J. Nig. Ass. Math. Phys.
Vol. 1 (1992) 61 - 69

SOLAR RADIATION MODELS IN NIGERIA: A CASE FOR VALIDATION OF RESULTS

by

A. A. L. Maduekwe
Department of Physics
University of Lagos
Lagos, Nigeria

ABSTRACT

Solar radiation regression models developed for the Nigerian environment are rarely validated. Authors of such models appear to ignore the dangers of publishing models which are not validated. A case study is made with monthly mean data for a period of ten years in Sokoto, Nigeria. Regression models were created for predicting solar radiation which is horizontal when it reaches the earth's surface. Two methods of validation were used: (a) the collection of fresh data, and (b) data splitting or cross-validation. The results depend on whether prediction data sets are different from estimation data sets. The method of data splitting introduces difficulties when the estimation data set and the prediction data set differ in predictive performance and coefficient estimates.

1. INTRODUCTION

Regression models are used extensively to predict horizontal solar radiation reaching the earth's surface. Several of such models have been developed for different locations in Nigeria by different workers [1,2,3]. These regression models, which have been developed for daily and monthly mean solar radiation, are based on the Angstrom formulation [4], which can be written as

$$Kt = a + bS \qquad (1)$$

where Kt is the ratio of the measured to the extraterrestrial solar radiation, S is the ratio of the measured sunshine duration to the calculated day length, and the constants a and b have to be evaluated for the specific location since they are site dependent. Regression models are used for prediction or estimation, data description, parameter estimation, and control. Model validation provides a measure of protection for both model developer and user. There are several methods in use for the validation of models, which include (a) fresh data collection, (b) data splitting or cross-validation, and (c) data from planned experiments. In this paper, techniques (a) and (b) are examined which, in our opinion, are adequate for most of the model development in the field of solar energy in Nigeria.

2. VALIDATION TECHNIQUES

2.1 FRESH DATA COLLECTION

In this technique, model predictions are compared directly with the

fresh data. If the model gives accurate predictions of new data, the user will have greater confidence in both the model and the model building process. At least 15 to 20 new observations are required, in order to give a reliable assessment of the model's prediction performance. In situations in which two or more models have been developed from the data, comparison with fresh data may provide a basis for final model selection.

2.2 DATA SPLITTING

Several stations in Nigeria, for example, are unable to continue with the measurement of solar radiation, due to the breakdown of equipment. When collection of fresh data is not possible, an acceptable procedure is to split the available data into two parts, which are called estimation data and prediction data [5]. Estimation data are used to build the regression model and, following this, prediction data are used to study the predictive ability of the model. A variety of methods can be used for data splitting, and these are described in [5]. The procedure described in this paper is that of splitting the data into estimation and prediction data sets.

A disadvantage of data splitting is that it reduces the precision with which the regression coefficients are estimated. That is, the standard errors of the regression coefficients obtained from the estimation data set will be larger than they would have been if all the data had been used to estimate the coefficients. In large data sets, the standard errors may be small enough that this loss in precision can be ignored. However, the percentage increase in the standard errors can be large. If the model developed from the estimation data set is a satisfactory predictor, one way to improve the precision of estimation is to re-estimate the coefficients using the entire data set. The estimates of the coefficients in the two analyses should be similar if the model is an adequate predictor of the prediction data set. Another method which is described in this paper is double cross-validation. In this procedure, the roles of the estimation and prediction data sets are reversed. An advantage of this procedure is that it provides two evaluations of model performance, but a disdavantage is that there are now three models to choose from: two developed from data splitting, and the model fitted to all the data. If the model is a good predictor, it will make little difference which one is used, except that the standard errors of the coefficients in the model fitted to the total data set should be smaller. If there are major differences in predictive performance, coefficient estimates, or functional form for these models, then further analysis is necessary in order to discover the reasons for the differences.

3. DATA COLLECTION AND ANALYSIS

The data sets used in the present work were collected from the meteorological station at the Sultan Bello International Airport, Sokoto, Nigeria. They include data on monthly mean values of solar radiation, measured with a Gunn-Bellani radiometer; sunshine hours, measured with a Stokes-Campbell sunshine recorder; and ambient temperature, recorded using a mercury-in-glass thermometer. The data

cover the period from 1981 to 1990 inclusive, except that data for 1983 were not available. The data used in the analysis, after reduction, were a total of 113. The Gunn-Bellani data were converted to W m^{-2} using the factor given in [6], before the clearness index Kt was calculated. A graph of Kt versus S is shown in figure 1. Linear regression analyses were performed on the data, and the results analyzed.

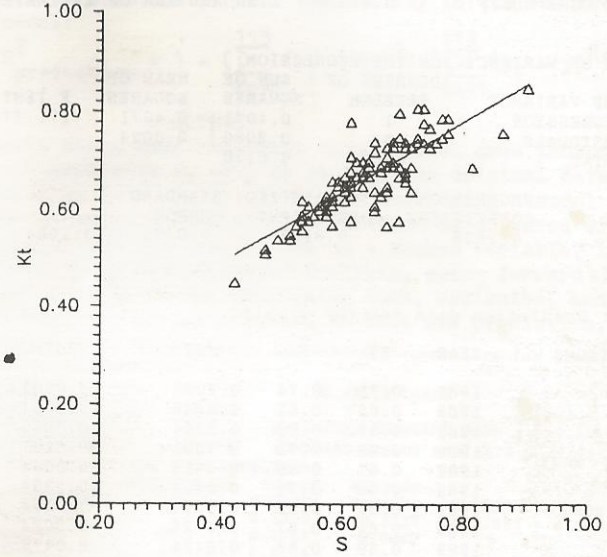

Fig. 1: Clearness index vs S for the complete data set used in the analysis.

4. REGRESSION ANALYSES AND VALIDATION

4.1 ANALYSIS WITH FRESH DATA COLLECTION

The data set was split such that data from 1981 to 1988 represented the estimation data set, while data for 1989 and 1990 were used as fresh data for prediction. The number of data used for building the model was 90, while the number used for prediction was 23. The results of the analysis are presented in table 1. The prediction data set is shown in table 2, along with the estimates and residuals. The average prediction error is

$$\left(\sum_{i=91}^{113} (y_i - x_i)\right)/23 = -0.0074 \qquad (2)$$

where y_i and x_i are, respectively, the measured and estimated values for the i-th observation. This value of average prediction error is quite low, implying that the model produces approximately unbiased predictions. Observation 10 (see table 2), has the highest error. This can be explained from the fact that the measured value

Table 1: Regression analysis for example on fresh data collection.

*** MULTIPLE LINEAR REGRESSION ***

DEPENDENT VARIABLE: KT		90 VALID CASES	
COEFF OF DETERM:	0.6612		
ADJUSTED R SQUARE:	0.6574	ESTIMATED CONSTANT TERM:	-0.0027
MULTIPLE CORR COEFF:	0.8132	STANDARD ERR OF ESTIMATE:	0.0487

ANALYSIS OF VARIANCE FOR THE REGRESSION:

SOURCE OF VARIANCE	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN OF SQUARES	F TEST	PROB
REGRESSION	1	0.4071	0.4071	171.777	0.0000
RESIDUALS	88	0.2086	0.0024		
TOTAL	89	0.6156			

VARIABLE	REGRESSION COEFFICIENT	STANDARDIZED COEFFICIENT	STANDARD ERROR	T	PROB
S	0.9628	0.8132	0.0735	13.1064	0.0000

Table 2: Prediction data set for s only

OBSERVATIONS	YEAR	KT	S	EST	RESIDUAL
1	1989	0.75	0.74	0.7098	0.0402
2	1989	0.65	0.61	0.5846	0.0654
3	1989	0.57	0.59	0.5654	0.0046
4	1989	0.68	0.73	0.7001	-0.0201
5	1989	0.67	0.69	0.6616	0.0084
6	1989	0.53	0.59	0.5654	-0.0354
7	1989	0.55	0.59	0.5654	-0.0154
8	1989	0.64	0.67	0.6424	-0.0024
9	1989	0.69	0.67	0.6424	0.0476
10	1989	0.69	0.80	0.7675	-0.0775
11	1989	0.58	0.66	0.6327	-0.0527
12	1989	0.73	0.81	0.7772	-0.0472
13	1990	0.64	0.69	0.6616	-0.0216
14	1990	0.73	0.74	0.7098	0.0202
15	1990	0.72	0.81	0.7772	-0.0572
16	1990	0.67	0.71	0.6809	-0.0109
17	1990	0.73	0.78	0.7483	-0.0183
18	1990	0.61	0.68	0.6520	-0.0420
19	1990	0.72	0.75	0.7194	0.0006
20	1990	0.67	0.65	0.6231	0.0469
21	1990	0.76	0.75	0.7194	0.0406
22	1990	0.51	0.55	0.5268	-0.0168
23	1990	0.57	0.62	0.5942	-0.0242

for this observation, shown in figure 1, is an outlier. The low prediction errors give the user some confidence in using the model. One can also check the residual mean square error (MSE) for the estimation data set against the average square prediction error. We have

$$MSE = \sum (y_i - x_i)^2 = 0.0024 \qquad (3)$$

The average square prediction error is

$$\left(\sum_{i=91}^{113} (y_i - x_i)^2\right)/23 = 0.0014 \qquad (4)$$

This result (eq (4)), which can be thought of as the average variance of the residuals from the fit, shows that, indeed, the least square model employed does predict the data as well as it fits the existing data, because of the lower value of the average square prediction. From this, we conclude that the least square model will be a successful predictor. It is also useful to compare R^2 from the least square fit (0.6574 and a correlation coefficient R of 0.8132 from table 1) with the percent variability in the new data explained by the model:

$$R^2_{\text{prediction}} = 1 - \left(\sum_{i=91}^{113} (y_i - x_i)^2\right) / \left(\sum_{i=91}^{113} (y_i - y)^2\right)$$

$$= 0.7298 \qquad (5)$$

This result shows that the least square model does indeed predict the new observations as well as it fits the original data. With a result such as this, is it worthwhile developing a model which incorporates an extra variable? This question is answered with the introduction of ambient temperature as a second variable, in addition to S. The results of regression analysis, using forward elimination, are shown in table 3. The prediction data, estimates, and errors are shown in table 4. Errors obtained in this new prediction, with tem-

Table 3: Regression analysis for S and ambient temperature for fresh data collection.

DEPENDENT VARIABLE: KT — 90 VALID CASES

COEFF OF DETERM:	0.7236		
ADJUSTED R SQUARE:	0.7172	ESTIMATED CONSTANT TERM:	0.1540
MULTIPLE CORR COEFF:	0.8506	STANDARD ERR OF ESTIMATE:	0.0442

ANALYSIS OF VARIANCE FOR THE REGRESSION:

SOURCE OF VARIANCE	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN OF SQUARES	F TEST	PROB
REGRESSION	2	0.4455	0.2227	113.880	0.0000
RESIDUALS	87	0.1702	0.0020		
TOTAL	89	0.6156			

VARIABLE	REGRESSION COEFFICIENT	STANDARDIZED COEFFICIENT	STANDARD ERROR	T	PROB
S	1.1053	0.9335	0.0741	14.9197	0.0000
TEMP	-0.0072	-0.2772	0.0016	-4.4301	0.0000

Table 4:Prediction data set for s and ambient temperature.

OBSERVATIONS	YEAR	KT	S	TEMP	EST	RESIDUAL
1	1989	0.75	0.74	35.5	0.7168	0.0332
2	1989	0.65	0.61	31.8	0.5997	0.0503
3	1989	0.57	0.59	35.1	0.5539	0.0161
4	1989	0.68	0.73	38.8	0.6821	-0.0020
5	1989	0.67	0.69	31.3	0.6917	-0.0217
6	1989	0.53	0.59	31.2	0.5819	-0.0519
7	1989	0.55	0.59	32.2	0.5747	-0.0247
8	1989	0.64	0.67	36.2	0.6344	0.0056
9	1989	0.69	0.67	28.5	0.6897	0.0003
10	1989	0.69	0.80	37.4	0.7695	-0.0795
11	1989	0.58	0.66	32.9	0.6471	-0.0671
12	1989	0.73	0.81	40.6	0.7575	-0.0275
13	1990	0.64	0.69	38.6	0.6393	0.0007
14	1990	0.73	0.74	33.0	0.7348	-0.0048
15	1990	0.72	0.81	40.9	0.7554	-0.0354
16	1990	0.67	0.71	37.4	0.6700	0.0000
17	1990	0.73	0.78	35.6	0.7603	-0.0303
18	1990	0.61	0.68	34.1	0.6606	-0.0506
19	1990	0.72	0.75	37.1	0.7164	0.0036
20	1990	0.67	0.65	33.5	0.6317	0.0383
21	1990	0.76	0.75	36.6	0.7200	0.0400
22	1990	0.51	0.55	31.6	0.5348	-0.0248
23	1990	0.57	0.62	32.0	0.6093	-0.0393

perature as an extra variable, are slightly higher in absolute value. The average prediction error in this case is -0.0118, which is further from zero than that for the model with S only. MSE = 0.0013 for the latter model, which is slightly better than that for the model with S only. The predicted R^2 = 0.7548. This value of R^2 is 3.43% better than the prediction value found with the first model. Because of the small increase in performance over the predictive ability of the first model, it can be said that not much is gained by the addition of temperature as a predictor. The same analysis has to be carried out with other variables before incorporating them into models. There are several methods of doing this such as the forward or backward selection methods or the stepwise regression method [5,7].

4.2 ANALYSIS WITH DATA SPLITTING

In the use of the data splitting technique for validation, it is assumed that the data represent the total available from the location, without the possibility of gathering fresh data. The data has therefore been split into two nearly equal parts, with the first set consisting of 57 data points and the second consisting of 56 data points. The data set with 57 data points has been used, initially, as the estimation data while that with 56 data points is used as the prediction data. Figure 2 shows graphs of these two data sets, with their regression lines. Following similar procedures as in the fresh data collection method, table 5 shows results of regression analysis for the 57 data points with S only. The MSE in this analysis is 0.0029, while the average square prediction error is 0.0015. Thus,

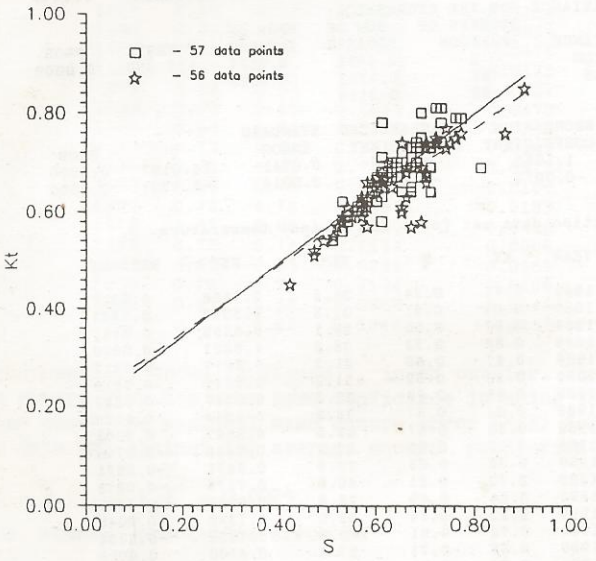

Fig. 2: Clearness index vs S for two data sets. The solid line is the best fit for the set with 57 data points. While the dashed line is for the set with 56 data points.

Table 5: Regression analysis for the estimation data with 57 data points in the data splitting example.

*** MULTIPLE LINEAR REGRESSION ***

DEPENDENT VARIABLE: KT 57 VALID CASES
COEFF OF DETERM: 0.4859
ADJUSTED R SQUARE: 0.4766 ESTIMATED CONSTANT TERM: 0.0563

MULTIPLE CORR COEFF: 0.6971 STANDARD ERR OF ESTIMATE: 0.0540

ANALYSIS OF VARIANCE FOR THE REGRESSION:

SOURCE OF VARIANCE	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN OF SQUARES	F TEST	PROB
REGRESSION	1	0.1516	0.1516	51.989	0.0000
RESIDUALS	55	0.1604	0.0029		
TOTAL	56	0.3120			

VARIABLE	REGRESSION COEFFICIENT	STANDARDIZED COEFFICIENT	STANDARD ERROR	T	PROB
S	0.8774	0.6971	0.1217	7.2104	0.000

Table 6: Regression analysis for the estimation data with 56 data points in the data splitting example.

*** MULTIPLE LINEAR REGRESSION ***

DEPENDENT VARIABLE: KT 56 VALID CASES
COEFF OF DETERM: 0.8191
ADJUSTED R SQUARE: 0.8157 ESTIMATED CONSTANT TERM: -0.0185

MULTIPLE CORR COEFF: 0.9050 STANDARD ERR OF ESTIMATE: 0.0376

ANALYSIS OF VARIANCE FOR THE REGRESSION:

SOURCE OF VARIANCE	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN OF SQUARES	F TEST	PROB
REGRESSION	1	0.3449	0.3449	244.489	0.0000
RESIDUALS	54	0.0762	0.0014		
TOTAL	55	0.4211			

VARIABLE	REGRESSION COEFFICIENT	STANDARDIZED COEFFICIENT	STANDARD ERROR	T	PROB
S	0.9769	0.9050	0.0625	15.6361	0.0000

the average variance of the residuals from the fit is low. The average prediction error is -0.0069, which is very near to zero, indicating approximately unbiased predictions. The value of R^2 for the estimation data set was found to be 0.4766 while that found from the prediction data set was 0.8042, a result which is 68.74% better than that for the estimation data. The method of cross-validation has been employed to examine whether the prediction data set would do as well when used as the estimation data set. Table 6 shows results of regression analysis performed with this data set of 56 data points. Analysis of the results using 57 data points as prediction data set shows that while the MSE for the estimation data set was 0.0014, average square prediction error was 0.0029, which is exactly the result when the roles of the data sets are reversed. The average prediction error using this data set as the prediction data was 0.0098, which is higher than what was obtained the other way round. The value of R^2 for the estimation data set was 0.8157 while that for the prediction data set was 0.4622. There is a reversal of roles, which is undesirable, as it introduces confusion as to which of the models should be used, although the standard error of the initial estimation data set of 57 data points (0.054) is higher than when the roles are reversed. In that case, the data set with 56 data

points gave a standard error of 0.038, which is 42.11% better. This means that the second model predicts the other data set as well as it fits the estimation data set. A possible explanation of the wide difference in the predictive performance, coefficient estimates, and functional forms of these models could be traced to the fact that several of the data points in the set with 56 data points are actually extrapolation points for the estimation model with 57 data points, as can be verified from figure 2. However, this is not the case for the set with 56 data points, when used as the estimation data, in which case the 57-point prediction data set are just interpolation points. A check is made of the regression analysis performed with the whole data set and the results are presented in table 7. R^2 for the whole data set is 0.6739, which lies somewhere in between those obtained when the data set was divided into estimation and prediction sets. If the model developed with the original estimation data set is a good predictor, it will make little

Table 7: Regression analysis for all the data points.

*** MULTIPLE LINEAR REGRESSION ***

DEPENDENT VARIABLE: KT — 113 VALID CASES

COEFF OF DETERM:	0.6768		
ADJUSTED R SQUARE:	0.6738	ESTIMATED CONSTANT TERM:	0.0179
MULTIPLE CORR COEFF:	0.8227	STANDARD ERR OF ESTIMATE:	0.0465

ANALYSIS OF VARIANCE FOR THE REGRESSION:

SOURCE OF VARIANCE	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN OF SQUARES	F TEST	PROB
REGRESSION	1	0.5024	0.5024	232.418	0.0000
RESIDUALS	111	0.2399	0.0022		
TOTAL	112	0.7423			

VARIABLE	REGRESSION COEFFICIENT	STANDARDIZED COEFFICIENT	STANDARD ERROR	T	PROB
S	0.9297	0.8227	0.0610	15.2453	0.0000

difference which of the three models is to be used, although the standard error of the model obtained with the whole data set should be smaller. But that was not the case, as a standard error of 0.046 was obtained for the whole data set. This is more than that found when the data set with 56 points was used as estimation data. From these results, we conclude that before a choice is made between the three contesting models, further analysis is necessary in order to uncover reasons for these differences.

5. CONCLUSIONS

The results obtained from the test case show that models should be validated before they are applied to scientific problems. Situations such as are encountered in the data splitting example are often met by model builders in the field of solar energy, but are often ignored to the detriment of the end users of such models. From our results, validation with the collection of fresh data is preferred over the method of data splitting. It is suggested that workers should plan for additional data collection, to extend beyond the initial period of data collection.

REFERENCES

1. A. S. Sambo, (1989) "The measurement and prediction of global and diffuse components of solar radiation for Kano in northern Nigeria", Solar Wind Technology, 5, pp1 - 6
2. C. I. Ezekwe & C. O. Ezeilo, (1981) "Measured solar radiation in a Nigerian environment compared with predicted data", Solar Energy, 26, pp181 - 186
3. J. O. Ojosu & L. K. Komolafe, (1987) "Models for estimating solar radiation availability in south western Nigeria", Nig. J. Solar Energy, 6, pp69 - 77
4. M. Iqbal, (1983) "An introduction to solar energy", Academic Press, Canada, pp231 - 235
5. D. C. Montgomery & E. A. Peck, (1982) "Introduction to linear regression analysis", John Wiley, New York, pp424 - 443
6. J. O. Ojosu, (1990) "On the bounds for global solar radiation estimates from sunshine hours for Nigerian cities", National Solar Energy Forum (NASEF 90), Sokoto, Nigeria, June 27 - 30, 1990
7. N. R. Draper & H. Smith, (1981) "Applied regression analysis", John Wiley, New York