MSc - Computer Science Research Proposal

A Machine Learning Approach to Quantifying and Relating the Determinants of Unemployment in South Africa

Author:
Rudzani Mulaudzi
(0601737R)

Supervisor:
Dr Ritesh Ajoodha

School of Computer Science and Applied Mathematics
The University of the Witwatersrand, Johannesburg

May 16, 2020

A research proposal submitted in fulfilment of the requirements for the degree of MSc in Computer Science offered by the School of Computer Science and Applied Mathematics

Abstract

According to Statistics South Africa's 2019 quarter 4 quarterly labour force survey, the unemployment rate of South Africa is currently 29%. This unemployment rate puts the country in the top ten countries with the highest unemployment rates in the world. Despite numerous policy interventions, the unemployment rate has been trending upwards since the dawn of democracy. Economists and politicians have warned that this failure to address unemployment makes South Africa a social unrest ticking time-bomb. This problem demands urgent attention and it is now labelled a humanitarian crisis.

The only known way to resolve unemployment sustainably is through public policy. Currently, public policy directions are informed by forecasts derived through economic (traditional statistical) models and expert judgement. These models are, however, suitable when the data is stationary and white noise generated. A preliminary analysis of South Africa's unemployment rate shows that the country's unemployment rate is asymmetric, seasonal, upward trending, and non-stationary. These attributes make traditional statistical models prone to error when trying to forecast South Africa's unemployment.

In the preliminary analysis, traditional statistical models were used to model unemployment in South Africa and the average Root Mean Squared Error (RMSE) across the models was 2,21 with the leading model being Holt, which had an RMSE of 1,99. This is a high RMSE as expected, especially when compared to neural networks and regression techniques which have demonstrated an RMSE of approximately 0,1 in North America, Asia, and Europe.

Therefore, this research will contribute to reducing South Africa's unemployment by improving South Africa's ability to forecast the country's unemployment rate. This will be achieved by introducing machine learning techniques to South African unemployment forecasting. The research will also use machine learning to determine the key drivers of unemployment and their relative influence on each other and unemployment.

Contents

Chapter 1

Introduction

South Africa's unemployment rate is currently 29%, this is the highest it has been since the global financial recession in 2008 [Statistics South Africa 2019; Dludla 2019]. This is viewed by economists and policymakers as one of South Africa's biggest problems [Jones 2019; Belling 2020]. According to Aiken [1996]; Pelaez [2006], many see the unemployment rate as a proxy measure for the health of a country. By this metric, South Africa is notably unhealthy. The country is in the top ten countries with the highest unemployment rates across the world [TradingEconomics 2019b; Meyer 2014].

There has been a wealth of research to understand this problem and some of its core drivers have been identified: yet the problem persists [Fourie 2011; Meyer 2014; De Lannoy *et al.* 2018]. Brynard [2011] states that a key reason for the persistence of South Africa's high unemployment is that the country's policy making is often too broad and not focused enough. Adding that, although there is a wealth of data relating to unemployment in South Africa, not all the available data is leveraged in the policy making process.

Solutions to unemployment can either be programmatic or systemic. The programmatic being targeted at particular groups of unemployed persons. These are typically executed by companies, government municipalities, and non-profit organisations. Shankar *et al.* [2016] studied programmatic interventions and some of the programmes studied are the IT management company, EOH's youth job creation initiative and Mentec Foundation's train and place initiative. These programmes were found to be effective for unemployed youths placing over 100 000 of them into jobs.

Systemic interventions are those that aim to put in place measures that sustainably reduce unemployment across the board with the aim being to resolve the problem permanently [Levinsohn 2007; Brynard 2011]. Examples of this would be changes in the education system to make it aligned to employment opportunities or changes in government policies to prioritise job creation [De Lannoy *et al.* 2018]. According to Levinsohn [2007], one the key policies to reduce unemployment has been the wage subsidy policy in South Africa. This policy enables more first-time job seekers to get employed because salaries to these employees

are subsidised by the government.

Even though programmatic interventions are beneficial because they get a number of people into jobs, these have limited impact when compared to systemic intervention [De Lannoy *et al.* 2018]. Therefore, where possible, systemic interventions are preferred over programmatic ones. In order to effect systemic changes, policymakers are reliant on forecasts to determine which policy options to put in place [Makridakis 1988; Levinsohn 2007]. Currently, these forecasts are generated through traditional statistical methods [Makridakis *et al.* 2018; Brooks 2014; Hyndman and Athanasopoulos 2018].

Brynard [2011] states that in developing unemployment policy the labour market and its dynamics must be understood. In South Africa, this understanding comes from Statistics South Africa's quarterly labour force survey. Which is a database that provides data on how unemployment changes yearly and quarterly. It also provides the demographic profile of the unemployed persons: their ages, location, race, gender, and education levels [Levinsohn 2007; Brynard 2011]. Once the labour market dynamics are understood previous policy interventions are studied for their success or failures [Brynard 2011]. Thereafter, scenarios are run for various policy options with judgments from experts used to determine possible reductions to unemployment for various policy options being considered [Brynard 2011; Levinsohn 2007]. Accurate forecasting is important throughout the policymaking process [Brynard 2011].

In order to remain consistent with previous forecasting research, traditional statistical methods are defined distinctly from machine learning methods. Makridakis [1988]; Makridakis *et al.* [2018]; Cerqueira *et al.* [2019] define traditional statistical models as "models that assume constancy of patterns and/or constancy of relationships or predictability in the way that changes occur". This distinction and definition is well established in forecasting and statistical literature [Makridakis *et al.* 2018; Cerqueira *et al.* 2019]. Concretely, Makridakis *et al.* [2020 2018] refer to a number of models as traditional statistical models: Naïve Forecast, Random Walks, Simple Exponential Smoothing (SES), Holt, Holt-Winters, Damped Exponential Smoothing, Theta Method, Autoregressive Integrated Moving Average (ARIMA), Vector Autoregressive (VAR), Threshold Autoregressive (TAR), and Generalized Autoregressive Conditional Heteroskedasticity (GARCH) models.

These statistical methods are historically the key methods for predicting unemployment [Aiken 1996; Olmedo 2014; Hall 2018; Katris 2019ab]. However, these models have limitations as they primarily forecast data that is stationary, linear, and symmetric [Olmedo 2014; Aiken 1996; Brooks 2014; Hyndman and Athanasopoulos 2018; De Gooijer and Hyndman 2006]. Being only able to cater for nonlinear data with significant manipulation from the data analyst [Cook and Hall 2017; Hyndman and Athanasopoulos 2018; De Gooijer and Hyndman 2006]. De Gooijer and Hyndman [2006]; Clements *et al.* [2004] add that nonlinear modelling using traditional statistical (i.e. TAR and GARCH) approaches is still in its 'infancy'. They found no evidence in performance improvement – measured by error rate –

when these models were compared to linear models on financial and macroeconomic data. This makes these traditional statistical methods limited when it comes to forecasting unemployment because unemployment data tends to be nonstationary, nonlinear, asymmetric, seasonal, and trendy [Hall 2018; Aiken 1996; Olmedo 2014; Katris 2019a].

Aiken [1996] was the first researcher to employ a multi-layer perceptron (MLP) to address the challenges associated with forecasting unemployment with traditional statistical methods. The MLP model outperformed traditional statistical methods, as measured by the mean average error (MAE). Since then, several researchers have demonstrated that machine learning models have a higher accuracy than traditional statistical methods when forecasting unemployment rates [Katris 2019a; Cook and Hall 2017; Mahipan *et al.* 2013; Atsalakis *et al.* 2007; Kouziokas 2019; Sermpinis *et al.* 2014; Sharma and Singh 2016]. Although, traditional statistical methods are still preferred when it comes to forecasting unemployment, machine learning models are showing a lot of promise. There is currently an increase in research to improve the forecasting accuracy and adoption of machine learning models [Makridakis *et al.* 2018; Hall 2018; Cook and Hall 2017; Sharma and Singh 2016; Sermpinis *et al.* 2014]. Much of this work is in North America, Europe, and Asia and not much in Africa[Makridakis *et al.* 2018; Hall 2018; Cook and Hall 2017; Sharma and Singh 2016; Sermpinis *et al.* 2014].

As the current application of machine learning models to forecasting unemployment is still at experimental stages [Makridakis *et al.* 2018; Hall 2018; Cook and Hall 2017]. Researchers thus far have focused on forecasting unemployment t-periods in the future with very few using the data to determine which features drive unemployment [Kreiner and Duca 2019; Hall 2018; Katris 2019a]. The majority of these models are univariate models using past unemployment rates as the only predictor of future unemployment [Makridakis *et al.* 2018; Cook and Hall 2017; Hall 2018; Olmedo 2014; Kouziokas 2019; Katris 2019ab]. Therefore, the motivation for this research is to apply machine learning models to forecast unemployment in South Africa as this could improve systemic interventions through policymakers. Furthermore, using a multivariate approach to determine which features drive unemployment in South Africa. This kind of work is currently limited in the country and yet very needed as South Africa's unemployment rate is high and rising.

The remainder of this chapter will discuss the problem, purpose, and research questions that this research intends to address. With the rest of the research proposal consisting of four sections. The first being the background literature, chapter 2, that provides a background to the traditional statistical methods that are used to forecast unemployment rates. The second section, chapter 3, is the related work section that provides a literature review on the use of machine learning techniques to forecast unemployment across the world. The third section, chapter 4, is the methodology section which discusses how this research will be carried out, the data to be used and some preliminary results for the South African context. Lastly, chapter 5 describes the overall plan to carry out the research over a eighteen-month period.

1.1 Problem Statement

According to Statistics South Africa [2019], South Africa's unemployment rate increased by 1,4% to 29,0% in the second quarter of 2019. They add that this figure is the highest unemployment rate since the global recession of 2008. Trending along the same trajectory, youth unemployment has also increased to 56,4% for 15 to 24-year-olds and 35,6% for 25 to 34-year-olds. South Africa's inability to resolve this social challenge is a humanitarian crisis and has been dubbed South Africa's biggest problem [Jones 2019; Belling 2020].

Figure 1.1 provides a visual depiction of South Africa's unemployment rate from 1998 to 2019. The figure shows that South Africa's unemployment rate has been persistently high since the dawn of democracy with the rate hovering between 20 and 30% from 1998 to 2019. The country has continually been ranked in the top ten of countries with the worst unemployment rates over the same time period [TradingEconomics 2019b; Meyer 2014].

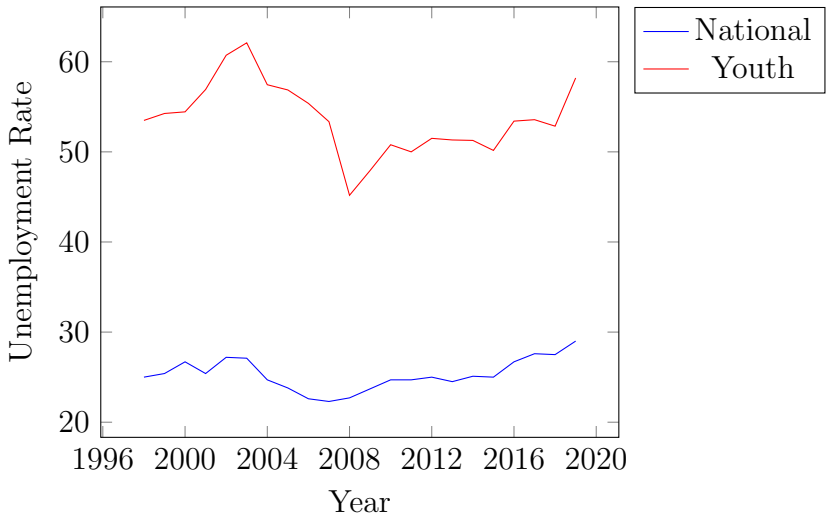


Figure 1.1: Unemployment in South Africa from 1998 to 2019 [Statistics South Africa 2019].

Forecasting techniques are essential in the process of attempting to address unemployment because they enable policymakers to develop systemic interventions. Currently, the dominant approaches to forecasting unemployment are traditional statistical models: ARIMA [Hall 2018; Mahipan *et al.* 2013], VAR [Olmedo 2014], and TAR [Katris 2019a]. These approaches have limitations because they preferably require the data to be nonseasonal, stationary, linear, symmetric, and without correlations in the dependant variables [Brooks 2014].

The movement of unemployment rates is typically nonstationary, nonlinear, asymmetric, seasonal, and trendy [Hall 2018; Aiken 1996; Olmedo 2014; Katris 2019a]. This makes traditional statistical models prone to error when forecasting unemployment because these are best when the data is stationary, linear and symmetric [Olmedo 2014; Aiken 1996; Brooks

2014; Hyndman and Athanasopoulos 2018; De Gooijer and Hyndman 2006]. These models are also biased as their autoregressive nature means that the models use a weighted average method where the most recently seen data points are given the largest weighting, which makes them biased towards recently seen data points [Makridakis 1988]. These models are also not not able to accommodate economic shocks which cause sudden changes in the movement of unemployment rates such as economic recessions and booms [Aiken 1996; Olmedo 2014; Makridakis 1988; Makridakis *et al.* 2009].

The appreciation for the importance of forecasting in solving unemployment and a need for accurate approaches thereof, has led a number of forecasters to deploy machine learning models instead of traditional statistical models alone. Right now, machine learning models are being explored as alternatives, but, they are yet to reach mainstream adoption [Aiken 1996; Katris 2019b; Moriwaki 2020; Pavlicek and Kristoufek 2015; Makridakis *et al.* 2018 2020]. These machine learning models do not only present an opportunity to improve unemployment forecast accuracy [Hall 2018]. They also enable key drivers of unemployment to be discovered from the data [Katris 2019a; Kreiner and Duca 2019]. Furthermore, through probabilistic machine learning models, the influence of various features on each other and unemployment can also be represented [Dahlhaus and Eichler 2003; Ajoodha 2018].

The usage of traditional statistical models, inherently requires that the researcher make upfront choices about the features to be considered in the model [Makridakis 1988]. These selections result in forecasting errors, as the selection of features is not done objectively and causes biases. Traditional statistical models are also not designed for big data settings and thus as the data grows in variety and volume they reach their computational limits [Cerqueira *et al.* 2019; Makridakis 1988]. There are machine learning models such as neural networks, elastic net, principal component analysis, and multivariate adaptive regression splines that offer an advantage over this as they they enable unemployment to be modelled with little to no restrictions on the data volume and feature variety [Katris 2019a; Hall 2018; Aiken 1996].

Machine learning models, as listed above, come with many benefits over traditional statistical models such as the ability to model nonlinear data, less stringent data preprocessing requirements, and long term memory capabilities. However, these models are applied to a limited extent in South Africa. These are particularly needed models in the country as the country's challenges with unemployment do not seem to have an end in sight. These models could assist by providing better forecasts and uncovering data determined factors that influence the unemployment rate's movements.

1.2 Purpose Statement

To date, traditional statistical models have proven to be a reliable mechanism of forecasting economic data, provided that the data is preprocessed to ensure it meets certain data quality standards. These standards typically include the data being detrended, made stationary,

without asymmetries, and preferably linear. However, unemployment data does not meet these data standards presenting a challenge for traditional statistical forecasting models. To overcome these challenges, neural networks and regression techniques are being explored to provide better forecasting predictions for unemployment [Hall 2018; Cook and Hall 2017].

This research intends to contribute to the currently underway attempts to use machine learning to improve the unemployment rate forecasting. The research will explore various machine learning techniques to determine which are best suited for predicting unemployment in South Africa. The techniques will include regression, neural networks, and probabilistic models. Furthermore, the research will investigate and establish which drivers of unemployment are most important: considering their relative contribution to unemployment.

To discover the key drivers of unemployment, a data-driven approaches will be taken with all possible drivers of unemployment being considered. Chandrashekar and Sahin [2013] states that there are three types of feature selection methods: Filter, wrapper, and embedded. Filter methods rank features based on a statistical score that represents their relative significance in predicting the target variable. One example is information gain, which ranks features based on the relative information that is gained by their inclusion or exclusion in the model. This technique is ideal for this research as there over 1400 possible features. These features will be ranked to determine which result in the highest information gain.

Chandrashekar and Sahin [2013] describes embedded methods as feature selection methods that allow subsets of the feature set to be selected and evaluated for performance with the one with the lowest error rate being selected as the best model. They add that these methods are more efficient than wrapper methods even though wrapper methods work in a similar fashion as embedded methods. Therefore, embedded methods will be used for this research. Kreiner and Duca [2019]; Hall [2018] used embedded methods, LASSO and Elastic Net, to determine features that best predict the USA unemployment rate's movements. Therefore, Elastic Net will be used for feature selection in this research as it offers all the advantages of LASSO as well as the Ridge regression technique. Elastic Net penalises models that overfit (LASSO) or has over reliance on a subset of features (Ridge).

1.3 Research Questions

To achieve the purpose of this research the following research questions will be explored:

1. Which features are key predictors of unemployment?
2. What information is gained from each feature to predict unemployment?
3. To what extent do the various features of unemployment influence each other?

The modelling of unemployment of this nature has thus far been limited in South Africa.

This research contributes to expanding the body of knowledge on applying machine learning models in the context of unemployment in South Africa. Therefore, this research will make the following contributions:

1. Use machine learning (i.e. neural networks, regression, and Bayesian models) algorithms to predict unemployment rates over multiple classes (i.e. youth, females and geographic) and various time horizons (i.e. 1, 2, 4, and 8 quarters ahead).

2. Use information gain and regression (i.e. Elastic Net) techniques to determine which features contribute significantly in predicting unemployment rates.

3. Use probabilistic techniques to discover how various features influence one another.

4. Advice policymakers on opportunities to reduce unemployment based on the findings of this research.

The rest of this paper is organised as follows. In chapter 2, unemployment is defined along with an overview of traditional statistical methods, this chapter serves as the theoretical background for this work. In chapter 3, a review of the literature is provided. In chapter 4, the methodology that will be employed in this research is discussed as well as the data to be used and some preliminary results for the South African context. The research proposal concludes with chapter 5, which describes the overall plan to carry out the research over a eighteen-month period.

Chapter 2

Background

Statistics South Africa [2019]'s definition of unemployment will be adopted for this research: *"Unemployment refers to a period or state where a person is not currently in a paid job opportunity and has taken active steps, although unsuccessful, to look for a job or to start a business and is available to take up a job should it be offered"*. This is the only definition that is used for unemployment in South Africa.

This chapter will provide a review of the background literature that is required for this research. In section 2.1 an economic perspective on unemployment is provided, this view is the foundational view through which unemployment is understood. Section 2.2 provides a review of traditional methods for forecasting unemployment: including traditional statistical methods. The chapter will close with known issues with traditional forecasting approaches, most of which are addressed by the models in chapter 3.

2.1 Economics of Unemployment

Through the years economic models have been developed in an attempt to provide tools that would enable us to better understand unemployment. Taylor *et al.* [2016] states that in order to understand how unemployment is modelled, it must be thought about over two periods: short run and long run. Economists do not define the short run and long run in terms of time periods but rather in the context of what companies can change or control and what they are unable to. In the short run costs are fixed but in the long run they are variable [LumenLearning 2020]. Firms in the short run are unable to easily respond to their environments whilst in the long run they are able to do so.

2.1.1 Unemployment in the Short Run

According to Taylor *et al.* [2016], unemployment in the short run is determined by the supply and demand for labour at certain wages. Firms, who hire people, will typically make decisions about how much labour they need based on the economic performance of their

firms and their perceptions of the probable macroeconomic performance of the economy as a whole. If the economy is performing well and their firms are growing then these firms will have an increased demand for labour which often leads to higher wages as the supply of labour remains fairly constant in the short run.

When an economy is not performing well, firms will reduce their demand for labour as well as their labour costs such as wages. This reduction in the demand for labour will result in some of those employed becoming unemployed and it will also keep others unemployed for longer as some might not be willing to work at lower wages. This kind of unemployment is referred to as cyclical unemployment, it is the unemployment that can be directly correlated to business cycles. Cyclical unemployment is what explains why employment increases in economic booms and drops in economic recessions. Figure 2.1 provides a visual description of cyclical unemployment [Taylor *et al.* 2016].

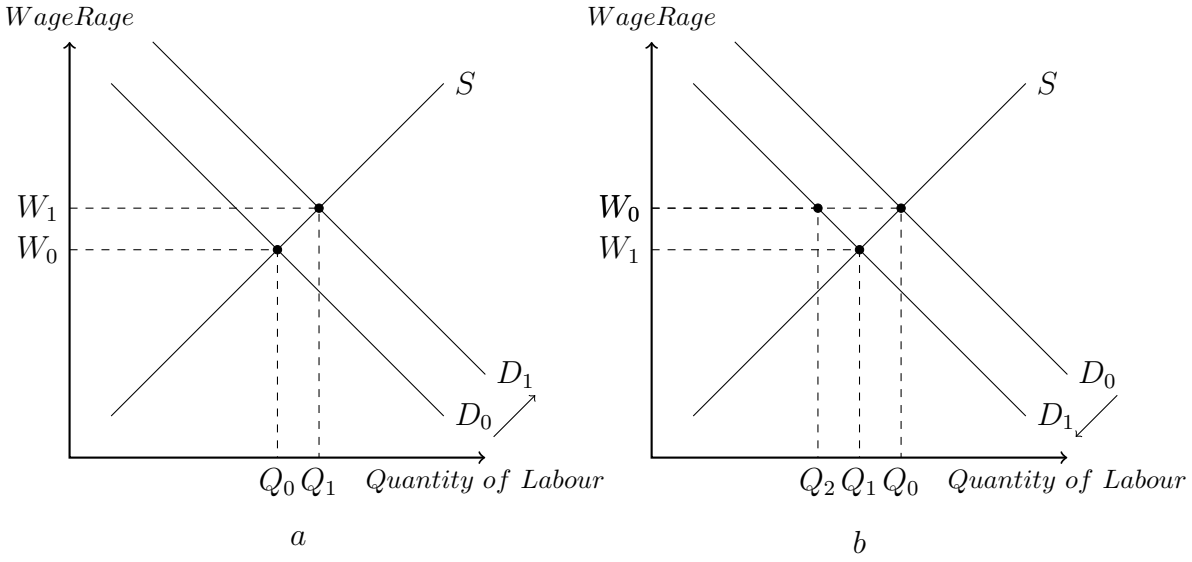


Figure 2.1: Cyclical unemployment illustration for a growing economy (shown in a, with rising demand for labour and wages) and a declining economy (shown in b, with a decline in demand for labour resulting unemployment: the gap between Q_0 and Q_2). Those employed when an economy is growing, enjoy the benefit of increased wages, however, during declines, their wages do not increase and those entering the market are offered lower wages for the same job.

The relationship between cyclical unemployment and cyclical output as described, is often referred to as Okun's Law [Marinkov and Geldenhuys 2007]. Output being a key determinant of growth in an economy and measured by gross domestic product (GDP). Okun's Law simply put, states that there is an inverse relationship between output and unemployment, thus when output increases we expect unemployment to decrease and vice versa [Marinkov and Geldenhuys 2007]. Therefore, output or growth is a key determinant of unemployment.

This relationship is represented by Equation 2.1 to Equation 2.5 [Marinkov and Geldenhuys 2007; Ball *et al.* 2019]:

$$\triangle UR_t = [\%\triangle LFPR_t + \%\triangle POP_s t] - [\%\triangle Y_t - \%\triangle APL_t], \tag{2.1}$$

$$y_t^c \equiv y_t - y_c^p, \tag{2.2}$$

$$u_t^c \equiv u_t - u_c^p, \tag{2.3}$$

$$u_t^c = \gamma y_t^c + \varepsilon_t, \ and \tag{2.4}$$

$$u_t^c = \sum_{i=1}^{m} \beta_i u_(t-i)^c + \sum_{i=1}^{m} \gamma_i y_(t-i)^c + \varepsilon_t. \tag{2.5}$$

In Equation 2.1, $LFPR$ denotes labour force participation rate, with POP denoting the working-age population, Y the real GDP, and APL the average product of labour (defined as Y/E). Equation 2.2 denotes the output gap, where y_t denotes the actual output (real GDP) and y_c^p the potential output and thus y_t^c is the output gap, similarly, Equation 2.3 denotes the employment gap with u_t denoting the actual unemployment rate, u_c^p is the potential unemployment rate and thus u_t^c is the unemployment rate gap. Furthermore, Equation 2.4 estimates Okun's Law in the short run, where γ is the estimated Okun's coefficient and ε_t is the error term. Lastly, Equation 2.5 is the equation for estimating Okun's Law in the medium run where β is the coefficient that enables a non-static relationship in Okun's Law in the medium term.

Marinkov and Geldenhuys [2007] investigated how strongly Okun's Law applies in South Africa and they found that for every 1% increase in real GDP, unemployment would decrease by between 0.164 and 0.772 percentage points. Thus, increases in output reduced cyclical unemployment, although negligibly so. They also found that in the medium run, Okun's Law is stronger in South Africa, with unemployment decreasing by twice as much as observed in the short run. Furthermore, there was proof of asymmetries in the movement of the unemployment rate as the relationship between unemployment and output proved to be strong during recessions than other economic times.

The relatively weak relationship between unemployment and output is due to structural changes that South Africa has undergone: apartheid, economic sanctions, transitions to democracy and democracy [Marinkov and Geldenhuys 2007]. A global study by Ball *et al.* [2019] found that although there was evidence of Okun's Law in South Africa, the relative $R - squared$ was extremely low compared to other countries. Therefore, Marinkov and Geldenhuys [2007]; Ball *et al.* [2019] stress that when researching unemployment in the developing world, cyclical unemployment is just one part of the puzzle, frictional and structural unemployment should also be considered. These are discussed next.

2.1.2 Unemployment in the Long Run

Long run unemployment in an economy can be thought of as the unemployment that is present even when the economy is performing well and growing each year. Nations like China have grown by 6% per year on average, however, their unemployment rate has been 4% on average over the last 15 years [Trading Economics 2019]. Therefore, this 4% unemployment rate that is present in growing China, is precisely what 'long run unemployment' attempts to explain. According to Taylor *et al.* [2016] four most common reasons for this unemployment are businesses' fluctuating demand for labour; skills mismatch between those hiring and those seeking employment; reduction in productivity levels of workers; and government interventions to try and improve the lives of workers. Taylor *et al.* [2016] explained these four:

- Businesses' fluctuating demand for labour
  The general economic performance of a firm will determine the quantum of employees that the firm will need. A firm that is growing will demand more employees whilst one that is in decline will demand less. Therefore, when those who are employed in a declining firm lose their job, this causes unemployment as a worker will not simply walk into a new job in a day but must instead search for work where they are most suitable. This type of unemployment is referred to as frictional unemployment. This is particularly complex in an economy like South Africa that is not growing because finding a new job is extremely hard as there are not enough jobs for those that are currently unemployed [IOL 2019]. Andolfatto [2006] adds that key to the prolonging of frictional unemployment or search unemployment is that there is imperfect information in the market. In other words, there is information asymmetry between employers and job seekers.

- Skills mismatch between those hiring and those seeking employment
  Often those who are unemployed possess particular skill sets that they desire to offer to employers. However, these employers have their own list of desired skills. When the skills that the unemployed person has is different from what employers want, this creates a mismatch of skills which prolongs the time that is spent in unemployment. The issue has been flagged as a key issue in South Africa and the suggested solution is to improve basic and higher education [BusinessTech 2019; Department of Higher Education and Training 2019; BCG 2019; Visser and Arends 2016]. Furthermore, South Africa's unemployed are often unskilled or semi-skilled persons.

- Reduction in productivity levels of workers
  In some cases, those who are employed get comfortable within their jobs and start dropping the standard that they would normally perform at. This productivity drop means that employers are not getting as much return from hiring those employees and therefore might reduce wages to try and discourage employees out of the organisation.

- Government interventions to try and improve the lives of workers
  Sometimes governments will create laws and policies that make it difficult for employ-

ers to hire and fire people. This results in employers being more cautious in their hiring decisions and processes. Some of these interventions by governments could be the introduction of minimum wages as was recently done in South Africa [De Lannoy *et al.* 2018]; increased unionisation of a particular industry; forced contributions from employers towards employee benefits; and labour laws that make employment termination legally expensive.

The relationship between unemployment and inflation has been heavily investigated and its application in the long run. In 1958, A.W Philips observed that as wages of employees increased, this would naturally be followed by more people joining the workforce and finding work which leads to an increase in employment. The increase being a consequence of a wage change which can be understood as a form of inflation. The phenomena became known as Philips Curve, even though the version of it today, which suggests that there is a trade-off between unemployment and inflation, is a modification on Philips' original thesis [Vermeulen 2017]. Figure 2.2 show's this relationship, the higher the inflation rate ($ir0$) the lower the unemployment rate ($ur0$).

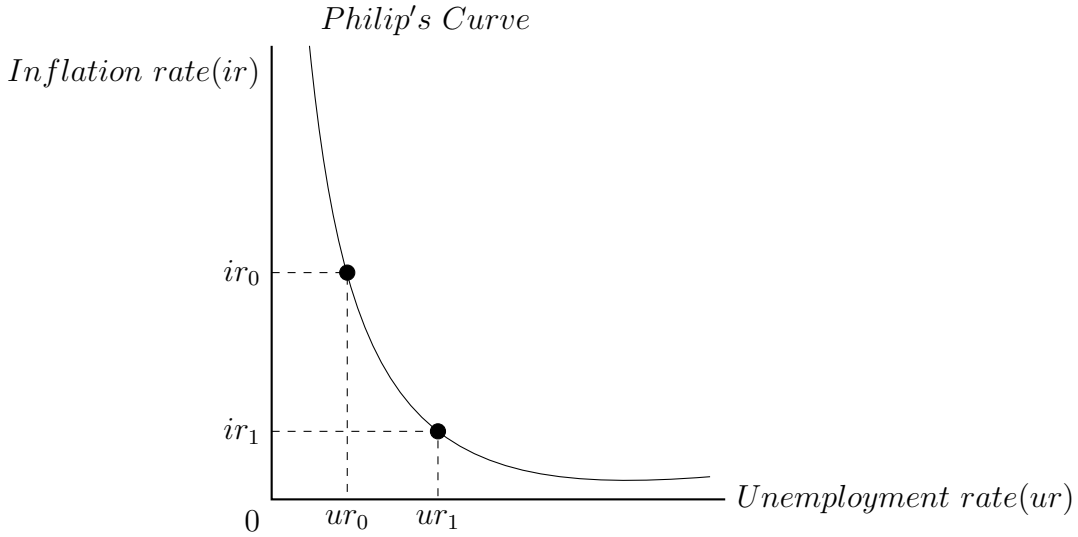


Figure 2.2: Philips Curve, showing that a decline in inflation leads to an increase in the unemployment rate. Simply put a decline in wages will result in an increase in unemployment as people are unwilling to work at lower wages.

Vermeulen [2017]'s seminal research on the applicability of Philips Curve in South Africa, found that the South African unemployment rate trends between 24% and 30% irrespective of what the inflation rate is. The unemployment rate in South Africa was in this range before the inflation targeting regime of the South African Reverse Bank (SARB) and has maintained the rate even now in the inflation targeting years [TradingEconomics 2019a]. Vermeulen [2017] demonstrate in their research that unemployment in South Africa seems to be inelastic to inflation.

Hodge [2002] noted that in South Africa, inflation and unemployment have been moving independently of each other for the recorded history. A relationship between these two variables was only observed during selected periods, particularly between 2008 and 2014, where a positive relationship between inflation and employment growth was observed [Vermeulen 2017]. It is also worth noting that a key challenge in South Africa, has been that, although jobs have been created in the country, the labour force has increased at a faster rate with new entrants to the labour force possessing inadequate skills for the jobs that are available [Vermeulen 2017; Marinkov and Geldenhuys 2007; Hodge 2002]. Thus pointing to the importance of considering the structural nature of the South African unemployment crisis.

The economic perspective on unemployment provides the global understanding of unemployment as well as its characteristics which are important for forecasting purposes. From the overview provided, it is clear that well established econometric models such as Okun's Law and Philips Curve weakly explain unemployment in South Africa [Marinkov and Geldenhuys 2007; Vermeulen 2017; Hodge 2002]. Furthermore, modelling business cycles is important when forecasting unemployment as these directly relate to the increase or decline in demand for labour [Taylor *et al.* 2016]. There is also evidence that unemployment is impacted by labour regulations and other well-meaning government interventions [Taylor *et al.* 2016]. This perspective is not often included in the statistical views to be discussed in section 2.2. Therefore, this research will use the economic perspective as key inputs to forecasting unemployment.

2.2 Traditional Unemployment Forecasting Models

Unemployment is considered by many as an indicator of the economic health of a nation [Aiken 1996; Pelaez 2006]. As such forecasting it has been a priority activity of econometricians since the dawn of forecasting. Okun's Law and Philips Curve are examples of economic models that can be used to forecast unemployment. Okun's Law stating that there is an inverse relationship between economic growth and unemployment, therefore, when economic growth is upwards unemployment declines [Marinkov and Geldenhuys 2007]. Philips Curve, on the other hand, states that inflation and unemployment have a trade-off relationship: increasing inflation leads to a reduction in unemployment [Vermeulen 2017].

Brooks [2014]; Cook and Hall [2017] state that there are two approaches to forecasting unemployment: structural and non-structural approaches. Structural being those approaches that are underpinned by a strong theoretical model or framework that relates variables to each other, whilst non-structural focus on properties and relationships within the data without explicit reliance on a particular economic theory. Okun's Law and Philips Curve are example of structural approaches, these are well established econometric frameworks for predicting unemployment as discussed in section 2.1.

This section will discuss non-structural approaches that are used to forecast unemployment. Sub-section 2.2.1 and sub-section 2.2.2 will provide the literature review of univariate and multivariate approaches. These models forecast unemployment based on properties discovered from the data.

2.2.1 Univariate Linear Models

Univariate models are linear models that forecast a variable based on its lag values and associated error terms only [Brooks 2014]. The ones that are most commonly used for forecasting unemployment will be discussed in this section.

Naïve Forecasts

Naïve forecasts are those that assume what has happened in the past will continue happening in the future. Therefore, future unemployment is forecasted based on recently observed unemployment data only. Barnichon and Nekarda [2013] used this technique in their research as a baseline to determine if their alternative unemployment forecasting model was better.

Autoregressive Integrated Moving Average (ARIMA)

Montgomery *et al.* [1998] state that the most commonly used univariate statistical model is the ARIMA model. The model states that the dependant variable, u_t, can be determined using a linear combination of past and present values of the same variable i.e. u_{t-i}, were $t-i$ is a time period i times before t. The ARIMA model is made of a combination of three common time series models:

1. *Autoregressive (AR)*, which according to Brooks [2014] is the linear regression part of the model, were the current variable, u_t, depends only on the variable's past values and an error term. This can be presented as $\phi(L)u_t = \epsilon$, were $\phi(L)$ is a linear combination as shown in Equation 2.6, with L being the lag operator, which returns a series' previous element, for example, L^2 and L^i would return two previous data points and i previous data points respectively. The ϵ is the error term.

2. *Integrated (I)*, Brooks [2014] defines the 'I' in ARIMA by stating that "If a non-stationary series, y_t must be differenced d times before it becomes stationary, then it is said to be integrated of order d." In other words, 'I', is the mechanism to ensure the data being considered is stationary.

3. *Moving Average (MA)*, are linear combinations of independent variables that are used to predict a dependant variable. These variables are generated independently of each other, are not correlated with one another, and have a constant mean and variance [Brooks 2014]. MA is given by $y_t = \theta(L)u_t + \mu$, were $\theta(L)$ a linear combination as shown in Equation 2.7, although it is in a negative form in this context [Brooks 2014].

Several researchers and institutions use the ARIMA model to forecast unemployment across the world [Montgomery *et al.* 1998; Barnichon and Nekarda 2013; Mahipan *et al.* 2013; Funke 1992; Katris 2019a; Hall 2018; Jelena *et al.* 2017]. In Equation 2.8 Montgomery *et al.* [1998]'s version of the ARIMA equation is shown where u_t is the unemployment rate being estimated. Variables p, d, and q are natural numbers used to set the autoregression order, the differencing order that makes the time series stationary and moving average window.

$$\phi(L) = 1 - \phi_1 L - ... - \phi_2 L^p, \tag{2.6}$$

$$\theta(L) = 1 - \theta_1 L - ... - \theta_q L^q, \ and \tag{2.7}$$

$$\phi(L)(1-L)^d u_t = c + \theta(L)\epsilon + t. \tag{2.8}$$

The ARIMA model can cater for different movements in time series data through modified versions of the ARIMA: Seasonal ARIMA (SARIMA) can cater for data with seasonality and Fractional ARIMA (FARIMA) allows for improved modelling of nonstationary data.

ARIMA models like all other univariate models are limited by the fact the models only rely on a variable's past values to determine its future value [Brooks 2014; Hyndman and Athanasopoulos 2018]. These models only model linear data and are unable to capture regime shifts (structural changes) [Brooks 2014; Hyndman and Athanasopoulos 2018]. Nor are they able to model data that is asymmetric [Montgomery *et al.* 1998]. Regime shifts and asymmetries are common in South Africa's unemployment rates' movements along with seasonality and an upward trend [Statistics South Africa 2019; Vermeulen 2017]. Furthermore, the ARIMA model requires extensive preprocessing when the underlying data is seasonal and trendy [Hyndman and Athanasopoulos 2018]. Therefore, multivariate models are explored as they address some of the limitations of univariate models

2.2.2 Multivariate Models

Multivariate models are generalisations of univariate models. They essentially allow for the forecasting of multiple dependent variables and one or more independent variables simultaneously. The common models used to model unemployment will be discussed in this sub-section.

Vector Autoregressive Models (VAR)

The generalisation of ARIMA models from single dependent variables to multiple is Vector Autoregressive Models (VARs) [Montgomery *et al.* 1998]. A VAR is a regression model with more than one dependent variable. The mathematical representation is the same as that of ARIMA (Equation 2.8, however the L, θ, ϕ, c, L, ϵ are matrices instead i.e. simultaneous equations). This model was used by Montgomery *et al.* [1998] to model unemployment trends in the United States of America (USA) and found that the model had a higher accuracy when compared to ARIMA.

Cook and Hall [2017], however, noted that VARs have a key limitation in that if one wants to model complex data or nonlinearities, it would be dependent on the researcher being able to manipulate the data to suit the model which requires the researcher to have deep knowledge of the underlying data structure. The VAR moving average (VARMA) model, a generalised version of VAR, is an alternative to VAR that is also employed in unemployment forecasting [Pelaez 2006]. However, this model does not solve the challenges associated with VARs.

Yang [2007] attempted to overcome these challenges by introducing the seasonal additive nonlinear VAR model (SANVAR) and successfully demonstrated that it could model unemployment rates in the USA with a higher accuracy in most cases than univariate models. The model was able to draw insights around how people of different races and genders are affected by different unemployment drivers. The researchers demonstrated that African Americans's unemployment rate movement require more variables to explain when compared to White Americas. The model, however, has drawbacks in that it requires extensive set up and expert knowledge from the researcher to successfully deploy.

Threshold Autoregressive Models (TAR)

Although ARIMA and VAR have proven successful in the past, these models, along with other univariate models are unable to accurately forecast asymmetric variables. Unemployment data is typically asymmetric in that it sometimes declines when it is expected to increase, such as during South Africa's "jobless growth era", depicted in Figure 2.3 [Vermeulen 2017; Marinkov and Geldenhuys 2007]. Therefore, nonlinear models are preferred over them for such variables. TAR is such a model, that has been shown to be adequate in addressing the asymmetric nature of unemployment [Montgomery *et al.* 1998].

TAR is able to accommodate asymmetries because the model allows for regime shifting [Montgomery *et al.* 1998; Fok *et al.* 2005; Fransesa *et al.* 2004]. Therefore, as the movement in unemployment rates change due to structural changes these models can accommodate the changes. TAR can cater for multiple regime shifts through its modified versions: Smooth TAR (STAR) and Self-Exciting TAR (SETAR) [Fok *et al.* 2005; Pelaez 2006].

Fransesa *et al.* [2004] used Root Mean Squared Error ($RMSE$) and Mean Absolute Percentage Error ($MAPE$) as performance measures in an experiment to forecast unemployment using two regime shifting models: TAR and Markov Switching Models. These models had lower error rates than the AR model in three of the four countries in the experiment. However, in Germany the AR models outperformed the regime shifting models. This is consistent with Clements *et al.* [2004]; De Gooijer and Hyndman [2006] who state that nonlinear models outside of neural networks, are still in their infancy and do not always produce better forecasts.

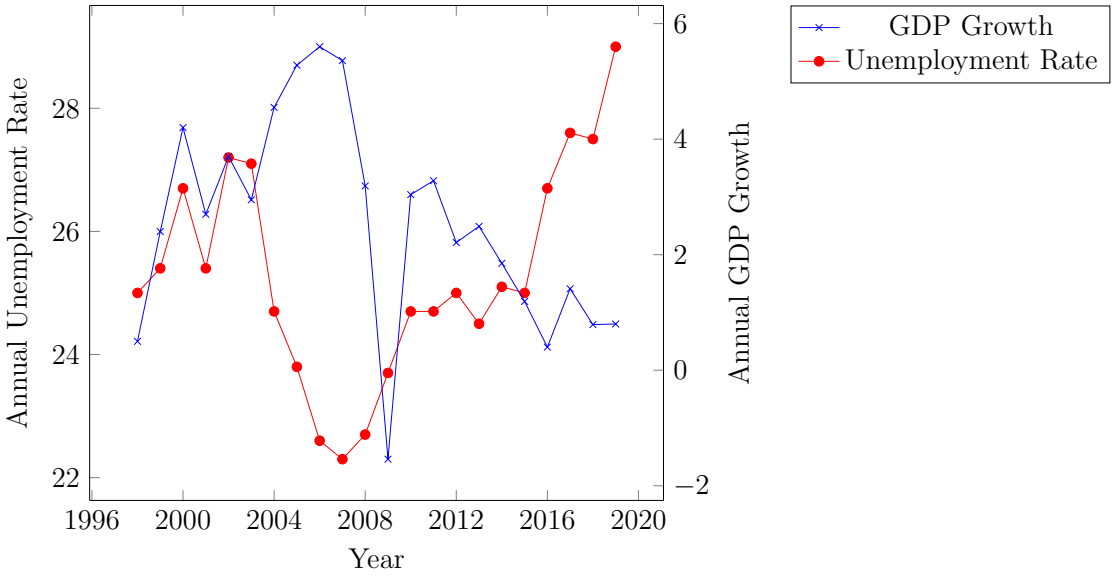


Figure 2.3: South Africa's asymmetric unemployment rate. Between 2003 and 2006 unemployment decreased as GDP increased. However, between 1998 and 2001 unemployment increased as GDP increased. [Statistics South Africa 2019; TradingEconomics 2020]

.

Fractional ARIMA (FARIMA) / Generalised Autoregressive conditionally heteroscedastic (GARCH)

Fractional ARIMA models are a generalised version of ARIMA that allows for long term memory to be incorporated into the model. According to Katris [2019ab]; Brooks [2014], FARIMA / GARCH extend the FARIMA and enable the modelling of data with differing variances (heteroskedastic time series). This model is preferred in modelling time series whose variance is not constant across its life span. The model can forecast unemployment even when its movements are affected by market shocks such as recessions with greater accuracy than ARIMA, SARIMA and FARIMA [Katris 2019ab].

Ewing *et al.* [2005] used the GARCH and the threshold ARCH model (TARCH) to model how different racial groups are impacted by changes in the national unemployment rates. Their research showed that different demographic groups, by gender and race, experience different unemployment rates. Furthermore demonstrating how each demographic group responds differently to economic shocks. White males experience the lowest unemployment rates and they are least impacted by economic shocks. Black males, on other hand, are vulnerable to economic shocks and have an unemployment rate that is twice as high as white males.

Holt-Winters

Hyndman and Athanasopoulos [2018] state that Holt is a model that takes trends into account in how it forecasts. Holt-Winters is an extension of Holt, allowing for seasonality to be captured in models. The model enables trends and seasons to be modelled, by using exponential smoothing to represent and capture past data. Katris [2019a] states that Holt-Winters and ARIMA are standard benchmarks when modelling time series data. Showing that the model in some instances is more accurate than FARIMA/GARCH when modelling unemployment rates.

2.2.3 Professional Forecasts

Across the world, there are several organisations whose role is to forecast macroeconomic indicators, unemployment being one of them. These organisations can be thought of broadly as professional forecasters, corporate forecasters, and international institutions. These organisations will typically release economic analysis and economic outlooks on a monthly or quarterly or annual basis with their forecasts often a month, a quarter or a few years ahead. These organisations typically look beyond the data and look at country-level risk such as political and social risks in making their forecasts. This means they can leverage both quantitative and qualitative data in their forecasting [Makridakis 1988]. However, the selection of variable they include and exclude leads many of these forecasts to have high error rates: higher than traditional statistical model [Makridakis 1988].

Professional Forecasters

In South Africa, the Bureau for Economic Research (BER) is seen as a leading institute of economic research and macroeconomic forecasts. They produce research and forecasts which cover the major macroeconomic indicators including the unemployment rate. Their forecasts are typically one to three-year horizons, they have currently forecasted that South Africa's unemployment rate will reach 32% by the year 2024.

Corporate Forecasters

Financial institutions, particularly banks and insurance firms, typically employ economists who forecast key macroeconomic indicators for them. Almost all the banks in South Africa have this function [Investec 2020; Nedbank 2020]. They use data from the BER, Statistics South Africa (StatsSA) and the South African Reserve Bank (SARB) to produce their forecasts. Some of these banks even have their own macroeconomic indices such as ABSA/BER's Purchasing Managers' Index, RMB/BER's Business Confidence Index and FNB/BER's Consumer Confidence Index.

International Development Institutions

At a global level, there are several highly influential international institutions whose forecasts, often inform business decisions across the world. These are the International Monetary Fund (IMF), the World Bank (WB), and the OECD:

- IMF [2020] forecasts every major economic indicator across the world. Their forecasts are relied upon by both governments and business leaders as key to determining the direction the world is going. However, an analysis conducted by McIntyre and Cedric [2019] found that IMF's forecasts are wrong most of the time, stating "In 6,1 percent of cases, the IMF was within a 0,1 percentage-point margin of error. The rest of the time, its forecasts underestimated GDP growth in 56 percent of cases and overestimated it in 44 percent". The IMF release their forecasts quarterly and annually some of which are for five-year horizons. Their current forecast is that South Africa's unemployment rate will continue increasing and reaching over 30% by 2024.

- WorldBank [2020b]'s forecasts are typically focused on economic growth but they have a wealth of data that covers all major economic indicators and as such they forecast them as well. Part of the World Bank's strategy is developing indices which then informs particular behaviours from countries, one such index is the Ease of Doing business [WorldBank 2020a]. Their outlook is that South Africa will experience growth of $\sim$1% on average for the next three years. Given that South African's unemployment rate responds minimally to growth, we can expect a minimal decline in the unemployment rate as well.

- OECD's forecasting methodology is a combination of opinion from experts, economic models (such as Okun's Law, Philips Curve, and so forth), policy analysis, and other statistical models [OECD 2019]. Their forecast horizon is a year and they expect South Africa's unemployment rate to continue increasing through 2020.

2.2.4 Leading Indicator of Employment in South Africa

In 2009, Davies *et al.* [2009] worked with the SARB to attempt to develop a leading indicator of employment in South Africa. This work was motivated by the existence of a business cycle leading indicator by the SARB. They found that although output (GDP) and employment do not seem to move in the same direction, by smoothing both the GDP and employment data, these variables appear to be moving in the same direction. Leading to a conclusion that business cycle indicators should be coincident indicators of employment, stating that "...leading indicators of the business cycle in South Africa provide the candidates for leading indicators of the employment cycle in South Africa".

Davies *et al.* [2009] tested their model's forecasting ability by arbitrarily selecting some leading business cycle indicators, namely, Continuous Commodity Price Index, Retail Sales, Manufacturing Employment and Trading-partner countries.

In Davies *et al.* [2009]'s model they leveraged the following equations:

$$Y_t^j = X_t^j - X_{t-1}^j, \tag{2.9}$$

$$S_t^j = \frac{Y_t^j}{\frac{1}{T-1}\sum_{t=1}^{T} \mid Y_t^j \mid}, \tag{2.10}$$

$$I_t = \sum_{j=1}^{J} w^j S_t^j, \ and \tag{2.11}$$

$$I_t^s = \frac{I_t}{\frac{\sum_{t=1}^{T} I_t}{\sum_{t=1}^{T} E_t}}. \tag{2.12}$$

where Equation 2.9 is a variable's difference over each quarter, X_t^j is variable j's value at time t. These differences are then standardized by diving them by their historical averages, as shown in Equation 2.10. A linear combination of the standardized variables is then computed to create a composite variable, I_t (Equation 2.11), with "the weights w_j determined by the concordance of each series with the reference series. The concordance is established by taking the maximum number of quarters over four lags during which the component moves in the same direction as the reference series". These composite variables are then standardized by Equation 2.12, where E_t is the quarterly change in employment. Therefore this I_t^s is the Leading Indicator of Employment in South Africa (LIESA) index.

LIESA was able to show the direction in which employment would move beforehand, although, there was a fair amount of error when employment changes were due to structural issues. The research intended to guide more researchers to follow in the same direction and conduct more indepth experiments. Davies *et al.* [2009]'s method was based on the New Zealand version of the study that preceded it by [Claus 2007]. Claus [2007] created six leading indicators for unemployment in New Zealand using six different approaches with a variation of the same 95 features. They found that their model was able to provide early warning indicators for New Zealand with models that used 'exogenously determined weights' being the most successful.

Aiken [1996] however, warns against the reliance on economic indices as he found that they are not as accurate when business cycles change or when there are economic turns. Indices also assume the source data structure will remain the same over the years, however, economic data is always being reviewed and revised continually. Adding that, composite indicators are not able to capture complex relations amongst the variables. Therefore, where there are structural changes in an economy indices are prone to errors.

2.2.5 Known Challenges with Traditional Forecasting Models

This section provided a view of the traditional approaches to forecasting which have been in use for decades. These models have several limitations which the machine learning approaches, discussed in chapter 3, attempt to overcome. The key challenges being:

- Hall [2018]; Aiken [1996] state that the models are best catered to forecast linear data and often require a significant amount of tweaking to enable them to model nonlinear data. The tweaks themselves are complex as they require the researcher to have intimate knowledge of the data itself.
- The data is often required to meet certain data quality requirements such as white noise (randomly generated uncorrelated variables with zero mean and constant variance) and stationarity [Aiken 1996; Brooks 2014; Katris 2019a; Olmedo 2014].
- The models assume that the data is symmetric and are prone to error when the data is asymmetric as unemployment data is, although multivariate approaches try to address this [Marinkov and Geldenhuys 2007].
- The highly relied upon forecasts from professional forecasters, who use univariate and multivariate models amongst other approaches as well, have proven to be extremely inaccurate [McIntyre and Cedric 2019].
- As there are hundreds of thousands of possible variables to consider in forecasting unemployment, the limitation in the method and tools demand that researcher select particular variables which introduced selection-based biases [Katris 2019b; Davies *et al.* 2009; Liang 2005].
- Unbiased selection of variables and data that will enable a model to produce highly accurate results are not available in traditional approaches as each model demands both data and variable selection before modelling [Katris 2019ba].
- The models are prone to error when there are economic shocks which cause the observed variable to behave in a manner which is not synonymous with more recent observations [Hall 2018; Aiken 1996; Olmedo 2014].

This section provided a review of traditional methods that are used for forecasting unemployment across the world. The next chapter, chapter 3, provides a literature review of the machine learning approaches that have thus far been employed to forecast unemployment. These models address the limitations of traditional unemployment forecasting model and offer an improvement in forecasting accuracy.

Chapter 3

Related Work

In South Africa, there is a limited body of literature on unemployment forecasting techniques [Burger and Fourie 2015]. Even more so is literature relating to South African machine learning approaches toward the unemployment problem. These techniques are mostly applied in North America, Europe, and Asia. The use of these machine learning techniques to forecast unemployment has been of interest to researchers since the 1990s.

Aiken [1996] was one of the first to demonstrate how neural networks could be used to forecast unemployment. His research was an attempt to address issues associated with the accuracy and reliability of traditional macroeconomic approaches to forecasting unemployment. He noted that the biggest problem is that traditional forecasting techniques require certain conditions to be met by the data before being applied: constant variance, constant mean, and no correlations amongst the independent variables.

Both univariate and multivariate models have extensive data requirements that need to be met before being effectively used. However, this is not the case with machine learning approaches as these techniques learn the underlying pattern in the data that map features to the target variable.

This section will provide an overview of machine learning approaches that have been used to forecast unemployment. The section will start with regression techniques, followed by neural networks and ending with nearest neighbour and ensemble techniques. The chapter will end with the literature gap that this research intends to fill.

3.1 Forecasting Unemployment

3.1.1 Regression

Hyndman and Athanasopoulos [2018] state that regression is an approach for estimating the relationship between one or more variables. Where the variables are either dependent and

independent. The models inherently assume that there is a relationship between dependent and independent variables. As such, they are models that try to estimate the influence of dependent variables on the independent variables. This section will discuss the common regression techniques used to model unemployment rates.

Elastic Net (EN)

Hall [2018] states that core to machine learning algorithms is minimizing errors relating to bias and variance, which coincidentally have a trade-off relationship. The Elastic Net (EN) model was designed to specifically minimize both these errors through regularization. The model has penalties for overfitting by penalizing the model if it is overly complex (too many variables) or has over-reliance on particular variables. Through this process, the EN learns which features are most important in the data without a need for the researcher to make assumptions as is required in the traditional forecasting approaches.

Hall [2018] demonstrated that the EN can forecast unemployment over 3, 6, 9, 12, and 24-month horizons more accurately than traditional forecasting models and professional forecasters which were used as baseline models. This model was able to identify unemployment shifts over recessions and booms more accurately than the baseline models. Through the regularization process of EN, the model was able to identify key variables that predict unemployment by setting all other coefficients to zero. In Hall [2018], the most important features identified by the EN model for the United States of America (USA) were: housing, manufacturing, and interest rates.

Least Absolute Shrinkage Selector Operator(LASSO) Regression

Kreiner and Duca [2019] used LASSO regression to forecast unemployment rates in the USA. This regression model minimizes prediction error through a regularization process. The model penalizes the reliance on particular variables for prediction, therefore, reducing overfitting. Kreiner and Duca [2019] demonstrated that LASSO can improve the forecasting accuracy of unemployment in the USA as well as identifying variables which are most important in forecasting unemployment. Interestingly, the model suggested that international data such as German job vacancies and Australian treasury rates were amongst the ten most important features in predicting the USA unemployment rates. The meaningfulness of these features is questionable because there is little the USA can do about them but it provides a starting point in thinking about how to discover features that drive unemployment.

Support Vector Regression (SVR)

Support Vector Machines (SVMs) are an approach to supervised learning that enables nonlinear data to be classified using a hyperplane [Goodfellow *et al.* 2016]. The approach outputs a class identity, categorising the inputs into varying classes. These models are very useful for modelling nonlinear data. There are two types of SVMs: Support Vector Classifiers (SVC)

and Support Vector Regression (SVR). To use SVMs for regression problems an implementation of the ϵ-sensitive loss function is required which enables nonlinear regression modelling [Sermpinis *et al.* 2014]. The loss function is shown in Equation 3.1 where y_i is the correct output, $f(x_i)$ is the predicted output, and ε is the error. Cook and Hall [2017]; Katris [2019b] demonstrated that the usage of nonlinear decision boundaries in SVRs makes them effective in unemployment forecasting tasks with an accuracy that is above linear regression.

$$L_\varepsilon(x_i) = \begin{cases} 0, & \text{if } \mid y_i - f(x_i) \mid \leq \varepsilon \ , \ \varepsilon \geq 0 \\ \mid y_i - f(x_i) \mid - \varepsilon, & \text{if other} \end{cases} \tag{3.1}$$

Multivariate Adaptive Regression Splines (MARS)

Multivariate Adaptive Regression Splines (MARS) is an extension of linear regression models, which enable the modelling of nonlinearities. The model is non-parametric and assumes no particular relationships amongst the independent variables and dependent variable(s) [Katris 2019b]. Katris [2019b] demonstrated that MARS models have a higher forecasting accuracy than SVRs although their performance was still lower than neural networks.

Diana *et al.* [2014] also used a multivariate regression model to forecast unemployment in Indonesia. They found similar results as Katris [2019a]: improvement in forecasting accuracy. Diana *et al.* [2014]'s implementation was, however, semi-parametric and it used Bayesian techniques to determine the forecasting confidence intervals. The researchers state that semi-parametric multivariable regression models offer an improvement over standard non-parametric MARS approaches by removing assumptions on how the regression function should look. The data is used to derive the regression function.

3.1.2 Neural Networks

Neural networks offer a great advancement in time series forecasting because these techniques do not require particular data assumptions to be met before being used and they can function with incomplete or imperfect data, unlike other models that present issues when data completeness is a requirement [Aiken 1996]. Therefore, neural networks are a superior choice when forecasting nonlinear time series data. Aiken [1996] was the first to demonstrate that neural networks can produce significantly better unemployment forecast estimates compared to traditional forecasting models. Table A.1, in Appendix A, provides more information on [Aiken 1996]'s results along with other machine learning models discussed in this section. Since Aiken [1996], several other researchers have investigated various neural network architectures for forecasting unemployment, these are discussed in this sub-section.

Feed Forward Neural Networks (FFNN)

Feed Forward Neural Networks (FFNN) are neural networks where the data flows from input to output in a 'forward' direction without moving backwards. Their goal is to find a predictor function, f, which maps an input x to some output y i.e. $y = f(x)$ [Goodfellow *et al.* 2016]. FFNNs are foundations to more advanced neural networks and they are referred to by multiple names: Multi-layer Perceptron (MLP) or Deep Feed Forward Networks. FFNNs enable nonlinear relationships to be effectively modelled, which is important for unemployment data.

Cook and Hall [2017] investigated the use of fully connected FFNNs to predict unemployment rates in the USA and they found them to perform better than Surveys of Professional Forecasters. Aiken [1996]; Kreiner and Duca [2019]; Pelaez [2006] were able to demonstrate that FFNNs offer a significant improvement in forecasting results when compared to traditional unemployment forecasting techniques. Aiken [1996]'s research inspired several similar investigations for other countries.

Katris [2019a] found that in Mediterranean countries, FFNN where sometimes outperformed by FARIMA / GARCH and Holt-Winters models on preprocessed data that showed evidence of nonlinearity, heteroskedasticity, and non-normality. The key reason given for the underperformance by the neural network was the simplicity of the FFNN structure: a single hidden layer model with 1 to 10 nodes and a sigmoid activation function. Cerqueira *et al.* [2019] also observed that large data sets are required for neural networks to outperform traditional statistical methods. Therefore, it is clear that neural networks need sufficient data and an appropriate architecture to perform effectively.

Recurrent Neural Networks (RNN)

Recurrent Neural Networks (RNN) are a type of neural network architectures that enable effective modelling of sequences of data as they possess an internal memory structure [Goodfellow *et al.* 2016]. At their core they are essentially FFNNs with feedback loops [Goodfellow *et al.* 2016]. RNNs were developed for natural language processing requirements such as language translation. Cook and Hall [2017] demonstrated that these models can forecast unemployment accurately by using a variation of RNNs refereed to as long short term memory (LSTM) networks. The model outperformed the ARIMA model, using only past unemployment rates in the USA as the input to the model.

Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNNs) are a class of neural networks that are used primarily for image classification [Goodfellow *et al.* 2016]. These networks employ a mathematical procedure called convolution. Through convolution, different filters and aspects of an image can be detected such as edges. Although CNNs are not typically used for time series data, they were used to predict unemployment rates and were found to perform poorly when compared to other neural network architectures but with higher accuracy rates than ARIMA

[Cook and Hall 2017]. The results are consistent with the intention of the architecture which is to extract features from images and not to perform regression type activities.

Encoder Decoder (ED)

Encoder Decoder (EDs) networks are a specialised form of LSTM networks, that were designed for predicting words that where likely to be used within a particular sentence structure [Cook and Hall 2017]. These models have the ability to 'remember' things over a long period of time which makes them ideal to predict data over long horizons. Cook and Hall [2017] found the models to perform better than all other neural network models that were used as benchmarks (FFNN, CNN, and RNN). Part of the reason for this superior performance is that EDs are a combination of two LSTM models that have been tweaked to extend the long-term memory capacity.

Psi Sigma Neural Network (PNN)

Psi Sigma Neural Networks (PNN) are a specialised kind of FFNNs where the way in which weights are updated is modified to instead use a partial update methodology. Sharma and Singh [2016] used the PNN with six inputs, five hidden layers and one output, where the weight of the last layer was fixed to 1 and was never updated throughout the training process. The minimised function used was,

$$E(c, w_j) = \frac{1}{N} \sum_{n=1} N(y_n - \hat{y_n}(w_k, c)^2) \tag{3.2}$$

where y_n is the actual value, $\hat{y_n}$ is the estimated value, w_k are the adjustable weights, and c is the adjustable term. The adjustable term enables the model to provide better estimates: improved accuracy and fitting. The term can be compared to the traditional sigmoid functions in normal neural networks. The model was used to forecast unemployment rates and was compared to FFNN and RNNs. The results showed that the PNN had a higher error rate which was five times higher that FFNNs and RNNs.

Radial Basis Function Neural Network (RBFNN)

A Radial Basis Function (RBF) is a function that is determined by the distance between the input value and a predetermined fixed point. According to Sharma and Singh [2016], RBFNN are neural networks that use the RBF as their activation function as opposed to the traditional sigmoids used in other neural networks. The output of these models is a linear combination of RBFs. This model, when compared to RNN and FFNN, had a mean average error that was four times higher [Sharma and Singh 2016].

Learning Vector Quantization (LVQ)

Sharma and Singh [2016] found that although the mean average error of RBFNN was five times higher than RNNs and FFNNs when these models are combined with Learning Vector

Quantization (LVQ) the error was reduced by close to 50%. Mathworks [2020] states that LVQs are classification neural networks that use unsupervised learning techniques to extract features which are then classified into individual classes. Adding that the unsupervised technique leverages competitive learning, where all neurons compete to be activated resulting in only a single one being activated at each layer. Hence the approach is sometimes called a 'winner takes all' approach [Sharma and Singh 2016].

Adaptive Neural Fuzzy Inference System (ANFIS)

Atsalakis *et al.* [2007] define a Neural Fuzzy System as "a combination of Artificial Neural Networks (ANN) and Fuzzy Inference System (FIS) in such a way that neural network learning algorithms are used to determine the parameters of FIS". One such combined system is the Adaptive Neural Fuzzy Inference System (ANFIS). Where a FIS is a system that uses fuzzy set theory to map inputs to outputs, it uses 'IF...THEN..' rules as well as logical operators AND and OR to achieve its mapping. Atsalakis *et al.* [2007] investigated the use of ANFIS to forecast unemployment in Greece and these models performed better than ARIMA models. Therefore, demonstrating that these models are legitimate alternative for forecasting time series data.

Yolcu and Bas [2016] used several fuzzy logic techniques to forecast unemployment rates in Turkey: one of which is the ANFIS (discussed in Appendix A along with their other models). Their results are similar to that of Atsalakis *et al.* [2007], in that their fuzzy time series forecasts had low error rates: affirming the suitability of these techniques to forecast unemployment rates. Yolcu and Bas [2016] therefore, demonstrate that the fuzzy logic approaches are appropriate for forecasting unemployment rates. Adding that these techniques offer advantages over others because they are able to perform optimally with small data sets.

3.1.3 Nearest Neighbour

Yang [2007] used a technique similar to K-nearest neighbour to predict unemployment rates in the USA. Their model is not named, however, it is a nonlinear and non-parametric nearest neighbour method. The method works by splitting the unemployment time series into a subset referred to as the time series' history and a subset referred to as the time series' future. This is similar to the train and test data splits that are common in machine learning. The algorithm then estimates the future from the history, by lopping through the historical values and calculating their distance from the future value. A subset of n historical values that are closest to the future values are selected. If the n values form a simplex - a triangular shape - that contains the future value, then the n values can be used to forecast future values. The forecasted future values are predicted using the weighted average of the nearest neighbours.

Yang [2007]'s nearest neighbour technique was able to forecast unemployment rates more accurately (as measured by MSE) than ARIMA, TAR, BVAR, and professional forecast-

ers for multiple horizons: 3,4,5.., and 10 months ahead. The ARIMA was, however, more accurate in one and two months ahead forecasts. These results show that Yang [2007]'s technique is can cater for uncertainly that longer horizon timelines present than traditional statistical methods. The technique, however, did not perform as well when it was predicting unemployment using quarterly unemployment rates as compared to monthly ones.

3.1.4 Ensemble Learning

Cook and Hall [2017] noted that when comparing forecasting models to Surveys of Professional Forecasters (SPF) it is important to realise that their forecast is in an average forecast of a number of different forecasters who use different methods and approaches. On that basis, Cook and Hall [2017] demonstrated that training four different neural networks independently, these could then be combined to form an Ensemble model (EM) which combines predictions from the individual architectures to a new improved model. The combined model was able to out perform each individual model significantly, therefore, showing that such approaches can be leveraged to improve accuracy of unemployment rate forecasts. Makridakis *et al.* [2018] also noted that ensemble techniques have grown in popularity and accuracy over the past few years and are now the best performing forecasting techniques.

3.2 Literature Gap

Chapter 3 showed that several economic and statistical forecasting models have been applied to unemployment time series data. These models, however, were limited when modelling asymmetric and nonlinear data as well as data that was not preprocessed to meet a particular standard such as white noise data [Marinkov and Geldenhuys 2007; Mahipan *et al.* 2013; Katris 2019a]. Furthermore, the models performed dismally when there were economic shocks due to recessions and booms that make unemployment rates behave inconsistently with past observations, this is largely because these models do not have long term memory [Olmedo 2014; Makridakis 1988].

To address these challenges, machine learning models, mostly regression models and neural networks, were employed to unemployment forecasting problems. Their performance has proved that they are more successful than traditional approaches in forecasting unemployment rates. Most of the current literature on the topic compares the forecasting performance, as measured by Root Mean Squared Error ($RMSE$) or Mean Average Error (MAE), of machine learning models with traditional statistical models.

Few researchers provide insight to the key features driving unemployment and with the researchers who did, key insights were uncovered such as how German job vacancies and Australian treasury rates affected the USA unemployment rates [Kreiner and Duca 2019]. Which is an odd finding as there is little action that the USA can take to mitigate this finding. Furthermore, the machine learning models implementation were often simple variations

of a particular architectures [Katris 2019a; Cook and Hall 2017; Hall 2018]. For example, most of the neural network implementations only had a one to three hidden layers and in most cases the input layer was made of past unemployment rates only. Katris [2019a] observed that in some cases simple architectures could be outperformed by more sophisticated traditional statistical models such as FARIMA / GARCH.

It is also worth noting that the vast majority of the literature on modelling unemployment rates using machine learning was in North America, followed by Europe and Asia. There were few to no similar models found for the South African context. In both the United States and Europe, key economic models such as Okun's Law and Philips Curve hold strongly resulting in GDP and inflation being standard features when forecasting unemployment [Aiken 1996; Jelena *et al.* 2017]. However, Ball *et al.* [2019] observed that in developing nations and particularly South Africa, these models are often weak (low R^2) in explaining unemployment. Meaning that South Africa is well suited for machine learning approaches as traditional economic models do not always provide the required explanations.

Machine learning models are often preferable if the model is to be inferred from the data, as is needed in South Africa. In the country, the frequency of unemployment forecasting is currently quarterly, whilst key macroeconomic indicators are monthly and weekly. This is unlike other countries where machine learning models have been applied to forecast unemployment, where unemployment rates are reported in similar frequencies as other macroeconomic indicators. Therefore, this is a motivation for the importance of South Africa having local versions of machine learning-based unemployment forecasts. It is an opportunity to use these models in forecasting contexts with varying data frequencies.

Therefore, this research will close the gaps that have been identified through the literature view by:

1. Applying machine learning models to forecast unemployment in South Africa. Thus far machine learning models have not been applied to forecast unemployment in South Africa. In other nations where they are applied, key economic models, Okun's Law and Philips Curve, which are used to predict unemployment's directional movement are strong predictors of unemployment which is not the case in South Africa [Marinkov and Geldenhuys 2007; Ball *et al.* 2019; Vermeulen 2017].

2. Predicting how various factors affect different unemployment classes. The body of literature reviewed generally focused on aggregate unemployment rates, whereas in reality the concern is often for various classes of unemployment rates such as youth unemployment, graduate unemployment, female unemployment, and so forth [Yang 2007; Ewing *et al.* 2005].

3. Using feature selection techniques such as information gain and regression to determine which features predict unemployment rates. Very few researchers thus far focused on

determining features that have the biggest impact on the magnitude and direction unemployment rates [Hall 2018; Kreiner and Duca 2019; Katris 2019b; Yang 2007].

4. Use probabilistic models to discover from the data how various features of unemployment impact each other. There is currently limited literature that attempts to use machine learning to determine a data-driven causality model for unemployment changes, even though there are a number literature contributions to on how to infer causality from time series data [Dahlhaus and Eichler 2003; Ullah *et al.* 2017].

Therefore, this research will apply machine learning techniques to forecasting unemployment in South Africa, determine which key features are the most important drivers of unemployment in South Africa, and attempt to show how these features relate to one another.

This chapter provided a literature review of the machine learning approaches that have thus far been employed to forecast unemployment across the world. Appendix A provides a table that compares the performance and architecture of the machine learning models discussed. In the next chapter, chapter 4, the methodology that will be employed in this research is discussed as well as the data to be used and some preliminary results for the South African context.

Chapter 4

Methodology

This section discusses the research methodology to be employed in carrying out this research. In section 4.1 the research aims and objectives is provided. In section 4.2 the motivation for this research is discussed. In section 4.3 the limitation of the research is provided. Section 4.4 provides an overview of the machine learning models that will be used in this research. Section 4.5 provides an overview of the data sources that will be used for this research. Section 4.7 provides preliminary analysis that was conducted with univariate traditional unemployment forecasting models. The section ends with section 4.8, which discusses ethical consideration that are important for this research.

4.1 Research Aims and Objectives

Unemployment is one of the most important macroeconomic indicators as it is used to signal the economic health of a particular country [Aiken 1996; Pelaez 2006]. In South Africa, the unemployment rate has been trending upwards for the past 12 years and there is no indication that it will stop doing so [Statistics South Africa 2019]. Numerous policies have been explored in an attempt to address this issue but none have yielded the necessary fruit [Levinsohn 2007; Brynard 2011; De Lannoy *et al.* 2018]. Therefore, accurately predicting the unemployment rate is an extremely important task. This forecasting is a key input in the process of policy making [Brynard 2011].

This research aims to contribute to solving unemployment in South Africa by achieving the following objectives:

1. Use machine learning (i.e. neural networks, regression, and Bayesian model) algorithms to predict unemployment rates over multiple classes (i.e. youth, females and geographic) and various time horizons (i.e. 1, 2, 4, and 8 quarters ahead).

2. Use information gain and regression (i.e. Elastic Net) techniques to determine which features contribute significantly in predicting unemployment rates.

3. Use probabilistic techniques to discover how various features influence one another.

4. Advice policymakers on opportunities to reduce unemployment based on the findings of this research.

5. Develop a website that can be publicly available post the research for policymakers to use as a reference.

4.2 Significance and Motivation

It is no longer a surprising fact when StatsSA announces the increase in unemployment rates in South Africa. This has been a constant theme for 12 years. Additionally, the advent of the so-called 'Fourth Industrial Revolution' is estimated to lead to higher unemployment rates in the immediate future. Therefore, this means that South Africa is a social unrest ticking time bomb because there is currently no obvious solution to the current unemployment crisis and there is already evidence of frustration amongst the unemployed [Meyer 2014].

Across the world, COVID-19, a highly contagious respiratory disease, has been center focus as the virus has brought countries to a stand still. In South Africa, a national lock down was instituted on 26 March 2020 to 30 April 2020. This period alone is estimated to result in 1,5 million job losses, increasing South Africa's unemployment rate above the 30% mark [Mkhabela 2020]. The impact of COVID-19 is likely to be felt for months if not years in South Africa and the world.

It is therefore expediently required that unemployment be studied to accurately determine features that influence it. As these are key inputs into the business of policymaking. Machine Learning techniques are capable of performing this task with significantly higher forecasting accuracy than traditional statistical modelling approaches. As well as capturing regime shifts such as the ones being caused by COVID-19. These models are applied to a limited extent in South Africa, therefore, this research will contribute significantly to the current body of literature on unemployment forecasting using machine learning.

4.3 Limitations and Assumptions

The research will enable greater accuracy in forecasting South Africa's unemployment rate, along with an indication of which features contribute most meaningfully to the problem. The output of this research will be useful for policymakers, however, South African policy making is not often data-driven and thus policymakers might ignore this output.

The intention of the research is to improve prediction accuracy for unemployment rates, identify key drivers of unemployment, and make policy recommendations to reduce unemployment. In policy settings interpretability of models is often very important, however, the most accurate models might not be interpretable. This is a trade-off that will have to be

made as the purpose of this research is not to build explainable models but rather accurate models.

4.4 Research Design

The study is a mixture of exploration and confirmation. The confirmatory part of it will be using known machine learning techniques that have been used to forecast unemployment and applying those to South Africa. New techniques and approaches that have yet to applied will be explored to determine their suitability to predict unemployment in South Africa. Through this exploration, a data-discovered hypothesis will be uncovered that shows which features are most meaningful in predicting unemployment.

Chapter 2 and chapter 3 showed that unemployment rates tends to be nonlinear and asymmetric. Therefore, the following will be models will be used in this research as they have demonstrated success in forecasting unemployment rates in other regions. Some are selected because they aid in feature selection:

- Artificial Neural Networks - Neural networks have demonstrated significantly higher accuracy rates in forecasting unemployment than traditional statistical approaches. This will be explored on South Africa's unemployment rate: with and without feature reduction for performance comparison. Sharma and Singh [2016]; Cook and Hall [2017]; Aiken [1996] used ARIMA and VAR as benchmarks for their feed forward, convolutional, and recurrent neural networks. Recurrent neural networks (RNNs) have been demonstrated as able to produce the most accurate unemployment forecasts. However, these have thus far been tested primarily on univariate data and this research will apply them to forecasting unemployment with multiple variables.

- Bayesian Models - although there is an increase in machine learning models being applied to unemployment forecasting, there is however, limited to no quantification of uncertainty in these models. Bayesian learning makes provision for such and therefore these will be considered along with neural network i.e. Bayesian Neural Networks (BNNs). Although Bayesian approaches have been applied with VAR and regression techniques for forecasting unemployment, BNNs have not been applied to in this context [Yang 2007]. BNNs have demonstrated that they are able to better capture nonlinear time series data when compared to traditional statistical methods and other neural networks [Liang 2005]. Therefore, these will be leveraged to forecast unemployment rates in South Africa.

- Information Gain - Currently, Genetic Algorithms (GA), LASSO, MARS, and Elastic Net are the three approaches that have been applied for feature selection in unemployment forecasting research [Sermpinis *et al.* 2014; Hall 2018; Cook and Hall 2017]. Therefore, GAs, LASSO, MARS, and Elastic Net will be applied to this research along with information gain techniques to compare which provide more accurate and predictive features.

- Regime Shifting Models - business cycles have a great impact on unemployment and multivariate models such as TAR have been used to cater for such changes [Montgomery *et al.* 1998; Fok *et al.* 2005; Makridakis 1988; Taylor *et al.* 2016]. Hidden Markov Models (HMMs) will be explored as these are able to model regime shifts [Fransesa *et al.* 2004]. The HMMs models will be benchmarked against the TAR models that have thus far been used to forecast unemployment rates in other regions.

- Ensemble Learning - hybrid or combination models have proven to be more successful than single models in several instances [Makridakis *et al.* 2018 2020]. Cook and Hall [2017]; Kouziokas [2019]; Sharma and Singh [2016] have demonstrated that combined neural networks can improve the accuracy in predicting unemployment rates. Therefore, because of their demonstrated success these models will also be considered for this research.

It was noted in chapter 2 and chapter 3 that to model unemployment, its asymmetric, nonlinear, and structural nature should be taken into account. Therefore, models that will considered in the exploratory phase will be those that are geared toward these data characteristics.

4.5 Data Sources

The data used in this research is all publicly available, regularly reported, and accessible without cost. In some instances a registration is required. Primarily, data from Statistics South Africa will be used as it contains the Quarterly Labour Force Survey which is South Africa's only database that tracks unemployment rates across the country. The South African Reserve Bank, and Bureau of Economic Research's macroeconomic databases track South Africa's macroeconomic indicators. The Investment Map and UNCTAD provide an open database of global trade and investment data, for which South Africa is covered.

From these databases, a 'general-to-specific' feature selection approach will be employed. Pelaez [2006] describes this method as a feature selection approach that starts with all possible features. Features are then iteratively removed based on how they contribute towards predicting unemployment. This is opposed to a 'specific-to-general', where one starts with a specific subset of features and adds as more are discovered. This method ensures parsimony but it is computationally expensive and time-consuming [Pelaez 2006]. It is not an ideal method for South Africa where the standard features in unemployment forecasting, GDP and inflation, do not strongly explain unemployment in the country [Vermeulen 2017; Marinkov and Geldenhuys 2007].

Table 4.1 shows the key databases for the data that will be utilized throughout the research. A significant amount of preprocessing will be required as the data is reported in different frequencies and the data is available for different intervals. The target variable, unemployment, is reported quarterly and annually by Statistic South Africa.

Table 4.1: Research data sources

Data Source	Description	Features	Preprocessing needed
Statistics South Africa	"Statistics South Africa is the national statistical service of South Africa, to produce timely, accurate, and official statistics to advance economic growth, development, and democracy" [Statistics South Africa 2020].	StatsSA has 49 time series variables that they track: vehicle sales, industry sales, industry price indices, GDP, and so forth. The data is typically reported monthly, except for GDP and unemployment rates which are reported quarterly.	Some data is available from 2002, whilst others only from 2008. There is also different reporting frequencies, monthly and quarterly.
South African Reserve Bank	"The South African Reserve Bank is the central bank of the Republic of South Africa. The primary purpose of the Bank is to achieve and maintain price stability in the interest of balanced and sustainable economic growth in South Africa. Together with other institutions, it also plays a pivotal role in ensuring financial stability" [South African Reserve Bank 2020].	There are 149 macroeconomic indicators such as: interest rates, exchange rates, gross domestic expenditure, production volumes, and other macroeconomic indicators. The data is available in quarterly and monthly frequencies from 1960 to date.	The data is available in different frequency formats: monthly and quarterly. Therefore, there is a requirement to align these dates. Furthermore, some of the data is only available for the last ten years, whilst other is available from 1960, so the missing data needs to be addressed.
United Nations Conference on Trade and Development	"The United Nations Conference on Trade and Development was established in 1964 as a permanent intergovernmental body. UNCTAD is the part of the United Nations Secretariat dealing with trade, investment, and development issues" [UNCTAD 2020].	The data is focused on trade, including: trade flows, population changes, commodity prices, and marine travel data.	The data is reported on annual basis. The data will need to be preprocessed in the context of the rest of the data sets.

Table 4.1 continued from previous page

Data Source	Description	Features	Preprocessing needed
Investment Map	"The Investment Map database collects yearly Foreign Direct Investment (FDI) statistics for about 200 countries and detailed FDI sectoral and/or country breakdown for about 115 countries" [InvestmentMap 2020].	The database has inward and outwards flows of FDI across the world.	The data is reported annually at an aggregate and sector level. The data will need to be preprocessed in the context of the rest of the data sets.

4.6 Instruments and Analysis

To accomplish the tasks of this research, the Python programming language will be used. Along with the Scikit-learn library which provides a rich toolbox of machine learning libraries, with the Keras library providing a rich API for neural networks and Statsmodels enabling easy access to statistical models. Therefore, these three libraries will be the core tools for this research with Tensorflow being used to improve the speed of development.

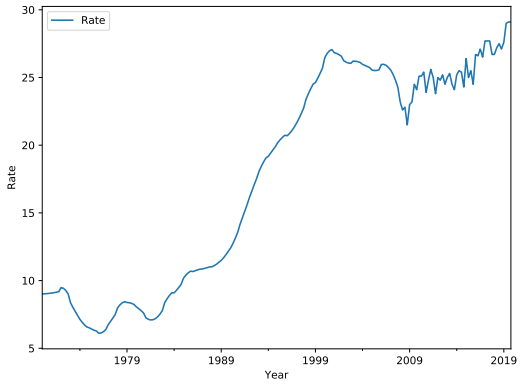
The research will involve experimentation and thus not all machine learning techniques to be used are yet fully known. However, regression, neural networks, ensemble learning, genetic algorithms, and Bayesian approaches will be considered as articulated in section 4.4. Previous researches used similar techniques in either R or Python programming languages. These languages have a wealth of existing libraries that are widely adopted and used in academia.

4.7 Analysis

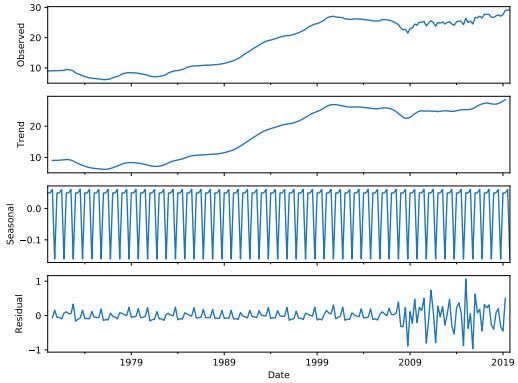
Some preliminary data analysis was conducted, where univariate traditional statistical time series models (discussed in section 2.2) were leveraged to build models to forecast South Africa's unemployment rate based on its past movements. This is a single series time series model, similar to what was used by several researchers to forecast unemployment [Cook and Hall 2017; Katris 2019a; Mahipan *et al.* 2013; Olmedo 2014]. The model's performance was measured based on their root mean squared error ($RMSE$): a common performance measure in time series analysis. This subsection will provide the descriptive statistics of unemployment as well as the results of the preliminary analysis.

Statistical Description of South Africa's Unemployment Rate
The unemployment data used for this preliminary analysis was accessed from the Bureau of Economic Research (BER). The time series had a total of 209 quarterly observations of the South African unemployment rate from January 1970 to January 2020. Figure 4.1 (a) depicts this data. It can be seen that South Africa's unemployment has an upward trend and high variance over the last 10 years. The average unemployment rate over the time period is 17.7%, the lowest was 6.1% in the 1970s, and the highest rate was 29.1% which is the current unemployment rate.

Time series data can be decomposed into three key parts: the trend, seasonality and residuals [Hyndman and Athanasopoulos 2018]. It was observed that South Africa's unemployment rate trends upwards and it is seasonal where repeated movements in the data are observed each year: depicted in Figure 4.1 (b). The change in variation is visible by studying the residuals, where it can be seen that between 2009 and 2019 there was greater variability in the data than all the previous periods.



(a) Unemployment rate from January 1970 to January 2020



(b) Seasonal Decomposition

Figure 4.1: South Africa's unemployment rate

From Figure 4.2, it is clear that South Africa's unemployment rate does not follow a normal distribution. The autoregressive nature of the univariate traditional statistical methods generally assumes that the data follows a normal distribution with zero mean and constant variance. Kolmogorof ($kstest$) and Lillietest ($lilliefors$) tests were also conducted, which are well-known tests for normality, and they both confirmed that South Africa's unemployment rate is not normally distributed. Figure 4.2 shows a kernel distribution function whose shape suggests that the unemployment rate is bi-modal.

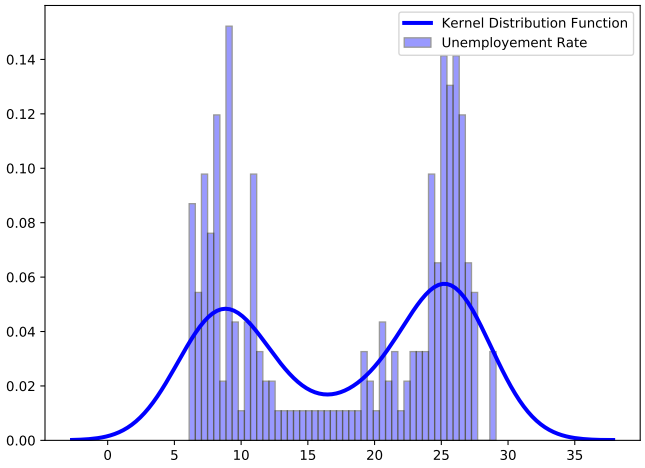


Figure 4.2: South Africa's unemployment rate probability distribution

It was also observed that the unemployment rate is also non-stationary over the observed period. Stationarity was tested through Augmented Dickey–Fuller (ADF) test, whose null-hypothesis is that the series is non-stationary. The $ADF\ test$ showed a $p-value$ of 0.923100 which means that we accept that null-hypothesis that South Africa's unemployment rate is non-stationary. This results differs from Burger and Fourie [2019] who found South Africa's unemployment rate to be stationary. However, Burger and Fourie [2019]'s data was only between a select period from 2008 to 2017, whilst these preliminary results are from 1970 to 2020. This differing results indicate that there are regime shifts within South Africa's unemployment: in some regimes the South African unemployment rate is stationary whilst in others it is not.

Preliminary Analysis of South Africa's Unemployment Rate

Six traditional statistical models were trained to forecast South Africa's unemployment rate: the naïve model which assumes the last observation will continue. The model simply assumes that the 25% observed on 1st of October 2004 will persist over the forecast period; the moving average is a three-period rolling average, where the three was selected through trial and error and was found to be the best performing average; Simple Exponential Smoothing (SES), Holt, and Holt-Winters are similar to each other and they work by adding weights to past observations with more recent observations weighing the most. Holt also leverages trends in the data whilst Holt-Winters leverages both trends and seasonality and SES just adds weights to the observations; lastly, ARIMA, the most popular traditional regression model: it is a combination of the moving average model and smoothing techniques with differencing to detrend the series.

The models were trained and tested on time series data comprising of 209 quarterly observations of the South African unemployment rate from January 1970 to January 2020. This data was split into training and testing, where, January 1970 to October 2004 was the testing data and January 2005 to January 2020 the testing data. The performance measure used was $RMSE$ over the test period. $R-squared$ (R^2) was also evaluated to test each model's ability to explain South Africa's unemployment rate.

Table 4.2 shows the preliminary results obtained. The results showed that of the traditional statistical models, Holt had the lowest error rate as measured by the $RMSE$. Holt also had the highest R^2, meaning that the model is the most suitable given the data. The model is closely followed by a version of ARIMA: SARIMA. SARIMA was used instead of plain ARIMA because the time series decomposition showed that the unemployment rate is seasonal.

Table 4.2: Preliminary data analysis results

Model	**Root Mean Squared Error** ($RMSE$)	**R-squared**(R^2)
Naïve Method	2.110266	-0.00140052
Moving Average	2.108790	-0.00000001
Simple Exponential Smoothing	2.108942	-0.00014383
Holt	*1.988286*	*0.1110222*
Holt Winters	2.936220	-0.9386988
ARIMA (SARIMA)	2.001990	0.09872572

The performance of Holt as compared to Holt-Winters, suggests that the South African unemployment rate is 'more trendy than it is seasonal': its trend provides greater explanation for it than its seasonality. Figure 4.3 provides a depiction of the forecast of the six models that were used for this preliminary analysis.

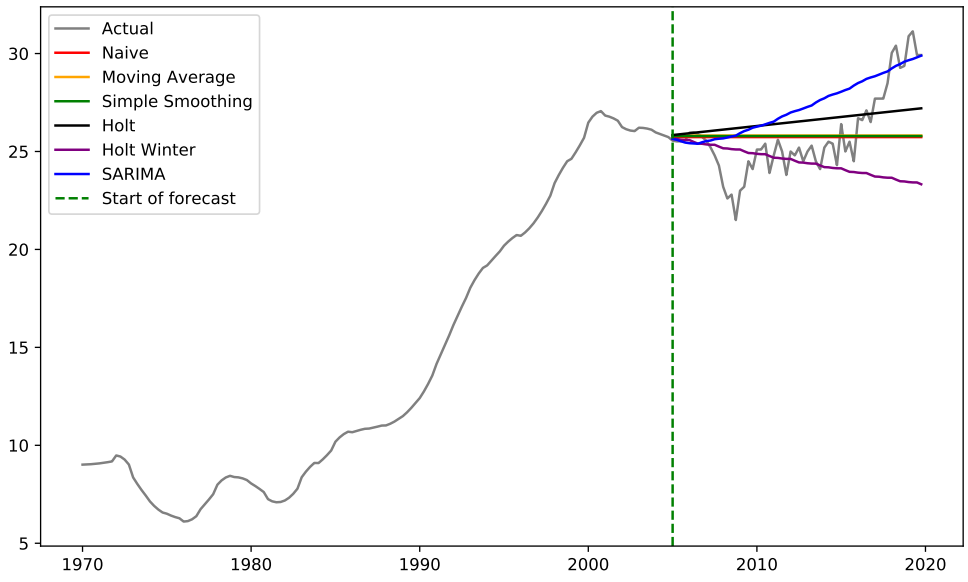


Figure 4.3: Preliminary forecasts with traditional statistical models

The results of this preliminary analysis will be used as a benchmark for this research. These models are orthodox benchmarks in unemployment forecasting [Aiken 1996; Hall 2018; Cook and Hall 2017; Makridakis *et al.* 2018]. Mean Squared Error (MSE), Root Mean Squared Error ($RMSE$), and Mean Absolute Error (MAE) are typically used for evaluating the

performance of unemployment rate forecasting [Brooks 2014; Hyndman and Athanasopoulos 2018]. This research will use these standard measures along with other measures:

- $RMSE$, MSE, and MAE will be used to evaluate the ability of models to forecast South Africa's unemployment rates on the test data. Makridakis *et al.* [2018] suggests using both MAE and MSE simultaneously because MAE penalizes large errors whilst MSE does not. MSE penalised predictions that are far from the mean.
- The Confusion Matrix will be used to determine the model's ability to rightly classify the unemployment rates of over different groups of people: youths, females, and geographic location.
- AIC, BIC and R^2 will be used to evaluate which of various machine learning models best fit South Africa's unemployment data.

4.8 Ethical Considerations

The data to be used for this research is publicly available and widely used for economic research and business reporting. Therefore, because of the orthodox nature of this data there is no ethical clearance required to access the selected data sets.

This chapter provided the methodology that will be employed in this research as well as the data to be used and some preliminary results for the South African context. The next chapter, chapter 5, provides a research plan that will be utilized to deliver this research.

Chapter 5

Research Plan

The research will be executed over eighteen months, including the research proposal phase. The key parts of the research are the data preprocessing, experimentation phase, and results writing. Public holidays and university closures have been factored into the project plan. Table 5.1 provides details of this plan.

Table 5.1: Proposed research schedule

Task	**Start Week**	**Duration**	**Review 1**	**Review 2**	**Sign-Off**
Research proposal	13/01/2020	3.5 months	10/02/2020	27/04/2020	4/05/2020
Proposal submission and presentation	11/05/2020	1 month	N/A	N/A	25/05/2020
Analysis and interpretation					
Data collection and pre-processing	11/05/2020	2 months	8/06/2020	13/06/2020	20/07/2020
Feature selection	27/07/2020	2 weeks	10/07/2020		10/08/2020
Model exploration (regression, neural networks, bayesian networks, PGMs, etc)	17/08/2020	3 months	21/09/2020	2/11/2020	23/11/2020
Model evaluation	17/08/2020	3 months	21/09/2020	16/11/2020	7/12/2020
Labour Economist input for interpretation	23/11/2020	2 weeks	16/11/2020		7/12/2020
Results interpretation write up	14/12/2020	1 month	4/01/2021	25/01/2021	1/02/2021
Experimentation write up	14/12/2020	2 months	25/01/2021	15/02/2021	01/03/2021
Research report finalisation					
Literature review update	01/03/2021	1 month	15/03/2021	29/03/2021	12/04/2021
Research report write (full thesis)	29/03/2021	1 month	26/04/2021		3/05/2021

Table 5.1 continued from previous page

Task	Start Week	Duration	Review 1	Review 2	Sign-Off
Professional review (grammar and flow)	3/05/2021	1 month	N/A	N/A	3/06/2021
Thesis submission	7/06/2021				7/06/2021

5.1 Research Risks

There are currently no key risk that pose a signifact threat to the project except that the researcher works full time in Pretoria whilst the University of Witwatersrand is situated in Johannesburg, which limits the researcher's access the supervisor. Electronic communication will be fully exploited to ensure that this risk does not materialise.

COVID-19 has changed the economy and education systems across the world. As the university and students adjust to the new way of working. This research might not meet the intended timelines as everyone adjust to a new way of working.

This chapter described the overall plan that will be used in carrying out this research over a eighteen-month period. The next chapter is the final chapter, which will summarize this research proposal.

Chapter 6

Conclusion

According to Statistics South Africa [2019], the unemployment rate of South Africa is currently 29.1%. This unemployment rate puts the country in the top 10 countries with the highest unemployment rates [Meyer 2014]. Despite numerous policy interventions, the unemployment rate has been trending upwards since the dawn of democracy [Levinsohn 2007; Brynard 2011]. Currently, policy direction is informed by forecasts derived through economic (traditional statistical) techniques and expert judgement [Levinsohn 2007; Makridakis 1988]. However, these techniques are mostly suitable for data that is stationary and white noise generated.

South Africa's unemployment rate is asymmetric, nonlinear, non-stationary, seasonal, and trendy [Vermeulen 2017]. Hence, traditional statistical methods are prone to error when forecasting unemployment. In 1996, Aiken [1996] showed that neural networks could be used to improve the accuracy of forecasting unemployment in the United States of America. Since then several researchers have attempted the same experiments in their countries, demonstrating that machine learning is a suitable alternative to forecasting unemployment [Hall 2018; Sermpinis *et al.* 2014; Olmedo 2014; Atsalakis *et al.* 2007; Katris 2019ab; Sharma and Singh 2016]. These experiences are primarily in North America, Asia and Europe. These have not yet been applied to South Africa despite this being needed [Fourie 2011].

In North America, Asia and Europe there are few structural impacts to their economies, unlike South Africa where the last 30 years has been marked with structural changes [De Lannoy *et al.* 2018]. Due to these structural changes, well established economic models such as Okun's Law and Philips Curve are not strong predictors of unemployment in the country [Ball *et al.* 2019; Vermeulen 2017; Marinkov and Geldenhuys 2007]. Therefore, South Africa would benefit from machine learning models to forecast unemployment as these models use the underlying data to detect predictive patterns.

This research, therefore, intends to apply machine learning techniques to forecast unemployment in South Africa. As well as determining the key drivers of unemployment and their relative influence on each other and unemployment. The research will use Neural Net-

works, Regression, Information Gain, Hidden Markov Models, and Probabilistic Graphical Models. This research is part of the global attempts to use machine learning to predict unemployment [Hall 2018; Cook and Hall 2017]. Therefore, the research contributes to an important conversation globally and it is important for South Africa: a country where the unemployment rate keeps rising without a clear end in sight.

Bibliography

[Aiken 1996] Milam Aiken. *A neural network to predict civilian unemployment rates*. Technical report, 1996.

[Ajoodha 2018] Ritesh Ajoodha. *Influence modelling an dlearning between dynamic bayesian networks using score based structure learning*. PhD thesis, University of Witswatersrand, 2018.

[Andolfatto 2006] David Andolfatto. *Search Models of Unemployment*. Technical report, 2006.

[Atsalakis *et al.* 2007] George S. Atsalakis, Camelia Ioana, and Christos H. Skiadas. Forecasting Unemployment Rate Using a Neural Network with Fuzzy Inference System. *ICAP*, 2007.

[Ball *et al.* 2019] Laurence Ball, Davide Furceri, Daniel Leigh, and Prakash Loungani. Does One Law Fit All? Cross-Country Evidence on Okun's Law. *Open Economies Review*, nov 2019.

[Barnichon and Nekarda 2013] Regis Barnichon and Christopher J Nekarda. *The Ins and Outs of Forecasting Unemployment: Using Labor Force Flows to Forecast the Labor Market*. Technical report, Board of Governors of the Federal Reserve System, 2013.

[BCG 2019] BCG. *Mass uniqueness a global challenge for one billion workers*. Technical report, BCG, 2019.

[Belling 2020] Ayal Belling. *South Africa's State of Unemployment Disaster*, 2020.

[Brooks 2014] Chris Brooks. *Introductory Econometrics for Finance*. Cambridge University Press, 3rd edition, 2014.

[Brynard 2011] P A Brynard. *The Implementation of Unemployment Policies in South Africa*. Technical Report 2, 2011.

[Burger and Fourie 2015] Philippe Burger and Frederick Fourie. Macroeconomic policy and South African unemployment: developing a three-segment macroeconomic model. 2015.

[Burger and Fourie 2019] Philippe Burger and Frederick Fourie. The unemployed and the formal and informal sectors in South Africa: A macroeconomic analysis. *South African Journal of Economic and Management Sciences*, 22(1), 2019.

[BusinessTech 2019] BusinessTech. *More than half of South Africa's labour force is affected by skills mismatch*, 2019.

[Cerqueira *et al.* 2019] Vitor Cerqueira, Luis Torgo, and Carlos Soares. Machine Learning vs Statistical Methods for Time Series Forecasting: Size Matters. *arXiv*, 2019.

[Chandrashekar and Sahin 2013] Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, 2013.

[Claus 2007] Edda Claus. *Centre for applied macroeconomic analysis six leading indexes of new zealand employment*. Technical report, 2007.

[Clements *et al.* 2004] Michael P Clements, Philip Hans Franses, and Norman R Swanson. Forecasting economic and financial time-series with non-linear models. *International Journal of Forecasting*, 20(2):169–183, 2004.

[Cook and Hall 2017] Thomas R. Cook and Aaron Smalter Hall. Macroeconomic Indicator Forecasting with Deep Neural Networks. 2017.

[Dahlhaus and Eichler 2003] Rainer Dahlhaus and Michael Eichler. *Causality and graphical models in time series analysis*. Technical report, 2003.

[Davies *et al.* 2009] Rob Davies, Dirk Van Seventer, and Miriam Altman. *Leading indicators of employment in South Africa*. Technical report, Human Sciences Research Council, Pretoria, 2009.

[De Gooijer and Hyndman 2006] Jan G De Gooijer and Rob J Hyndman. *25 Years of Time Series Forecasting*. Technical report, 2006.

[De Lannoy *et al.* 2018] Ariane De Lannoy, Lauren Graham, Leila Patel, and Murray Leibbrandt. *What drives youth unemployment and what interventions help? A systematic overview of the evidence and a theory of change high-level overview report*. Technical report, 2018.

[Department of Higher Education and Training 2019] Department of Higher Education and Training. *Skills supply and demand in South Africa*. Department of Higher Education and Training, 2019.

[Diana *et al.* 2014] Rita Diana, I. Nyoman Budiantara, Purhadi, and Satwiko Darmesto. Statistical modeling for unemployment rate using smoothing spline in semiparametric multivariable regression model with Bayesian approach. *Model Assisted Statistics and Applications*, 9(2):159–166, 2014.

[Dludla 2019] Siphelele Dludla. *At least 6.7 million South Africans jobless with unemployment rate at an 11-year high*, 2019.

[Ewing *et al.* 2005] Bradley T. Ewing, William Levernier, and Farooq Malik. Modeling Unemployment Rates by Race and Gender: A Nonlinear Time Series Approach. *Eastern Economic Journal*, 31(3):333–347, 2005.

[Fok *et al.* 2005] Dennis Fok, Dick van Dijk, and Fransesa Hans Philip. Forecasting aggregates using panels of nonlinear time series. *International Journal of Forecasting*, 21, 2005.

[Fourie 2011] Frederick Fourie. The South African unemployment debate: three worlds, three discourses? 2011.

[Fransesa *et al.* 2004] Hans Philip Fransesa, Richard Paapa, and Björn Vroomen. Forecasting unemployment using an autoregression with censored latent effects parameters. *International Journal of Forecasting*, 20(2):255–271, 2004.

[Funke 1992] Michael Funke. Time-series forecasting of the German unemployment rate. *Journal of Forecasting*, 11(2):111–125, feb 1992.

[Goodfellow *et al.* 2016] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[Hall 2018] Aaron Smalter Hall. *Machine Learning Approaches to Macroeconomic Forecasting*. Technical report, Economic Review (Kansas City, MO), 2018.

[Hodge 2002] D Hodge. Inflation Versus Unemployment in South Africa: Is There a Trade-Off? *The South African Journal of Economics*, 70(3):193–204, mar 2002.

[Hyndman and Athanasopoulos 2018] Rob J Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, Melbourne, Australia, 2 edition, 2018.

[IMF 2020] IMF. *World Economic Outlook (October 2019) - Unemployment rate*, 2020.

[Investec 2020] Investec. *Investec macro-economic outlook*. Technical report, Investec, Sandton, 2020.

[InvestmentMap 2020] InvestmentMap. *Foreign investment opportunities, FDI statistics, company data, trade and tariff data*, 2020.

[IOL 2019] IOL. *There aren't enough jobs in Africa for population*, 2019.

[Jelena *et al.* 2017] Mladenovic Jelena, Ilic Ivana, and Kostic Zorana. Modeling The Unemployment Rate At The Eu Level By Using Box-Jenkins Methodology. *KnE Social Sciences*, 1(2):1, mar 2017.

[Jones 2019] Aidan Jones. *Unemployment: our biggest problem*, 2019.

[Katris 2019a] Christos Katris. Forecasting the Unemployment of Med Counties using Time Series and Neural Network models Forecasting the Unemployment of Med Counties using Time Series. *Journal of Statistical and Econometric Methods*, 8(2):37–49, 2019.

[Katris 2019b] Christos Katris. Prediction of Unemployment Rates with Time Series and Machine Learning Techniques. *Computational Economics*, 2019.

[Kouziokas 2019] Georgios N. Kouziokas. Unemployment Prediction in UK by Using a Feed-forward Multilayer Perceptron. pages 65–74. 2019.

[Kreiner and Duca 2019] Aaron Kreiner and John V. Duca. Can machine learning on economic data better forecast the unemployment rate? *Applied Economics Letters*, 2019.

[Levinsohn 2007] James Levinsohn. *Two Policies to Alleviate Unemployment in South Africa*. Technical report, 2007.

[Liang 2005] Faming Liang. Bayesian neural networks for nonlinear time series forecasting. *Statistics and Computing*, 15(1):13–29, jan 2005.

[LumenLearning 2020] LumenLearning. *Production Cost — Boundless Economics*, 2020.

[Mahipan *et al.* 2013] Kanlapat Mahipan, Nipaporn Chutiman, and Bungon Kumphon. A forecasting model for thailand's unemployment rate. *Modern Applied Science*, 7(7):10–16, 2013.

[Makridakis *et al.* 2009] Spyros Makridakis, Robin M. Hogarth, and Anil Gaba. Forecasting and uncertainty in the economic and business world. *International Journal of Forecasting*, 25(4):794–812, oct 2009.

[Makridakis *et al.* 2018] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PLOS ONE*, 13(3):e0194889, mar 2018.

[Makridakis *et al.* 2020] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The M4 Competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1):54–74, jan 2020.

[Makridakis 1988] Spyros Makridakis. Metaforecasting. Ways of improving forecasting accuracy and usefulness. *International Journal of Forecasting*, 4(3):467–491, jan 1988.

[Marinkov and Geldenhuys 2007] Marina Marinkov and Jean-pierre Geldenhuys. Cyclical unemployment and cyclical output: an estimation of okun's coefficient for South Africa. *The South African Journal of Economics*, 75(3):373–390, sep 2007.

[Mathworks 2020] Mathworks. *Learning Vector Quantization (LVQ) Neural Networks - MATLAB & Simulink*, 2020.

[McIntyre and Cedric 2019] Alex McIntyre and Sam Cedric. *IMF Forecasts Show It's Hard to Predict the Global Economy*, 2019.

[Meyer 2014] Daniel Francois Meyer. Job creation, a mission impossible? The South African case. *Mediterranean Journal of Social Sciences*, 5(16):65–77, 2014.

[Mkhabela 2020] Miyelani Mkhabela. *Covid-19 puts at least 1.5 million South African jobs at risk*, 2020.

[Montgomery *et al.* 1998] Alan L. Montgomery, Victor Zarnowitz, Ruey S. Tsay, and George C. Tiao. Forecasting the U.S. Unemployment Rate. *Journal of the American Statistical Association*, 93(442):478, jun 1998.

[Moriwaki 2020] Daisuke Moriwaki. Nowcasting Unemployment Rates with Smartphone GPS Data. pages 21–33. 2020.

[Nedbank 2020] Nedbank. *Forecast and data*, 2020.

[OECD 2019] OECD. *Forecasting methods and analytical tools - OECD*, 2019.

[Olmedo 2014] Elena Olmedo. Forecasting Spanish Unemployment Using Near Neighbour and Neural Net Techniques. *Computational Economics*, 43(2):183–197, jan 2014.

[Pavlicek and Kristoufek 2015] Jaroslav Pavlicek and Ladislav Kristoufek. Nowcasting Unemployment Rates with Google Searches: Evidence from the Visegrad Group Countries. *PLOS ONE*, 10(5):e0127084, may 2015.

[Pelaez 2006] Rolando F. Pelaez. *Using neural nets to forecast the unemployment rate*, jan 2006.

[Sermpinis *et al.* 2014] Georgios Sermpinis, Charalampos Stasinakis, Konstantinos Theofilatos, and Andreas Karathanasopoulos. Inflation and Unemployment Forecasting with Genetic Support Vector Regression. *Journal of Forecasting*, 33(6):471–487, sep 2014.

[Shankar *et al.* 2016] Vignesh Shankar, Ansulie Cooper, and Harvey Koh. *Reducing Youth Unemployment in South Africa*. Technical report, 2016.

[Sharma and Singh 2016] Saloni Sharma and Sanjay Singh. Unemployment rates forecasting using supervised neural networks. In *Proceedings of the 2016 6th International Conference - Cloud System and Big Data Engineering, Confluence 2016*, pages 28–33. Institute of Electrical and Electronics Engineers Inc., jul 2016.

[South African Reserve Bank 2020] South African Reserve Bank. *Welcome to the South African Reserve Bank - South African Reserve Bank*, 2020.

[Statistics South Africa 2019] Statistics South Africa. *Statistical release P0211*. Technical report, Statistics South Africa, Pretoria, 2019.

[Statistics South Africa 2020] Statistics South Africa. *Statistics South Africa — The South Africa I Know, The Home I Understand*, 2020.

[Taylor *et al.* 2016] Timothy Taylor, Steven A. Greenlaw, and OpenStax. Principles of Economic - Unemployment. In OpenStax, editor, *Principles of Economic*, chapter 21. OpenStax, 2016.

[Trading Economics 2019] Trading Economics. *China Unemployment Rate*, 2019.

[TradingEconomics 2019a] TradingEconomics. *South Africa Unemployment Rate*, 2019.

[TradingEconomics 2019b] TradingEconomics. *Unemployment rate - Countries list*, 2019.

[TradingEconomics 2020] TradingEconomics. *South Africa GDP Growth Rate*, 2020.

[Ullah *et al.* 2017] Muhammad Najeeb Ullah, Kim Ki Su, and Bahrawar Jan. Forecasting, Cointegration and Causality Analysis of Unemployment Using Time Series Models. *SSRN Electronic Journal*, jun 2017.

[UNCTAD 2020] UNCTAD. *UNCTAD — Home*, 2020.

[Vermeulen 2017] J.C. Vermeulen. Inflation and unemployment in South Africa: Is the Phillips curve still dead? *Southern African Business Review*, 21, 2017.

[Visser and Arends 2016] Mariette Visser and Fabian Arends. *Skills Supply and Demand in South Africa*. Technical report, 2016.

[WorldBank 2020a] WorldBank. *WorldBank Ease of Doing Business Rankings*, 2020.

[WorldBank 2020b] WorldBank. *WorldBank Global Economic Prospects*, 2020.

[Yang 2007] Lijian Yang. *Nonparametric Modelling of Quarterly Unemployment Rates*. Technical report, 2007.

[Yolcu and Bas 2016] Ufuk Yolcu and Eren Bas. The forecasting of labour force participation and the unemployment rate in Poland and Turkey using fuzzy time series methods. *Comparative Economic Research*, 19(2):5–25, jun 2016.

Appendix A

Model Comparison

Table A.1 shows a performance comparison of machine learning models used to forecast unemployment compared to traditional models. Most the models were discussed in the chapter 2 and chapter 3. This comparison gives a comprehensive review of machine learning models used to forecast unemployment, their performance and implementation specification. Most of these models are applied to univariate data, benchmarked against the ARIMA model with few being multivariate. In cases where multivariate data is considered, dimension reduction techniques such as PCA are used as well as feature selection techniques are employed to reduce the model's computational requirements [Sermpinis *et al.* 2014; Kreiner and Duca 2019; Hall 2018].

Katris [2019a] shows that model architecture makes a difference in performance outcomes. Their finding is that simple neural network architectures can be outperformed by traditional statistical models by preprocessing the data to meet the statistical requirements. Furthermore, feature selection is fairly arbitrary and often based on the domain knowledge of the researcher [Aiken 1996; Atsalakis *et al.* 2007; Kouziokas 2019; Diana *et al.* 2014]. Economic laws such as Okun's Law and Phillips Curve as discussed in section 2.1 are often used as standard features for multivariate models. Therefore, Table A.1, serves as a foundation on which this research is built on.

Table A.1: Comparison of Machine Learning Approaches to Forecasting Unemployment

	Models															Evaluation *			
Author	FFNN	CNN	RNN	PNN	RBFNN	LVQ	ED	ANFIS	FL	SVR	EN	GA	GA-SVR	MARS	LASSO	*MAE*	*RMSE*	*MSE*	Model specifications
Aiken [1996]	■															0,12			A FFNN was used with a genetic algorithm. Training was run for millions of iterations until the error rate was 3%. The researcher considered 250 macroeconomic indicators over a period of 50 years, after analysis seven features where chosen from economic theory and statistical analysis of the data. The features included: Money growth (money supply Ml,money supply M2), interest rates, exports, consumer expectations, consumer sentiment, the composite index of 11 leading indicators, and the Center for International Business Cycle Research's composite index.
Katris [2019a]	■															0,12			A FFNN was used with a single hidden layer with a variations in the number of nodes from 1 to 10. The sigmoid activation function used. The R package AMORE was used in the implementation with 500 epochs. The data was preprocessed to ensure nonlinearity, heteroskedasticity and non-normality, which is a prerequisite for FARIMA / GARCH model (the baseline). Only past unemployment rates were used as input to the model. The data was accessed from the Eurostat database.
Cook and Hall [2017]	■	■	■				■									45,50			Data was split into training, validation and testing set, where, the training was from 1963 to 1996 with every tenth observation added to the validation set. The remaining data was used for testing. A stochastic gradient descent was applied with dropout for regularization. The researchers opted for a univariate approach, were only past unemployment rates where used as the features.

Table A.1 continued from previous page

Author	FFNN	CNN	RNN	PNN	RBFNN	LVQ	ED	ANFIS	FL	SVR	EN	GA	GA-SVR	MARS	LASSO	*MAE*	*RMSE*	*MSE*	Model specifications
Mahipan *et al.* [2013]	■															65,35			Four different FFNNs were considered as candidate models each using the tangent activation function, with changes in the number of hidden layers and the size of the network which varied as follows: 12 input variables, 11 hidden layers and 1 output; 2 input variables, 2 hidden layers and 1 output; and 6 input variables, 11 hidden layers and 1 output. 30% of the data was used for testing purposes. In the preprocessing, the data was normalized to ensure each value was in the range [0.1 - 0.9]. The learning process was stopped after 50 000 epochs were reached. The data was accessed from the UNational Statistical Office of Thailand database, where only past unemployment rates were used. It is important to note that MAPE was used instead of MAE.
Olmedo [2014]	■																	0,02	FFNN with backpropagation method was employed with past unemployment rates being the only input to the model. The data was accessed from Eurostat database.
Atsalakis *et al.* [2007]								■										0,34	FFNN techniques combined with Fuzzy Logic techniques was run over 250 000 epochs. Features used where: population, population change, education, popularity of education courses, types of jobs by education, pay, percentage of labour force in government job, past unemployment, private consumption expenditure, government expenditure, Exports of Good / Services, current unemployment rate and the gross national product. Trial and error was used to determine the final feature list. The data was accessed from Bureau of Labor Statistics in Greece.

Table A.1 continued from previous page

Author	FFNN	CNN	RNN	PNN	RBFNN	LVQ	ED	ANFIS	FL	SVR	EN	GA	GA-SVR	MARS	LASSO	*MAE*	*RMSE*	*MSE*	Model specifications
Yolcu and Bas [2016]								■	■								0,67		Four different fuzzy logic techniques where used. The standard steps in their fuzzy logic algorithms are fuzzification, determination of fuzzy relations, and defuzzification. They employed four different approaches for the 'determination of fuzzy relations' step: standard matrix operations, linguistically defined relationships, feed forward neural networks, and a clustering algorithm. The data used was Turkey's unemployment rate data from January 2005 to December 2013 with 11 different age groups.
Kouziokas [2019]	■																	0,25	The Levenberg-Marquardt algorithm was used as the learning algorithm for the FFNN, it is fast and iterative algorithm. It combines Gauss-Newton algorithm and the steepest descent algorithm used for solving nonlinear least squares problems. The data collected was split into training (60%), validation (20%) and testing (20%). Through trial and error, it was found that a FFNN with 14 neurons was most effective with the linear activation function for the first hidden layer and nine neurons and Tanh-Sigmoid transfer function in the second hidden layer as the activation function. 8 epochs were used. The past unemployment rates, GDP growth, exports of goods and services were used as the model's features. The data accessed from the UK Office for National Statistics.

Table A.1 continued from previous page

Author	FFNN	CNN	RNN	PNN	RBFNN	LVQ	ED	ANFIS	FL	SVR	EN	GA	GA-SVR	MARS	LASSO	*MAE*	*RMSE*	*MSE*	Model specifications
Sermpinis *et al.* [2014]	■											■	■			0,01			A Genetic Algorithm algorithm with high cross over probability of 0,85 was used to determine the feature set. The Support Vector Regression was then used to forecast unemployment rates. The initial population size was 400 chromosomes with maximum generations set at 5 000 with roulette wheel selection and the best member of each generation is maintained (elitism). Essentially, the best features are discovered through evolution. 11 predictors were used: HOUSE, INDP, M1, EMPL, PCE, PI, WAGE, and DJIA with WAGE and INDP being the consistently preferred features. The data was accessed from Federal Reserve Bank of St Louis (FRED) and Bloomberg. The preprocessing step included seasonal adjustments to the data.
Hall [2018]											■					0,66			The Elastic Net was used with 50% of the data was used for training, leveraging a rolling forecast framework, which enables period $t+1$ to be forecasted using data until period t. The model started with 138 macroeconomic variables drawn from a number of economic categories in the FRED-MD database from March 1959 to April 2017. EN was able to discard features through regularization process. The data was accessed from Bureau of Labor Statistics, Federal Reserve Bank of St. Louis FRED.

Table A.1 continued from previous page

Author	FFNN	CNN	RNN	PNN	RBFNN	LVQ	ED	ANFIS	FL	SVR	EN	GA	GA-SVR	MARS	LASSO	*MAE*	*RMSE*	*MSE*	Model specifications
Kreiner and Duca [2019]	■																0,13		Size of FFNN was selected through trial and error, with one to five hidden layers being considered each with 1 to 150 neurons. The model started with 600,000 variables in the FRED database, structured as a recursive tree with data binned into eight categories. PCA was then applied to reduce the variables to 185. Three hidden layers where used each with 97 neurons per layer with a logistic activation function. The data was accessed from FRED.
Kreiner and Duca [2019]															■		0,17		Through regularisation LASSO was able to set all that coefficients to zero with only 10 remaining. The 10 most important features were housing starts, the change in non-financial corporate commercial mortgages, unfilled German job vacancies, retail vehicle registrations, weekly aggregate payrolls, household net lending, 3-month Australian Treasury rates, and producer prices for pharmaceuticals and tractors. The data was accessed from FRED.
Sharma and Singh [2016]	■															0,34			The FFNN had 8 input neurons, 6 hidden layer and 1 output layer with a learning rate of 0,005 and 60 000 epochs. The data was accessed from FRED.
Sharma and Singh [2016]			■													0,5			The RNN had 7 input neurons, 6 hidden layer and 1 output layer were used with a learning rate of 0,003 and 50 000 epochs. The data was accessed from FRED.
Sharma and Singh [2016]				■												2,62			The PSN had 6 input neurons, 5 hidden layers and 1 output layer with a learning rate of 0,002 and of 75 000 epochs. The weight of the first layer was fixed to 1, whilst the rest are updated through the iterations. The data was accessed from FRED.

Table A.1 continued from previous page

Author	FFNN	CNN	RNN	PNN	RBFNN	LVQ	ED	ANFIS	FL	SVR	EN	GA	GA-SVR	MARS	LASSO	*MAE*	*RMSE*	*MSE*	Model specifications
Sharma and Singh [2016]					■											2,14			The RBFN had 8 input neurons, 4 hidden layers and 1 output layer with a learning rate of 0,003 and 45 000 epochs. The data was accessed from FRED.
Sharma and Singh [2016]						■										1,22			In each epoch, the neurons 'compete' to be activated and only one is activate at each iteration. The data was access from FRED.
Katris [2019a]						■										0,21			The FFNN was implemented in the R package AMORE. A single hidden layer was used and trial and error was used to determine number of nodes: from 1 to 10. The model used an adaptive gradient descent with adaptive backpropagation for 500 epochs with the sigmoid activation function. The model was implemented with only past unemployment rates being the only input to the model. The data was accessed from the Eurostat database.
Katris [2019a]										■						0,25			A standard SVR was implemented from the R programming language without any modifications. The model was implemented with only past unemployment rates being the only input to the model. The data was accessed from the Eurostat database.
Katris [2019a]														■		0,24			A standard MARS was implemented from the R programming language. The model with lagged 1, 2, 3 and 4 variables was implemented. The model was implemented with only past unemployment rates being the only input to the model. The data was accessed from the Eurostat database.

Table A.1 continued from previous page

Author	FFNN	CNN	RNN	PNN	RBFNN	LVQ	ED	ANFIS	FL	SVR	EN	GA	GA-SVR	MARS	LASSO	*MAE*	*RMSE*	*MSE*	Model specifications
Diana *et al.* [2014]														■				0.32	A semi-parametric multivariable regression model was used with Generalized Maximum Likelihood to estimate the smoothing spline function. Features were pre-determined for the model: percentage of educated above high school, economic growth, population density, the ratio of investment to labor force, regional minimum wage, and ratio of large and medium industries to labor force. The data was accessed from the Indonesian Central Bureau of Statistics.

Legend:

- FFNN - Feed Forward Neural Network
- CNN - Convolutional Neural Network
- RNN - Reccurent Neural Network
- PNN - Psi Sigma Neural Network
- RBFNN - Radial Basis Function Neural Network
- LVQ - Learning Vector Quantization
- ED - Encoder Decoder
- ANFIS - Adaptive Neural Fuzzy Inference System
- FL - Fuzzy Logic / Fuzzy Time Series
- SVR - Support Vector Regression
- EN - Elastic Net
- GA - Genetic Algorithm
- MARS - Multivariate Adaptive Regression Splines

* The performance measures are in most cases average performance as the researchers often had error rates calculated over multiple horizons. In other cases models are run across multiple countries, in such cases the best country's results are displayed: the results were often significantly close to one another.