Forecasting Daily River Flow Using SVR and EWMA: An Analysis of Saugeen River Data

Jiaze Wang Zijian Wang Qiuhan Li

Abstract

Forecasting daily river flow is essential for effective water resource management and environmental planning. This study utilizes the Saugeen River Flow dataset, spanning 65 years and comprising 23,741 daily observations, to evaluate and compare the performance of traditional and machine learning models in predicting river flow. Time series decomposition shows seasonal patterns account for 34.41% of variance. Three EWMA models were tested, with the best achieving an R-square of 0.31 after tuning. An SVR model with an RBF kernel outperformed, achieving an R-square of 0.85 without explicit seasonal features. The results highlight machine learning's advantage in capturing nonlinear patterns, suggesting future exploration of hyperparameter tuning and alternative models to improve accuracy.

1. Introduction

Forecasting river flow is a critical task for effective water resource management, flood prevention, and environmental sustainability. Reliable predictions support decision-making in infrastructure planning, agricultural water use, and ecosystem preservation. This study aims to identify methodologies that can effectively capture seasonal, trend, and irregular components of river flow data. The analysis can help improve the accuracy of hydrological predictions, which is important for water management in the context of climate variability and increasing demand.

2. Data

This study employed the Saugeen River Flow dataset, available in the Monash Time Series Forecasting Repository [1]. The link of the data is https://doi.org/10.5281/zenodo.4656058
The dataset contains an extensive time series that records the daily average flow of the Saugeen River in cubic meters per second. The data spans a period of 65 years, covering observations from January 1, 1915, to December 31, 1979. With a total of 23,741 data points, the dataset provides a comprehensive representation of long-term river flow patterns. It was originally sourced from the R package deseasonalize and is formatted as a .tsf file with a size of 107.5 KB. Statistical analysis of the dataset reveals a mean daily flow of 30.06 cubic meters per second, a standard deviation of 39.16, a minimum value of 2.30, and a maximum value of 640.00, reflecting a wide range of variability in river flow. In addition, the dataset does not contain any missing values.

In this project, the first 80 percent of data are used to train our models, and the left 20 percent of data are the test data. By comparing our forecast and the true value of test data, the forecast accuracy can be calculated. The criteria used here are R-square and RMSE.

3. Method

Firstly, this study employs time series decomposition to analyze the components of the Saugeen River dataset, using the seasonal_decompose method from the statsmodels library. The decomposition splits the data into three components: trend, seasonal, and residual. The results reveal the following variance proportions: the trend component accounts for 3.38% of the total variance, reflecting long-term changes in the data. The seasonal component explains 34.41% of the variance, which captures repetitive annual patterns in the series. Meanwhile, the residual component dominates, accounting for 62.56% of the variance, representing irregular variations and noise.

Building on those findings from decomposition, an exponential-weighted moving average (EWMA) model is employed. Given the large contribution of the seasonal component, incorporating this pattern into the model is crucial to effectively capture cyclical changes, while the trend still provides valuable information about the long-term behaviour of the series. Therefore, both the trend and seasonal components are incorporated into the EWMA model.

Three types of EWMA models are examined in our project, which are the non-seasonal addictive EWMA model, the multiplicative and addictive seasonal EWMA models. Their forecast accuracy is compared based on our test data and the corresponding

Apart from that, we also employ a Support Vector Machine model to forecast the daily average flow of Saugeen River. This is a regression problem in essence, and a regression version of the Support Vector Machine model is employed. The radial basis function(RBF) kernel and large value of C is set to better fit the complex non-linear nature of our data.

The code link is https://github.com/wangjiaze/Time-Series-Final-Project

4. Results

We employed a machine learning model(SVR) and Exponential Weighted Moving average(EWMA) to forecast the daily mean flow of the Saugeen River at Walkerton. The unit is cubic meters per second. To begin with, the results of the machine learning model are introduced. The prediction results for the SVR model with RBF kernel, with and without seasonal features are as follows,

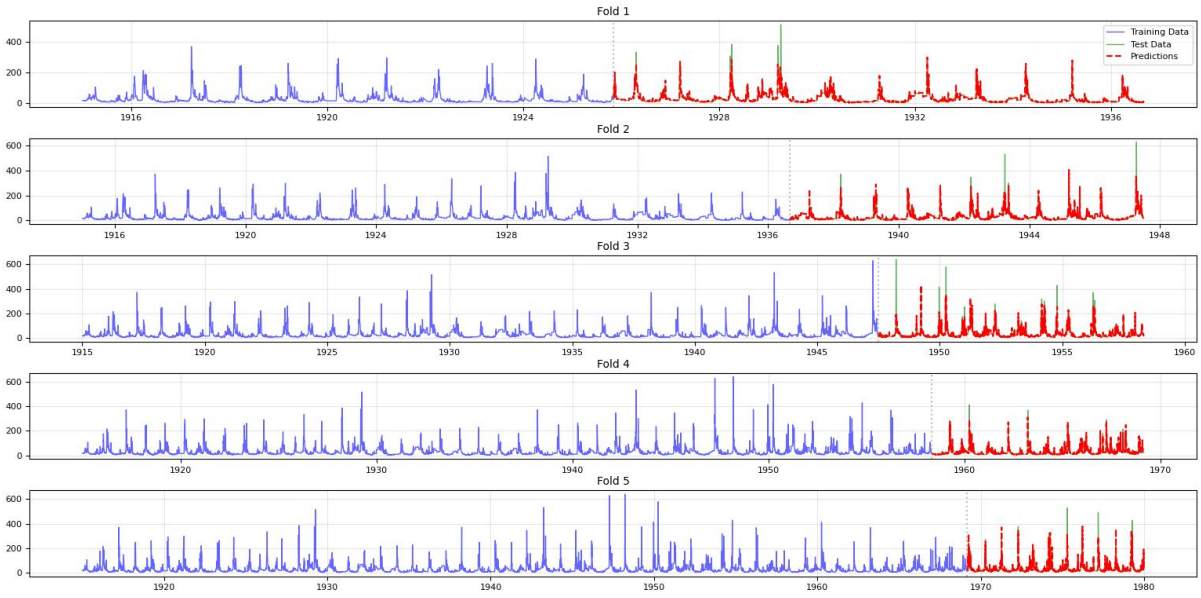

Figure 1. SVR prediction with 5 Fold Cross-validation and five daily lagged features Average(RMSE: 15.14, Average R2: 0.85).

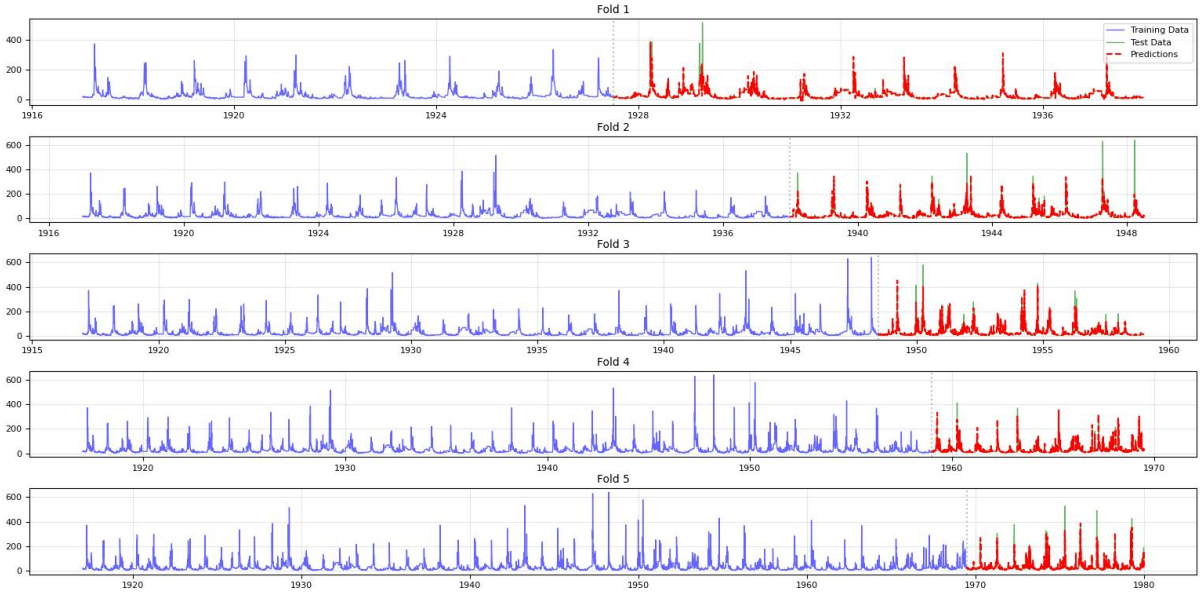

Figure 2. SVR prediction with 5 Fold Cross-validation, five daily lagged features and two seasonal features (Average RMSE: 16.92, Average R2: 0.82).

SVM CV-Results Comparison with/without Seasonal features

Fold	RMSE No Seasonal	R2 No Seasonal	RMSE With Seasonal	R2 With Seasonal
Fold 1	14.70	0.84	16.21	0.79
Fold 2	16.93	0.83	22.36	0.76
Fold 3	19.26	0.80	16.19	0.84
Fold 4	10.29	0.91	13.57	0.86
Fold 5	14.52	0.88	16.27	0.85
Mean	15.14	0.85	16.92	0.82

Table 1. The Cross-Validation results of SVM with and without seasonal features.

The SVR model that considers the seasonal pattern does not outperform the SVR without seasonal features. The R-square even decreases from 0.85 to 0.82. This may be because the radial basis function（RBF） kernel has already grasped the complex long-term seasonality of our data, it is not necessary to include the seasonal features. The setting of C=100 forces the SVR model to learn more patterns from the training data, and this may cause the model to suffer from overfitting problems. However, as a result, this setting helps SVR learn the seasonality of our data without the need to provide seasonal features.

Furthermore, the results of the EWMA models are presented as follows,

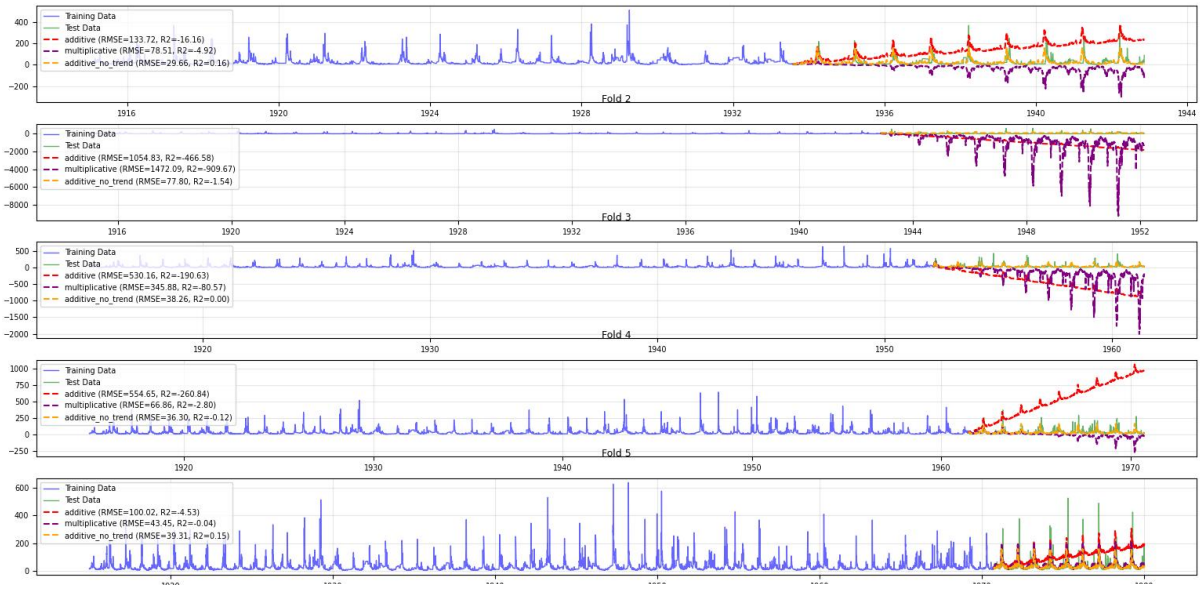

Figure 3. The 5 Fold Cross-Validation forecast results of three types of EWMA models.

It can be seen from Figure 3. that the Addictive seasonal EWMA model without trend has the smallest RMSE and highest R-square. Therefore, we further improve this model by tuning the value of the smoothing constant.

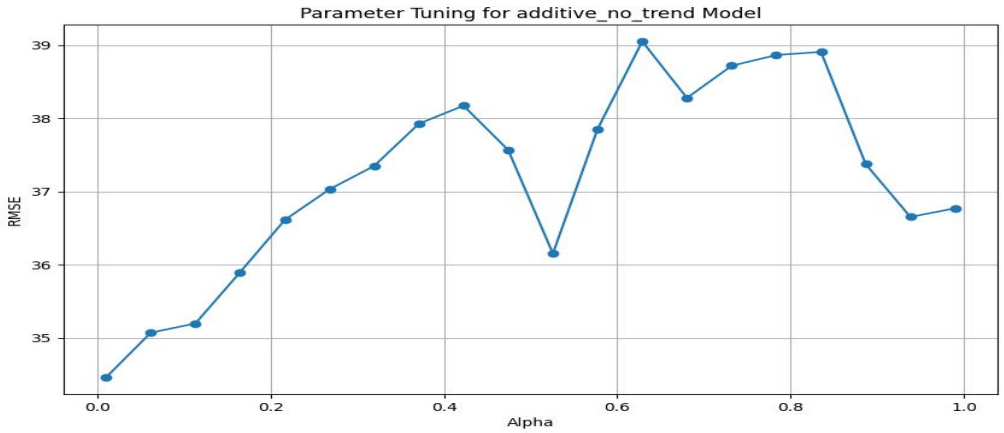

Figure 4. The RMSE results for the Addictive Seasonal EWMA Model(not considering trend) with different values of Alpha.

The best value of alpha we obtained from Figure 4 is 0.01, and hence the graph of the best Addictive Seasonal EWMA Model with alpha equal to 0.01 is plotted in Figure 5. below.

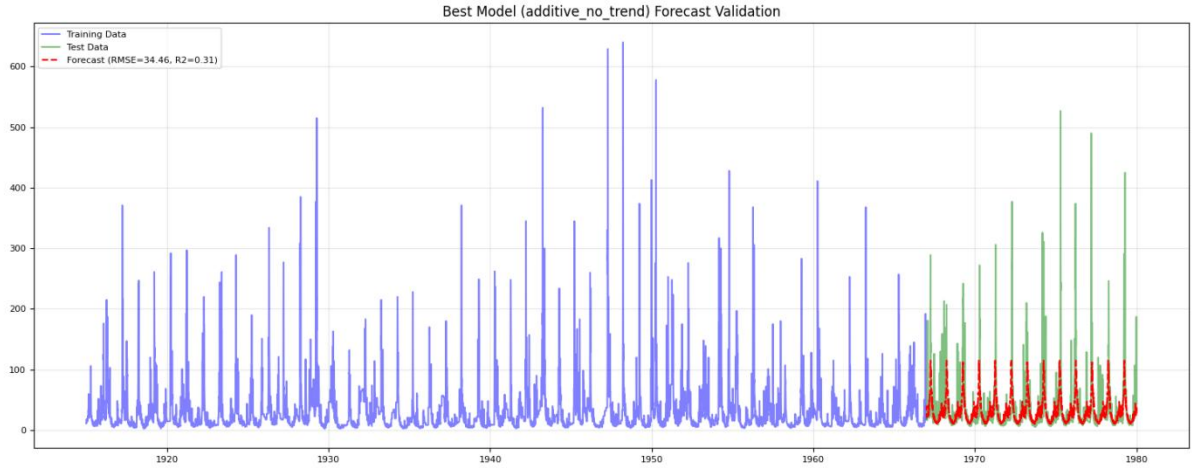

Figure 5. The forecast results of the Addictive Seasonal EWMA Model (not considering trend) with Alpha=0.01.

By tuning the smooth constant, the performance of our EWMA model almost doubled from around 15 percent to more than 30 percent. With the alpha value of 0.01, our best Addictive Seasonal EWMA Model (not considering trend) achieves an R-square of 0.31.

5. Discussion and Conclusion

In this project, we explored both Support Vector Regression (SVR) and Exponential Weighted Moving Average (EWMA) models to forecast the daily flow of the Saugeen River.

The traditional time series model EMWA performs comparatively weaker than SVR.
To begin with, among all three types of EWMA models, the best EWMA model with an optimized alpha value of 0.01, achieves R-square of 0.31. On the other hand, without feeding the seasonal features, SVR can learn the complex non-linear pattern of data, achieving an R-square of 0.85.

More hyperparameters of SVR can be tuned to further improve its performance. We can test the effectiveness of more machine learning models and traditional time series models on this problem.

Reference

1. Godahewa R, Bergmeir C, Webb G, Hyndman R, Montero-Manso P. Saugeen River Flow (SaugeenDay) Dataset. Zenodo; 2020.