Exploring new procedures to deal with time-varying relationships in forecasting: A study in the context of forecasting youth unemployment with Google searches.

Nathan de Bruijn – author
Fons Wijnhoven – 1st supervisor
Robin Effing – 2nd supervisor

n.b.debruijn@student.utwente.nl
S2875616

MSc Business Administration
Version two
September 13, 2023

Abstract

Forecasting has evolved dramatically over time, from humans looking at the sky to predict weather conditions to an era characterized by an increasing imperative to forecast, a surge in data, and advanced computational capabilities. Despite these developments, forecasting methodologies have barely adapted. Today, forecasters continue to rely on procedures for out-of-sample forecasting that were introduced over five decades ago. In this study a crucial limitation that arises with the use of these traditional procedures in today's dynamic world is addressed: their assumption that relationships between phenomena are constant. To address this issue, this study follows a method comparison approach, comparing the traditional procedures against innovated procedures. Results show that the innovated procedures, tailored to the use of big data in forecasting and leveraging increased computational capabilities, are better able to capture time-varying relationships between variables. As a result, forecasts of the youth unemployment rate are up to 44% more accurate. Moreover, when applying the innovated procedures with a simple regression model, forecasts relying on more complex models are outperformed. With the innovated procedures, practitioners are assisted in building big data capability, resulting in improved forecasts and better decision-making making. Additionally, this study makes labour market forecasting an accessible endeavour to all organizations by sharing the algorithm to accurately forecast the youth unemployment rate with publicly available data. Moreover, this study furnishes forecasting literature with a novel perspective, stressing a reconsideration of methodologies before exploiting more complex models in a search for greater forecasting accuracy.

Keywords: Forecasting, forecasting accuracy, forecasting evaluation, time series, rolling window, expanding window, unemployment, google trends, parameter instability

Table of contents

1. Introduction

Forecasting has evolved substantially since humans first looked at the sky to predict weather conditions for hunting. Today, thanks to technological and methodological advancements, anyone can access detailed weather forecasts through their smartphones. Despite advancements over time, the essence of forecasting remains rooted in data analysis, driven by computational capabilities. Approaching 2025, it is expected that data generation will reach an astounding 180 trillion gigabytes (Rydning, 2022). This exponential surge in data generation, coupled with advances in computational power, has expanded the forecaster's toolbox (Petropoulos et al., 2022). However, as the forecaster's toolbox grows, so does our approach to data, shifting from descriptive analytics to predictive analytics (Jeyanthi et al., 2022).

Considering these developments, a variety of organizations stand on the verge of an analytics revolution (Petropoulos et al., 2022). The imperative to forecast is pronounced in today's dynamic environment, where accurate forecasts empower businesses with actionable insights and improved decision-making (Li et al., 2022). The imperative to forecast is especially prevalent in domains that are confronted with heightened instability, as seen in human resources management (Gurchiek, 2022). As unemployment rates decline in Europe (Eurostat, 2023a), there is a simultaneous surge in vacant jobs (Eurostat, 2023b). These developments caused an unforeseen mismatch on the labour market, accentuating the urgency of forecasting in human resources management (Turulja et al., 2023). Both Brunellu & Wruuck (2021) and the European Commission (Turulja et al., 2023) single out big data as the transformative force in anticipating the increasingly complex labour markets.

Despite an evolving landscape that underscores the imperative to forecast with big data, methodologies to leverage big data remain largely unexplored (Maroufkhani et al., 2019; Mikalef et al., 2018), especially in the field of forecasting where conventional data science norms do not always apply (Hewamalage et al., 2022). Currently, the field of forecasting is mainly occupied with developing increasingly complex forecasting models with the promise of enhanced accuracy. However, in today's data-driven landscape with millions of potential predictor variables, the search for suitable variables emerges as a more pressing concern, outpacing the desire to overly complicate forecasting models. This search for suitable variables becomes even more pressing as the issue of parameter instability, previously addressed by Stock & Watson (1996) and Swanson (1998), reappears amidst the era of big data. Parameter instability refers to varying relationships between variables as time evolves.

Hardly withstanding the reemerging issue of parameter instability, the field of forecasting continues to rely on the out-of-sample forecasting procedures introduced by Armstrong & Grohman (1972) over five decades ago. These out-of-sample forecasting procedures replicate the way a practitioner sequentially forecasts over a period of time, serving as a way to cross-validate the accuracy of a forecasting model. Two out-of-sample forecasting procedures possess this capability, the expanding window and the rolling window, with the latter recommended when parameter instability is at play. Still, a crucial consideration remains unaddressed with these traditional out-of-sample forecasting procedures: they rely on model recalibration. Model recalibration entails that while the coefficients of variables are allowed to be adjusted based on new data, the fundamental structure of the forecasting model remains static. Consequently, the forecasting model's inability to responds to time-varying relationships leads to the persistence of variables with little predictive value, resulting in inaccurate forecasts. Moreover, the procedures do not accurately replicate reality, as a practitioner is unlikely to retain a weakened predictor variable.

Supported by Mulero & Garcia-Hiernaux (2021), today's vast amount of data, and increased computational power, this study advocates for a fundamental shift, moving from model recalibration to model respecification. Different from model recalibration, the novel concept of model respecification involves the continuous search for a superior forecasting model by substituting weakened predictor variables. This is expected to be a solution that is more potent in mitigating the effect of parameter instability on forecasting accuracy. Supporting this rationale, the brightest brain in the room, ChatGPT, stipulates five limitations that arise from using the out-of-sample forecasting procedures that rely on model recalibration in today's world. ChatGPT also puts forward five potential solutions to address these limitations, all of them suggesting to allow the addition or removal of variables based on new data. To quote ChatGPT: "while using rolling window and expanding window techniques can help capture temporal patterns, the inflexibility of a fixed variable set can limit the adaptability and accuracy of your forecasting model. Employing adaptive strategies or hybrid approaches can help address these limitations and improve the model's performance in dynamic environments."

The concept of respecification is operationalized through the development of a fully-automated R-based algorithm. This algorithm is applied to a real-world scenario that involves the forecasting of youth unemployment rates in the Netherlands, exploiting the big data captured by Google Trends. In doing so, this study empirically validates the exploratory nature of its hypotheses. To systematically work towards a conclusion, the following research question is utilized: "What impact does model respecification have on the forecasting accuracy of Dutch youth unemployment rates when there are time-varying relationships between variables?"

Supporting the chosen forecasting exercise, it is precisely big data taken from internet sources like Google Trends that the European Commission recognizes as having transformative potential in labour market forecasting (Turulja et al., 2023). Moreover, the complexity of exploiting big data in forecasting becomes evident when applying Google Trends data, making the forecasting exercise well suited to the research rationale. The complexity is reflected in the task of selecting a few suitable keywords from an array of millions of potential keywords (Li et al., 2017; Varian, 2014), the complexity of which is amplified with Google Trends data having the tendency to structurally change over time (Stephens-Davidowitz & Varian, 2014; Nagao et al., 2019), giving rise to the topic parameter instability (Suhoy, 2009; Zagheni & Weber, 2015).

In answering the research question, this study aims to make several academic contributions. First, the main academic contributions of this study is to address the future research suggestion put forward by Mulero & Garcia-Hiernaux (2021). This is done by developing and testing a respecification algorithm that furnishes literature with a validated out-of-sample forecasting procedure that is suited to an ever-evolving landscape characterized by data generation, computational advancement, and increased instability. Second, this study aims to enrich forecasting literature by presenting a novel perspective on recent developments. Contrasting the trend of introducing more complex forecasting models to improve forecasting accuracy, this study aims to highlight that forecasting accuracy can also be improved through methodological advancements that are applied to basic forecasting models. This insight aligns with the growing demand for explainable AI and a departure from black-boxes like neural networks in the forecasting domain (Rozanec et al., 2022). Third, this study aims to demonstrate how out-of-sample forecasting procedures affect forecasting accuracy. Thereby answering the research call put forward by Rossi & Inoue (2012) and Fezzi & Mosetti (2020).

Fourth, this study aims to contribute to forecasting literature by strengthening previous studies that underscored the potential of leveraging big data from internet sources like Google Trends in forecasting. Different from previous studies, this study also demonstrates how Google Trends data behaves during crises like COVID-19. The relevance of which is underscored with literature claiming that COVID-19 is precisely what triggered organizations to engage in big data analysis and forecasting (Sheng et al., 2021). Finally, to promote research on the concept of model respecification, the algorithm developed by this study is made available upon request.

The practical contributions of this study closely related to the academic contributions. First, by providing practitioners with an innovated out-of-sample forecasting procedure, this study aims to assist organizations in developing the capabilities necessary to leverage big data in today's volatile business environment, in which the necessity to forecast is underscored (Sheng et al., 2021). Assisting organizations in building big data capability becomes particularly relevant as an increasing amount of organizations are forced to lean more toward data-driven decisions to effectively respond to future events (Sheng et al., 2021). Given that forecasting is a key-component to decision-making processes (Makridakis et al., 1998), this study intends to significantly increase the accuracy of forecasts, thereby increasing the quality of strategic decision-making across businesses, government institutions, and other organizations.

Second, while literature claims that COVID-19 has triggered many organizations to start engaging in big data analytics and forecasting (Sheng et al., 2021), the necessity of forecasting labour markets has been a point of discussion for years (Louch, 2014; Akram & Nymoen, 2016; Muehlemann & Leisser, 2018). Yet, small organizations may not have access to costly labour market intelligence tools. This study aims to show practitioners how they could leverage big data, that is publicly available, through relatively straightforward techniques to forecast youth unemployment rates. In doing so, this study contributes to making labour market forecasting an accessible endeavor to a broad range of organizations.

This study starts with a discussion of literature related to out-of-sample forecasting procedures, Google Trends data, and the use of Google Trends data for unemployment forecasting. Subsequently, the methodology chapter outlines the research design and the techniques employed to answer the research question. Following the methodology chapter, the results chapter presents the forecasts and hypothesis tests. Afterwards, the analysis chapter compares the obtained results with other studies. Finally, a conclusion is presented.

2. Literature review

This chapter introduces the theoretical concepts required for a comprehensive understanding of this study and its addressed research gap. First, the reader will be familiarized with the literature on out-of-sample forecasting procedures, which inspired the research gap of this study. Subsequently, the applications and limitations of forecasting with Google Trends data are discussed. Finally, studies conducted in a similar context are discussed, focusing on forecasting unemployment with Google Trends data.

2.1 Out-of-sample procedures

Despite few studies exploring the out-of-sample forecasting procedures, they are an essential consideration in every forecasting study. This section first discusses the basics of out-of-sample testing, later extending it to the field of forecasting, followed by a discussion of fixed-origin and rolling-origin procedures for out-of-sample forecasting. Finally, two crucial considerations inherent to the procedures for out-of-sample forecasting are discussed.

In most data analysis applications, a model is trained and subsequently tested. This is done by dividing the dataset into a training set (in-sample) and a testing set (out-of-sample/unseen data). In forecasting, the training set is also known as the window or the origin (Tashman, 2000). The training set is used to create (train) the model, while the testing set its function is to test how the model reacts to data it is unfamiliar with. If done correctly, any claims regarding accuracy of a model are reported based on the testing data (Tashman, 2000). Typically, a rule of thumb is to allocate 70% of the dataset for training and set-aside 30% for testing. The separation of training and testing aims to detect overfitting, where a model is able to perfectly explain the training data while poorly explaining unseen data (Makridakis et al., 1982). In most data analysis applications, a model that is overfitting is rendered useless.

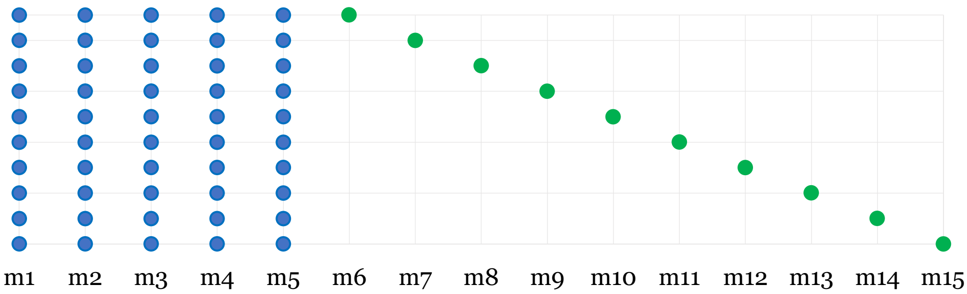
Reacting to the problem of overfitting, in a great deal of data applications, model testing is considered the most crucial part in generalizing the performance of a model (Cerquiera et al., 2020). Similarly, most forecasters agree that the performance of a forecasting model should be assessed on unseen data (Tashman, 2000). However, due to the critical role of chronological order in forecasting, testing forecasting models becomes complex (Hewamalage, 2022). To deal with this complexity, various out-of-sample forecasting procedures exist, each with its advantages and disadvantages. Yet, most forecasting studies neglect to make a well-informed decision about the out-of-sample forecasting procedure to apply (Tashman, 2000). Moreover, most studies don't even mention the out-of-sample forecasting procedures that was used. Consequently, the replication and comparison of forecasting studies becomes a challenge (Tashman, 2000). Below, the out-of-sample forecasting procedures are discussed in detail.

Two primary procedures for out-of-sample forecasting exist; the fixed-origin and the rolling-origin (Tashman, 2000; Hewamalage, 2022). While the fixed-origin procedure for out-of-sample forecasting has long been used, the rolling-origin procedure for out-of-sample forecasting is today's preferred choice. This preference stems from the rolling-origin's ability to test forecasting models recursively, allowing for cross-validation. Complicating things even more, two variations of the rolling-origin procedure for out-of-sample forecasting exist; the expanding window and the rolling window. First, the fixed-origin procedure is discussed.

In figure 1, the fixed-origin procedure for out-of-sample forecasting is illustrated. The fixed-origin is likely sufficient for most data analysis applications, where chronological order is no concern. In forecasting, the fixed-origin is deemed the most limited procedure for out-of-sample forecasting (Hewamalage, 2022). With the fixed-origin, the complete dataset is divided into a training set and testing set, following some rule of thumb, and this division remains constant. As depicted in figure 1, the complete dataset consists of 15 monthly values that are in chronological order. In the top row of figure 1, the five blue dots represent the training set (window), consisting of five monthly values. Using a forecasting model trained on this window, an one month ahead forecast is done, forecasting month six, as shown with the green dot.

However, as shown in the second row from the top in figure 1, month seven is forecasted using the same window that was used to forecast month six, resulting in a two month ahead forecast. This is due to the window that is used for training remaining fixed. Consequently, the ability of a forecasting model to forecast any number of months ahead can only be assessed based on single test value. Assessing a forecasting model's performance on a single test value is hardly representative, therefore, the fixed-origin procedure for out-of-sample forecasting is rarely used nowadays (Liu & Yang, 2020; Hewamalage, 2022). Next, the rolling-origin procedures for out-of-sample forecasting are introduced.

Figure 1: Fixed-origin

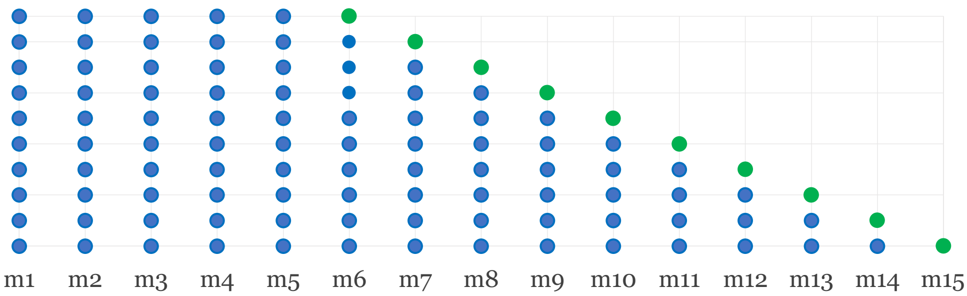


Note: Figure depicts the fixed-origin procedure for out-of-sample forecasting. Blue dots are values that are used to train the forecasting model. Green dots represent values that are forecasted. The horizontal axis represents months, with m1 representing month one. The figure should be read as follows: to forecast month six, a window consisting of month one till five is used for training. To forecast month seven a window consisting of month one till six is used for training. This figure is replicated from Hewamalage (2022).

The rolling-origin is different from the fixed-origin in one specific regard. With the rolling-origin, the window that is used for training is continuously being updated with the actual value for a month that was forecasted as soon as it becomes available (Tashman, 2000). Subsequently, the forecasting model is retrained based on the updated window, thus incorporating up-to-date data, and allowing to cross-validate a forecasting model's ability to forecast any number of months ahead (Hyndman & Athanasopoulos, 2021). Thanks to the cross-validation of forecasts, the obtained error metric is more representative of the forecasting model's performance (Berrar, 2019). Below, both rolling-origin procedures for out-of-sample forecasting are discussed separately, starting with the expanding window.

In the expanding window, the window used for training is allowed to become larger over time. Depicted in the top row of figure 2, the window initially contains five monthly values that are used for training. Using a forecasting model trained on this window, an one month ahead forecast is then done, forecasting month six. Unlike the fixed-origin, after month six is forecasted, the window is updated with the actual value for month six as soon as it is known. Based on the new window, the same forecasting model is updated or recalibrated, and another one month ahead forecast can be done, forecasting month seven. Consequently, the expanding window allows to recursively forecast any number of month ahead. The concept of updating and recalibrating are discussed later. First, the rolling window is discussed below.

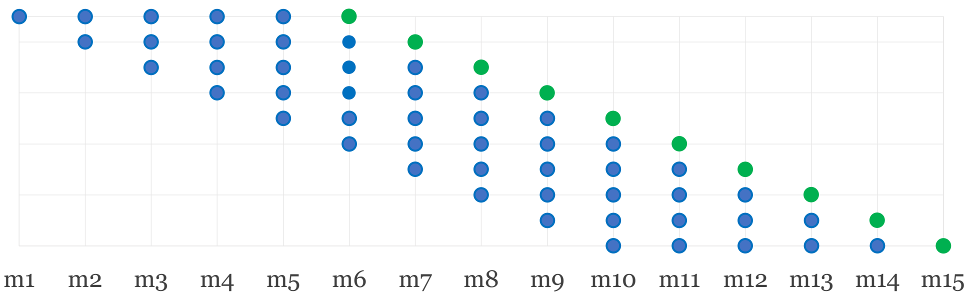
Figure 2: Expanding window



Note: Figure depicts the expanding window. Blue dots are values that are used to train the forecasting model. Green dots represent values that are forecasted. The horizontal axis represents months, with m1 representing month one. The figure should be read as follows: to forecast month six, a window consisting of month one till five is used for training. To forecast month seven a window consisting of month one till six is used for training. This figure is replicated from Hewamalage (2022).

Figure 3 shows that the rolling window operates similar to the expanding window. The only difference being that when the window is updated with the actual value for a month that was forecasted, the oldest value is pruned in the rolling window. This is depicted in the second row from the top in figure 3. As a result, the size of the window remains constant. Next, the advantages and disadvantages of the expanding window and rolling window are discussed.

Figure 3: Rolling window



Note: Figure 3 depicts the rolling window. Blue dots are values that are used to train the forecasting model. Green dots represent values that are forecasted. The horizontal axis represents months, with m1 representing month one. The figure should be read as follows: to forecast month six, a window consisting of month one till five is used for training. To forecast month seven a window consisting of month two till six is used for training. This figure is replicated from Hewamalage (2022).

As shown in figure 2, the expanding window captures all data that is available without pruning any data. Consequently, the expanding window captures more historic data which allows the forecasting model to better recognize long-term trends (Tashman, 2000). However, capturing all historic data may lead to the inclusion of old, noisy data that obscures the accuracy of forecasts (Tashman, 2000; Hewamalage, 2022). Taking into consideration Schnaars (1986), that becomes a pressing issue as the further away the past, the less nuances from the past can be translated to the future. The rolling window offers remedy to this by systematically pruning data that is too old and potentially noisy (Tashman, 2000; Hewamalage, 2022). Moreover, the rolling window tends to be more responsive to sudden changes compared to the expanding window. This responsiveness is due to the practice of pruning old observations, resulting in a smaller window that better highlights sudden, extraordinary changes. However, pruning data may result in essential data not being captured. Nevertheless, the rolling window is strongly recommended in the field of econometrics, where structural breaks are likely to occur (Inoue et a., 2017; Pesaran & Timmermann, 2002). The topic of structural breaks, and the closely related topic of parameter instability, is discussed below.

Structural breaks in data come into play due to abrupt changes that leave a major impact, such as; wars, epidemics, recessions, and major changes in government policy (Hyndman, 2014). Additionally, when the relationships between variables suddenly change, it's also considered to be a structural break (Hyndman, 2014), however, more often referred to as parameter instability (Inoue et al., 2017). Parameter instability, as introduced to econometric literature by Stock & Watson (1996) and Swanson (1998), entails that predictors with predictive value in one period may not have predictive value in subsequent periods. In fact, predictors may not even have predictive value in the successive period (Stock & Watson, 1996). The topic of parameter instability remains a rare focus of studies in the field of forecasting. Meanwhile, it also remains an open discussion point to this day, as per Wang & Hao (2023), Zhang et al. (2019), Pettenuzzo & Timmermann (2017), Catania et al. (2019). Finally, two critical consideration with the out-of-sample forecasting procedures are discussed below, starting with the window size consideration, followed by updating and recalibrating.

Like any data analytics application, the minimum sample size to use in the training of a model is an important consideration with the procedures for out-of-sample forecasting. Even more so with the rolling window and expanding window. The importance of sample size lies in its influence on the reliability, validity and generalizability of results. A smaller sample size tends to undermine validity, while a large sample size may find very small differences to be significant (Faber & Fonseca, 2014). In this context, sample size is synonymous with window size. The window size becomes a pressing topic of discussion when considering the study performed by Rossi & Inoue (2012) and Fezzi & Mosetti (2020). Both Rossi & Inoue (2012) and Fezzi & Mosetti (2020) found that forecasting accuracy is notably sensitive to the window size that is used with out-of-sample forecasting procedures. Yet, the window size consideration is rarely discussed by studies, and some studies even experiment with various window sizes till significant results are finally found (Rossi & Inoue, 2012). Consequently, Rossi & Inoue (2012) plea for robust solutions. Pesaran & Timmermann (2002), both experts in this field, suggest that finding the optimal window size is incredibly complex, and that robustness can only be ensured in one way. To obtain robust results, it is recommended to average forecasts done by the same model but for various window sizes (Pesaran & Timmermann (2002). Below, the second consideration when using procedures for out-of-sample forecasting is discussed, the updating or recalibrating of forecasting models.

When the window is updated with new data and possibly pruned of data, the question is how the forecasting model should adapt to this. Generally, literature recognizes that there are two options in adapting the forecasting model to the updated window (Tashman, 2000; Hewamalage et al., 2022). One is updating the forecasting model, which involves feeding new data to the forecasting model without re-estimating the model parameters (Petropoulos et al., 2022; Tashman, 2000; Hewamalage et al., 2022). The second, more comprehensive, option is to recalibrate the forecasting model each time the window is updated. With model recalibration, all model parameters remains constant, meaning that the forecasting model does not fundamentally change. However, unlike the updating of the forecasting model, recalibrating the forecasting model results in all parameters to be re-estimated based on the updated window (Petropoulos et al., 2022; Tashman, 2000; Hewamalage et al., 2022). As a result, the forecasting model is ensured to have the best possible fit with the new data (Tashman, 2000; Hewamalage et al., 2022). Disadvantageous to model recalibration is the computational power that is needed. However, thanks to increased computational power, model recalibration has become a common practice in forecasting (Hewamalage, 2022).

However, before the parameters of forecasting model can be re-estimated, the forecasting model needs to be specified. Specification refers to the formulation of a model that captures relationships between variables, while estimation is about estimating the value of those variables (Greene et al., 2002). More to the point, Allen (1997) refers to model specification as: “The determination of which independent variables should be included or excluded from a regression equation”. According to Mulero & Garciá-Hiernaux (2021), the distinction between estimation and specification is significant in forecasting. Consequently, Mulero & Garciá-Hiernaux (2021) suggest future research to take the next hurdle in adapting forecasting models when windows are updated, being model respecification.

This section has discussed the out-of-sample forecasting procedures, focussing specifically on the rolling-origin procedures for out-of-sample forecasting. Moreover, the reader has been familiarized with the advantages and disadvantages of the various procedures for out-of-sample forecasting, and two important considerations have been addressed. Next, the application of Google Trends data in forecasting is discussed.

2.2 Google Trends forecasting

This section starts with a brief introduction to Google Trends data and Google's unsuccessful attempt to forecast the flu. Then, various applications of using Google Trends data for forecasting are briefly addressed, leading to a discussion of the advantages and disadvantages of its utilization in forecasting.

Google Trends is an online tool provided by Google. While originally launched as an experiment to explore trends like ice cream popularity and singing competition contestants, it has grown to be a valuable resource for journalists, researchers, and scholars (Rogers, 2021). In an age where data is increasingly locked behind pay-walls, Google Trends data remains publicly available and the only source that offers real-time data on the search behaviour of users. As a result, the potential of Google Trends data continues to be leveraged to this day.

The first studies exploiting Google Trends data are those of Ginsberg et al. (2009) and Choi & Varian (2009). These studies revealed that Google Trends data is able to detect patterns before they actually occur. Ginsberg et al. (2009) showed the ability to detect influenza epidemics in real-time by leveraging Google Trends data. Contributing to another field, Choi & Varian (2009) pursued an econometric forecasting study and demonstrated that the popularity of Google searches like "apply for unemployment" are useful in forecasting future unemployment. The next paragraph addresses Google's failed attempt to forecast the flu.

In response to Ginsberg et al. (2009)'s finding that Google Trends can detect influenza epidemics in real time, Google launched a tool called Google Flu Trends. It did not take long for Google Flu Trends to face notable criticism, when it overestimated doctor visits for influenza related illness by a factor of more than two (Lazer et al., 2014). One of the causes was Google Flu Trends' continuous search for the most correlated keywords, without theorizing ex-ante which keywords are appropriate (Lazer et al., 2014). Moreover, Google Flu Trends was conceived as a substitute, rather than a supplement to traditional data sources (Lazer et al., 2014). Finally, algorithmic changes caused Google Flu Trends to become less accurate over time (Lazer et al., 2014). Yet, Google Flu Trends was not adapted to these algorithmic changes and was simply phased out by Google (Lazer et al., 2014). This led to misconceptions about the reliability of Google Flu Trends, forgetting that it were mainly the methodological decisions and misperceptions that caused the overestimation. More recent research proofs that the flu and various other diseases can be successfully forecasted with Google Trends data (Mavragani & Gkillas, 2020; Dugas et al., 2013; Teng et al., 2017; Chen et al., 2022; Kandula & Shaman, 2019). Below, forecasting studies done in other contexts are briefly addressed.

Undoubtedly, the studies performed by Ginsberg et al. (2009) and Choi & Varian (2009) kickstarted the research on forecasting with Google Trends data. Since then, Google Trends data has been leveraged for forecasting exercises in numerous contexts. Varying from forecasts of car sales in Germany (Fantazzini & Toktamysova, 2015) and Chile (Carriére-Swallow & Labbé, 2011), to tourism in South Korea (Park et al., 2017), Belgium, Austria, Barcelona, and Vienna (Önder et al., 2017). Google Trends data is most prominently used in macroeconomic forecasting, with many studies having forecasted unemployment (Askitas & Zimmerman, 2009; D'Amuri & Marucci, 2017), private consumption (Vosen & Schmidt, 2011; Woo & Own, 2018), consumer price index (Li et al., 2015), gross domestic product (Ferrara & Simone, 2022), and many other macroeconomic variables. Next, the advantages and disadvantages of forecasting with Google Trends data are discussed, starting with the advantages.

The advantages of leveraging Google Trends data in forecasting include full anonymization, thus addressing privacy concerns (Ginsberg et al., 2009). Additionally, Google Trends collects data without effort needed from the user, contrasting traditional tools like surveys (Mclaren & Shanbhogue, 2011). Users may not even be consciously aware that data is being recorded, ensuring that the recorded data accurately reflects natural behaviour (Mclaren & Shanbhogue, 2011; Scharkow & Vogelsang, 2011). However, data collection without users being aware of it could give rise to privacy concerns. Other advantages include the real-time availability of pre-processed data at no cost. Consequently, Google Trends data facilitates research possibilities without the need to invest substantial financial or human resources (Askitas & Zimmermann, 2009; Zhu et al., 2012). Nevertheless, pre-processing by Google does come with the risk of "garbage in, garbage out" (Cebrián & Domenech, 2022). Moreover, it does not allow to critically assess the sampling error (Cebrián & Domenech, 2022). This, and other disadvantages of using Google Trends data are discussed in detail below.

Literature suggests that the out-of-sample forecasting potential of Google Trends data is largely overlooked, with most studies focusing on merely the in-sample fit (Niesert et al., 2020). Moreover, Google Trends data is criticized to be unreliable, due to the homogeneity of internet users having an influence on keyword popularity (Stephens-Davidowitz & Varian, 2014). Also, some individuals may lack internet access or use alternative sources of search, resulting in coverage bias when using Google Trends data (Cebrián & Domenech, 2022).

Recent literature increasingly focusses on the inconsistency of Google Trends data. Various studies have demonstrated that Google Trends data for the same keyword, during the same period, may be different when again collected tomorrow, or next week (Cebrián & Domenech, 2022). Eichenauer et al. (2021) attribute this inconsistency to sampling variation, which is more noticeable for less popular keywords and smaller regions due to smaller samples. Remedy can be taken by repeatedly obtaining Google Trends data for the same keyword and period, spread out over various days, to then take the mean (Eichenauer et al., 2021). Furthermore, Google Trends data has the tendency to structurally change over time (Stephens-Davidowitz & Varian, 2014; Nagao et al., 2019; Behnen et al., 2021), giving rise to the issue of parameter instability in forecasting applications. Suhoy (2009) and Zagheni & Weber (2015) also emphasize that the relationship between Google Trends data and phenomena that needs to be forecasted only hold for a limited period. Although Google is likely to have explanations for the deviations, they are not publicly shared (Behnen et al., 2021).

Furthermore, spurious correlations between Google Trends data and phenomena that need to be forecasted are easily found (Stephens-Davidowitz & Varian, 2014). For example, Google Trends data for a popular drink was highly correlated to housing sales in the US (Stephens-Davidowitz & Varian, 2014). Forecasting studies have paid little attention to this, which is criticized by Tran et al. (2017), who suggest to follow a structured approach. Returning to Google Flu Trends, it was precisely these spurious correlations that contributed to its failure. Finally, there are very few published studies that conducted real-world exercises and found no improvement in forecasts when using Google Trends data. Instead, most criticism appears to stem from review articles. Consequently, anyone interested in forecasting with Google Trends data should be aware of the possibility of a file drawer effect (Chatfield & Xing, 2019).

This section has familiarized the reader with the application of Google Trends data in forecasting. Moreover, various advantages and disadvantages involved in the use of Google Trends data are discussed. The next section focusses specifically on leveraging Google Trends data to forecast unemployment rates.

2.3 Google Trends unemployment forecasting

This section starts by discussing a few prominent studies that used Google Trends data to forecast unemployment. Then, an intricate issue that is often raised is discussed, which is the selection of keywords. Moreover, the consideration of procedures for out-of-sample forecasting are discussed in the context of forecasting unemployment with Google Trends data. To summarize some of the discussion points in this section, table 1 presents the forecasting accuracy and outlines the out-of-sample forecasting procedures used in the discussed studies.

As previously discussed, forecasting with Google Trends data finds its roots in Ginsberg et al., (2009) and Choi & Varian (2009) their studies. However, forecasting unemployment with Google Trends data has particularly become an popular research endeavour in recent times, with 30% of the studies in this field being published in the last two years. The study of D'Amuri & Marucci (2017) may be the most prominent study so far. Numerous forecasting models with traditional variables were tested against a forecasting model using Google Trends data for the keyword "jobs" (D'Amuri & Marucci, 2017). Results demonstrate that the forecasting model with Google Trends data outperformed all the other forecasting models. Notably, the study of D'Amuri & Marucci (2017) reported the procedure that was used for out-of-sample forecasting and the window size, although they did not motivate it.

Most studies that exploited Google Trends data to forecast unemployment start with a simple forecasting model that serves as a benchmark. Often, this is an autoregressive model that solely uses historic unemployment data to forecast future unemployment. Subsequently, this benchmark model is augmented with Google Trends data for one keyword or multiple keywords. If the augmented forecasting model, with Google Trends data, proofs to be more accurate than the benchmark, Google Trends data is considered to have added value in forecasting. Examples of studies that followed this approach are; Mclaren & Shanbogue (2011) for the UK, Tuhkuri (2016) for the US, Te Brake (2017) for the Netherlands, González-Fernández & González-Verlasco (2018) for Spain, Pavlicek & Kristoufek (2015) for Hungary, Poland, Slovakia, and the Czech Republic, Barreira et al. (2013) for Portugal, Spain, France, and Italy, and Simionescu & Cifuentes-Faura (2022) for Spain and Portugal. The obtained forecasting accuracy off all these studies are mentioned in table 1, nearly all studies found Google Trends data to enhance unemployment forecasts.

Although most studies followed the above-mentioned approach, other modelling techniques have also been exploited. For example, Singhania & Kundu (2021) developed a neural network model that leverages Google Trends data for over 500 potential keywords. Results show the superiority of the neural network over an autoregressive model that leveraged Google Trends data for a single keyword (Singhania & Kundu (2021). However, it is unknown in what lies the improvement. Is the use of Google Trends data for multiple keywords superior to using Google Trends data for a single keyword? Or is the neural network modelling technique superior to the autoregressive modelling technique? This boils down to an uncertainty in literature, being the selection of keywords. A detailed discussion of this uncertainty follows on the next page.

The selection of keywords to obtain Google Trends data for is incredibly challenging, as claimed by Li et al. (2017) and Varian (2014) who speak from experience. Yet, keyword selection is also considered to be the most crucial aspect in successfully leveraging Google Trends data in forecasting. The selection of suitable keywords becomes an intricate task because there are millions of keywords that could have potential in forecasting unemployment (Varian, 2014). In selecting keywords, researchers are not only challenged with which keywords to collect Google Trends data for, but also how many keywords to collect Google Trends data for. Collecting Google Trends data for multiple keywords comes with the benefit of more heterogeneity and thus a higher likelihood of capturing patterns. Conversely, collecting Google Trends data for multiple keywords may lead to overfitting or multicollinearity (Varian, 2014). While some studies leveraged Google Trends data for a single keyword, others undertook a multiple keyword approach. Both approaches are discussed below.

Many studies that leveraged Google Trends data to forecast unemployment focused on a single keyword. Examples being: D'Amuri & Marucci (2017) and Choi & Varian (2009) that used just the keyword "jobs", Mclaren & Shanbogue (2011) that used "jobseeker's allowance", González-Fernández & González-Verlasco (2018), Barreira et al. (2015), and Simionescu & Cifuentes-Faura (2022) that used "unemployment", Naccarato et al. (2018) that used "job offers", Fondeur & Karamé (2013) that used "employment", and Vicente et al. (2015) that used "job offer". Others constructed an index by averaging over Google Trends data for multiple keywords. For example, Tuhkuri (2016) created an index by averaging over thirteen keywords with weights based on the search volume that is known by the Google algorithm. Askitas & Zimmermann (2009), Smith (2016), and Te Brake (2016) followed a similar methodology. Askitas & Zimmermann (2009) mentions the use of multiple keywords to "weed out the noisy activity and get to the signal in any kind of effective way". However, this methodology solely relies on intuition, and any formal keyword selection techniques are not utilized.

Recent literature has neglected keyword selection based on intuition and adopted formal techniques (Borup & Schutte, 2022; Singhania & Kundu, 2021; Mulero & Garcia-Hiernaux, 2021). Although each of the aforementioned studies undertook a slightly different approach, all started by obtaining Google Trends data for a substantial set of keywords that are related to unemployment. Li et al. (2014) refers to this as the use of domain ontology, being the phenomena that: "Captures concepts and their relationships to a specific domain as well as representing the axioms and constraints that define the prominent features of the domain". After obtaining the Google Trends data, dimension reduction techniques were applied. Dimension reduction techniques serve as a way to remove noisy and irrelevant data in order to improve accuracy and reduce computational time (Velliangiri et al., 2019). Mulero & Garcia-Hiernaux (2021) and Li et al. (2014) concluded that forward selection is best suited for dimension reduction and is preferred over other dimension reduction techniques, like; principal component analysis, backward selection, and the genetic algorithm.

Borup & Schutte (2022) introduced regularization in their dimension reduction approach, a technique that was not considered by Mulero & Garciá-Hiernaux (2021), Li et al. (2014) or Singhania & Kundu (2021). However, Borup & Schutte (2022) merely concluded that using Google Trends data for multiple keywords is preferable to using Google Trends data for a single keyword. Borup & Schutte (2022) do not compare regularization to other dimension reduction techniques, nor did they conduct robustness checks. Mulero & Garcia-Hiernaux (2021) do provide robust results, consequently, their suggestion to use forward selection inspires more confidence. The next page presents a discussion relating back to the consideration of window size, updating vs. recalibration, and parameter instability.

Remarkably, Mulero & Garcia-Hiernaux (2021) are one of the few studies that explicitly address the sensitivity of forecasting accuracy to the window size and provide results that are robust to the window size. Most studies do not even mention the procedure that was followed for out-of-sample forecasting procedure, nor is the window size explicitly mentioned. Barreira et al. (2013) is the only study coming close, as they avoided out-of-sample forecasts during periods characterized by structural breaks. Considering literature's recommendation to use the rolling window with economic data, it is surprising that a majority of the discussed studies used the expanding window in their forecasts of unemployment. Examples being; Mclaren & Shanbogue (2011), González-Fernández & González-Verlasco (2018), Naccarato et al. (2018), Mulero & Garcia-Hiernaux (2021), and Smith (2016). Moreover, considering that Google Trends data is criticized for being unstable, it is equally surprising that none of the studies brought up the issue of parameter instability. Astonishingly, one recent study did not even recalibrate its forecasting model as the window was updated.

Nevertheless, a few studies did use the rolling window, as is shown in table 1 (D'Amuri & Marucci, 2017; Tuhkuri, 2016; Te Brake, 2017). Unfortunately, none of these studies motivated their decision for the out-of-sample forecasting procedure or the window size. The studies may be engaging in what Rossi & Inoue (2012) brought up, the experimentation with various procedures and window sizes until significant results are finally found.

Moreover, numerous studies use a small out-of-sample, which may lead to a biased assessment of forecasting accuracy (Liu & Yang, 2020; Hewamalage, 2022). For example, Simionescu & Cifuentes-Faura (2022) use an out-of-sample that consists of merely six out-of-sample forecasts (n=6), Chadwick & Sengul (2012) use n=15, Vicente et al. (2015) use n=12, Barreira et al. (2013) use n=12, and Singhania & Kundu (2021) use n=20. Although there are no formal guidelines for the required sample size, such small sample sizes are hardly valid when performing hypothesis tests for forecasting accuracy. Consequently, many studies also did not perform hypothesis tests. Examples being; Mclaren & Shanbogue (2011), González-Fernández & González-Verlasco (2018), and Singhania & Kundu (2021).

As mentioned earlier, the table presented on the next page serves as a summary of some of the discussion points of this section. For example, table 1 shows that D'Amuri & Marucci (2017) their forecasting model that leverages Google Trends data had an accuracy of .1958 as measured by RMSFE. Moreover, table 1 shows that a rolling window with a window size of 37 was used as out-of-sample forecasting procedure. Finally, table 1 shows that the accuracy of the forecasting model was determined based on 84 one month ahead forecasts in the US.

Rounding off, this chapter has discussed the various out-of-sample forecasting procedures and revealed that studies fail to motivate their choice for an out-of-sample forecasting procedure and the considerations that come with it. This part of the literature review inspires this study's research rationale. Moreover, the advantages and disadvantages of using Google Trends data were discussed, and the use of Google Trends data in unemployment forecasting were discussed. This provided valuable insights into various important considerations when using Google Trends data for forecasting. Next, the methodology of this study is addressed.

Table 1. Summary of discussed literature

Study	Country	Model	Procedure	Window size	Out-of-sample	Metric	Accuracy
D'Amuri & Marucci (2017)	US	AR	Rolling	37	84	RMSFE	.1958
Mclaren & Shanbogue (2011)	UK	AR	Expanding	49	31	RMSE	35.3
Mclaren & Shanbogue (2011)	UK	AR	Expanding	49	31	Adj. R^2	85%
Pavlicek & Kristoufek (2015)	CZ	AR	-	96	24	-	-
Pavlicek & Kristoufek (2015)	Hungary	AR	-	96	24	-	-
Pavlicek & Kristoufek (2015)	Poland	AR	-	96	24	-	-
Pavlicek & Kristoufek (2015)	Slovakia	AR	-	96	24	-	-
Barreira et al. (2013)	Portugal	AR	-	17	12, 24, & 36	Adj. R^2	55.4%
Barreira et al. (2013)	Portugal	AR	-	17	12, 24, & 36	RMSFE	.00126
Barreira et al. (2013)	Spain	AR	-	17	12, 24, & 36	Adj. R^2	68.3%
Barreira et al. (2013)	Spain	AR	-	17	12, 24, & 36	RMSFE	.00122
Barreira et al. (2013)	France	AR	-	17	12, 24, & 36	Adj. R^2	75.5%
Barreira et al. (2013)	France	AR	-	17	12, 24, & 36	RMSFE	.000509
Barreira et al. (2013)	Italy	AR	-	17	12, 24, & 36	Adj. R^2	69.3%
Barreira et al. (2013)	Italy	AR	-	17	12, 24, & 36	RMSFE	.00182
Tuhkuri (2016)	US	AR	Rolling	48	130	MAPE	7.01%
Te Brake (2017)	NL	AR	Rolling	132 & 48	28 & 112	MAPE	4.09%
González-Fernández & González-Verlasco (2018)	Spain	AR	Expanding	84	83	RMSE	.3343
Simionescu & Cifuentes-Faura (2022)	Spain	AR	-	204	6	RMSE	.261
Simionescu & Cifuentes-Faura (2022)	Spain	AR	-	204	6	RMSE	.266
Chadwick & Sengul (2012)	Turkey	BMA	-	82	15	RMSE	.0176
Vicente et al. (2015)	Spain	ARIMA	-	108	12	RMSE	59.056
Vicente et al. (2015)	Spain	ARIMA	-	108	12	MAPE	1.075%
Mulero & Garcia-Hiernaux (2021)	Spain	ARIMA	Expanding	144	33	RMSE	25% lower
Smith (2016)	UK	MIDAS	Expanding	48	71	RMSFE	36.7
Singhania & Kundu (2021)	US	LSTM	Expanding	166	20	MAPE	6.24%
Singhania & Kundu (2021)	US	VAR	Expanding	166	20	MAPE	9.12%
Naccarato et al. (2018)	Italy	VAR	Expanding	80	44	Theil	.775

Note: This table shows the approach and results of the afore mentioned studies in forecasting unemployment with Google Trends data. The first column describes the country for which unemployment forecasts were done. The model column describes which modelling technique was used by the study, the procedure refers to the procedure that was used for out-of-sample forecasting. Related to the procedure is the window size column, this describes the size (amount of data points) of the training window. In the case of the expanding window, window size refers to the initial size of the window. The out-of-sample column refers to the number of out-of-sample forecasts that were done to test the forecasting model. Finally, the last two columns show the accuracy of the forecasting model in forecasting unemployment one month ahead, measured by some metric.

3. Methodology

This chapter starts by delving into the research question and the hypotheses underpinning this study. Subsequently, the data that is leveraged to address the research question, and the pre-processing approach is elaborated upon. After discussing the data, this chapter proceeds to cover the methodology that was used to transition from in-sample analysis to out-of-sample analysis and error diagnostics. By addressing the methodology, the aim is to clarify the results that are presented later and to enhance this study's replicability.

3.1 Research design

In this section the research rationale of this study is first addressed, followed by a motivation for the context in which the research rationale is placed. Then, the research question of this study and the subsequent hypothesis are elaborated upon based on literature.

This quantitative study engages in an comprehensive exploration of the procedures commonly applied for out-of-sample forecasting. This exploration specifically focusses on the two rolling-origin procedures for out-of-sample forecasting; the rolling window and the expanding window. The fixed-origin procedure for out-of-sample forecasting is not considered due to many abandoning it (Hewamalage, 2022). The rolling window and expanding window stem from Armstrong & Grohman's (1972) research, which is over fifty years old and has barely been innovated since then, with model recalibration being the only notable improvement. Nowadays, close to every forecaster uses model recalibration in their out-of-sample forecasts (Hewamalage, 2022), and this study argues the time has come to further innovate the traditional procedures for out-of-sample forecasting.

Together with Mulero & Garciá-Hiernaux (2021), this study argues that in today's increasingly dynamic world, characterized by booming data generation and unparalleled computational power, model respecification is a promising innovation of the traditional procedures for out-of-sample forecasting, substituting model recalibration. While model recalibration allows the forecasting model to adjust to changes in the external environment, it does so to a limited degree. These limitations stem from the assumption that the initial creation (specification) of a forecasting model lasts indefinitely, with adjustments of the model limited to the coefficients of parameters as new data is fed to the model. However, with economic data where parameter instability is likely at play, adjusting coefficients may not be sufficient. Parameter instability can cause the relationships between variables to structurally change, reducing the predictive value of predictor variables. Nevertheless, even when a predictor variable loses all of its predictive value, model recalibration will maintain the predictor variable because the model specification remains static. In contrast, the concept of model respecification continuously adapts the model specification, substituting weakened predictor variables with superior ones. In this study, the novel concept of model specification is compared to model recalibration.

Comparing model respecification to model recalibration is done by means of a real-world forecasting exercise. The approach that yields the highest forecasting accuracy is considered to be superior. To engage in this forecasting exercise, two algorithms will be developed. One algorithm initially specifies a forecasting model and then recalibrates the model parameters when the rolling window or expanding window is updated. The other algorithm is unique in literature, it specifies a new forecasting model each time the rolling window or expanding window is updated. Both algorithms are driven by a dimension reduction technique.

The real-world forecasting exercise that this study engages in, involves one month ahead forecasting the unemployment rate of Dutch 15-25-years-olds. Not only is unemployment data notorious for exhibiting structural breaks, making it fitting to the research rationale, it is also allows this study to contribute to the increasing complexity of forecasting labour markets (Turulja et al,. 2023). Moreover, the practical relevance of forecasting the youth unemployment rate was stipulated by a multinational that operates in the retail industry and Westerhuis (2022). To forecast the youth unemployment rate, data taken from Google Trends is leveraged. With the use of Google Trends data comes the complexity of keyword selection and parameter instability. Both complexities fit the research rationale and are representative of the complexity that arises from leveraging big data in forecasting. Moreover, the European Commission calls for more research on how to effectively leverage big data taken from internet sources to forecast labour markets (Turulja et al., 2023). To structurally explore the potential of model respecification, the following research question is used:

RQ: "What impact does model respecification have on the forecasting accuracy of Dutch youth unemployment rates when there are time-varying relationships between variables?"

Using the research question as the guiding point of this study, four hypothesis are drafted that systematically work their way to answering the research question. Ensuring that each separate hypothesis has added value to literature, all hypotheses are inspired by the literature that was discussed in the literature review. Below, the use of each hypothesis is motivated and the hypotheses are presented in the order of testing.

Motivated by the studies of Inoue et al. (2017) and Pesaran & Timmermann (2002), the first hypothesis examines whether the rolling window is more accurate than the expanding window in this study's forecasting exercise. Although it has already been proven that the rolling window is recommended with economic data, no study has questioned this finding in a context similar to this study's forecasting exercise. This is confirmed by Hewamalage (2022) who claims that practically all forecasting studies forget to consider the impact of their chosen out-of-sample forecasting procedure. By testing hypothesis one, literature's attention can be directed to the importance of this consideration. Moreover, it may raise questions about the fairness of cross-comparing studies that used different procedures for out-of-sample forecasting. Motivated by the afore-mentioned, hypothesis one is drafted and shown below.

H1: "Model recalibration in a rolling window is significantly more accurate than model recalibration in an expanding window when forecasting Dutch youth unemployment rates"

Hypothesis one also serves as the starting point for hypothesis two and three. To assess the impact of model respecification on forecasting accuracy, it is essential to first quantify the forecasting accuracy when model recalibration is used. Hypothesis one allows to do so by providing the forecasting accuracy when using model recalibration. Except for Mulero & Garciá-Hiernaux (2021) their study there is no prior literature to motivate hypothesis two and three. Mulero & Garcia-Hiernaux (2021) apply model recalibration in their algorithm and argue that model respecification will improve forecasting accuracy. Motivated by Mulero & Garciá-Hiernaux (2021) and this study's own theorizing, hypothesis two and three are drafted:

H2: "Model respecification in a rolling window is significantly more accurate than model recalibration in a rolling window when forecasting Dutch youth unemployment rates"

H3: "Model respecification in an expanding window is significantly more accurate than model recalibration in an expanding window when forecasting Dutch youth unemployment rates"

Most significant to the research rationale of this study are hypothesis two and three. Nevertheless, this study also aims to determine which procedure for out-of-sample forecasting is more accurate when model respecification is used. Literature argues that the rolling window will yield the most accurate results in a forecasting exercise like the one presented in this study. However, this argument is based on procedures for out-of-sample forecasting that rely on model recalibration (Inoue et al., 2017; Pesaran & Timmermann, 2002). By testing hypothesis four, this study identifies the out-of-sample forecasting procedure that yields the highest accuracy when model respecification is used. Hypothesis four is given below.

H4: “Model respecification in a rolling window is significantly more accurate than model respecification in an expanding window when forecasting Dutch youth unemployment rates”

This section has outlined the research rationale, context, research question, and hypotheses of this explorative study. The next chapter will outline how this research design is implemented by addressing how the data is pre-processed, and how various techniques are used to test the hypotheses and ultimately answer the research question.

3.2 Data

In this section, the two data sources that are used for this study's forecasting exercise, unemployment data and Google Trends data, are discussed. Both data sources are discussed in terms of reliability, validity, and representativeness. The Google Trends data is discussed in more detail, due to the complexity of collection and pre-processing Google Trends data. First, the unemployment data is discussed.

The unemployment dataset, containing the unemployment rate of 15-25-years-olds in the Netherlands, has been sourced from the Dutch Central Bureau for Statistics (CBS). This dataset contains the monthly unemployment rate for a period spanning from April 2008 till December 2022. Two datasets are available, one adjusted for seasonal effects, and one that is not adjusted for seasonal effects. The CBS recommends statisticians to use the non-seasonal dataset because it is not subject to revisions after publication (Van den Brakel & Krieg, 2010). This recommendation is followed. The unemployment rate is calculated by taking the proportion of the working population that is unemployed, as shown in formula 1 (CBS, 2023). In the case of youth unemployment, the unemployed and working population is limited to the 15-25-years-olds. Forecasting the unemployment rate may present a limitation for practitioners, as it can't reflect the absolute number of unemployed without knowing the working population. Conversely, if the absolute number of unemployed would need to be forecasted, forecasts of population growth would need to be taken into account, introducing additional uncertainty.

$$\text{Unemployment rate:} \quad \left(\frac{U}{W}\right) * 100 \tag{1}$$

U = Unemployed
W = Working population

The CBS collects the unemployment data by means of a monthly vacancy, called the "Enquête Beroepsbevolking", with an approximate sample size of 20,000 individuals (CBS, 2023). As acknowledged by the CBS, a sample size of 20,000 is not adequate to represent a working population of approximately 10,000,000 individuals (Van den Brakel & Krieg, 2010). To deal with the lack of representativeness, the unemployment data is inferenced with a time series model (Van den Brakel & Krieg, 2010). Although the CBS deems the inferencing to be satisfactory, representativeness is not guaranteed. Little information is available regarding the reliability and validity of the unemployment data. Concerning the reliability of the unemployment data, it is only known that repeated sampling is avoided, that should not necessarily put stress on the reliability if each sample has a similar distribution. Validity may be at risk if an individual engaging in undeclared work is not truthful in its response.

Despite raising some doubts, the unemployment data is considered reliable, valid, and representative. This conclusion is based on the alignment of CBS their methodology with the guidelines prescribed by the European Union (Van den Brakel & Krieg, 2010). Moreover, the CBS remains to be the only, and official, source that publishes unemployment data. Consequently, unemployment data that is published by the CBS is often used in the decision making processes of policymakers. That motivates the use of the unemployment data in this study. On the next page, the Google Trends data is discussed in great detail.

A dataset containing the popularity of keywords, that are inputted by users, is taken from Google Trends. To quantify the popularity of keywords, a score between 0-100 is used. This score is known as the search volume index (SVI). A low SVI indicates a low search volume, and thus low popularity, compared to when the search volume was highest. Either weekly or monthly data can be obtained from Google Trends, with the Google Trends data always being monthly when the selected time span is more than 5 years. Google Trends data spanning from April 2008 till December 2022 is collected, which is in line with the unemployment dataset. Formula 2 shows the calculation of the SVI (Tuhkuri, 2016; Google, n.d.).

$$\text{Search Volume Intensity (SVI)}: \left(\frac{\left(\frac{K_{t,i}}{T_{t,i}}\right)}{\max_t\left(\frac{K_{t,i}}{T_{t,i}}\right)}\right) * 100 \qquad (2)$$

K = Number of Google searches for a keyword
T = Total number of Google searches
t = Time unit
i = Geographical unit

Formula 2 demonstrates that the SVI depends on the chosen time span, this is because the SVI is calculated relative to the moment within the chosen time span that the search volume was highest. If the same time span is always used, this should not necessarily put stress on the reliability of the Google Trends data. However, representativeness is at risk as Google Trends data is sampled, yet the sampling error is not disclosed by Google. To overcome the issue of representativeness, this study applied a novel approach that is inspired by the field of online marketing. Precisely, monthly search volume is estimated with the Google Ads Keyword Planner, as search volumes are not disclosed by the Google Trends tool. Any keywords with a too low search volume are considered not representative and are dropped.

As revealed in the literature review, Google Trends data exhibits high inconsistency, and the SVI obtained for the same keyword, during the same period, may vary when obtained at a different time. Consequently, repeating this study with the same Google Trends data, but obtained at a different time, will likely yield slightly different results. This leads to conclude that Google Trends data is fundamentally unreliable. However, by following the suggestion of Eichenauer et al. (2021) the issue of unreliability is mitigated. Eichenauer et al. (2021) suggest to collect the same Google Trends data at twelve random moments, resulting in twelve datasets per keyword. By averaging over all twelve datasets the sampling variation should be reduced with 90%, enhancing the possibility that future replications of this study will yield the same results. Still, the unreliability of Google Trends data remains an important consideration.

Validity is a questionable issue with Google Trends data. The number of searches that are done are likely accurately captured by Google Trends. Therefore, Google Trends data should measure exactly what it aims to measure, the popularity of a keyword. Herein lies no invalidity. However, validity becomes problematic when Google Trends data is related to other phenomena, like unemployment. This is because spurious correlations are easily found with Google Trends data, as shown in the literature review. Consequently, the Google Trends data would not actually measure the phenomena it should measure, thus causing invalidity. To deal with this invalidity, this study starts from a domain ontology, to then only select keywords based on literature and economic intuition. Below, the exact steps are discussed.

First, 26 studies that forecasted unemployment with Google Trends data were selected, to then obtain the keywords these studies used. 329 keywords were obtained and manually checked for fitting into the domain ontology of unemployment forecasting. To get an overview of the keywords, coding was done. This led to the identification of 20 main themes/categories, six of these themes were most prominently used in the studies. These six themes are: (1) job search, (2) unemployment interest, (3) employment agency, (4) job platform, (5) unemployment benefits, and (6) unemployment claims. These six themes were then used to find the Dutch equivalent of keywords that fit to those themes. Additionally, Dutch keywords are taken from Te Brake (2016). Other studies like Tuhkuri (2016) are followed by including some misspelled keywords. Keywords with less than 1,000 monthly searches are removed to ensure representativeness. The aforementioned resulted in only 58 keywords to remain.

Using these 58 keywords, Google's algorithm is prompted to return closely related keywords. The search volumes of the closely related keywords are also checked and after doing so another 21 keywords are added, resulting in a total of 79 keywords that together account for 2,799,400 monthly Google searches. These 2,799,400 monthly searches are not unique searches however, meaning that each search does not represent one individual. Each step of the keyword selection process is documented and shown in appendix 1, enhancing replicability of this study.

Following Eichenauer et al. (2021), the Google Trends data for each keyword is obtained at 12 different moments across 9 days to reduce sampling variation. The IP address used to obtain the Google Trends data has been kept constant in doing so. This resulted in 12 datasets for each keyword, summing to a total of 948 datasets. For each dataset a SVI value of <1 indicates that the search volume is too low to be quantified, whereas a SVI value of 0 indicates there is no search volume at all. Considering a SVI value of <1 still represents some search activity, SVI values of <1 are recoded into 1. The recoding is necessary to satisfy the conditions that statistical software puts on data that is inputted to the software application.

Then, the mean correlation between various Google Trends data samples, for the same keyword but collected at twelve different models was checked. This revealed that in the case of 18 keywords, the mean correlation was .90 or lower. This confirms literature's claim of inconsistency. All keywords with a mean correlation that was lower than .90 were dropped. This was done because practitioners can't be expected to collect Google Trends data at 12 different intervals. Consequently, 63 keywords remained that are deemed fit for analysis. Google Trends data for these 63 keywords is unlikely to significantly fluctuate when collected at different moments, thus practitioners can collect the Google Trends data for these keyword only at a single interval. For each of the 63 keywords that remained, the 12 datasets obtained for these keywords were averaged, as per the suggestion of Eichenauer et al. (2021).

The quality of the Google Trends data is still argued to be questionable, despite the approaches taken to minimize violations of reliability, validity, and representativeness. Yet, the Google Trends data is considered adequate for use in this study. With the discussion of the data coming to an end, the next page discusses the techniques that are used to analyse the unemployment data and the Google Trends data. Starting with the in-sample analysis.

3.3 In-sample analysis

The in-sample analysis of this study consists of the specification of appropriate forecasting models to subsequently use these models for out-of-sample forecasting. Therefore, this section will merely describe the modelling technique that is used, the approach to specify the forecasting model, and the in-sample that is used.

In this study, a linear regression model is used as modelling technique. Linear regression models are simple to interpret and have been used as a starting point for various new approaches (James et al., 2021). In regression analysis, there is a distinction between independent variables and the dependent variable. In forecasting, the independent variables serve as predictor variables, while the dependent variable is the phenomena that is forecasted. In this study, each independent variable represents the Google Trends data for a single keyword. Following literature's suggestion, multiple independent variables are used to forecast youth unemployment. Thus, a multiple linear regression model is used.

The multiple linear regression model, as shown in formula 3, assumes a linear relationship between the independent variables and dependent variable (James et al., 2021). When the relationship is not linear, the regression model struggles to quantify how changes in independent variables translate to variations in the dependent variable. In forecasting, it is crucial that past changes in the independent variables translate to today's value of the dependent variable. Conversely, changes in the independent variable occurring today must translate to the dependent variable's future value. Therefore, when applying regression in the context of forecasting, the independent variables must be lagged before the dependent variables (Hyndman & Athanasopoulos, 2018). In this study, each of the 63 variables are lagged up to 6 months before the dependent variable, resulting in 372 potential independent variables.

In addition to the assumption of linearity, there are other assumptions to ensure a linear regression model is reliable, valid, and representative. The linearity assumptions and all other assumptions are checked with the errors that are obtained from the out-of-sample analysis and are discussed later. The multicollinearity assumption is not considered because it is only relevant when inferencing coefficients, something that is irrelevant in forecasting (Hyndman & Athanasopoulos, 2018; Hardy, 2017).

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i + \varepsilon \tag{3}$$

y = Dependent variable
x = Independent variable
β_0 = Intercept
β_1 = Slope
ε = Error

This section now progresses to the specification of the multiple linear regression model. Drawing from literature, the multiple linear regression model will be specified by selecting independent variables with forward stepwise selection (Hyndman & Athanasopoulos, 2018). The forward stepwise selection is introduced on the next page.

Forward stepwise selection starts with a model containing just the intercept (β_0), known as the null model (James et al., 2021). Subsequently, each independent variable is separately added to the null model and tested on the in-sample, retaining the independent variable with the most predictive value. This retained independent variable, together with the intercept, results in a new model. Consequently, the first "step" is done. Then, each independent variable that is left is again separately added, now to the new model, and tested for predictive value. This process repeats itself until some stopping rule is met (James et al., 2021).

Stopping rules are often based on the size of the in-sample, to avoid overfitting (James et al., 2021). That raises the question of the appropriate window size to use. Remember, the window refers to the in-sample that is used to specify and train the forecasting model. Literature highlights the notable impact of window size on forecasting accuracy, suggesting to average over forecasts done for various window sizes (Pesaran & Timmerman, 2002). This study, in pursuit of robust results, follows this approach and considers window sizes spanning from 48 to 96. This is inspired by a window size of 48 being frequently used by literature (Tuhkuri, 2016; Borup & Schutte, 2022; Te Brake, 2017; Smith, 2016). Consequently, the forward stepwise selection is used to specify a linear regression model for numerous in-samples that vary in size. The forward stepwise selection is stopped after five steps are done, thereby limiting the number of independent variables that are included in the multiple linear regression model to five. Considering the smallest window size will be 48, this stopping rule conforms to the 10k rule of thumb for determining minimum sample in multiple linear regression, as shown in formula 4 (Riley et al., 2018). For the hypothesis testing, a window with a size of 96 will start on 01/10/2008, while a window with a size of 48 will start on 01/10/2012, this is done to ensure the same out-of-sample period is forecasted. The final forecast done, see section 5.3, uses an initial window size of 48 that starts on 01/10/2008.

$$\text{Max. steps:} \quad k = \frac{n}{10} \tag{4}$$

k = number of independent variables
n = sample size used for training

To identify the independent variable with the most predictive value, Akaike's Information Criteria (AIC) is used. This is in line with Mulero & Garciá-Hiernaux (2021) and the suggestion of Hyndman & Athanasopoulos (2018). A low AIC is always preferred over a high AIC, consequently, the independent variables that have the lowest AIC are selected in the forward stepwise selection. AIC is a strictly in-sample metric, evaluating model fit by means of maximum likelihood and penalizing additional parameters with little predictive value (Chatfield & Xing, 2019). Formula 5 shows how AIC is obtained (Chatfield & Xing, 2019).

$$\text{AIC:} \quad -2\ln(\text{ml}) + 2\text{r} \tag{5}$$

r = number of independent variables
ml = maximum likelihood
ln = natural logarithm

This section has outlined how the in-sample analysis will be conducted. However, in-sample analysis and out-of-sample analysis are intertwined in this study, adding complexity. On the next page, the methodology transitions to the out-of-sample analysis, hoping to further contribute to the reader's understanding of this study.

3.4 Out-of-sample analysis

So far, the methodology chapter has addressed the research question, data that is used, and the in-sample analysis. In this section, the focus shifts to the out-of-sample analysis, considered to be the most important aspect in generalizing forecasting models. Before delving exclusively into the out-of-sample analysis, this section first briefly introduces the transition from in-sample analysis to out-of-sample analysis.

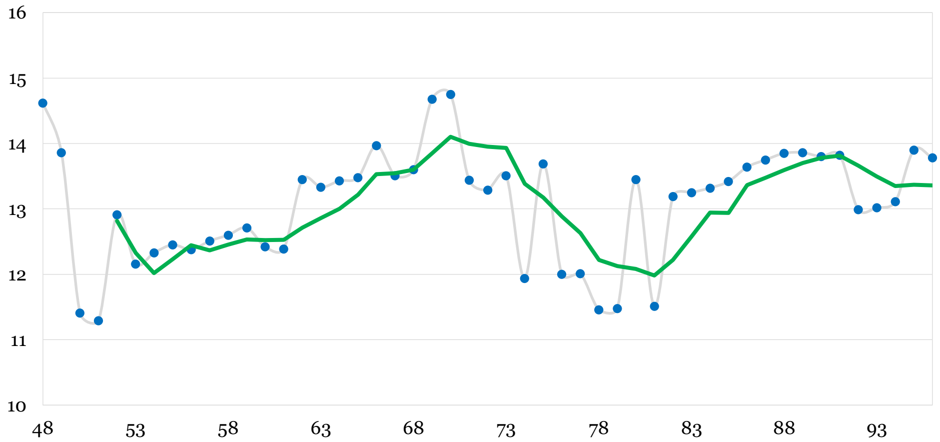
To test the hypotheses drafted for this study, two approaches are employed, both transitioning from in-sample analysis to out-of-sample analysis, to then arrive back to in-sample analysis. In the first approach, the initial window is used to specify the optimal multiple linear regression with forward stepwise selection (in-sample analysis). The specified multiple linear regression model is then used to forecast one month ahead (out-of-sample analysis). Subsequently, when the window is updated with new data, the multiple linear regression model is recalibrated (in-sample analysis). The recalibrated multiple linear regression model is then used to again forecast one month ahead (out-of-sample analysis). This process continues until new data is exhausted. This model recalibration approach is applied to both the rolling window and the expanding window procedure for out-of-sample forecasting. Moving forward, the rolling window is called RC and the expanding window EC.

The second approach involves model respecification. In this approach, the initial window is used to specify the optimal multiple linear regression model with forward stepwise selection (in-sample analysis). The specified multiple linear regression model is then used to forecast one month ahead (out-of-sample analysis). Subsequently, when the window is updated with new data, a new multiple linear regression model is specified with forward stepwise selection (in-sample analysis). The respecified multiple linear regression model is then used to again forecast one month ahead (out-of-sample analysis). This process continues until new data is exhausted. This model respecification approach is applied to both the rolling window and the expanding window procedure for out-of-sample forecasting. Moving forward, the rolling window is called RS and the expanding window ES. Moreover, from now on RC, EC, RS, and ES are referred to as procedures for out-of-sample forecasting.

As mentioned before, various window sizes are used to produce forecasts. Originally, to obtain robust results for the hypothesis testing, forecasts computed with various window sizes would simply be averaged. However, this study decided to also demonstrate the sensitivity of forecasting error (and thus accuracy) to the window size. This sensitivity analysis is done for the RC, EC, RS, and ES procedures for out-of-sample forecasting. Scatterplots are employed to illustrate the relationship between window size and forecast error, as shown in figure 4.

In figure 4, the x-axis represents the window size, while the y-axis represents the corresponding forecasting error. The green line represents the moving average of the relationship. Various metrics to quantify forecasting error exist, all indicating slightly different results. However, for these scatterplots, merely the MAPE is used as a measure of forecasting error because it is the only metric that is cross-study comparable. Other metrics like MSE and RMSE are used later on in the study, and will be discussed on the next page.

Figure 4. Window size X error



Note: Figure shows the sensitivity of forecast errors to window size. The forecasts errors, quantified by MAPE, are shown on the y-axis. The window size is shown on the x-axis. Green line concerns moving average. Out-of-sample is October 2016 till December 2022.

This study employs the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE) to quantify the forecasting error, and thus forecasting accuracy. The use of these metrics aligns with most other forecasting studies (Hyndman & Athanasopoulos, 2018). The MSE and RMSE, both scale-dependent metrics, are able to identify outliers by accentuating large errors (Chicco et al., 2021). RMSE, as opposed to MSE, presents forecast errors in units more closely related to the observed values, helping with interpretation. To obtain both the MSE and the RMSE, formula 6 and 7 are used.

$$\text{MSE:} \quad \frac{1}{m}\sum_{i=1}^{m}(x_i - y_i)^2 \tag{6}$$

$$\text{RMSE:} \quad \sqrt{\frac{1}{m}\sum_{i=1}^{m}(x_i - y_i)^2} \tag{7}$$

m = number of forecasted values
x_i = actual i^{th} value
y_i = forecasted i^{th} value

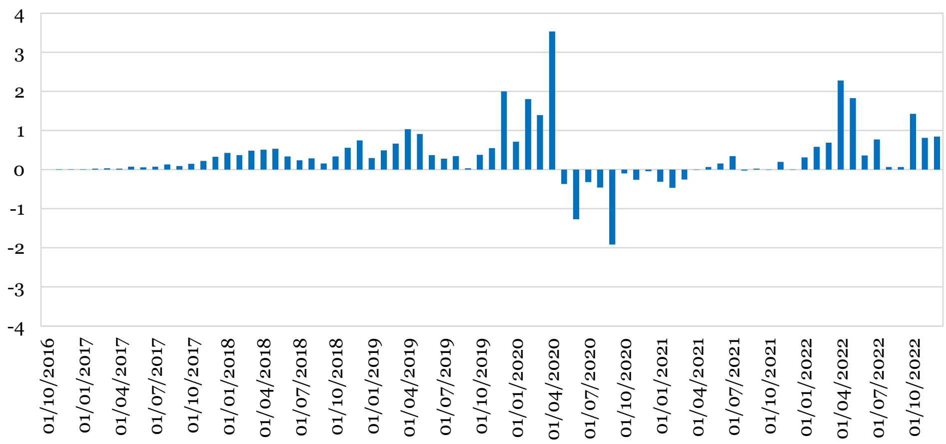
MAPE quantifies the forecasting error with a percentage, making it more intuitive to interpret and allowing cross-study comparison (Chicco et al., 2021). For example, a MAPE of 5% indicates that, on average, the deviation between forecast and reality is 5%. Outliers are harder to identify with MAPE because equal weight is put on all errors. Moreover, MAPE can be sensitive to values close to zero (Hyndman & Athanasopoulos, 2018), though this is not a concern in this study. To obtain the MAPE, formula 8 is used (Chicco et al., 2021).

$$\text{MAPE:} \quad \frac{1}{m}\sum_{i=1}^{m}\left|\frac{y_i - x_i}{y_i}\right| \tag{8}$$

m = number of forecasted values
x_i = actual i^{th} value
y_i = forecasted i^{th} value

Out-of-sample forecasts for the period from October 2016 till December 2022 are produced 49 times with each of the out-of-sample forecasting procedures (RC, ES, RS, ES), corresponding to the window sizes ranging from 48 to 96. After averaging over the 49 forecasts for the out-of-sample October 2016 till December 2022, a single robust forecast is obtained for each out-of-sample forecasting procedure. The forecast errors from the four final forecasts are used in the hypothesis testing. However, before the forecast errors are subjected to hypotheses tests, they are visually analysed with some figures. In relation to the hypotheses, the figures that visualize the loss differential of the forecast errors are most significant. The loss differential demonstrates how, over time, the errors of one forecast compare to the errors of a competing forecast. Consequently, the superior forecast can be found. Figure 5 shows an example, where the x-axis represents time, and the y-axis represent the loss differential.

Figure 5. Loss differential



Note: Figure shows the loss differential of the forecast errors for two competing forecasts over time. Time units are shown on the x-axis. The loss differential of the forecast errors per time unit is shown on the y-axis.

Various loss differentials can be used to quantify the superiority of one forecast over another. Most commonly used is the loss differential of the squared forecasts errors, as it aligns with the loss differential that is used in the Diebold-Mariano test (Diebold, 2015). The Diebold-Mariano test is also used in this study, the next paragraph discusses this test in more detail. To obtain the loss differential of the squared forecast errors for the hypotheses of this study, formula 9 is used (Diebold, 2015). For each hypothesis a loss differential is obtained and visualized. Taking as an example hypothesis one, the loss differential is obtained by subtracting the squared forecast errors produced by the RC out-of-sample forecasting procedure, from the squared forecast error produced by the EC out-of-sample forecasting procedure. Positive values then show that the RC procedure for out-of-sample forecasting produces more accurate forecasts, while negative values show it does not produce more accurate forecasts.

Loss differential: $e_{1t}^2 - e_{2t}^2$ (9)

e_{1t}^2 = squared error of forecast hypothesized to be inferior at time t
e_{2t}^2 = squared error of forecast hypothesized to be superior at time t

To formally test this study's hypotheses, the Diebold-Miarano test is performed under the null hypothesis of equal forecasting accuracy (Diebold, 2015). To perform an unbiased Diebold-Mariano test, assumption DM must be satisfied. Assumption DM states that the loss differential must exhibit covariance stationarity, meaning that the mean and variance are stable over time (Diebold, 2015; Box et al., 2015). In line with literature, the Augmented Dickey-Fuller (ADF) test is used to test for covariance stationarity (Rothman, 1998; Vicente et al., 2015; Barreira et al., 2013). Following Simionescu & Cifuentes (2022) and Vicente et al. (2015) the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test is also used.

The ADF test conducts a hypothesis test under the null hypothesis of, the time series has a unit root (Brockwell, 2016). This is considered a test for stationarity because the presence of unit root indicates non-stationarity (Chatfield & Xing, 2019). To fulfil assumption DM, the null hypothesis of the ADF test must be rejected. To reject the null hypothesis with 95% of confidence, the test statistic must be smaller than the -1.95 critical value. To perform the ADF test the lag order must be chosen, the lag order helps to account for autocorrelation. In this study the lag order was chosen based on AIC, aligning with Barreira et al. (2019).

The KPSS test performs a hypothesis test under the null hypothesis of, the time series is stationary (Kwiatkowski et al., 1992). To fulfil assumption DM, the null hypothesis of the KPSS test must be accepted. To accept the null hypothesis with 95% of confidence, the test statistic must be smaller than the .463 critical value (Pfaff, 2022). In choosing the lag order for the KPSS test, Kwiatkowski et al. (1992) their suggestion to choose the lag order based on sample size is followed, the appropriate lag order is obtained with formula 10.

Lag order: $$\sqrt[4]{4 * \left(\frac{n}{100}\right)} \tag{10}$$

n = sample size

Once assumption DM is satisfied, the Diebold-Mariano test is used to formally test the hypotheses. In addition to the Diebold-Mariano test, the non-parametric Wilcoxon Signed Rank test is utilized. The Wilcoxon Signed Rank test has a null hypothesis of, the difference between paired observations is zero. The Wilcoxon Signed Rank test has no assumptions, reducing the possibility of obtaining a biased test statistic. Other studies like Terregrossa & Ibadi (2021) and Yu & Schwartz (2006) followed a similar approach.

After performing the out-of-sample analysis and testing all hypotheses, the combination of the most optimal out-of-sample forecasting procedure and window size were identified. Using this knowledge, a forecast for a larger out-of-sample was performed and the errors of this forecast were subjected to extensive scrutiny. The methodology that was followed to analyse these forecast errors is discussed in the next section.

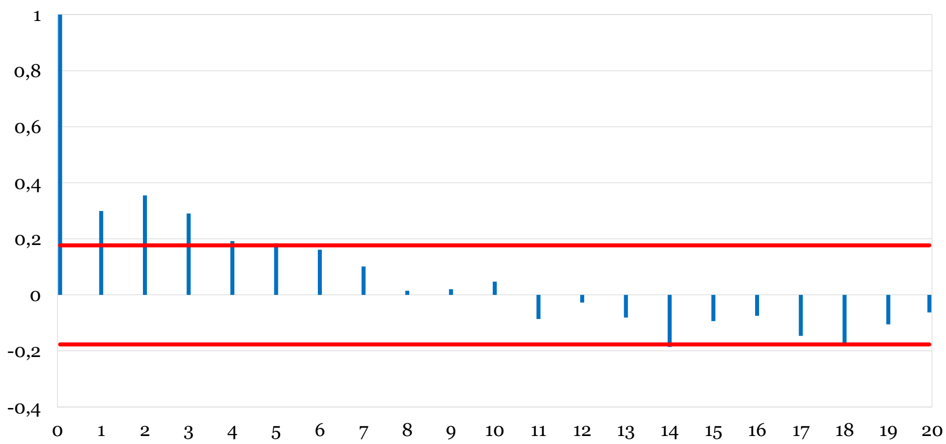
3.5 Error diagnostics

This section addresses how error diagnostics were performed to detect biases in the forecast produced by the optimal out-of-sample forecasting procedure and window size. First, the out-of-sample period that is used is discussed. Subsequently, this section discusses how the assumptions of multiple linear regression were leveraged to assess the forecast errors.

The out-of-sample that is subjected to error diagnostics spans from 01/10/2012 till 01/12/2022. The initial in-sample spans from 01/10/2008 till 01/12/2012, obviously the in-sample period will vary by either expanding or rolling the window. The main objective of the error diagnostics was to investigate whether the optimal out-of-sample forecasting procedure and optimal window size introduce any biases. To do so, the assumptions of multiple linear regression are used, being: (1) uncorrelatedness, (2) homoskedasticity, (3) approximately linear, (4) no outliers, (5) normally distributed (James et al., 2021; Hair et al., 2018).

The assumption of uncorrelatedness refers to the abundance of autocorrelation in the forecast errors. Autocorrelation, in this context, is the linear relationship between a forecast error obtained today and forecast errors obtained in the past (Hyndman & Athanasopoulos, 2018). When the errors exhibit autocorrelation it suggests that the data is not optimally leveraged by the forecasting model (Hyndman & Athanasopoulos, 2018). To detect autocorrelation, plots for the autocorrelation function (ACF) and the partial autocorrelation function (PACF) are used. These plots represent the time lag on the x-axis and the correlation on the x-axis. Taking figure 6 as an example, the .30 correlation at time lag one implies that the forecast error for the current month has a .30 correlation with the forecast error from the previous month. Also, the correlation is significant as it exceeds the upper (red) significance band.

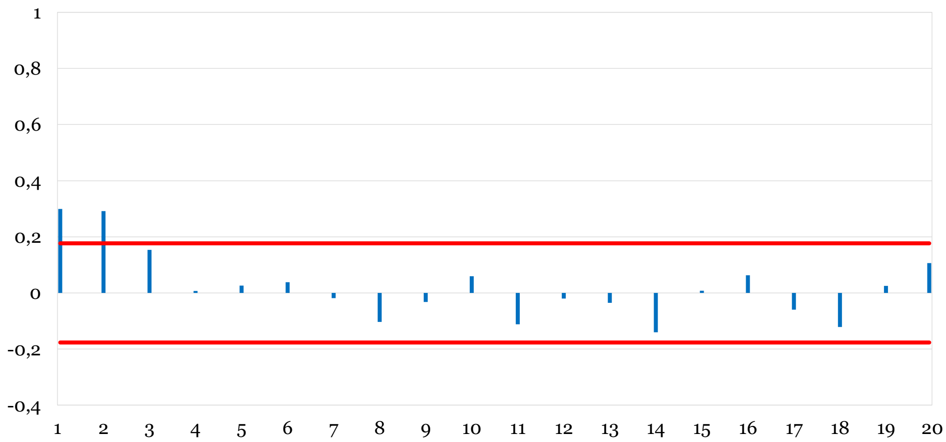
Figure 6. ACF



Note: Figure 6 shows the autocorrelation of the forecast errors as measured by ACF. Time lags are shown on the x-axis. The correlation coefficient for a time lag is shown on the y-axis. Red lines represent significance.

The difference between ACF and PACF is grounded in the inclusion of intermediate correlations. ACF does account for intermediate correlations, for example, when calculating the correlation at lag two, ACF also considers the correlation at lag one (Flores et al., 2012). Contrasting this, PACF calculates correlation only from the direct effect at a certain time lag (Flores et al., 2012). Comparing figure 6 with figure 7 shows this difference. Figure 6 shows an autocorrelation at lag one that is similar to figure 7, the autocorrelation at lag two is higher however in figure 6, this is due to ACF also considering the intermediate correlation of lag one.

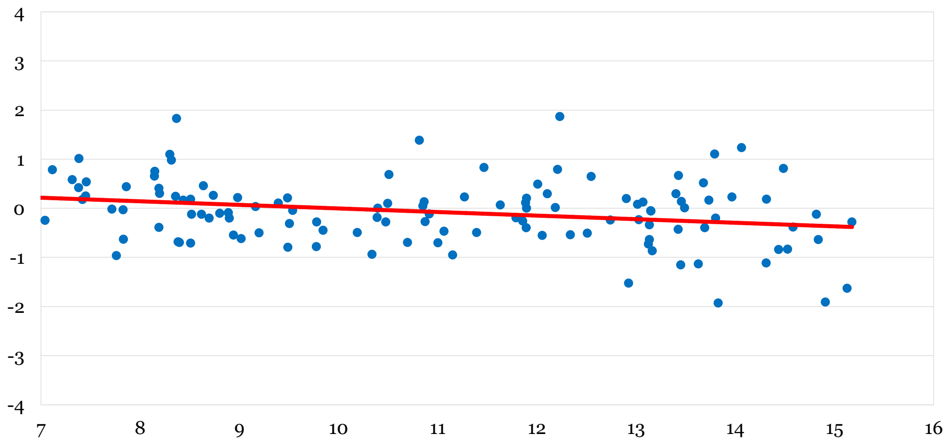
Figure 7. PACF



Note: Figure 7 shows the autocorrelation of the forecast errors as measured by PACF. Time lags are shown on the x-axis. The correlation coefficient for a time lag is shown on the y-axis. Red lines represent significance.

The second assumption is that of homoscedasticity, this assumption refers to an equal variance of errors across the entire range of values the dependent variable takes (Hair et al., 2018). This assumption is checked with an error plot, often called a residual plot. (James et al., 2021). Figure 8 shows an error plot, representing the forecasted values on the x-axis and the forecast errors on the y-axis. Significant variation in the spread of errors implies that the dependent variable is not always predicted with similar accuracy (Hair et al., 2018; Hyndman & Athanasopoulos, 2018). Consequently, the assumption of homoscedasticity is violated.

Figure 8. Error plot



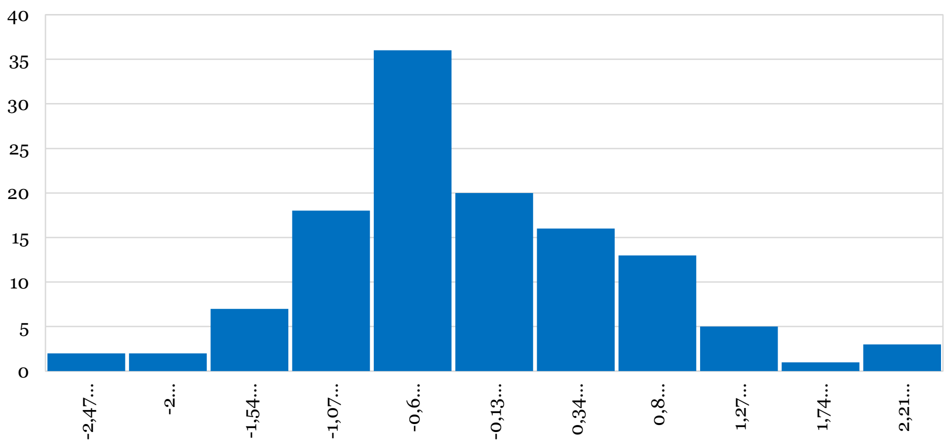
Note: Figure shows the error plot. The forecasted values are shown on the x-axis. The forecasting errors are shown on the y-axis. Each blue dot is an forecasting error x forecasted value combination. Red line depicts linear relationship.

The third assumption is that of linearity. In the context of multivariate approaches, it is a common practice among statisticians to check the linearity assumption post-hoc with an error plot (James et al., 2021). This study follows the same approach. If the relationship between forecasting errors and forecasted values is non-linear, the effect of the independent variables on the dependent variable is not well captured by linear regression (James et al., 2021). Thus, the assumption of linearity would be violated and other modelling techniques or transformation techniques should be applied (James et al., 2021). Additionally, certain variables that cause non-linearity could be omitted (James et al., 2021).

The error plot is also used to investigate the outlier assumption (James et al., 2021; Hyndman & Athanasopoulos, 2018). Outliers are data points that deviate significantly from the distribution of other data points. The strict definition of outliers are incorrectly recorded observations during data collection (James et al., 2021). This study is unlikely to identify outliers that conform to this definition. However, identifying forecast errors that deviate substantially from others could help to identify weaknesses and strengths of the forecast.

Finally, there is the assumption that forecast errors follow a normal distribution. However, Hyndman & Athanasopoulos (2018) argue that deviations from non-normality do not invalidate forecasts and that it's only important for statistical tests or confidence intervals. Checking normality can help to better understand the model performance however. For example, non-normality could indicate that the forecasting model is systematically underestimating or overestimating. The most commonly applied techniques to investigate normality is the histogram (Hyndman & Athanasopoulos, 2018). Figure 9 show an histogram where the x-axis represents ranges of values, called bins. The y-axis represents the frequency with which a value corresponding to the bin's value range is present in the data. In forecasting, the bins contain ranges of forecast errors. For example, the first bin in figure 9 illustrates that there are two instances where the forecast error lies between -2.47 and -2.

Figure 9. Histogram



Note: Figure shows the histogram of the forecast errors. The bins are shown on the x-axis.
The frequency with which values that fit within a bin occur are shown on the y-axis.

Histograms can be biased, by changing the bin width the histogram may start to appear more normally distributed. Therefore, normality is also assessed with the Shapiro-Wilk test under the null hypothesis of normality (Hair et al., 2018). Also, the skewness value of the forecast errors is obtained, where a value falling outside the range of -1 to +1 indicates non-normality (Hair et al., 2018). The skewness value is obtained with formula 11 (Bono et al., 2019).

$$\text{Skewness:} \quad \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{s}\right)^3 \qquad (11)$$

n = sample size
x_i = i'th value of x
$\bar{x}$ = mean of x
s = standard deviation of x

After diagnosing the forecast errors as produced by the optimal procedure for out-of-sample forecasting and optimal window size, the multiple linear regression model that is used in this procedure is extended with an autoregressive (AR) component. An AR component simply refers to a lagged version of the dependent variable that is used to forecast the dependent variable (Hyndman & Athanasopoulos, 2018). Specifically, an AR(1) component is added to the regression model, as is done by many other studies like D'Amuri & Marucci (2017), Mclaren & Shanbogue (2011), Tuhkuri (2016). Adding this AR(1) component shows whether assumptions that have been violated, can be improved upon after adding the AR component. Also, it demonstrates whether the results of this study can easily be extended to an AR model.

Hoping to have enhanced the readers understanding of this study, and allowing replication of this study, the methodology chapter is concluded and the next page presents the result chapter. The results chapters promises to unravel novel insights that are useful for practitioners and researchers around the world.

4. Results

This chapter presents all the findings of this study. First, the sensitivity of forecasting accuracy to the window size that is used in out-of-sample forecasting procedure is explored. Then, the forecasts produced by various out-of-sample forecasting procedures are visualized and assessed in terms of accuracy. Subsequently, the forecasts are used in hypothesis tests to find support or rejection for this study's hypotheses. Finally, the most accurate forecasts that this study was able to identify is used to perform error diagnostics on.

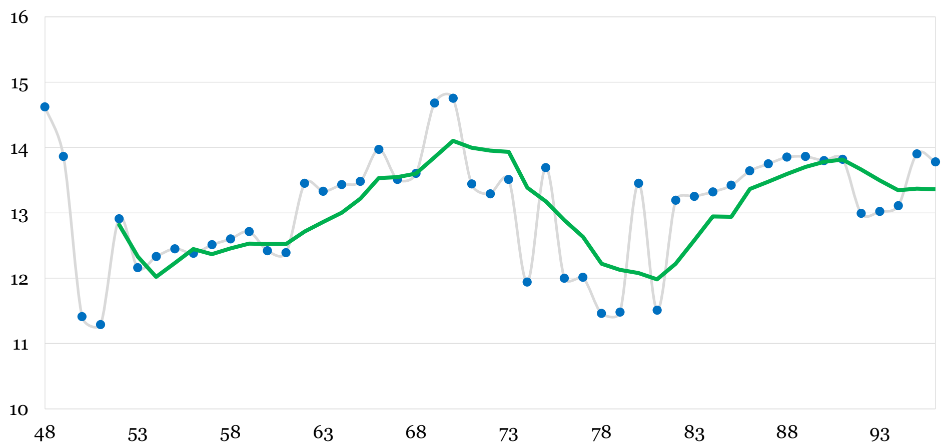
4.1 Window sensitivity

In the section, the sensitivity of forecasting accuracy to the window size is investigated. This sensitivity analysis is performed for each out-of-sample forecasting procedure outlined in this study, being: the rolling window with calibration (RC), the expanding window with recalibration (EC), the rolling window with respecification (RS), and the expanding window with respecification (ES). The first sensitivity analysis focuses on the RC procedure for out-of-sample forecasting, as depicted below. It is worth noting that the out-of-sample period used for each window size is October 2016 till December 2022, while the in-sample period varies.

Figure 10 illustrates that there is no linear relationship between the window size and the forecast error produced by the RC procedure for out-of-sample forecasting. Instead, figure 10 reveals that initially, there is a positive and almost linear relationship between window size and forecasting error. This positive relationship transform into a negative one when the window size exceeds 73, followed by a return to a positive relationships for window sizes larger than 81. The relationship between window size and forecasting error is not necessarily cyclic, nor can it be described as a sine wave due to inconsistent wave magnitudes.

Although the relationship is hard to characterize, figure 10 reveals that the forecast error is notably sensitive to the window size. For example, a window size of 51 results in a MAPE of 11,29%, whereas increasing the window size to 70 leads to a MAPE almost 4% higher. Additionally, figure 10 indicates that there are clusters of closely packed together data points, indicating that windows of more or less equal size are unlikely to produce different results.

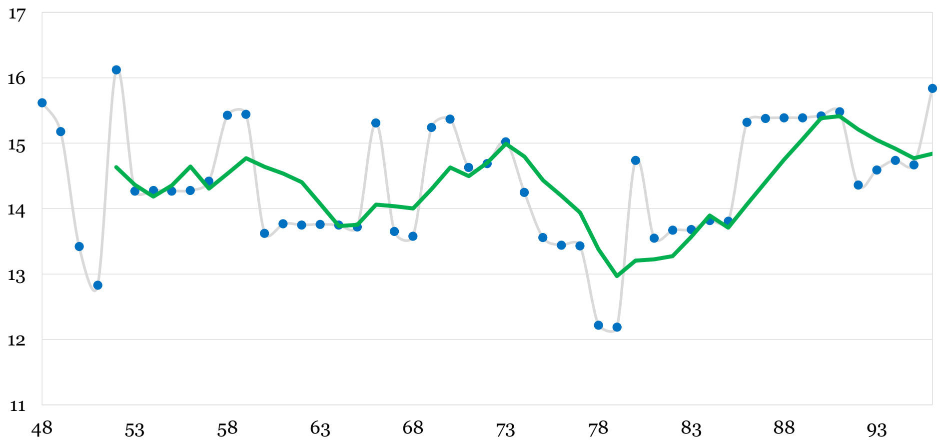
Figure 10. Window size X forecast error | RC



Note: Figure shows the sensitivity of forecast errors to window size. The forecast errors, quantified by MAPE, are shown on the y-axis. The window size is shown on the x-axis. Green line concerns moving average. Out-of-sample is October 2016 till December 2022.

Contrasting figure 10, figure 11 outlines that forecast errors respond slightly different to window size when using the EC procedure for out-of-sample forecasting. Notably, there is a difference for window size 48 till 73. While figure 11 exhibits a wave-like relationship within this window size range, the relationship was positive and nearly linear in figure 10. However, from a window size of 73 onwards, figure 10 and figure 11 align quite well, even in terms of a similar discrepancy between the lowest forecast error and highest forecast error. Figure 11 offers few other insights, leading to the sensitivity analysis for respecification approaches.

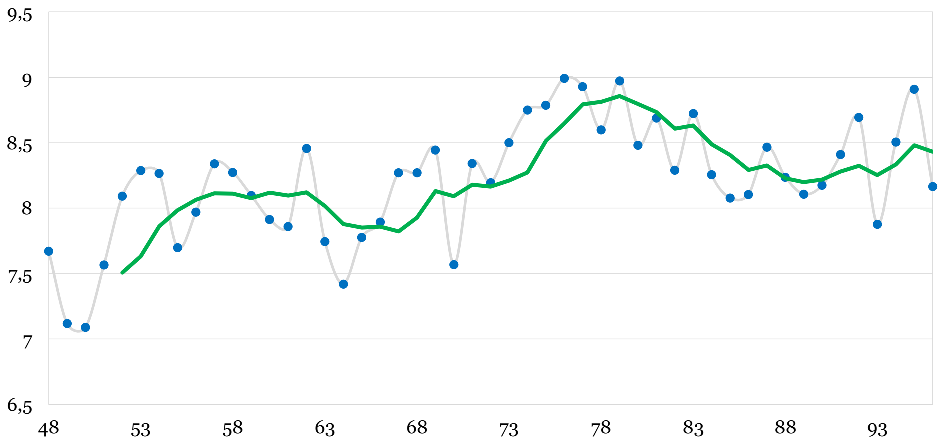
Figure 11. Window size X forecast error | EC



Note: Figure shows the sensitivity of forecast errors to window size. The forecast errors, quantified by MAPE, are shown on the y-axis. The window size is shown on the x-axis. Green line concerns moving average. Out-of-sample is October 2016 till December 2022.

Figure 12 reveals that the relationship between window size and forecast error is quite different compared to figure 10 or figure 11. The forecast errors almost consistently increases until the window size becomes larger than 78, after which the forecast errors decline very slightly. Also, figure 12 reveals that the discrepancy between the lowest forecast error and the highest forecast error is much smaller, with a difference of about 2%. This may be due to the forecast error generally being much smaller with the RS procedure for out-of-sample forecasting. Moreover, no clusters of closely packed together data points are observable in figure 12.

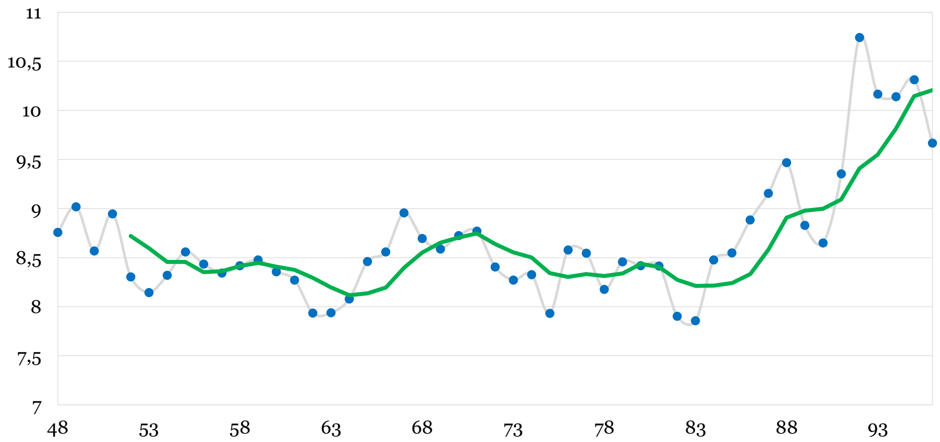
Figure 12. Window size X forecast error | RS



Note: Figure shows the sensitivity of forecast errors to window size. The forecast errors, quantified by MAPE, are shown on the y-axis. The window size is shown on the x-axis. Green line concerns moving average. Out-of-sample is October 2016 till December 2022.

Arriving at the last out-of-sample forecasting procedure that is subjected to a sensitivity analysis, figure 13 reveals a pattern distinct from the previously discussed ones. While the window size impacts forecast errors for window sizes between 48 and 83, the impact is almost half as strong than observed in figure 12. However, as soon as the window size becomes larger than 83, the forecast errors increase at a pace that is unmatched by the previously discussed procedures for out-of-sample forecasting. Consequently, the discrepancy between the lowest forecast error and highest forecast error is higher when compared to figure 12.

Figure 13. Window size X forecasting error | ES



Note: Figure shows the sensitivity of forecast errors to window size. The forecast errors, quantified by MAPE, are shown on the y-axis. The window size is shown on the x-axis. Green line concerns moving average. Out-of-sample is October 2016 till December 2022.

Before drawing any major conclusions, it is evident from these results that the forecast error, and thus accuracy, is quite sensitive to the window size that is used. However, the relationship between window size and forecast error is complex to comprehend, and can hardly be captured with an equation. Moreover, for each out-of-sample forecasting procedure, the relationship between window size and forecast error is different. Not only does the pattern differ, but the magnitude of the effect is also different. This has led to the decision to average over the 49 different windows that were used in producing forecasts with an out-of-sample forecasting procedure. In the next section, the accuracy of these averaged forecasts are addressed, and the hypotheses of this study are put to the test.

4.2 Hypothesis testing

This section presents the accuracy of forecasts that are produced by various out-of-sample forecasting procedure. To begin, the accuracy of the forecasts is visualized with figures, highlighting differences between the various forecasts. Subsequently, error metrics that quantify the forecasting accuracy are provided, before all hypotheses of this study are tested.

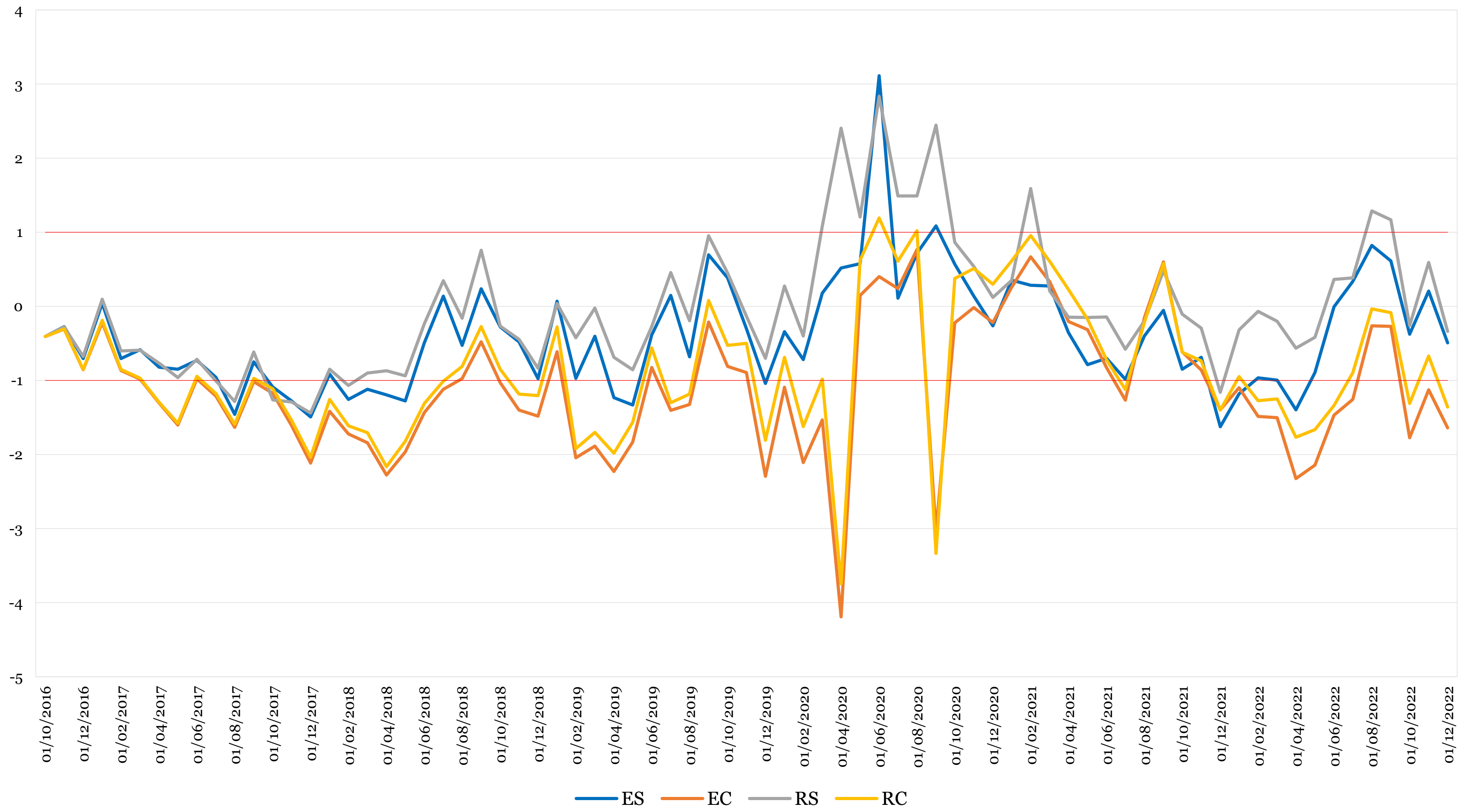
Figure 14, presented on the next page, displays the forecast errors of the (averaged) forecasts produced by each out-of-sample forecasting procedure. The forecast errors given in figure 11 are unadjusted, precisely indicating how much a forecasted youth unemployment rate deviates from the actual youth unemployment rate. For example, an error of 1 means that the forecast predicted youth unemployment rate to be 11%, while in reality, it was 12%.

Figure 14 illustrates that the two out-of-sample forecasting procedures relying on model recalibration produce very similar forecasts errors, with the RC procedure for out-of-sample forecasting being slightly more accurate. The forecast errors produced by both out-of-sample forecasting procedures relying on model respecification behave slightly different to each other. Substantial differences can be observed in figure 14 when comparing an out-of-sample forecasting procedure relying on model recalibration with the same one but relying on model respecification. For example, figure 14 shows that the rolling window relying on model recalibration (RC) produced forecast errors very different to the same rolling window but relying on model respecification (RS).

The fact that both out-of-sample forecasting procedures relying on model recalibration produced very similar forecast errors is easily explained. Both the EC and RC procedure for out-of-sample forecasting use the same variables in producing forecasts. Yet, precisely the variables that are used tend to strongly impact forecasting error. The only difference is that the EC procedure for out-of-sample forecasting potentially captures more noise. On the other hand, the model recalibration on a small and fixed window, as used in the RC procedure for out-of-sample forecasting could be suboptimal by pruning valuable data.

Generally, figure 14 illustrates that the forecast errors produced by each out-of-sample forecasting procedure are reasonably stable. The period characterized by COVID-19 has distorted the results however, it is precisely this period where major discrepancies between the various out-of-sample forecasting procedure are apparent. Surprisingly, the ES procedure for out-of-sample forecasting responded more adequately to start of the COVID-19 period than the RS procedure for out-of-sample forecasting. Moving to the loss differentials on page 39, the results chapter will now demonstrate more concisely which out-of-sample forecasting procedures are superior in terms of forecasting accuracy.

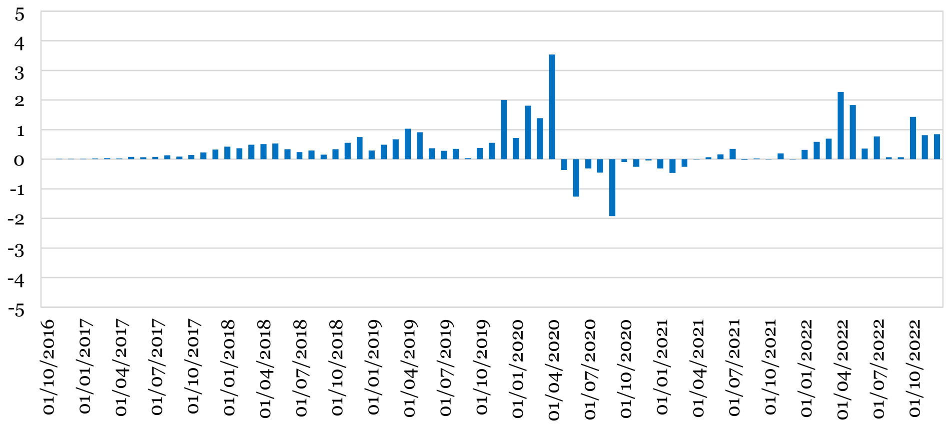
Figure 14. Forecast error for all forecasts



Note: This figure shows the raw, unadjusted, forecast errors that are obtained from each out-of-sample procedure for forecasting. Time units are represented on the x-axis. The corresponding forecasting errors for a time unit is represented on the y-axis. The out-of-sample is October 2016 till December 2022, the in-sample varies as this concerns forecasts that are averaged over multiple window sizes. The red horizontal line is merely used for visual assistance and does not concern a significance level of any sorts.

Figure 15 shows that the RC procedure for out-of-sample forecasting produces more accurate forecasts for a majority of the months, when compared to the EC procedure for out-of-sample forecasting. Unexpectedly, the period characterized by COVID-19 changed this, and the EC out-of-sample forecasting procedure became superior in terms of forecasting accuracy. This is surprising, as the rolling window used should be more capable at adapting to sudden changes than the expanding window. Another peculiarity, shown in figure 15, is that both out-of-sample forecasting procedures produced forecasts with similar accuracy for the first twelve months. That leads to believe that it takes about 12 one month ahead forecasts before the rolling window starts capturing data so different, that it affects forecasting accuracy.

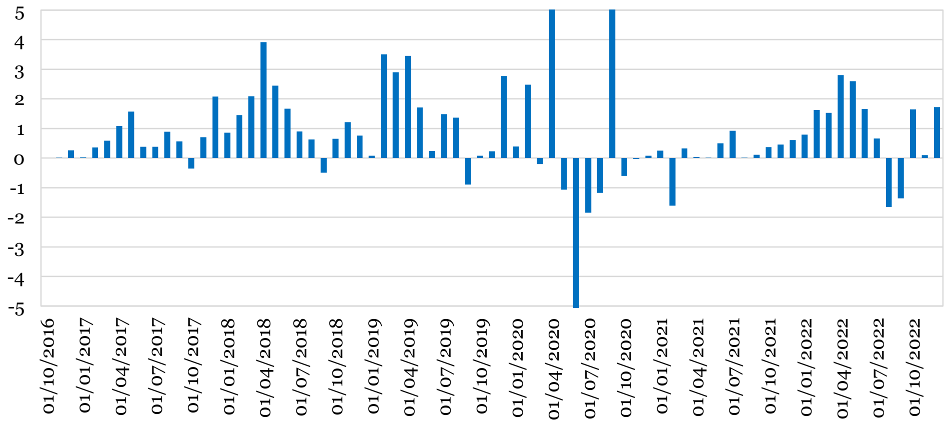
Figure 15. Loss differential H1 | EC-RC



Note: Figure shows the loss differential of the squared forecast errors for hypothesis one. Squared forecast errors produced by RC are substracted from those produced by EC. Positive values indicate RC is more accurate. Time is shown on the x-axis. Loss differential is shown on the y-axis. Out-of-sample is October 2016 till December 2022.

Figure 16 illustrates that the RS procedure for out-of-sample forecasting produces more accurate forecasts for nearly all the months, when compared to the RC procedure for out-of-sample forecasting. Moreover, the improvement is quite large. Not observable in figure 16 but shown in figure 14, the forecasts produced by the RS out-of-sample forecasting procedure were more than triple as accurate for some months. Notably, for the month 01/06/2020, the RC out-of-sample forecasting procedure was much more accurate.

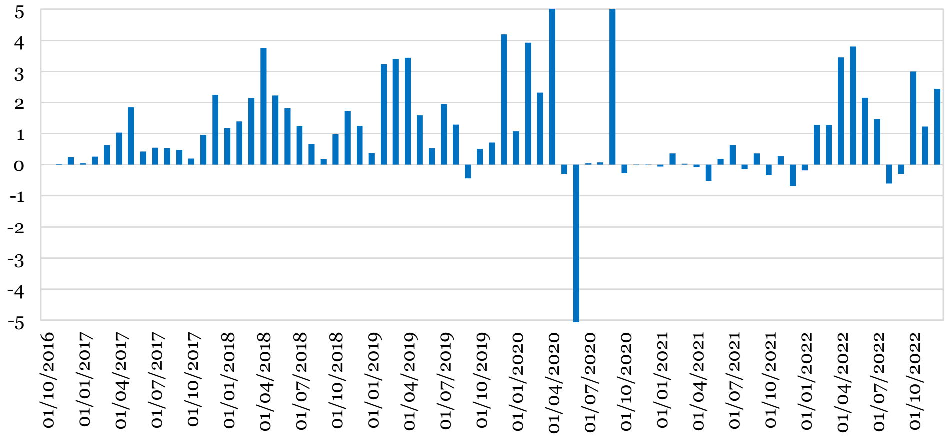
Figure 16. Loss differential H2 | RC-RS



Note: Figure shows the loss differential of the squared forecast errors for hypothesis one. Squared forecast errors produced by RS are substracted from those produced by RC. Positive values indicate RS is more accurate. Time is shown on the x-axis. Loss differential is shown on the y-axis. Out-of-sample is October 2016 till December 2022.

Similar to figure 16, figure 17 demonstrates that model respecification is preferred over model recalibration, this time showing it for an expanding window procedure for out-of-sample forecasting. Like figure 16, figure 17 shows that the forecast for 01/06/2020 was substantially less accurate when using model respecification as opposed to model recalibration. Different from figure 16, the improvement of model respecification over model recalibration appears more consistent in figure 17. For the period starting 01/06/2020 and ending around 01/12/2021, the produced forecasts are similarly accurate.

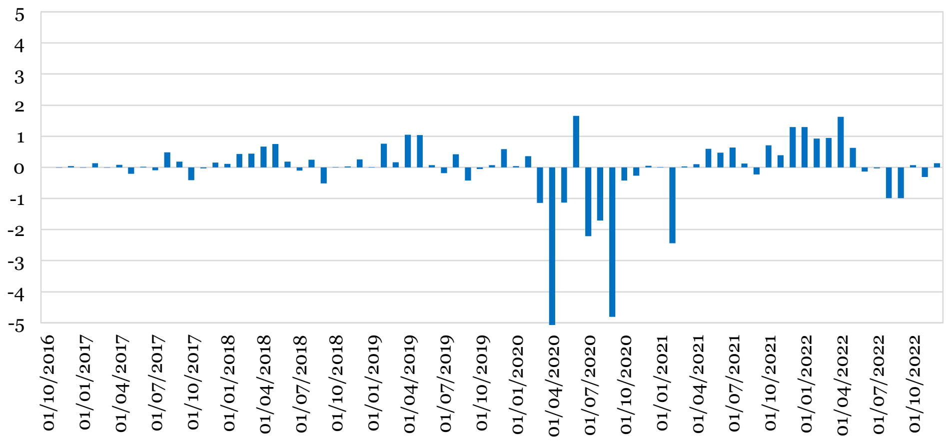
Figure 17. Loss differential H3 | EC-ES



Note: Figure shows the loss differential of the squared forecast errors for hypothesis one. Squared forecast errors produced by ES are substracted from those produced by EC. Positive values indicate ES is more accurate. Time is shown on the x-axis. Loss differential is shown on the y-axis. Out-of-sample is October 2016 till December 2022.

The last loss differential discussed in this study is visualized in figure 18. Figure 18 shows that the RS procedure for out-of-sample forecasting very slightly improves upon the ES procedure for out-of-sample forecasting. Yet, this improvement is offset by the ES procedure for out-of-sample forecasting producing substantially more accurate forecasts for the period 01/03/2020 till 01/02/2021, leading to believe that, on average, similarly accurate forecasts are produced.

Figure 18. Loss differential H4 | ES-RS



Note: Figure shows the loss differential of the squared forecast errors for hypothesis one. Squared forecast errors produced by RS are substracted from those produced by ES. Positive values indicate RS is more accurate. Time is shown on the x-axis. Loss differential is shown on the y-axis. Out-of-sample is October 2016 till December 2022.

All previously discussed loss differentials reveal that there is not a single instance where one out-of-sample forecasting procedure consistently produced forecasts that were more accurate than those produced by another out-of-sample forecasting procedure. This inconsistency appears to occur due to the period 01/05/2020 till 01/01/2021, characterized by great instability and uncertainty due to COVID-19. In general, all loss differentials show that when a procedure for out-of-sample forecasting produced superior forecasts for the period 01/10/2016 till 01/04/2020, it also produced inferior forecasts for the period 01/05/2020 till 01/2021. This demonstrates the impact of COVID-19 on forecasts. Below, various error metrics are given that present the obtained forecasting error more concisely.

Table 2 shows various error metrics that quantify the average forecast error produced by an out-of-sample forecasting procedure. Like demonstrated in the loss differentials, table 2 reveals that out-of-sample forecasting procedures relying on model respecification result in substantially more accurate forecasts than those relying on model recalibration. Thus, some evidence is presented that already points to support for hypothesis two and three.

Regarding hypothesis one, table 2 shows that the forecasts produced by the RC procedure for out-of-sample forecasting are more accurate than those produced by the EC procedure for out-of-sample forecasting. However, the improvement is small, as measured by RMSE and MAPE, while being much larger when measured with MSE. Most likely, this is due to the squaring done in the MSE, thus accentuating a few large improvements. In any case, table 2 provides some evidence that points to the support of hypothesis one.

Figure 18 has demonstrated that the RS and ES procedures for out-of-sample forecasting produce very similar forecast errors. Table 2 confirms this by showing that the MSE, RMSE and MAPE error metrics barely differ between both out-of-sample forecasting procedures. However, on the one hand, MAPE points to the support of hypothesis four by indicating that the RS procedure for out-of-sample forecasting is more accurate. On the other hand, RMSE and MSE oppose hypothesis four by indicating that the ES procedure for out-of-sample forecasting is more accurate. The discrepancy between the error metrics is due to large errors being accentuated by MSE and RMSE. While some evidence is already provided, the formal tests of the hypotheses are shown on the next page.

Table 2. Error metrics

	MSE n=75	RMSE n=75	MAPE n=75
ES	.7302	.8545	7.807%
EC	1.936	1.391	13.34%
RS	.7816	.8841	7.213%
RC	1.589	1.260	11.94%

Note: Table shows metrics for the one month ahead forecasts produced with the different out-of-sample forecasting procedures. Metrics are calculated from the forecast errors that are averaged over each window spanning from 48 till 96. The out-of-sample based is October 2016 till December 2022. The in-sample varies.

Before testing the hypotheses with the Diebold-Mariano test, each hypothesis must have a loss differential of the squared forecast errors that is covariance stationary. To this end, the ADF-test and KPSS-test are used, the results of which are given in table 4. The ADF-test and KPSS-test proof, with 95% of confidence, that the loss differential of the squared forecast error for hypothesis one, two, and three are stationary. The loss differential for hypothesis four is considered non-stationary by the ADF-test. This is not problematic, as discussed below.

To test the null hypothesis of equal forecasting accuracy, the Diebold-Mariano test uses the squared forecast errors. However, the MSE, which also uses squared forecast errors, has already shown the exact opposite of hypothesis four becomes a reality when squared forecast errors are used to test for equal forecasting accuracy. Hypothesis four will be tested with the Wilcoxon Signed Rank test, which uses the raw forecast errors. MAPE has indicated that when the raw forecast errors are used, it points into the direction of support for hypothesis four.

Table 3. Stationarity tests

	Augmented Dickey-Fuller n=75		Kwiatowski-Phillips-Schmidt-Shin n=75	
	t	lag	t	lag
Hypothesis four	-2.965***	1	.1291	3
Hypothesis one	-4.349***	1	.1272	3
Hypothesis three	-2.107**	4	.0890	3
Hypothesis four	-1.771*	4	.1473	3

Note: Table shows the results of the tests for stationarity. The ADF-test is performed under the null hypothesis of the timeseries having unit root. Stationarity is assumed when the null hypothesis of the ADF-test is rejected with 95% of confidence, to do so the test statistic (t) must be further away than the -1.95 critical value. The KPSS-test is performed under the null hypothesis of stationarity. Stationarity is assumed when the null hypothesis of the KPSS-test is not rejected with 95% of confidence, to do so the test statistic must be smaller than the .463 critical value. Significance is indicated with asterisks, where ***p<.01, **p<.05, and *p<.1.

With the loss differential of the squared forecast errors for hypothesis one, two, and three considered stationarity, the Diebold-Mariano test is now applied. Additionally, all hypotheses are tested with the Wilcoxon Signed Rank test. The results of both tests are shown in table 4. Both the Diebold-Mariano test and the Wilcoxon Signed Rank test provide evidence to accept hypothesis one, two, and three with 99% of confidence. The support for hypothesis one indicates that when model recalibration is used, that the rolling window produces more accurate forecasts than the expanding window. The support for hypothesis two and three indicates that in the case of the rolling window and expanding window, the use of model respecification results in more accurate forecasts than the use of model recalibration.

According to the Wilcoxon Signed Rank test, hypothesis four can’t be supported with 95% of confidence, although it can be accepted with 94% of confidence. Consequently, there is only weak evidence that when model respecification is used, that the rolling window produces more accurate results than the expanding window. Essentially, no conclusive evidence is found for either the rejection or support of hypothesis four, although support appears more likely.

Table 4. Hypothesis tests

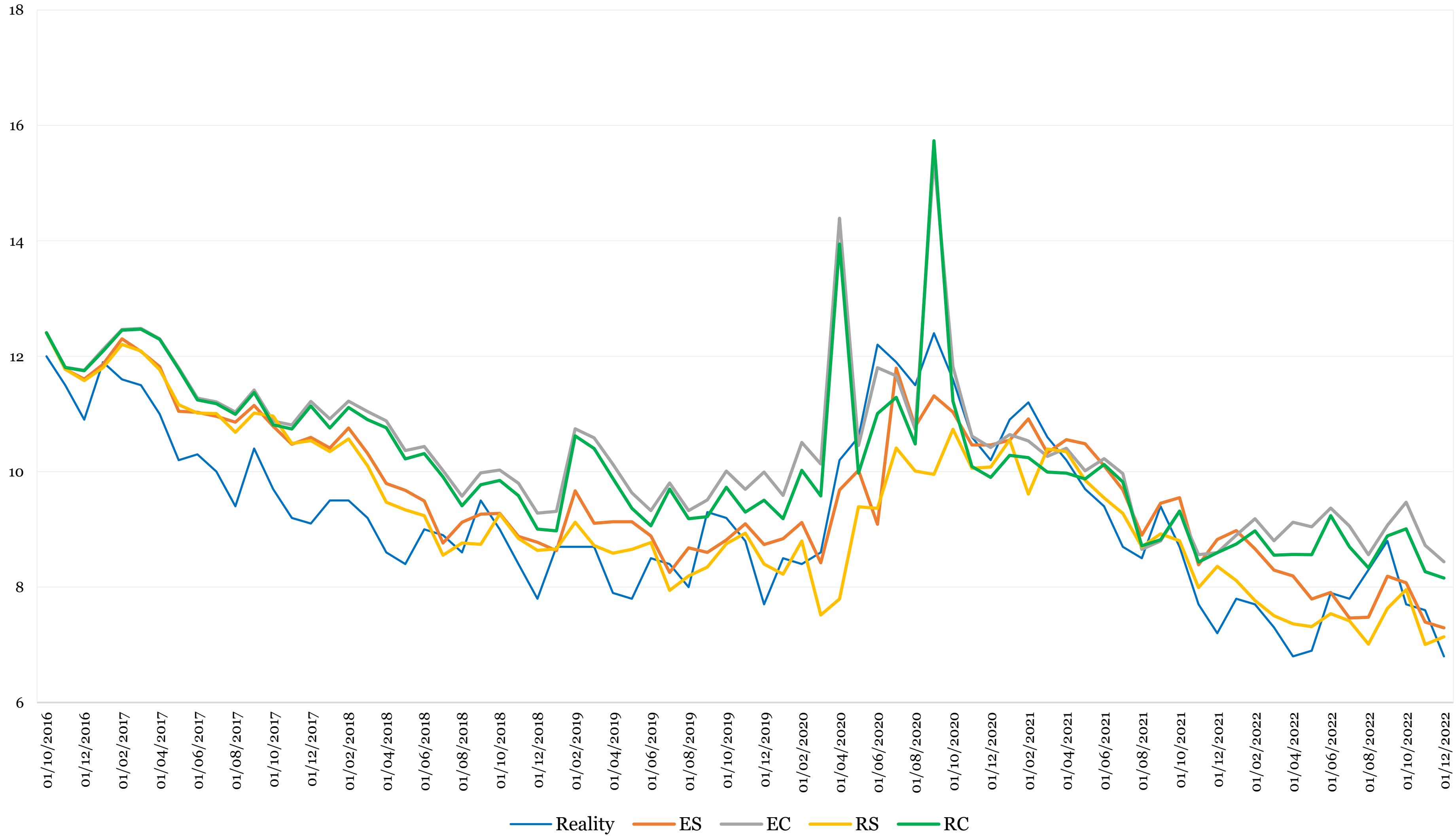
	Diebold-Mariano n=75		Wilcoxon Signed Rank n=75	
	t	sig	t	sig
Hypothesis one	4.144***	<.001	2143***	<.001
Hypothesis two	3.923***	<.001	2262***	<.001
Hypothesis three	3.889***	<.001	2412***	<.001
Hypothesis four	-	-	1673*	<.06

Note: Table shows the results of the hypothesis tests. The Diebold-Mariano test and the Wilcoxon Signed Rank test are both performed under the null hypothesis of equal forecasting accuracy, with the alternative hypothesis conforming to the hypotheses drawn for this study. Significance is indicated with an asterisk, where ***p<.01, **p<.05, and *p<.1. The Diebold-Mariano test is not applied to hypothesis four as discussed earlier.

Before concluding this section, figure 19 is presented on the next page. Figure 19 presents the youth unemployment rate forecasts that are produced with each out-of-sample forecasting procedure. Moreover, the youth unemployment rate forecasts are aligned with the actual youth unemployment rate. Due to the presence of four forecasts and the actual youth unemployment rate in figure 19, appendix 2 shows four additional graphs that are easier to interpret. Figure 19 is not interpreted, as the discussion of figure 14 has covered must peculiarities.

With all hypotheses addressed, the most important results have been obtained and this section comes to an end. However, before the results chapter can be concluded, the next section will discuss the forecast errors that are produced by the optimal out-of-sample forecasting procedure and window size. This should reveal any potential biases.

Figure 19. Forecasts aligned with reality



Note: This figure shows the forecasts that are produced by each out-of-sample forecasting procedure. Time units are represented on the x-axis. The corresponding youth unemployment rates are represented on the y-axis. The out-of-sample is October 2016 till December 2022, the in-sample varies.

4.3 Error diagnostics

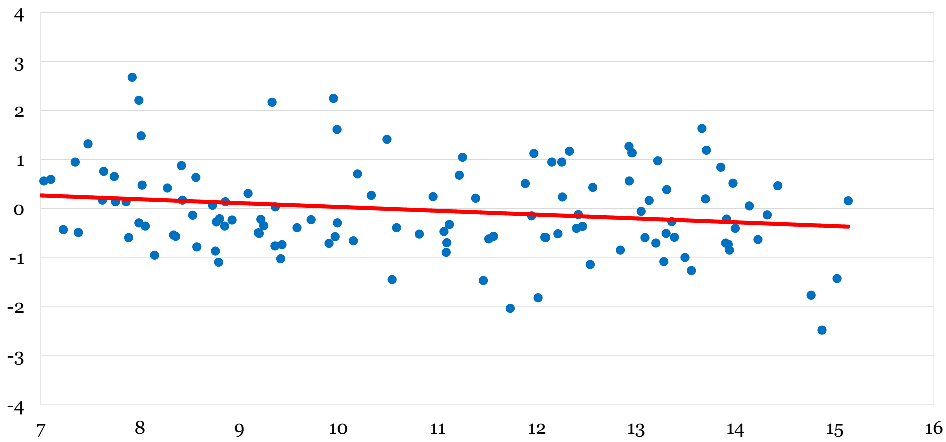
Following up on the previous section, this section starts by selecting the out-of-sample forecasting procedure and respective window size that produces the most accurate forecasts. Using this, a forecast is produced for an even larger out-of-sample than previously used. Then, the errors of this out-of-sample forecast are investigated in detail. Finally, another forecast is produced, leveraging an autoregressive component in addition to independent variables.

The previous section has pointed out that, generally, the out-of-sample forecasting procedures relying on model respecification will produce the most accurate forecasts. However, it is unclear whether a rolling window or expanding is preferred when model respecification is applied. Considering that MAPE is a more generalizable metric than MSE or RMSE, the out-of-sample forecasting procedure that is most accurate according to MAPE is selected. Consequently, the forecast errors produced by the RS procedure for out-of-sample forecasting will be investigated. However, the window size to use must also be selected.

A window with a size of 48 is used in the RS procedure for out-of-sample forecasting. This decision is based on section 5.1, which revealed that, generally, smaller window sizes produce more accurate forecasts when using the RS procedure for out-of-sample forecasting. Moreover, using a window size of 48 aligns with other studies like Tuhkuri (2016), Borup & Schutte (2022), Te Brake (2017), and Smith (2016). Using a similar window size as other studies may allow for some cross-comparison. Due to the window size being quite small, the out-of-sample period is allowed to be large. The out-of-sample period that is now used to produce one month ahead forecasts of the youth unemployment rate spans from October 2012 till December 2022. Below, the forecast errors produced by the RS procedure for out-of-sample forecasting with a window size of 48 are investigated.

Figure 20 reveals a linear relationship between forecasted values and forecast errors. The relationship is negative, implying that when forecasted values increase, the forecast errors decrease. This is confirmed by a linear regression between forecasted values and forecast errors ($\beta = -.0782$, $p = .026$). While the assumption of linearity is considered fulfilled, the assumption of homoscedasticity appears not fulfilled. Yet, it's important to note that the relationship between forecasted values and forecast errors is extremely weak. When forecasted values increase with one, the forecast errors decrease with just -.0782. Moreover, figure 20 illustrates that there is not necessarily unequal variance, but that lower forecasted values are underestimated, while higher forecasted values are overestimated. This suggests that the assumption of homoscedasticity is only weakly violated. In addition to the linearity assumption and the homoscedasticity assumption, figure 20 may show some systematic bias with more negative errors than positive errors being present. No obvious outliers are detected.

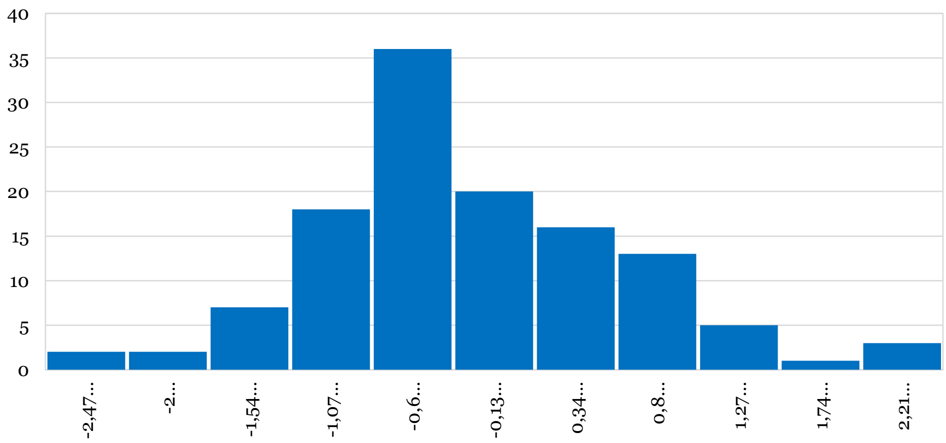
Figure 20. Forecast error plot



Note: Figure shows the error plot. The forecasted values are shown on the x-axis. The errors are shown on the y-axis. Each blue dot is an error x forecast combination. Out-of-sample is October 2012 till December 2022. Red line depicts linear relationship.

Moving on to the assumption of normality, Shapiro Wilk's null hypothesis of normality is supported (Shapiro Wilk= .98, n= 123, p= .0564), indicating that the forecast errors are normally distributed. However, the Shapiro Wilk test is very close to rejecting it's null hypothesis of normality, which would indicate that the forecast errors are not normally distributed. Figure 21 also reveals that the distribution of the forecast errors is right skewed, which is confirmed by a skewness value of .3882. This right skew indicates that forecasts are, on average, overestimated, resulting in a departure from normality.

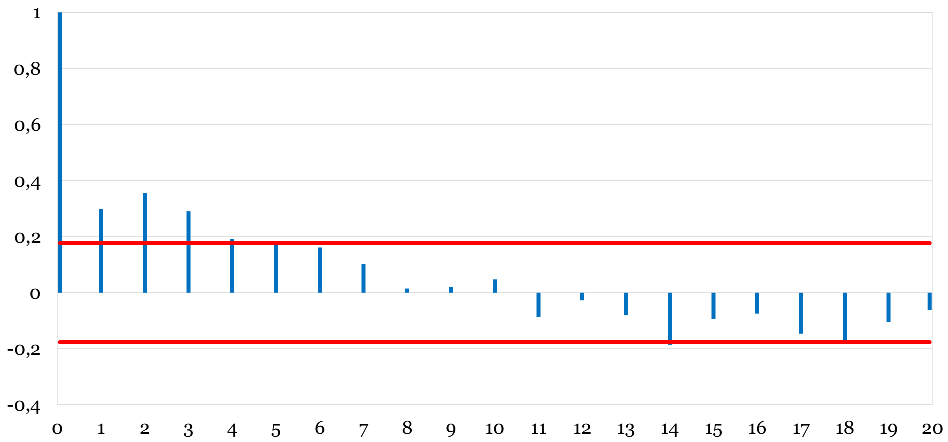
Figure 21. Histogram of forecast errors



Note: Figure shows the histogram of the forecast errors. The bins are shown on the x-axis and represent a range of values. The frequency with which forecast errors that fit within a bin occur, are shown on the y-axis. The out-of-sample is October 2012 till December 2022.

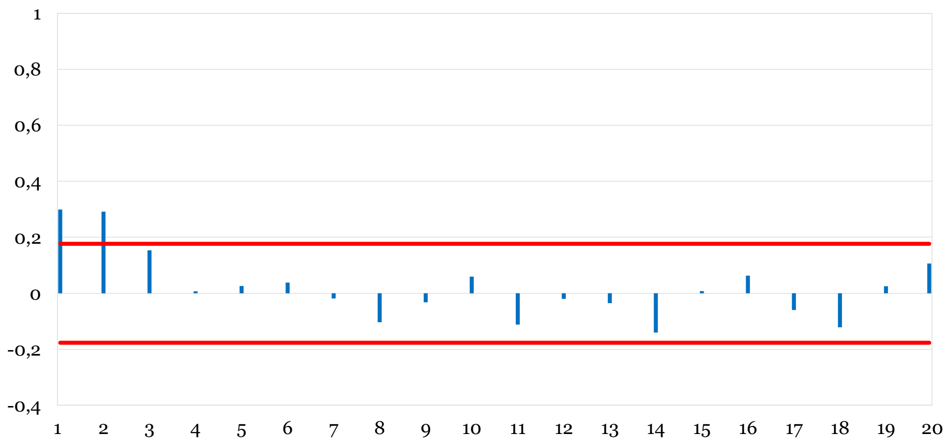
Figure 22 and 23 reveal that some autocorrelation remains in the forecast errors. According to figure 22, the ACF indicates that five lags are significantly autocorrelated. Meaning that the error of next month's forecast is correlated, and thus predictable, with the forecast error obtained for the last five months. The PACF nuances this finding by showing significant autocorrelation at lag one and two, as illustrated in figure 23. Still, the assumption of uncorrelatedness is violated according to both the ACF and PACF. Consequently, there is evidence that some important information is excluded from the forecasting model. Interestingly, seasonality is not at play according to ACF and PACF.

Figure 22. ACF of forecast errors



Note: Figure shows the autocorrelation function of the forecast errors. Time lags are shown on the x-axis. The correlation coefficient for a time lag is shown on the y-axis. Out-of-sample is October 2012 till December 2022. Red lines represent significance.

Figure 23. PACF of forecast errors



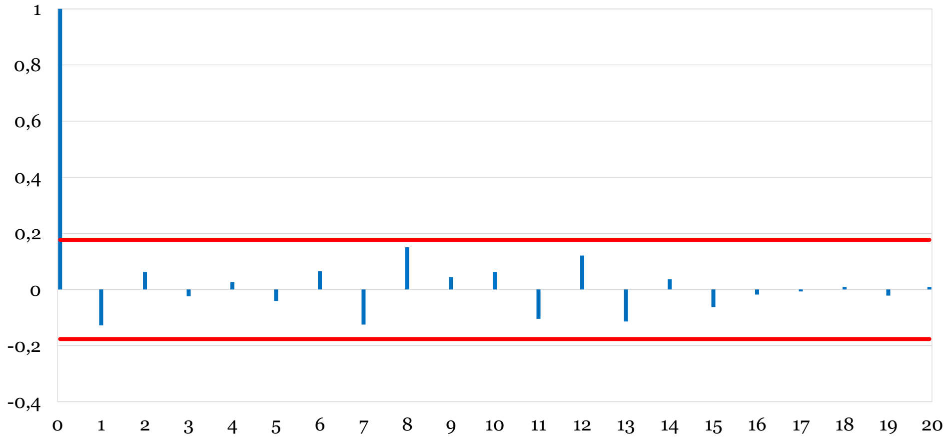
Note: Figure shows the partial autocorrelation function of the forecast errors. Time lags are shown on the x-axis. The correlation coefficient for a time lag is shown on the y-axis. Out-of-sample is October 2012 till December 2022. Red lines represent significance.

With figure 22 and 23 indicating that some information is not captured by the multiple linear regression model, it is extended. Whereas the multiple linear regression model relied solely on independent variables, it is now extended with an autoregressive component. In contrast to independent variables that are constantly being re-selected with each window update, the autoregressive component is always included in the multiple linear regression model. By adding the autoregressive component, the forecasting accuracy is expected to increase while simultaneously reducing the autocorrelation of the forecast errors.

In line with many other studies, an autoregressive component with a lag order of 1 is used, referred to as AR(1). Using an AR(1) component means that last month's youth unemployment rate is used to predict this month's youth unemployment rate. The errors that are now discussed stem from the same multiple linear regression model and out-of-sample forecasting procedure previously used, only now the multiple linear regression model is extended with the fixed inclusion of an AR(1) component. The forecast is produced for the same out-of-sample previously used, spanning from October 2012 till December 2022. Note that all figures that now follow, and the discussion of them, regard this new forecast.

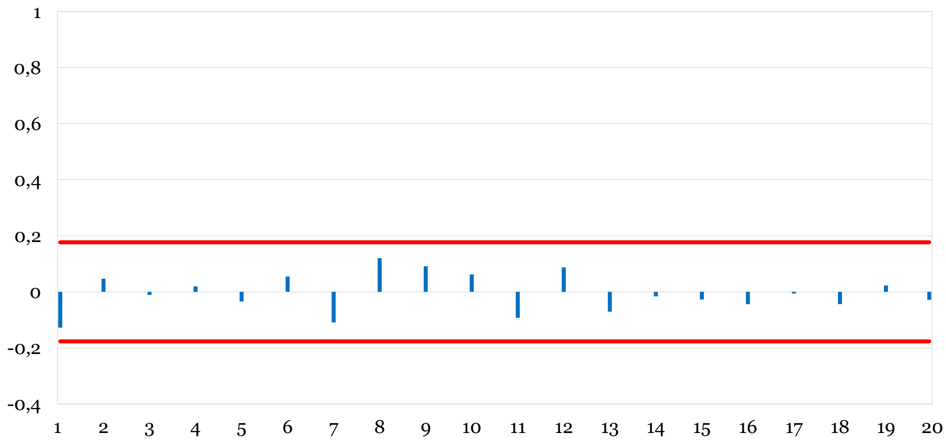
In line with the previously investigated forecast errors, the errors of this new forecast are now investigated. Figure 24 and 25 show that the autocorrelation of the forecast errors is drastically reduced when the AR(1) component is added to the multiple linear regression model. There is no significant autocorrelation at any lag. These results suggest that all important information is optimally leveraged to produce one month ahead forecasts of the youth unemployment rate.

Figure 24. ACF of forecast errors | AR(1) added



Note: Figure shows the autocorrelation function of the forecast errors. Time lags are shown on the x-axis. The correlation coefficient for a time lag is shown on the y-axis. Out-of-sample is October 2012 till December 2022. Red lines represent significance.

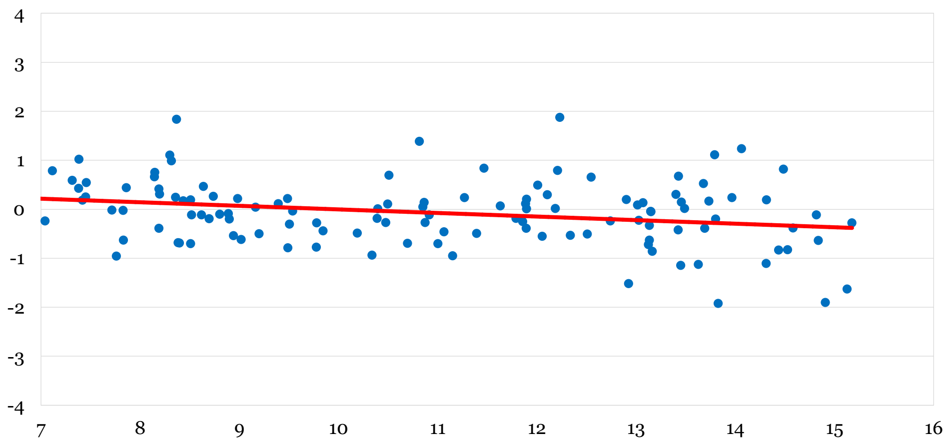
Figure 25. PACF of forecast errors | AR(1) added



Note: Figure shows the partial autocorrelation function of the forecast errors. Time lags are shown on the x-axis. The correlation coefficient for a time lag is shown on the y-axis. Out-of-sample is October 2012 till December 2022. Red lines represent significance.

Figure 26 shows that the forecast errors, when the AR(1) component is added, are less dispersed and lower than when the AR(1) component is not added. However, improvements seem most prominent for lower forecasted values, indicating that for periods with a high youth unemployment rate, the AR(1) component did not add much predictive value. Again, a negative linear relationship is found between forecasted values and forecast errors, as is shown in figure 26. The relationship is equally strong (β= -.07229, p= .0031) as in the forecast that did not utilize the AR(1) component. Consequently, the homoscedasticity assumption is again weakly violated. Homoscedasticity does appear more prevalent when the AR(1) component is added, as a cone appears to take form with higher fitted values. The linearity assumption is fulfilled and no obvious outliers are detected in figure 26.

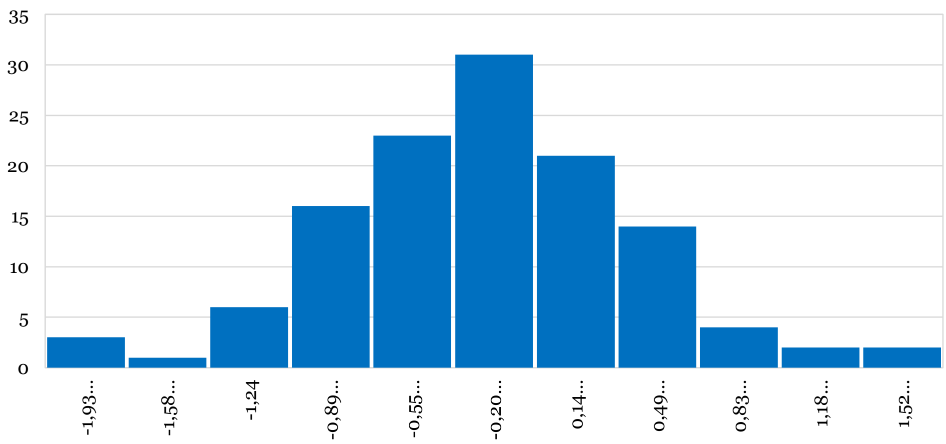
Figure 26. Forecast error plot | AR(1) added



Note: Figure shows the error plot. The forecasted values are shown on the x-axis. The errors are shown on the y-axis. Each blue dot is an error x forecast combination. Out-of-sample is October 2012 till December 2022. Red line depicts linear relationship.

When the AR(1) component is added, the distribution of the forecast errors appears to be more normally distributed, as illustrated in figure 27. This is confirmed by a skewness value of just .0463, whereas the skewness value was .3882 when the AR(1) component was not used. Moreover, a Shapiro-Wilk test shows that the null hypothesis can now be supported with more confidence (Shapiro-Wilk= .99, n=123, p= .2134). With the forecast errors showing to be approximately normal, there is an indication that no systematic bias is present.

Figure 27. Histogram of forecast errors | AR(1) added



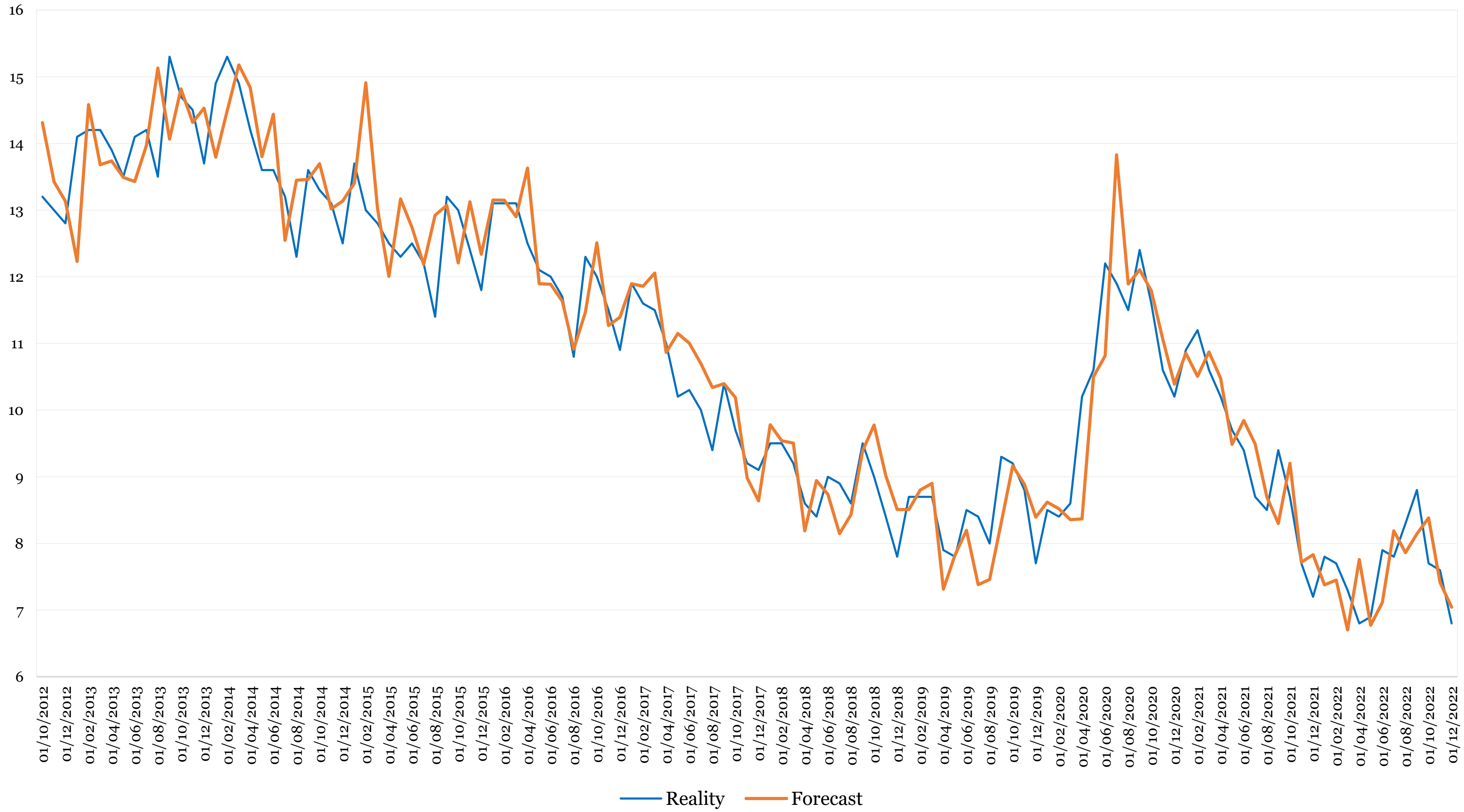
Note: Figure shows the histogram of the forecast errors. The bins are shown on the x-axis and represent a range of values. The frequency with which forecast errors that fit within a bin occur, are shown on the y-axis. Out-of-sample is October 2012 till December 2022.

So far, it has been proven that the assumptions of multiple linear regression are fulfilled to a greater extent when an AR(1) component is used in addition to the independent variables Moreover, the one month ahead forecasts of the youth unemployment rate are more accurate when supplementing the multiple linear regression model with an AR(1) component, with MSE being .4375 as opposed to .8079. Similarly, RMSE is substantially down from .8988 to just .6615. Finally, MAPE is also lower when the AR(1) component is used in addition to the independent variables, decreasing from 6.75% to just 4.73%.

Figure 28, on the next page, presents the one month ahead forecasts of the youth unemployment rate. This regards the forecasts that are produced by using the AR(1) component in addition to the independent variables. Figure 28 shows that the forecasts align incredibly well with the actual youth unemployment rate, even for the period characterized by COVID-19. An overall well fit of the forecasting model with reality is confirmed by an Adjusted R-Squared of 91.33%. This section has revealed that the multiple regression model, relying on the RS procedure for out-of-sample forecasting, is relatively unbiased. However, this section also revealed that it is important to supplement independent variables, that represent Google Trends data, with additional information, like the AR(1) component that was added.

The results chapter has presented some insights that promise to be of great relevance for forecasting practitioners and researchers. First, the importance of selecting an appropriate window size when using out-of-sample forecasting procedures is highlighted. Moreover, the results revealed the increased forecasting accuracy that can be obtained by using model respecification instead of model recalibration. Finally, the results section has shown the incredible potential of forecasting the youth unemployment rate with Google Trends data. Yet, the results have also revealed that when Google Trends data is used for forecasting, it should be supplemented with some additional information. In the next chapter, the results are compared to other studies and the academic contributions are addressed.

Figure 28. Forecasts aligned with reality | AR(1) added



Note: This figure shows the forecasts that are produced by a multiple linear regression model, relying on independent variables (that reflect Google Trends data for keywords), and an AR(1) component. The RS procedure for out-of-sample forecasting was used, with a window size of 48. Time units are shown on the x-axis. The corresponding youth unemployment rates are represented on the y-axis. The out-of-sample is October 2012 till December 2022. The in-sample varies.

5. Analysis

In this chapter, the key findings of this study are identified and discussed in relation to the literature that was covered in the literature review. Subsequently, this chapter identifies the academic contributions that are made by this study.

The literature review revealed that studies often neglect to consider the window size in out-of-sample forecasting procedures (Rossi & Inoue, 2012; Fezzi & Mosetti, 2020). This finding applies not only to studies that forecasted unemployment Google Trends data, but to forecasting studies in general. Even the overall choice for an out-of-sample procedure is hardly considered by forecasting studies (Tashman, 2000; Rossi & Inoue, 2012; Fezzi & Mosetti, 2020). Shen et al. (2020) is one of the few that studied the impact of window size on forecasting accuracy, focusing solely on the rolling window. The results of this study align with Shen et al. (2020) their findings, revealing that a larger window size leads to lower forecasting accuracy, both for the rolling window and the expanding window. Consequently, Shen et al. (2020) their findings are not only strengthened by this study, but also extended by examining whether the relationship holds with an expanding window.

In addition to strengthening and extending Shen et al. (2020)'s study, this study explicitly contributes to two more studies. This study's answers to Rossi & Inoue (2012) and Fezzi & Mosetti (2020) their call for more research on the window size consideration in out-of-sample forecasting procedures. Furthermore, by investigating the sensitivity of forecasting accuracy to the window size, this study contributes to the broader forecasting literature in three ways: (1) Raising awareness on the importance of window size, (2) invalidating cross-comparisons between studies with different window sizes, (3) demonstrating that studies their forecasts can be improved by adjusting just the window size. The second contribution is particularly significant, as studies in forecasting literature often seek to outperform others. For example, D'Amuri & Marucci (2017) claim to have outperformed the study of Barnichon and Nekarda (2012) without considering the differences in window size. While Barnichon and Nekarda (2012) used a rolling window spanning 15 years of data, D'Amuri & Marucci (2017) used a rolling window spanning approximately 3 years of data. That does not necessarily invalidate D'Amuri & Marucci (2017) their claim. But, it begs the question; could Barnichon and Nekarda (2012) have produced forecasts with similar accuracies similar to D'Amuri & Marucci (2017) with a smaller window size?

Various out-of-sample forecasting procedure exist, each requiring the window size to be set appropriately. The literature review revealed two out-of-sample forecasting procedures that are primarily used nowadays; the rolling window and the expanding window (Tashman, 2000; Hewamalage, 2022; Liu & Yang, 2020). Both procedures come with their own advantages and disadvantages, as discussed in the literature review. However, literature's main focus regarding the rolling and expanding window is grounded in their ability to deal with sudden changes, also known as structural breaks. Structural breaks can substantially affect the relationship between variables, causing relationships to either weaken or strengthen over time. This time-varying relationship between variables is also known as parameter instability (Stock & Watson, 1996; Swanson, 1998). The next substantial finding of this study is related to the rolling and expanding window their ability to deal with parameter instability.

Inoue et al. (2017) and Pesaran & Timmermann (2002) suggest to use the rolling window in the field on econometrics, where structural breaks are likely to occur. The rolling window is considered more able to respond to sudden changes by pruning old and potentially noisy data (Tashman, 2000; Hewamalage, 2022). This study's strengthens the suggestion of Inoue et al. (2017) and Pesaran & Timmermann (2002), while also testing it in another daylight.

The results of this study reveal that the rolling window outperforms the expanding window when out-of-sample forecasting the youth unemployment rate. Specifically, this study finds support for hypothesis one: "Recalibration in a rolling window is significantly more accurate than recalibration in an expanding window when forecasting Dutch youth unemployment". Consequently, this study's strengthens the suggestion of Inoue et al. (2017) and Pesaran & Timmermann (2002) to use the rolling window for-out-of-sample forecasting econometrics. However, hypothesis one merely addresses the advantage of using the rolling window when model recalibration is used, following Inoue et al. (2017) and Pesaran & Timmermann (2002). With hypothesis four, this study goes a step further by also investigating whether the rolling window is more accurate than the expanding window when model respecification is used. Unfortunately, this study fails to either support or reject hypothesis four: "Respecification in a rolling window is significantly more accurate than respecification in an expanding window when forecasting Dutch youth unemployment rate". The conflicting results obtained for hypothesis four shed a different light on the suggestions of Inoue et al. (2017) and Pesaran & Timmermann (2002) when novel techniques like model respecification are used.

While strengthening and challenging the study of Inoue et al. (2017) and Pesaran & Timmermann (2002), another substantial result was obtained. This study reveals that during the structural break characterized by COVID-19, the rolling window did not perform better than the expanding window in forecasting the youth unemployment rate. Consequently, this study's finding demonstrates that the rolling window is not always superior to the expanding window during structural breaks. Moreover, it may reveal that no structural breaks is like another, invalidating the suggestion to always use the rolling window during structural breaks. This has not been researched by literary works preceding this one. This study's results regarding hypothesis one and four also answers to Tashman (2000) and Hewamalage (2022) their research call. In a similar manner to the contributions made by analyzing the sensitivity of forecasting accuracy to the window size, the findings related to hypothesis one and four contribute to literature by: (1) Raising awareness on the importance of picking the correct out-of-sample forecasting procedure, (2) invalidating the cross-comparisons between studies that used different out-of-sample forecasting procedures, (3) demonstrating that studies their forecasts can be improved by using a different out-of-sample forecasting procedure. Next, the main finding and contribution of this study is discussed.

The literature review revealed that the rolling and expanding window procedures for out-of-sample forecasting rely on the concept of model calibration. Meaning that with each window update, the forecasting model's parameters are re-estimated. However, as previously discussed, this study embarked on an exploration of the novel concept of model respecification, where model parameters are respecified with each window update, essentially, building a new forecasting model each time the window is updated. This research endeavor was motivated by Mulero & Garciá-Hiernaux (2021)'s their future research suggestion.

This study’s finding demonstrates that model respecification substantially improves the accuracy when out-of-sample forecasting the youth unemployment rate. Compared to model recalibration, model respecification yields 44% more accurate forecasts of the youth unemployment rate. This finding is supported with 99% of confidence by accepting hypothesis two: “Respecification in a rolling window is significantly more accurate than recalibration in a rolling window when forecasting Dutch youth unemployment” and hypothesis three: “Respecification in an expanding window is significantly more accurate than recalibration in an expanding window when forecasting Dutch youth unemployment rate”. Consequently, the dominance of model respecification, as opposed to model recalibration, is generalizable for both the rolling window and the expanding window procedure for out-of-sample forecasting.

In demonstrating the significance of model respecification, Mulero & Garciá-Hiernaux (2021) their research call is answered. In addition to this study, Mulero & Garciá-Hiernaux (2021) is the only other study discussing model respecification, while none have actually tested model respecification. Consequently, the results of this study can’t be verified with other studies. However, this study does extend the findings of Li et al. (2014), Singhania & Kundu (2021), and Borup & Schutte (2022) by showing that dimension reduction techniques, such as the ones suggested in their studies, can be leveraged to empower model respecification. Moreover, this study makes a valuable contribution to forecasting literature at large by demonstrating how model respecification has the ability to significantly improve forecasts. Thus, enlarging the toolbox from which researchers can tap to obtain more accurate forecasts. Furthermore, this study puts forward a novel perspective on the recent trend of introducing more complex forecasting models to improve forecasting accuracy. This study highlights that refining methodologies, while using basic models, may exert a more positive impact on forecasting accuracy than introducing increasingly complex forecasting models. This contribution aligns with a growing demand for explainable AI and a departure from black-boxes like neural networks in the forecasting domain (Rozanec et al., 2022). Next, this study’s findings and academic contributions related to forecasting with Google Trends data are discussed.

The methodology section of this study revealed that Google Trends data substantially varies when collected at different points in time, suggesting that Google Trends data is fundamentally inconsistent. This finding aligns with Cebrián & Domenech (2022) and Eichenauer et al. (2021), thus strengthening their findings. Consequently, this study reminds forecasting literature of the limitations inherent to Google Trends data. Moreover, this study found that forecasts relying on solely Google Trends data are substantially improved when additional information, like an autoregressive component, is added. With this finding, forecasting literature is reminded that Google Trends data should never be used as the sole predictor. This finding aligns with Lazer et al. (2014), who claim that Google Flu Trends failed simply due to the misconception that Google Trends data is the only relevant variable in forecasting the flu. Sticking with Google Trends, this study’s forecasts of the youth unemployment rate by leveraging Google Trends data is compared to other studies.

The results section concluded with a final out-of-sample forecast of the youth unemployment rate for the period October 2012 till December 2022. Now, this forecast is compared to the forecasts done by studies presented in the literature review. However, since the studies did not all use the same window size and out-of-sample forecasting procedure, the cross-study comparison is not entirely valid. Additionally, most studies forecasted the overall unemployment rate, while this study focused on the youth unemployment rate. Furthermore, each study focused on a different country. Only studies that reported the MAPE or (Adjusted) R-Squared are used for comparison, as only these metrics are cross-study comparable. Despite these limitations, some comparisons are made based on the studies that are shown in table 1.

Barreira et al. (2013) and Mclaren & Shanbogue (2011) both used an autoregressive model with seasonal effects and Google Trends data to forecast the unemployment rate one month ahead, although in different contexts. Mclaren & Shanbogue (2011) forecasted the unemployment rate in the UK, obtaining an Adjusted R-Squared of 85%. Barreira et al. (2013) applied a similar model in various contexts, yielding an Adjusted R-Squared of 55.4% for Portugal, 68.3% for Spain, 75,5% for France, and 69.3% for Italy. This study's forecast, characterized by an Adjusted R-Squared of 91.33%, substantially outperforms Barreira et al. (2013) and Mclaren & Shanbogue (2011). Noteworthy, the out-of-sample used by both Mclaren & Shanbogue (2011) and Barreira et al. (2013) was small, potentially not representative of the actual performance. Their out-of-sample did include the 2008 crisis, which may negatively affected performance.

Two other studies also forecasted the unemployment rate one month ahead using an autoregressive model with seasonal effect and Google Trends data, although in different contexts than Barreira et al. (2013) and Mclaren & Shanbogue (2011) and reporting on the more reliable error metric MAPE. Tuhkuri (2016) forecasted the unemployment rate in the US, resulting in a MAPE of 7.01% for a relatively large out-of-sample that included the 2008 crisis. Te Brake (2016) followed an approach to Tuhkuri (2016) but focused on the Netherlands and considered subperiods. Te Brake (2016)'s forecast for the first subperiod yielded a MAPE of 4.59%, for the second subperiod 2.51%, and for the third subperiod 3.19%. In comparison, this study's forecast yielded a MAPE of 4.73%, outperforming Tuhkuri (2016) while also being outperformed by Te Brake (2016). However, when this study also considers a subperiod of 24 months, the forecast errors are cut in half and Te Brake (2016) is outperformed.

Singhania & Kundu (2021) and Vicente et al. (2015) forecasted the unemployment rate one month ahead with Google Trends data and also reported on the MAPE error metric. However, models different from an autoregressive model were used. Vicente et al. (2015) used an autoregressive integrated moving average model, to forecast the unemployment rate in Spain. Results show that Vicente et al. (2015) obtained an incredibly accurate forecast with a MAPE of just 1.08%, outperforming this study's forecast. Notably, the out-of-sample period used by Vicente et al. (2015) was small and concerned a relatively stable period. Surprisingly, Singhania & Kundu (2021) used a neural network model to forecast the unemployment rate in the US and obtained a MAPE of just 6.24%. Causing the forecast of this study, done with a relatively simple forecasting model, to be more accurate than a neural network. Like most other studies, Singhania & Kundu (2021) used a small out-of-sample.

While this study's forecast outperformed most of the studies presented in the literature review, this cross-study comparison has massive limitations. Yet, this cross-study comparison provides an insight into the incredibly accurate forecast that this study realized by leveraging Google Trends data. An especially impressive performance considering that the forecasting model used by this study is inferior to the forecasting models used by aforementioned studies. Moreover, this study used a very large out-of-sample that included a structural break, being the COVID-19 period, further highlighting the impressive forecasting accuracy. By successfully forecasting the youth unemployment rate with Google Trends data, this study strengthens the findings of many other studies that emphasized the potential of Google Trends data. In doing so, forecasting literature is again urged to not abandon Google Trends for forecasting applications, even though there are some limitations inherent to using Google Trends data.

This chapter has discussed all the main findings off this study and motivated academic contributions that are made with these findings. Turning to the next page, this study draws to a close with a conclusion. This conclusion reflects back on the research rational and question, before moving to the practical contributions, limitations, and future research suggestions.

6. Conclusion

This chapter presents the conclusion of this study, reflecting back on the research rationale and research question. Additionally, this conclusion summarizes the key findings of this study, addresses its practical contributions and limitations, and concludes with various research suggestions that can be exploited by future research.

This study aimed to explore the procedures that are used for out-of-sample forecasting, and to extend these procedures with the novel concept of model respecification in the context of forecasting youth unemployment with Google Trends data. The motivation for this exploration was the pressing need for accurate forecasts in the field of human resource management, where rising job vacancies and declining unemployment rates came unforeseen. Consequently, labour market forecasters are faced with the incredible challenge of leveraging online big data, which is singled-out as the transformative force in labour market anticipation. However, the utilization of big data in labour market forecasting re-ignites the long-standing issue of parameter instability and introduces the intricate challenge of identify suitable predictor variables among millions of potential predictor variables. Yet, recent literature has primarily focused on developing more complex forecasting models, neglecting to consider the methodologies requisite for leveraging big data in forecasting. In fact, procedures for out-of-sample forecasting, introduced fifty years ago, continue to be used despite their limitations in leveraging big data. To address this limitation, this study advocates to extend the out-of-sample forecasting procedures with model respecification. This approach promises to be more potent at mitigating the effect of parameter instability. Additionally, model respecification is suited to the intricate search for suitable predictor variables. To guide this research endeavor, the following research question was drafted:

“What impact does model respecification have on the forecasting accuracy of Dutch youth unemployment rates when there are time-varying relationships between variables?”

To answer the research question, forecasts of the youth unemployment rate are produced with a multiple linear regression model that leverages Google Trends data. Google Trends data, known for its instability, perfectly suits a study surrounding the topic of parameter instability. Four forecasts are done, first the rolling window and expanding window relying on model recalibration are used to produce forecasts. The same two procedures, but relying on model respecification, are also used to produce forecasts. The results reveal that both the rolling window and expanding window will produce significantly more accurate results when model respecification is used. The increases in forecasting accuracy average to approximately 44% and hold with more than 99% of confidence. Consequently, the answer to the research question is concise: model respecification has a positive impact on the forecasting accuracy of Dutch youth unemployment rates when there are time-varying relationships between variables.

This study contributes to forecasting literature in several ways. First, this study introduces an out-of-sample forecasting procedure suited to an era characterized by data generation, computational advancements, and increased instability. In doing so, the forecasting toolbox from which researchers can tap is enlarged and researchers are assisted in obtaining more accurate forecasts. Second, this study puts forward a perspective that contrasts the trend of developing more complex forecasting models to improve forecasting accuracy. This study highlights that forecasting accuracy can also be improved through methodological advancements while using simple forecasting models. Third, this study strengthens other studies by underscoring the potential of Google Trends data, even during structural breaks.

Fourth, this study also contributes to forecasting literature by showing that forecasting accuracy is sensitive to the window size that is used in out-of-sample forecasting procedures. With this finding, this study contributes by: (1) Raising awareness on the importance of window size, (2) invalidating cross-comparisons between studies with different window sizes, (3) demonstrating that studies their forecasts can be improved by adjusting just the window size. Fifth, this study contributes to forecasting literature by strengthening previous studies their finding that that the rolling window is preferred in econometrics.

In addition to academic contributions, this study makes several practical contributions. First, this study equips practitioners with an innovative out-of-sample forecasting procedure that is tailored for the use of big data. Thereby, organizations are assisted in leveraging big data amidst an era where big data is singled-out as being the transformative force in (labour market) forecasting. Consequently, organizations are better able to navigate volatile business environments through accurate forecasts of future reality. This, in turn, enhances the quality of strategic decision-making across businesses, government institutions, and other organizations. Second, this study democratizes labour market forecasting by demonstrating how simple forecasting models effectively leverage publicly available big data. Consequently, labour market forecasting becomes a more accessible endeavor for a broad range of organizations. With these forecasts, human resource managers are in turn assisted in getting timely insight into the misalignment between strategy and the workforce availability.

Although contributing to literature and practice, this study also comes with limitations. First, model respecification requires enormous computational power, especially when more complex modelling techniques are used. For example, an autoregressive integrated moving average model takes around 60 times longer to process out-of-sample forecasts than a multiple linear regression model. Second, this study merely established correlation, and not causation, between Google Trends data and the youth unemployment rate. It could be the case that the youth unemployment rate is explaining Google Trends data more strongly than the other way around. While individuals may engage in unemployment-related searches before getting unemployed, they are equally likely to do so after getting unemployed. Third, the keywords that are used could be subject to noise. For example, keywords like “werkloosheid” (unemployment) do not solely reflect searches done by the unemployed. Rather, this keyword simply reflects an interest in the current state of the economy. Fourth, this study introduced coverage bias by using merely Dutch keywords, therefore excluding individuals that don’t speak Dutch. Fifth, this study only forecasted one month ahead, limiting the practical value.

Inspired by the limitations, there are various future research suggestions. First, more research should be dedicated to the out-of-sample forecasting procedures and the considerations, like window size, that come with it. Second, future research is suggested to readdress the concept of model respecification, paying extra attention to spurious correlations and causation. Consequently, a more profound basis for the use of model respecification is established. Third, future research should apply model respecification alongside more complex modeling techniques to assess whether the hypotheses of this study still hold. Moreover, this will reveal whether the obtained increases in accuracy justify the needed computational power. Fourth, future research is also urged to apply model respecification in forecasts that span more than one month ahead. Thus, potentially increasing the practical value of model respecification. Fifth, future research is suggested to further automate this study’s model respecification algorithm. Whereas this study assumed the initial selection of keywords/variables is done by hand, it should be fully automated to better align with the needs of big data. For example, ChatGPT could be prompted to return keywords related to a certain domain ontology. Subsequently, Google Tends data for these keywords could be obtained automatically.

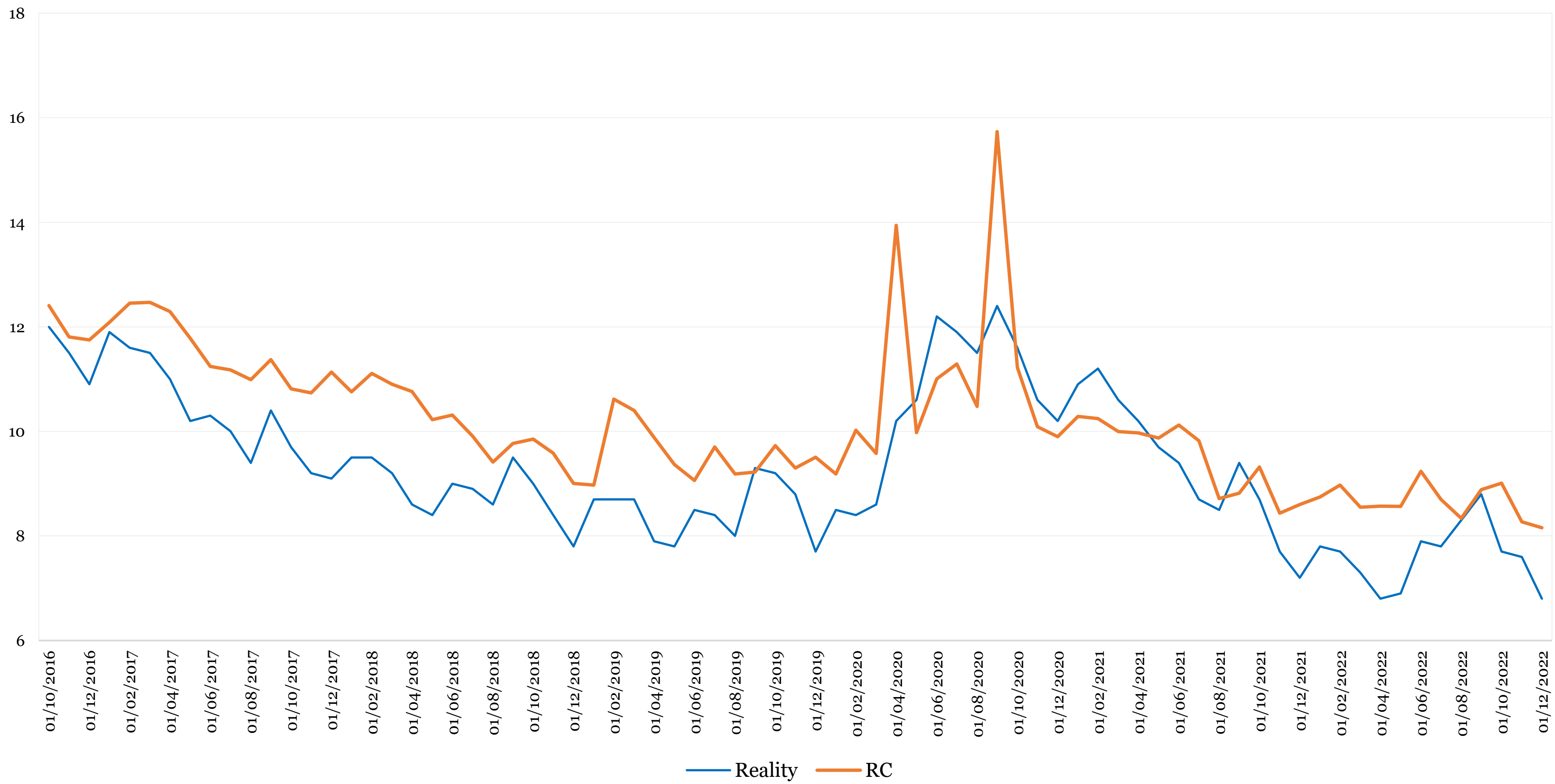
Appendices

Appendix 1

See attached Excel document.

Appendix 2

Figure 29. Forecast aligned with reality | RC

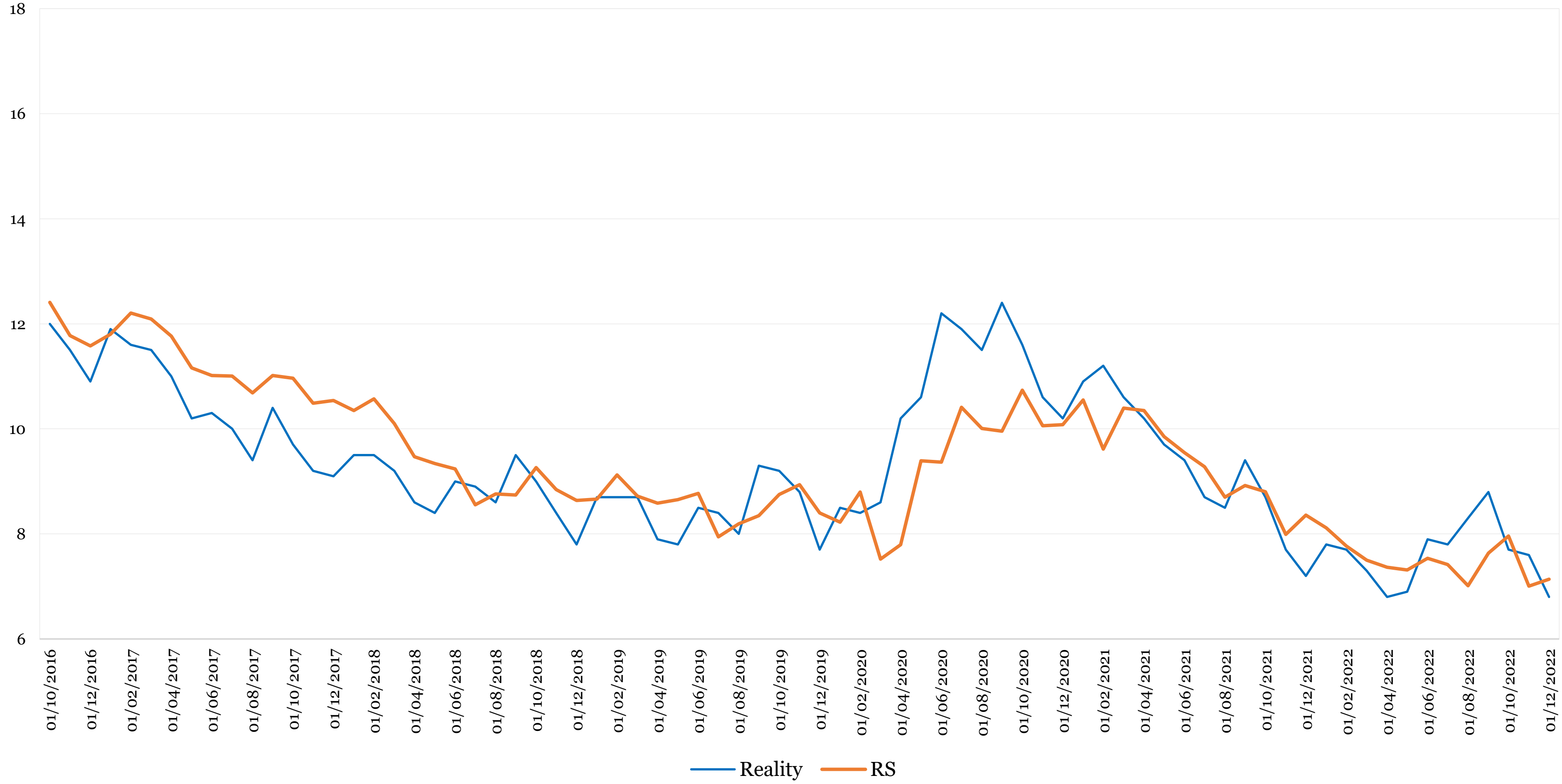


Note: This figure shows the one month ahead forecasts that are produced by the RC out-of-sample forecasting procedure. Time units are represented on the x-axis. The corresponding youth unemployment rates are represented on the y-axis. The out-of-sample is October 2016 till December 2022, the in-sample varies.

Figure 30. Forecasts aligned with reality | EC

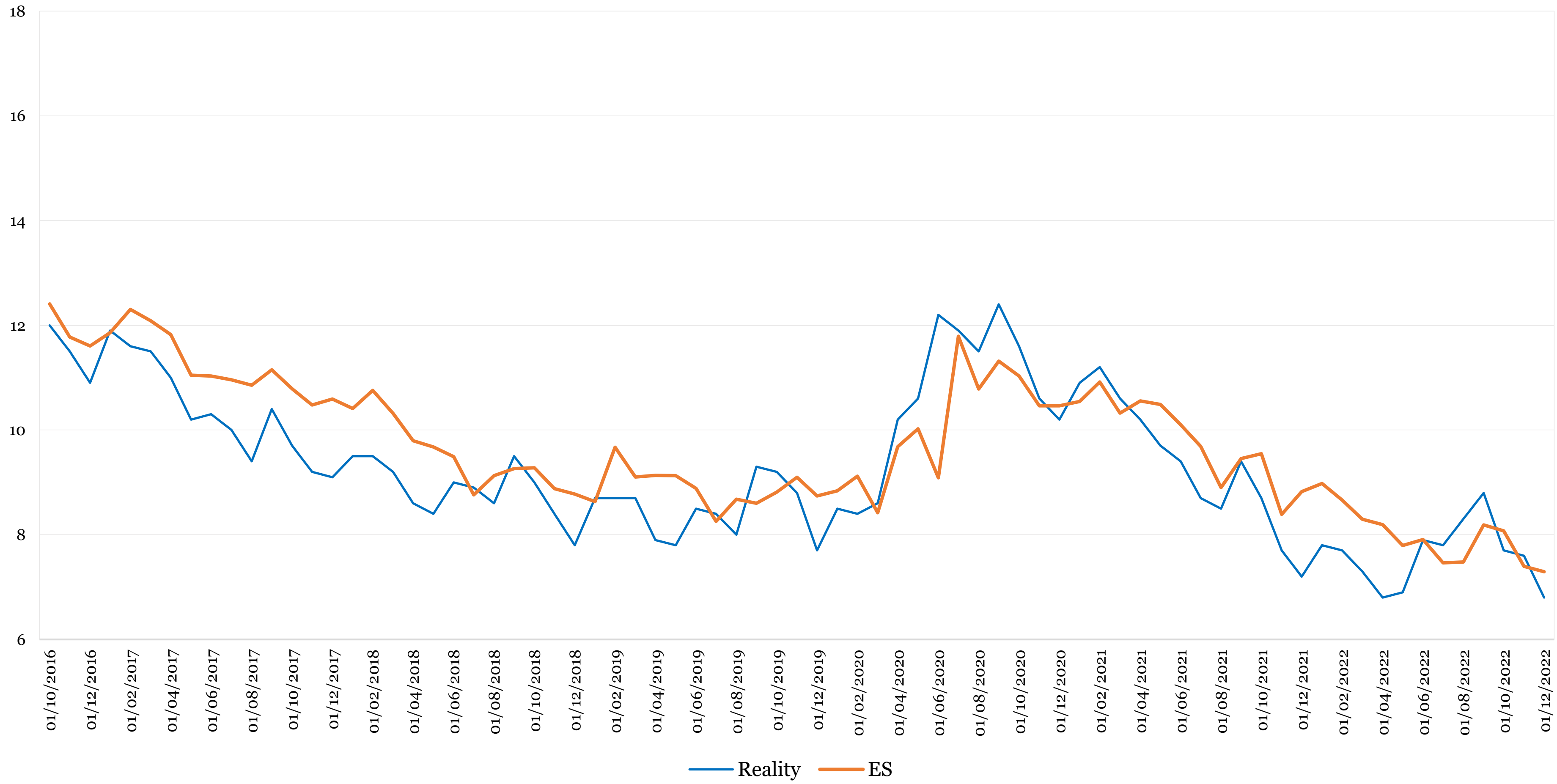
Note: This figure shows the one month ahead forecasts that are produced by the EC out-of-sample forecasting procedure. Time units are represented on the x-axis. The corresponding youth unemployment rates are represented on the y-axis. The out-of-sample is October 2016 till December 2022, the in-sample varies.

Figure 31. Forecasts aligned with reality | RS



Note: This figure shows the one month ahead forecasts that are produced by the RS out-of-sample forecasting procedure. Time units are represented on the x-axis. The corresponding youth unemployment rates are represented on the y-axis. The out-of-sample is October 2016 till December 2022, the in-sample varies.

Figure 32. Forecasts aligned with reality | ES



Note: This figure shows the one month ahead forecasts that are produced by the ES out-of-sample forecasting procedure. Time units are represented on the x-axis. The corresponding youth unemployment rates are represented on the y-axis. The out-of-sample is October 2016 till December 2022, the in-sample varies.

References

Akram, Q.F., & Nymoen, R. (2006). Econometric modelling of slack and tight labour markets. *Economic Modelling, 23*(4), 579-596. https://doi.org/10.1016/j.econmod.2006.03.001

Allen, M.P. (1997). Model specification in regression analysis. *Understanding Regression Analysis*, 166-170. https://link.springer.com/chapter/10.1007/978-0-585-25657-3_35

Armstong, J.S., Grohman, M.C. (1972). A Comparative Study of Methods for Long-Range Market Forecasting. *Management Science, 19*(2), 211-221. http://dx.doi.org/10.1287/mnsc.19.2.211

Askitas, N., Zimmerman, K.F. (2009). Google Econometrics and Unemployment Forecasting. *Applied Economics Quarterly, 55*(2), 107-120. https://doi.org/10.3790/aeq.55.2.107

Barnichon, R., Nekarda, C.J. (2012). *The Ins and Outs of Forecasting Unemployment: Using Labor Force Flows to Forecast the labor market (FEDS Working Paper 2013-19).* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2976716

Barreira, N., Godinho, P., & Melo, P. (2013). Nowcasting unemployment rate and new car sales in south-western Europe with Google Trends. *NETNOMICS: Economic Research and Electronic Networking, 14*, 129-165. https://doi.org/10.1007/s11066-013-9082-8

Behnen, P., Kessler, R., Kruse, F., & Gómez, J.C.M. (2021). Experimental Evaluation of Scale and Patterns of Systematic Inconsistencies in Google Trends data. *Communications in Computer and Information Science*. http://doi.org/10.1007/978-3-030-65965-3_25

Berrar, D. (2019). Cross-Validation. *Encyclopedia of Bioinformatics and Computational Biology, 1*, 542-545. https://doi.org/10.1016/B978-0-12-809633-8.20349-X

Bono, R., Arnau, J., Alarcón, R., & Blanca, M.J. (2019). Bias, Precision, and Accuracy and Skewness and Kurtosis Estimators for Frequently Used Continuous Distributions. *Symmetry, 12*(1), 19. https://doi.org/10.3390/sym12010019

Borup, D., & Schütte, E.C.M. (2022). In Search of a Job: Forecasting Employment Growth Using Google Trends. *Journal of Business & Economics Statistics, 40*(1), 186-200. https://doi.org/10.1080/07350015.2020.1791133

Box, G.E.P., Jenkins, G.M., Reinsel, G.C., & Ljung, G.M. (2015). *Time Series Analysis* (5th edition). Hoboken, New Jersey, John Wiley & Sons, Inc.

Brockwell, P.J., & Davis, R.A. (2016). *Introduction to Time Series and Forecasting* (3rd edition). Cham, Springer.

Brunello, G., & Wruuck, P. (2021). Skill shortages and skill mismatch: A review of the literature. *Journal of Economic Surveys, 35*(4), 1145-1167. https://doi.org/10.1111/joes.12424

Catania, L., Grassi, S., Ravazzolo, F. (2019). Forecasting cryptocurrencies under model and parameter instability. *International Journal of Forecasting, 35*(2), 485-501. https://doi.org/10.1016/j.ijforecast.2018.09.005

Carrière-Swallow, Y., & Labbé, F. (2011). Nowcasting with Google Trends in an Emerging Market. *Journal of Forecasting, 32*(4), 289-298. https://doi.org/10.1002/for.1252

CBS. (2023). *Arbeidsdeelname en werkloosheid per maand*. Retrieved January 23, 2023, from https://opendata.cbs.nl/#/CBS/nl/dataset/80590ned/table?dl=770B2

Cebrián, E., & Domenech, J. (2022). Is Google Trends a quality data source? *Applied Economics Letters*. https://doi.org/10.1080/13504851.2021.2023088

Cerqueira, V., Torgo, L., Mozetic, I. (2020). Evaluating time series forecasting models: an empirical study on performance estimation methods. *Machine Learning, 109*, 1997-2028. https://doi.org/10.1007/s10994-020-05910-7

Chadwick, M.G., & Sengul, G. (2012). *Nowcasting Unemployment Rate in Turkey: Let's Ask Google*. (Working Paper No. 1218). Central Bank of the Republic of Turkey. https://ideas.repec.org/p/tcb/wpaper/1218.html

Chatfield, C., & Xing, H. (2019). *The Analysis of Time Series – An introduction with R* (7th edition). Boca Raton, Florida, CRC Press.

Chen, J., Mi, H., Fu, J., Haitian, Z., Zhao, H., Yuan, R., Guo, H., Zhu, K., Zhang, Y., Lyu, H., Zhang, Y., She, N., & Ren, X. (2022). Construction and validation of a COVID-19 pandemic trend forecast model based on Google Trends data for smell and taste loss. *Frontiers in Public Health, 10*. https://doi.org/10.3389/fpubh.2022.1025658

Chicco, D., Warrens, M.J., Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*. https://doi.org/10.7717/peerj-cs.623

Choi, H., & Varian, H. (2009). *Predicting Initial Claims for Unemployment Benefits* [Google Research Paper]. https://ssrn.com/abstract=1659307

D'Amuri, F., & Marucci, J. (2017). The predictive power of Google searches in forecasting US unemployment. *International Journal of Forecasting, 33*(4), 801-816. https://doi.org/10.1016/j.ijforecast.2017.03.004

Diebold, F.X. (2015). Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold-Mariano Tests. *Journal of Business & Economic Statistics, 33*(1). https://doi.org/10.1080/07350015.2014.983236

Dugas, A.F., Jalalpour, M., Gel, Y., Levin, S., Torcaso, F., Igursa, T., & Rothman, R.E. (2013). Influenza Forecasting with Google Flu Trends. *PLOS ONE, 8*(2). https://doi.org/10.1371/journal.pone.0056176

Eichenauer, V.Z., Indergand, R., Martínez, I.Z., & Sax, C (2021). Obtaining consistent time series from Google Trends. *Economic Inquiry, 60*, 694-705. https://doi.org/10.1111/ecin.13049

Eurostat. (2023a). *Unemployment statistics*. Retrieved August 30, 2023, from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Unemployment_statistics

Eurostat. (2023b). *Job vacancy statistics*. Retrieved August 30, 2023, from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Job_vacancy_statistics

Faber, J., Fonseca, L.M. (2014). How sample size influences research outcomes. *Dental Press Journal of Orthodontics, 19*(4), 27-29. https://doi.org/10.1590%2F2176-9451.19.4.027-029.ebo

Fantazzini, D., Toktamysova, Z. (2015). Forecasting German car sales using Google data and multivariate models. *International Journal of Production Economics, 170*, 97-135. https://doi.org/10.1016/j.ijpe.2015.09.010

Ferrara, L., & Simoni, A. (2022). When are Google Data Useful to Nowcast GDP? An Approach via Preselection and Shrinkage. *Journal of Business & Economics Statistics*. https://doi.org/10.1080/07350015.2022.2116025

Fezzi, C., & Mosetti, L. (2020). Size Matters: Estimation Sample Length and Electricity Price Forecasting Accuracy. *Journal of the IAEE's Energy Economics Education Foundation*. https://doi.org/10.5547/01956574.41.4.cfez

Flores, J.H.F, Engel, P.M., Pinto, R.C. (2012). Autocorrelation and partial autocorrelation functions to improve neural networks models on univariate time series forecasting. *The 2012 International Joint Conference on Neural Networks (IJCNN).* https://doi.org/10.1109/IJCNN.2012.6252470

Fondeur, Y., & Karamé. (2013). Can Google data help predict French youth unemployment? *Economic modelling, 30*, 117-125. https://doi.org/10.1016/j.econmod.2012.07.017

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature, 457*, 1012-1014. https://doi.org/10.1038/nature07634

Google. (n.d.). *FAQ about Google Trends data*. Retrieved January 24, 2023, from https://support.google.com/trends/answer/4365533

González-Fernández, M., & González-Velasco, C. (2018). Can Google econometrics predict unemployment? Evidence from Spain. *Economics Letters, 170*, 42-45. https://doi.org/10.1016/j.econlet.2018.05.031

Greene, W.H. (2002). *Econometric Analysis (5th edition).* Prentice Hall, New Jersey. http://www.mim.ac.mw/books/William%20H.%20Green's%20Econometric.pdf

Gurchiek, K. (2022, December 28). *HR Battered by an Unstable Economy, Return to Workplace, Union Action in 2022*. SHRM. Retrieved August 30, 2023, from https://www.shrm.org/hr-today/news/hr-news/pages/unstable-economy-return-to-workplace-union-action-in-2022.aspx

Hair, J.F., Babin, B.J., Anderson, R.E., & Black, W.C. (2018). *Multivariate Data Analysis* (8th edition). Cengage.

Hardy, R. (2017). *Multicollinearity of lagged terms in ARIMA model*. StackExchange. https://stats.stackexchange.com/questions/278027/multicollinearity-of-lagged-terms-in-arima-model

Hewamalage, H., Ackermann, K., Bergmeir, C. (2022). Forecast evaluation for data scientisats: common pitfalls and best practices. *Data Mining and Knowledge Discovery, 37*, 788-832. https://doi.org/10.1007/s10618-022-00894-5

Hyndman, R.J. (2014). *Structural breaks*. Retrieved on July 3, 2023, from https://robjhyndman.com/hyndsight/structural-breaks/

Hyndman, R.J., & Athanasopoulos, G. (2018). *Forecasting: principles and practices* (2nd edition). OTexts. https://OTexts.com/fpp2

Hyndman, R.J., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice* (3rd edition). https://otexts.com/fpp3/

Inoue, A., Jin, L., Rossi, B. (2017). Rolling window selection for out-of-sample forecasting with time-varying parameters. *Journal of Econometrics, 196*(1), 55-67. https://doi.org/10.1016/j.jeconom.2016.03.006

James et al. (2021). *An introduction to Statistical Learning with Applications in R (2nd edition)*. Cham, Springer.

Jeyanthi, P.M., Choudhury, T., Hack-Polay, D., Singh, T.P., & Abujar, S. (2022). Decision Intelligence Analytics and the Implementation of Strategic Business Management. *EA/Springer Innovations in Communication and Computing*. https://doi.org/10.1007/978-3-030-82763-2_2

Kandula, S., & Shaman, J. (2019). Reappraising the utility of Google Flu Trends. *PLoS COMPUTIONAL BIOLOGY, 15*(8). https://doi.org/10.1371/journal.pcbi.1007258

Kwiatkowski, D., Phillips, P.C.B., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics, 54*(1-3), 159-178. https://doi.org/10.1016/0304-4076(92)90104-Y

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). *The Parable of Google Flu: Traps in Big Data Analysis*. *PolicyForum*. Retrieved on September 6, 2023, from https://gking.harvard.edu/files/gking/files/0314policyforumff.pdf

Li, L., Lin, J., Ouyang, Y., Luo, X. (2022). Evaluating the impact of big data analytics usage on the decision-making quality of organizations. *Technological Forecasting and Social Change,* 175. https://doi-org.ezproxy2.utwente.nl/10.1016/j.techfore.2021.121355

Li, X., Shang, W., Wang, S., & Ma, J. (2015). A MIDAS modelling framework for Chinese inflation index forecast incorporating Google search data. *Electronic Commerce Research and Applications, 14*(2), 112-125. https://doi.org/10.1016/j.elerap.2015.01.001

Li, X., Pan, B., Law, R., & Huang, X. (2017). Forecasting tourism demand with composite search index. *Tourism Management, 59*, 57-66. https://doi.org/10.1016/j.tourman.2016.07.005

Li, Z., Xu, W., Zhang, L., & Lau, R.Y. (2014). An ontology-based Web mining method for unemployment rate prediction. *Decision Support Systems, 66*, 114-122. https://doi.org/10.1016/j.dss.2014.06.007

Liu, Z., Yang, X. (2020). Cross validation for uncertain autoregressive model. *Communications in Statistics – Simulation and Computation. 51*(8), 4715-4726. https://doi.org/10.1080/03610918.2020.1747077

Louch, P. (2014). *Workforce planning is essential to high-performing organizations*. SHRM. Retrieved February 16, 2022, from https://www.shrm.org/resourcesandtools/hr-topics/technology/pages/louch-workforce-planning.aspx

Makridakis, S., Anderson, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, P., Winkler, R. (1982). The accuracy of extrapolation (time series) methods: results of a forecasting competition. *Journal of Forecasting, 1*, 111-153. https://doi.org/10.1002/for.3980010202

Makridakis, S.G., Wheelwright, S.C., & Hyndman, R.J. (1998). *Forecasting: Methods and Applications* (3rd edition). New York, John Wiley & Sons.

Maroufkhani, P., Wagner, R., Ismail, W.K.W., Baroto, M.B., & Nourani, M. (2019). Big Data Analytics and Firm Performance: A Systematic Review. *Information, 10*(7), 226. https://doi.org/10.3390/info10070226

Mavragani, A., & Gkillas, K. (2020). COVID-19 predictability in the United States using Google Trends time series. *Scientific Reports, 10*. https://doi.org/10.1038/s41598-020-77275-9

Mclaren, N., & Shanbhogue, R. (2011). *Using internet search data as economic indicators*: Quarterly Bulletin Q2. Bank of England. https://www.bankofengland.co.uk/-/media/boe/files/quarterly-bulletin/2011/using-internet-search-data-as-economic-indicators.pdf?la=en&hash=CE738D37B3A15477EA70225F302BC94E63EA68BD

Mikalef, P., Pappas, I.O., Krogstie, J., Giannakos, M. (2018). Big data analytics capabilities: a systematic literature review and reserach agenda. *Information Systems and e-Business Management, 16*, 547-578. https://doi.org/10.1007/s10257-017-0362-y

Muehlemann, S., & Leiser, M.S. (2018). Hiring costs and labor market tightness. *Labour Economics, 52*, 122-131. https://doi.org/10.1016/j.labeco.2018.04.010

Mulero, R., & García-Hiernaux, A. (2021). Forecasting Spanish unemployment with Google Trends and dimension reduction techniques. *SERIEs, 12*, 329-349. https://doi.org/10.1007/s13209-021-00231-x

Naccarato, A., Falorsi, S., Loriga, S., Pierini, A. (2018). Combining official and Google Trends data to forecast the Italian youth unemployment rate. *Technological Forecasting and Social Change, 130*, 114-122. https://doi.org/10.1016/j.techfore.2017.11.022

Nagao, S., Takeda, F., & Tanaka, R. (2019). Nowcasting of the U.S. unemployment rate using Google Trends. *Finance Research Letters, 30*, 103-109. https://doi.org/10.1016/j.frl.2019.04.005

Niesert, R.F., Oorschot, J.A., Veldhuisen, C.P., Brons, K., & Lange, R. (2020). Can Google search data help predict macroeconomic series? *International Journal of Forecasting, 36*(3), 1163-1172. https://doi.org/10.1016/j.ijforecast.2018.12.006

Önder, I. (2017). Forecasting tourism demand with Google trends: Accuracy comparison of countries versus cities. *International Journal of Tourism Research, 19*(6), 648-660. https://doi.org/10.1002/jtr.2137

Park, S., Lee, J., & Song, W. (2017). Short-term forecasting of Japanese tourist inflow to South Korea using Google trends data. *Journal of Travel & Tourism Marketing, 34*(3), 357-368. https://doi.org/10.1080/10548408.2016.1170651

Pavlicek, J., & Kristoufek, L. (2015). Nowcasting Unemployment Rates with Google Searches: Evidence from the Visegrad Group Countries. *PLoS ONE, 10*(5), https://doi.org/10.1371/journal.pone.0127084

Pesaran, M., Timmerman, A. (2002). Market timing and return prediction under model instability. *Journal of Empirical Finance, 9*, 495-510. https://doi.org/10.1016/S0927-5398(02)00007-5

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M.Z., Barrow, D.K., Taieb, S.B., Bergmeir, C., Bessa, R.J., Bijak, J., Boylan, J.E., Browell, J., Carnevale, C., Castle, J.L., Cirillo, P., Clements, M.P., Cordeiro, C., Oliveira, F.L.C., De Beats, S., Dokumentov, A.,… Ziel, F. (2022). *Forecasting: theory and practice. International Journal of Forecasting, 38*(3), 705-871. https://doi-org.ezproxy2.utwente.nl/10.1016/j.ijforecast.2021.11.001

Pettenuzzo, D., Timmermann, A. (2017). Forecasting Macroeconomic Variables Under Model Instability. *Journal of Business & Economic Statistics, 35*(2), 183-201. https://doi.org/10.1080/07350015.2015.1051183

Pfaff, B., Wuetz, D. (2022). *Unit Root Time Series Tests. R Documentation*. Retrieved June 20, 2023, from https://search.r-project.org/CRAN/refmans/fUnitRoots/html/UnitrootUrcaInterface.html

Riley, R.D., Snell, K.I., Ensor, J., Burke, D.L., Harrell, F.E., Moons, K.G.M., Collins, G.S. (2018). Minimum sample size for developing a multivariate prediction model: PART II – binary and time-to-event outcomes. *Statistics in Medicine, 38*(7). https://doi.org/10.1002/sim.7992

Rogers, S. (2021). *Catching up on 15 years of Google Trends*. Google. Retrieved on September 6, 2023, from https://blog.google/products/search/catching-15-years-google-trends/

Rossi, B., Inoue, A. (2012). Out-of-Sample Forecast Tests Robust to the Choice of Window Size. *Journal of Business & Economic Statistics, 30*(2), 432-453. https://doi.org/10.1080/07350015.2012.693850

Rothman, P. (1998). Forecasting Asymmetric Unemployment Rates. *The Review of Economics and Statistics, 80*(1), 164-168. https://doi.org/10.1162/003465398557276

Rozanec, J.M., Fortuna, B., Mladenic, D. (2022). Knowledge graph-based rich and confidentiality preserving Explainable Artificial Intelligence (XAI). *Information Fusion, 81*, 91-102. https://doi.org/10.1016/j.inffus.2021.11.015

Rydning, J. (2022). *Worlwide IDC Global DataSphere Forecast, 2022-2026: Enterprise Organizations Driving Most of the Data Growth*. IDC Corporate. Retrieved August 30, 2023, from https://www.idc.com/getdoc.jsp?containerId=US49018922

Scharkow, M., & Vogelsang, J. (2011). Measuring the Public Agenda using Seach engine Queries. *International Journal of Public Opinion Research, 23*(1), 104-113. https://doi.org/10.1093/ijpor/edq048

Schnaars, S.P. (1986). A comparison of extrapolation models on yearly sales forecasts. *International Journal of Forecasting, 2*(1), 71-85. https://doi.org/10.1016/0169-2070(86)90031-2

Shen, Z., Zhang, Y., Lu, J., Xu, J., & Xiao, G. (2020). A novel time series forecasting model with deep learning. *Neurocomputing, 396*, 302-313. https://doi.org/10.1016/j.neucom.2018.12.084

Sheng, J., Amankwah-Amoah, J., Khan Z., & Wang, X. (2021). COVID-19 Pandemic in the New Era of Big Data Analytics: Methodological Innovations and Future Research Directions. *British Journal of Management, 32*(4), 1164-1183. https://doi.org/10.1111/1467-8551.12441

Simionescu, M., & Cifuentes-Faura, J. (2022). Can unemployment forecasts based on Google Trends help government design better policies? An investigation based on Spain and Portugal. *Journal of Policy Modeling, 44*(1), 1-21. https://doi.org/10.1016/j.jpolmod.2021.09.011

Singhania, R., & Kundu, S. (2021). Forecasting the United States Unemployment Rate by Using Recurrent Neural Networks with Google Trends data. *International Journal of Trade, Economics and Finance, 11*(6). https://ssrn.com/abstract=3801209

Smith, P. (2016). Google's MIDAS Touch: Predicting UK Unemployment with Internet Search Data. *Journal of Forecasting, 35*(3), 263-284. https://doi.org/10.1002/for.2391

Stephens-Davidowitz, S., & Varian, H. (2014). *A Hands-on Guide to Google Data*. Mountain View, CA, Google. https://people.ischool.berkeley.edu/~hal/Papers/2015/primer.pdf

Stock, J.H., & Watson, M.W. (1996). Evidence on structural instability in macroeconomic time series relations. *Journal of Business and Economic Statistics, 14*(1), 1-30. https://doi.org/10.1080/07350015.1996.10524626

Suhoy, T. (2009). *Query Indices and a 2008 Downturn: Isreali Data*. Bank of Israel. https://www.boi.org.il/media/m2ujchfj/dp0906e.pdf

Swanson, N.R. (1998). Money and output viewed through a rolling window. *Journal of Monetary Economics, 41*(3), 455-474. https://doi.org/10.1016/S0304-3932(98)00005-1

Tashman, L.J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting, 16*(4), 437-450. https://doi.org/10.1016/S0169-2070(00)00065-0

Te Brake, G. (2017). *Unemployment? Google It! Analyzing the usability of Google queries in order to predict unemployment*. [Master thesis, Universitat de Barcelona]. Dipòsit Digital. http://hdl.handle.net/2445/113761

Teng, Y., Bi, D., Xie, G., Jin, Y., Huang, Y., Lin, B., An, X., Feng, D., & Tong, Y. (2017). Dynamic Forecasting of Zika Epidemics Using Google Trends. *PLoS ONE, 12*(1). https://doi.org/10.1371/journal.pone.0165085

Terregrossa, S.J., & Ibadi, M.H. (2021). Combining Housing Price Forecasts Generated Separately by Hedonic and Artificial Neural Network Models. *Asian Journal of Economics, Business and Accounting, 21*(1), 130-148. https://doi.org/10.9734/ajeba/2021/v21i130345

Tran, U.S., Andel, R., Niederkrotenthaler, T., Till, B., Ajdacic-Gross, V., & Voracek, M. (2017). Low validity of Google Trends for behavioral forecasting of national suicide rates. PLoS ONE. https://doi.org/10.1371/journal.pone.0183149

Tuhkuri, J. (2016). *Forecasting Unemployment with Google Searches*. (ETLA Working Paper No. 35). The Research Institute of the Finnish Economy (ETLA), Helsinki. https://www.econstor.eu/handle/10419/201250

Turulja, L., Vugec, D.S., & Bach, M.P. (2023). Big Data and Labour Markets: A Review of Research Topics. *Procedia Computer Science, 217*, 526-535. https://doi.org/10.1016/j.procs.2022.12.248

Van den Brakel, J., & Krieg, S. (2010). *Schatten van maandcijfers over de beroepsbevolking*. CBS. Retrieved January 23, 2023, from https://www.cbs.nl/-/media/imported/onze-diensten/methoden/dataverzameling/korte-onderzoeksbeschrijvingen/documents/2010/27/ebb-maandcijfers-schatten.pdf?la=nl-nl

Varian, H.R. (2014). Big Data: New Trick for Econometrics. *Journal of Economic Perspectives, 28*(2), 3-28. https://www.aeaweb.org/articles?id=10.1257/jep.28.2.3

Velliangiri, S., Alagumuthukrishnan, S., & Iwin Thankumar Joseph, S. (2019). A Review of Dimensionality Reduction Techniques for Efficient Computation. *Procedia Computer Science, 165,* 104-111. https://doi.org/10.1016/j.procs.2020.01.079

Vicente, M.R., Menéndez, J.L., & Pérez, R. (2015). Forecasting unemployment with internet search data: Does it help to improve predictions when job destruction is skyrocketing? *Technological Forecasting and Social Change, 92*, 132-139. https://doi.org/10.1016/j.techfore.2014.12.005

Vosen, S., & Schmidt, T. (2011). Forecasting Private Consumption: Survey-Based Indicators vs. Google Trends. *Journal of Forecasting, 30*(6), 565-578. https://doi.org/10.1002/for.1213

Wang, Y., & Hao, X. (2023). Forecasting the real prices of crude oil: What is the role of parameter instability. *Energy Economics, 117*, 106482. https://doi.org/10.1016/j.eneco.2022.106483

Westerhuis, G. (2022). *Detailhandel moet lonen en welzijn verhogen in strijd om personeel.* ABN Amro. Retrieved February 18, 2023, from https://www.abnamro.nl/nl/media/detailhandel-moet-lonen-en-welzijn-verhogen-in-strijd-om-personeel_tcm16-151626.pdf

Woo, J., Owen, A.L. (2018). Forecasting private consumption with Google Trends data. *Journal of Forecasting, 38*(2), 81-91. https://doi.org/10.1002/for.2559

Yu, G., Schwartz, Z. (2006). Forecasting Short Time-Series Tourism Demand with Artificial Intelligence Models. *Journal of Travel Research, 45*(2). https://doi.org/10.1177/0047287506291594

Zagheni, E., & Weber, I. (2015). Demographic research with non-representative internet data. *International Journal of Manpower, 36*(1), 13-25. https://doi.org/10.1108/IJM-12-2014-0261

Zhang, H., He, Q., Jacobsen, B., Jiang, F. (2019). Forecasting stock returns with model uncertainty and parameter instability. *Journal of Applied Econometrics, 35*(5), 629-644. https://doi.org/10.1002/jae.2747

Zhu, J.J.H., Wang, X., Qin, J., & Wu, L. (2012). Assessing Public Opinion Trends based on User Search Queries: Validity, Reliability, and Practicality. *65th Annual Conference of the World Association for Public Opinion Research.* https://scholars.cityu.edu.hk/en/publications/publication(54cc3372-239d-40cd-8e3f-1a6e875ecfd7).html