OPEN ACCESS

International Journal Of Modern Engineering Research (IJMER)

Data Analytics on Solar Energy Using Hadoop

K.Akila[1], B.Manasavani[2], Dr.Gopikrishnan M[3]

[1]Department Of ComputerEngineering Prathyusha Engineering College Thiruvallur, B.E.,India.

[2]Department Of ComputerEngineeringPrathyusha EngineeringCollege

ABSRACT: *Missing data is one of the major issues in data miningand pattern recognition. The knowledge contains in attributeswith missing data values are important in improvingregression correlationprocessofanorganization.Thelearningprocesson each instance is necessary as it may containsome exceptional knowledge. There are various methods tohandle missing data in regression correlation. Analysis of photovoltaic cell, Sunlight striking on different geographical location to know the defective or connectionless photovoltaic cell plate. we mainly aims To showcasing the energy produced at different geographical location and to find the defective plate.*

And also analysis the data sets of energy produced along with current weather in particular area to know the status of the photovoltaic plates. In this project, we used Hadoop Map-Reduce framework to analyze the solar energy datasets.

Keywords: *missing data, genetic algorithm, regression correlation.*

INTRODUCTION

Missing data is the missing form of informationabout phenomena, which is important, and it is the informationin which we are interested. The existence of missing data isone significant problem in data quality. Data quality playsmajor role in machine learning, data mining andknowledge discovery from databases. Machine learningalgorithmshandle missing data in a quite naive way. To avoid biasingin induced hypothesis missing data treatment shouldbe carefullyhandled.Imputationisaprocessthatreplacesthe missing values in instance by some reasonable values.The case substitution is the method developed for dealingwith missingdataininstancesanditishavingsomedrawbackswhen applied to the data mining processes. Themethods,such as substitution of missing values by the attributemean or mode should be cautiously handed to avoid inclusionof bias.

1.1 Randomness of MissingData

Missing data randomness is classified [1] in threeclasses.

Missing completely at random (MCAR): Missing values are scattered randomly across all instances. Inthistype of randomness, any missing data handling methodcan be applied without risk of introducing bias on the data.It occurs when the probability of an instance having amissing valueforanattributedoesnotdependoneithertheknown values or the missingdata.differences on attributes establish that the two groups donot differsignificantly.

Missing at random (MAR): Missing atrandom (MAR) is a condition, which occurs when missing valuesare not randomly distributed across all observations butare randomly distributed within one or more classes (ex.missing more among whites than non-whites, but randomwithin each). The probability of an instance with a missing valuefor an attribute may depend on the known values, and not onthe value of the missing dataitself.

Not missing at random (NMAR): Not missingat random is the most challenging form, occurs whenmissing values are not randomly distributed across observations. Itisalso called as non-ignorable missingness. The probabilityof an instance with a missing value for an attributemight depend on the value of thatattribute.

1.2 Handling MissingData

Missing data handling methods are categoriesed asfollows**Ignoring data:**This method throw-outs all instanceswith missingdata.Therearetwocoremethodstodiscarddata with missing values. The first one is known as completecase analysis.Itisavailableineveryoneofstatisticalpackagesand is the default method in manyprograms.

The next method is recognized as discarding instancesor attributes. This method determines the level of missingdata oneachinstanceandattribute,anddeletestheinstancesor attributes with high extents of missing data.

Prior todeletingany attribute, it is vital to evaluate its connotation tothe investigation. The methods, complete case analysisand discarding is executed only if missing data aremissing completely at random. The missing data that are notmissing completely at random contain non-random elements thatmay prejudice theresults.

Imputation:In imputation-based proceduresmissing valuesareimputedwithreasonable,probablevaluesrather than being deleted totally. The objective is to useknown associations that can be recognized in the valid rangevaluesofthedatasettofacilitateinestimatingthe missing values

Data quality playsmajor role in machine learning, data mining andknowledge discovery from databases. Machine learningalgorithmshandle missing data in a quite naive way. To avoid biasingin induced hypothesis missing data treatment shouldbe carefullyhandled.Imputationisaprocessthatreplacesthe missing values in instance by some reasonable values.In first method limitation of ΔT was adjusted, above equation was used for $\Delta T<8^oC$. Table 1 shows criteria for the decision of τ value. τ value of 0.6-0.7 are commonly used for clear sky atmospheric transmittance coefficient value. In this study τ value of 0.69 was used for clear sky, assumed that the clear sky condition occurred when RH<40% and ambient temperature more than 8^oC. Calculation algorithm was built based on decision matrix and the τ value was locally determined using the training of data set to get minimum error.

The second method used in this study is by finding the correlation between RH, clearness index and beam transmittance. The data used to find correlation between beam transmittance and clearness index is measured data from new radiometer set that was installed in 2010 on the rooftop of Block P, UniversitiTeknologiPetronas, which located about 30 km from Ipoh city. About 1 month, 5 minutes time step data of global, beam and diffuse radiation from June to July 2010 was used. Before find the correlation of beam transmittance and clearness index, RH-clearness index correlation was obtained from Ipoh city available data as can be seen in Figure 3.

Then beam transmittance-clearness index correlation can be obtained by scatter plot as can be seen in Figure 4. To plot Figure 4 some data were rejected due to obvious error that can be analyzed from measurement results, and the basic concept of terrestrial solar radiation characteristicsTherearetwocoremethodstodiscarddata with missing values. The first one is known as completecase analysis.Itisavailableineveryoneofstatisticalpackagesand is the default method in manyprograms

MultipleImputations: ItisamethodbyRubin[1]for making multiple simulated values for eachincomplete information, and then iteratively examining datasets witheach simulated value substituted in every turn. The intentionis,possibly, to generate estimates that better indicatetrue variance and uncertainty in the data than doregression methods. This permits expert staff and software to be usedto create imputed datasets that can be analyzed byrelatively naive users equipped with standard software. It can bevery effective particularly for small to moderate levelsof missingness, where the missing data mechanism isorganized, and for datasets that are to be placed in the publicdomain

Not missingat random is the most challenging form, occurs whenmissing values are not randomly distributed across observations. Itisalso called as non-ignorable missingness. The probabilityof an instance with a missing value for an attributemight depend on the value of thatattribute

Following constraint were used as data rejection criteria:

- Reject night data
- Reject data if clearness index >1
- Reject data if beam transmittance >1
- Reject data if beam transmittance > clearness index
- Reject data if clearness index >0.6 and beam transmittance <0.1
- Reject data if clearness index <0.2 and beam transmittance>0.15

Correlation between RH and beam transmittance was obtained from above correlation and plotted in Figure 5. Balaras et al studied the relationship between beam transmittance and clearness index in Athens, Greece [10], the results of the study was adopted to carry out second method in this study. Regression results were presented as follows:

Therearevariousmethodstoestimatesolarradiation. Satisfactoryresultforhourlysolarradiationestimationwas obtainedbyusingatmospherictransmittancemodel[1]while otherauthorshaveuseddiffusefraction[2]andclearness indexmodels[3].Parametricoratmospherictransmittance modelrequiresdetailsatmosphericcharacteristicinformation [4].Thismodelgiveshigh-accuracyforclearsky/cloudless conditions,whichisleadingsomeauthor tousethis modelto evaluatetheperformanceofanempiricalmodelunder cloudlessconditions[5].Therearenumerousauthors proposedthiskindofmodelasmentionedin[6].However, pureparametricmodelwasnotused in thisstudy,since there is no detail atmospheric condition data for the site.

Meteorologicalparametersfrequentlyusedaspredictorsof atmosphericparameterssinceacquiringdetailatmospheric conditionsrequireadvancemeasurement.Meteorological parameterssuchassunshineduration,cloudcover,ambient temperature,relativehumidity,andprecipitationdatahave beenusedtoestimateatmospherictransmittancecoefficient inparametricmodel.Thiskindofmodeliscalled meteorological model

Therearevariouswaystoestimatesolarradiationon certainareaontheearth.Ambienttemperaturebased estimation is widely used sinceambienttemperaturedataare measuredinmanyweatherstations.Inthisstudy,missing datawereestimatedbasedonambienttemperature measurementandusedmeasuredRHdataasatmospheric transmittance of the determination criteria.thatestimate hourlysolarradiationbasedondevelopedCampbelland Norman method [2], was adapted in this study.

Results of the new model then compared with the existing temperature-based solar radiation prediction model as follow:

A. H-S model

Hargreaves and Samani [14] conducted an initial study on using T_{max} and T_{min} to estimate solar radiation by the following equation:

$$G_{Th} = K_r (T_{max} - T_{min})^{0.5} \ G_{0h} \quad (9)$$

$K_r$is an empirical coefficient, which was recommended tobe 0.16 for interior regions and 0.19 for coastal regions. In this study K_r was locally determined using training data set.

B. H-S-A model

Annandale et al [15] modified H-S model by introducing correction factor as follow:

$$G_{Th}=K_r \ (1+2.7 \ 10^{-5}Z) \ (T_{max}-T_{min})^{0.5} \ G_{0h} \quad (10)$$

Z is elevation in *m* and $K_r$was locally determined.

V. STATISTICAL ANALYSIS FOR MODEL VALIDATION

Estimation results validated using statistical parameters. Pearson correlation coefficient was calculated as routine correlation indicator. Residual error was calculated using RMSE (Root Mean Square Error) and also presented in NRMSE (Normalized Root Mean Square Error) as follows:

$$RMSE = [\Sigma\{Y_c - Y_0\}^2/n]^{0.5} \quad (11)$$

$$NRMSE = RMSE/y_{max} - y_{min} \quad (12)$$

where, Y_c is predicted variable Y_0 is measured variable, n is number of data, y_{max} is maximum measured data y_{min} is minimum measured data.

As an addition, index of agreement was calculated using equation below:

$$d = 1 - [\Sigma(x_i-y_i)^2/\Sigma(|x_i-\overline{x}_i|+|y_i-\overline{y}_i|)]^{0.5} \quad (13)$$

where, x_i is predicted variable y_i is measured variable, $\overline{x_i}$ is averaged predicted variable and $\overline{y_i}$ is averaged measured variable.

VI. RESULTS AND DISCUSSIONS

Calculation have been carried out using methods 1 and 2, statistical calculation analysis also has been performed. Table 2 shows statistical analysis results of both methods and results of existing method (H-S and H-S-A method).The most satisfactory results were obtained using method 1. Figure 3 shows scatter plot of predicted and measured data for first method. The minimum RMSE value of 87.6 Watt/m^2 was obtained with 0.95 correlation coefficient and 0.97 index of agreement value. Previous method which use precipitation data obtained averaged index of agreement of 0.95, thus the model presented in this study also performed well.

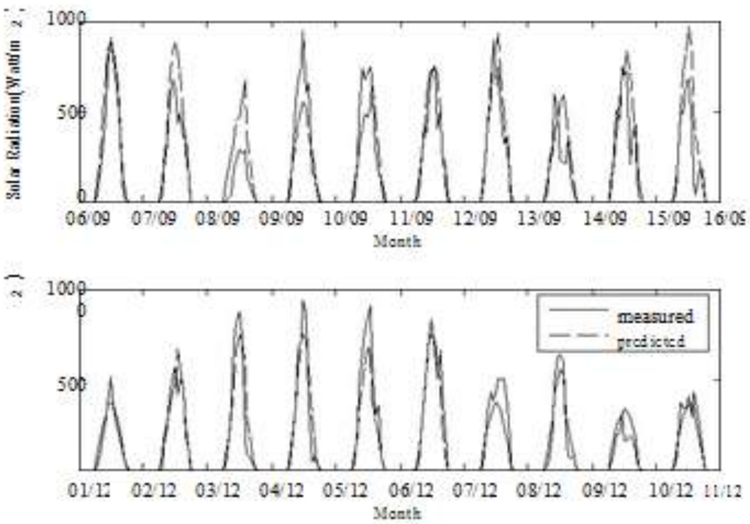
Table II. Statistical analysis results

Themethodsarebasedonthedecompositionmodelthatiscalculatingeachofsolarradiationcomponents,which dependonatmospherictransmittance.Twomethodstopredictatmospherictransmittancevalueusingavailablemeteorol ogicaldatawereproposed.Inthefirstmethod,adecisionmatrixwasused,whileinthesecondmethod,regressioncorrelati onofmeteorologicalparameterswasused.Thecalculationsresultswereevaluatedusingstatisticalparameter.Thoughth eresultshowsbothofthemethodsperformwell,moresatisfactoryresultswereobtainedfromfirstmethodwithRootMean SquareErrorof87.6Watt/m^2,NormalizedRootMeanSquareErrorof8.29%,correlationcoefficientof0.95andindexofa greementof0.97.Furthermore,thefirstmethodonlyneedsambienttemperatureandrelativehumiditydatathatcommonl ymeasuredinmeteorologicalstations.graphical comparison of measured and predicted solar radiation for random dates. As can be seen most of the error occurred in low radiation or cloudy sky which often occurred in rainy season (Fig. 7-b). Then after the results has been validated using available data, the predicted results in the missing data days can be occupied in the measured data set. Now the time series composite data of solar radiation is completed and can be used for other purpose such as prediction of PV system performance for respected location.

Missing data is one of the major issues in data miningand pattern recognition. The knowledge contains in attributeswith missing data values are important in improvingdecision-makingprocessofanorganization.Thelearningprocesson each instance is necessary as it may containsome exceptional knowledge. There are various methods tohandle missing data in decision tree learning. Analysis of photovoltaic cell, Sunlight striking on different geographical location to know the defective or connectionless photovoltaic cell plate. we mainly aims To showcasing the energy produced at different geographical location and to find the defective plate.

And also analysis the data sets of energy produced along with current weather in particular area to know the status of the photovoltaic plates. In this project, we used Hadoop Map-Reduce framework to analyze the solar energy datasets.

Not missingat random is the most challenging form, occurs whenmissing values are not randomly distributed across observations. Itisalso called as non-ignorable missingness. The probabilityof an instance with a missing value for an attributemight depend on the value of thatattribute



(b) Rainy season

Figure 7: Graphical comparison between measured and predicted solarradiation for random dates (Method 1)

Figure 8 shows prediction results in the days which solar radiation measurement were absent. In our case the amount of missing solar radiation data is 21 full days and 2 half days.

Data quality playsmajor role in machine learning, data mining andknowledge discovery from databases. Machine learningalgorithmshandle missing data in a quite naive way. To avoid biasingin induced hypothesis missing data treatment shouldbe carefullyhandled.Imputationisaprocessthatreplacesthe missing values in instance by some reasonable values.The case substitution is the method developed for dealingwith missingdataininstancesanditishavingsomedrawbackswhen applied to the data mining processesThe method developed in this study may be applied for any location on the earth with notes, for first method the assignment criteria of atmospheric transmittance using RH and ambient temperature should be adjusted to the available solar radiation data of the area to get minimum error. To generate general criteria of atmospheric transmittance assignment using RH and ambient temperature further research is required with sufficient large amount of data for various areas. For the second method the correlation should be rebuild based on available measurement nearest from the location to give satisfactory estimation results.

VII. CONCLUSIONS

This paper represents to estimate solar radiation from available weather data. The methods presented in this paper are based on atmospheric transmittance determination using available meteorological data. In this method regression correlation of meteorological parameter was used.Thekeyfortheaccuracyofabovemethodisinthe determinationofbeamatmospherictransmittance(τ).Beam atmospherictransmittanceisthepercentageofthebeam (direct)radiationthatwillpenetratetheatmospherewithout beingscattered.KurtandSpokasusedprecipitationdata to built decisionmatrix ofatmospheric transmittance .temperature data. The second method is by using RH-clearness index, clearness index-beam atmospheric transmission and beam atmospheric transmission-RH correlation. The result shows that both methods perform well. Method 1 provided better results with minimum correlation coefficient of 0.95, RMSE of 87.6 Watt/m^2, NRSME of 8.29% and index of agreement of 0.97. The prediction was intended to fill missing data in solar radiation data set to get complete time series data. However, in this study only one year of one area data have been used. Validation using sufficient large amount of data is required for wider application of the method.

REFERENCES

[1]. G. S. Campbell and J. M. Norman, “Introduction to Environmental Biophysics. 2nd ed. New York: Springer-Verlag. Pp. 167–183, 1998
[2]. Guofeng Wu, et al. “Methods and strategy for modeling daily global solar radiation with measured meteorological data – A case study in Nanchang station, China”, Energy Conversion and Management 48, 2447-2452, 2007
[3]. Yang K, Koike T. Estimating surface solar radiation from upper-airhumidity. Solar Energy, 72(2):177–86, 2002
[4]. N. Mohan Kumar et al, “An empirical model for estimating hourly solar radiation over the Indian seas during summer monsoon season”, Indian Journal of Marine Sciences, Vol. 30, pp 123-131, 2001
[5]. Reindl DT, Beckman WA, Duffie JA, “Diffuse fraction correlations”, Solar Energy 1990; 45:1-7
[6]. F.J. Batlles et al, “Empirical modeling of hourly direct irradiance by means of hourly global irradiance”, Energy 25: 675-688, 2000
[7]. Kurt Spokas and Frank Forcella, “Estimating hourly incoming solar radiation from limited meteorological data”, Weed Science, 54:182– 189, 2006
[8]. Hunt LA, Kuchar L, Swanton CJ. Estimation of solar radiation for use in crop modelling. Agric. Forest Meteorol. 91(3–4):293–300, 1998
[9]. Louche et al, “Correlations for direct normal and global horizontal irradiation on French Mediterranean site”, Solar Energy 46: 261-266, 1991 as cited by [8]
[10]. Athens, Greece. et al., “On the relationship of beam transmittance on clearness index for Athens, Greece”, Int. J. Solar Energy, Vol. 7, pp 171-179, 1989