American Journal of Computational Mathematics, 2014, 4, 406-413
Published Online December 2014 in SciRes. http://www.scirp.org/journal/ajcm
http://dx.doi.org/10.4236/ajcm.2014.45035

Scientific Research

Statistical Modelling of Soybean Crop Yield in Regions of Central India through Mathematical and Computational Approach

Sarvraj Singh[1], Dilpreet Tuteja[1], Param Tripathi[1], Chirag Basavaraj[2]

[1]Jaypee University of Engineering & Technology, Guna, India
[2]R.V. College of Engineering, Bangalore, India
Email: sarvraj.5@hotmail.com

Received 7 August 2014; revised 15 September 2014; accepted 1 October 2014

Abstract

In this paper, we have discussed a number of fitting methods to predict crop yield of soybean depending on the nature of environment and a comparison is done between them on the basis of available data set. Later we have suggested a suitable method for the prediction of the crop yield on the basis of residual (error) terms. Statistical analysis is also used for getting the relationships between different components (variables) of available data set. At last, we have discussed about Chaos that can distort the whole mathematical analysis and a computational approach.

Keywords

Climate Change, Prediction, Chaos, Uncertainty

1. Introduction

Climate describes the ensemble sum of typical conditions of temperature, relative humidity, cloudiness, precipitation, wind speed and direction and innumerable other meteorological factors that prevail regionally for extended periods [1]. Weather of a demographic region is defined by the hourly description of the climatic conditions experienced by the inhabitants of that region. Here we discuss the soybean yield as a function of these environmental parameters.

Many different approaches are used for constraining climate based crop yield predictions based on observations of past empirical change in the yield [2]. Here we setup distinct models based on the environmental model

How to cite this paper: Singh, S., Tuteja, D., Tripathi, P. and Basavaraj, C. (2014) Statistical Modelling of Soybean Crop Yield in Regions of Central India through Mathematical and Computational Approach. *American Journal of Computational Mathematics*, **4**, 406-413. http://dx.doi.org/10.4236/ajcm.2014.45035

parameters; significant correlations are calculated based on the inferred outputs. Meteorologists say that if only they could design an accurate mathematical model of the atmosphere with all its complexities, they could forecast the weather with real precision. But this is an idle boast, immune to any evaluation, for any inadequate weather forecast would obviously be blamed on imperfections in the model. Catering to the often glitches in the models prepared the fidelity of the dynamics governing the respective models can be doubted. With the introduction of computer simulations the weather predictions can be done in just a few minutes. We make use of such a technique to generalize the crop yield, and make prediction on the basis of the environmental factors like wind speed, wind direction, temperature and humidity. These factors are trivial when considering crop yield however, makes a difference as suggested by the models ahead.

Since the sensors of the parameters mentioned above are respect to one region in Central India, so we consider the crop that this region has lavishly produced, soybean. Soybean is one of the important crops of the world [3]. In India the production of soybean is currently restricted to mainly Madhya Pradesh, Uttar Pradesh, Maharashtra and Gujarat. Himachal Pradesh, Punjab and Delhi are other states with some marginal produce. According to 2010 estimates of soybean production India produces 4.4% of the total production; central India is the largest contributor of soybean yield. This brings us to concentrate more over this region for our fitting models.

Soybean is a crop that grows in warm and moist climate. An optimum yield requires a temperature ranging between 26.5˚C to 30˚C. For rapid germination and vigorous seedling growth soil temperatures of 15.5˚C or above are most suitable. A lower temperature delays flowering. Although, moisture enhances the yield of the crop but excess of moisture can make it prone to foliar diseases like frogeye leafs spot and septoria brown spot. Therefore, an optimum amount of humidity is required for the crop.

Wind direction and velocity also have a significant influence on crop growth [4]. While it has a few benefits, gusty winds blowing in one direction can harm the crop. Beneficial impacts include increasing the supply of carbon dioxide by increasing turbulence in the atmosphere. It also alters the balance of hormones. Strong winds in a region may uproot the crop or be an inevitable carrier of dispersive seeds that may hamper the yield. **Table 1** elucidates the conditions prevalent in Central India, state of Madhya Pradesh that monitor the soybean growth.

As far as the prediction of the yield on a larger perspective is considered, the simulations carried out by supercomputers are based on curve fitting methods. Curve fitting is the process of constructing a curve that has the best fit to a series of data points, possibly subject to constraints. Curve fitting involves interpolation [5], where an exact fit to the data is required in which a “smooth” function is constructed that approximately fits the data. A related topic is regression analysis, which focuses more on questions of statistical inference which includes the uncertainty present due to the random errors in the observed data. Fitted curves can be used as approximate data visualization for a model to which it is applied and to summarize the relationships among two or more variables.

Table 1. Varying environmental parameters dependent for yield of soybean in Central India.

Month	Temperature (X)	Humidity (Y)	Wind speed (Z)	Wind direction (W)
January	142.188	14.992	50.708	4.510
February	159.590	19.230	34.500	4.610
March	191.820	25.080	13.190	4.486
April	202.441	30.287	17.602	5.180
May	252.712	33.287	17.541	5.503
June	255.880	32.440	56.000	9.268
July	238.640	28.380	64.650	6.770
August	245.000	24.100	81.910	4.708
September	203.300	25.210	71.035	4.000
October	143.916	25.330	32.250	5.267
November	148.460	20.600	31.234	3.065
December	159.660	17.970	40.830	2.972

Extrapolation refers to the use of a fitted curve beyond the range of the observed data, and is subject to a greater degree of uncertainty since it may reflect the method used to construct the curve as much as it reflects the observed data. In order to fit a polynomial up to three degree which exactly fits four constraints, each constraint can be a point, angle, or curvature (which is the reciprocal of the radius of an osculating circle). Angle and curvature constraints are most often added to the ends of a curve, and in such cases are called end conditions. Identical end conditions are frequently used to ensure a smooth transition between polynomial curves contained within a single spline. If we have more than $n + 1$ constraints (n is the degree of the polynomial), we can still run the polynomial curve through those constraints. An exact fit to all constraints is not certain (but it might happen, for example, in the case of a first degree polynomial exactly fitting three collinear points). In general, however, some method is then needed to evaluate each approximation. The least squares method is one way to compare the deviations.

Low-order polynomials tend to be smooth and high order polynomial curves tend to be lumpy. To define this more precisely, the maximum number of inflection points possible in a polynomial curve is $n-2$, where n is the order of the polynomial equation. An inflection point is a location on the curve where it switches from a positive radius to negative. It is only possible that high order polynomials will be lumpy; they could also be smooth, but there is no guarantee of this, unlike with low order polynomial curves. A fifteenth degree polynomial could have, at most, thirteen inflection points, but could also have twelve, eleven, or any number down to zero.

2. Fitting a Polynomial Function

When a given set of data does not appear to satisfy a linear equation, we can try a suitable polynomial as a regression curve to fit data. The least squares technique can be readily used to fit the data to a polynomial.

Consider a polynomial of degree $m-1$

$$y = a_1 + a_2 x + a_3 x^2 + \cdots + a_m x^{m-1} = f(x). \tag{1}$$

If the data contains n sets of x and y values, then the sum of squares of the errors is given by

$$Q = \sum_{i=1}^{n} \left[y_i - f(x_i) \right]^2. \tag{2}$$

Since $f(x)$ is a polynomial and contains coefficients a_1, a_2, a_3 etc. we have to estimate all m coefficients. As before, we have the following m equations that can be solved for these coefficients.

$$\begin{aligned}
&\frac{\partial Q}{\partial a_1} = 0, \\
&\frac{\partial Q}{\partial a_2} = 0, \\
&\cdots \\
&\cdots \\
&\frac{\partial Q_m}{\partial a_m} = 0.
\end{aligned}$$

Consider a general term,

$$\begin{aligned}
&\frac{\partial Q}{\partial a_j} = -2 \sum_{i=1}^{n} \left[y_i - f(x_i) \right] \frac{\partial f(x_i)}{\partial a_j} = 0, \\
&\frac{\partial f(x_i)}{\partial a_j} = x_i^{j-1}.
\end{aligned}$$

Thus we have

$$\begin{aligned}
&\sum_{i=1}^{n} \left[y_i - f(x_i) \right] x_i^{j-1} = 0 \\
&\sum \left[y_i x_i^{j-1} - x_i^{j-1} f(x_i) \right] = 0
\end{aligned} \qquad j = 1, 2, \cdots, m.$$

Substituting for $f(x_i)$

$$\sum_{i=1}^{n} x_i^{j-1}\left(a_1 + a_2 x^i + a_3 x_i^2 + \cdots + a_m x_i^{m-1}\right) = \sum_{i=1}^{n} y_i x_i^{j-1}.$$

These are *m* equations $(j = 1, 2\cdots, m)$ and each summation is for $i = 1$ to *n*.

$$\begin{aligned}
&a_1 n + a_2 \sum x_i + a_3 \sum x_i^2 + \cdots + a_m \sum x_i^{m-1} = \sum y_i,\\
&a_1 \sum x_i + a_2 \sum x_i^2 + a_3 \sum x_i^3 \cdots + a_m \sum x_i^m = \sum y_i x_i,\\
&\vdots\\
&a_1 \sum x_i^{m-1} + a_2 \sum x_i^m + a_3 \sum x_i^{m-1} \cdots + a_m \sum x_i^{2m-2} = \sum y_i x_i^{m-1}.
\end{aligned}$$

The set of *m* equations can be represented in a matrix notation as follows:

$$CA = B$$

where

$$C = \begin{bmatrix} n & \sum x_i & \sum x_i^2 & \cdots & \sum x_i^{m-1} \\ \sum x_i & \sum x_i^2 & \sum x_i^3 & \cdots & \sum x_i^m \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sum x_i^{m-1} & \sum x_i^m & \cdots & \cdots & \sum x_i^{2m-2} \end{bmatrix}, \quad A = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_m \end{bmatrix}, \quad B = \begin{bmatrix} \sum y_i \\ \sum y_i x_i \\ \sum y_i x_2 \\ \vdots \\ \sum y_i x_i^{m-1} \end{bmatrix}.$$

The element of matrix *C* is

$$C(j,k) = \sum_{i=1}^{n} x_i^{j+k-2},\ j = 1, 2, \cdots, m, \text{ and } k = 1, 2, \cdots, m$$

$$B(j) = \sum_{i=1}^{n} y_i x_i^{j-1},\ \ j = 1, 2, \cdots, m.$$

The first model which we fit the yearly soybean yield is the linear model described by

$$L = a_0 + a_1 x + a_2 y + a_3 z + a_4 w \tag{3}$$

where a_0 being a constant term, *w* is the wind direction in degree, *x* being temperature parameter in degree Celsius, "*y*" the percentage humidity, "*z*" is the speed of wind in km/hr.

The error in the generalisation

$$\left[L - \left(a_0 + a_1 x + a_2 y + a_3 z + a_4 w\right)\right]. \tag{4}$$

And squaring the error term for Minimum Squared Error

$$E = \left[\left[L - \left(a_0 + a_1 x + a_2 y + a_3 z + a_4 w\right)\right]^2\right]. \tag{5}$$

Differentiating with respect to various factors, similar to equation for the weighted coefficients for the parameters that determine the yield, given by

$$\begin{aligned}
&24\sum L = 284a_0 + 9892.52a_1 + 1233.43a_2 + 46602.534a_3 + 15964.916a_4,\\
&899.32\sum L = 9892.52a_0 + 404388.23a_1 + 50420.375a_2 + 1905026.85a_3 + 652616.73a_4,\\
&112.13\sum L = 1233.43a_0 + 50420.37a_1 + 6286.56a_2 + 237524.64a_3 + 81370.27a_4,\\
&4236.594\sum L = 46602.53a_0 + 1905026.85a_1 + 237524.64a_2 + 8974364.36a_3 + 3074369.168a_4,\\
&1451.356\sum L = 15964.916a_0 + 725184.539a_1 + 81370.274a_2 + 3074403.016a_3 + 1053217.119a_4.
\end{aligned}$$

The yield that is $\sum L$ as per statistics available from the first estimate of soybean crop from Soybean Processor Association of India [6] (SoPA 2012) is 1150 kg/hectare.

Solving the equations to get the values of the weighted coefficients

$$a_0 = 104.545,$$
$$a_1 = 2.145,$$
$$a_2 = -0.000001139,$$
$$a_3 = -8.368,$$
$$a_4 = 0.0000013472.$$

Generalizing the model the yield can be predicted by

$$L = 104.545 + 0.0000013472w + 2.145x - 0.000001139y - 8.368z$$

where the *w*, *x*, *y*, *z* are the parameters discussed above.

The second model which we fit the yearly soybean yield is the linear model described by

$$L = a_0 + a_1x^2 + a_2y^2 + a_3z^2 + a_4w^2 \tag{6}$$

where a_0 being a constant term, *w* is the wind direction in degrees, *x* being temperature parameter in degree Celsius, "*y*" the percentage humidity, "*z*" is the speed of wind in km/hr.

The error in the generalisation

$$\left[L - \left(a_0 + a_1x^2 + a_2y^2 + a_3z^2 + a_4w^2\right)\right]. \tag{7}$$

And squaring the error term for Minimum Squared Error

$$E = \left[\left[L - \left(a_0 + a_1x^2 + a_2y^2 + a_3z^2 + a_4w^2\right)\right]^2\right]. \tag{8}$$

Differentiating with respect to various factors, similar to equation for the weighted coefficients for the parameters that determine the yield, given by

$$24\sum L = 284a_0 + 7751.249a_1 + 25624.804a_2 + 317.6214a_3 + 479550.034a_4,$$
$$15502.498\sum L = 170527.478a_0 + 120163722.1a_1 + 397248472.8a_2 + 4923925.118a_3 + 7434223443a_4,$$
$$51249.608\sum L = 563745.688a_0 + 397248472.8a_1 + 1313261160a_2 + 16277972.24a_3 + 24576751260a_4,$$
$$635.2428\sum L = 6987.6708a_0 + 4923965.424a_1 + 16277972.24a_2 + 201766.7074a_3 + 304630706.4a_4,$$
$$479550.034\sum L = 5275050.374a_0 + 3717111721a_1 + 12288375630a_2 + 152315353.2a_3 + 22996823510a_4.$$

The yield that is $\sum L$ as per statistics available from the first estimate of soybean crop from Soybean Processor Association of India (SoPA 2012) is 1150 kg/hectare.

Solving the equations to get the values of the weighted coefficients

$$a_0 = -30.304,$$
$$a_1 = 2.247,$$
$$a_2 = 0.06088,$$
$$a_3 = -0.242,$$
$$a_4 = -1.545.$$

Generalizing the model the yield can be predicted by

$$L = -30.304 + 2.247x^2 + 0.06088y^2 - 0.242z^2 - 1.545a_4w^2$$

where the *w*, *x*, *y*, *z* are the parameters discussed above.

The third model which we fit the yearly soybean yield is the linear model described by

$$L = a_0 + a_1x^3 + a_2y^2 + a_3z^4 + a_4w \tag{9}$$

where a_0 being a constant term, *w* is the wind direction in degrees, *x* being temperature parameter in degree Celsius, "*y*" the percentage humidity, "*z*" is the speed of wind in km/hr.

The error in the generalisation

$$\left[L-\left(a_0+a_1x^3+a_2y^2+a_3z^4+a_4w\right)\right]. \tag{10}$$

And squaring the error term for Minimum Squared Error

$$E=\left[\left[L-\left(a_0+a_1x^3+a_2y^2+a_3z^4+a_4w\right)\right]^2\right]. \tag{11}$$

Differentiating with respect to various factors, similar to equation for the weighted coefficients for the parameters that determine the yield, given by

$$\begin{aligned}
24\sum L &= 284a_0+128601055.4a_1+600745.167a_2+309524.226a_3+46697.354a_4,\\
111691004.5\sum L &= 12860105.41a_0+68339793020000a_1+319241581600a_2\\
&\quad +164484050800a_3+24815407980a_4,\\
54613.197\sum L &= 600745.167a_0+319241565800a_1+1491300643a_2\\
&\quad +768368524.2a_3+115922354.2a_4,\\
28138.566\sum L &= 309522.466a_0+164484050800a_1+768368524.2a_2\\
&\quad +3958889448a_3+59727117.16a_4,\\
4245.214\sum L &= 46697.354a_0+24815407980a_1+115922354.2a_2\\
&\quad +59727117.16a_3+9010920.94a_4.
\end{aligned}$$

The yield that is $\sum L$ as per statistics available from the first estimate of soybean crop from Soybean Processor Association of India (SoPA 2012) is 1150 kg/hectare.

Solving the equations to get the values of the weighted coefficients

$$\begin{aligned}
a_0 &= -0.00004890265298,\\
a_1 &= 0.0006236823179,\\
a_2 &= -0.21479748791427158,\\
a_3 &= 0.2395042332638745,\\
a_4 &= -5.32059174909421e-8.
\end{aligned}$$

Generalizing the model the yield can be predicted by

$$L=-30.304+2.247x^2+0.06088y^2-0.242z^2-1.545a_4w^2$$

where the *w*, *x*, *y*, *z* are the parameters discussed above.

3. Chaos

Chaos is associated with complex and unpredictable behavior of phenomena over time [7]. Such behavior can arise in deterministic dynamical systems. These processes are intriguing in that the realizations corresponding to different, although extremely close, initial conditions typically diverge. The practical implication of this phenomenon is that, despite the underlying determinism, we cannot predict, with any reasonable precision, the values of the process for large time values; even the slightest error in specifying the initial condition eventually ruins our attempt. The chaos in terms of correlation coefficient within various environmental factors (say *n*) is given by

$$C=\left(r_{12}r_{23}r_{34}\cdots\right)^{1/n}.$$

4. Conclusion

Table 2 describes the possible correlation permutation and **Table 3** elucidates the variability of the yield amongst the different models under scrutiny. The results suggest, about the dependence of the yield on the environmental factors more under the variable weighted powers rather than being in linearly or quadratic fashion. **Figure 1** shows the proper harvesting time of the season for maximising the yield of soybean. The data are indeed direct acceptance of the model variable power model as the data match with the conventional values of

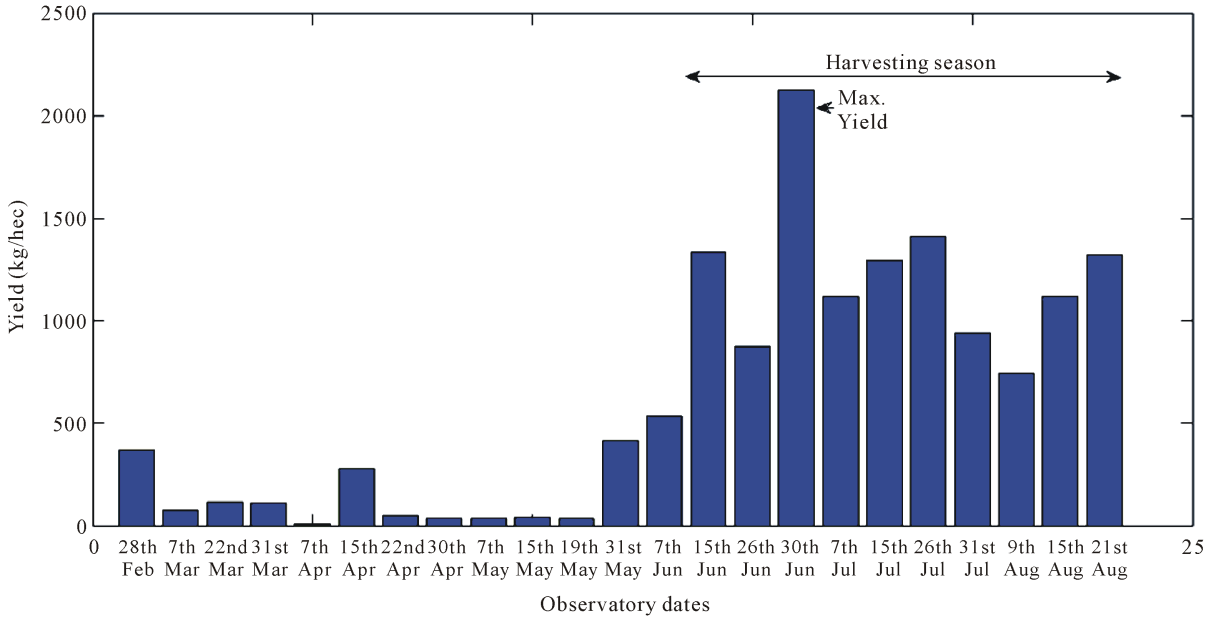

Figure 1. Statistical yield suggested by the third model for the yield of soybean in Central India.

Table 2. Correlation permutations amongst various environmental factors.

	Wind direction & temperature	Wind direction & humidity	Wind direction & wind speed	Temperature & humidity	Temperature & wind speed	Humidity & wind speed
Linear model	0.7919	0.2105	0.7554	0.6390	−0.2465	0.0168
Quadratic model	0.7930	0.2086	0.7420	−0.2465	0.6390	0.0168
Variable power model	0.7831	0.4155	0.5424	−0.1147	0.6038	0.1638

Table 3. Correct prediction percentage amongst the three models under consideration.

	Chaos
Linear model	0.2622
Quadratic model	0.2634
Variable power model	0.3554

the 2012 estimate, thereby proving the legitimacy of the accuracy of the computational calculation of yield using hidden environmental parameters.

References

[1] Schwartz, M.D. (1995) Detecting Structural Climate Change: An Air Mass-Based Approach in the North Central United States, 1958-1992. *Annals of the Association of American Geographers*, **85**, 553-568. http://dx.doi.org/10.1111/j.1467-8306.1995.tb01812.x

[2] Fahd, T. (1996) Botany and Agriculture. In: Morelon, R. and Rashed, R., Eds., *Encyclopedia of the History of Arabic Science*, Routledge, London, 815.

[3] Saxena, M.C. and Pandeny, R.K. (1972) Characteristics and Performance of Some Promising Varieties of Soybean. *Indian Journal of Agricultural Sciences*, **41**.

[4] Bansil, P.C. (1984) A Strategy for Self Sufficiency in Vegetable Oils. *Quarterly Economic and Agricultural Report*, **27**.

[5] Ahn, S.-J. (2008) Geometric Fitting of Parametric Curves and Surfaces. *Journal of Information Processing Systems*, **4**.

[6] (2013) All India State Wise Production and Yield of Soybean, First Estimate, Soybean Production Association of India.

[7] Devaney, R.L. (2003) An Introduction to Chaotic Dynamical Systems. 2nd Edition, Westview Press, Boulder.