# Practical Estimation of Missing Phosphorus Values in Pyhäjärvi Lake Data

Alexander Grigorievskiy[1], Anton Akusok[1], Marjo Tarvainen[2], Anne-Mari Ventelä[2], and Amaury Lendasse[1,3,4]

[1] Aalto University, Department of Information and Computer Science, PO Box 15400, FI-00076 Aalto , Finland
{alexander.grigorevskiy,amaury.lendasse}@aalto.fi
[2] Pyhäjärvi Institute, Sepäntie 7, FIN-27500 Kauttua, Finland
{marjo.tarvainen,anne-mari.ventela}@pji.fi
[3] IKERBASQUE, Basque Foundation for Science, 48011 Bilbao, Spain
[4] Computational Intelligence Group, Computer Science Faculty, University of the Basque Country, Paseo Manuel Lardizabal 1, Donostia-San Sebastián, Spain

**Abstract.** Practical problem of missing values estimation of phosphorus concentration is addressed in this paper. There are several covariates which can be used to estimate phosphorus in Pyhäjärvi lake, however some of them also contain missing data. In addition, variable selection needs to be done in order to increase accuracy of modeling and facilitate understanding of underlying dependencies. We address the problem by first, Delta test variable selection and then by regression approach with Ridge Regression, SVM and LS-SVM accompanied with *wrapper* variable selection. It is shown that for some time periods it is possible to improve estimations from regression by averaging them with missing values imputation methods like Empirical Orthogonal Functions (EOF).

**Keywords:** Environmental Modeling, Missing values, Regression, Support Vector Machine, SVM, Least-Squares Support Vector Machine, LS-SVM, Empirical Orthogonal Functions, EOF

## 1 Introduction and Work Motivation

Pyhäjärvi lake is a large lake located on the south-west of Finland. The lake plays an important role in the local agriculture and fishing industries. Due to the human activity and changing climate the ecology of the lake has been challenged [1]. The main substance that influence the ecological balance in the lake is phosphorus. Therefore, it is very important to model and analyze phosphorus concentration in order to develop adequate measures for the lake protection.

Complication, which is frequently encountered when dealing with environmental data, is the presence of missing values. Measurements are often taken manually by humans and cases like spoiling the sample or sickness of a particular person are not exceptions. Selecting the best subset of covariates is also an important step in the modeling. In this work, the goal is to estimate concentration of phosphorus in various locations of Pyhäjärvi lake for the time period

26.03.1991 - 21.04.2008. Some values of phosphorus are given but many are missing. Some covariates also contain missing values. To shorten the exposition, data only for one location is considered in what follows.

The paper organization is the following: in the next section description of the dataset is provided. In the Section 3 regression approach to phosphorus concentration estimation is given. Then follows the missing values approach and finally conclusions.

## 2 Dataset Description

The dataset is shown in the Table 1. There are 16 variables (columns) and 1230 rows in the dataset. Each row correspond to averaged value of corresponding variable over 5 day interval. This interval is called "Week" for brevity. In the column "Complete dataset" number of present values of different variables is given. The variable to estimate is the second one - "Total P S11".

**Table 1.** Dataset and amounts of present values in variables

| | | Dataset | | |
|---|---|---|---|---|
| No. | Variable name | Complete dataset (1230 rows) | Part 1 (351 rows) | Part 2 (271 rows) |
| 1 | "Flow S11" | 1230 (full) | 351 (full) | 271 (full) |
| 2 | "Total P S11" | 227 | 59 | 58 |
| 3 | "Total P S10" | 225 | 60 | 58 |
| 4 | "Total P S12" | 226 | 58 | 72 |
| 5 | "Temperature" | 1228 | 351 (full) | 271 (full) |
| 6 | "Integrated Flow S11" | 1230 (full) | 351 (full) | 271 (full) |
| 7 | "Smoothed Flow S11" | 1230 (full) | 351 (full) | 271 (full) |
| 8 | "Rains" | 1230 (full) | 351 (full) | 271 (full) |
| 9 | "Sin Week" | 1230 (full) | 351 (full) | 271 (full) |
| 10 | "Cos Week" | 1230 (full) | 351 (full) | 271 (full) |
| 11 | "Time shift 1 Ph. S11" | 226 | 58 | 57 |
| 12 | "Time shift 2 Ph. S11" | 226 | 58 | 57 |
| 13 | "Time shift 1 Ph. S10" | 225 | 59 | 57 |
| 14 | "Time shift 2 Ph. S10" | 225 | 59 | 57 |
| 15 | "Time shift 1 Ph. S12" | 225 | 57 | 71 |
| 16 | "Time shift 2 Ph. S12" | 225 | 57 | 71 |

The sparsity of the dataset is 46% i. e. almost half of all the values are absent. However, missing values are distributed in time non uniformly. For variable "Total P S11" there are large periods (up to a year) when no data is present and periods where gaps are relatively small (several "Weeks"). Preliminary tests showed that missing values imputation methods provide good estimation of "Total P S11", when there are no large gaps between given values of this variable. Therefore, missing values imputation methods are applied only to datasets named "Part 1" (03.1991-02.1996) and "Part 2" (03.1997-12.2000) which correspond to time intervals with no big gaps between given values of "Total P S11". Regression modeling is conducted for the complete dataset.

Not all variables might be useful for phosphorus concentration estimation. One goal of this work is to select relevant variables and discard irrelevant.

**Regression Dataset.** Regression dataset (Table 2) is constructed from the complete dataset in the Table 1 by taking covariates where no missing data occurs (including "Temperature").

**Table 2.** Regression dataset

| No. | Variable name |
|---|---|
| **To predict** | **"Total P S11"** |
| 1 | "Flow S11" |
| 2 | "Temperature" |
| 3 | "Integrated Flow S11" |
| 4 | "Smoothed Flow S11" |
| 5 | "Rains" |
| 6 | "Sin Week" |
| 7 | "Cos Week" |
| 8 | "Rains Int. 1" |
| 9 | "Rains Int. 2" |
| ⋮ | ⋮ |
| 17 | "Rains Int. 10" |

The variable to predict is "Total P S11". The number of training samples is 227 and equals the number of present values in "Total P S11" variable. Having trained the regression model, it is possible to estimate phosphorus concentration on all other "Weeks" when it is missing. This is called regression approach and it is compared to missing values approach described in details in Section 4. Since missing values approach is studied only during periods "Part 1" and "Part 2", for all other "Weeks" regression approach is used to estimate "Total P S11". Ten additional variables No. 8-17 are added to the regression dataset. They are integrated values of "Rains" over 1 "Week" and so forth up to 10 "Weeks". The motivation for including these variables is to check possibility that phosphorus concentration depends on accumulated precipitation intensity during a long period.

In the following sections, we consider regression and then missing values approaches.

## 3 Regression Approach to Phosphorus Concentration Estimation

Regression approach has been applied to the data in Table 2. Three regression models are evaluated, and the best one which has smallest normalized mean square error (NMSE) is selected. First model is a linear one - Ridge Regression, and the other two are nonlinear Support Vector Regression (SVR) and Least-Squares Support Vector Regression (LS-SVR). Nonlinearity is obtained by using Gaussian kernel.

One thing that can deteriorate regression models is the presence of irrelevant, redundant, or too noisy input variables. Those can increase computational time, contribute to the curse of dimensionality and, finally, reduce accuracy of the regression [2]. In addition, selecting of only useful variables facilitates interpretability of the model.

### 3.1 Variable Selection

There exist many methods for variable selection. Overview of some of them is presented in [2] and [3]. These methods can be divided into three main categories: filters, wrappers and embedded methods. Filter methods optimize some external criteria and select a subset of input variables which corresponds to the optimum. Advantage of filter methods is that they are usually faster to compute than other types of methods, but disadvantage is that they doesn't take into account data model used during learning process. Wrapper methods utilize learning machine as a black box method to score different subsets of input variables. Multiple retraining of learning algorithm and measuring performance on

a separate validation set are usually required. This is a main disadvantage of this class of methods.

In this work, hierarchical variable selection is applied. On the first step less accurate but more computationally efficient filter method is used - Delta test [4],[5], on the second step, when less variables are left for analysis, wrapper method is utilized.

Based on the results of Delta test variables are divided onto 3 groups. First group is completely irrelevant variables which are discarded from subsequent investigation. The second group is important variables which are always kept. And finally to the third group attributed variables which are investigated through wrapper approach by passing all possible their combinations through the regression algorithm and measuring NMSE on validation set.

**Table 3.** Variable selection via Delta test for regression datasets

| No. of samples | Relevant variables | Variables to be investigated further |
|---|---|---|
| 227 | "Flow S11" , "Temperature", "Integrated Flow S11" | "Smoothed Flow S11", "Rains", "Sin Week", "Cos Week", "Rain int 1", "Rain int 4" |

Variables in the right most column are investigated further through a wrapper approach. Actually, three regression models are considered and selection of the best subset of variables is done along with selection of the best model.

### 3.2 Regression Models

Three regression models have been analyzed in this work. One linear - Ridge regression, and two nonlinear Support Vector Regression (SVR) and Least Squares Support Vector Regression (LS-SVR) [6].

Regularization parameter $\lambda$ in Ridge regression is adjusted via second internal cycle of cross validation. Gaussian kernel functions are used in SVR and LS-SVR. There are three hyper-parameters to adjust in SVR formulation: $C$ - regularization parameter, $\epsilon$ - width of a tube inside which no penalty for a point occurs, and $\sigma$ - width of a Gaussian kernel. We have utilized method of Cherkassky and Ma [7] followed by *pattern search* [8] to tune these parameters. LS-SVM Toolbox for Matlab has been used for LS-SVR modelling. Parameter optimization in this toolbox is done through coupled simulated annealing algorithm [9] and fine tuning through simplex method and cross-validation.

### 3.3 Regression Results

Before applying regression modeling all input variables and output variable have been normalized to have zero mean and unit variance. Generalization error of

different models and different subsets of input variables is measured by Monte-Carlo 15-fold cross-validation which is repeated 50 times. Number of folds is increased in comparison with standard 10 because number of samples in each dataset is small, and there is a need to increase number of samples for training. Regression model and subsets of variables which have the smallest NMSE are presented in the Table 4. We see that the best model is LS-SVM and the worst

**Table 4.** Relevant variables and best models for the regression dataset

| Best model | Relevant variables | $NMSE \pm (std)$ |
|---|---|---|
| LS-SVR | "Flow S11" , "Temperature", "Integrated Flow S11", "Smoothed flow S11", "Sin Week", "Cos Week" | $0.530 \pm (0.312)$ |
| Ridge R. | "Flow S11" , "Temperature", "Integrated Flow S11", "Sin Week", "Int. Rain 2", "Int. Rain 5" | $0.675 \pm (0.394)$ |
| SVM | "Flow S11" , "Temperature", "Integrated Flow S11", "Smoothed flow S11", "Sin Week", "Cos Week" | $0.570 \pm (0.359)$ |

one is Ridge Regression. This indicates the fact that dataset is highly nonlinear. SVM is the second best model. We suppose that the hyper-parameter selection strategy of LS-SVM toolbox is superior over the method we use for SVM.

The most relevant variables are the same for LS-SVM and SVM. Except "Rains" variable, all relevant variables form the application domain point of view are selected as important. Several subsets of variables are analyzed further in the missing values imputation approach.

## 4 Missing Values Approach to Phosphorus Concentration Estimation

Missing values datasets have been described in the Section 2. There are two datasets named "Part 1" and "Part 2". They correspond to time intervals when measurements of phosphorus are not very sparse. They include all 16 variables from the Table 1.

Regression modeling allows estimating phosphorus concentration when it is unknown. However, in regression modeling the sequential nature of the data is not taken into account. By utilizing missing values approach we are able to account for this and also include additional predictors (covariates) which themselves contain missing values. Importance of several subsets (Table 5) of input variables has been analyzed as well. Therefore, for periods for which missing values datasets are constructed, improved estimation of phosphorus concentration is obtained.

Generally, missing values imputation is a wide area of research with many applications [10], so it is hardly possible to try all the methods. Therefore, only

a subset from different classes of methods is selected and subsequent ensemble averaging is utilized to lighten possible disadvantages of a single method. Each method takes as input a matrix with missing values, fills missing values and returns the complete matrix. Due to the space constraints we describe only one method - Empirical Orthogonal Functions (EOF) [11]. It is a widely used method in meteorology and climate research for missing values imputation and is based on Singular Value Decomposition (SVD) (Algorithm 1). For other two: Mixture of Gaussians (MoF) [12],[13] and Singular Value Thresholding (SVT) [14] we redirect to the original articles.

**Algorithm 1** Empirical Orthogonal Functions

Given the incomplete matrix $\boldsymbol{X} \in \mathbb{R}^{m,n}$

1: Make initial imputation $\boldsymbol{X}^0$, for example, by column means
2: i = 0 (iteration number)
3: **repeat**
4: Perform SVD: $\boldsymbol{X}^i = \boldsymbol{U}^i \boldsymbol{D}^i (\boldsymbol{V}^i)^T$ to obtain $\boldsymbol{U}^i, \boldsymbol{D}^i$and$(\boldsymbol{V}^i)^T$
5: Nullify $K$ smallest singular values of $\boldsymbol{D}^i$. Denote this modified matrix as $\boldsymbol{D}_0^i$
6: Do inverse transformation: $\boldsymbol{X}_0^i = \boldsymbol{U}^i \boldsymbol{D}_0^i (\boldsymbol{V}^i)^T$
7: Restore exactly known values: $known(\boldsymbol{X}_0^i) = known(\boldsymbol{X}^0)$
8: i = i + 1 (iteration number)
9: **until** Convergence

### 4.1 Model Selection for Missing Values Approach

**Combining different models.** It is possible to select only one model based on the lowest NMSE of cross-validation, however there is s reason to keep all three and do an ensemble (*e.g.* see [15, p. 656]) Since regression can provide estimations of phosphorus it is also included in the ensemble.

Ensemble is done via arithmetic averaging of predictions from different models. However, even further improvement can be achieved if we choose which of the models to include in the averaging. There are five models we are investigating, namely "Regression", "Mixture of Gaussians 1 component" (MM1), "Mixture of Gaussians 2 components" (MM2), "SVT", "EOF".

**Experimental Setup.** Experiments are done in the similar way as regression experiments. Accuracy of imputation is characterized by Normalized Mean Squared Error (NMSE) and is measured by Monte-Carlo 15-fold cross-validation. There are 50 iterations in total, on each of those dataset is randomly permuted. The final estimation of NMSE is an average over folds within one iteration and total average over all iterations. Iterations of cross-validation are required because datasets are very small - only about 225 samples.

There are two missing values datasets "Part 1" and "Part 2" as described in Section 2. They are processed simultaneously in the cross-validation cycle. For each dataset, averaging estimations of all possible combinations of five models

and three subsets of variables (Table 5) is analyzed in terms of NMSE and standard deviation (STD) of NMSE.

### 4.2 Model Selection Results

Results of the model selection are presented in the Table 5. Three groups of variables which are interesting from the interpretation point of view have been analyzed. In particular, usefulness of "Rains" variable which has been rejected on the regression phase, as well as time shifted versions of phosphorus "Time shift 1 Ph. S11", "Time shift 2 Ph. S12".

**Table 5.** Groups of variables which have been tested for missing values imputation

| Missing Values Imputation Results | | | | |
|---|---|---|---|---|
| No. | Variable Name | Group 1 | Group 2 | Group 3 |
| 1 | "Flow S11" | ✓ | ✓ | ✓ |
| 2 | "Total P S11" | ✓ | ✓ | ✓ |
| 3 | "Total P S10" | ✓ | ✓ | ✓ |
| 4 | "Total P S12" | ✓ | ✓ | ✓ |
| 5 | "Temperature" | ✓ | ✓ | ✓ |
| 6 | "Integrated Flow S11" | ✓ | ✓ | ✓ |
| 7 | "Smoothed Flow S11" | ✓ | ✓ | ✓ |
| 8 | "Rains" | ✓ | | |
| 9 | "Sin Week" | ✓ | ✓ | ✓ |
| 10 | "Cos Week" | ✓ | ✓ | ✓ |
| 11 | "Time shift 1 Ph. S11" | ✓ | ✓ | |
| 12 | "Time shift 2 Ph. S11" | ✓ | ✓ | |
| 13 | "Time shift 1 Ph. S10" | ✓ | ✓ | |
| 14 | "Time shift 2 Ph. S10" | ✓ | ✓ | |
| 15 | "Time shift 1 Ph. S12" | ✓ | ✓ | |
| 16 | "Time shift 2 Ph. S12" | ✓ | ✓ | |
| Best model combination 10001: "Regression LS-SVM", "EOF" | | | | |
| $NMSE \pm std$, Part 1 | | $\mathbf{0.503 \pm 0.599}$ | $0.504 \pm 0.634$ | $0.503 \pm 0.637$ |
| $NMSE \pm std$, Part 2 | | $\mathbf{0.343 \pm 0.611}$ | $0.340 \pm 0.665$ | $0.340 \pm 0.662$ |

It turns out that the best model combination is an average of estimations of LS-SVM Regression and EOF. Actually, the best variable subset and best model combination is selected as compromise between two "Part 1" and "Part 2" datasets. The reason is that sometimes one model combination is better for "Part 1" and another one is better for "Part 2". So, resulting table is produced by manually inspecting NMSE and STD for various sets of variables and model combinations, and choosing the one with good results for both "Part 1" and "Part 2". It is seen that the first group of variables is the best one in terms of NMSE and STD, which means that all variables included in the dataset carry some useful information.

It is seen that STD is higher than NMSE in all cases. This is an indicator of the fact that some extreme values of phosphorus concentration is very hard to predict.

## 5 Conclusions

Practical problem of phosphorus concentration estimation has been addressed in this article. Two stage approach has been developed where on the first stage regression problem with only complete covariates have been solved and on the second stage improvements by missing values method has been made. Selection of the best regression model and variable selection have been done along.

Empirical Orthogonal Functions(EOF) method in combination with LS-SVM achieved the best accuracy for predicting phosphorus concentration.

In the future, other classes of methods are intended to be applied for the problem. We plan to use existing methods or develop new ones which can do nonlinear regression with missing values in the covariates.

## References


1. Ventelä, A.M., Kirkkala, T., Lendasse, A., Tarvainen, M., Helminen, H., Sarvala, J.: Climate-related challenges in long-term management of säkylän pyhäjärvi (SW finland). Hydrobiologia **660** (2011) 49–58
2. François, D.: High-Dimensional Data Analysis. VDM Publishing (2008)
3. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3** (March 2003) 1157–1182
4. Guillén, A., Sovilj, D., Mateo, F., Rojas, I., Lendasse, A.: Minimizing the delta test for variable selection in regression problems. International Journal of High Performance Systems Architecture **1**(4) (2008) 269–281
5. Jones, A.: New tools in non-linear modelling and prediction. Computational Management Science **1**(2) (07 2004) 109–149
6. Suykens, J.: Least Squares Support Vector Machines. World Scientific (2002)
7. Cherkassky, V., Ma, Y.: Practical selection of svm parameters and noise estimation for svm regression. Neural Netw. **17**(1) (2004) 113–126
8. Marin-Galiano, M., Luebke, K., Christmann, A., Rüping, S.: Determination of hyper-parameters for kernel based classification and regression. Technical Report / Universität Dortmund, SFB 475 Komplexitätsreduktion in Multivariaten Datenstrukturen 2005,38 (2005)
9. de Souza, S.X., Suykens, J.A.K., Vandewalle, J., Bollé, D.: Coupled simulated annealing. IEEE Transactions on Systems, Man, and Cybernetics, Part A **40**(2) (2010) 320–335
10. Little, R.J., Rubin, D.B.: Statistical Analysis with Missing Data. 2 edn. Wiley (2002)
11. Preisendorfer, R., Mobley, C.: Principal component analysis in meteorology and oceanography. Developments in atmospheric science. Elsevier (1988)
12. Ghahramani, Z., Jordan, M.I.: Learning from incomplete data. Technical report, Cambridge, MA, USA (1994)
13. Eirola, E., Lendasse, A., Vandewalle, V., Biernacki, C.: Mixture of gaussians for distance estimation with missing data. In: Machine Learning Reports 03/2012. (2012) 37–45 Proceedings of the Workshop - New Challenges in Neural Computation 2012.
14. Cai, J.F., Candès, E.J., Shen, Z.: A singular value thresholding algorithm for matrix completion. SIAM J. on Optimization **20**(4) (March 2010) 1956–1982

15. Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag New York, Inc., Secaucus, NJ, USA (2006)