FORECASTING OF WOLF NUMBER SERIES USING THE MACHINE LEARNING METHODS

Alexander Shibaev

Moscow State University
e-mail: alexshibaev@yandex.ru

Keywords*: Wolf number series, machine learning, XGBoost*

Abstract*: Recently the methods of machine learning, deep learning (neural networks) have been used intensively in scientific research and to suit many applications. This paper attempts to analyse and forecast the Wolf number series cycles using machine learning algorithms. The applied class of algorithms is decision-tree ensembles like the Random Forest and the Gradient Boosting (the XGBoost realisation). The quality of designed algorithms was evaluated based on the three suspended cycles 22–24, the forecast for cycle 25 was provided. The comparison of algorithm forecasting results for the old and new versions of WSN revealed the improved forecasting quality for the old version of the series compared to the new one.*

ПРОГНОЗИРОВАНИЕ РЯДА ЧИСЕЛ ВОЛЬФА МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

Александр Шибаев

МГУ имени М. В. Ломоносова, Москва, Россия
e-mail: alexshibaev@yandex.ru

Ключевые слова*: числа Вольфа, машинное обучение, XGBoost*

Резюме*: В последние годы методы машинного обучения (ML), глубокое обучение (нейронные сети) все активнее внедряются и используются в широком круге научных исследований и прикладных задач. В данной работе, применяя алгоритмы машинного обучения, предпринята попытка анализа и прогнозирования циклов ряда чисел Вольфа WSN v.2. Используемый класс алгоритмов – ансамбли деревьев решений: случайный лес (Random Forest) и Gradient Boosting(реализация XGBoost). Качество построенных алгоритмов оценивается на трех отложенных циклах: 22-24, также построен прогноз для 25 цикла WSN v.2. При сравнении результатов прогнозирования алгоритмов на WSN v.1 и v.2 отмечено улучшение качества предсказаний для старой версии ряда(v.1) по сравнению с v.2.*

Data preprocessing

The most crucial stage of data analysis and model development in the ML is data pre-processing. For better discerning of regularities, the WSN v.2 [1] series without quasi-two-year components (periods less than 2 years) averaged by 13 months was used. The obtained smoothed WSN_smooth series is shown in Fig. 1.

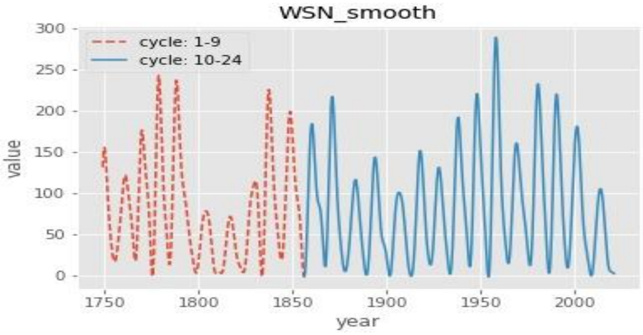

Fig. 1. WSN v.2 series without quasi-two-year components

Due to incompletion and gaps in the data, cycles from 1 to 9 were discarded and not used in the analysis and forecasting [2]. To increase the dataset amount, the WSN_smooth series was duplicated twice from the cycle 10. For example, to forecast the cycle 22, the part of the series from cycle 10 to cycle 21 was repeated twice.
In the ML, objects are characterised with the set of numeric parameters (features), to which target variables (targets) are associated. The general problem is to learn to restore the desired regularity using the set of known objects (the training dataset) and the corresponding target vector. As the new set of objects (the test dataset) is obtained, the ML algorithms forecast the target vector.

In this paper, the algorithm for the creation of the training dataset is as follows: the current value of the series (target) x_t is forecasted using the previous ***k*** values: x_{t-k} , … , x_{t-1}. To help algorithms discern long-period components of the series, the ***k=831*** was selected (it corresponds to about 6 cycle lengths). The training dataset was created using the slide along the WSN_smooth series. The training dataset and the target vector are shown schematically below.

$$\text{Dataset: } \begin{pmatrix} x_1 & \cdots & x_{831} \\ \vdots & \ddots & \vdots \\ x_{t-831} & \cdots & x_{t-1} \end{pmatrix}, \qquad \text{target: } \begin{pmatrix} x_{832} \\ \vdots \\ x_t \end{pmatrix}$$

Now will be considered in more detail the process of the solution to forecast the cycle 22. To create the training dataset and target vector, the WSN_smooth series from double-length cycles 10 to 21, which was processed to the end with the slide algorithm, was used as a base. The ML algorithms were adjusted based on the results obtained. To forecast the first value x_1^{pred} of suspended cycle 22, the ***k*** of previous values was used. To forecast the n point of suspended cycle x_n^{pred}, the already forecasted $x_1^{pred}, \dots, x_{n-1}^{pred}$ values of cycle and $k-n+1$ values of previous cycles were used. To forecast cycle 23, the twice-duplicated WSN_smooth series from cycles 10 to 22 was processed to the end with the slide algorithm; the further solution scheme is the same. To forecast the cycle 24, the twice-duplicated WSN_smooth series from cycles 10 to 23 was processed to the end with the slide algorithm, and so on.

Data analysis algorithms

Parameters of machine learning models are of two types: internal and hyperparameters. The model seeks the internal those automatically based on the dataset and target vector. Hyperparameters should be set up by a researcher who should vary the values; models will discern regularities in the data better or worse, and the error function will be larger or smaller respectively for the new (or suspended) data. In this paper, adjustable hyperparameters of the Random Forest Regressor algorithm were: n_estimators (the number of trees in the algorithm), max_features (the number of flags to choose the splitting), min_samples_leaf (the limitation for the number of samples in the leaf). Parameters of the XGBoost Regressor algorithm were: n_estimators (the number of trees in the algorithm), learning_rate (the learning rate), subsample (the part of the dataset used for learning), max_depth (the maximum depth of trees), min_samples_leaf (the limitation for the number of samples in the leaf). To evaluate the forecasting quality of the suspended cycle model, the RMSE error function (the root-mean-square error, $\sqrt{1/m \sum_{i=1}^{m} \left(x_i^{true} - x_i^{pred}\right)^2}$, m is the number of points in the cycle) was used. The optimum parameter values for the models were adjusted by varying the hyperparameter values and obtaining the forecast error function in cycles 22 to 24. For example, the RMSE distribution for cycle 22 of one of the Random Forest models with various combinations of value pairs for max_features and min_samples_leaf parameters is shown in Fig. 2.

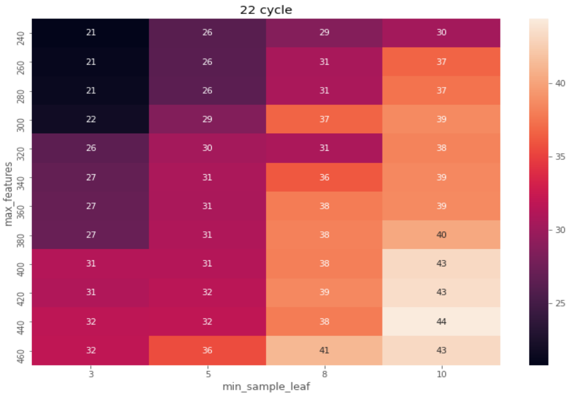

Fig. 2. RMSE error matrix at cycle 22

Cycle's predictions

For steadier forecasting, several models were developed, and forecasts for them were averaged. Five models were used in forecasting; three of them (rf1, rf3, rf4) relate to Random Forest Regressor and two (xgb5, xgb6) to XGBoost Regressor. Besides the rf4, all models were trained using duplicated series; the rf4 was trained based on the dataset created from the WSN_smooth series with no repetitions, which increased the variety of algorithms. As a rule, the forecasting averaging force for several algorithms (the ensemble) increases as the variety in the ensemble rises. The forecasting results for models and their averaged forecasts for cycles 22 to 24 respectively are shown in Fig. 3-5.

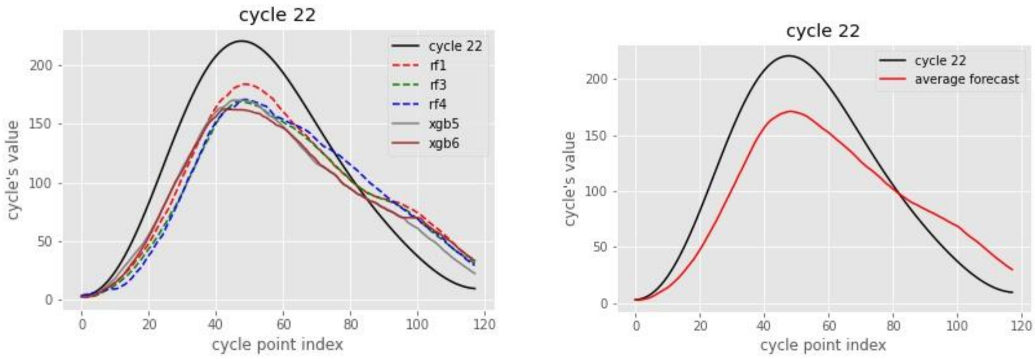

Fig. 3. Model predictions for cycle 22 (left) and their average (right)

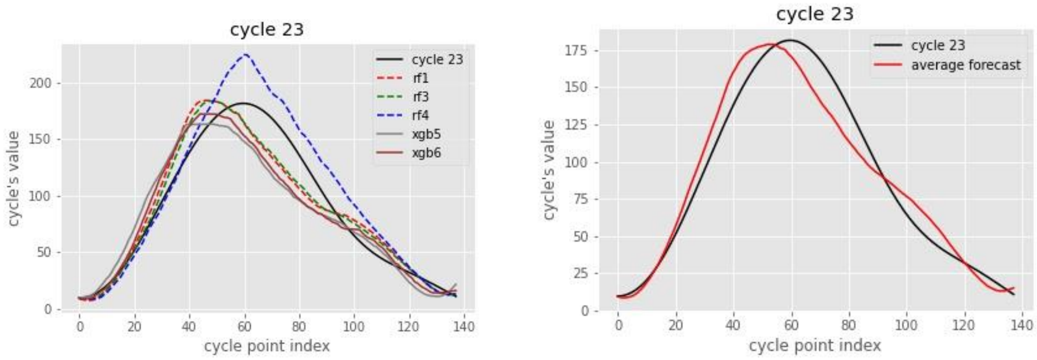

Fig. 4. Model predictions for cycle 23 (left) and their average (right)

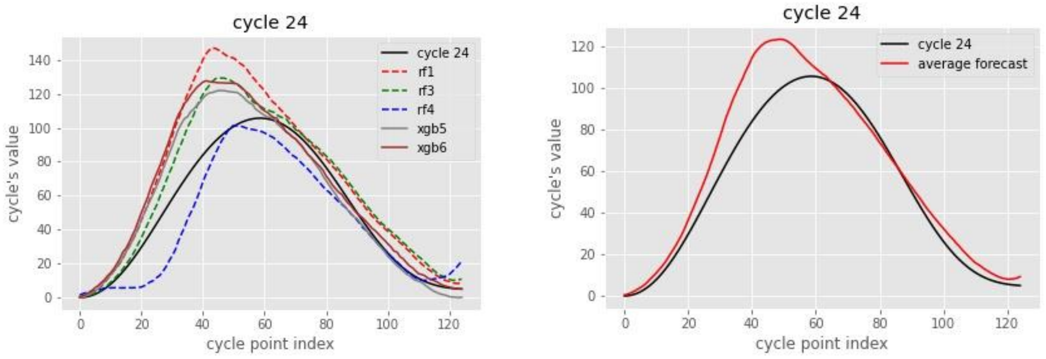

Fig. 5. Model predictions for cycle 24 (left) and their average (right)

It is worth noting that the minimum point between the previous and new cycles was accepted as the new cycle start in the smoothed WSN_smooth series. Even insignificant shifting of the cycle start by 4-6 points often results in the notable changes in the model forecast.
Model forecasts for cycle 25 are shown in fig.6.

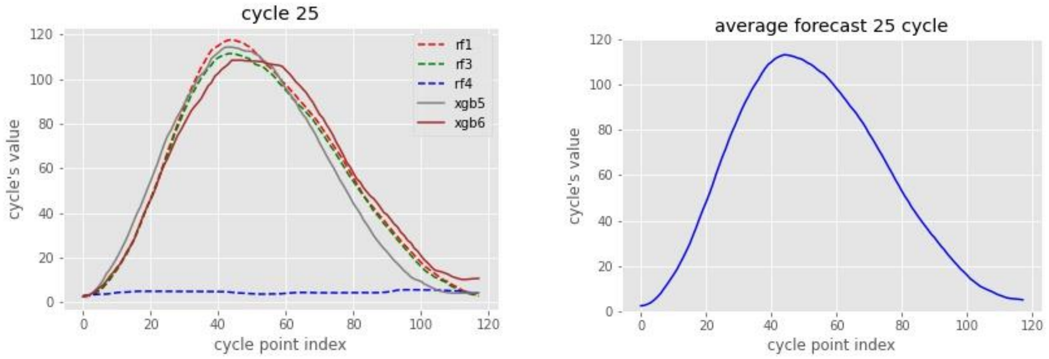

Fig. 6. Model predictions for cycle 25 (left) and their (no rf4) average (right)

As it is seen, the rf4 model fails on cycle 25, but if it is adjusted using the dataset obtained from the WSN_smooth series using the duplication, then the forecast complying to those for other models will be obtained. The averaged forecast of the new cycle without including the forecast for the rf4 model is shown in fig.6.

Comparing WSN v.1 to v.2

Besides the analysis of the WSN v.2 series with no quasi-two-year components (WSN_smooth), the machine learning methods were used to study the classic WSN v.1 and WSN v.2 series. The training dataset creation scheme and the adjustment of parameters using suspended cycles are the same. For WSN v.1, the Random Forest-class algorithms are steadier in hyperparameters. It means that it is easy to adjust the parameter value ranges, in which the error is minimised at once for all three suspended cycles. For WSN v.2, hyperparameter value ranges that minimise the error in cycles 22 and 24 do not match. Fig.7-9 shows distributions of error (RMSE) in cycles 22 to 24 for WSN v.1 and v.2 by values of max_features, min_samples_leaf parameters of one of the Random Forest models.

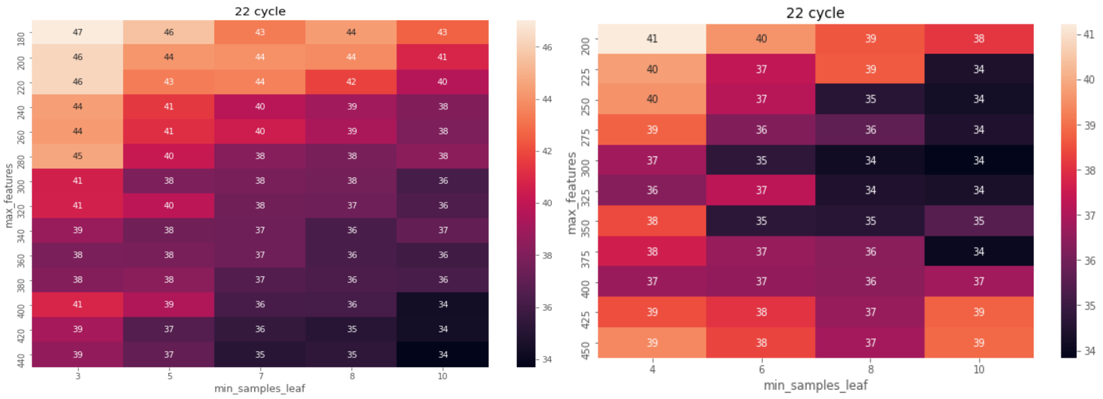

Fig. 7. Error matrix on cycle 22 WSN v.1 (left), WSN v.2 (right)

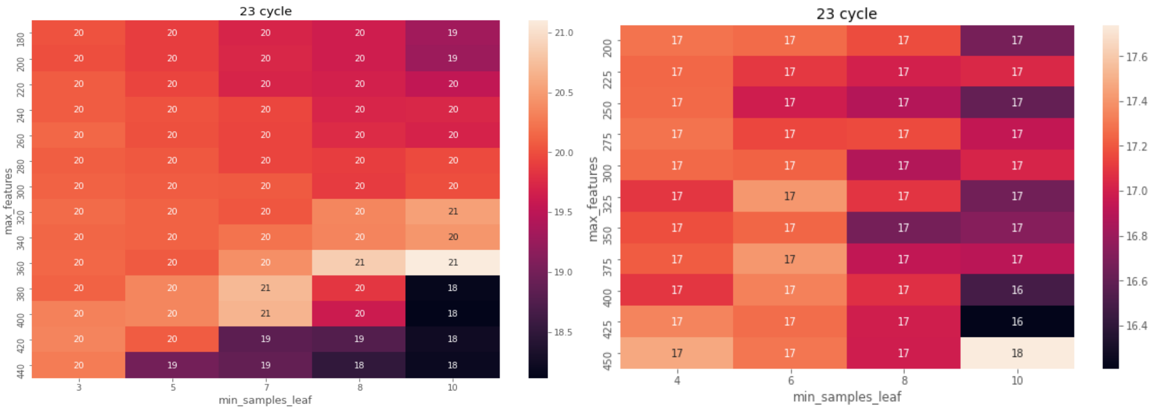

Fig. 8. Error matrix on cycle 23 WSN v.1 (left), WSN v.2 (right)

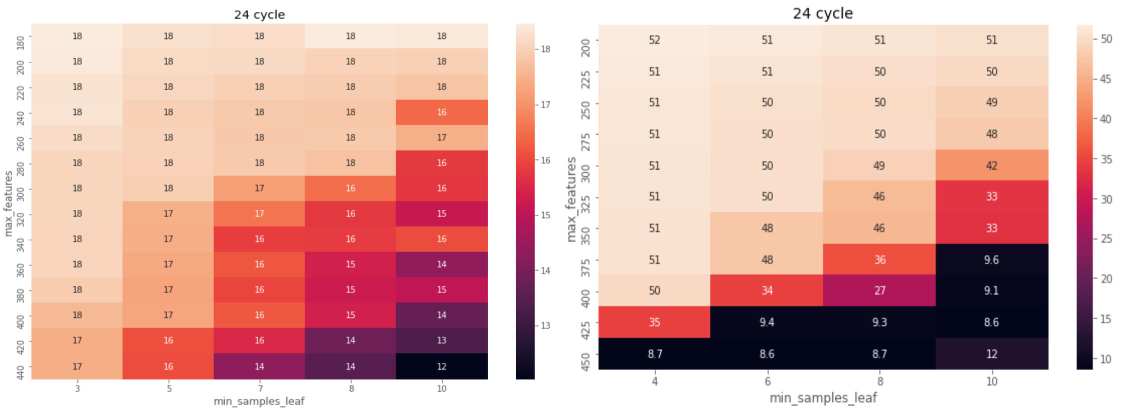

Fig. 9. Error matrix on cycle 24 WSN v.1 (left), WSN v.2 (right)

It is clearly seen that for v.1, the minimum points in suspended cycles are reached in the same range (below the secondary diagonal of the error matrix). For v.2, these ranges almost do not intersect, which is a characteristic pattern. Analysing the forecast results for different models, it may be said that, on average, for v.1, algorithms restore better the regularities of the series and are steadier than for v.2. It might be associated with the supplementary noisiness of the new version of series after the transformation of v.1 into v.2 [3] (fig.10).

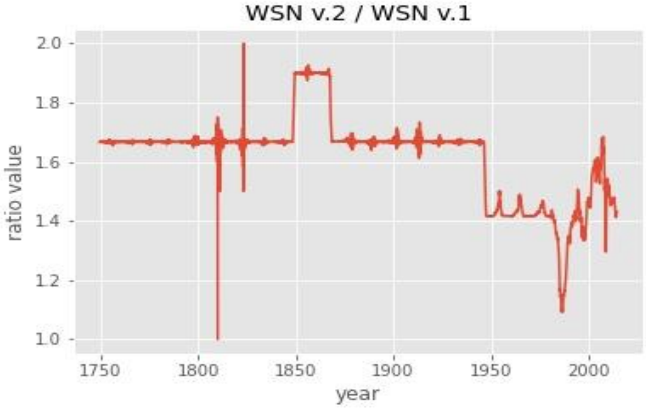

Fig. 10. Ratio WSN v.2 to WSN v.1

Results

This paper attempted to restore the dynamics of smoothed Wolf number series (with no two-year components) using the machine learning algorithm. The quality of forecasting models was evaluated based on cycles 22 to 24, and algorithms demonstrated quite good results for suspended data. Moreover, the forecast for the current cycle 25 was performed too. Also, the ability of algorithms to restore regularities for WSN v.1 and v.2 was analysed. In average, the reviewed algorithms based on decision trees restore better the regularities for WSN v.1 than for v.2. Perhaps it occurs due to the supplementary noisiness of the new version of the Wolf number series after the transformation of v.1 into v. 2.

References:

1. Clette, F., L. Svalgaard, J. M. Vaquero, E. W. Cliver. Revisiting the Sunspot Number Space Science Reviews. 2014.
2. David, H. Hathaway. The Solar Cycle Living Rev. Solar Phys. 2015.
3. Shibaev, A. I. The characteristics of old and new versions of month Wolf numbers range and there uniformity are compared. XIV Young Scientists Conference “Fundamental and Applied Space Researches” papers. 2017.