Wind Farms Conclusions

Teo Zeng

This document attempts to summarize all the observations from the exploratory data analysis done the forcast and actual wind farm production in Texas.

Notations

J	Number of assets
$d = 1, \ldots, D; h$	Days, Hours
C_i	Capacity for Asset i
$\mathbf{g}_d = (g_{d,h})$	Actual Production MWh
$\mathbf{f}_d = (f_{d,h})$	forcasted Production MWh
$\tilde{\beta}_{d,h}$	prediction bias
i	assets indices
$\alpha_{d,h}$	actuals fraction $\in [0, 1]$
$\gamma_{d,h}$	forecast fraction $\in [0, 1]$
$\beta_{d,h}$	bias fraction
$q_{d,h}$	95 quantile fraction
$\mathbf{q}_{d,h}$	quantile
$\tau_{d,h}$	kendall's tau
r^2	corelation coefficient
σ	normalized standard deviation of bias
$\tilde{\sigma}$	unnormalized standard deviation of bias
S	the set of all assets
n	number of occurence as outliers

- Days are indexed by $d = 1, 2, ..., 365$, and hours are indexed by $h = 1, 2, ..., 24$
- $\mathbf{g}_{d,h}$, $\mathbf{f}_{d,h}, \alpha_{d,h}, \beta_{d,h}, q_{d,h}$ are vectors of Dimension J Where

$$\alpha_{d,h} \in [0, 1] = \frac{\mathbf{g}_{d,h}}{C} \quad \text{and} \quad \gamma_{d,h} \in [0, 1] = \frac{\mathbf{f}_{d,h}}{C}$$

- The bias fraction is then

$$\beta_{d,h} = \alpha_{d,h} - \gamma_{d,h}$$

- For a population, of discrete values or for a continuous density, the k^{th} q-quantile is the data value where the cumulative distribution function crosses k/q. x is a k^{th} q-quantile for a variable X if

$$\Pr[X < x] \leq k/q \quad \text{and} \quad \Pr[X \leq x] \geq k/q$$

- For the entire article, we let $q = 95$, therefore, the quantile for an asset at a given time interval is calculated as

$$\mathbf{q}_{d,h} = q_{d,h} - \bar{\beta}_{d,h}$$

Bias

Bias at hourly interval

For a given asset, let

$$\beta_h = \frac{1}{365} \sum_{d=1}^{365} \beta_{d,h}$$

and β_h is the prediction bias for a given hour of the day. For each $h \in [1, 24]$, we plot out β_h in a violin plot and a box plot.

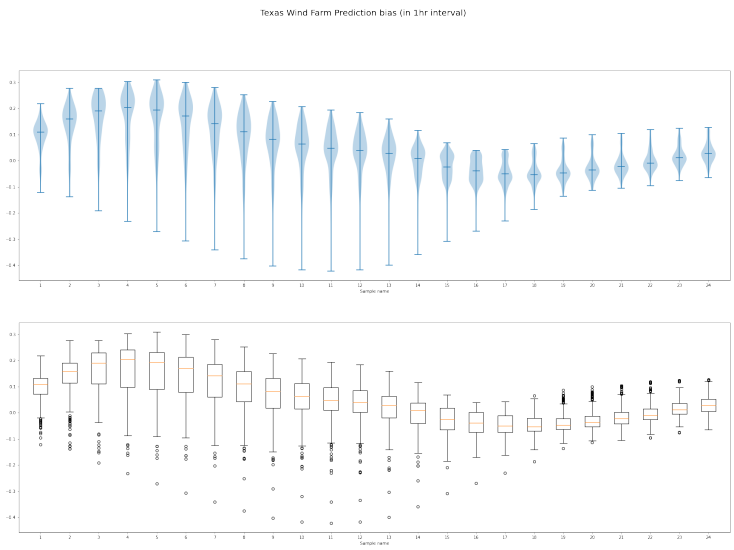

Figure 1: Violin and Box of Bias at 1 Hour Interval

The blue line on the violin plot and the red line on the box plot represent the median of the data. It was found that, overall in the morning the prediction bias tend to be positive, with negatively skewed distribution. In the afternoon and evening the prediction bias tend to be negative, with less skewed distribuiton. With closer analysis on the outliers, two outliers at the bottom from hour 3 to hour 19 are due to **Canadian Breaks Wind** and **Desert Sky repower**.

Bias at 24 Hours interval

We are also interested in the geographic distribution of the prediction bias.

For a given asset, let

$$\beta = \frac{1}{365 \times 24} \sum_{d=1}^{365} \sum_{h=1}^{24} \beta_{d,h}$$

and β is also a vector of length J. Plotting β on the map of Texas,

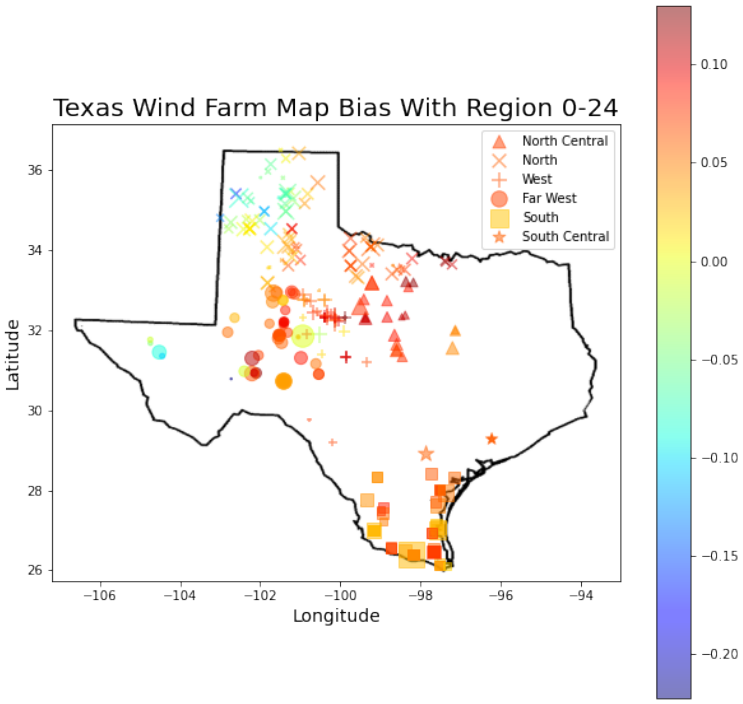

Figure 2: Bias at 24-hour Interval

As observed from the scatter plot, the bias, across all hours and all year, is relatively higher for Southern and central wind farms and is lower for Northern and Western windfarms.

Stardard Deviation of Bias

Standard Deviation of Bias at Hourly Interval

We are also interested in the standard deviation of the Bias. Let

$$\sigma_h = \sqrt{\frac{\sum_{d=1}^{365} \left(\beta_{d,h} - \bar{\beta_h}\right)^2}{365}}$$

where $\bar{\beta_h} = \frac{1}{365} \sum_{d=1}^{365} \beta_{d,h}$ is the mean of the bias at a given hour across all year. For each $h \in [1, 24]$, we plot out σ_h in a violin plot and a box plot.

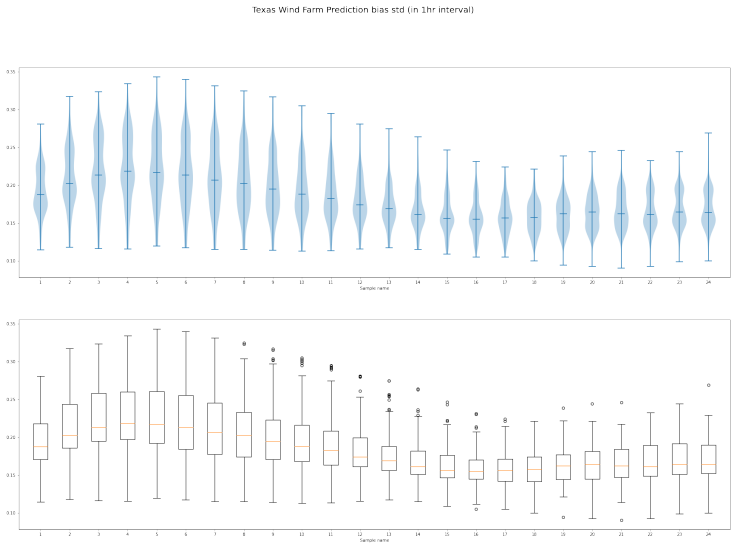

Figure 3: Violin and Box of std at 1 Hour Interval

It was found that, from the above violin/box plot, the mean and median of the standard deviation of bias tend to be stable across 24 hours. There are some outliers during sunlight time and few outliers at night. It was also found that the standard deviation of bias tend to be larger in the early morning comparing to the afternoon.

Standard Deviation of Bias at 24 hour interval

We are also interested in the geographic distribution of the standard deviation of bias.

For a given asset, let

$$\sigma = \sqrt{\frac{\sum_{h=1}^{24} \sum_{d=1}^{365} \left(\beta_{d,h} - \bar{\beta}\right)^2}{365}}$$

where $\bar{\beta} = \frac{1}{365 \times 24} \sum_{h=1}^{24} \sum_{d=1}^{365} \beta_{d,h}$ is the mean of the bias across all year and all hours. σ is also a vector of length J.

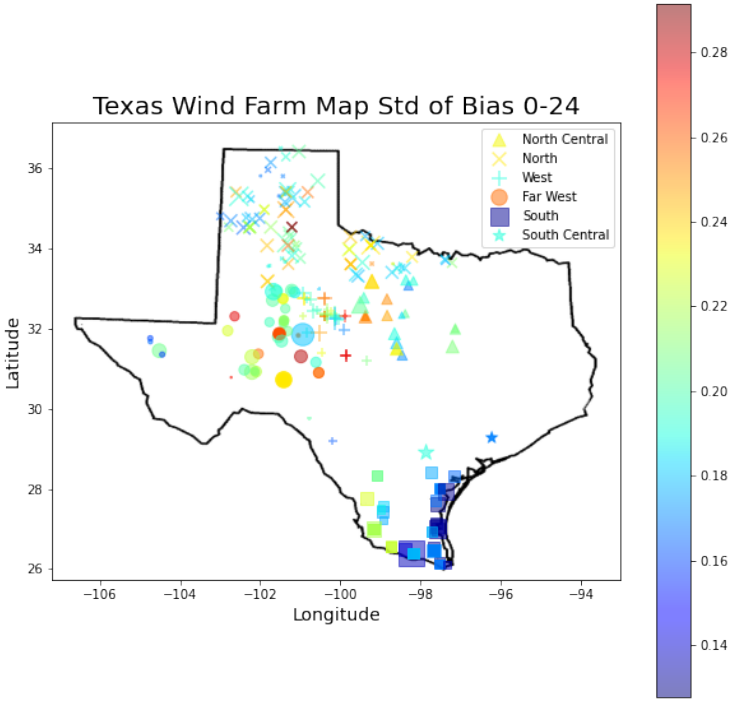

Figure 4: Standard Deviation of Bias at 24 hour interval

From this scatter plot, it was found that the stardard devaition of bias is relatively lower for the Southern assets.

Quantile of Bias

Quantile of Bias at Hourly Interval

We are also interested in how the quantile of bias of each assets behave across the day. For each $h \in [1, 24]$, we plot out $\mathbf{q}_h$ in a violin plot and a box plot.

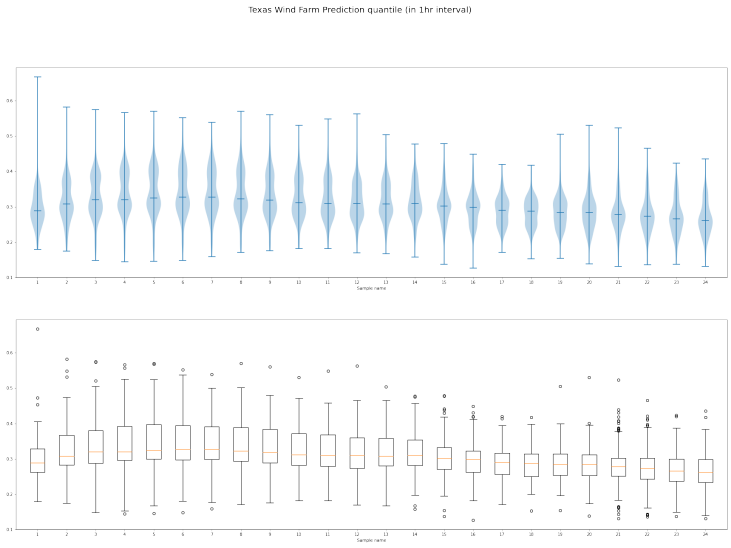

Figure 5: Violin and Box of Quantile at 1 Hour Interval

We see that, from the above violin and box plot, the distribution of quantiles are similar across most of the day time.

Quantile at 24 Hour Interval

We are also interested in how $\mathbf{q}$ is distribution geographically across all assets, we know that $\mathbf{q}$ is vector of length J. Plotting $\mathbf{q}$ on the map of Texas, we get the following

As shown on the plot below, we see that there is no apparent spatial relationship across assets on quantile. Some assets at the Southwest boarder of texas tend to have lower quantile.

kendall's Coefficient for Each Region

We are also interested in the ordinal correlation of σ between each asset across hours. Since there are a total of 6 regions texas and S is the entire set of assets. Let

$$S = \{S_1, S_2, S_3, S_4, S_5, S_6\}$$

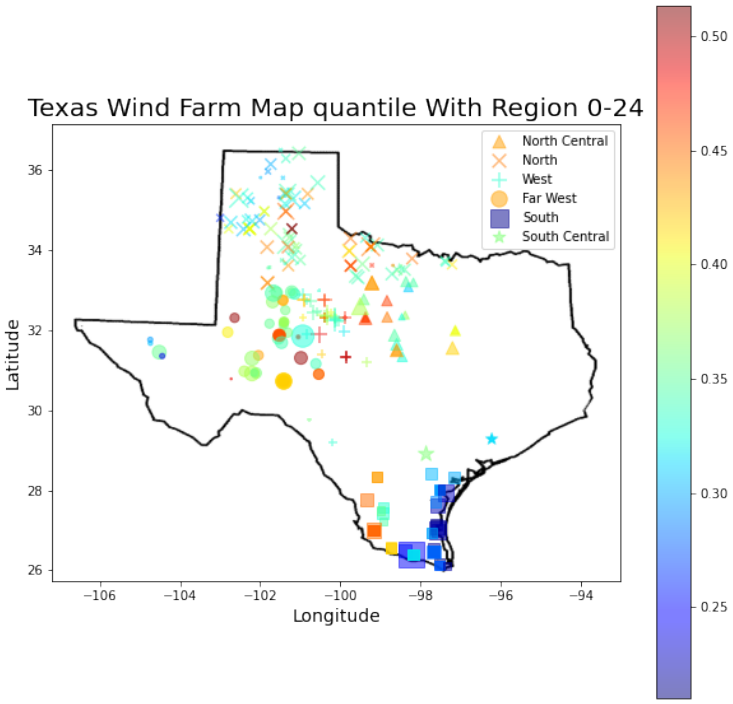

Figure 6: Quantile of Bias at Hourly Interval

Where $S_i, i \in \{1,2,3,4,5,6\}$ is the set of assets in a particular region. And let

$$\tau_i^h = \frac{2}{|S_i|(|S_i|-1)} \sum_{m<n, m,n \in S_i} \operatorname{sgn}\left(\sigma_m^h - \sigma_n^h\right) \operatorname{sgn}\left(\sigma_m^{h+1} - \sigma_n^{h+1}\right)$$

So τ_i^{h+1} is the kendall correlation coefficient for asset S_i during hour interval $[h, h+1]$. Then for $h \in [1,23]$ and for$i \in \{1,2,3,4,5,6\}$ we plot out the kendall correlation for each region across 23 hour intervals,

We also calculated the overall kendall correlation coefficient τ^h by

$$\tau^h = \frac{2}{|S|(|S|-1)} \sum_{m<n, m,n \in S} \operatorname{sgn}\left(\sigma_m^h - \sigma_n^h\right) \operatorname{sgn}\left(\sigma_m^{h+1} - \sigma_n^{h+1}\right)$$

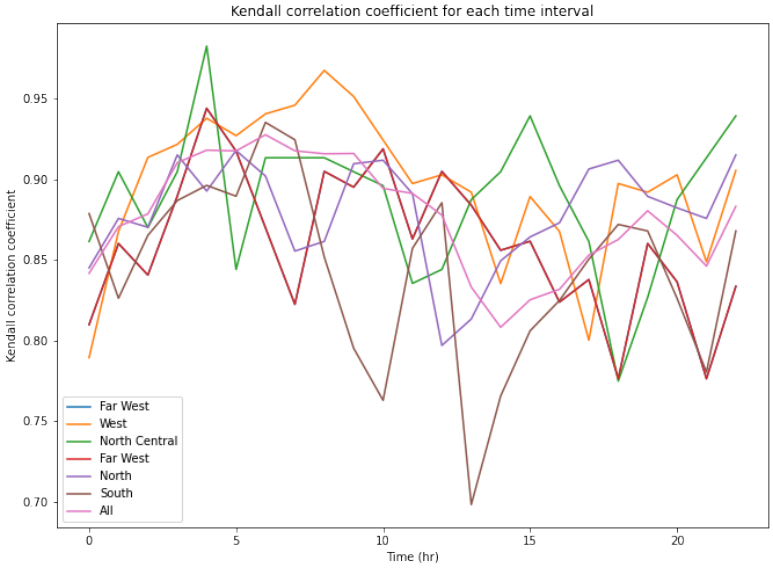

Figure 7: Kendall Coefficient by Area

As seen from the plot, overall, the kendall's τ tend to be lower in the afternoon comparing to the mornings. It was also found that the assets in the South behave unsual comparing to other regions, its kendall's τ is lowest at hour 13.

Capacity vs. std of unnormalized bias

We are also interested in the correlation between an asset's capacity and the standard deviation of unormalized bias. Let

$$\tilde{\sigma} = \sqrt{\frac{\sum_{h=1}^{24} \sum_{d=1}^{365} \left(\tilde{\beta}_{d,h} - \bar{\tilde{\beta}}\right)^2}{365}}$$

Plotting Capacity C on the x-axis and $\tilde{\sigma}$ on the y axis, we calculate the corelation coefficient r^2

$$r = \frac{\sum_{i=1}^{J} (\sigma_i - \bar{\sigma}) \left(C_i - \bar{C}\right)}{\sqrt{\sum_{i=1}^{J} (\sigma_i - \bar{\sigma})^2 \sum \left(C_i - \bar{C}\right)^2}}$$

where $\bar{\sigma} = \frac{1}{J} \sum_{i=0}^{J} \sigma_i$ and $\bar{C} = \frac{1}{J} \sum_{i=0}^{J} C_i$

It was calculated that $r^2 = 0.897$, indicating a relatively strong linear relationship. However, from the scatter plot we can see that the linear relationship is strong for $C < 450$. It was also observed that for $C > 450$ There is a clearly non linear pattern. Besdies, the assets in the south has lower std/capacity ratio comparing to other regions. Therefore, we use a local regression to make a better fit of the data.

With the weight function

$$w(x) = \left(1 - |d|^3\right)^3$$

where d is the distance of a given data point from the point on the curve being fitted, scaled to lie in the range from 0 to 1. We also specify the loss function as

$$\mathrm{RSS}_x(A) = \sum_{i=1}^{N} (y_i - A\hat{x}_i)^T w_i(x) (y_i - A\hat{x}_i) .$$

Here, A is an $(n+1) \times (n+1)$ real matrix of coefficients, $w_i(x) := w(x_i, x)$ and the subscript i enumerates input and output vectors from a training set.

Outliers

Let outliers to be any oberservation outside the range

$$\left[q^{25} - k\left(q^{75} - q^{25}\right), q^{75} + k\left(q^{75} - q^{25}\right)\right]$$

Where q^{25}, q^{75} are the 25 quantile and 75 quantile respectively, k is some non-negative constant. We use $k = 2$ across all the analysis

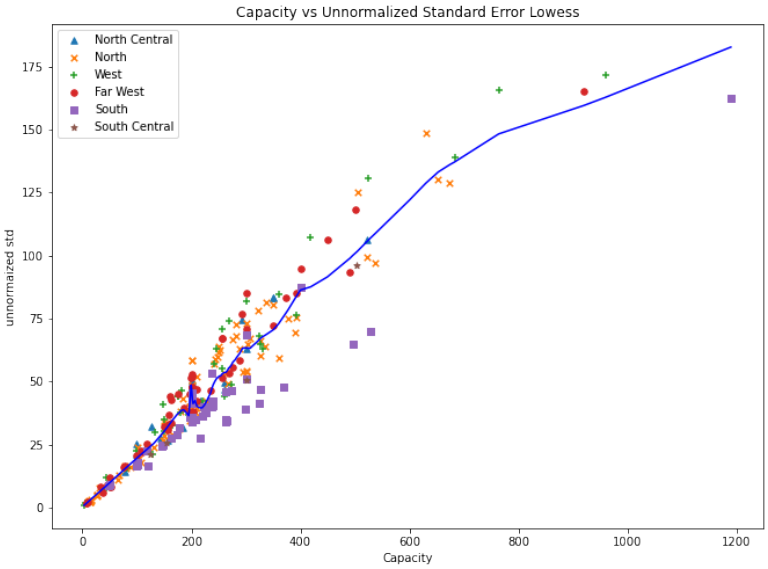

Figure 8: Capacity vs. Unnormalized std lowess

Bias Outliers

We count the number of occurence as outliers of each assets across assets for bias.

Bias Butliers

Asset	n_β
Desert Sky repower	19
Canadian Breaks Wind	18
Tierra Blanca W	16
Broadview Energy JN LLC	16
Northdraw Wind	16
Wildrose Wind	15
Wind Power Partners '94 Wind Farm	14
Delaware Mountain Wind Farm	14
Gusty Hill Wind	14
Pantex Plant Wind Project	13

Standard Deviation Outliers

We count the number of occurence as outliers of each assets across assets for standard deviation of bias.

Standard Deviation Outliers

Asset	n_σ
Cameron	10
Anacacho Wind Farm	10
San Roman	8
Bruenning's Breeze	8
Magic Valley	8
Gulf Wind Farm	7
Penascal II	7
Baffin	6
Tecovas 1 W	5
Papalote Creek	5

Quantile Outliers

We count the number of occurence as outliers of each assets across assets for quantile.

Quantile Outliers

Asset	$n_{\mathbf{q}}$
Wind Power Partners '94 Wind Farm	15
Delaware Mountain Wind Farm	13
Anacacho Wind Farm	10
Cameron	10
San Roman	9
Gusty Hill Wind	9
Penascal II	6
Bruenning's Breeze	5
Magic Valley	5

For the above tables, we see that **Wind Power Partners '94 Wind Farm** has leading outlier occurence in bias and quantile, but not standard deviation. **Desert Sky** is most outlier occurences in bias but not in quantile. The assets on the bias and quantile outliers tend to align, but an asset is outlier in quantile or bias does not mean that it is a outlier in standard deviation.

Standard Deviation vs. Quantile

We are interested in the relationship between standard deviation and the quantile. If each assets are independent and identically distributed the ratio of quantile to

standard deviation should be constant, that is

$$\frac{\mathbf{q}}{\sigma} \sim 2$$

Now we plot a scatter plot for all the assets, with the standard deviation on the x-axis and the quantile on the y-axis.

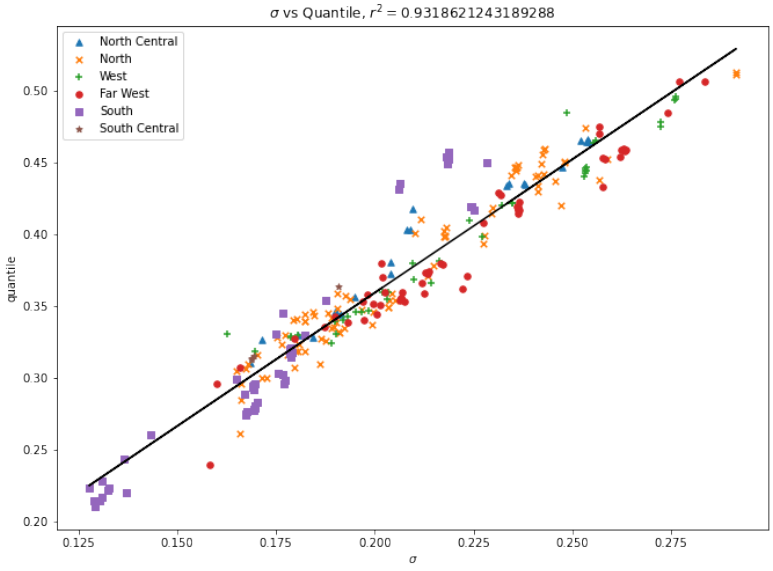

Figure 9: Standard Deviation vs. Quantile

As shown on the above figure, a clear linear relationship is observed with $r^2 = 1.85$. And we found out that the slop of the fitted line is 1.85, which is close to the value of 2.