Regression Estimation Modelling Techniques on Static Solar Photovoltaic Module

Emmanuel B. Balogun¹, Xu Huang², Dat Tran³, Yun-Chuan Lin⁴, Mingyu Liao⁵, Michael F. Adaramola⁶

¹,²,³ Faculty of Education Science Technology and Mathematics University of Canberra, Australia, ACT 2617.
⁴,⁵ Graduate Institute of Applied Science, National Kaohsiung University of Applied Sciences, Kaohsiung 80778, Taiwan, R.O.C.
⁶ Dept. of Electrical-Electronics Eng. Lagos State Polytechnic, Ikorodu, Lagos, Nigeria

Abstract - This paper aims to evaluate three different regression modelling techniques using the sparse based regression estimation algorithm on the static solar photovoltaic module (SSPVM). The three regression models were used to predict on a very large scale the most suitable model based on manufacturer solar photovoltaic parameters in determining its performance, efficiency and reliability. The calibration coefficients of the regression models used for evaluating are the mean square errors (MSE), root mean square errors (RMSE) and standard error of the estimate in determining the maximum power point tracking for current, voltage and power respectively. The Matlab fitting tool is used for performance evaluation using the sparse based regression algorithm. The parameters evaluated in this comparative analysis were obtained from the general non-linear mathematical equation of the solar photovoltaic module. Computational comparison with simulated solar photovoltaic manufacturer’s data on the standard error of the estimate shows that these regression methods are extremely promising.

Keywords-- Correlation coefficients, robust fit regression, lasso, static solar photovoltaic module; maximum power points, ordinary least squares regression, least trimmed square regression model

I. INTRODUCTION

With recent advancement in the renewable energy industry and interest focus on the solar photovoltaic energy industry, optimisation and design of high quality, efficient and long lasting solar photovoltaic module panels are top priority of various solar photovoltaic manufacturers. For this reason the regression based modelling techniques of the static solar photovoltaic module have become a very important tool. There has been growing interest in the research field in design, application and simulation processes in ensuring a high performance, superlative output consistency, meticulous tolerance and quality checks. This result in excellent quality production of solar photovoltaic modules [1-2]. This branch of the renewable energy industry provides clean, noise pollution free power generation, an energy security assurance guarantee and is environmentally friendly [3].

The static solar photovoltaic module installation on a large hectare of land space usually referred to as solar farms has low maintenance and the operational costs are also minimal due to absence of rotating parts in its design [4]. There have been recurrent challenges faced by the solar photovoltaic manufacturing industry in ensuring optimal choice in the selection of materials and design, improving the overall maximum output efficiency of the solar photovoltaic module [5]. These challenges are resolved by using accurate modelling algorithms to predict the performance. This reduces the error measures and margin norms. The regression approach has continued to receive a high level of interest in various research fields. It has resulted in the development of various regression models for time series, growth detection, image and pattern recognition, determination of missing data, cloud computing, making predictions and casual inferences with regression models [6].

To date, relatively few papers have addressed the issues of regression errors in solar photovoltaic modules. Although not common in recent literature in the solar research field, the sparse based algorithm can be adopted in the determination of maximum power point (MPP) errors for the simulated current, voltage and power parameters of the solar photovoltaic module respectively compared to the expected maximum specification designed output [7-8]. Simulation and modelling of a solar photovoltaic module have been presented by several authors using the general mathematical expression to predict the I-V characteristics, determination of the maximum power point for the current, voltage and power curves under varying environmental conditions (such as temperature, wind speed and solar irradiance), using the manufacturer parameter model data sheet. The general non-linear characteristics equation adopted for a solar photovoltaic module is represented based on a single diode circuit having one end connected across by a single current source and its other end with two resistors connected in series and parallel respectively.
Other related solar photovoltaic equations are used in the regression analysis for predicting the maximum power points for the current, voltage and power distribution variables plotted on the vertical axis against the solar irradiance on the horizontal level.

The regression based approach under review in this paper for the static solar photovoltaic module is the ordinary least squares regression, robust fit regression and the least trimmed squares regression model. The result of these models are compared and validated based on the non-linear characteristics equation of the solar photovoltaic module. The regression model helps to predict the reliability, efficiency and performance compared with the solar photovoltaic manufacturer’s parameter.

1.1. Static Solar Photovoltaic Model

The static solar photovoltaic array configuration involves the connection of series and parallel combination of solar photovoltaic modules to form an array that gives the desired I-V characteristics and electrical output energy. The solar photovoltaic energy world has been experiencing tremendous growth over the last decade in its awareness, research and improving the efficiency on the quality of absorption of the solar photovoltaic module. There have also been various establishments of large hectares of solar farms in many parts of the world. The solar energy industry has been regarded as one of the most eco-friendly renewable energy sources, noise pollution-free and reliable means of power generation [26-27]. Figure 1 depicts the aerial view of a solar photovoltaic array on a hectare of land space for power generation.

![Figure 1. Solar photovoltaic array on a hectare of land](image)

1.2 Description of the Solar Photovoltaic Model

A solar photovoltaic model is represented by an equivalent circuit model analogous to a four component parameter model having a current source, a diode, a parallel and series resistor connected together. The photovoltaic model relationship can be best described by a mathematical expression giving the relation between the voltage and current at its terminals.

![Figure 2. The equivalent circuit model of a solar photovoltaic model](image)

The characteristics mathematical equation is non-linear and therefore an analytical solution is implicitly impracticable [28-29]. Figure 2 depicts the equivalent circuit model of a solar photovoltaic model.

The output characteristics of the solar photovoltaic model essentially depend on these three significant factors: solar radiation, $S$, wind speed, $w$, and temperature, $T$. The temperature of solar photovoltaic model is a determinant factor and constantly varies depending on the other two parameters. The mathematical expression showing the relationship is given as:

$$T = 3.12 + 0.25S + 0.899$$

The leakage or reverse saturation current, $I_o$ in the above equation (1) is a measure of recombination in the solar photovoltaic module, and thus, inversely related to the quality of the material and changes in temperature of the solar photovoltaic module according to the equation given in (3) as [33]:

$$I_o = I_{or} \left( \frac{V}{V_T} \right)^3 \cdot e^{\left( \frac{qE_g}{kT} \right)}$$

The photocurrent generator, $I_L$ in equation (1) is a function of the incident solar radiation and cell temperature and is given as:

$$I_L = (I_{scr} + k(T - T_e)) \frac{S}{T_{max}}$$

$V_{oc}$, open circuit voltage of the photovoltaic module is given as [13]:

$$V_{oc} = \frac{A_{KT}}{q} \ln \left( \frac{I_L}{I_{max}} \right) = aV_T \ln \left( \frac{I_L}{I_{max}} \right)$$

$V_{mp}$, maximum power voltage is given as

$$V_{mp} \approx V_{oc} \left( 1 - \frac{ln(c)}{c} \right)$$
The output power, \( P \) of a solar photovoltaic array is given as:

\[
P = n_p I_{ph} V - n_p I_0 [V/n_s + IR_s] - (V/n_s - IR_s) V/R_{sh}
\]  

where \( I_0 = q/AkT \).  

The rest of the paper is organised as follows: Section 2 focuses on the three regression model methods under consideration for validating the static solar photovoltaic module. In Section 3, experiments and results are described and Section 4 presents data analysis and validation for the three chosen models. Finally, the conclusion is presented in Section 5.

II. METHODS

2.1. Regression Concept and Theory

Regression is one of the most commonly used statistical techniques applied to various research fields. It is useful for analysing, predicting, forecasting, investigating, hypothesis testing and modelling the relationship between a scalar dependent variable and an independent variable [9-10]. Regression analysis tries to fit a model to one dependent variable based on one or more independent variables. One of the most common regression models is the simple linear regression model. Simple linear regression is a statistical method that provides a summary between two quantitative variables. It assumes a linear relationship between an independent variable (predictor) and a dependent variable (response). It fits the straight line through the set of points in such a way making the sum of the squared residuals of the model have the smallest possible error. The best fitting line is called the regression line. The popular procedure for determining the best fit is called the least square method. Multi-linear regression exhibits linear equation characteristics with two or more predictor variables. Non-linear regression depends on multiple independent variables modelled by non-linear functional combination model parameters. A good regression model should have a high regression standard error estimate, \( R^2 \) value fairly close to unity, in order to predict the best fitting line for the data and overall summary of the model performance [10]. The plotting of residuals is a tool used in checking the adequacies of the fitted multiple linear regressions based on the determined standard error of the estimate. The data used for analysis should cover the entire range of response value for analysis and prediction.

The simple linear regressions are classified into three groups namely linear, quadratic and cubic [11]:

- Linear: \( Y = a_0 + a_1 x \)
- Quadratic: \( Y = a_0 + a_1 x_1 + a_2 x_2 \)  
  \[ (13) \]
- Cubic: \( Y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 \)
The linear parameters \( a_0 \) and \( a_1 \) represent the y-intercept and the slope of the relationship, respectively. The multi-linear regression model performs linear and nonlinear regressions with two or more independent variables. It is classified into two groups namely, multiple and involving interactions [12-13]:

- **Multiple:** \( Y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 \)
- **Involving interactions:** \( Y = a_0 + a_1x_1 + a_2x_2 + a_3x_1x_2 \)

(14)

The non-linear regression model can assume any type of relationship between the dependent variable and independent variables. This function uses the least squares approach for computing the regression. It is classified into two groups namely sine wave and exponential [14]:

- **Sine Waves:** \( Y = a_0 + \sin(a_1x_1) \)
- **Exponential:** \( Y = a_0 + e^{(a_1x_1)} \)

(15)

2.2. **Regression-type Model**

The regression algorithm model considered in the estimation technique of the static solar photovoltaic module is the sparse based regression algorithm. The sparse based algorithm has established itself among the state of the art in supervised learning. It learns classifiers constructed as weighted linear combinations of basis functions. The weights are estimated in the presence of training data to be either significantly large or exactly zero which automatically removes irrelevant basis functions. The sparse based regression algorithm is embedded into our previous Levenberg Marquardt algorithm in the neural network training [34]. The primary emphasis of the introduction of the sparse based algorithm was based on its flexibility and performance efficiency. The sparse based algorithm incorporates an iterative linear system solver for minimizing the number of iterations for the model.

Considering a data matrix \( P \) and a column vector \( k \), the optimisation problem is solved by obtaining a weight of vector \( w \) of variables for a sparse based algorithm.

\[
\min \| w \|_1, \text{ s.t. } P w = k
\]

(16)

Where \( k \in R^N \) is a given signal vector and \( P \in R^{N \times M} \) (N < M) is a basis matrix.

The objective of the sparse based algorithm is to find a solution \( w \in R^M \) for eq. (16).

Where \( 1 \)-norm \( \| w \|_1 \) is defined as \( \sum_{i=1}^{M} |w_i| \)

The assumption for the sparse based algorithm is set for the non-negatives as

\[
w = u - v, \text{ where } u, v \in R^M \text{ are non-negatives}
\]

The non-negative expressions can be converted to equivalent linear solvable programming problems using the Matlab optimisation toolbox.

\[
\min \sum_{i=1}^{K} (u_i + v_i), \text{ s.t. } [P - P][u^T, v^T]^T = k, \ u, v \geq 0 \]

(17)

The sparse weight identifies the relevant variables or the supposed regression vectors. The sparse based algorithm can be categorised into two sets: one time series correlated with the output and secondly time series not significantly correlated with output. There are quite a few parameter estimators developed in linear regression differing in the simplicity of the algorithm, robustness in heavy tailed distributions, theoretical assumption and close form solutions to validate asymptotic efficiency and consistency [15-17]. A brief review of the three regression estimation techniques used in the static solar photovoltaic module by the sparse based algorithm are described in the sections below.

2.3. **Ordinary Least Squares Regression**

The ordinary least squares (OLS) is also referred to as linear least squares. This regression estimator method is used in analyzing both experimental and observational data and in the estimation of unknown parameters in a linear regression model. The OLS minimizes the sum of squared residuals between the observed and predicted responses from the dataset by linear approximation. In the OLS, several assumptions are made to validate the predicted and the estimated model. The residuals, known as the difference between the predicted and real data, when largely distorted are referred to as outliers. These outliers influence the error variance, the standard errors and the final estimation becomes asymptotically inconsistent. A common basic assumption made is the normality of the residual along the parallel line. Another critical assumption made is the normality of the residuals making the estimation of significance impaired. In the OLS regression model, it is primarily assumed that there is zero or negligible errors in the independent variables. The sum squared errors is a measure of the overall model fit given by [18-19]:

\[
S(e) = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = (y - \hat{y})^T(y - \hat{y})
\]

(18)

Where \( y_i - \hat{y}_i \) is called the residual for the \( i \)-th observation and the hyper-plane \( y = \hat{y} \) measuring the vertical distance between the data point \((x_i,y_i)\). \( T \) in the above equation denotes the matrix transpose. The value of \( e \) minimises this sum called the OLS estimator for \( \beta \).
The expression for the estimated unknown parameter $\beta$:

$$\beta = (X^TX)^{-1}X^Ty$$

Or equivalently the explicit expression for the estimated unknown parameter $\beta$ is given by:

$$\beta = \left( \sum_{i=1}^{n} x_i x_i^T \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} x_i y_i$$

(20)

The predicted value from the regression is given by the expression:

$$\hat{y} = X\hat{\beta} = Ay$$

(21)

Where $A = (X^TX)^{-1}X^T$ is the projection matrix. The residual from the regression is given by:

$$e = y - X\beta$$

(22)

A simple linear regression model containing two variables, a constant and a scalar regressor, $X$ is given in equation (23). The vector of parameters in the model is a two-dimensional and denoted by $\beta_0, \beta_1$ respectively;

$$Y = \beta_0 X + \beta_1$$

(23)

$$\beta_1 = \frac{\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{N \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

(24)

$$\beta_0 = \frac{N \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{N \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

(25)

The correlation coefficient, $r$ and the standard error, $s$ for the simple linear regression respectively given by:

$$r = \frac{N \sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sqrt{[N \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2][N \sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2]}}$$

(26)

$$s = \frac{\sum_{i=1}^{n} y_i^2 - \beta_1 \sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i}{N}$$

(27)

2.4. Robust Fit Regression

The robust fit regression is an advanced statistical technique developed to overcome the influence of outliers. It enhances resistant in the presence of outliers achieving the expected stability. The robust fit regression approach employs a fitting criterion that is not vulnerable as the least squares to unusual data. The robust regression is basically resolved to address these three key challenges:

I. Solving problems with outliers in the y-direction.

II. Solving problems with multivariate outliers in the covariant space (x-space).

III. Solving problems with outliers in both the y-direction and the x-space.

The robustness of data is measured by two criterion, first the breakdown point and secondly the influence curve. The breakdown point of an estimate is the smallest fraction of the data that can be damped by an arbitrarily large amount and still cause an arbitrarily large change in the estimate. The influence curve for a statistical function measures how much an individual observation changes the value of the estimator. It is a measure of the dependence of the estimator on the values of one of the points in the sample.

The theoretical background for the robust estimates is described as follows:

Let $X = (x_{ij})$ denote an $n \times p$ matrix,

$$y = (y_1, \ldots, y_n)^T$$

a given n-vector of responses, and

$$\theta = (\theta_1, \ldots, \theta_p)^T$$

an unknown p-vector of parameters or coefficients whose components have to be estimated

The matrix $X$ is called a design matrix.

Considering the normal linear model.

$$y = X\theta + e$$

(28)

Where $e = (e_1, \ldots, e_n)^T$ is an n-vector of unknown errors.

It is assumed that (for given $X$) the components $e_i$ of $e$ are independent and identically distributed according to a distribution $L(\cdot; \delta)$, where $\delta$ is a scale parameter (usually unknown). Often $L(\cdot; \delta) = \Phi(\cdot)$, the standard normal distribution with density is given as:

$$\phi(s) = \frac{1}{\sqrt{2\pi}} e^{-s^2/2}$$

(29)

Where $r = (r_1, \ldots, r_n)^T$ denotes the n-vector of residuals for a given value of $\theta$ and by $x_{i1}^T$ the i-th row of the matrix $X$. The robust distance is defined as:

$$R_d(x_i) = \left[ (x_i - T(X))^T C(X)^{-1} (X_i - T(X)) \right]^{1/2}$$

(30)

Where $T(X)$ and $C(X)$ are the robust location and scatter matrix for the multivariates. In the determination of the high leverage points the following assumptions are made:

Let $C(p) = \sqrt{\chi^2_{p, 1-a}}$ be the cut-off value

(31)

Therefore, the variable for the leverage is defined as:

$$\text{Leverage} = \begin{cases} 0 & \text{if } R_d(x_i) \leq C(p) \\ 1 & \text{otherwise} \end{cases}$$

(32)
The response of the outliers for the robust regression residuals \( r_i \), \( i =1,\ldots,n \) are detected based on the robust estimates. Therefore the variable outlier is defined as:

\[
\text{Outlier} = \begin{cases} 
0 & \text{if } |r| \leq k \sigma \\
1 & \text{otherwise}
\end{cases}
\]  

(33)

The robust measure of goodness of fit and model selection is defined by the expression as:

\[
R^2 = \frac{\sum \rho(y_i - \mu)^2}{\sum \rho(y_i - s)^2}
\]  

(34)

And the robust deviance is defined as the optimal value of the objective function on the \( \sigma^2 \)-scale:

\[
D = 2s^2 \sum \rho \left( \frac{y_i - x_i^T \theta}{s} \right)
\]  

(35)

Where \( \rho \) is the objective function for the robust estimate, \( \mu \) is the robust location estimator, and \( s \) is the robust scale estimator in the full model [30].

There are a few known robust estimators such as least mean squares (LMS), least trimmed squares (LTS), M-estimators, S-estimators and many others. The LMS and LTS possess breakdown points of zero. The introduction of the weight functions on the influence curve result in estimators bounded influence or generalized M-estimators (GM-estimators). The significance of the weight function effect is to reduce the impact of a high leverage point corresponding to an increase in the efficiency of the estimate. The weight function most times is chosen to minimize the asymptotic variance of the estimators. This leads to weights of the form, for matrix \( A \):

\[
W(x) = \|Ax\|^{-1}
\]  

(36)

Often, the breakdown points of these estimates is better than an M-estimate, but cannot exceed \( 1/p \), where \( p \) is the rank of X. The higher the breakdown point of an estimator, the more robust performance is achieved. The weighting functions provided in the robust fit Matlab software application gives a coefficient estimates that are approximately 95% as statistically efficient as the OLS estimates, provided the response has a normal distribution with outliers. This fundamental principle applies to the overall output efficiency. As the tuning constant decreases the weight function increases and vice-versa [20-22].

2.5. Least Trimmed Squares

The least trimmed squares (LTS) is a robust statistical technique that protects against large errors both in explanatory and dependent variables and it is considered as one of the most efficient regression estimators.

The LTS fits a function to a set of data and is unduly affected by the presence of outliers. The LTS estimator belongs to the class of affine-equivariant estimators that converges at the rate of \( n^{-1/2} \) with the same asymptotic efficiency under normal conditions. It has received a lot of attention because of its strong consistency, sensitivity analysis, asymptotic distributions, small-sample corrections, bootstrap and computational methods. The LTS in nonlinear regression model \( (i = 1,\ldots,n) \) is defined by:

\[
y_i = h(x_i, \beta^0) + \varepsilon_i
\]  

(37)

Where \( y_i \) represents the dependent variable, \( h(x_i, \beta^0) \) is a regression function, and \( \beta^0 \in \mathbb{R}^p \) denotes the underlying parameter value. The vector \( \beta \) of unknown parameters is assumed to belong to a parametric space \( S \subseteq \mathbb{R}^p \). Therefore, the nonlinear least trimmed squares estimator \( \beta_{(lts,h,n)} \) is then defined by:

\[
\beta_{(lts,h,n)} = \text{arg min} \sum_{i=1}^{n} r^2(||(x_i, \beta)||) = \text{arg min} \sum_{i=1}^{n} r^2(x_i, \beta) \leq r^2_h(\beta)
\]  

(38)

(39)

Where \( r^2(\beta) \) represents the ordered absolute residuals \( r^2(\beta) = (y_i - h(x_i, \beta))^2 \). The trimming constant \( h \) must satisfy \( n/2 < h \leq n \) and determines the breakdown point of the LTS estimator implies that \( n \rightarrow h \) observations with the largest residuals do not affect the estimator. The choice of the trimming constant \( h \) should vary with the sample size \( n \). However, the least trimmed squares estimator achieves robustness by trimming away observations with large residuals [23-25]. Applying the assumed standard regression model

\[
y_i = x_i \beta + \varepsilon_i
\]  

(40)

Where the regression parameter is \( \beta = (\beta_1,\ldots,\beta_p)' \) and \( \varepsilon_i \) have zero expected value. With a penalty parameter \( \lambda \), the least absolute shrinking operator and selector operator of \( \beta \) is given as:

\[
\beta = \text{arg min} \sum_{i=1}^{n} (y_i - x_i \beta)^2 + n \lambda \sum_{j=1}^{p} |\beta_j|
\]  

(41)

\( L_1 \) penalty allows some coefficients to shrink to zero. The LTS estimator has a simple definition which is quite fast to compute and is the most popular robust regression estimator. The vector of squared residuals is denoted by:

\[
r^2(\beta) = (r^2_1, \ldots, r^2_n) = (y_i - x_i \beta)^2, i = 1, \ldots, n
\]  

(42)
The LTS estimator is defined as
\[ \beta_{\text{LTS}} = \sum_{i=1}^{h} \left( r^2(\beta) \right)_{i:n} \]  
\[ (43) \]

Where \( r^2(\beta) \) is the order statistic of the squared residuals and \( h \leq n \).

A regularized sparse version of the LTS is obtained by adding a penalty with the penalty parameter \( \lambda \) of eq.3 to give the sparse LTS estimator.

\[ \beta_{\text{sparse LTS}} = \text{argmin} \sum_{i=1}^{n} (r^2(\beta))_{i:n} + h \lambda \sum_{j=1}^{p} |\beta_j| \]
\[ (44) \]

The sparse LTS has a high breakdown point which makes it resistant to multiple regression outliers and leverage points. The sparse LTS is efficiently robust compared to the other estimators based on these following factors:

I. The prediction performance is much improved through variance reduction relative to the data size.
II. The provision of simultaneous model selection leading to higher interpretability.
III. The exclusion of computational probability of traditional robust regression methods.

It is observed from (44) when \( h = n \) it yields a lasso solution. The least absolute deviation (LAD) type of estimator called LAD-lasso is given as:

\[ \beta_{\text{LAD-lasso}} = \text{argmin} \sum_{i=1}^{n} |y_i - x'_i \beta| + n \lambda \sum_{j=1}^{p} |\beta_j| \]
\[ (45) \]

The breakdown point of the sparse LTS estimator is also known as the replacement finite simple breakdown point, and is given as:

Let \( Z = (X,y) \)
\[ (46) \]

For a regression estimator \( \beta \), the breakdown point is defined as:

\[ \varepsilon^*(\beta; z) = \min \left\{ \frac{m}{n} \cdot \sup_{\mathcal{I}} \| \beta(z) \|_2 = \infty \right\} \]
\[ (47) \]

Where \( z \) are computed data obtained from \( z \) by replacing \( m \) of the original \( n \) data points by arbitrary values. For a fixed penalty parameter \( \lambda \), the objective function is defined as:

\[ Q(H, \beta) = \sum_{i \in H} (y_i - x'_i \beta)^2 + h \lambda \sum_{j=1}^{p} |\beta_j| \]
\[ (48) \]

The \( L_1 \) penalized sum of squares based on the assumption is given as:

\[ H \subseteq \{1, \ldots, n\} \quad \text{with} \quad |H| = h \]
\[ \beta_H = \text{argmin}_H Q(H, \beta_H) \]
\[ (49) \]

In order to increase the efficiency, a reweighting step that weights down the outliers detected by the sparse LTS is introduced. The sparse LTS estimator is known to be biased, therefore the centre of residuals are determined. A natural estimate for the center of the residuals is given as:

\[ \mu_{\text{raw}} = \frac{1}{h} \sum_{i \in \text{H}_{\text{opt}}} r_i \]
\[ (50) \]

Where \( r_i = y_i - x'_i \beta_{\text{sparse LTS}} \) and \( H_{\text{opt}} \) is the optimal subset from (9). Therefore, the residual scale estimate associated to the raw sparse LTS estimator is given by

\[ \sigma_{\text{raw}} = k_a \left( \frac{1}{h} \sum_{i=1}^{h} (r_i^2)_{i:n} \right) \]
\[ (51) \]

Where \( \sigma_{\text{raw}} \) is a consistent estimate of the standard deviation at the normal model. The squared central residuals \( r_i^2 = ((r_1 - \mu_{\text{raw}})^2, \ldots, (r_n - \mu_{\text{raw}})^2)' \), and

\[ k_a = \left( \frac{1}{a} \int \phi_{-1} \left( \frac{a+1}{2} u^2 \right) d\Phi(u) \right)^{-1/2} \]
\[ (52) \]

The binary weights are defined by this formulation:

\[ w_i = \begin{cases} 1, & \text{if } (r_i - \mu_{\text{raw}})/\sigma_{\text{raw}} \leq \Phi^{-1}(1-\delta), \\ 0, & \text{if } (r_i - \mu_{\text{raw}})/\sigma_{\text{raw}} > \Phi^{-1}(1-\delta), \end{cases} \quad i = 1,\ldots,n \]
\[ (53) \]

The reweighted sparse LTS estimator is given by the weighted lasso fit as;

\[ \beta_{\text{reweighted}} = \text{argmin}_{\beta} \sum_{i=1}^{n} w_i (y_i - x'_i \beta)^2 + n_w \lambda \sum_{j=1}^{p} |\beta_j| \]
\[ (54) \]

with \( n_w = \sum_{i=1}^{n} w_i \) is the sum of the weights.

Considering other weighting schemes, the residual center estimates is given as:

\[ \mu_{\text{reweighted}} = \frac{1}{n_w} \sum_{i=1}^{n} w_i (y_i - x'_i \beta_{\text{reweighted}}) \]
\[ (55) \]

Therefore, the residual scale estimate of the re-weighted sparse LTS estimator is given as;

\[ \sigma_{\text{reweighted}} = k_{a,w} \left( \frac{1}{n_w} \sum_{i=1}^{n} w_i (y_i - x'_i \beta_{\text{reweighted}})^2 \right) \]
\[ (56) \]

Where \( k_{\alpha,w} \) is the consistency factor from \( k_\alpha \) with \( a_w = n_w/n \)

The objective of the sparse based model estimation is to improve prediction performance using different estimators to determine the root mean squared prediction error (RMSPE). The respective sampling data without outliers are generated as test data for each simulation experiment. The RMSPE is given as;

\[ \text{RMSPE} (\beta) = \frac{1}{n} \sqrt{\sum_{i=1}^{n} (y'_i - x'_i \beta)^2} \]
\[ (57) \]
2.6. Computational Regression Experiments

In this section we empirically evaluate the three regression model performance analysis on the static solar photovoltaic array using the sparse based algorithm (SBA) as described in Section 2. We integrated and tested the proposed algorithm using the data set already trained by the neural network algorithm in our previous paper [34].

2.6.1. Computational Experiments

In this section we describe how the data set for the regression model performance was achieved. The Levenberg-Marquardt algorithm was used in the neural network training to solve the nonlinear and implicit equation for the solar photovoltaic array. The input parameters for the experimental setup were obtained by varying the solar radiation, S and wind speed, w, to obtain a new temperature value, T. The solar radiation was increased in the steps of 10 units from 10 to maximum of 1000W/m² and the wind speed increased in the steps of 2 units from 1 to maximum of 20m/s². A data set of 1000 input-output pairs was generated for the experimental setup. The experimental data set is subdivided into three categories. The data categories are training, validation and test data sets. For the experiment, 75% of the total samples randomly were used for our training set, 15% of the remaining samples were used for validation and test respectively. A tenfold cross validation procedure was used for training and testing the solar photovoltaic array model. To study the performance of the proposed regression estimation algorithm the least square loss with L¹–norm regularization and L²–norm regularization parameters have been employed.

\[
\text{min } \frac{1}{2} \| A X - Y \|^2 + \frac{1}{2} \lambda_1 \| X \|^2 + \lambda_2 \| X \|_1 \quad (58)
\]

**Input parameters**

| A-       | Matrix of size m x n |
| Y -      | Response vector (of size mx1) |
| \lambda_2 | L¹ norm regularization parameter (\lambda_2 >=0) |
| \lambda_1 | L² norm regularization parameter (\lambda_1 =0) |

**Output parameters**

<table>
<thead>
<tr>
<th>X -</th>
<th>Co-efficient Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>funVal-</td>
<td>Function value during iterations</td>
</tr>
</tbody>
</table>

The maximum number of iterations used during the experiments is 100. The expected maximum output current, I_{mp}, voltage, V_{mp} and power, P_{mp} characteristics for solar photovoltaic array for the regression model were obtained respectively.

### III. Results

The performance of the proposed regression estimation technique on the static solar photovoltaic array was implemented using the sparse based regression algorithm to compare the three different models. The results are illustrated in this section. The three regression methods includes the ordinary least squares (OLS), robust fit and least trimmed squares (LTS) were compared. In this paper, we reported only the performance based on these two benchmarks; the mean square error (MSE) and root mean squared error (RMSE). The maximum number of iterations is set at 100 for the experiment, while the dimension for each regression input is 910. The values of solar irradiance are randomly obtained from the corresponding temperatures given as:

\[
T = 3.12 + 0.25 S + 0.89 T_a - 1.3 w + 273 \quad (59)
\]

The evaluation for the regression method is obtained by a simple linear regression equation that gives the corresponding \( P_{mp}, V_{mp} \) and \( I_{mp} \) values versus solar irradiance, S. The dispersion diagram for the OLS for \( P_{mp}, V_{mp} \) and \( I_{mp} \) versus solar irradiance and S respectively are shown in figures 3, 4 and 5. The regression standard error estimate values are 0.96015, 0.0047247, 0.96587 for \( P_{mp}, V_{mp} \) and \( I_{mp} \) respectively. As it was observed the standard error estimate value for \( V_{mp} \), the linear OLS is not accurate enough due to the presence of outliers not filtered. The obtained equation for \( P_{mp}, V_{mp} \) and \( I_{mp} \) with OLS, are:

\[
\begin{align*}
P_{mp} & = -26.9195 + 0.0849S \\
V_{mp} & = -0.1823 + 0.0015S \\
I_{mp} & = -37.7989 + 0.1320S
\end{align*} \quad (60)
\]

And the mean square error (MSE) and root mean squared error (RMSE) for \( P_{mp} \) are 0.9219 and 0.9602 respectively; and the MSE and RMSE for \( V_{mp} \) are 2.232 x 10⁻⁵ and 0.0047 respectively; and the MSE and RMSE for \( I_{mp} \) are 0.9329 and 0.9659 respectively.

The dispersion diagram for the Logistic robust fit regression method for \( P_{mp}, V_{mp} \) and \( I_{mp} \) versus solar irradiance, S respectively are shown in figures 6, 7 and 8. The regression standard error estimate values are 0.96134, 0.004774, 0.96593 for \( P_{mp}, V_{mp} \) and \( I_{mp} \) respectively.
As it was observed the standard estimate value for $V_{mp}$, the linear OLS is not accurate enough due to the presence of outliers not filtered. The obtained equations for $P_{mp}$, $V_{mp}$ and $I_{mp}$ with OLS, are:

$$\begin{align*}
P_{mp} &= -26.9941 + 0.0850S \\
V_{mp} &= -0.01780 + 0.0015S \\
I_{mp} &= -37.7299 + 0.1318S
\end{align*}$$

(61)

And the MSE and RMSE for $P_{mp}$ are 0.9242 and 0.9613 respectively; and the MSE and RMSE for $V_{mp}$ are 2.2824 x $10^{-3}$ and 0.0048 respectively; and the mean square error (MSE) and root mean squared error (RMSE) for $I_{mp}$ are 0.9330 and 0.9659 respectively.

The regression standard error estimate values is 0.9871, 0.4763, and 0.41815 for $P_{mp}$, $V_{mp}$ and $I_{mp}$ respectively. As it was observed the standard estimate value for $V_{mp}$, the linear OLS is not accurate enough due to the presence of outliers not filtered. The obtained equation for $P_{mp}$, $V_{mp}$ and $I_{mp}$ with OLS, are:

$$\begin{align*}
P_{mp} &= -0.0296 + 0.0849S \\
V_{mp} &= -0.0002 + 0.0015S \\
I_{mp} &= -0.0415 + 0.1320S
\end{align*}$$

(62)

And the MSE and RMSE for $P_{mp}$ are 0.9756 and 0.9877 respectively; and the MSE and RMSE for $V_{mp}$ are 0.2286 and 0.4764 respectively; and the MSE and RMSE for $I_{mp}$ are 0.1749 and 0.4182 respectively.

IV. DISCUSSION

4.1 Performance Evaluation

Table 2. presents the comparison of the standard error estimates on the static solar photovoltaic module results of the regression methods. It is observed from the experimental results, that the $P_{mp}$ standard error estimate is relatively comparable. The $V_{mp}$ standard error estimate is relatively low compared to $P_{mp}$ and $I_{mp}$ respectively. Lastly, the $I_{mp}$ standard error estimate results for the OLS and logistic robustfit are better than that of the LTS obtained from the experiment. Table 3. presents the comparison of the MSE and RMSE of the static solar photovoltaic module results of regression methods. It is observed from the experimental results, the $P_{mp}$ MSE and RMSE are relatively comparable. The $V_{mp}$ standard error estimate is relatively low compared to $P_{mp}$ and $I_{mp}$ respectively. Finally, the $I_{mp}$ MSE and RMSE results for the OLS and logistic robustfit are better than the LTS obtained from the experiment. We observed from other literatures the $V_{mp}$ characteristics have not been reported which calls for further studies.

Fig 3. Dispersion diagram of Pmp vs. S
In this paper, we evaluated the comparative performance of the static solar photovoltaic module using the sparse based algorithm. For training, testing and validation processes the neural network algorithm was employed before introducing the sparsity based algorithm for evaluating the performance of the three regression models. The comparative performance of the sparse based algorithm for the three different regression methods, the OLS and logistic robustfit prediction based on the simulation results demonstrated from the experiment shows that the first two are superior to the LTS. We chose the mean square error and the root mean square error as performance measures for the regression classification of the methods and the results were presented in Tables 2 and 3 respectively.
REFERENCES


[34] Balogun E., Xu Huang and Dat Tran. "Comparative study of different artificial neural networks methodologies on static solar photovoltaic model."