Forecasting on Equipment Manufacturing Industry Development in View of Big Data

Xiaofei Xu[1] and Yanjuan Cui[2,*]

[1]School of Business, Beijing Language and Culture University, Beijing, China, 100083
[2]School of Management, Dalian Polytechnic University, Dalian, China, 116034
[*]Corresponding author

***Abstract*—Big data are widely used today, whether and how to use big data in economic variables forecast has become a new field of economic research. The equipment manufacturing industry is the foundation of the national economic development, so forecasts the development of equipment manufacturing industry is a very important research. In equipment manufacturing industry development forecasting, two types of data can be applied, namely traditional government statistical data and online data. Government statistical data are well-structured, whereas online data is unstructured information. This paper explores whether online data can help us to forecast equipment manufacturing industry development and analyze the best model to forecast. We find that traditional government statistical data and online data both can help forecast, so when we doing the forecasting, we should use the traditional government statistical data and online data at the same time.**

Keywords- forecasting; equipment manufacturing industry; big data; metal products industry; Automobile manufacturing industry

I. INTRODUCTION

Big data should not be static, but should be the collision and aggregation with the surrounding data, and it is the embodiment of government or corporate insight and action. Big data has four basic elements, which are early warning, forecasting, decision-making and intelligence. For the early warning, through the data acquisition, data mining, data analysis forecasts and warns the already existing risk. For the forecasting, based on the vertical axis of time, forms guidance for some relatively long time judgment. The decision making refers to form the data analysis and decision making conclusions based on all relevant data linkage. Intelligence is when we based on the analysis and judgment of the real problem, the realization of intelligent behavior by means of technology. Four elements interpret the core of big data from function, but ultimately achieve these features need to return to the big data applications, only through the application can realize the real function of big data.

The equipment manufacturing industry is the advanced industry producing technology and equipment manufactures for the national economy and national defense construction, is the core component of the manufacturing sector, and is the foundation of the national economic development especially industrial development, so forecasts the development of equipment manufacturing industry is a very important research. This paper selects the representative industries in the equipment manufacturing industry, and forecasts the development of them by using big data.

II. DATA DESCRIPTION

In this research, the evaluation variable (i.e., forecast variable) is export delivery value of equipment manufacturing industry; monthly data (i.e., Y) are available from Jan. 2012 to Dec.2016 from the People's Republic of China Statistics Bureau network. The explanatory variables in this research are divided into two categories. One category represents government statistical indicators; these data are available monthly from the People’s Republic of China Statistics Bureau network from 2012 to 2016. We choose 9 indicators which are closely related to the equipment manufacturing industry, including inventory, gross profit, total liabilities, etc., the statistical characteristics are shown in Table 1.

TABLE I. CHARACTERISTICS OF GOVERNMENT STATISTICAL VARIABLES

Variable	Mean	Std. Dev.	Min	Max
Y	6005.92	1574.89	5082.83	7614.80
X1	195.03	575.26	-1746.90	1103.60
X2	1718.87	1355.62	-1181.40	5183.87
X3	104.56	55.77	-17.43	222.30
X4	1791.31	797.97	1096.76	3404.70
X5	1877.15	1433.71	-1571.44	4810.60
X6	820.61	977.75	-1976.83	2377.80
X7	23746.74	8330.51	19760.82	36397.70
X8	27780.03	9851.89	26471.97	43904.60
X9	3209.28	1689.73	971.81	7041.40

This paper introduces another category of explanatory variables which is using online data provided by Baidu index. The Baidu index scientifically analyzes and calculates the weighted sum of the frequency of each keyword in Baidu Webpage searching. The variables of online data selected in this study are divided into four categories; those are related to the equipment manufacturing industry. These four categories are investment, government purchase, macroeconomic policies and the situation of equipment manufacturing industry. According to the association of online data and the equipment manufacturing industry, representative searching words are selected, and the number of searching words for each type of information is determined. The total number of the selected searching words is 40, and the searching words profiles are shown in Table 2.

The monthly data of this study are from Jan. 2012 to Dec.2016, a total of 60 periods of data; of these, the training set data are from Jan. 2012 to Dec.2015, a total of 48 periods of data. The forecast set data are from Jan. 2016 to Dec.2016, a total of 12 periods of data.

TABLE II. INTERNET SEARCHING WORDS INFORMATION

Variable	Measurement	Index Containing
I	Investment	I_1～I_{10}
G	Government purchase	G_1～G_{10}
M	Macroeconomic policies	M_1～M_{10}
E	Situation of equipment manufacturing industry	E_1～E_{10}

Note: Baidu search index variables are using the corresponding Chinese words to search.

III. MODEL CONSTRUCTION

In equipment manufacturing industry evaluation, two different types of information can be applied, namely government statistical data and online data. The advantages of the former are low noise and standardization; however, the disadvantage is that the data has a certain time lag. The advantages of the latter are the fast information update and real-time availability; however, the disadvantage is the large information noise.

The model construction formulas are expressed as follows. Set y as the explained variable, $Y_t^T = \{y_t\}_{t=t}^T$ is a time series of explained variables and $t(t=1,2,...,T)$ on behalf of the period; this study uses a monthly time series in value. In general, the information of $\{y_t\}_{t=1}^{T-1}$ is available, T is the period that we want to forecast; that is, we need to predict y_T based on existing information. y_{T-i} is the explained variable lag i period, ε_T is residual. $y_{T|T-1}$ means we will forecast y of T period based on the $(T-1)$ period information.

Explanatory variables are divided into two categories, the government statistical indicators and online data. X represents the government statistical indicators. $X_t = (x_1, x_2, ..., x_{mt'})'$ is an $(m \times 1)$ vector; each component represents a government statistical indicator for t period, and the number of government statistical indicators is m. $X^T = \{X_t\}_{t=1}^T$ is the time sequence of government statistical indicators. In general, the information of $\{X_t\}_{t=1}^{T-1}$ is available at period T. Similarly, Z are indicators measuring online data, $Z_t = (z_1, z_{2t}, ..., z_{nt'})'$ is an $(n \times 1)$ vector, each component represents one online data for t period, and the number of online data is n. $Z^T = \{Z_t\}_{t=1}^T$ is the time sequence of online data because its information is available real-time; the information of $Z^T = \{Z_t\}_{t=1}^T$ is available at period T. In practice, we can use the limited information for prediction, for example, for the government statistical indicators, we can use $(T-p) \sim (T-1)$ period information; for the online data, we can use $(T-p) \sim T$ period information. p and q are determined through model selection. The model is as follows:

$$\hat{y}_{T|T} == c + \beta_1 y_{T-1} + ... + \beta_i y_{T-i} + \alpha_1 X_{T-1} + ... + \alpha_p X_{T-p} + +\gamma_1 Z_T + ... + \gamma_q Z_{T-q+1} + \varepsilon_T \tag{1}$$

For mode 1, place the explained variable's information, government statistical indicators and online data in the model together; a_t is the coefficient row vector of $\{1 \times m\}$ (the meaning of a_t is the same in the following content). Here, we expect to find an optimal subset of independent variables; for this set, its subscript set is $\{(j_1,t_1),(j_2,t_2),...,(j_r,tr),\} \subseteq \{1,...,m\} \times \{T-p,...,T-1\}$. In addition, we expect to find the optimum lag combination of explained variable y to ensure that y_t can be well-fitting and predictive by using the multivariate linear model with the independent variables' optimal subset and the explanatory variables' optimum lag combination. γ_t is the coefficient row vector of $\{1 \times n\}$ (the meaning of γ_t is the same in the following content). Here, we expect to find an optimal subset of $\{(j_1,t_1),(j_2,t_2),...,(j_r,tr),\} \subseteq \{1,...,n\} \times \{T-q+1,...,T\}$ independent variables; for this set, its subscript set is. Also we expect to find the optimum lag combination of explained variable y to ensure that y_t can be well-fitting and predictive by using the multivariate linear model with the independent variables' optimal subset and the explanatory variables' optimum lag combination.

Because of the limited number of samples, the number of selected variables should not be excessively large, or else the selection of an excessive quantity of variables may lead to a training set fits well but performs poorly in the forecast set. It is assumed that the best forecast model is produced by variables whose lags are not more than 4 periods, that is, the longest lag period is 4.

In this model, explained variable's information, government statistical indicators and online data are put into the model evenly at the same time, and the variables are selected. The automatic model selection module in OxMetrics software can realize the evaluation.

IV. Empirical Results

This section evaluates the situation of equipment manufacturing industry. We put the explained variable's information, the government statistical and online data together into the model, in order to determine a final optimal model to evaluate the equipment manufacturing industry situation in the future. In this section, we apply OxMetrics, one of the most popular software for variables selection and model decision.

A. Forecast on Metal Products Industry

Here we put Y with lag 1-4, government statistical data with lag 1-4 and online data with lag 1-4 equally into the model, selecting the automatic model selection procedure in OxMetrics, the results are presented in Table 3.

From Table 3 we can see, based on the regression and forecast results of the model, we put altogether249 variables equally into the model, finally 10 variables are left, and the forecast figure is as follows. From the figure we can know that for the forecast set of 12 periods, the forecast result is so good.

TABLE III. EMPIRICAL RESULTS OF METAL PRODUCTS INDUSTRY DEVELOPMENT FORECASTING

variables	**Coefficient**	**Std.Error**	**t-value**	**t-prob**
X6_2	-0.284764	0.1376	-2.07	0.0462
X7	-0.394185	0.1160	-3.40	0.0017
X7_1	0.817170	0.1949	4.19	0.0002
X7_3	-0.758883	0.1899	-4.00	0.0003
X7_4	-0.0207867	0.006442	-3.23	0.0028
X8	0.400538	0.09436	4.24	0.0002
X8_1	-0.661447	0.1586	-4.17	0.0002
X8_3	0.670765	0.1683	3.99	0.0003
M3_2	-2.31909	0.6842	-3.39	0.0018
G3_3	0.162038	0.07711	2.10	0.0431

Note: variable mantissa 1, 2,..., indicates lag 1 period, 2 period,..., respectively; variable mantissa 0 indicates the current period.

From the above table, we can see the result:

(1) In the prediction of the development of the metal products industry, only 10 variables were retained in the 249 variables of the model.

(2) The value of the export value of the metal products, that is the explanatory variable, the lagged 1-4 stage of it has not been retained, indicating that the lagged variables of the explanatory variables do not play a significant role in the prediction of the development of the metal products industry.

(3) Of the 45 government statistical variables, 8 variables were retained. Respectively, X6 lag 2, X7 and its lag phase 1, 3 and 4, X8 and its lag phase 1 and 3. This shows that in the 9 indicators involving the development of equipment manufacturing industry, only three indicators have a role in forecasting; the current and lagging indicators of the period one and three play the biggest role in forecasting.

(4) In the 200 Baidu search index variables, 2 variables were retained. Of these, 1 of the Baidu search index variable involved in macroeconomic policies was retained; 1 variable related to government purchases related to Baidu search index was retained.

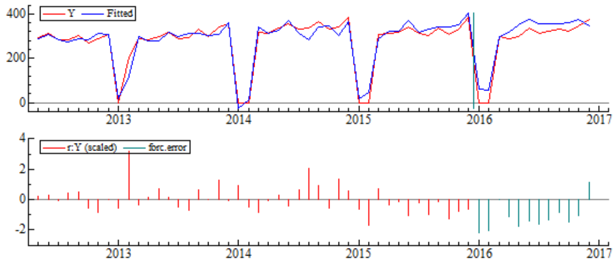

FIGURE I. MODEL FITTING RESULT

Next, the 10 explanatory variables retained in the model filter are put into the prediction model, and the results of model fitting are shown in figure 1. As can be seen from the diagram, the model constructed by the 10 explanatory variables retained in the model selection works well with the values of the variables being interpreted. These 10 variables are sufficient to predict the explanatory variables. However, because of the lack of statistical data in January each year, the graph is incoherent.

The 10 explanatory variables retained in the model selection are put into the prediction model, and the prediction of the development of the metal products industry of the next 12 months is predicted. The forecast results are shown in Figure2.

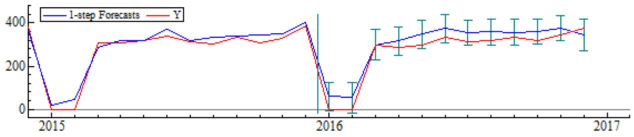

FIGURE II. DEVELOPMENT FORECAST OF METAL PRODUCTS INDUSTRY

B. Forecast on Automobile manufacturing industry

Here we put Y with lag 1-4, government statistical data with lag 1-4 and online data with lag 1-4 equally into the model, selecting the automatic model selection procedure in OxMetrics, the results are presented in Table 4.

From the above table, we can see the result:

(1) In the prediction of the development of the Automobile manufacturing industry, only 11 variables were retained in the 249 variables of the model.

(2) The value of the export value of the metal products, that is the explanatory variable, the lagged 1-4 stage of it has not been retained, indicating that the lagged variables of the explanatory variables do not play a significant role in the prediction of the development of the metal products industry.

TABLE IV. EMPIRICAL RESULTS OF AUTOMOBILE MANUFACTURING INDUSTRY DEVELOPMENT FORECASTING

variables	Coefficient	Std.Error	t-value	t-prob
X8	0.0543482	0.002210	24.6	0.0000
I5	0.0903182	0.01533	5.89	0.0000
M2_2	0.648348	0.07260	8.93	0.0000
M4	-0.246430	0.03019	-8.16	0.0000
M5_1	-0.231406	0.08012	-2.89	0.0068
M5_3	0.567630	0.09273	6.12	0.0000
M8_1	0.0903523	0.01527	5.92	0.0000
M8_2	-0.0756260	0.01419	-5.33	0.0000
G2_3	-0.329330	0.06830	-4.82	0.0000
G8_3	-0.00428508	0.001426	-3.00	0.0050
G9_4	0.270621	0.06219	4.35	0.0001

(3) Of the 45 government statistical variables, only 1 variable was retained, that is X8. This shows that only 1 of the 9 indicators involved in the development of the automotive industry are predictive.

(4) In the 200 Baidu search index variables, 10 variables were retained. Among them, involving investment related Baidu search index variable 1 is reserved; related to Baidu macroeconomic policy related search index variable 6 is reserved; related to the government to buy related Baidu search index variable 3 is reserved. Next, the 11 explanatory variables retained in the model filter are put into the prediction model, and the results of model fitting are shown in figure 2. As can be seen from the diagram, the model constructed by the 11 explanatory variables retained in the model selection works well with the values of the variables being interpreted. These 11 variables are sufficient to predict the explanatory variables. However, because of the lack of statistical data in January each year, the graph is incoherent.

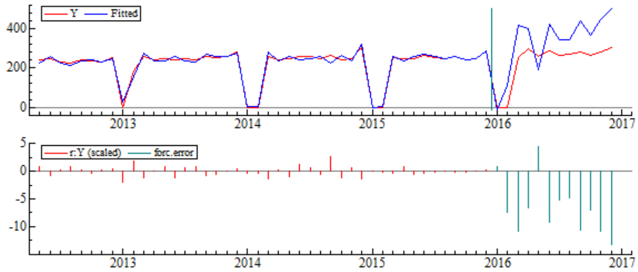

FIGURE III. MODEL FITTING RESULT

The 11 explanatory variables retained in the model selection are put into the prediction model, and the prediction of the development of the metal products industry of the next 12 months is predicted. The forecast results are shown in Figure 4.

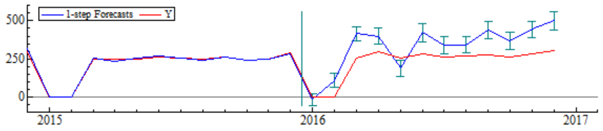

FIGURE IV. FDEVELOPMENT FORECAST OF AUTOMOBILE MANUFACTURING INDUSTRY

V. CONCLUSION

In the contemporary era of popular “Big Data”, this study uses traditional statistical data and online data to forecast the future development of the equipment manufacturing industry. In this study, the final evaluation model of the development of equipment manufacturing industry is established by putting the government statistical data and the online data into the model to screen variables.

This study also shows that traditional government statistical data along with the online data can forecast the development of equipment manufacturing industry well, but the selected variables in the final evaluation model are the current and lag 1 period variables, which means that the current and lag 1 period variables can give a good evaluation to the situation of equipment manufacturing in the future.

ACKNOWLEDGEMENTS

This study is sponsored by Liaoning BaiQianWan Talents Program (【2015】70), Liaoning Scientific Research Staring Foundation for Doctor (201601276) and Dalian High Level Talent Program for Innovation, Entrepreneurship (2015R086). This research project is supported by Science Foundation of Beijing Language and Culture University (supported by “the Fundamental Research Funds for the Central Universities”) (17PT04).

REFERENCES

[1] Hendry, D. F., and Hubrich, K., 2011, “Combining disaggregate forecasts or combining disaggregate information to forecast an aggregate”, Journal of Business and Economic Statistics, 29, 216-227.

[2] Hyunyoung Choi, Hal Varian, 2009a, “Predicting the present with Google Trends”. Technical report, Google,

[3] Jennifer L. Castle, David F. Hendry & Oleg I. Kitov, 2013,”Forecasting and Nowcasting Macroeconomic Variables:A Methodological Overview”, Economics Series Working Papers 674, University of Oxford, Department of Economics.

[4] Nii Ayi Armah, 2013, “Big Data Analysis: The Next Frontier”, Bank of Canada Review (Summer), 32-39.