Environment. Technology. Resources. Rezekne, Latvia
Proceedings of the 12th International Scientific and Practical Conference. Volume II, 104-108

Forecasting Missing Data Using Different Methods for Road Maintainers

Jānis Pekša
Institute of Information Technology
Faculty of Computer Science and
Information Technology
Riga Technical University
Kalku street 1, Riga, Latvia
Janis.Peksa@rtu.lv

***Abstract*—Observations collected from meteorological stations that are available to road maintainers and used for experimental purposes in this paper. Unfortunately, these observations are insufficient to make good forecasting that is needed for road maintainers. Those meteorological stations are located next to the road surface in the territory of the Republic of Latvia. The road maintainers can make forecasting using this data what is needed for the winter months. It is up to the road maintainers in winter months to process decision-making on road surface smudging with anti-slip chemical materials. The missing data in each meteorological station exists from time to time. The paper represents the possibility of using several approaches to fill out these missing data. This process is needed to be more accurate in predicting specific parameters aggregated from meteorological stations. These approaches are compared between the three closest meteorological stations available in the Republic of Latvia. The relevant data are for the winter months of 2017-2018. To conclude which is more accurate with VAS "Latvijas valsts celi" data set.**

***Keywords*— missing data, time-series, forecasting.**

I. Introduction

Meteorological stations for road maintainers today's data is needed without missing observations collected from several stations located in one region, such as the Republic of Latvia, where meteorological measurements are recorded. Using these complete data, conclusions and decisions can be drawn by observations with the same spacing, where system statistics information is stored sequentially [1]. However, the missing information in the time-series of meteorological stations is unavoidable, owing to the full observation of all the continuous processes is almost impossible [2]. If observation stops for any reason, then there is a problem with observations, in other words, if time-series data are missing they need to be filled for incomplete data. Information flows often result in missing data for many reasons, including sensor failures in meteorological stations, recording of observations errors and network lags [3]. Whereas the weather series contains total time and space characteristics, reconstruction of the missing period needs to be done carefully without characteristics of time-series statistics for interference. Widely for this purpose, the method used in the literature is the average value calculating the value in the time-series [4]. There are other methods which also considers the fleeting behavior of the time-series.

Currently, Latvia has 30 actives of 52 total meteorological stations near the road surface. However, a significant number of these stations suffering from incomplete data. Also, some time-series are suffering from a problem of inhomogeneity. These problems affect not only in the Republic of Latvia but also for the other countries' collecting meteorological time-series data [1]. Therefore, working with meteorological data must cope with these problems before any analysis is carried out. Road maintenance is very urgent in Latvia directly during the winter months, when the country roads are treated with anti-slip material. For road maintainers, forecasting from meteorological station data would allow for better decision-making, whether there is a need for anti-slip materials at a section of the road surface in future. The maintainers to make a clear-cut on the decision; therefore, requires a series of actions. Time-series should consider a specific arch. The road maintainers need predictions in the near future, which is aimed 30 minutes ahead [5]. First, there needs to be handled missing data problem. A few methods can be applied like:

- Normal ratio method – estimated by weighing the data points at various meteorological stations by the ratios of normal annual observations [6];
- k-nearest neighbors algorithm – the closest neighbor algorithm (k-NN) is a non-parameter method used for classification and regression. The input is made up of examples of the nearest training in the function space [7];
- Multilayer perceptron neural network – a class of feedforward artificial neural network. A multilayer perceptron neural network consists of at least three layers of node: the input layer, hidden layer, and the output layer. Except for the input nodes, each node is a neuron that uses a non-linear activation function [8].

Print ISSN 1691-5402
Online ISSN 2256-070X
http://dx.doi.org/10.17770/etr2019vol2.4120

Filling in missing data with existing relevant methods can continue with the next step. Second, choose the particular method due to it is most effective in predicting missing data observations. Accuracy is different, and it is fluctuating mainly of time-series windowing. Finally, the method applied can fill in missing data for meteorological stations that have missing data, forming indirect observations from VAS "Latvijas valsts celi" data set. Road maintainers can carry out future decisions by merging the observation and forecast data in their forecasting models.

The objective of this paper is to fit in the missing periods of measurement data from road surface maintenance meteorological stations dataset then requirements for the forecasting model set-up.

The paper is structured as follows. Section 2 background, Section 3 forecasting approach. Section 4 concludes.

II. Background

In order to meet the objective is the need for mathematical formulas and methods that are available in the literature. The meteorological stations in the Republic of Latvia are shown in Fig. 1.

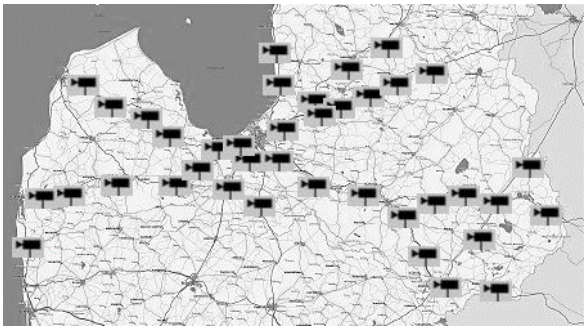


Fig. 1. Stations next to the road surface.

This paper used dataset consists of the destination station, indicated as *M*, and reference station group of all data set denoted as *Y*. The aim of the station is the one whose *x* observation(s), Y_{Mx+1}, Y_{Mx+2}, Y_{Mx+3}, …, Y_{Mx+n}. It is assumed that there is a lack of Y_{Mnone}, and they are needed to be estimated. However, there is a lack of value which is missing and needs to be filled. Thus, the missing value data capture period is artificially created. It is also assumed that the missing observations sequentially succeed one another if there are more than one of the following observations [9]. The reference stations are those that surround the target, with similar traits. Usually, for this purpose, it is used a nearby station, which has a high correlation with the objective[10]. Presuming that *M* reference stations and purpose, *t*, and if the same number of observations and including missing, *m* is recorded at all stations, the study used datasets terminology and the layout shown in Table 1.

TABLE I. The terminology of imputation methods

Denotation	Description
M	destination station
Y	all data set
x+1, x+2, …, x+n	observations
Y_{Mx+1}, …, Y_{Mx+n}	all data set with the destination station in each observation
m_1, m_2, …, m_n	missing observations
t	the target station
$m_k t$	subscript to represent missing observations recorded at the target (k=1, …, m)
$Y_{Mnone} = Y_{Mx+t}$, …, Y_{Mx+n}	missing observations at the target station, t

The following methods described below:

- Normal ratio method;
- k-nearest neighbors algorithm;
- Multilayer perceptron neural network.

A. Normal ratio method

Normal ratio method is used when the normal annual precipitation at any of the index station differs from that of the interpolation station by more than 10%. In this method, the precipitation amounts at the index stations are weighted by the ratios of their normal annual precipitation data in a relationship of the form[6], precipitation with dew point is replaced as follows:

$$P_m = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{N_m}{N_i}P_i\right) \quad (1)$$

- Pm – value at the missing location;
- Pi – value at meteorological index station;
- Nm – average annual observations at "missing data" gauge;
- Ni – average annual observations at gauge;
- n – some gauges.

Using procedure from reciprocal inverse weighting factor approach [11]:

- Divide area around gauge of interest into three parts;
- Using entries at the nearest station on each quadrant;
- Compute the missing value amount, where:

$$P_m = \frac{1}{\sum_{i=1}^{3} 1/X_i}\left(\sum_{i=1}^{3}\frac{P_i}{X_i}\right) \quad (2)$$

P_i – observations recorded by gauge *i*;

X_i – distance from gauge *i* to missing data point.

Normal ratio method is used to be able to compare with other meteorological stations in the reliability of data and their weight; the spreadsheet is used for this method calculations [12].

B. k-nearest neighbors algorithm

The k-nearest neighbors' algorithm (k-NN) is a non-parameter method used for classification and regression. The input is made up of examples of the nearest training in the function space. In the classification phase, *k* is a user-defined constant, and an unlabeled vector (query or test point) is classified by assigning a label that is most often between *k* training samples that are closest to this query point.

Usually, the distance metric used for continuous variables is the Euclidean distance [13]. The algorithm is as follows: Firstly, load the training and test data. Next, choose the value of *K*. Finally, for each point in test data a few steps need to be done: a) find the Euclidean distance to all training data points, then b) store the Euclidean distances in a list and sort it, then c) choose the first *k* points, in the end, d) assign a class to the test point based on the majority of classes [14]. For regression, k-NN predictions are the average of the k-nearest neighbors' outcome by:

$$y = \frac{1}{K}\sum_{i=1}^{k} y_i \quad (3)$$

where is the *i*th case of the examples sample and *y* is the prediction (outcome) of the query point. In contrast to regression, in classification problems, k-NN predictions are based on a voting scheme in which the winner is used to label the query. The distance weighting is formulated:

$$W(x, p_i) = \frac{\exp(-D(x,p_i))}{\sum_{i=1}^{k}(-D(x,p_i))} \quad (4)$$

Set of weights *W*, one for each nearest neighbor, determined by the relative closeness of each neighbor to the query point. Where is the distance between the query point *x* and the *i*th case of the example sample [15].

The application of the corresponding algorithm can replace the missing data from the meteorological station dataset. Since the region in question is a vast territory of the Republic of Latvia with several meteorological stations located at 64,589 km^2 of the total territory of the country [16]. The k-nearest neighbors' algorithm helps to get higher certainty comparison with other methods. The next method, which is viewed is the multilayer perceptron neural network.

C. Multilayer perceptron neural network

Multilayer perceptron (MLP) neural network uses a supervised learning method called backpropagation training. Its multilayer and nonlinear activation distinguish MLP from linear perception. It can differentiate between non-linear data.

Learning occurs in the perception by changing the weights after each of the expected results. Generalization of the least mean squares algorithm in the linear perception represent the error in output node *j* in the *n*th data point by [18]:

$$e_j(n) = d_j(n) \quad y_j(n) \quad (5)$$

where *d* is the target value and *y* are the value produced by the perception. The weight of the node is adjusted based on the corrections that reduce the error across the output that is indicated:

$$(n) = \frac{1}{2}\sum_j e_j^2(n) \quad (6)$$

With gradient descent, each weight change is:

$$\Delta w_{ji}(n) = \eta \frac{\partial \varepsilon(n)}{\partial v_j(n)} y_i(n) \quad (7)$$

where *y* is the output of the previous neuron and *n* is the training speed chosen to ensure that the scales quickly converge with the response, without fluctuations. To change the weight of a hidden layer, the weight of the output layer changes according to the activation function derivative and this algorithm denotes the activation function backpropagation [17]. Using the training can control the reliability of the data and provide for partial predictions data credibility.

The next chapter outlines the approach to the forecasting for the VAS "Latvijas valsts celi" dataset.

III. Forecasting approach

Initially, to understand the distance between meteorological stations, they can be combined in a single location, represented in Table 2.

TABLE II. The distance of the nearest station

Station name	Location (°)	Distance (km, name)
Adazi	57.0903, 24.3133	10.98 (Garkalne)
Admini	56.5931, 25.5464	36.77 (Livani)
Annenieki	56.6671, 23.0608	36.71 (Kalnciems)
Apvedcels	56.8436, 24.0192	12.84 (Dalbe)
Dalbe	56.7511, 23.8937	12.84 (Apvedcels)
Daugavpils	56.7511, 26.6438	68.90 (Kraslava)
Durbe	56.5752, 21.2939	37.10 (Rudbarzi)
Garkalne	57.0688, 24.4900	10.98 (Adazi)
Inciems	57.2741, 24.8949	12.53 (Sigulda)
Kaibala	56.6608, 24.8981	36.31 (Saulkalne)
Kalnciems	56.8360, 23.5761	10.53 (Laci)
Kraslava	55.8976, 27.2933	68.90 (Daugavpils)
Laci	56.8945, 23.7118	7.07 (Sloka)
Livani	56.4365, 26.0723	35.54 (Nicgale)
Ludza	56.3850, 28.0688	50.49 (Rezekne)
Melturi	57.2212, 25.2308	16.90 (Sigulda)
Nica	56.2145, 21.1329	41.38 (Durbe)
Nicgale	56.1551, 26.3433	35.54 (Livani)
Rezekne	56.5401, 27.2991	50.49 (Ludza)
Rudbarzi	56.6452, 21.8846	25.02 (Saldus)
Saldus	56.6710, 22.2898	25.02 (Rudbarzi)

Station name	Location (°)	Distance (km, name)
Saulkalne	56.8465, 24.4101	23.86 (Apvedcels)
Sigulda	57.1689, 24.9683	12.53 (Inciems)
Sloka	56.9295, 23.6150	7.07 (Laci)
Smiltene	57.3881, 25.9739	40.91 (Valmiera)
Stalbe	57.3781, 25.0512	14.93 (Inciems)
Strenci	57.6512, 25.8638	36.08 (Valmiera)
Talsi	57.1522, 22.6902	58.55 (Annenieki)
Valmiera	57.5035, 25.3271	21.67 (Stalbe)
Vircava	56.6401, 23.7763	14.30 (Dalbe)

The average value between the distances nearest to all distances between meteorological stations is 29.09 km. As well as the fact that standard deviation is 17.95 km not particularly surprising because a large proportion of meteorological stations do not work. Marking the smallest and largest distance of 68.90 km largest and the smallest is 7.07 km. As already acknowledged to predict from the meteorological station, the higher the distance, the more inaccurate the forecast, while splitting distances between meteorological stations, may reduce the RMSE when forecasting. In other words, windowing allows to take meteorological station time-series data and transform it into a cross-sectional format. When essentially convert time-series values into cross-sectional attributes, therefore, apply a predictive modeling algorithm to predict future values.

For experiment purposes, observations are taken between three meteorological stations that are close to each other, compared to all meteorological stations. One of the three meteorological stations will be adopted as a source of missing values, which will allow the use of the methods mentioned above to obtain missing values, or their forecasts, between two existing meteorological stations. In other words, methods will be used and then compared results with actual observations. In this way, they are getting the RMSE for each method. The interval used is 5 minutes, which consists predominantly of data entering between stations at 5 minutes interval [19]. The selected meteorological stations' names and coordinates are "Inciems" (57.2741, 24.8949), "Stalbe" (57.3781, 25.0512) and "Sigulda" (57.1689, 24.9683) located average 13.33 km from each other.

Of these meteorological stations, 392 observations are taken average 130 observation per meteorological station for a given parameter that is a dew point (C°) on winter month 15th of January 2018. The missing data meteorological station "Stalba" with 131 observations for this period is selected; part of the data is represented in Table 3.

TABLE III. "Inciems", "Stalbe" and "Sigulda" partly observation of dew point (C°) on 15th of January 2018

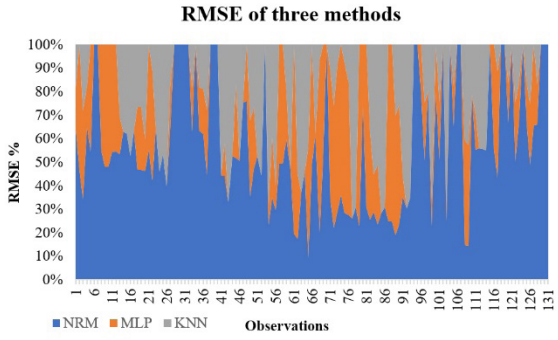
Station name	Dew point (C°)	Date time
…	…	…
Sigulda	-12.1	15.01.2018 11:41
Inciems	-10.6	15.01.2018 11:41
Stalbe	-9.9	15.01.2018 11:51
Inciems	-10.7	15.01.2018 11:52
Sigulda	-12.3	15.01.2018 11:53
Stalbe	-10.0	15.01.2018 12:02
…	…	…

The missing data is calculated using the following methods: Normal ratio method using spreadsheet for calculation, on the other hand, k-nearest neighbors algorithm is made up of examples of the nearest training in the function space using Orange Visual Programming tool (version 3.18.0) for prediction and RMSE estimation and next method MLP neural network is used with parameters neurons in hidden layers 100 then activation function for the hidden layer called "ReLu" the rectified linear unit function and finally solver called "SGD" stochastic gradient descent[20]. RMSE for each method is calculated the results are represented in Table 4.

TABLE IV. Accuracy for each method

Method	RMSE
Normal ratio method	0.35
k-nearest neighbors algorithm	0.23
MLP neural network	0.21

For Normal ratio method, RMSE is 35%, on the other hand, the k-nearest neighbors' algorithm is 23%, and finally, the MLP neural network is 21% accurate. Between k-NN and MLP is only 2% difference in accuracy. The results are also displayed graphically in order to be more readily perceived. Fig. 2 represents all three methods for the period 15th of January 2018.



15th of January 2018 RMSE for three methods of missing data prediction (dew point).

As shown in the results, the MLP neural network method has proved to be best compared to both methods.

IV. Conclusions

Meteorological stations will be required for methods that will be able to predict missing data. One of the most important reasons is sensor failures and potential interference. As well as a long distance between meteorological stations. Which makes long distances and makes significant errors in calculations, and forecasts with a high probability of error. There are a few methods to tackle this problem. Each time-series has its best or most efficient method, which can be used in practice. Before making any future forecasts, it is necessary to verify the use of the best method.

One of the well-known methods is multilayer perceptron neural network which was able to prove between the three methods as the best — presenting good results with 21% accuracy, which is relatively good but not enough. In the future, there is a need for better methods to show accuracy at the 10% mark. The road surface maintenance work in the winter months is compelled to minimize the response time on arrival at the specified road stage. The method used in the literature is the average value calculation.

Thus, the paper identifies further research direction on the generalized hybrid method on real-time forecasting for time-series. Compared to existing results and using a repeating method that can process real-time data and be able to adapt to the current situation, using modern solutions with programming language capabilities.

References

[1] Yozgatligil, C., Aslan, S., Iyigun, C. and Batmaz, I., 2013. Comparison of missing value imputation methods in time series: the case of Turkish meteorological data. *Theoretical and applied climatology*, 112(1-2), 143-167, https://doi.org/10.1007/s00704-012-0723-x

[2] Jeffrey, S.J., Carter, J.O., Moodie, K.B. and Beswick, A.R., 2001. Using spatial interpolation to construct a comprehensive archive of Australian climate data. *Environmental Modelling & Software*, 16(4), 309-330.

[3] Spadavecchia, L. and Williams, M., 2009. Can spatio-temporal geostatistical methods improve high resolution regionalization of meteorological variables?. *Agricultural and Forest Meteorology*, 149(6-7), 1105-1117.

[4] Dibike, Y.B. and Coulibaly, P., 2005. Hydrologic impact of climate change in the Saguenay watershed: comparison of downscaling methods and hydrologic models. *Journal of hydrology*, 307(1-4), 145-163, https://doi.org/10.1016/j.jhydrol.2004.10.012

[5] Taylor, J.W., De Menezes, L.M. and McSharry, P.E., 2006. A comparison of univariate methods for forecasting electricity demand up to a day ahead. *International Journal of Forecasting*, 22(1), 1-16, https://doi.org/10.1016/j.ijforecast.2005.06.006

[6] Subramanya, K., 2013. Engineering Hydrology, 4e. Tata McGraw-Hill Education.

[7] Dudani, S.A., 1976. The distance-weighted k-nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics*, (4), 325-327.

[8] Haykin, S.S., 2009. Neural networks and learning machines/Simon Haykin. New York: Prentice Hall.

[9] Giannone, D., Reichlin, L. and Small, D., 2008. Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, 55(4), 665-676.

[10] Bagheri, M., Nurmanova, V., Abedinia, O., Naderi, M.S., Naderi, M.S. and Ghadimi, N., 2018, June. A Novel Wind Power Forecasting Based Feature Selection and Hybrid Forecast Engine Bundled with Honey Bee Mating Optimization. In *2018 IEEE International Conference on Environment and Electrical Engineering (EEEIC/I&CPS Europe)*, 1-6.

[11] Sharp, J.J. and Sawden, P.G., 2013. BASIC hydrology. Elsevier.

[12] Cuypers, W., Van Gestel, N., Voet, A., Kruth, J.P., Mingneau, J. and Bleys, P., 2009. Optical measurement techniques for mobile and large-scale dimensional metrology. *Optics and Lasers in Engineering*, 47(3-4), 292-300, https://doi.org/10.1016/j.optlaseng.2008.03.013

[13] Fabbri, R., Costa, L.D.F., Torelli, J.C. and Bruno, O.M., 2008. 2D Euclidean distance transform algorithms: A comparative survey. *ACM Computing Surveys* (CSUR), 40(1), 2.

[14] Acuna, E. and Rodriguez, C., 2004. The treatment of missing values and its effect on classifier accuracy. In *Classification, clustering, and data mining applications*. Springer, Berlin, Heidelberg, 639-647, https://doi.org/10.1007/978-3-642-17103-1_60

[15] Barrientos, R.J., Gómez, J.I., Tenllado, C., Matias, M.P. and Marin, M., 2011, August. kNN query processing in metric spaces using GPUs. In *European Conference on Parallel Processing*. Springer, Berlin, Heidelberg, 380-392, https://doi.org/10.1007/978-3-642-23400-2_35

[16] Gylfason, T. and Hochreiter, E., 2011. Growing Together: Croatia and Latvia. *Comparative Economic Studies*, 53(2), 165-197.

[17] Zhang, Z., 2018. Artificial neural network. In *Multivariate Time Series Analysis in Climate and Environmental Research*, 1-35, https://doi.org/10.1007/978-3-319-67340-0_1

[18] Murata, N., Yoshizawa, S. and Amari, S.I., 1994. Network information criterion-determining the number of hidden units for an artificial neural network model. *IEEE Transactions on Neural Networks*, 5(6), 865-872, https://doi.org/10.1109/72.329683

[19] Park, T. and Lee, S., 2004. A Bayesian approach for estimating link travel time on urban arterial road network. In *International Conference on Computational Science and Its Applications*, 1017-1025, https://doi.org/10.1007/978-3-540-24707-4_114

[20] Chaudhari, P. and Soatto, S., 2018. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In *2018 Information Theory and Applications Workshop (ITA)*, 1-10