第三届“泰迪杯”

全国大学生数据挖掘竞赛

优

秀

作

品

“泰迪杯”优秀作品

作品名称：基于电商平台家电设备的消费者评论数据挖掘分析

荣获奖项：二等奖

作品单位：华南师范大学

作品成员：刘天虹　吴慧敏　邱舒琪

指导教师：

基于数据挖掘技术的市财政收入分析预测模型

摘　要：本文以广州市财政收入为研究对象，主要分成两部分，第一部分是分析和识别影响财政收入的关键影响因素，第二部分是对广州市 2015 年财政总收入及各类别收入进行预测。第一部分中，将财政收入组成成分分成四大类，并找出各自的初始影响因素。然后汇总各类初始影响因素，对财政收入做回归筛选出九个主要影响因素，再对这九个主要影响因素进行主成分分析，得到三个主成分。第二部分中，对财政收入的各个项目建立 ARMA 模型，据此预测 2014 年、2015 年的收入。最后收集 2014 年的数据与其预测值作比较，计算预测误差，发现预测模型是有效的。整个挖掘过程主要通过 R、SAS、SPSS 软件实现。

关键词：财政收入　聚类分析　主成分分析　应用回归　时间序列预测　SAS

“泰迪杯”优秀作品

Analysis and Forecast Model of Financial Revenue Based on Data Mining Technology

Abstract: The thesis based on the datas of financial revenue of Guangzhou government is divided into two parts.The first part is to analyse and identify key elements that influence financial revenue.The second one is to forcast the sum and every single component of the year 2015. In the first part,we separate financial income into four categories to find out their own equations,after which it comes to remain nine main factors according to the results of applied regression.Then Principal Component Analysis is used to summarize out three principal components.Besides,we establish several ARMA models to predict each revenue item of 2014 and 2015 in the second part.Finally,comparisons are made between the predicted and the true values to find out the validity of the models.R,SAS and SPSS softwares are utilized through the whole process of data mining.

Key words: financial revenue Cluster Analysis Principal Component Analysis Applied Regression Prediction of Time Series SAS

“泰迪杯”优秀作品

目 录

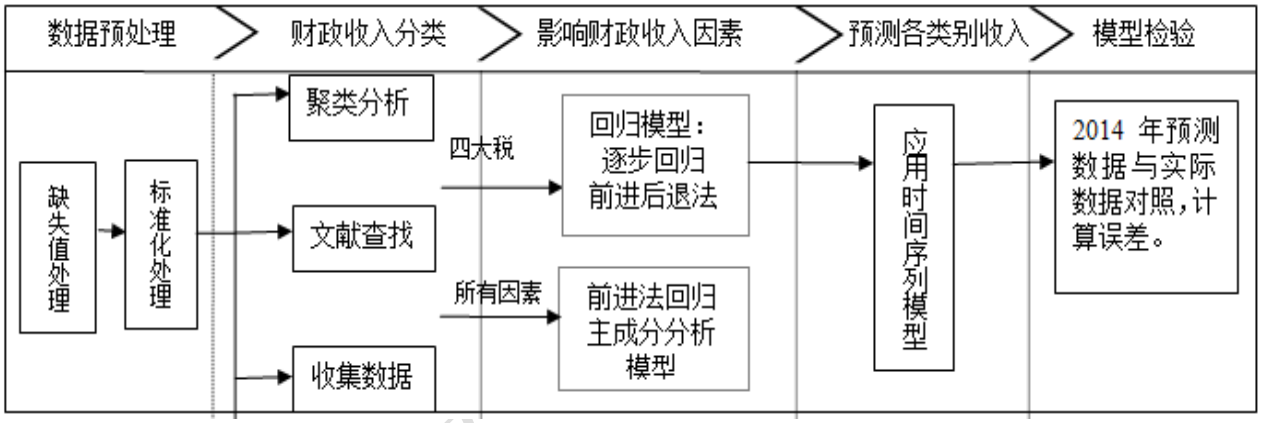
“泰迪杯”优秀作品

1. 挖掘目标

本次挖掘目标是利用广州市在 1999-2013 年财政收入、经济、教育等方面的真实数据，采用数据挖掘技术，分析、识别影响财政收入的关键因素，构建广州市财政收入分类体系以及财政收入预测模型，实现对广州市 2015 年的财政总收入及各个类别收入的预测，从经济因素与非经济因素出发，在财政收入以及支出预算方面向广州财政局提出建议，帮助其做出下一年有效的财政收入预算，为下一年的政策提供指导依据。

2. 分析方法与过程

2.1. 总体流程



【步骤一】对源数据进行缺失值补充及标准化；

【步骤二】运用聚类分析的思想，将财政收入各个组成成分分成四大类：三大经济税（包括增值税、营业税和企业所得税）、对与社会生产有关的对象所征收税、税外收入和政府性基金；

【步骤三】通过文献查找，定性分析影响对与社会生产有关的对象所征收税、税外收入和政府性基金的其他潜在因素（附件提供的因素除外），并到广州市统计信息网等网站搜集相关数据；

【步骤四】尝试运用前进法、后退法、逐步回归法筛选影响增值税、营业税、企业所得税、个人所得税四大税的关键因素；

【步骤五】汇总财政收入的十九个影响因素，运用前进法回归，得到影响总财政收入的九个主要因素，将这九个主要因素建立主成分分析模型并检验；

【步骤六】运用时间序列模型分别对广州市 1999-2013 年各个项目建立相应的 ARMA 模型，并据此对 2014、2015 年财政收入的各个项目收入进行预测；

【步骤七】将用时间序列模型预测得到的广州市 2014 年财政收入的各项目收入数据与实际数据对比，计算预测误差，对模型可行性进行检验。

2.2. 具体步骤

2.2.1 数据预处理

（1）缺失值处理

缺失值处理是指对样本由于各种原因导致的数据缺失进行的一种补救，方法包括删除法、随机插补、均值插补、回归法等。在我们得到的数据中，由于营业税中批发零售业增加值与批发零售业零售额所反映的对象相同，都是体现批发零售业的发展，而且批发零售业增加值有 5 个缺失值，所以我们决定剔除批发零售业增加值样本。

另外，针对样本中的其他缺失值，为了避免样本的进一步减少和信息的丢失，我们采用填补的方法进行补救。由于各个变量的取值不是随机数据，且变量之间具有一定的相关关系，所以，拟合含有缺失值的变量和其相关变量之间的一个回归会是一个比较合理和有效的补救方法。将有缺失值的变量作为因变量，相关变量作为自变量建立回归方程，最终以得到的预测值作为我们所需要的填补值。

根据以上分析，利用 R 软件，我们对企业所得税中的规模以上国有及国有控股工业企业企业亏损面、建筑业企业利润总额和限额以上连锁店(公司)零售额以及个人所得税中的城镇非私营单位从业人员的缺失值进行填补。结果如表 1 和表 2 所示（加粗部分为填补值）：

表 1 部分企业所得税表

规模以上国有及国有控股工业企业企业亏损面	建筑业总产值	建筑业企业利润总额	限额以上连锁店(公司)零售额
32	2470523	**91780**	**1316593**
31	2561326	51918	**2219297**
31.25	3403870	144803	**7059236**
31.99	3733922	80922	1053156
29.87	4785787	167217	1154425
30.69	5459314	154958	1434440
31.63	6331382	186678	3621757
28.95	6870406	219390	4196301
24.88	7507109	376839	7068265
30.85	8754491	458096	17829885
23.16	10134050	485760	17019222
20.42	12805288	653736	26192835
22.55	15613171	668043	21639131
20.9	17417072	703733	21396742
19.7	21828895	877889	22659148

表 2 部分个人所得税表

第二产业增加值	城镇非私营单位从业人员数	地方财政收入	个人所得税
9310691	**1873488**	1881388	133621
10216241	1755512	2199077	185625
11122943	1997579	2719058	254892

12113416	2071574	2690984	159684
14859261	2236902	3005475	153080
17880638	2255380	3384477	167379
20452183	2351538	4088545	198017
24415160	2039164	4767231	231794
28257805	2190420	8389925	295316
32278717	2205292	8431405	353372
34051588	2284029	11076649	389824
40022658	2387940	13991612	472154
45769763	3102356	15351387	462098
47206504	3268488	15796804	439592
52273431	3245858	20881374	489777

（2）标准化处理

因为不同的数据之间具有不同的量纲，这会影响模型的建立和模型的精确度。所以，在建立回归模型之前我们需要对各大税影响因素的样本数据进行均值为 0，方差为 1 的标准化处理，以达到消除量纲的目的。

2.2.2 对财政收入的组成成分进行分类

为进一步分析和研究财政收入的组成成分和影响因素，我们对财政收入的各个组成成分（包括营业税、增值税、企业所得税、个人所得税等 16 个组成成分）进行分类。运用聚类分析样本距离分类的思想和 SAS 聚类模块，得到如下聚类结果：

Cluster History

NCL	--Clusters Joined--		FREQ	SPRSQ	RSQ	ERSQ	CCC	PSF	PST2	Flex Dist
15	OB9	OB15	2	0.0001	1.00	.	.	476	.	0.2598
14	OB7	OB8	2	0.0005	.999	.	.	223	.	0.4929
13	CL14	CL15	4	0.0008	.999	.	.	169	2.3	0.5828
12	OB6	OB10	2	0.0009	.998	.	.	153	.	0.636
11	OB13	OB14	2	0.0009	.997	.	.	150	.	0.6512
10	OB5	CL12	3	0.0011	.996	.	.	150	1.2	0.7028
9	CL10	OB12	4	0.0025	.993	.	.	125	2.5	1.0689
8	CL13	OB11	5	0.0035	.990	.	.	108	7.1	1.3421
7	CL9	CL11	6	0.0055	.984	.	.	92.9	4.0	1.4706
6	OB4	CL7	7	0.0049	.979	.	.	94.3	2.2	1.5583
5	OB1	OB2	2	0.0102	.969	.	.	86.1	.	2.1404
4	CL6	CL8	12	0.0515	.918	.	.	44.5	24.8	4.6695
3	CL5	OB3	3	0.0574	.860	.774	2.62	40.0	5.6	5.0629
2	CL3	OB16	4	0.2350	.625	.608	0.23	23.4	7.0	10.648
1	CL2	CL4	16	0.6252	.000	.000	0.00	.	23.4	18.644

图 1　聚类分析历史图

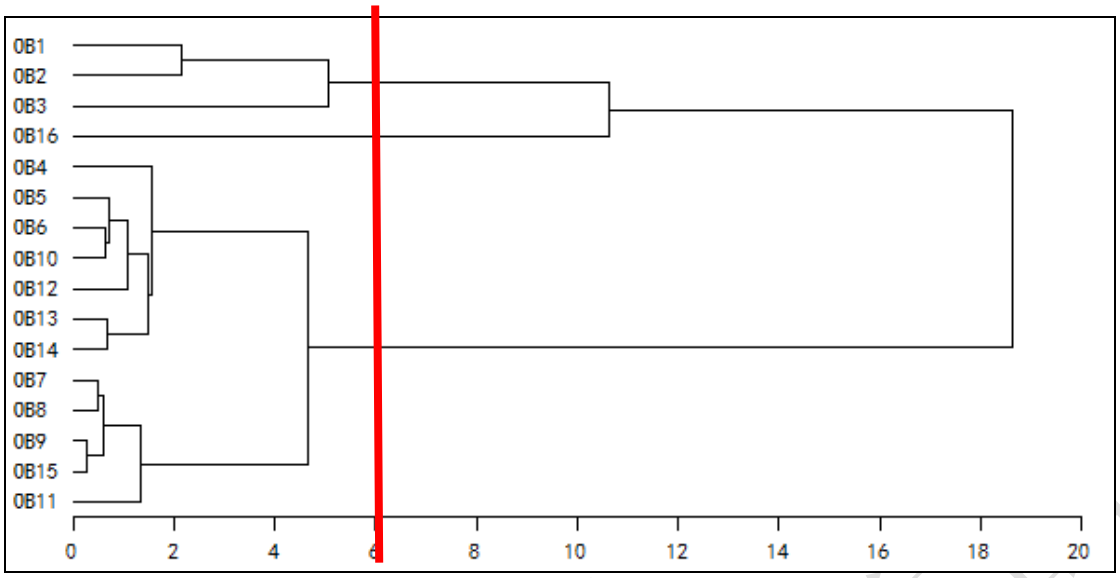


图 2　聚类分析树状图

由聚类的历史图看到，当分类数为 3 时，CCC 达到峰值，所以认为将财政收入的组成成分分为三类为宜。即第一类：增值税、营业税、企业所得税；第二类：政府性基金收入；第三类：其他收入。又因为第三类中的其他收入包含税收和非税收，为了更好地解释其经济意义，我们把第三类其他收入再进行细分，分为对与社会生产有关的对象所征收税和税外收入，所以，最终分类结果如下：

第一类：增值税、营业税、企业所得税

第二类：政府性基金收入

第三类：对与社会生产有关的对象所征收税（个人所得税、城市建设维护税、房产税、印花税、城镇土地使用税、车船使用税、契税）

第四类：税外收入（国有企业计划亏损补贴、行政性收费收入、罚没收入、专项收入、其他收入）

其中，第一类反映实体经济的主要收入；第二类反映国家通过向社会征收以及出让土地、发行彩票等方式取得收入，表示国有资源的经营收入。在社会主义市场经济下，政府性基金收入大部分需要在全社会范围内进行统一筹集和使用，市财政分配的主要对象；第三类反映的是在社会建设、拥有个人资产等时候对个人征收的各种税；第四类反映的是税外收入，是在一般生产经营活动中征收的其他收入，具有较大的灵活性和相对不稳定性。

2.2.3 定性分析和寻找影响财政收入组成类别的潜在因素

财政收入的组成成分众多而复杂，要研究影响财政收入的影响因素，就要从财政收入的各个组成成分入手。根据上面聚类分析的结果，我们对这四类组成成分进行定性分析，以寻找和挖掘出影响各类财政收入的因素（第一类反映实体经济的税收的影响因素已经给出，见附件）。

第二类政府性基金预算收入是为实现特定经济社会领域的政策目的，各级人民政府及其所属部门按照规定程序批准，依法向特定群体无偿征收的具有专项用途的一种非税收入，主要有出让土地、发行彩票等方式，政府性基金种类繁多，与一般税、特殊类型税、规费、受益费等有着明显区别，其基本特征表现为特别政策性、被课征群体特定性、特殊的法律关联性、非对待给付性和专款专用性。

正是由于这些特性，使得每年政府性基金收入的涨跌基本与当时的政策干预有关，并且大多数资料显示，政府性基金收入很大程度上与土地出让收入直接相关，而这又与当地政府出台的有关土地出让方面的政策相联系。因此，政府性基金收入基本与当地相关政策和社会整体情况有关。

第三类财政收入是对与社会生产有关的对象所征收税，包括个人所得税、城市建设维护税、房产税等。其中，房产税、契税主要与房屋买卖、房屋交换等行为有关，故反映房地产发展、商品房销售情况的因素均会对它们产生影响，如商品房销售面积、商品房销售合同金额等。城市维护建设税是指为加强城市的维护建设，扩大和稳定城市维护建设所征收的一种税，其纳税人是有经营收入的单位和个人，故与经济有关的因素、就业人数、教育水平以及人口数都会对其产生影响。土地使用税是指对土地使用权征收的一种税，它和车船使用税都在一定程度上反映了社会固定资产的总量。而印花税则关系到经济、生产活动的方方面面，居民消费水平是衡量经济发展的重要指标，故它会对财政收入产生一定的影响。最后，政府财政支出是与财政收入相辅相成的因素，正是因为政府要为实现一定职能而投出资金，才有了财政收入的必要。

第四类财政收入是指税外收入。企业计划亏损补贴主要是指国家为了使国有企业能够按照国家计划生产、经营一些社会需要，但由于客观原因使生产经营出现亏损的产品，而向这些企业拨付的财政补贴。而专项收入，是指根据特定需要设置有专门用途的收入。行政性收费是指国家机关、司法机关和法律、法规授权的机构，依据国家法律法规、相关规定行使其管理职能，向公民、法人和其他组织收取的费用。罚没收入是指执法、司法机关依照法律规定，对违法违章者实施经济的罚款收入。这些收入均属于税外收入，是对税收的补充收入，具有较大的灵活性和相对不稳定性，所以影响因素主要由经济发展水平和政策变动，即地区生产总值和政策干预。

根据上面分析，我们便可以确定影响财政收入组成成分（增值税、营业税、企业所得税、个人所得税除外）的其他潜在因素（数值型变量），通过广州市统计信息网等网站搜集相关数据，得到数据汇总见表 3：

表 3　财政收入第二三四类组成成分的影响因素**错误！链接无效。2.2.4　影响增值税、营业税、企业所得税、个人所得税四大税的关键因素**

根据附件二提供的数据，我们尝试运用前进法、后退法和逐步回归法进行筛选变量，建立影响四大税的因素模型。

◆增值税

表 4　增值税原始影响因素

指标变量	变量名称
x_1	商品进口总值
x_2	地区生产总值
x_3	工业增加值
x_4	批发零售业零售额
x_5	工业增加值占 GDP
Y	增值税

（1）数据预处理

1）剔除重复变量

由于营业税表格中批发零售业增加值与批发零售业零售额所反映的指标相同，都是体现批发零售业的发展，并且批发零售业增加值有 5 个缺失值，所以决定剔除批发零售业增加值样本。

2）数据标准化

对x_1-商品进口总值、x_2-地区生产总值、x_3-工业增加值、x_4-批发零售业零售额、x_5-工业增加值占 GDP、Y-增值税数据进行均值为 0，方差为 1 的标准化。

（2）模型建立与检验

由经济背景知识得到，增值税和城商品进口总值等五个变量存在着显著的相关关系，所以五个变量之间可能存在着严重的多重共线性，我们尝试建立回归模型。

表 5　模型参数表

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-2.4199E-16	0.03482	8.78366E-31	0.00	1.0000
x1	-0.51521	0.20765	0.11197	6.16	0.0289
x3	1.49553	0.20765	0.94347	51.87	<.0001

表 6　方差分析表

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	13.78174	6.89087	378.86	<.0001
Error	12	0.21826	0.01819		
Corrected Total	14	14.00000			

得到模型如下：

$$Y=-0.5152X_1+1.4956X_3$$

（6.16）　（51.87）　$R^2=0.9844$

由方差分析表中得到，模型通过 F 检验，拒绝原假设，认为模型总体是显著的。由模型参数表得知变量X_1，X_3系数均通过 t 检验，拒绝原假设，参数均显著。VIF 小于 10，说明变量之间不存在着多重共线性。所以，回归方程成立。

（3）对回归结果进行理论分析

在影响增值税的因素中，由线性模型结果得到，商品进口总值、工业增加值可以作为代表因素。增值税作为一个流转税，其水平受进出口额影响较大，为鼓励出口，我们国家实行出口退税优惠政策，所以商品进口总值是影响增值税的一个重要因素。而工业的发展促进商品生产、流通、劳务服务等的发展，这其中需要对新增价值或商品附加值征收增值税，所以工业增加值是另外一个重要因素。

◆营业税

表 7　营业税原始影响因素

指标变量	变量名称
x_1	公路货运量
x_2	公路客运量
x_3	建筑业增加值
x_4	第三产业增加值
x_5	全社会房地产开发投资额
x_6	全社会住宅投资额
x_7	建筑业总产值
x_8	住宿和餐饮业零售额
x_9	限额以上餐饮业主营业务收入
Y	营业税

（1）数据预处理

缺失值处理和标准化处理：见步骤一

（2）模型建立与检验

为研究营业税与公路运货量、公路客运量等九个变量之间的关系，找出影响营业税的关键因素，我们建立回归模型加以探究。

表 8　模型参数表

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	9.46533E-17	0.06085	1.34389E-31	0.00	1.0000
x6	-0.88758	0.23748	0.77574	13.97	0.0028
x9	1.80290	0.23748	3.20066	57.64	<.0001

表 9　方差分析表

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	13.33362	6.66681	120.05	<.0001
Error	12	0.66638	0.05553		
Corrected Total	14	14.00000			

得到模型如下：

$$y = -0.8876x_6 + 1.8029x_9$$

(13.97)　(57.64)　R^2=9524

由方差分析表中得到，模型通过 F 检验，拒绝原假设，认为模型总体是显著的。由模型参数表得知变量 X_6，X_9 系数均通过 t 检验，拒绝原假设，参数均显著。所以，回归方程显著成立。

（3）对回归结果进行理论分析

在众多影响营业税的因素中，通过逐步回归筛选得出了两个比较重要的因素，即全社会住宅投资额、限额以上餐饮业主营业务收入。全社会住宅投资额越大，房地产业发展越快，对营业税起到

一定的拉动作用。虽然限额以上餐饮业主营业务收入的系数为负数，但是全社会住宅投资额、限额以上餐饮业主营业务收入的系数相加为正，而房产业和餐饮业都属于第三产业，可以认为第三产业的增长使得营业税也呈现上升趋势。

◆企业所得税

表 10　企业所得税原始影响因素

指标变量	变量名称
x_1	第二产业增加值
x_2	第三产业增加值
x_3	全社会固定资产投资额
x_4	城市商品零售价格指数（1978=100）
x_5	规模以上工业企业盈亏相抵后的利润总额
x_6	规模以上国有及国有控股工业企业企业亏损面
x_7	建筑业总产值
x_8	建筑业企业利润总额
x_9	限额以上连锁店(公司)零售额
Y	企业所得税

（1）数据预处理

缺失值处理和标准化处理：见步骤一

（2）模型建立与检验

同理，通过建立逐步回归模型，研究 X_1-X_9 对企业所得税的影响并筛选关键因素。

表 11　模型参数表

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	3.35196E-16	0.03713	1.68535E-30	0.00	1.0000
x4	0.27989	0.10390	0.15006	7.26	0.0209
x8	0.47043	0.13987	0.23395	11.31	0.0063
x9	0.26563	0.11908	0.10291	4.98	0.0475

表 12　方差分析表

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	13.77252	4.59084	221.99	<.0001
Error	11	0.22748	0.02068		
Corrected Total	14	14.00000			

得到回归方程：

$$y = 0.2799x_4 + 0.4704x_8 + 0.2656x_9$$

（7.26）　（11.31）　（4.98）　R^2=0.9838

由此可见，方程总体和各系数均通过显著性检验，故该方程显著成立。

（3）对回归结果进行理论分析

由逐步回归结果可以看出，在众多影响企业所得税的因素中，通过筛选得出了三个有代表性的因素，即城市商品零售价格指数、建筑业企业利润总额和限额以上连锁店（公司）零售额。其他变量和这三个变量均呈高度相关关系，故将它们从模型中剔除。商品零售价格指数是反映一定时期内商品零售价格变动趋势和变动程度的相对数，它与国家宏观调控和国民经济息息相关，所以它对企业所得有着重要影响。而企业所得税是对我国内资企业和经营单位的生产经营所得和其他所得征收的一种税，建筑业企业和限额以上连锁店（公司）又是其中的重要组成部分，因此，它们的利润总额和零售额直接影响着企业所得税。

◆个人所得税

表 13　个人所得税原始影响因素

指标变量	变量名称
x_1	城市居民年人均可支配收入
x_2	城镇单位职工年平均工资
x_3	城镇居民储蓄存款余额
x_4	地区生产总值
x_5	第二产业增加值
x_6	城镇非私营单位从业人员数
Y	个人所得税

（1）数据预处理

缺失值处理和标准化处理：见步骤一

（2）模型建立与检验

同理，通过建立逐步回归模型，我们研究 X_1-X_6 对个人所得税的影响并筛选关键因素。

表 14　模型参数表

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	3.06055E-16	0.06948	1.40505E-30	0.00	1.0000
x1	1.36517	0.16623	4.88473	67.45	<.0001
x6	-0.46329	0.16623	0.56257	7.77	0.0164

表 15　方差分析表

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	13.13094	6.56547	90.66	<.0001
Error	12	0.86906	0.07242		
Corrected Total	14	14.00000			

得到模型如下：

$$Y=1.3651X_1-0.4633X_6$$

（67.45）　（7.77）　$R^2=0.9379$

由方差分析表中得到，模型通过 F 检验，拒绝原假设，认为模型总体是显著的。由模型参数表

得到，变量X_1，X_6的系数均通过 t 检验，拒绝原假设，参数均显著。综上，回归模型是有效的。

（3）对回归结果进行理论分析

在影响个人所得税的因素中，由线性模型结果得到，城市居民年人均可支配收入和城镇非私营单位从业人员数可以作为代表因素。可支配收入与就业水平是衡量社会福利水平高低的标准，经济发展，就业率提高，劳动者的可支配收入增加，从而影响经济发展，对个人所得税收产生重大的影响。

综上，得到影响增值税、营业税、企业所得税、个人所得税四大税种的影响因素汇总如下：

表 16　四大税种的影响因素汇总

税种	增值税	营业税	企业所得税	个人所得税
影响因素	商品进口总值	限额以上餐饮业主营业务收入	城市商品零售价格指数	城市居民年人均可支配收入
	工业增加值	全社会住宅投资额	建筑业企业利润总额	城镇非私营单位从业人员数
			限额以上连锁店（公司）零售额	

2.2.5　建立影响财政收入总模型

（1）汇总影响因素

根据 2.2.3 和 2.2.4 得到影响财政收入的 19 个因素，运用前进法回归进行筛选和剔除，最终得到影响总财政收入的九个代表因素，见表 17。

表 17　财政收入影响因素汇总

	组成类别	各类别影响因素（影响总财政收入的 19 个因素）	回归后得到 9 个代表因素
财政收入四个组成类别	第一类：增值税、营业税、企业所得税	商品进口总值	商品房销售合同金额 全社会固定资产投资额 普通高等院校毕业生数 政府财政支出 限额以上餐饮业主营业务收入 城市居民年人均可支配收入 工业增加值 限额以上连锁店(公司)零售额 建筑业企业利润总额
		工业增加值	
		限额以上餐饮业主营业务收入	
		全社会住宅投资额	
		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	
	第二类：政府性基金收入	教育、经济水平 政策影响	
	第三类：对与社会生产有关的对象所征收税	商品房销售面积	
		商品房销售合同金额	
		房屋施工面积	
		全社会固定资产投资额	
		普通高等院校毕业生数	
		普通高等学校学校数	
		地区生产总值	
		政府财政支出	

		总人口	
		城镇居民消费水平	
	第四类：税外收入	地区生产总值	
		政策变动	

（2）建立主成分分析模型

由于影响财政收的各个变量之间存在很大的相关性，信息重叠严重，所以，我们采用主成分分析的方法进行降低维数，用较少的几个综合变量来代替原来较多的变量，使这些独立的、互不相关的综合变量尽可能地代表原来的信息量。主成分是原变量的线性组合，是对原来变量信息的一种提取，既不增加总信息量，也不减少总信息量，只是对原信息进行了重新分配，它解决了原始数据的相关性问题，实现数据的降维。借助 SPSS 软件，得到以下结果：

表18 解释的总方差

成份	初始特征值			提取平方和载入		
	合计	方差的 %	累积 %	合计	方差的 %	累积 %
1	8.688	96.529	96.529	8.688	96.529	96.529
2	0.155	1.724	98.254	0.155	1.724	98.254
3	0.085	0.946	99.200	0.085	0.946	99.200
4	0.047	0.526	99.725	0.047	0.526	99.725
5	0.011	0.123	99.848	0.011	0.123	99.848
6	0.007	0.075	99.923			
7	0.004	0.042	99.965			
8	0.003	0.029	99.994			
9	0.001	0.006	100.000			

提取方法：主成份分析。

表19 成份得分系数矩阵

	成份				
	F1	F2	F3	F4	F5
商品房销售合同金额	0.112	-0.974	1.232	3.186	0.545
全社会固定资产投资额	0.114	-0.312	0.762	-0.525	5.344
普通高等院校毕业生数	0.113	-0.037	-2.346	1.377	0.236
政府财政支出	0.114	-0.240	1.182	-1.866	-3.238
限额以上餐饮业主营业务收入	0.114	-0.307	-1.012	-1.800	0.648
城市居民年人均可支配收入	0.115	-0.296	0.177	-1.243	3.194
工业增加值	0.114	-0.214	-1.049	0.109	-2.473
限额以上连锁店零售额	0.108	2.248	0.446	0.714	1.508
建筑业企业利润总额	0.115	0.238	0.624	0.171	-5.668

提取方法 :主成份，构成得分。

从解释的总方差表可以知道，当我们提取前五个主成分时，累积贡献率达到 99.8%以上，即前五个主成分几乎包含了样本的全部信息，其中第一主成分包含了 96.529%的信息。

从成份得分系数矩阵可以看到，第一主成分前面的系数均在 0.11 左右，各因素所占比例相等，故可认为 F1 为影响财政收入**综合指标**；第二主成分中限额以上连锁店(公司)零售额前面系数明显大于其他变量，连锁店主要以服务业为主，而第三产业又以服务业为代表，故 F2 代表**第三产业影响指标**；第三主成分中普通高等院校毕业生数系数明显大于其他变量，且为负数。通过分析，普通高等院校毕业生数越大多，该地区教育程度越高，一个地区的教育和经济又有着相互促进的作用，故普通高等院校毕业生数越多，地方财政收入越大。因此，F3 为负向指标，代表该地区**教育影响指标**；第四主成分中商品房销售合同金额占绝大部分比例，因此可以认为 F4 为**房地产影响指标**；第五主成分全社会固定资产投资额占较大比重，故可认为 F5 是**社会固定资产影响指标**。

综上所述，主成分分析概括了所有影响财政收入的因素。从宏观上来讲，影响财政收入的因素是众多而复杂的，一个地区的经济发展水平、教育程度、就业状况、房地产发展以及社会固定资产都是影响地方财政收入的重要原因。主成分分析的结果也从侧面验证了我们通过定性分析找到的九个影响财政收入的因素（商品房销售合同金额、全社会固定资产投资额、普通高等院校毕业生数、政府财政支出、限额以上餐饮业主营业务收入、城市居民年人均可支配收入、工业增加值、限额以上连锁店零售额和建筑业企业利润总额，相关定性分析见 2.2.3 和 2.2.4）的全面性和准确性。

（3）主成分分析模型的检验

1）首先对主成分分析适用性进行检验，即检验各个变量之间是否有将强的线性相关关系。如果原始变量之间的线性相关关系程度很小，它们之间不存在简化的数据结构，进行主成分分析是没有意义的。通过相关系数矩阵我们可以看出，各变量之间确实存在显著的相关性。另外，从SPSS输出的KMO检验和巴特莱特球形检验的结果中也可以看出，$KMO = 0.79 \approx 0.8$，$P = 0.000$,说明这些变量是适合做主成分分析且效果是不错的。

表20 KMO 和 Bartlett 的检验

取样足够度的 Kaiser-Meyer-Olkin度量		0.790
Bartlett 的球形度检验	近似卡方	343.115
	df	36
	Sig.	0.000

表21 KMO检验标准

适用于主成分分析的程度	KMO 取值范围
非常适合	0.9<KMO
适合	0.9<KMO<0.8
一般	0.8<KMO<0.7
不太适合	0.7<KMO<0.6
不适合	KMO<0.6

2）其次进行主成分方差的检验，即检验相关矩阵的特征值不同且均为正值。从表 15 解释的总方差中可以简单看出，五个主成分均有不同的方差且显著大于 0，故通过该检验。

3）最后进行主成分个数的选取和检验。我们采用累积贡献率的方法，当 k=5 时，累积贡献率达到 99.8%以上，故我们选取前五个主成分已经涵盖了几乎全部样本信息，足以替代原始变量刻画总财政收入的影响因素。综上所述，我们建立的影响财政收入主成分模型是合理的。

2.2.6 建立各项税收入的时间序列模型并预测

时间序列是指同一种现象在不同时间上的相继连续的观察值排列而成的一组数字序列。时间序列预测方法的基本思想是：预测一个现象的未来变化时，用该现象的过去行为来预测未来。即通过时间序列的历史数据就可以揭示现象随时间变化的规律，将这种规律延伸到未来的一段时间，从而对该现象的未来做出预测。

表 22 中所提及的地方财政收入等 20 种项目都是具有滞后性的数据，根据这一特点，我们可以利用广州市 1999-2013 年的财政总收入及各个类别收入数据，运用时间序列分析的方法对这 20 种项目进行合理拟合（但不排除有误差的存在），从而对 2015 年广州市财政收入的各项项目做出合理的预测。

首先，分别对不同项目进行差分处理后得到平稳序列，再进行 $\mathrm{AR}(p)$、$\mathrm{MA}(q)$ 的尝试，最终确定模型后对其进行 t 检验与白噪声检验，发现所有模型都通过检验，并由模型得出每种项目 2014、2015 年的预测值，结果见表 22（其中，ε_t 为随机扰动项，显著性水平为 α =0.05）：

表22 广州市20种财政收入相关项目模型与预测

项目	模型	2014 年 预测值	2015 年 预测值
地方财政收入合计	$\nabla^2 x_t = \varepsilon_t - 0.969\varepsilon_{t-1}$	23903002.1	23881374
一般预算收入	$\nabla^2 x_t = 1330221.1 + \varepsilon_t - 0.7726\varepsilon_{t-1}$	12263140.2	12748265.1
增值税	$\nabla^2 x_t + \nabla^2 x_{t-1} = \varepsilon_t$	2400498	2858204
营业税	$\nabla^3 x_t + 0.91185\nabla^3 x_{t-1} = \varepsilon_t$	1745923.9	162213808
企业所得税	$\nabla^3 x_t + 0.92362\nabla^3 x_{t-1} = \varepsilon_t$	1166921.9	1155923
个人所得税	$\nabla^2 x_t = \varepsilon_t - 0.76747\varepsilon_{t-1}$	528608	562515.2
城市维护建设税	$\nabla^2 x_t = \varepsilon_t - 0.9542\varepsilon_{t-1}$	1003324.2	1013703
房产税	$\nabla x_t + 0.66197\nabla x_{t-1} = 32971 + \varepsilon_t$	682827.9054	717020.1734
印花税	$\nabla^2 x_t = 28679 + \varepsilon_t + 0.91409\varepsilon_{t-1}$	271828.4941	289625.0207
城镇土地使用税	$\nabla^2 x_t = \varepsilon_t - 0.97588\varepsilon_{t-1}$	186491.6671	178772
车船使用税	$\nabla x_t = \varepsilon_t - 1.79771\varepsilon_{t-1} - \varepsilon_{t-2}$	222007.2651	260686.4938
契税	$\nabla^2 x_t = \varepsilon_t - 0.99449\varepsilon_{t-1}$	655368.3272	698657

国有企业计划亏损补贴	$\nabla x_t - 0.8709\nabla x_{t-1} = 47466.1 + \varepsilon_t$	777029.6	582109.2
行政性收费收入	$\nabla^2 x_t + 0.64712\nabla^2 x_{t-1} = \varepsilon_t$	799933.7	856874
罚没收入	$\nabla^2 x_t = \varepsilon_t - 0.67945\varepsilon_{t-1}$	259259.7	234994
专项收入	$\nabla^2 x_t = \varepsilon_t - 0.9195\varepsilon_{t-1}$	472385.0635	493419
其他收入	$\nabla^2 x_t + 0.91143\nabla^2 x_{t-1} = \varepsilon_t$	437458.1668	508331.9882
基金预算收入	$\nabla^4 x_t = \varepsilon_t - 0.78069\varepsilon_{t-1}$	9935105	10463330
附：上级补助收入	$\nabla^2 x_t = \varepsilon_t - 0.93225\varepsilon_{t-1}$	2984366.2	3041249
税收返还收入	$\nabla^3 x_t = \varepsilon_t - 0.68117\varepsilon_{t-1}$	409556.9	406982

注： 由于篇幅限制，此处只给出增值税的模型建立和检验结果，如下图：

Conditional Least Squares Estimation

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
AR1,1	1.00000	0.21399	4.67	0.0005	1

Variance Estimate	3.395E10
Std Error Estimate	184262.6
AIC	353.0789
SBC	353.6438
Number of Residuals	13

* AIC and SBC do not include log determinant.

Autocorrelation Check of Residuals

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	4.31	5	0.5054	-0.074	-0.272	0.009	0.059	0.097	-0.305
12	17.30	11	0.0992	0.257	0.274	-0.145	-0.086	0.012	0.181

图 3　增值税时间序列模型分析

2.2.7　预测模型的检验

从广州市 2014 年预算执行情况和 2015 年预算草案的报告中得到 2014 年财政收入部分数据的真实值，将 2014 年的真实值和 2014 年的预测值进行比对，计算预测精度=（2014 预测值-2014 真实值）/2014 年真实值，检验预测模型的的有效性，检验结果如表 23：

表 23　预测精度检验

项目	2015 年预测值	2014 年预测值	2014 年真实值	预测精度
地方财政收入合计	23881374	23903002.1	24539000	-2.59%
一般预算收入	12748265.1	12263140.2	12415000	-1.22%
增值税	2858204	2400498	2537000	-5.38%
营业税	162213808	1745923.9	1480000	0.36%
企业所得税	1155923	1166921.9	1372000	-7.77%
个人所得税	562515.2	528608	560000	-5.61%
城市维护建设税	1013703	1003324.2	1117350	-9.21%
房产税	717020.1734	682827.9054	744900	-8.33%
印花税	289625.0207	271828.4941	——	——

城镇土地使用税	178772	186491.6671	124150	**50.21%**
车船使用税	260686.4938	222007.2651	——	——
契税	698657	655368.3272	869050	-8.28%
国企计划亏损补贴	582109.2	777029.6	——	——
行政性收费收入	856874	799933.7	744900	7.39%
罚没收入	234994	259259.7	124150	**108.83%**
专项收入	493419	472385.0635	516600	-8.56%
其他收入	508331.9882	437458.1668	476600	-8.21%
基金预算收入	10463330	9935105	10740000	-7.49%
附：上级补助收入	3041249	2984366.2	——	——
税收返还收入	406982	409556.9	——	——
所得税基数返还	——	——	——	

由表格可以看到，除了城市土地使用税、罚没收入以外，其他财政收入的各个项目的预测误差都小于 10%，但是城市土地使用税、罚没收入真实值和预测值相差甚远。

1）增值税等其他财政收入的各个项目的预测误差都小于 10%，说明对各个项目建立的预测模型比较准确，由此可得，据此预测 2015 年的数据具有一定的有效性。

2）专项收入、其他收入、政府性基金收入 2014 年的真实值都比预测值稍高，可以看出非税收入增速较高，而且近两年来增速较高，在一定程度上表明目前广州地区的经济形势稍稍严峻，需要用非税收入来弥补差额。

3）罚没收入在 2014 年远远低于以前的水平，相对于其他收入，罚没收入具有更大的不确定性，而且，随着税制改革的推进，逐步规范各种非税收入，逐步推广“费改税”，同时，公民素质也在不断提高，所以罚没收入大幅度下降，和预测值相去甚远，仅仅依靠定量分析得到罚没收入的预测值似乎不是很合理。

4）城镇土地使用税在 2014 年预测值远高于当年的真实值，着很大程度上可能归结于真实值获取时所存在的误差。我们利用 2014 年广州市一般公共预算收入各项目的比重得到各个项目的真实值，这过程中比重数据处理可能存在误差，导致得到的数据不准确。

综上，对各个项目建立的预测模型总体上是准确的，得到 2015 年广州市财政收入各个项目的预测值可以认为是有效的。

2.3. 结果分析

2.3.1 分析、识别财政收入的影响因素

通过建立财政收入总模型，得到五个影响财政收入的主成分因子，分别是综合指标、第三产业影响指标、教育影响指标、房地产影响指标和社会固定资产影响指标，另外，通过定性分析，政策变动也是其一影响因素。

第一，综合指标对财政收入的影响。财政收入是一个国家各项收入得以实现的物质保证。一个国家财政收入规模大小往往是衡量其经济实力的重要标志。其次，财政收入是国家对经济实行宏观调控的重要经济杠杆。宏观调控的首要问题是社会总需求与总供给的平衡问题，实现社会总需求与

总供给的平衡。财政收入的杠杆既可通过增收和减收来发挥总量调控作用，也可通过对不同财政资金缴纳者的财政负担大小的调整，来发挥结构调整的作用。因此，财政收入受各个方面综合影响，以经济影响尤为甚，其发展趋势和当前经济大环境发展趋势息息相关，经济发展会促进国民收入，从而会提高居民个人收入水平，直接影响储蓄水平，使得和财政收入增长保持一定的同向性。在经济体制及政策不变的情况下，财政收入会随着经济繁荣而增加，随着经济衰退而下降。

第二，第三产业对财政收入的影响。市场经济体制日趋完善，非公有制经济快速发展，成为吸收社会就业的主要渠道。在扩大内需引导下，以公有制经济为主、多种经济成分并存的第三产业的新格局已经呈现。从第三产业创造的价值看，第三产业增加值居前三位分别是零售业、交通运输业和金融业。广州作为华南地区的商业活动中心，商品零售业在经济运行中都发挥着不可替代的作用，随着城市软实力的增强，越来越多的游客、商人等到广州来进行商务、旅游活动，促进旅游业、餐饮业等发展，从而促进零售业的发展，带动第三产业的增长。另外，近年来，随着地铁 12 条线路同时开工，广汕铁路外绕线、南沙港铁路建设、白云机场扩建、广州港出海航道南北段维护等工程的开展，交通设施建设投入不断增加，促进交通运输业的发展，拉动第三产业的增长。金融业方面随着广州南沙新区、中新知识城等重大平台建设持续推进，城市基础设施建设全面铺开，企业投融资需求不断增加，广州金融业发展加快。新兴产业不断壮大，产业结构日趋合理，成为第三产业发展的主体。随着广东华南新药创新中心、浙江大学华南工业技术研究院等项目的引进，意味着新一代战略性新兴产业发展，这对第三产业的发展，将有着巨大的内推力。

尽管我国目前三个产业之间的界限不太明确，但是会相互渗透的，所以第三产业在快速发展的同时也能促进第一、第二产业的发展，而第一、第二产业也能带动第三产业的发展，从而推动三大产业的结构转型。另外，随着经济的发展，第一产业和第二产业的增长空间逐渐减小，而第三产业的增长空间还相对较大，所以第三产业在未来经济发展中所起到的作用会越来越明显，促进财政收入的增长，尤其是高科技和服务业这两个领域，将成为未来经济发展中新的增长点

第三，教育指标对财政收入的影响。21 世纪初，高等院校逐渐扩大招生规模，2004 年随着大学城的全面建成与投入使用，十一年来，累计培养高校毕业生数达到近五十万，另外，国家“千人计划”、“长江学者”、广州市创新创业领军人才等项目的实施对吸纳人才具有一定的影响力。增大教育投入促进人才培养，从而促进创新产业的发展，近年来在支持新能源汽车研发、超级计算机中心、广汽菲亚特等重大创新能力建设项目和重大产业化发展项目，支持交通节能和生物医药、光伏产业、新材料、新一代信息技术等战略性新兴产业发展上崭露头角。教育是兴国之本，教育事业的发展对经济发展是可持续的，从长远角度来看，对财政收入的影响是深远而且持久的。

第四，房地产业对财政收入的影响。广州时华南地区的中心城市，处于改革开放的前沿，房地产业发展较为成熟。自 21 世纪以来，广州房地产业由“幼稚期”向“成熟期”转变，房产业是属于第三产业中的生产服务业，房产业的发展能带动建材、装饰材料、家用电器等行业的发展，成为继金融保险业后又一经济增长亮点，是财政收入的重要源泉，财政收入的增加，使得城市建设资金有了明显增加，城市整体环境有了巨大改变，从而土地以及房地产业的收益得到提高，形成了以地养

地的良性循环。房地产业能提供的财政收入有房产税、土地使用税、契税营业税等等。另外，自 2004 年广州申亚成功，房地产业的投资信息和消费信心大大增强，房屋漏雨的消费需求增加，亚运村、新场馆和酒店建设等项目的开展不仅促进了服务业的发展，而且促进了城市基础设施建设，推动区域房地产业的发展，形成了亚运经济周期，拉动财政收入。

第五，全社会固定资产投资对财政收入的影响。固定资产投资是建造和购置固定资产的经济活动，即固定资产再生产活动，主要通过投资来促进经济增长，扩大税源，进而拉动财政收入整体增长。

第六，政策变动对财政收入的影响。政策出台是政府宏观调控的一个手段，无论是 2004 年的废除农业税、2014 年“营改增”试点改革、“限购令”的出台，还是南沙港成立经济自贸区，都会影响政府性基金收入、营业税、增值税、农业税等收入，对财政收入起到干预影响。

2.3.2 预测广州市 2015 年的财政总收入及各个类别收入

利用广州市 1999-2013 年的财政总收入及各个类别收入数据，运用时间序列分析的方法对这 20 种项目进行合理拟合，从而对 2015 年广州市财政收入的各项项目做出合理的预测。特别地，2015 年将进一步扩大“营改增”试点改革，到 2015 年下半年，全国将全面告别营业税，所以 2015 年广州市财政收入中的营业税收入会大幅下降，结合政策变动，最终预测值=0.5*初始预测值。

表 24 2015 年财政收入预测表

项目	2015 年预测值
地方财政收入合计	23881374
一般预算收入	12748265.1
增值税	2858204
营业税	**81106904**
企业所得税	1155923
个人所得税	562515.2
城市维护建设税	1013703
房产税	717020.1734
印花税	289625.0207
城镇土地使用税	178772
车船使用税	260686.4938
契税	698657
国企计划亏损补贴	582109.2
行政性收费收入	856874
罚没收入	234994
专项收入	493419
其他收入	508331.9882
基金预算收入	10463330
附：上级补助收入	3041249
#税收返还收入	406982
所得税基数返还	——

3. 结论

综上所述，财政收入受综合指标、第三产业影响指标、教育影响指标、房地产影响指标、社会固定资产影响指标、政策变动影响，对财政收入的预测对于克服年度地方预算收支规模确定的随意性和盲目性，正确处理地方财政与经济的相互关系具有十分重要的意义，为此，我们从经济因素和非经济因素角度向广州市财政局提出以下建议。

第一，增收。财政收入方面，随着央行出台下降银行准存利率，限贷政策有所松动，但受限购政策的影响，房地产成交销售持续低迷，房地产相关税收和土地出让形势不容乐观。明年起广州市全面停征堤围防护费，进一步扩大“营改增“改革试点范围等，都将对财政收入造成较大的冲击。但是另一方面，随着南沙成立自贸区，对南沙的金融、航运、物流和制造业发展有很大的推动作用通过自贸区这个渠道进入南沙，则可辐射珠三角地区乃至全中国，服务对象得到扩大，这将会是一个财政增收的突破口，政府应该抓住机遇，推进建设南沙港自贸区。

第二，节支。财政支出方面，发展社会公共事业、推进地铁、南沙港铁路等建设都需要一定的财政资金，财政支出压力不少。建议继续贯彻落实中央八项规定，严格控制公务接待费、办公经费、展会等一般性支出，坚持从严从简，勤俭办一切事业，优化财政支出结构，集中财力保民生。

第三，统筹兼顾。统筹安排一般公共预算、政府性基金等财政收入，充分发挥各类财政资金使用效率。对各项民生事业的投入要量入为出，保障基础民生，大力振兴服务业，增加科技投入力促科技强市。

第四，建设宜居城市。改善城市环境空气质量，增大环保投入，推进宜居城市建设，有利于城市名片的建设，从而促进旅游、服务业、商品零售业等的发展，完善城市功能。

第五，信息公开。公开统一部门预决算内容，进一步提高部门预决算的透明度与公开性。

4. 参考文献

[1]新华网.广州市出台土地出让金新政策
[EB/OL].http://news.xinhuanet.com/fortune/2010-05/15/c_12104783.htm ，2010-05-15/2015-05-20

[2]每日经济新闻.土地收入减少 今年多地政府性基金预算收缩
[EB/OL].http://business.sohu.com/20120221/n335316939.shtm，2012-02-21/2015-05-20

[3]证券时报网.上半年全国政府性基金收入降 20% 因土地收入减少
[EB/OL].http://finance.eastmoney.com/news/1350,20120713221947216.html ，2012-07-13/2015-05-20

[4]百度百科.政府性基金[EB/OL].http://baike.baidu.com/view/1453893.htm

[5]张新燕.四川省财政一般收入预测与分析[M] 2007

[6]樊元,朱卫明.从地方税收收入看甘肃三次产业的发展对地方财政收入的影响[M] 2007

[7]唐洁,张景. 申亚成功对广州房地产业发展的影响[M]2010

[8]李宗伟.房地产业与广州城市发展[M]2002

[9]骆翠敏.“营改增“对广州经济发展的影响[M]2014

[10]李璐.基于 R 语言的缺失值填补方法[M]2012

[11]傅德印. 主成分分析中的统计检验问题[J]2007

“泰迪杯”优秀作品