2015 8th International Symposium on Computational Intelligence and Design

Prediction of crop yield using big data

Wu Fan[1], Chen Chong[2], Guo Xiaoling[2], Yu Hua[*]
College of Engineering and Information Technology
University of Chinese Academy of Sciences
Beijing, China
wufan13@mails.ucas.ac.cn, ccdxo@126.com
guoxiaoling@ucas.ac.cn, yuh@ucas.ac.cn

Wang Juyun[2]
School of Science
Communication University of China
Beijing, China
wangjuyun@cuc.edu.cn

Abstract— **Quantifying the yield is essential to optimize policies to ensure food security. This paper aims at providing a new method to predict the crop yield based on big-data analysis technology, which differs with traditional methods in the structure of handling data and in the means of modeling. Firstly, the method can make full use of the existing massive agriculture-relevant datasets and can be still utilized with the volume of data growing rapidly, due to big-data friendly processing structure. Secondly, the "nearest neighbors" modeling, which employs results gained from the former data processing structure, provides a well-balanced result on the account of accuracy and prediction time in advance. Numerical examples on actual crop dataset in China from 1995-2014 have showed a better performance and an improved prediction accuracy of the proposed method compared with traditional ones.**

Keywords— ***crop yield prediction; big data; food security; nearest neighbors***

I. INTRODUCTION

As a comprehensive issue, food security is composed of several aspects, including producing sufficient food and stabling food supply in the market, etc. As the foundation of formulating policies to ensure food security, the prediction of crop yield, which is crucial to guarantee the well-balance between supply and demand, also raises much attention.

Every aspect of our lives is influenced by the abundance of message due to the fact that we are going through an era of data explosion. The information, characterized by alarming volume, velocity and variety, is often referred to as "Big Data" (Beyer and Laney, 2012[1]). As one fundamental element of national economy security, food security also confronts the promises and challenges brought about by the big data era, which can be made use of to enhance experts' understanding of the whole system.

Accuracy and the earliest time (time in advance) to offer the result have long been deemed as priorities in the prediction of crop yield. Extensive works have been conducted to identify contributing factors of accuracy and time in advance respectively. But these studies failed to take advantage of the massive data loads produced every day and required complex combination of information which are not derived directly from measuring equipment. Also, they lead to gathering consensus for those two facets have played two sides of a coin and ameliorating either will have a negative effect on the other. With the advancement of big data, improvement of both disadvantages proactively becomes an urgent call.

This study focuses on the yield prediction using agricultural data (mainly weather data) in China, which are mainly collected from 825 meteorological stations located in 34 districts. This information is collected continuously from different sources over a vast geographic scale. Since 1951, weather data have been generated and collected, including air pressure, temperature, relative humidity, precipitation, evaporation, wind speed, sunlight intensity and temperature at ground level. Yield data have also been archived through years, which comprises yield of varied species from all provinces. As a result, the continuous large geographic scale of data gathering provides a full view of the system and serves as the source of Big Data. In the paper, a MapReduce weather data processing structure and a nearest neighbors modeling are carried out in the hope to shed some light on the big data applications in the food security and agriculture area.

This paper is arranged as follows. Section 2 presents the progress in big data, data applications and crop yield prediction. In section 3, we propose a weather data processing structure and a model based on prior structure. Then experiment is conduct using already existing weather data and we discuss advantages of this new method. Finally in section 5 we make a conclusion and put forward our future work.

II. LITERATURE REVIEW

Effective strategies to improve the performance of crop yield prediction and methodologies to take advantage of massive data related to agriculture and food security require profound understanding. In the age of information, these objectives could be realized via Big Data applications.

A. Big Data in Agriculture and Food Security Arena

Current big data sources are:

1) Monitoring either the growing status of the crops or the condition of soil, which makes use of image recognition[2] technology;

2) Implementing Radio Frequency Identification (RFID) tags on the crops to build intelligent agriculture and IoT (internet of things)[3];

3) Remote sensing data from satellite[4];

4) Weather stations situated across the nation - the amount of data including air pressure, temperature, relative humidity, precipitation, evaporation, wind speed, sunlight intensity and temperature at ground level, which are generated and saved every minute during 50 years period, and can be over TB.

Data coming from these sources is in structural, semi-structural, or no-structural forms. Even though we have traditional methods like data mining or machine learning that can deal with massive datasets[5], few agricultural applicable scenes exist where we can coalesce these algorithm and data, which means these information cannot be effectively utilized.

978-1-4673-9587-8/15 $31.00 © 2015 IEEE
DOI 10.1109/ISCID.2015.191

B. Data Applications in Agriculture and Food Security Arena

From latest trends we can see that on one hand, crop yield forecast is among the most addressed topics (6.1%) in agriculture and food security arena, following climate change impact assessment (34.4%), crop growth simulation (18.5%), water resources (8.1%) and climate attribution (6.9%)[6]. On the other hand, we can also find that big data application is a trendy topic. However, little research is conducted in the mentioned subjects.

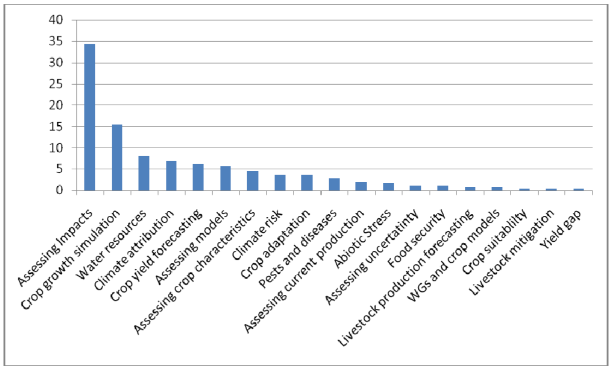


Fig. 1. Percentage of topics treated in the analyzed agricultural studies

C. Yield Prediction

Well-developed methods can be categorized as follows:

1) Statistics methods

Researches for statistics methods usually resolve the problem into "trend yield" and "weather yield" using smoothing method and build regression model after finding the key influential factors[7].

2) Remote sensing

$$Q = m \times p \tag{1}$$

Here Q is the total yield, p is yield per area and m is the crop planted area which is gained from analyzing images taken by remote sensing[8].

3) Crop growth simulation

This method combines numerical modeling with plant physiology for prediction.

4) Econometrics

This method applies input-output technique to estimate total yield using equations derived from economy area, whose figures cannot be got directly from sensors[9].

By synthesizing the discussions above, this study underlines the Big Data applications in agriculture and food security. A new structure to put the generated massive datasets into use is proposed and an analytical framework is raised to predict the crop yield.

III. Methodology

A. Data Preparation

The weather data used in this paper are collected by China Meteorological Administration (CMA) from 756 weather stations from 1951 up to now, contain varied agro-meteorological index, and pass strict quality inspection and control, for example, statistical values, especially the extreme parts, are tested and verified. Data are also proved to be time consistent by CMA. The relative humidity has an accuracy of 99.4% and other related indexes have an accuracy of more than 99.99%. The distribution of weather stations can be found in Figure 2.

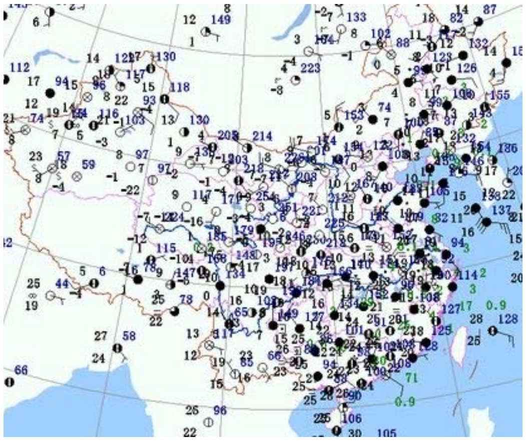


Fig. 2. Weather stations situated across China, from which data are collected

B. MapReduce Weather Data Processing Structure

Several traditional means have been developed to deal with weather data, mainly depending on statics. For example, when we are deciding if two years have similar weather patterns, we usually build models from mean values. However, this cannot reflect the accumulation of tiny differences in weather conditions, which may result in the departure in actual weather patterns. Therefore, big data technology is used to compute the whole datasets that we were not using before, and brings us to build models accounting for all the micro distinctions.

The new data processing structure divides the computing process into 2 phases, which can be described as Map and Reduce. In Map phase, the function manages all the input key/value (year/weather data) and produces median key/value. In Reduce phase, the function combines all the values which have the same median key, and then produces the final result.

The implementation process can be shown as in Figure 3. Firstly the weather data are partitioned into multiple sections. Secondly, the Map function is executed and the data are classified according to certain rules and then written to local hard drive. After Map phase, the Reduce function is executed, where the intermediate data having the same year value will be shuffled and consolidated, and the output is written to distributed file systems. The final result can be obtained by merging all the Reduce phase output at last.

The advantage of this new weather data processing structure is its high scalability, which allows users to compute and analyze large datasets[10].

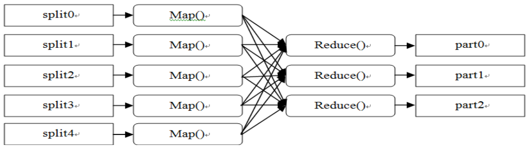


Fig. 3. MapReduce weather data processing structure

C. *Weather Similarity – Nearest Neigbors*

Weather similarity is defined by weather distances[11], which make use of multivariable analysis and varied weather indexes. It is based on such an assumption that each climatic index can be treated as one-dimensional space. Then for *m* kinds of indexes, an *m*-dimensional space is found. In this way, every year in the system can be seen as a spot in the *m*-dimensional space, and the valuation of similarity between two years can be calculated through reckoning the distance between the two years. Apparently, the smaller the distance is, the more similar the two years can be. Therefore, the similarity between any two years can be quantified, hence, nearest neighbors of a specific year can be found.

A P	A_1	A_2	...	A_n
P_1	X_{11}	X_{12}	...	X_{1n}
P_2	X_{21}	X_{22}	...	X_{2n}
...	......			
P_m	X_{m1}	X_{m2}	...	X_{mn}

Here, P stands for each year and A is the different properties of one specific year.

The measure for distances between any 2 spots P_s and P_t commonly uses Euclidean distance, which can be described as:

$$D_{(P_s-P_t)} = \sqrt{\sum\nolimits_1^n (X_{si} - X_{ti})} \tag{2}$$

D. *The ARMA Model*

The Box-Jenkins methodology using autoregressive moving average model (ARMA) has dominated time series forecasting[12].

An AR process can be described as a general linear model. One produces the output whose input is white noise, that is, the output variable depends linearly on its own previous values and a stochastic term:

$$y_t = a_0 + \sum_{i=1}^{p} a_i y_{t-i} + \varepsilon_t \tag{3}$$

where a_i is the parameter of model, a_0 is a constant, and ε_t is white noise.

Likewise, the moving-average (MA) model is a common approach for modeling univariate time series models:

$$x_t = \sum_{i=1}^{q} \beta_i \varepsilon_{t-i} \tag{4}$$

where β_i is the parameter of the model and ε_{t-i} is white noise error term.

Combining AR and MA models, ARMA models provide a parsimonious description of a stationary stochastic process in terms of two polynomials, which can be used to understand and predict future values given a time series of data[13].

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \tag{5}$$

The notation ARMA (p, q) refers to the model with p autoregressive terms and q moving-average terms.

IV. Experiments and Discussions

Crop yield prediction is actually an application based on a new processing structure. This new structure is used to manage data in order to search for similar year. And the prediction model is built on account of the result.

A. *Weather Data Processing*

Due to the large volume of weather data, traditional method cannot compute such datasets. The MapReduce weather data processing structure mentioned in 3.2 is implemented. 3 variables are selected to measure the differences: precipitation, intensity of sunshine and temperature at ground level, which are able to reflect the growing environment of crops. Daily mean and monthly mean is obtained through concurrent computing.

We utilize 3 computers, each of which has a 2G RAM and 3.2GHz CPU. One computer operates as the tasktracker and the other 2 operate as job trackers. The process can be divided as 3 steps:

1) Map - Calculate monthly mean value, using precipitation as an example, each line represents the precipitation of that month of the year:

PRE195101	5
PRE195102	11
PRE195103	17
PRE195104	36
PRE195105	28

PRE201406	47
PRE201407	48

2) Reduce - Combine intermediate data with the same key:

PRE19510648 PRE19510755 PRE19510848
PRE19510211 PRE19510317 PRE19510436
PRE19510528 PRE19510933 PRE19511023
PRE1951015 PRE19511118 PRE1951126

…

PRE2014012 PRE20140213 PRE20140317
PRE20140422 PRE20140536 PRE20140647
PRE20140748

3) Results stored in each part:

1951 - 5 11 17 36 28 48
55 48 33 23 18 6

1952 - 4 12 21 29 49 40
60 56 43 21 7 3

…

2013 - 3 9 15 24 42 45
55 45 31 13 12 9

2014 - 2 13 17 22 36 47
48

B. Search for Similar Years

Using weather similarity in 3.C, we can deduce the years, the weather condition of which is similar to the one of the target year. The process can be divided into 3 steps:

1) Having all the variables from 1955 till now, three 59× 12 matrices are established, each row of which represents the value of a variable in 12 months. A column can show the change of the variable over the same period. Then combing these 3 matrixes after conducting normalization, a 59×36 matrix is gained.

2) As discussed, differences can be gained by computing the norm of the target year row minus each row in the matrix:

$$G(I)=\sqrt{[G(traget)-G(other)]^2} \qquad (6)$$

(6) is the distance between a specific year and target year.

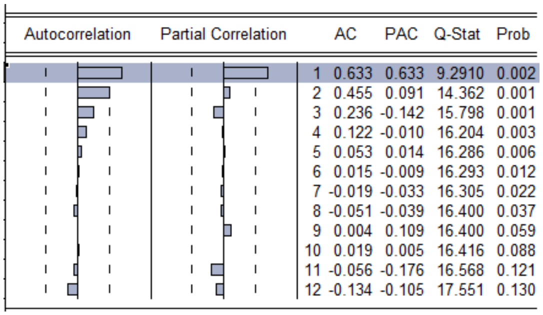
3) The distances between all years from 1955 to 2012 are derived from Step 3, and after sorting, we can have the 20 nearest neighbors–the weather condition of which is the most similar to the target year (2013).

TABLE I. RESULTS IN 4.B – 20 NEAREST NEIGHBORS

Year	R-square	Yield
1961	0.561272	13650.9
1975	0.52105	28451.5
1977	0.579961	28272.5
1983	0.539398	38727.5
1990	0.518831	44624.3
1993	0.541497	45648.8
1994	0.552663	44510.1
1995	0.464472	46661.8
1998	0.515252	51229.53
2002	0.530545	45705.75
2003	0.474187	43069.53
2004	0.50952	46946.95
2005	0.47894	48402.19
2006	0.511161	49804.23
2007	0.437145	50160.28
2008	0.541738	52870.92
2009	0.448476	53082.08
2010	0.415369	54647.71
2011	0.549987	57120.85
2012	0.553025	58957.97

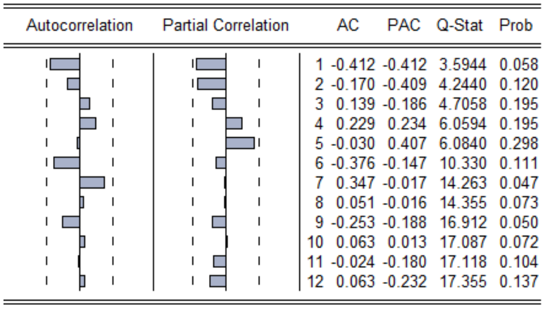
C. Modeling for Prediction

Analysis of the Augmented Dickey-Fuller(ADF) unit root test which can decide the series' constancy shows that second-order of original time series is stable, hence an ARMA (2, 1) model based on nearest neighbors is founded.



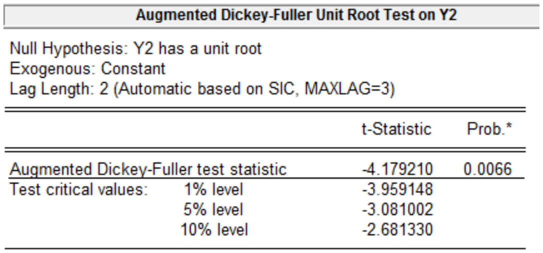
Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.633	0.633	9.2910	0.002
		2	0.455	0.091	14.362	0.001
		3	0.236	-0.142	15.798	0.001
		4	0.122	-0.010	16.204	0.003
		5	0.053	0.014	16.286	0.006
		6	0.015	-0.009	16.293	0.012
		7	-0.019	-0.033	16.305	0.022
		8	-0.051	-0.039	16.400	0.037
		9	0.004	0.109	16.400	0.059
		10	0.019	0.005	16.416	0.088
		11	-0.056	-0.176	16.568	0.121
		12	-0.134	-0.105	17.551	0.130

(a)



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.412	-0.412	3.5944	0.058
		2	-0.170	-0.409	4.2440	0.120
		3	0.139	-0.186	4.7058	0.195
		4	0.229	0.234	6.0594	0.195
		5	-0.030	0.407	6.0840	0.298
		6	-0.376	-0.147	10.330	0.111
		7	0.347	-0.017	14.263	0.047
		8	0.051	-0.016	14.355	0.073
		9	-0.253	-0.188	16.912	0.050
		10	0.063	0.013	17.087	0.072
		11	-0.024	-0.180	17.118	0.104
		12	0.063	-0.232	17.355	0.137

(b)

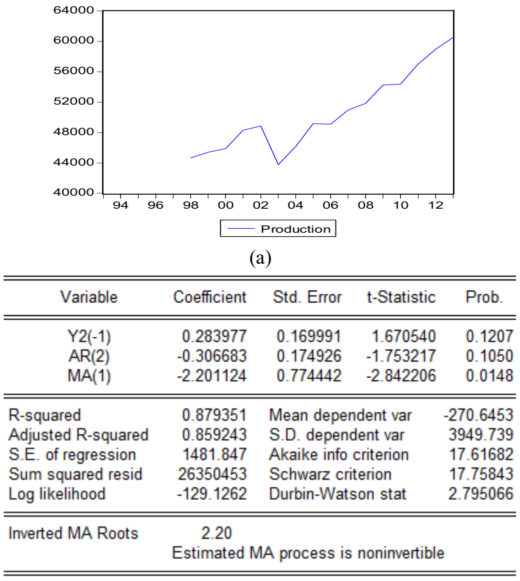


```
Augmented Dickey-Fuller Unit Root Test on Y2

Null Hypothesis: Y2 has a unit root
Exogenous: Constant
Lag Length: 2 (Automatic based on SIC, MAXLAG=3)

                                              t-Statistic    Prob.*

Augmented Dickey-Fuller test statistic        -4.179210     0.0066
Test critical values:      1% level           -3.959148
                           5% level           -3.081002
                           10% level          -2.681330
```

(c)

Fig. 4. Correlogram and Unit root test, (b)(c) show that second-order of origianl time series is stationary, and (a) indicates that ARMA (2, 1) is suitable for modeling[14]

We use the model to predict crop yield for year 2013 as an example. As can be seen in the diagram below, the relative deviation is 0.5% (actual yield is 60501.32 and the prediction is 60193.84).



(a)

```
Variable        Coefficient   Std. Error   t-Statistic    Prob.

Y2(-1)           0.283977     0.169991      1.670540     0.1207
AR(2)           -0.306683     0.174926     -1.753217     0.1050
MA(1)           -2.201124     0.774442     -2.842206     0.0148

R-squared            0.879351   Mean dependent var    -270.6453
Adjusted R-squared   0.859243   S.D. dependent var     3949.739
S.E. of regression   1481.847   Akaike info criterion  17.61682
Sum squared resid    26350453   Schwarz criterion      17.75843
Log likelihood      -129.1262   Durbin-Watson stat     2.795066

Inverted MA Roots         2.20
                    Estimated MA process is noninvertible
```

(b)

Fig. 5. ARMA(2,1) Modeling Result, (a) is the fitting cure of sereis and (b) shows the degree fit

D. *Evaluation and Discussions*

We calculate the crop yield from 2009-2013, the deviation off the actual number is below 5%, with an average accuracy of 97%. The traditional time series has an accuracy of 93%, when used in predicting the same year's crop yield, which can prove that it will lead to better result when building model using time series based on nearest neighbors. And it also indicates that crop yield relates closely to weather patterns.

We also compare our method with other existing ones included in 2.c and the results are listed as in Table 2 and Figure 6. As mentioned in introduction, accuracy and earliest time to provide the result (time in advance) are 2 priorities, thus the comparison use these 2 as basic standard.

TABLE II. COMPARISONS OF NEAREST NEIGHBORS METHOD AND TRADITIONAL METHODS

Method	**Accuracy**	**Time in advance**
Nearest Neighbors	>95%	6 months
Statistical method	>90%	12 months
Remote sensing	<90%	>6 months
Econometrics	>95%	<1 months

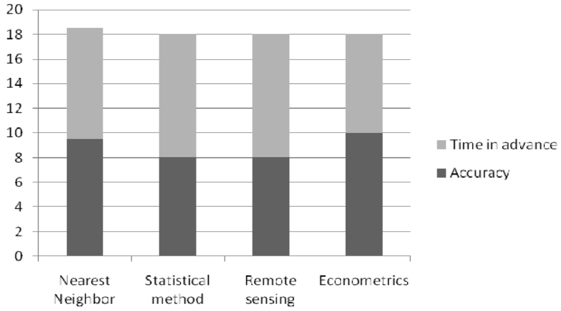


Fig. 6. Comprehensive measurement for crop yield prediction methods

As can be concluded from the chart, the nearest neighbors method using MapReduce weather data processing structure can reach a balance that both accuracy and time in advance can have preferable performances.

V. CONCLUSIONS

This paper proposes a solid architecture for managing big data in agriculture area. Using this method we can take advantage of the massive datasets already existed today and thus put into proper use. The architecture contains 3 parts. The first part is a MapReduce weather data processing structure, which runs and calculates big datasets on a group of computers. The second part is to find similar years, hence "nearest neighbor", which uses weather distances. The last part is to build ARMA model base on the "nearest year" and obtain the prediction number.

Our experimental evaluation reports a good performance of the "nearest neighbors" method, which also indicates the fact that crop yield relates closely to weather patterns. The method stands out on 2 dimensions:

Long "time in advance" - If we are to predict the yield of 2013, we use the data from March, 2012 to February, 2013, so that we can obtain the result about 5-6 months in advance. Accuracy is beyond average and therefore the method can be used to support decision and policy makers to guarantee food security.

The conclusions above summarize our trends of future work. First, with the faster accumulation of data, our new weather data processing structure can still be qualified to analyze these data. Second, if meticulous comparison is needed, the weather similarity calculation can also be integrated into the data processing section in order to control the computation time. Last but not least, this paper mainly focuses on agricultural data mining from time aspect, other scenarios can also be exploited like agricultural data mining from geographic aspect, that is to develop more applications based on the MapReduce weather data processing structure.

Acknowledgements

This research was supported by Y5110391A1/71450009 of National Natural Science Foundation of China. Authors also appreciate the support from National Development and Reform Commission(NDRC) comprising Fang Yuanyuan for the help in data collection.

References

[1] M. A. Beyer and D. Laney, “The importance of ‘big data’: a definition,” Stamford, CT: Gartner.

[2] Z. F. Sun, K.M. Du, F. X. Zheng and S. Y. Yin, “Outlook of big data applied in Intelligent Agriculture,” Journal of Agricultural Science and Technology, 2013, vol. 15(6), pp. 63-71.

[3] Z. F. Sun, K. M. Du, and S. Y. Yin, “ Future of IoT and its application in agricultrue,” Agriculture Network Information, 2010, vol. 5, pp. 5-8.

[4] T. Goddard, L. Kryzanowski, K. Cannon, C. Izaurralde and T. Martin, “Potential for integrated GIS-agriculture models for precision farming systems,” University of Alberta, Edmonton Canada, 1996.

[5] X. Wu, X. Zhu, G. Q. Wu and W. Ding, “Data mining with big data,” IEEE Transactions on Knowledge & Data Engineering, 2014, vol.26(1), pp. 97-107.

[6] J. Ramirez-Villegas and A. Challinor, “Assessing relevant climate data for agricultural applications,” Agricultural & Forest Meteorology, 2012, vol. 161(3), pp. 26–45.

[7] F. T. Wang, Y. Z. Li and S. L. Wang, “Introduction to climatic simulation and modeling of crop yeild,” Beijing: Science Press, 1990, pp. 58.

[8] C. O. Stockle., S. A. Martin and G. S. Campbell, “CropSyst, a cropping systems simulation model: water/nitrogen budgets and crop yield,” Agricultural Systems, 1994, vol. 46(3), pp. 335-359.

[9] X. K. Chen and C.H. Yang, “Characteristic of agricultural complex giant system and national grain output prediction,” System Engineering Theory and Practice, 2002, vol. 6(6), pp. 120-125.

[10] J. Dean and S. Ghemawat, “Mapreduce: simplified data processing on large clusters,” Operating Systems Design & Implementation, 1989, vol. 51(1), pp. 147–152.

[11] X. E. Wang and W.L. Decker, “The use of distance coefficient in the research on agroclimatological resemblance,” Journal of Nanjing Institute of Meteorology, 1989, vol. 12(2), pp. 187-199

[12] J. Durbin, “Introduction to state space time series analysis,” State Space & Unobserved Component Models, 2004, pp. 3-25.

[13] B. Gep and G. Jenkins, “Time series analysis: forecasting and control,” IEEE Transactions on Automatic Control, vol. 17(2), 1976, pp. 281 - 283.

[14] D. L. Mo, “Using ADF to build time series model,” Times Finance, vol. 4, 2010, pp. 46-48, doi:10.3969/j.issn.1672-8661.2010.04.024.