Filling missing meteorological data with Computational Intelligence methods

Joanna Kajewska-Szkudlarek* and *Justyna* Stańczyk

Wroclaw University of Environmental and Life Sciences, Institute of Environmental Engineering, Grunwaldzki Square 24, 50-363 Wroclaw, Poland

Abstract. Estimates of temperature and humidity values at a specific time of day, from hourly to monthly profiles, are needed for a number of environmental, ecological, agricultural and technical applications, ranging from natural hazards assessments, crop growth forecasting to designing solar energy systems. In climatology, they constitute the basis for drawing conclusions about climate variability. Data used in such analyses should be complete and reliable. Therefore, effective methods for filling missing values are sought. The initial scope of this research is to investigate the efficiency of computational intelligence methods in filling missing daily temperature and humidity parameters values. For this reason, a number of experiments have been conducted with Artificial Neural Networks and Support Vector Regression using meteorological data from the city of Wroclaw in Poland. The performance of these methods has been evaluated using standard statistical indicators, such as Correlation Coefficient and Root Mean Squared Error. Finally, certain computational intelligence techniques are proposed that can be used to predict daily temperature and humidity values more accurately in order to fill the missing data.

1 Introduction

The automation of meteorological stations and the use of electronic sensors result in collecting vast databases, whose potential is often not used properly, as their analysis is strenuous and time-consuming. Such data are still more commonly analysed with use of various Data Mining methods, including Computational Intelligence methods, such as Artificial Neural Networks, Support Vector Machines or Evolutionary Algorithms. Various fields of science are witnessing the development of prediction methods. In meteorology and climatology, such methods are used not only to develop weather forecasts, but also to estimate the value of parameters that are difficult to measure (such as evapotranspiration) or to generate predictions of future values of meteorological elements based on the possessed data. One of their advantages is the fact that they do not require knowledge about correlations between data or about the existence of such correlations. In general, such methods use input data to recreate the most often recurring patterns. They are able to learn on examples of entered data, to generalise certain phenomena and to find links between them.

Homogeneous and complete time series of basic meteorological parameters are used in numerous sectors of industry and fields of science. They also constitute the foundation for climatologic research. As they are the basis for drawing conclusions on climate changeability, the measurement series used in climatology should be reliable. It happens sometimes that long-term data series obtained from meteorological stations and measurement sites do not meet the criterion of homogeneity, due to changes in the location of the site, in measuring equipment, measurement times or the environment of the station. Such changes always affect data quality [15, 16].

The introduction of automated meteorological measurements has doubtlessly facilitated the process of collecting and processing meteorological data. However, in comparison to traditional stations, where the observer supervises the measurements personally, measurement series obtained from AWS (Automatic Weather Stations) much more often contain gaps resulting, for example, from breaks in power supply. It is quite often necessary to supplement them with data from other measurement sites, which cannot be substituted freely, as they do not meet the criterion of comparability of measurement sites, or, often, of time and measurement equipment.

The homogeneity of meteorological measurement series is evaluated with use of several tests, including the Mann–Kendall, Wald–Wolfowitz runs, Von Neumann ratio, Pettitt, Buish and range test, and Standard Normal Homogeneity Test (SNHT). One of the most commonly used tests is the SNHT, which was first applied by Alexandersson (1986) to detect changes in precipitation series [2, 10]. It was used mainly for this purpose [5, 11, 13], but it may also be successfully applied to determine the homogeneity of series of other parameters, such as surface solar radiation [3, 12], air temperature [14] and wind speed [17].

Missing data in series are supplemented in a determinist or stochastic way or with use of Computational

* Corresponding author: joanna.kajewska-szkudlarek@upwr.edu.pl

Intelligence methods [4]. The most commonly used methods among the latter include Artificial Neural Networks and Support Vector Machines [8, 9]. However, the methods of filling in missing data are selected discretionally, due to the fact that none of the methods used has been generally accepted or recommended by the World Meteorological Organisation [1]. As a result, it is necessary to conduct further research in this area, with the aim to improve the methods of predicting basic meteorological parameters basing on data from various measurement sites, recorded by different types of equipment.

1.1 The aim of the study

The aim of the study was to assess the possibility to use selected Computational Intelligence methods (Artificial Neural Networks, Support Vector Regression) for the homogenisation and completion of incomplete data series obtained from meteorological measurements.
The authors attempted to create predictive models for 3.5 year time series of daily values of temperature parameters (maximum *tmax*, minimum *tmin*, and average temperature *tm*) and humidity parameters (relative air humidity *f*, saturation deficit *d*) of the air.
The extent to which daily air temperature and humidity values obtained from one measurement site may be used to predict the value of these parameters at a different location was assessed.

2 Material and methods

Data were obtained from measurements taken in the area of Wrocław in the period from the 1st of June 2014 to the 31st of December 2017. Artificial Neural Networks and Support Vector Regression were used to attempt to create a model of prediction of the series of analysed parameters measured by the Agro- and Hydrometeorological Observatory of the Wrocław University of Environmental and Life Sciences in Swojec (output data). For this purpose, daily values of air temperature and relative humidity were used along with influence parameters, including atmospheric pressure, wind speed as well as the day of the year and length of the day (input data). The meteorological data used for forecasting originated from the automated measurement station belonging to the monitoring system of the Municipal Water and Sewer Company MPWiK, located in the city centre, at Na Grobli Street.
Additionally, input and output data were obtained with use of different measurement methods. Measurements in the Wrocław-Swojec Observatory were conducted in the standard way. Daily *tm, f* and *d* values were calculated basing on measurements with use of classical equipment, i.e. liquid thermometers, while *tmax* and *tmin* were obtained from extreme thermometers, read once a day.
The automated meteorological station (AWS model 1302) equipped with a set of sensors measuring the basic elements of weather was located at Na Grobli Street and programmed for sampling at one-minute intervals that constituted the basis for preparing hourly and daily reports.

2.1. Computational Intelligence models

The methodology of tests consisted in using modern Computational Intelligence algorithms to predict the daily values of air temperature and humidity parameters. It was conducted with use of two most commonly used types of Artificial Neural Networks: MLP (Multi-Layer Perceptron) and RBF (Radial Basis Function), which differ in terms of modelling the input-output relation and SVR (Support Vector Regression). In order to eliminate the excessive adjustment phenomenon, data were divided into two sub-sets: learning and testing set. The learning set covered the period from June 1st, 2014 to February 6th, 2017 and it included 75% of all values. The remaining 25% instances constituted the test sub-set.
The networks selected for further analyses and comparison with the SVR model from among a series of created Artificial Neural Networks were those of MLP 6-6-5 and RBF 6-28-5 architecture, where the number of neurons (6) in the input layer represents the variables used for modelling, and the output layer neurons (5) represent the predicted variables.
For MLP, the BFGS (Broyden-Fletcher-Goldfarb-Shanno) learning algorithm was used, with an SOS (Sum of Squares) error function and linear neuron activation function in the hidden layer and tanh in the output layer. For RBF learning, the RBFT (Reduced Breadth-First Search) algorithm was used, with the SOS error function, Gauss function in the hidden layer and linear function in the output layer.
The SOS error function was used to assess the current quality of MLP and RBF networks in the learning process. The error function measures the consistency of network prediction with the defined value and it is used to determine the size of necessary neuron weight changes in each iteration. The learning process strives to minimise the size of SOS error:

$$E_{sos} = \sum_{i=1}^{N} (y_i - t_i)^2 \qquad (1)$$

where:
N – number of instances (input-output) pairs used for learning,
y_i – network prediction (network output),
t_i – observed value (output based on data) for the i^{th} instance.
The higher the differences between the actual and predicted values, the larger the error and the more adjustment required for the network [7].
To compare the results with those obtained with use of Artificial Neural Networks; prediction was conducted with use of the SVR (Support Vector Regression), which is an element of the SVM (Support Vector Machine) method originating from the classical perceptron, whose construct are the Artificial Neural Networks. Both SVM and SVR enable to find a hyperplane that reflects the nature of the data in the best way and they control whether it meets the binding requirements of correct

element classification for SVM and computational error in a specific range for SVR. In comparison to many other learning methods, both SVM and SVR are characterised by the fact that their aim is not only to adapt to the learning data, but to do so in such a way that will allow to predict future behaviour of the modelled phenomenon in the best way possible. For SVM, the so-called *kernel trick* is used as a standard approach that allows to introduce non-linearity to the model and thus to improve its ability to cope with data that are not separated in a linear way [6]. The non-linear nature of these data may be captured as a result of using Radial Basis Functions (RBF). For SVR, the values of the co-ordinates of the vector *w* and the value of the absolute term *b* are found by solving the following optimisation problem:

min ||w|| maintaining:

$$\begin{cases} y_i - \vec{w} \cdot \vec{x}_i - b \leq \varepsilon \\ \vec{w} \cdot \vec{x}_i + b - y_i \leq \varepsilon \end{cases} \quad (2)$$

where:
||*w*|| – length of vector *w*,
w – weight vector determining the hyperplane that separates points belonging to different classes,
b – absolute term determined during optimisation,
x_i – set of data belonging to two classes defined by y_i, variables
y_i – actual values corresponding to x_i, input vectors
ε non-optimised method parameter that defines the acceptable level of error, i.e. of the difference between predicted values and those that exist in teaching data.
In this research, prediction of selected meteorological parameters was performed with use of the SVR method with Radial Basis Function. The RBF was characterised by the following parameters: gamma, defining the width of kernel function, equal to γ= 0.167 and C= 10.0, concerning the width of the margin of trust and maximum value of weight that may be determined for the given vector.
Model quality assessment was based on the value of Pearson Linear Correlation Coefficients for the actual and prognosed values and RMSE (Root Mean Squared Error) errors of the models.

3 Research results

The first stage of works involved calculating the correlation and degree of adjustment of temperature and relative air humidity obtained from both measurement sites in Wrocław.
The correlation coefficient for *tm* was very high and reached 0.986, while for *f* it was 0.894. The aim of the study was to assess the possibility to replace the daily values obtained from measurements on one station with data from the other one, so the RMSE values for both parameters were also determined. It was 1.8°C for *tm* and 6.5% for *f*. The aim of modelling the thermal and humidity parameters of the air with use of advanced, intelligent methods was to improve correlation between data and to reduce the RMSE error.
The evaluation of modelling results was based on r and RMSE, both in each sub-set separately and in both samples together (Table 1).
From the user's point of view, the results obtained in tests are the most important. The test sample is a set of independent data that is never used in the learning process. This is why the modelling results are usually worse than in the learning sub-set. Separating test data enables to evaluate the progress of modelling of the input-output relation by analysing the prediction accuracy and quality of the model based on information contained in data that were obtained from the learning dataset.
All the created models in the test dataset were characterised by high adjustment of predicted and observed values r≈0.9. The highest values were obtained for *tm* (0.979; 0.992; 0.954. respectively for SVR. MLP and RBF) and *tmax* (0.973; 0.973; 0.913), while lower ones were noted for *tmin* (0.940; 0.937; 0.894) and *f* (0.924; 0.926 and 0.884) (Table 1).
Very similar predicting results were obtained with use of SVR and MLP 6-6-5. On the other hand, the correlation between actual data and those modelled with use of RBF 6-28-5 network was lower.
In the test dataset, the lowest value of RMSE among all the analysed thermal parameters was obtained for *tm* and MLP network (1.4°C), while the prediction error of SVR was 1.7°C. The errors in extreme temperature models were higher: 1.9°C for *tmax* (SVR. MLP) and 2.5°C (SVR) and 2.7°C (MLP) for *tmin*. Air humidity parameters were predicted with 6.0 and 6.6% error, respectively, for MLP and SVR (*f*) and 1.7 and 1.6 hPa (*d*).

Table 1. Modelling results.

SVR	r			RMSE		
	Learning	Test	Both	Learning	Test	Both
tmax, °C	0.987	0.973	0.984	1.6	1.9	1.7
tmin, °C	0.954	0.940	0.948	2.1	2.5	2.2
tm, °C	0.987	0.979	0.985	1.3	1.7	1.4
f, %	0.919	0.924	0.908	5.3	6.6	5.6
d, hPa	0.965	0.955	0.958	1.3	1.7	1.4
MLP 6-6-5	**r**			**RMSE**		
	Learning	Test	Both	Learning	Test	Both
tmax, °C	0.984	0.973	0.983	1.7	1.9	1.7
tmin, °C	0.951	0.937	0.948	2.2	2.7	2.3
tm, °C	0.985	0.992	0.987	1.4	1.4	1.3
f, %	0.914	0.926	0.908	5.4	6.0	5.6
d, hPa	0.966	0.951	0.964	1.3	1.6	1.3
RBF 6-28-5	**r**			**RMSE**		
	Learning	Test	Both	Learning	Test	Both
tmax, °C	0.965	0.913	0.944	2.5	3.5	3.0
tmin, °C	0.924	0.894	0.893	2.7	2.8	3.1
tm, °C	0.964	0.954	0.942	2.1	2.1	2.6
f, %	0.900	0.884	0.877	5.8	6.8	6.3
d, hPa	0.940	0.918	0.928	1.7	2.1	1.8

Significantly higher values were noted for prediction with use of RBF 6-28-5 network: from 2.1°C for *tm* to 2.8 °C for *tmin and* 3.5 °C for *tmax.* Higher errors were also obtained for *f and d* (6.8%, 2.1 hPa).

Due to limited space, only selected diagrams were included in the paper.
The scatter plots of daily actual and modelled data confirm the existence of correlation between them.

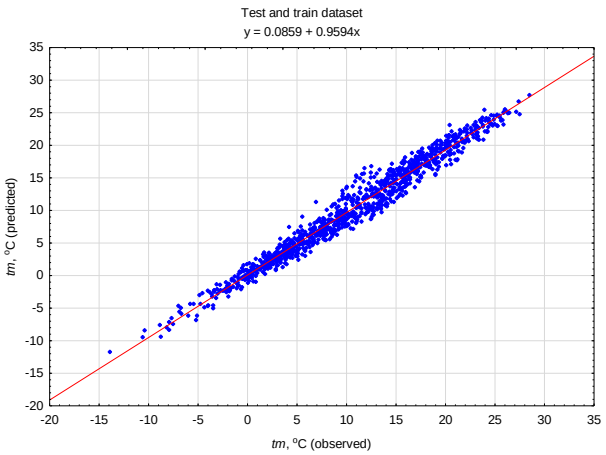

Fig. 1. Relation between observed values of *tm* (°C) and those predicted by SVR.

For all the analysed meteorological elements, the points were scattered along simple regression lines (Fig. 1-9).
The best adjustment was obtained for *tm* (Fig. 1, 2), with use of SVR and MLP. However, the results obtained from RBF network were considerably different (Fig. 3).

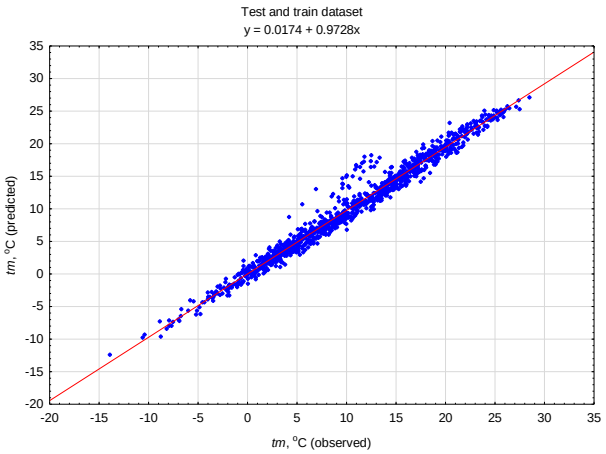

Fig. 2. Relation between observed values of *tm* (°C) and those predicted by MLP 6-6-5.

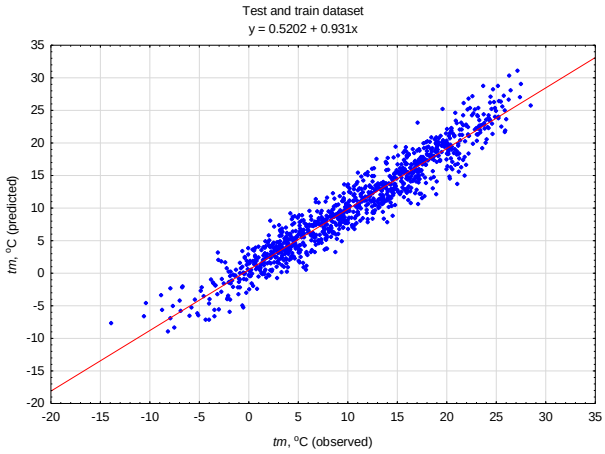

Fig. 3. Relation between observed values of *tm* (°C) and those predicted by RBF 6-28-5.

The minimum temperature was characterised by more scattered values along simple regressions. In this case, the differences between results obtained with use of SVR or MLP and RBF were not that noticeable on the diagrams (Fig. 4–6).

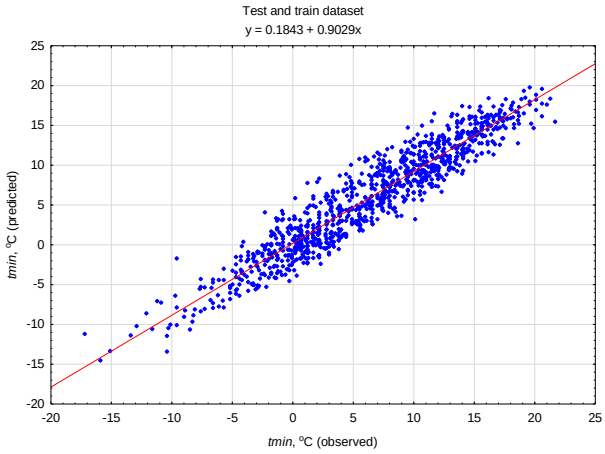

Fig. 4. Relation between observed values of *tmin* (°C) and those predicted by SVR.

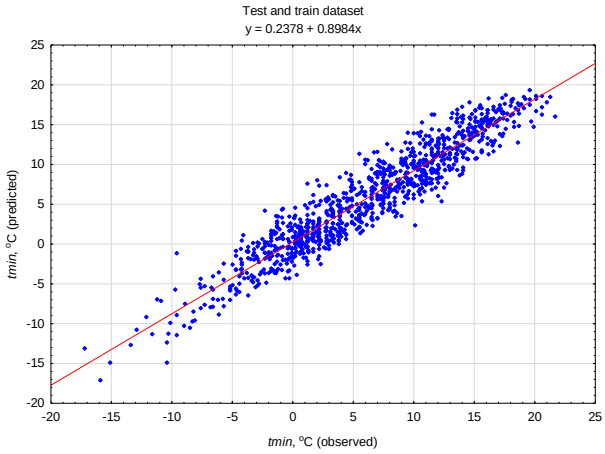

Fig. 5. Relation between observed values of *tmin* (°C) and those predicted by MLP 6-6-5.

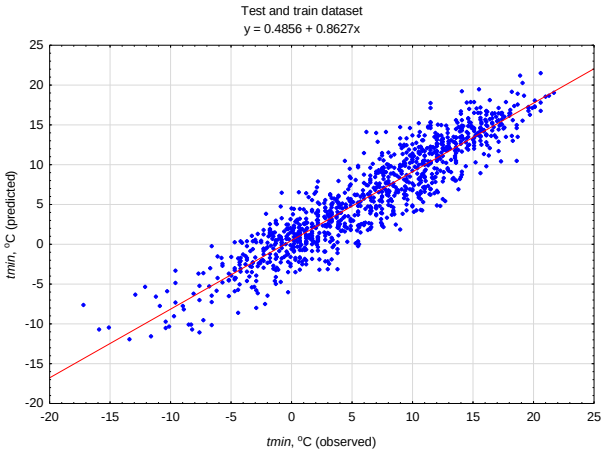

Fig. 6. Relation between observed values of *tmin* (°C) and those predicted by RBF 6-28-5.

The worst results (in spite of high correlation coefficients: ≥0.9) were obtained for relative air humidity. Here, the difference between the prediction and actual data was the highest, which is confirmed by much larger scattering of values along the lines. As it was in the case of minimum temperature, the significant difference between results obtained with SVR or MLP and RBF, based on the correlation r and RMSE errors was not visible on the diagrams (Fig. 7–9).

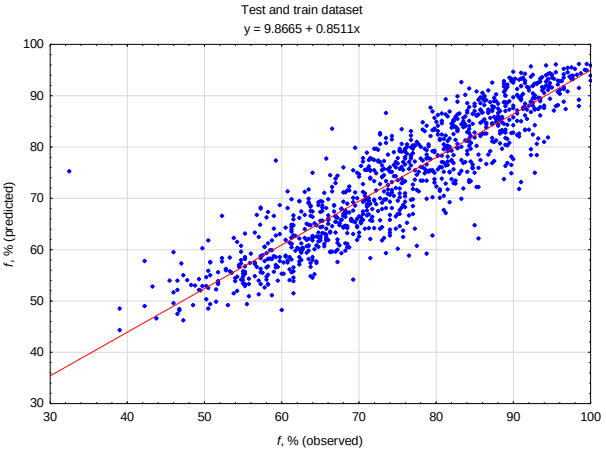

Fig. 7. Relation between observed values of *f* (%) and those predicted by SVR.

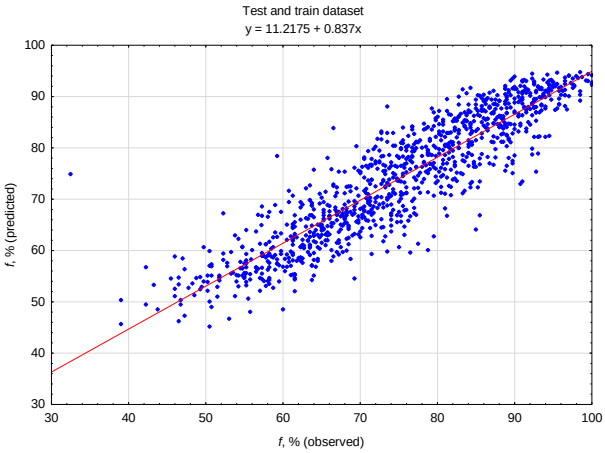

Fig. 8. Relation between observed values of *f* (%) and those predicted by MLP 6-6-5.

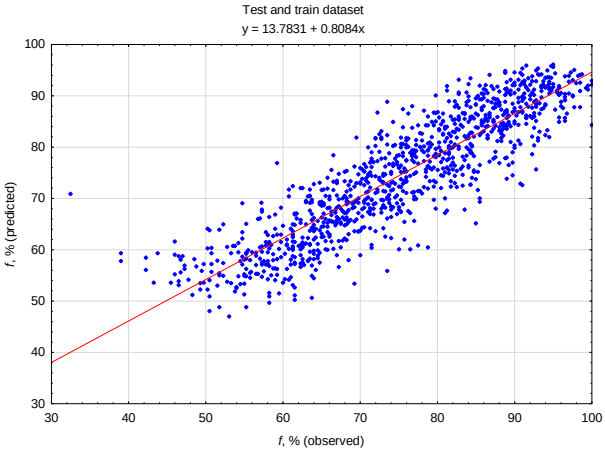

Fig. 9. Relation between observed values of *f* (%) and those predicted by RBF.

The residual analysis of the obtained models demonstrates that their distribution is very similar to normal distribution, which corresponds to the general assumption of the normality of noise contained in data [7] (Fig. 10–18).

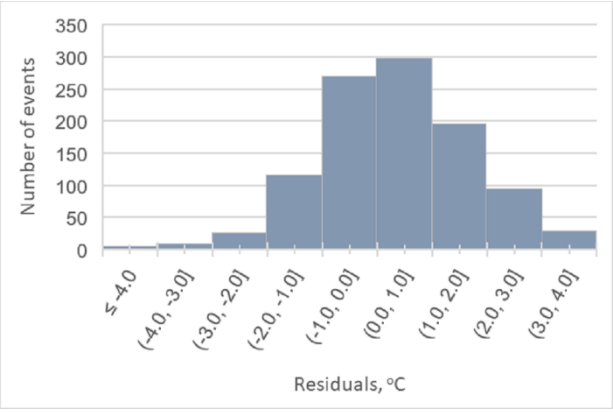

Fig. 10. Residual histogram of the *tm* (°C) model obtained using SVR.

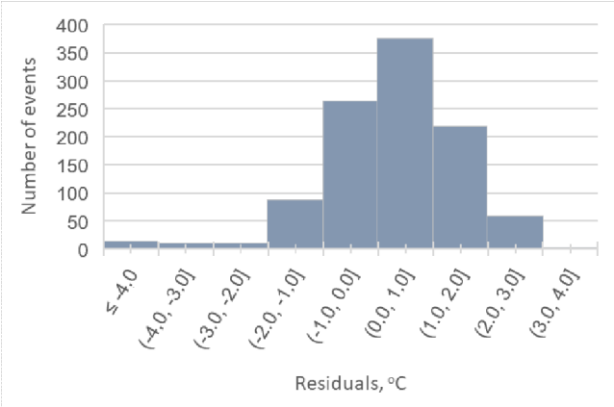

Fig. 11. Residual histogram of the *tm* (°C) model obtained using MLP 6-6-5.

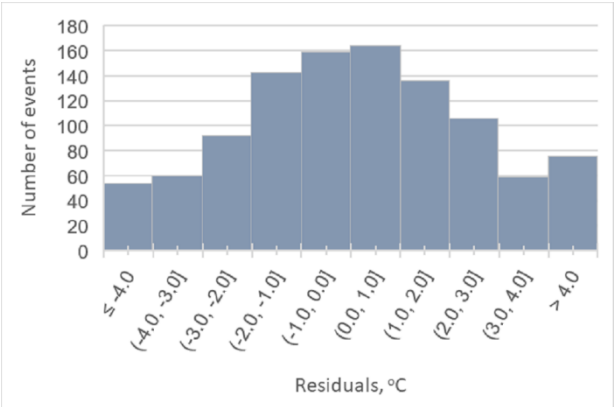

Fig. 12. Residual histogram of the *tm* (°C) model obtained using RBF 6-28-5.

The most numerous residual ranges were (0.0; 1.0]°C for *tm*. *tmin* (SVR), *tmax* (RBF). (-1.0; 0.0]°C for *tmax* (SVR and MLP), *tmin* (RBF). (1.0; 2.0]°C for *tmin* (MLP). The most numerous class included approx. 300 residual values of SVR and MLP models of mean and maximum air temperature and approx. 200 for minimum temperature. Residual histograms of prediction models obtained with use of RBF differed from the above distribution, and the most numerous class contained approx. 170 cases.

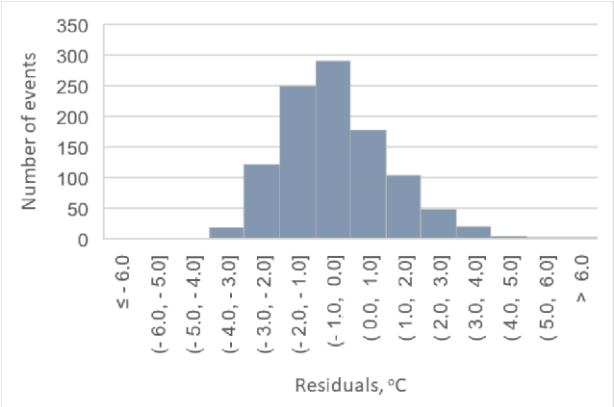

Fig. 13. Residual histogram of the *tmax* (°C) model obtained using SVR.

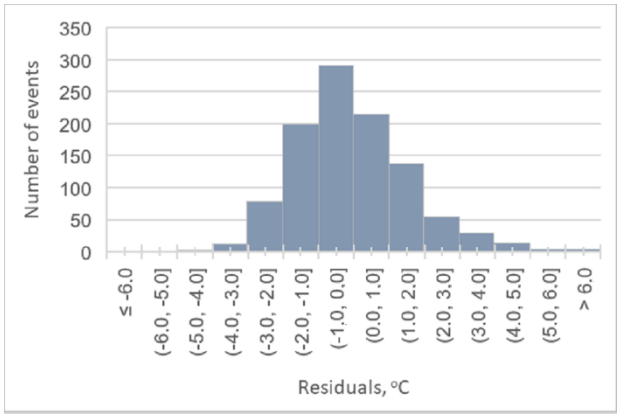

Fig. 14. Residual histogram of the *tmax* (°C) model obtained using MLP 6-6-5.

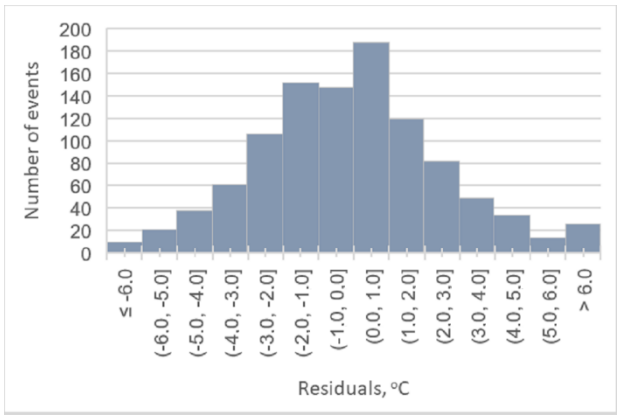

Fig. 15. Residual histogram of the *tmax* (°C) model obtained using RBF 6-28-5.

Histograms for relative air humidity show that the most frequently occurring differences were in the range (0; 5]%, in 320 (RBF) and approx. 400 cases (SVR and MLP), while for saturation deficit the most numerous class was (0.0; 1.0] hPa – SVR and (-1.0; 0.0) hPa – MLP, RBF and it included 340 values for SVR, 390 for MLP and 290 for RBF.

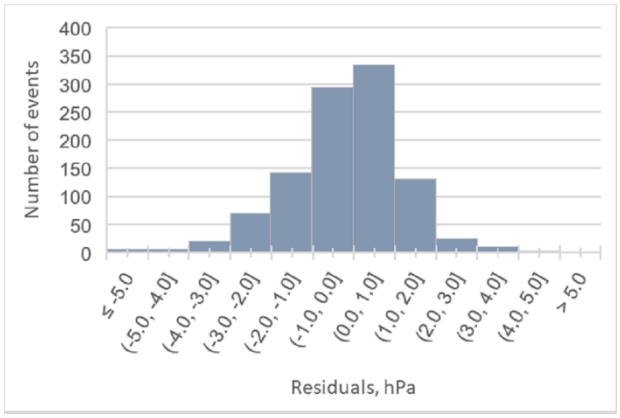

Fig. 16. Residual histogram of the *d* (hPa) model obtained using SVR.

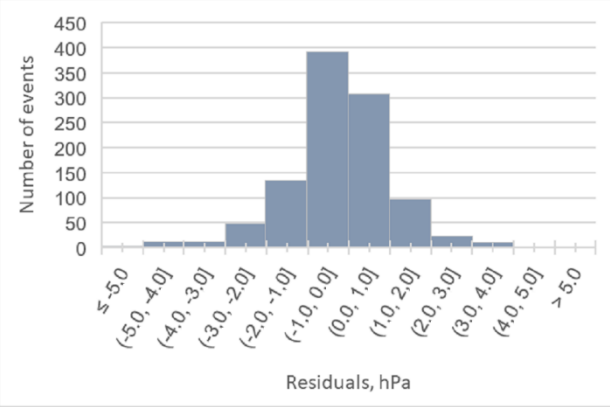

Fig. 17. Residual histogram of the *d* (hPa) model obtained using MLP 6-6-5.

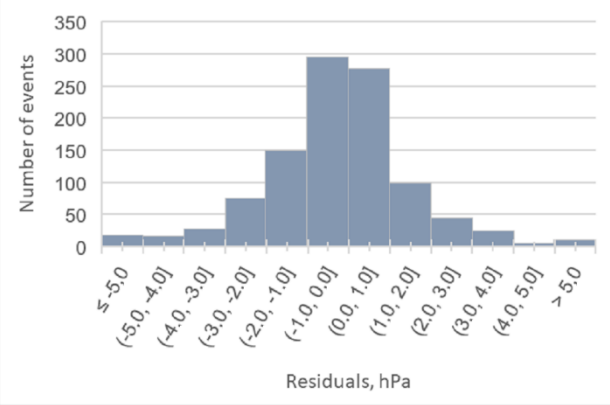

Fig. 18. Residual histogram of the *d* (hPa) model obtained using RBF 6-28-5.

The scope of histograms obtained for models of the analysed meteorological parameters created with use of 6-28-5 network was much wider than for the other models. This means that the residuals were characterised by much higher variability, which was also confirmed by the results obtained earlier.

4 Conclusions

Computation Intelligence Models are modern tools that are becoming more and more popular in supporting the processes of collecting, processing and analysing data obtained from meteorological and climatological measurements.

The research presented herein used the Multi-Layer Perceptron and Radial Basis Function networks and Support Vector Regression to predict the daily values of average, maximum and minimum air temperature as well as relative air humidity and saturation deficit in a 3.5 year period of measurement. One of the aims of the research work was to evaluate the possibility to apply the aforementioned Computational Intelligence methods to supplement incomplete time series basing on data obtained from different stations, with use of various measurement equipment types.

All the created models were characterised by good adjustment of the predicted values to observed ones, r>0.9. The highest correlation coefficients, and thus the smallest modelling errors were obtained for mean and maximum air temperature, while weaker correlations were noted for minimum temperature and relative air humidity.

The differences in quality of models obtained for specific elements result mainly from two reasons. First of all, the station that provided input data, is located in the city centre, within the Urban Heat Island, while output data were obtained from a station located outside it. The correlation between thermal parameters of the air (*tm*, *tmax*, *tmin*) in both stations is stronger than between humidity parameters (*f* and *d*), whose value is strongly affected by the type of environment.

For the purposes of the present paper, the authors intentionally selected two stations located in completely different areas in order to highlight the differences between them and to verify the hypothesis that computational intelligence methods are applicable in such cases as well. According to the methodology recommended by the WMO, incomplete series of measurement data are supplemented by data from stations located in close vicinity and in a similar environment.

Moreover, the discrepancies between the models obtained for the analysed parameters result from different methods of measurement and of calculating daily averages. The element whose modelling results were the most similar to actual values among all the parameters analysed herein is the average air temperature. Larger differences noted for extreme temperatures measured with use of liquid thermometers (input data) and automatic sensors (output data) in comparison to *tm* can be explained by lower sensitivity of extreme thermometers and higher likelihood of errors while reading the measurements. This refers mainly to the minimum thermometer and the location of the rod in the capillary tube. Moreover, in standard stations the noted extreme temperatures are the actual absolute minimum and maximum temperatures of the given day, while *tmax* and *tmin* in automatic stations are averaged values from one hour.

Better prediction capabilities were shown by SVR and MLP 6-6-5, which yielded similar results. The model obtained with use of RBF 6-28-5 network, although its quality was satisfactory, was considerably divergent.

The obtained results demonstrate that the Computational Intelligence methods used to supplement incomplete measurement series of thermal and humidity parameters of air may provide an alternative to standard solutions.

In the article meteorological data from the Faculty of Environmental Engineering and Geodesy Wroclaw University of Environmental and Life Sciences Observatory of Agro- and Hydrometeorology Wroclaw-Swojec (WOAiHW-S) as well as from Municipal Water Supply and Sewage Company in Wroclaw (MPWiK) were used.

The research was carried out with the statutory activity funds of the Faculty of Environmental Engineering and Geodesy Wroclaw University of Environmental and Life Sciences in 2017. Task *Prediction of water demand using computational intelligence methods* with agreement number B030/0103/17.

References

1. N. Alavi, J. S. Warland, A. A. Berg, Agric. For. Meteorol. **141 (1)**, 57–66, (2006)
2. H. Alexandersson, J. Clim. **6**, 661-675, (1986)
3. G. Alexandri, A. K. Georgoulias, C. Meleti, D. Balis, K. A. Kourtidis, A. Sanchez-Lorenzo, J. Trentmann, P. Zanis, Atmos. Res. **188**, 107–121, (2017)
4. L. Campozano, E. Sánchez Cordero, A. Avilés, E. Samaniego, MASKANA **5** (2014)
5. Y. Chi, C. Zhang, C. Liang, H. Wu, Acta Ecol. Sin. **33 (4)**, 217-226 (2013)
6. C. Cortes, V. Vapnik, Mach. Learn. **20(3)**, 273–297 (1995)
7. Dell Inc. Dell Statistica (data analysis software system) **13**, software.dell.com (2016)
8. M. Mazhar, M.T. Ikram, N.A. Butt, A.J. Butt (2015)
9. F.D. Mwale, A.J. Adeloye, R. Rustum, Phys. Chem. Earth **50**–**52**, 34–43 (2012)
10. S. Ribeiro, J. Caineta, A. C. Costa, Phys. Chem. Earth, Parts A/B/C **94**, 167–179 (2016)
11. S. Ribeiro, J. Caineta, A. C. Costa, R. Henriques, A. Soares, Atmos. Res. **171**, 147–158 (2016)
12. M. Sanchez-Lorenzo, M. Wild, J. Trentmann, Remote Sens. Environ. **134**, 355-366 (2013)
13. M. Santos, M. Fragoso, Atmos. Res. **131**, 34–45 (2013)
14. Q. Wang, Quat. Int. **279–280**, 527 (2012)
15. WMO **100** (2011)
16. WMO **8** (2014 edition, updated in 2017)
17. J. Zha, J. Wu, D. Zhao, Q. Yang, Atmos. Sci. Lett. **17**, 264-269 (2016)