ENG Transactions

Journal homepage: http://www.engtransactions.com

ENG Transactions, vol. 2, pp. 1-8, May 2021
ID: ENGT-2105212112332

Research Article

Statistical Analysis of Rainfall Data (Case Study: TRNC)

Mohammadbagher Amjadi[1*], Pooria Mostafapoor[2] , Parinaz Chegin[1]

[1]*Department of Electronic Engineering, Royal Holloway University of London, London, United Kingdom*
[2]*Department of Business, The University of Texas at Arlington, Arlington, USA*

Keywords

Rainfall,
Statistical analysis,
TRNC,
Normality Test

Abstract

The rainfall data can be help to engineers to predict the events and models to decide in other engineering subject. Therefore, in this paper, the monthly rainfall data of six meteorological regions of TRNC as a whole for the hydrologic years from September 1975 to August 2014 period were gathered. In order to study these gathered monthly data's statistically, other than the minimum required sample sizes for each region, the quality tests (consistency, normality, stationarity, and trend) were as well carried out based on parametric and non-parametric tests. To determine the most representative probability distribution function for each region, Normal distributions were used. The result show that Normality, Homogeneity, Consistency, Trend analysis, and Stationarity tests were used as quality test for each meteorological region and TRNC as a whole and found that they are all within the acceptable range.

1. Introduction

It is a known fact that, many quantities encountered in all phases of life are treated as random variables in statistical sense. Theoretically, there should be a scientific explanation as to the occurance of every sensible being, so the physical and the engineering quantities should be mathematically formulated. Because of the three dimensional complexity and time necessarily being the fourth dimension of some phenomena, however, even the most developed organizations or individuals of exceptional dexterity are unable to mathematically depict some events such as hurricanes, many meteorological incidents, and severe earthquakes [1, 2]. For example, aside from snow melt, everybody knows that, an intense rainfall exeeding the infiltration capacity of a particular area causes direct overland flow ultimately results in a flood.

The physical mechanism of direct runoff beginning from a thin sheet flow [3,4], passing through the rest of the drainage paths, and finally continuing its travel in a mix qualitatively explainable. There are qualitative models that accounts these

* Corresponding Author: Mohammadbagher Amjadi
E-mail address: mohammadbagher.amjadi.2020@live.rhul.ac.uk; mb.amjadi88@gmail.com

Received: 16 March 2021; Revised: 18 April 2021; Accepted: 27 April 2021

Please cite this article as: M. Amjadi, P. Mostafapor, P. Chegin, "Statistical analysis of rainfall data (case study:TRNC)", *ENG Transactions*, vol 2, pp. 1-8, 2021.

complicated mechanisms with respectable accuracy through appropriate computer programs and packages but the unpredictability of the meteorological events however, brings about a serios difficulty for realistic calculations of the magnitude and spatial and temporal variation of the hydro-meteorological (i.e. rainfall, snow, evaporation, etc.) input, in the first place. Most of the case study problems in engineering dealt with these uncertainties [5]. Even the conditions of similar cases look common and similar, their effects may be different [6]. This is mainly due to the randomness characteristic that involves during the occurrence of the natural (real case) problems and the inappropriateness of the suggested model as well as the gathered data that is used to express this occurred phenomenon mathematically. Naturally, mankind will keep up the endeavour of making accurate meteorological forecasts for longer periods in the coming future. In this paper, by use of various statistical analysis methods, trend in the rainfall data will be analyzed.

2. Methodology

2.1. Study Area

Cyprus is an island, being located in the north-eastern part of the Mediterranean Sea, and is the third largest island with a surface area of 9251 km^2. It is bounded by latitudes of 35^{0}45’ and 34^{0}15’ N, and by longitudes of 32^{0}15’ and 34^{0}30’ E. The island lies about 64 km south of Turkey, 97 km west of Syria and 402 km north of Egypt’s Nile Delta and 380 km south east of Greece. Islands total coastline is 782 km in length [7]. Along the north, TRNC meteorology department, with simple regional modifications along the regional boundaries and renumbering of the existing meteorologically divided map, establishes its own meteorological regions. Hence, along the geographical occupation of TRNC, there are 6 meteorologically grouped geographical regions as shown in figure 1 which are investigated in this study.

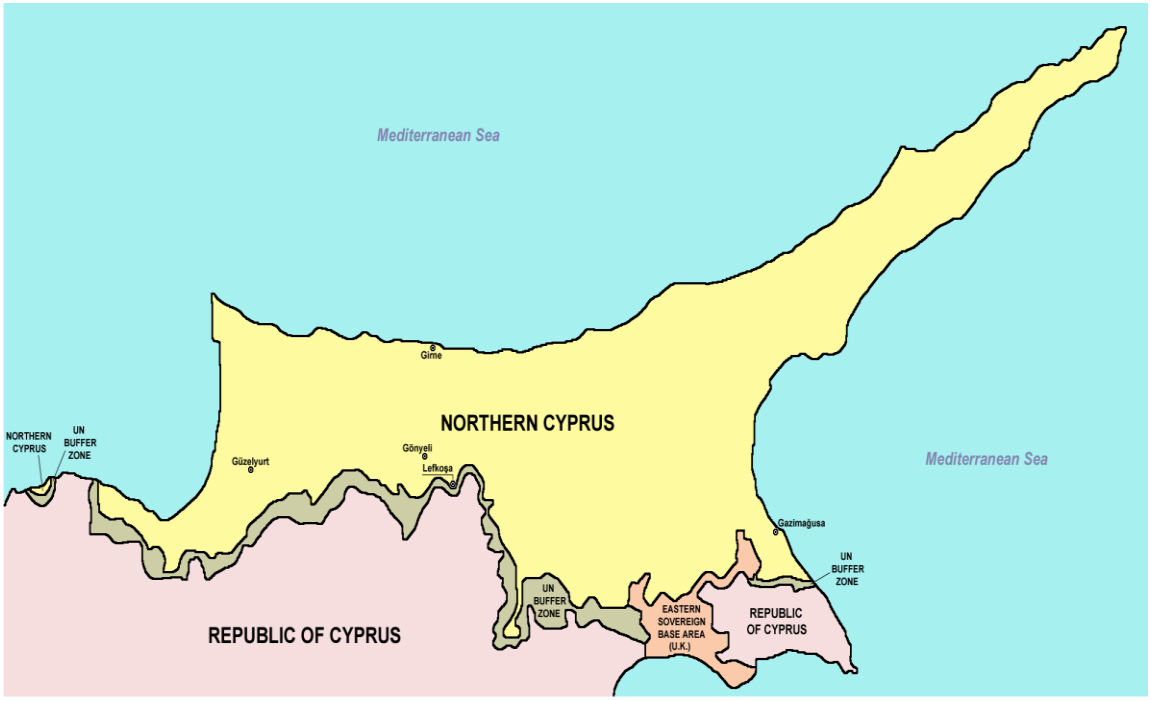

Figure 1. Geographical regions map of TRNC (obtained from Wikimedia)

2.2. Statistical analysis methods

2.2.1. Normality Test

In statistics, normality tests are used to determine if a data is well-modelled by a normal distribution and to compute how likely it is for a random variable underlying the data to be normally distributed. Hence assessment of the normality of data is a prerequisite for many statistical tests because normally distributed data is an underlying assumption in all the parametric testing. In other words, application of most of the statistical methods requires the data to behave in a Gaussian fashion [8,9].

2.2.2. Cumulative Distribution Function

This test compares the CDF (cumulative distribution function) of sample data with the distribution expected if the data were normal. If the observed difference is adequately large, it will be rejected the null hypothesis of population normality [10,11].

Because this test for each region is done by Minitab 16 software, the theory is not explained here. If the P value that is given by software will be equal or greater than 5%, then it is concluded that, the time series is normally distributed. In this paper, Anderson Darling test is selected for testing normality in Minitab.

2.2.3. Consistency Test

Consistency is another desired property for any data. It checks whether or not any data within the data is reasonable. In other words, it checks if there is a surprise data (outlier) compared with the similar family of data. For example, records for rainfall within an area might be increased in three ways: records for additional time periods; records for additional sites with a fixed area; records for extra sites obtained by extending the size of the area. In such cases, the property of consistency may be limited to one or more of the possible ways a sample size can grow [12, 13].

Double mass curve is a fundamental tool in data analysis. It is a plot of cumulative values of one variable against the accumulation of another quantities during the same time period. The theory of double mass curve is that, when accumulation of two quantities is drawn, they represent straight line. If there is a break in this continuous line, it means that there is a systematic error and it requires to be corrected. Conversely if, there is no break or change of slope within the line, it could be concluded that, the two sets of compared data are consistent. Correction of the data can be done by multiplying a constant ratio based on slopes [14].

$$P_{adjusted} = \frac{M_a}{M_0} P_{observed} \tag{1}$$

where M_a is the slope of the line before the abrupt change and M_0 is the slope of the systematic errors line.

2.2.4. Trend Test

It is a change in the level of data series, usually overtime but sometimes in space. It is a general increase or decrease in the observed values of random variable over a time. In most cases, it is not generally possible to detect trends that are not apparent by inspection, especially for data records of short to moderate length - say 20 years or less. Testing the existence of linear (monotonic) trend (serial correlation) within the whole time series is important in hydro-meteorological datas. Testing for the existence of linear (monotonic) trend within the whole time series can be done by parametric and nonparametric methods [15-19].

Mann-Kendall test is a nonparametric test that is used to find trend in time series. It was suggested by Mann (1945) and Kendall (1975). Mann-Kendall test also referred as Kendall's Tau 'τ' test. Mann-Kendall test is used to measure the connection of two sets of data. When one set of data is time then this test is used to point out the trend [20-29]. The test statistic is founded by

$$\tau = \sum_{i=2}^{n} \sum_{j=1}^{i-1} \text{sign}(X_i - X_j) \tag{2}$$

where 'τ' is approximately converging to normal distribution stated as $N_{(0,s^2)}$, if 'τ' is positive, it illustrates that the trend is increasing and if it is negative, it means the trend is decreasing. Standard deviation s_x is also described as

$$s_x = \sqrt{n(n-1)(2n+5)/18} \tag{3}$$

After obtaining the 'τ', Median slope should be obtained through Sen's method. Sen's method for the approximation of slope needs a time series of equally spaced data. Sen's method proceeds by calculating the slope as a change in measurement per change in time. The equation is given as below:

$$Q = \frac{X_j - X_i}{j - i} \tag{4}$$

3. Result and Discussion

It is essential to check if the data is consistent and lies within the data collected from neighboring regions or not. So as to check consistency of the data, double mass curve is used. Steps of applying double mass curve are as follows:

- The accumulation of the desired parameter in the studied region (station) is found.
- Then the accumulation of the average of the desired parameter over the nearby regions (station) is calculated.
- A graph is drawn of which its x-axis is cumulative average of the parameter over nearby regions and its y-axis represents the cumulative of desired parameter over the studied region.

Figure 2 shows the preceding steps are given. The studied area was Central Mesaria region and the desired parameter was rainfall. The consistency of rainfall data between Central Mesaria and average of other 5 regions in TRNC is checked.

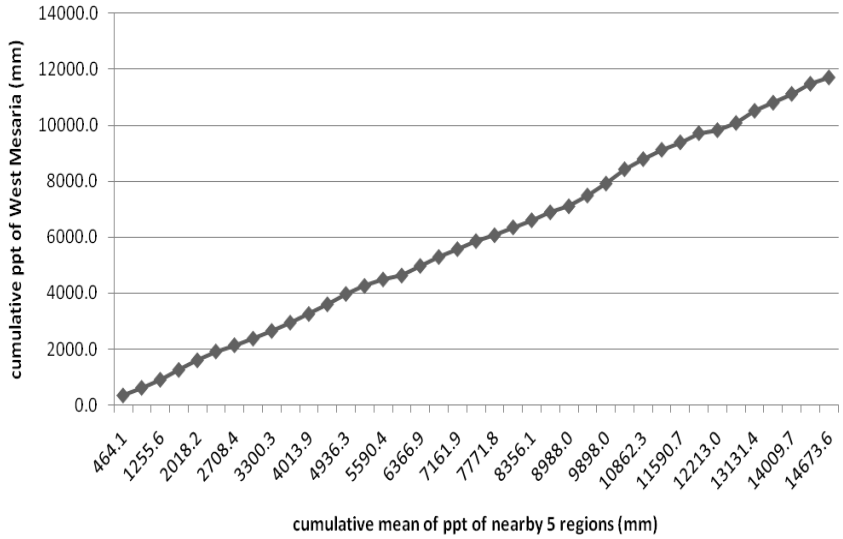

Figure 2. Central Mesaria Double Mass Curve for rainfall data with respect to nearby 5 other regions

It is found that the Central Mesaria region Rainfall is consistent with the mean Rainfall data of nearby 5 regions. In order to check if there are any trends in data Mann-Kendal test is used. This test is done by XLSTAT software. In this test, the p-value of Mann-Kendall will be computed. The test has Hypothesis test:

H_0= There is no trend in the series,

H1: There is a trend in the series,

and the confidence interval alpha is 5%, therefore if the P-value of Mann-kendall is greater than 5% then it will be concluded that there is no trend in series. Here the sample of Mann-kendall and Sens Median Slope test for Central Mesaria is illustrated. Tables 1 and 2 show the result of Mann-Kendall. Besides, mostly used formulas for CDF in hydrology are given in the table 3.

Table 1. Summary of Mann-Kendall trend tests

Variable	Observations	Obs. with missing data	Obs. without missing data	Minimum	Maximum	Mean	Std. deviation
351.8	38	0	38	107.500	510.300	298.574	76.437

Table 2. Mann-Kendall trend test

Kendall's tau	0.046
S	32.000
Var (S)	6326.000
p-value (Two-tailed)	**0.697**
alpha	0.05

As it can be seen from the test result, the P-value of Mann-Kendall trend test is 0.697 that is greater than 5%, therefore based on the null hypothesis of this paper (H0) there is no trend in the rainfall data time series of Central Mesaria. The Sens slope is 0.521.

Table 3. Mostly used CDF equations with comments

Distribution	Equations	Comments
Normal	$x = \bar{x} + zs_x$	By using table of normal distribution given in appendix, z is obtained.
Log Normal	$y = Log(x),\ y = \bar{y} + zs_y$	After finding the *Log*, using the normal distribution table z is obtained.
Extreme - Value (Gumble)	$q = (1-p) = e^{-e^{-y}}$	$y = a(x - x_0),\ a = \frac{\sigma_n}{s_x},\ x_0 = \bar{x} - y_n \frac{s_x}{\sigma_n}$ y_n ,σ_n coefficients are obtained from tables given in appendix
Log Gumble	$q = (1-p) = e^{-e^{-y}}$	$y = a(\log x - \log x_0),\ a = \frac{\sigma_n}{s_x},\ x_0 = \log \bar{x} - y_n \frac{s_{\log x}}{\sigma_n}$ y_n ,σ_n coefficients are obtained from tables given in appendix
Pearson Type III	$x = \bar{x} + Ks_x$	Referring to appropriate table given in appendix, K is obtained.
Log-Pearson Type III (Gamma)	$\log x = \overline{\log x} + Ks_{\log x}$	Referring to appropriate table given in appendix, K is obtained.

In this paper by using Minitab, 2 types of distribution model are Compared for best fitting, Normal and Log-Normal which are shown according to figure 3. Also, Equations of the probability distribution with their confidence intervals of Central Mesaria are given in the table 4.

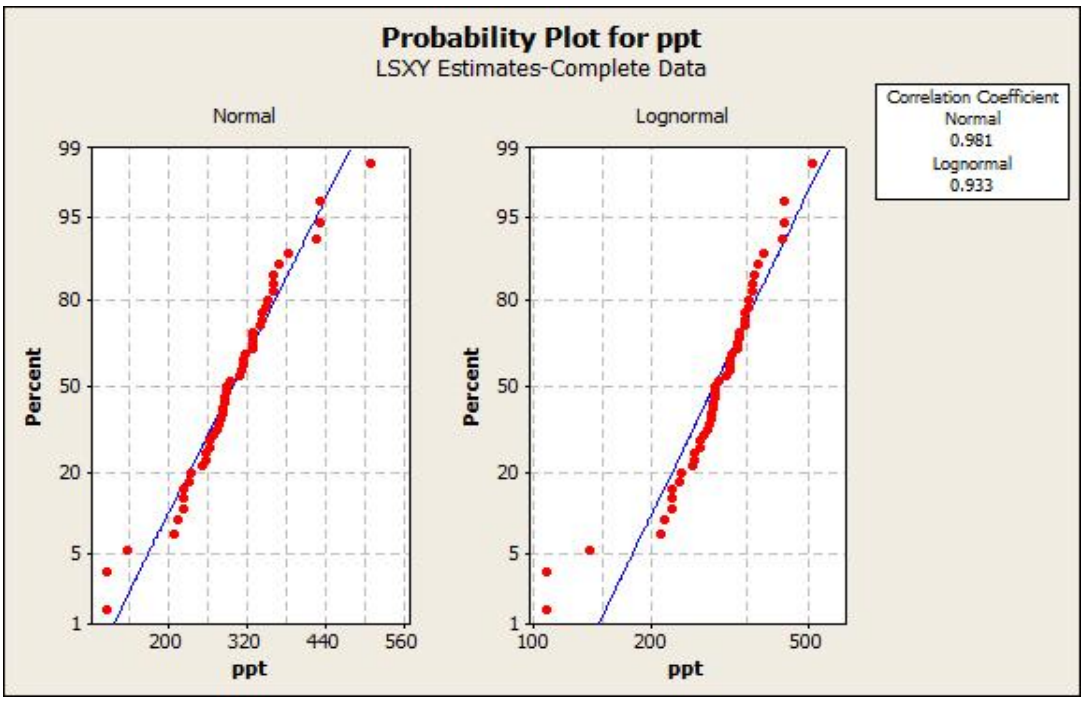

Figure 3. Probability distribution of Central Mesaria

Table 4. Equations of the probability distribution with their confidence intervals of Central Mesaria

Name	Equation	Correlation Coefficient
Normal	$x = 299.9 + 74.9\ Z$	0.981
Log Normal	$y = \log x = 2.5 + 0.1\ Z$	0.933

4. Conclusion

In this paper, by investigating of rainfall data of 6 meteorologically grouped geographical regions of TRNC, the trend in the rainfall data was analyzed. The result show that the P-value of Mann-Kendall trend test is 0.697 that is greater than 5%, therefore based on the null hypothesis of this paper (H0) there is no trend in the rainfall data time series of Central Mesaria. Also, by comparing Correlation Coefficients of all probability distribution of Central Mesaria, it is concluded that the best fitted model is Normal distribution. Normality, Homogeneity, Consistency, Trend analysis, and Stationarity tests were used as quality test for each meteorological region and TRNC as a whole and found that they are all within the acceptable range.

Conflict of interest

The authors declare no conflict of interest.

References

[1] M. B. Movahhed, J. Ayoubinejad et al., "The Effect of Rain on Pedestrians Crossing Speed", *Computational Research Progress in Applied Science & Engineering (CRPASE)*, vol. 6, no. 3, 2020.

[2] I. Bargegol and V. N. M. Gilani, "The effect of rainy weather on walking speed of pedestrians on sidewalks", *Bulletin Teknol. Tanaman*, vol. 12, pp. 217-222, 2015.

[3] V. N. M. Gilani, M. Ghasedi, et al., "Estimation delay variation and probability of occurrence of different level of services based on random variations of vehicles entering signalized intersections", In *IOP Conference Series: Materials Science and Engineering*, vol. 245, no. 4, pp. 042023, 2017

[4] I. Bargegol, et al., "Modeling pedestrian flow at central business district", *Journal UMP Social Sciences and Technology Management*, vol. 3, no. 3, 2015.

[5] M. L. Neshaei, V. N. M. Gilani, "Investigation of Cross Shore Sediment Transport Using Physical and Numerical Methods", *Journal of Applied Science*, vol. 8, pp. 795-805, 2013.

[6] P. Mobtahej, "Psychology of Change Management in Development Process within Software Industry", *Computational Research Progress in Applied Science & Engineering*, vol. 6, no. 4, 2020.

[7] F. J. Golrokh and A. Hasan, "A comparison of machine learning clustering algorithms based on the DEA optimization approach for pharmaceutical companies in developing countries," ENG Transactions, vol. 1, pp. 1–8, 2020.

[8] S. Ashraf et al., "Multibiometric Sustainable approach for human Appellative", *CRPASE Trans. Electr. Electron. Comput. Eng*, vol. *6, no.* 3, pp. 146-152, 2020.

[9] H. Behbahani, V. N. M. Gilani et al., "Analysis of Crossing Speed of the Pedestrians in Marked and Unmarked Crosswalks in the Signalized and Un-Signalized Intersections (Case Study: Rasht city)", In *IOP Conference Series: Materials Science and Engineering*, vol. 245, no. 4, pp. 042014, 2017.

[10] A. G. Mahani, P. Bazoobandi et al., "Experimental investigation and multi-objective optimization of fracture properties of asphalt mixtures containing nano-calcium carbonate", *Construction and Building Materials*, vol. 285, pp. 122876, 2021.

[11] M. Mahjoob et al., "Green Supply Chain Network Design with Emphasis on Inventory Decisions", *arXiv preprint arXiv: 2104.05924*, 2021.

[12] H. Ziari et al., "Investigation of the Effect of Crumb Rubber Powder and Warm Additives on Moisture Resistance of SMA Mixtures", *Advances in Civil Engineering*, 2021.

[13] G. Azeem, "Exploring the impacts of COVID-19 pandemic on risks faced by infrastructure projects in Pakistan", *International Journal of Applied Decision Sciences*, 2021 <hal-03213837>.

[14] A. Rahman et al., "Parkinson's Disease Detection Based on Signal Processing Algorithms and Machine Learning", *CRPASE: Transactions of Electrical, Electronic and Computer Engineering*, vol. *6*, pp. 141-145, 2020.

[15] M. Mirmozaffari, "Filtering in image processing", *ENG Transactions*, 2020 <hal-03213844>.

[16] J. N. Jerin et al., "Climate change effects on potential evapotranspiration in Bangladesh", *Arabian Journal of Geosciences*, vol. 14, no. 8, pp. 1-15, 2021.

[17] R. Yazdani et al., "VCS and CVS: New Combined Parametric and Non-parametric Operation Research Models", *Sustainable Operations and Computers*, 2021.

[18] M. Mirmozaffari, "A novel artificial intelligent approach: comparison of machine learning tools and algorithms based on optimization DEA Malmquist productivity index for eco-efficiency evaluation", *International Journal of Energy Sector Management*, 2021.

[19] M. Mirmozaffari et al., " Machine Learning Algorithms Based on an Optimization Model" 2020,

[20] M. Mirmozaffari, M. Yazdani et al., "A novel machine learning approach combined with optimization models for eco-efficiency evaluation., " *Applied Sciences*, vol. 10, no. 15, pp. 5210, 2020.

[21] M. Mahjoob et al., "A Green Multi-period Inventory Routing Problem with Pickup and Split Delivery: A Case Study in Flour Industry", *Sustainable Operations and Computers*, (2021), doi: https://doi.org/10.1016/j.susoc.2021.04.002.

[22] M. Yazdani, K. Kabirifar et al., "Improving construction and demolition waste collection service in an urban area using a simheuristic approach: A case study in Sydney, Australia", *Journal of Cleaner Production*, vol. *280*, pp. 124138, 2021.

[23] N. A. Golilarz et al., "Optimized wavelet-based satellite image de-noising with multi-population differential evolution-assisted harris hawks optimization algorithm", *IEEE Access*, vol. 8, pp. 133076-133085, 2020.

[24] M. Mirmozaffari and A. Alinezhad, "Ranking of Heart Hospitals Using cross-efficiency and two-stage DEA", *in proc. of* the *7th International Conference on Computer and Knowledge Engineering (ICCKE),* pp. 217-222, 2017.

[25] M. Mirmozaffari, " Developing an expert system for diagnosing liver diseases", *European Journal of Engineering and Technology Research*, vol. 4, no. 3, pp. 1-5, 2019.

[26] M. Mirmozaffari, "Eco-Efficiency Evaluation in Two-Stage Network Structure: Case Study: Cement Companies", *Iranian Journal of Optimization*, vol. 11, no. 2, pp. 125-135, 2019

[27] M., Zandieh, M., & Hejazi, S. M. (2017, October) , " A Cloud Theory-based Simulated Annealing for Discovering Process Model from Event Logs", *in proc. of the* 10th International Conference on Innovations in Science, Engineering, Computers and Technology (ISECT-2017), pp. 70-75, 2017

[28] A. Alinezhad, " Window analysis using two-stage DEA in heart hospitals", *in proc. of the 10th International Conference on Innovations in Science, Engineering*, Computers and Technology (ISECT-2017), pp. 44-51, 2017.

[29] F. J. Golrokh, Gohar Azeem, and A. Hasan, "Eco-efficiency evaluation in cement industries: DEA malmquist productivity index using optimization models," ENG Transactions, vol. 1, pp. 1–8, 2020.