recent literature (Ferrara, and Simoni, 2022; Heikkinen, and Heimonen, 2023; Katris, 2020; Kohns, and Bhattacharjee, 2023; Scott, and Varian, 2014). Among these, notable applications include the OECD Weekly Tracker from Woloszko (2020), which leverages neural networks and GT data across a panel dataset of 46 countries for high-frequency GDP prediction.

Heikkinen and Heimonen (2023) present a particularly novel approach, using principal component analysis and partial least squares for dimension reduction to extract information from 181 GT subcategories, enabling effective nowcasting of the purchasing manager's index (PMI) and non-manufacturing survey index (NSI) from *ISM Report On Business* (2025). This innovative methodology demonstrates the potential of GT data to predict traditional leading indicators of macroeconomic activity, which themselves serve as early signals of broader economic conditions.

Despite these promising applications, the relationship between model complexity and forecasting performance remains nuanced. Katris (2020) provides important evidence, from 22 countries, that sophisticated machine learning approaches – such as Artificial Neural Networks, Support Vector Regression, and Multivariate Adaptive Regression

Splines – do not universally outperform simpler methods in unemployment forecasting. Additionally, the high-dimensional nature of GT data, with potentially hundreds of search terms, presents significant challenges for traditional econometric methods.

2.2 Google Trends Data: Selection and Construction Approaches

Studies incorporating both traditional leading indicators and GT data provide more compelling evidence of search data's incremental value, as they test whether internet search variables add more predictive power beyond established economic indicators already used in practice. This approach better reflects the information environment faced by real-world forecasters and policymakers. Woloszko (2020) makes the distinction between two different types of leading indicators: (1) hard indicators which are collected by national administrations/statistical agencies; and (2) soft indicators often timelier but less reliable during crises, for example, survey data or Purchasing Manager's Indexes (PMIs).

Common hard indicators found in the nowcasting literature include unemployment benefits claims and production/sales indexes. Whereas common soft indicators include: consumer/employer confidence surveys, the economic policy uncertainty index (Baker et al., 2016), stock price indexes, and PMIs. GT data would be regarded as a soft indicator in this instance but benefiting from real-time, high-frequency advantages.

The literature traditionally presents three main approaches to selecting what exact internet search data is included in the models. The first approach uses manually selected single search terms, such as "jobs", then including this term the augmented regression (Askitas, and Zimmermann, 2009; D'Amuri, 2009; D'Amuri, and Marcucci, 2017; Fondeur, and Karamé, 2013; McLaren, and Shanbhogue, 2011; Pavlicek, and Kristoufek, 2015). The second approach aggregates many terms into categories or indexes (Askitas, and Zimmermann, 2009; Choi, and Varian, 2012; Heikkinen, and Heimonen, 2023; Smith, 2016; Vosen, and Schmidt, 2011; Woloszko, 2020). The third approach tests multiple individual search terms to identify which performs best at explaining the dependent variable (Ettredge et al., 2005; McLaren, and Shanbhogue, 2011; Preis et al., 2013).

For example, D'Amuri and Marcucci (2017) uses the single keyword "jobs", Vosen and Schmidt (2011) use factor analysis to reduce 56 search categories into four common factors (explaining 61% of the variance), while Preis et al. (2013) test 98 different individual search terms to identify the most predictive ones. On the other hand,

Koop and Onorante (2019) argue that the best way to include the Google data is in the form of model probabilities rather than linear predictors. In this paper, the authors create *Google Probabilities* – exploiting the key finding in the literature that GT is particularly good at signalling turning points – to control their Dynamic Model Selection (DMS) framework.

Smith (2016) highlights that selecting specific search terms is a key challenge in unemployment forecasting. A popular approach, suggested by the literature, is thinking of the typical search behaviours of someone who has recently become unemployed or who anticipates to be unemployed in the near future. For example, searches for jobs, financial support, résumé writing assistance, or benefit eligibility requirements may capture different aspects of the job loss and job search process. However, Smith also notes that individual terms are *vulnerable to change* due to policy reforms such as Universal Credit. This is because such reforms directly change the public's search vocabulary, as terms for benefits are replaced by new ones, diminishing the predictive power of the original keywords.

To address this challenge, modern GT provides a *topics* feature that groups together related search terms with similar meanings. Unlike searching for a single keyword like "unemployment", a topic captures all searches related to that concept, regardless of the specific words people use or the language they search in (Google, 2023a). For example, the "unemployment" topic would include searches for "jobless", "out of work", and equivalent terms in other languages used by multilingual residents in the UK. Google determines which searches belong together by analysing user behaviour (specifically where people click after searching) to ensure that grouped terms reflect genuine shared intent rather than merely similar-sounding words (Woloszko, 2020). This approach provides a more complete and accurate picture of public interest in labour market issues and job-seeking behaviour than tracking individual keywords alone.

Furthermore, GT also provides a *website* feature, which allows users to track searches for specific websites. This is particularly useful for tracking searches for job-search websites such as Reed, Indeed, and CV-Library. This study adopts a hybrid approach, incorporating a set of individual search terms, topics, and websites to maximize the information content while using regularization techniques to systematically identify the most predictive variables from this high-dimensional set.

While GT offers a powerful tool for social science research, its use is not without significant methodological challenges. A primary concern is the reliability of the data; because Google provides a random sample of its searches, different downloads can

produce largely different values. Medeiros et al. (2021) demonstrate this instability, showing that different random samples for the same search term can have correlations as low as 0.496. This sample instability can lead to arbitrary conclusions drawn merely by chance. Despite this, in a systematic review of 360 studies, Hölzl et al. (2025) found that only 7% of studies accounted for this issue. To address this critical problem, in this paper we adopt the recommended approach from both papers: collecting multiple samples over several days and averaging them to produce a more reliable and consistent measure.

2.3 The Lasso and Extensions

Given the high-dimensional nature of GT data, there is a clear need for techniques that can effectively perform variable selection. The Least Absolute Shrinkage and Selection Operator (Lasso), first proposed by Tibshirani (1996), provides a powerful solution through its ability to perform automatic variable selection by setting the coefficients of irrelevant predictors exactly to zero. This functionality is the core reason for its use in this paper's post-selection OLS framework. It allows the algorithm to objectively identify a sparse set of the most informative predictors from a large dataset, which can then be used for unbiased estimation and interpretation in a subsequent OLS model.

It allows the algorithm to objectively identify a sparse set of the most informative predictors from a large dataset, which can then be used for unbiased estimation and interpretation in a subsequent OLS model. As demonstrated in the macroeconomic forecasting literature (Ahrens et al., 2021; Audrino, and Knaus, 2016; Bai, and Ng, 2008; Lee et al., 2022; Li, and Chen, 2014; Mei, and Shi, 2024; Scott, and Varian, 2014), this shrinkage approach is particularly valuable for unemployment nowcasting. It maintains model parsimony by automatically selecting the most informative GT variables, a feature essential for real-time policy applications.

Scott and Varian (2014) pioneered this approach, using Bayesian "spike-and-slab" priors with Markov chain Monte Carlo (MCMC) sampling to automatically select the most relevant predictors. Their model drew from a high-dimensional set of 700 variables, which combined 100 terms from the now-discontinued Google Correlate service with 600 predefined GT categories. This approach eliminates the need for subjective term selection or manual aggregation, instead allowing the statistical model to determine which search variables provide meaningful predictive power through probabilistic inclusion indicators. Similarly, applying Lasso/Elastic Net in a rolling-window manner offers a robust data-driven methodology to identify the variables in the active set without requiring *a priori* judgements about term relevance.

Building on the concept of "targeted predictors" from Bai and Ng (2008), this study incorporates crisis-specific search terms such as "Brexit" and "Financial Crisis". The inclusion of these event-driven variables is crucial for our rolling-window analysis, as it allows the model to identify and leverage time-varying predictive relationships to enhance nowcasting accuracy during economic turning points.

In standard Lasso regression, when faced with groups of correlated predictors, the method tends to arbitrarily select just one variable from each correlated group and shrink the others to zero. Zou and Hastie (2005) propose Elastic Net to address this limitation, combining the ℓ_1 (Lasso) and ℓ_2 (Ridge) penalty terms. Unlike pure Lasso, Elastic Net tends to select entire groups of correlated variables rather than individual representatives. This approach distributes the predictive weight across correlated variables instead of concentrating it in a single variable, resulting in more stable coefficient estimates and more interpretable models. Additionally, the Ridge component reduces sensitivity to small perturbations in the data, making variable selection more consistent across different samples or time periods.

To address the coefficient shrinkage bias from regularization, this study follows the post-Lasso OLS methodology, first proposed by Belloni and Chernozhukov (2013), applying the same logic to our Elastic Net models. While the ℓ_1 and ℓ_2 penalties are effective for variable selection, they result in biased coefficient estimates for the selected predictors. Our two-stage approach corrects for this: first, we use Lasso or Elastic Net for variable selection, and second, we run an OLS regression on the selected subset to obtain unbiased coefficients. This method combines the selection strength of regularization with the unbiased estimation of OLS.

These extensions are particularly relevant to our unemployment nowcasting framework with GT data. Our dataset will contain groups of naturally correlated search terms – capturing job search intensity, economic uncertainty, or financial assistance – making Elastic Net's grouped variable selection potentially valuable for identifying thematic search behaviours rather than individual terms. Post-selection OLS allows for robust comparisons between regimes and the reliable identification of which GT variables significantly contribute to unemployment prediction during various economic periods. Ultimately, this addresses the core research question regarding the time-varying predictive power of search data.

2.4 Empirical Findings and Performance

Many papers find that models incorporating internet search data in any form outperform traditional econometric methods in combination with traditional leading indi-

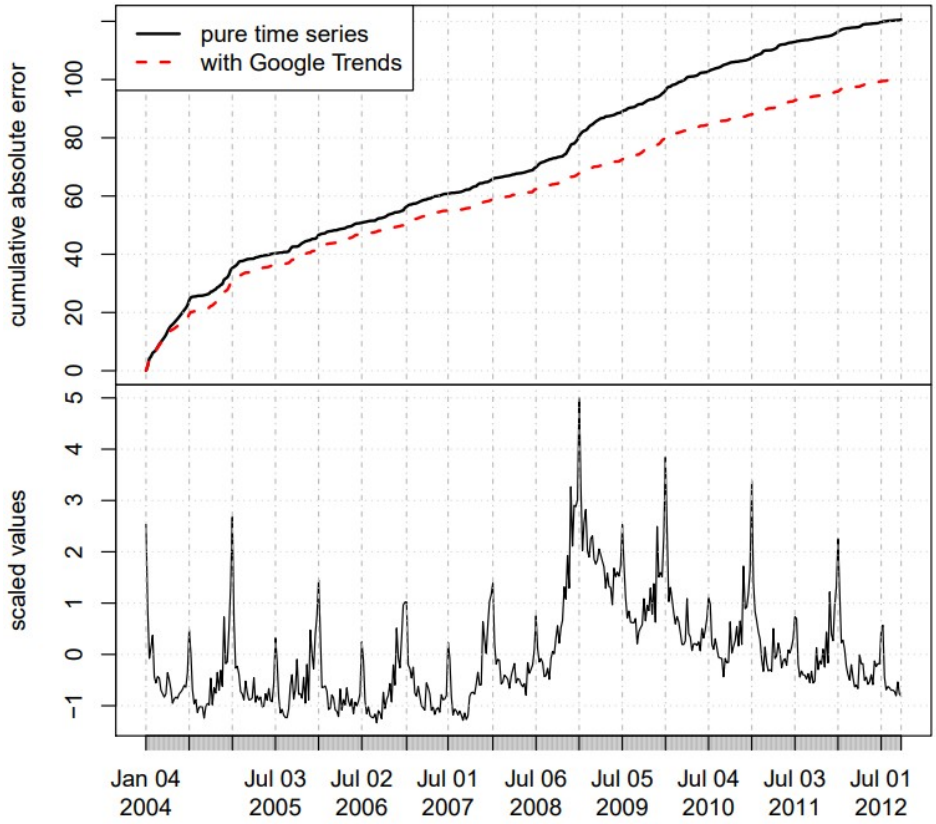
cators (D'Amuri, 2009; D'Amuri, and Marcucci, 2017; Fondeur, and Karamé, 2013; Koop, and Onorante, 2019; Pavlicek, and Kristoufek, 2015; Smith, 2016; Vosen, and Schmidt, 2011). However, the effectiveness of Google search data varies across both countries and time periods. Pavlicek and Kristoufek (2015) demonstrates that the utility of Google search data for nowcasting is country-specific, finding that Google searches enhance nowcasts for the Czech Republic and Hungary but yield limited or mixed results for Poland and Slovakia. The suggested mechanism behind this variation is the prevalent labour mobility in these countries, where many workers seek jobs abroad, creating a mismatch between domestic Google search patterns and local labour market conditions.

There is broad consensus in the literature that internet search data is especially useful for nowcasting during economic turning points (Askitas, and Zimmermann, 2009; Choi, and Varian, 2012; Kohns, and Bhattacharjee, 2023; Koop, and Onorante, 2019; McLaren, and Shanbhogue, 2011; Scott, and Varian, 2014; Smith, 2016; Woloszko, 2020). The underlying mechanism is that, while economic variables can change rapidly during turbulent periods, publication delays for traditional indicators remain constant, substantially increasing the relative value of real-time internet search data for capturing contemporaneous economic conditions.

Scott and Varian (2014) provide a particularly clear visualization of this phenomenon, showing a dramatic divergence in cumulative absolute errors for their Bayesian structural time series models with versus without GT data during 2008–2009 (see Figure 1). The GT models maintained relatively stable error accumulation while traditional time series models experienced substantial performance degradation during the financial crisis. Corroborating this finding, Smith (2016) show that the predictive power of their Google Redundancy Index was most pronounced during economic turbulence; while it improved the root mean square forecasting error (RMSE) by 8% over a no-change benchmark for 2009–2014, this figure rose to 12% during the Great Recession 2009–2011.

Furthermore, GT data has been proven to reduce errors during, and successfully predict the impact of, the Covid-19 pandemic (Heikkinen, and Heimonen, 2023; Kohns, and Bhattacharjee, 2023). The pandemic prompted the use of novel timelier indicators to better gain an understanding of the size of the fall in economic activity. For example, the Bank of England (2020) published an article showcasing high-frequency indicators of economic activity including: moving averages of flight departures; transport data from the Department for Transport; GT searches for cars, hotel bookings, and theatre tickets; UK retail footfall; and seated diners in restaurants. This publication demonstrates the practical adoption of these high-frequency indicators by major

Figure 1: Cumulative Absolute Errors for Equivalent Time Series Models with and without Google Trends Data



Notes: The top panel shows cumulative absolute one-step-ahead prediction errors for weekly US unemployment claims, and the bottom panel shows the initial claims data (2004-2012). Source: Scott and Varian (2014).

economic institutions during periods of crisis.

Conversely, several studies find that GT data offers little significant improvement to nowcasting models once traditional leading indicators become available (Choi, and Varian, 2012; Kohns, and Bhattacharjee, 2023; McLaren, and Shanbhogue, 2011; Woloszko, 2020). This suggests its main value is not in replacing these indicators, but in providing timely signals for them before they are officially published (Heikkinen, and Heimonen, 2023). Therefore, search data is best viewed as a complement to traditional data, not a substitute, serving as a crucial 'bridge' over the information gaps that exist during publication lags (Ferrara, and Simoni, 2022; Smith, 2016). This established finding directly motivates this study's use of a rolling-window analysis, which is specifically designed to assess how the predictive power of GT evolves, paying particular attention to the key crisis periods identified in our methodology (the Great Recession, Brexit uncertainty, and the Covid-19 pandemic).

3 Methodology

3.1 Explaining the Dataset

The dependent variable in this study is the first-differenced, seasonally adjusted unemployment rate for people over the age of 16 in the UK (available at ONS, 2025c). To account for its autoregressive (AR) characteristics (see Figure A1), three lagged values of the first-differenced unemployment rate are included as predictors (see Table 1). Reporting to only one decimal place may limit the model's ability to detect subtle changes in labour market conditions, particularly during periods of relatively stable unemployment, but the data's official status makes it the most reliable measure of UK unemployment for the analysis period.

To build the nowcasting models, several traditional economic and financial indicators were incorporated, as detailed in Table 1. These variables are chosen for their established use as leading indicators of economic activity. While the unemployment rate typically experiences substantial reporting lags of two–three months, alternative measures such as the Claimant Count, Retail Sales Index, and EPU Index generally have shorter delays of around one month. In contrast, financial market indicators like the FTSE 100 are available nearly in real time. This timing advantage allows the leading indicators to provide more current insights into labour market conditions that have not yet been reflected in the official unemployment statistics.

GT provides normalized search data from a representative sample of queries (Google, 2023b). It applies a three-step procedure: dividing searches by the total within a given geography and time, indexing results on a 0–100 scale with 100 as peak intensity, and sampling to protect privacy while preserving representativeness. Consequently, GT reflects relative rather than absolute search volumes, making it useful for tracking real-time temporal patterns and comparing search intensities across terms or periods.

To ensure a transparent and replicable data collection process, as recommended by Hölzl et al. (2025), we programmatically obtained monthly search data from GT for the period January 2004 to March 2025. Using the `pytrends` Python library. This process was repeated daily over a ten-day period to mitigate sampling noise inherent in GT data. From these collections, we extracted unemployment and employment-related queries in the UK, which included specific search terms, websites, and broader topics as detailed in Table A1.

After cleaning, the dataset contains 48 predictors: 39 GT variables, six traditional indicators, and three autoregressive terms. With a rolling window of 48 observations, this creates a high-dimensional setting where $p = n = 48$, making standard OLS

Table 1: Independent Variables (predictors)

Indicator	Description	Source	Treatment in Model
AR Terms of ΔUnemployment Rate	Three lagged values of the first-differenced unemployment rate included to account for autoregressive dynamics.	Engineered from ONS, 2025c data	Lagged (1–3 months)
Claimant Count	The seasonally adjusted number of people claiming unemployment-related benefits.	ONS, 2025a	Lagged by 1 month
UK EPU Index	An index quantifying economic policy uncertainty based on newspaper coverage of policy-related terms.	Baker et al., 2025	Lagged by 1 month
Retail Sales Index	The monthly percentage change in the total value of retail sales. Seasonally adjusted.	ONS, 2025b	Lagged by 1 month
FTSE 100 Returns	The monthly percentage change in the FTSE 100 closing price, capturing market performance.	Investing.com, 2025	Contemporaneous (no lag)
FTSE 100 Volatility	A 3-month rolling standard deviation of the monthly FTSE 100 returns, measuring market risk and uncertainty.	Engineered from Investing.com, 2025 data	Lagged by 1 month
Google Trends	40 search-based indicators including 26 employment terms, 4 job sites, and 10 related topics capturing real-time job-seeking behaviour.	Google, 2025; Hogue, and DeWilde, 2016	Contemporaneous (no lag)

infeasible and illustrating the "big data" challenge in macroeconomic forecasting. Regularization methods such as Lasso or Elastic Net are therefore required. The cleaned dataset spans 249 monthly observations (01/07/2004–01/03/2025). Due to the rolling-window design, the end date of the final out-of-sample window is adjusted slightly to maintain consistent window lengths.

To aid the analysis, we have also identified key economic periods, such as the Great Recession and the Covid-19 pandemic (see Table A2), in order to test a central hypothesis from the literature: that the predictive value of search data is most pronounced during such turbulent periods. Identifying these event periods allows for a targeted evaluation of model robustness and provides crucial context for interpreting

the evolution of the models' performance over time.

To accurately simulate a real-world forecasting scenario, this study explicitly incorporates the known publication delays of key economic indicators. This is achieved by lagging the Claimant Count, Retail Sales Index, and EPU Index by one month, while financial market indicators such as FTSE 100 returns, FTSE 100 volatility, and Google Trends variables are included contemporaneously. This approach ensures that the model's information set realistically reflects what would have been known at the point of prediction, thereby preventing data leakage from sources with inherent reporting lags.

3.2 Preprocessing and Initial Testing

3.2.1 Seasonal Adjustment of Predictors

In the literature, many papers find seasonal patterns in both predictors and dependent variables. Therefore, these variables are seasonally adjusted either implicitly (Askitas, and Zimmermann, 2009; D'Amuri, 2009; Fondeur, and Karamé, 2013; Kohns, and Bhattacharjee, 2023; Pavlicek, and Kristoufek, 2015) or explicitly (Choi, and Varian, 2012; D'Amuri, and Marcucci, 2017; Ferrara, and Simoni, 2022; McLaren, and Shanbhogue, 2011; Smith, 2016; Woloszko, 2020). Implicit seasonal adjustment relies on the econometric model itself to capture seasonal patterns through components such as seasonal dummy variables, Bayesian structural time series seasonal components, or harmonic terms, while using raw (non-seasonally adjusted) data.

In contrast, explicit seasonal adjustment involves pre-processing both the target variable and predictors using statistical procedures. Common methods used are ARIMA-based seasonal adjustment, Seasonal and Trend decomposition using Loess (STL), or creating annual growth rates. Some studies employ hybrid approaches that combine both implicit and explicit methods. For example, Scott and Varian (2014) use implicit seasonal modelling for the dependent variable, through structural time series components, while explicitly seasonally adjusting GT predictors via STL decomposition.

Since our dependent variable is already seasonally adjusted, we decided to seasonally adjust our predictors using STL decomposition. First proposed by Cleveland et al. (1990), STL is a robust non-parametric method which we implement via Python's `statsmodels` package. For variables that remain at zero or near-zero values (below 1) for more than six consecutive months at the beginning of the sample, we apply a "conditional STL" approach: preserving the original zero values and applying STL decomposition only from when the variable first becomes non-zero, preventing the addition of noise to variables that are naturally zero for extended periods. Furthermore,

event-driven GT variables – such as `Furlough`, `Brexit`, and `Financial Crisis` – are not seasonally adjusted as their fluctuations are understood to be tied to specific events rather than a recurring calendar cycle (see Figure A2).

3.2.2 Stationarity and Cointegration

Following the guidance of Li and Chen (2014), and established practice in macroeconomic econometrics, all data series are transformed to achieve stationarity to avoid spurious regression problems, outlined by Granger and Newbold (1974). Therefore, we employ the Augmented Dickey-Fuller (ADF) test, first proposed by Said and Dickey (1984), with AIC automated lag length selection, using both constant-only and constant-plus-trend specifications to account for different data generating processes across variables.

When variables show conflicting results between specifications at the 5% significance level, we conduct visual inspection to determine the presence of deterministic trends. Visual inspection reveals that, out of the variables with conflicting test results, only `Employment Rights` and `Learn New Skills` showed clear deterministic trends, leading us to adopt the constant-plus-trend specification for these variables and constant-only for all others (see Figure A3).

Based on these results, Table A3 categorises the variables by their order of integration. Of the variables, 14 are I(0) (stationary in levels), 29 are I(1) and require first-differencing, and three are I(2), requiring second-differencing. Among the I(1) variables was the dependent, `Unemployment Rate`, which is then differenced to achieve stationarity. Therefore, instead of predicting the level of unemployment, the model will predict the change in unemployment from one period to the next.

To maintain methodological rigour, we employ a cointegration test for the I(1) variables. This involves an augmented Engle and Granger (1987) cointegration tes between each I(1) variable and `Unemployment Rate` (also being I(1)). We discover `Jobseekers Allowance (term)` to be the only cointegrated variable at the 5% significance level (see Table A4), which is robust across both trend specifications. However, given our methodological framework does not employ an error correction model, we opt to use the `Jobseekers Allowance (topic)` instead of the term, since this is not cointegrated but likely holds similar, if not superior, information.

Beyond stationarity considerations, another potential issue with GT data is a systematic downward drift in search volume indices from 2004 onwards, reflecting the expanding user base of Google rather than genuine declines in interest (Stephens-Davidowitz, and Varian, 2015). Woloszko (2020) suggests addressing this by ex-

tracting the common trend through principal component analysis. In our dataset, however, no such bias was observed and the average gradients were slightly positive (see Figure A4). Moreover, the differencing applied to achieve stationarity would, in any case, remove residual trends, making additional adjustment unnecessary.

3.3 The Models

Before estimating the models, overlapping rolling windows of 48 months are created, with each new window starting six months after the previous one (for the in-sample analysis) and one month after the previous one (for the out-of-sample analysis). For example, in the in-sample analysis, Window 1 spans months 1–48, Window 2 spans months 7–54, Window 3 spans months 13–60, and so on, resulting in 34 equally sized windows.

We apply a two-stage Post-shrinkage (OLS) procedure to the rolling windows. In the first stage, regularization is used purely for variable selection within each rolling window. In the second stage, a standard OLS model is estimated using only the subset of variables selected in the first stage. This approach combines the strength of regularization for variable selection in high-dimensional settings with the desirable properties of OLS for unbiased coefficient estimation.

For each rolling window t, the variable selection is performed by solving a general optimization problem where the matrix of predictors, $\mathbf{Z}_t^{(m)}$, is defined by the model specification m. The general forms of the Lasso (equation 1) and Elastic Net (equation 2) estimations are:

$$\hat{\boldsymbol{\beta}}_t^{(m)} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2T} \left\| \Delta u_t - \mathbf{Z}_t^{(m)} \boldsymbol{\beta} \right\|^2 + \lambda_t \|\boldsymbol{\beta}\|_1 \right\}, \tag{1}$$

$$\hat{\boldsymbol{\beta}}_t^{(m)} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2T} \left\| \Delta u_t - \mathbf{Z}_t^{(m)} \boldsymbol{\beta} \right\|^2 + \alpha_t \left[\rho_t \|\boldsymbol{\beta}\|_1 + \frac{(1-\rho_t)}{2} \|\boldsymbol{\beta}\|_2^2 \right] \right\}, \tag{2}$$

where:

- Δu_t is the first difference of the unemployment rate,
- $T = 48$ is the rolling window length in months,
- λ_t, α_t, and ρ_t are the regularization parameters.

The predictor matrix $\mathbf{Z}_t^{(m)}$ for model m can take one of three nested sets of predictors: (1) `Benchmark`, a simple AR(3) model using only lagged values of the dependent variable ($\mathbf{Z}_t^{(AR)}$); (2) `AR+GT`, the benchmark model augmented with a suite of GT

variables ($\mathbf{Z}_t^{(AR+GT)}$); and (3) `Full` (AR+Econ+GT), which is further augmented with other traditional economic indicators ($\mathbf{Z}_t^{(FULL)}$).

Although the AR(3) benchmark is low-dimensional, the Lasso selection step is applied to it as well to maintain methodological consistency across models and ensure a fair basis of comparison with the richer specifications. This prevents performance differences from being driven by the estimation procedure rather than the predictor set. In practice, the algorithm is expected to retain all three autoregressive terms in nearly every window.

Let S_t be the set of indices for the variables selected in the first stage (i.e., where $\hat{\boldsymbol{\beta}}_t^{(m)} \neq 0$). The second-stage OLS regression is then estimated using only this subset of predictors for each window, $\mathbf{Z}_{t,S_t}^{(m)}$. An OLS regression taking the following form is then estimated:

$$\Delta u_t = \mathbf{Z}_{t,S_t}^{(m)} \boldsymbol{\beta}_{t,S_t}^{(m)} + \boldsymbol{\epsilon}_t. \tag{3}$$

The resulting post-selection OLS coefficient vector, $\hat{\boldsymbol{\beta}}_{t,S_t}^{(m)}$, serves two primary purposes in this study. For the in-sample analysis, the evolution of these coefficients is tracked across each rolling window to assess how the predictive importance of the selected variables changes over different economic regimes. For the out-of-sample analysis, this coefficient vector is used to generate the h-step-ahead point forecasts, which are then used to evaluate the model's nowcasting performance.

3.4 Hyper-parameter Tuning and Evaluation Metrics

3.4.1 In-Sample Analysis

The in-sample analysis is designed primarily to evaluate variable selection stability and goodness-of-fit. Following the rolling window structure, a single set of optimal global regularisation parameters is determined for each of the three model specifications (`Benchmark`, `AR+GT`, and `Full`). These parameters are found by applying a time-series cross-validation procedure to all observations and each model's corresponding variable sets, which identifies the values that minimise the mean squared error. Using these fixed, *globally optimal parameters* ensures that the variable selection process is consistent across all 34 windows, making the evolution of predictor importance directly comparable over time.

Within each 48-month window, the competing models are estimated using these predetermined parameters. Following the two-stage approach, regularization is first used for variable selection, after which a final OLS model is estimated using only the selected subset of variables. In-sample performance metrics – such as R^2, in-sample

root mean squared error (RMSE), and Adjusted-R^2 – are derived exclusively from this final OLS model to ensure the evaluation is based on unbiased coefficient estimates. This process generates a time series of selected variables and in-sample performance metrics, allowing for an analysis of how predictor importance evolves.

3.4.2 Out-of-Sample Analysis

The out-of-sample analysis provides a more rigorous test of the models' real-world predictive capabilities across multiple forecast horizons ($h \in \{1, 3, 6, 12\}$ months). For each horizon h, a model is trained on a fixed 48-month period to generate a single direct forecast for h months into the future. The window is then rolled forward by one month, and the process is repeated to generate a continuous series of out-of-sample forecasts.

In contrast to the in-sample analysis, the regularisation parameters are *adaptively re-optimised* within each 48-month training window. For each window, the optimal hyperparameters (λ, α, ρ) are selected via time-series cross-validation on the most comprehensive predictor set (`Full` model). These window-specific parameters are then applied uniformly across the `Benchmark`, `AR+GT`, and `Full` specifications, ensuring that any performance differences reflect the predictors rather than the regularisation strength. This adaptive approach allows model complexity to respond to prevailing economic conditions. Consistent with the two-stage design, the final h-step-ahead forecast is produced by post-shrinkage OLS using only the variables retained by this procedure.

The performance of these forecasts is then evaluated against the actual realised values of the unemployment rate using metrics such as RMSE and Cumulative Squared Forecast Error (CSFE). Finally, to determine if the performance differences between the models are statistically meaningful, the Diebold and Mariano (1995) test is employed to compare the forecast accuracy of the GT-enhanced models against the traditional benchmark.

To ensure the robustness of these findings, an alternative hyper-parameter tuning strategy was also tested, wherein the optimal parameters for each of the three models were estimated independently within each window using only their own specific set of predictors. The results of this model-specific approach were found to be very similar to the main analysis (see Table A5), with no significant changes to the overall model rankings or conclusions. This indicates that the study's findings are robust and not an artifact of the specific, centralized tuning methodology employed.

4 In-Sample Results

4.1 Descriptive Statistics

As expected, the `Benchmark` model retains almost all of its predictors, selecting on average 2.97 of the 3 autoregressive terms across the 34 windows (there is only one window in which one of the terms is not selected). The `AR+GT` model selects an average of 9.88 variables, while the `Full` model selects 15.03. This stepwise increase indicates that the GT variables contribute explanatory power beyond the AR terms, and that the traditional economic indicators add further value beyond the GT data.

Furthermore, the Elastic Net selection specification tends to select more predictors than Lasso. As shown in Table 2, the `AR+GT` model selects, on average, 2.77 more variables (28.0%) between estimations, while the `Full` model selects 1.94 more (12.9%). This pattern supports the presence of correlated groups in the GT data, consistent with Elastic Net's tendency to select entire groups of correlated predictors (Zou, and Hastie, 2005). In addition, Elastic Net models demonstrate slightly superior or highly comparable performance to Lasso across key metrics such as RMSE and Adjusted-R^2. Accordingly, the remainder of the in-sample analysis in this study is conducted using Elastic Net, ensuring both robust variable selection and stable coefficient estimation overtime.

Table 2: In-Sample Model Performance Summary

Model	**Variables Selected**	**RMSE**	**R-Squared**	**Adj. R-Squared**
Post-Lasso OLS Specification				
1. Benchmark	2.971 (0.169)	0.092 (0.012)	0.197 (0.128)	0.143 (0.128)
2. AR+GT	9.882 (1.859)	0.062 (0.007)	0.614 (0.125)	0.516 (0.135)
3. Full	15.029 (1.740)	0.050 (0.008)	0.745 (0.104)	0.628 (0.140)
Post-Elastic Net OLS Specification				
4. AR+GT	12.647 (1.998)	0.059 (0.007)	0.661 (0.107)	0.540 (0.132)
5. Full	16.971 (2.022)	0.049 (0.008)	0.760 (0.105)	0.624 (0.147)

Notes: The table displays the average of key performance metrics across all rolling windows, with standard deviations in parentheses. Models were estimated using the Python package `scikit-learn`, a library for machine learning and predictive data analysis (Pedregosa et al., 2011).

Looking at Table 2, both the AR+GT and Full models provide a superior in-sample fit to the Benchmark, as shown by higher R^2/Adjusted-R^2 and lower RMSE. This indicates that GT models better capture the data, though this does not guarantee stronger out-of-sample performance. The Full model further outperforms AR+GT, with an Adjusted-R^2 of 0.624 versus 0.540, reflecting improved fit even after accounting for greater complexity.

The risk of overfitting is evident from the gap between R^2 and Adjusted-R^2. For the AR+GT model, these fall from 0.661 to 0.540, and for the Full model from 0.760 to 0.624. Such declines highlight the penalty from many predictors, suggesting both models fit in-sample noise as well as the underlying signal.

4.2 Top Selected Variables

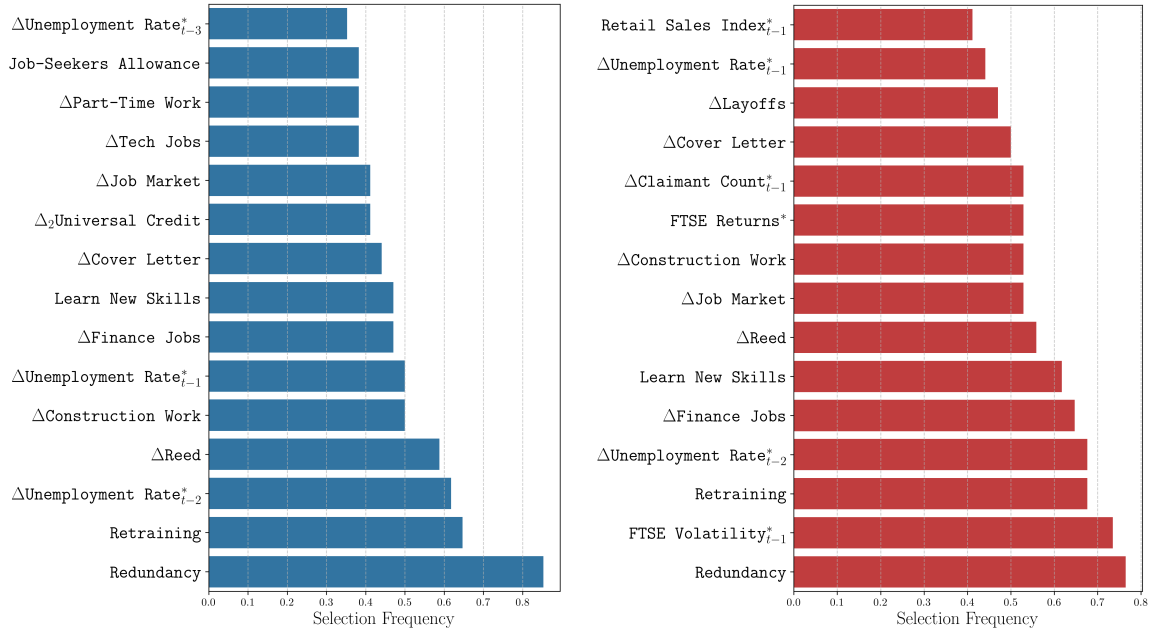
The GT variable Redundancy emerges as the most important predictor, selected in approximately 85% of windows for the AR+GT model and 76% for the Full model. The AR terms are key predictors in both models, demonstrating strong persistence in the data. Specifically, the second autoregressive term, ΔUnemployment Rate$_{t-2}$, alongside the GT search term Retraining, are consistently ranked highly.

The first autoregressive lag appears among the top 15 variables in both models, whereas the third autoregressive lag only makes the top 15 in the AR+GT model and is absent in the Full model. This indicates that the third lag carries less predictive value and, in the Full model, its role is effectively replaced by the broader set of traditional economic indicators.

The addition of traditional variables in the Full model reshapes the set of top predictors. Notably, FTSE Volatility$_{t-1}$ becomes the most important traditional indicator and the second-most selected variable overall, appearing in 73% of windows. These leading indicators reduce the relative importance of many GT variables, reflecting overlapping information. This is most evident for Redundancy, whose selection frequency drops by 11 percentage points – from 88% in the AR+GT model to 77% in the Full model.

Conversely, some GT variables are selected more frequently once traditional indicators are introduced. For example, ΔFinance Jobs rises from 47% to 65%, and Learn New Skills from 47% to 62%. This likely reflects the role of traditional variables as controls for overall economic conditions: once broad market movements are captured (e.g., via FTSE Volatility$_{t-1}$), the behavioural signals in GT terms become more distinct and valuable. This supports the view that GT variables complement, rather than substitute, traditional indicators by bridging specific information gaps.

Figure 2: Top 15 Most Frequently Selected Variables by Model



(a) `AR+GT Model` (b) `Full Model`

Notes: The top 15 variables for the `AR+GT` and `Full` models are ranked by selection frequency across 34 rolling windows in the Elastic Net estimations. Plots generated with `matplotlib`, a Python package for creating publication-quality figures (Hunter, 2007).

4.3 Coefficient Evolution

Figure 3 shows the coefficient evolution for the eight most frequently selected variables in the `Full` model, highlighting their time-varying predictive power. For example, the GT term `Retraining` has a large positive coefficient in the first half of the sample, indicating that increased search interest in retraining predicted rising unemployment, but its effect diminishes after 2016. Similarly, Δ`Claimant Count`$_{t-1}$ is important and stable early on, being selected in all windows up to 2016, before largely disappearing until a single selection in 2024, suggesting the model favours GT terms during this period. Its positive coefficient aligns with expectations, as higher claimant counts signal higher unemployment.

Several GT variables exhibit pronounced parameter instability, with coefficients switching sign over the sample period, reflecting shifts in their relationship with unemployment. For example, Δ`Finance Jobs` is negative during crises such as the financial crisis and Covid-19 pandemic, but positive during the post-crisis and Brexit uncertainty period. This suggests a shift from signalling recovery in the financial sector to reflecting job losses or economic anxiety. Similarly, `Learn New Skills` reverses from negative pre-2015 to positive during the post-crisis/Brexit period, consistent with searches for upskilling being associated with job transitions rather than labour

market strength.

The first-differenced series of Reed, a popular job-search website, also shows regime-dependent behaviour but is most informative during the Covid-19 pandemic. While its coefficient fluctuates between positive and negative in earlier periods, it exhibits a notably larger positive value during Covid-19, highlighting its predictive power in times of labour market disruption.

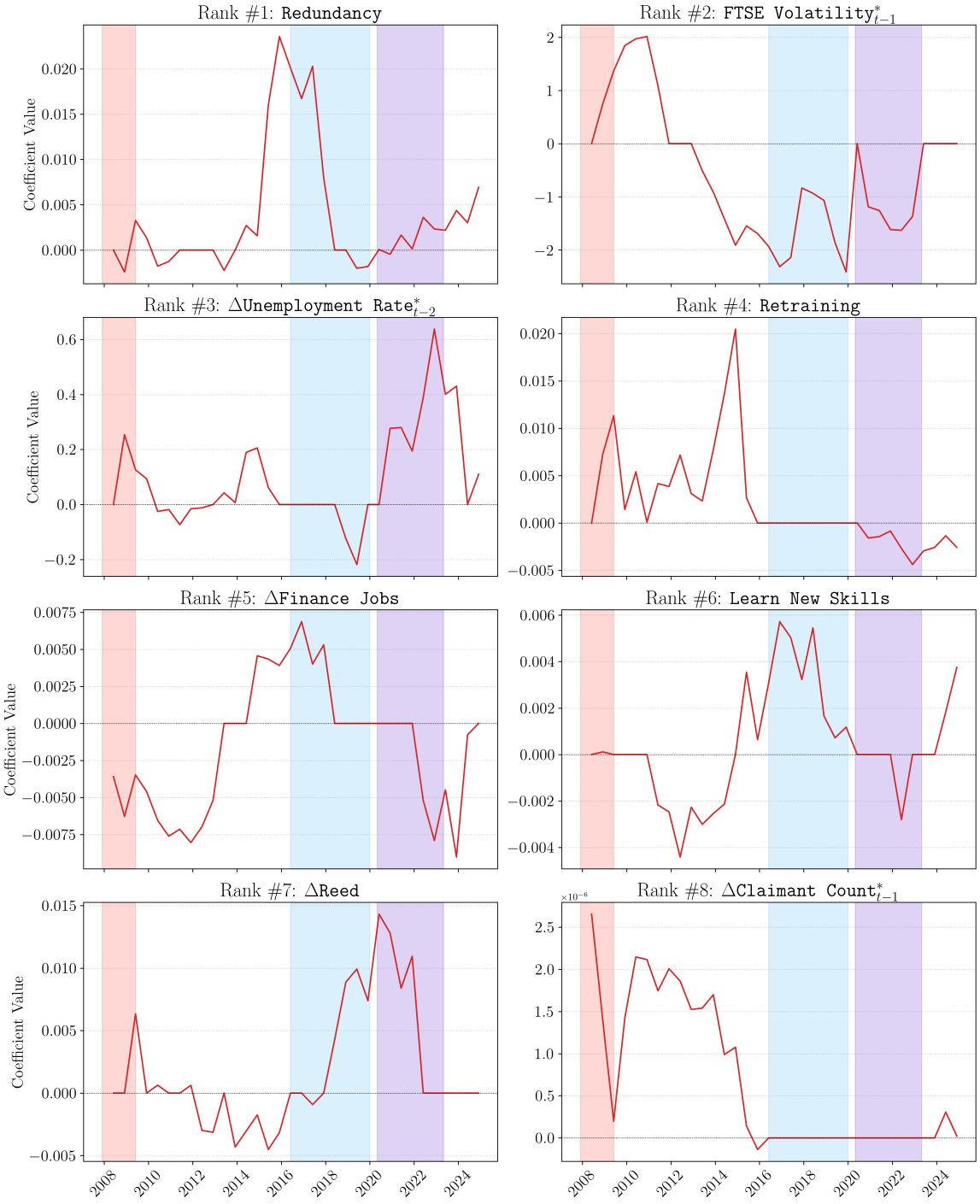
Overall, several variables – including FTSE Volatility$_{t-1}$, ΔFinance Jobs, Learn New Skills, ΔReed, and FTSE Returns – exhibit pronounced sign instability across the sample. As the preceding discussion shows, their coefficients often reverse depending on the prevailing economic regime, making them unreliable as broad, general indicators. Instead, their value lies in capturing dynamics during crises or periods of heightened uncertainty, but in transitional phases their shifting relationships risk introducing more noise than signal, limiting their usefulness for consistent nowcasting.

Several of the top-ranked variables exhibit coefficient signs that are more or less consistent across the sample period. These include Redundancy, ΔClaimant Count$_{t-1}$, Retraining, the first two AR terms, and ΔConstruction Work (see Figure A5). This stability provides strong evidence of a more robust relationship with the unemployment rate, reinforcing their potential as reliable leading indicators.

Figure A6 shows the OLS coefficient evolution for the eight most frequently selected variables in the AR+GT model. Comparing this with Figure 3 suggests that introducing traditional indicators does not substantially change the magnitude or sign of most GT variable coefficients. Some GT terms, such as ΔFinance Jobs and Learn New Skills, gain relative importance, while others (like ΔConstruction) fall out of the top eight, replaced by traditional predictors such as FTSE Volatility$_{t-1}$ and ΔClaimant Count$_{t-1}$.

In summary, the analysis shows that GT data provides substantial, time-varying predictive power for the unemployment rate, and this contribution remains robust even when traditional economic indicators are included. The results point to two distinct groups of predictors. First, a core set of variables – such as Redundancy and Retraining – exhibit stable, consistent relationships, making them reliable leading indicators. Second, a group including Learn New Skills, ΔFinance Jobs, Reed, and several financial indicators displays pronounced coefficient instability. These variables are less useful for general forecasting but become highly informative during periods of economic stress or transition, where their regime-specific signals can capture shifts that more stable indicators miss.

Figure 3: Evolution of Top 8 Post-Elastic Net OLS Full Model Coefficients



Notes: The shaded regions reflect specific economic regimes: light red for the Great Recession (Dec 2007–Jun 2009), light blue for Brexit Uncertainty (Jan 2016–Dec 2019), and light purple for the Covid-19 Pandemic (Jan 2020–Dec 2023). Plots generated with `matplotlib`.

5 Out-of-Sample Results

5.1 Descriptive Statistics

Table 3 shows the different average root mean squared forecast errors (RMSE) for each model at various forecast horizons, as well as their cumulative squared forecast errors (CSFE) over the whole sample. The 1-month ahead column is the best indicator for the nowcasting performance of the models, whereas the other horizons experiment with forecasting and trend detecting ability of internet search data. The table also shows the Diebold-Mariano tests of equal forecast accuracy – indicating whether the differences in forecast accuracy are statistically significant.

Table 3: Out-of-Sample Forecasting Performance by Model and Horizon

Model / Statistic	**1-Month**	**3-Month**	**6-Month**	**12-Month**
Panel A: Root Mean Squared Error (RMSE)				
Benchmark	0.1100	0.1108	0.1132	0.1087
AR+GT Model				
Lasso	0.1096	0.1138	0.1149	0.1144
Elastic Net	0.1088	0.1146	0.1159	0.1151
Full Model				
Lasso	0.1087	0.1083	0.1117	0.1102
Elastic Net	0.1071	0.1098	0.1122	0.1118
Panel B: Diebold-Mariano Test p-values (vs. Benchmark)				
AR+GT Model				
Lasso	0.910	0.267	0.625	0.085*
Elastic Net	0.744	0.175	0.427	0.062*
Full Model				
Lasso	0.785	0.358	0.704	0.663
Elastic Net	0.529	0.731	0.801	0.376
Panel C: Cumulative Squared Forecast Error (CSFE)				
Benchmark	2.1189	2.1283	2.2530	2.0249
AR+GT Model				
Lasso	2.0573	2.1752	2.2852	2.1521
Elastic Net	2.0292	2.1932	2.3082	2.1531
Full Model				
Lasso	2.0177	2.0579	2.2216	2.1268
Elastic Net	1.9729	2.0941	2.2223	2.1750
Panel D: Best Performing Model (Lowest RMSE)				
Model	Full (*Elastic*)	Full (*Lasso*)	Full (*Lasso*)	Benchmark

Notes: Models were again estimated using the Python package `scikit-learn`. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

The `Full` model, particularly when paired with Elastic Net selection, demonstrates

the strongest nowcasting performance. It achieves an average RMSE of 0.1071, a 2.6% improvement compared to the `Benchmark` model, and shows a more substantial 6.9% improvement in terms of CSFE. While this trend of the `Full` model consistently outperforming in the short and medium term holds, the Diebold-Mariano test confirms that these gains are not statistically significant at any forecast horizon.

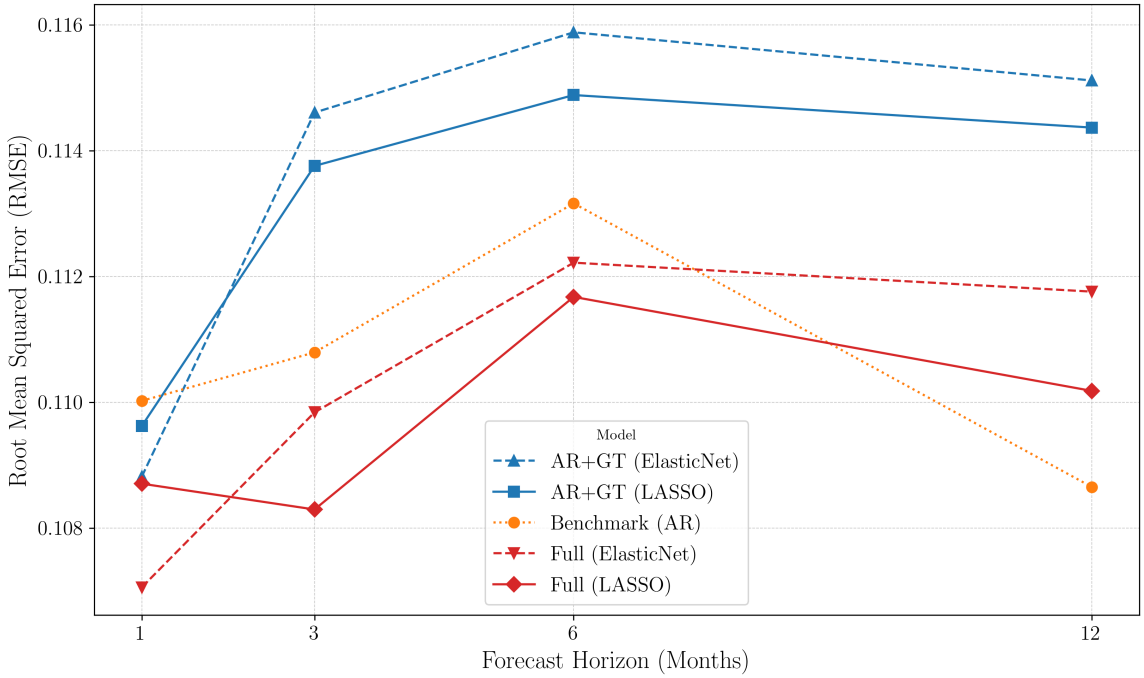
The choice between Lasso and Elastic Net selection reveals a bias-variance trade-off that is dependent on the forecast horizon. For longer-term forecasts (3 to 12 months), the more parsimonious Lasso models consistently outperform their Elastic Net counterparts in both RMSE and CSFE. This finding implies that Elastic Net's tendency to retain more correlated variables leads to overfitting on the training data, degrading its out-of-sample performance as the horizon extends. In contrast, for immediate 1-month-ahead forecasts, Elastic Net's more complex specification is superior, indicating it captures relevant short-term signals that the sparser Lasso models discard.

The relative performance of the `AR+GT` model compared to the `Benchmark` is highly horizon-dependent, reflecting a trade-off between short-term gains and longer-term deterioration. At the 1-month horizon, the `AR+GT` models deliver modest improvements – for example, the Elastic Net variant reduces CSFE by 4.2%. However, these gains are not statistically significant according to the Diebold-Mariano test. Beyond this point, performance declines sharply: as shown in Figure 4, the `AR+GT` model's RMSE rises above that of the `Benchmark` from the 3-month horizon onward. This suggests that GT data alone offers limited predictive value for medium or long-term forecasts.

Further analysis of forecast errors across horizons reveals an irregular pattern in predictability. As shown in Figure 4, errors peak at the 6-month horizon – rather than at 12 months as might be expected – with all models showing a distinct rise in RMSE before stabilising or improving at the one-year mark. Perhaps the most striking finding is the sharp reversal in relative performance at this 12-month horizon. While the more complex `Full` model holds an advantage at shorter horizons, the simple `Benchmark` model emerges as the most accurate predictor after one year, with its RMSE falling significantly below all other competitors. This shift is not driven by a substantial deterioration in the complex models, but rather by a marked improvement in the `Benchmark`'s accuracy.

These results suggest that the predictive signal from both GT and traditional economic variables decays quickly. At a 12-month horizon, the additional predictors contribute more noise than signal, leading the more complex models to underperform. By contrast, the parsimonious `Benchmark` model effectively captures the strong

Figure 4: Model RMSE verses Forecast Horizon



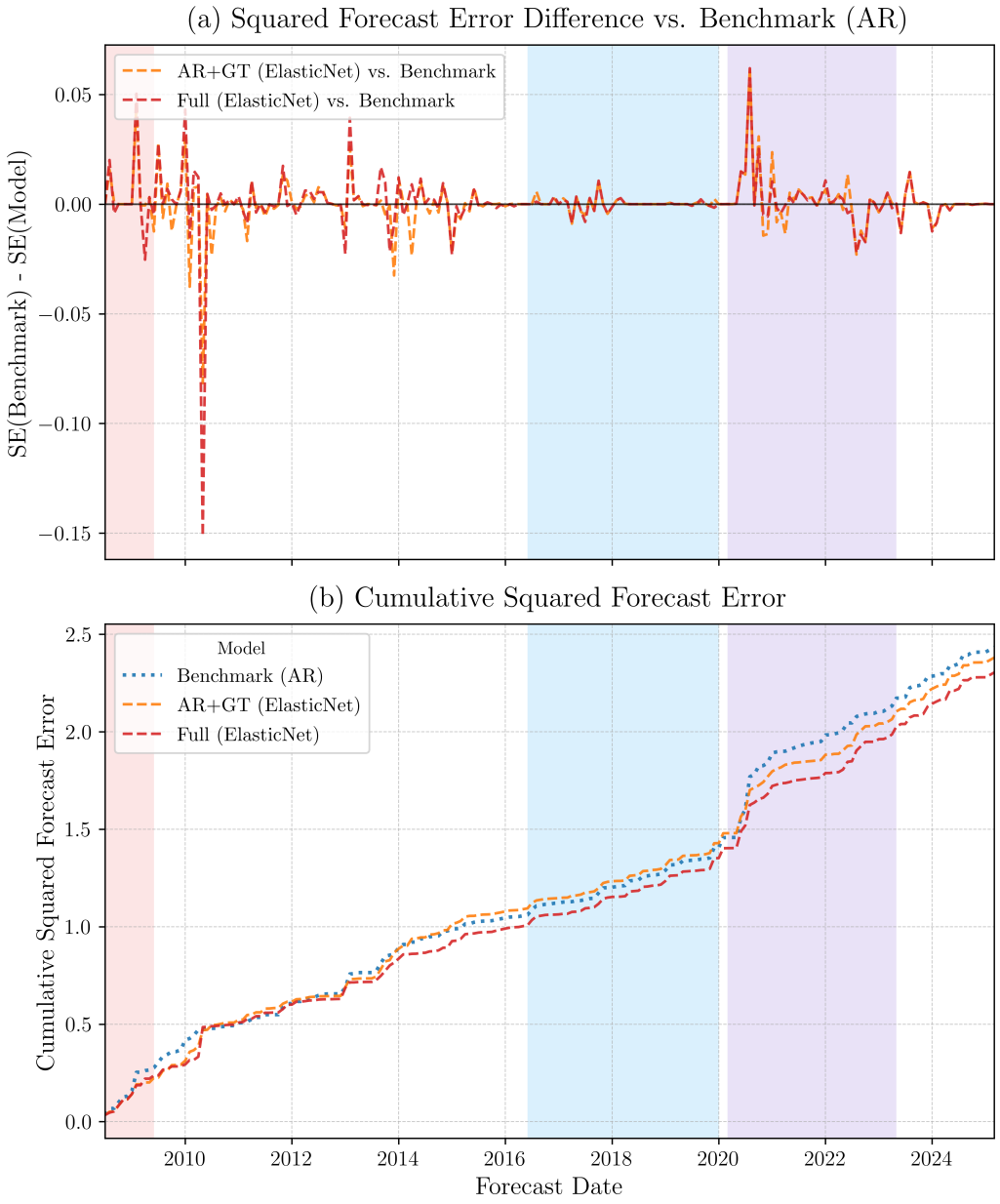
mean-reverting nature of the unemployment rate. This highlights that as horizons lengthen and uncertainty grows, simpler models can often provide more robust and reliable forecasts than complex alternatives.

5.2 When is Google Trends Effective?

Figure 5 illustrates the relative performance of the nowcasting (1-month ahead forecasting) results for each model. Panel (a) displays the difference in squared forecast error (SE) between the `Benchmark` model and GT models, where a positive reading indicates that the `Benchmark` model's error was larger, meaning the `AR+GT` or `Full` model performed better in that period. Panel (b) tracks the cumulative squared forecast error over time, where a model with a flatter line is accumulating less error and is therefore more consistently accurate.

At the start of the period, during the 2008 financial crisis, our GT models perform well, exhibiting various positive spikes in Panel (a) and remaining below the benchmark in cumulative errors in Panel (b). However, around 2010, during the transition from the Great Recession to post-crisis, there are multiple large negative spikes in Panel (a) for both the `AR+GT` and `Full` models. This indicates that, in the period between regimes, the GT model is struggling to adapt to the changing economic conditions.

Figure 5: Out-of-Sample Performance for One-Month Ahead Forecasts



This period of underperformance is likely a direct consequence of the rolling window methodology employed in the analysis. For forecasts made in 2010 and 2011, the 48-month training window is still heavily populated by data from the Great Recession. During this crisis period, the model correctly learns that certain crisis-related variables are powerful predictors of a deteriorating labour market. However, as the economy transitions into a recovery phase, these learned relationships become unstable and even misleading.

Following the period of significant underperformance around 2010, the initial advantage gained by the GT-augmented models during the Great Recession is entirely

eroded. As shown in Panel (b), the CSFE paths of all models converge, and from approximately mid-2010 to 2014, they perform almost identically with their cumulative error lines overlapping.

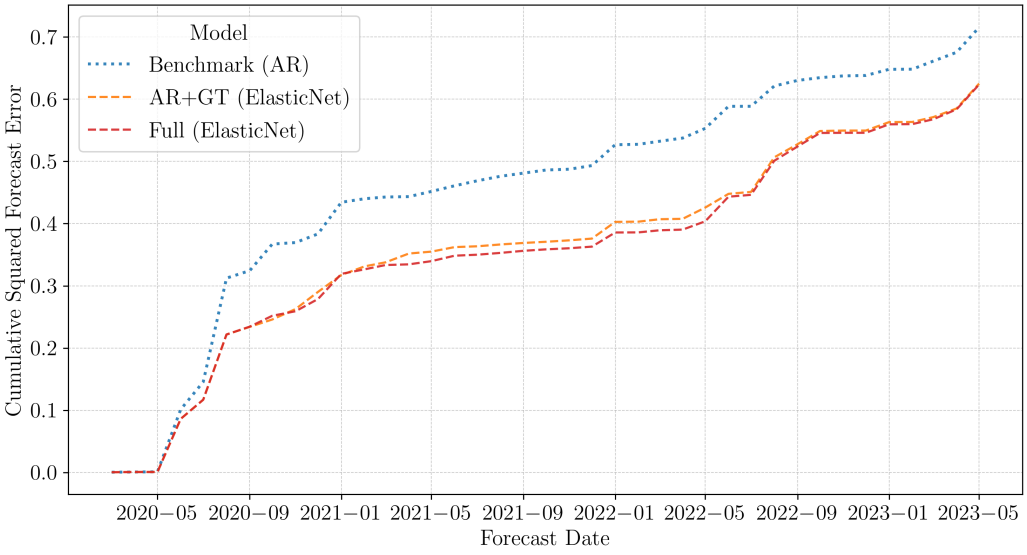
A minor divergence occurs post-2014, where the `Full` model establishes and maintains a consistent, albeit small, advantage over the other models. Critically, the gap between the CSFE lines in Panel (b) remains stable and does not widen between 2014 and the onset of the Covid-19 pandemic. This indicates that, on average, the models had equal forecast accuracy during this multi-year period; the `Full` model was merely preserving an earlier-gained lead. This is corroborated by Panel (a), where the squared error difference fluctuates symmetrically around the zero line throughout this period of relative economic calm.

During the final five years of the sample a new pattern emerges. First, all of the models experience a steep increase in CSFE, with the `Benchmark` being hit the hardest. Looking at Panel (a), at the start of the highlighted Covid period, the `AR+GT` and `Full` models outperform the `Benchmark` in several windows at the start of the period – shown by the large positive spikes from 2020 through to 2022. This creates a gap between the models' CSFEs, with the `Benchmark` having the largest, followed by the `AR+GT`, and then the `Full` model.

Looking at this period more closely in Figure 6, which isolates model performance exclusively during the Covid-19 pandemic, reveals that the `AR+GT` and `Full` models performed almost identically for the whole of this period, as evidenced by their nearly entirely overlapping CSFE paths. This implies that during this period of high volatility, the traditional economic indicators included in the `Full` model provided little to no additional predictive power beyond the GT variables. This finding offers compelling evidence for the utility of search data during turbulent periods; its real-time nature allows the more parsimonious `AR+GT` model to match the performance of the most comprehensive model precisely when traditional economic relationships may be breaking down.

A similar divergence in model performance is observed during the Great Recession period (2007-2009), as illustrated in Figure A7. Consistent with the findings from the Covid-19 crisis, the `Full` and `AR+GT` models again exhibit nearly identical performance, with their CSFE paths closely tracking one another. In contrast, the `Benchmark` model underperforms, as its cumulative error gradually separates and rises above that of the GT-augmented models throughout this period.

Figure 6: Cumulative Squared Forecast Error during Covid-19 Pandemic (1-Month Horizon)



5.3 Do these findings hold for non-linear methods?

To assess whether a non-linear functional form can outperform the linear models, we conduct a robustness check using a Random Forest model. This serves as a check to see if the lack of statistically significant improvements from the linear models over the `Benchmark` is due to a misspecification of the functional form. Using the eight most selected variables identified in the coefficient analysis from the `Full` model, shown in Figure 3, we run a Random Forest model forecasting the same one-month ahead rolling windows as before (in Table 3).

Table 4: Random Forest Out-of-Sample Nowcasting Performance

Metric	**Random Forest**	**Benchmark**
Panel A: Performance Metrics		
Root Mean Squared Error (RMSE)	0.1028	0.1100
Mean Absolute Error (MAE)	0.0808	0.0864
R-squared (R^2)	0.2200	0.1074
Panel B: Diebold-Mariano Test vs. Benchmark		
DM Statistic	-2.1206	
p-value	0.0340*	

Notes: Random Forest was implemented with `scikit-learn`. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$.

Table 4 shows that this model outperforms the linear models, with an RMSE improvement of 4% over the best-performing linear model. Furthermore, this is statistically significant over the `Benchmark`, suggesting that the non-linear approach captures relationships that the linear models could not. This provides a promising result and

highlights the need for further research into combining non-linear methods with GT data for macroeconomic forecasting.

6 Conclusions

The in-sample results initially suggest that the inclusion of GT and traditional economic variables provides substantial explanatory power, as evidenced by the significantly higher R^2 statistics and lower RMSE of the `AR+GT` and `Full` models. However, the large discrepancy between the standard R^2 and Adjusted-R^2 statistics signals the critical issue of overfitting. The out-of-sample analysis confirms this, demonstrating that the impressive in-sample fit does not translate into statistically significant out-of-sample predictive power. This highlights the primary challenge of using high-dimensional data in forecasting: the risk of modelling in-sample noise rather than the true underlying signal.

The in-sample coefficient evolution plots reveal significant time-varying relationships and parameter instability, which helps to explain the models' inconsistent out-of-sample performance. The analysis identifies two distinct types of predictors: stable indicators like `Redundancy`, Δ`Claimant Count`$_{t-1}$ and the autoregressive terms, and unstable, 'event-driven' indicators like `Learn New Skills` and financial metrics. Nowcasting success during the Great Recession and Covid-19 pandemic was driven by the activation of these volatile predictors. However, this strength became a liability during post-crisis recoveries, as models trained on crisis-era data struggled to adapt to the new regime. This asymmetry – better at predicting crises than recoveries – highlights the challenges of using a fixed rolling-window approach amid structural breaks.

The out-of-sample analysis reveals three main insights. The `Full` model generally provides the most accurate forecasts, but its advantage over the simple `Benchmark` is marginal and statistically insignificant, except when using Random Forest at the one-month horizon. GT-augmented models are useful primarily for very short-term forecasts and during volatile periods, such as the Great Recession and Covid-19 pandemic, consistent with evidence that search data excel at nowcasting turning points rather than long-term trends (Choi, and Varian, 2012; Koop, and Onorante, 2019; Smith, 2016). Over longer horizons, the `Benchmark` model proves most robust, as the predictive signal from high-frequency indicators decays quickly, supporting the view that simpler models often outperform complex approaches in unemployment forecasting (Katris, 2020).

Despite these challenges, the findings provide some evidence for the unique value

of GT data during periods of unprecedented economic shock. During the Covid-19 pandemic, the performance of the more parsimonious `AR+GT` model was nearly identical to that of the `Full` model. This implies that the traditional economic indicators added little value beyond the real-time information provided by search data. This supports the consensus in the literature that GT data is especially useful for nowcasting when the relationships captured by traditional indicators may lag or break down entirely (Kohns, and Bhattacharjee, 2023; Scott, and Varian, 2014; Smith, 2016).

Table A6 shows the unemployment data has relatively low month-to-month volatility for extended periods. For the full sample, the standard deviation of monthly changes in the unemployment rate is 0.1103. However, during the Post-Recession Recovery (2009–2016) and Brexit Uncertainty (2016–2020) periods, volatility was even lower, at 0.0945 and 0.0912 respectively. In such stable environments, where changes are minimal, it is naturally difficult for any model to meaningfully outperform a simple benchmark. By contrast, during periods of heightened uncertainty – the Great Recession (0.1170) and COVID-19 pandemic (0.1256) – models incorporating GT and traditional indicators provide the most predictive value. This pattern highlights that the effectiveness of forecasting models is strongly tied to the volatility of the target variable.

Several limitations should be considered when interpreting these findings. First, the selection of GT predictors, while extensive, is not exhaustive, and the results are contingent on this specific set of variables. Second, this study primarily employs a linear framework (post-shrinkage OLS). The relationship between search behaviour and unemployment may be non-linear, and more complex machine learning approaches could potentially yield different results. Third, the use of a fixed 48-month rolling window, while a standard technique, has shown limitations in adapting to regime changes, as evidenced by the poor performance during the post-crisis recovery. Alternative specifications, such as an expanding window or different window lengths, could alter the findings. Finally, the results are specific to the UK; as noted by Pavlicek and Kristoufek (2015), the utility of search data for unemployment forecasting is highly country-specific and may not be generalisable.

Based on the study's conclusions, several avenues for future research could yield valuable insights. First, exploring non-linear machine learning models could better capture the complex, time-varying relationships that a linear framework may miss. Second, to address the limitations of the fixed rolling-window, more adaptive frameworks could be developed; for instance, Markov-switching models could formally identify different economic regimes, while a Dynamic Model Switching (DMS) approach

could use stable predictors during calm periods and activate event-driven ones during crises. Finally, conducting a cross-country comparative analysis would be essential to test the external validity and generalizability of this study's key findings across different economies.

Reference List

Ahrens, A, Aitken, C, Ditzen, J, Ersoy, E, Kohns, D, and Schaffer, M (2021). A Theory-Based Lasso for Time-Series Data. Springer, pp. 3–36. DOI: 10.1007/978-3-030-48853-6_1.

Askitas, N, and Zimmermann, KF (2009). "Google econometrics and unemployment forecasting". *Applied Economics Quarterly* 55.2, pp. 107–120.

Audrino, F, and Knaus, SD (2016). "Lassoing the HAR Model: A Model Selection Perspective on Realized Volatility Dynamics". *Econometric Reviews* 35.8-10, pp. 1485–1521.

Bai, J, and Ng, S (2008). "Forecasting economic time series using targeted predictors". *Journal of Econometrics* 146.2, pp. 304–317. DOI: 10.1016/j.jeconom.2008.08.010.

Baker, SR, Bloom, N, and Davis, SJ (2016). "Measuring Economic Policy Uncertainty". *The Quarterly Journal of Economics* 131.4, pp. 1593–1636. DOI: 10.1093/qje/qjw024.

Baker, SR, Bloom, N, and Davis, SJ (2025). *UK Monthly Index*. Economic Policy Uncertainty. Available from: https://www.policyuncertainty.com/uk_monthly.html (Accessed: 4 June 2025).

Bank of England (2020). *How are we monitoring the economy during the Covid-19 pandemic?* Available from: https://www.bankofengland.co.uk/bank-overground/2020/how-are-we-monitoring-the-economy-during-the-covid-19-pandemic (Accessed: 11 June 2025).

BBC News (2016). *EU referendum timeline: Countdown to the vote*. BBC News Online. British Broadcasting Corporation. Available from: https://www.bbc.co.uk/news/uk-politics-33141819 (Accessed: 11 July 2025).

Belloni, A, and Chernozhukov, V (2013). "Least squares after model selection in high-dimensional sparse models". *Bernoulli* 19.2, pp. 521–547.

Centers for Disease Control and Prevention (2023). *CDC Museum COVID-19 Timeline*. Available from: https://www.cdc.gov/museum/timeline/covid19.html (Accessed: 11 July 2025).

Choi, H, and Varian, H (2012). "Predicting the Present with Google Trends". *Economic Record* 88, pp. 2–9. DOI: 10.1111/j.1475-4932.2012.00809.x.

Cleveland, RB, Cleveland, WS, McRae, JE, and Terpenning, I (1990). "STL: A seasonal-trend decomposition procedure based on loess". *Journal of Official Statistics* 6.1, pp. 3–73.

D'Amuri, F (2009). *Predicting unemployment in short samples with internet job search query data*. MPRA Paper 18403. Munich, Germany: University Library of Munich. Available from: https://ideas.repec.org/p/pra/mprapa/18403.html (Accessed: 10 June 2025).

D'Amuri, F, and Marcucci, J (2017). "The predictive power of Google searches in forecasting US unemployment". *International Journal of Forecasting* 33.4, pp. 801–816. DOI: 10.1016/j.ijforecast.2017.03.004.

Diebold, F, and Mariano, R (1995). "Comparing Predictive Accuracy". *Journal of Business & Economic Statistics* 13.3, pp. 253–63. Available from: https://EconPapers.repec.org/RePEc:bes:jnlbes:v:13:y:1995:i:3:p:253-63 (Accessed: 26 June 2025).

Engle, RF, and Granger, CWJ (1987). "Co-integration and error correction: Representation, estimation, and testing". *Econometrica* 55.2, pp. 251–276. DOI: 10.2307/1913236.

Ettredge, M, Gerdes, J, and Karuga, G (2005). "Using web-based search data to predict macroeconomic statistics". *Communications of the ACM* 48.11, pp. 87–92. DOI: 10.1145/1096000.1096010.

Faberman, RJ, and Kudlyak, M (2019). "The Intensity of Job Search and Search Duration". *American Economic Journal: Macroeconomics* 11.3, pp. 327–357. DOI: 10.1257/mac.20170315.

Federal Reserve History (2013). *The Great Recession*. December 2007-June 2009. Federal Reserve History. Available from: https://www.federalreservehistory.org/essays/great-recession-of-200709 (Accessed: 11 July 2025).

Ferrara, L, and Simoni, A (2022). "When are Google Data Useful to Nowcast GDP? An Approach via Preselection and Shrinkage". *Journal of Business & Economic Statistics* 41.4, pp. 1188–1202. DOI: 10.1080/07350015.2022.2116025.

Fondeur, Y, and Karamé, F (2013). "Can Google data help predict French youth unemployment?" *Economic Modelling* 30, pp. 117–125.

Giannone, D, Reichlin, L, and Small, D (2008). "Nowcasting: The real-time informational content of macroeconomic data". *Journal of Monetary Economics* 55.4, pp. 665–676. DOI: 10.1016/j.jmoneco.2008.05.010.

Google (2023a). *Compare Trends search terms. Google Trends Help*. Available from: https://support.google.com/trends/answer/4359550?hl=en (Accessed: 29 May 2025).

Google (2023b). *FAQ about Google Trends data*. Google Trends Help Center. Available from: https://support.google.com/trends/answer/4365533 (Accessed: 29 May 2025).

Google (2025). *Google Trends*. Available from: https://trends.google.com/trends/ (Accessed: 10 May 2025).

Granger, CWJ, and Newbold, P (1974). "Spurious regressions in econometrics". *Journal of Econometrics* 2.2, pp. 111–120. DOI: 10.1016/0304-4076(74)90034-7.

Heikkinen, J, and Heimonen, K (2023). "Nowcasting the nowcasting - Forecasting ISM Business surveys (PMI and NSI) with weekly Google trends". *Applied Economics* 56.51, pp. 6300–6313. DOI: 10.1080/00036846.2023.2273235.

Hogue, J, and DeWilde, B (2016). *Pytrends: Pseudo API for Google Trends*. Version 4.9.2. Available from: https://github.com/GeneralMills/pytrends (Accessed: 23 July 2025).

Hölzl, J, Keusch, F, and Sajons, C (2025). "The (mis)use of Google Trends data in the social sciences - A systematic review, critique, and recommendations". *Social Science Research* 126, p. 103099. DOI: 10.1016/j.ssresearch.2024.103099.

Hunter, JD (2007). "Matplotlib: A 2D graphics environment". *Computing In Science & Engineering* 9.3, pp. 90–95. DOI: 10.1109/MCSE.2007.55.

Investing.com (2025). *FTSE 100 Historical Data*. Available from: https://uk.investing.com/indices/uk-100-historical-data (Accessed: 4 June 2025).

ISM Report On Business (2025). Institute for Supply Management. Available from: https://www.ismworld.org/supply-management-news-and-reports/reports/ism-pmi-reports/ (Accessed: 13 Aug. 2025).

Katris, C (2019). "Forecasting the unemployment of med counties using time series and neural network models". *Journal of Statistical and Econometric Methods* 8, pp. 37–49.

Katris, C (2020). "Prediction of Unemployment Rates with Time Series and Machine Learning Techniques". *Computational Economics* 55.2, pp. 673–706. DOI: 10.1007/s10614-019-09908-9.

Kohns, D, and Bhattacharjee, A (2023). "Nowcasting growth using Google Trends data: A Bayesian Structural Time Series model". *International Journal of Forecasting* 39.3, pp. 1384–1412. DOI: 10.1016/j.ijforecast.2022.05.002.

Koop, G, and Onorante, L (2019). Macroeconomic Nowcasting Using Google Probabilities. *Topics in Identification, Limited Dependent Variables, Partial Observability, Experimentation, and Flexible Modeling: Part A*. Vol. 40. Advances in Econometrics. Emerald Group Publishing Limited, pp. 17–40.

Kuhn, P, and Mansour, H (2022). *How new job search technologies are affecting the U.S. labor market*. Tech. rep. Washington Center for Equitable Growth. Available from: https://equitablegrowth.org/how-new-job-search-technologies-are-affecting-the-u-s-labor-market/ (Accessed: 2 June 2025).

Lee, JH, Shi, Z, and Gao, Z (2022). "On LASSO for predictive regression". *Journal of Econometrics* 229.2, pp. 322–349. DOI: 10.1016/j.jeconom.2021.02.002.

Li, J, and Chen, W (2014). "Forecasting macroeconomic time series: LASSO-based approaches and their forecast combinations with dynamic factor models". *International Journal of Forecasting* 30.4, pp. 996–1015. DOI: 10.1016/j.ijforecast.2014.03.016.

McLaren, N, and Shanbhogue, R (2011). "Using internet search data as economic indicators". *Bank of England Quarterly Bulletin* 2011 Q2, pp. 134–140. DOI: 10.2139/ssrn.1865276.

Medeiros, MC, Vasconcelos, GFR, Veiga, A, and Zilberman, E (2021). "Forecasting with high-dimensional panel data". *Journal of Business & Economic Statistics*. Forthcoming. arXiv: 2104.03065 [econ.EM].

Mei, Z, and Shi, Z (2024). "On LASSO for High Dimensional Predictive Regression"". *Journal of Econometrics* 242.2, p. 105809. DOI: 10.1016/j.jeconom.2024.105809.

ONS (2025a). *Claimant count QMI*. Available from: https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/methodologies/claimantcountqmi (Accessed: 29 May 2025).

ONS (2025b). *Retail Sales Index Reference Tables*. Available from: https://www.ons.gov.uk/businessindustryandtrade/retailindustry/datasets/retailsalesindexreferencetables (Accessed: 5 June 2025).

ONS (2025c). *Unemployment rate (aged 16 and over, seasonally adjusted)*. Available from: https://www.ons.gov.uk/employmentandlabourmarket/peoplenotinwork/unemployment/timeseries/mgsx/lms (Accessed: 29 May 2025).

Pavlicek, J, and Kristoufek, L (2015). "Nowcasting unemployment rates with google searches: Evidence from the visegrad group countries". *PLoS ONE* 10.5, e0127084. DOI: 10.1371/journal.pone.0127084.

Pedregosa, F, Varoquaux, G, Gramfort, A, Michel, V, Thirion, B, Grisel, O, Blondel, M, Prettenhofer, P, Weiss, R, Dubourg, V, et al. (2011). "Scikit-learn: Machine learning in Python". *Journal of Machine Learning Research* 12, pp. 2825–2830.

Preis, T, Moat, HS, and Stanley, HE (2013). "Quantifying trading behavior in financial markets using Google Trends". *Scientific Reports* 3.1, pp. 1–6. DOI: 10.1038/srep01684.

Said, SE, and Dickey, DA (1984). "Testing for unit roots in autoregressive-moving average models of unknown order". *Biometrika* 71.3, pp. 599–607. DOI: 10.1093/biomet/71.3.599.

Scott, SL, and Varian, HR (2014). "Predicting the present with Bayesian structural time series". *International Journal of Mathematical Modelling and Numerical Optimisation* 5.1-2, pp. 4–23. DOI: 10.1504/IJMMNO.2014.059942.

Seabold, S, and Perktold, J (2010). statsmodels: Econometric and statistical modeling with python. *9th Python in Science Conference*. Austin, TX, pp. 92–96.

Smith, P (2016). "Google's MIDAS Touch: Predicting UK Unemployment with Internet Search Data". *Journal of Forecasting* 35.3, pp. 263–284. DOI: 10.1002/for.2391.

StatCounter (2025). *Search Engine Market Share United Kingdom*. Available from: https://gs.statcounter.com/search-engine-market-share/all/united-kingdom (Accessed: 2 June 2025).

Stephens-Davidowitz, S, and Varian, H (2015). *A Hands-on Guide to Google Data*. Tech. rep. Google, Inc. Available from: https://people.ischool.berkeley.edu/~hal/Papers/2015/primer.pdf.

Tibshirani, R (1996). "Regression shrinkage and selection via the lasso". *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58.1, pp. 267–288. DOI: 10.1111/j.2517-6161.1996.tb02080.x.

Vosen, S, and Schmidt, T (2011). "Forecasting private consumption: survey-based indicators vs. Google trends". *Journal of Forecasting* 30.6, pp. 565–578. DOI: 10.1002/for.1213.

Woloszko, N (2020). *Tracking activity in real time with Google Trends*. Economics Department Working Papers 1634. Paris: OECD. DOI: 10.1787/6b9c7518-en. (Accessed: 10 June 2025).

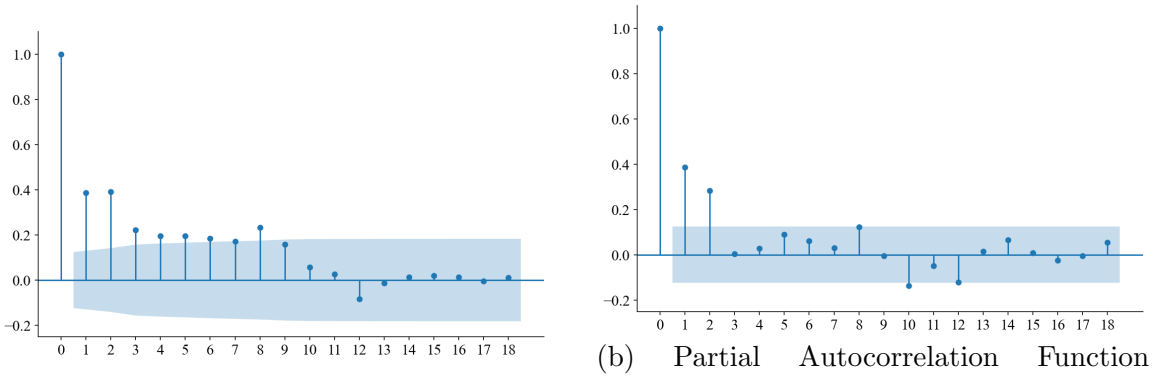
Zou, H, and Hastie, T (2005). "Regularization and Variable Selection Via the Elastic Net". *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67.2, pp. 301–320. DOI: 10.1111/j.1467-9868.2005.00503.x.

Appendix

Replication Package

The replication package for this dissertation, including all code, Python packages, and data used for the analyses, is publicly available at:
https://github.com/Arthur-Grainger/MSc-Dissertation.
This repository contains all scripts, datasets, and documentation necessary to reproduce the results presented in this dissertation.

Figure A1: ACF and PACF plots for the time series data.



(a) Autocorrelation Function (ACF)

(b) Partial Autocorrelation Function (PACF)

Notes: The shaded area represents the 95% confidence interval. Graphs were created using the `statsmodels` (Seabold, and Perktold, 2010) package in Python.

Table A1: Google Trends Search Variables

Category	**Variables**
Search Terms (26)	`Apprenticeships`, `Bankruptcy`, `Birmingham Jobs`, `Construction Work`, `Employment Rights`, `Employment`, `Finance Jobs`, `Free Courses`, `Hospitality Jobs`, `Jobs`, `Job Centre`, `Job Market`, `Jobseekers Allowance`, `Layoffs`, `Learn New Skills`, `London Jobs`, `Manchester Jobs`, `Manufacturing Jobs`, `Part Time Work`, `Recession`, `Recruitment Agencies`, `Redundancy`, `Retail Jobs`, `Retraining`, `Tech Jobs`, `Work From Home Jobs`
Websites (4)	`Indeed`, `CV Library`, `Reed`, `LinkedIn`
Topics (10)	`Brexit`, `Citizens Advice`, `Cover Letter`, `Financial Crisis`, `Furlough`, `Jobseeker's Allowance`, `Remote Work`, `Unemployment`, `Unemployment Benefits`, `Universal Credit`

Figure A2: Event Indicators Plots

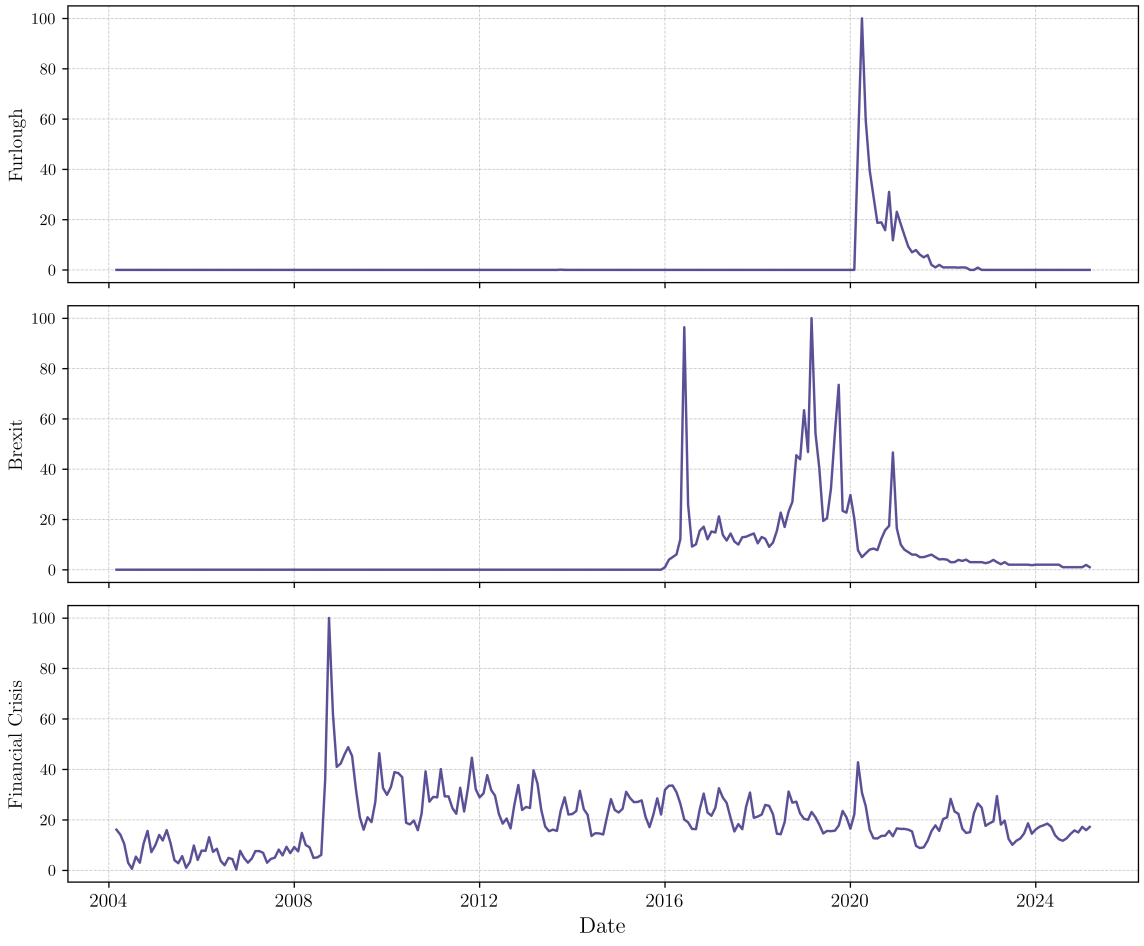


Table A2: Definitions of Core Crisis Periods

Event	**Core Period**	**Source**
Great Recession	Dec 2007 – Jun 2009	(Federal Reserve History, 2013)
Brexit Uncertainty	Jun 2016 – Jan 2020	(BBC News, 2016)
COVID-19 Pandemic	May 2020 – May 2023	(Centers for Disease Control and Prevention, 2023)

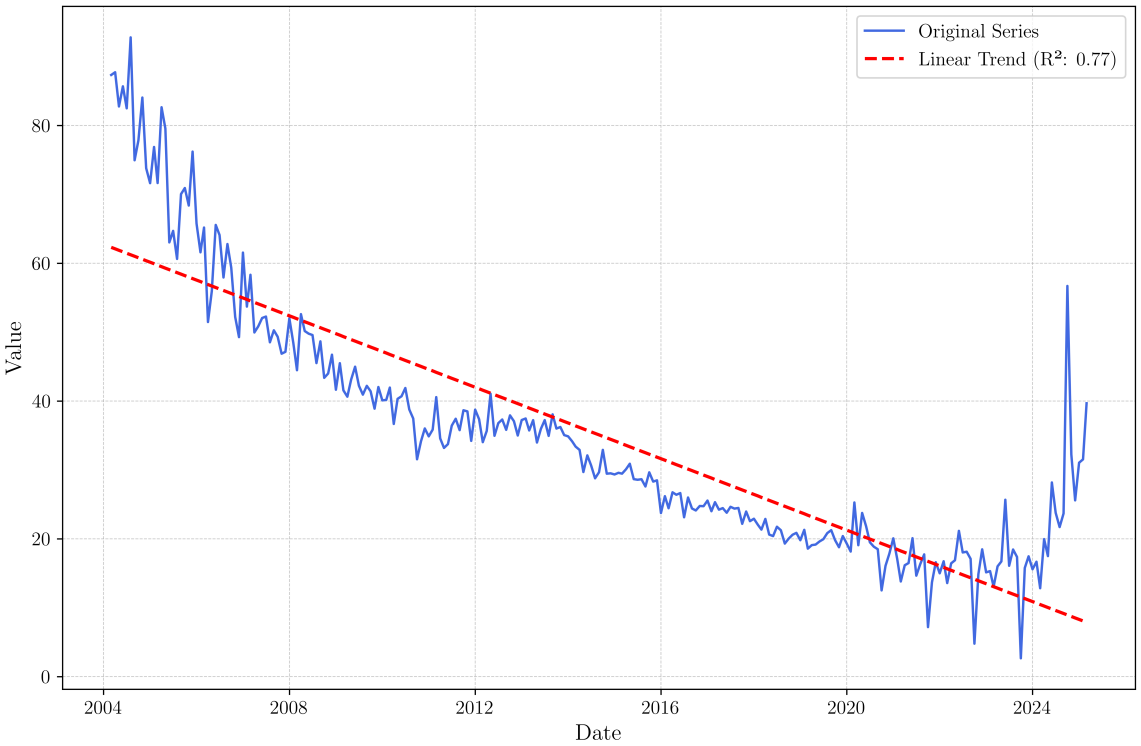
Table A3: Summary of Stationarity Test Results and Integration Orders

Integration Order	**Count**	**Variables**
I(0)	14	`Brexit`, `Employment`, `FTSE Returns`*, `Furlough`, `Job Seekers Allowance`, `Learn New Skills`, `Manufacturing Jobs`, `Recruitment Agencies`, `Redundancy`, `Retraining`, `Retail Sales Index`*, `Unemployment`, `Unemployment Benefits`, `FTSE Vol 3m`*
I(1)	29	`Apprenticeships`, `Bankruptcy`, `Birmingham Jobs`, `Citizens Advice`, `Claimant Count`*, `Construction Work`, `Cover Letter`, `CV Library`, `Employment Rights`, `Finance Jobs`, `Financial Crisis`, `Free Courses`, `Hospitality Jobs`, `Jobseekers Allowance`, `Jobs`, `Job Centre`, `Job Market`, `Layoffs`, `London Jobs`, `Manchester Jobs`, `Part Time Work`, `Recession`, `Reed`, `Remote Work`, `Retail Jobs`, `Tech Jobs`, `UK EPU Index`*, `Unemployment Rate`*, `Work From Home Jobs`
I(2)	3	`Indeed`, `LinkedIn`, `Universal Credit`

Notes: Variables marked with * are traditional economic indicators. Variables without asterisks are GT search terms, topics, or websites. See Table A7, Table A8, and Table A9 for detailed test results.

Figure A3: Trend Inspection Plots

(a) Employment Rights Trend



(b) Learn New Skills Trend

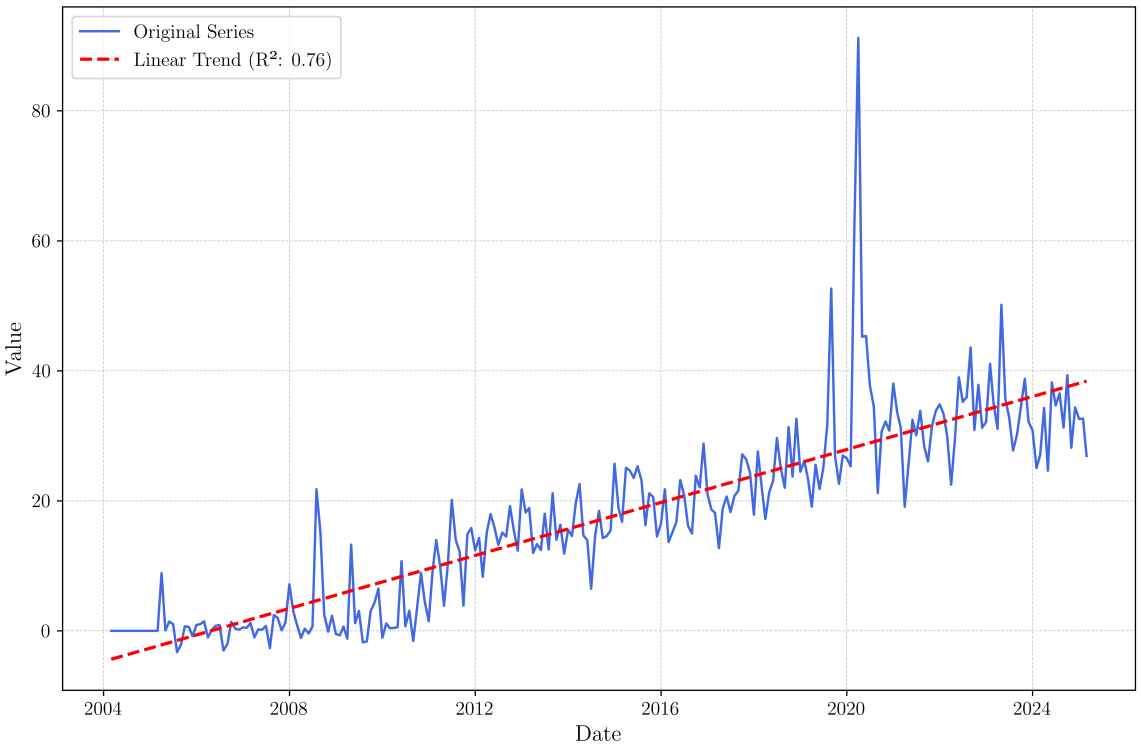


Figure A4: Distribution of linear trends for GT variables

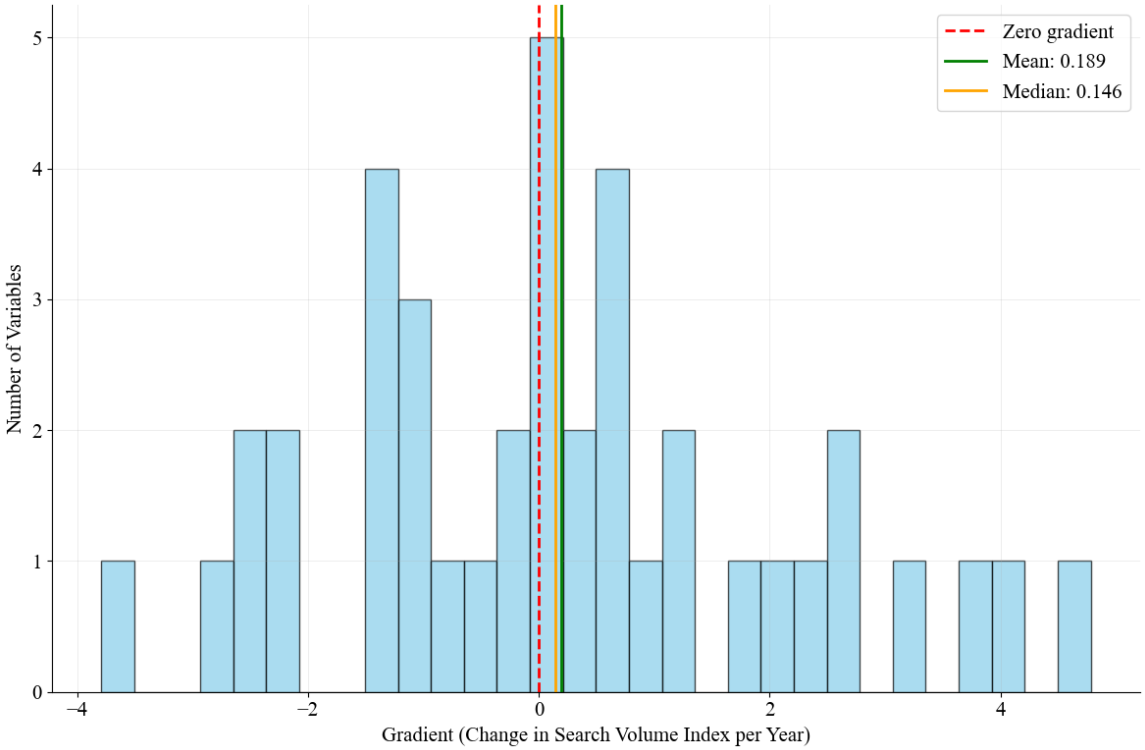


Table A4: Engle-Granger Cointegration Test Results with `unemp_rate` as Target

Variable	EG Statistic	P-value	Cointegrated at 5%
claimant_count	-1.4086	0.7946	No
UK_EPU_Index	-1.5391	0.7459	No
apprenticeships	-1.1942	0.8594	No
bankruptcy	-2.0492	0.5029	No
birmingham_jobs	-1.0415	0.8947	No
citizens_advice	-1.3803	0.8042	No
construction_work	-2.0255	0.5153	No
cover_letter	-1.4271	0.7881	No
cv-library	-1.1491	0.8707	No
employment_rights	-1.4377	0.7843	No
finance_jobs	-1.1180	0.8781	No
Financial_crisis	-1.4211	0.7902	No
free_courses	-2.1180	0.4670	No
hospitality_jobs	-1.2014	0.8575	No
jobseekers_allowance	**-4.1973**	**0.0037**	**Yes**
jobs	-1.4776	0.7697	No
job_centre	-1.9552	0.5518	No
job_market	-1.4444	0.7819	No
layoffs	-1.4979	0.7620	No
london_jobs	-1.7964	0.6315	No
manchester_jobs	-2.1672	0.4414	No
part_time_work	-1.1218	0.8772	No
recession	-1.4294	0.7873	No
reed	-2.0639	0.4952	No
remote_work	-1.8846	0.5878	No
retail_jobs	-2.4014	0.3242	No
tech_jobs	-1.7513	0.6530	No
work_from_home_jobs	-1.8690	0.5956	No

Notes: The Engle-Granger (EG) test for cointegration tests the null hypothesis of no cointegration. The null is rejected if the EG statistic is more negative than the critical value. The 5% and 1% critical values for these tests were -3.3605 and -3.9404, respectively.

Table A5: Out-of-Sample Forecasting Performance by Model and Horizon

Model / Statistic	**1-Month**	**3-Month**	**6-Month**	**12-Month**
Panel A: Root Mean Squared Error (RMSE)				
Benchmark	0.1097	0.1126	0.1165	0.1089
AR+GT Model				
Lasso	0.1113	0.1187	0.1182	0.1202
Elastic Net	0.1095	0.1178	0.1178	0.1213
Full Model				
Lasso	0.1088	0.1083	0.1117	0.1102
Elastic Net	0.1071	0.1098	0.1122	0.1118
Panel B: Diebold-Mariano Test p-values (vs. Benchmark)				
AR+GT Model				
Lasso	0.672	0.306	0.717	0.109
Elastic Net	0.954	0.358	0.780	0.078*
Full Model				
Lasso	0.831	0.164	0.261	0.731
Elastic Net	0.571	0.382	0.297	0.431
Panel C: Cumulative Squared Forecast Error (CSFE)				
Benchmark	2.421	2.520	2.659	2.255
AR+GT Model				
Lasso	2.491	2.802	2.737	2.746
Elastic Net	2.412	2.767	2.718	2.796
Full Model				
Lasso	2.375	2.334	2.444	2.307
Elastic Net	2.304	2.401	2.468	2.373
Panel D: Best Performing Model (Lowest RMSE)				
Model(s)	Full (*Elastic*)	Full (*Lasso*)	Full (*Lasso*)	Benchmark

Table A6: Descriptive Statistics for the First-Differenced Unemployment Rate (Δu_t)

Period	**Start Date**	**End Date**	**Mean**	**Std. Dev.**	**Variance**
Overall Sample	2004-07	2025-03	−0.0008	0.1103	0.0122
Pre-Recession	2004-01	2007-11	0.0098	0.0735	0.0054
Great Recession	2007-12	2009-06	0.1421	0.1170	0.0137
Post-Recession Recovery	2009-07	2016-05	−0.0361	0.0945	0.0089
Brexit Uncertainty	2016-06	2020-02	−0.0178	0.0912	0.0083
COVID-19 & Aftermath	2020-03	2025-03	0.0082	0.1256	0.0158

Figure A5: Evolution of Top 9 to 16 Post-Elastic Net OLS Full Model Coefficients

Figure A6: Evolution of Post-Elastic Net OLS `AR+GT` Model Coefficients

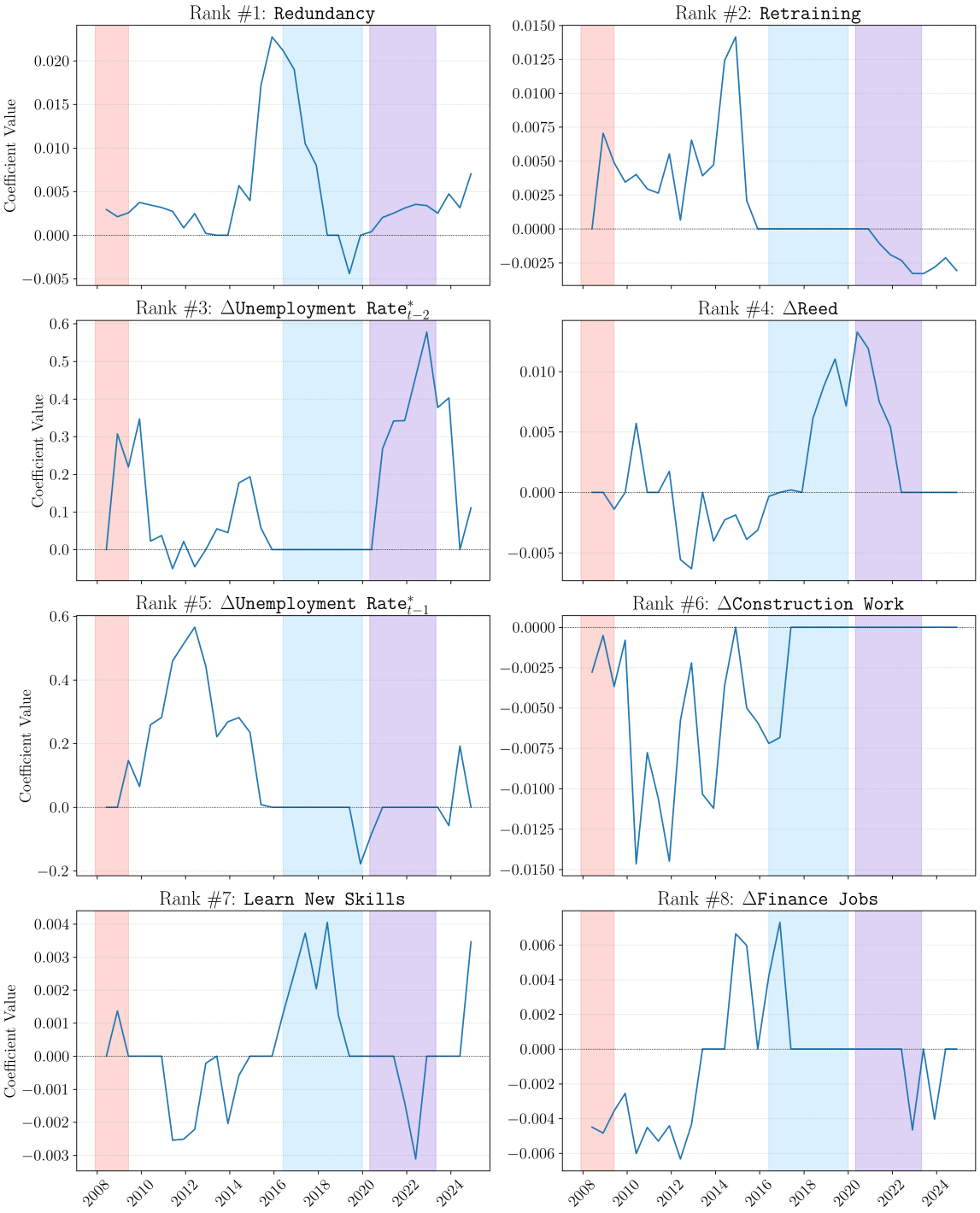


Figure A7: Cumulative Squared Forecast Error during Great Recession (1-Month Horizon)

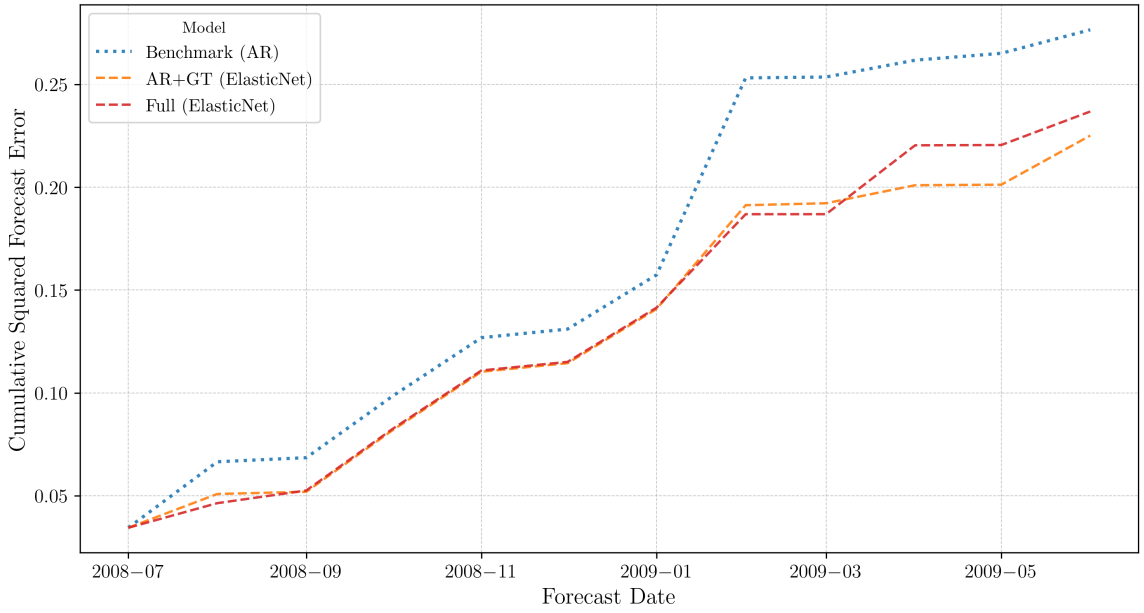


Table A7: Augmented Dickey-Fuller Test for Stationarity: Levels

Variable	Test Type	Lags	ADF Statistic	P-value	Stationary at 5%
claimant_count	Constant	9	-2.6711	0.0792	No
UK_EPU_index	Constant	3	-2.5083	0.1135	No
ftse_returns	Constant	0	-16.3209	0.0000	Yes
retail_sales_index	Constant	8	-7.6049	0.0000	Yes
unemp_rate	Constant	2	-1.2002	0.6735	No
apprenticeships	Constant	3	-1.3619	0.6004	No
bankruptcy	Constant	2	-1.5991	0.4840	No
birmingham_jobs	Constant	5	-2.2870	0.1762	No
brexit	Constant	3	-3.2799	0.0158	Yes
citizens_advice	Constant	16	-1.9800	0.2954	No
construction_work	Constant	11	-2.0703	0.2566	No
cover_letter	Constant	12	-1.7644	0.3983	No
cv-library	Constant	3	-2.0615	0.2603	No
employment_rights	Constant & Trend	6	-0.6526	0.9761	No
employment	Constant	2	-6.2103	0.0000	Yes
finance_jobs	Constant	5	-1.4271	0.5692	No
Financial_crisis	Constant	13	-2.2866	0.1763	No
free_courses	Constant	9	-2.7026	0.0736	No
furlough	Constant	9	-3.3226	0.0139	Yes
hospitality_jobs	Constant	7	-0.0247	0.9564	No
indeed	Constant	13	-1.4983	0.5343	No
job-seekers_Allowance	Constant	6	-2.8829	0.0474	Yes
jobseekers_allowance	Constant	2	-1.2218	0.6642	No
jobs	Constant	15	-1.3574	0.6025	No
job_centre	Constant	15	-0.5899	0.8733	No
job_market	Constant	4	-1.9005	0.3318	No
layoffs	Constant	5	-2.3745	0.1491	No
learn_new_skills	Constant & Trend	6	-3.7045	0.0221	Yes
linkedin	Constant	9	-1.6192	0.4732	No
london_jobs	Constant	3	-1.8046	0.3782	No
manchester_jobs	Constant	2	-1.8724	0.3451	No
manufacturing_jobs	Constant	4	-3.5497	0.0017	Yes
part_time_work	Constant	4	-2.2345	0.1940	No
recession	Constant	12	-2.7741	0.0621	No
recruitment_agencies	Constant	4	-5.5131	0.0000	Yes
redundancy	Constant	3	-4.6977	0.0001	Yes
reed	Constant	2	-0.2490	0.9324	No
remote_work	Constant	2	0.9084	0.9932	No
retail_jobs	Constant	5	-1.3970	0.5837	No
retraining	Constant	0	-13.3279	0.0000	Yes
tech_jobs	Constant	3	0.2587	0.9754	No
unemployment_benefits	Constant	3	-3.2040	0.0198	Yes
unemployment	Constant	3	-2.8875	0.0468	Yes
universal_credit	Constant	9	-0.4819	0.8955	No
work_from_home_jobs	Constant	14	0.3034	0.9775	No
ftse_vol_3m	Constant	9	-4.0134	0.0013	Yes

Table A8: Augmented Dickey-Fuller Test for Stationarity: First Difference

Variable	Test Type	Lags	ADF Statistic	P-value	Stationary at 5%
claimant_count	Constant	2	-6.3446	0.0000	Yes
UK_EPU_index	Constant	2	-13.4433	0.0000	Yes
unemp_rate	Constant	1	-6.4367	0.0000	Yes
apprenticeships	Constant	2	-12.2954	0.0000	Yes
bankruptcy	Constant	1	-14.6433	0.0000	Yes
birmingham_jobs	Constant	4	-10.2658	0.0000	Yes
citizens_advice	Constant	15	-5.0047	0.0000	Yes
construction_work	Constant	10	-5.8486	0.0000	Yes
cover_letter	Constant	11	-3.4293	0.0100	Yes
cv-library	Constant	2	-12.7004	0.0000	Yes
employment_rights	Constant	4	-10.7048	0.0000	Yes
finance_jobs	Constant	4	-10.7187	0.0000	Yes
Financial_crisis	Constant	11	-7.0126	0.0000	Yes
free_courses	Constant	8	-6.4903	0.0000	Yes
hospitality_jobs	Constant	6	-10.9451	0.0000	Yes
indeed	Constant	12	-2.4755	0.1216	No
jobseekers_allowance	Constant	1	-13.9529	0.0000	Yes
jobs	Constant	14	-3.9461	0.0017	Yes
job_centre	Constant	14	-4.8247	0.0000	Yes
job_market	Constant	3	-11.4635	0.0000	Yes
layoffs	Constant	9	-6.5875	0.0000	Yes
linkedin	Constant	16	-2.4059	0.1401	No
london_jobs	Constant	2	-11.7225	0.0000	Yes
manchester_jobs	Constant	1	-15.2321	0.0000	Yes
part_time_work	Constant	3	-11.8913	0.0000	Yes
recession	Constant	11	-6.4244	0.0000	Yes
reed	Constant	1	-17.4278	0.0000	Yes
remote_work	Constant	1	-16.2265	0.0000	Yes
retail_jobs	Constant	4	-9.8982	0.0000	Yes
tech_jobs	Constant	2	-13.4286	0.0000	Yes
universal_credit	Constant	16	-1.9397	0.3136	No
work_from_home_jobs	Constant	13	-4.6539	0.0001	Yes

Table A9: Augmented Dickey-Fuller Test for Stationarity: Second Difference

Variable	Test Type	Lags	ADF Statistic	P-value	Stationary at 5%
indeed	Constant	10	-11.6195	0.0000	Yes
linkedin	Constant	15	-6.9232	0.0000	Yes
universal_credit	Constant	15	-8.2080	0.0000	Yes

Notes: the 5% critical value for tests with a 'Constant' is approximately -2.87, and for tests with 'Constant & Trend' it is approximately -3.43. The exact values can vary slightly with the sample size of each specific test.

Figure A8: Secondary Data Checklist

Project Name: *Shrinkage and Selection in ``Big Data'': Nowcasting the UK Labour Market with Google Trends*

Which organisation /person is providing the data?	Office for National Statistics, Google Trends, Investing.com, policyuncertainty.com.

Are you going to only use secondary data for your project? *(i.e. you will not be collecting any new data)*	Yes
Are you going to take (scrape / mine) data from social media feeds?	No
Are the data publicly accessible / available? *(this includes datasets which are available to subscribers or can be purchased)*	Yes
If the data is not publicly accessible / available, is there an agreement in place for using the data? *(e.g. see agreement next page)*	NA
Has the data been anonymised?	Yes
Is there no risk of re-identification of individuals? *(This could be for instance by combining the data with other data on same groups, or rare, unusual or low number data. In general, this would be if re-identification is possible by "means reasonably likely to be used")*	Yes
Are the data being managed and shared by reputable external organisations or institutions (that comply with all legal requirements concerning data protection)? *(examples are NHS, central government, local authorities, public bodies, for example accessed via https://www.data.gov.uk or https://opendatascience.com/15-open-datasets-for-healthcare or https://www.england.nhs.uk/)*	Yes
Is the proposed research in line with the website / organisation / institution sharing agreement, terms and conditions? *(the terms and conditions that the organisation / person is sharing the data will vary, and may have conditions on sharing [e.g. not for commercial use, must always reference original paper/author])*	Yes
Does consent exist in the original study / data collection to share the data, and is the proposed research in line with the participants original consent? *(if possible, include a blank copy of the original consent form in your dissertation from original study that participants have agreed for their anonymised data to be shared for further research [or some other evidence])*	Yes
Is your data management plan adequate for the data? *(this will be access, storage, processing and preservation/deletion of data)*	Yes
Does the organisation / person sharing the data require evidence of an ethical review for use of the data?	No

Supervisors: Please sign below to state that 1) only secondary data used for the project, 2) no data are being taken from social media feeds, 3) the data being used contain no personal data (is de-identified/ anonymous), 4) has existing consent from participants for sharing and use in line with the research, 5) is able to be used legally in the way you propose (in terms and conditions / agreement of data use), 6) has no risk of re-identification, 7) data management is adequate for the data, and 8) the organisation / person / institute managing and sharing the data complies with data processing regulations and does not require an ethical review for sharing.

Supervisor Name: LUCIANO RISPOLI Signature: Luciano Rispoli Date: 02/09/2025

Student Name: Arthur Grainger Student Level (e.g. UG/ PGT): PGT

Secondary Data Checklist & Usage Agreement v.6 16.2.25 ethics@surrey.ac.uk