Modeling the H-Index Based on the Total Number of Citations and the Duration from the First Publication

Mohammad Reza Mahmoudi, Marzieh Rahmati, Zulkefli Mansor, Amir Mosavi and Shahab S. Band

April 12, 2022

Modeling the H-index based on the Total Number of Citations and the Duration from the First Publication

Mohammad Reza Mahmoudi[1], Marzieh Rahmati[2], Zulkefli Mansor[3], Amir Mosavi[4,5], Shahab S. Band[6,7]

[1]Department of Statistics, Faculty of Science, Fasa University, Fasa, Fars, Iran
[2]Department of Computer Engineering, Yazd University, Yazd, Iran
[3]Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia
[4]Environmental Quality, Atmospheric Science and Climate Change Research Group, Ton Duc Thang University, Ho Chi Minh City, Vietnam
[5]Faculty of Environment and Labour Safety, Ton Duc Thang University, Ho Chi Minh City
[6]Institute of Research and Development, Duy Tan University, Da Nang 550000, Vietnam
[7]Future Technology Research Center, College of Future, National Yunlin University of Science and Technology 123 University Road, Section 3, Douliou, Yunlin 64002, Taiwan

Abstract. The productivity of researchers and the impact of the work they do is a preoccupation of universities, research funding agencies and sometimes even researchers themselves. The *h*-index is the most popular of different metrics to measure these activities. This research deals to present a practical approach to model the *h*-index based on the total number of citations and the duration from the publishing of the first article. To determine the effect of every factor (C and D) on H, we applied a set of simple nonlinear regression. The results indicated that both C and D had significant effect on H ($p<0.001$). The power of these equations to estimate of *H* was 93.4% and 39.8%, respectively, that verified the model based on C had a better fit. Then, to investigate the simultaneous effects of C and D on H, multiple nonlinear regression were applied. The results indicated that C and D had significant effect on H ($p<0.001$). Also, the power of this equation to estimate of *H* was 93.6%. Finally, to model and estimate the *h*-index, *h*, as a function of C and D, the multiple nonlinear quartile regression was used. The goodness of fitted model also was also assessed.

Keywords: H-index, Citation, Duration, Modelling, Relationship, Regression.

1. Introduction

The productivity of researchers and the impact of the work they do is a preoccupation of universities, research funding agencies and sometimes even researchers themselves. Various metrics have been used to measure these including journal impact factors, citation counts and publication rates. At present, however, the *h*-index is the most popular of these metrics (Hirsch, 2005; Braun et al., 2006; Schubert and Glänzel, 2007; Harzing and van der Wal, 2009). Hirsch's definition of the index is that $h = m$ if *m* of a researcher's *p* papers have at least *m* citations each and each of the other papers has no more than *m* citations. As a guide, Hirsch (2005) suggested that a 'successful' scientist would have $h = 20$ after 20 years of work, whereas outstanding and 'truly unique' individuals would have $h = 40$ and $h = 60$, respectively, after 20 years of work. Subsequent work has shown that this is too great a generalisation, if only because *h* is highly discipline-specific and depends on circumstance, the comprehensiveness of the literature databases used to calculate the index and many other factors (Vinkler, 2007; Ruch and Ball, 2010). For example, very eminent mathematicians often have $h < 10$ and some Nobel laureates also have very small *h*-indices (Yong, 2014). The inevitable inference is an individual's *h*-index should be considered in the context of these factors and of the distribution of *h* for a given number of papers and citations appropriate to the individual researcher. Some researchers introduced alternative versions of the h-index (Bar-Ilan, 2010). Generally, all of the given indices consider the number of citations received by articles. Recently, scientists have studied and developed theoretical models to estimate and model these indices based on other indicators, for example based on the total number of citations C (Hirsch, 2005), based on the total number of publications T (Egghe and Rousseau, 2006), based on the total number of publications with minimum one citation T_1 (Burrell 2013a), based on C and T (Glänzel, 2006; Iglesias and Pecharroman, 2007; Schubert and Glänzel, 2007; Bletsas and Sahalos, 2009; Egghe et al., 2009; Egghe and Rousseau, 2012), based on C, T_1 and the total number of citations for the 1 most cited papers C_1 (Bertoli-Barsotti and Lando, 2015). Burrell (2013b) and Bertoli-Barsotti and Lando (2015) respectively applied standard and shifted geometric distribution to predict and estimate the h-index of scientists. Bertoli-Barsotti and Lando (2017a) empirically studied the basic and improved Lambert-*W* formula for estimating the *h*-index and compare them with the well-known previous models. Bertoli-Barsotti and Lando (2017b) presented a new formula to estimate the *h*-index when we do not have information about the whole set of citation dataset.

This research deals to present a practical approach to model the *h*-index based on the total number of citations and the duration from the publishing of the first article

2. Methodology

This section is devoted to discuss about details of data collection, samples and statistical techniques that have applied to analyze dataset.

2.1. Data Collection

The dataset of this research contains the information of articles for 29470 Iranian scientists that have indexed in Google Scholar.

2.2. Data Analysis

Statistics, data analysis and data mining are popular approaches to extract knowledge from dataset. The data gathered from the Google Scholar were fed and analyzed using the SPSS 25, and R 3.3.2 software. First, the descriptive statistics about the values of h-index, C and D is provided.

To determine the effect of every factor (C and D) on H, we applied a set of simple nonlinear regression. Also to investigate the simultaneous effects of C and D on H, multiple nonlinear regression were applied. Finally, to model and estimate the h-index based on C and D, the multiple nonlinear quartile regression (MNLQR) was used. The goodness of fitted model also was assessed by the coefficient of determination (R^2), and comparing actual values with predicted values.

2.2.1. Simple Nonlinear Regression

To model a quantitative response variable Y based on a predictor variable X, simple nonlinear regression (SNLR) model is a powerful technique. The general equation of SNLR is presented by

$$Y = \beta_0 + \beta_1 X^{\beta_2} + \boldsymbol{\varepsilon},$$

where β_0, β_1 and β_2 are model parameters (coefficients) and $\boldsymbol{\varepsilon}$ is the random components of the model which follow independent normal distribution. The estimated equation of SLR model is presented by

$$\hat{Y} = b_0 + b_1 X^{b_2},$$

where, b_0, b_1, b_2 and $\hat{Y}$ are estimations of β_0, β_1, β_2, and Y, respectively.

2.2.2. Multiple Nonlinear Regression

To model a quantitative response variable Y based on predictor variables $X_1, \ldots, X_k$, multiple nonlinear regression (MNLR) is a powerful technique. The general equation of MNLR with two predictors X_1 and X_2 is presented by

$$Y = \beta_0 + \beta_1 {X_1}^{\beta_2} + \beta_3 {X_2}^{\beta_4} + \beta_5 {X_1}^{\beta_6} {X_2}^{\beta_7} + \boldsymbol{\varepsilon},$$

where β_0, …, β_7 are model parameters (coefficients) and $\boldsymbol{\varepsilon}$ is the random components of the model which follow independent normal distribution. The estimated equation of MNLR model is also presented by

$$\hat{Y} = b_0 + b_1 {X_1}^{b_2} + b_3 {X_2}^{b_4} + b_5 {X_1}^{b_6} {X_2}^{b_7},$$

where b_0, …, b_7 are estimations of β_0, …, β_7, and $\hat{Y}$ is the estimated value of Y.

2.2.3. Multiple Nonlinear Quartile Regression

In multiple nonlinear quartile regression (MNLQR), first the quartiles of response variable have been computed. Then, based on the values of quartiles, the observations categorized in 4 distinct categories. Finally, a separate MNLR is run, on each category.

3. Results

The descriptive statistics of research variables contained C and D is given the first subsection. The Subsection 2 reports the SNLR results to predict the separate effects of every factor (C and D) on *h*. The Subsection 3 is regards to MNLR results to investigate the simultaneous effects of C and D on H. The Subsection 4 reports the MNLQR results to model the effects of factors on *h*, in each quartile.

3.1. Descriptive Statistics

The descriptive statistics of research variables contained minimum, maximum, mean, standard deviation, and quartiles are summarized in Table 1. As Table 1 indicates the means of *h*, C and D for Iranian scientists are 5.74, 248.78, and 7.98, respectively. Also, the value of *h* for at least 25%, 50% and 75% of them is at most 2 (Q_1=2), 4 (Q_2=4), and 7 (Q_3=7), respectively.

Table 1: Descriptive statistics of research variables

	Mean	Std. Deviation	Minimum	Maximum	Quartile		
					First (Q_1)	Second (Q_2)	Third (Q_3)
h	5.74	5.79	1	84	2.00	4.00	7.00
C	248.78	828.84	1	37570	15.00	58.00	200.00
D	7.98	4.59	1	42	5.00	7.00	10.00

3.2. SNLR Results

This part is regard to study the impact of each factor (C and D) on *h*. In this research, *h* was the response variable. Also the variables C and D were continuous predictors. Tables 2 and 3 summarize the results of SNLR models for the variables C and D. As Table 2 indicates, the C and D factors had significant effect on *h* ($p<0.001$). Figure 1 also shows the plot of fitted curve with data.

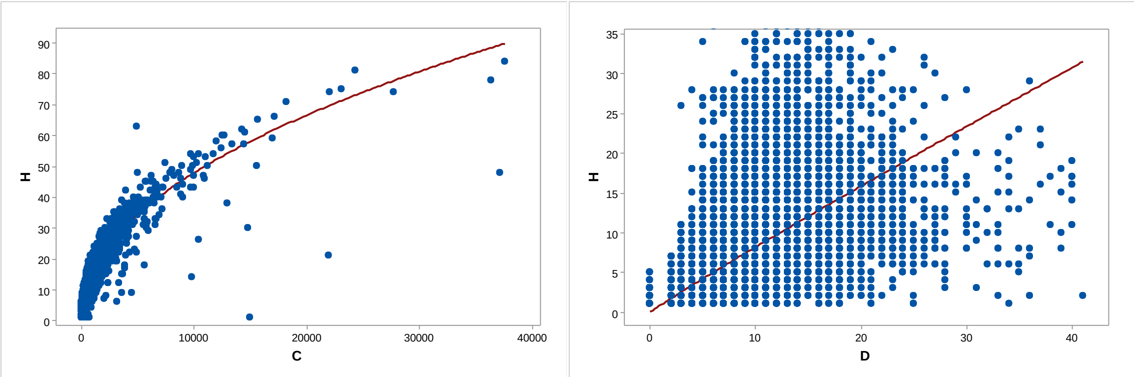

Figure 1: Plot of fitted curve with data SNLR models

Table 3 shows the parameter estimates of SNLR models for C and D, respectively. Based on the results of Table 3, we can estimate the h as a function of C and D, by

$$\hat{h}_C = 0.600C^{0.476},$$

and

$$\hat{h}_D = 0.667D^{1.041},$$

respectively. Also, the power of these equations to estimate of h is 93.4% and 39.8%, respectively. Figure 2 and Table 4 show the plot of actual values versus predicted values and the correlations between them. As can be seen the SNLR model based on C had a better fit.

Table 2: The results of SNLR models to study the effect of C and D on h

Factor	Source	Sum of Squares	df	Mean Squares	F	R^2	p
C	Regression	1892965	2	946482.5	429143.66	0.934	<0.001
	Residual	64992.09	29468	2.205514			
	Uncorrected Total	1957957	29470				
	Corrected Total	987069.3	29469				
D	Regression	1364231	2	682115.4	33854.95	0.398	<0.001
	Residual	593726.3	29468	20.14817			
	Uncorrected Total	1957957	29470				
	Corrected Total	987069.3	29469				

Table 3: The parameter estimates of SNLR models for C and D

Factor	Parameter	Estimate	Std. Error	95% Confidence Interval		p
				Lower Bound	Upper Bound	
C	b_1	0.600	0.003	0.595	0.606	<0.001

	b_2	0.476	0.001	0.474	0.477	<0.001
D	b_1	0.667	0.014	0.640	0.694	<0.001
	b_2	1.041	0.008	1.025	1.057	<0.001

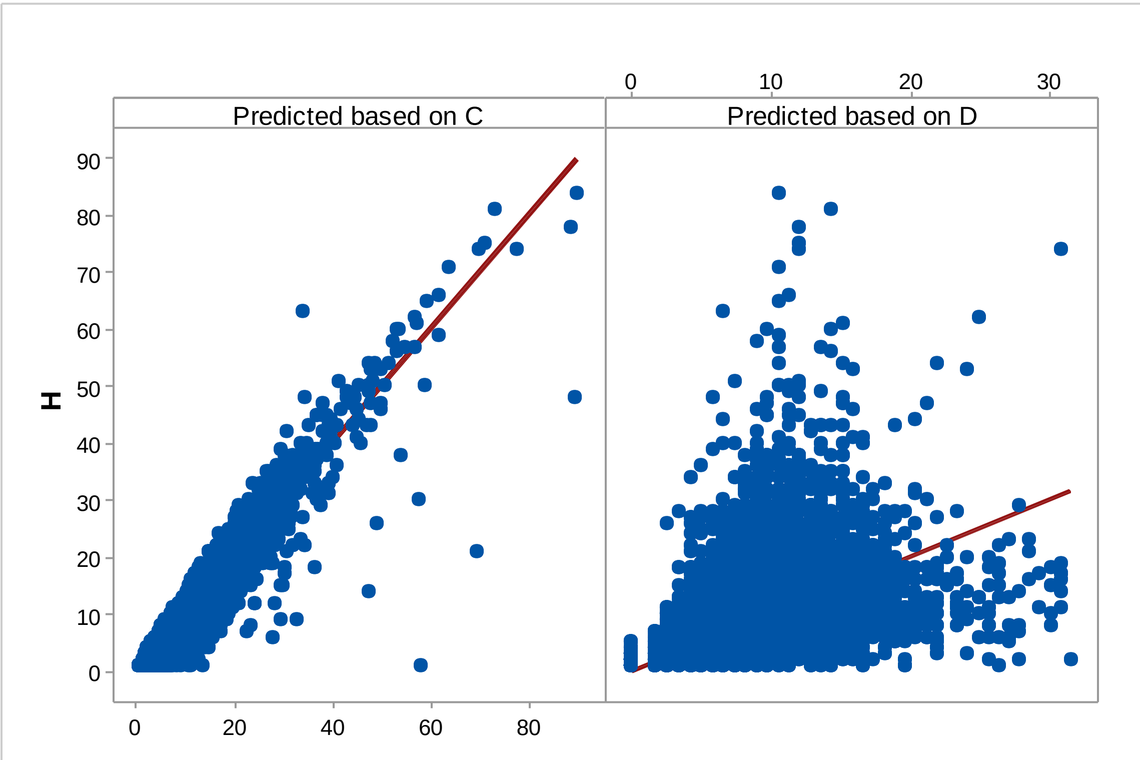

Figure 2: Plot of actual values versus predicted values

Table 4: Pearson and Spearman correlations between actual values and predicted values

	Spearman's rho		Pearson	
	Correlation Coefficient	p	Correlation Coefficient	p
Predicted Values (based on C)	0.954	<0.001	0.967	<0.001

Predicted Values (based on D)	0.779	<0.001	0.632	<0.001

3.3. MNLR Results

This part is regard to study the simultaneous impacts of C and D on h. Tables 5 and 6 summarize the results of MNLR model. As Table 5 indicates, the C and D factors had significant effect on H ($p<0.001$). Table 6 shows the parameter estimates of MNLR model.

Table 5: The results of MNLR model to study the effect of C and D on h

Factor	Source	Sum of Squares	df	Mean Squares	F	R^2	p
C, D	Regression	1895013.156	7	270716.1652	126717.88	0.936	<0.001
	Residual	62943.8438	29463	2.136369134			
	Uncorrected Total	1957957	29470				
	Corrected Total	987069.2904	29469				

Based on the results of Table 6, we can estimate the H as a function of C and D, by

$$\hat{h}_{C,D} = 0.673C^{0.419} - 0.183D^{0.939} + 0.129C^{0.424}D^{0.370}.$$

Also, the power of this equation to estimate of h is 93.6% that is not significantly more than 93.4% ($\hat{h}_C$). Figure 3 and Table 7 show the plot of actual values versus predicted values and the correlations between them. As can be seen the MNLR model can nicely estimate the values of h.

Table 6: The parameter estimates of MNLR model

Parameter	Estimate	Std. Error	95% Confidence Interval	

			Lower Bound	Upper Bound	p
b_1	0.673	0.020	0.633	0.712	<0.001
b_2	0.419	0.014	0.392	0.445	<0.001
b_3	-0.183	0.028	-0.238	-0.128	<0.001
b_4	0.939	0.061	0.819	1.058	<0.001
b_5	0.129	0.028	0.073	0.184	<0.001
b_6	0.424	0.027	0.372	0.477	<0.001
b_7	0.370	0.084	0.206	0.534	<0.001

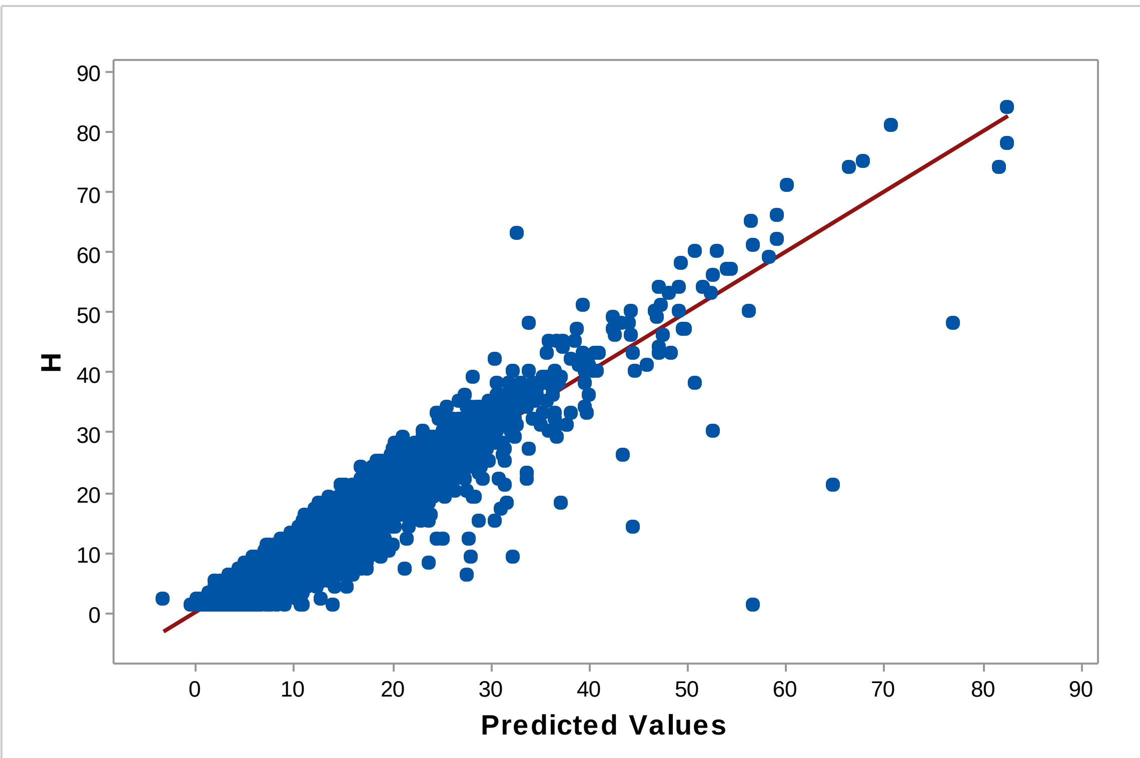

Figure 3: Plot of actual values versus predicted values

Table 7: Pearson and Spearman correlations between actual values and predicted values

	Spearman's rho	Pearson

	Correlation Coefficient	p	Correlation Coefficient	p
Predicted Values (based on C and D)	0.968	<0.001	0.953	<0.001

3.4. MNLQR Results

This part is regard to study the simultaneous impacts of C and D on different quartiles of h. We divide the observations in 4 groups as follow: First group: Observations with $h \leq 2$; Second group: Observations with $2 < h \leq 4$; Third group: Observations with $4 < h \leq 7$; Fourth group: Observations with $h > 7$. Based on the results of Table 8, we can conclude that the C and D factors had significant effect on h (p<0.001), in every category. Based on the results, the h can be estimated as a function of C and D, by

$$\hat{h}_{C,D} = b_1 C^{b_2} + b_3 D^{b_4} + b_5 C^{b_6} D^{b_7},$$

in categories 1 to 4, respectively.

Table 8: The parameter estimates of MNLQR model

Category	Parameter	Estimate	Std. Error	95% Confidence Interval		p
				Lower Bound	Upper Bound	
1	b_1	.929	.010	.909	.948	<0.001
	b_2	.230	.008	.214	.246	<0.001
	b_3	.104	.020	.064	.144	<0.001
	b_4	.813	.079	.658	.968	<0.001
	b_5	-.057	.015	-.087	-.027	<0.001
	b_6	.322	.020	.284	.361	<0.001

	b_7	.729	.079	.574	.883	<0.001
2	b_1	11.837	322.351	-620.073	643.748	<0.001
	b_2	.211	1.558	-2.844	3.266	<0.001
	b_3	-4.951	14.521	-33.416	23.514	<0.001
	b_4	.021	.182	-.335	.377	<0.001
	b_5	-5.983	335.988	-664.626	652.661	<0.001
	b_6	.288	1.739	-3.122	3.698	<0.001
	b_7	-.006	.256	-.508	.496	<0.001
3	b_1	1.682	.191	1.307	2.057	<0.001
	b_2	.319	.036	.247	.390	<0.001
	b_3	.082	.194	-.297	.462	<0.001
	b_4	.554	.564	-.552	1.659	<0.001
	b_5	-.069	.051	-.168	.030	<0.001
	b_6	.672	.057	.561	.783	<0.001
	b_7	.093	.036	.022	.164	<0.001
4	b_1	.414	.032	.352	.476	<0.001
	b_2	.523	.009	.505	.541	<0.001
	b_3	4.709	.643	3.449	5.969	<0.001
	b_4	-.494	.101	-.693	-.295	<0.001
	b_5	-.001	.001	-.003	.000	<0.001
	b_6	1.275	.055	1.167	1.383	<0.001
	b_7	-1.348	.148	-1.637	-1.059	<0.001

4. Conclusion

This research dealt to present a practical approach to model the h -index (h) based on the total number of citations (C) and the duration from the publishing of the first article (D). To determine the effect of every factor (C and D) on h, we applied a set of simple nonlinear regression. The results indicated that both C and D had significant effect on h ($p<0.001$) and we can estimate the h as a function of C and D, by

$$\hat{h}_C = 0.600C^{0.476},$$

and

$$\hat{h}_D = 0.667D^{1.041},$$

respectively. Also, the power of these equations to estimate of h was 93.4% and 39.8%, respectively, that verified the model based on C had a better fit.

Then, to investigate the simultaneous effects of C and D on h, multiple nonlinear regression were applied. The results indicated that C and D had significant effect on h ($p<0.001$) and we can estimate the h as a function of C and D, by

$$\hat{h}_{C,D} = 0.673C^{0.419} - 0.183D^{0.939} + 0.129C^{0.424}D^{0.370}.$$

Also, the power of this equation to estimate of H was 93.6% that is not significantly more than 93.4% ($\hat{h}_C$).

Finally, to model and estimate the h, as a function of C and D, the multiple nonlinear quartile regression was used. The goodness of fitted model also was also assessed.

References

Bar-Ilan, J. (2010). Ranking of information and library science journals by JIF and by *h*-type indices. *Journal of Informetrics*, ***4***, 141–147.

Bertoli-Barsotti, L., & Lando, T. (2015). On a formula for the *h*-index. *Journal of Informetrics*, ***9*****(4)**, 762–776.

Bertoli-Barsotti, L., & Lando, T. (2017a). A theoretical model of the relationship between the h-index and other simple citation indicators. *Scientometrics*, ***111*****(3)**, 1415–1448.

Bertoli-Barsotti, L., & Lando, T. (2017b). The h-index as an almost-exact function of some basic statistics. *Scientometrics*, ***113*(2)**, 1209–1228.

Bletsas, A., & Sahalos, J. N. (2009). Hirsch index rankings require scaling and higher moment. *Journal of the American Society for Information Science and Technology*, ***60***, 2577–2586.

Braun, T., Glänzel, W., & Schubert, A. (2006). A Hirsch-type index for journals. *Scientometrics*, ***69***, 169–173.

Burrell, Q. L. (2013a). Formulae for the *h*-index: A lack of robustness in Lotkaian informetrics? *Journal of the American Society for Information Science and Technology*, ***64***, 1504–1514.

Burrell, Q. L. (2013b). The *h*-index: A case of the tail wagging the dog? *Journal of Informetrics*, ***7***, 774–783.

Egghe, L., Liang, L., & Rousseau, R. (2009). A relation between *h*-index and impact factor in the power-law model. *Journal of the American Society for Information Science and Technology*, ***60***, 2362–2365.

Egghe, L., & Rousseau, R. (2006). An informetric model for the Hirsch-index. *Scientometrics*, ***69***, 121–129.

Egghe, L., & Rousseau, R. (2012). The Hirsch-index of a shifted Lotka function and applications to the relation with the impact factor. *Journal of the American Society for Information Science and Technology*, ***63***, 1048–1053.

Glänzel, W. (2006). On the *h*-index—a mathematical approach to a new measure of publication activity and citation impact. *Scientometrics*, ***67***, 315–321.

Harzing, A. W. K., & van der Wal, R. (2009). A google scholar h-index for journals: An alternative metric to measure journal impact in economics & business? *Journal of the American Society for Information Science and Technology*, ***60***, 41–46.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the USA*, ***102***, 16569–16572.

Iglesias, J., & Pecharroman, C. (2007). Scaling the *h*-index for different scientific ISI fields. *Scientometrics*, ***73***, 303–320.

Ruch, S. & Ball, R. (2010). Various correlations between the H-Index and citation rate (CPP) in neuroscience and quantum physics: new findings. *International Journal of Information Science and Management*, **8(1)**, 1-19.

Schubert, A., & Glänzel, W. (2007). A systematic analysis of Hirsch-type indices for journals. *Journal of Informetrics, **1***, 179–184.

Vinkler, P. (2007). Eminence of scientists in the light of the *h*-index and other scientometric indicators. *Journal of Information Science*, **33(4)**, 481-491.

Yong, A. (2014). Critique of Hirsch's citation index: a combinatorial Fermi problem. *Notices of the American Mathematical Society*, **61(9)**, 1040-1050.

Samadianfard, Saeed, et al. "Wind speed prediction using a hybrid model of the multi-layer perceptron and whale optimization algorithm." Energy Reports 6 (2020): 1147-1159.

Taherei Ghazvinei, Pezhman, et al. "Sugarcane growth prediction based on meteorological parameters using extreme learning machine and artificial neural network." Engineering Applications of Computational Fluid Mechanics 12.1 (2018): 738-749.

Qasem, Sultan Noman, et al. "Estimating daily dew point temperature using machine learning algorithms." Water 11.3 (2019): 582.

Mosavi, Amir, and Atieh Vaezipour. "Reactive search optimization; application to multiobjective optimization problems." Applied Mathematics 3.10A (2012): 1572-1582.

Shabani, Sevda, et al. "Modeling pan evaporation using Gaussian process regression K-nearest neighbors random forest and support vector machines; comparative analysis." Atmosphere 11.1 (2020): 66.

Ghalandari, Mohammad, et al. "Aeromechanical optimization of first row compressor test stand blades using a hybrid machine learning model of genetic algorithm, artificial neural networks and design of experiments." Engineering Applications of Computational Fluid Mechanics 13.1 (2019): 892-904.

Mosavi, Amir. "Multiple criteria decision-making preprocessing using data mining tools." arXiv preprint arXiv:1004.3258 (2010).

Karballaeezadeh, Nader, et al. "Prediction of remaining service life of pavement using an optimized support vector machine (case study of Semnan–Firuzkuh road)." Engineering Applications of Computational Fluid Mechanics 13.1 (2019): 188-198.

Asadi, Esmaeil, et al. "Groundwater quality assessment for sustainable drinking and irrigation." Sustainability 12.1 (2019): 177.

Mosavi, Amir, and Abdullah Bahmani. "Energy consumption prediction using machine learning; a review." (2019).

Dineva, Adrienn, et al. "Review of soft computing models in design and control of rotating electrical machines." Energies 12.6 (2019): 1049.

Mosavi, Amir, and Timon Rabczuk. "Learning and intelligent optimization for material design innovation." In International Conference on Learning and Intelligent Optimization, pp. 358-363. Springer, Cham, 2017.

Torabi, Mehrnoosh, et al. "A hybrid machine learning approach for daily prediction of solar radiation." International Conference on Global Research and Education. Springer, Cham, 2018.

Mosavi, Amirhosein, et al. "Comprehensive review of deep reinforcement learning methods and applications in economics." Mathematics 8.10 (2020): 1640.

Ahmadi, Mohammad Hossein, et al. "Evaluation of electrical efficiency of photovoltaic thermal solar collector." Engineering Applications of Computational Fluid Mechanics 14.1 (2020): 545-565.

Ghalandari, Mohammad, et al. "Flutter speed estimation using presented differential quadrature method formulation." Engineering Applications of Computational Fluid Mechanics 13.1 (2019): 804-810.

Ijadi Maghsoodi, Abteen, et al. "Renewable energy technology selection problem using integrated h-swara-multimoora approach." Sustainability 10.12 (2018): 4481.

Mohammadzadeh S, Danial, et al. "Prediction of compression index of fine-grained soils using a gene expression programming model." Infrastructures 4.2 (2019): 26.

Sadeghzadeh, Milad, et al. "Prediction of thermo-physical properties of TiO2-Al2O3/water nanoparticles by using artificial neural network." Nanomaterials 10.4 (2020): 697.

Choubin, Bahram, et al. "Earth fissure hazard prediction using machine learning models." Environmental research 179 (2019): 108770.

Emadi, Mostafa, et al. "Predicting and mapping of soil organic carbon using machine learning algorithms in Northern Iran." Remote Sensing 12.14 (2020): 2234.

Shamshirband, Shahaboddin, et al. "Developing an ANFIS-PSO model to predict mercury emissions in combustion flue gases." Mathematics 7.10 (2019): 965.

Salcedo-Sanz, Sancho, et al. "Machine learning information fusion in Earth observation: A comprehensive review of methods, applications and data sources." Information Fusion 63 (2020): 256-272.