第三届“泰迪杯”

全国大学生数据挖掘竞赛

优

秀

作

品

作品名称：基于数据挖掘技术的市财政收入分析预测模型

荣获奖项：一等奖

作品单位：汕头大学

作品成员：林西西　陈炎君　王莎莎

指导教师：李健

基于 BP 神经网络的地方财政收入预测模型

本文针对广州市财政收入及影响财政收入关键因素的问题，以题目提供的各类税收收入及宏观经济和非经济指标数据为基础，利用典型相关分析、熵权系数法、灰色预测、主成分多元回归、BP 神经网络预测等方法，对众多复杂的数据进行多元统计分析和预测，得到对广州市财政评价的更为深层次的探究结果。

针对问题一，通过分析原数据，可以得出了历年地区财政收入为公共财政收入与基金预算收入之和的结论，并且历年的政府性基金收益率固定，每年收入也固定，所以我们把研究影响地方财政关键因素的问题转化为研究影响公共财政收入关键宏观因素问题。

我们通过典型相关分析，即利用宏观因素和对应关联的税种收入的相关关系来衡量两组指标的关联度，得出城市居民年人均可支配收入、第二产业增加值、城市商品零售价格指数、建筑企业利润总额、第三产业增加值、住宿和餐饮业零售额、全社会房地产开发投资额、地区生产总值、批发零售业增加值以及工业增加值是影响公共财政收入的关键因素的结论，而这些因素也是影响地方财政的关键因素。另一方面，利用熵权系数模型求得与公共财政收入关联宏观因素的权重，并确定关联度较大的指标。通过比较两个模型的结论基本一致。

针对问题二，我们把财政总收入分成公共财政收入类以及基金预算收入类两类。首先，对于公共财政收入类的预测，一方面，根据影响公共财政收入的关键宏观因素，采用灰色预测模型对原始数据做累加生成得到规律性较强的近似指数序列，再对各个宏观因素作预测；另一方面，根据题目给出的历年数据，我们利用主成分回归法建立公共财政收入关于主成分的回归方程，进而预算出公共财政收入。其次，对于基金预算收入类的预测，我们采用多项式拟合的方法对历年基金预算收入拟合，并作相应预测。最后，加总公共财政收入与基金预算收入预测值得到财政总收入，我们得出 2014 年和 2015 年地方财政总收入分别为 2453.9 亿元和 2843.6 亿元。为了优化模型和克服灰色预测—主成分回归模型在处理反馈信息时的缺陷，采用 BP 神经网络构建地方公共财政收入预测模型，以充分挖掘公共财政收入、支出与宏观经济活动的反馈关系，最后得出 2014 年和 2015 年公共财政收入的预测值分别为 1369 亿元和 1496.6 亿元。

针对问题三，我们通过对比历年财政支出情况，给出了 2015 年广州市财政预算草案一些分析和建议，并提出有效支配财政收入的策略。

关键词：典型相关分析；灰色预测；BP 神经网络；主成分回归分析

Local Financial Revenue Forecast Model Based on BP Neural Network

Abstract

Aiming Guangzhou revenue and revenue key factor which affect the Government revenue problems, we base on various types of tax revenue to provide the title and non-economic indicators and macroeconomic data. The Canonical Correlation Analysis, Entropy Coefficient, Gray Prediction, the main ingredient Multiple regression, BP neural network forecasting method, are used to analysis and forecast the complex data statistical tests, we get more in-depth financial evaluation of Guangzhou exploration results.

As for the question one, we can draw the calendar year revenue areas of public revenue and income fund budget and the conclusions of government funds through analying the original data , since annual income is fixed, so we change Local financial issues into finding key factors which affect public revenue study macroeconomic factors critical issue.

By using Canonical Correlation Analysis, using the correlation between macroeconomic factors and the corresponding revenue, to measure the correlation degree between two sets, we conclude the key factors is as follows, urban residents per capita disposable income, secondary industry, urban retail price index, construction enterprises total profits, the tertiary industry, accommodation and catering retail sales, total investment in real estate development, GDP, retail sales, and industrial added value. These factors are also key factors in the local financial . On the other hand, Entropy Coefficient Model is used to obtain and macroeconomic factors associated with heavy public revenue and identify indicators related degree. The two conclusions are similar through comparing two models.

As for the question two, we have divided the total government revenue into fund budget revenue and public income. Firstly, for the public revenue prediction, on one hand, based on the impact of public revenue according to key macroeconomic factors, the Gray Prediction Model turn the raw data to accumulated generating strong regularity approximate exponential sequence, and then make prodictions for various macroeconomic factors. On the other hand, based on the given historical data, we use Principal Component Regression to establishment public revenue on principal component regression equation, and then the budget of the public revenue. Secondly, for the kind of fund budget revenue forecast, we use polynomial fitting method to fit the calendar year fund budget revenue, and forecast accordingly. Finally, we add the total public revenue and fund budget revenue forecast to total revenue worth. The local fiscal revenue forecast in 2014 and 2015 are 245.39 billion yuan and 284.36 billion yuan respectively. In order to overcome the gray prediction - Principal Component Regression Model defects in dealing with feedback information, the use of BP neural network to build local public financial revenue forecast model, in order to fully tap the feedback between public revenue and expenditure and macroeconomic activity.Finally,the result of public revenue in 2014 and 2015 are 1369 billion yuan and 1496.6 billion yuan respectively.

As for question three, we compare the financial expenditure over the years and give the 2015 draft budget in Guangzhou, some analysts and recommendations and propose effective strategies disposable revenue.

KeyWords:Canonical Correlation Analysis;GrayForecasting; BP neural network Principal Component Regression Analysis

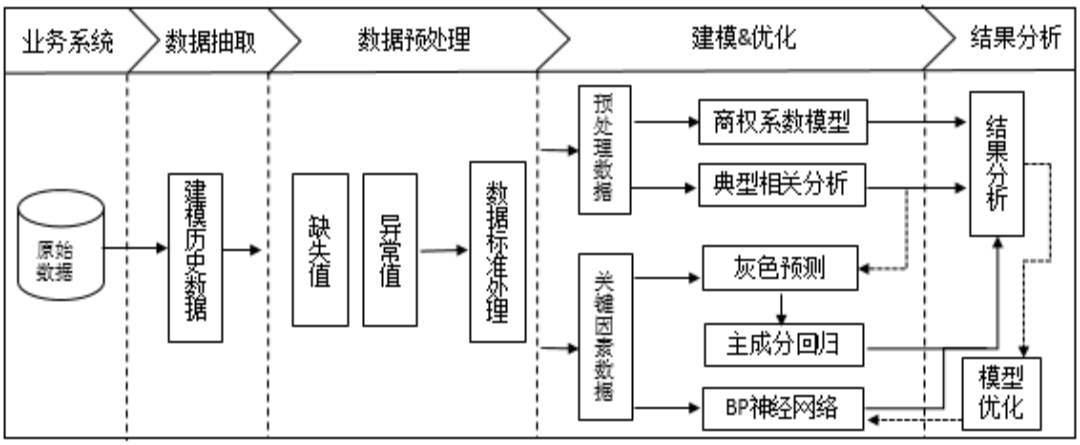
目 录

1. 挖掘目标

本文是基于 1999 年-2013 年地方财政收入以及其关联指标的数据，采用典型相关分析研究不同市辖区企业所得税、营业税、增值税和个人所得税和相应关联宏观指标指标的相关，确定影响地方财政收入的关键因素。并基于上述因素分析，我们把财政总收入的预测化分成公共财政收入类预测以及基金预算收入类预测两个类别，分别进行公共预算收入的预测和政府性基金收入的预测。对于公共预算收入使用灰色模型预测关联指标的值，通过主成分降维的方法构建公共预算收入的回归方程；由于经济活动和财政收入这两个体系是存在交互影响的，因此我们提出了 BP 神经网络用于改进缺乏处理反馈信息能力的灰色预测—主成分回归模型；对于政府性基金收入，则通过拟合政府性基金收入离散序列来预测短期的数值，最后求出地方财政收入。同时通过研究近几年广州市财政支出、民生工程、城市维护建设支出等项目以及结合 2015 年预算草案提出建议。

2. 分析方法与过程

2.1. 总体流程



本用例主要包括如下步骤：

步骤一：数据预处理

步骤二：建模与优化

步骤三：模型结果分析

2.2. 具体步骤

2.2.1 数据预处理

样本数据预处理主要包括缺失值处理、异常值处理、数据整理和数据标准化等。

1. **异常值处理**

在原始样本数据中，存在一些异常值的情况，如增值税表中 2012 年、2013 年批发零售业增加值的数据，其较之往常的数据都异常小，我们将其定义为异常值。考虑到国家政策调整等因素的影响，首先我们在广州市统计信息网（http://www.gzstats.gov.cn/）挖取 1999 年至 2013 年批发零售业增加值的数据，发现 2012 年和 2013 年的数据是错误的，直接采用网站数据替代。

商品进口总值	地区生产总值	工业增加值	批发零售业零售额	工业增加值占GDP	批发零售业增加值	增值税		
93.18	21391758	7980207	6661555	0.373050546	0	288972		
115.6	24927434	8779835	7839516	0.352215756	0	350495		
114.13	28416511	9554676	8803979	0.336236774	0	443213		
141.49	32039616	10509450	9733195	0.328014231	0	526377		
180.52	37586166	13141254	11833760	0.349630074	0	581898		
233.14	44505503	15941538	14030973	0.358192514	4636933	528365		
268.07	51542283	18439550	16171817	0.357755787	5574275	816119		
313.85	60818614	22270093	18921556	0.36617232	5907373	967265		
355.91	71403223	26029310	22841850	0.364539707	6875421	1115007		
389.47	82873816	29724781	27953721	0.358675182	9328615	1287226		
392.82	91382135	31173422	31565720	0.341132564	11237237	1375085		
553.89	107482828	36449611	38835933	0.33912032	13541607	1594182		
596.94	124234390	41405926	45444614	0.333288762	15947698	1573830		
582.52	135512072	42641557	51685711	0.314669803	3837376	1758311		
560.89	154201434	47548175	59858717	0.308351056	4168317	2216017		

缺失值

异常值

图 1-原始数据的部分异常值与缺失值

2. 缺失值处理

在企业所得税表和增值税表中，发现数据缺失的现象，究其原因可归为类：第一类原因是数据丢失；第二是国家政策更改；第三是统计局没有录入相应的数据。因此，对于丢失的数据我们用数据挖掘的方法找出；我们要研究个人所得税、企业所得税、增值税和营业税与相应关联指标的影响，因此我们需要的是缺失指标的一个趋势，且国家的经济指标在一个经济稳定的环境中是缓慢变化的，因此第二类和第三类缺失值不会出现剧增或者剧减，因此可以采用数据处理方法求出缺失值。常用的求缺失值的方法有平均法、移动平均法、时间序列推测和加权调整，考虑到平均法、移动平均法和加权平均会消除序列的趋势，而此些经济指标往往是随着经济发展而增大，因此并不适合采用上述方法，同时通过时间序列推测法检测第三产业增加值、地区生产总值等具有比较接近真实值的结果，因此我们采用时间序列推测法推测出在稳定的值。

对缺失数据列采用时间序列分析，利用 ARMA 模型推算序列的值.

3. 数据整理

选取并合并个人所得税表、企业所得税表、增值税表和营业税表所有宏观经济的数据，并提取地方财政收入表各市辖区个人所得税、企业所得税、增值税和营业税的值。

4. 数据标准处理

由于数据存在不同的量纲，我们采用 Z-Score 值标准方法对数据进行标准化处理。

设数据为$(x_1, x_2, \cdots, x_n)$,其均值为$\bar{x}$，标准差为σ，标准化公式如下：

$$x_i^* = \frac{x_i - \bar{x}}{\sigma}$$

2.2 模型建立与优化

2.2.1 问题一的模型

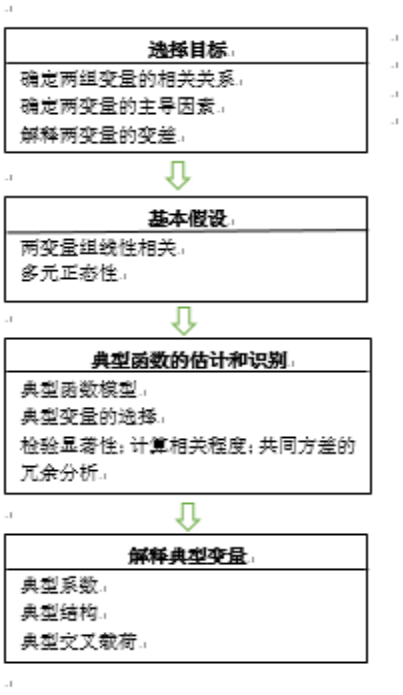
历年地区财政收入为公共财政收入与基金预算收入之和的结论，并且历年的政府性基金收益率固定，每年收入也固定，所以我们把研究影响地方财政关键因素的问题转化为研究影响公共财政收入关键宏观因素问题。

MODEL1 典型相关分析

典型相关分析是研究多个变量与多个变量之间的相关关系，利用主成分分析的思想将多个变量与多个变量的相关化为两个新变量的相关，即求$\alpha = (\alpha_1, \alpha_2, \cdots \alpha_p), \beta = (\beta_1, \beta_2, \cdots \beta_q)$使得新变量$u = \alpha X, v = \beta Y$之间有最大可能的相关。

基于这个思想我们将所有关联宏观经济的指标设为X，将广州市区、增城市、从化市的个人所得税、企业所得税、增值税和营业税设为Y，研究影响X、影响Y的主要表现因素，在通过研究典型系数、典型结构分析宏观经济组和税收组的相关关系。

1. 典型相关性逻辑分析图



2. 典型相关系数和典型结构

设有两组变量 $X=(X_1,X_2,\cdots,X_p)',Y=X=(Y_1,Y_2,\cdots,Y_q)'$，假设其都经过标准化处理，即有 $E(X_i)=0,D(X_i)=1,E(Y_j)=0,D(Y_j)=1,i=1,2,\cdots p,\ j=1,2,\cdots,q$.设总体为 $Z=\begin{bmatrix}X & Y\end{bmatrix}$。

已知总体 Z 的 n 次观测数据为：

$$Z_{(t)}=\begin{bmatrix}X_{(t)}\\ Y_{(t)}\end{bmatrix}_{(p+q)\times 1} \qquad (t=1,2,\cdots,n),$$

假定 $Z\ \square\ N_{p+q}(\mu,\Sigma)$,其协方差阵 Σ 的最大似然估为：

$$S=\frac{1}{n-1}\sum_{t=1}^{n}(Z_{(t)}-\overline{Z})(Z_{(t)}-\overline{Z})',$$

S 为样本协方差阵，其中 $\overline{Z}=\frac{1}{n}\sum_{t=1}^{n}Z_{(t)}$，可将 S 相应的分割为：

$$S=\begin{bmatrix}S_{11} & S_{12}\\ S_{21} & S_{22}\end{bmatrix}$$

使 $u_k=\alpha X,v_k=\beta Y$ 达到最大相关的 α,β 是 $\square\square'$ 的特征值 $\hat{\lambda}_k^2$ 所对应的特征向量 $\square_k$，其中 $\square\square'=S_{11}^{-1/2}S_{12}S_{22}^{-1/2}$，其特征值依次为

$$\hat{\lambda}_1^2\geq\hat{\lambda}_2^2\geq\cdots\geq\hat{\lambda}_p^2>0,(\hat{\lambda}_i>0,i=1,2,\cdots,p)。$$

令

$$\begin{cases}\alpha_k = S_{11}^{-1/2}\hat{l}_k \\ \beta_k = \hat{\lambda}_k^{-1}S_{22}^{-1}S_{21}\alpha_k\end{cases}$$

则 $\hat{u}_k = \alpha_k X, \hat{v}_k = \beta_k Y$ 为样本 X 和 Y 的第 k 对样本典型相关变量；而 $\hat{\lambda}_k^2$ 为样本 X 和 Y 的第 k 个样本典型相关系数。

3.典型冗余分析

在典型相关分析中，因所提取的每对典型成分保证其相关程度达到最大，故每一个典型成分不仅解释了本组变量的信息，还解释了另一组变量的信息。典型冗余分析是包括典型变量解释本组变量的的总变差百分比和典型变量解释另一组变量总变差百分比。

第 k 个典型变量解释本组变量 X （或 Y ）总变差百分比如为下式：

$$R_d(X;u_k) = \frac{1}{p}\sum_{j=1}^{p} r^2(X_j,u_k),$$

$$R_d(Y;v_k) = \frac{1}{q}\sum_{j=1}^{q} r^2(Y_j,v_k).$$

第 k 个典型变量解释另一组变量 Y （或 X ）总变差百分比如为下式：

$$R_d(X;v_k) = \frac{1}{p}\sum_{j=1}^{p} r^2(X_j,v_k),$$

$$R_d(Y;u_k) = \frac{1}{q}\sum_{j=1}^{q} r^2(Y_j,u_k).$$

典型相关系数越大，典型成分解释对方变量组变差的信息也越多，就能衡量两组变量之间的影响程度大小。

通过 SAS 中的 proc cancorr 函数运行程序 Que1.sas 得到各税收组与宏观指标组的相关关系，输出源文件为附件资料问题一中。

MODEL2 熵权系数模型

在上一个模型我们提出了典型相关分析分析关联宏观因素与不同地区的不同税种的相关关系，在典型相关分析中我们主要根据典型结构及其冗余结构分析关联宏观指标的影响，这启发我们利用各个税种相应关联宏观指标的变异程度来衡量指标的影响程度，因此我们采用研究指标熵值的熵权系数模型来作财政预算收入因素分析，通过实践两种分析方法的结果具有相似性，下面是熵权系数模型。

1.理论知识

熵的概念源于热力学, 表示不能用来做功的热能,其计算方式为热能的变化量除以温度所得的熵,后由申农(C. E. S hannon)引入信息论, 现已在工程技术、社会经济等领域得到了广泛地应用。在信息论中,信息是系统有序程度的一个度量,熵是系统无序程度的一个度量,两者绝对值相等,符号相反,当系统可能处于几种不同状态、每种状态出现的概率为 $P_i, i = 1,2,\cdots,m$,该系统的熵定义为：

$$\varphi = -\sum_{i=1}^{m} P_i \ln(P_i)$$

显然，当 $P_i = P_j, \forall i \neq j$ 时，即 $P_i = \frac{1}{c}, i = 1,2,\cdots,c$ 时 3，其熵取得最大值，其最大值为：

$$\varphi_{\max} = \ln(c)$$

设有 n 组观测值，c 个指标，矩阵 $H = (h_{ik})_{n\times c}$ 为原始数据数据，通过以下方法对 H 进行归一化处理：

$$P_{i,k} = \frac{h_{i,k}}{\sum_{j=1}^{n} h_j}$$

第 k 个指标的信息熵为：

$$\varphi_k = \sum_{t=1}^{n} P_{ik} \ln(P_{ik})$$

某一个指标的信息熵越小，表明其指标的变异程度越大，并记

$$e_k = \frac{\varphi_k}{\ln c}$$

第 k 个指标的的客观权重为：

$$\omega_k = \frac{1-e_k}{\sum_{t=1}^{c}(1-e_t)}$$

通过 EXCEL 对每一组宏观指标变量进行上述操作，输出结果见附录资料熵权系数模型的 EXCEL 结果。

2.2.2 问题二的模型

我们把财政总收入的预测化分成公共财政收入类预测以及基金预算收入类预测。首先，对于公共财政收入类的预测，我们根据影响公共财政收入的关键宏观因素，采用灰色预测模型对原始数据做累加生成得到规律性较强的近似指数序列，再对各个宏观因素作预测。在灰色预测的基础上考虑利用主成分回归以预测公共财政收入。再者，对于基金预算收入类的预测，我们则采用拟合基金收入历史离散数据得到基金预算收入预测。

MODEL1 主成分回归和灰色预测模型

灰色预测模型（Gray Forecast Model，GM）是通过少量的、不完全的信息，建立数学模型并做出预测的一种预测方法。预测是根据客观事实的过去和现在的发展规律，借助于科学的方法对未来的发展趋势和状况进行描述和分析，并形成科学的假设和判断。

在问题 1 中，我们已经分析出影响地方财政收入的关键影响因素有：城市居民人均可支配收入、第二产业增加值、城市商品零售价格指数、建筑业企业利润总额、第三产业增加值、住宿和餐饮业零售额、全社会房地产开发投资额、地区生产总值、工业增加值、批发零售业增加值以及政府性基金收入，其中前面十个因素是影响公共财政收入的关键性指标。

我们选取 1999-2013 年 15 年的指标数据作为历史数据，由于我们的历史数据少、系列的完整性及可靠性比较低，而灰色预测模型能利用微分方程来充分挖掘系统的本质且精度高从而有效地解决这类问题，所以我们采用该模型对十个关键影响因素进行短期预测，并采用主成分回归分析模型得到以这十个因素作为自变量，公共财政收入为因变量的回归方程。

Step1. GM(1,1)预测模型

1.GM(1,1)模型预测方法

定义：已知参考序列（即 1993-2013 年这 15 年间 10 个关键影响因素的数据列）$x^{(0)}=(x^{(0)}(1),x^{(0)}(2),...,x^{(0)}(\mathrm{n}))$，1 次累加生成序列（1-AGO）

$$x^{(1)}=(x^{(1)}(1),x^{(1)}(2),...,x^{(1)}(\mathrm{n}))$$
$$=(x^{(1)}(1),x^{(1)}(1)+x^{(1)}(2),...,x^{(1)}(1)+...+x^{(1)}(\mathrm{n}))$$

其中：$x^{(1)}(k)=\sum_{i=1}^{k}x^{(0)}(i),k=1,2,...,n$。$x^{(1)}$ 的均值生成序列

$$z^{(1)}=(z^{(1)}(2),z^{(1)}(3),...,z^{(1)}(n))$$

其中：$z^{(1)}(k)=0.5x^{(1)}(k)+0.5x^{(1)}(k-1),k=2,3,...,n.$

建立灰微分方程 $x^{(0)}(k)+az^{(1)}(k)=b,k=2,3,...,n.$

相应的白化微分方程为

$$\frac{dx^{(1)}}{dt}+ax^{(1)}(t)=b \qquad (*)$$

记 $u=[a,b]^T,Y=[x^{(0)}(2),x^{(0)}(3),...,x^{(0)}(n)]^T,B=\begin{bmatrix}-z^{(1)}(2) & 1\\ -z^{(1)}(3) & 1\\ \mathrm{M} & \mathrm{M}\\ -z^{(1)}(n) & 1\end{bmatrix}$，则由最小二乘法，求得使 $J(u)=(Y-Bu)^T(Y-Bu)$ 达到最小值 u 的估计值为 $\hat{u}=[a,b]^T=(B^TB)^{-1}B^TY$

于是求解方程（*）得

$$\hat{x}^{(1)}(k+1)=(\hat{x}^{(0)}(1)-\frac{\hat{b}}{\hat{a}})e^{-\hat{a}k}+\frac{\hat{b}}{\hat{a}},k=0,1,...,n-1,...$$

2.GM(1,1)模型预测步骤

1) 数据的检验与处理

首先，为了保证建模方法的可行性，需要对 1993-2013 年这 15 年间 10 个关键影响因素的数据列作必要的检验处理。设参考数据为 $x^{(0)}=(x^{(0)}(1),x^{(0)}(2),...,x^{(0)}(\mathrm{n}))$，其中 $n=10$，计算序列的级比

$$\lambda(k)=\frac{x^{(0)}(k-1)}{x^{(0)}(k)},k=2,3,...,n.$$

如果所以的级比都落在可容覆盖 $\Theta=(e^{-\frac{2}{n+1}},e^{\frac{2}{n+2}})$ 内，则序列 $x^{(0)}$ 可以作为模型 GM(1,1)的数据进行灰色预测。否认，需要对序列 $x^{(0)}$ 做必要的变幻处理，使其

落入可容覆盖内。即取适当的常数c，做平移变换

$$y^{(0)}(k)=x^{(0)}(k)+c,k=1,2,...,n.$$

s.t. 序列 $y^{(0)}=(y^{(0)}(1),y^{(0)}(2),...,y^{(0)}(n))$的级比

$$\lambda_y(k)=\frac{y^{(0)}(k-1)}{y^{(0)}(k)}\in\Theta,k=2,3,...,n.$$

2） 建立模型

按式（*）建立 GM(1,1)模型，则可以得到预测值

$$\hat{x}^{(1)}(k+1)=(\hat{x}^{(0)}(1)-\frac{\hat{b}}{\hat{a}})e^{-\hat{a}k}+\frac{\hat{b}}{\hat{a}},k=0,1,...,n-1,...$$

而且 $\hat{x}^{(0)}(k+1)=\hat{x}^{(1)}(k+1)-\hat{x}^{(1)}(k),k=1,2,...,n-1,...$

3)：检验预测值

□ 残差检验。令残差为$\varepsilon(k)$，计算

$$\varepsilon(k)=\frac{x^{(0)}(k)-\hat{x}^{(0)}(k)}{x^{(0)}(k)},k=1,2,...,n.$$

这里 $\hat{x}^{(0)}(1)=x^{(0)}(1)$，如果$\varepsilon(k)<0.2$，则可认为达到一般要求；如果$\varepsilon(k)<0.1$，则认为达到较高要求。

□ 级比偏差值检验。首先由参考数据$x^{(0)}(k-1),x^{(0)}(k)$，计算出级比$\lambda(k)$，再用发展系数a求出相应的级比偏差

$$\rho(k)=1-(\frac{1-0.5a}{1+0.5a})\lambda(k)$$

如果 $\rho(k)<0.2$，则可认为达到一般要求；如果 $\rho(k)<0.1$，则认为达到较高要求。

3.预测预报

由 GM(1,1)模型得到指定时区内的预测值，根据实际问题的需要，给出相应的预测预报。

在 MATLAB 中运行 Que2_1_1 等文件即可得到各个主要因素的预测值。

Step2. 主成分回归

假设进行主成分分析的指标变量有m个，共有n个评价对象，分别为$a_1,a_2,\cdots,a_m$，且它们经过标准化处理化为$a_1^*,a_2^*,\cdots,a_m^*$。

1.计算相关系数矩阵R。

相关系数矩阵$R=(r_{ij})_{m\times m}$，有

$$r_{ij}=\frac{\sum_{k=1}^{n}a_{ki}^*\cdot a_{kj}^*}{n-1},i,j=1,2,\cdots,m,$$

其中：$r_{ii}=1,r_{ij}=r_{ji},r_{ij}$是第个指标与第个指标的相关系数。

2.计算特征值和特征向量。

计算相关系数矩阵R的特征值$\eta_1\geq\eta_2\geq\cdots\geq\eta_m\geq0$，以及对应的特征向量$u_1,u_2,\cdots,u_m$，其中$u_j=[u_{1j},u_{2j},\cdots,u_{mj}]^{\mathrm{T}}$，由特征向量组成$m$个新的指标变量：

$$z_1 = u_{11}a_{11}^* + u_{21}a_{21}^* + \cdots + u_{m1}a_{m1}^*$$
$$z_2 = u_{12}a_{12}^* + u_{22}a_{22}^* + \cdots + u_{m2}a_{m2}^*$$
$$\vdots$$
$$z_m = u_{1m}a_{1m}^* + u_{2m}a_{2m}^* + \cdots + u_{mm}a_{mm}^*$$

其中 z_1 是第 1 主成分，z_2 是第二个主成分，…，z_m 是 m 第个主成分。

3.计算特征值的信息贡献率

选择 s 个主成分，计算其特征值 $\eta_j(j=1,2,\cdots,m)$ 信息贡献率和累计贡献率。称

$$w_j = \frac{\eta_j}{\sum_{k=1}^{m}\eta_k}, j = 1,2,\cdots,m$$

为主成分的信息贡献率，同时有

$$\alpha_s = \frac{\sum_{k=1}^{s}\eta_k}{\sum_{k=1}^{m}\eta_k}$$

为主成分 $z_1, z_2, \cdots, z_s$ 的累计贡献率。当 α_s 接近于 1 时（一般取 $\alpha_s = 0.85, 0.90, 0.95$）时，则选择前 s 个指标变量 $z_1, z_2, \cdots, z_s$ 作为 s 个主成分，代替原来 m 个指标变量，从而可对 s 个主成分进行回归分析。

4.建立回归方程

用 s 个主成分与目标预测变量建立回归模型：

$$\begin{cases} I = b_0 + b_1 Z_1 + \cdots + b_s Z_s + \varepsilon (s \le m) \\ E(\varepsilon) = 0, Var(\varepsilon) = \sigma^2, Cov(\varepsilon_i, \varepsilon_j) = 0 (i \ne j) \end{cases}$$

利用最小二乘法使得 $Q(\beta) = \sum_{t=1}^{n}\varepsilon_t^2 = \sum_{t=1}^{n}[I_t - (\beta_0 + \beta_1 x_{t1} + \cdots + \beta_s x_{ts})]^2$ 达到最小

可得向量 β 的最小二乘估计 $\hat{\beta}$ 为

$$\hat{\beta} = (C'C)^{-1}C'I$$

其中

$$C = \begin{bmatrix} 1 & z_{11} & z_{12} & \cdots & z_{1s} \\ 1 & z_{21} & z_{22} & \cdots & z_{2s} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & z_{n1} & z_{n2} & \cdots & z_{ns} \end{bmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

在 MATLAB 中运行程序 Que2.m，即可得到相应主要影响因素的预测值和灰色预测效果图。

MODEL2. 拟合

由于政府性基金收入是独立于公共预算收入的变量，因此我们对这组数据采多项式用拟合，在 MATLAB 的环境下利用平 polyfit 函数进行拟合，得到政府性基金收入关于年份的函数表达式，并利用此函数表达式推算出 2014 年和

2015 年政府性基金收入的预算值

通过加总公共预算收入预测值和政府性基金收入指得到地方性财政收入预测值。

我们通过运行程序，得到 2014 年和 2015 年公共预算收入和政府性基金收入以及地方财政收入的预测值。

2.2.3 模型优化

MODEL3 BP 神经网络

上述的灰色预测—主成分回归模型其对变量信息反馈处理能力较弱，而很多经济活动不仅影响财政收入，同时它们也与其他经济活动相关联，即经济活动与财政收入是存在反馈活动和交互影响的，因此上述模型无法处理财政收入影响各个指标变量的信息以及无法衡量此信息对财政收入潜在影响。由于神经网络具有学习和存贮大量的输入-输出模式映射关系以及不断修正传播误差的能力，为充分挖掘公共财政收入、支出与宏观经济活动的正反馈和负反馈关系以及不断缩小模型误差，我们利用 BP 神经网络构建地方公共预算收入预测模型。将每一个变量看作一个神经元，将公共预算收入预测值作为输出变量。

BP（Back Propagation）网络是 1986 年由 Rumelhart 和 McCelland 为首的科学家小组提出，是一种按误差逆传播算法训练的多层前馈网络，是目前应用最广泛的神经网络模型之一。BP 网络能的学习规则是使用最速下降法，通过反向传播来不断调整网络的权值和阈值，使网络的误差平方和最小。BP 神经网络模型拓扑结构包括输入层（input）、隐层(hide layer)和输出层(output layer)（如图 2 所示）

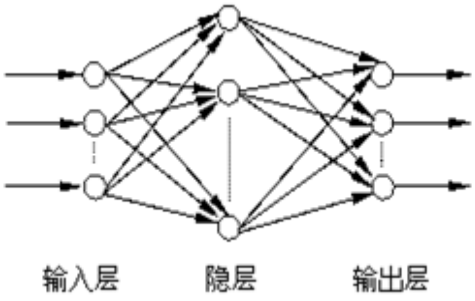


图 2-BP 神经网络模型拓扑结构图

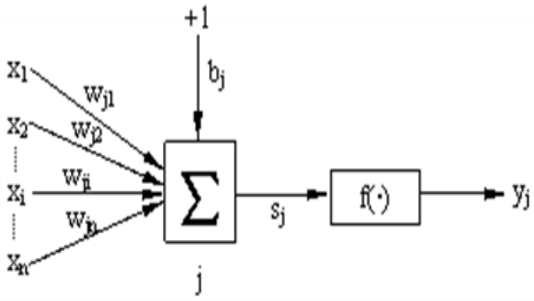


图 3-第 *j* 个基本 BP 神经图

1. 公共预算收入 BP 神经元

图 3 给出了第 j 个基本 BP 神经元（节点），它模仿了生物神经元所具有的三个最基本也是最重要的功能：加权、求和与转移。其中 $x_1, x_2, ..., x_i, ..., x_n$ 分别代表来自神经元 $1, 2, ..., i, ..., n$ 的输入； $w_{j1}, w_{j2}, ..., w_{ji}, ..., w_{jn}$ 则分别表示神经元 $1, 2, ..., i, ..., n$ 与第 j 个神经元的连接强度，即权值； b_j 为阈值； $f(\cdot)$ 为传递函数； y_j 为第 j 个神经元的输出。第 j 个神经元的净输入值 S_j 为：

$$S_j = \sum_{i=1}^{n} w_{ji} \cdot x_i + b_j = W_j X + b_j$$

其中

$$X = [x_1, x_2, ..., x_i, ..., x_n]^T, W_j = [w_{j1}, w_{j2}, ..., w_{ji}, ...w_{jn}]$$

若视 $x_0 = 1, w_{j0} = b_j$ ，即令 X 及 W_j 包括 x_0 及 w_{j0} ，则

$$X = [x_0, x_1, x_2, ..., x_i, ..., x_n]^T, W_j = [w_{j0}, w_{j1}, w_{j2}, ..., w_{ji}, ...w_{jn}]$$

于是节点 j 的净输入 S_j 可表示为

$$S_j = \sum_{i=0}^{n} w_{ji} x_i = W_j X$$

净输入 S_j 通过传递函数（Transfer Function） $f(\cdot)$ 后，便得到第 j 个神经元的输出 y_j ，

$$y_j = f(S_j) = f(\sum_{i=0}^{n} w_{ji} \cdot x_i) = F(W_j X)$$

式中 $f(\cdot)$ 是单调上升函数，而且必须是有界函数，因为细胞传递的信号不可能无限增加，必存在最大值。

2. 财政收入 BP 网络

BP 算法由数据流的前向计算（正向传播）和误差信号的反向传播两个过程构成。正向传播时，传播方向为输入层→隐层→输出层，每层神经元的状态只影响下一层神经元。若在输出层得不到期望的输出，则转向误差信号的反向传播流程。通过这两个过程的交替进行，在权向量空间执行误差函数梯度下降策略，动态迭代搜索一组权向量，使网络误差函数达到最小值，从而完成信息提取和记忆过程。

- **正向传播**

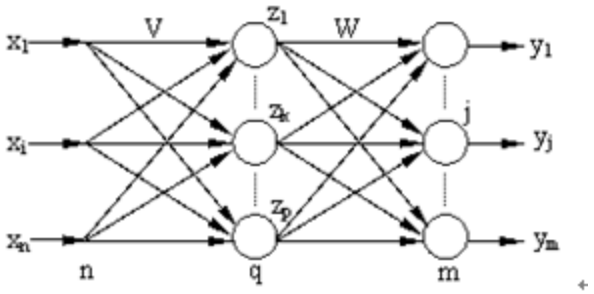


图 4-BP 网络的图谱结构图

设 BP 网络的输入层有 n 个节点，隐层有 q 个节点，输出层有 m 个节点，输入层与隐层之间的权值为 v_{ki}，隐层与输出层之间的权值为 w_{jk}，如图 4 所示。隐层的传递函数为 $f_1(\cdot)$，输出层的传递函数为 $f_2(\cdot)$，则隐层节点的输出为（将阈值写入求和项中）：

$$z_k = f_1(\sum_{i=0}^{n} v_{ki} x_i), k = 1, 2, ..., q$$

输出层节点的输出为：

$$y_j = f_2(\sum_{k=0}^{q} w_{jk} z_k), j = 1, 2, ..., m$$

至此 B-P 网络就完成了 n 维空间向量对 m 维空间的近似映射。

- **反向传播**

Step1 定义误差函数

输入 P 个学习样本，用 $x^1, x^2, ..., x^i, ..., x^P$ 来表示。第 p 个样本输入到网络后得到输出 $y_j^p (j = 1, 2, ..., m)$。采用平方型误差函数，于是得到第 p 个样本的误差 E_p：

$$E_p = \frac{1}{2}\sum_{j=1}^{m}(t_j^p - y_j^p)^2$$

式中：t_j^p 为期望输出。

对于 P 个样本，全局误差为：

$$E = \frac{1}{2}\sum_{p=1}^{P}\sum_{j=1}^{m}(t_j^p - y_j^p) = \sum_{p=1}^{P} E_p$$

Step2 输出层权值的变化

采用累计误差 BP 算法调整 w_{jk}，使全局误差 E 变小，即

$$\Delta w_{jk} = -\eta \frac{\partial E}{\partial w_{jk}} = -\eta \frac{\partial}{\partial w_{jk}}(\sum_{p=1}^{P} E_p) = \sum_{p=1}^{P}(-\eta \frac{\partial E_p}{\partial w_{jk}})$$

式中：η 表示学习率

定义误差信号为：

$$\delta_{yj} = -\frac{\partial E_p}{\partial S_j} = -\frac{\partial E_p}{\partial y_j} \cdot \frac{\partial y_j}{\partial S_j}$$

其中第一项：$\frac{\partial E_p}{\partial y_j} = \frac{\partial}{\partial y_j}[\frac{1}{2}\sum_{j=1}^{m}(t_j^p - y_j^p)^2] = -\sum_{j=1}^{m}(t_j^p - y_j^p)$

第二项：

$$\frac{\partial y_j}{\partial S_j} = f_2^{'}(S_j)$$

是输出层传递函数的偏微分。
于是：

$$\delta_{yj} = \sum_{j=1}^{m}(t_j^p - y_j^p) f_2^{'}(S_j)$$

由链定理得：$\frac{\partial E_p}{\partial w_{jk}} = \frac{\partial E_p}{\partial S_j} \cdot \frac{\partial S_j}{\partial w_{jk}} = -\delta_{yj} \cdot z_k = -\sum_{j=1}^{m}(t_t^p - y_j^p) f_2^{'}(S_j) \cdot z_k$

于是输出层各神经元的权值调整公式为：$\Delta w_{jk} = \sum_{p=1}^{P}\sum_{j=1}^{m}\eta(t_t^p - y_j^p) f_2^{'}(S_j) z_k$

Step3 隐层权值的变化

$$\Delta v_{ki} = -\eta \frac{\partial E}{\partial v_{ki}} = -\eta \frac{\partial}{\partial v_{ki}}(\sum_{p=1}^{P} E_p) = \sum_{p=1}^{P}(-\eta \frac{\partial E_p}{\partial v_{ki}})$$

定义误差信号为：

$$\delta_{zj} = -\frac{\partial E_p}{\partial S_j} = -\frac{\partial E_p}{\partial z_k} \cdot \frac{\partial z_k}{\partial S_j}$$

其中第一项：

$$\frac{\partial E_p}{\partial z_k} = \frac{\partial}{\partial z_k}[\frac{1}{2}\sum_{j=1}^{m}(t_j^p - y_j^p)^2] = -\sum_{j=1}^{m}(t_j^p - y_j^p)\frac{\partial y_i}{\partial z_k}$$

依链定理有：

$$\frac{\partial y_j}{\partial z_k} = \frac{\partial y_j}{\partial S_j} \cdot \frac{\partial S_j}{\partial z_k} = f_2^{'}(S_j) w_{jk}$$

第二项：

$$\frac{\partial z_k}{\partial S_k} = f_1^{'}(S_k)$$

是隐层传递函数的偏微分。
于是：

$$\delta_{zk} = -\sum_{j=1}^{m}(t_j^p - y_j^p) f_2^{'}(S_j) w_{jk} f_1^{'}(S_j)$$

由链定理得：

$$\frac{\partial E_p}{\partial v_{ki}} = \frac{\partial E_p}{\partial S_k} \cdot \frac{\partial S_k}{\partial v_{ki}} = -\delta_{zk} x_i = -\sum_{j=1}^{m}(t_j^p - y_j^p) f_2^{'}(S_j) w_{jk} f_1^{'}(S_j) \cdot x_i$$

从而得到隐层各神经元的权值调整公式为：

$$\Delta v_{ki} = \sum_{p=1}^{P}\sum_{j=1}^{m}\eta(t_j^p - y_j^p) f_2^{'}(S_j) w_{jk} f_1^{'}(S_k) x_i$$

借助于 MATLAB 神经网络工具箱来实现十个关键影响因素与公共财政收入的转换。神经网络的实际输出值与输入值以及各权值和阈值有关，为了使实际输出值与网络期望输出值相吻合，可用含有一定数量学习样本的样本集和相应期望输出值的集合来训练网络。训练时为 1999-2013 年这 15 年城市居民人均可支配

收入、第二产业增加值、城市商品零售价格指数、建筑业企业利润总额、第三产业增加值、住宿和餐饮业零售额、全社会房地产开发投资额、地区生产总值、工业增加值、批发零售业增加值的实测样本数据。即该神经网络中自变量有 10 个，因变量有 1 个，输入神经元为 4，输出神经元的个数为 1。

通过多次运行程序（附件资料/问题二/神经网络/Que2_3_1, Que2_3_2）并求出程序中的值平均值作为 2014 年和 2015 年公共财政收入的预测值。

2.2.3：结果分析

由典型相关分析和熵权系数模型找出不同税种的关键影响因素，查找资料研究这些因素如何影响财政收入；利用灰色预测—主成分回归预报模型以及通过 BP 神经网络建立的预报系统估计 2014 年相应财政收入类别，并与真实值进行比较以评估我们的模型；检验主成分回归与神经网络的效果图是否满足要求。

2.3. 结果分析

2.3.1 典型相关分析

通过 SAS 运行得到结果，通过整合输出信息得到以下结果（程序见资料附件/问题一/Que1.sas）

- **显著性水平检测**

四组典型变量分析中在 0.1 的显著水平下都只有一个典型相关系数是显著的，后文的叙述都选择一对典型变量。

表 1-典型系数和相关系数检验表

	典型相关	特征值			H_0：当前行和随和所有行相关值都是零			
		特征值	比例	累计比例	近似 F 值	分子自由度	分母自由度	Pr > F
个人所得税组	0.97299	17.1611	0.9584	0.9584	2.05	18	14.627	0.0855
企业所得税组	0.997169	175.8912	0.8997	0.8997	5.21	27	6.4834	0.0191
营业税组	0.99719	176.9195	0.8922	0.8922	5.01	27	6.4834	0.0212
增值税组	0.997399	191.4787	0.9930	0.9930	6.61	18	14.627	0.00003

我们通过相关系数的 χ^2 统计量检验确定典型变量相关性的显著程度，在 0.1 的显著性水平上，第一典型相关系数通过了检验，说明第一典型变量相关性显著，能够用“个人所得税宏观指标组”来解释“市辖区个人所得税组”。同理，在 0.1 的显著性水平上，我们能够用“企业所得税宏观指标组”来解释“市辖区企业所得税组”，用“营业税宏观指标组”来解释“市辖区营业税组”，用“增值税宏观指标组”来解释“市辖区增值税组”。

- **典型变量**

通过四次典型相关分析，我们得到每一次分析的典型变量，如下显示。鉴于原始变量的计量单位不同，不宜直接比较，本文采用标准化的典型系数，给出典

型相关模型，如下表列所示，表中 $x_i^*(i=1,...,6)$ 和 $y_j^*(j=1,...,3)$ 是标准化变量。

个人所得税组典型变量：

$$u_{11}=2.5089x_{11}^*+0.0711x_{12}^*-1.5488x_{13}^*-0.6412x_{14}^*+1.3572x_{15}^*-0.8010x_{16}^*$$

$$v_{11}=2.9914y_{11}^*-1.006y_{12}^*-1.2976y_{13}^*$$

企业所得税组典型变量：

$$u_{21}=1.0234x_{21}^*-0.0016x_{22}^*-0.4642x_{23}^*+0.2983x_{24}^*-0.6128x_{25}^*+0.5267x_{26}^*-1.1385x_{27}^*+2.1428x_{28}^*+0.1104x_{29}^*$$

$$v_{21}=1.0316y_{21}^*-0.1887y_{22}^*-0.1459y_{23}^*$$

营业税组典型变量：

$$u_{31}=-2.8284x_{31}^*+0.5877x_{32}^*-0.9473x_{33}^*+5.3132x_{34}^*+0.7817x_{35}^*-0.0554x_{36}^*-0.3160x_{37}^*+0.2885x_{38}^*-1.9058x_{39}^*$$

$$v_{31}=-0.9668y_{31}^*+1.6110y_{32}^*+0.2544y_{33}^*$$

增值税组典型变量

$$u_{41}=0.2227x_{41}^*+2.2587x_{42}^*+0.8736x_{43}^*-4.9533x_{44}^*+0.0490x_{45}^*+2.6606x_{46}^*$$

$$v_{41}=0.3219y_{41}^*+0.0166y_{42}^*+0.6662y_{43}^*$$

典型系数的大小可以反映出原变量与典型变量的关系，观测 X 组的典型变量系数，我们得到以下信息：“个人所得税宏观指标” 组中的主要因素是 x_{11}^*、x_{15}^*（典型系数为 2.5089，1.3572）；“企业所得税宏观指标”组的主要因素是 x_{21}^*、x_{24}^*、x_{26}^*、x_{28}^*；“营业税宏观指标”组中的主要因素是 x_{32}^*、x_{34}^*、x_{35}^*、x_{38}^*；“增值税宏观指标”组中主因素是 x_{42}^*、x_{43}^*、x_{46}^*。由此表明“个人所得税的宏观指标组”中影响“市辖区个人所得税”的主要是城市居民年人均可支配收入，其次是第二产业增加值；第二产业增加值、城市商品零售价格指数、规模以上国有及国有控股工业企业亏损面和建筑业企业利润总额对应因素对各地区企业所得税的影响较大；“营业税宏观指标”中影响“市辖区营业税”的主要是第三产业增加值、全社会房地产开发投资额、建筑业总产值以及住宿和餐饮业零售额；“增值税宏观指标”中影响“市辖区增值税”的主要是地区生产总值、工业增加值、批发零售业增加值。

考虑 Y 组的典型系数，可以得到以下信息：“市辖区个人所得税”的典型变量 v_{11} 与 y_{11}^* 呈高度相关（典型系数为 2.9914）；“市辖区企业所得税”的典型变量 v_{21} 与 y_{21}^* 呈高度相关（典型系数为 1.0316）；“市辖区营业税”的典型变量 v_{41} 与 y_{42}^* 呈高度相关（典型系数为 1.6110）；“市辖区增值税”的第一典型变量 v_{41} 与 y_{41}^* 呈高度相关（典型系数为 0.3219）。因此有以下结论：市区的个人所得税、企业所得税、增值税分别在全市的个人所得税中、全市的企业所得税和全市的增值税占主导地位；增城区的营业税在全市的营业税中占主导地位。

- **典型结构**

表 2 -X 组（宏观指标组）的典型结构

个人所得税宏观指标与典型变量的相关性			企业所得税宏观指标与其典型变量的相关性			营业税宏观指标与其典型变量之间的相关性			增值税宏观指标与典型变量之间的相关性		
	U11	**V21**		**U21**	**V21**		**U31**	**V31**		**U41**	**V41**
X11	0.9652	0.9391	**X21**	0.8867	0.8842	**X31**	0.9194	0.9168	**X41**	0.9747	0.9721
X12	0.9500	0.9244	**X22**	0.8690	0.8665	**X32**	0.9315	0.9288	**X42**	0.9943	0.9917
X13	0.9420	0.9166	**X23**	0.8491	0.8467	**X33**	0.9249	0.9223	**X43**	0.9967	0.9941
X14	0.9595	0.9336	**X24**	0.9325	0.9299	**X34**	0.9365	0.9338	**X44**	0.9859	0.9833
X15	0.9616	0.9356	**X25**	0.8259	0.8236	**X35**	0.9460	0.9433	**X45**	-0.5464	-0.5449
X16	0.8987	0.8744	**X26**	-0.7154	-0.7134	**X36**	0.9395	0.9368	**X46**	0.9684	0.9659
			X27	0.8297	0.8274	**X37**	0.9605	0.9578			
			X28	0.9032	0.9007	**X38**	0.9528	0.9501			
			X29	0.9257	0.9231	**X39**	0.8934	0.8908			

表 3- Y 组（不同市辖区税收组）的典型结构

各市辖区个人所得税与典型变量的相关性			各市辖区企业所得税与典型变量的相关性			各市辖区营业税与典型变量的相关性			各市辖区增值税与其典型变量的相关性		
	V11	**U21**		**V21**	**U21**		**V31**	**U31**		**V41**	**U41**
Y11	0.8321	0.8097	**Y11**	0.8867	0.8842	**Y11**	0.7874	0.7852	**Y11**	0.9917	0.9891
Y12	0.5315	0.5171	**Y12**	0.8690	0.8665	**Y12**	0.9537	0.9510	**Y12**	0.9658	0.9632
Y13	0.7379	0.7180	**Y13**	0.8491	0.8467	**Y13**	0.8834	0.8809	**Y13**	0.9978	0.9952

由表 2 知，$x_{11},x_{12},x_{14},x_{15}$与“个人所得税宏观指标”的典型变量$u_{11}$均高度相关，相关系数在 95%以上，且典型变量$u_{11}$和$v_{11}$高度相关性，说明城市居民年人均可支配收入、城镇单位职工年平均工资、地区生产总值、第二产业增加值是影响地区个人所得税收入的主要因素。 其中城市居民年人均可支配收入、第二产业增加值是最大影响因素。 $x_{21},x_{24},x_{28},x_{29}$与“企业所得税宏观指标”的变量$u_{21}$均高度相关，相关系数在 88%以上，且典型变量高度相关，因此可以认为第二产业增加值、城市商品零售价格指数、建筑业企业利润总额、限额以上连锁店（公司）零售额是影响各地区企业所得税收入的主要因素。$x_{34},x_{35},x_{36},x_{37},x_{38}$与“营业税宏观指标”变量$u_{31}$均高度相关，相关系数在 93%以上，表明第三产业增加值、全社会房地产开发投资额、全社会住宅投资额、建筑业总产值以及住宿和餐饮业零售额是影响地区营业税收入的主要因素。$x_{41},x_{42},x_{43},x_{44},x_{46}$与“增值税宏观影响因素组”的典型变量$u_{41}$均高度相关，相关系数在 96%以上，说明商品进口总值、地区生产总值、工业增加值占 GDP 以及批发零售业增加值是影响地区增值税收入的主要因素。

由表 3 知，通过分析典型结构并研究相关系数大小，我们得到以下信息：“市辖区个人所得税组”的典型变量v_{11}与y_{11}相关系数高达 0.8321；“市辖区企业所得税组”的典型变量v_{21}与y_{21}相关系数高达0.9980；“市辖区营业税组”的典型变量v_{31}与y_1相关系数高达 0.9537；“市辖区营业税组”的典型变量v_{41}与y_3相关系数高达0.9978。由此可得，市区的个人所得税、企业所得税分别在全市个人所得税收入和企业所得税中占主导地位；增城区营业税在全市营业税收入中占主导地位；从

化区增值税在全市增值税收入中占主导地位。

综上分析，可以得到以下影响财政收入的关键因素和反映一般财政收入的市辖区影响分析。

表 4-不同税组的关键因素

	个人所得税组	企业所得税组	营业税组	增值税组
X 组主要指标	城市居民年人均可支配收入	建筑业企业利润总额	全社会房地产开发投资额	地区生产总值
	第二产业增加值	限额以上连锁店(公司)零售额	全社会住宅投资额	商品进口总值
	地区生产总值	第二产业增加值	建筑业总产值以及住宿和餐饮业零售额	工业增加值占 GDP
	城镇单位职工年平均工资	城市商品零售价格指数	第三产业增加值	工业增加值
Y 组主导体现	市区的个人所得税	市区的企业所得税	增城区营业税	从化区增值税

- **冗余分析**

表 5-被典型变量解释的 X 组原始变量的方差

	被本组的典型变量解释			典型相关系数平方	对方组典型变量解释	
		比例	累计比例		比例	累计比例
个人所得税组	U11	0.9211	0.9211	0.9467	0.8270	0.8270
企业所得税组	U21	0.7621	0.7621	0.9944	0.7578	0.7578
营业税组	U31	0.8794	0.8794	0.9943	0.8744	0.8744
增值税组	U41	0.9860	0.9860	0.9948	0.9809	0.9809

表 6-被典型变量解释的 Y 组原始变量的方差

	被本组的典型变量解释			典型相关系数平方	对方组典型变量解释	
		比例	累计比例		比例	累计比例
个人所得税组	V11	0.6919	0.6919	0.9467	0.6550	0.6550
企业所得税组	V21	0.9955	0.9955	0.9944	0.9899	0.9899
营业税组	V31	0.6224	0.6224	0.9943	0.6189	0.6189
增值税组	V41	0.9834	0.9834	0.9948	0.9783	0.9783

可以看出，通过检验的第一对典型变量u_{11}、v_{11}较好的预测了对应的那组变量，而且交互解释能力也比较强。来自“地区个人所得税组”的方差被“个人所得税宏观影响因素组”的典型变量u_{11}解释的比例为 65.50%；来自“个人所得税宏观影响因素组”的方差被“地区个人所得税组”的典型变量v_{11}解释的比例和为 87.20%。两个变量组被其自身以及对立典型变量解释的百分比均较高，反映了

两者之间较高的相关性。

典型变量u_{21}、v_{21}较好的预测了对应的那组变量，而且交互解释能力也比较强。来自“地区企业所得税组”的方差被“企业所得税宏观影响因素组”的典型变量u_{21}解释的比例高达 98.99%；来自“个人所得税宏观影响因素组”的方差被“地区个人所得税组”的典型变量v_{21}解释的比例为 75.78%。两个变量组被其自身以及对立典型变量解释的百分比均较高，反映了两者之间较高的相关性。

典型变量u_{31}、v_{31}较好的预测了对应的那组变量，而且交互解释能力也比较强。来自“地区营业税组”的方差被“营业税宏观影响因素组”的典型变量u_{31}解释的比例为 61.89%；来自“个人所得税宏观影响因素组”的方差被“地区个人所得税组”的典型变量v_{31}解释的比例为 87.44%。两个变量组被其自身以及对立典型变量解释的百分比均较高，反映了两者之间较高的相关性。

典型变量u_{41}、v_{41}较好的预测了对应的那组变量，而且交互解释能力也比较强。来自“地区增值税组”的方差被“增值税宏观影响因素组”的典型变量u_{41}解释的比例高达 97.83%；来自“增值税宏观影响因素组”的方差被“地区增值税组”的典型变量v_{41}解释的比例也高达 98.09%。两个变量组被其自身以及对立典型变量解释的百分比均较高，反映了两者之间较高的相关性。
得到与个人所得税、企业所得税、营业税、增值税的主导宏观指标。

2.3.2 熵权系数模型

□ 运行结果(可见资料附件/问题一/熵权系数模型 EXCEL 结果)

表 7-熵权系数法公共预算收入的主要指标

	个人所得税组		企业所得税组		营业税组		增值税组	
排名	指标	熵权值	指标	熵权值	指标	熵权值	指标	熵权值
1	城市居民年人均可支配收入	0.2071	建筑业企业利润总额	0.2884	全社会房地产开发投资额	0.2014	地区生产总值	0.2499
2	第二产业增加值	0.2005	限额以上连锁店(公司)零售额	0.2574	建筑业总产值	0.1809	工业增加值	0.2203
3	地区生产总值	0.1876	建筑业总产值	0.1016	住宿和餐饮业零售额	0.1540	批发零售业增加值	0.2120
4	城镇单位职工年平均工资	0.1854	规模以上工业企业盈亏相抵后的利润总额	0.1004	第三产业增加值	0.1460	批发零售业零售额	0.1730

从表 4 和表 8 知，典型相关分析与商权系数模型具有相似的结果。

□ **检验结果**

一部分指标的效果预测图，其他的指标可通过运行程序得到（程序见资料附件/问题二/灰色预测/Que2_1_1—Que2_1_10）。

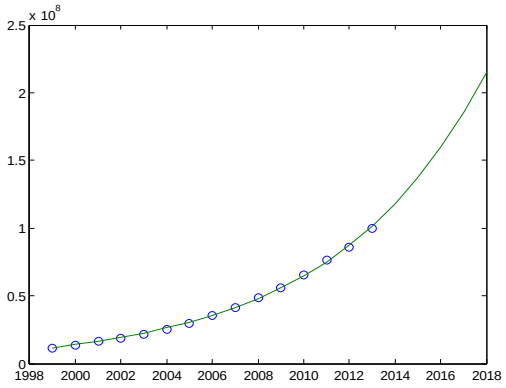


图 5-第三产业增加值预测效果

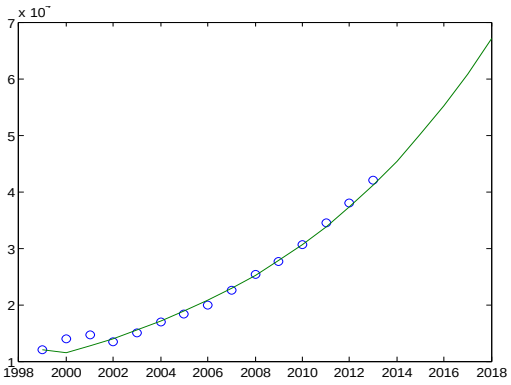


图 6-城市居民人均可支配收入预测效果

表 8 -2014 年财政收入主要关联指标的预测值

2014	城市居民年人均可支配收入 45375.085	第二产业增加值 63377150.99	城市商品零售价格指数（1978=100） 681.39	建筑业企业利润总额 1179928.179	第三产业增加值 118049333.4
2014	住宿和餐饮业零售额 9469432.733	全社会房地产开发投资额 17081117.41	地区生产总值 183601899.7	工业增加值 58163230.83	批发零售业增加值 24684195.46

2. 3. 3 主成分回归分析

通过 MATLBA 运行程序得到结果。

- **主成分分析结果**

我们选取 5 个主成分，其特征根及其贡献率见表 9。

表 9-主成分分析结果

序号	特征根	贡献率	累计贡献率	序号	特征根	贡献率	累计贡献率
1	9.7500	0.9750	0.9750	4	0.0195	0.002	0.9987
2	0.1339	0.134	0.9884	5	0.0078	0.0007	0.9994
3	0.0833	0.0083	0.9967	6	0.0781	0.0003	0.9997

主成分回归方程

$$I = -6101788.392360 - 104.660570a_1 + 0.088328a_2 + 11421.028053a_3 + 1.191784a_4 + -0.002039a_5 + 0.895880a_6 + -0.263509a_7 + 0.009532a_8 + 0.081285a_9 + 0.008857a_{10}$$

- 回归效果图

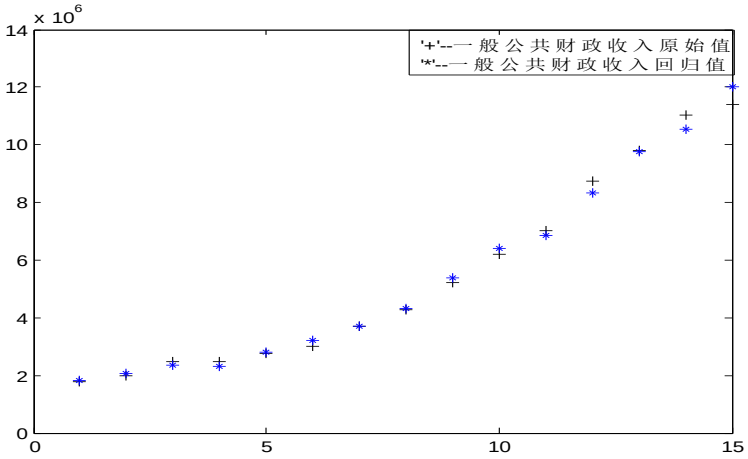


图 7-公共财政预算收入的回归图

用回归模型得到的公共财政预算收入的值与实际的公共预算收入值的比较，可以看出，我们的回归还是比较准确的，我们得到 2014 年和 2015 年公共预算收入(亿元)预测值：

$$\hat{I}_{2014}=1437.4\times,\hat{I}_{2015}=1643.4\text{。}$$

2.3.4 拟合模型

采用拟合模型预测政府性基金收入，我们利用多项式拟合 1993 年-2013 年政府性基金收入的离散数据，并延迟 2 期得到 2014 年和 2015 年得到政府性基金收入(亿元)：

$$\hat{T}_{2014}=985.3,\hat{T}_{2015}=1157.2$$

综合 2.3.3 中的公共预算收入预测值和上述拟合数值得到 2014 年、2015 年地方财政收入(亿元)的预测值：

$$\hat{Y}_{2014}=\hat{I}_{2014}+\hat{T}_{2014}=2422.7,$$

$$\hat{Y}_{2015}=\hat{I}_{2015}+\hat{T}_{2015}=2800.6$$

实际 2014 年地方财政收入和 2015 年的预算收入(亿元)为：

$$Y_{2014}=2453.9,$$

$$\hat{Y}_{2015}=2843.6$$

2.3.5 BP 神经网络预测模型

由于 BP 神经网络对于神经元个数大于 4 的运算比较不稳定，它的每一次运行结果相差很大，因此我们采用多次运行并取其均值的结果作为我们最后的预测值，预测的平均值会趋于一个稳定水平。经过运行 1000 次后并取其平均值，得到 $\hat{I}^{*}_{2014}=1.3642\times10^{8}$

3. 结论

3.1 结论

我们得出了影响政府财政收入的主要影响因素为：城市居民年人均可支配收入、第二产业增加值、城市商品零售价格指数、建筑企业利润总额、第三产业增加值、住宿和餐饮业零售额、全社会房地产开发投资额、地区生产总值、批发零售业增加值以及工业增加值。

城市居民年人均可支配收入是家庭扣除缴纳所得税以及其他费用的平均年收入，是居民用于日常生活的开支，其往往用来衡量居民生活水平。对于个人所得税现行使用划分级距的超额累计税率，近些年中国城市居民年人均可支配收入持续增加，个人所得税增大，公共财政收入也随之增加。

城市商品零售价格指数是反映城市零售商品价格变动趋势的一种经济指数，其与企业所得税关联比较大。零售价格指数调整变动直接影响到城乡居民的生活支出和国家的财政收入，影响居民购买力和消费与积累的比例。

中国经济正处于工业化的中期，主要由第二和第三产业带动国民生产总值。。第二产业主要指加工业，第三产品主要指服务业。第二产品增加值和第三产业增加值对财政收入的影响是多方面、深刻的。产业的增加的扩增带来就业增

加，据估计同期第二产业增加一个点带动就业人数 70 万，同期第二产业增加一个点带动就业人数 61 万；产业增加值的增加，财政收入项中的个人所得税和企业所得税、营业税、增值税都会相应扩增，由于此种经济现象还存在乘数效应，因此第二产增加值和第三产业增加值对财政收入的影响具有更加深刻的影响。

地产生产总值是衡量地区经济状况的指标，其往往能反映出一个地区经济的发展程度和居民生活水平，地区生产总值高，经济活动活跃，居民生活水平高。地区生产总值对财政收入的关联程度是双向的，其往往与财政收入保持相对稳定的增长趋势。地区生产总值增高，财政收入随之提高；财政收入的扩增，政府有更多的资金反馈到经济活动、民生工程、教育活动中，带动经济发展，拉动地区生产总值增长。

建筑企业利润总额是影响企业所得税的主导因素、住宿和餐饮业零售额是影响营业税的主导因素。全社会房地产开发投资额的增长，降低社会闲置资金再投资的积极性，引起地区生产总值持续增长，提高财政收入。

3.2 预测结果

表 10 为 2014 年-2015 年公共预算收入、政府基金收入、地方财政收入的灰色—主成分回归数值结果以及 BP 神经网络的公共预算收入的结果

表 10-不同模型的预测值和实际值

		灰色预测-主成分回归			BP 神经网络
		公共预算收入（亿元）	政府性基金收入（亿元）	地方性财政收入（亿元）	公共预算收入（亿元）
2014	预测	1437.4	985.3	2422.7	1369
	实际	1241.5	1074	2453.9	1241.5
2015	预测	1643.4	1157.2	2843.6	1496.6
	预算	1357.9	1094.1	2843.6	1357.9

由于财政管理体制改革，2014 年地方财政收入包括一般公共预算方面、政府性基金预算方面、国有资本经营预算方面、财政专户管理资金方面；而 1999-2013 年，地方财政收入只包括一般公共预算方面和政府性基金预算方面，其托有资本经营项和财政专户管理资金项是与一般公共预算方面合并，而我们的灰色主成分回归模型预测 2014 年和 2015 年是在考虑了国有资本经营预算项和财政专户管理资金项。因此我们预测的公共预算收入会较之于 2014 年单项公共预算的值要大，它们的差距刚好是国有资本金英预算和财政庄户管理基金项的和量纲和数量级别的。

3.3 建议

关于 2015 年广州市财政预算支出草稿报告一些分析和建议

表 11-近几年不同项目的财政收入

	2013 年市本级执行情况	2013 年各支出占总支出比例	2014 年市本级执行情况	2014 年各支出占总支出比例	2015 年市本级预算草稿	2015 年各支出占总支出比例
一般公共预算支出（亿元）	606.5	52%	571.1	42%	679	45%
政府性基金预算支出（亿元）	460.8	40%	683.1	51%	716.3	47%

国有资本经营预算支出（亿元）	51	4%	61.8	5%	91.6	6%
财政专户管理资金支出（亿元）	44.7	4%	32.4	2%	26.3	2%
市本级支出（总和）（亿元）	1163	100%	1348.4	100%	27.6	100%

从表 11 来看，各年的一般公共预算支出占市本级支出的为 50%左右，基本稳定；在政府性基金预算支出方面和国有资本经营预算支出也分别稳定在 40%~60%之间和 5%左右；而 2015 年的财政专户管理资金预算与去年一样。总的来说，2015 年各财政支出预算于往年基本持平，做出的各类预算支出是比较合理的。

根据地方财政收入=公共财政收入+政府性基金收入，我们从 1999~2013 年的地方财政收入数据表中可以得到税收收入约占公共财政收入的 80%，所以，在 2015 年的预算草案中，税收收入占据公共预算收入的 79.93%是合理的。近几年来，广州市税收收入从 2010 年的 6029663 万元到 2013 年 8525558 万元的突破，大大增强了广州市财政实力，增加了公共产品和服务的供给，促进经济发展。

税收作为宏观调控的重要工具，对经济运转产生调节作用。正是因为税收收入对广州市财政收入有着如此举足轻重的作用，税收征管监管工作就显得很重要。对此我们所提出的建议是：（1）税收征收部门应该提高税收征管执行力度。税务机关要加强税源监控，强化税收业务知识培训、转变工作作风、依法治税、依法征税，通过加强各部门配合，不断提高税收征收率，保持税收随经济的发展平稳增长。（2）地方财政“量出为入”表明，与政策目标相反，省级政府在财政收入方面存在较大自主权，这实际上体现了地方财政征收体制还存在一些不足，各项财政税收法规在实际执行中存在一些灰色区域，不够严格规范。财税征收体制不健全容易导致寻租和不公平。因此，财政征收体制需要进一步改革，加强税收纪律，实现财政收入规范化、公平税负。

对于一般公共预算支出方面，有几点问题，下面选取了公共财政收入的一些情况来分析：

- **民生及各项公共事业支出方面**

表 12-近几年民生及各项公共事业支出方面

	2013 年执行情况	2013 年各支出占一般公共财政支出比例	2014 年执行情况	2014 年各支出占一般公共财政支出比例	2015 年预算草稿	2015 年各支出占一般公共财政支出比例
教育（亿元）	122.2	20.1%	104	18.2%	83.3	12.2%
社会保障和就业（亿元）	93.4	15.4%	92.7	16.2	101.7	15.0%
医疗卫生与计划生育（亿元）	50.2	8.3%	53.1	9.3%	55.5	8.2%
交通运输（亿元）	149.8	24.7%	144.4	25.3%	44	6.5%
住房保障（亿元）	29.6	4.8%	37.7	6.6%	21.3	3.1%
“三农事业”（亿元）	70	11.5%	74.9	13.1%	73.4	10.8%

我们通过比较隔年各类支出占一般公共财政支出的比例不难发现，2015 年交通运输预算和住房保障预算分别比往年是 1/4 倍和 1/3 倍左右，说明广州市政府

对这两方面的投入支出是相比往年来说是非常小的。

政府紧急缩减交通运输方面的支出的话可能会导致公共交通方面秩序混乱，由于用于检查和维修公共交通运输工具的资金少了，可能会引发高频率的交通事故，对于交通运输方面的支出骤然减少太多是不合理的。

另外，2015 年广州将严控一般性财政支出，集中财力保民生、保重点及落实政府性债务偿还资金，突出对包括民生公共服务、支持“三农”、战略性主导产业发展、区域协调发展、宜居城市建设、科技创新和政府性债务偿还 7 个重点方面。同时，民生十事投入对比于 2014 年增加 25.9%。

12 条地铁线同时施工，可谓广州史上同时施工地铁线最多的一年，去年市本级财政投入地铁建设资金也达 75.4 亿元，这一投入今年将大幅降至 22 亿元。

2014 年财政预算执行情况及 2015 年预算草案的报告中特别说明，市本级财政安排 2015 年十件民生实事资金 54.5 亿元，较去年投入的 43.3 亿元增加 25.9%。今年投入的方面包括加强住房保障 29.2 亿元、提高垃圾处理能力 10 亿元、促进和稳定就业 5.5 亿元等。实现充分就业，促进经济快速增长。促进就业作为中央和地区经济工作会议的重要议题，其重要性可想而知。广州市政府相关部门应优化就业环境，建立公平竞争的劳动力市场，引进高端技术人才，开展各种类型的教育培训，提高各层次就业人员的职业素质和专业技能，加大投入发展职业教育，从而增加财政收入。

表 13-2014 年和 2015 年财政支出项目执行状况

2014 年执行状况	**（单位：亿元）**	**2015 年预算**
2453.9	全市财政总收入	2843.6
2655.6	全是财政总支出	2843.6
104	公共教育体系	126.2
92.7	社会保障和就业	115.1
53.1	公共卫生和医疗	64.2
144.4	公共交通建设	95
37.7	保障性住房及配套	
74.9	“三农”事业	66
40	战略性主导产业发展	40
16.4	扶贫开发和对口援建	
24.8	创新性城市建设	
206.6	偿债资金	198
49	维持政府行政运行支出	51.1
0.97	公务接待费	0.7
3.1	公务用车	3.2
0.68	因公出国（境）费	0.7
2.26	会议费	0.6
政府还债（单位：亿元）		
206.6	还债资金	198
66.6	地铁公司	28
39.7	水投集团	40.6
53.2	城投集团	98
43.2	交投集团	28
3.9	地方政府债券本息	3.4

我们查看了广州市财政局提交市人大审议的预算草案报告中提到“2014年本级财政收支矛盾十分突出，财力紧张”，今年这一现象将持续。不同的是，去年预算体现了“钱紧”这一特点，今年的主题则重点在“节约”。

创新是一项必须长期坚持的工作，2014年支持创新性城市建设24.8亿元，但在2015年预算草案中并没有提及对此项目的投入费用，我们认为这是不科学的，对此我们建议保持对创新性城市建设项目的支持。

- **三公经费较2014年下降了24%**

2015年广州财政预算的节约支出，主要盯紧政府行政开销。今年市本级财政用于维持政府行政运行支出51.1亿元，占一般公共预算支出总额7.5%。除公务用车购置及运行费预算增加500万元外，因公出国、公务接待和会议费均明显比上年下降。三公经费作为政府日常职能运转的必经过程，它的下降程度并不是越大越好，经费的下降可能会影响政府职能的正常运转，随着日后预算的细化及公开透明力度的加大，三公经费的下降空间会越来越有限，所以三公经费需要一个确定的标准便于分析对社会经济造成非积极的影响。

- **安排198亿元为国企还债**

备受关注的政府债，首次被列为市本级财政预算草案重点项目。从上表中可知，2014年广州市本级财政共投入206.6亿元偿还政府性债务本息(其中18亿元属提前偿还)，其中地铁公司最高，达66.6亿元。而2015年的预算草案中突出“保障政府性债务偿还资金”，市本级财政安排偿债资金198亿元，比去年减少8.6亿元。各类还款主体所得还债资金也跟去年有较大差异，今年城投集团所得还债资金最高，达98亿元。

自2007年起，我国将宏观调控目标主要放在节能减排和污染治理方面，达到促进能源的节约和有效利用，这也有利于长期的投资结构优化和经济增长方式的转变。报告中突出支出宜居城市建设正式这一目标的具体体现。

结合广州市当前经济情况，对广州市财政局提出如下相关政策建议：

（1）提高经济发展水平，以经济增长带动财政收入。无论是长期还是短期，经济发展水平对财政收入的增加都具有显著影响，因此广州市政府应该大力发展生产力，才能通过增加广州市生产总值，为财政收入增长提供财源。

（2）优化产业结构，构建现代产业体系。产业结构优化升级与经济的持续发展具有强相关性，加快结构调整步伐，坚持产业高端发展，全力建设现代产业体系。广州市加快第三产业发展，尤其是现代物流、金融保险、商贸会展和文化旅游等税收量高的现代服务业蓬勃发展，积极培育新的财源增长点。

4. 参考文献

[1]司守奎,孙玺菁. 数学建模算法与应用[M] 国防工业出版社，2014.

[2]李春林,陈旭红. 应用多元统计分析[M] 清华大学出版社，2013

[3]汤瑞凉,郭存芝,董晓娟. 灌溉水资源优化调配的熵权系数模型研究[J]. 河海大学学报,2000, 28(1):15~ 21.

[4] 杨树佳,王爱萍,郑新奇. 基本农田指标分解的熵权系数法研究[J]. 资源开发与市场,2006,22(4) :305~ 306.

[5] 何晓群.多元统计分析（第三版）[M].北京:中国人民大学出版社.2011.

[6]王少平. 宏观计量经济学研究现状与展望[J]. 学动态经济,2003(9),52-56

[7] 吕宁 地方财政一般预算收入预测模型研究[D] (2006)

[8] 关于广州市 2013 年预算执行情况和 2014 年预算草案的报告[R] 广州人大网 (2014.1.31)
[9]王荣. 地方政府财政收入适度规模研究[D]. (2007.11)
[10]王宁. 灰色组合预测模型研究[D]. (2009.3)
[11]许为夷. 我国财政收入增长的影响因素及对策分析[D]. (2007.10)
[12]袁飞.任立良.姜红梅.季成康. MATLAB 神经网络工具箱在径流量模拟中的应用[J]. 人民长江期刊.
[13]金菊良，基于神经网络的年径流量预测模型[J]人民长江期刊(1999. 10)

附录

1 程序运行环境
本文的解决主要在MATLAB和SAS中完成代码的编写和编译，计算机操作系统：64 位 Win7 操作系统。软件环境：MATLAB R2014b；SAS 9.3。

2.附件资料有两个问题以及所有的模型的数据，其中，附件文件“问题一”中有用于典型相关分析的 SAS 程序（Que1.sas）,运行结果时需将源程序放在 D 盘中。“问题一”还有熵权系数模型的权重系数，各个表格代表不同税种的求解熵权系数结果。文件“问题二”是灰色预测模型、主成分回归以及 BP 神经网络预测模型的程序包；运行 MATLAB 时，需将相关 EXCEL 表格放在 MATLAB 的当前目录中。