Desalination and Water Treatment
www.deswater.com

doi: 10.1080/19443994.2014.1002282

57 (2016) 964–970
January

Taylor & Francis
Taylor & Francis Group

The use of interpolation methods for the modelling of environmental data

Rafał Jasiński

Faculty of Environmental Engineering and Biotechnology, Czestochowa University of Technology, Dabrowskiego Str. 73, Czestochowa 42-200, Poland, email: raphael@is.pcz.pl

Received 22 October 2014; Accepted 3 December 2014

ABSTRACT

The purpose of the study was to provide some practical examples of using interpolation methods for modelling one-dimensional environmental data. Based on the measurement data acquired from the automatic air monitoring station in Warsaw (Poland), simulation of different variants of interpolation of the 1 h average values of air temperature and SO_2 concentration was performed. The interpolation was done by three methods, namely, by a linear method with a cubic polynomial and with a cubic spline. The simulation of supplementing the missing values was conducted for test vectors with a varying measurement gap length and the variable margin of extreme adjacent values. Comparison of the obtained modelling results with previously removed actual data was also made. For the assessment of the modelling error, the mean absolute error, the root mean squared error and Willmott's Index of Agreement were used. High accuracy of modelling was obtained, both for the short and the longer test vector variants, whereas the best environmental data modelling accuracy was obtained in short time intervals. It has been found that the use of interpolation for modelling a given type of environmental data should be preceded by the assessment of the accuracy of the employed methodology. In the framework of this study, a function for carrying out simulation of the accuracy of environmental data modelling was prepared within the MATLAB software program.

Keywords: Modelling; Interpolation; Linear; Cubic; Spline; MATLAB; Environmental; Data

1. Introduction

Environmental data are often characterized by an incompleteness of measurement series. A common practice to deal with the missing values is to omit them while analysing the results [1,2]. However, removing entire data records from the analysis of results due to the absence of a few or even a single value might result in failing to utilize a considerable amount of information, the acquisition of which, would be very costly. The analysis of an incomplete data-set might also affect the correctness of formulated conclusions [3]. In addition, some of the data analysis methods may only be used for complete data [4–6]. Therefore, some environmental data analysts use modelling methods to substitute the missing values with alternative ones, in accordance with presumed modelling accuracy criteria [7–10].

Presented at the 12th Scientific Conference on Microcontaminants in Human Environment
25–27 September 2014, Czestochowa, Poland

One of the simplest, though fairly accurate, methods for substituting missing data with modelling data is interpolation [8–11]. Interpolation is an analytical technique consisting in searching for intermediate points amongst the existing ones, provided that the measurement data should form a logical sequence of numbers, e.g. a time series [11].

In order to be able to use interpolation methods for modelling environmental data, one must know the extreme values. In the case of the linear interpolation, the graphic interpretation of modelling is a straight line connecting extreme values. In the case of the interpolation with a cubic polynomial or a cubic spline, the graphic interpretation represents curves connecting the points of known values. However, for non-linear methods, modelling results are dependent on both the employed interpolation method and on successive extreme adjacent values. This study undertakes to examine how the results of modelling by interpolation methods are influenced by the type of the employed interpolation method, the measurement gap length and the extreme adjacent value margin. Observed environmental data were used, and a computation algorithm was prepared in the MATLAB computer program in order to carry out simulation of the accuracy of modelling by interpolation methods.

On automatic air monitoring stations, the concentration levels of selected air pollutants and the meteorological parameters are determined in a continuous manner. These data are averaged to 1 h mean values and then stored in extensive databases [12,13]. Environmental data of this type form times series with a constant 1 h time interval. Therefore, it is justifiable to use automatic air monitoring databases for the validation of linear and non-linear interpolation methods, so that the missing data with modelling values could be substituted.

2. Materials and methods

Environmental data coming from the Warszawa-Targówek air monitoring station in Warsaw (Poland) were used for the investigation. The air monitoring station under examination is an urban background station belonging to the Provincial Environmental Protection Inspectorate in Warsaw. The air pollutant concentrations values determined in this station are used by EuroAirnet—the European Air Quality monitoring network. Simulation of different variants of using environmental data interpolation methods was done using the 1 h mean values of air temperature—as a meteorological parameter—and SO_2 sulphur dioxide concentration—as one of the air quality-defining parameters. The measurement data were derived from three years' period, i.e. 2009–2011.

The computation methodology involved the removal of some data fragments from the input data in such a manner as to obtain apparent (fictitious) gaps in observed data. Then, those removed values were modelled by interpolation methods. The simulation was conducted in different variants, which were arranged into test vectors moved sequentially along the data series. Test vectors with missing measurement value intervals of 1, 3, 6, 9 and 12, along with a 1-, 2- and the 3-element extreme adjacent value margin were adopted (Table 1). The test vectors were moved sequentially by 23 values, so that the simulated missing values would occur in different times of the day. If at least one missing value occurred amongst the observed ones in a given sequence, then that step was omitted. As a result of the procedure described above, a list of all removed fragments of observed values and their corresponding simulated values was obtained. Thus, so prepared set of observed and simulated values was subjected to modelling error assessment. For the modelling error assessment, the following three statistical parameters were used:

(1) mean absolute error (MAE),
(2) root mean-squared error (RMSE),
(3) Willmott's Index of Agreement (*d*).

The Index of Agreement (*d*) developed by Willmott as a standardized measure of the degree of model prediction error varies between zero and one. The value one indicates a perfect match, whereas zero indicates no agreement at all [14,15].

The modelling of the missing values was conducted by the following three methods: 1—linear method, 2—cubic polynomial and 3—cubic spline. The

Table 1
Test vectors adopted for simulation of environmental data modelling: 0—extreme adjacent value margin, 1—substituted value

[010]	[01110]	[01111110]	[01111111110]	[01111111111110]
[00100]	[0011100]	[0011111100]	[011111111100]	[0011111111111100]
[0001000]	[000111000]	[000111111000]	[000111111111000]	[000111111111111000]

Table 2
Statistical parameters defining the accuracy of modelling of air temperature and SO_2 concentration, depending on the adopted test vector variants (Warszawa-Targówek 2009–2011)

		Temperature				SO_2 concentration			
Test vector variants	Methods of interpolation	No. of sequences	MAE (℃)	RMSE (℃)	*d*	No. of sequences	MAE (μg/m^3)	RMSE (μg/m^3)	*d*
[010]	Linear	1,094	0.22	0.33	0.9997	1,066	1.33	3.84	0.9510
	Cubic		–	–	–		–	–	–
	Spline		–	–	–		–	–	–
[00100]	Linear	1,091	0.19	0.30	0.9998	1,051	1.14	2.99	0.9686
	Cubic		0.19	0.30	0.9998		1.13	3.05	0.9675
	Spline		0.19	0.30	0.9998		1.22	3.28	0.9636
[0001000]	Linear	1,090	0.19	0.30	0.9998	1,039	1.18	2.94	0.9666
	Cubic		0.19	0.30	0.9998		1.14	2.86	0.9688
	Spline		0.19	0.30	0.9998		1.21	3.17	0.9626
[01110]	Linear	1,091	0.43	0.63	0.9989	1,051	1.69	4.04	0.9401
	Cubic		–	–	–		–	–	–
	Spline		–	–	–		–	–	–
[0011100]	Linear	1,090	0.42	0.62	0.9990	1,039	1.73	4.21	0.9341
	Cubic		0.35	0.53	0.9993		1.69	4.33	0.9316
	Spline		0.33	0.50	0.9993		2.04	5.51	0.8996
[000111000]	Linear	1,085	0.47	0.70	0.9987	1,022	1.71	4.17	0.9335
	Cubic		0.38	0.58	0.9991		1.65	4.17	0.9348
	Spline		0.35	0.57	0.9991		2.02	5.09	0.9100
[01111110]	Linear	1,087	0.85	1.22	0.9959	1,032	2.20	4.94	0.9074
	Cubic		–	–	–		–	–	–
	Spline		–	–	–		–	–	–
[0011111100]	Linear	1,083	0.85	1.22	0.9959	1,013	2.16	4.69	0.9201
	Cubic		0.69	1.01	0.9972		2.12	4.77	0.9198
	Spline		0.60	0.90	0.9978		2.90	6.88	0.8613
[000111111000]	Linear	1,079	0.87	1.25	0.9957	998	2.34	5.34	0.8957
	Cubic		0.69	1.03	0.9971		2.26	5.34	0.8965
	Spline		0.61	0.92	0.9977		3.19	7.78	0.8170
[01111111110]	Linear	1,082	1.29	1.86	0.9904	1,010	2.66	5.91	0.8754
	Cubic		–	–	–		–	–	–
	Spline		–	–	–		–	–	–
[0011111111100]	Linear	1,078	1.29	1.86	0.9904	988	2.75	6.24	0.8568
	Cubic		1.07	1.59	0.9930		2.69	6.28	0.8577
	Spline		0.90	1.34	0.9952		4.10	9.69	0.7519
[000111111111000]	Linear	1,072	1.29	1.85	0.9905	974	2.76	6.42	0.8428
	Cubic		1.06	1.57	0.9932		2.71	6.47	0.8424
	Spline		0.92	1.37	0.9949		4.38	10.19	0.7110
[01111111111110]	Linear	1,074	1.74	2.50	0.9825	978	2.86	6.33	0.8481
	Cubic		–	–	–		–	–	–
	Spline		–	–	–		–	–	–
[0011111111111100]	Linear	1,069	1.73	2.49	0.9826	965	2.97	6.59	0.8352
	Cubic		1.50	2.20	0.9865		2.94	6.72	0.8321
	Spline		1.22	1.78	0.9915		4.93	11.64	0.6650
[000111111111111000]	Linear	1,060	1.77	2.53	0.9821	953	2.96	6.55	0.8328
	Cubic		1.54	2.25	0.9859		2.87	6.54	0.8337
	Spline		1.20	1.78	0.9915		5.05	12.68	0.6120

Note: MAE—mean absolute error; RMSE—root mean squared error; D—Willmott's Index of Agreement.

function *valid_int.m* was prepared in the MATLAB computer program (Appendix 1) in order to carry out the simulation of the accuracy of environmental data modelling by interpolation methods within this study. The developed function makes it possible to adopt arbitrary test vector length values as well as to select the margin of extreme adjacent values and the length of the test vector substitution sequence. The function *will.m* for the computation of the Willmott's Index of Agreement was also prepared (Appendix 2).

3. Results

Table 2 provides the statistical parameters defining the modelling accuracy, depending on the adopted test vectors. The number of sequences defines the number of repetitions of observed value removal in particular test vector variants. For test vectors containing single extreme adjacent values, neither the cubic polynomial nor cubic spline methods were used in modelling since the results could be reduced to the linear method.

As a result of the performed computations, high modelling quality was achieved, both for the short and long test vector variants. For single-element measurement gaps, Willmott's index values close to the unit were obtained. It means that almost complete agreement between the simulation values and the observed ones was achieved. With an increase in the length of the measurement gaps tested, the modelling accuracy decreased, while for air temperature, the lowest Willmott's index value was as high as 0.9821. The interpolation accuracy was higher for air temperature than for SO_2 concentration. For modelling the air temperature values, slightly better results were achieved using non-linear methods in particular test vector variants, with the best results being obtained for interpolation with the cubic spline. In turn, for SO_2 concentration modelling, better results were achieved using the linear interpolation method. Increasing the extreme adjacent value margin length had no effect in the case of modelling with linear interpolation. On the other hand, when using non-linear methods for air temperature modelling, the modelling accuracy increased with increasing extreme adjacent value margin length. An opposite relationship was observed for modelling longer test gaps for SO_2 concentrations. With an increase in the extreme adjacent value margin length, the modelling accuracy distinctly decreased, especially in the case of using interpolation with the cubic spline.

4. Conclusions

Environmental data used in this study, such as the average 1 h values of air temperature (a meteorological parameter) and SO_2 concentrations (one of the air quality-defining parameters), usually do not undergo any rapid fluctuations in the successive hours of the day. This feature favours the accuracy of modelling data using the linear and non-linear interpolation methods described above. The obtained good modelling results presented in this study give a good reason to conclude that environmental data of this type can be successfully modelled by interpolation methods. Knowing the extreme values of measurements gaps in one-dimension environmental data, it is possible do substitute missing values with interpolated values at least up to 12 missing measurement values with high accuracy. However, in order to choose the best interpolation method for a given environmental data type, making up for missing data should be preceded by the assessment of the accuracy of the methodology used. Within this study, a simple algorithm for the validation of the accuracy of modelling environmental data has been developed, whereby one-dimensional interpolation is done by the "line", "cubic" and "spline" methods for different lengths of measurement gaps and the extreme adjacent value margin. A finished procedure for assessing the quality of modelling of one-dimensional environmental data by interpolation methods within the MATLAB software program is provided in Appendix 1.

As a result of the performed simulation using interpolation methods for modelling environmental data, the following conclusions have been drawn:

(1) By using interpolation methods for modelling environmental data, very high modelling quality can be achieved, while the modelling accuracy decreases with an increase in the length of the modelled value interval.

(2) For short time intervals, the use of the adjacent value margin is of little significance for the accuracy of modelling. For longer time intervals, increasing the adjacent value margin may result in either an increase or a decrease in the modelling accuracy, depending on the nature of environmental data.

(3) Using interpolation methods in order to make up for missing values should be preceded by the analysis of modelling accuracy aimed at selecting the best methodology for a given type of environmental data.

Acknowledgement

This work was supported by Czestochowa University of Technology project no. BS-402-301/11.

References

[1] M.P. Gómez-Carracedoa, J.M. Andradea, P. López-Mahíaa, S. Muniateguib, D. Pradab, A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets, Chemom. Intell. Lab. Syst. 134 (2014) 23–33.

[2] K. Muteki, J.F. MacGregor, T. Ueda, Estimation of missing data using latent variable methods with auxiliary information, Chemom. Intell. Lab. Syst. 78 (2005) 41–50.

[3] J. Han, J. Pei, Y. Yin, R. Mao, Mining frequent patterns without candidate generation: A frequent-pattern tree approach, Data Min. Knowl. Discov. 8 (1) (2004) 53–87.

[4] A. Smolinski, B. Walczak, J.W. Einax, Exploratory analysis of datasets with missing elements and outliers, Chemosphere 49 (2002) 233–245.

[5] I. Stanimirova, M. Daszykowski, B. Walczak, Dealing with missing values and outliers in principal component analysis, Talanta 72 (2007) 172–178.

[6] S. Serneels, T. Verdonck, Principal component analysis for data containing outliers and missing elements, Comput. Stat. Data Anal. 52 (2008) 1712–1727.

[7] S. Hoffman, R. Jasiński, Classification of air monitoring data gaps, Pol. J. Environ. Stud. 18(2B) (2009) 177–18.

[8] H. Junninen, H. Niska, K. Tuppurainen, J. Ruuskanen, M. Kolehmainen, Methods for imputation of missing values in air quality data sets, Atmos. Environ. 38 (2004) 2895–2907.

[9] A. Plaia, A.L. Bondi, Single imputation method of missing values in environmental pollution data sets, Atmos. Environ. 40 (2006) 7316–7330.

[10] S. Hoffman, R. Jasiński, Completing missing data in air monitoring stations using diurnal courses of regional pollution concentrations, Arch. Environ. Prot. 34 (3) (2008) 133–142.

[11] J. Joseph, H.O. Sharif, T. Sunil, H. Alamgir, Application of validation data for assessing spatial interpolation methods for 8-h ozone or other sparsely monitored constituents, Environ. Pollut. 178 (2013) 411–418.

[12] R. Jasiński, Multidimensional analysis of daily variations in air pollutants and meteorological parameters derived from the upper Silesian urban area, Pol. J. Environ. Stud. 20(4A) (2011) 104–109.

[13] R. Jasiński, Directions of air pollution inflows as a method for evaluation of representativeness of automatic air monitoring stations area, Environ. Prot. Eng. 38(2) (2012) 99–108.

[14] C.J. Willmott, On the validation of models, Phys. Geogr. 2 (1981) 184–194.

[15] C.J. Willmott, S.M. Robeson and K. Matsuura, Short communication A refined index of model performance, Int. J. Climatol. 32 (2012) 2088–2094.

Appendix 1

Function *valid_int.m* by MATLAB to be used for the assessment of the accuracy of modelling of one-dimensional environmental data (single variable) by interpolation methods

```
function [uxy]=valid_int(u,v,k);
% The assessment of the accuracy of modelling of one-dimensional environmental data
     by interpolation methods;
% Caution! Missing data must be substituted with the designation NaN;
% u - observed values (single variable);
% v - test vector, e.g. [0 1 1 1 0], where 0 – extreme adjacent value margin, 1 –
     substituted value;
% k – test set step length, np. 23;
% uxy – the list of all removed fragments of observed values (u1x) and their
     corresponding simulated values (u1y);

vl=length(v);    % vl - single test vector length, e.g. for [0 1 1 1 0] v=5;
z=1:length(u);   % z - position vector generation;
x=[];            % x - test interval positions (excluding NaN containing intervals);
for i=1:floor(length(u)/k)
    if sum(isnan(u(z(1:vl)+k*(i-1))))==0
        x=[x; z(1:vl)+k*(i-1)];
    end
end

ux=u(x); % ux - test intervals with dimension v1 acquired from data u every step w

p=find(v==1); % p - position of units in vector v
up=ux(:,p); % up - test values in the form of a table (defined by positions p, e.g.
              p=[2 3]);
u1x=up';
u1x=u1x(:);  % u1p - up in a vector form;

ux(:,p)=NaN; % ux - data test intervals u with dimension v1 with test values
              substituted with NaN;
uxn=ux';
uxn=uxn(:); % uxn - ux in a vector form;

xn=x(:,p)'; % xn - positions of substituted NaN in vector u;
xn=xn(:);

vn=findstr(isnan(uxn)',v); % vn - vector with the initial positions of places found
                           corresponding to v sought for;
n=length(vn);
for i=1:n
    v1=find(v==1);
    x1=vn(i)+v1-1;
    v0=find(v==0);
    x=vn(i)+v0-1;
    y=uxn(x);
    y1=interp1(x,y',x1,'line'); % y1 - interpolated(missing) values (possible
                     interpolation methods: 'line','cubic','spline');
    uxn(x1)=y1;
end
str=[ ended interpolation of ',num2str(n),' measurement gaps type v
',' [',num2str(v),']'];
disp(str)
u1=reshape(uxn,vl,n)';
u1y=u1(:,p)';
u1y=u1y(:); % u1y - simulated values;
uxy=[u1x, u1y];

MAE=mae(u1x-u1y);
MSE=mse(u1x-u1y);
RMSE=sqrt(MSE);
WILL=will(u1y,u1x);

['MAE','RMSE','WILL']
[MAE,RMSE,WILL]
```

Appendix 2

Function *will.m* by MATLAB to be used for the calculation of Willmott’s Index of Agreement

```
function [d]=will(X,Y);
% Function to be used for the calculation of Willmott’s Index of Agreement (C.J.
Willmott, On the validation of models, Physic. Geogr., 1981, 2, 184-194);
% X - matrix or data frame with observed values (obs);
% Y - matrix or data frame with simulated values (sim);
% Caution! Size X and Y must be the same. Missing data can not occur;
% d = 1 - [ ( sum( (obs - sim)^2 ) ] / sum( ( abs(sim - mean(obs)) + abs(obs -
mean(obs)) )^2 );

[n,m]=size(X);

li=(X-Y).^2;
sy=nanmean(Y);

for i=1:m
    a=abs(X(:,i)-sy(i));
    b=abs(Y(:,i)-sy(i));
    mi(i)=sum((a+b).^2);
end

d=1-(sum(li)./mi);
```