A time prediction model for residents consumption level based on ARIMA and PCA

Zhongyu Su*, Wei* Li*, Yu* Sun *and Pengcheng* Guo

School of Information Engineering, Dalian Ocean University, 116023 Dalian, China

Keywords: Resident consumption level, Principal component analysis, Data dimension reduction, ARIMA model.

Abstract. It is necessary but difficult to make a large number of observations on multiple variables reflecting the residents consumption level and collect a large amount of data for analysis to search for the rules. In this paper, a time prediction model for residents consumption level based on ARIMA and principal component analysis is proposed to solve this problem. Principal component analysis is firstly used to effectively reduce the number of indicators reflecting the residents consumption level. Combined with the ARIMA model, the residents consumption level is predicted. The results reflect the trend of residents consumption level towards the need for enjoyment and development materials on the basis of obtaining basic survival data.

1 Introduction

In the era of big data, researchers are faced with the huge challenges in processing massive data and unstructured data [1-2]. It is necessary but difficult to make a large number of observations on multiple variables reflecting the residents consumption level and collect a large amount of data for analysis to search for the rules. Principal Component Analysis (PCA), as an effective multivariate statistical method, reduces the problems in high-dimensional space to those in low-dimensional space, making the problem simple and intuitive. The fewer composite indicators obtained by PCA are irrelevant and provide most of the information on the original indicators [3-4]. In this paper, the PCA method is used to effectively reduce the 12 indicators data reflecting the residents consumption level from 1952 to 2010 to two main components. Combined with the ARIMA model, we construct a time prediction model to predict the residents consumption level from 2011 to 2020 and observe the changes between the residents consumption levels. The results reflect the trend of residents consumption level towards the need for enjoyment and development materials on the basis of obtaining basic survival data [5].

The level of residents consumption refers to the extent to which residents meet the needs of people's survival, development and enjoyment in the process of consumption of physical products and services. It is reflected by the quantity and quality of the physical goods

* Corresponding author: 18340897429@163.com

consumption and labour services consumption. Residents consumption level is calculated based on the gross domestic product which is the total consumption including labour consumption. Due to the limitations of the basic data, the data from 1952 to 2010 is selected for analysis in this paper.

This paper consists of the following four parts. Section 1 is the introduction including data background and research significance. The second part is the theoretical introduction of the time prediction model, including PCA and ARIMA model. In section 3, a time prediction model for residents consumption level using ARIMA model based on feature extraction is established. Section 4 is the summary.

2 Method

This paper presents an efficient combination analysis method. Firstly, PCA is applied to effectively reduce the multiple indicators reflecting the residents consumption level. Based on the PCA, combined with the ARIMA model, the changes of residents consumption levels are predicted.

2.1 PCA

PCA converts multiple indicators into several comprehensive indexes by using the idea of dimension reduction and minimizing the loss of information [6-7]. Usually called main component, the comprehensive indicators of the transformation to generate each principal component are a linear combination of the original variables and unrelated and have better properties than the original variables. So only using a few principal components without loss of too much information, can solve the complex high-dimensional problems [8-9].

The algorithm steps are as follows:

(1) select initial analysis variables according to the research questions;

(2) according to the characteristics of initial variables, determine whether the principal component is determined by covariance matrix or correlation matrix;

(3) find the eigenvalues and eigenvectors of covariance matrix or correlation matrix;

(4) get the expression of principal components and determine the number of principal components, and select the principal components;

(5) analyze and deeply study the data with the combination of principal components.

2.2 ARIMA model

ARIMA model, the full name is auto regressive moving average model, which is a time series prediction method. It is a model for fitting stationary series, which is convenient for analyzing the structure and inherent properties of data, and also convenient for optimal prediction and control in the sense of minimum variance [10-11].

The algorithm steps are as follows:

(1) Perform first-order difference and seasonal difference on the data, convert unstable data into stable data, and determine d.

(2) The values of p and q are determined by the ACF and PACF maps after the difference.

(3) The goodness of fit is compared using the ARIMA model, the AIC model and the BIC model. Finally, p, d, q are determined, substituted into the formula, and the model is established [12].

3 Result

First, PCA is applied to reduce the dimensions of residents consumption data. Based on the PCA, the ARIMA model is established.

3.1 PCA

This part is the data on changes in residents consumption levels. The original data is from "China Statistical Summary 2011" [13]. The data on changes in residents consumption levels are shown in Table 1.

Table 1. Residents consumption level data.

year	Absolute number (yuan)			...	Index (1978=100)		
	National	rural	town	...	National	rural	town
1952	80	65	154	...	55.9	63.5	46
1953	91	72	188	...	60.1	65.2	52.9
...	...	...	...	...	...	...	...
2009	9098	4021	15025	...	1001.6	616.8	712.2
2010	9963			...	1062.1		

Table 2. KMO and bartlett tests.

KMO sampling adequacy		.818
Bartlett test of sphericity	The approximate chi-square	2761.227
	df	66
	Sig	.000

3.1.1 Statistical test

The results of the KMO (Kaiser-Meyer-Olkin) and bartlett tests are obtained using SPSS in Table 2. It can be seen from table 2 that KMO statistic =0.818 > 0.7, chi-square statistic of spherical test =2761.227, P=0.000 < 0.01, suitable for PCA. (note: KMO is a ratio of correlation coefficient and partial correlation coefficient. It is used to compare a statistic of simple correlation coefficient and partial correlation coefficient between variables. The closer its value is to 1, the more suitable it is for PCA. Bartley sphericity test is a statistic to determine whether the correlation coefficient matrix of variables is a unit matrix. Its associated probability is less than significance (0.05 or 0.1), indicating that the correlation coefficient matrix is not an unit matrix and suitable for PCA [14].)

3.1.2 Extraction factor

The main component analysis method is to solve the factor solution. The basis for determining the number of factors is that the cumulative contribution rate of variance exceeds 85%, as shown in table 3.

Table 3. Total variance interpretation.

composition	Initial eigenvalue			Extract the sum of squares of the load.		
	aggregate	Percentage variance	Cumulative percentage	aggregate	Percentage variance	Cumulative percentage
1	9.271	77.259	77.259	9.271	77.259	77.259
2	1.539	12.822	90.081	1.539	12.822	90.081
...	...	...	...			
12	2.657E-5	.000	100.000			

3.1.3 Screen plot

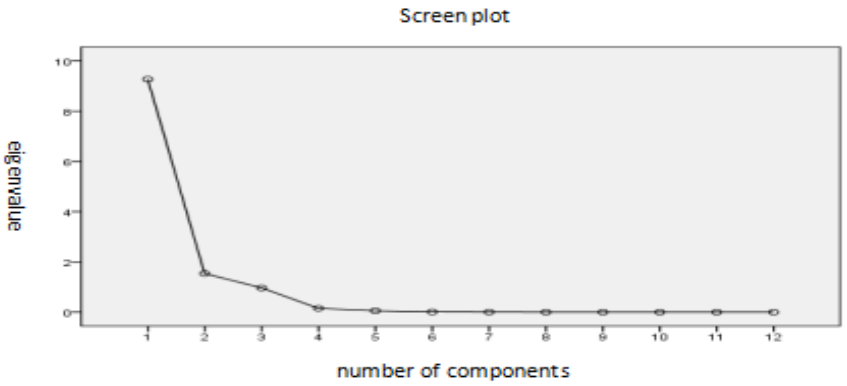


Fig.1. Principal component lithotripsy.

According to figure 1, till the third principal component eigenvalue shows a steep inflection point, the first principal component eigenvalue is significantly different from the rest, and the rest are basically equal. As can be seen from table 3, the first two principal component eigenvalues account for 90.081% of all information, accounting for the majority of all information. So we obtain two principal components. In this way, only two principal components can be used to express the majority of the original information that all indicators can express [15].

3.1.4 Calculate the principal component scores of each variable

The score of principal components is obtained by multiplying the number of factors by the arithmetic square root of their respective variances, as shown in table 4.

Table 4. The main component of residents consumption level data.

year	factor1	Com1	factor2	Com 2	year	factor1	Com1	factor2	Com2
1952					...	...	...		
1953	-0.72630	-2.21	0.43508	0.54	2008	2.73120	8.32	-0.20335	-0.25
1954	-0.82628	-2.52	-0.54046	-0.67	2009	2.96550	9.03	-0.63038	-0.78
1955	-0.67941	-2.07	1.03449	1.28	2010	2.35810	7.18	-0.01450	-0.02

3.2 ARIMA model

ARIMA model (Autoregressive Integrated Moving Average model), the difference autoregressive integrated moving average model, is one of time series prediction method. By

means of PCA, residents consumption level data is effectively reduced to two principal components. ARIMA model is established to observe and predict changes in residents consumption level [16-17]. First, the ARIMA model of principal component I is established and analyzed.

3.2.1 Stationary sequence

The original sequence is processed by difference. The original sequence is shown in figure 2. The stationary sequence is obtained after 3 orders of difference, as shown in figure 3.

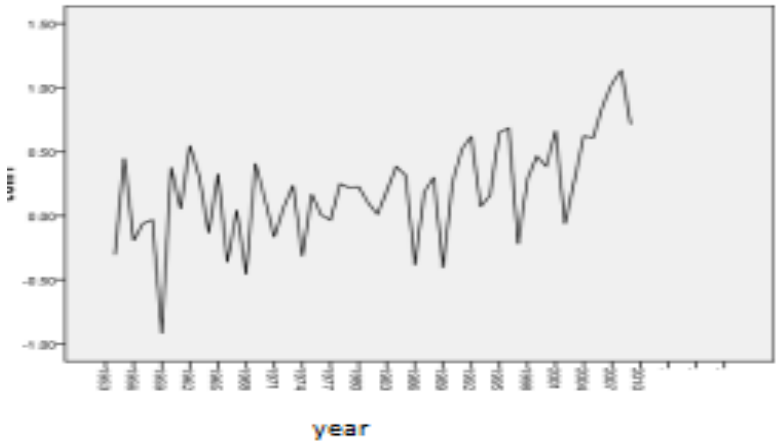


Fig 2. The original sequence.

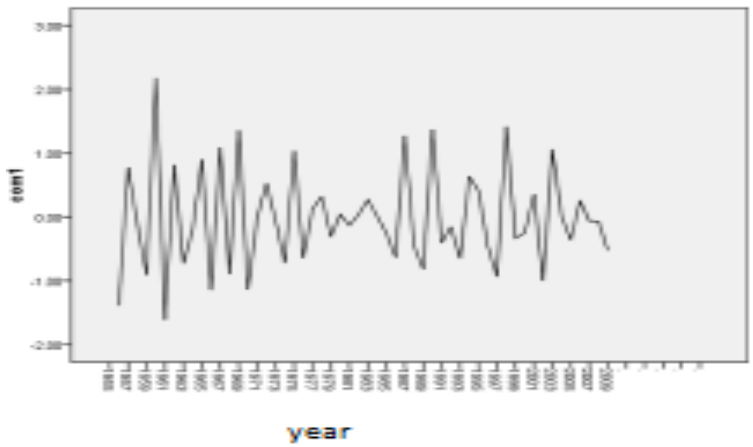
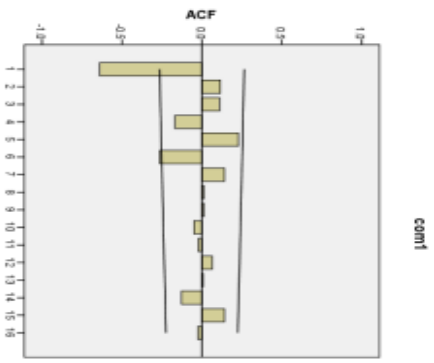


Fig. 3. Stationary series.

3.2.2 Autocorrelation and partial autocorrelation function graphs

After repeated confirmation, the adjusted sequence autocorrelation function graph and partial correlation function graph confirm that the model is ARIMA (1,3,4), as shown in figure 4 and 5.



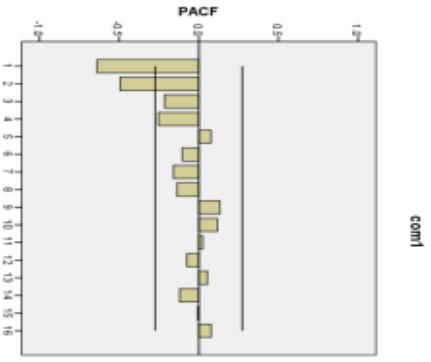


Fig. 4. Graph of autocorrelation function. **Fig 5.** Graph of partial autocorrelation function.

Therefore, the results of ARIMA model are as follows:

$$\Delta X_t = -0.331 - 0.875\Delta X_{t-1} + \varepsilon_t + 1.075\varepsilon_{t-1} + 0.852\varepsilon_{t-2} - 1.076\varepsilon_{t-3} + 0.142\varepsilon_{t-4}$$

3.2.3 Model curve

According to the time series data of residents consumption level in several years, the ARIMA model is established to predict the change trend of residents consumption level from 2010 to 2020. The curve is shown in figure 6. ARIMA model is established for principal component two. The curve is shown in figure 7.

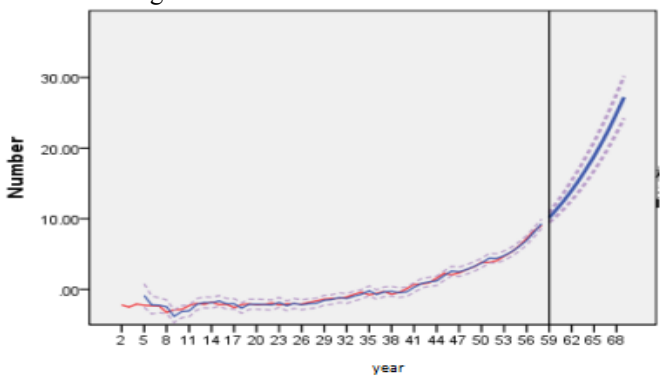


Fig. 6. Forecast trend curve.

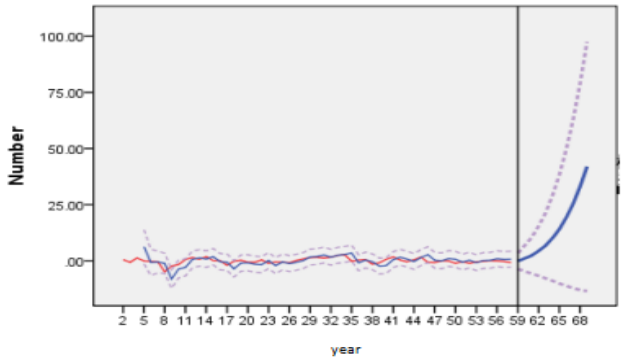


Fig. 7. Forecast trend curve.

ARIMA models are established for principal components 1 and 2 respectively, as shown in figure 6 and figure 7. The horizontal axis 1-68 represented 1952 to 2020, respectively, and the vertical axis represented each value of extracted principal components. In the figure, the red curve represents the observed value. The blue thin line represents the predicted value. The two dotted lines LCL and UCL represent the control down line and control upper line respectively. The blue bold line represents the predicted value. From the trend of the curve, we can see that residents consumption level shows an upward trend and the upward trend is relatively obvious.

4 Summary

In this paper, a time prediction model is constructed to analyze the change of residents consumption level. Since there are many indicators reflecting the change of residents consumption level, it is impossible to establish ARIMA model to predict the change of residents consumption level in the next few years. Therefore, PCA is firstly used to effectively reduce the data into two main components. Based on two principal components, ARIMA model is established to predict the change of residents consumption level. According to the predicted change curve, it is found that residents consumption level shows a relatively obvious upward trend. With the deepening of China's economic restructuring and the gradual establishment of socialist market economy, residents consumption level has been gradually improved.

This research was financially supported by the Scientific Research Foundation of Liaoning Province of China (JL201919) and Natural Science Foundation of Liaoning Province of China (20170540103).

References

1. D Wang, H Shen Y Truong. Efficient dimension reduction for high-dimensional matrix-valued data [J]. Neurocomputing, 2016, 190: 25-34.
2. Q Pan. Research on dimensionality reduction method in high dimensional longitudinal data analysis [J]. The financial times, 2017, 672(9): 19-20.
3. W Hou. An improved method for comprehensive evaluation by principal component analysis[J]. Journal of liaoning normal university (natural science edition), 2004(04):403-406..
4. S W Meng. The problems that should be paid attention to in multi-index evaluation by principal component analysis[J]. Statistical study, 1992(4): 86-87.
5. J Li. Based on ARIMA model, this paper analyzes and predicts the consumption level of residents in anhui province[J]. Modern business, 2017(01): 195-196.
6. X Q He. Multivariate statistical analysis (fourth edition) [M]. Renmin university of China press, 2015, 3: 113-127.
7. W H Song, Q Zhang. The application of PCA algorithm in image feature reduction [J]. Journal of huangshan college, 2014, 16(05): 20-22.
8. W H Su. Study on the theory and method of multi-index comprehensive evaluation [D]. Xiammen: xiamen university, 2000.
9. Y J Guo. Comprehensive evaluation theory and method [M]. Beijing: science press, 2002: 111-117.

10. G E P Box, G M Jenkins, G C Reinsel. Time Series Analysis: Forecasting and Control (Renvised Edition) [J]. Journal of Marketing Research, 1994, 14(2): 353-412.
11. R A Fildes.The analysis of time series: Theory and practice: by C. Chatfield, Chapman and Hall, London(1975) [J]. Long Range Planning, 1976, 9(6): 113-114.
12. Y Peng. Introduction of ARIMA model [J]. Electronics World, 2014, 10: 259.
13. Summary of Chinese statistics 2011. http://www.yearbookchina.com/naviBooklist-YCDES-0. html, 2011.
14. P Z Li. The application of cluster analysis and principal component analysis in regional comprehensive consumption level evaluation[C]. Nankai university, 2008(11).
15. A E USORO, I U MOFFAT. Principal component analysis of nigeria nalue of major imports [J]. Am J Econ, 2015: 5(5): 508-512.
16. X F pan, X X Peng. Time series analysis [M]. beijing: Tsinghua university press, 2016: 40-62.
17. Z H Xiao, M Y Guo. Time series analysis and SAS application [M]. 2012: 44-65.