On statistical analysis and prediction of sap flow density for smart urban tree monitoring

Anastasia Safargalieva[1], Irina Kochetkova[1,2], Elena Makeeva[1] and Sergey Shorgin[2]

[1]*Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation*

[2]*Institute of Informatics Problems, Federal Research Center "Computer Sciences and Control" of the Russian Academy of Sciences, 44-2 Vavilova St, Moscow, 119333, Russian Federation*

Abstract

The use of IoT technologies in various areas of our life, including environmental monitoring of green spaces, is increasing every year. One such solution is the TreeTalker sensor-based monitoring system, which collects data on various parameters of trees. One of the most important parameters is the rate of tree sap flow. Predicting the density of sap flow and studying the relationship between the parameters of trees and the environment is an urgent task. In this work, a statistical analysis of the data collected using the TreeTalker monitoring system was carried out. The data was pre-processed: outliers in the data were removed using mean value replacement, z-score replacement and cumulative moving average replacement. Groups of trees that were homogeneous in time were identified, and regression models were built to predict the sap flow parameter using auto-regressive moving average and linear modeling. The results obtained can be used for further studies of the dependence of the state of the tree on external factors.

Keywords

Smart Urban Nature, Smart Urban Tree, TreeTalker, time series, sap flow density, statistical analysis, prediction,

1. Introduction

Monitoring of the health of the trees helps to achieve a comprehensive view of ecosystems. Nowadays environment is stressed by human activities. Providing a monitoring of trees health can answer a lot of questions about the effectiveness of the measures to maintain ecosystem's health. TreeTalker(TT) is an IoT device that collects information about the health state of the trees based on various internal and external factors. The main factor of the tree which is considered the most important is sap flow [1], [2], [3], [4], [5].

This work has the following structure: the section 2 is devoted to a primary statistical analysis, work with the outliers in data with three methods: Mean Replacement, Z-score replacement and Cumulative Mean Average. In section 3 we perform prediction of the sap flow using linear models: auto-regressive moving average and linear regression.

Workshop on information technology and scientific computing in the framework of the XI International Conference Information and Telecommunication Technologies and Mathematical Modeling of High-Tech Systems (ITTMM-2021), Moscow, Russian, April 19–23, 2021

ansafargalieva@mail.ru (A. Safargalieva); gudkova-ia@rudn.ru (I. Kochetkova); elena-makeeva-96@mail.ru (E. Makeeva); sshorgin@ipiran.ru (S. Shorgin)

2. Initial Data Analysis

2.1. Time Series Description

TreeTalker sensors were installed on 195 trees in seven territories located in the center of Moscow, on the RUDN University campus and in parks in the Moscow region. During 2019, measurements were made of the parameters of 22 different tree species, different in age, as well as in different states of "health". Every hour, data was collected on eight parameters – sap flow, air temperature and humidity, negative pressure of water vapor in the leaves of a tree, angles of tree deviations from the axis, wood moisture, temperature inside the trunk, and the normalized relative vector of vegetation (NDVI) (Tab.1 and 2). Age group and VTA score are constants. For the following work the data from the Troitsk Territory was selected.

Table 1
Parameters changing in time

Designatiton	Parameter	Range of values	Units of measurement
F	Flux	0.01 – 5.9	l*m⁻2 h−1
t	Air temperature	10-27	℃
rh	Pressure	27-28	Pa
v	Vapour-pressure deficit	0.5 - 2.5	g*m−3
Th, psi, phi	Tree trunk axis movement	-60 – -56, 3 – 9, -31 – -30	°
w	Stem Humidity	25-40	ton*m−3
nt	Temperature inside trunk	10-20	℃
nd	Normalized vegetation vector	0-100	%

Table 2
Parameters not changing in time

Designatiton	Parameter	Range of values
AG	Age group	I – VI, where I – the youngest tree, VI – the oldest tree.
VTA	VTA score	1 – 7, where 1 – good condition, 7 – bad condition.

The primary statistical analysis shows heterogeneity – there is no data for some periods of time due to damage to the electronics after heavy rain (Fig.1) and abnormally high-values (Fig.2). The presence of gaps in measurements leads to false statistical analysis, as well as incorrect modeling of dependencies. Therefore, the next task was to identify time-homogeneous groups of data [6].

2.2. Working with Unevenly Spaced Data

The presence of time gaps, that Fig.1 showed, makes it impossible to build models. The set of time values $T = \{\tau_1, \tau_2, ..., \tau_n\}$ consists of time-homogeneous subgroups $T_j = \{\tau_{i_j}, \tau_{i_{j+1}}, ..., \tau_{i_{j+k}}\}$, selected according to the algorithm [6]. For our model, we will consider as homogeneous data those values, the difference in arrival between which is 1 hour (Alg.1).

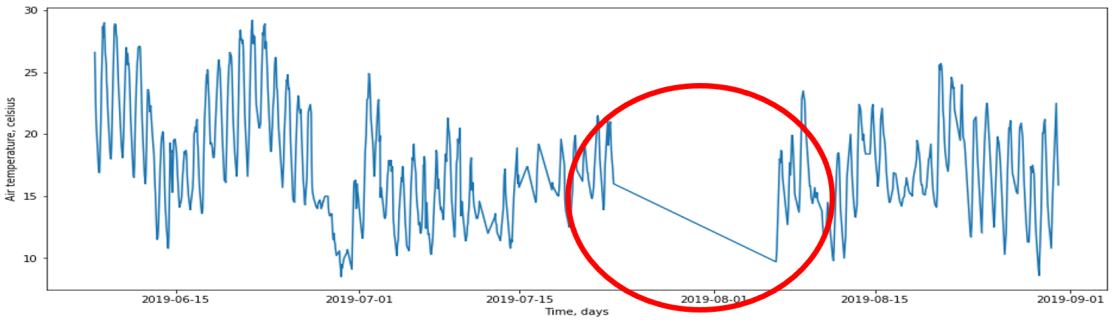

Figure 1: Air temperature: unevenly spaced data

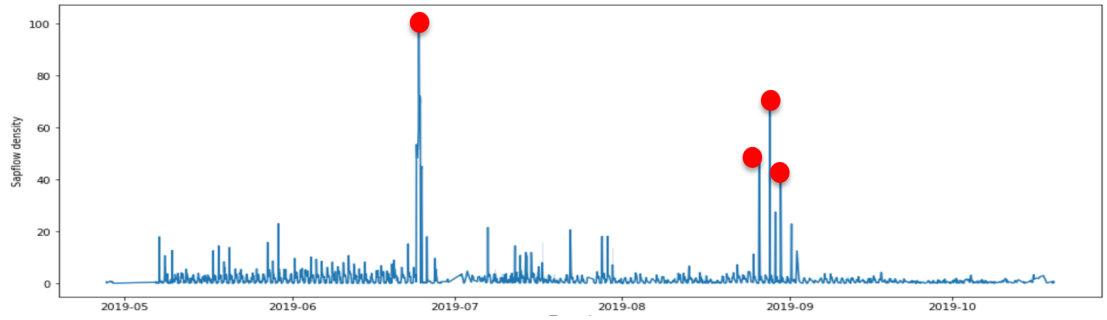

Figure 2: Sap flow density: outliers

Data: $T = \{\tau_1, \tau_2, ..., \tau_n\}$ - set of time values
Result: $T_j = \{\tau_{i_j}, \tau_{i_{j+1}}, ..., \tau_{i_{j+k}}\}$ - time-homogeneous subgroups
for $t = 1, 2, ...$ **do**
  if $t[i+1] - t[i] > 1$ **then**
    put $t[i+1]$ in the new group;
  else
    Leave $t[i+1]$ in the same group
  end
end

Algorithm 1: Algorithm to reveal time-homogeneous data

After data selection we get homogeneous data regarding flux parameter (Fig.3). The graph of the dependence of sap flow on time within a homogeneous group showed the presence of abnormally high values of sap flow (Fig.4).

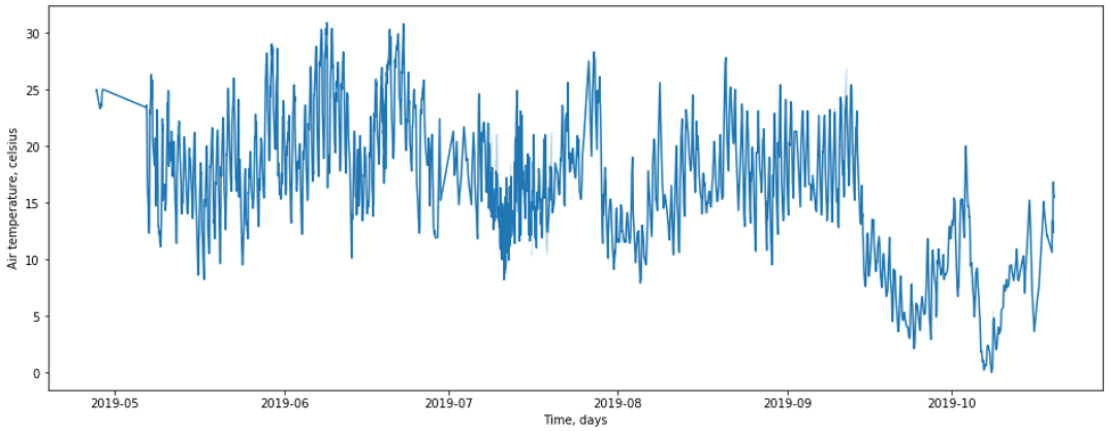

Figure 3: Air temperature: equally spaced data

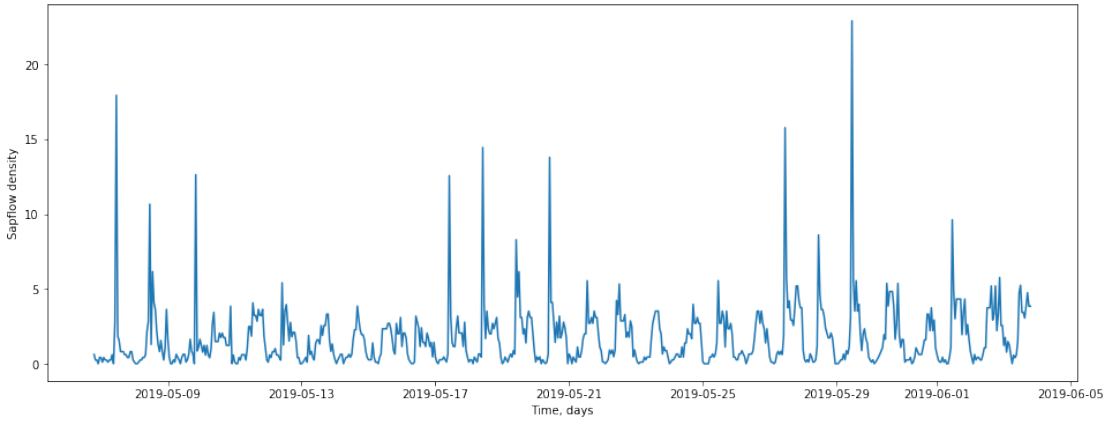

Figure 4: Sap flow density: equally spaced data with outliers

2.3. Working with Outliers

Mean Replacement Method. The first way to work with outliers in your data is to replace outliers with mean values. x_i – source row of one of eight parameters. y_i – row after processing from outliers after applying the following algorithm:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i - \text{mean value} \tag{1}$$

$$y_i = \begin{cases} x_i, & \text{if } x_i \leq \bar{X} \\ \bar{X}, & \text{if } x_i > \bar{X} \end{cases} \tag{2}$$

The results of the replacement of outliers with mean values can be seen on Fig.5. From

the graph it is clear that the values are no more than 5 points of the sup tree flux units of measurement.

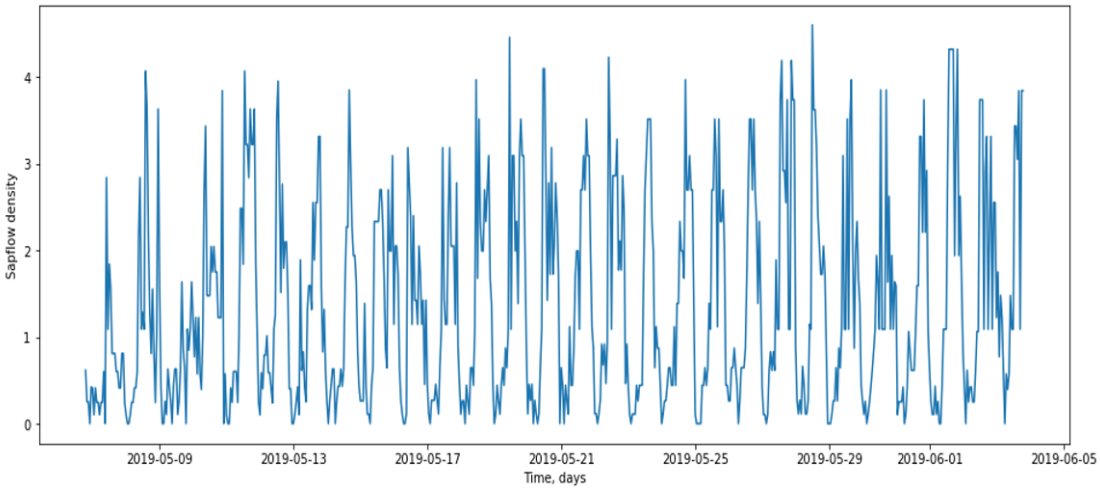

Figure 5: Sap flow density: applying mean replacement method

Z-score Substitution Method. The second way to process data from outliers is preliminary analysis of values using z – estimation and subsequent processing of abnormally high values of the parameter [7]. For the z – estimate, calculate the mean $\bar{X}$ and standard deviation s_x calculated for the set of processed data

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2} \tag{3}$$

$$z = \frac{x_i - \bar{X}}{s_x}, \tag{4}$$

$$y_i = \begin{cases} x_i, & \text{if } x_i \leq z \\ z, & \text{if } x_i > z \end{cases} \tag{5}$$

The results of the replacement of outliers with mean values can be seen on Fig.6. From the graph it is clear that the values are no more than 5 points of the sup tree flux units of measurement as it was with the mean replacement method. However, the structure of the curves is different.

Cumulative Moving Average Method. The third method used to deal with outliers in this work is the cumulative moving average method. It is used for smoothing time series [6]. This method smooths outliers using the arithmetic mean of the original function x_i over the entire period:

$$y_i = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_n + x_{n-1} + ... + x_2 + x_1}{n}, \tag{6}$$

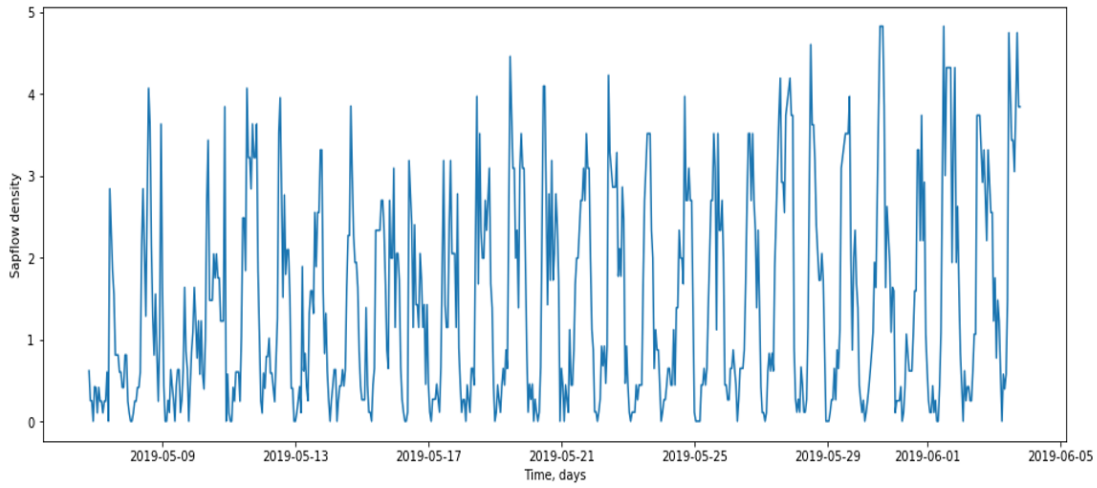

Figure 6: Sap flow density: applying Z-score replacement method

where y_i is a new series smoothed using the cumulative moving average at the moment n (Fig.7), n is the number of intervals available for calculation, x_i - the value of the original function at points. After using three methods mentioned above, we selected the data processed with z-score (Fig.6) as this method provides the better structure of the data - smooths it, the box-plots showed no outliers in the data.

The methods of working with outliers are presented in the form of the algorithm (Alg.2).

3. Sap Flow Density Prediction

3.1. Preliminary Considerations

The construction of a mathematical model of sap flow and prediction of sap flow will help to find out whether there is really a direct relationship between the sap flow and the air temperature, whether other factors affect the sap flow parameter. To analyze the obtained dependencies, 4 parameters for assessing the quality of the models were investigated: R^2, F – statistics, root-mean-square error (RMSE) and mean absolute error (MAE) [8].

3.2. ARMA Model

The ARMA model is an auto-regressive moving average model. The formula is as follows:

$$y_i = 1.4220 + 0.7150x_i + 1.39\theta + 0.134, \tag{7}$$

where $a = 0.7150$ is the parameter of the model, x is the parameter of the regression model, $b = 1.39$ is the coefficient of the moving average, θ is the parameter of the moving average, $c = 1.4240$ is a constant. The graph of the sup flow prognostication shows deceleration of the flow (Fig.8).

Data: x_i - original series, $\bar{X}$ - mean value of original series, z - z-score of original series
Result: y_i - processed series
Case 1: Mean-value Replacement
for *i = 1, 2, ...* **do**
  if $x[i] > \bar{X}$ **then**
    $y[i] = \bar{X}$;
  else
    $y[i] = x[i]$
  end
end
Case 2: Z-score Replacement
for *i = 1, 2, ...* **do**
  if $x[i] > z$ **then**
    $y[i] = z$;
  else
    $y[i] = x[i]$
  end
end
Case 3: Cumulative Moving Average
for *i = 1, 2, ...* **do**
  for *n = 1, 2, ...* **do**
    $y[i] = \frac{\sum_{j=1}^{n} x[j]}{n}$
  end
end

Algorithm 2: Algorithm of Replacement of Outliers

3.3. Linear Regression

We will forecast 25 observations ahead. We will draw the plot of the result, which turned out as a result of applying the linear regression model (Fig.9) [9]. Simulation of sap flow with different combinations of factors made it possible to identify the most effective models for describing the dependence of the tree sap flow. In the equation of the dependence of aspen sap flow on the territory of the Trotsk green spaces the parameter of negative pressure of water vapor in the leaves of the tree has the greatest influence:

$$\begin{aligned} y_F = 0.78 + 1.0538x_t + 0.3458x_{rh} - 4.7587x_v - \\ - 0.1761x_t h - 0.8449x_{nt1} + 1.4893x_{nd} - 0.1945x_W \end{aligned} \tag{8}$$

4. Conclusion

It was found in the work that the negative pressure of water vapor in the leaves is significantly correlated with the parameter of tree sap flow. After analyzing the data, it was found that the

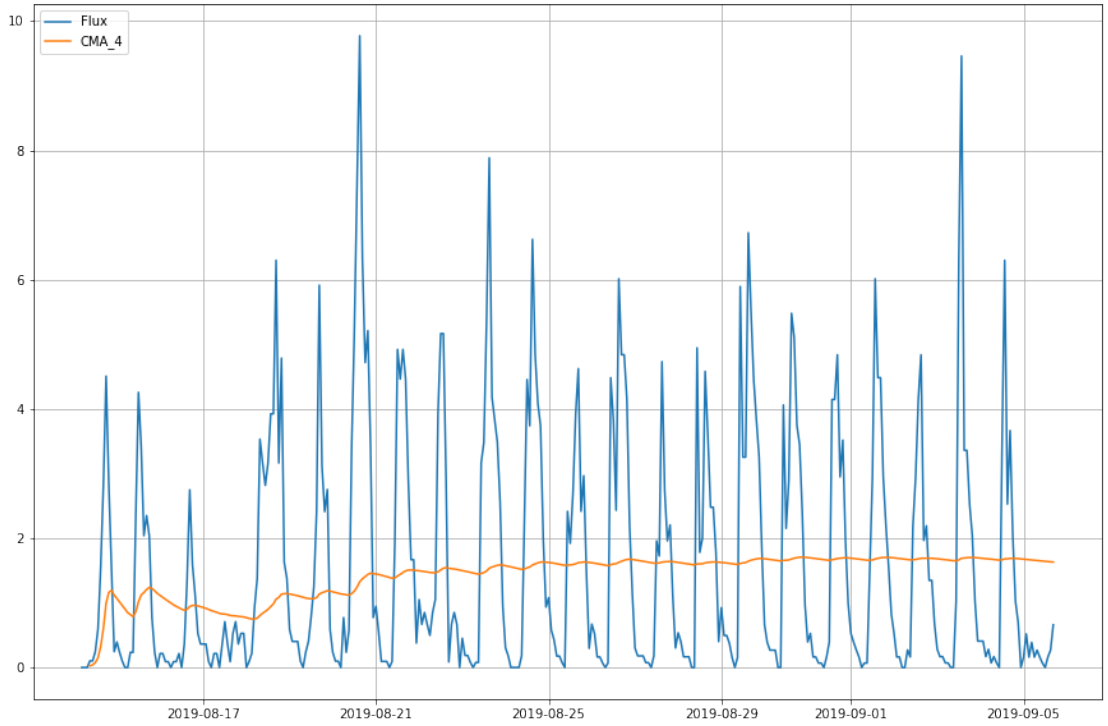

Figure 7: Sap flow density: applying cumulative moving average method

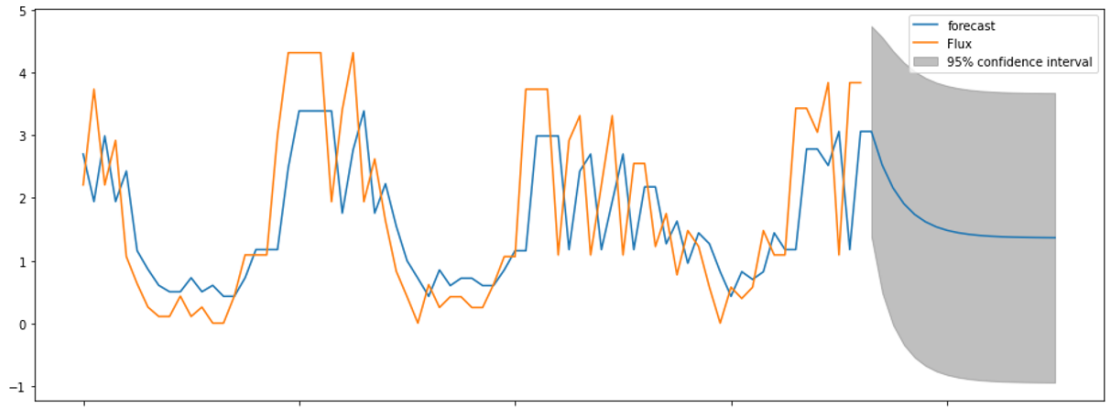

Figure 8: Sap flow density: applying ARMA model

sap flow of trees depends on 7 factors. The results obtained during the work showed which parameters should be taken into account when analyzing the state of the tree and predicting time-dependent factors. This study will provide a starting point for more sophisticated modeling approaches. For example, predicting a model using Fourier series can provide more accurate

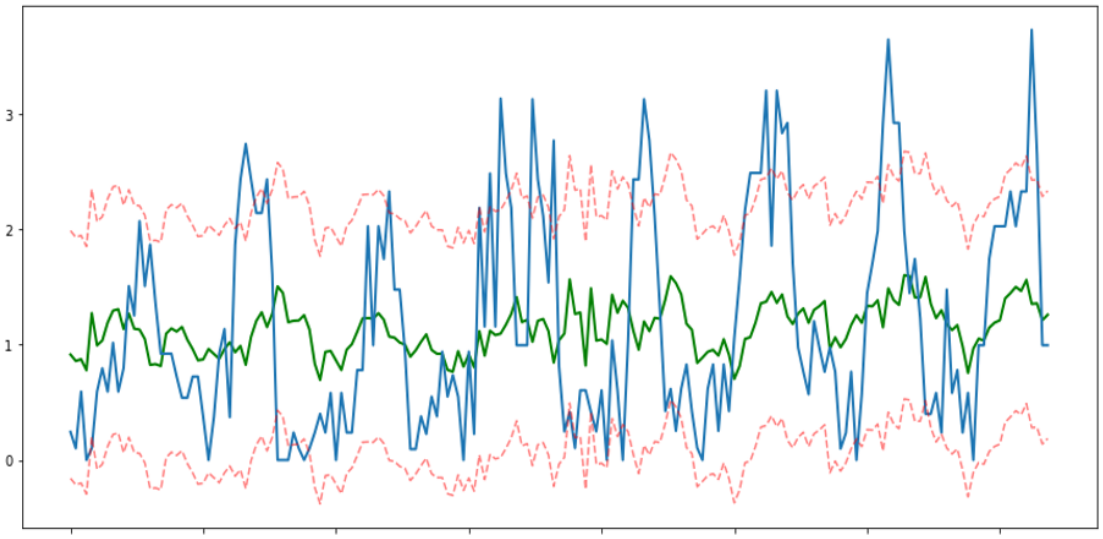

Figure 9: Sap flow density: applying linear regression

Table 3
Model quality metrics

Model	R^2	Prob (F-statistic)	RMSE	MAE
Flux: t, rh, v, th, nt, W, nd	0.86	5.50E-16	0.79	0.6844
Flux: t, rh, v, th, psi, phi, nt, W, nd	0.68	6.02E-10	0.4152	0.3585
Flux: t, rh, v, nt, W	0.39	1.50E-23	0.3457	0.2903

parameter estimates. In addition, assessing the flow density of sap flow is the main goal of researching the health of green spaces to predict changes in their health status.

The authors grateful to Dr. Alexey Yaroslavtsev (RUDN University) for providing the dataset from TreeTalker system.

Acknowledgments

The work was supported by the Russian Science Foundation, project 19-77-30012 (recipient Irina Kochetkova). This paper has been supported by the RUDN University Strategic Academic Leadership Program (recipient Elena Makeeva).

References

[1] V. Matasov, L. B. Marchesini, A. Yaroslavtsev, G. Sala, O. Fareeva, I. Seregin, S. Castaldi, V. Vasenev, R. Valentini, Iot monitoring of urban tree ecosystem services: Possibilities and challenges, Forests 11 (2020). doi:10.3390/f11070775.

[2] V. Riccardo, B. L. Marchesini, S. Giovanna, A. Yaroslavtsev, V. Vasenev, S. Castaldi, New tree monitoring systems: from industry 4.0 to nature 4.0 (2019). doi:10.12899/asr-1847.
[3] M. Fidino, S. Magle, Using fourier series to estimate periodic patterns in dynamic occupancy models, Ecosphere 8 (2017). doi:10.1002/ecs2.1944.
[4] D. Efrosinin, I. Kochetkova, N. Stepanova, A. Yarovslavtsev, K. Samouylov, R. Valentini, The fourier series model for predicting sapflow density flux based on treetalker monitoring system, Lecture Notes in Computer Science 12526 LNCS (2020) 198–209. doi:10.1007/978-3-030-65729-1_18.
[5] D. Efrosinin, I. Kochetkova, N. Stepanova, A. Yarovslavtsev, K. Samouylov, R. Valentini, Trees classification based on fourier coefficients of the sapflow density flux, Annales Mathematicae et Informaticae 53 (2021) 109–123. doi:10.33039/ami.2021.03.002.
[6] G. Box, G. Reinsel, G. Ljung, Time Series Analysis: Forecasting and Control, volume 68, 2016. doi:10.2307/2284112.
[7] W. McKinney, Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython, O'Reilly Media, Inc, Massachusetts, 2017.
[8] A. Granier, A new method of sap flow measurement in tree stems, Annales Des Sciences Forestieres 42 (1985) 193–200.
[9] J. A. Rice, Mathematical Statistics and Data Analysis, Duxbury Original Series, Massachusetts, 2010.