Expectation Maximization Imputation in Forecasting the Rate of Unemployment in South Africa

Andile Thabethe
School of Computer Science and Applied Mathematics
University of The Witwatersrand
Johannesburg, South Africa
2092108@students.wits.ac.za

***Abstract*—According to statistics from the fourth quarter of 2021, South Africa's unemployment rate grew from 34.9 percent in the previous period to 35.3 percent. Unemployment rates are widely recognized as crucial indicators of a nation's labor market success. Particularly during difficult economic times and recessions, it is an economic indicator that is closely watched. Previous studies have used a variety of data imputation techniques to impute missing values in the SARB and BER data sets, including constant imputation, last known value imputation, multivariate imputation using chained equations, k-nearest neighbor imputation, forward imputation, and mean value imputation. Last known value imputation imputation produced the best results. Past research improved the prediction of the unemployment rate in South African through using machine learning approaches over traditional statistical models, deep learning, feature selection and engineering but not through data imputation. This study investigated the performance of expectation maximization imputation against last known value imputation to better forecast the unemployment rate South African unemployment rate. Traditional statistical models performed better when the expectation maximization imputed data was used. The deep learning methods performed better when the last known value data set was used. This study demonstrated that there is a role for expectation maximization imputation in the prediction of the rate of unemployment in South Africa when traditional statistical methods are used.**

***Index Terms*—machine learning, forecast, unemployment rate, expectation maximisation imputation, last known value imputation**

I. Introduction

This study compares last known value imputation and expectation maximization imputation to see how accurately the unemployment rate in South Africa will be predicted.

The unemployment rate is globally acknowledged as an important indication of a country's labor market performance. It is an economic indicator that is carefully monitored, particularly during difficult economic times and recessions. This is because the unemployment rate has an impact on the entire economy, not just people who are unemployed. The amount and persistence of unemployment variables have far-reaching consequences.

Predicting the rate of unemployment with as much precision as feasible might help with decision-making regarding the economy and policy formation, allowing for the identification and reduction of the core causes.

Previous work in the prediction of the rate of unemployment in South Africa has used constant imputation, last known value imputation, k-nearest neighbor imputation, forward imputation, mean value imputation and, multivariate imputation by chained equations in predicting the rate of unemployment in South African [8]. Expectation maximization has not been employed for missing data imputation.

The feature set accessed from the South African Reserve Bank (SARB) has values that are missing at random (MAR). MAR arises when the input vector's missing value depends on other variables, making it possible to identify the pattern through which the data becomes missing [5]. Expectation maximization is applicable whenever the data is missing completely at random (MCAR) or MAR.

Prior research has improved the accuracy of predicting the rate of unemployment in South African through the use of machine learning methods over traditional statistical methods [9], deep learning [8] and feature selection [10]. An improvement through a better data imputation technique has not been attempted.

This study brings a novel approach in imputation of missing data in the SARB data set using expectation maximization imputation. Therefore, the aim of this study is to compare expectation maximization imputation and last known value imputation in order to better forecast the unemployment rate in South African.

This study tests the hypothesis,

- **H1**: Expectation maximization imputation produces better performing models compared to the last known value imputation.
- **H0**: Expectation maximization imputation *does not* produce better performing models compared to the last known value imputation.

The hypothesis is tested by training and testing traditional statistical methods, machine learning models and deep learning techniques using last known value imputed data and expectation maximization imputed data and comparing the performance. The mean absolute error (MAE), the standard deviation (SD) over errors and root mean square error (RMSE) of each model will be compared.

The document structure is as follows: In section 2, the background is detailed where the expectation maximization algorithm is discussed. The related work is detailed in section 3. Section 4 elaborates on the experimental design with subsections detailing the data preparation, methodology and performance measures. In section 5, the results are analyzed and discussed.

II. Background

A. Expectation Maximization

Arthur Dempster *et al* introduced the EM algorithm in 1997. It is utilized to determine the local maximum likelihood parameters of a statistical model in the event that the latent variables are available or the data is missing or incomplete. It does so by iterating between the E-step (Expectation) and the M-step (Maximization) to estimate those parameters [3].

In order to determine the model's parameters when there are latent variables, the following phases are executed in the expectation maximization algorithm:

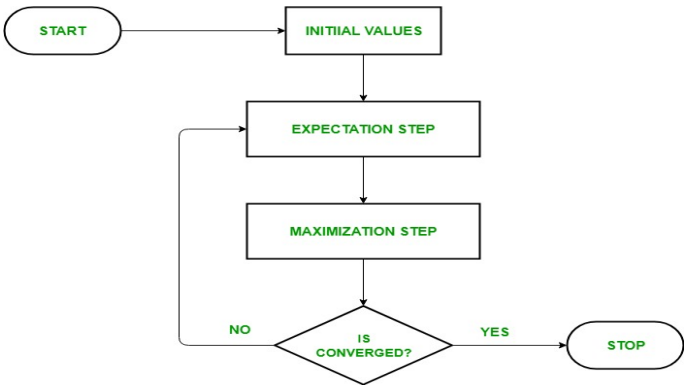


Figure 2.1 Diagram of the EM algorithm
Adapted from: Raman (2019) Expectation-Maximization Algorithm

1) An initial set of parameters is considered at this step. The system is provided a collection of incomplete and unseen data with the presumption that the observed data comes from a particular model.
2) **Expectation Phase**: In this step, the values of the data's missing values are estimated. To essentially infer the values in the missing data, it uses the observed data.
3) **Maximization Phase**: After the expectation phase updates the data's missing values, the maximization step creates complete data generated in the E-step.
4) The values are examined in the final phase to determine whether or not they are convergent. If the values are same, nothing is done; otherwise, steps 2 and 3 are repeated until convergence is achieved.

Convergence: In probability, the idea of convergence is founded on intuition. Let's assume we have two random variables if the probability of their difference is very tiny, it is considered to be converged. It converges when the estimates of the parameters are nearly identical [4].

The EM algorithm's fundamental principle is to estimate missing data using the observed data before changing the parameter values.

III. Related Work

The rate of unemployment in South Africa increased from 34.9 percent in the prior period to 35.3 percent in the last quarter of 2021 [1]. The unemployment rate is globally acknowledged as an important indication of a country's labor market performance. With speculations of a global recession, it is vital that the rate of unemployment be predicted as accurately as possible to assist policy makers in alleviating the problem.

Data used in forecasting the unemployment rate is typically sourced from a governmental or federal body of a country or state. In South Africa, the data comprising of 147 features with varied frequencies, was accessed from the BER and the SARB [9], [8]. The features range across vital sectors of economy of South African: real, fiscal, financial, and external sectors, including demographic data. For other countries, [12] sourced data from the open-access data repositories: Federal Reserve Economic Data sets and the Organisation for Economic Co-operation and Development data repository for the unemployment rates.

Feature selection has been employed to improve the precision of the models that are used to predict the rate of unemployment in South Africa. Feature selection methods utilized are filter, wrapper, and embedded techniques. The accuracy of forecasting is improved by feature selection [10].

Machine learning projections are frequently compared to classical statistical models for comparison [9]. Mulaudzi and Ajoodha bench marked seven machine learning methods, some of which were deep learning methods, against six classic statistical methods to predict the rate of unemployment in South Africa and the underlying patterns [9]. They showed that machine learning methods, specifically deep learning methods, can better predict the rate of unemployment the South Africa with increased accuracy than traditional statistical methods.

Six data imputation methods were used to fix the missing values in the data set [2]. Data imputation techniques employed include: k-nearest neighbor imputation, constant imputation, multivariate imputation by chained equations, last known value imputation, forward imputation and mean value imputation. Compared to other imputation procedures, the last known value data imputation technique produced reduced error rates [9].

Past research has show that using machine learning techniques rather than conventional statistical ones [9], deep learning techniques [8] and feature selection [10] have increased the precision of forecasting the rate of unemployment in South Africa. No attempt has been made to make improvements using improved data imputation methods.

A study done in Ontario, Canada demonstrated that expectation maximization imputation provided a better alternative to the standard mean imputation [11].

A particular advantageous data imputation technique that has not been used for missing data imputation in the SARB and BER data sets used for the forecasting of the South African rate is expectation maximization imputation.

IV. Experimental Design

To test the hypothesis: ***H1****: Expectation maximization imputation produces better performing models compared to the last known value imputation*, six models (three traditional statistical models, one machine learning model and two deep learning) were trained and tested. This section explains the procedures used to carry out these tests.

A. Data Preparation

In this study, the data utilized will be sourced from the SARB and the BER. The data set has 147 macroeconomic predictors, which range across the South African economy including the target variable, the South African rate of unemployment. The data spans January 1922 to January 2020 with 1432 observations. Due to varying frequencies being mixed there is some missing data.

During data pre-processing, the rows that had missing unemployment rates (target variable) were removed. After that, 794 observations were left with dates dating from March 1970 to January 2020. The data was then up-sampled to convert the data from a monthly frequency to a quarterly frequency such that they all have the same frequency.

For the remaining 201 observations, the missing data were imputed using last known value imputation and expectation maximization imputation. The data set was then separated into training and testing subsets, with 194 observations used for training the models and 7 observations used for testing the models. Similar test sizes were utilized in prior literature.

B. Methodology

As discussed above, two data imputation techniques were used for the missing values: last known value imputation and expectation maximization maximization.

Three traditional statistical models were selected to be trained and tested for the experiment: Holt-Winters, autoregressive integrated moving average (ARIMA), simple exponential smoothing (SES). These three models were chosen as they are the three best performing traditional statistical models in the prediction of the unemployment rate of South Africa [9].

In addition, three machine learning methods were also selected to be utilized for the experiment: long short-term memory (LSTM) network, gated reccurent units (GRU), and multilayer perceptron (MLP). LSTMs and GRUs were selected as they have been demonstrated to be the best performing deep learning models, and best performing models overall [8] in the forecast of the South African unemployment rate. The MLP was chosen as a comparative machine learning model.

Grid search was not possible using neural network techniques since it required too much computer power. To find the hyperparameters and parameters, the LSTM, GRU, ARIMA, MLP, SES and Holt-Winters models were optimized using trail and error. The hyperparameters and parameters for each model were kept constant for each data set.

C. Performance Measure

This study utilized the standard deviation (SD) over errors, root mean square error (RMSE) and, mean absolute error (MAE) to measure the performance of the models and imputation techniques. MAE is advised for evaluating accuracy on a single series [6]. In continuous data, the goodness of fit of an imputation technique is normally evaluated by measuring how far the imputed value was from the original value using RMSE. The formula for MAE is given as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}^i - y_i)^2 \tag{1}$$

The formula for RMSE is given as:

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\hat{y}^i - y_i)^2}{n}} \tag{2}$$

and the formula for SD is given as:

$$SD = \sqrt{\sum_{i=1}^{n}\frac{(\hat{y}^i - y_{mean})^2}{n-1}} \tag{3}$$

where $\hat{y}_i$ is the forecast value of the i-th observation and n is the number of observations and y_i is the analogous true values.

V. Results and Discussions

The residuals, trend, and seasonality are the three main components time-series data may be separated into [7]. The rate of unemployment in South African is seasonal and tends upward, with yearly recurring fluctuations in the data (as shown in Figure 4.1 and Figure 4.2). The expectation maximization imputed data was able to capture the trend, seasonality and residuals of the South African rate of unemployment as well as the standard last known value imputation technique. Both

figures depict the increased variability ranging from 2009 to 2020.

Figure 4.1 The breakdown of the unemployment rate in South Africa into trends, seasonality, and residuals after last known value imputation.

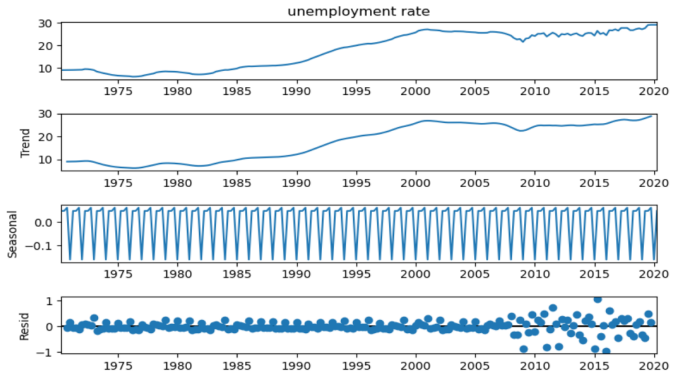
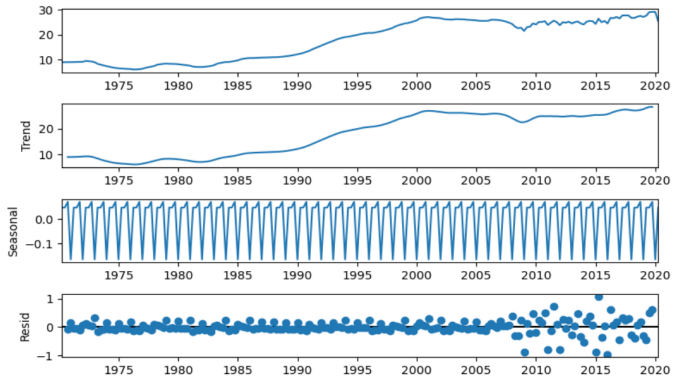


Figure 4.2 The breakdown of the unemployment rate in South Africa into trends, seasonality, and residuals after expectation maximization imputation.



The performance of the data imputation techniques were analyzed with three machine learning models and three traditional statistical models. These experiments allow for the acceptance or rejection of the hypothesis to be decided. This section discusses the findings.

When the models are used to predict the rate of unemployment for each dataset, the results depict that the MAE and RMSE scores of the LSTM and GRU are significantly lower and quite close. This is expected as prior research has shown that the LSTM and GRUs performed best at forecasting the South African unemployment rate [8].

It is noted that the LSTM and GRU had lower MAE scores when the last known value imputed data set was used. It is probably worth noting that these are deep learning learning methods. While in comparison, the expectation maximization imputed data set returned slightly higher MAE scores in the deep learning methods. The RMSE scores for the deep learning models follow the same trend with lower RMSE scores when the last known value imputation is used and slightly higher RMSE scores when expectation maximization is used.

The traditional statistical models (ARIMA, Holt-Winters and SES) had lower MAE and RMSE scores for the dataset that utilized expectation maximization imputation while the scores in the last known value imputed data set were slightly higher. The RMSE and MAE scores of the traditional statistical models were higher overall compared to the deep learning models. The poor performance of traditional statistical models against deep learning techniques has been demonstrated in prior work [8].

The standard deviations over the errors are lower in the deep learning models in the expectation maximization imputed data. While in the traditional statistical models, the standard deviations over errors are lower for the last known value imputed data.

The results acquired are limited as only a few prediction models were employed.

Table 1: MAE of ARIMA, Holt-Winters, SES, MLP, GRU and LSTM models

Model	Last-Known Value	Expectation Maximization
LSTM	0.093	0.268
GRU	0.074	0.306
MLP	8.075	7.608
ARIMA	9.147	6.202
Holt-Winters	8.483	6.102
SES	8.485	5.890

Table 2: RMSE of ARIMA, Holt-Winters, SES, MLP, GRU and LSTM models

Model	Last-Known Value	Expectation Maximization
LSTM	0.101	0.352
GRU	0.082	0.403
MLP	8.351	8.366
ARIMA	10.158	8.100
Holt-Winters	9.352	7.749
SES	9.219	7.625

Table 3: SD of ARIMA, Holt-Winters, SES, MLP, GRU and LSTM models

Model	Last-Known Value	Expectation Maximization
LSTM	0.404	0.200
GRU	0.418	0.272
MLP	1.151	5.528
ARIMA	3.311	4.725
Holt-Winters	2.816	3.854
SES	2.459	3.563

VI. Conclusion

The hypothesis: ***H1****: Expectation maximization performs better at missing data imputation compared to the last known value imputation*, was tested by predicting the South African unemployment rate using macro-economic variables from the SARB data set. It was tested by training and testing six models that utilized expectation maximization imputed data and last known value imputed data. In 8 out of 36 of the evaluations, that is approximately 17%, the expectation maximization imputed data set trained models had lower error rates.

Therefore we fail to reject the null hypothesis: ***H0****: Expectation maximization imputation does not perform better at missing data imputation compared to the last known value imputation*. 83% of the models had lower error rates when the last known value data set was used.

The study has contributed to the existing body of work done in the forecast of the unemployment South African by showing that there is a place for expectation maximization imputation in the forecast of the South African unemployment rate when traditional statistical methods are used.

The study's approach paves the path for more study that might concentrate on expanding the application of imputation to the SARB data set. Further research in the area may be able to support forecasting techniques by analyzing how various forecasting models are impacted by data imputation accuracy.

References

[1] Statistics South Africa. Quarterly labour force survey (qlfs) – q4:2021. *Pretoria: Statistics SA [producer], 2021. Cape Town: DataFirst [distributor], 2021*, 2021.

[2] D. Bertsimas, C. Pawlowski, and Y. Zhuo. From predictive methods to missing data imputation: An optimization approach. *Journal of Machine Learning Research, 1-39,*, 2018.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society, Series B (methodological), pages 1–38*, 1977.

[4] Y. Dong and C. Joanne Peng. Principled missing data methods for researchers. *SpringerPlus, 2(1):222*, 2013.

[5] S. Mohamed F. V. Nelwamondo and T. Marwala. Missing data: A comparison of neural network and expectation maximisation techniques. *Current Science. 93.*, 2006.

[6] R. J. Hyndman. Another look at forecast-accuracy metrics for intermittent demand. *Unpublished*, 2006.

[7] R. J. Hyndman and G. Athanasopoulos. Forecasting: Principles and practice. *OTexts, Melbourne, Australia, 2 edition*, 2018.

[8] R. Mulaudzi and R. Ajoodha. Application of deep learning to forecast the south african unemployment rate: A multivariate approach. *7th IEEE CSDE 2020, the Asia-Pacific Conference on Computer Science and Data Engineering*, 2020.

[9] R. Mulaudzi and R. Ajoodha. An exploration of machine learning models to forecast the unemployment rate of South Africa: A univariate approach. *The International Multidisciplinary Information Technology and Engineering Conference*, 2020.

[10] R. Mulaudzi and R. Ajoodha. Demonstration that the use of feature selection on high dimensional south african macroeconomic data results in improved performance with lower compute requirements. *Sixth International Congress on Information and Communication Technology (6th ICICT 2021) — 25 - 26 February 2021 — London, United Kingdom. SPRINGER*, 2021.

[11] H. M. K. Ghomrawi *at al*. Is there a role for expectation maximization imputation in addressing missing data in research using womac questionnaire? comparison to the standard mean approach and a tutorial. *BMC Musculoskeletal Disorders 2011, 12:109*, 2011.

[12] T. Chakraborty *et al*. Unemployment rate forecasting: A hybrid approach. *Computational Econonomics 57, 183–201*, 2021.