PREDICTION OF RAINFALL WITH A LINEAR REGRESSION FOR MULTIPLE WEATHER DATA-VARIABLES BY INCORPORATING THE WEIGHTED MOVING AVERAGE FILTER

Ruhiat Sultana, Mehveen Mehdi Khatoon and Muneeba Zuha
Department of Information Technology, Bhoj Reddy Engineering College for Women, India

Abstract

Factors like traffic volume, the make and model of the vehicles, and driver behaviour are just as crucial to a road's operational performance and safety as the weather. Weather conditions like fog and rain have an impact on visibility, which in turn affects how frequently accidents happen on a given road or the likelihood of getting into an accident. In this study, meteorological data were analysed using multiple linear regression to forecast precipitation and visibility for the benefit of various stockholders. For the purpose of analysing the data (10 years with 4018 samples), the mean square error (MSE) and its rooted version (RMSE), mean absolute error (MAE), and R-squared are used. As a result, the only way to judge the precision of models is through residuals. The empirical findings can be used by practitioners. The findings that were discovered to produce forecasts for the visibility and the rainfall.

Keywords:

Linear Regression, Prediction Accuracy, Weather Data, Correlation

1. INTRODUCTION

The weather is just as important for analysing a road's operational performance and safety as other factors like the volume of traffic, the characteristics of the vehicles, and the behaviour of the drivers. weather circumstances like Fog and rain impair visibility, which raises the likelihood of being in an accident as well as its frequency and severity. To be more exact, visibility diminished. Weather-related accidents might involve wet or slippery pavement after rain, dew, haze, mist, or snow, or in severely low conditions, mushy or snowy pavement after it has snowed.

A lot of studies have looked into how the weather affects traffic accidents. Fog contributes to around 2% of all road accidents [1], whereas on slick roads, weather-related crashes account for around 75% of accidents. As per author [2] studies on weather-related fatal collisions discovered that the majority of these occurrences occurred on roadways in the wee hours of the morning during the coldest seasons of the year. Also, they discovered that if the relevant government agency hasn't issued a visibility-related caution, a startling 70% of fatal weather-related incidents occur. Various further research has been done on how fog or smoke affect accidents between vehicles; They include [1]-[7], these studies' findings suggest that driving in the fog might be risky. Since drivers have a harder difficulty seeing their surroundings, including other cars, pedestrians, cyclists, and immovable objects, reduced visibility may be to blame for the increase in hazard. The fog makes it difficult to comprehend information and take proper action. Wet, slippery, slushy, snowy, muddy pavement or pavement polluted with debris, dust, or sand can have catalytic effects on braking performance and an increased chance of crashes. When eyesight abruptly fades, the likelihood of being in an accident on a road increase.

Past studies and statistics that are generally available indicate that driving through fog can increase the frequency or severity of collisions. delays in visibility-related advisories, a sudden drop in visibility from fog, driving behaviour, and breaking the performance of tyres and pavement conditions all influence collision frequency and crash involvement probability, which both rise in the presence of fog. These problems might be avoided and road safety increased by giving early visibility-related advice. It requires time and money to install state-of-the-art weather monitoring equipment at regular intervals to record the condition of the atmosphere (including visibility) or accurately estimate visibility, which restricts the timely nature of information on visibility.

Before embarking on a journey, or at any time while travelling, drivers could be notified of the current or expected visibility situation to alter the driver's actions. A model for predicting visibility at the level of a road link, however, has not been adequately assessed. It is crucial that practitioners build and assess the applicability of the visibility prediction model to serve as a visibility-related travel suggestion to drivers because they cannot regularly install technology or weather monitoring devices.

The time of day, recent precipitation, and the existence of water close to a road connection are all potential reasons of fog or decreased visibility. Dynamic changes in visibility over short time intervals are explained by the rainfall in the preceding hours, whereas the other two explanatory variables account for temporal and geographical dimensions. The historical lack of investigation into the impacts of these explanatory variables and other meteorological data on visibility prevented the development of a strong, with good predictability.

Prediction techniques recommended by [8]-[11] and reviewed by [12]-[13] are some of the best. By learning from the vast amount of meteorological data, some algorithms can model complex relationships and forecast visibility (hourly data from weather monitoring stations). As a result, the main goal of this research is to compare decision tree and linear regression models for precipitation prediction and assess how predicTable.it is on a short time scale. Practitioners can make use of the obtained empirical findings and conclusions to forecast visibility and precipitation at the road network, disseminate information via variable/dynamic message signs on highways or travel advisory software, and influence travel behaviour.

Weather forecasting has an impact on many areas of life, including business, transportation, crisis management, and energy management; these applications are addressed as: a) An industry is a group of businesses or manufacturers that produce certain

goods for the market and provide necessary services. The principal source of income for a company or group determines the industry classification. b) Agriculture/Food, Unquestionably, the need for food production increases as the population expands. Big data analytics and weather predictions can be used to raise the calibre of agricultural products. Farmers who can predict the weather can receive warnings about drought, overwatering, and soil erosion. Farmers can plan their crops precisely and foresee food prices by predicting precipitation. Additionally, weather forecasting assists supermarkets in arranging efficient stock control.

2. STATISTICAL SUMMARY

A vital process, weather forecasting affects many aspects of people's daily life and may have an impact on industries like agriculture, irrigation, and sea trade. It also has the potential to prevent many fatal incidents and save countless lives.

We may now talk about how these characteristics affect our climate. We know from our study of fluid mechanics that when discharged into the vacuum of space, air particles do not remain static but instead move quickly. More spreading occurs when there is greater space, which lowers pressure and promotes cooling. The radiation from the surface heats the air above the ground to a higher temperature than the upper atmosphere as the day goes on and the Earth warms.

The structure's height rises, and the hot air with low pressure rises and spreads fast over the area. The surrounding cold air rushes in to fill the void formed on the ground. A powerful windstorm develops when the air temperature climbs too high, creating a vacuum that cold air rushes in to fill. In the event of a nearby body of water, the heated air will rise along with the moisture it has absorbed.

Low temperatures in the upper atmosphere prevent dry air from travelling very far. Its dew point is distinct from the relative humidity at which it reaches full saturation with moisture (and can no longer hold any more). Rain will occur if the water droplets are large enough; otherwise, it will produce fog in the tropics or snow in the poles. The weather remains pleasant when the dew point and humidity are low. Therefore, the air's transparency can serve as a weather indicator, depending on factors like wind and precipitation.

2.1 STATISTICAL SUMMARY

The datasets use 4018 sample attributes in total as examples. As a result, it might occasionally be beneficial to express the entire data set as a single number. It represents the collection's average value. In statistical analysis, this average is referred to as the "central tendency."

The mean, median, or mode can all be used to calculate central trends. We place equal weight on those that describe the range or variability of the data values. We employ the standard deviation approach for this reason. A five-number summary, as shown in Table.1, is particularly useful in descriptive analysis and the initial examination of large data sets.

Five numbers make up a summary: the upper and lower quartiles, the median, the minimum and maximum values, and the mean each corresponding to six variable of weather data like cloud-cover, humidity, pressure, temperature, visibility and precipitation.

Table.1. Statistical summary of parameters

	Cloud cover	**Humidity**	**Pressure**	**Temp °C**	**Visibility**	**Precip MM**
mean	15.953	44.303	1009.86	30.96	9.757	3.093
std	18.418	18.934	6.721	7.449	0.615	8.731
min	0	5	993	11	5	0
25%	2	29	1004	25	10	0
50%	9	43	1010	32	10	0
75%	25	59	1013	36	10	1.9
max	98	96	1026	50	10	172.2

Summary statistics are extremely useful for comparing two separate projects or taking a before-and-after snapshot of time. Central tendency and dispersion measures are the two most typical categories of summary statistics used in evaluation. The compilation of a data set's key features is the definition of summary statistics. In addition to providing accompanying metrics, the values are defined. They are helpful for putting the data acquired in a study into context. Descriptive statistics can be used to get the mean of a set of variables.

2.2 CORRELATION CHART

We have identified the connections between the variables with the aid of the correlation matrix. The correlation matrix is an excellent tool for researching untested data. It has the capacity to determine the strength of any two variables' correlation. Whether or not there is an output is immaterial at this point; nonetheless, the system will compare each input to every other input. By measuring their correlation coefficient, as shown in Table.2, two numerical vectors can be compared to determine how similar they are to one another.

Table.2. Correlation matrix of parameters

	Cloud cover	Humidity	Pressure	Temp °C	Visibility	Precip MM
Cloud cover	1					
Humidity	0.547	1				
Pressure	-0.23	-0.0384	1			
Temp °C	-0.078	-0.348	-0.817	1		
Visibility	-0.5	-0.528	0.253	0.015	1	
Precip MM	0.445	0.469	-0.232	-0.008	-0.813	1

The Fig.1 to Fig.3 is symmetrical and shows all conceivable pairs of variable covariances. Multivariate data distribution magnitude and direction in n-dimensional space values in the covariance matrix serve as indicators. We can discover more about the two-dimensional dispersion of data by changing these factors. The covariance between two variables sheds light on how

they are related. Covariance measures how likely it is for two random variables in a data collection to change simultaneously.

Positive covariance between two variables shows that the two variables are positively correlated and fluctuate in the same way. By allowing for the influence of other co-varying continuous variables and testing for main and interaction effects of categorical variables on a continuous dependent variable, analysis of covariance is a powerful statistical technique which is mentioned in Table.3.

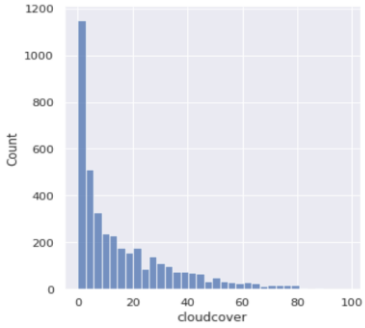

Fig.1. Distribution plot of cloud parameter

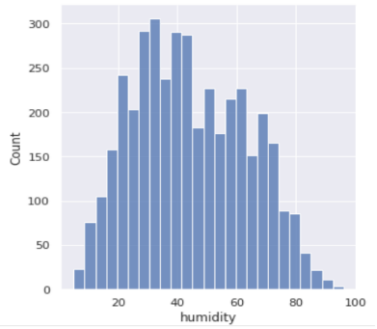

Fig.2 Distribution plot of humidity parameter

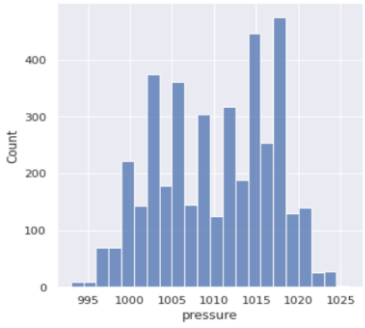

Fig.3. Distribution plot of pressure parameter

Table.3. Covariance Matrix of Parameters

	Cloud cover	Humidity	Pressure	Temp °C	Visibility	Precip MM
Cloud cover	339.15					
Humidity	190.77	358.41				
Pressure	-28.58	-4.89	45.17			
Temp °C	-10.8	-49.04	-40.94	55.47		
Visibility	-5.67	-6.154	1.04	0.07	0.379	
Precip MM	71.58	77.57	-13.61	-0.57	-4.38	76.21

3. REGRESSION LEARNER

The feature of Matlab known as Regression Learner instructs regression models to make accurate predictions based on data. We are able to explore pattern in data by following the steps like choosing features, establish validation schemes, train models, and evaluate results.

We are able to conduct automated training to search for the best regression model type, and the types of regression models that are eligible for this search such as linear regression models, regression trees, Gaussian process regression models etc.

Preprocessing and cleaning the data is essential before utilising this feature, as it is with every machine learning tool. This encompasses anything from ensuring the data is in the right format to ensuring it is all of the same magnitude to ensuring data is distributed fairly to handling duplicates or missing data. In the Parallel Computing Toolbox, we can try the basic and easiest option initially: training a regression learner model in parallel.

By selecting this option, a parallel pool will be created so that many models can be trained in parallel as we continue to work. If we check this box, the History list will display progress bars for each trained and queued model, and we'll be able to cancel specific models at any time. As a result of taking this strategy, less time will be needed for training. Verify that we have selected the appropriate validation method if the data set remains sizable.

Cross-validation is chosen automatically whenever a new session is opened in the Regression Learner feature once the data has been selected. When using cross-validation, the data is split into a certain number of groups (folds), the model is trained, and the average test error is determined over all folds. Even though this method involves repeated fits, it offers superior protection against over-fitting compared to other solutions and is hence suitable for small to medium data sets.

With validation, we can pick a subset of the data to utilise as a test group. The feature will use the training set to educate a model, then compare it against the testing set to determine how well it performed. Holdout validation works well with big data since the testing model is based on a subset of the total.

The feature allows us to train only the group models that we want to train, so if we know that the data is only function well in one type of group model or if one sort of group model is taking too long to train, we may disable or enable it as needed. We may then train all the models in the subset with the lowest RMSE to determine which one is the best.

A lot of attention is paid to amassing a large amount of data, but the real challenge lies in ensuring that the data we collect is accurate. There's a chance that you're holding on to old, irrelevant data that was collected years ago but has no use now.

4. NUMERICAL COMPUTATION

The proffered approach is done in MATLAB (2019b) with Regression Learner feature by selecting the models as need of

multiple linear regression evaluation. Estimating precipitation requires the usage of the simulation environment provided by MATLAB. For the purpose of this investigation, three distinct estimating models are developed by employing three distinct delay steps and three distinct filtering strategies on the weather data for one year. There is a range of possibilities for the number of delay steps, from three to ten. When it comes to filtering procedures, the moving average filter, the weighted moving average filter, and the exponential moving average filter are each applied in turn to the data.

Linear regression and tree techniques with linear and Gaussian kernels are used to create estimation models. These filtering procedures are performed to the dataset before the training phase of regression learning algorithms, with the normalisation step resulting in precipitation expressed as values between [0,1]. A cross-validation procedure that is multiplied by 10 is used during the training phase. Error values are measured across all scenarios, and the model with the smallest error is presented. x and y are D dimensional vectors, and x_i denotes the value on the i^{th} sample of x which is shown in Table.4 to Table.6 .

Table.4. Error Criteria

Criteria	Description	Formula
MSE	Mean Absolute Error	$\sum_{i=1}^{D} \lvert x_i - y_i \rvert$
MAE	Mean Squared Error	$\sum_{i=1}^{D} (x_i - y_i)^2$
RMSE	Root Mean Squared Error	$\sqrt{\sum_{i=1}^{D} (x_i - y_i)^2}$
R^2	Coefficient of Determination	1-(sum squared regression (SER)/total sum of squares (SST))

Table.5. Performance Measures of Tree Models

Tree Models			
Error Metrics	**Fine Tree**	**Medium Tree**	**Coarse Tree**
RMSE	5.3207	4.9568	5.0247
R^2	0.63	0.68	0.67
MSE	28.31	24.57	25.248
MAE	1.6343	1.5173	1.5107

Table.6. Performance Measures of Linear Regression Models

Error Metrics	Linear Regression Models			
	Linear	**Interactions Linear**	**Robust Linear**	**Stepwise Linear**
RMSE	5.0329	4.7344	7.9715	**4.7254**
R^2	0.67	0.71	0.17	0.71
MSE	25.33	22.414	63.544	22.33
MAE	1.5816	1.5409	4.7254	1.5393

As mentioned in Table.2 and Table.3, Error quantified by its root mean square. The root-mean-squared error (RMSE) is always positive, and its units are consistent with those of our answer. Measure of reliability or accuracy. R^2 is never equal to 1, and it's always greater than 0. It evaluates the trained model against one where the response is always the same and always equals the mean of the training response. Having a negative R-Squared value indicates that our model is inferior to the constant model. Error sum squared If you square the RMSE, we get the MSE. Error standard deviation The MAE is equivalent to the RMSE, but it is less susceptible to outliers and always returns a positive value.

5. CONCLUSION

This work assesses the prediction of rainfall with a linear regression for multiple weather data- variables by incorporating the weighted moving average filter, and an exponential moving average filter based on linear and Gaussian kernels in Regression Learners feature of MATLAB environment. The potential of a 10 years data set with six variables is investigated for improved forecasting. During the model training phase, the number of delay steps is varied from 3 to 10, and the effectiveness of MAE, MSE, and RMSE criteria is compared using the 5-fold cross-validation approach. The technique using the filter for 10 delay steps yields the lowest MSE error number when all filtering and estimation models are compared head-to-head. For the specified minimum error model, a graph showing the estimated versus the actual regression coefficients is generated. The correctness of the model is also supplied, along with the coefficient of Determination of correlation (i.e., R-squared), measurement of the strength of the relationship between input and output.

REFERENCES

[1] P.A. Pisano, L.C. Goodwin and M.A. Rossetti, "Us Highway Crashes in Adverse Road Weather Condition", *Proceedings of International Conference on Interactive Information and Processing Systems*, pp. 1-5, 2008.

[2] W.S. Ashley, S. Strader, D.C. Dziubla and A. Haberlie, "Driving Blind: Weather-Related Vision Hazards and Fatal Motor Vehicle Crashes", *Bulletin of the American Meteorological Society*, Vol. 96, No. 5, pp. 755–778, 2015.

[3] L.M. Trick, R. Toxopeus and D. Wilson, "The Effects of Visibility Conditions, Traffic Density, and Navigational Challenge on Speed Compensation and Driving Performance in Older Adults", *Accident Analysis and Prevention*, Vol. 42, No. 6, pp. 1661-1671, 2010.

[4] H.M. Hassan and M.A. Abdel-Aty, "Predicting Reduced Visibility Related Crashes on Freeways using Real-Time Traffic Flow Data", *Journal of Safety Research*, Vol. 45, pp. 29-36, 2013.

[5] A.S. Mueller and L.M. Trick, "Driving in Fog: The Effects of Driving Experience and Visibility on Speed Compensation and Hazard Avoidance", *Accident Analysis and Prevention*, Vol. 48, pp. 472-479, 2012.

[6] M.M. Ahmed, M. Abdel-Aty, J. Lee and R. Yu, "Real-Time Assessment of Fog-Related Crashes using Airport Weather Data: A Feasibility Analysis", *Accident Analysis and Prevention*, Vol. 72, pp. 309-317, 2014.

[7] A. Theofilatos and G. Yannis, "A Review of the Effect of Traffic and Weather Characteristics on Road Safety",

Accident Analysis and Prevention, Vol. 72, pp. 244-256, 2014.

[8] S. Kavitha, S. Varuna and R. Ramya, "A Comparative Analysis on Linear Regression and Support Vector Regression", *Proceedings of International Conference on Green Engineering and Technologies*, pp. 1-5, 2016.

[9] A. Vlachogianni, P. Kassomenos, A. Karppinen, S. Karakitsios and J. Kukkonen, "Evaluation of a Multiple Regression Model for the Forecasting of the Concentrations of Nox and Pm10 in Athens and Helsinki", *Science of the Total Environment*, Vol. 409, No. 8, pp. 1559-1571, 2011.

[10] T. Fang and R. Lahdelma, "Evaluation of a Multiple Linear Regression Model and Sarima Model in Forecasting Heat Demand for District Heating System", *Applied Energy*, Vol. 179, pp. 544-552, 2016.

[11] A. Mahabub, A.Z.S.B. Habib, M. Mondal, S. Bharati and P. Podder, "Effectiveness of Ensemble Machine Learning Algorithms in Weather Forecasting of Bangladesh", *Proceedings of International Conference on Innovations in Bio-Inspired Computing and Applications*, pp. 267-277, 2021.

[12] D. Maulud and A.M. Abdulazeez, "A Review on Linear Regression Comprehensive in Machine Learning", *Journal of Applied Science and Technology Trends*, Vol. 1, No. 4, pp. 140-147, 2020.

[13] I. Gad and D. Hosahalli, "A Comparative Study of Prediction and Classification Models on NCDC Weather Data", *International Journal of Computers and Applications*, Vol. 44, No. 5, pp. 414-425, 2022.