Vector AutoRegression model - forecasting of statistical climate data

Paweł Siemiginowski[a], BEng (Student), Paweł Cedzich[a], BEng (Student), Marek Pietrowicz[a], BEng (Student) and Jan Kłek[a], BEng (Student)

[a]*Opole University of Technology, Prószkowska 76, Opole, 45-758, Poland*

ARTICLE INFO

Keywords:
VAR
prediction
forecast
time-series

ABSTRACT

This report focuses on the use of the Vector AutoRegression (VAR) model to forecast the average energy parameters of buildings based on hourly data. The aim is to present the theoretical foundations of the VAR model and its applications in time series analysis, as well as the practical use of the VAR model to forecast the average energy parameters of buildings based on average real hourly data.

1. Introduction and Objective

This report focuses on the use of the Vector AutoRegression (VAR) model to forecast the average energy parameters of buildings based on hourly data. These data are sourced from the government website https://www.gov.pl/web/archiwum-inwestycje-rozwoj/dane-do-obliczen-energetycznych-budynkow. The VAR model, an extension of the classical autoregressive model to multiple variables, allows for the analysis of dynamic relationships between multiple time series variables simultaneously.

The VAR model is a statistical model used to capture linear interdependencies among multiple time series. In the VAR model, each variable is modeled as a function of the lags of all variables in the system, allowing for the analysis of mutual influences between variables. This is particularly useful in forecasting, where the relationships between different variables can provide valuable insights into future values.

The objectives of this report are:

- **Learning and understanding the VAR model**: Presenting the theoretical foundations of the VAR model and its applications in time series analysis.
- **Practical implementation of the VAR model**: Using the VAR model to forecast the average energy parameters of buildings based on real hourly data. This analysis will help understand how the VAR model can be used to predict future values based on historical data.

2. Literature Review

The Vector AutoRegression (VAR) model was introduced by Christopher A. Sims in the 1980s and has since found wide application in various scientific fields, including economics, finance, and engineering [2]. VAR is a natural extension of the autoregressive (AR) model to multiple variables, allowing each variable to be modeled as a function of the lags of all variables in the system [4]. VAR captures dynamic relationships between time series variables, which is particularly useful in the analysis of multivariate time series [5].

In the scientific literature, the VAR model is often used for forecasting macroeconomic variables such as inflation, interest rates, and GDP [3]. VAR is used to assess the effects of various economic policies. For example, it can analyze the impact of interest rate changes on inflation and production [4]. VAR models are also used to identify structural economic shocks and their impact on the economy [5].

Traditional VAR models can be sensitive to the presence of outliers, which can lead to incorrect conclusions [3]. In response to these challenges, various robust methods have been developed, such as reweighted multivariate least trimmed squares (RMLTS) and multivariate MM-estimation, which improve estimation accuracy in the presence of outliers [1].

In recent years, VAR models have been used to analyze volatility in financial markets, including the impact of extreme market events on the prices of safe-haven assets such as gold, silver, and currencies [2]. The VAR model is a powerful tool for analyzing multivariate time series, offering extensive capabilities in forecasting, policy analysis, and identifying structural shocks [4]. Despite certain challenges, such as sensitivity to outliers, the development of robust methods has significantly increased its usefulness and accuracy [5].

3. Methodology

3.1. Dataset

The dataset utilized in this study includes meteorological data from three Polish cities: Kołobrzeg, Opole, and Warszawa Okęcie. The dataset comprises the following columns:

- **Hour of the year (N)**: Sequential hour number within the year.
- **Month (M)**: Month of the year.
- **Day (D)**: Day of the month.

*This document is a report made for the course "Pattern Recognition" in a master's degree program.

*Authors' contact information:
p.siemiginowski@student.po.edu.pl (P. Siemiginowski);
p.cedzich@student.po.edu.pl (P. Cedzich);
m.pietrowicz@student.po.edu.pl (M. Pietrowicz);
j.klek@student.po.edu.pl (J. Kłek)

- **Hour UTC (H)**: Hour in Coordinated Universal Time (UTC).
- **Dry bulb temperature (DBT)**: Temperature measured by a dry thermometer in degrees Celsius (°C).
- **Relative humidity (RH)**: Percentage of moisture in the air relative to the maximum amount the air can hold at that temperature.
- **Humidity ratio (HR)**: Mass of water vapor per unit mass of dry air, measured in grams per kilogram (g/kg).
- **Wind speed (WS)**: Speed of the wind in meters per second (m/s).
- **Wind direction (WD)**: Direction from which the wind is blowing, categorized into 36 sectors (0 - calm, N - 36, E - 9, S - 18, W - 27, 99 - variable).
- **Total solar radiation on a horizontal surface (ITH)**: Total solar energy received per unit area on a horizontal surface, measured in watts per square meter (W/m²).
- **Direct solar radiation on a horizontal surface (IDH)**: Solar radiation received directly from the sun on a horizontal surface, measured in W/m².
- **Diffuse solar radiation on a horizontal surface (ISH)**: Solar radiation received from the sky (excluding direct sunlight) on a horizontal surface, measured in W/m².
- **Sky radiation temperature (TSKY)**: Temperature of the sky as perceived by a radiometer, measured in degrees Celsius (°C).

Columns 14-47, which contain solar radiation data for various orientations and inclinations, were excluded from the analysis due to a high amount of missing data, which negatively impacted the model's performance. The dataset was assessed for stationarity using the Augmented Dickey-Fuller (ADF) test, confirming its suitability for the VAR model.

3.2. Method

A Vector Autoregression (VAR) model was employed using the 'statsmodels' library. The VAR model is a statistical model used to capture the linear interdependencies among multiple time series. The model selection was based on the Akaike Information Criterion (AIC), which helps in selecting the model with the best fit by balancing the complexity and goodness of fit. The maximum number of lags considered was 50, ensuring that the model captures the temporal dependencies adequately.

3.3. Calculations

The calculations were performed using the VAR model from 'statsmodels'. The analysis considered 5000 hours of the year, and the model was used to forecast the next 24 hours. This approach allows for understanding the short-term dynamics and dependencies between the variables over a one-day horizon.

3.4. Evaluation Metrics

The prediction accuracy was evaluated by comparing the actual values of the next 24 hours with the forecasted values using the following metrics:

- **Mean Squared Error (MSE)**: This metric measures the average of the squares of the errors, indicating the average squared difference between the estimated values and the actual value. It is calculated as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

  where y_i are the actual values and $\hat{y}_i$ are the predicted values.

- **Mean Absolute Error (MAE)**: This metric measures the average magnitude of the errors in a set of predictions, without considering their direction. It is calculated as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

  where y_i are the actual values and $\hat{y}_i$ are the predicted values.

- **R-squared (R^2)**: This metric represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. It is calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

  where y_i are the actual values, $\hat{y}_i$ are the predicted values, and $\bar{y}$ is the mean of the actual values.

4. Results

The results of the VAR model predictions for the three cities (Opole, Kołobrzeg, and Warszawa Okęcie) are presented in this section. The predictions include dry bulb temperature (DBT), relative humidity (RH), and wind speed (WS) for the next 24 hours, based on the previous 100 hours of data.

4.1. Kołobrzeg

Figures 1-3 show the predictions and actual values for DBT, RH, and WS in Kołobrzeg. The error metrics for Kołobrzeg are summarized in Table 1, and the correlation matrix of residuals is presented in Table 2.

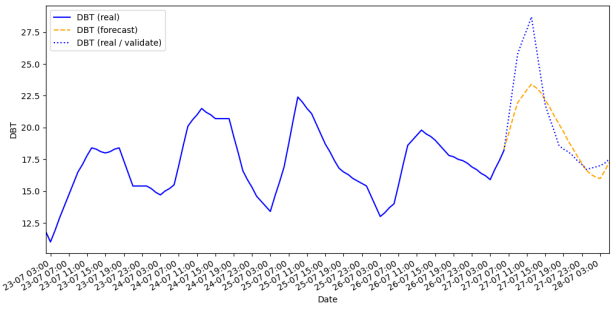

Figure 1: Kołobrzeg - DBT predictions vs actual values

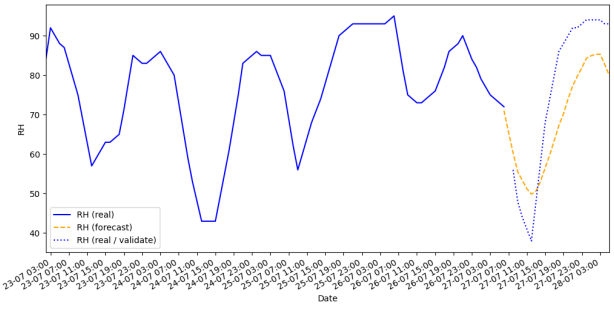

Figure 2: Kołobrzeg - RH predictions vs actual values

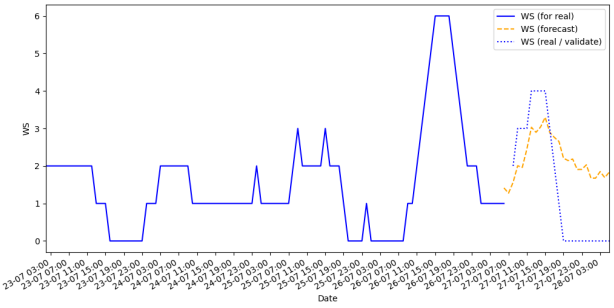

Figure 3: Kołobrzeg - WS predictions vs actual values

4.2. Opole

Figures 4-6 show the predictions and actual values for DBT, RH, and WS in Opole. The error metrics for Opole are summarized in Table 3, and the correlation matrix of residuals is presented in Table 4.

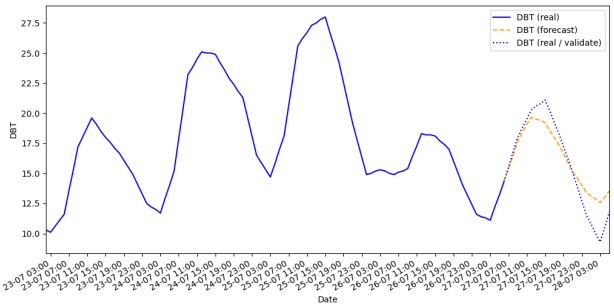

Figure 4: Opole - DBT predictions vs actual values

Table 1
Kołobrzeg - error metrics

Column	MSE	MAE	R2
DBT	4.79	1.55	0.69
RH	127.98	10.24	0.67
HR	1.48	1.07	-1.33
WS	2.11	1.29	0.16
WD	130.47	10.63	-0.06
ITH	8602.91	67.68	0.92
IDH	32323.88	111.52	0.55
ISH	9158.39	63.84	-0.04
TSKY	26.63	4.34	-4.09

Table 2
Kołobrzeg - correlation matrix of residuals

	DBT	RH	HR	WS	WD	ITH	IDH	ISH	TSKY
DBT	1.00	-0.59	0.17	0.02	-0.02	0.10	0.03	0.09	0.37
RH	-0.59	1.00	0.59	-0.01	-0.01	-0.09	-0.03	-0.09	-0.14
HR	0.17	0.59	1.00	0.00	-0.03	-0.02	-0.02	-0.01	0.13
WS	0.02	-0.01	0.00	1.00	0.09	-0.00	0.01	-0.01	0.05
WD	-0.02	-0.01	-0.03	0.09	1.00	0.01	0.03	-0.02	0.01
ITH	0.10	-0.09	-0.02	-0.00	0.01	1.00	0.63	0.51	-0.41
IDH	0.03	-0.03	-0.02	0.01	0.03	0.63	1.00	-0.34	-0.26
ISH	0.09	-0.09	-0.01	-0.01	-0.02	0.51	-0.34	1.00	-0.21
TSKY	0.37	-0.14	0.13	0.05	0.01	-0.41	-0.26	-0.21	1.00

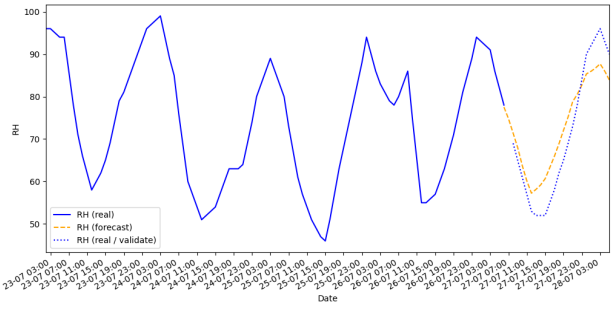

Figure 5: Opole - RH predictions vs actual values

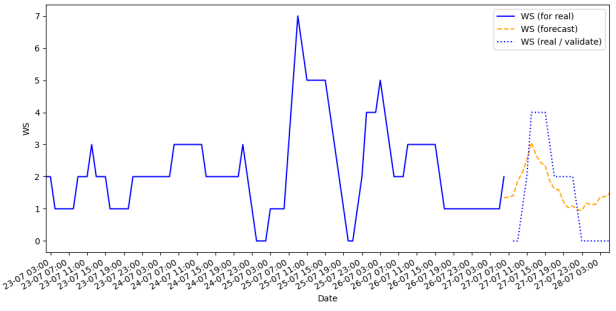

Figure 6: Opole - WS predictions vs actual values

4.3. Warszawa Okęcie

Figures 7-9 show the predictions and actual values for DBT, RH, and WS in Warszawa Okęcie. The error metrics for Warszawa Okęcie are summarized in Table 5, and the correlation matrix of residuals is presented in Table 6.

Table 3
Opole - error metrics

Column	MSE	MAE	R2
DBT	2.20	1.17	0.85
RH	32.96	5.21	0.85
HR	0.18	0.33	-0.49
WS	1.25	1.03	0.40
WD	831716.77	745.28	-4390.30
ITH	16837.69	102.70	0.66
IDH	6383.39	59.97	0.28
ISH	4243.60	52.02	0.79
TSKY	58.72	6.84	-0.46

Table 4
Opole - correlation matrix of residuals

	DBT	RH	HR	WS	WD	ITH	IDH	ISH	TSKY
DBT	1.00	-0.62	0.20	0.02	-0.02	0.19	0.09	0.13	-0.01
RH	-0.62	1.00	0.49	-0.03	0.00	-0.16	-0.06	-0.11	-0.00
HR	0.20	0.49	1.00	-0.01	-0.06	-0.02	-0.01	0.02	-0.05
WS	0.02	-0.03	-0.01	1.00	0.01	-0.02	-0.03	0.00	0.02
WD	-0.02	0.00	-0.06	0.01	1.00	-0.03	-0.32	-0.17	0.87
ITH	0.19	-0.16	-0.02	-0.02	-0.03	1.00	0.59	0.49	-0.06
IDH	0.09	-0.06	-0.01	-0.03	-0.32	0.59	1.00	-0.27	-0.40
ISH	0.13	-0.11	0.02	0.00	-0.17	0.49	-0.27	1.00	-0.19
TSKY	-0.01	-0.00	-0.05	0.02	0.87	-0.06	-0.40	-0.19	1.00

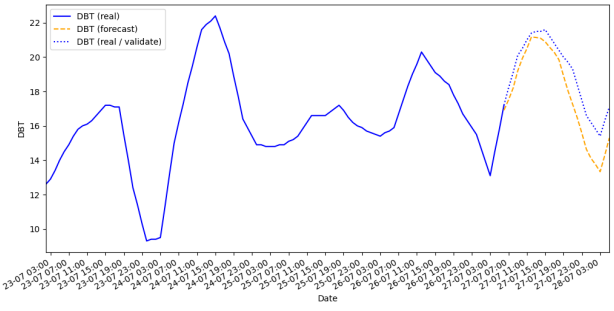

Figure 7: Okęcie - DBT predictions vs actual values

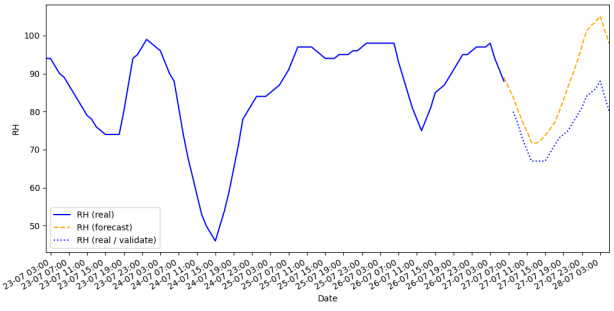

Figure 8: Okęcie - RH predictions vs actual values

5. Discussion of Results

The results indicate that the VAR model performs reasonably well in predicting dry bulb temperature (DBT) and relative humidity (RH) across all three cities. However, the predictions for wind speed (WS) are less accurate, which can be attributed to the smaller changes and fewer data points available for this variable.

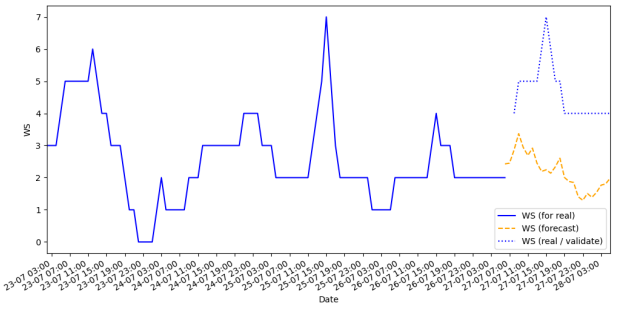

Figure 9: Okęcie - WS predictions vs actual values

Table 5
Okęcie - error metrics

Column	MSE	MAE	R2
DBT	1.84	1.16	0.55
RH	124.94	9.57	-1.57
HR	0.32	0.47	-0.09
WS	6.24	2.32	-5.25
WD	2508.54	49.45	-54.95
ITH	2055.11	38.00	0.93
IDH	1660.53	37.87	-0.80
ISH	1127.78	27.89	0.94
TSKY	2.88	1.43	0.80

Table 6
Okęcie - correlation matrix of residuals

	DBT	RH	HR	WS	WD	ITH	IDH	ISH	TSKY
DBT	1.00	-0.59	0.29	0.05	-0.01	0.23	0.09	0.15	0.37
RH	-0.59	1.00	0.45	-0.07	0.02	-0.16	-0.06	-0.12	-0.14
HR	0.29	0.45	1.00	-0.04	0.01	0.02	-0.02	0.04	0.16
WS	0.05	-0.07	-0.04	1.00	-0.00	0.00	-0.01	0.01	0.04
WD	-0.01	0.02	0.01	-0.00	1.00	-0.01	0.01	-0.03	-0.00
ITH	0.23	-0.16	0.02	0.00	-0.01	1.00	0.63	0.37	-0.38
IDH	0.09	-0.06	-0.02	-0.01	0.01	0.63	1.00	-0.48	-0.26
ISH	0.15	-0.12	0.04	0.01	-0.03	0.37	-0.48	1.00	-0.13
TSKY	0.37	-0.14	0.16	0.04	-0.00	-0.38	-0.26	-0.13	1.00

In Kołobrzeg, the model achieved an R^2 of 0.69 for DBT and 0.67 for RH, indicating a good fit. However, the R^2 for WS was only 0.16, suggesting that the model struggled to capture the variability in wind speed. The error metrics (Table 1) and the correlation matrix (Table 2) further highlight these discrepancies.

For Opole, the model performed better with an R^2 of 0.85 for both DBT and RH, but again, the R^2 for WS was lower at 0.40. The error metrics (Table 3) and the correlation matrix (Table 4) show similar patterns to those observed in Kołobrzeg.

In Warszawa Okęcie, the model's performance was mixed. While the R^2 for DBT was 0.55, indicating a moderate fit, the R^2 for RH was negative, suggesting poor predictive power. The R^2 for WS was also negative, further confirming the model's limitations in predicting wind speed. The error metrics (Table 5) and the correlation matrix (Table 6) provide additional insights into these results.

Overall, the VAR model demonstrated its capability to predict temperature and humidity with reasonable accuracy, but improvements are needed for wind speed predictions.

The correlation matrices suggest that there are significant interdependencies between the variables, which the model partially captures. Future work could focus on enhancing the model's ability to predict wind speed by incorporating additional data or using more sophisticated modeling techniques.

6. Conclusions

This study explored the application of the Vector AutoRegression (VAR) model to forecast meteorological parameters relevant to building energy calculations, using data from three Polish cities: Kołobrzeg, Opole, and Warszawa Okęcie. The analysis focused on predicting multiple variables, including dry bulb temperature (DBT), relative humidity (RH), and wind speed (WS), for the next 24 hours.

The results demonstrated that the VAR model is effective in capturing the temporal dependencies and interrelationships among a wide range of variables. The model performed well in predicting DBT and RH, with relatively high R-squared values indicating a good fit. However, the predictions for WS were less accurate, which can be attributed to the smaller changes and fewer data points available for this variable.

In Kołobrzeg, the model achieved an R-squared of 0.69 for DBT and 0.67 for RH, indicating a good fit. However, the R-squared for WS was only 0.16, suggesting that the model struggled to capture the variability in wind speed. Similar patterns were observed in Opole and Warszawa Okęcie, where the model performed better for DBT and RH but showed limitations in predicting WS.

The error metrics and correlation matrices provided further insights into the model's performance. The high correlation between certain variables, such as DBT and RH, suggests that these variables are strongly interdependent, which the VAR model was able to capture effectively. However, the lower correlation and higher error metrics for WS indicate that additional data or more sophisticated modeling techniques may be needed to improve predictions for this variable.

The VAR model's ability to handle multiple interrelated time series variables simultaneously is one of its key strengths. In this study, the model was applied to a comprehensive dataset that included not only DBT, RH, and WS but also other meteorological parameters such as humidity ratio (HR), wind direction (WD), total solar radiation (ITH), direct solar radiation (IDH), diffuse solar radiation (ISH), and sky radiation temperature (TSKY). This extensive dataset allowed the VAR model to capture the complex dynamics and interactions between these variables, providing a robust framework for forecasting.

An important aspect of the VAR model used in this study was its ability to automatically select the optimal number of lags (hours) to include in the prediction model. Although the model was initially set to consider up to 50 lags, it utilized the Akaike Information Criterion (AIC) to determine the most appropriate number of hours to look back for making accurate predictions. This adaptive feature ensured that the model was not overfitted and could generalize well to new data.

The analysis showed that the VAR model requires a wide range of correlated features to improve prediction accuracy. The inclusion of multiple variables helps the model to better understand the underlying relationships and dependencies, leading to more accurate forecasts. This highlights the importance of having a comprehensive dataset with multiple interrelated variables to fully capture the dynamics of the system.

Overall, the VAR model demonstrated its capability to predict temperature and humidity with reasonable accuracy, making it a valuable tool for building energy calculations. The model's performance highlights the importance of having a wide range of correlated features to improve prediction accuracy. The VAR model requires a comprehensive dataset with multiple interrelated variables to fully capture the dynamics of the system.

Future work could focus on enhancing the model's ability to predict wind speed by incorporating additional data or using more advanced modeling approaches. Additionally, addressing the issue of missing data in the dataset could further improve the model's performance. Exploring other robust methods and integrating more features could also enhance the predictive power of the VAR model.

In conclusion, the VAR model provides a robust framework for forecasting meteorological parameters, and its application in this study highlights its potential for use in building energy analysis. Despite some limitations, the model's ability to capture the dynamic relationships between variables makes it a powerful tool for predicting future values based on historical data. The findings underscore the importance of comprehensive datasets and the inclusion of multiple interrelated variables to achieve accurate and reliable forecasts.

References

[1] L. Chang and Y. Shi. A discussion on the robust vector autoregressive models: novel evidence from safe haven assets. *Annals of Operations Research*, 339:1725–1755, 2022.

[2] L. J. Christiano. Christopher a. sims and vector autoregressions. *New York University*, 2012.

[3] R. B. Litterman. Forecasting with bayesian vector autoregressions—five years of experience. *Journal of Business & Economic Statistics*, 4(1):25–38, 1986.

[4] C. A. Sims. Macroeconomics and reality. *Econometrica*, 48(1):1–48, 1980.

[5] J. H. Stock and M. W. Watson. Vector autoregressions. *Journal of Economic Perspectives*, 15(4):101–115, 2001.