A statistical approach to estimate seasonal crop production in India

Anil Ramakrishna

Department of Computer Science,
University of Southern California,
Los Angeles, CA
{akramakr}@usc.edu

Abstract. Estimation of crop production can be a very useful tool while planning a nation's agriculture. The use of statistical techniques in developing such tools is a popular and well studied problem. In this paper, we present a statistical model that predicts the total crop production for a geographical region, given the season of harvest and the total land used. We apply this model on four major crops of India and demonstrate the effectiveness of the system.

1 Introduction

Agriculture has a major contribution to India's GDP and is often referred to as the backbone of India. Hence a careful planning of the agricultural output is critical in maintaining the nation's economy. Estimation of crop production is a vital part of such a planning and statistical techniques provide a perfect solution to this estimation problem. In this paper, we present a system trained on the Indian agricultural data set [1] which uses these techniques to report the estimated yield of different crops based on a few input variables.

We initially apply multivariate polynomial regression to this problem and uncover the problems associated with such an approach. We also demonstrate the problems with using a traditional performance measure such as the Mean Squared Error (MSE) and propose to use a simple yet powerful alternative measure of performance for the prediction task. We then apply regression trees to this problem and demonstrate its superiority over simple regression through the alternate measure.

The remainder of the paper is organized as follows: In section two we discuss related work. In section three we explain our approach along with the experimental results. We conclude about the utility of the approach in section four.

2 Related Work

Numerous systems have been proposed to estimate the production of crops around the world using statistical techniques. [2] apply the technique of robust regression to estimate crop production as an alternate to least squares regression to avoid the problem with outliers. [3] use the Canadian satellite RADARSAT to monitor and estimate rice production in China. [4] examine the applicability of the Generalized Regression Neural Network model to forecast crop production. [5] examine the correlation between crop production and factors such as soil nitrogen content, etc. and identify the ones with high correlation.

In the Indian context however, there seem to be very few attempts to estimate the production of crops. [6] apply regression to estimate the effect of the total rainfall on the produced yield. [7] evaluate the effect of altering salinity on the production of crops using segmented linear regression. Given the importance of agriculture to India's overall economy, there is a huge incentive to develop systems capable of estimating the crop production for a given season and a given geographic location. We intend to address this exact issue in this work.

3 Approach

In this section we give a brief overview about the dataset followed by the actual approach.

3.1 Dataset

The dataset [1] was obtained from the recently created data portal of the Indian government [8] and it contains around 11 million entries of crop production between the years of 1998 to 2000. Each record contains 7 features: 'year', 'state', 'district', 'crop', 'season', 'area' of the agricultural land used to grow each crop and 'the final yield' in tonnes. The system we present uses regression to build a model using this dataset, which can be used to make predictions about new unseen data. We limit our analysis to four major crops: Rice, Sugarcane, Wheat and Maize, but note that the work can be directly extended to any other crop.

Data preprocessing Of these, both the 'state' and 'district' attributes provide geographic information at different granularity. We discard the 'state' attribute and use only the 'district' attribute for reasons explained later. We also discard the 'year' attribute as we suspect that the production depends mainly on the season. We use the final yield as the dependent variable for regression along with three independent variables: district, season and area. Additionally, since the 'district' and 'season' attributes are categorical, we expand each of these into a k-dimensional binary vector, where 'k' is the number of possible values for the corresponding attribute. Since the yield and the area are on different scales, we apply mean normalization on these for faster convergence. To prevent high bias, we also create 399 additional features from the 'area' attribute by raising its power to values ranging from 1 to 400.

3.2 Simple Regression

Our first attempt to build a predictor system was to use simple regularized multivariate regression on the processed dataset. We chose the best value for the regularization parameter lambda through cross-validation. The resultant training and test errors for the four crops are shown in table 1.

Table 1. Test Error for Simple Regression

Crop	Training Error	Test Error
Rice	0.0011	0.0014
Sugarcane	0.0065	0.0111
Wheat	0.000484	0.000487
Maize	0.000550	0.000543

When this model was applied on a few randomly selected test points, the predicted values for production are shown in table 2 below, along with the actual values.

Table 2. Predicted values for Simple Regression (Rice)

District	Season	Area (in 1000 hectares)	Production (in 1000 tonnes)	Predicted Value (in 1000 tonnes)
VISAKHAPATNAM	Rabi	3.00	6.00	33.30
KARIMNAGAR	Rabi	77.44	239.21	274.00
KORAPUT	Winter	90.00	161.00	148.58
KOLASIB	Rabi	4.75	11.80	31.23
DANTEWARA	Kharif	130.26	144.89	184.61

Table 3. Predicted values for Simple Regression (Sugarcane)

District	Season	Area (in 1000 hectares)	Production (in 1000 tonnes)	Predicted Value (in 1000 tonnes)
RUDRAPRAYAG	Kharif	5.07	269.25	253.20
CHANDAULI	Whole year	8.42	570.46	340.07
NAGPUR	Whole year	1.61	136.96	250.75
LUCKNOW	Whole year	1.00	60.00	217.14
MAU	Whole year	1.81	219.76	235.38

As seen, the simple regression model severely under performs at many data points. Even though the model reported low training and test errors, the prediction performance remains low. This is probably due to the fact that the output variable has very high variance and a little variation in the parameters can result in huge prediction differences. This also suggests that the simple error alone cannot be a strong indicator of the system's performance. To address this, we propose to use the **Mean Percentage Error (MPE)**. We define MPE as the

Table 4. Predicted values for Simple Regression (Wheat)

District	Season	Area (in 1000 hectares)	Production (in 1000 tonnes)	Predicted Value (in 1000 tonnes)
BELGAUM	Autumn	65.34	166.59	158.83
RANGAREDDY	Autumn	27.20	64.34	94.28
MAHASMUND	Autumn	52.18	120.10	136.60
HAZARIBAGH	Autumn	22.33	37.85	82.87
North (Tripura)	Autumn	63.15	196.44	167.36

Table 5. Predicted values for Simple Regression (Maize)

District	Season	Area (in 1000 hectares)	Production (in 1000 tonnes)	Predicted Value (in 1000 tonnes)
CHAMARAJANAGAR	Kharif	7.14	30.04	31.88
THIRUPPUR	Kharif	1.32	1.50	18.60
BHARATPUR	Summer	3.43	8.49	4.28
GULBARGA	Kharif	1.00	3.00	95.22
CHAMARAJANAGAR	Kharif	13.13	33.85	33.46

mean of the percentage change between the actual values and the predicted values of the model. Such a measure is independent of the data pre-processing steps and units. The test MPE values for simple regression are shown below.

Table 6. Test Mean Percentage Error for Simple Regression

Crop	Test MPE
Rice	126.4311
Sugarcane	142.5857
Wheat	58.7573
Maize	143.7319

Since the MPE values are very high, we need a better approach than simple regression to solve this problem. Here, we make use of the fact that the dataset includes records from different geographical regions of the country and that different regions could have different yields depending on several unobserved attributes such as soil fertility, etc. Hence the dataset can be grouped into several coherent clusters, where each cluster can be solved using a different regression problem. This was also the main reason for our decision of choosing district over state attribute as we believe that using district offers better precision. We turn to regression trees as they are an obvious solution here.

3.3 Regression Trees

Regression Trees are special decision trees that are constructed such that each internal node of the tree denotes a test on the attribute and the leaf nodes contain one or more data points that share some features. For each leaf node, a regression model is constructed using the data points associated with that node. For each new test point, we start at the root and arrive at a particular leaf node based on the attributes of the data point and make a prediction using the regression model of that leaf node. We use MATLAB's Regression Tree class to construct the decision trees from the dataset. The test MPE values for regression trees are shown below in table 7.

Table 7. Test Mean Percentage Error for Regression Trees

Crop	Test MPE
Rice	27.7266
Sugarcane	41.5930
Wheat	28.0755
Maize	36.8425

As seen above, the test MPE values are significantly better than those from simple regression. When the above model was applied on the same test points as before, the predicted values are shown below.

Table 8. Predicted values for Regression Trees (Rice)

District	Season	Area (in 1000 hectares)	Production (in 1000 tonnes)	Predicted Value (in 1000 tonnes)
VISAKHAPATNAM	Rabi	3.00	6.00	4.53
KARIMNAGAR	Rabi	77.44	239.21	241.23
KORAPUT	Winter	90.00	161.00	153.89
KOLASIB	Rabi	4.75	11.80	11.80
DANTEWARA	Kharif	130.26	144.89	155.35

Table 9. Predicted values for Regression Trees (Sugarcane)

District	Season	Area (in 1000 hectares)	Production (in 1000 tonnes)	Predicted Value (in 1000 tonnes)
RUDRAPRAYAG	Kharif	5.07	269.25	271.01
CHANDAULI	Whole year	8.42	570.46	561.08
NAGPUR	Whole year	1.61	136.96	139.08
LUCKNOW	Whole year	1.00	60.00	61.03
MAU	Whole year	1.81	219.76	219.76

The above examples illustrate the superior performance of regression trees over simple regression for this task.

Table 10. Predicted values for Regression Trees (Wheat)

District	Season	Area (in 1000 hectares)	Production (in 1000 tonnes)	Predicted Value (in 1000 tonnes)
BELGAUM	Autumn	65.34	166.59	162.55
RANGAREDDY	Autumn	27.20	64.34	61.14
MAHASMUND	Autumn	52.18	120.10	115.18
HAZARIBAGH	Autumn	22.33	37.85	35.79
North (Tripura)	Autumn	63.15	196.44	191.55

Table 11. Predicted values for Regression Trees (Maize)

District	Season	Area (in 1000 hectares)	Production (in 1000 tonnes)	Predicted Value (in 1000 tonnes)
CHAMARAJANAGAR	Kharif	7.14	30.04	28.89
THIRUPPUR	Kharif	1.32	1.50	2.02
BHARATPUR	Summer	3.43	8.49	8.25
GULBARGA	Kharif	1.00	3.00	3.55
CHAMARAJANAGAR	Kharif	13.13	33.85	37.83

4 Conclusion

We examined the problem of constructing a regression system to predict the crop production for four major crops, given the district, season and area of land used. We applied multivariate regression to the problem and identified the problems associated with this approach. We also used a unified performance metric that gives a better measurement of the model's performance compared to traditional metrics. We finally applied regression trees to the problem and demonstrated their superiority over simple regression.

Bibliography

[1] District-wise, season-wise crop production statistics from 1998 in india. `http://data.gov.in/dataset/district-wise-season-wise-crop-production-statistics-1998`.

[2] Robert Finger and Werner Hediger. The application of robust regression to a production function comparison-the example of swiss corn. 2009.

[3] Yun Shao, Xiangtao Fan, Hao Liu, Jianhua Xiao, S Ross, B Brisco, R Brown, and G Staples. Rice monitoring and production estimation using multitemporal radarsat. *Remote sensing of Environment*, 76(3):310–325, 2001.

[4] Jin Miaoguang and Jin Chaochong. Forecasting agricultural production via generalized regression neural network. In *Advanced Management of Information for Globalized Enterprises, 2008. AMIGE 2008. IEEE Symposium on*, pages 1–3. IEEE, 2008.

[5] Hari Dahal and JK Routray. Identifying associations between soil and production variables using linear multiple regression models. *Journal of Agriculture and Environment*, 12:27–37, 2013.

[6] B Parthasarathy, AA Munot, and DR Kothawale. Regression model for estimation of indian foodgrain production from summer monsoon rainfall. *Agricultural and Forest Meteorology*, 42(2):167–182, 1988.

[7] RJ Oosterbaan, DP Sharma, KN Singh, and KVGK Rao. Crop production and soil salinity: evaluation of field data from india by segmented linear regression with breakpoint.

[8] Indian data portal. `http://data.gov.in/`.

[9] Sushila Kaul. Bio-economic modelling of climate change on crop production in india. *New Delhi, India: Indian Agricultural Statistics Research Institute*, 2007.

[10] Josette Murphy and Leendert H Sprey. *Introduction to farm surveys*, volume 33. International Institute for Land Reclamation and Improvement, 1983.

[11] Indian Agricultural Statistics Research Institute. Research projects. `http://iasri.res.in/IASRIWEBSITE/SAMPLE_SURVEY.HTM#projects`.