Short thesis for the degree of doctor of philosophy (PhD)

Developing Hybrid Forecasting Models for Large-Scale Heavy Manufacturing Industries

by Herry Kartika Gandhi

Supervisor: Prof. Dr. Márton Ispány

UNIVERSITY OF DEBRECEN
Doctoral School of Informatics
Debrecen, 2025

Contents

Chapter 1
Introduction

Large-scale heavy manufacturing industries are crucial for global economic growth. Fulfilling the needs of people in various sectors requires manufacturers to provide products to customers on time and in the right quantities. The supply chain's key areas are connectivity with suppliers, manufacturers, transportation, and logistics, which require balancing demands and supplies, raw material availability, and machine equipment readiness. Planning using accurate forecasting prevents shortages and excessive products and ensures a smooth and efficient supply chain (SC) process.

It is essential for industries to avoid overproduction to prevent the wasteful accumulation of goods and maintain the quality of their products. Prolonged storage in the warehouse can result in a loss of quality for products. In addition, production numbers that are lower than demand can lead to less availability of goods in the community, causing negative sentiment in the community and leading to loss of profit. It represents a significant challenge for the production and planning inventory control (PPIC) department, as they are required to accurately analyze the demand for goods and predict future needs using forecasting techniques.

Forecasting is a crucial decision-making method in various industry operational areas, including production and planning inventory control (PPIC), raw material purchasing, quality control (QC), inventory management, and goods delivery. Accurate forecasting is essential for avoiding shortages, which can lead to delivery delays and negative customer satisfaction. Therefore, it is vital to enhance forecasting accuracy and reduce errors to provide balancing between product demand and industry supplies (Rosienkiewicz, 2021).

Forecasting is the foundation of the production planning hierarchy and influences several critical decisions, including raw material procurement, machine scheduling, bill of material composition, machine maintenance and delivery schedule. It is vital to

guarantee that the forecasting results are as precise as possible.

The following benefits of using forecasting techniques have been observed in the industrial sector:

1. Forecasting enables the anticipation of forthcoming changes in demand and reduces uncertainty.
2. Forecasting serves to reinforce the collaborative efforts of various departments.
3. Managements are able to assess opportunities, prospects, and potential for growth.
4. The remaining operations are scheduled on the basis of predictions.
5. It impacts on saving costs by reducing inventories.
6. It offers understanding of the market and the development of new products.

1.1 Scope of the dissertation

The present dissertation will utilize three categories of heavy industries:

1. The paper industry
2. The electric power industry (EPI)
3. The energy industry

1.2 Scope of the datasets

The dissertation employs three kinds of datasets:

1. The number of defective products
2. Demand products
3. Product prices

Chapter 2

Prediction of product defects using Poisson and negative binomial (INAR(1)) models

2.1 Introduction

The initial study in this dissertation is concerned with the issue of defective products. It can be observed that all product quality problems can be attributed to two principal categories: controllable and uncontrollable factors. The former category encompasses factors influenced by the production components, such as machines, workers, tools and raw materials. In contrast, the latter category comprises factors that come from the environment, such as temperature, humidity, and contamination and are difficult to control (Montgomery, 2017). These factors are challenging to identify; they can also be called noise factors. Forecasting aims to establish product tolerance, preventing potential shortages at the end of the production process.

2.2 Literature review

Let $(x_t)_{t\in\mathbb{N}_0}$ be a discrete, i.e., $\mathbb{N}_0$-valued time series where $\mathbb{N}_0$ denotes the non-negative integers.

2.2.1 First order integer-valued AR model (INAR(1))

Let the innovations $(\epsilon_t)_{t\in\mathbb{N}_0}$ be independent identically distributed $\mathbb{N}_0$ valued random variables. The INAR (1) model follows the recursion

$$x_t = \alpha \circ x_{t-1} + \epsilon_t \tag{2.1}$$

where $\circ$ is a thinning operation and α is the inflation parameter with $\alpha \in (0;1)$. We use binomial thinning for the thinning operation defined as the random sum

$$\alpha \circ X := \sum_{i=1}^{X} Z_i \tag{2.2}$$

where $Z_i, i = 1, 2, ...,$ is a counting sequence of Bernoulli random variables with mean α (Weiß, 2018).

2.2.2 First order Poisson INAR model (PINAR(1))

The data follow PINAR(α, λ) model if the innovations (ϵ_t) follow Poisson Poi(λ) distribution. I use the index of dispersion (I) to analyze equidispersion of a distribution using the formula of

$$I := \frac{\sigma^2}{\mu} \in (0, \infty) \tag{2.3}$$

where $I \in (0, \infty)$. The mean and variance of $\epsilon_t \sim$ Poisson(λ) are $\mu_\epsilon = \sigma_\epsilon^2 = \lambda$, thus $I_\epsilon = 1$ (equidispersion).

2.2.3 First order negative binomial INAR model (NBINAR(1))

The data follow NBINAR(α, n, p) model if the innovations (ϵ_t) follow negative binomial NB(n, p) distribution, which show overdispersion property since $I_\epsilon = \frac{1}{p} > 1$.

2.3 The dataset

I use an object dataset from a paper industry in Banten, Indonesia. The products are jumbo paper rolls weighing 1.5 to 2 tonnes. The defects come from dirt, pebbles or sharp objects, which can damage the paper layer during production. From the PACF plot, the three datasets can be categorized as AR(1) models.

2.4 Results and discussion

2.4.1 Analysis of Index of Dispersion

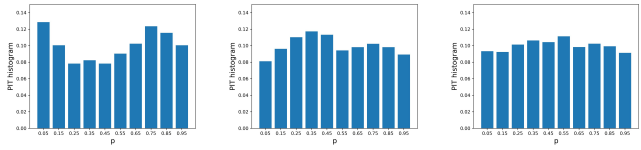
Machine 1 (1.899) has overdispersion ($I > 1$), while machines 2 (0.951) and 3 (0.93) have equidispersion ($I \approx 1$). The value of I is outside the upper (1.138) and lower bounds (0), so machine 1 indicates overdispersion. The value of p_0 on machine 1 (0.133) is much larger than the Poisson value (0.041), so the model cannot be modelled by Poisson model.

2.4.2 Model results and error estimation

Method of moments (MM) estimators give NBINAR(1) model parameters n_{est}=2.033, p_{est}=0.46 and $\hat{\alpha}$=0.254 for machine 1. For machines 2 and 3, PINAR(1) model is fitted with ($\hat{\lambda}$=0.133, $\hat{\alpha}$=0.125) and ($\hat{\lambda}$=0.133, $\hat{\alpha}$= 0.148).

Comparison using AIC and BIC shows that NBINAR(1) model (AIC=1377; BIC=1388) is better than negative binomial model (AIC=1401; BIC=1409) for machine 1.

2.4.3 Probability integral transform (PIT) model comparison



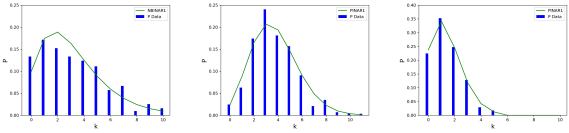
(a) Machine 1 (b) Machine 2 (c) Machine 3

Figure 2.1: PIT for all three machines.

Using NBINAR(1) model in figure 2.1 is likely to be uniform. PIT of data shows that NBINAR(1) model of machine 1 fits significantly better than PINAR(1) model. Figures 2.1b and 2.1c are

based on PINAR(1) model and they are likely to be uniform as well.

2.4.4 Model fitting



(a) Machine 1 (b) Machine 2 (c) Machine 3

Figure 2.2: Model fitting for all three machines.

Figure 2.2a shows that NBINAR(1) model (green line chart) with the proportion data (blue bar chart) shows a similar movement, indicating positive skewness. The fitting model has a peak at event 2 (18.89%).

In machine 2, the highest percentage is in event 3, with 20.7% in PINAR(1) model. Machine 3 shows very well fitting in all events. The largest percentage is in event 1, at 35%.

2.4.5 Forecasting analysis

In figure 2.3, the probabilities of 5+ product defects per day for machines 1(green), 2(blue), and 3(red) are 23.51%, 32.09%, and 17.2%, respectively. The percentages of machines with less than or equal to two defects are 46.28%, 27.79%, and 57.9% for machines 1, 2, and 3.

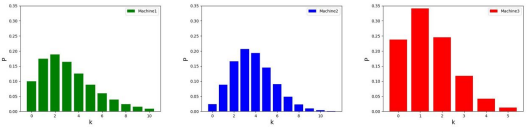


Figure 2.3: Forecasting for all three machines.

2.5 Conclusion

The best model for machine 1 is NBINAR(1) model, while machines 2 and 3 are PINAR(1) models. In the forecasting analysis, machine 1 shows that there is a 46.28% chance of less than or equal to 2 defects per day, machine 2 is 27.79%, and machine 3 is 57.9%. Machines 1 and 3 perform better than machine 2.

Thesis 1

I focus on industrial quality problems using Poisson and negative binomial first-order integer-valued autoregressive (INAR(1)) models to fit and predict the product defect dataset from paper machines containing discrete numbers. I provide some statistical theories to detect which model performs better than others. I also perform comparison tests of Poisson and negative binomial models with and without INAR(1) model, where adding INAR(1) model brings lower error measurements.

Related publications:

- Gandhi, Herry Kartika, and Ispány Márton. Analyzing uncontrollable factors that cause defective products by Poisson and negative binomial INAR(1) for fitting model. Proceedings on Engineering. doi:10.36055/jiss.v5i1.6494. (SJR: Q4) [Status: Accepted]

Chapter 3

Forecasting of daily energy consumption using hybrid linear nonlinear model: SARIMA-SVR

3.1 Introduction

In this chapter, the study use Electricity consumption (EC) dataset from the electric power industry (EPI) as an object. The planning manager must predict the daily electrical supply so that the supply and demand are balanced. If there is too much energy supplied to consumers, the efficiently will reduce, raw materials will be wasted and production costs will increase. If there is not enough electricity supply, there may be blackouts.

EC data is difficult to understand, but new forecasting methods make predictions more accurate. This study investigates the use of the SARIMA-SVR hybrid forecasting model for daily EC. SARIMA model employs a linear equation for non-stationary datasets with seasonal indications (Shumway and Stoffer, 2019), whereas SVR model uses kernel techniques for handling nonlinearity.

3.2 Literature review

3.2.1 SARIMA model

Multiplicative seasonal ARIMA models are based on seasonal backshift operator B^s defined as $B^s x_t = x_{t-s}$ and seasonal difference operator $\nabla_s x_t = x_t - x_{t-s}$. The multiplicative SARIMA$(p,d,q)\times(P,D,Q)_s$ model is

$$\Phi(B^s)\phi(B)\nabla_s^D\nabla^d x_t = \Theta(B^s)\theta(B)w_t, \tag{3.1}$$

where ϕ and θ are the autoregressive and moving average polynomials with orders p and q, d is the order of differencing, and Φ and Θ are the seasonal autoregressive and moving average polynomials with orders P and Q.

3.2.2 Support vector regression (SVR) model

SVR solves the following optimization problem:

$$\min_{\boldsymbol{w},b,\zeta,\zeta^*} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{n}(\zeta_i + \zeta_i^*) \tag{3.2}$$

subject to

$$\begin{gathered} y_i - \boldsymbol{w}^T\phi(\boldsymbol{x}_i) - b \leq \epsilon + \zeta_i, \\ \boldsymbol{w}^T\phi(\boldsymbol{x}_i) + b - y_i \leq \epsilon + \zeta_i^*, \\ \zeta, \zeta_i^* \geq 0, i = 1, \ldots, n \end{gathered}$$

where vectors $\boldsymbol{x}_i \in \mathbb{R}^p, i = 1, \ldots, n$ and y_i is output value.

3.2.3 Hybrid forecasting SARIMA-SVR model

The proposed model uses SVR model to forecast ε_t, where ε_t are the innovations or residuals between the actual values and the SARIMA predictions, $\varepsilon_t = x_t - \hat{x}_t$.

I describe the input of SVR model as (ε_1, ε_2, ...,ε_t) as described in Algorithm 1 for n-step forecasting.

3.2.4 SARIMA-(G)ARCH models

In these models, the error term of SARIMA model is modeled by ARCH and GARCH models.

The GARCH (p, q) model equation is

$$\sigma_t^2 = \alpha_0 + \sum_{k=1}^{p} \alpha_k \cdot r_{t-k}^2 + \sum_{k=1}^{q} \beta_k \cdot \sigma_{t-k}^2, \tag{3.3}$$

Algorithm 1 Recursive multi-step SARIMA-SVR model

Input: $\{x_t\}$, $\{\varepsilon_t\}$, n_{step}, k
Output: $prediction = [\hat{x}_{t+1}, \hat{x}_{t+2}, \ldots, \hat{x}_{t+n_{step}}]$

$i \leftarrow 1$
$X \leftarrow \{x_t\}$
$\varepsilon \leftarrow \{\varepsilon_t\}$
$prediction \leftarrow [\ ]$
while $i \leq n_{step}$ **do**
    $\boldsymbol{X}_{svr}, Y_{svr} \leftarrow$ to-SVR-input(ε, k)
    $\text{SVR}_{model} \leftarrow \text{SVR}.fit(\boldsymbol{X}_{svr}, Y_{svr})$
    $\varepsilon\_hat \leftarrow \text{SVR}_{model}.predict(\varepsilon[:-k])$
    $\text{SARIMA}_{model} \leftarrow \text{SARIMA}.fit(X)$
    $x\_hat \leftarrow \text{SARIMA}_{model}.predict(X)$
    $X.append(x\_hat + \varepsilon\_hat)$
    $\varepsilon.append(\varepsilon\_hat)$
    $prediction.append(x\_hat + \varepsilon\_hat)$
    $i \leftarrow i + 1$
end while

where r_t is defined from SARIMA's error term, σ_t^2 is the variance and the following conditions should be held: $\alpha_0 > 0$, $\alpha_k \geq 0$ for $k = 1, \ldots, p$, $\beta_k \geq 0$ for $k = 1, \ldots, q$, and $\sum_{k=1}^{q} \beta_k + \sum_{k=1}^{p} \alpha_k < 1$.

3.3 Dataset

The dataset is the electrical consumption in the eastern part of the United States. It is daily from '2004-10-01' to '2018-08-02' and has a size of 5054. The metric is 1/20000 MW.

3.4 Results and discussion

3.4.1 Normality check

Using $\alpha = 0.05$, only the SVR and SARIMA-SVR models indicate normal distribution, where Shapiro-Wilk (SW) and Jarque-Bera

(JB) normality tests p-values are higher than α. For the SARIMA model, two statistical tests show that the residuals do not indicate normal distribution.

Table 3.1: Normality check using SW and JB tests

Model	SW stat	SW p-value	JB stat	JB p-value
SARIMA	0.9515	0.001	18.8466	8.082E-05
SARIMA-ARCH	0.9554	0.0019	17.6552	0.0001
SARIMA-GARCH	0.9572	0.0025	18.4483	9.86E-05
SVR	0.9942	0.3059	1.7564	0.4155
SARIMA-SVR	0.9792	0.1144	3.9277	0.1516

3.4.2 Homoscedasticity check

The p-value for the Engle ARCH (EARCH) statistic is less than the α level, indicating that the SARIMA residuals do not have constant residual variance. Of the four benchmark models, only SVR and SARIMA-SVR model show constant variances. SARIMA-ARCH and SARIMA-GARCH models do not indicate homoscedasticity.

Table 3.2: Homoscedasticity check using EARCH tests

Model	EARCH statistic	EARCH p-value
SARIMA	32.0846	0.0004
SARIMA-ARCH	31.8654	0.0004
SARIMA-GARCH	31.0137	0.0006
SVR	11.9212	0.2904
SARIMA-SVR	7.876	0.6409

3.4.3 Residual dependency test

Using $\alpha = 0.05$, the SARIMA model shows that the first six lags indicate residual dependencies. Unlike SVR and SARIMA-SVR models, where all p-values are above the significance level (α), this indicates that the residuals of the two models show independence.

3.4.4 Error measurement test

I use five error measurements: mean squared error (MSE), mean absolute error (MAE), root mean squared error (RMSE), mean absolute percent error (MAPE) and r squared (R2). The SARIMA-SVR model is the first rank, exhibiting the best model. However, the R2 value remains relatively similar to that of the SARIMA model. The SARIMA model is the second-highest rank. The SARIMA model outperforms the SARIMA-GARCH and SARIMA-ARCH models, which rank third and fourth, respectively. Conversely, the single SVR model demonstrates bad performance compared to other models.

Table 3.3: Benchmark Error Measurement

Model	MSE	MAE	RMSE	MAPE	R2
SARIMA	0.0051	0.0545	0.0713	2.8546	0.7588
SARIMA-ARCH	0.0059	0.0611	0.0772	3.1911	0.7173
SARIMA-GARCH	0.0055	0.0578	0.0739	3.0121	0.7443
SVR	0.0082	0.0718	0.0906	3.8209	0.6109
SARIMA-SVR	0.0048	0.0436	0.0694	2.3214	0.75715

3.4.5 Diebold-Mariano (DM) test

The SARIMA-SVR model is demonstrated to have p-values that are all smaller than α in comparison with all benchmark models. Meanwhile, the SARIMA model is observed to have a considerably smaller p-value than the SARIMA-ARCH model, yet is not as effective when compared to the SARIMA-GARCH model.

Table 3.4: DM test between the models (p-value)

Models	SARIMA	SARIMA-GARCH	SARIMA-ARCH	SVR
SARIMA-SVR	0.0325	0.0119	0.014	0.024
SARIMA		0.2868	0.0055	0.098
SARIMA-GARCH			0.0159	0.0934
SARIMA-ARCH				0.0515

3.4.6 N-step horizon forecasting

The SARIMA-SVR model continues to demonstrate the most minimal measurement error at [1, 2, 3]-step in comparison to the other models. However, for [5, 10, 15]-step, the SARIMA-GARCH model demonstrates superior performance compared to other models, including the SARIMA-SVR model. The SVR model exhibits less robust performance compared to other models.

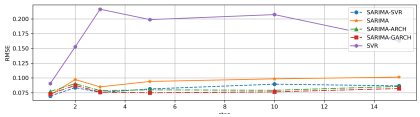


Figure 3.1: Line Plot RMSE for n-step forecasting models.

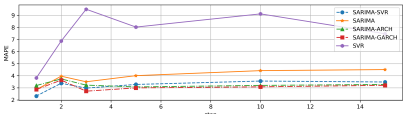


Figure 3.2: Line Plot MAPE for n-step forecasting models.

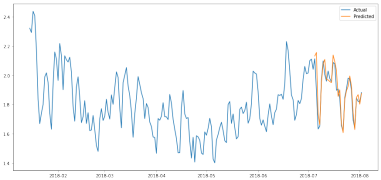


Figure 3.3: SARIMA–SVR prediction values plot.

3.5 Conclusion

The DM test demonstrates that the SARIMA-SVR model significantly outperforms the four benchmark models under the significance level $\alpha = 0.05$. The residuals of the SARIMA-SVR model also demonstrate independence and normality. Finally, the proposed model exhibits consistently low measurement error values for the [1, 2, 3] step, whereas the SARIMA-GARCH model demonstrates superior performance than the other models for the [5, 10, 15] step.

Thesis 2

I draw attention to industrial demand forecasting by introducing a novel SARIMA-SVR model, which belongs to the hybrid linear and nonlinear forecast (HLNF). After preprocessing the daily electrical consumption dataset, I leverage the SARIMA model to incorporate the seasonal relationship inside the model. To address the residual SARIMA's non-normality and heteroscedasticity properties, I enhance the model with SVR, which better captures nonlinear relationships.

Related publications:

- Gandhi, Herry Kartika. Applying hybrid forecasting model SARIMA-SVR for daily energy consumption data. The 2024 IEEE 3rd Conference on Information Technology and Data Science (CITDS 2024), Debrecen, Hungary, 2024, doi: 10.1109/CITDS62610.2024.10791394.

Chapter 4

Energy price mid-term forecasting using hybrid decomposition and deep learning models

4.1 Introduction

Energy sources (ENS) are a fundamental necessity for all, residential, commercial and industrial use, as well as for transportation, education, healthcare and other facilities across the globe. Up to 35% of total energy consumption (ENC) is accounted for by industries, which include crude oil (CO), natural gas (NG), coal and heating oil (HO) (Energy Information Administration, 2023). Some industries generate their own energy sources (ENS) through the use of generators.

The oil industry is significantly impacted by several other sectors, including oil refining, petrochemicals, and furniture manufacturing.

The objective of this study is to utilize mid-term forecasting techniques in order to provide a longer-term view of energy price prediction, extending up to three months or more at each stage of the process. The proposed hybrid decomposition forecasting (HDF) models are employed with deep learning (DL) models.

4.2 Decomposition methods

4.2.1 Hodrick-Prescott decomposition (HPD)

HPD is used to separate the time series raw data (x_t) into trend components (τ_t) and cyclical components (c_t). We express this relationship through

$$x_t = \tau_t + c_t. \tag{4.1}$$

The method is used to minimize the following objective function:

$$\min_{\tau_1,\tau_2,\ldots,\tau_t} \left\{ \sum_{t=1}^{T} (x_t - \tau_t)^2 + \lambda \cdot \sum_{t=2}^{T-1} ((\tau_{t+1} - \tau_t) - (\tau_t - \tau_{t-1}))^2 \right\} \tag{4.2}$$

where T is the sample size and λ is the smoothing parameter. The original data is rescaled by taking natural logarithm.

4.2.2 Wavelet decomposition (WD)

The calculation of the wavelet transform comprises two distinct types: scaling function $\varphi(t)$ and wavelet function $\psi(t)$. We can describe as

$$f(t) = V_j + W_j = \sum_k c_k \varphi(2^j t - k) + \sum_k d_{j.k} \psi(2^j t - k) \tag{4.3}$$

where j corresponds to scales, k is index which is vary from 0 to 2^{j-1}; $k, j \in \mathbb{Z}$, and $c_k, d_{j.k} \in \mathbb{R}$.

For the higher number of j, and if $f(t) \in V_{j+1}$, the formula becomes

$$f(t) = \sum_k c_k \varphi(t - k) + \sum_k \sum_j d_{j.k} \psi(2^j t - k) \tag{4.4}$$

where $V_0 = \sum_k c_k \varphi(t - k)$.

4.2.3 Empirical mode decomposition (EMD)

The outputs of EMD are the intrinsic mode functions ($imfs$) and a residual as shown in the following formulation

$$x_t = \sum_i imf_i + res. \tag{4.5}$$

To get imf_i, there are several stages in EMD (Huang et al., 1998).

Step 1. Find local extrema of (x_t) and put them into envelopes.

Step 2. Generate $E_{up}(t)$ and lower envelope $E_{low}(t)$ with cubic splines interpolation.

Step 3. Generate $E_{mean}(t)$ by calculating point-by-point average of upper and lower envelopes.

Step 4. Find the residual, $res(t) = x_t - E_{mean}(t)$.

Step 5. If the difference between the number of zero crossings and extrema at most by one and $E_{mean}(t)$ is close to zero, go to step 6. else $f(t) = res(t)$, and repeat steps 1-6.

Step 6. Define $imf_i = res(t)$ and calculate $f^{new}(t) = f(t) - res(t) = f(t) - imf_i$.

Step 7. Repeat steps 1-6 until stopping criterion (ϵ)

$$\sum \frac{(res(t) - f(t))^2}{f(t)^2} < \epsilon. \tag{4.6}$$

4.2.4 Complete ensemble EMD (CEEMD)

CEEMD generates two datasets using noise signals ($M_+(t)$ and $M_-(t)$) using formula:

$$\begin{bmatrix} M_+(t) \\ M_-(t) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \cdot \begin{bmatrix} f(t) \\ w(t) \end{bmatrix}, \tag{4.7}$$

where $w(t)$ is a noise process.

At the end of step 7, we get imf_{+ji} and imf_{-ji}. After M repetition, we get

$$imf_i = \frac{1}{2M} \sum_{j=1}^{M} (imf_{+ji} + imf_{-ji}). \tag{4.8}$$

4.3 Deep learning (DL) models

4.3.1 Artificial neural network (ANN) model

ANN model is an extension of Neural Network (NN) (Lazzeri, 2020). The single neuron will repeat the training process until it gets an acceptable weight vector. The initial weight vector ($\boldsymbol{w}$) is taken from random weights, and then linear combinations are calculated with the addition of bias (b) as follows

$$z = \sum_{i=1}^{n} w_i x_i + b_i = \boldsymbol{x}^T \boldsymbol{w} + b \tag{4.9}$$

Using activation function (ϕ), the formula:

$$h(\boldsymbol{x}) = \phi(\boldsymbol{x}^T \boldsymbol{w} + b). \tag{4.10}$$

where the examples of ϕ are sigmoid, tanh, or swish function (Vishwas and Patel, 2020).

4.3.2 Long short-term memory (LSTM) model

LSTM model simultaneously uses long-term dependencies and short-term memories about previous sequential data (Brownlee, 2018).

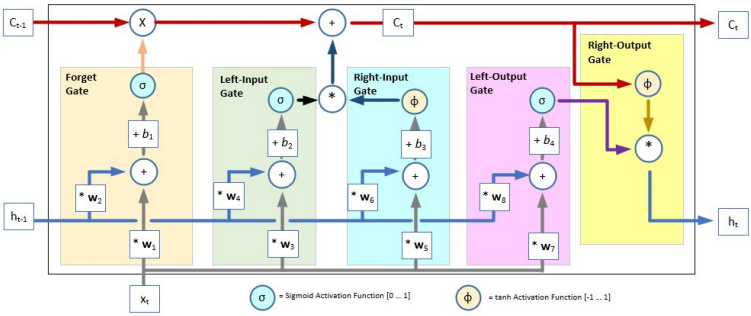


Figure 4.1: Long short-term memory model.

The forget gate has equation of

$$f_t = \sigma(\boldsymbol{w}_t \cdot [h_{t-1}, x_t] + b_1) \tag{4.11}$$

where h_{t-1} is the previous short-term memory, $\boldsymbol{w}_t$ is the weight vector with components $[\boldsymbol{w}_1, \boldsymbol{w}_2]$, where $\boldsymbol{w}_1$ relates to x_t and $\boldsymbol{w}_2$ relates to h_{t-1}. The b_1 is the bias at the forget gate. The σ is a sigmoid function.

The subsequent gate is the input gate, which consist of left-input gate (4.12) and right-input gate (4.13). The gate determines the long-term memory (C) to be retained.

$$i_t = \sigma(w_i \cdot [h_{t-1}, x_t] + b_2) \tag{4.12}$$

$$\tilde{C}_t = tanh(w_c \cdot [h_{t-1}, x_t] + b_3) \tag{4.13}$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{4.14}$$

where $\boldsymbol{w}_i$ is the weight of the input i in left-input gate $(\boldsymbol{w}_3, \boldsymbol{w}_4)$, b_2 is the bias of the left-input gate. The C_{t-1} is the previous long-term memory, $\boldsymbol{w}_c$ is the weight of the input c in right-input gate $(\boldsymbol{w}_5, \boldsymbol{w}_6)$, and b_3 is the bias of the right-input gate.

Furthermore, the formulas of output gate are

$$o_t = \sigma(\boldsymbol{w}_o \cdot [h_{t-1}, x_t] + b_4) \tag{4.15}$$

$$h_t = o_t * tanh(C_t) \tag{4.16}$$

where $\boldsymbol{w}_o$ is the weight of the input o in left-output gate $(\boldsymbol{w}_7, \boldsymbol{w}_8)$, and b_4 is the bias of the left-output gate (Hochreiter and Schmidhuber, 1997).

4.3.3 Convolutional neural network-LSTM (CNN-LSTM) model

The mechanism of CNN-LSTM model is as shown in Figure 4.2. CNN model reduces dimension using n-dimension kernel tricks using CNN techniques: filtering, shreding and pooling.

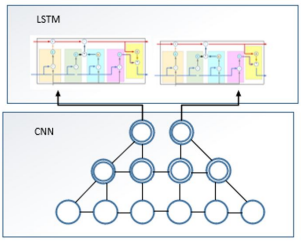


Figure 4.2: CNN-LSTM model.

4.3.4 Proposed hybrid decomposition model

The dataset splits into multiple $imfs$. Every imf_i enters the forecasting model, which is a deep learning (DL) model. After the forecasting model predicts, the results are combined into one number with simple addition.

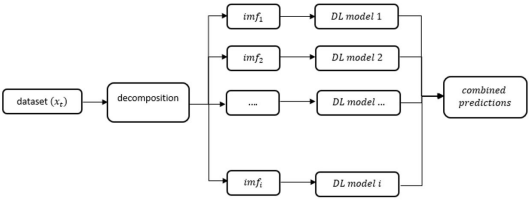


Figure 4.3: Hybrid decomposition.

Algorithm 2 Hybrid decomposition

Input: $\{x_t\}$, n_{test}, n_{train}, n_{step}
Output: prediction = $[\hat{x}_{t+1}, \hat{x}_{t+2}, \ldots, \hat{x}_{t+n_{step}}]$

train $\leftarrow x_t[: -n_{test}]$
prediction $\leftarrow$ [ ]
for h in range $(1, n_{test})$ **do**
    $imfs \leftarrow$ [ ]
    $total\_xhat \leftarrow [n_{step}]$ of zeros
    If h % n_{step} $!= 0$ **then** continue
    else
    $imf_i \leftarrow$ decompose(train)
    $imfs.insert(imf_i)$
    for j in range $(1, \text{len}(imfs))$ **do**
        forecasting.fit$(imf_j[: -n_{step}], imf_j[-n_{step} :])$
        $xhat \leftarrow$ forecasting.predict(n_{step})
        $total\_xhat$ += xhat
    end for
    prediction.append$(total_xhat)$
    train = $x_t[: (-n_{test} + h)]$
end for

4.4 Datasets

4.4.1 Data description

I use two pivotal sources of oil prices: Brent (BCOP) and West Texas Intermediate (WTI) crude oil price (WCOP). Natural gas price (NGP) data comes from Yahoo Finance. The heating oil price (HOP) dataset based on two sources, financial contracts and over-the-counter (OTC) in USD/GAL units, comes from Yahoo Finance as well.

4.4.2 Decomposition results

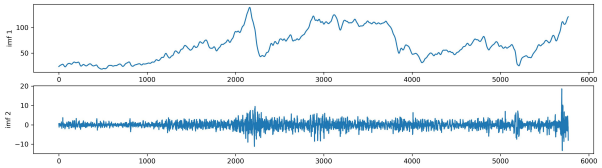


Figure 4.4: HP decomposition of BCOP.

Hodrick-Prescott decomposition (HPD) gives two results, expressed as trend and cycle. We also get two $imfs$ from WD (Figure 4.5), where $imf1$ is the retransformed process, and $imf2$ is the residual from WD ($imf1$).

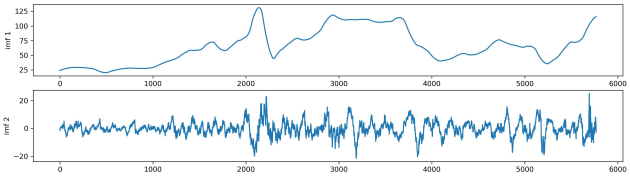


Figure 4.5: Reconstruct WD of BCOP.

The result of EMD decomposition is shown in Figure 4.6. The result of CEEMD on BCOP is shown in Figure 4.7 with eight $imfs$.

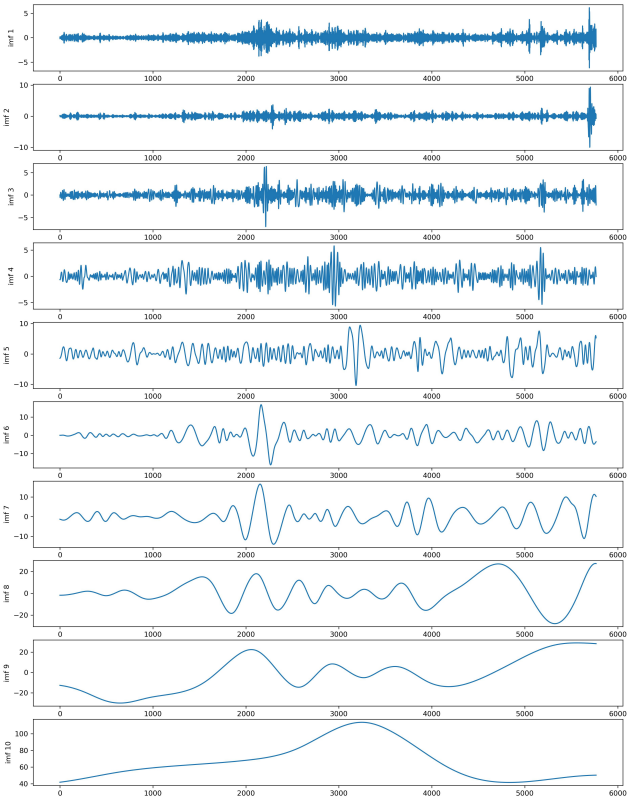


Figure 4.6: EMD decomposition of BCOP.

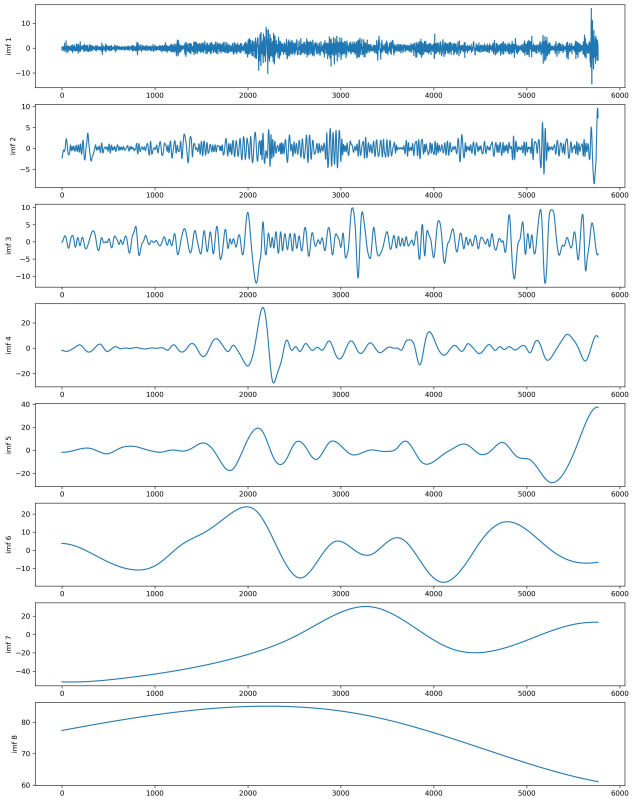


Figure 4.7: CEEMD decomposition of BCOP.

4.5 Results and discussion

4.5.1 Analysis Brent Crude Oil Price (BCOP)

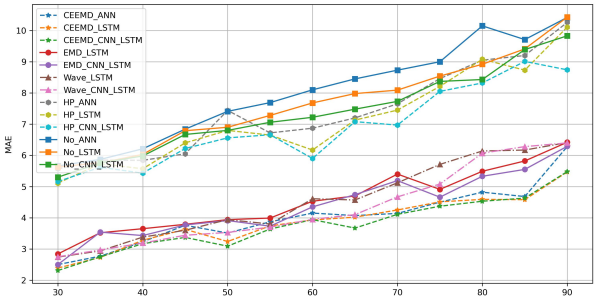


Figure 4.8: Benchmark Models using MAE for BCOP.

As illustrated in Figure 4.8, the MAE of CEEMD-CNN-LSTM model exhibits a lower line plot than that of other models. The plots of CEEMD-ANN and CEEMD-LSTM models exhibit relatively low MAE, though not at a level lower than that of CEEMD-CNN-LSTM model. All MAE values of models without decomposition and with HPD demonstrate considerable gap values at the upper end of the line plot.

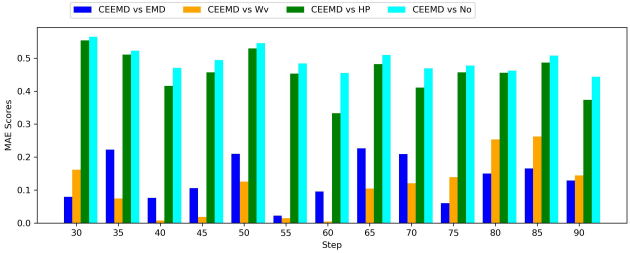


Figure 4.9: Percentage Decreasing MAE of the Models for BCOP.

Figure 4.9 compares the optimal model (CEEMD-CNN-LSTM) with a similar model (CNN-LSTM) but with distinct decomposition techniques. The CEEMD is capable of reducing the MAE value to approximately 50% when compared to the 'no-decomposition' approach. When CEEMD is employed in place of HP decomposition, the reduction is approximately 30-55%. Furthermore, when CEEMD is compared with EMD and WD, the MAE reduction at certain steps exhibits fluctuations between 0-30%.

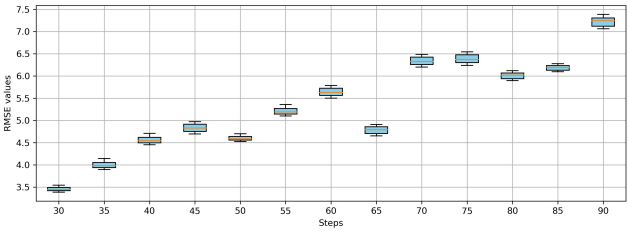


Figure 4.10: BoxPlot RMSE of CEEMD-CNN-LSTM for BCOP.

Figure 4.10 depicts the root mean square error (RMSE) outputs of 30 simulations from the optimal model (CEEMD-CNN-LSTM) in a boxplot graph. The interquartile range (IQR) of the box plot of all steps in range from 0.07 to 0.18. Meanwhile, the difference between the lower and upper whiskers ranges from 0.21 to 0.6. The difference in output from each iteration of the forecasting model is still relatively low and tolerable. The percentage comparison between the difference whiskers and the mean is below 10%.

4.5.2 Analysis WTI Crude Oil Price (WCOP)

The three models with CEEMD decomposition (CEEMD-ANN, CEEMD-LSTM and CEEMD-CNN-LSTM) demonstrate lower MSE values in comparison to the remaining models. As with the preceding dataset, the CEEMD-CNN-LSTM model provides the lowest value. The MSE values for the EMD and WD models are

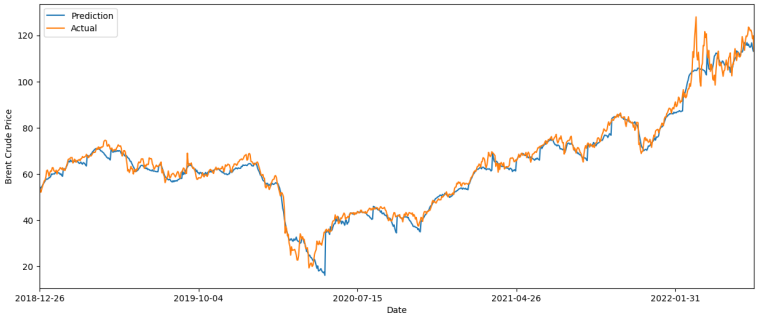


Figure 4.11: Line Plot of Actual vs Predictions (BCOP) using CEEMD-CNN-LSTM.

larger than those for the CEEMD model, while the values for HPD and No Decomposition are considerably higher.

CEEMD-CNN-LSTM model brings MAE reduction against the no-decomposition and HPD with ranges from 45 to 60%. While against WD, the ranges is between 10 and 40%. For EMD, replacing it with CEEMD is more effective, reducing MAE by 3 to 30%.

4.5.3 Analysis Natural Gas Price (NGP)

The CEEMD and EMD models have lower MAE values than other decompositions for natural gas prices. WD has a slightly larger MAE value, while HPD and no decomposition have much larger values. EMD can outperform CEEMD at some steps, but CEEMD-CNN-LSTM model provides more accurate values. MAE and RMSE are similar in this case.

In NGP, EMD gives better results than CEEMD in some steps. For other steps, CEEMD is better than EMD, with percentages between 1-12%. CEEMD is better than HPD and No-Decomposition with different percentages above 50%.

The box plot shows that the Q2 value rises from 0.26 to 0.41. The mean of the RMSE also gets bigger. The whisker length is between 4 and 8% of the mean RMSE.

4.5.4 Analysis Heating Oil Price (HOP)

The HOP results differ from the previous datasets. WD gives better results than the other four decompositions. WD has the lowest MAPE value, indicating that it forecasts well. CEEMD, EMD and HPD have similar results. The other four models have a majority of MAPE values above 10.

In HOP, wavelet-CNN-LSTM model provides the best forecasting compared to other models. WD reduces MAE by 25-40% compared to EMD. CEEMD can be improved by 15-35% with WD. Replacing HPD and No-Decomposition with WD also decreases the results by 30-55%.

The best model (wavelet-CNN-LSTM) has low variation, with values between 0.01 and 0.02. The means are between 0.16 and 0.23, and the whisker length is between 6 and 8% of the mean.

4.6 Conclusion

I use the multiple output forecasting method with 30 to 90 steps and five intervals. Three out of four datasets show that the CEEMD-CNN-LSTM model is better than the other models. The CEEMD decomposition technique is better than others. EMD and WD give better results than HPD and no decomposition, but not as good as CEEMD. CEEMD reduces MAE values by 50 to 30% compared to HPD or no decomposition. For WD or EMD, CEEMD can reduce MAE values by up to 20%. The best model in the RMSE trials has a variation of 5 to 8% of the mean RMSE.

Using HOP dataset, wavelet-CNN-LSTM model works better than the other models for the whole n-step forecasting. CEEMD and EMD are the second-best models, HPD is third, and no decomposition is the worst. WD decreases MAE for CEEMD and EMD by 20 to 40%, and for HPD and no decomposition, wavelet can decrease MAE by 30 to 50%. The best model predictions are close to the actual data.

Thesis 3

The study concerned about price forecasting and has successfully applied a novel combination of decomposition techniques and deep learning models to predict four datasets of daily energy prices. The decompositions could split the datasets into several intrinsic mode functions ($imfs$) with different frequencies. I compare using all paired models of four decomposition techniques and three deep learning models. I also perform mid-level multi-step forecasting, where the prediction ranges from 30 to 90 days ahead.

Related publications:

- Gandhi, Herry Kartika, and Ispány, Márton. Multi-step Natural Gas Price Forecasting using Ensemble Empirical Mode Decomposition and Long Short-Term Memory Hybrid Model. International Journal of Energy Economics and Policy. 14(4) (2024), 590–598. doi:10.32479/ijeep.16053. (SJR: Q2)

- Gandhi, Herry Kartika. Mid-term forecasting of crude oil prices using hybrid CEEMDAN and CNN-LSTM deep learning model. Polityka Energetyczna. 27(4) (2024), 19-38, doi: 10.33223/epj/190486. (SJR: Q3)

Chapter 5

Conclusions

The first study is applied to paper roll defect products in the paper industry. The results of the model fitting analysis demonstrate that both the PINAR(1) and NBINAR(1) models are capable of accurately representing the actual dataset. Furthermore, Probability Integral Transform (PIT) visualisation is employed to facilitate the selection of an appropriate model. The selected model exhibits lower AIC and BIC error values in comparison to other models.

The second study is a demand forecasting exercise. I utilize daily electrical consumption data from the eastern region of the United States. The hybrid model combining SARIMA and SVR models demonstrates superior performance in (1, 2, 3)-step ahead forecasting compared to benchmark models. The proposed model shows a notable superiority in terms of the DM test, with a significance level of $\alpha = 0.05$. Nevertheless, for longer steps (5, 10, 15), the SARIMA-GARCH model exhibits superior accuracy compared to the SARIMA-SVR model.

The third study is concerned with price forecasting. The four highest-priority energy price datasets are employed in this analysis. In three cases, CEEMD decomposition yields favourable results in comparison to the other decompositions. In the case of the heating oil price, however, wavelet decomposition proved to be the most effective among the alternatives. The optimal model has the potential to reduce mean absolute error (MAE) values by up to 50% in comparison to alternative decompositions. Despite the inherent variability in deep learning outputs, with 30 simulations conducted for each step, the changes in root mean square error (RMSE) values are approximately 5 to 8% around the mean RMSE, which remains relatively low. The plot of prediction versus actual demonstrates that the prediction is in close alignment with the actual value.

References

Brownlee, J. (2018). *Deep learning for time series forecasting*. Machine Learning Mastery.

Energy Information Administration (2023). Energy use in industry. `https://www.eia.gov/energyexplained/use-of-energy/industry.php` [Accessed on 7 February 2024].

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Yen, N., Tung, C. C., and Liu, H. H. (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society A*, 454(1971):903–995.

Lazzeri, F. (2020). *Machine learning for time series forecasting with Python*. John Wiley & Sons.

Montgomery, D. C. (2017). *Design and analysis of experiments*. John Wiley & Sons.

Rosienkiewicz, M. (2021). Artificial intelligence-based hybrid forecasting models for manufacturing systems. *Eksploatacja i Niezawodność*, 23(2):263–277.

Shumway, R. and Stoffer, D. (2019). *Time series: A data analysis approach using R*. CRC Press.

Vishwas, B. V. and Patel, A. (2020). *Hands-on time series analysis with Python: From basics to bleeding edge techniques*. Springer.

Weiß, C. H. (2018). *An introduction to discrete-valued time series*. Wiley.

List of publications

Journal papers

Gandhi, Herry Kartika, and Ispány, Márton. Analyzing uncontrollable factors that cause defective products by Poisson and negative binomial INAR(1) for fitting model. *Proceedings on Engineering*. doi:10.36055/jiss.v5i1.6494. (**SJR: Q4**) [Status: Accepted]

Gandhi, Herry Kartika, and Ispány, Márton. Multi-step Natural Gas Price Forecasting using Ensemble Empirical Mode Decomposition and Long Short-Term Memory Hybrid Model. *International Journal of Energy Economics and Policy*. **14(4)** (2024), 590–598. doi:10.32479/ijeep.16053. (**SJR: Q2**)

Gandhi, Herry Kartika. Mid-term forecasting of crude oil prices using hybrid CEEMDAN and CNN-LSTM deep learning model. *Polityka Energetyczna*. 27(4) (2024), 19-38, doi: 10.33223/epj/190486. (**SJR: Q3**)

Conferences

Gandhi, Herry Kartika. Application of time-series method in company revenue. *The 2022 IEEE 2nd Conference on Information Technology and Data Science (CITDS 2022)*, Debrecen, Hungary, May 16 – 18, 2022. (Online Conference) [Note: Presentation only]

Gandhi, Herry Kartika. Applying hybrid forecasting model SARIMA-SVR for daily energy consumption data. *The 2024 IEEE 3rd Conference on Information Technology and Data Science (CITDS 2024)*, Debrecen, Hungary, 2024, pp. 1-6, doi: 10.1109/CITDS62610.2024.10791394.

UNIVERSITY of DEBRECEN

UNIVERSITY AND NATIONAL LIBRARY
UNIVERSITY OF DEBRECEN
H-4002 Egyetem tér 1, Debrecen
Phone: +3652/410-443, email: publikaciok@lib.unideb.hu

Registry number: DEENK/588/2024.PL
Subject: PhD Publication List

Candidate: Herry Kartika Gandhi
Doctoral School: Doctoral School of Informatics
MTMT ID: 10095536

List of publications related to the dissertation

Foreign language scientific articles in international journals (3)

1. **Gandhi, H. K.**, Ispány, M.: Analyzing Uncontrollable Factors that Cause Defective Products by Poisson and Negative Binomial INAR(1) for Fitting Model.
   [Epub ahead of print] 7 (1), 10, 2025. ISSN: 2620-2832.
   DOI: http://dx.doi.org/10.24874/PES07.01.010
2. **Gandhi, H. K.**: Mid-term forecasting of crude oil prices using hybrid CEEMDAN and CNN_LSTM deep learning model.
   Polityka Energetyczna. accepted for publication (-), [1-36], 2024. ISSN: 1429-6675.
3. **Gandhi, H. K.**, Ispány, M.: Multi-step Natural Gas Price Forecasting using Ensemble Empirical Mode Decomposition and Long Short-Term Memory Hybrid Model.
   IJEEP. 14 (4), 590-598, 2024. EISSN: 2146-4553.
   DOI: http://dx.doi.org/10.32479/ijeep.16053

The Candidate's publication data submitted to the iDEa Tudóstér have been validated by DEENK on the basis of the Journal Citation Report (Impact Factor) database.

11 December, 2024

DEBRECENI EGYETEM * Egyetemi és Nemzeti Könyvtár *