PERBANDINGAN MODEL SARIMA, ARIMA, SARIMAX, DAN TIME SERIES REGRESSION DALAM PERALAMAN DATA KARBON DIOKSIDA DARI OBSERVATORIUM MAUNA LOA

Penulis Rendika Nurhartanto Suharto, dan Dosen Regita Putri Permata, S.Stat., M. Stat.
Jurusan Sains Data, Fakultas Informatika, Telkom University Surabaya
Jl. Ketintang No.156, Ketintang, Kec. Gayungan, Surabaya, Jawa Timur, Indonesia
e-mail: regitapermata@ittelkom-sby.ac.id

Konsentrasi karbon dioksida (CO2) dalam atmosfer bumi memiliki peran krusial dalam perubahan iklim global. Observatorium Mauna Loa (MLO) di Hawaii memainkan peran penting dalam pemantauan pertumbuhan konsentrasi CO2 selama beberapa dekade terakhir. Data deret waktu konsentrasi CO2 yang dikumpulkan sejak tahun 1958 di MLO telah menjadi landasan penting bagi pemahaman kita tentang perubahan lingkungan. Dalam upaya meramalkan konsentrasi CO2, empat model statistik utama digunakan: Autoregressive Integrated Moving Average (ARIMA), Seasonal ARIMA (SARIMA), Seasonal ARIMA with Exogenous Factors (SARIMAX), dan Time Series Regression. Penelitian ini menunjukkan bahwa model terbaik untuk peramalan adalah SARIMA (0, 1, 0) (2, 1, 1, 12) dengan RMSE sebesar 3.15 dan rentang interval kepercayaan 95% yang cukup sempit, menandakan ketepatan model dalam memprediksi hasil. Penggunaan model hybrid SARIMAX, yang menggabungkan TSR OLS dan SARIMA, juga menunjukkan peningkatan yang positif dibandingkan dengan penggunaan model TSR OLS saja. Hasil peramalan konsentrasi CO2 di MLO menunjukkan tren kenaikan dan pola bulanan yang konsisten dengan bulan-bulan sebelumnya.

***Kata Kunci*— CO2, TSR , ARIMA, SARIMA, SARIMAX**

I. PENDAHULUAN

Kadar karbon dioksida (CO2) dalam atmosfer bumi adalah salah satu aspek kritis dalam perubahan iklim global. Karbon dioksida, lazim disebut gas asam arang yang merupakan senyawa kimia yang sangat penting bagi kehidupan organisme dan udara bersih mengandung kira-kira 0,03% karbon dioksida [1]. Salah satu pusat pengamatan penting untuk merekam pertumbuhan konsentrasi CO2 dalam atmosfer selama beberapa dekade terletak di *Mauna Loa Observatory* (MLO).

MLO terletak di 19.5°N, 155.6°W di pulau Hawaii, merupakan salah satu dari empat garis pangkal stasiun pemantauan atmosfer yang dioperasikan oleh *Geophysical Monitoring for Climatic Change* (GMCC) dari *National Oceanic and Atmospheric Administration* (NOAA) [2]. MLO adalah tempat untuk salah satu rangkaian pengukuran atmosfer dengan waktu terlama. Sejak tahun 1958, MLO telah ada mengumpulkan data deret waktu tentang konsentrasi CO2 di atmosfer, yang menjadi tulang punggung *kurva keeling* yang terkenal merupakan landasan penting dalam pemahaman kita terhadap perubahan lingkungan yang sedang terjadi. [3].

Salah satu pendekatan yang penting dalam menganalisis data deret waktu seperti ini adalah menggunakan model statistik, termasuk model seperti *Autoregressive Integrated Moving Average* (ARIMA), *Seasonal* ARIMA (SARIMA), *Seasonal* ARIMA *with Exogenous* Factors (SARIMAX), dan *Time Series Regression*. Artikel ilmiah ini bertujuan untuk membandingkan keefektifan dan keakuratan empat model tersebut dalam meramalkan berdasarkan jangka pendek dan panjang dari data konsentrasi CO2 di MLO.

Pada dasarnya, setiap model memiliki keunggulan dan kelemahan tersendiri dalam menangani aspek kompleks dari data deret waktu. Beberapa model, seperti ARIMA, cenderung kompatibel dengan linearitas data, namun tidak mampu menangani data musiman [4]. Oleh karena itu, ARIMA dimodifikasi menjadi SARIMA untuk mengakomodasi komponen musiman dalam analisis. Selanjutnya, SARIMA berkembang menjadi SARIMAX agar dapat menyertakan variabel eksternal atau faktor eksogen.

Pada penelitian ini, model SARIMAX adalah *hybrid model* yang bertujuan untuk meningkatkan akurasi prediksi pada data time series, pendekatan yang diadopsi adalah penggabungan informasi dari dua jenis model, yakni *Time series Regression* (TSR) dengan metode *Ordinary Least Squares* (OLS) dan SARIMA. Hal ini memungkinkan peningkatan kemampuan model dalam menghadapi beragam jenis data. Penelitian ini akan fokus pada perbandingan kinerja berbagai model ini, yang akan memberikan wawasan mendalam tentang seberapa baik kemampuan setiap model dalam memprediksi tren konsentrasi CO2. Prediksi ini memiliki peranan fundamental dalam memahami perubahan iklim global. Hasil perbandingan ini diharapkan memberikan gambaran yang jelas mengenai

kehandalan masing-masing model dalam mendukung upaya pemahaman dan proyeksi terkait dampak perubahan iklim.

II. TINJAUAN PUSTAKA

A. *Time Series forecasting*

Time series forecasting adalah proses menggunakan model statistik atau pembelajaran mesin untuk memprediksi nilai masa depan dalam sebuah seri waktu berdasarkan informasi historis [5]. *Time series forecasting* melibatkan penggunaan model statistik atau *machine learning* untuk meramalkan nilai di masa depan dalam sebuah *time series*, didasarkan pada informasi historis. *Time series* atau deret waktu merujuk pada rangkaian data yang terukur secara berurutan sepanjang periode waktu tertentu [6]. Dalam bidang-bidang seperti bisnis, ekonomi, keuangan, dan ilmu cuaca, analisis peramalan *time series* sangat penting karena memberikan prediksi yang berguna untuk pengambilan keputusan yang berbasis data. Tujuan utamanya adalah untuk memperkirakan tren, pola, dan hubungan masa depan dari data yang diamati.

Tabel 1. Struktur ACF dan PACF [7]

Model	ACF	PACF
AR (*p*)	Turun cepat secara eksponensial (*dies down*)	Terpotong setelah *lag* ke-*P*
MA (*q*)	Terpotong setelah *lag* ke-*Q*	Turun cepat secara eksponensial (*dies down*)
ARMA (*p,q)*	Turun cepat secara eksponensial (*dies down*)	Turun cepat secara eksponensial (*dies down*)
AR (*p*) atau MA (*q*)	Terpotong setelah *lag* ke-*q*	Terpotong setelah *lag* ke-*p*

* Lag – *urutan angka yang mewakili jarak atau interval antara nilai variabel pada waktu tertentu dengan nilai-nilai pada waktu sebelumnya*

* PACF – *Partial Autocorrelation Function*

* ACF - *Autocorrelation Function*

Dalam analisis *time series*, terdapat beberapa komponen penting yang menjadi fokus utama untuk memahami dinamika data tersebut. Pertama, ada *trend* yang menggambarkan arah perubahan jangka panjang dalam data, bisa berupa peningkatan atau penurunan yang konsisten. Selanjutnya, *Seasonality* merujuk pada pola berulang yang terjadi pada interval waktu tertentu seperti harian, mingguan, bulanan, atau tahunan, yang dapat diprediksi. Selain itu, *Cyclic* adalah fluktuasi jangka panjang yang tidak teratur, seringkali terkait dengan fenomena ekonomi atau tren jangka panjang lainnya. Terakhir, ada *Noise* atau variabilitas acak yang merupakan bagian dari data *time series* yang tidak dapat dijelaskan oleh model, sering kali dianggap sebagai data yang tidak terstruktur.

B. *Uji Stationeritas Augmented Dicky-Fuller*

Sebelum melakukan pemodelan *time series*, penting untuk memastikan bahwa data telah mencapai kondisi stasioner terhadap *varians* dan *mean*. Dalam konteks deret waktu, stasioner merujuk pada kondisi di mana fluktuasi data berada dalam kisaran nilai rata-rata yang tetap dan variansnya konstan. Dengan kata lain, deret waktu dianggap stasioner jika tidak terdapat tren perubahan yang signifikan dalam nilai rata-rata (*mean*) dan variabilitas (*varians*) [10]. Hal ini menjadi aspek krusial sebelum melakukan analisis atau pemodelan lebih lanjut terhadap data deret waktu.

Pada penelitian ini, pengujian stationeritas data dilakukan menggunakan metode *Augmented Dickey-Fuller* (ADF). Tujuan dari pengujian stationeritas ini adalah untuk mencegah terjadinya regresi palsu (*spurious regression*). Pada pengujian ADF, nilai statistik yang dihitung untuk koefisien y tidak menggunakan nilai t dari tabel distribusi t biasa dengan derajat kebebasan dari jumlah observasi dan tingkat signifikansi tertentu. Sebaliknya, nilai statistik ADF yang relevan digunakan untuk menghindari terjadinya over-rejection terhadap hipotesis nol. Penggunaan nilai kritis dari tabel distribusi t dalam konteks ini dapat menyebabkan kesimpulan yang keliru, di mana data dianggap stasioner padahal sebenarnya tidak. Ini menunjukkan pentingnya menggunakan nilai statistik yang tepat seperti ADF dalam menentukan stationeritas data deret waktu.

Dalam uji akar unit ADF pada level bila menghasilkan kesimpulan bahwa data tidak stasioner maka diperlukan proses diferensi data. persamaan diferensi data adalah sebagai berikut:

$$X'_t = X_t - X_{t-1} \quad (1)$$

Langkah-langkah pengujian ADF sebagai berikut:
Hipotesis:
H_0: data tersebut tidak stasioner.
H_a: data tersebut stasioner.

Pengambilan keputusan dilakukan dengan kriteria:

1. Jika Augmented Dickey-Fuller (ADF) test statistic > Test Critical Values (*critical value* α = 5%) maka H0 ditolak.
2. Jika Augmented Dickey-Fuller (ADF) test statistic < Test Critical Values (*critical value* α = 5%) maka H0 diterima.

C. *TSR Metode Ordinary Least Square (OLS)*

Time Series Regression (TSR) pada dasarnya memiliki struktur serupa dengan regresi umum, namun perbedaannya terletak pada variabel dependen dan independennya yang merupakan data berurutan dalam waktu [9]. Model TSR digunakan ketika parameter yang menjelaskan tren naik atau turun dari data berurutan tetap konstan. Selain itu, TSR juga digunakan saat data menunjukkan pola berulang dari waktu ke waktu yang disebut sebagai pola musiman. Model TSR dituliskan dalam bentuk yang memungkinkan untuk memodelkan tren serta pola musiman dalam data seri waktu. Model TSR dapat dituliskan sebagai berikut:

$$Y_t = T_t + S_t + \varepsilon_t \quad (2)\ [9]$$

Y_t: data pengamatan model TSR pada periode waktu ke-t
T_t: komponen *trend* pada periode waktu ke-t
S_t: komponen *seasonal* pada periode waktu ke-t

εt: residual pada periode waktu ke-*t*

Secara umum, model TSR mengadopsi pendekatan regresi linear yang terintegrasi dengan konsep deret waktu. Variabel dependen dalam model TSR adalah variabel yang ingin diprediksi atau dijelaskan, sedangkan variabel independen adalah faktor-faktor yang digunakan untuk menjelaskan variasi dalam variabel dependen.

Model TSR dengan trend linear adalah metode pemodelan time series regression di mana data observasi menunjukkan pola linear naik atau turun dalam tren. Persamaan model TSR trend linear dapat dirumuskan sebagai berikut:

$$Y_t = T_t + \varepsilon_t$$
$$Y_t = \beta 0 + \beta 1 t + \varepsilon_t \quad (3)\ [9]$$

di mana
Y_t : data pengamatan model TSR pada periode waktu ke- *t*
T_t : komponen *trend* pada periode waktu ke-*t*
β_0 : parameter *constant*
β_1 : parameter periode waktu
t: periode waktu
ε_t: residual pada periode waktu ke-*t*

Metode Ordinary Least Squares (OLS) merupakan teknik yang digunakan dalam penelitian ini untuk mengestimasi parameter model TSR. OLS bertujuan untuk menemukan estimasi parameter regresi dengan cara mengurangi sebanyak mungkin jumlah kuadrat dari perbedaan antara nilai-nilai yang diamati dan nilai-nilai yang diprediksi oleh model regresi [9]. Tujuan utama dari OLS adalah meminimalkan jumlah kuadrat residual (JKR), yang menggambarkan jumlah dari kuadrat kesalahan antara nilai aktual variabel dependen dengan nilai yang diprediksi oleh model regresi. Proses ini melibatkan serangkaian langkah-langkah dalam menyesuaikan model terhadap data observasi untuk menemukan parameter-parameter yang sesuai dengan model yang paling optimal.

Langkah-langkah OLS meliputi inisialisasi estimasi parameter awal pada model regresi, perhitungan kesalahan prediksi antara nilai aktual dari variabel dependen dan nilai yang diprediksi oleh model, kuadratkan setiap kesalahan prediksi dan jumlahkan untuk mendapatkan jumlah kuadrat residual (JKR), menggunakan teknik turunan parsial untuk memperbarui estimasi parameter agar meminimalkan JKR, iterasi dilakukan secara berulang untuk menyesuaikan nilai parameter hingga ditemukan nilai-nilai yang menghasilkan JKR yang minimal. Proses ini memungkinkan estimasi parameter regresi yang optimal untuk menjelaskan hubungan antara variabel dependen dan independen dalam model regresi.

D. Autoregressive Integrated Moving Average(ARIMA)

Model Autoregressive Integrated Moving Average (ARIMA) mengombinasikan elemen dari model Autoregressive (AR) dan model Moving Average (MA) melalui proses yang melibatkan langkah-langkah differencing. Melalui proses differencing sebanyak d kali, model ini terbentuk menjadi ARIMA (p, d, q) yang dijelaskan oleh persamaan 4[7].

$$\phi\phi_p(B)(1-B)^d Z_t = \theta\theta_0 + \theta\theta_q(B)a_t \quad (4)$$

Identifikasi model dugaan ARIMA dapat diperoleh dari plot *Autocorrelation Function* (ACF) dan *Partial Autocorrelation Function* (PACF) [8]. Kriteria dugaan model ARIMA dapat dilihat pada Tabel 1.

E. Seasonal-ARIMA(SARIMA)

SARIMA, yang merupakan singkatan dari Seasonal Autoregressive Integrated Moving Average, adalah sebuah metode peramalan yang digunakan untuk menganalisis dan memprediksi data deret waktu yang menunjukkan pola musiman atau ketergantungan pada waktu sebelumnya. SARIMA merupakan perluasan dari model ARIMA (Autoregressive Integrated Moving Average), dengan penambahan komponen musiman untuk menangkap fluktuasi yang terjadi pada interval waktu tertentu, seperti bulanan atau tahunan [11].

Model SARIMA didefinisikan dengan tiga parameter utama dan tiga parameter musiman yaitu AR (Autoregressive) Parameter yang menunjukkan hubungan antara nilai saat ini dengan nilai-nilai sebelumnya dalam deret waktu. I (Integrated) Parameter yang menunjukkan jumlah diferensiasi yang diperlukan untuk membuat deret waktu menjadi stasioner, yaitu deret waktu yang sifat statistiknya tidak berubah seiring waktu. MA (Moving Average) parameter yang menunjukkan hubungan antara nilai saat ini dengan kesalahan prediksi dari nilai-nilai sebelumnya.

Parameter musiman dari SARIMA menambahkan aspek musiman ke dalam model ARIMA dengan mempertimbangkan pola yang berulang pada interval waktu yang sama setiap tahun atau periode lainnya. Ini memungkinkan model untuk menyesuaikan diri dengan pola musiman yang mungkin tidak terdeteksi oleh model ARIMA non-musiman.

Bentuk umum dari model SARIMA (p, d, q) (P, D, Q, S) dapat dinyatakan seperti persamaan 5 berikut:

$$\phi_p(B)\Phi_P(B^S)(1-B)^d(1-B^S)^D X_t = \theta_q(B)\Theta_Q(B^S)\alpha_t \quad (5)\ [12]$$

F. Seasonal-ARIMA with Exogenous Factors(SARIMAX)

Model SARIMAX adalah pengembangan dari model SARIMA yang memungkinkan integrasi variabel eksogen atau variabel penjelas tambahan untuk meningkatkan performa atau kinerja dalam meramal. Model SARIMAX ini merupakan versi multivariat dari model ARIMA Musiman dengan adanya faktor eksogen. Secara matematis, model ini sering diungkapkan sebagai:

$$\phi_p(B)\Phi_P(B^s)\nabla^d\nabla_s^D y_t = \beta_k x'_{k,t} + \theta_q(B)\Theta_Q(B^s)\varepsilon_t \quad (6)\ [13]$$

Dimana $x'_{k,t}$ adalah vektor termasuk variabel masukan penjelas k^{th}pada waktu ke t dan β^k adalah nilai koefisien dari k^{th}

variabel masukan eksogen. Kondisi stasioneritas dan invertibilitas sama dengan model ARMA.

Dalam implementasi yang diselidiki, variabel respons dari model SARIMAX adalah produksi PV, sementara variabel penjelasnya adalah deret waktu dari prakiraan radiasi matahari yang diperoleh dari model NWP. Karena data memiliki karakteristik nonstasioner dan termasuk pola musiman dalam interval 24 jam, satu aspek kunci untuk membangun model SARIMAX yang berhasil adalah melakukan differencing terhadap rangkaian waktu respons dan eksogen sebelum melakukan estimasi model. Jika langkah differencing ini tidak dilakukan, ada risiko yang disebut sebagai *spurious regression* [13]. Pemilihan orde musiman (P, D, Q) dan non-musiman (p, d, q) dalam model SARIMAX didasarkan pada beberapa kriteria seperti kriteria informasi seperti AIC (Akaike Information Criterion) dan FPE (Final Prediction Error). Selain itu, plot autokorelasi (ACF) dan autokorelasi parsial (PACF) juga diobservasi untuk membantu mengevaluasi dan memilih model yang sesuai. Meskipun demikian, dalam menentukan model akhir SARIMAX yang memiliki kinerja terbaik, keputusan biasanya didasarkan pada perbandingan nilai rata-rata NRMSE (Normalized Root Mean Squared Error) secara tahunan di antara berbagai model yang dievaluasi. Model yang memberikan NRMSE tahunan terendah dipilih sebagai model yang paling sesuai untuk analisis dan prediksi yang lebih akurat.

G. Optimizer Hyperparameters ARIMA dan SARIMA

Proses optimasi dalam model ARIMA dan SARIMA adalah langkah krusial dalam memilih parameter terbaik untuk model. Dalam konteks ARIMA, optimasi melibatkan penentuan nilai optimal untuk parameter non-musiman, yaitu p, d, dan q, yang mewakili order dari komponen *autoregressive, differencing, dan moving average*. Pada penelitian ini metode *greedy search* digunakan untuk mengevaluasi berbagai kombinasi nilai p, d, dan q dengan tujuan menemukan model yang memberikan kinerja terbaik berdasarkan kriteria seperti Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), serta Mean Squared Error (MSE) atau Root Mean Squared Error (RMSE). Sementara itu, pada model SARIMA, proses optimasi juga melibatkan penentuan nilai parameter tambahan untuk komponen musiman, yaitu P, D, dan Q. Langkah ini juga memanfaatkan teknik serupa yakni *greedy search* untuk menemukan kombinasi nilai terbaik untuk P, D, dan Q serta parameter non-musiman. Fokus utamanya adalah mencari model SARIMA yang paling sesuai dengan deret waktu, mempertimbangkan karakteristik musiman dan non-musiman. Dengan mengevaluasi berbagai kombinasi nilai parameter ini, proses optimasi membantu dalam menentukan model yang tepat dan sesuai dengan data deret waktu yang dianalisis, sehingga dapat memberikan prediksi yang akurat dan relevan.

H. Kriteria Model Terbaik

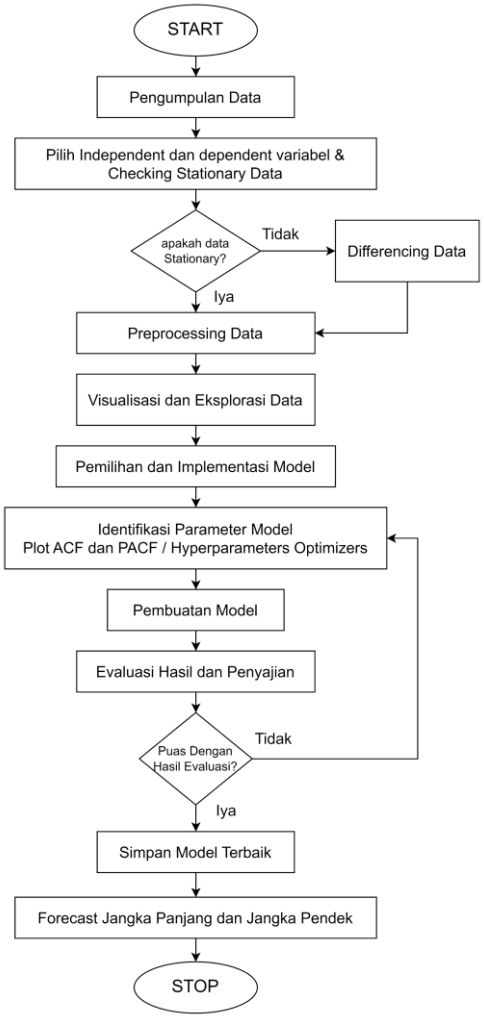
Pada penelitian ini berkaitan dengan pemodelan deret waktu konsentrasi CO2 di Mauna Loa Observatory (MLO), penting untuk memahami kriteria yang digunakan untuk memilih model terbaik. Beberapa model statistik seperti ARIMA, SARIMA, SARIMAX, dan Time Series Regression dievaluasi untuk meramalkan konsentrasi CO2.

Parameter evaluasi seperti AIC, BIC, serta MSE atau RMSE menjadi landasan untuk mengevaluasi kinerja model. Kriteria-kriteria ini memberikan gambaran tentang kualitas dan kemampuan masing-masing model dalam meramalkan tren konsentrasi CO2 dengan akurat, baik pada jangka pendek maupun panjang.

Tabel 2. Bentuk Matematis parameter evaluasi

RMSE	$\sqrt{\frac{1}{n}}\sum_{i=1}^{n}(o_i - p_i)^2$	(7) [15]
AIC	$2\left[l(\hat{\theta}_2) - l(\hat{\theta}_1)\right] - 2(p_2 - p_2)$	(8) [14]
BIC	$2\left[l(\hat{\theta}_2) - l(\hat{\theta}_1)\right] - \log n(p_2 - p_2)$	(9) [14]

III. METODE PENELITIAN

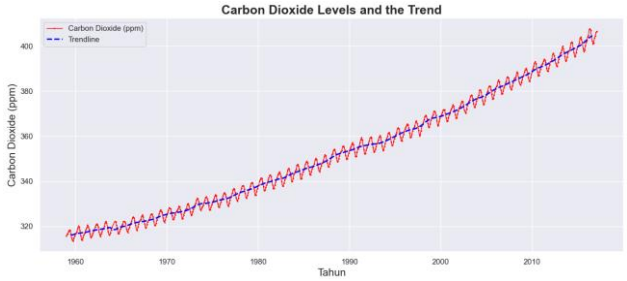


Gambar 1. *Flowchart pengerjaan pengerjaan peneltian terhadap Data dan Model yang dipilih*

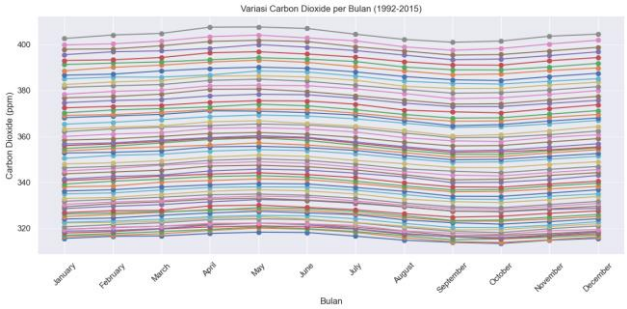
A. Deskripsi Data

Data yang menjadi fokus penelitian ini adalah konsentrasi karbon dioksida (CO2) yang telah tercatat sejak tahun 1958 hingga 2017 dari MLO. Deret waktu yang dihasilkan dari

pengukuran ini menjadi fondasi utama dalam memahami perubahan konsentrasi CO2 dalam atmosfer selama beberapa dekade. Pengumpulan data ini menjadi penting karena memberikan gambaran yang jelas tentang perubahan konsentrasi CO2, yang merupakan komponen penting dalam analisis dampak perubahan iklim global.



Gambar 2. *Line chart* visual data *Carbon Dioxide Level* beserta *trend line* dari data secara tahunan



Gambar 3. *Line chart* visual data *Carbon Dioxide Level* tiap bulannya

Gambar 2 menggambarkan data tingkat Karbon Dioksida yang menunjukkan kecenderungan peningkatan dengan nilai rata-rata yang stabil sepanjang rentang data. Di sisi lain, Gambar 3 menampilkan grafik garis yang menggambarkan pola bulanan dalam data ini, menegaskan keberadaan sifat musiman dalam data tersebut. Terlihat gelombang periodik pada setiap bulan dalam setiap tahunnya, dengan konsistensi kenaikan mulai dari Januari menuju puncaknya pada April, diikuti oleh penurunan dari April hingga September. Lalu, dari September hingga Desember, terjadi kenaikan kembali yang mengikuti pola ini secara berulang setiap tahunnya.

Tabel 3. Data setelah dilakukan *prepocessing* yang akan menjadi fokus utama penelitian

Date	Carbon Dioxide (ppm)	Tahun	Bulan	Hari
1959-01-01	315.0411	1959	January	Thursday
1959-02-01	316.4800	1959	February	Sunday
⋮	⋮	⋮	⋮	⋮
2016-11-01	403.6400	2016	November	Tuesday
2016-12-01	404.5500	2016	December	Thursday

[a]ppm – *parts per million.*

B. *Prepocessing Data*

Pada tahap ini, data konsentrasi CO2 dari MLO disiapkan untuk analisis lebih lanjut. Langkah-langkah melibatkan pemeriksaan kualitas data, penanganan data yang hilang, transformasi data, dan penyesuaian format. Dilakukan pengecekan terhadap entri data yang hilang, duplikat, dan outlier menggunakan metode Interquartile Range (IQR). Hasilnya menunjukkan data sangat bersih dengan hanya 2.36% data yang hilang, tanpa adanya outlier atau duplikasi. Transformasi data akan dilakukan menggunakan metode One-hot encoding untuk penggunaan pada model SARIMAX.

Data dalam tabel 3 telah dibagi menjadi dua bagian, yakni data latih (*training* data) dan data uji (*test* data), menggunakan metode *Hold-Out*. Metode *Hold-Out* ini memisahkan dataset menjadi dua bagian terpisah, di mana satu bagian digunakan untuk melatih model (data latih) dan bagian lainnya untuk menguji performa model (data uji) [16]. Pembagian data dilakukan dengan rasio 80:20, dimana data latih menyertakan 80% dari dataset (meliputi rentang waktu selama 46 tahun dari 1959-01-01 hingga 2004-12-01), dan data uji terdiri dari 20% sisanya (meliputi rentang waktu 11 tahun dari 2005-01-01 hingga 2016-12-01).

C. *Varibel Penelitian*

Variabel penelitian yang digunakan terdapat dalam tabel 3. Variabel "Carbon Dioxide (ppm)" adalah variabel dependen yang ingin diprediksi atau dijelaskan oleh variabel lain dalam suatu model. Variabel independen atau eksogen merupakan variabel yang digunakan untuk menjelaskan atau memprediksi variabel dependen.

Dalam dataset yang digunakan, hanya terdapat satu variabel yang diberikan, yaitu "Carbon Dioxide (ppm)," bersama dengan informasi tentang tanggal (Date), tahun (Tahun), bulan (Bulan), dan hari (Hari). Dalam kasus ini, tidak ada variabel independen yang diberikan secara eksplisit. Namun, dalam analisis deret waktu, seringkali variabel waktu (Date) dapat dijadikan representasi dari variabel eksogen dengan melakukan teknik seperti *one hot encoding*. Dengan demikian, variabel waktu (Date) dapat dianggap sebagai variabel independen yang mewakili faktor eksogen dalam analisis deret waktu.

D. *Langkah Analisis*

Pemodelan data konsentrasi CO2 di MLO menggunakan ARIMA, SARIMA, TSR OLS, dan SARIMAX diuraikan dalam flowchart pada Gambar 1. Terdapat perbedaan dalam persiapan data dan penentuan parameter/orde dari keempat model tersebut. Untuk SARIMA dan ARIMA, masing-masing menggunakan satu variabel dependen dan independen. Penentuan orde atau identifikasi parameter model ini memanfaatkan plot acf dan pacf, dengan syarat bahwa data harus bersifat stationer terlebih dahulu.

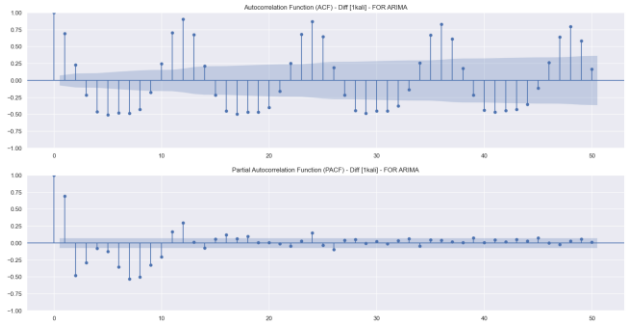
TSR metode OLS tidak memerlukan penentuan orde seperti pada tiga model sebelumnya. Pada SARIMAX, pendekatan menggabungkan hasil dari OLS dan SARIMA, di mana OLS meramalkan data konsentrasi CO2 di MLO sementara SARIMA model meramalkan residual dari OLS. Meskipun tidak secara eksplisit disebut sebagai model SARIMAX, kombinasi antara SARIMA dan OLS dapat dianggap sebagai model SARIMAX hybrid. SARIMAX sendiri menggunakan

hasil prediksi residual dari SARIMA ditambah dengan hasil prediksi dari OLS.

Setelah pembuatan model, dilakukan peramalan untuk jangka pendek dan panjang. Peramalan jangka pendek dilakukan untuk melihat tren atau pola data dalam waktu satu tahun ke depan, sementara peramalan jangka panjang dilakukan untuk mengamati tren atau pola dalam lima tahun ke depan. Diharapkan bahwa model yang baik mampu mengikuti dan merepresentasikan pola yang ada pada data sebelumnya, memungkinkan untuk memberikan estimasi yang akurat untuk kedua jangka waktu tersebut.

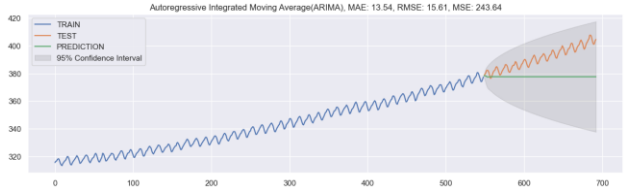
IV. ANALISIS DAN PEMBAHASAN

A. Peralaman data CO2 (ppm) menggunakan metode ARIMA dan Optimizer hyperparameters ARIMA



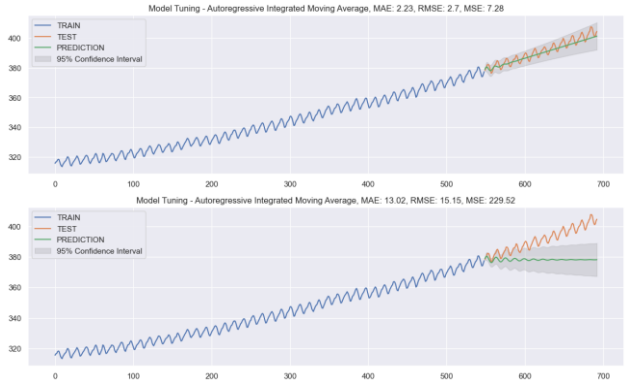
Gambar 4. *Plot ACF dan PACF terhadap data differencing(1) data yang stationer*

Berdasarkan analisis dari plot gambar 4, data asli terlihat pola musiman sehingga jika kita ingin mencari order dari Non Seasonal tidak akan ada pola yang cocok untuk model. Plot PACF dengan *Cut Off* pada lag 2 pada PACF dan lonjakan pada lag kelipatan 12. Ini menunjukkan bahwa data ini memiliki sifat musiman, sehingga asumsinya model ARIMA cocok. Plot ACF menunjukkan pola *Diesdown sinusoidal extremely slowly*, menandakan bahwa data cenderung tidak stasioner meskipun sudah dilakukan differencing dan uji ADF. Meski demikian, untuk memeriksa kecocokan, saya akan mempertimbangkan ARIMA dengan asumsi terdekat, adalah *Model AR*(2) tanpa adanya MA karena ACF *Diesdown Sinusoidal not extremely slowlydan*. PACF Cut off lag ke-2.



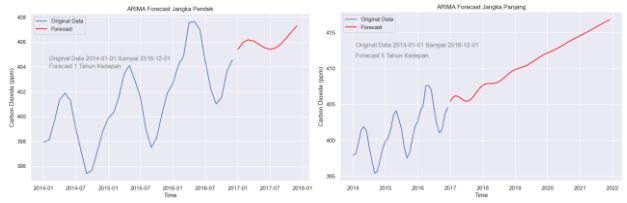
Gambar 5. *Hasil prediksi base model ARIMA(2,1,0) Terhadap data test*

Terlihat pada gambar 5 hasil prediksi base model ARIMA dengan menggunakan orde pengamatan dari plot ACF dan PACF sangat jauh dari kata baik, dengan RMSE 15.61 dan pada 95% Confidene Interval. Untuk menginterpretasikan interval kepercayaan, kita harus memahami bahwa interval tersebut memberikan kisaran di mana nilai sebenarnya diharapkan berada pada tingkat kepercayaan yang ditentukan. Semakin sempit margin suatu interval, semakin tinggi keakuratan perkiraannya [17].



Gambar 7. *Optimizer ARIMA (Atas) berdasarkan MAE, (bawah) berdasarkan AIC*

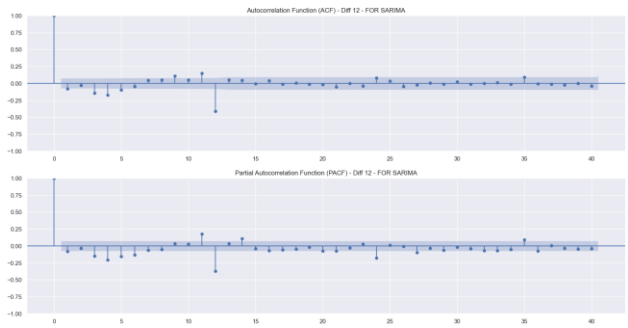
Optimizer pada model ARIMA adalah komponen kunci yang bertanggung jawab untuk menemukan parameter terbaik yang sesuai dengan data dalam analisis deret waktu. Pada gambar 7 atas terlihat bahwa RMSE lebih baik dengan nilai 2.7 dari gambar 7 bawah sekitar 15.15. Pada gambar atas mendapatkan orde ARIMA (3, 2, 2) dan gambar bawah mendapatkan orde ARIMA (3, 1, 3) dengan percobaan random dari p, d, q masing masing 0 hingga 3.



Gambar 8. *Forecasting model terbaik ARIMA terhadap data CO2 (kiri) jangka pendek, (kanan) jangka panjang*

Gambar 8 adalah hasil plot dari forecasting data untuk ARIMA terhadap CO2 (ppm). Model yang digunakan untuk forecasting adalah model terbaik yang telah didapatkan sebelumnya yakni ARIMA (3, 2, 2). Meskipun hasil kurang memuaskan dikarekan data asil CO2(ppm) adalah seasonal dan ARIMA tidak bisa menangani data seasonal maka dari kita butuh bantuan SARIMA.

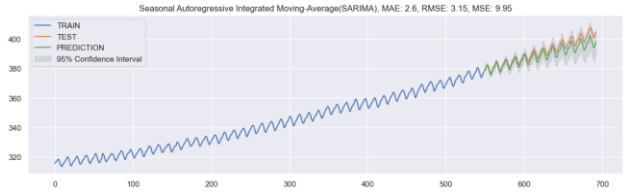
B. Peralaman data CO2 (ppm) menggunakan metode SARIMA dan Optimizer hyperparameters SARIMA



Gambar 9. *Plot ACF dan PACF terhadap data differencing(12) data yang stationer*

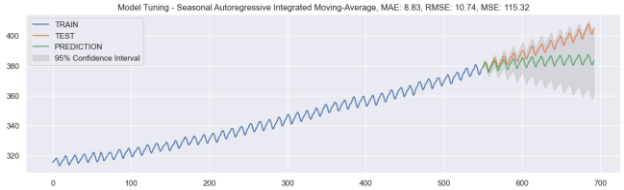
Berdasarkan plot ACF dan PACF dari data yang telah di-differencing dengan interval 12, terlihat adanya pola yang menunjukkan sifat *stationary data* dengan pola musiman tahunan (dengan nilai m = 12).

Dari plot PACF, terdapat peningkatan yang signifikan pada lag 12 dan 24, menunjukkan adanya pola *Spike Lag* pada setiap kelipatan 12. Ini mengindikasikan bahwa P = 2 dapat dipertimbangkan sebagai orde *autoregressive* (AR) untuk model SARIMA. Sementara pada plot ACF, terdapat spike hanya pada lag 12, mengindikasikan bahwa Q = 1 bisa menjadi orde *moving average* (MA) untuk model SARIMA.



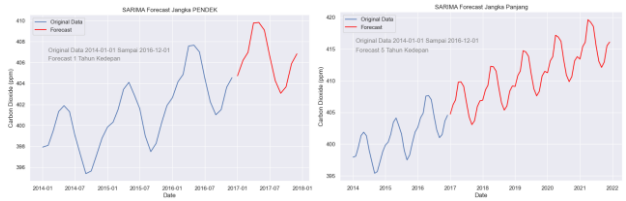
Gambar 10. *Hasil prediksi base model SARIMA(0,1,0)(2,1,1,12) Terhadap data test*

Terlihat pada gambar 10 hasil prediksi SARIMA sangat jauh lebih baik dari ARIMA pada gambar 5, dengan RMSE 3.15 dan pada 95% Confidene Interval. Lalu disini terlihat juga untuk range dari interval kepercayaannya cukup kecil sehingga merepresentasikan bahwa model meliki kepastian dalam menentukan peramalannya. Jika dibandingkan dengan ARIMA, terlihat bahwa garis yang dihasilkan juga memiliki lengkungan yang berarti memiliki tren musiman pada hasil *forecast.*



Gambar 11. Forecast *Optimizer SARIMA berdasarkan AIC terhadap data test*

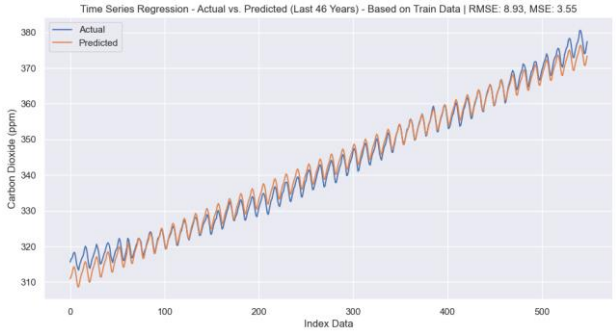
Telihat pada gambar 11 telihat bahwa RMSE dan juga rentang kepercayaan yang dihasilkan sangat buruk dengan nilai RMSE sebesar 10.74 namun meski demikian, masih lebih baik dari ARIMA pada gambar 7 bawah dan gambar 5. Hal tersebut karena SARIMA dapat memasukkan komponen seasonal pada model saat memprediksi data.



Gambar 12. *Forecasting model terbaik SARIMA terhadap data CO2 (kiri) jangka pendek, (kanan) jangka panjang*

Gambar 12 adalah hasil plot dari forecasting data untuk SARIMA terhadap CO2 (ppm). Model yang digunakan untuk forecasting adalah model terbaik yang telah didapatkan sebelumnya yakni SARIMA (0,1,0) (2,1,1,12). Terlihat bahwa SARIMA dapat meramalkan data musiman dengan sangat baik, Garis merah yang merepresentasikan hasil *forecast*, dan garis biru yang mererpesentsaikan aktual data memiliki kesinambungan dalam bentuk pola garis dan trend naik nya. Sehingga terbukti bahwa SARIMA lebih baik dari ARIMA dalam kasus peralaman jangka panjang dan pendek untuk data CO2 (ppm) MLO.

C. Peralaman data CO2 (ppm) menggunakan model TSR metode OLS



Gambar 13. *Hasil prediksi model TSR dengan metode OLS terhadap data train*

Pada model machine learning seperti TSR OLS, berbeda dengan model statistik ARIMA dan SARIMA yang menggunakan satu variabel independen dan satu variabel dependen, TSR OLS menggunakan one hot encoding dari bulan 1 hingga 12 sebagai variabel independen. Hasil plot aktual dan prediksi pada data latih menunjukkan kualitas yang baik karena model dilatih dengan data yang sama. RMSE yang dihasilkan adalah sebesar 8.93.

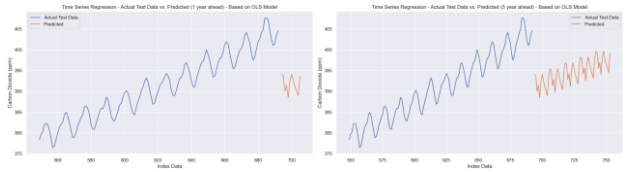
Gambar 14. *Hasil prediksi model TSR dengan metode OLS terhadap data test*

Namun, terdapat perbedaan yang signifikan antara plot data prediksi dan aktual pada gambar 14 berdasarkan data uji. Untuk mengatasi hal ini, penelitian ini akan menggunakan pendekatan hybrid model. Residual dari hasil prediksi TSR OLS terhadap data CO2 (ppm) akan dimasukkan ke dalam model SARIMA untuk meramalkan residual tersebut baik dalam jangka panjang maupun jangka pendek. Dengan demikian akan ada persamaan dibawah ini sebagai bentuk SARIMAX hybrid model:

$$SARIMAX(zt) = Prediction_t + \varepsilon_t \quad (10)$$

$Prediction_t$ = Hasil prediksi TSR OLS terhadap data ke t
ε_t = prediksi residual TSR OLS berdasakan model SARIMA

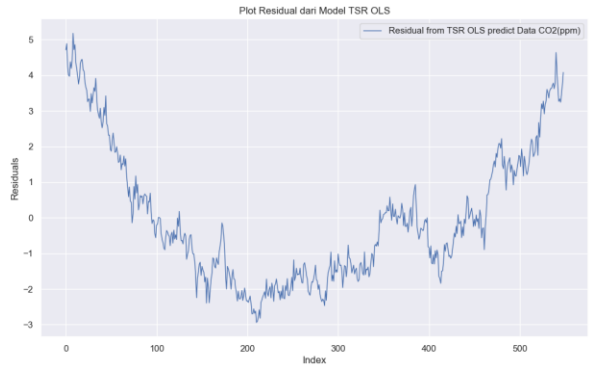
Diharapkan pendekatan ini akan meningkatkan performa prediksi dan menghasilkan hasil prediksi yang lebih akurat.



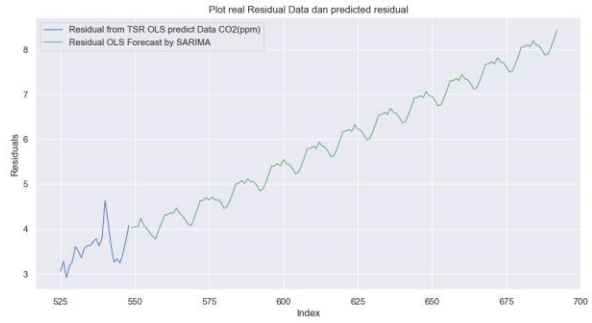
Gambar 15. *Forecasting model TSR OLS terhadap data CO2 (kiri) jangka pendek, (kanan) jangka panjang*

Pada gambar 15 terlihat bahwa model OLS kurang efektif dalam memprediksi jangka panjang maupun jangka pendek. Garis merah menunjukkan hasil prediksi sedangkan garis biru adalah data aktual. Kedua garis tersebut tidak menunjukkan pola yang konsisten, bahkan terputus-putus tanpa adanya kelanjutan tren. Diharapkan dengan penerapan hybrid model, akan meningkatkan akurasi dalam meramalkan data CO2 (ppm) ke depan.

D. Peralaman data CO2 (ppm) menggunakan model hyrbrid OLS + SARIMA(SARIMAX model)

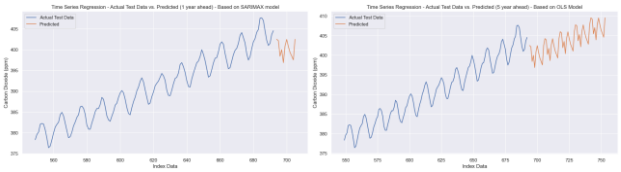


Gambar 16. *Plot residual TSR OLS*



Gambar 17. *Plot hasil perdiksi SARIMA residual dengan residual asli*

Gambar 16 menampilkan residual yang dihasilkan oleh model TSR OLS saat meramalkan data CO2 (ppm). Sedangkan pada gambar 17, terlihat hasil prediksi SARIMA terhadap data residual dengan langkah-langkah yang serupa seperti pada sub bab sebelumnya. Data residual ini diuji untuk dijadikan stasioner dan kemudian dianalisis dengan plot ACF dan PACF untuk menentukan orde yang sesuai. Setelah proses tersebut, diperoleh orde SARIMA yaitu (1,1,1) (4,1,1,12).



Gambar 18. *Forecasting SARIMAX terhadap data CO2 (kiri) jangka pendek, (kanan) jangka panjang*

Hasil pada gambar 18 diatas didapatkan dari persamaan 10 untuk pemodelan SARIMAX. Hasil yang didapatkan cukup sesuai dengan harapan, lebih baik dari gambar 15. Sesuai dengan harapan ketika menggunakan hybrid model. Namun tetap saja terlihat untuk peramalan data CO2(ppm) lebih baik menggunakan model statistik SARIMA. Dengan hasil yang lebih memuaskan dan sesuai.

V. KESIMPULAN/RINGKASAN

Kesimpulan dari analisis perbandingan antara model SARIMA, ARIMA, SARIMAX, dan Time Series Regression (TSR) metode OLS dalam meramalkan konsentrasi karbon dioksida (CO2) dari Observatorium Mauna Loa adalah bahwa model SARIMA unggul dalam memprediksi konsentrasi CO2 dibandingkan dengan ARIMA. Meskipun demikian, pendekatan hybrid SARIMAX dengan penggabungan residual dari model TSR OLS ke dalam model SARIMA berhasil meningkatkan akurasi prediksi, meskipun masih kalah efektif dibandingkan dengan SARIMA murni. Rekomendasi untuk penelitian selanjutnya adalah mengeksplorasi lebih banyak orde dalam penggunaan optimizer hyperparameters serta memperluas komposisi dalam metode hold-out. Jika data yang digunakan bersifat musiman, saran tersebut juga menyarankan untuk menghindari penggunaan model ARIMA sebagai model prediksi.

UCAPAN TERIMA KASIH

Penulis, R.N.S. ingin mengungkapkan rasa terima kasih yang mendalam kepada laptop saya dan diri saya sendiri telah kuat dan berhasil menyelesaikan artikel hingga tuntas. Tidak lupa, terima kasih kepada dosen pengampu, Ibu Regita Putri Permata, S.Stat., M. Stat., atas bimbingan, arahan, serta ilmu yang berharga dalam mata kuliah Analisis Deret & Waktu yang telah sangat membantu dalam kelancaran penyelesaian tugas besar ini. Semoga upaya dan dedikasi ini dapat memberikan kontribusi yang bermanfaat dalam bidang ilmu pengetahuan.

DAFTAR PUSTAKA

[1] T. Susana, "KARBON DIOKSIDA," *Oseana,* vol. XIII, no. 1, 1988.

[2] K. W. THONING, P. P. TANS and W. D. KOMHYR, "Atmospheric Carbon Dioxide at Mauna Loa Observatory to Analysis of the NOAA GMCC Data, 1974-1985," *Journal of Geophysical Research: Atmospheres,* vol. XCIV, pp. 8549-8565,, 20 June 1989.

[3] L. Tipton, G. Zahn, E. Datlof, S. N. Kivlin, P. Sheredan, A. S. Amend and N. A. Hynson, "Fungal aerobiota are not affected by time nor environment over a 13-y time series at the Mauna Loa Observatory," *PNAS*, December 2019.

[4] A. H. Adineh, Z. Narimani and S. C. Satapathy, "Importance of data prepocessing in time series prediction using SARIMA: A case study," *International Journal of Knowledge-based and Intelligent Engineering Systems*, pp. 331-342, 2020.

[5] C. Borui, S. Yang, L. Gao and . X. Yong, "Hybrid variational autoencoder for time series forecasting," *Elsevier B.V.*, 20 October 2023.

[6] C. M. Douglas , L. J. Cheryl and M. Kulahci, Introduction to Time Series Analysis and Forecasting, Hoboken: Library of Congress Cataloging-in-Publication Data applied for., 2015.

[7] A. Fahrila and M. Prastuti, "Peramalan Konsumsi Energi Listrik untuk Sektor Industri di PT PLN (Persero) Area Gresik Menggunakan Metode Time Series Regression," *JURNAL SAINS DAN SENI ITS*, vol. XII, no. 1, 2023.

[8] W. Wei, Time Series Analysis: Univariate and Multivariate Methods, Pearson Addison Wesley, 2006.

[9] K. Ramadani, S. Wahyuningsih and M. Nor, "Pemodelan Harga Saham PT. Telekomunikasi Indonesia Tbk," *Jurnal EKSPONENSIAL*, vol. XIII, no. 1, May 2022.

[10] Z. S. Jingga and M. Prastuti, "Peramalan Jumlah Penumpang Pesawat Domestik di Bandara Soekarno-Hatta pada Masa Pandemi Covid-19 Menggunakan ARIMAX dengan Model Intervensi," *JURNAL SAINS DAN SENI ITS*, vol. XII, no. 1, 2023.

[11] A. E. Permanasari, I. Hidayah and I. A. Bustoni, "SARIMA (Seasonal ARIMA) Implementation on Time Series to Forecast The Number of Malaria Incidence," *IEEE*, 2013.

[12] A. Zaki, M. S. Wahyuni, I. W. Nari and A. Real, "Peramalan Jumlah Penderita Demam Berdarah Dengue Menggunakan Metode Seasonal-ARIMA," *ARRUS Journal of Mathematics and Applied Science*, vol. III, no. 2, 2023.

[13] I. V. Stylianos, . I. C. G., G. K. E., K. S. C. and G. B. A., "Comparison of SARIMAX, SARIMA, Modified SARIMA and ANN-based Models for Short-Term PV Generation Forecasting," *IEEE International Energy Conference (ENERGYCON)*, April 2016.

[14] J. Kuha, "AIC and BIC Comparisons of Assumptions and Performance," *SOCIOLOGICAL METHODS & RESEARCH*, vol. XXXIII, no. 2, pp. 188-229, November 2004.

[15] R. Pelanek, "Metrics for Evaluation of Student Models," *Journal of Educational Data Mining,*, vol. VII, no. 7, 2015.

[16] N. A. C.A, D. H. Citra, W. Purnama, C. Nisa and A. R. Kurnia, "Implementasi Algoritma Naive Bayes untuk Analisis Sentimen Ulasan Shopee pada Google Play Store," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. II, no. 1, pp. 47-54, April 2022.

[17] A.-M. r. Šimundic, "Confidence Interval," *Lessons in biostatistics*, p. 54–61, 1 April 2008.