STUDY ON UNCERTAIN WEATHER DATA USING DIFFERENT SAMPLING METHODS

[1,2] Santhi B, [1,3]Mohamed Yasir M and [1,4]Nithin Christopher S
[1]Dept. of Information and Communication Technology
SASTRA University
Thanjavur, India
[2]shanthi@cse.sastra.edu [3] 113014073@sastra.ac.in [4]113014085@sastra.ac.in

Abstract— The most entangled system always gives pretty trouble in decision making. One such system is weather prediction with perfect set of inputs. Two sources of jangle need to consider here. One is uncertainty in the inputs and second is imperfection in mathematical model which is supposed to solve the forecast equations. As the weather forecast is a chaotic system, tiny error in the initial state can lead to the large error in the output. The uncertainty falls in two categories. Stochastic uncertainty, which is the changing behavior of the system and the epistemic uncertainty which is hard to set the initial condition of a parameter in the mathematical model due to dynamic nature. This proposed forecast model has the answer for the ongoing issue by means of incorporating sampling method to calculate the true set of inputs to be used in the forecast model. Applying Regression equation, this takes the available weather statistics and gives linear and pellucid relationship of the forecast with least error. Our model has taken several Atmospheric parameters as predictors and the weather is decomposed to several features to get accuracy to its peak. This model exhibit best accuracy, however the predictor/feature is by identifying and discarding the unwanted data and predictor which made least impact on the result without compromising the quality of the prediction. Further identified predictor may refine using neural network to get more reliable result. This study facilitates the duo to bring out notable difference in prevailing method to render the uncertainty along with addressing mathematical imperfection.

Keywords— Uncertainty % Multiple linear regression (MLR) % Sampling % Predictor % Feature.

INTRODUCTION

Weather forecasting is a complex system, handling uncertain inputs. Since the atmosphere fetches the input here, the underlying system is dynamic. An entangled system will not deal with just a single input but it has to handle inputs from diverse components and each component may have different scales of measurements with some error. Weather forecasting being an entangled system is affected by several factors and each has their own range of values and collected through various instruments. The main factors affecting the weather are temperature, wind flow, wind pressure, wind direction, humidity, time and visibility[5]. There may be several other factors available out in the atmosphere but this paper tries to bring out a best sample to find reliable output with these parameters. The major factors are considered as predictors for the mathematical model formulation. Due to the changing temperament of the dynamic system and the error that occur during initialization of the dynamic variables. This system should handle both stochastic uncertainty and epistemic uncertainty [2]. Since weather prediction needs more accuracy and it has to deal with large volume of statistical data, formulation of mathematical model and choosing appropriate sampling method turns out to a challenge.

A series of historical data[5] is collected and stored in a database over a period of time. This statistical data helps to contrive a mathematical model. These inputs should be considered with care since it may lead to a catastrophic error if they are formulated improperly. And hence they should be pre-processed to predict their behavior so that we tend to bring out the reaction in our hand.

But processing a huge set of data costs more time and money. So sampling should be done, wherein we have to choose appropriate sampling that should suffice our needs in less model runs[11]. The proposed paper handles these challenges. This paper finds the solution to stochastic uncertainty through multiple linear regressions and the epistemic uncertainty through comparison of sampling methods. So the final model will be tantamount in answering all the parameters with reliability, less cost and less time. The study uses weather data[5] from which a sample portion is depicted in table 1 and The work flow diagram is shown in fig1.

Table 1 : Statistical Data for Weather Report

DATE	TIME	TEMP (C^o)	SPEED	DIR	HUM	PRE	VIS	DAY
14-jan	13.40	29	9	8	0.51	1012	6	2
14-jan	14.40	28	6	8	0.58	1012	6	2
14-jan	15.40	28	0	0	058	1012	6	2
14-jan	16.40	27	0	0	0.66	1012	6	1
14-jan	17.40	27	0	0	0.66	1012	6	1
14-jan	18.40	24	0	0	0.78	1013	6	1
14-jan	19.40	23	0	0	0.83	1013	6	1

The Number of predictors(parameters) taken here is 7
TEMP = Temperature (C^o)
DIR=Direction of wind
HUM=Humidity
PRE=Wind pressure
VIS=Visibility
DAY=Day/Night

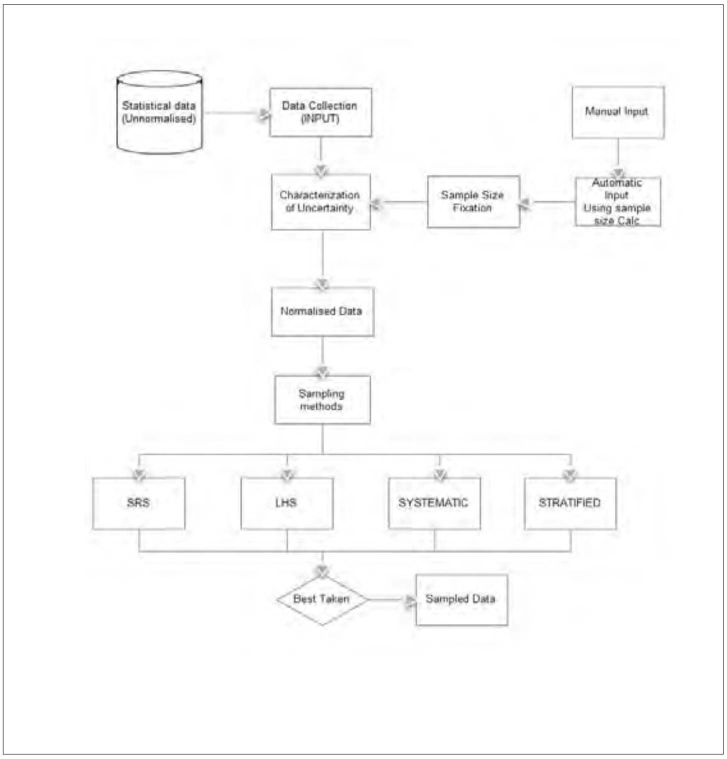

Fig 1: Flow diagram of Data Selection

RELATED WORK:

Study on uncertain data and prediction of future outcomes are plenty. Also developing further study over such real time problems brings more wonder. The works on several sampling methods are notable.

Latin hypercube sampling: It is used to The NUREG-1150 analyses, the analyses carried out in support of the Compliance Certification Application[1] for the Waste Isolation Pilot Plant, and analyses carried out in support of the Yucca Mountain Project’s development of a facility for the deep geologic disposal of high level radioactive waste provide examples of complex analyses that have used Latin hypercube sampling in the propagation of epistemic uncertainty.

Sampling [4]: An opinion poll on America’s health concern was conducted by Gallup Organization between October 3- 5, 1997, and the survey reported that 29% adults consider AIDS is the most urgent health problem of the US, with a *margin of error* of +/- 3%. The result was based on telephone interviews of 872 adults.

Drug addict estimation [13]: A survey is conducted by Frerichs, R.R. Rapid Surveys to estimate the drug addict of total nine people taken for investigation to explain simple random sampling. Our intention is to sample three addicts from the population of nine, assuming that the whole population cannot be examined. So randomly 3 people are selected and are studied.

Multiple Linear Regression[14]: This article estimated the temperature all around the United States of America after acquiring the necessary details from the West stations and East stations by the Latitude, Longitude and Elevation of these respected areas by multiple linear regression.

METHODS :

MULTIPLE LINEAR REGRESSION

The multiple linear regression (MLR) [14] is the technique used in this paper to devise a mathematical model. The statistical data contains several predictors which are fed as inputs and their corresponding outputs are derived. A certain relationship exists between the input and output variables. A generalized form is as follows:

$$Y = \alpha + \beta X + \varepsilon \qquad (1)$$

Where, Y and ε are output and standard error matrices respectively of size 1xN and X is the input matrix with size pxN, α is the constant and β is a 1xp matrix containing regression coefficients[7]. Here N is the population size and p is the number of predictors. Thus the expanded formula is as:

$$Y_1 = \alpha + \beta_1 X_{1,1} + \beta_2 X_{1,2} + ... + \beta_k X_{1,p} + \varepsilon_1 \qquad (2)$$

$$Y_2 = \alpha + \beta_1 X_{2,1} + \beta_2 X_{2,2} + ... + \beta_k X_{2,p} + \varepsilon_2 \qquad (3)$$

$$Y_N = \alpha + \beta_1 X_{N,1} + \beta_2 X_{N,2} + ... + \beta_k X_{N,p} + \varepsilon_N \qquad (4)$$

By solving the matrices formed by the inputs and outputs, the regression coefficients can be obtained. The generalized formula for obtaining regression coefficients is:

$$\beta \quad = (X^1 X)^{-1} X^1 Y \qquad (5)$$

Once the regression coefficients are obtained we can obtain any output value until the inputs are fed. Also once the model is formulated we can obtain how a single input affects an output. But the model will not be fully accurate unless we add an error term, ε. When dealing with small systems which require less precision, this error term may be neglected. But in the case of weather forecasting system the model should be aware of the error in order to produce better results. The standard error matrix is formed by the relation

$$^2 (X^1 X)^{-1} \qquad (6)$$

where, ^{2} is the error variance. And thus our fitted values can be estimated.

$$\hat{Y} = X \beta \qquad (7)$$

The mathematical model formed is analyzed with some parameters such as R-squared value, estimation of error variance, F-statistics and p-value. The R-squared value shows how closely our formed model matches with the original data[6][7]. And F-statistics and p-value helps in hypothesis tests. The estimation of error variance is important because they are variances which are unrelated to the predictors and should not be considered in our investigation. The confidence intervals are obtained which has a set of two values for each predictor. The lower and upper bound values of the confidence intervals indicate the reliability of the estimate for a particular confidence level.

The next process is to relate the fitted values with the observed values to get the residual matrix which is of size 1xN. The residual value is denoted here by the symbol, r and it is formed by the following equation:

$$r = Y - \hat{Y} \qquad (8)$$

After a full model has been framed, confusion matrix is formed between the observed outputs and the predicted outputs. The fidelity of the model can be found out by this confusion matrix. We can't fully rely on this. We may need to refine some data or remove if it is irrelevant. A subset model is formed by removing each predictor and the resulting fidelity is noted. This is done through trial and error method. The subset model which yields the lower fidelity than the full model shows that the removed predictor is important for our model formulation and further process. If the subset model's fidelity is comparatively higher than the full model then it implies either the predictor needs some refinement or else it needs to be removed from the model. Thus we arrive at improving fidelity in mathematical model formulation using the multiple linear regression method.

The output of the model can be increased by breaking the single output to a possible extent until they can be divided using the divide and conquer strategy. For example, consider "passing clouds mild" as a single output. It can be broken into passing + clouds + mild forming three separate outputs. Now the complexity gets reduced when the output got split. Now our fidelity can be increased to a greater extent. Confusion matrices are formed separately. This has increased the output's fidelity to a greater extent. Now a refined model is formed which can be used for further processing. Whatever may be the behavior of the system, it is just observed and model is updated. Thus stochastic uncertainty is well handled through this proposed technique. To check the fidelity and suggestion may improve the fidelity features are collected and portion of a possible values are mentioned in table 2.

Table 2: Portion of a Statistical Weather Chart with feature and their numbers.

ACTUAL WEATHER	ACTUAL WEATHER Value	F 1	F 1 Value	F 2	F 2 Value	F 3	F 3 Value	F 4	F 4 Value
passing clouds warm	6	warm	2	clouds	2	passing	2	-	2
passing clouds warm	6	warm	2	clouds	2	passing	2	-	2
passing clouds warm	6	warm	2	clouds	2	passing	2	-	2
passing clouds warm	6	warm	2	clouds	2	passing	2	-	2
passing clouds warm	6	warm	2	clouds	2	passing	2	-	2
passing clouds mild	5	mild	1	clouds	2	passing	2	-	2
clear mild	7	mild	1	Clear	3	-	4	-	2

Types of weather observed and their Respective Numbers assumed to solve the weather equations.

Fog Mild	1
Scattered clouds warm	4
Passing Clouds warm	6
Clear warm	8
Clear Mild	7
Passing clouds Mild	5
Broken Clouds warm	12
Drizzle Broken clouds Mild	13

Feature 1 (F 1) can takes Mild, Warm

Feature 2 (F 2) can takes Fog, Clouds, Clear, Haze

Feature 3 (F 3) can takes Scattered, Passing, Broken, None

Feature 4 (F 4) can takes Drizzle, Nothing

COMPARISON OF SAMPLING METHODS

Sampling is a process which subset is derived from the population where subset or sample gives the nature of the entire population. There are several sampling methods available but they are picked depending on the purpose of the problem[9]. This paper enlist four different kinds of sampling:

- Simple Random Sampling
- Stratified sampling
- Systematic sampling
- Latin Hypercube sampling

These sampling methods are studied separately and congruous sampling method is chosen for the input set of data. Thus the model is not immured to just forecasting weather but it responds to any given input data set. Now to find the best sampling method for weather forecasting some standard procedures are followed

A. Determining the Sample Size

Defining the sample size is the fundamental part of sampling. Initially the population size is evaluated from the statistical data. For a large sample the population size should be greater than 200. From that large voluminous data samples need to be taken. In this paper the sample size may be determined either manually or automatically. The sample size is calculated automatically through a user interactive sample calculator which is embedded along with the main software. This sample calculator collects confidence level, marginal error and variance from the user [3][10]. The confidence level is converted into z-value by looking up the z-table which is programmed inside.

B. Formation of Samples using Simple Random sampling

Simple Random Sampling just picks random samples from the population [1][13]. This sampling is unbiased and easy to code and implement. But it cannot be applied for heterogeneous data. For example if the population is distributed over different geographical locations, this sampling does not holds good.

C. Formation of Samples using Stratified Sampling

Stratified sampling is an improvement over random sampling. This can handle heterogeneous data because it classifies data into separate strata[1]. Then from each strata the samples are randomly picked. The condition for picking random samples here is the sample should contain same proportion of data from each stratum as they were in the entire population set. This sampling is unbiased but the complexity of code is higher comparative to simple random sampling. And this sampling method holds good even for heterogeneous data. So it is well suited for data that are geographically distributed. It is important to note that stratified sampling is different from cluster sampling, where cluster sampling requires a prior knowledge about the distribution of data.

D. Formation of Samples using Systematic Sampling

Systematic sampling is another version of random sampling. In random sampling the data are selected in random. Here it follows a systematic approach. A sample is selected at regular interval thus following a regular flow. This sampling can be coded easily. But the sampled data may be biased because there is more probability that the same data repeats over the same interval.

E. Formation of Samples using Latin Hypercube Sampling

Latin hypercube sampling is an advancement of Monte-Carlo Sampling[1][3][11]. In Latin Hypercube sampling the range of each variable is proportioned into n non overlapping intervals of equal probability 1/n. Then from each interval one value is selected randomly according to the probability density of the order to make our sample meaningful.

F. Pairing the Samples

The samples are thus formed for individual predictors. Now every predictor value should be paired with the other in order to form a complete set of inputs. One value of temperature is randomly selected from the sample and it is paired randomly with one wind flow value. The pairing is done for all the predictors used forming a set of inputs[3]. Now our sampled input is formed with a matrix of size nxp, where n is the sample size and p is the number of predictors.

G. Comparison of sampling methods

The samples thus formed are compared [1][11] by various metrics such as measures of central tendency, dispersion of the parameters and variance in the parameters. By comparing the mean value of the population with the sample mean obtained we can conclude how close the sample data set is centered on the mean. Therefore the best method is chosen depending upon our requirements. For the weather forecasting, the sample should resemble the true set of population. So the Latin Hypercube Sampling is chosen because of its mean and variance which converges very close to the population even for a small sample size. And also the elapsed time and the total number of model runs are checked. Latin Hypercube Sampling surpass these challenges and it proves to cost effective and less time since it produces sample in very less number of model runs.

ALGORITHM:

The following steps should be carried out to identify the irrelevant predictor in the Net predictors taken count at that moment.

Step 1: Get the statistical weather data of one station by means of various measures (predictors).

Step 2: Normalize the collected data.

Step 3: Determine the sample size to deduce the sample needed from the entire statistical data.

Step 4: Apply different sampling methods.

Step 5: Identify irrelevant feature.

Step 6: Check fidelity.

Step 7: Repeat step 5 and step 6 for the entire feature.

Step 8: Design the model.

Step 9: Compute performance metrics.

RESULTS AND ANALYSIS

The following tables display the result of feature 1 (F1) which takes Mild and Warm for population of size 197*7

i.e. 197 Observation and 7 predictors. Performance of this study can be evaluated using confusion matrix which is given in table 3 with sample size of 20.

Table 3: Confusion matrix

	Mild	Warm
Mild	8	0
Warm	0	12

Regression co-efficient helps to understand the variation in the function variable with respect to other independent variable. It is shown in table 4.

Table 4: The Regression co-efficient

Alpha value	2.021742
Temp Co-efficient	2.021742
Wind flow Co-efficient	3.126024
Wind speed Co-efficient	-0.519969
Wind Direction Co-efficient	-1.741550
Humidity Co-efficient	-6.165916
Visibility Co-efficient	2.945706
Daylight Co-efficient	0.151614

The significant metrics helps to understand the power of the model that provide evidence to trust the model, which is given in table 5.

Table 5: Performance Analysis

R-Square Value	0.929503
Adjusted R-square	0.888380
F stat Value	28.567628
P value	0.000001
Estimation of Error Variance	0.026030

The respective sum of squares of error and regression is depicted in table 6:

Table 6: Estimation of Sum of Squares

Sum of Squares of	Value	Mean value	Degrees of Freedom
Error	0.000000	0.000000	12
Regression	4.800000	0.685714	7
Total	4.800000	0.685714	19

The respective mean, variance and standard deviation of each predictors is displayed in table 7.

Table 7: Parameters of the predictor

Predictor	Mean	Variance	Std Deviation
Temperature	0.008661	0.924837	0.961684
Wind flow	0.000000	0.000000	1.000000
Wind pressure	0.011180	0.223596	0.950003
Wind Direction	0.000000	0.000000	1.000000
Humidity	0.014358	0.287160	0.535873
Visibility	0.002538	0.050767	0.225316
Day/Night	0.013019	0.260378	0.510272
Total Population	0.049756	1.746737	1.321642

Before applying proposed technique, this model is 80 % accurate.

The fidelity rate of full system after removing respective predictor is given in table 8.

Table 8: Parameter Influence

S.No	Removed parameter	Fidelity after Removing
1	Temperature	60.000000
2	Wind Flow	65.000000
3	Wind Direction	70.000000
4	Wind Pressure	75.000000
5	Humidity	75.000000
6	Visibility	80.000000
7	Day/Night	85.000000

After applying proposed technique, this model is 85 % accurate

CONCLUSION

This model exhibit efficient effort in order to lessen the current uncertain effects in elucidating forecasting prediction through Multiple Linear Regression and also provides the exact set of sample inputs through Latin Hypercube Sampling. At present this model could patch up the imperfection in the numerical prediction model for weather predicting system and this study which takes 197 observation of 7 predictors i.e. 197*7 size of population in which most relevant predictors are taken in to consideration, the study with sampling of proposed techniques provides 85% fidelity whereas other existing methods gives fidelity 80% only. Hence this proposed model meets and mitigates the problems faced by uncertain data in real time application to give an uncanny and amiable result.

REFERENCES

[1] Helton, J. C., Johnson, J. D., Sallaberry, C. J., & Storlie, C. B. (2006). Survey of sampling-based methods for uncertainty and sensitivity analysis. Reliability Engineering & System Safety, 91(10), 1175-1209.

[2] Cacuci, Dan Gabriel. "Sensitivity and Uncertainty Analysis of Models and Data." Nuclear Computational Science. Springer Netherlands, 2010. 291-353.

[3] Wyss, Gregory D., and Kelly H. Jorgensen. "A user’s guide to LHS: Sandia’s Latin hypercube sampling software." SAND98-0210, Sandia National Laboratories, Albuquerque, NM (1998).

[4] Saifuddin Ahmed on “Statistical Methods for Sample Surveys” John Hopkins Bloomberg school of public health.

[5] www.timeanddate.com

[6] www.mathworks.in/help/stats/understanding-linear-regression-outputs.html

[7] www.kellogg.northwestern.edu/faculty/weber/emp/_Ses sion_2/Regression.htm

[8] http://www.palisade.com/downloads/help/risk35/faq_html/ latinhypercubevs.montecarlo.htm

[9] http://www.ccl.rutgers.edu/~ssi/thesis/thesis-node15.html#SECTION00553000000000000000

[10] Fox N., Hunn A., and Mathers N. “ Sampling and sample size calculation “
The NIHR RDS for the East Midlands / Yorkshire & the Humber 2007.

[11] Anna Matala 60968U (20.5.2008) on “Sample Size Requirement for Monte Carlo – simulations using Latin Hypercube Sampling “ HELSINKI UNIVERSITY OF TECHNOLOGY, Department of Engineering Physics and Mathematics, Systems Analysis Laboratory Mat-2.4108 Independent Research Projects in Applied Mathematics.

[12] www.stat.ufl.edu/~ssaha/3024.html

[13] www.ph.ucla.edu/epi/rapidsurveys/RScourse/RSbook_ch3.pdf

[14] http://www.stat.lsa.umich.edu/~kshedden/Courses/Stat401/Notes/401-multreg.pdf