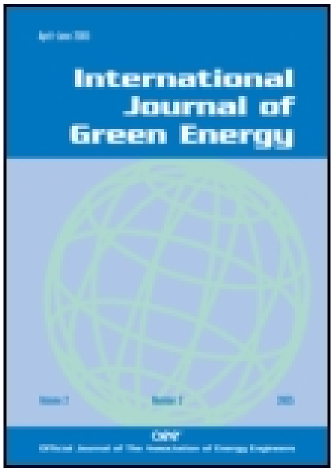

International Journal of Green Energy
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ljge20

The Importance of the Chosen Technique to Estimate Diffuse Solar Radiation by Means of Regression

Talha Arslan[a], Arzu Altin Yavuz[a] & Emin Açikkalp[b]

[a] Department of Statistics. Science and Literature Faculty. Eskisehir Osmangazi University, Eskisehir, Turkey

[b] Department of Mechanical and Manufacturing Engineering. Engineering Faculty. Bilecik S.E. University, Bilecik. Turkey

Accepted author version posted online: 03 Mar 2014.Published online: 30 Sep 2014.

International Journal of Green Energy (2015) **12**, 23–27

ISSN: 1543-5075 print / 1543-5083 online
DOI: 10.1080/15435075.2014.889010

The Importance of the Chosen Technique to Estimate Diffuse Solar Radiation by Means of Regression

TALHA ARSLAN[1], ARZU ALTIN YAVUZ[1], and EMIN AÇIKKALP[2]

[1]*Department of Statistics. Science and Literature Faculty. Eskisehir Osmangazi University, Eskisehir, Turkey*
[2]*Department of Mechanical and Manufacturing Engineering. Engineering Faculty. Bilecik S.E. University, Bilecik. Turkey*

The Ordinary Least Squares (OLS) is one of the widely used methods which is used for estimating the diffuse solar radiation. However, in order to use the OLS method in the estimation, the dataset must provide certain assumptions. In this study, alternative robust methods have been described and they were compared with the OLS method, which is used for estimating diffuse radiation frequently in an application. At the end of the analysis, the R^2 value obtained by the OLS method is less than the values obtained by M regression models. In other words, the explanation of the dependent value is weak when the OLS method is used. Finally, it can be said that the most appropriate method is Andrews for estimating the diffuse solar radiation.

Keywords: Solar energy, Ordinary least Square, Robust methods, Huber, Andrews, Tukey

Introduction

Countries need clean energy sources to maintain their development (Tarhan and Sari 2005). Clean energy sources don't harm the environment while being produced and consumed. Solar and wind energies are examples of clean energy sources. Turkey has big solar energy potential because of its location. (Gunes 2001; Aras, Balli, and Hepbasli 2006; Ulgen and Hepbasli 2009).

The efficient and effective use of sustainable energy resources depends on observation values and correction of model derived for prediction (Sozen and Arcaklioglu 2005; Aras, Balli, and Hepbasli 2006; Li et al. 2011). Incorrect solar estimation models cause inefficiency and financial loss at the solar energy facility.

In the literature, there are a lot of models for estimating diffuse solar radiation. While investigating these models, the variables under consideration are; clearness index ($K_T = H/H_0$), relative sunshine duration or sunshine fraction (S/S_0), diffuse coefficient ($K_{dd} = H_d/H_0$); and diffuse fraction or cloudiness index ($K_d = H_d/H$) (Oliveira et al. 2002; Aras, Balli, and Hepbasli 2006; Bakirci 2009; Koussa, Malek, and Haddani 2009; Pandey and Katiyar 2009; Ulgen and Hepbasli 2009; Ruis-Arias et al. 2010; Torres et al. 2010).

The Ordinary Least Squares (OLS) is one of the widely used methods which is used for estimating the diffuse solar radiation. However, in order to use the OLS method in the estimation, the dataset must provide certain assumptions. The most important among these, error terms must fit to the normal distribution. In most of the applications, the dataset does not fit to the normal distribution (Geary 1947; Scheffe 1959). In addition to that, the dataset could have values different from the rest of the data, which are called outliers. Outliers can change the result of our analysis about parameter estimation concerning distribution. If error terms don't fit to the normal distribution, OLS estimators lose their efficiency. In other words, unbiased and minimum variance characteristics of this method disappear. If error terms don't fit to normal distribution or they have outliers, the regression equation provided with the OLS method will be incorrect. Every conclusion based on this equation will be far from reality.

In the absence of normal distribution due to the existence of outliers or other reasons, using an alternative robust estimator to obtain parameter estimations will be useful in order to achieve more effective results. Also, robust estimators are not affected from outliers or deflection from the model as much as the traditional estimators (Tiku, Tan, and Balakrishnan 1986).

In this study, alternative robust methods have been described and they were compared with the OLS method, which is used for estimating diffuse radiation frequently in an application. Investigated dataset, supplied from Turkish State Meteorological Service, includes the measures for Bilecik city of Turkey, in 2000. The analysis has been conducted by using the OLS and M regression (Huber, Andrews, and Tukey) techniques. In this study, the dependent variable was chosen as the monthly average daily diffuse fraction (or the monthly average daily cloudiness (clearness) index (dimensionless) denoted by K_d) and the independent variable was chosen as the monthly average daily clearness index (dimensionless and denoted by K_t).

Address correspondence to Emin Açıkkalp, Department of Mechanical and Manufacturing Engineering, Engineering Faculty, Bilecik University, Bilecik, Turkey. E-mail: acikkalp@gmail.com

Methodology

Model Used for Estimating the Diffuse Solar Radiation

The monthly average daily extraterrestrial radiation on a horizontal surface (H_o) was obtained from the following equations (1) (Iqbal 1983; Aras, Balli, and Hepbasli 2006; Ulgen and Hepbasli 2009);

$$H_o = \frac{24}{\pi} G_{\text{on}} \left[(\cos\phi)(\cos\delta)(\sin\omega_s) + \frac{\pi}{180}\omega_s (\sin\phi)(\sin\delta) \right], \tag{1}$$

where G_{on} is the extraterrestrial radiation, computed from (2) (Helwa et al. 2000; Aras, Balli, and Hepbasli 2006)

$$G_{\text{on}} = G_{\text{sc}} \left[1 + 0.034 \cos\left(\frac{360 n_{\text{day}}}{365} \right) \right], \tag{2}$$

where G_{sc} is the solar constant (= 1367 W/m^2), ϕ is the latitude of the site, δ is the solar declination angle, ω_s is the sunshine hour angle for the month, and n_{day} is the number of the days of the year, starting from the first of January. The solar declination (δ), the main sunshine hour angle for the month (ω_s) and the maximum possible sunshine duration day length (S_o) calculated from (3) – (5) (Cooper 1969; Aras, Balli, and Hepbasli 2006);

$$\delta = 23.45 \sin\left[\frac{360}{365} \left(284 + n_{day} \right) \right], \tag{3}$$

$$\omega_s = \cos^{-1}\left[\tan(\phi) \tan(\delta) \right], \tag{4}$$

$$S_o = \frac{2}{15}\omega_s. \tag{5}$$

By using the diffuse fraction or cloudiness index, the function of the clearness index is given in (6) (Ulgen and Hepbasli 2009)

$$\left(K_d = \frac{H_d}{H} \right) \cong f \left(K_t = \frac{H}{H_{\text{o}}} \right). \tag{6}$$

Similarly, a function of the relative sunshine duration or sunshine fraction is (Ulgen and Hepbasli 2009);

$$\left(K_d = \frac{H_d}{H} \right) \cong f \left(\frac{S}{S_o} \right). \tag{7}$$

The function of the clearness index including the diffuse coefficient is given in (8) (Ulgen and Hepbasli 2009);

$$\left(K_{dd} = \frac{H_d}{H_o} \right) \cong f \left(K_t = \frac{H}{H_o} \right). \tag{8}$$

From the diffuse coefficient, the function of the relative sunshine duration or sunshine fraction is as in (9) (Ulgen and Hepbasli 2009)

$$\left(K_{dd} = \frac{H_d}{H_{\text{o}}} \right) \cong f \left(\frac{S}{S_o} \right). \tag{9}$$

In the preceding equations, H_{o} is the monthly average daily extraterrestrial radiation (MJ/m^2), H is the monthly average daily global solar radiation (MJ/m^2), H_d is the monthly average daily diffuse solar radiation (MJ/m^2), S is the monthly average daily measured sunshine duration (h), and S_{o} is the monthly average daily maximum possible sunshine duration (h) (Ulgen and Hepbasli 2009).

Statistical Analysis Method

Simple regression analysis can be used in defining cause–effect relation among variables. In regression analysis, a mathematical model is derived by using variables. The obtained model contains two types of variables: dependent and independent. Independent variables are not affected by any another variables in the model. However, dependent variables are generally affected from the other variables in the model. Dependent variable values can be predicted from the values of the independent variables.

In the derived model, correct estimation of parameters has the most important role in detemining consistency of the estimated models. There are different techniques to estimate parameters. The most frequently used technique in regression analysis is the OLS method (Johnson and Wichern 2002; Deniz, Atik, and Bugutekin 2006). The OLS method is valid under certain assumptions. One of the most important assumptions is that the error terms should have **0** mean and $\boldsymbol{\sigma^2 I}$ variance with normal distribution. In the case when the normal distribution assumption for error terms is not met, the efficiency of the estimation rapidly decreases on using the OLS method and such consequences do not reflect the reality. In addition to this, if the dataset has outlier(s), OLS estimation provides weak results. In contrast to OLS estimators, robust estimators are not affected from the normality assumption of error terms and existence of outlier in the dataset.

In this study, our dataset includes outliers and the distribution of error terms does not fit to the normal. We compared the efficiencies of some robust estimators (Huber, Andrews, and Tukey regression methods) with OLS estimators. In the following sections, OLS, Huber, Andrews, and Tukey methods are provided.

Ordinary least square (OLS)

In this study, there is one dependent and one independent variable and the simple linear regression equation is defined as given in (10)

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, 2, ..., n. \tag{10}$$

The estimated regression equations given in (11) are

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, i = 1, 2, ..., n. \tag{11}$$

Error terms are calculated as in the following:

$$\varepsilon_i = y_i - \hat{y}_i, i = 1, 2, ..., n. \tag{12}$$

The OLS equation is obtained as given in (13) and parameter estimation as given in (14). (David 1981; Mutan 2004; Ergul 2006)

$$n\hat{\beta}_o + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \quad \hat{\beta}_o \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} {x_i}^2 = \sum_{i=1}^{n} y_i x_i, \tag{13}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \quad \hat{\beta}_o = \bar{y} - \hat{\beta}_1 \bar{x}. \tag{14}$$

M-*regression*

M-Regression was developed for cases if error terms don't fit to normal distribution. The purpose of the OLS method is to minimize the sum of error terms. In OLS, weights of error terms are assigned as1. The M-Regression technique minimizes error terms function given in (15) instead of the sum of error terms (Tasker and Granato 2000)

$$\sum_{i-1}^{n} \rho \left[\left(\frac{y_i - \sum_{j=i}^{k} x_{ij}\beta_j}{d} \right) \right], \tag{15}$$

where ρ is the objective, differentiable, and symmetrical function, with a constant single minimum at zero, d is a Median Absolute Deviation (MAD). Using MAD, standardized error terms can be calculated as in (16) and where 0.675 is constant, they are used to obtain a robust estimate of the standard deviation of Gaussian (Normal) distribution

$$MAD = \frac{\text{median}\{|\varepsilon_i - \text{median}(\varepsilon_i)|\}}{0.675}. \tag{16}$$

Various functions have been proposed for the objective. In this study, we consider Huber, Andrews, and Tukey functions. These functions are most widely used in application. Huber's weight function can be defined as in (17) (Huber 1973), where $k = 1.345$ and

$$w = \begin{cases} 1, & |r| \le k, \\ \left(\frac{k}{r}\right), & |r| > k. \end{cases} \tag{17}$$

The weighted function proposed by Andrews is given in (18) (Andrews 1974)

$$w = \begin{cases} \frac{\sin\left(\frac{k}{r}\right)}{r}, & |r| \le k\pi, \\ 0, & |r| > k\pi, \end{cases} \tag{18}$$

where k values are generally selected as $k = 1.5$ or $k = 2.1$ (Şanlı 2005).

The bisquare weight function proposed by Tukey is defined as in (19) (Tukey 1970)

$$w = \begin{cases} \left[1 - \left(\frac{k}{r}\right)^2\right]^2, & |r| \le k, \\ 0, & |r| > k, \end{cases} \tag{19}$$

where k values are generally selected as $k = 4.685$ (Fox 2002). In Huber, Andrew, and Tukey M -Regression techniques, the procedure to be followed in estimation of parameters are;

(i) β_0 and β_1 values are estimated with OLS.
(ii) Using these estimations ε_i values are calculated.
(iii) MAD values are calculated.
(iv) Weight values are calculated.
(v) $\hat{\beta}_{00}$ and $\hat{\beta}_{10}$ values are calculated with weighted OLS.
(vi) $\hat{\beta}_{00}$ and $\hat{\beta}_{10}$ with $\hat{\beta}_0$ and $\hat{\beta}_1$ values are compared.
(vii) If differences among estimations are less than 0.001, procedure can be finished.
(viii) If differences among estimations are greater than 0.001, $\hat{\beta}_{00}$ and $\hat{\beta}_{10}$ values are assigned as $\hat{\beta}_0$ and $\hat{\beta}_1$ values and return to step 2.

Determination of the Model Fit

There are different types of methods in statistics to determine whether observation values fit to the obtained regression model. In this study, model fit is examined by using the determination coefficient (R^2). If the R^2 value is close to 1, then the obtained model is fit with the observation values. In other words, if R^2 values are close to 1, then the sum of the square error is close to 0. In addition, R^2 shows how the total value of a dependent variable can be explained by independent variables (Gujarati 2004; Ergul 2006). The determination coefficient can be calculated as in (20)

$$R^2 = \frac{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2 - \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2}. \tag{20}$$

Results and Discussion

Before starting analysis, we have to check the normality of residuals and outlier(s). The Q–Q plot which is used for the normality test for residuals is given in Figure 1. From Figure 1, one can see that the dataset does not fit to the normal distribution, as it includes outliers. Outliers were marked with the circle in the figure. Therefore, the normality assumption in the method is not satisfied. That's why, it's better to use the robust techniques instead of the OLS technique.

The estimated regression equation is given in (21) as

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, i = 1, 2, ..., n. \tag{21}$$

As a result of the analysis, the estimated OLS and M regression (Huber, Andrews, and Tukey) equations are given in (22), (23), (24), and (25), respectively

$$\widehat{K_d} = 5.67 - 8.9566(\widehat{K_t}), \tag{22}$$

$$\widehat{K_d} = 4.2857 - 6.2800(\widehat{K_t}), \tag{23}$$

$$\widehat{K_d} = 3.1163 - 4.1131(\widehat{K_t}), \tag{24}$$

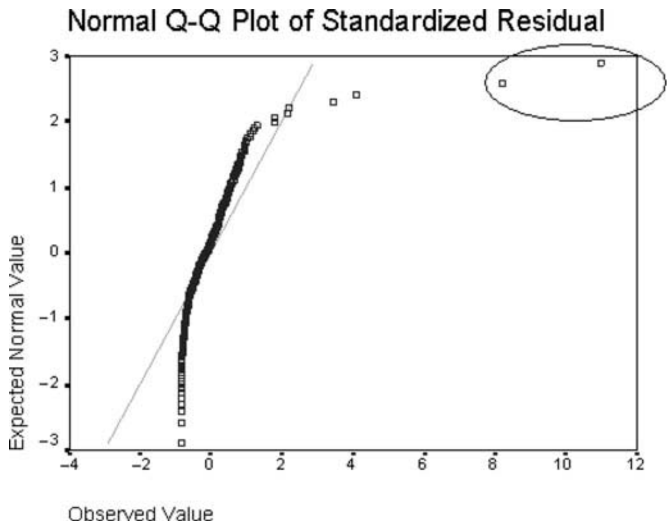

Fig.1. *Q–Q* plot for standardized residual.

Table 1. Regression coefficients for estimating the diffuse solar radiation

Techniques	$\hat{\beta}_0$	$\hat{\beta}_1$	R^2
OLS	5.6700	−8.9566	0.6907
Huber	4.2857	−6.2800	0.8920
Andrews	3.1163	−4.1131	0.9222
Tukey	3.0980	−4.0803	0.9090

$$\widehat{K_d} = 3.0980 - 4.0803(\widehat{K_t}). \tag{25}$$

For determining distribution fitting (normal distribution) about residual, the Kolmogorov–Smirnov (K-S) test was used. The K-S test statistics was found to be 0.1417 ($p < 0.0001$). Figure 1 and K–S test showed that error terms did not fit normal distribution.

The R^2 values and parameter estimations are given in Table 1. In this table, the R^2 value obtained by the OLS method is less than the values obtained by M-regression models. In other words, the explanation of the dependent value is weak when the OLS method is used. Also, the equations obtained by robust regression techniques have smoother explanation values. As referred, the usage of more independent variables can generate more certain results.

Nevertheless, it can be seen that robust methods supply quietly certain results even if they have only one explanation variable. Finally, it can be said that the most appropriate method is Andrews for estimating the diffuse solar radiation.

Conclusions

Solar energy is one of the most important alternative energy sources. For designing any solar energy device, solar energy parameters and components have an important role. Solar energy offers us clean and sustainable energy for the future. Due to energy demand of the world these days, correct predictions using solar energy models have assumed importance. It was aimed to supply correct model calculation in this study for using solar energy truly.

Acknowledgment

The authors acknowledge the considerable help of the Turkish Meteorological Office for access to archives.

Nomenclature

G_{on}	extraterrestrial radiation (W/m^2)
G_{sc}	solar constant (W/m^2)
H	monthly average daily global solar radiation (MJ/m^2)
H_d	monthly average daily diffuse solar radiation (MJ/m^2)
H_{o}	monthly average daily extraterrestrial radiation (MJ/m^2)
K_d	($= H_d/H$) monthly average daily diffuse fraction or monthly average daily cloudless (clearness) index (dimensionless)
K_{dd}	($= H/H_{\text{o}}$) monthly average daily clearness index (dimensionless)
K_t	($= H/H_{\text{o}}$) monthly average daily clearness index (dimensionless)
n_{day}	number of day of the year, starting from the first of January (dimensionless)
S	monthly average daily measured sunshine duration (h)
S_0	monthly average daily maximum possible sunshine duration (h)
n	number of observations,
y_i	observed value about dependent variable
$\hat{y}_i$	predicted value about dependent variable
x_i	observed value about independent variable
β_0	intercept coefficient about population
$\hat{\beta}_0$	intercept coefficient about sample
β_1	regression coefficient about population
$\hat{\beta}_0$	regression coefficient about sample
ε_i	residual value
R^2	Coefficent of determination

Greek Letters

ϕ latitude of site (°)
δ solar declination angle (°)

References

Andrews, D. 1974. A robust method for linear regression. *Technometrics* 16:523–31.

Aras, H., O. Balli, and A. Hepbasli. 2006. Estimating the horizontal diffuse solar radiation over the Central Anatolia Region of Turkey. *Energy Conversion and Management* 47:2240–49.

Bakirci, K. 2009. Correlations for estimation of daily global solar radiation with hours of bright sunshine in Turkey. *Energy* 34: 485–501.

Cooper, P.I. 1969. The absorption of solar radiation in solar stills. *Solar Energy* 12:333–46.

David, A. 1981. *Introduction to statistics*. St. Paul: West Publishing Company.

Deniz, E., K. Atik, and A. Bugutekin. 2006. Bazı Meteorolojik verilerin regresyon analizi metodu kullanılarak tahmin edilmesi. *Teknoloji* 9:271–76.

Ergul, B. 2006. Robust regression and applications. Master of Science Thesis, Department of Statistics, Eskisehir Osmangazi University, Eskisehir, Turkey.

Fox, J. 2002. *Robust regression. Appendix to an R and S-PLUS companion to applied regression*. S-PLUS Software.

Geary, R.C. 1947. Testing for normality. *Biometrika* 34:209–20.

Gunes, M. 2001. Analysis of daily total horizontal solar measurements in Turkey. *Energy Source* 23:563–70.

Gujarati, N. 2004. *Basic econometrics*. Translated by; Ü. Şenesen, İstanbul: Literatür Publishing.

Helwa, N.H., A.B.G. Baghdad, A.M.R. El Shafee, and E.T. El Shenawy. 2000. Computation of the solar energy captured by different solar tracking systems. *Energy Source* 22:35–44.

Huber, P. 1973. Robust regression: Asymptoties, conjections and Monte Carlo. *Annals of Statistics* 1:799–821.

Iqbal, M. 1983. *An introduction to solar radiation*. New York, USA: Academic Press.

Johnson, R.A., and D.W. Wichern. 2002. *Applied multivariate statistical analysis*. New Jersey: Pearson Education International.

Koussa, M., A. Malek, and M. Haddani. 2009. Statistical comparison of monthly mean hourly and daily diffuse and global solar irradiation models and a simulink program development for various Algerian climates. *Energy Conversion and Management* 50:1227–35.

Li, H., W. Ma, X. Wang, and Y. Lian. 2011. Estimating monthly average daily diffuse solar radiation with multiple predictors: A case study. *Renewable Energy* 36:1944–48.

Mutan, C.O. 2004. A comparison of regression techniques via Monte Carlo simulation, Master of Science Thesis. Middle East Technical University, Ankara, Turkey.

Oliveira, A.P., J.F. Escobedo, A.J. Machado, and J. Soares. 2002. Correlation models of diffuse solar-radiation applied to the city of Sao Paulo. *Brazil, Applied Energy* 71:59–73.

Pandey, C.K., and A.K. Katiyar. 2009. A comparative study to estimate daily diffuse solar radiation over India. *Energy* 34:1792–96.

Ruiz-Arias, J.A., H. Alsamamra, T. Pescador, and D. Pozo-Vazquez. 2010. Proposal of regressive model for the hourly diffuse solar radiation under all sky conditions. *Energy Conversion and Management* 51:881–93.

Scheffe, H. 1959. *The Analysis of variance*. New York: John Wiley.

Sozen, A., and E. Arcaklioglu. 2005. Solar potential in Turkey. *Applied Energy* 80:35–45.

Şanlı, K. 2005. Bulanık robust regresyon çözümlemesi. Doktora Tezi, Ankara Üniversitesi Ankara, Turkey.

Tarhan, S., and A. Sari. 2005. Model selection for global and diffuse radiation over the Central Black Sea (CBS) region of Turkey. *Energy Conservation and Management* 46:605–13.

Tasker, G., and G. Granato. 2000. Statistical approaches to interpretation of local, regional and national highway -Runoff and urban- stormwater data. *U.S. Geological Survey*:1–47.

Tiku, M.L., W.Y. Tan, and N. Balakrishnan. 1986. *Robust inference*. New York: Marcel Dekker.

Tukey, J. 1970. *Exploratory data analysis*. Boston, MA: Addison-Wesley.

Torres, J.L., M. De Blas, A. Garcia, and A. de Francisco. 2010. Comparative study of various models in estimating hourly diffuse solar irradiance. *Renewable Energy* 35:1325–32.

Ulgen, K., and A. Hepbasli. 2009. Diffuse solar radiation estimation models for Turkey's big cities. *Energy Conversion and Management* 50:149–56.