Forecasting of rice yields in Guangxi Province China via Bi-LSTM

1st Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

3rd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

4th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

5th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

6th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

***Abstract*—To predict the rice yield of 81 counties in Guangxi province China, this paper picks out the algorithms with best performance in predicting the agricultural yield by testing several regression algorithms on winter wheat yield in the U.S which holds more data. The result of algorithm test shows that random forest and LSTM network are the best performance algorithms to predict the agricultural yield whose accuracy reach 85% more. Using the results of the algorithm test, the random forest and LSTM RNN with the best performance are used to predict the yield of early and second rice in Guangxi province. The results show that the accuracy of random forest without location information is 83.7%, while the accuracy of LSTM RNN is 87.7%. Both algorithms can use climate data to effectively predict crop yields for the current season.**

***Index Terms*—Prediction, Agricultural Yield, Random Forest, LSTM RNN**

I. Introduction

Precision agriculture, first proposed by the United States in the 1990s, points out a new direction for the development of modern agriculture. Precision agriculture is a modern agricultural technology based on 3S technology(GPS, GIS and RS), decision support technology and intelligent equipment technology to implement precise timing, positioning and quantitative control of agriculture, agricultural resources and farming. Its core lies in the collection and processing of farmland information, and combined with climate, satellite, geography and other external conditions for fine management and guidance of agricultural cultivation, so as to improve agricultural output and quality. Many developed countries and regions have advanced agricultural yield forecasting system, such as Monitoring Agricultural ResourceS(MARS) of EU.

MARS using remote sensing started in 1988, initially designed to apply emerging space technologies for providing independent and timely information on crop areas and yields. Since 1993, this activity has contributed towards a more effective and efficient management of the common agricultural policy(CAP) through the provision of a broader range of technical support services to DG Agriculture and Member-State Administrations. Since 2000, the expertise in crop yields has been applied outside the EU. Services have been developed to support EU aid and assistance policies and provide building blocks for a European capability for global agricultural monitoring and food security assessment.

Crop yield forecasting is undertaken to provide monthly bulletins forecasting crop yields to support the EU's Common Agriculture Policy(CAP). Providing early warning of crop shortages or failure provides rapid information for EU development aid activities to support food insecure countries, as part of the JRC work on global food security.

II. Related works

The research on corp yield forecasting mainly use satellite remote sensing data and sensor network data. Satellite remote sensing data are used to study the relationship between crop yield and sunshine and surface conditions in a large scope, such as country and river basin. The sensor network data is closer to the plants, which is more used to monitoring the condition within a farm.

A. Satellite remote sensing data

The basic idea of satellite remote sensing technology is that the wavelength and frequency of reflected waves are different in different growth stages of different crops, which result in different total energy and radiation of reflected waves. The vegetation growth can be obtained by monitoring the reflected waves on the ground through satellites, and then the yield of crops can be predicted. However, Such data will be restricted by resolution of satellites and other factors like cloud, which results in the increment of the cost.

Bastiaanssen(2003) [1] measure corp rotation cycle and predict corp yield in Indus basin base on satellite remote sensing data, their research finds that this model has better prediction accuracy on wheat, rice and sugarcane comparing

to cotton. Becker-Reshef(2010) [2] use regression model base on combination of corp data and daily surface reflections data to predict winter wheat yield in Ukraine, which can give alert on production shortage. De Wit(2007) [3] use Kalma Filter to assimilate the soil water content reflected by satellite remote sensing data, which improve the prediction accuracy of winter wheat yield.

B. Sensor network data

Comparing to the satellite data, sensor network data is both more accessible and more economical, which provides more accurate data on a small scale, always within a farm, and a better overview of local envirmental conditions, which can guide better farming.

Mkhabela(2011) [4] uses sensors like Advanced Very High Resolution Radiometer(AVHRR), Moderate-resolution Imaging Spectroradiometer(MODIS) e.t to get NDVI data, with such data they use regression model to predict the yield of soybean and spring wheat from 2000 to 2006 and get good performance which contral the error under 10%. Prasad(2006) [5] use NDVI, Vegetation Condition index(VCI) and Tempreture Condition index(TCI) data to monitor the drought and assess vegetayion health and yield, with piecewise linear regression method they predicted the soybean yield in lowa for 19 years.

III. Data

Dataset includes data in winter wheat yield in U.S with local clomate conditions in 2013 and 2014, which contains 150 counties from 5 states and has 26 features, 360042 entries.

A. Missing items

There are 654 missing items in data, accounting for 1.81% of the total data volume, among which 2013 accounts for 43.42%. Deleting an entry with missing items may result in the decline of continuity due to the characteristics of crop growth. Filling to missing items with the nearest and most recent entry due to the continuity of climate data in both space and time. A single miss item being fulled by its adjacent items, while continuous missing data can be fulled from two ends to the middle one.

B. Abnormal data

By looking at the data, apparent anomaly found in the yield. There are entries having nonzero yield with relatively short period of time from being seeded to harvest which was recorded by column *DayInSeason*.

Fig. 1a shows the distribution of duration between first record and the last in different locations, locations have 185 days of records account for 93.6%. Therefore, delete data from locations whose planting duration less than 185 days. However, these positions not all were fully recorded for 185 days shown by Fig. 1c, there are days not to be recorded for some reason. After all position with miss records being deleted, all the data had consistent planting duration shown by Fig. 1d.

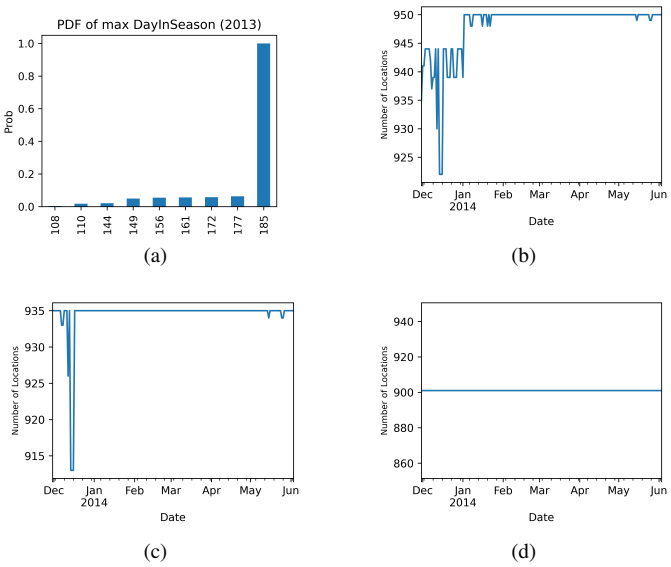


Fig. 1. Missing data processing

C. Map visualization

For data with location, map visualization can always help us find some intuitive relations between location and target. According to Fig. III-C-(b), there are two intuitive conclusions:

- The yield in the northern growing areas was significantly higher than that in the south.
- In the south growing areas, the more east the locations are the high the yields.

D. Feature selection

There are significant multicollinearity between some features in the data, which can lead to errors and distortions in the results [6]. By calculating the correlation coefficient between the features, findding that there are features with correlation higher than 90% with another, some even reach 99%. Table. I shows the features correlation high than 90%, these features accompanied with some other features unrelated to target such as *State*, *Date* e.t.

TABLE I
FEATURES CORRELATION HIGH THAN 90%

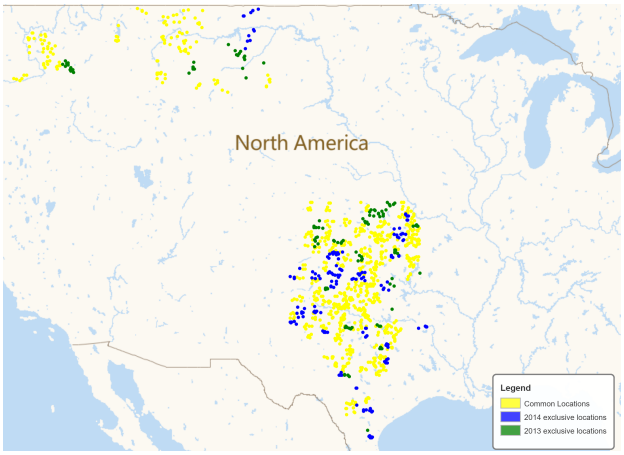
Feature 1	Feature 2	cor1	cor2	cor3
apparentTempMin	TempMin	0.99	-0.04	-0.06
apparentTempMax	TempMax	0.99	-0.14	-0.15
precipIntensity	precipIntensityMax	0.91	0.04	0.03
dewPoint	tempMin	0.91	0.01	-0.06
apparentTempMin	dewPoint	0.90	-0.04	0.01

[a]cor1: correlation between feature1 and feture2
[b]cor2: correlation between feature1 and yeild
[c]cor3: correlation between feature2 and yeild

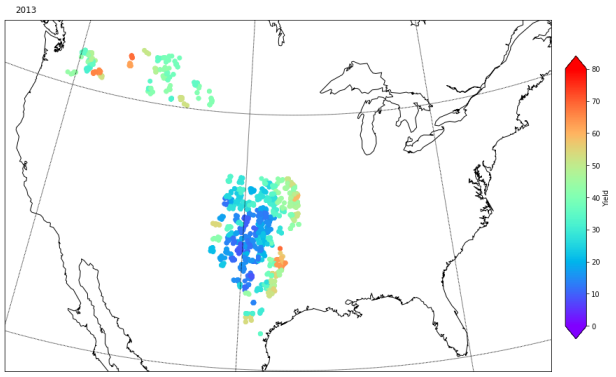
IV. Bi-LSTM Model

A. LSTM

Recurrent Neural Network is a kind of neural network that is good at processing time series data. It can well process



(a)



(b)

Fig. 2. (a)shows the location of all data being collected, yellow points represent locations monitered for two years which can be used for forecasting next year's yields via data of previous years. (b)shows the yeild varying by locations

task with continuous time and contextual relationship wich is exactly in line with the characteristics of the crop growth. The connections existing between each layer make the network better in learing time series relations, but on the other hand, this leads to gradient explosion in backpropagation. [7]

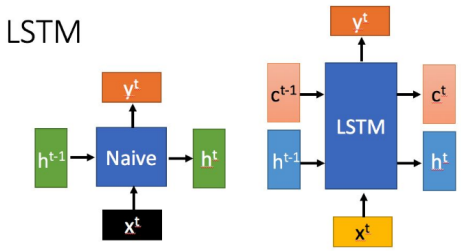


Fig. 3. LSTM and RNN structure

Fig. 3 shows the structure of LSTM and RNN. The structure of t-th RNN node is on the first image, x^t is the input of t-th node, h^{t-1} is the input of t-th node recived from $t-1$th node, $y^t$is the output and h^t is the output which transfer to next node, while LSTM have two more states c^t(cell state) and h^t(hidden state), state c in LSTM is the counterpart of state h in RNN, and normaly, state c changes slowly in front propagation which typicaly to add a number to c^{t-1}. On the contrary, h^t changes drasticly.

For a LSTM node shown in Fig. 4, four states f_t(forget gate), i_t(input gate), c_t(current state) and o_t(output gate) are generated by concating x^t and h^{t-1} via (1). Forget gate f_t contral the remain of information transtered from c_{t-1}, $[h_{t-1}, x_t]$ means to concat vector h_{t-1} and vector x_t, w_f and b_f are the weight and bias of forget gate f_t, σ is the Sigmoid activate function, expression of Sigmoid is (2).

$$\begin{cases} f_t = \sigma\left(w_f \cdot [h_{t-1}, x_t] + b_f\right) \\ i_t = \sigma\left(w_i \cdot [h_{t-1}, x_t] + b_i\right) \\ o_t = \sigma\left(w_o \cdot [h_{t-1}, x_t] + b_i\right) \end{cases} \tag{1}$$

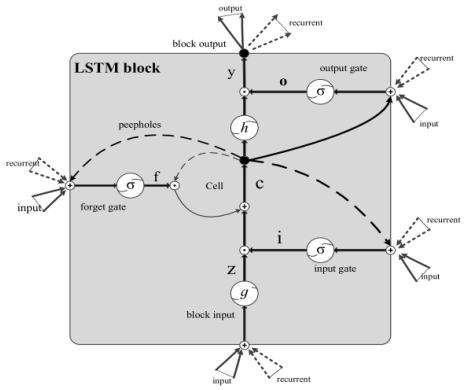


Fig. 4. LSTM node

$$\sigma\left(x\right) = \frac{1}{1+e^{-x}} \tag{2}$$

To calculate c_t, the input state $\widetilde{c}_t$ are needed to be calculate first. Use activate function $\tanh$ via (3), (4) and (1) to calculate c_t, the $\odot$ in (4) means corresponding element multiplication.

$$\tilde{c}_t = \tanh\left(W_c \cdot [h_{t-q}, x_t] + b_c\right) \tag{3}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot c_t \tag{4}$$

All these features of LSTM meet the corp growing process, there are weather fluctuatuons or natural disasters in short time, while there are stable natural cinditions over a long period of time.

B. Bi-LSTM Model

As shown in Fig. 5, build Bi-LSTM network using Keras and Tensorflow, there are two parts in model. The left part is three layers of stacked Bi-LSTM use to predict the yield by the weather data, right part use stacked fully connected network to link the county and the predicted yield.

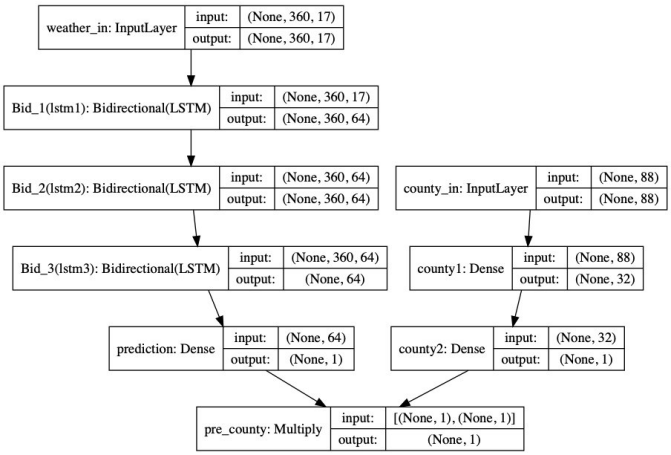


Fig. 5. Bi-LSTM Model

V. Experiment

A. Test method

Before testing, use *Min-Max* method shown by (5) to standarlize data and shuffle data due to the data is arranged by location. If not doing so, there will be data from new locations in train data while spliting the test set and train set.

$$\frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)} \tag{5}$$

Because the time complexity of SVR is more than twice of the data size [8], extract 10000 entries of data in order to reduce time, randomly extract and run multiple times to avoid accidental exceptions caused by extraction.

B. Test Result

The result of algorithm test is as shown in Table. III, and RF shows a severe overfiting among all the algoritms. The prediction contribution of features shown in Table. II, it is obviously that yield prediction is dominated by loction.

TABLE II
Feature contribution fo RF

Feature	Contribution
Longitude	0.7656
Latitude	0.2342
cloudCover	0.0001

Group the data by location and check the yield finding that the number of unique yields not match that of locations, which means there are entries share the same yield. Look further into data, finding that the yield of a day is not the real yield of current day, but the yield of final yield. What's more, the locations from the same conunty also share the same yield. That derectly results in loctions' neighborhoods get same yields, and the algorithm can give the exact yield according to the yield.

C. Overfitting in Random Forest

Simply remove longitude and latitude form data and run test again, get the learning curve shown by Fig. 6a. Clearly, it is not the "location" leads to the overfitting, but the structure of data.

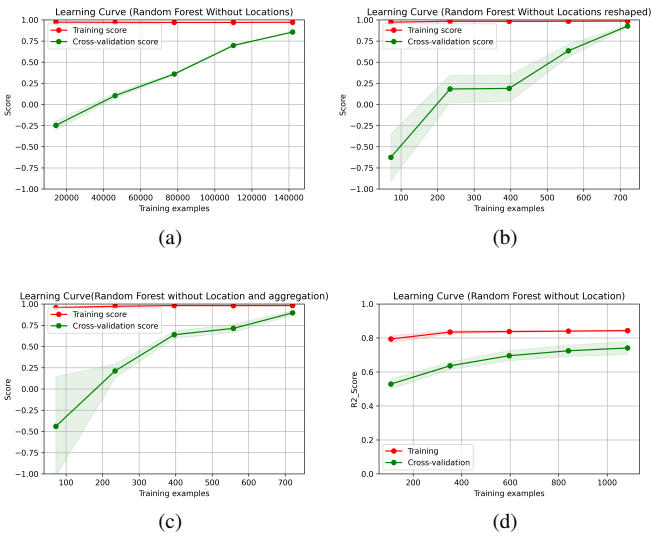


Fig. 6. Learning Curve

Reshape data to shrink the disparity between yield and data entries. For data in 2013, the origin shape is $(166685, 16)$, means that there are 186 days every location for 16 features. Transform the shape of data in to (901×2960) and run test again get the learning curve shown in Fig. 6b. These two methods can't solve the overfitting problem.

Try to aggregate the data to shrink the number of features. Calculate the mean values of ever 60 days and shrink features to 48, run test again and get the learning curve shown in Fig. 6c. There are some improvement after half of the dataset geting involved.

Extract some indirect features from the origin features, calculate the minimum and maximum value of features in 30 days, minimum maximum NDVI in 30 days, mean temperature difference and variance e.t using rolling window method from pandas. Then, run test base on these indirect features get the learning curve in Fig. 6d.

D. LSTM

Use EarlyStopping callback function to alleviate overfiting and set training epoch to 1000. The process of training shown in

The MSE of both training and cross validation go lower acompanied with the incresment of taining epochs. This procedure has been run several times, it always triggers the EarlyStopping callback function at between 100 to 300 epochs. The final mean cross validation MSE for all runs is 12.3793 comparing to RF's 23.72, the performance of LSTM is 47.84% better than RF which is quite accurate to forecast the yield of

TABLE III
PERFORMANCE OF ALGRORITHMS ON CORP YIELD PREDICTION(TOP5)

Directly run		Data Reshape		Data Compressed		Data aggregated	
Algorithm	R2 Error	Algorithm	R2 Error	Algorithm	R2 Error	Algorithm	R2 Error
RFR	0.8436	RFR	0.8983	RFR	0.8901	RFR	0.8361
KNN	0.7867	KNN	0.7867	KNN	0.7860	KNN	0.7835
PRR	0.6320	LRR	0.4954	LR	0.7710	LR	0.4303
PR	0.6258	GBR	0.0082	LRR	0.7303	LRR	0.4293

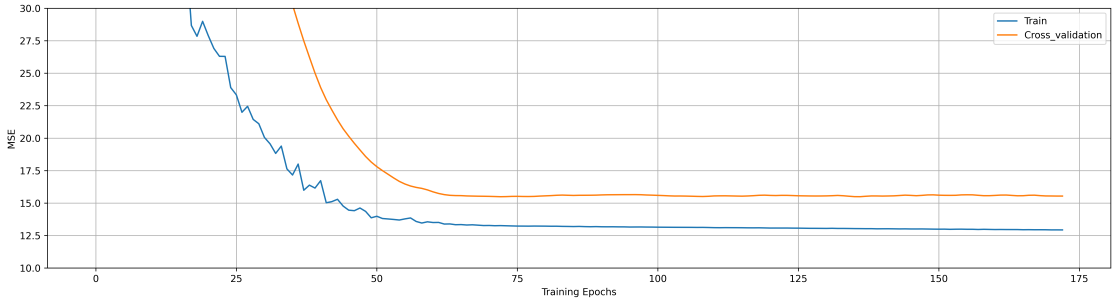


Fig. 7. Training Process of LSTM

corps whose R^2 error is 83.61% after solving the overfitting problem.

VI. CONCLUSION

It is found that both LSTM and RF algorithm do well in prediction of agricultural production, but there are some difference between them.

The implementation of LSTM is more comlpex, and many parameters need to be test and decided, and it is not easy to change the data's structure when there are new data get involved. There will be more time needed to config the model to meet the change of data either on features or recorded days. What's more, the training of a neural network consumes a lot of computing resources comparing to random forest. The random forest, on the contrary, less work needed to fit the change of data but more work needed to get pretreatment data, which alleviate the overfitting problem.

VII. FUTURE WORK

Some features which are crucial to the growing of crops, such as how farmers plant the crop, the condition of sow in a certain location e.t, are not involved to the models due to the lack of data source. The weather, undoubtedly, can influence the yield of corps, but some subjective behaviors can greatly improve the yield. The corps are not planted in an environment free of human intervention, water the crops when precipitation is in a low level is a natural behavior. If can get access such sort of data , the accuracy of prediction will be higher.

REFERENCES

[1] W. G. Bastiaanssen and S. Ali, "A new crop yield forecasting model based on satellite measurements applied across the Indus Basin, Pakistan," *Agriculture, Ecosystems & Environment*, vol. 94, no. 3, pp. 321–340, mar 2003. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0167880902000348

[2] I. Becker-Reshef, E. Vermote, M. Lindeman, and C. Justice, "A generalized regression-based model for forecasting winter wheat yields in Kansas and Ukraine using MODIS data," *Remote Sensing of Environment*, vol. 114, no. 6, pp. 1312–1323, jun 2010. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0034425710000325

[3] A. J. de Wit and C. A. van Diepen, "Crop model data assimilation with the Ensemble Kalman filter for improving regional crop yield forecasts," *Agricultural and Forest Meteorology*, vol. 146, no. 1-2, pp. 38–56, sep 2007.

[4] M. S. Mkhabela, P. Bullock, S. Raj, S. Wang, and Y. Yang, "Crop yield forecasting on the Canadian Prairies using MODIS NDVI data," *Agricultural and Forest Meteorology*, vol. 151, no. 3, pp. 385–393, mar 2011.

[5] A. K. Prasad, L. Chai, R. P. Singh, and M. Kafatos, "Crop yield estimation model for Iowa using remote sensing and surface parameters," *International Journal of Applied Earth Observation and Geoinformation*, vol. 8, no. 1, pp. 26–33, jan 2006.

[6] D. E. Farrar and R. R. Glauber, "Multicollinearity in regression analysis: the problem revisited," *The Review of Economic and Statistics*, pp. 92–107, 1967.

[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[8] S. M. Clarke, J. H. Griebsch, and T. W. Simpson, "Analysis of support vector regression for approximation of complex engineering analyses," 2005.