# Kalman Filter Linear Regression for Energy Prediction

Thesis

Submitted in partial fulfillment of the requirements of

BITS F424T Thesis

**By**

Prashant Tak

(2018B4A81050P)

**Under the Supervision of**

Dr. Sumanta Pasari

Assistant Professor

Department of Mathematics


Birla Institute of Technology and Science, Pilani

Rajasthan – 333031

# Acknowledgement

I would like to thank the Mathematics Department of BITS Pilani for providing me with this incredible opportunity to pursue my undergraduate thesis and for providing me with the background necessary to start working on the project.

I'd also like to thank Dr. Sumanta Pasari who has guided me in the whole endeavour from the beginning. At every step, he has provided me with valuable inputs and motivation to do better. Alongwith that, I'd also like to mention Ms. Sakshi Shukla, who has provided me with suggestions regarding through the whole process and clarified any doubts I had.

Last but not the least, I would like to thank my parents for providing me with lessons throughout my life and never ending moral and financial support.

# Abstract

Short-term energy forecasting influences energy plants operations a lot and it can also have environmental impact. This report looks into studying Dynamic Linear Models for assessing solar irradiance values and estimating them. The Global Horizontal Irradiance (GHI) data for Bhadla Solar Park (Gujarat) from 2001 to 2021 is considered for the analysis. The stationarity of the data is verified and a dynamic linear regressive framework with seasonal terms is proposed to model the GHI values. The parameters for the model are obtained via Kalman Filter. The residuals for the predicted values are investigated.

# Introduction

A country's ability to advance economically and technologically is greatly influenced by its access to electricity. The amount of power produced globally has significantly increased as a result of society's growing reliance on energy. Conventional fossil fuels are frequently the main contributors to the production of electricity globally. They take a long time to create, and the rate at which current reserves are being used up is much higher than the rate at which they are being produced. As a result, they are close to going extinct.

Fossil fuels are also one of the main sources of greenhouse gas emissions, which significantly contribute to global warming and, as a result, pose a threat to the environment and human-dependent living conditions. In recent years, the emphasis has shifted from the use of fossil fuels to the use of renewable energy sources for the production of electricity. Accurate renewable energy forecasting aids in planning and projecting energy output from the short to the long term.

The contribution of wind and solar energy is notable among other renewable energy sources. For a variety of practical uses, including estimating plant energy outputs, marketing renewable energy, and planning maintenance for wind farms and solar plants, wind speed and GHI estimates are helpful. While monthly weather forecasts can be utilised for long-term planning of power plants, daily forecasts can be used to determine the optimal months for solar and wind energy production.

GHI (Global Horizontal Irradiance) refers to the amount of radiant power from sunlight received by a particular surface, perpendicular to the suns rays. It is useful when monitoring a solar power plant, nding optimal placement location of solar plants and for assessing solar power plant feasibility. The higher the GHI value, the more power a system will produce. Its measurement unit is Watts per Sq Meter ($W/m^2$).

# Background Literature

Statistical analysis of time series data is usually faced with the problem that we have only one realization of a process whose properties we might not fully understand. A Dynamic Linear Model (DLM) is a time series forecasting method that combines linear regression with a state-space model. It allows for modeling time-varying parameters and can incorporate both exogenous variables and latent variables that are unobserved but influence the time series. The key advantage of DLM over other forecasting methods is its ability to model time-varying parameters, which makes it suitable for modeling complex dynamic systems. In linear trend analysis, we assume that there is an underlying change in the background that stays approximately constant over time. In dynamic regression systems, by explicitly allowing for variability in the regression coefficients we let the system properties change in time. Also, unlike ARMA models, they can be applied to non-stationary data without transformation. Furthermore, the use of unobservable state variables allows direct modelling of the processes that are driving the observed variability, such as seasonality or external forcing, and we can explicitly allow for some modelling error.

## Bayesian Inference

Bayesian inference is a statistical approach to data analysis and decision making that involves updating probabilities based on new data and prior knowledge. In Bayesian inference, probabilities are treated as measures of uncertainty rather than frequencies. The key idea is to start with an initial prior probability distribution that represents our beliefs about the unknown parameters of interest, and then update this distribution based on observed data using Bayes' theorem. It states that

$$Posterior = \frac{Likelihood * Prior}{Evidence}$$

Bayesian inference allows for incorporating prior knowledge or assumptions about the parameters, which can improve the estimation and prediction accuracy.

## Dynamic Linear Models

State space models consider a time series as the output of a dynamic system perturbed by random disturbances. They allow a natural interpretation of a time series as combination of trend, seasonal or regressive components. In a state space model we assume that there is an unobservable Markov chain $(x_t)$, called the *state process*, and that $y_t$ is an *imprecise measurement* of $x_t$. A trivial DLM consists of two sets

of equations:

$$y_t = F_t x_t + v_t$$

$$x_t = G_t x_{t-1} + w_t$$

Here $y_t$ represents the observation at time t, $v_t$ and $w_t$ are sequences of independent gaussian random errors (*observation error and evolution error*) and $x_t$ corresponds to the unobserved state of the system having a *prior distribution* for $x_0 \sim N(m_0, C_0)$. $F_t$ and $G_t$ are the *observation* and *system matrices*.

## Kalman Filters

Model building can be a major difculty: there might be no clear identication of physically interpretable states, or the state space representation could be non unique, or unsuitable choice of parameters could result in an inadequate model. To estimate the state vector we compute the conditional densities $\pi(x_s|y_{1:t})$. We distinguish between problems of filtering (when s = t), state prediction (s > t) and smoothing (s < t).

In a DLM, the Kalman lter provides the formula for updating our current inference on the state vector as new data become available, that is, for passing from the filtering density $\pi(x_1|y_{1:t})$ to $\pi(x_{t+1}|y_{1:t+1})$. It allows us to compute the predictive and ltering distributions recursively, starting from $x_0 \sim N(m_0, C_0)$ then computing $\pi(x_1|y_1)$, and proceeding recursively as new data becomes available. This is the usual Bayesian sequential updating, in which the posterior at time t takes the role of a prior distribution for what concerns the observations after time t.

### Filtering

Taking the vector of observations $y_{1:t}$, the filtering distribution $\pi(x_t|y_{1:t})$ is computed recursively as:

1. Start with $x_0 \sim N(m_0, C_0)$
2. One step forecast for the *state*:

$$x_t|y_{1:t} \sim N(a_t, R_t) \text{ where } a_t = G_t m_{t-1} \text{ and } R_t = (G_t C_{t-1} G_t') + W_t$$

3. One step forecast for the *observation*:

$$y_t|y_{1:t}\ N(f_t, Q_t) \text{ where } f_t = F_t a_t \text{ and } Q_t = (F_t R_{t-1} F_t') + V_t$$

4. Compute the posterior at time t:

$$x_t|y_{1:t} \sim N(m_t, C_t) \text{ where } m_t = a_t + R_t f_t' Q_t^{-1}(y_t - f_t) \text{ and } C_t = R_t - (R_t F_t' Q_t^{-1} F_t R_t)$$

# Formulation and Methodology

To estimate the GHI values, a Linear Regressive Dynamic Model is chosen with seasonal factors. A linear regression model (with lagged values of observation as regression variable) looks like

$$y_t = y_{t-1}x_t + v_t$$

$$x_t = G_t x_{t-1} + w_t$$

The available data for GHI and wind speed are hourly data without a date-time index. Therefore, a date-time index for the data is created and any missing values are dealt with. Due to the absence of sun radiation, values from nighttime to early morning are minimal in GHI data. Therefore, those values are omitted, as the forecast is only necessary for times when there is sufficient sun irradiation.

Then, the Augmented Dickey-Fuller(ADF) test is ran to determine whether the time series is stationary or not. The null hypothesis of the ADF is that the underlying series is nonstationary, whereas the alternative hypothesis is that the series is stationary but lacks a unit root. If the p-value of the ADF test is less than the critical value, then the data is considered stationary. When p-values are large, however, the null hypothesis cannot be rejected, indicating that the data is not stationary.

Afterwards, the time series is decomposed using an additive model (since it has no trend with time) into trend, seasonality and residuals. This allows one to infer about the underlying characteristics of the data and provides initial ideas regarding the formulation of the DLM.

The DLM implementation is performed with the help of `pyDLM` library. The DLM is built upon two layers. The first layer is the fitting algorithm. DLM adopts a modified Kalman filter with a unique discounting technique from Harrison and West (1999). The second layer of DLM is its modeling feature. The DLM can easily incorporate most modeling components and turn them into the corresponding transition matrices and other quantities to be supplied to the Kalman filter. Examples are trend, seasonality, holidays, control variables and auto-regressive, which could appear simultaneously in one model.

After creating the regressive model (with lagged values of data) and applying the Kalman Filter, the estimated plots are generated and the residuals are computed to verify the model accuracy.

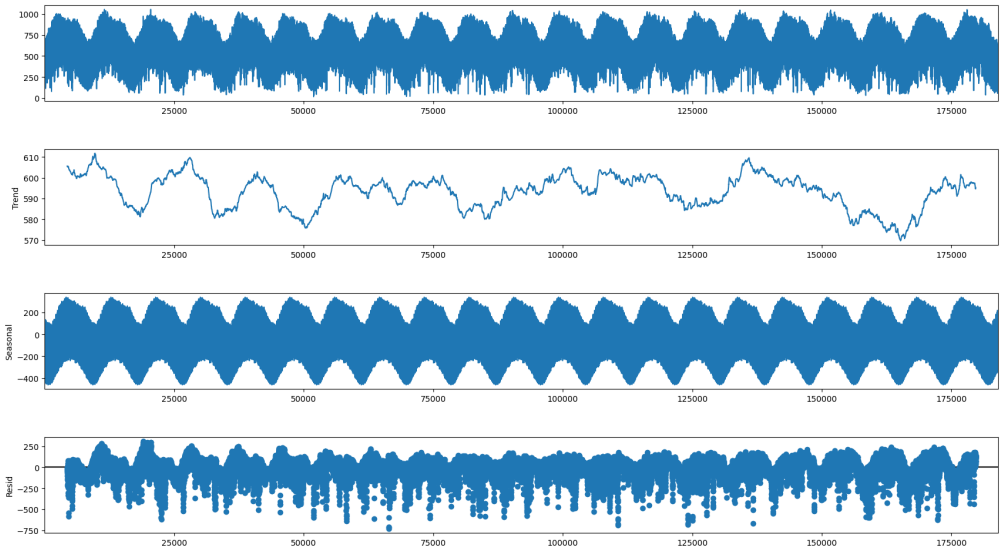
# Results

Upon performing the ADF test, the resultant ADF statistic and p-value were as follows:

- ADF Statistic : -8.77047
- p-value : $2.529 \text{x} 10^{-14}$

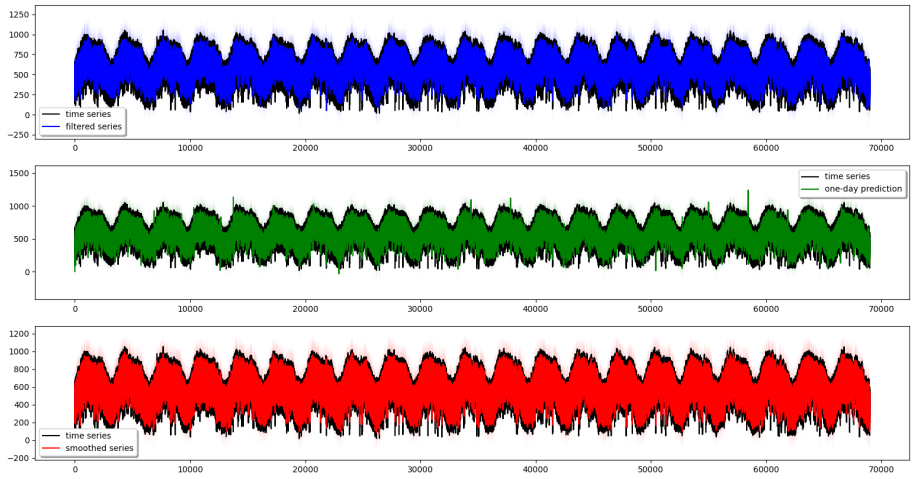
On testing our hypothesis on 1% confidence interval, since the resultant p-value is much lesser than the critical value of 0.01, the null hypothesis is rejected and it is concluded that there is no unit root and that the time series is stationary.

Below is the time series decomposition of the data, as is clearly seen that there is a periodic seasonal component however when it comes to trend there's no continuous clear rise or fall hence we can avoid having a trend component in our DLM but still consider an yearly seasonal component. On checking the residuals, it points towards the naive additive decomposition not taking into account the outliers.

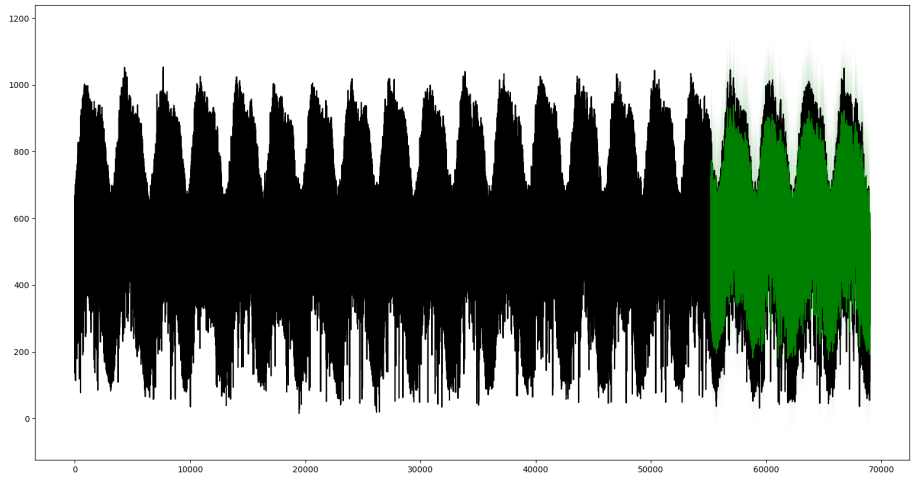



GHI Decomposition

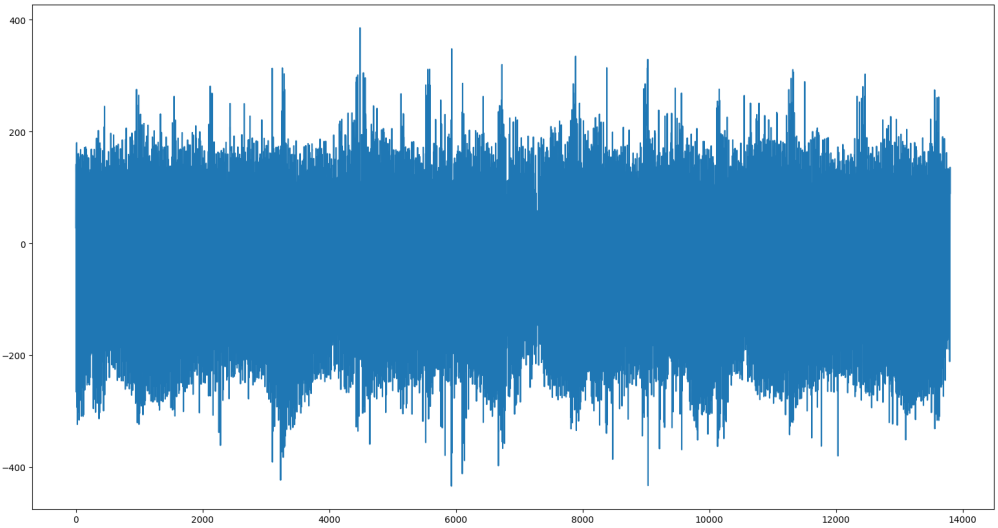
After the model is constructed by adding the dynamic regressive and seasonal components, the Kalman Filter is run. The key output from the model are the filtered time series, smoothed time series and one-step ahead prediction.




The GHI data was divided into 80-20 where 80% of the data was used for model fitting, on the 20% testing data the data prediction was performed.



Upon getting the estimates, the residuals of the predicted values were plotted. The residuals of a forecast model should exhibit Gaussian distribution with zero mean and a constant variance. However the mean of the residuals was less than zero which could be attributed to ineffective model formation or incorrect parameter estimation.




Residuals

# References

1. Petris, G., Petrone, S., & Campagnoli, P. (2009). Dynamic Linear Models with R. Springer Science & Business Media.

2. West, M., & Harrison, J. (2013). Bayesian Forecasting and Dynamic Models. Springer Science & Business Media.

3. Wang, Y. (2017). Bayesian-based Methodology for Progressive Structural Health Evaluation and Prediction by Use of Monitoring Data.

4. Nagi, A. (2014, September 1). Linear State Space Linear Models, and Kalman Filters.

5. Laine, M. (n.d.). Dynamic linear model tutorial.

6. PyDLM PyDLM 0.1.1 documentation. (n.d.).