LUNDS UNIVERSITET

Lund University

Department of Economics

Forecasting energy consumption in Sweden

Can Machine Learning outperform standard techniques?

Author

Jonathan Ferdinand-Dreyfus

Supervisor

Luca Margaritella

Bachelor's thesis

June 2022

Acknowledgements

I would like to express my gratitude to my supervisor Luca Margaritella. Without his extensive expertise and supporting guidance this work would not have been possible.

Abstract

Machine learning has acquired a lot of attention in the economic forecasting literature in recent years. In this thesis we forecast Swedish energy consumption and compare the forecasting performance of a machine learning technique with that of more traditional time series models. In fact, the LSTM neural network is compared with ARIMA and VAR forecasts. We conclude that in our setting, while these newer techniques perform well under some conditions and are able to outperform the ARIMA forecast, they are not found to outperform the VAR model which remains the best modelling choice among those considered here.

Contents

Abbreviations

ML: Machine Learning

ANN: Artificial Neural Network

RNN: Recurrent Neural Network

LSTM: Long Short-Term Memory

ARIMA: Autoregressive (AR) Integrated (I) Moving Average (MA)

VAR: Vector Autoregression

RMSE: Root Mean Squared Error

MAPE: Mean Average Percentage Error

1 Introduction

This paper aims to compare benchmark time series forecasting models with a newly popularized machine learning model to forecast energy consumption in Sweden. We use the benchmark ARIMA and VAR models and compare the forecasting accuracy with that of a newly popularized machine learning technique, namely the LSTM neural network model. We examine the benefits as well as problems of ML in forecasting and evaluate this by using total energy consumption in Sweden as the variable to forecast. In many contexts, ML has shown superior prediction accuracy compared to traditional econometric methods, and the benefits of the ML approach have led to much interest in economics forecasting (Ghoddusi, Creamer, and Rafizadeh, 2019).

ML models also have their complications in economics applications. One aspect of ML models is that they generally require large datasets, which can pose an issue in areas such as macroeconomics. In such fields, observations are generally available in the 50-100 range, which in machine learning terms often is the bare minimum. We will evaluate empirically the LSTM neural network's ability to forecast a macroeconomic variable with a small number of observations using yearly data between 1971-2020. We will then compare the results to benchmark econometric models to examine whether this particular neural network can enhance economic forecasts.

The dependent variable was chosen due to its importance to the Swedish economy. As energy supply is essential for households, the manufacturing sector, and the transportation industry, precise predictions are crucial for policymakers. Sweden has committed to lowering its carbon footprint by joining the Paris agreement (United Nations, 2015) and pledging to be leading in Agenda 2030 and become the first fossil fuel-free welfare country in the world (Regeringskansliet, 2015). One way Sweden has reduced fossil fuel usage is by shifting energy production towards solar and wind (Regeringskansliet, 2018). Change in the structure of energy production requires an accurate forecast of future energy demand so that households and industries can continue to have access to their required energy supply.

The thesis is organized as follows: Chapter 2 outlines the research which has previously been made in energy consumption forecasting, as well as some theoretical background.

Chapter 3 explains the methodology behind the three models used in this paper and Chapter 4 explains the variables used in the models. Chapter 5 then presents the results of the models and Chapter 6 discusses the results of the different models.

2 Previous Research

In this section, an overview of previous research in forecasting energy-related variables is presented. We focus on applications using ML as well as the benchmark ARIMA and VAR models. The methodology of these will be presented in Section 3.

2.1 Energy consumption forecasting, benchmark models

2.1.1 ARIMA

ARIMA is a benchmark model in time series forecasting, especially in univariate forecasting. The ARIMA model was generalized by Box and Jenkins who developed a method for time series forecasting using the model (further explained in section 3.1) (Zivot and Wang, 2006). The application of the ARIMA model has since seen a great increase in its applications. It has in forecasting become a standard benchmark method (Studenmund and Johnson, 2016) and has also been used in energy-related forecasting.

Nichiforov, Stamatescu, Fagarasan & Stamatescu (2017) compare energy consumption forecast results using the ARIMA model and a non-linear Autoregressive Neural Network (NAR) model and find that the ARIMA model outperforms the NAR model. Jahanshahi, Jahanianfard, Mostafaie, & Kamali (2019) were able to predict household energy consumption in the Euro-area using the Box-Jenkins methodology. Elsaraiti, Musbah, Merabet & Little (2021) apply the ARIMA model to electricity consumption and were able to accurately predict energy consumption with high accuracy. Dritsaki, Niklis, & Stamatiou (2021) forecast oil consumption in Greece using the ARIMA model and find that the oil consumption in the country will decrease after 2020 following the coronavirus as well as regulatory measures to decrease oil use. These papers are part of a wide range of academic research which successfully forecasts energy-related variables using the ARIMA framework.

2.1.2 VAR

The use of VAR in macroeconomic forecasting was popularised after the work published in 1980 by Cristopher A. Sims. The then-new framework of VARs was provided by Sims as a critique of the models used in that period, which Sims claimed illogically excluded key aspects of data analysis (Sims, 1980). Sims then proposed the VAR model which incorporates lagged values of the dependent variable as well as lagged values of the other variables. Since its introduction by Sims, the VAR has continued to provide a systematic and straightforward approach to forecasting time series and evaluating economic models (Christiano, 2012). The use of the VAR model has also found application in energy economics, as many works have found this approach robust and able to increase forecasting accuracy.

Jin & Chen (2013) apply the VAR model to energy consumption in China and find that total energy supply, improvements in living standards, and economic growth contribute to the country's energy consumption. The authors also conclude that economic growth has a smaller impact than the other two. Singh & Vashishtha (2020) apply a bivariate VAR model and conclude that per capita GDP impacts per capita energy consumption, however, any long-run equilibrium between the two variables could not be found. Yu & Qayyum (2022) use a Panel VAR (PVAR) and find a directional cause from GDP per capita on energy consumption in industrialized countries. Interestingly, the paper found a unidirectional relationship *from* energy consumption *to* GDP per capita in non-industrialized countries, implicating that non-industrialized countries rely on energy consumption for economic development.

2.2 Energy demand forecasting using ML and LSTM models

Machine learning (ML) is a type of data science where an algorithm learns from a dataset and is able to automatically improve itself, i.e. the "machine" is learning. Machine learning emerged in the 1950s as an attempt from the scientific community to replicate human learning behavior. The applications of machine learning models in the last decades have shown to be serious competitors to traditional forecasting models, especially for processing and forecasting complex data. Therefore, ML has grown increasingly useful in

energy economics and the energy industry (Ghoddusi, Creamer, & Rafizadeh 2019).

In one of the first success stories of price prediction in energy economics Moshiri & Foroutan (2006) were able to accurately predict crude oil prices using an Artificial Neural Network (ANN), and the model has since then expanded to being used in different areas. Lu, Sun, Duan & Wang (2021) published an extensive paper comparing several different variable selection and forecasting methods and finding that an LSTM forecasting model outperformed other competing models in crude oil price forecasting. Papadimitriou, Gogas & Stathakis (2014) were able to predict electricity prices with 76% accuracy over a 200-day period using a type of ML called a Support-vector machine.

ML techniques have also been used in predicting energy consumption and energy demand. Geem and Roper (2009) employ an ANN to forecast energy consumption in South Korea and find that their approach outperforms standard linear and nonlinear regressions. Ozturk and Ceylan (2005) use a Genetic Algorithm to forecast electricity consumption in the Turkish industrial sector. Ghoddusi, Creamer, and Rafizadeh (2019) conclude in a review of 130 published papers between 2005 and 2018 that ML is foremost used in energy economics for forecasting prices, and mostly the price of crude oil and power. Li (2019) used an LSTM neural network to estimate energy consumption in China.

3 Method

3.1 ARIMA

The ARIMA model is a univariate model frequently used in statistics and econometrics, which effectively forecasts future values of a variable by only using past values of that variable. The ARIMA model is a composite process that builds upon the past values of the dependent variable, a differencing process in order to achieve stationarity, and past values of the error terms (Kennedy, 2008). The ARIMA(p, d, q) model is explained in this section.

The *autoregressive* (AR) part of the ARIMA model states that the dependent variable y_t depends on earlier values of itself, y_{t-p}. The model uses the earlier (lagged) values of

the dependent variable and bases the forecasting on these earlier data points. The AR(p) process is formulated as:

$$y_t = \delta + \theta_1 y_{t-1} + \theta_2 y_{t-2} + \ldots + \theta_p y_{t-p} + \epsilon_t \tag{1}$$

Where δ is a constant, θ is the autoregressive coefficient, ϵ_t is the white noise error term. How many earlier values of y_t to include in the AR model, i.e. the amount of lags, is denoted by p. A model with no past values of y_t is an AR(0) process and is modeled as: $y_t = \delta + \epsilon_t$. In an AR(0) process only the error term contributes to the value of the dependent variable and the process is therefore white noise (Studenmund and Johnson, 2016). An AR(1) process includes one lagged period and is therefore modeled as:

$$y_t = \delta_0 + \theta_1 y_{t-1} + \epsilon_t \tag{2}$$

The I part in the ARIMA model stands for *Integrated* and denotes the fact that the data has been differenced in order to achieve stationarity. A time-series differenced once is denoted by equalizing d to 1 in the ARIMA(p, d, q) model (Kotu and Deshpande 2018). A time-series variable is stationary if the mean and variance of the variable are constant over time, and the autocovariance of two observations in time only depends on the length between observations (Studenmund and Johnson, 2016).

The *moving-average process* (MA) part of the ARIMA model states that the dependent variable y_t is defined as a linear combination of past values of the white noise error term ϵ_t. In the MA(q) process the dependent variable y_t is a function of past and current values of the random error term, and the amount of previous error terms included in the estimation is denoted by q. The MA(q) process is formulated as:

$$y_t = \epsilon_t + \Phi_1 \epsilon_{t-1} + \Phi_2 \epsilon_{t-2} + \ldots + \Phi_q \epsilon_q \tag{3}$$

Where ϵ_t is the error term, or random noise, Φ is the moving average coefficient (Kotu and Deshpande, 2018).

These three components, the AR(p), I(d), and MA(q) process create the ARIMA(p, d, q) model. For example, the ARIMA(1,1,2) model is shown below:

$$y_t^* = \delta_0 + \theta_1 y_{t-1}^* + \epsilon_t + \epsilon_{t-1} + \epsilon_{t-2} \tag{4}$$

We can observe that the dependent variable is once-differenced, denoted by *, and the model includes one autoregressive term and two moving-average terms as well as a constant. We will follow the Box-Jenkins methodology for an ARIMA model. It consists of (1) applying necessary transformations to achieve stationarity, (2) identifying the autoregressive component p and the moving average component q, (3) fit the time series ensuring that there is no autocorrelation in the residuals, (4) forecast future values of the variable (Kennedy, 2008).

3.2 VAR

The VAR model was popularised by Sims and has shown to be one of the more effective models in multivariate financial and economic time-series forecasting. The VAR model extends upon univariate autoregression and allows for multiple time-series variables to be included in the model with the lagged values of these variables as regressors (Hashimzade & Thornton, 2013). The VAR(p) model is used in this paper as a multivariate benchmark model to be used in comparison to the other models presented.

A simple one-lag univariate AR model would take the form $y_t = \theta_1 y_{t-1} + \epsilon_t$ where the current value of y depends on the previous value of y multiplied with the autoregressive coefficient θ and an error term (Kotu and Deshpande, 2018). However, it is possible that the value of y_t is dependent on other exogenous variables. Because of the possibility of two or more variables having a causal effect on each other in both directions, we can create a VAR(p) model with several variables, where p denotes the number of lags to include in the estimation. For example, a VAR(p) model with two variables y_t and x_t is formulated as follows:

$$\begin{aligned} y_t &= \delta_1 + \theta_{11} y_{t-1} + \ldots + \theta_{1p} y_{t-p} + \gamma_{11} x_{t-1} + \ldots + \gamma_{1p} x_{t-p} + \epsilon_{1t} \\ x_t &= \delta_2 + \theta_{21} y_{t-1} + \ldots + \theta_{2p} y_{t-p} + \gamma_{21} x_{t-1} + \ldots + \gamma_{2p} x_{t-p} + \epsilon_{2t} \end{aligned} \tag{5}$$

The VAR(p) model can also be expressed in matrix form (Zivot and Wang, 2006). A bivariate VAR(1) model is formulated as:

$$\begin{bmatrix} y_t \\ x_t \end{bmatrix} = \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix} + \begin{bmatrix} \theta_{11} & \gamma_{12} \\ \theta_{21} & \gamma_{22} \end{bmatrix} \begin{bmatrix} y_{t-1} \\ x_{t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix} \tag{6}$$

These equations show that in the VAR model framework both variables are endogenous and the matrix formulation shows that the addition of an additional variable will generate one additional equation. The coefficients θ and γ are estimated using OLS and therefore the Gauss Markow assumptions on ϵ_1, ϵ_2, need to be satisfied. The number of lags included i.e. how many values from past years to include in the model, are determined using information criteria (IC). The three most common IC are AIC, BIC, and HQ. These test specified lags $p = 0, 1, ..., p_t$ and then select the lag which minimizes one of the selection criteria (Zivot and Wang, 2006). The mathematics of these is beyond the scope of this paper and will not be further discussed here.

The benefits of VAR models over univariate models as the ARIMA model is that the model can consider the interconnectedness of several variables over time (Kim, Shim and Park, 2022), which has led to the model becoming one of the most used tools in applied finance and macroeconomics (Miao, Phillips and Su, 2022).

3.3 LSTM neural network

The LSTM neural networks have in recent years become a popular neural network system in Artificial Intelligence and ML. The network is highly useful in tasks such as classification, processing, and time series prediction. The LSTM network builds upon a special type of Artificial Neural Network (ANN) called Recursive Neural Network (RNN) (Yu, Si, Hu and Zhang, 2019). The mathematical aspects of ANNs, RNNs, and LSTMs are beyond the scope of this paper and will not be discussed in detail. Rather, this section aims to give the reader intuition of the basic structure of these algorithms.

Artificial Neural Networks:

An Artificial Neural Network is a computer algorithm inspired by biology. The structure of an ANN simulates human brain processes by mimicking the way that neurons in a

brain signal each other. The structure of an ANN consists of at least three layers: 1) an input layer, 2) one or more "hidden" layers, 3) an output layer. In each of these layers, there are nodes (imitated neurons), which connect all the layers together (Gibbs et al., 2006). The nodes are connected to each other by weighted links. If the information sent to a neuron from another is important enough to pass a certain threshold, the data is passed along. Otherwise, it doesn't. If the information passes, the node has a weight that adjusts the information that flows to the next node (Vanneschi and Castelli, 2019).

The network learns and improves by comparing its output to the actual observed values and adjusting the weights of its neurons accordingly. The errors are sent back through the hidden layer[1] and every node changes its own weight depending on how "wrong" it was, and the procedure is then repeated with a new observation. The number of times the data is sent through the network is called *epochs* and for every epoch, the weights of the neurons are adjusted to minimize the errors. Neural networks generally use the sigmoid function to scale inputs between 0 and 1, by doing this they minimize the impact of a single variable on the output of the neural network (Dreyfus, 2005). A typical ANN with one hidden layer is illustrated in Figure 1.

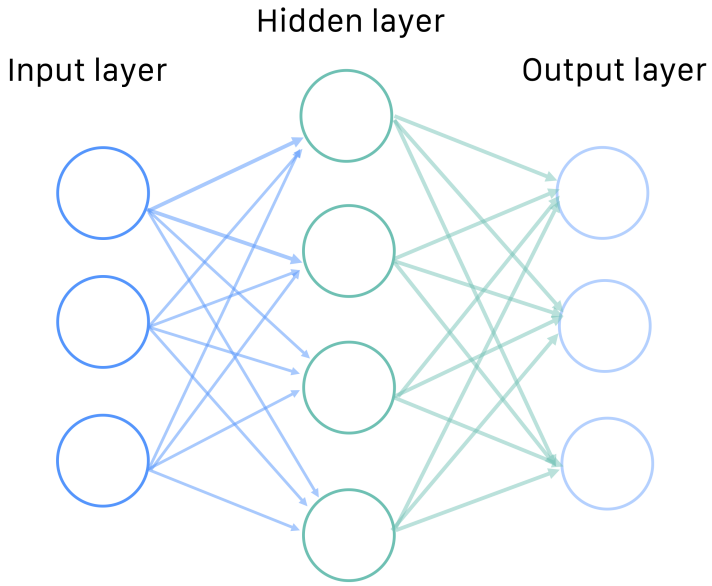


Figure 1: Artificial Neural Network

Recursive Neural Networks:

A Recursive Neural Network builds upon the idea that the quality of processing informa-

[1]This process is called backpropagation

tion is improved when recalling earlier inputs of that information. Just as humans are able to watch a movie and process a frame depending on the frames that came earlier, RNNs try to mimic this process by including long-term memory. Traditional ANNs do not remember information from earlier processes, thus lowering their efficiency compared to RNNs. The hidden layers in a RNN acts as memory storage and transmit information from earlier processing of the data in the sequence. The hidden layer cell takes an input x_t and from the neural network transmits this information to the next cell (Dreyfus, 2005), as illustrated in Figure 2. Every A denotes an artificial neural network as the one illustrated in Figure 1.

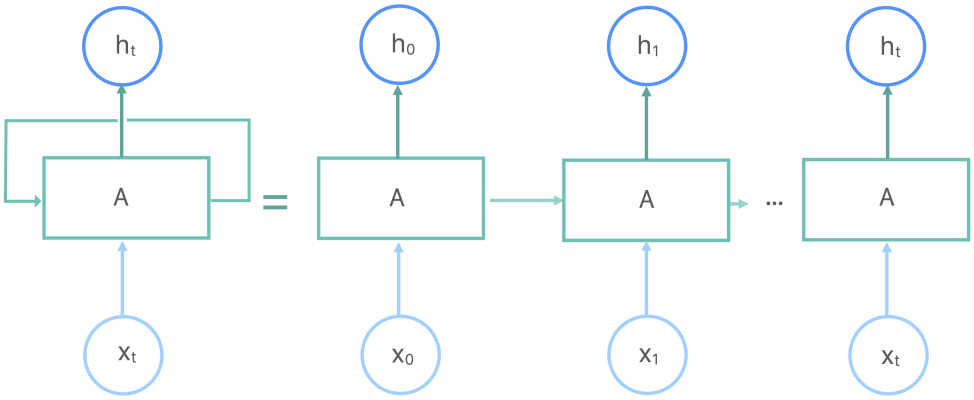


Figure 2: Recurrent Neural Network Loop

RNNs mainly have two issues that arise in the transmission of information between cells. The first issue is that the RNN only remembers a few steps back in the data sequence and is therefore not a good fit for longer sequences of data with long memory. The second issue is that of exploding and vanishing gradients. As inputs flow through the memory cells of the network the weight of an input can increase or decrease with every iteration and therefore "explode" or "vanish" (Schmidt, 2019). LSTMs were envisioned to address this issue and have shown success in its applications.

LSTM:

The LSTM neural network aims to resolve the two issues related to RNNs. The LSTM network adds additional features in every neural network, denoted by A in Figure 2, to

handle the transmission of data between cells. Three gates are implemented in every cell and the data has to pass through this gate in order to continue on to the next cell. The main addition of LSTMs can be described as a transport line, passing data from one cell to the next cell only if that data is found to be significant enough to pass through the gates (Lu, Sun, Duan and Wang, 2021). The gates are based upon the sigmoid function, a central function in ML which outputs a number between 0 and 1 and has the equation:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{7}$$

The sigmoid function reduces all real numbers $\mathbb{R}$ into the range $(0, 1)$ which makes processing data between cells in the neural network easier (Liu, 2021). The three gates added in every cell using the sigmoid function are:

- A Forget Gate which outputs a number between 0 and 1. An output of 1 in this gate means "let everything through" and a 0 means "let nothing through".
- A Memory Gate which chooses which new data needs to be stored in the cell.
- An Output Gate which decides what will be the output of each cell.

The LSTM network is structured as a transport line, where the data moves on the transport line and passes through cells that can filter and change the data, completely stop the data from continuing on the transport line, or let all the data through. The previously processed data flowing on the memory line is then remembered and processed to make predictions (Al-jabery, Obafemi-Ajayi, Olbricht and Wunsch II, 2020). The LSTM model has several parameters which can be tuned to increase accuracy, such as *Layers, Batch size, Epochs*, which are explained in Table 1 (Dreyfus, 2005) (Li, 2019). The parameters chosen in the neural network for this paper are explained in Table 9.

Parameter	Explanation
Layers	Layer of nodes between input and output layer, where each node has an assigned weight
Loss	Function chosen to minimize loss or maximize gain, ex. Mean Squared Error Loss
Optimizer	Function which changes the learning attributes of neural networks
Epochs	Each time the *entire* dataset is passed through the network layers counts as one epoch
Batch size	Nº observations from the dataset which passes trough the network at a time before updating weights
Activation	Function used to determine the output of the neural network, giving the output in the range (0,1) or (-1,1) as an example

Table 1: LSTM model parameters

3.4 Evaluation metric

For the comparison of the three different forecasting models, the RMSE and MAPE will be used. Both metrics are frequently used to evaluate model prediction accuracy. The equations are:

$$\begin{aligned} RMSE &= \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \\ MAPE &= \frac{1}{n}\sum_{t=1}^{n} 100\frac{|y_t - \hat{y}_t|}{|y_t|} \end{aligned} \tag{8}$$

Where n is the number of time periods observed, y_t is the actual value and $\hat{y}_t$ is the predicted value from the model. The RMSE measures the differences between the actual values and the predicted values; the square root penalizes large errors. The MAPE measures the percentage error for every predicted value (standardizes unit of measurment) (Ji and Gallo, 2006). The model with the lowest RMSE and MAPE is then selected as the most efficient.

3.5 Time series properties

3.5.1 Cointegration and Stationarity

Before we can begin creating our models we have to test for stationarity. A time series variable is covariance stationary if:

1. The mean of the variable is constant over time, $E[y_t] = \mu \forall t$
2. The variance of the variable is constant over time, $Var(y_t) = \sigma^2 \forall t$
3. The autocovariance function between two observations separated by time k, $Cov(y_t, y_{t-k})$, only depends on k

(Studenmund and Johnson, 2016). A time series that does not fulfill such requirements is non-stationary and has trends (Zivot and Wang, 2006). Testing for the presence of unit roots in the series is fundamental to avoiding spurious regression issues. If no unit root is found in the series, they are said to be stationary (Studenmund and Johnson, 2016). To achieve stationarity one has to apply differencing, which consists of taking the difference of the variables of interest (Kotu and Deshpande, 2018):

$$y_t^* = \Delta y_t = y_t - y_{t-1} \tag{9}$$

If the time series is not stationary, then any inference or regression runs the risk of spurious correlation. Spurious correlation occurs when two variables that do not have any real underlying correlation show a strong relationship because of their shared drift in time (Smith, 2015).

However, if we have non-stationary variables we could have cointegration which could avoid us taking the differences of the series altogether, but we need to test this. Two variables that are cointegrated share a trend. If cointegration is found between two or more variables the variables should not be differenced, as they do not run the risk of spurious correlation and we can consistently fit our model (Kennedy, 2008).

3.5.2 Autocorrelation in residuals

For the time series to be correctly modeled, the residuals should be white noise i.e. have a constant mean and variance. Testing for autocorrelation is done to ensure that there is no misspecification in the model (Cryer and Chan, 2008). We test for autocorrelation in this paper by using the Ljung-Box test, a type of Portmanteau test.

4 Data and Software

4.1 Variables

The variables used in this paper are presented in Table 2 and justification of the selected variables is presented in 4.1.1. All variables are measured in a yearly frequency and the observations are for the years between 1971-2020.

Variable name	Description and measurement	Source
Energy consumption	Energy Consumption in Petajoules, Sweden	International Energy Agency & Swedish Energy Agency
Real GDP	GDP in 2015 US$, Sweden	World Bank national accounts
Population	Swedish Population, all residents	United Nations & Statistics Sweden
Energy intensity	Energy consumption per unit of gross domestic product, Sweden	Our World in Data
Oil Price	Crude Oil Price ($ per barrel), US	U.S. Energy Information Administration
CPI	Consumer Price Index Energy Products, Sweden	OECD

Table 2: Variable description

Energy consumption is the variable being predicted by the different models. It is measured in Petajoules. One Petajoule is equal to 31.6 million m^3 of natural gas or 278 million

kilowatt-hours of electricity. The data is obtained from the International Energy Agency (IEA) and encompasses the total energy used (industry, household, transportation, etc.) of all energy types (coal, oil, electricity, etc.) in Sweden per year (IEA, 2021) (Swedish Energy Agency, 2021).

Real GDP is the value of all goods and services produced in Sweden each year. The variable is measured in 2010 dollars and the dataset is obtained from the World Bank (World Bank national accounts, 2022).

Population is data on the total population of Sweden regardless of resident status. The data is obtained from the United Nations Population Division (United Nations, 2019) (Statistics Sweden, 2022).

Energy intensity is included as an indicator of the technological advancements and effectiveness of the country's energy infrastructure. As energy use decrease with more energy-efficient technology in manufacturing, transportation, etc. the inclusion of this variable could have a significant relationship with total energy consumption. In fact, it measures the energy consumed per unit of GDP, thus measuring how effective a country is in producing economic output. It can also represent industry structure, as a shift in industry from manufacturing to services can decrease energy use but maintain or increase economic output (Our World in Data, 2022).

Oil Price is measured as the price per barrel of crude oil traded in the US. Crude oil prices for Sweden are not available for the time period in this paper, but we can assume that prices in the US and Sweden will follow a similar pattern. As oil is frequently used in manufacturing and heating, an increase in oil prices can have a significant connection with energy consumption (U.S. Energy Information Administration, 2022).

Consumer Price Index (CPI) measures the increase in consumer products related to energy, such as an increase in heating oil and electricity prices. This variable is included as increases in energy products could cause consumers to consume less energy and thus lowering aggregate energy consumption. The price of consumer energy products is assumed to be closely related to production-side prices of energy products and can therefore represent both consumer and manufacturing energy product prices (OECD, 2022).

4.1.1 Variable choice

The selection of relevant variables for energy consumption is central to the multivariate models used in this paper. Here we draw inspiration from other papers' conclusions on relevant variables to explain energy consumption and focus on the case of Sweden. Several papers have found a causal relationship between population and economic output on energy consumption in developed and developing countries. Li (2019) found that GDP, population, secondary sector, and tertiary sector were significantly linked to energy consumption in China. Geem and Roper (2009) found that gross domestic product (GDP), population, import, and export amounts were significantly linked to energy consumption in South Korea. Camarero et al. (2015) applied a variable selection model to energy consumption determinants and found that variables such as economic growth, energy prices, government spending, energy efficiency, and source of energy production all contribute to explaining energy consumption.

Various potential variables had to be excluded due to them not being available for the time period researched. Data on industry sectors is only available from 1981 onward. In an attempt to address any omitted variable bias this paper included variables such as *Energy Intensity* and *CPI*, which are indices representing various parts of the economy (explained in Section 4.1). We perform a Granger causality test to aid this paper in understanding which variables have predictive power with respect to energy consumption. The test found that *Real GDP, Energy Intensity, Oil Price, and CPI* all have a significant predictive relationship on energy consumption and are therefore used in the VAR and LSTM models. Table 3 clarifies which variables were used for which model.

ARIMA	VAR	LSTM
Energy consumption	*Energy consumption*	*Energy consumption*
	Real GDP	*Real GDP*
	Energy Intensity	*Population*
	Oil Price	
	CPI	

Table 3: Variables in models

4.2 Software

For this paper the statistical open-source software R was used, version 4.2.0 (2022-04-22). Some of the packages used in the paper for standard operations were *tidyverse, readxl, tseries, dplyr,* and *graphics*. For the ARIMA and VAR modeling the packages *forecast* respectively *vars* was used. The package necessary for the LSTM neural network is the *keras* package (R Core Team, 2021).

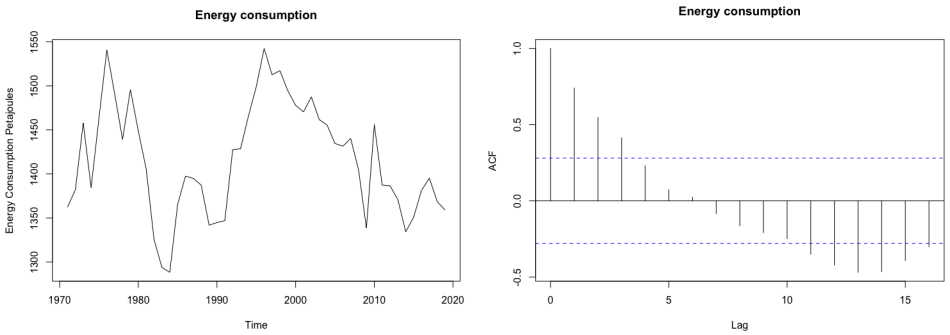
5 Results

5.1 ARIMA model

ARIMA is used in this paper as the benchmark univariate model for comparison with other models. We evaluate its accuracy in predicting the dependent variable *Energy consumption*. In forecasting, we follow the Box-Jenkins method for ARIMA forecasting.

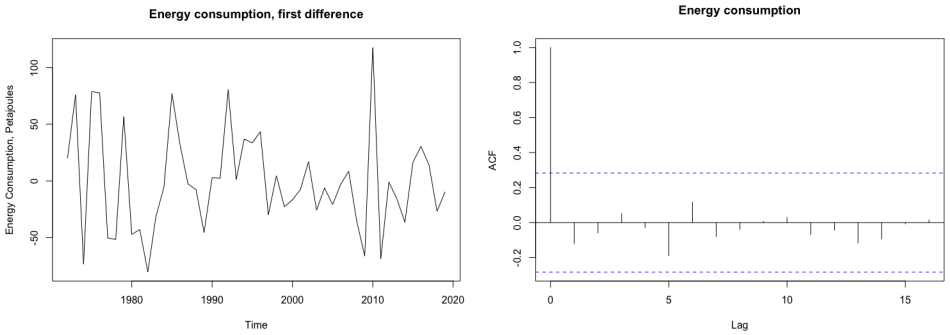
(1)*Stationarity*: To ensure stationarity we test the *Energy consumption* variable for unit root. The unit root test was performed by using the Augmented Dickey-Fuller test at the 5% level which tests the null hypothesis that the data is nonstationary. After performing this test we find the presence of a unit root in *Energy consumption*. After taking the first difference of the dataset we can state that we have removed the unit root

and made the time series stationarity. The autocorrelation observed in the time series has been accounted for by taking the first difference, as shown in Figure 3 and Figure 4.



(a) Total Energy Consumption (b) Autocorrelation Function

Figure 3: Energy Consumption 1971-2020



(a) Energy consumption (b) Autocorrelation Function

Figure 4: Energy Consumption, first difference 1972-2020

ADF-test after differencing, *Energy Consumption*	
Dickey-Fuller statistic	p-value
-3.5269	0.04844^*
H_0:	Unit root is present (non-stationary)
H_a:	Unit root is not present (stationary)

Table 4: Augmented Dickey-Fuller test, *Energy consumption*

(2) *Identifying the required parameters*:

The Box-Jenkins approach involves fitting the number of autoregressive lags p and the number of moving averages q. These components can be determined by analyzing the

Autocorrelation Function (ACF) and the Partial Autocorrelation Function (PACF), but we can automatically obtain these parameters in the auto.arima function in the R package "forecasts". The auto.arima function uses a variation of the Hyndman-Khandakar algorithm (Hyndman & Khandakar, 2008) to estimate the model parameters. The function finds that ARIMA$(3, 1, 0)$ is the optimal fit for the dataset, meaning that the model includes 3 autoregressive terms, the time series is once differenced, and no moving averages are included. Our model is thus fitted with the equation:

$$y_t^* = \delta + \theta_1 y_{t-1}^* + \theta_2 y_{t-2}^* + \theta_3 y_{t-3} + \epsilon_t \tag{10}$$

(3) *Evaluating the residuals*:

To ensure that the ARIMA model is appropriate for the data, we need to evaluate the residuals. It is necessary that the residuals are uncorrelated and normally distributed with a constant mean and variance. For the residual diagnostic, we use the plots of the ACF and PACF to qualitatively identify any autocorrelation, then we quantitatively test for autocorrelation using the Ljung-Box Q-test.

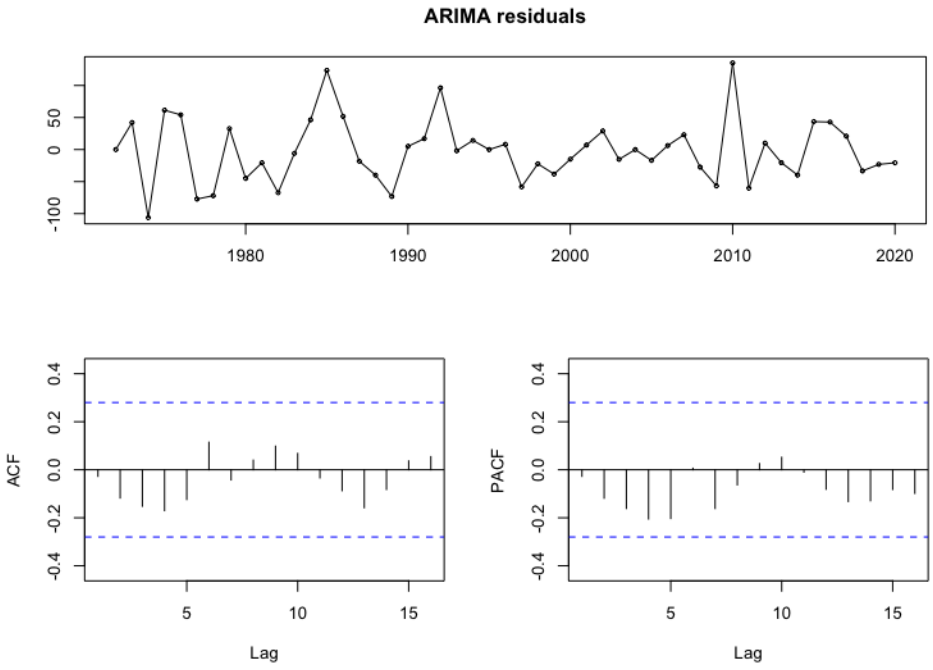


Figure 5: Residual diagnostics ARIMA

We can observe in Figure 5 that there is no autocorrelation as more than 95% of obser-

vations fall between the bounds $\pm 1.96/\sqrt{n}$. We have therefore ensured homoscedasticity in the model and can therefore trust the results to be unbiased. We can also test for autocorrelation by using the Ljung-Box Q-test at the 5% level, shown in Table 5. The Ljung-Box test generates a Q test statistic and is used for ensuring that the residuals are evenly distributed. A large value of Q signals that the autocorrelation in the sample data is too large for the data to be independent and identically distributed (Cryer and Chan, 2008).

Ljung-Box Test		
Q*	DF	p-value
6.3839	7	0.4957
H_0:	Residuals are evenly distributed	
H_a:	Residuals show signs of autocorrelation	

Table 5: Ljung-Box test, residuals

We have now satisfied the stationarity requirement, estimated the model parameters (p, q), and ensured the absence autocorrelation in the model residuals.

(4) *Forecasting*

The fourth step in the Box-Jenkins methodology is forecast. For the comparison between models, we have chosen the time frame between 2005 and 2020. We evaluate the ARIMA models' effectiveness by obtaining its RMSE and MAPE. The evaluation metrics are obtained by comparing the actual values to the estimated values for the period 2005-2020 and presented in Table 6.

ARIMA Model evaluation	
RMSE	46.819
MAPE	2.063

Table 6: ARIMA evaluation metrics

The ARIMA model forecast for the time period is presented in Figure 6, where the forecasted values are illustrated as the red line and the actual values are presented as the black line (the forecasted values for the other models are presented in the same way).

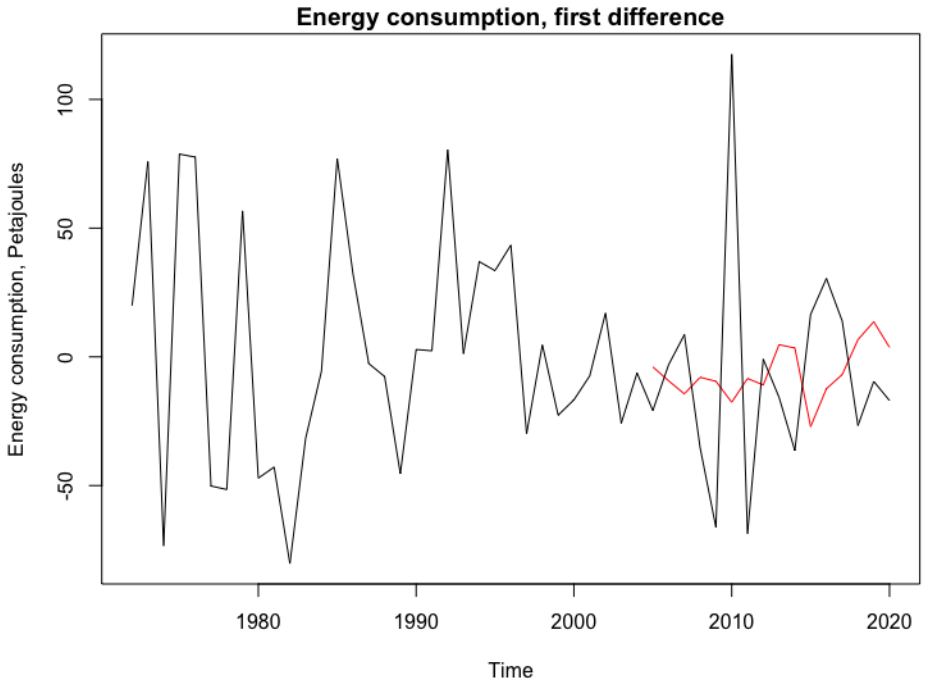


Figure 6: ARIMA prediction

5.2 VAR model results

The VAR model used in this paper contained 5 variables, *Energy Consumption, Real GDP, Energy Intensity, Crude Oil Price, and CPI*. We test the variables for a unit root. The p-value at the 5% level of the Augmented Dickey-Fuller test for the non-stationary variables is presented in Table 7. The variables are non-stationary but if they are cointegrated we can still fit our model. Because of this, we test for cointegration by using the Johansen test (Johansen, 1988). We find no evidence of significant cointegration between the time

series variables in the Johansen test. The variable *Energy Intensity* had no unit root, but the variable exhibited autocorrelation and was therefore differentiated.

Augmented Dickey-Fuller test

	Levels	Δ
Energy consumption	0.4321	0.04844^{*}
Real GDP	0.7589	0.04306^{*}
Crude Oil Price	0.4495	$< 0.01^{**}$
CPI	0.6501	0.02197^{*}

H_0: Unit root is present (non-stationary)

H_a: Unit root is not present (stationary)

Table 7: Augmented Dickey-Fuller test results

In the table, we can see that we have successfully achieved stationarity by taking the first-order difference, and we can now construct our VAR model. As mentioned in section 3.2 we will select the number of lags to include by using the most common selection criteria. The number of lags selected by the AIC and HQ selection criteria is 7 lags, which means that energy consumption is estimated by using the 7 lagged values of the variables. We fit the model by using the VAR() function from the "vars" package in R which estimates the model using OLS (Lütkepohl, 2006). The results from the VAR(7) model show that the model can very accurately predict energy consumption. The evaluation metrics and graphic illustration are presented in Table 8. In Table 8 "Levels" denotes that the time series is not differenced, and the symbol Δ denotes that the series has been differenced once.

VAR model evaluation	
RMSE	8.565
MAPE	0.489

Table 8: VAR evaluation metrics

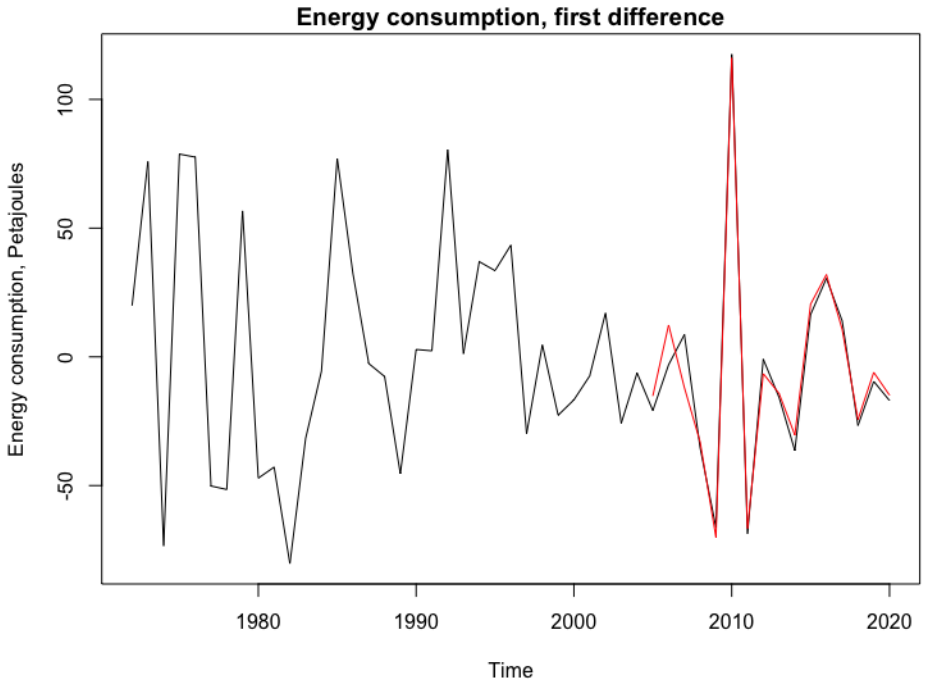


Figure 7: VAR prediction

As stated in Section 3.2 the VAR model produces one equation for every variable in the model and incorporates lagged values of every variable in the equation. Apart from forecasting, we can also analyze which variables at which lag was significant for determining energy consumption. The variables significant at the 5% level were *Oil Price L1* and *CPI L6*. The variables significant at the 10% level were *Real GDP L3, L5, L6, L7* and *Oil Price L6* (the number after the variable denotes the lag order for the variable). We analyze the model residuals by using a Portmanteau test to ensure that we have correctly fitted the model. We find no autocorrelation in the residuals at the 5% level in the Portmanteau test.

5.3 LSTM neural network

In the prediction for energy consumption using the LSTM network, we find that the network most accurately predicts using three variables, *Real GDP*, *Population*, and one lagged value of *Energy consumption*, i.e. predicting today's energy consumption using the value of yesterdays energy consumption. Even though *Population* was omitted in the VAR model, we find that including the variable improves the accuracy of the LSTM forecast. The tuning of the LSTM model also showed that including any additional explanatory

variables in the model overfits the model and worsens the prediction accuracy. We split the dataset in a 70/30 window, where we use 70% of the observations for training the model and 30% of the observations for evaluating the model fit. For machine learning algorithms to properly function different types of scaling are often used to improve the accuracy of the model (Dreyfus, 2005). We proceed by normalizing the data.

After normalizing and splitting our data we can then form our LSTM network where our input is *Real GDP*, *Population*, *Energy consumption (lagged)*, and our output is forecasted energy consumption. With ARIMA and VAR models any time series data needs to be stationary, however, with LSTM this is not required as it can "remember" trends over time. Our forecast however shows higher accuracy when our time series is stationary and we, therefore, use the differentiated series. We create an LSTM neural network with the parameters presented in Table 9.

Parameter	Value
Layers	4
Loss	Mean squared error
Optimizer	adamax
Epochs	100
Batch size	7
Activation	Relu

Table 9: LSTM model parameters

The model inputs the training data 100 times (epochs) through the 4 layers, 7 observations at a time. The network then adjusts the weights and evaluates the result by mean squared error. For the reader interested in the process of the neural network, Figure 8 illustrates how the mean squared error (Y-axis) is reduced with every epoch (X-axis).

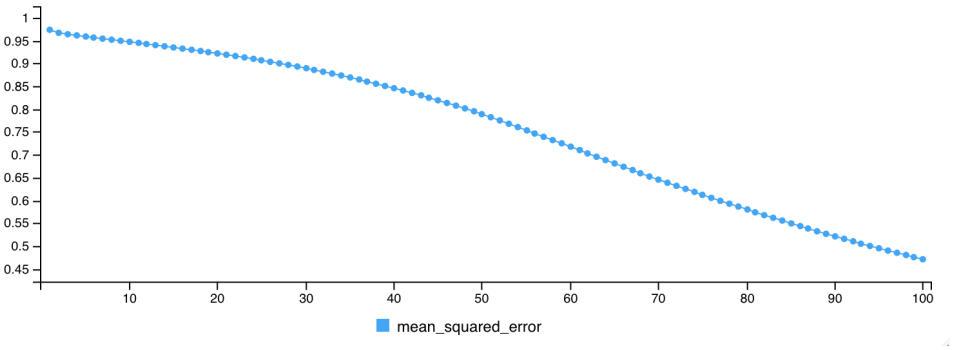


Figure 8: Accuracy improvement with every epoch

After adjusting the weights the model then predicts a value for the remaining years in the dataset and reverts the scaling. We use the predicted data and compare it to the actual values to obtain the evaluation metrics. The RMSE and MAPE for the LSTM predictions are presented in Table 10. The LSTM predictions are shown in Figure 9, where we can observe the neural networks' ability to accurately predict energy consumption using only two input variables and one autoregressive term.

LSTM network evaluation	
RMSE	26.616
MAPE	1.495

Table 10: LSTM evaluation metrics

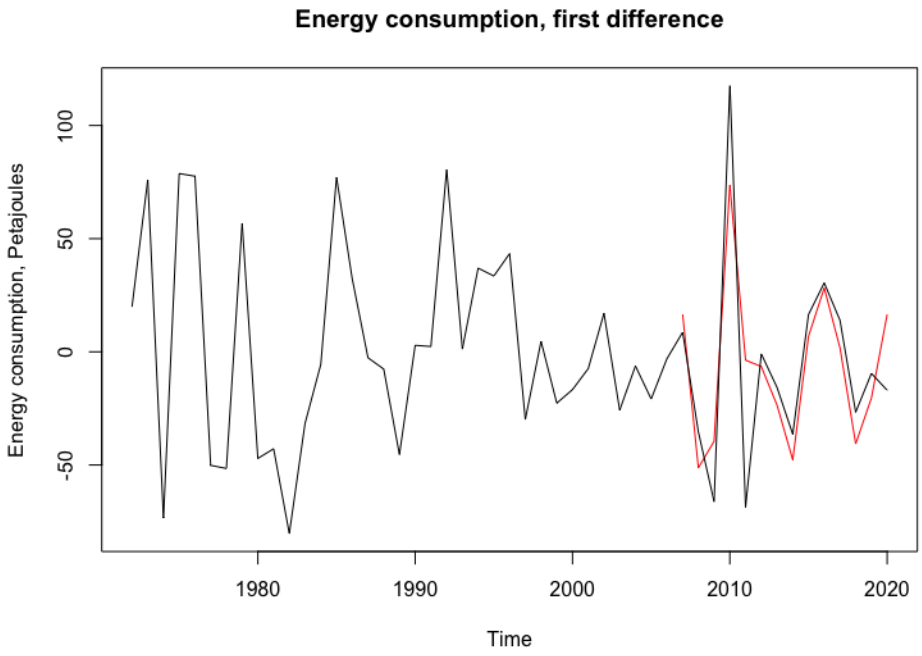


Figure 9: LSTM prediction

6 Concluding Remarks and Future Research

This paper set out to compare the forecasting accuracy of the ARIMA and VAR models and the newer LSTM neural network. While many might expect any model which employs ML to be the superior model in any aspect, this paper interestingly showed that this is not the case in energy consumption forecasting. The most accurate model in this paper showed to be the VAR model, which had better prediction accuracy by both evaluation metrics. While both the VAR and LSTM could predict and capture shocks, such as the large increase in energy consumption in the year 2010, the VAR model showed a much more refined ability to capture these shocks.

We forecasted energy consumption to evaluate how well the LSTM neural network could be fitted with only 50 observations, and we can conclude that this paper showcased how ML is dependent on large datasets as compared to traditional benchmark models. In the training phase of the LSTM model, we split the dataset of 49 variables in a 70/30 split. This means that only 34 observations for each input variable are available for the neural network to use for training. We can conclude that with these limited observations the LSTM network was nevertheless able to provide an accurate forecast. We can hypothesize that if the input variables were available for a longer time period, or on a more frequent basis, the LSTM model could increase its forecast accuracy.

The VAR model showed great prediction accuracy in this energy consumption forecasting. The model was able to accurately forecast the variable by using 5 variables and lagged values of itself. With these accurate results, the VAR model could be used by companies, governments, and organizations as a powerful tool for forecasting energy consumption and other forecasts. The VAR model also generates forecasts for every variable in the model. For example, we could also analyse how *Energy consumption* affects *Real GDP*, or how *Real GDP* affects *Energy intensity*. However, this is beyond the scope of this paper and was not presented in the model results.

The ARIMA model could not produce as accurate results as the other two models. This was at some level expected since the model only uses past values of itself and will therefore have issues in anticipating shocks. However, with only past values of itself the ARIMA model was able to adequately forecast energy consumption.

While ML is increasing in application in economics, it could not outperform the VAR model in this case. Future research could investigate whether this is due to the small number of observations, or whether the LSTM model is simply not as accurate as the VAR model for energy consumption forecasting. As both the LSTM and VAR models showed great forecasting accuracy, future research could expand upon the comparison between the two in other areas of economics.

7 Sources

Al-jabery, K., Obafemi-Ajayi, T., Olbricht, G. and Wunsch II, D., 2020. Selected approaches to supervised learning. *Computational Learning Approaches to Data Analytics in Biomedical Applications*, pp.101-123.

Camarero, M., Forte, A., Garcia-Donato, G., Mendoza, Y. and Ordoñez, J., 2015. Variable selection in the analysis of energy consumption–growth nexus. *Energy Economics*, 52, pp.207-216.

Christiano, L., 2012. Christopher A. Sims and Vector Autoregressions. *The Scandinavian Journal of Economics*, 114(4), pp.1082-1104.

Cryer, J. and Chan, K., 2008. *Time series analysis.* 2nd ed. New York, NY: Springer, pp.175-179.

Dreyfus, G., 2005. *Neural networks: methodology and applications.* 1st ed. Berlin [etc.]: Springer Science & Business Media.

Dritsaki, C., Niklis, D. and Stamatiou, P., 2021. Oil Consumption Forecasting using ARIMA Models: An Empirical Study for Greece. *International Journal of Energy Economics and Policy*, 11(4), pp.214-224.

Elsaraiti, M., Ali, G., Musbah, H., Merabet, A. and Little, T., 2021. Time Series Analysis of Electricity Consumption Forecasting Using ARIMA Model. *2021 IEEE Green Technologies Conference (GreenTech),*.

Geem, Z. and Roper, W., 2009. Energy demand estimation of South Korea using artificial neural network. *Energy Policy*, 37(10), pp.4049-4054.

Ghoddusi, H., Creamer, G. and Rafizadeh, N., 2019. Machine learning in energy economics and finance: A review. *Energy Economics*, 81, pp.709-727.

Gibbs, M., Morgan, N., Maier, H., Dandy, G., Nixon, J. and Holmes, M., 2006. Investigation into the relationship between chlorine decay and water distribution parameters

using data driven methods. *Mathematical and Computer Modelling*, 44(5-6), pp.485-498.

Hashimzade, N. and Thornton, M., 2013. *Handbook of research methods and applications in empirical macroeconomics*. 1st ed. Cheltenham: Edward Elgar Publishing Limited, pp.139-164.

Hyndman, RJ and Khandakar, Y (2008) "Automatic time series forecasting: The forecast package for R", *Journal of Statistical Software*, 26(3).

International Energy Agency (IEA). 2021. World Energy Balances. *International Energy Agency*. [Online] [Accessed 13 April 2022] Available from: https://www.iea.org/data-and-statistics/data-product/world-energy-balances

Jahanshahi, A., Jahanianfard, D., Mostafaie, A. and Kamali, M., 2019. An Auto Regressive Integrated Moving Average (ARIMA) Model for prediction of energy consumption by household sector in Euro area. *AIMS Energy*, 7(2), pp.151-164.

Jin, J. and Chen, Y., 2013. VAR-Based Research on Energy Consumption in China. *2013 International Conference on Computational and Information Sciences*,.

Ji, L. and Gallo, K., 2006. An Agreement Coefficient for Image Comparison. *Photogrammetric Engineering & Remote Sensing*, 72(7), pp.823-833.

Johansen, S. (1988), Statistical Analysis of Cointegration Vectors, *Journal of Economic Dynamics and Control*, 12, 231–254.

Kennedy, P., 2008. *A guide to econometrics*. 6th ed. Malden: Blackwell.

Kim, S., Shim, S. and Park, D., 2022. Dynamic interactions between trade globalization and financial globalization: A heterogeneous panel VAR approach. *Journal of International Money and Finance*, 122, p.102547.

Kotu, V. and Deshpande, B., 2018. *Data Science: Concepts and Practice*. 2nd ed. Cambridge: Morgan Kaufmann.

Li, Y., 2019. Prediction of energy consumption: Variable regression or time series? A case in China. *Energy Science & Engineering*, 7(6), pp.2510-2518.

Liu, H., 2021. Single-point wind forecasting methods based on deep learning. *Wind Forecasting in Railway Engineering*, pp.137-175.

Lu, Q., Sun, S., Duan, H. and Wang, S., 2021. Analysis and forecasting of crude oil price based on the variable selection-LSTM integrated model. *Energy Informatics*, 4(S2).

Lütkepohl, H., 2006. *New introduction to multiple time series analysis.* Berlin: Springer.

Miao, K., Phillips, P. and Su, L., 2022. High-dimensional VARs with common factors. *Journal of Econometrics*,.

Munshi, J., 2016. Spurious Correlations in Time Series Data: A Note.*SSRN Electronic Journal*,.

Nichiforov, C., Stamatescu, I., Fagarasan, I. and Stamatescu, G., 2017. Energy consumption forecasting using ARIMA and neural network models. *2017 5th International Symposium on Electrical and Electronics Engineering (ISEEE)*,.

Organization for Economic Co-operation and Development (OECD). 2022. Consumer Price Index: Energy for Sweden. *OECD*. [Online]. [Accessed 12 June 2022]. Available from: https://fred.stlouisfed.org/series/SWECPIENGMINMEI

Our World in Data, Energy Intensity. *Our World in Data* [Online]. [Accessed June 6 2022] Available from: https://ourworldindata.org/search?q=sweden+energy+intensity.

Ozturk, H. and Ceylan, H., 2005. Forecasting total and industrial sector electricity demand based on genetic algorithm approach: Turkey case study. *International Journal of Energy Research*, 29(9), pp.829-840.

Papadimitriou, T., Gogas, P. and Stathakis, E., 2014. Forecasting energy markets using support vector machines. *Energy Economics*, 44, pp.135-142.

Park, Y. and Lek, S., 2016. Artificial Neural Networks. *Developments in Environmental Modelling*, pp.123-140.

R Core Team, 2021. R: *A Language and Environment for Statistical Computing*, Vienna, Austria. Available at: https://www.R-project.org/.

Regeringskansliet, 2015. *Sveriges arbete med Agenda 2030.* [online] Available at: https://www.regeringen.se/regeringens-politik/globala-malen-och-agenda-2030/globala-mal-for-hallbar-utveckling/ [Accessed June 17 2022].

Regeringskansliet, 2018. *Handlingsplan Agenda 2030.* [online]. Regeringskansliet. Available at: https://www.regeringen.se/rapporter/2018/06/handlingsplan-agenda-2030 [Accessed June 17 2022].

Schmidt, R., 2019.*Recurrent Neural Networks (RNNs): A gentle Introduction and Overview.* Eberhard-Karls-University Tübingen,.

Singh, K. and Vashishtha, S., 2020. Does any relationship between energy consumption and economic growth exist in India? A var model analysis. *OPEC Energy Review*, 44(3), pp.334-347.

Smith, G., 2015. The Art of Regression Analysis. *Essential Statistics, Regression, and Econometrics*, pp.261-299.

Statistics Sweden. 2022. Befolkningsstatistik. *Statistics Sweden.* [Online]. [Accessed April 18 2022]. Available from: https://www.scb.se/hitta-statistik/statistik-efter-amne/befolkning/befolkningens-sammansattning/befolkningsstatistik/

Studenmund, A. and Johnson, B., 2016. *Using Econometrics: A Practical Guide.* 7th ed. Boston: Pearson, pp.456-460.

Swedish Energy Agency. 2021. Energy in Sweden 2021: An overview. *Swedish Energy Agency.* [Online]. [Accessed April 29 2022]. Available from: http://www.energimyndigheten.se/en/facts-and-figures/statistics/

United Nations, 2015. *Historic Paris Agreement on Climate Change.* [online] Available at:

https://web.archive.org/web/20160117141004/http://
newsroom.unfccc.int/unfccc-newsroom/finale-cop21/ [Accessed June 20 2022].

United Nations. 2019. World Population Prospects 2019. *United Nations*. [Online]. [Accessed April 18 2022]. Available from: https://population.un.org/wpp/

U.S. Energy Information Administration. Year. Crude Oil Prices: West Texas Intermediate (WTI). *U.S. Energy Information Administration*. [Online]. [Accessed April 29 2022]. Available from: https://fred.stlouisfed.org/series/DCOILWTICO

Vanneschi, L. and Castelli, M., 2019. Multilayer Perceptrons. *Encyclopedia of Bioinformatics and Computational Biology*, pp.612-620.

World Bank national accounts. 2022. World Development Indicators. *World Bank* [Online]. [Accessed April 13 2022]. Available from: http://data.worldbank.org/data-catalog/world-development-indicators

Yu, Y., Si, X., Hu, C. and Zhang, J., 2019. A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. *Neural Computation*, 31(7), pp.1235-1270.

Yu, Y. and Qayyum, M., 2022. Dynamics between carbon emission, imported cultural goods, human capital, income, and energy consumption: renewed evidence from panel VAR approach. *Environmental Science and Pollution Research*,.

Zivot, E. and Wang, J., 2006. *Modeling Financial Time Series with S-PLUS*. 2nd ed. New York, NY: Springer.