Modelling and Forecasting of Commodity Trading Price

Jun Liu\textsuperscript{1,2,3,4,*}, Yuan Tian\textsuperscript{1,2,3,4,b} and Qing Yan\textsuperscript{1,2,3,4,c}

\textsuperscript{1}Center for Data Science, BUPT, Beijing 100876, China.
\textsuperscript{2}School of Information and Communication Engineering, BUPT, Beijing 100876, China.
\textsuperscript{3}Beijing Laboratory of Advanced Information Networks, BUPT, Beijing 100876, China.
\textsuperscript{4}Beijing Key Laboratory of Network System Architecture and Convergence, BUPT, Beijing 100876, China.

\*\texttt{liujun@bupt.edu.cn}, \texttt{lance6716@bupt.edu.cn}, \texttt{yq@haohandata.com.cn}

Abstract. This paper proposes and improves a model for China’s cotton reserves trading market, generates a more accurate price level table in contrast to the widely used China Cotton Association(CCA)’s table, and predicts the future trading price with this model. Data used is all 29895 trade records of cotton reserves from May, 2016 to Sept, 2016. In this paper, we firstly give a briefly introduction to the data as well as basic knowledge of cotton, especially the CCA’s price level table as the target to improve accuracy. We then present a multiple linear regression with dummy variable model, which can reduce the error on predicting the trading price of current month, but still remains some problems such as the result shows a batch of cotton with better quality may get a lower price. So we finally advance the model with multidimensional isotonic regression and more factors taken into consideration. Our last model could be used to predict the trade price of both current month and next month with approximately 4\% mean absolute percentage error(MAPE).

1. Introduction

Commodity trading servers an important role in the national economy and has been widely studied by economics, statistics and many other subjects\cite{1}\cite{2}\cite{3}. Meanwhile China’s commodity market has received increasingly attention and research these years for its unique economic system \cite{4}\cite{5}. In this paper, we focus on cotton as a representative of the commodity, study the modelling and forecasting of China’s cotton price with the data provided by Information Centre of China National Cotton Reserves Corp(CNRC). China has sold from its cotton reserves of 12’ and 13’ and put up for auction, whose trading price is determined of bidding of textile enterprises as well as guided by some policies and plans to some extent. The final trading price is a good indicator of cotton industry. We have studied and proposed models to analysis and find hidden patterns of cotton reserves trading price, which aim to improve policy effectiveness, advance accuracy of cotton index and predict trading price of next month.

2. Data Structure

Data used in this paper was collected from May, 2016 to Sept, 2016 by CNRC’s cotton reserve trading system, with total amount of 29895 cotton reserves transactions. The data could be introduced as three parts as follow.
2.1. Basic cotton knowledge
CNCRC provided cotton index of basic cotton type 3128B (meaning of this code will be interpreted in next section) every week. This index reflects the ups and downs of cotton market and is used as well as the other index to calculate the base price of cotton auction.

CCA also publish a price level table every month to calculate the exact price of cotton of different properties. The price then can be calculated by finding the according difference of price of each properties, adding or subtracting it from the basic type 3128B’s price. Table 1 shows part of CCA’s price level table.

| Staple length | Difference of price | Micronaire level | Difference of price |
|---------------|---------------------|------------------|---------------------|
| 30            | 600                 | A                | 200                 |
| 29            | 300                 | B1               | 0                   |
| 28            | 0                   | B2               | 0                   |
| 27            | -400                | C1               | -600                |
| 26            | -850                | C2               | -100                |

The calculated result should be referenced during the cotton trading as CCA stated.

2.2. Public quality inspection data
Public quality inspection was made before the cotton reserves was listed for sale. It inspects many properties but nine important ones are listed in Table 2. Some of these properties are involved in CCA’s price level table and others are concerned by the cotton corporations in actual trading. Others properties beyond those nine, such as production year and production area could be reflected by those nine properties and thus omitted.

Table 2. Cotton properties and values.

| Properties         | Values and explanations |
|--------------------|-------------------------|
| Colour             | ‘11’, ‘21’, ‘31’, ‘41’, ‘51’, ‘12’, ‘22’, ‘32’, ‘13’, ‘23’, ‘33’, ‘14’, ‘24’.  
(First character represents brightness and last character represents whiteness. Small value indicates high quality.) |
| Average micronaire | Decimal and according levels  
(Value of 3.4-and-lower: C1, 3.5~3.6: B1, 3.7~4.2: A, 4.3~4.9: B2, 5.0-and-higher: C2, A indicate moderate value and high quality.) |
| Rolling quality    | ‘P1’, ‘P2’, ‘P3’.  
(P1 indicate high quality) |
| Staple length      | Integer from 21 to 33. |
| Fibre strength     | Integer from 21 to 33 and according levels.  
(Value of 31~33: S1, 29~30: S2, 26~28: S3, 24~25: S4, 21~23: S5.) |
| Length uniformity  | Integer from 75 to 87 and according levels.  
(Value of 86~87: U1, 83~85: U2, 80~82: U3, 77~79: U4, 75~76: U5.) |
| Other fibre        | ‘N’ for none and ‘L’ for low. |
| Heterogeneity      | Percentage (Not included in CCA’s table) |
| Moisture regain    | Percentage (Not included in CCA’s table) |

So the basic type 3128B represents for ‘colour 31, staple length 28, average micronaire B1 or B2’.

2.3. Trading records
Many information was recorded but trading date, price of 3128B on that day and trading price are merely involved in this paper. Price of 3128B is used for auction base price, the auction is twice a day with over 100 batches of cotton on sale. Bid for any batch of cotton will reset global countdown for every cotton batch, so the final trade price is an equitable and systematic indicator of cotton value.

3. Linear regression model

3.1. Trading records
CCA’s price level table was widely used in actual cotton trade these years, but its accuracy and room for improvement remain to be studied. Our first goal is to propose a model which could generate a more accurate price level table. Part of CCA’s table (May, 2016) shows as table 1.

Linear regression is a widely used method with a simple form and strong ability to interpret, and can be easily found in most statistic and economic textbooks. So we build our first model on linear regression. Cotton trading data and CCA’s table indicate a lot of information which specifies the exact form of linear regression.

We could infer from the price level table that difference price of each cotton property could be calculated independently, thus the correlation among them must be weak. That is a prerequisite for linear regression. We subsequently perform a significance tests of Kendall’s Tau-b, the result shows most properties have slightly correlation (absolute value below 0.3) except staple length and length uniformity have a tau value of 0.58. A possible explanation is cotton of high quality will have both long and neat staple. These two properties will be treated carefully in the followi

Another point is though all properties are order variables, the table shows that difference price between two adjacent level of some properties are not constant. And some properties such as micronaire are complex to find a linear relationship, so it is suitable to use dummy variables for all properties. Standard values of each property are dropped to avoid multi-collinearity, which is known for “dummy variable trap”.

We take first seven properties mentioned in table 2 into consideration and calculate the regression for each month using 80% data as training set. The regression has a form of

\[ y = \beta_0 + \beta_{\text{colour}1} \times x_{\text{colour}1} + \beta_{\text{colour}2} \times x_{\text{colour}2} + \cdots + \beta_{\text{rolling}1} \times x_{\text{rolling}1} + \cdots \]

The coefficient of \( x \) is the difference of price in table.

\( \beta_0 \) represents the average trade price of each month, which is similar to China Cotton Index 3128B (CC Index 3128B). The validity of the model can be checked from the relationship between that two variables. Table 3 shows average values of \( \beta_0 \) and CC Index 3128B from May, 2016 to Sept, 2016:

|       | May  | June | July | Aug  | Sept |
|-------|------|------|------|------|------|
| \( \beta_0 \) | 12011 | 12185 | 15076 | 14377 | 14355 |
| CC Index 3128B | 12510 | 12694 | 14316 | 14673 | 14257 |

Mean absolute percentage error(MAPE) of model and CCA’s price level table are listed in Table 4

|       | May  | June | July | Aug  | Sept |
|-------|------|------|------|------|------|
| CCA’s MAPE | 4.3%  | 4.0%  | 5.6%  | 4.2%  | 4.6%  |
| Model’s MAPE | 2.3%  | 2.9%  | 3.9%  | 2.3%  | 3.5%  |

From table 3 and 4 we could draw a conclusion that our linear regression model does improve the accuracy in contrast to CCA’s price level table.

3.2. Problems remained
However, there are still some points needed further improvement. Table 5 shows part of the price level table generated by our model.

**Table 5.** Part of model generated table.

| Staple length | Difference of price | Length uniformity | Difference of price |
|---------------|---------------------|-------------------|---------------------|
| 30            | 342                 | U1                | No data             |
| 29            | 207                 | U2                | 187                 |
| 28            | 0                   | U3                | 0                   |
| 27            | -349                | U4                | -96.5               |
| 26            | -567                | U5                | 258                 |

One main problem is this price table doesn’t follow the prior knowledge of “higher price better quality”, for example, U5 is the worst length uniformity level but gets highest price. This is both a true reflection of the trading record to some extent after we take a drill-down analysis under current assumption, and a common representation of multi-collinearity as we mentioned in section 3.1[6]. There are other regression methods such as lasso regression, ridge regression, to solve this problem[7], but they are not suitable in our situation of dummy variables for order variable.

Another problem is price of a certain property varies a lot among these five months, which may lead to a low acceptance of public and cotton corporations.

After making more research and having a better understanding of cotton trade business, we proposed an improved model using multidimensional isotonic regression.

4. **Isotonic regression model**

4.1. **Improvement to previous model**

The improvements we made could be roughly divided into three part: a) using isotonic regression to keep high-quality cotton evaluated high price, b) using CNCRC’s 3128B week price to eliminate the error of base price changing within a month, c) taking more properties into model factors.

First improvement is using isotonic regression. It is a widely used regression method of fitting a sequence of variables \( \{x_i\} \) into \( \{y_i\} \) under some partial ordering constrains, for example \( y_i < y_{i+1} \), with a minimum of \( \sum (y_i - x_i)^2 \). Many isotonic regression algorithms have been proposed such as Pool Adjacent Violators Algorithm(PAVA)[8], Least square isotonic regression in two dimensions[9], Merge and Chop Algorithm[10] etc. For our condition, isotonic regression is a constrain to combine the “higher price better quality” prior knowledge with our model. In order to apply isotonic regression, further processing is required to indicate the partial order relationship of properties, which will be introduced in next section.

Next improvement is replacing trade price with difference price between trade price and base price as target variable of regression, in order to eliminate the error caused by market goes up and down within a month. This could account for the above-average MAPE 3.9% of July, because CNCRC’s 3128B week price shows there was a sharp ascending in July.

Third improvement is adjustment on taking more cotton properties into model’s factors to fit the actual cotton trading business. Levels of properties are split into integer values to make a more accurate prediction, and heterogeneity and moisture regain are used to calculate the net price of cotton which is more concerned by cotton corporations.

4.2. **Improvement to previous model**

The origin data is firstly calculated the auxiliary variables such as net price or variables used for isotonic regression, then processed through multidimensional isotonic regression with seven properties as input and difference of trading price and base price as output, finally the output prices of each combinations of properties values are decomposed into seven dimensions representing each property to generate the price level table.
The multidimensional isotonic regression algorithm we used is multiisotonic component of Python Scikit-learn package (https://github.com/alexfields/multiisotonic). The interface of this component only accept multidimensional vector \( X_i \) as constrains and corresponding \( y_i \) as input, returns \( y'_i \) with minimum squared distance from \( y_i \), for \( y'_i \geq y_j \) if all components of \( X_i \) are greater than or equal to corresponding components of \( X_i \). So colour and micronaire characteristics should be replaced by auxiliary variables since they don’t follow an increasing numeric value with increasing quality. As shown in Table 2, first character of colour represents brightness and the last character represents whiteness, and both represent a better quality with lower value. So auxiliary variables for colour could be generated by splitting colour measurements values and negating it. Quality of micronaire follows umbrella order \( C1 \preceq B1 \preceq A \succeq B2 \succeq C2 \). Auxiliary variables suitable for that Python package interface shows in Table 6.

**Table 6.** Auxiliary variables for micronaire

| Average micronaire | Auxiliary variables |
|--------------------|---------------------|
| A                  | (3, 3)              |
| B1                 | (3, 2)              |
| B2                 | (2, 3)              |
| C1                 | (3, 1)              |
| C2                 | (1, 3)              |

Then the “higher price better quality” result of cotton property combinations appeared in each month is returned, which need deduplication and decomposition into seven dimensions for generating the table. This could be easily done using multiple linear regression with dummy variables from our first model. Table 7 shows the result generated by isotonic regression model.

**Table 7.** part of table of isotonic regression model.

| Staple length | Difference of price | Length uniformity | Difference of price |
|---------------|---------------------|-------------------|---------------------|
| 30            | 345                 | 83                | 349                 |
| 29            | 169                 | 82                | 182                 |
| 28            | 0                   | 81                | 0                   |
| 27            | -258                | 80                | -158                |
| 26            | -396                | 79                | -345                |

MAPE of this model as well as CCA’s table and our first model lists for convenience in Table 8.

**Table 8.** MAPE of different tables.

|                   | May  | June | July | Aug  | Sept |
|-------------------|------|------|------|------|------|
| CCA’s MAPE        | 4.3% | 4.0% | 5.6% | 4.2% | 4.6% |
| First model’s MAPE| 2.3% | 2.9% | 3.9% | 2.3% | 3.5% |
| Second model’s MAPE| 3.2% | 3.1% | 3.7% | 2.1% | 4.0% |

Error slightly improved in contrast with our first linear regression model, one possible reason is that there is a trade-off between the accuracy and “higher price better quality” knowledge, which is very similar to the trade-off between empirical error and generalization error. Considering the model will be announced to guide cotton’s trade, there is no strong reason to violate the widely accepted rule of “higher price better quality”. So it is agreed to us all that isotonic regression model is more acceptable.

5. Prediction of future trade price
Isotonic regression model generates price level tables that have steadier values among months, so it is possible that price level table from previous month could be used to predict the price of next month. That is a valuable application in cotton reserves trade. The accuracy of next-month prediction could be evaluated by the MAPE of difference price between CNCRC’s 3128B week price and predicted trading price by the table of every trade cotton. We then apply the price level table of previous month to data of next month to calculate average error of a month. The result shows in Table 9.

|       |       |       |       |
|-------|-------|-------|-------|
| June  | July  | Aug   | Sept  |
| 3.3%  | 3.5%  | 3.1%  | 4.2%  |

Table 9. MAPE of prediction on next month's data.

Contrast the prediction result with isotonic regression result of each month, we could find that the error only improves slightly on most month, which indicates that our price level table is general through two adjacent months, and the effect of the prediction is acceptable.

6. Future works
The isotonic regression model generates price level table which keeps a balance between accuracy and “higher price better quality”. But there are still some points needed study and improvement.

CCA’s price level table has a simple form as well as some limits. Combines of cotton properties, such as staple length and length uniformity, may generate a new auxiliary property such as staple length level, to make use of correlation of current seven properties and advance accuracy.

CCA’s monthly table may be advanced to weekly table or a period of floating windows, in order to make a quick response to cotton market changes.

Intercept $\beta_0$ of linear regression model and CC Index 3128B may also be used to analysis and predict the cotton market ups and downs. Future cooperation with CNCRC is planning and studies on cotton will be further explored to produce more valuable results.

References
[1] Bazzi S and Blattman C 2014 Economic Shocks and Conflict: Evidence from Commodity Prices American Economic Journal: Macroeconomics 6 1-38.
[2] Brooks C, Prokopczuk M and Wu Y 2013 Commodity futures prices: More evidence on forecast power, risk premia and the theory of storage The Quarterly Review of Economics and Finance 53 73-85.
[3] Hodgson G 2016 Karl Polanyi on economy and society: a critical analysis of core concepts Review of Social Economy 75 1-25.
[4] Hammoudeh S, Nguyen D, Reboredo J and Wen X 2014 Dependence of stock and commodity futures markets in China: Implications for portfolio investment Emerging Markets Review 21 183-200.
[5] Wang X and Zhang C 2014 The impacts of global oil price shocks on China’s fundamental industries Energy Policy 68 394-402.
[6] Yoo W, Mayberry R, Bae S, Singh K, He Q P and Lillard Jr J W 2014 A study of effects of multicollinearity in the multivariable analysis International journal of applied science and technology 4 9
[7] Sagaert Y, Aghazzaf E, Kourentzes N and Desmet B 2018 Tactical sales forecasting using a very large set of macroeconomic indicators European Journal of Operational Research 264 558-569.
[8] Leeuw J, Hornik K and Mair P 2009 Isotone Optimization inR: Pool-Adjacent-Violators Algorithm (PAVA) and Active Set Methods Journal of Statistical Software 32 1-24.
[9] Spouge J, Wan H and Wilbur W 2003 Least Squares Isotonic Regression in Two Dimensions Journal of Optimization Theory and Applications 117 585-605.
[10] Hu X and Hansohn J 2008 Merge and chop in the computation for isotonic regressions Journal of Statistical Planning and Inference 138 3099-3106.