A Combinatorial Forecasting Model Research Based on GMDP for Small Sample Data of ICT Service Consumption

Xu Han

College of Economics & Management, China Agriculture University, Beijing, 100083, China

hx@cau.edu.cn

Abstract

Combinatorial forecasting model used for small sample data prediction was discussed in this paper, we use SPSS curve estimation to pre-estimate the curve suitable to the sample data, and found out that Cubic Curve is the most accurate model to match the small sample in this research. We used GMDH auto-regressive model to predict in steps. Finally, we used combinatorial GMDH forecasting model containing both autoregressive GMDH and Cubic Curve model to forecast data and trend based on the small sample data. The results show that the combinatorial GMDH forecasting model is better and more stable than Cubic Curve model and autoregressive GMDH model.

Keywords

GMDH, forecasting, small sample data, ICT, consumption

1. Introduction

Modern macroeconomic forecasting models such as regression forecasting model, Markov prediction model, grey system prediction model, are usually based on the analysis of the previously data. Through finding out the internal rules and dependencies, they can result in the ability to predict the unknown data. The more sample data, the more accurate are the prediction. However, when the sample data set is small, the accuracy of prediction method above is not high yet.

Regression forecasting model need lots of historic data in prediction, the modeling of multivariate nonlinear regression forecasting is complex and difficult [1]. The Markov model can work in small amount data sample, but it has low accuracy and high storage complexity. Grey system prediction model suite for small sample prediction, but the modeling process is difficult and complex.

Due to the drawbacks of the above-described models, high accuracy forecasting methods and models for small sample data set are needed in reality. In this research, we will explore the utility of SPSS Curve estimation model, autoregressive GMDH model in small sample data forecasting. At the end, we will combine the two models into a combinatorial GMDH model, and compare the accuracy among the three models.

2. Organization of the Text

2.1 SPSS Curve Estimation Model

Relationships between variables include linear as well as non-linear. Relationship of linear is easier to find out, but no-linear relationship is difficult to figure out. We have different methods in processing variables of linear and non-linear.

Variables in linear can be modeled by linear regression. For non-linear, we cannot analyze the variables by linear regression and linear model. In order to judge the relationship of variables properly, we can use SPSS to estimate the curve model through sample data fitting, for example, quadratic, cubic, exponential, inverse, power, compound, logarithmic, growth, logistic, and etc.

By fitting the sample data, the parameter, such as f value of the regression equation significance test, p value of probability, and R-squared coefficient of determination, can be calculated in different curve model mentioned before. We can compare the parameter in models and choose the optimum

model mainly based on R-squared coefficient of determination. The optimum model will be used in the follow-up combination forecasting.

2.2 GMDH Method

Group Method of Data Handling (GMDH) is a self-organize modeling technology raised by A.G Ivakhnenko in 1968[2]. Self-organization is a method for adaptive synthesis of complex system, proceeding from the existed sample data, and generates candidate models by calculating. With certain external criteria, an optimal complexity model can be determined [3].

GMDH determines the coefficients and parameters directly from data sampling through an iterative and inductive sorting process. In a nutshell, GMDH is a controlled optimization process between the inputs and the outputs which gradually generates more complicated models and retains those candidate models that have relatively low forecasting errors [4].

Compared with the traditional neural network algorithm, GMDH algorithm has the following characteristics [5]:

Firstly, Analytic functions can be clearly expressed in the model results. In many cases, the modeler hopes that through the model structure reveals the interaction and dependencies between variables. However, the traditional BP network model is difficult to give physical meaning. It could not answer the questions like why and how kinds of problem, which limit the neural network analyzes the factors in the system. GMDH, combination of neural networks and statistical modeling of thought, can give the results of functional expression, or even other modeling methods are difficult to reach higher multivariate regression equation

Secondly, the GMDH modeling is a self-organized control process without any initial assumptions. Statistical models and neural network modeling process generally, often require the modeler to model input variables based on experience and assumptions, and through trial and error to find satisfaction model. GMDH algorithm allows hundreds of input variables, and then to drill a large number of variables to produce a large number of candidate models, based on data-driven algorithms to find have a substantial impact on the dependent variable entries, self-organizing generate optimal network architecture to minimize the build subjective factors influence mold.

Thirdly, GMDH algorithm has optimal complexity and precision forecast. In small sample data set or noisy data set, generally the neural network will produce excessive noise fitting and lower generalization capabilities. However, GMDH network ensures optimal complexity it from similar, uncertainty and even mutual contradictory in knowledge decisions environment, while avoiding over-fitting and fit the model structure insufficiently. The model is closer to the real situation, and has high forecast reliability.

The observation sample in GMDH is divided to at least two sub samples: training sample and testing sample [6]. The selection models are built based on the inner criterion of training sample, and the external criterion characteristic of testing sample determines the choose rules in selection models. When the external criterion reaches its minimum, the corresponding model is the optimal complex model.

GMDH algorithms give possibility to find automatically interrelationship in data, generate optimal model and increase the accuracy of existing algorithms. After 30 years development, GMDH has become a powerful forecasting modeling method for small sample data and was used widely in data mining, knowledge discovery, and prediction.

The GMDH network is shown in Fig.1. The y is the predicted value of GMDH network, which is also the output variable of the network. The x_{il} is the i^{th} input variable in the l^{th} sample, i=1, 2, 3…, n, n is the number of independent variables. y_{jkl} is the predicted value of k-th neuron, which is in the layer j of the l-th sample, k=1, 2, … ,m. r2jk is the RMS threshold collection of kth neuron in layer j. R_j is the max number of neuron in layer j.

Suppose an abject investigated with GMDH is represented by multiple inputs and at least on output. The object can be modeled by a certain subset of components of the base function (1):

(1)

In the upper function, (x_1, x_2, …, x_M) is input variable, (a_1, a_2, …, a_M) is coefficient matrix, y is the output variable. Generally, GMDH network uses multi-layer neurons iterative algorithm to choose modeling processes, gets the nonlinear mapping between input and output through learning, and finally selects the optimal model with the smallest deviation criteria.

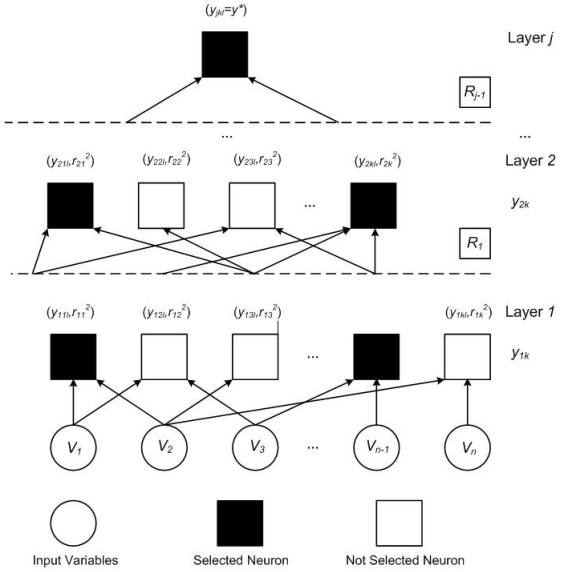

Fig. 1 GMDH Network

The GMDH algorithm makes the following steps:

Data collection. Divides data sample ω(number of sample data is N) onto training sample set A and testing sample set B, ω=A∪B. When building forecasting model, divides data sample onto learning sample set A, testing sample set B and forecasting sample set C, ω=A∪B∪C. Setting the original number of input variables d_0 and the max number of neuron R_j in every layer.

Select a external criteria as the target function. Generaly, GMDH set the smallest deviation criteria as the neuron selection criteria, Y is the output value of dependentive variable, N is the number of samples.

(2)

Initialize the GMDH network. Establishes the common relationship between inputs and output like (1), generates d_0 neurons in layer 1, and initialize the original network.

Compute and check r^2_{jk}. Firstly, sort the r^2_{jk} in ascending order, select the former Rj neurons corresponding r^2_{jk} and retain them, switch to step 2 and delete the rest of neruons. Secondly, find the minimal r^2_{jk} of retaining neurons in layer j, and compare r^2_{jk} with $r^2_{j-1,k}$, which is the minimal in layer $j-1$, if $r^2_{jk}<r^2_{j-1,k}$, switch to step 5, otherwise, switch to step 6.

Generate next layer neurons, $j=j+1$.Keep the neurons generated from step 4, and generate new neurons of next layer, then switch step 4.

Complete the GMDH training. When the minimal r^2_{jk} in the layer j biger than the minimal $r^2_{j-1,k}$ in the layer $j-1$, it is considered that the optimal parameter of the kth neuron has been found in layer $j-1$, the training in layer j is over.

2.3 Combinatorial Forecasting Method

Combinatorial forecasting refers to compose several forecasting methods properly in one forecasting task, and enhance the degree of accuracy as far as possible based on the parameters getting from the combinatorial forecasting methods.

Weight coefficient combinatorial forecasting method is the linear combination of single forecasting method. However, the single forecasting method in combination is nonlinear. GMDH combinatorial forecasting model based on weight coefficient can avoid over-fitting in prediction and enhance the accuracy degree, which is choose to be the main forecasting method in this research.

3. Experiment result and discussion

The small sample data in experiment is total amount of the telecommunication service (one hundred million Yuan) in China from 1994 to 2010, which contains 17 annual data and is getting from the China Statistical Yearbook. The annual data from 1994 to 2008 is used to generate the forecasting model, and the annual data of 2009, 2010 is used to test the accuracy of forecasting.

3.1 SPSS Curve Estimation Forecasting

Curve Estimation in SPSS software can figure out the certain style curves based on the imported annual data from 1994 to 2008, such as quadratic, cubic, exponential, inverse, power, compound, logarithmic, growth, logistic, and etc.

The principle of selecting optimal curve is mainly based on the R square value of the curve, when the R square value is close to 1, the bigger value of F in testing is better when the sig value less than 0.05.

Comparing the parameter value in table I, we choose cubic curve shown in Fig. 2 as the optimal curve.

Table 1 parameter of SPSS Curve Estimation

	Model summary					Parameter Estimation Value			
	R2	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	0.851	74.14	1	13	0.000	-4128.6	1399.63		
Logarithmic	0.589	18.62	1	13	0.001	-5317.3	6659.19		
Inverse	0.288	5.26	1	13	0.039	10344.4	-14808.9		
Quadratic	0.984	370.90	2	12	0.000	2413.60	-909.37	144.3	
Cubic	0.997	1044.03	3	11	0.000	-480.61	965.72	-139.4	11.82
Compound	0.987	975.67	1	13	0.000	582.56	1.28		
Power	0.925	160.40	1	13	0.000	328.66	1.38		
S	0.637	22.81	1	13	0.000	9.17	-3.64		
Growth	0.987	975.67	1	13	0.000	6.37	0.25		
Exponential	0.987	975.67	1	13	0.000	582.56	0.25		
Logistic	0.987	975.67	1	13	0.000	.002	0.78		

The model of cubic curve is:

(3)

The value of annual 2009 and 2010 forecasted by cubic curve are 27703 and 33726.

3.2 Autoregressive GMDH Forecasting

Based on the principles of autoregressive GMDH forecasting model, we use Knowledge Miner software to forecast the annual value in 2009 and 2010.

The selection of optimal model depends on parameters value of R square, mean absolute percentage error (MAPE), and prediction error sum of squares (PESS) on the forecasting model. The optimal

principles include R square value closed to 1, the smaller value of MAPE and PESS is better, the MAPE value less than 0.05.

The model of autoregressive GMDH is:

(4)

The R square value is 0.993, MAPE value is 0.025, PESS value is 0.14%.

The value of annual 2009 and 2010 forecasted by Autoregressive GMDH are 25972 and 28986.

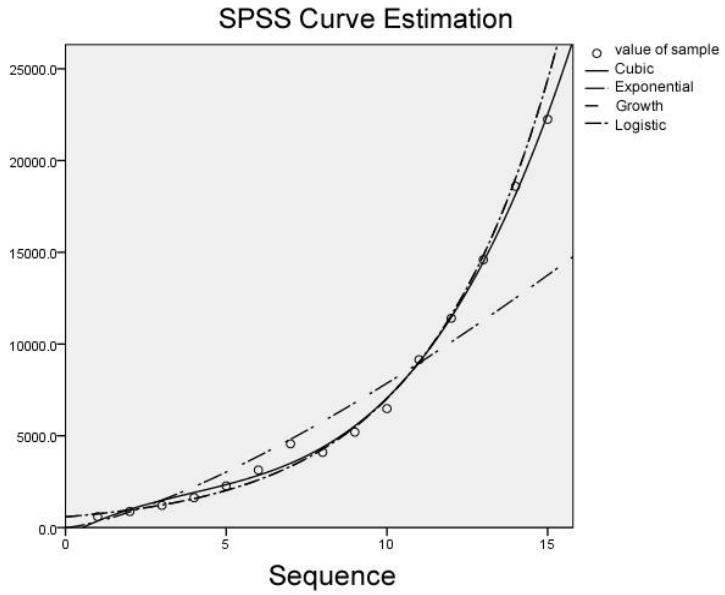

Fig. 1 SPSS Curve Estimation

3.3 Combinatorial GMDH Forecasting

Set the same parameter, we combine autoregressive GMDH forecasting model and cubic curve forecasting model into combinatorial GMDH forecasting model, which is shown in equation (5).

(5)

The R square value is 0.993, MAPE value is 0.014, PESS value is 0.03%.

The value of annual 2009 and 2010 forecasted by combinatorial GMDH are 25942 and 29634.

Table 2 The results of forecasting model

Year	Real Value	Cubic Curve		Auto-regression GMDH		Combinatorial Forecasting	
		Predicted Value	Relative Error	Predicted Value	Relative Error	Predicted Value	Relative Error
2009	25554	27703	8.41%	25972	1.64%	25942	1.52%
2010	29993	33726	12.45%	28986	-3.36%	29634	-1.20%
PESS		2.25%		0.14%		0.03%	
MAPE		0.104		0.025		0.014	

3.4 Discussion

The results of forecasting model in table II show the accuracy of prediction in three models. Though there is an optimal fitting in every model, the overall error is high because the data number of sample set is small. The predicted results in combinatorial forecasting model are better than the values in the other two models. It shows the advantage of combinatorial GMDH in enhancing the accuracy of prediction in small data sample.

4. Conclusion

We use cubic curve model, autoregressive GMDH model and combinatorial GMDH forecasting model to predict the value in small data sample.

The results show that the combinatorial GMDH forecasting model is better than autoregressive GMDH model in this research. The curve forecasting methods in SPSS are not suitable in small data sample forecasting.

For different learning sample set, the prediction results is different with GMDH forecasting method, so it is important to select proper learning sample set to ensure we can get optimal forecasting results. The result of combinatorial GMDH forecasting model depends on the accuracy of models which make up the combinatorial GMDH forecasting model.

References

[1] Shi guang-ren, et al, "Application of artificial neural network and multiple regression analysis to optimization of exploration prospects," Acta Petrolei Sinica, Vol.23, No.5, 2002, pp.19-23.

[2] Ivakhnenko, A. G, "Polynomial Theory of Complex Systems," IEEE transactions on Systems, Man, and Cybernetics, Vol. SMC-1, No.4, 1971, pp.364-378.

[3] Ivakhnenko, A.G, "Heuristic self-organization in problems of engineering cybernetics", Automatica. Vol.6, 1970, pp.207-219.

[4] Simon Fong, Zhou Nannan, Raymond K. Wong, Xin-she Yang, "Rare Evens Forecasting Using a Residual-Feedback GMDH Neual Network," Proceeding of 2012 IEEE 12th International Conference on Data Mining Workshops. 2012, pp.464-473, December 2012.

[5] Liu Guangzhong, Yan Keqi and Kang Yinlao, "GMDH-type Neural Network Algorithm and its Application," Mathematics in Practice and Theory, Vol.31, No.4, 2001, pp.464-469.

[6] Nan Li, Shuyong Liu and Xiangwei Mu, "Trade Surplus Analysis Using Self-organizing Data Mining Based on GMDH Principle," Proceeding of 2009 World Congress on Computer Science and Information Engineering. 2009, pp.28–32.