Dissertation

Assignment 001

Student Name: Kurtis Jackson

Student Number: s4105903

Course: Computer Science

Submission Date: 19/04/2024

**University of Gloucestershire**

Contents

[1 Abstract 3](#_Toc163551495)

[2 Introduction 3](#_Toc163551496)

[2.1 Overview 3](#_Toc163551497)

[2.2 Problem Statement 3](#_Toc163551498)

[2.3 Questions 3](#_Toc163551499)

[2.4 Objectives 3](#_Toc163551500)

[3 Literature Review 3](#_Toc163551501)

[3.1 Introduction 3](#_Toc163551502)

[3.2 Evaluating Time Series Methods 3](#_Toc163551503)

[3.3 Predicting Greenhouse Gas Emissions 4](#_Toc163551504)

[4 Methodology 4](#_Toc163551505)

[4.1 Discussion of Data Sets 4](#_Toc163551506)

[4.2 Discussion of Time-Series Analysis 6](#_Toc163551507)

[5 Results and Discussion 6](#_Toc163551508)

[6 Conclusion 6](#_Toc163551509)

[7 References 6](#_Toc163551510)

[7.1 References for code used within implementation 6](#_Toc163551511)

[8 Appendix 7](#_Toc163551512)

# Abstract

# Introduction

## Overview

## Problem Statement

Could the use of time-series analysis be effective in predicting future turnover rates and emissions of greenhouse gases in varying sectors?

## Questions

1. What methods of time-series analysis can be used to predict the data?
2. What method is the most appropriate with both given data sets (GHG emissions, and historical turnover rates)?
3. What considerations need to be taken when considering the sector which is being analysed?
4. What conclusions can be made from the results?

## Objectives

1. To investigate time-series analysis methods in respects to accuracy.
2. To develop a model which predicts the future historical turnover rates and greenhouse gas emissions (GHGe).
3. To evaluate how accurate the proposed model is, quality of data sets, viability of methods such as simulation.
4. To evaluate the results given from this model and draw conclusions on the state of GHGe going into the future, and how this may relate to the size and purpose of a given sector.

# Literature Review

## Introduction

There exists a large number of literatures pertaining to time-series analysis best practices and its use in relevant fields. This section compiles many of these sources into a number of categories including:

* Time-Series Methods (and which ones produce the best results) for general data sets.
* Existing literature in GHGe with respects to future predictions.
* Existing literature in financial turnover rates with respects to future predictions.

## Evaluating Time Series Methods

Parmezan, Souza and Batista (2019) propose a selection of methods from a series of statistical techniques, deep and shallow learning methods which can be seen as the most optimal depending on the nature of the data set. Overall, shallow learning method SVM (Support Vector Machine) is found to be best method of the sample, though other methods such as SARIMA, MLP and kNN-TSPI are also suggested as close competitors. An addressed limitation of this study was the decision to limit it to univariate analysis, when conducting multivariate analysis (factoring in explanatory variables) there may be different results.

In support of the research above, a later study setting out to compare multiple shallow and deep learning methods for time series analysis within the pharmaceutical industry found the shallow learning methods (notably GR-NN, or General Regression Neural Network) to be generally superior to the deep learning methods (LSTM) and classical statistical methods such as ARIMA. The paper goes further to suggest that the reason for this advantage is due to the higher noise of and the smaller sample size of the data set which deep learning methods can struggle with (Rathipriya *et al.*, 2023).

Results of a forecasting model are also dependent on the choice of parameters used with each method, these can include factors such as learning rate, number and size of hidden layers, and number of epochs. To find the best parameters, and therefore reduce the loss of the method, parameter estimation methods have been created which include Cross Validation methods, Out-of-sample (also known as Holdout) methods and Prequential methods. A number of studies have set up to find which methods are most accurate for a given data set, for example one by Cerqueira, Torgo and Mozetič (2020) which empirically ranks a number of these methods and finds the Blocked Cross-Validation method to be the most appropriate in most situations, and also provides some other methods which should be used in case of certain characteristic cases in the dataset (such as if the data set is not stationary, or has a low inter-quartile range). Information from this study is also corroborated by other studies which highlight that Cross-Validation is the most versatile category of method compared to the other two (Bergmeir, Hyndman and Koo, 2018).

## Predicting Greenhouse Gas Emissions

A number of studies provide in depth prediction models for GHGe in a specific industry or sector (Hamrani, Akbarzadeh and Madramootoo, 2020). A study by Alfaseeh et al. (2020) is notable for evaluating the model using 9 methods, including statistical methods as well as shallow and deep learning, in order to find the most fitting one for this study into agriculture. In this study, LSTM is shown to be most accurate model listed of the 9. This is in contrast to earlier pieces of literature (Parmezan, Souza and Batista, 2019; Rathipriya *et al.*, 2023) which showed LSTM to be a relatively poor performer. This discrepancy may be due to the usage of exogenous factors within the agriculture study, or it may be to do with the usage of a Cross Validation method being used for LSTM in the agriculture study whilst the other uses a Holdout method (which is shown to be worse).

# Methodology

## Discussion of Data Sets

A number of assumptions have been made about the usage of the prediction model that will affect the nature of the datasets used:

* An equivalent model, or models, will be used by businesses to compare their emissions per unit of turnover to a known baseline.
* Businesses wishing to use the model are unlikely to have collected large amounts of data on their greenhouse gas emissions.
* Businesses wishing to use the model are likely to have a robust amount of financial data.

Following these assumptions, the requirements of the GHGe model and turnover model will vary. The prediction of GHGe will utilise a sparse dataset (in terms of time-series prediction) in order to approximate a likely, use case, this can heavily affect the quality of different prediction models so it is important to test under this situation to ensure the envisioned end user will get the most accurate results. While the turnover forecasting does not have the same restriction, data relating to turnover per industry sector is limited as only yearly recordings from 2016 to 2021 were found (as provided by the Office for National Statistics) and will need to be simulated as to give a quality prediction.

For the datasets themselves, the GHGe data has been provided by Balance Eco Ltd which documents the carbon dioxide equivalent emissions of 18 sectors in the 30-year period of 1990 to 2019. Whilst 30 data points is sparce for time-series analysis, lack of demand in having businesses publicly release GHGe records will likely mean most businesses will not have as large a data set and individual breakdowns of each sector can provide information in analysing trendlines when it comes to discussing the results which all makes this a suitable dataset for this requirement.

The turnover dataset given from the Office of National Statistics is insufficient for prediction on its own, containing only a 6-year period of 12 sectors. Using another, larger, dataset to approximate the 6-year period is a potential idea in order to simulate a more usable dataset of 12 sectors. A potential approximator is the Gross Domestic Product of the UK and giving every sector a percentage share of the GDP. Whilst the sum of all percentages may not equal, and does in fact exceed, 100% the overall trend may be close enough to the original data as to provide a useful approximation.

Figure 4.1, shown below, expresses the percentage share of the GDP every sector has, but instead of showing the share itself the values are presented as the percentage of the 2016 value per sector (hence every 2016 value is 100%). This has been done to evaluate the trend between year’s, in order to evaluate the effectiveness of GDP as an approximator for turnover. A theoretical perfect approximator would have every year listed as 100% as a 1% increase in the approximator would translate to a 1% increase in every sector’s turnover rate resulting in each percentage share in a sector being the same.

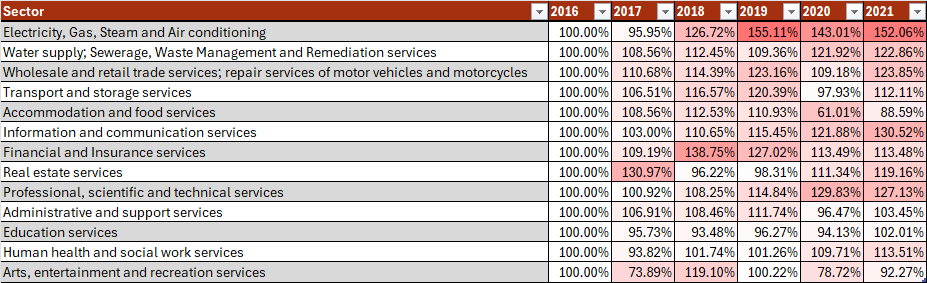


Figure 4.1: UK GDP over Sector Turnover, values are given as a percentage of the 2016 value.

Whilst 2020 appears to be an inaccurate year for using this as an approximator, this is likely due to many sectors experiencing a large decrease in production due to COVID-19 related lockdowns in the same year, which is less pronounced on the GDP dataset. This often results in a large reduction in the percentage figure, which is most apparent in sectors such as wholesale and retail; accommodation and food services; and arts, entertainment and recreation services. Some sectors, such as professional, scientific and technical services, have no apparent effect during 2020 resulting in their percentage increasing.

Whilst this instability likely affects the 2021 results as well, the results before this year appear to be far more consistent which means that predicting values backwards from 2016 will likely yield usable results. For results from 2020 onwards, the resulting decline in turnover is accounted for by the true results from this period and need not to be approximated whilst future results (from 2022 onwards) can be assumed to have recovered from this decline and more closely follow the country’s GDP.

It should be noted that the particular sector of electricity, gas, steam and air conditioning has a poor approximation compared to other sectors. Whilst it will be approximated the same as the other sectors, the affect of this simulation on the final result will be discussed in more detail as it is presented.

## Discussion of Time-Series Analysis

A number of regression models, usable for time-series analysis, have been discussed during the literature review. However, a methodology is needed to evaluate the effectiveness of all models in both predictions independently. A number of other studies have wished to evaluate the effectiveness of time-series analysis models in other use cases and have used a variety of performance metrics such as Mean Squared Error (sometimes Root Mean Squared Error), Mean Absolute Error, POCID, TU Coefficient, Coefficient of determination (R2), and Mean Absolute Percentage Error (Parmezan, Souza and Batista, 2019, pp. 324-332; Hamrani, Akbarzadeh and Madramootoo, 2020, pp. 9-12). It has been noticed that there is generally not a large variance in rankings when using these scores, as in a model that places fourth best in Mean Absolute Error is likely to place fourth best in any other metric. Due to this, only some metrics have been used to test the models. The metrics are as explained below:

* R2, or coefficient of determination, measures the proportion of variation of the variable to predict (in this case GHGe and financial turnover) that is explained by available predictors (time) (Zhang, 2017). In certain settings that are prone to outliers, R2 may not be the most applicable metric and should not be used as a sole predictor (Renaud and Victoria-Feser, 2010). However, as one metric in a selection of others, any disadvantages should be mitigated.
* RMSE, or root mean squared error, is similar to the next performance metric in a way. Both summarise how far the predicted value is from the true value. Whilst MAE gives all areas the same weighting, RMSE will provide a higher (worse) score disproportionately to larger absolute values (Chai and Draxler, 2014) and as such will always provide a higher result then MAE. Whilst similar to MAE, it is important to note that they are not defined identically and should be used in conjunction with one and another rather then being seen as alternatives to each other.
* MAE, or mean absolute error, has been discussed with RMSE. Instead of squaring the error (to remove negatives) then square rooting the result, MAE merely takes the absolute value of the error (to remove negatives). Using MAE in conjunction with RMSE provides more insight than any one metric on its own, for example a RMSE score that is significantly higher than a MAE indicates that errors within a prediction are more “extreme” and individually very far from the true value.
* MAPE, or mean absolute percentage error, is the average percentage difference between the predicted value and the true value. Whilst MAE gives the average absolute difference across all data points, MAPE gives a relative difference from point to point. MAPE is commonly used in finance and is applicable whenever the value to predict is known to always be significantly over 0 (De Myttenaere *et al.*, 2016).

With the metrics to discern the best model outlined, is it now important to outline the steps taken in order to create the presented models and then evaluate them with the provided metrics. All processing was done using Python code, in Jupyter notebooks, using libraries Pandas, NumPy, TensorFlow, SciKit Learn and MatPlotLib with Seaborn.

1. Data from the GHGe dataset is processed in a way that is more ideal for machine learning algorithms, using a Minmax scaler, and then split into sequences where y, the current value, is given with x, the 9 former values.
2. Data is split into an 80/20% split between training data and testing data, where applicable 20% of the training data is turned into validation data.
3. Data is passed through all time-series analysis algorithms presented in the literature review; each being given hyper parameter tuning independently to maximise the accuracy of the model. This can be in the form of increasing or decreasing learning rates, epsilon values or the number of nodes in a deep learning model.
4. Each model is performed 10 times per sector as some types of models, such as LSTM, can have run to run variance.
5. The financial turnover of each sector is simulated using the UK’s GDP, as this data is quarterly the annual turnover data is interpolated between given year values. Data before 2016 is extrapolated using the 2016 value, and data after 2021 is extrapolated using the 2021 value.
6. Using this new data, steps 1-4 are repeated with a new wave of hyperparameter tuning being used for the new dataset.
7. With all models created the best of the 10 models in each sector is selected by using a combined ranking of all proposed performance metrics.
8. For GHGe and turnover, every type of model is compared to each other with the same ranking to find the best type of model to use for every sector.
9. Using the best GHGE and turnover models, a GHGe per pound (measured in tonnes of carbon dioxide equivalent per Pound Sterling) can be created, forecasting future results up to 5 years from writing (2029).

# Results and Discussion

All datasets and files used in making these results are available within Appendix A.

# Conclusion

# References

Alfaseeh, L. *et al.* (2020) ‘Greenhouse gas emission prediction on road network using deep sequence learning’, *Transportation Research Part D: Transport and Environment*, 88, p. 102593.

Bergmeir, C., Hyndman, R.J. and Koo, B. (2018) ‘A note on the validity of cross-validation for evaluating autoregressive time series prediction’, *Computational Statistics & Data Analysis*, 120, pp. 70–83.

Cerqueira, V., Torgo, L. and Mozetič, I. (2020) ‘Evaluating time series forecasting models: an empirical study on performance estimation methods’, *Machine Learning*, 109(11), pp. 1997–2028. Available at: https://doi.org/10.1007/s10994-020-05910-7.

Chai, T. and Draxler, R.R. (2014) ‘Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature’, *Geoscientific model development*, 7(3), pp. 1247–1250.

Hamrani, A., Akbarzadeh, A. and Madramootoo, C.A. (2020) ‘Machine learning for predicting greenhouse gas emissions from agricultural soils’, *Science of the Total Environment*, 741. Available at: https://doi.org/10.1016/j.scitotenv.2020.140338.

De Myttenaere, A. *et al.* (2016) ‘Mean absolute percentage error for regression models’, *Neurocomputing*, 192, pp. 38–48.

Parmezan, A.R.S., Souza, V.M.A. and Batista, G.E. (2019) ‘Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model’, *Information sciences*, 484, pp. 302–337.

Rathipriya, R. *et al.* (2023) ‘Demand forecasting model for time-series pharmaceutical data using shallow and deep neural network model’, *Neural Computing and Applications*, 35(2), pp. 1945–1957. Available at: https://doi.org/10.1007/s00521-022-07889-9.

Renaud, O. and Victoria-Feser, M.-P. (2010) ‘A robust coefficient of determination for regression’, *Journal of Statistical Planning and Inference*, 140(7), pp. 1852–1862.

Zhang, D. (2017) ‘A coefficient of determination for generalized linear models’, *The American Statistician*, 71(4), pp. 310–316.

## References for code used within implementation

# Appendix