




# Application of Machine Learning Algorithms to Handle Missing Values in Precipitation Data

Andrey Gorshenin<sup>1,2</sup> , Mariia Lebedeva<sup>2</sup>, Svetlana Lukina<sup>2</sup>,  
and Alina Yakovleva<sup>2</sup>

<sup>1</sup> Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences, Moscow, Russia

[agorshenin@frccsc.ru](mailto:agorshenin@frccsc.ru)

<sup>2</sup> Faculty of Computational Mathematics and Cybernetics,  
Lomonosov Moscow State University, Moscow, Russia

[mash.lebedeva2010@yandex.ru](mailto:mash.lebedeva2010@yandex.ru), [svetl.luckina2016@yandex.ru](mailto:svetl.luckina2016@yandex.ru),

[alyna\\_yakovleva@mail.ru](mailto:alyna_yakovleva@mail.ru)

**Abstract.** The paper presents two approaches to filling gaps in precipitation based on classification (Support-Vector Machines) and regression (EM, Random Forests, k-Nearest Neighbors) machine learning algorithms as well as the pattern-driven methodology. These methods are among of the most powerful tools for data mining in a wide range of research areas including meteorology and climatology due to the presence of a large amount of temporal and spatial observations. When collecting observations from weather stations, there are a lot of missing records. Data processing algorithms are often very sensitive to the presence of incomplete data, so missing values should be firstly imputed and only after that the complete samples can be analyzed. The possibility of a correct filling data even for high missing levels based on suggested methods is demonstrated. The observations in Potsdam and Elista for about 60 years were used. Also, comparison of various algorithms for data imputation taking into account different missing levels is presented. The proposed methodology can be successfully used for real-time data processing of information flows.

**Keywords:** Precipitation · Missing values · Support-Vector Machines · XGBoost · Patterns · EM algorithm · Random Forests

## 1 Introduction

Methods of time series analysis are one of the most powerful tools for data mining in a wide range of research areas. Different approaches based on vari-

All ideas for imputation methodology based on patterns and machine learning techniques were suggested by Andrey Gorshenin whose work was supported by the **Russian Science Foundation** (project **18-71-00156**). Software tools were written in Python and tested on real data by MSc students (M. Lebedeva, S. Lukina, A. Yakovleva).

© Springer Nature Switzerland AG 2019

V. M. Vishnevskiy et al. (Eds.): DCCN 2019, LNCS 11965, pp. 563–577, 2019.

[https://doi.org/10.1007/978-3-030-36614-8\\_43](https://doi.org/10.1007/978-3-030-36614-8_43)

ous autoregressive models as well as a family of machine learning algorithms, including artificial neural networks, are traditionally used. These methods are very effective for meteorological problems due to the presence of a large amount of temporal and spatial data and, as a consequence, significant problems of their processing. In this paper, we focus on precipitation, but similar tools and approaches can equally well be used for another types of observations.

Precipitation is one of the most difficultly analyzed atmospheric parameters due to the nature of its variability. When collecting data from weather stations, there are a lot of missing records by various reasons. Data processing algorithms are often very sensitive to the presence of missing values, since they can seriously distort the results of analysis [20]. In particular, statistical tests and thresholds for extreme precipitation events [13] based on incomplete data can be incorrect.

Under the circumstances, in practice researchers should firstly fill in gaps, and only after that the complete set of observations can be analyzed and used as parameters of hydrologic models [16]. It is also worth noting that one can use an approach based on exclusion from the processing of subsamples with gaps until the time moment when the data becomes complete. However, in meteorological data a missing value may appear at any random time moment. Another approach may be based on reanalysis, but such data often exhibits variations compared with observations obtained by other techniques.

A well-known filling approach is inverse distance weighting method [24], including its extensions based on artificial neural networks [3]. Precipitation is usually weakly correlated with other meteorological parameters collected by weather stations. Moreover, the neighboring stations, which observations could be used to fill gaps correctly, are not in all geographic locations. It is worth noting, there are some suitable processing methods for case of a cluster of stations [21]. Having summarized, filling methods that use only initial data themselves without invoking any additional features are of particular interest.

This paper uses such well-known [27] machine learning algorithms as  $k$ -Nearest Neighbors ( $k$ -NN) [1], Expectation-Maximization ( $EM$ ) [18], Support-Vector Machines ( $SVM$ ) [7], and Random Forests ( $RFs$ ) [4]. They are still actively used to solve applied scientific problems in various fields. For example, researches related to the software [28] and big data problems [19, 26] can be mentioned. The theoretical aspects are also improved, and results are published in the worlds leading high-ranking journals, see [2, 8]. The novelty of our paper is based on a joint use of the pattern-based methodology [11] and machine learning algorithms to fill a significant amount (up to 40%) of missing values in data. The effectiveness of our approach using the precipitation data in Potsdam and Elista for about 60 years [30] is also demonstrated.

The paper is organized as follows. Section 2 introduces the approach to imputation missing values based on classification. It implies that each observation is replaced by “D” (no rain) or “W” (rain), and the imputation does not take into account the exact precipitation daily volume. SVM is used as a classification machine learning (ML) algorithm. Section 3 describes the results of regression ML methods to handle missing values in precipitation data, i.e., “continuous”

forecasts are implemented. The k-NN, EM algorithm and Random Forests are involved. Section 4 is devoted to discussion of the obtained results.

## 2 Approach Based on Classification

In this section, a classification approach to handling the missing values is introduced. The initial non-negative precipitation data should be modified as follows: if any positive value is observed in the current day, it is replaced by “W” (i. e., wet day), otherwise the symbol “D” (i. e., dry day) is used. Thus, the continuous time-series become discrete. Subsequences of some length of such modified sample are data patterns. The so-called wet and dry periods are defined as follows:

$$\begin{aligned} \dots - D - \underbrace{W - W - W - \dots - W}_{\text{wet period}} - D - \dots \\ \dots - W - \underbrace{D - D - D - \dots - D}_{\text{dry period}} - W - \dots \end{aligned}$$

### 2.1 Simulation of Incomplete Data

There are two complete precipitation data sets for the period from 1950 to 2009 for analysis. Therefore, the missing values should be artificially inserted into the samples. Then, we can jointly use patterns and machine learning algorithms for imputations and subsequent comparison of our results with true values (in terms of “D–W” classification). It is possible to determine the frequencies of appearance of each pattern as the ratio of the number of such sets of fixed length  $N$  to the total number of possible chains (obviously,  $2^N$ ). Patterns with size  $N = 5$  will be used throughout this work. The detailed description of the corresponding numerical characteristics for this case is given in [12].

In this section, the cases of 1, 2 and 3 consecutive missing values (MVs) are considered, and their total number varies from 5% to 40% of the sample size. That are so called missing levels (for example, see paper [17] where the threshold methodology is used to filling gaps at levels 5%, 10%, and 15–18%). The procedure of inserting the missing values is described using pseudocode, see Algorithm 1. Within this approach, the upper bounds of missing levels are as follows:

- up to 20% for only one MV in the window which size equals 5;
- up to 33% for two consecutive MVs;
- up to 43% for three consecutive MVs.

It explains the different limits for the curves on Figs. 1 and 2 in Sects. 2.2 and 2.3.

**Algorithm 1.** Simulation of incomplete data

---

```

1: INPUT(MV); //Number of consecutive MVs
2:  $\mathcal{I}_0 = \text{DROPOUT}(\text{Sample}, \text{MissLvl})$ ; //Array of indices for possible insertion of MVs
3: if  $\mathcal{I}_0(\text{end}) \geq \text{LENGTH}(\text{Sample}) - N$  then
4:    $\mathcal{I}_0(\text{end}) = []$ ; //MV cannot be inserted into the end of data
5: end if
6:  $k = 0$ ;
7: for  $i = 1 : \text{length}(\mathcal{I}_0) - 1$  do
8:   if  $\mathcal{I}_0(i+1) - \mathcal{I}_0(i) \geq N + MV$  then
9:      $k = k + 1$ ;
10:    for  $j = 0 : MV - 1$  do
11:       $\mathcal{I}(k+j) = \mathcal{I}_0(i) + j$ ; //Array of indices of MVs
12:    end for
13:  end if
14: end for

```

---

**2.2 Imputation Methodology Based on Binary Patterns**

In this section, an algorithm of “pure probabilistic” pattern-based filling is given.

- Let us consider the subsample with a missing value (further it is denoted by symbol  $\mathcal{X}$ ):

$$\dots - D - W - D - D - D - \boxed{\mathcal{X}} - W - D - D - W - \dots$$

- All subsamples of pre-selected length that contain this gap are chosen (the value  $N = 5$  is still used):

$$(a) \dots - D - \boxed{W - D - D - D - \mathcal{X}} - W - D - D - W - \dots$$

$$(b) \dots - D - W - \boxed{D - D - D - \mathcal{X} - W} - D - D - W - \dots$$

$$(c) \dots - D - W - D - \boxed{D - D - \mathcal{X} - W - D} - D - W - \dots$$

$$(d) \dots - D - W - D - D - \boxed{D - \mathcal{X} - W - D - D} - W - \dots$$

$$(e) \dots - D - W - D - D - D - \boxed{\mathcal{X} - W - D - D - W} - \dots$$

- There are only two possible “D–W” patterns in each situation. For example, the subsample from item 2a can be only as follows:

$$W - D - D - D - \boxed{W} \quad \text{or} \quad W - D - D - D - \boxed{D}$$

- To fill the missing value, the pattern with maximum possible frequency should be chosen. The corresponding element in this pattern is considered as a decision:

$$\dots - D - \boxed{W - D - D - D - \mathcal{X}} - W - D - D - W - \dots \Rightarrow W - D - D - D - \boxed{D}$$

$$\dots - D - W - \boxed{D - D - D - \mathcal{X} - W} - D - D - W - \dots \Rightarrow D - D - D - \boxed{W} - W$$

$$\dots - D - W - D - \boxed{D - D - \mathcal{X} - W - D} - D - W - \dots \Rightarrow D - D - \boxed{D} - W - D$$

$$\dots - D - W - D - D - \boxed{D - \mathcal{X} - W - D - D} - W - \dots \Rightarrow D - \boxed{D} - W - D - D$$
$$\dots - D - W - D - D - D - \boxed{\mathcal{X} - W - D - D - W} - \dots \Rightarrow \boxed{W} - W - D - D - W$$

5. Then, from the set of such decisions, the most frequent element (“D” or “W”) should be selected, and then it is used to fill in data:

$$\dots - D - W - D - D - D - \boxed{D} - W - D - D - W - \dots$$

Table 1. Accuracy of pattern-based data imputation, Potsdam.

	Missing levels								
	1%	5%	10%	15%	20%	25%	30%	35%	40%
One MV	72.4%	71.8%	71.5%	71.08%	70.41%	–	–	–	–
Two MVs	68.73%	68.12%	67.98%	66.7%	66.5%	66.32%	66.1%	–	–
Three MVs	54.11%	48.94%	48.63%	47.2%	46.8%	45.87%	45.7%	45.1%	44.7%

On test data, it was found that presence of missing values do not significantly change the frequency for the patterns, so this method can be successfully applied to real incomplete data. Figure 1 and Table 1 demonstrate examples for various missing levels.

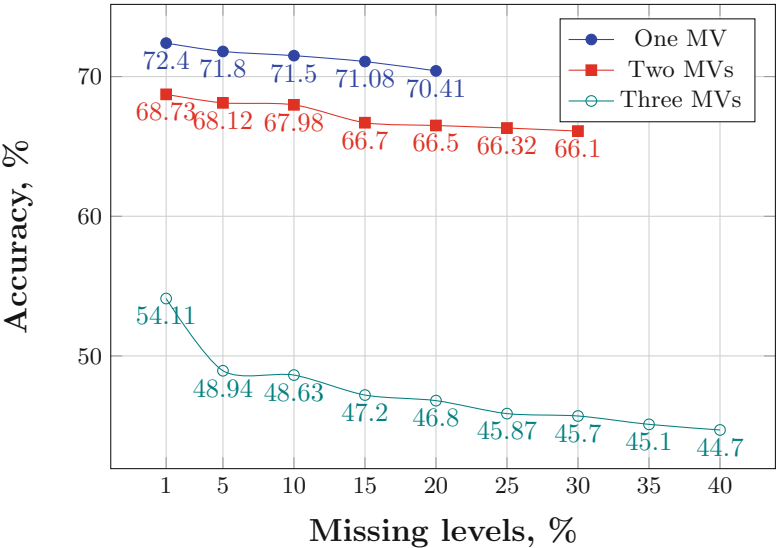


Fig. 1. Accuracy of pattern-based data imputation, Potsdam.

The described algorithm is a very simple and natural to use, however, the obtained values of the filling accuracy, especially for the case of several consecutive gaps, should be increased.

We demonstrate the case of only one consecutive missing value. In the remaining situations, one can act similarly with the corresponding modifications.

**2.3 SVM for Handling the Missing Values**

In this section, we use SVM, a supervised learning model with associated algorithms, to fill in the gaps in data based on “D–W” classification. The following features have been used: min, max and mean temperatures, mean dew point, mean wind speed. The corresponding results will be compared with decisions based on a probabilistic approach described in Sect. 2.2. Figure 2 and Table 2 demonstrate accuracy of SVM-based data imputation for 1, 2 and 3 consecutive missing values.

The x-axis presents missing levels from 1% to 40%, and the y-axis corresponds to a percentage of elements correctly filled by “D” or “W”.

**Table 2.** Accuracy of SVM-based data imputation for patterns, Potsdam.

	Missing levels								
	1%	5%	10%	15%	20%	25%	30%	35%	40%
One MV	80.17%	79.28%	78.41%	78.3%	77.52%	–	–	–	–
Two MVs	77.91%	76.34%	76.05%	75.12%	74.87%	74.31%	74.17%	–	–
Three MVs	75.68%	75.17%	73.69%	73.02%	73.32%	72.02%	71.87%	71.26%	70.19%

With increasing the total number of missing values, accuracy is reduced in all cases for both methods. But the values for a “pure probabilistic” approach are unsuitable for practical use even for 1% level especially if three consecutive MVs are allowed (see Fig. 1). The SVM accuracy even for the 40% missing level and the same number of consecutive MVs does not fall below 70%. Errors for the upper and lower levels differ by no more than 5.5%, that should also be considered as a good result for practice. It is also important to note that the SVM accuracies in the classification problem are close to each other for various numbers of consecutive MVs.

Table 3 presents the comparison of average prediction accuracies of SVM and pure probabilistic pattern-based filling.

For the case of one consecutive missing value, the difference is about 7%, for two and three consecutive MVs is about 8% and 25%, respectively. That is, when increasing the number of MVs, the accuracy of the probabilistic approach dramatically decreases, while SVM is less sensitive to this situation. Thus, SVM accuracy is higher than “pure probabilistic” one. So, these results indicate the possibility of practical applications in real problems. Some examples of using SVM-based classification for observations obtained in another climatic zone are also given in paper [23].

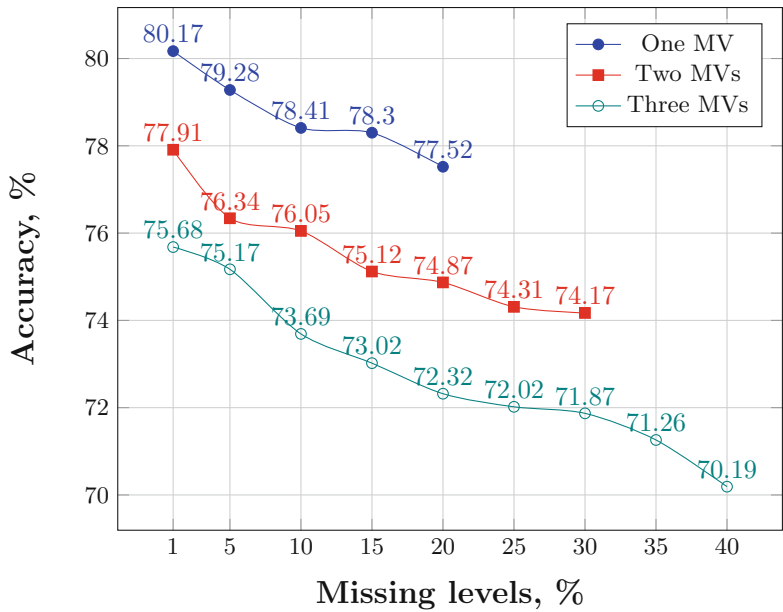


Fig. 2. Accuracy of SVM-based data imputation for patterns, Potsdam.

Table 3. Comparison of average prediction accuracies for “pure probabilistic” approach and SVM (Potsdam).

	Patterns	SVM
One MV	71.44%	78.74%
Two consecutive MVs	67.21%	75.54%
Three consecutive MVs	47.45%	72.8%

### 3 Approach Based on Regression

In this section, machine learning algorithms will be used to fill in the exact values of the MVs, that is, the regression problem will be solved. In this case, the patterns discussed in Sect. 2 will also be used as an auxiliary tool to improve the quality of the methods.

#### 3.1 Simulation of Incomplete Data

In this section, we use a slightly different method of randomly selecting the positions for insertion of missing values.

1. The sample are divided into  $K$  equal parts, where the length of each is 5:

$$\boxed{W - D - D - D - W} - \boxed{W - D - D - W - W} - \dots - \boxed{D - D - D - D - W}$$

2. Depending on the required percentage of the missing values,  $L$  subsamples are randomly selected from  $K$  ones mentioned above to replace with an unknown value:

$$W - D - D - D - W - \boxed{W - D - D - W - W} - \dots - \boxed{D - D - D - D - W}$$

3. Then, in each of the  $L$  subsamples, the position to replace is randomly selected:

$$W - D - D - D - W - \boxed{W - \boxed{\mathcal{X}} - D - W - W} - \dots - \boxed{\boxed{\mathcal{X}} - D - W - D - W}$$

This algorithm is simple to implement, and gaps are filled in a random but a controlled way. It is worth noting that this method also allows us to prevent the consecutive missing values. Using this algorithms, test samples for 1%, 5%, 10%, 15% and 20% levels were simulated using initial complete observations from Potsdam and Elista for about 60 years.

### 3.2 Algorithms for Handling the Missing Values

In this section, we briefly describe the regression machine learning algorithms that will be used to fill missing values.

**Mean imputation**, *Means*, is the simplest method of filling the missing values. All missing elements are replaced by an arithmetic mean of a selected subsample or the full time-series. The corresponding drawbacks are obvious. For example, this method can be not suitable well for the non-stationary time-series and data with outliers.

**EM algorithm** is one of the most popular regression algorithms that allows us to work effectively with large data volumes. It is assumed that the distribution of analyzed data can be approximated by a linear combination of multidimensional normal distributions.

**k-Nearest Neighbors algorithm** is one of the most used non-parametric method for data prediction. The missed observation should be filled by mean of values of  $k$  nearest neighbors. It is worth noting that k-NN is one of the simplest machine learning algorithms.

**Random Forests** are the ensemble learning methods for regression based on decision trees [15]. RFs present mean prediction as output in this problem. An ensemble training is used to obtain better results than could be obtained from any of the constituent methods alone. It is an attractive method for imputing missing data, wherein such data can be analyzed even without filling. Possible modification for better performance is discussed, for example, in [22].

Extreme gradient boosting **XGBoost** [6] is widely used to solve classification and regression problems in a wide range of applications (see, e. g., [5, 25, 29]).

Below, we consider the application of these algorithms to real precipitation data and compare the imputation accuracy for different missing levels. The volume of each test sample is more than 20000 observations.



3.3 Data Imputation with Pattern-Based Classification

Let us suppose that the filling procedure was initially carried out using the pattern-based classification. Thus, the “D–W” sample does not contain gaps. However, the exact values of daily volumes are still unknown. Thus, they need to be imputed using the regression approach.

First, it should be checked whether the missing value belongs to the dry or wet period. Indeed, if the MV is located in a dry period, then the best filling is realized by Means due to the zeros are correctly placed into the corresponding positions. In this case, other algorithms can forecast small but nonzero values, and thus it leads to increase the total error due to the test data contains a lot of zero values.

Table 4. Accuracy of data imputation for dry periods in Potsdam and Elista.

City	Missing levels				
	1%	5%	10%	15%	20%
Potsdam	89.4%	87.7%	84.2%	83.2%	77.3%
Elista	89.6%	87.9%	86.2%	85.1%	83.3%

Table 4 and Fig. 3 demonstrate the accuracies of data imputation for dry periods in Potsdam and Elista.

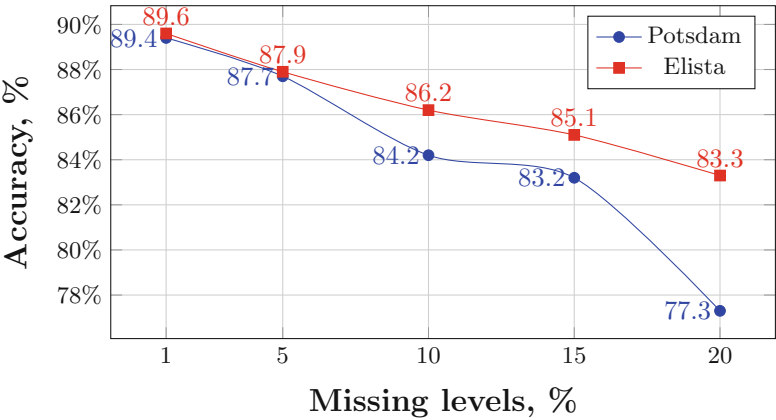


Fig. 3. Accuracy of data imputation for dry periods in Potsdam and Elista.

As a disadvantage, as mentioned earlier, it is worth noting that a pattern-based accuracy strongly decreases with increasing missing levels. The accuracy can be undoubtedly increased using SVM (see Sect. 2).

To evaluate the imputation accuracy for wet periods, the following metric is used:  $\varepsilon_m = V_m^{-1}RMSE_m$ , where  $RMSE_m$  is a Root Mean Square Error corresponding to a  $m$ -th wet period and  $V_m$  is a total precipitation volume of the same period. Note that values  $V_m$  are determined using complete data. In fact, this is the normalization of observations over the wet period. It makes possible to accurately compare errors  $\varepsilon_m$  for different wet periods, and also to compute their mean value for all intervals with gaps.

A row vector  $\varepsilon = \{\varepsilon_m\}_{m=\overline{1,L}}$  corresponds to the consecutive errors for all wet periods containing missing values. Then, the total error  $Err = L^{-1}\varepsilon\mathbf{1}_{L \times 1}$ , where  $\mathbf{1}_{L \times 1}$  is a column vector consisting of  $L$  ones.

**Table 5.** Accuracy of data imputation for wet periods in Potsdam.

Method	Missing levels				
	1%	5%	10%	15%	20%
Means	83.2%	82.4%	80.1%	78.8%	77.3%
k-NN	83.5%	82.3%	80.5%	79.1%	78.4%
EM algorithm	84.1%	83.6%	82.9%	80.4%	78.8%
Random forest	83.4%	82.3%	81.5%	79.7%	77.6%

Tables 5, 6 and Figs. 4, 5 present comparison of various ML algorithms for data imputation in Potsdam and Elista taking into account different missing levels. ML methods were implemented using *scikit-learn* library for the Python programming language.

**Table 6.** Accuracy of data imputation for wet periods in Elista.

Method	Missing levels				
	1%	5%	10%	15%	20%
Means	83.6%	82.4%	82.1%	80.2%	78.3%
k-NN	83.8%	82.9%	81.7%	80.3%	78.4%
EM algorithm	84.3%	83.7%	83.1%	81.2%	78.9%
Random forest	83.7%	82.5%	81.5%	80.3%	78.8%

In most cases, the Means accuracy is minimal among all compared methods. The exception is demonstrated for Elista on the 10% missing level (see Fig. 5). The differences for Potsdam on the 5% missing level (see Fig. 4) can be explained by computational errors.

The best results for all situations are given by EM algorithm. On test data, it turned out that EM and Means set the upper and lower bounds of accuracies. For Potsdam data on 20% missing level, EM accuracy is 1.5% more compared to

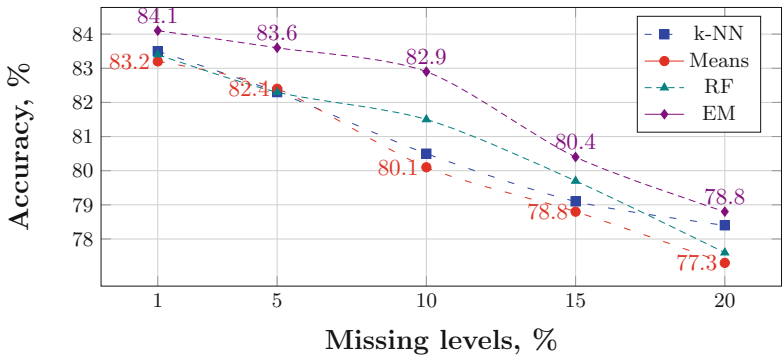


Fig. 4. Accuracy of data imputation for wet periods in Potsdam.

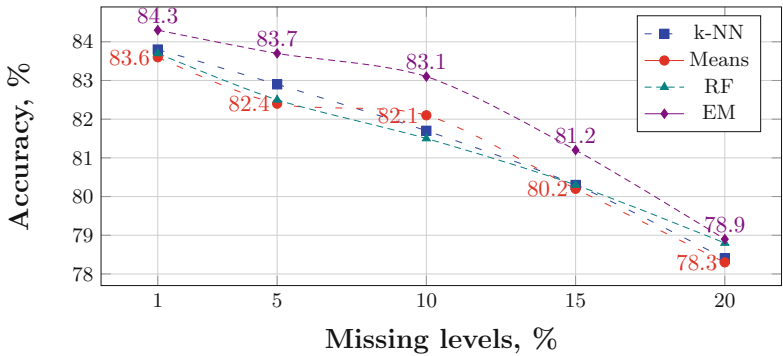


Fig. 5. Accuracy of data imputation for wet periods in Elista.

Means one (see Table 5). This is a significant advantage for real data. For Elista, this difference is about 1% (see Table 6). With the increasing missing levels, the RF results tends to EM values.

3.4 Data Imputation with SVM Classification

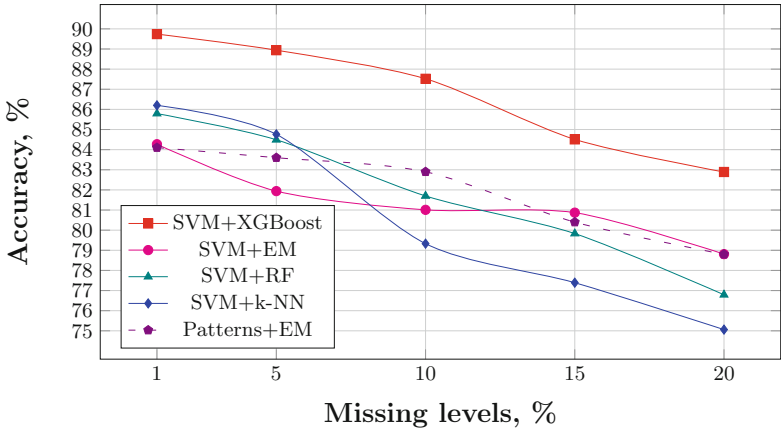
Let us suppose that the filling procedure was initially carried out using improved machine learning classification based on SVM (see Sect. 2.3). Tables 7 and 8 present improved results for various regression algorithms for one and two consecutive MV's.

Comparing Tables 5 and 7, it is easy to see that extreme gradient boosting provides a five percent advantage over the best results of Sect. 3.3. This is most clearly shown in Fig. 6.

The obvious advantage of this approach allows maintaining the accuracy of forecasting continuous data at a level of more than 80%. The results for two consecutive passes are presented in Table 8. It can be seen that the accuracy decreases significantly with increasing missing levels, but for SVM with XGBoost

**Table 7.** Improved accuracy of classification-regression data imputation: one MV, Potsdam.

Missing levels	Method			
	SVM+XGBoost	SVM+EM	SVM+RF	SVM+k-NN
1%	89.74%	84.27%	85.79%	86.20%
5%	88.94%	81.94%	84.49%	84.76%
10%	87.52%	81.01%	81.70%	79.33%
15%	84.51%	80.87%	79.83%	77.39%
20%	82.89%	78.81%	76.79%	75.06%



**Fig. 6.** Improved accuracy of data imputation in Potsdam.

**Table 8.** Improved accuracy of classification-regression data imputation: two MV's, Potsdam.

Missing levels	Method			
	SVM+XGBoost	SVM+EM	SVM+RF	SVM+k-NN
1%	82.63%	79.51%	71.96%	83.75%
5%	76.37%	75.15%	71.56%	74.87%
10%	75.92%	73.32%	72.04%	70.11%
15%	74.27%	72.67%	68.22%	67.14%
20%	73.74%	71.08%	66.22%	67.95%
25%	73.19%	69.88%	68.33%	70.59%
30%	71.39%	68.79%	65.99%	66.53%

**Table 9.** The learning rates in seconds.

Missing levels	Method			
	SVM+XGBoost	SVM+EM	SVM+RF	SVM+k-NN
1%	2.03	1.71	1.98	1.82
5%	1.99	1.93	1.88	1.74
10%	1.89	2.01	1.76	1.60
15%	1.64	2.55	1.54	1.39
20%	1.53	2.89	1.29	1.13
25%	1.18	3.05	0.98	0.84
30%	1.00	3.13	1.01	0.77

it remains above 70%. Moreover, it follows from Table 9 that the learning rate for this method is quite comparable with others, therefore, it can be used for solving real-time problems.

#### 4 Conclusion

The paper presents two approaches to filling gaps in precipitation based on classification and regression machine learning algorithms as well as the pattern-driven methodology. The possibility of correct imputation even for high missing levels is demonstrated. The Python implementations lead to the possibility of using the high-performance computing, in particular, to solve the problem of the effective selection of various hyperparameters. The obtained results are quite suitable for the analysis of real incomplete data. Therefore, the analysis of data sets from distributed weather stations in Russia and neighboring countries, Europe, Asia, etc., should be mentioned as a direction for further researches. In addition, these methods can be applied to data of a different nature, in particular, to various information systems.

For the observed data, the optimal algorithm is based on SVM classification and XGBoost regression. However, the MVs problem does not have a universal solution, and there is no the only method that would be superior in quality to all others for all situations. Each ensemble requires an individual approach taking into account the physical nature of data. The further research in this area can also be focused on involving neural networks in classification, regression or both stages, because an accuracy of about 97% was obtained [14] for one-step pattern-based forecasts.

It is worth noting that the suggested methodology can be successfully used for real-time data processing of information flows [9,10]. For example, it can be useful for telecommunication loads or traffic, where different states exist due to the packet loss or hacker attacks.

**Acknowledgments.** The authors are grateful to **Professor V. Yu. Korolev** for useful discussions and joint researches. The authors would like to thank **Professor**

**K. E. Samuylov** for careful reading of material and the valuable comments that helped us to improve the manuscript.

## References

1. Altman, N.: An introduction to kernel and nearest-neighbor nonparametric regression. *Am. Stat.* **46**(3), 175–185 (1992). <https://doi.org/10.1080/00031305.1992.10475879>
2. Athey, S., Tibshirani, J., Wager, S.: Generalized random forests. *Ann. Stat.* **47**(2), 1148–1178 (2019). <https://doi.org/10.1214/18-AOS1709>
3. Barrios, A., Trincado, G., Garreaud, R.: Alternative approaches for estimating missing climate data: application to monthly precipitation records in South-Central Chile. *For. Ecosyst.* **5**, 28 (2018). <https://doi.org/10.1186/s40663-018-0147-x>
4. Breiman, L.: Random forests. *Mach. Learn.* **45**, 5–32 (2001). <https://doi.org/10.1023/A:1010933404324>
5. Chatzis, S., Siakoulis, V., Petropoulos, A., Stavroulakis, E., Vlachogiannakis, N.: Forecasting stock market crisis events using deep and statistical machine learning techniques. *Expert Syst. Appl.* **112**, 353–371 (2018). <https://doi.org/10.1016/j.eswa.2018.06.032>
6. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794 (2016). <https://doi.org/10.1145/2939672.2939785>
7. Cortes, C., Vapnik, V.N.: Support-vector networks. *Mach. Learn.* **20**(3), 273–297 (1995). <https://doi.org/10.1007/BF00994018>
8. Fernandez-Gonzalez, P., Bielza, C., Larranaga, P.: Random forests for regression as a weighted sum of k-potential nearest neighbors. *IEEE Access* **7**, 25660–25672 (2019). <https://doi.org/10.1109/ACCESS.2019.2900755>
9. Gorshenin, A., Kuzmin, V.: Online system for the construction of structural models of information flows. In: *Proceedings of the 7th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops*, pp. 216–219 (2015). <https://doi.org/10.1109/ICUMT.2015.7382430>
10. Gorshenin, A., Kuzmin, V.: On an interface of the online system for a stochastic analysis of the varied information flows. *AIP Conf. Proc.* **1738**(220009) (2016). <https://doi.org/10.1063/1.4952008>
11. Gorshenin, A.: Pattern-based analysis of probabilistic and statistical characteristics of precipitations. *Informatika i ee Primeneniya* **11**(4), 38–46 (2017). <https://doi.org/10.14357/19922264170405>
12. Gorshenin, A.: Investigation of parameters of meteorological models based on patterns. In: *CEUR Workshop Proceedings*, vol. 2177, pp. 4–10 (2018). <http://ceur-ws.org/Vol-2177/paper-01-a005.pdf>
13. Gorshenin, A., Korolev, V.: Determining the extremes of precipitation volumes based on a modified “Peaks over Threshold”. *Informatika i ee Primeneniya* **12**(4), 16–24 (2018). <https://doi.org/10.14357/19922264180403>
14. Gorshenin, A., Kuzmin, V.: Neural network forecasting of precipitation volumes using patterns. *Pattern Recognit. Image Anal.* **28**(3), 450–461 (2018). <https://doi.org/10.1134/S1054661818030069>
15. Ho, T.: The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.* **20**(8), 832–844 (1998). <https://doi.org/10.1109/34.709601>

16. Kalteh, A., Hjorth, P.: Imputation of missing values in a precipitation-runoff process database. *Hydrol. Res.* **40**(4), 420–432 (2009). <https://doi.org/10.2166/nh.2009.001>
17. Kim, J., Ryu, J.: Quantifying a threshold of missing values for gap filling processes in daily precipitation series. *Water Resour. Manag.* **29**(11), 4173–4184 (2015). <https://doi.org/10.1007/s11269-015-1052-5>
18. Korolev, V.Y.: Probabilistic and Statistical Methods of Decomposition of Volatility of Chaotic Processes. Moscow University Publishing House, Moscow (2011)
19. Lulli, A., Oneto, L., Anguita, D.: Mining big data with random forests. *Cogn. Comput.* **11**(2), 294–316 (2019). <https://doi.org/10.1007/s12559-018-9615-4>
20. Sattari, M., Rezazadeh-Joudi, A., Kusiak, A.: Assessment of different methods for estimation of missing data in precipitation studies. *Hydrol. Res.* **48**(4), 1032–1044 (2017). <https://doi.org/10.2166/nh.2016.364>
21. Simolo, C., Brunetti, M., Maugeri, M., Nanni, T.: Improving estimation of missing values in daily precipitation series by a probability density function-preserving approach. *Int. J. Climatol.* **30**(10), 1564–1576 (2010). <https://doi.org/10.1002/joc.1992>
22. Tang, F., Ishwaran, H.: Random forest missing data algorithms. *Stat. Anal. Data Min.* **10**(6), 363–377 (2017). <https://doi.org/10.1002/sam.11348>
23. Teegavarapu, R., Aly, A., Pathak, C., Ahlquist, J., Fuelberg, H., Hood, J.: Infilling missing precipitation records using variants of spatial interpolation and data-driven methods: use of optimal weighting parameters and nearest neighbour-based corrections. *Int. J. Climatol.* **38**(12), 776–793 (2018). <https://doi.org/10.1002/joc.5209>
24. Teegavarapu, R., Chandramouli, V.: Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. *J. Hydrol.* **312**(1–4), 191–206 (2005). <https://doi.org/10.1016/j.jhydrol.2005.02.015>
25. Torres-Barran, A., Alonso, A., Dorronsoro, J.: Regression tree ensembles for wind energy and solar radiation prediction. *Neurocomputing* **326**, 151–160 (2019). <https://doi.org/10.1016/j.neucom.2017.05.104>
26. Wang, W., Du, X., Wang, N.: Building a cloud IDS using an efficient feature selection method and SVM. *IEEE Access* **7**, 1345–1354 (2019). <https://doi.org/10.1109/ACCESS.2018.2883142>
27. Wu, X., et al.: Top 10 algorithms in data mining. *Knowl. Inf. Syst.* **14**(1), 1–37 (2008). <https://doi.org/10.1007/s10115-007-0114-2>
28. Yang, N., Wang, Y.: Identify silent data corruption vulnerable instructions using SVM. *IEEE Access* **7**, 40210–40219 (2019). <https://doi.org/10.1109/ACCESS.2019.2905842>
29. Zhang, D., Qian, L., Mao, B., Huang, C., Huang, B., Si, Y.: A data-driven design for fault detection of wind turbines using Random Forests and XGboost. *IEEE Access* **6**, 21020–21031 (2018). <https://doi.org/10.1109/ACCESS.2018.2818678>
30. Zolina, O., Simmer, C., Belyaev, K., Kapala, A., Gulev, S.: Improving estimates of heavy and extreme precipitation using daily records from European rain gauges. *J. Hydrometeorol.* **10**, 701–716 (2009). <https://doi.org/10.1175/2008JHM1055.1>