European Water 57: 193-199, 2017.

Imputation of erosivity values under incomplete rainfall data by machine learning methods

K. Vantas and E. Sidiropoulos*
Faculty of Engineering, Aristotle University of Thessaloniki, Greece
* *e-mail: nontas@topo.auth.gr*

Abstract: In this article, a comparison is presented of empirical equations to machine learning methods for the estimation and imputation of rainfall erosivity values, associated with significant amounts of rainfall measurements that are missing in the available recording rain gauge data of the Greek Hydroscope database. The empirical equations are mainly based on exponential relations between erosivity and rainfall, while the machine learning methods employed in this paper are feed-forward neural networks with Bayesian regularization and ridge regression with nonlinear transformation. The data came from 81 measuring stations of the Ministry of the Environment and Energy. In the employed algorithms, the output was the weekly cumulative erosivity value, which resulted from processing the data of all rain gauges and pluviographs, while the input data consisted of the weekly cumulative rainfall, the month, the co-ordinates and the elevation of the station, as well as the number of days for which the rainfall was recorded. For validation, a method of nested cross-validation was employed. The machine learning methods gave significantly better results compared to the empirical equations, thus reducing the effects of estimating R from only weekly rainfall records.

Key words: machine learning, missing values, imputation, rainfall erosivity

1. INTRODUCTION

The soil loss problem of Greece dates back to prehistoric times (Vita-Finzi, 1969; Van Andel and Zangger, 1990), while during the last fifty years it has been aggravated due to the intensification and mechanization of the agricultural sector (Boardman and Poesen, 2006). In the year 2001, the Hellenic National Action Plan against Desertification was enacted, in which it was recognized that the country is inflicted by the phenomenon of desertification. The most significant process responsible for soil loss in Greece is related to rainfall erosivity.

The evaluation of rainfall erosivity is essential for the assessment of the soil loss risk, and the difficulty in small scale modelling has led to more tractable rainfall indices, such as the coefficient R of the Universal Soil Loss Equation (USLE) (Wischmeier and Smith, 1978). The computation of R requires pluviograph data for the determination of 30 minutes' maximum intensities of storms over time periods of more than 20 years (Renard and Freimund, 1994). In order to deal with the lack of such data, several models have been developed based on rainfall measurements at various time steps, on spatial parameters and on climatological data, such as maximum precipitation etc. (Diodato and Bellocchi, 2007; Angulo-Martínez and Beguería, 2009).

Pluviograph data, although available to a great extent in Greece, suffer from significant proportions of missing values. This fact necessitated imputation of the erosivity values that resulted from the data, so as to finally estimate R on a countrywide basis. For this imputation both empirical equations and machine learning methods were employed and comparisons between these two approaches are presented in this paper.

The problem of infilling hydro-meteorological data in general has been dealt within the literature, in terms of local averages (Pappas et al., 2014) and in relation to various forms of neural networks (Coulibaly and Evora, 2007; Nkiaka et al., 2016). Imputation in relation to erosivity is scarce. A recent publication (Diodato et al., 2017) concerns reconstruction of time series from

coarser rainfall records. The use of machine learning methods, such as Neural Networks can be found in the recent literature on the subject of rainfall prediction (Hellman et al., 2012; Sharma and Goyal, 2016), but not in relation to the special issue of erosivity, as presented in this paper.

2. DATA AND METHODS

2.1 Constitution of the data set

The data utilized in the analysis were taken from the Hydroscope Database (Koutsogiannis et al., 1995) and came from 81 meteorological stations of the Ministry of the Environment (Figure 1). The time series comprised a total of 2,333 years of pluviograph records with a time step of 30 minutes for the time period from 1953 to 1997, including the corresponding rain gauge 24-hour measurements with average record length 28 years. The time series were checked for gross errors and cleaned from noise that was due to the initial digitization of the pluviometers' bands. Missing values were marked out in a consistent way. In the pluviograph data, the percentage of these values was 44% on average, while the corresponding value in the rain gauges' data was 7%.

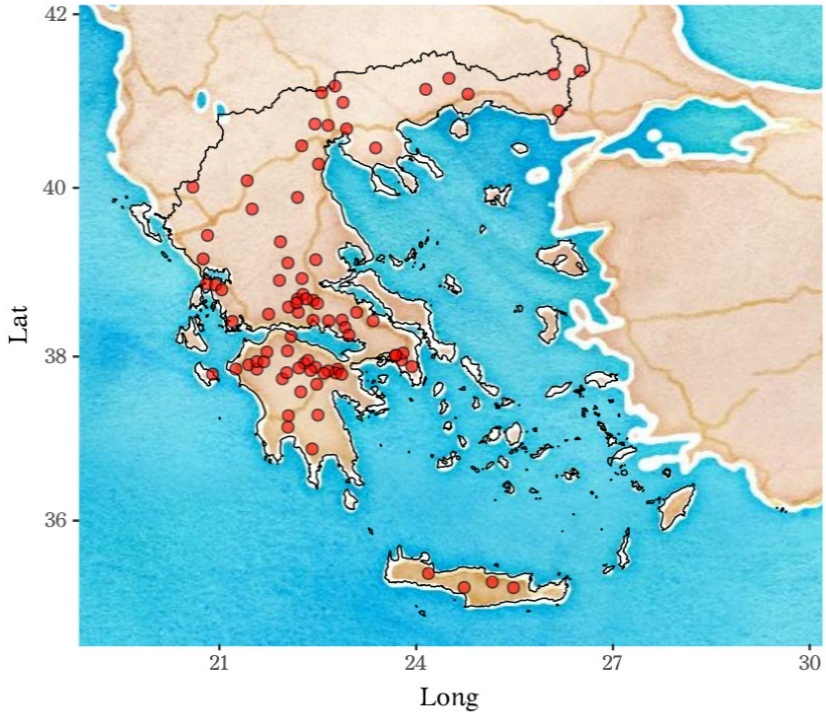

Figure 1. Meteorological stations.

For the above described data, it was deemed useful to change the time step and aggregate the data into weekly values, because 57% of the recordings were associated with storms occupying time periods covering parts of more than one calendar day, although only 17% of the storms had duration of more than 24 hours. Under the time step of one week, it was found out that 80% of the values emanated from a single storm. The storms that were crossed temporally by two consecutive weeks were assigned to the first of the two weeks. They comprised only 7% of the data. Thus, through the use of weekly instead of daily values, divisions of storms due to time-step were prevented.

2.2 Extraction of rainfall erosivity

The R coefficient (MJ.mm/ha/h/yr) is defined as the long-term average of the product of the

kinetic energy of a storm and the maximum 30 min intensity (Brown and Foster, 1987):

$$R = \frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{m_j}\left(EI_{30}\right)_k \tag{1}$$

where n is the number of years with rainfall records, m_j the number of storms during year j and EI_{30} the erosivity of storm k. The erosivity EI_{30} (MJ.mm/ha/h) is equal to:

$$EI_{30} = \left(\sum_{r=1}^{m} e_r v_r\right) I_{30} \tag{2}$$

where e_r is the energy of rainfall (MJ/ha/mm) and v_r the rainfall depth (mm) for the time interval r of the hyetograph, which has been divided into r = 1, 2, ..., m subintervals, such that each one of these is characterized by constant rainfall intensity. The quantity e_r is calculated for r from the relation:

$$e_r = 0.29 \cdot \left(1 - 0.72e^{-0.05 i_r}\right) \tag{3}$$

where i_r is the rainfall intensity (mm/h). The rules that apply in order to single out the storms causing erosion and to divide rainfalls of large duration are: (a) A rainfall event is divided into two parts, if its cumulative depth for duration of 6 hours at a certain location is less than 1.3 mm and (b) A rainfall is considered erosive, if it has a cumulative value greater than 12.7 mm or if during a time period of 15 mins a cumulative value of at least 6.4 mm is recorded.

2.3 Solution methods

The erosivity imputation problem is set up as a scheme of statistical learning, consisting of (a) data representing features of the studied physical processes, (b) measured outcomes of the same processes and (c) numerical models that simulate the physical processes as closely as possible based on the data of (a) and (b).

2.3.1 Outcomes and features

The outcomes, denoted as $y^{(i)}$, where (i) is the index of a sample consisting of m = 20,603 values, represented weekly cumulative rainfall erosivity, as it resulted from the pertinent calculations based on the data of all pluviographs. The features included the weekly cumulative rainfall P of erosive events, the month to which the above individual m values are referred, the longitude, latitude and altitude of the meteorological stations and the number of the days of the week (one, two, three or more), for which rainfall is recorded. Depending on the model used, a different subset of input variables was used.

2.3.2 Validation of models

A number of models were evaluated for the imputation problem, accompanied each time by a suitable algorithm. In all cases the estimation of the out-of-sample error and the optimal hyperparameters of the used models was performed using nested cross validation. In the outer cross validation, the data were divided into 10 sections for the purpose of estimating the out-of-sample error. In the inner cross validation, every set of the outer training was divided again into 10 sections in order to select the optimal hyperparameters of the models. The division of the data into sections

was executed on a yearly basis, so that sets of equal sizes could be formed. The coefficient of determination R^2 was used as a measure of the error:

$$R^2 = 1 - \frac{\sum_{i=1}^{m} \left(h\left(x_{test}^{(i)} \right) - y_{test}^{(i)} \right)^2}{\sum_{i=1}^{m} \left(\hat{y}_{train} - y_{test}^{(i)} \right)^2} \tag{4}$$

thus, each model was compared with a simplistic one which was equal to the average of the training set values $\hat{y}_{train}$.

2.4 Models and algorithms

2.4.1 Nonlinear least squares models

Two alternative exponential models (Richardson et al., 1983; Yu and Rosewell, 1996) were tried, given by Equations (5) and (6) and leading to optimal adjustments of two respective hypotheses, NLLS1 and NLLS2:

$$h_\theta(x) = \theta_0 \cdot P^{\theta_1} \tag{5}$$

$$h_\theta(x) = \theta_0 \left(1 + \theta_1 \cos\left(\frac{\pi}{6}(m - \omega) \right) \right) \cdot P^{\theta_2} \tag{6}$$

where ω is the month with the largest median of the y values as it derived from the training data and θ_0, θ_1 and θ_2 the parameters of the two hypotheses. The latter were determined by minimizing the following objective function by the trust region reflective method (Coleman and Li, 1994, 1996):

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2 \tag{7}$$

2.4.2 Ridge regression with nonlinear transformation

The hypothesis LR tried was the following:

$$h_\theta(x) = \theta_0 + \sum_{j=1}^{17} x_j \cdot \theta_j \tag{8}$$

where x_j, j=1, 2, ..., 17 stands for the input data and θ_j, for the parameters of the linear regression to be determined. The input data x were the normalized values of log(P), Long, Lat, Alt, 11 binary values for the corresponding month and 2 binary values representing the rainy days. The normalizing transformation of the data was the following:

$$N\left(x^{(i)} \right) = \frac{x^{(i)} - \bar{x}}{sd(x)} \tag{9}$$

where $\hat{x}$ denotes average and sd(x) standard deviation. The parameters θ_j were determined from the minimization of the following objective function by means of the conjugate gradient algorithm (Nocedal and Wright, 2006):

$$J(\theta)=\frac{1}{2m}\sum_{i=1}^{m} w^{(i)}\left(h_\theta\left(x^{(i)}\right)-y^{(i)}\right)^2+\frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2 \quad (10)$$

The regularizing hyperparameter λ prevented overfitting of the model to the training data (Abu-Mostafa et al., 2012). The weight of sample i, $w^{(i)}$ was set equal to $w^{(i)} = y^{(i)}/\max(y)$ so that higher y values had proportionately larger weights and z = log(y) was the nonlinear transformation of the outcome. The optimal value of $\lambda = 3\cdot10^{-7}$ resulted from a vector of values $\mathbf{\Lambda} = [10^{-9}, 3\cdot10^{-8}, 10^{-8}, ...,1]$.

2.4.3 Feed-forward neural networks with Bayesian regularization

Neural networks as powerful regression tools try to mimic the function of the human brain (Bishop, 1995). The learning algorithm used here is Bayesian regularization back-propagation, in combination with the Levenberg-Marquart method (Moré, 1978). Regularization and feature selection are automatically incorporated and the whole algorithm yields a network with good generalizing abilities (MacKay, 1992; Foresee and Hagan, 1997).

The architecture of the neural networks included one hidden layer with a linear output. The training of the networks was performed by the method of early stopping (Wang et al., 1994) by utilizing a random validation set consisting of 10% of the training data, so as to avoid overfitting of the neural networks. In order to prohibit negative values, the network output results were bounded based on the smallest outcome value of the training data. The input data x were the normalized values of P, Long, Lat, Alt, eleven binary values representing the month and two binary values representing the number of rainy days.

The hypothesis NNET employed consisted in the ensemble averaging of 20 neural networks, in order to deal with the inherent instability of neural networks due to locally optimal solutions for their parameters. The number of hidden neurons was treated as a hyperparameter of the algorithm and it was finally determined as l = 30 from a vector of possible values L = [20, 25, 30, 35, 40].

3. RESULTS

The comparison of the algorithms was based on the work Demšar (2006), Garcia and Herrera (2008) and García et al. (2010) on the use of non-parametric methods for the evaluation of results of machine learning algorithms, because parametric hypothesis testing methods (pairwise t-test and ANOVA) were not deemed suitable due to the nature of the algorithms. The procedure employed for hypothesis testing made use of the statistical programming language R (R Core Team, 2016).

The mean ranking of the algorithms is given on the basis of the 10 values of R^2 that resulted from the outer nested cross validation and is: NLLS1 = 3.9, NLLS2 = 3.0, LR = 1.8 and NNET = 1.3 with median values of R^2: 0.46, 0.50, 0.61 and 0.65 respectively. Figure 2 shows the box plots of the error measures, where the superior performance of NNET is demonstrated.

The Friedman test (Friedman, 1937) was performed, in order to determine whether an algorithm has a systematically better or worse performance. The obtained p-value $1.7\cdot10^{-5}$ indicated that the null hypothesis of all the algorithms perform the same could be safely rejected. Then post-hoc tests followed for all possible pairs of algorithms using the Wilcoxon signed rank test (Wilcoxon, 1945). Because of the multiple pair wise tests, the p-values that resulted were adjusted using Benjamini and Hochberg method (1995), which controls the false discovery rate. The results are given in Table 1, where it is shown that there were statistically strong evidences that the machine learning algorithms outperformed parametric methods. Between LR and NNET there were statistical evidences that the latter performed better.

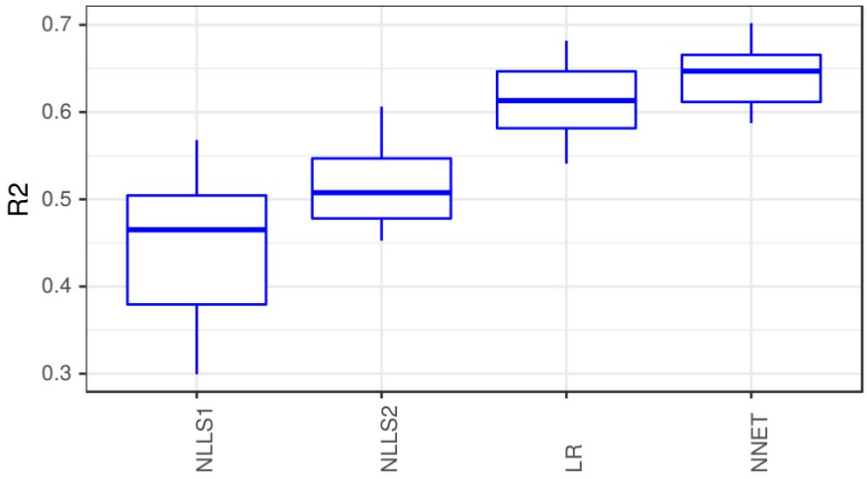

Figure 2. Outer cross validation R^2 boxplots.

Table 1. Adjusted Wilcoxon signed rank test p-values by Benjamini and Hochberg method.

	NLLS1	NLLS2	LR
NLLS2	0.011		
LR	0.004	0.006	
NNET	0.004	0.004	0.064

3. CONCLUSIONS

Because of the significant percentage of missing values in the time series, the use of the existing pluviograph data only would lead to an underestimation of R values. For this reason, a number of algorithms were employed for the imputation of erosivity values. From the performance evaluation of the algorithms by descriptive analysis and statistical inference, it was concluded that the presented machine learning algorithms outperformed the classical methods of estimation via parametric equations, reducing the effects of estimating R from weekly rainfall records. The NNET algorithm, besides having the best performance, presented the additional advantage of not necessitating prior knowledge of the form of its nonlinearity, in contrast to other tested algorithms.

REFERENCES

Abu-Mostafa, Y. S., Magdon-Ismail, M. and Lin, H.-T. (2012) Learning from data. Vol. 4, AMLBook Singapore.

Angulo-Martínez, M. and Beguería, S. (2009) Estimating rainfall erosivity from daily precipitation records: A comparison among methods using data from the Ebro Basin (NE Spain). Journal of Hydrology 379(1), 111–121.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) 57, 289–300.

Bishop, C. M. (1995) Neural Networks for Pattern Recognition. Oxford University Press.

Boardman, J. and Poesen, J. (2006), Soil Erosion in Europe, Wiley Online Library.

Brown, L. and Foster, G. (1987) Storm erosivity using idealized intensity distributions. Trans. ASAE 30(2), 379–386.

Coleman, T. F. and Li, Y. (1996) An interior trust region approach for nonlinear minimization subject to bounds. SIAM Journal on Optimization 6(2), 418–445.

Coleman, T. and Li, Y. (1994) On the convergence of Reflective Newton Methods for Large-scale Nonlinear Minimization Subject to Bounds. vol. 67, Ithaca, NY, USA: Cornell University.

Coulibaly, P. and Evora, N. (2007) Comparison of neural network methods for infilling missing daily weather records. Journal of Hydrology 341(1), 27–41.

Demšar, J. (2006) Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7(Jan), 1–30.

Diodato, N. and Bellocchi, G. (2007) Estimating monthly (R) USLE climate input in a Mediterranean region using limited data. Journal of Hydrology 345(3), 224–236.

Diodato, N., Borrelli, P., Fiener, P., Bellocchi, G. and Romano, N. (2017) Discovering historical rainfall erosivity with a parsimonious approach: A case study in Western Germany. Journal of Hydrology 544, 1–9.

Foresee, F. D. and Hagan, M. T. (1997) Gauss-Newton approximation to Bayesian learning. In: International Conference on Neural Networks, 1997., Vol. 3, IEEE, pp.1930–1935.

Friedman, M. (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association 32(200), 675–701.

García, S., Fernández, A., Luengo, J. and Herrera, F. (2010) Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. Information Sciences 180(10), 2044–2064.

Garcia, S. and Herrera, F. (2008) An extension on "Statistical Comparisons of Classifiers over Multiple Data Sets" for all Pairwise Comparisons. Journal of Machine Learning Research 9(Dec), 2677–2694.

Hellman, S., McGovern, A. and Xue, M. (2012) Learning ensembles of Continuous Bayesian Networks: An application to rainfall prediction. In: 'Intelligent Data Understanding (CIDU), 2012 Conference, IEEE, pp. 112–117.

Koutsogiannis, D., Tsakalias, G., Christofidis, A., Manetas, A., Sakelariou, A., Maurodimou, P., Papakostas, N., Mamasis, N., Nalbandis, I. and Ksanthopoulos, T. (1995) HYDROSCOPE: Creation of a National Databank for Hydrological and Meteorological Information. Technical report, National Technical University of Athens.

MacKay, D. J. (1992) Bayesian interpolation. Neural Computation 4(3), 415–447.

Moré, J. J. (1978) The Levenberg-Marquardt algorithm: Implementation and theory. In: Numerical Analysis, Springer, pp. 105–116.

Nkiaka, E., Nawaz, N. and Lovett, J. (2016) Using self-organizing maps to infill missing data in hydro-meteorological time series from the Logone catchment, Lake Chad basin. Environmental Monitoring and Assessment 188(7), 1–12.

Nocedal, J. and Wright, S. J. (2006) Conjugate Gradient Methods. Numerical Optimization pp. 101–134.

Pappas, C., Papalexiou, S. M. and Koutsoyiannis, D. (2014) A quick gap filling of missing hydrometeorological data. Journal of Geophysical Research: Atmospheres 119(15), 9290–9300.

R Core Team (2016) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Renard, K. G. and Freimund, J. R. (1994) Using monthly precipitation data to estimate the R-factor in the revised USLE. Journal of Hydrology 157, 287–306.

Richardson, C., Foster, G. and Wright, D. (1983) Estimation of erosion index from daily rainfall amount. Transactions of the ASAE 26(1), 153–156.

Sharma, A. and Goyal, M. K. (2016) Bayesian network for monthly rainfall forecast: A comparison of K2 and MCMC algorithm. International Journal of Computers and Applications 38(4), 199–206.

Van Andel, T. H. and Zangger, E. (1990) Landscape stability and destabilisation in the prehistory of Greece. In: Proceedings of the INQUA/BAI symposium on the impact of ancient man on the landscape of the Eastern Mediterranean region and the Near East, 12, 139–157.

Vita-Finzi, C. (1969) The Mediterranean valleys: Geological changes in historical times. Vol. 165, University Press.

Wang, C., Venkatesh, S. S. and Judd, J. S. (1994) Optimal Stopping and Effective Machine Complexity in Learning. Advances in Neural Information Processing Systems 6, 303–310.

Wilcoxon, F. (1945) Individual comparisons by ranking methods. Biometrics Bulletin 1(6), 80–83.

Wischmeier, W. H. and Smith, D. D. (1978) Predicting rainfall erosion losses - A guide to conservation planning. USDA, Science and Education Administration.

Yu, B. and Rosewell, C. (1996) An assessment of a daily rainfall erosivity model for New South Wales. Soil Research 34(1), 139–152.