山东财经大学

2015 年(第四届)全国大学生统计建模大赛参赛论文

（研究生组）

题目：大数据背景下的山东省主要景区动态客流及因素分析[1]

学　　校：山东财经大学

学　　院：统计学院

专　　业：统计学

参赛队员：王琳　金戈　万道侠

指导教师：杨冬梅

二〇一五年六月

[1]注:该论文获得由中国统计教育学会举办的“2015 年（第四届）全国大学生统计建模大赛”大数据统计建模类研究生组三等奖。

摘　要

大数据在 21 世纪已经融入到人类的生产生活中，对统计模型的建立和数据的处理提出了新的要求。随着居民生活水平的提高，山东省旅游业也迅猛发展起来。由于其具有景区数量大，景点类型繁和客流人数多等特点，山东重点景区的客流量的合理预测成为一个亟待解决的问题。

本文利用大数据思想，基于“ 全国重点景区动态流量监测和服务系统 ”，利用移动通讯信号所勘测到的动态客流量、客源地、驻留时间等原始数据及记录，首先对客流量等指标进行探索性分析，深入挖掘景区实时客流、游客来源地、驻留时间的特点及其关系；随后，运用 BP 神经网络和支持向量机算法模型分别建立回归模型，选取其结果最优者，创造性地对影响山东重点景区的实时客流量进行建模分析；最后，验证了所采用机器算法模型在处理实时变量时较之于传统预测方法的优越性，并提出了相应地政策建议，同时也在大数据的应用研究方面进行了展望。

关键词：大数据　算法模型　客流量　回归

Abstract

In the 21st century, Big Data has been integrated into our life, raising new demands in building statistical models and dealing with data. With the improvement of our living standards, the tourism in Shandong Province has a rapid development. Because of its features of large number of scenic spots, many types and numerous visitors, a reasonable forecast of the number of Shandong scenic visitors becomes a serious problem. In this paper, we use the thought of Big Data, based on the " Dynamic Flow Monitoring and Service System of National Scenic", using the data and records of the dynamic number of visitors, coming cities of tourists and the dwell time collecting from mobile communication signals to analyze. First ,we analyze the relationship between the dynamic number of visitors, coming cities of tourists and the dwell time; Second, we use BP neural network and support vector machine model to build models, respectively, select the modeling which has the better result to analyze the numbers of visitors of Shandong plots; Finally, we verify the superiority of the machine algorithm model in dealing with real-time variables compared to traditional forecasting methods, and put forward some policy recommendations accordingly, and also express our prospects of the application of Big Data.

Keywords: Big Data algorithmic model passenger volume regression

大数据背景下的山东省主要景区动态客流及因素分析

一、研究背景

21 世纪是一个信息时代，是大数据的时代。迄今为止，大数据已成为各领域内的热点词汇，不管是沃尔玛语义搜索技术，还是梅西百货的实时定价机制，或是 Morton 牛排店的品牌认知，都显示出大数据特有的魅力。可以说，只有乘着大数据的臂膀与大数据同行，才能争当引领时代的“弄潮儿”，才能创造出更丰硕的成果。大数据时代对我们来说既是机遇也是挑战，2014 年 3 月 5 日，李克强在十二届全国人大二次会议上作政府工作报告首次提出了“互联网”和行动计划，这是“大数据”首次进入政府工作报告，也表明大数据将作为一种新兴产业会得到国家政府的大力支持。诚然，我国作为世界上第一人口大国并拥有世界五分之一的人口，我国产生的数据是巨大的，这无疑对中国与大数据来说是一个契机，中国海量“大数据”必将是大数据技术的发展的基础，大数据技术也将会给中国带来更多的发展机会。

大数据具有数据量大、数据种类多、要求实时性强、数据所蕴藏价值大四大特点。在各行各业均存在大数据，但是众多的信息和咨询是纷繁复杂的，我们需要搜索、处理、分析、归纳、总结其深层次的规律。统计建模的目的是为了寻求规律，对未来进行预测。我们可以充分利用大数据建立富有价值、操作性强、多样性的统计模型，为分析社会经济发展服务。因此，本文立足与大数据的时代背景，以统计建模为方式，应用大数据对山东省主要景区客流量变动进行分析，将大数据“落地”，而不空谈。

二、问题的描述

（一）现状分析

山东省作为我国的旅游大国，近几年来发展迅速，“好客山东”经过多年的培育，品牌价值达到 170 亿元。目前为止，山东省共有国家 10 处 AAAAA 风景区，其数量仅次于江苏、浙江、河南，在全国排名第四。在收入方面，2014 年山东省旅游总收入突破 5000 亿元，但与旅游收入位居第一的广东省（8305 亿元）相比还尚有距离。本节基于山东省旅游收入水平、旅游接待人次等指标，对山东省旅游业整体的发展现状进行总结：

1.旅游人次总体增加但增长速度在波动中减小

2000 年山东省旅游总人次约为 0.71 亿，从 2000 年到 2014 年，山东省每年的旅游人数在不断增加，图 1 中山东省旅游总人次每年的增长速度除 2003 年特殊时期的影响外，其余年份均呈正增长揭示了山东省旅游人次是逐年增加，但其增长速度却在波动中减小，说明山东省旅游人次逐年增加的人数是越来越少的。

2.旅游收入总体增加但增长速度在波动中减小

2000 年旅游总收入为 412.65 亿元，2014 年山东省旅游总收入达到 5878 亿元约为 2000 年旅游收入的 14 倍，从图 1 中山东省旅游总收入的增长速度可以看出，除 2003 年负增长以外（非典 SARS 的影响），其余年份旅游收入增长速度觉大于 0，显然山东省旅游收入在不断增加。但观测其增长速度，2001 年省旅游总收入增长速度为 15.39%，2012 年却降至 9.7%，在波动中呈逐渐减小的态势。

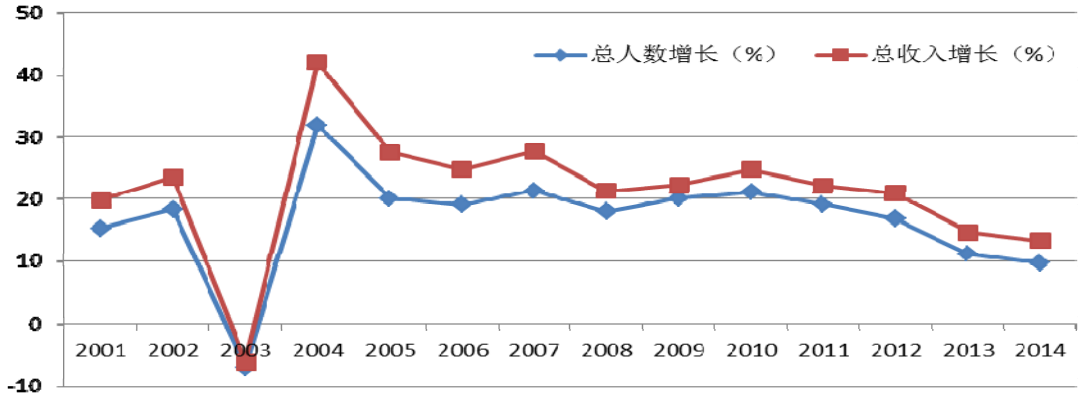

图 1　2000-2014 年山东省旅游总人数和总收入增长速度趋势图

3.旅游业总体发展不平衡

山东省 17 地市的旅游收入水平差异很大，如表 1 示，2014 年山东省 17 地市旅游总收入青岛市高达 1046.2 亿元，莱芜市却低至 44.1 亿元，省内旅游业发展及其不平衡；比较各地级市旅游总收入占全省的比重与其 GDP 比重，经济水平高的地区其旅游业也相对发达，或是旅游收入水平高的地区其经济发展水平也较高，二者有趋同之势。总的来说，旅游业发展良好与 GDP 比重较高的市区主要有青岛市、济南市、烟台市；经济水平较低的莱芜市，其旅游业发展水平也很低。

表 1　2014 年山东省 17 地级市旅游总收入、GDP 及其占全省比重

17 地级市	旅游总收入（亿元）	占全省比重 (%)	GDP (亿元)	占全省比重 (%)
青岛市	1046.2	16.16	8692.1	14.54
济南市	598.7	9.25	5770.6	9.66
烟台市	614.0	9.48	6002.1	10.04
淄博市	395.4	6.11	4029.8	6.74
潍坊市	499.5	7.71	4786.7	8.01
威海市	383.9	5.93	2790.3	4.67
泰安市	500.5	7.73	3002.2	5.02
东营市	98.9	1.53	3430.5	5.74
临沂市	469.7	7.25	3569.8	5.97
济宁市	443.4	6.85	3800.1	6.36
枣庄市	724.3	11.19	1980.1	3.31
日照市	237.9	3.67	1611.9	2.70
滨州市	93.4	1.44	2276.7	3.81
荷泽市	91.7	1.42	2222.2	3.72
德州市	117.6	1.82	2596.1	4.34
聊城市	115.7	1.79	2516.4	4.21
莱芜市	44.1	0.68	687.6	1.15

（二）问题分析

山东省经过十多年的快速发展, 旅游市场虽取得一定成效但有很多令人担忧的问题，加之现如今竞争日趋激烈，我们不得不思考，如何才能协调区域产业一体化的发展？如何才能增强本省旅游业的竞争优势？如何才能保持旅游业强劲的发展势头？基于大数据背景，我们提出本文的研究思路：利用大数据来分析山东省景区旅游客流量的细部特征，观测客流规律及分析影响客流因素，从而制定有效的营销策略，挖掘潜在客源，提高旅游管理水平。

移动通讯信号数据的可获得性，为我们研究这一问题了便利。移动通信技术自 1987 年进入我国以来，经历了一个爆炸性增长过程，手机从当年的奢侈品变成了如今寻常的通信工具。据工信部公布的数据，截止 2014 年末，全国大陆移动通信用户达到 15.36 亿户，而全国总人口为 13.6 亿人，移动电话普及率已达 112 部/百人，平均超越人手一部手机。移动通信技术的普及，使得手机用户与人口总体的统计特征已基本一致。移动通信运营商在为用户提供语音、短信、上网等各类服务的同时, 在其底层技术层面对人机交互过程中各种行为有着详细的记录，产生了海量数据。这些数据都具备了大数据的 4V（Volume、Variety、Value、Velocity）特征，形成了重要的大数据资源，为我们开发利用这些资源提供了可能。

因此，可以应用手机无线通信的特点，对景区游客流量、新增客流量、来源地及驻留时间进行实时动态监测，获取所需大数据，展开本文的研究。

三、数据的来源与处理

（一）数据的获取

本文在山东省旅游统计局的支持下，利用“ 全国重点景区动态流量监测和服务系统 ”取得了研究所需的原始数据及记录。该系统主要分景点信息获取、数据的存储、数据分析、数据呈现四个部分。具体而言，对列入计划的主要景区均设置了移动通讯监控设备，手机用户在监测范围内产生的通话、短信及位置变更等都会被实时采集，采集的信息通过专用网络传输到服务器进行存储并分析，最后，基于 24 小时勘测到的移动通讯信号得到每时点的客流信息。

山东省旅游局划定的主要景区包 7 个 AAAAA 景区、40 个 AAAA 景区及 6 个 AAA 景区，遍布在山东省的 17 地级市。本文将以监测系统中的数据作为基础，截取其中旅游客流数据进行分析，可以做到实时接收数据、实时处理数据和随时查询最新数据。所得数据主要包括新增客流量，实时客流量，客源以及驻留时间等数据。新增客流量反映的是每个时间隔内到达旅游景区人数的多少，为方便记录，在统计汇总中以一小时为间隔进行记录；实时客流量是在每一个时点所在景区的人数，在统计汇总的记录中也以一小时为间隔；客源是根据移动通讯的号码确定到达景区旅客的来源地，在统计中按照省份不同分类记录，反映的是各省每日到达景区的旅客数量的多少；驻留时间反映旅客在景区游玩时间的长短，从旅客到达景区开始，到旅客离开景区结束，在统计表中按照旅客的驻留时间长短不同分为 0-2 小时，2-4 小时，4-8 小时，8-24 小时，24-48 小时以及 48 小时以上 6 个类别，分别统计每个主要景点在各驻留时段的人数。

按照目前业界对大数据的普遍定义，我们所选用的数据源的特征与大数据的特征基本相吻合：一是数据体量巨大（Volume）。系统将一天分为 24 个整时点，每小时进行一次统计。移动网络在打包数据过程中以 1 分钟为周期，以适应景区动态信息的实时要求，这样产生的数据量无疑是巨大的，以 TB 来计量，普通的、传统的数据库已经无法满足处理的需要；二是处理速度快（Velocity）。这是大数据区分于传统数据挖掘的最显著特征，检测系统的记录包含了大量我们所需的数据信息且产生的速度很快，实时传送至终端；三是数据类型繁多（Variety）。类型的多样性也让数据分为结构化数据和非结构化数据，检测系统在处理便于存储的结构化数据的同时，也收录了图片和地理位置等非结构化数据；四是价值密度低（Value）。这是由于数据总量的巨大和价值密度高低与数据总量大小成反比的关系决定的，浩大的数据使我们对有效信息的“ 提纯 ”成为一个难题。

（二）数据的预处理

首先，我们对数据对象进行了统一化的处理与转化。比如，在统计汇总的过程中我们发现，极个别月份涉及到的景区会出现多一个景点的记录数据的状况，基于此类问题，我们选取系统所记录的相同景区的信息。其次，剔除异常信息。在对数据的处理中我们发现，每日都存在客源地未知的信息，这可能与检测系统的不完善及移动通讯信号的异常有关。对这类非人为异常数据我们给予剔除，只保留能够明确检测出旅客来源地的客源信息。第三，被检测的数据包括了普通游客、景区的工作人员、周边的商家、住户和路人等数据，除了正常游客外其他类型数据均不计入总量。我们基于被检测数据的时长来对数据进行甄别，比如根据景区特点，驻留时间超过 48 小时的一般为工作人员和常住的商户及住户，而驻留时间低于某个时段的一般为路人，通过对非游客人员的剔除我们得到的数据集更加清晰，准确度更加提高。

本研究利用该系统监测的山东省主要景点的实时客流人数、旅客来源地及驻留时间等数据，运用 EDA 统计方法和支持向量机、神经网络等算法，对数据进行了深度研究，下面将逐次展开。

四、基于 EDA 的特征分析

探索数据分析（EDA）的一些基本方法包括通过空间分位图、比例图、箱图等对数据非空间分布进行可视化处理。全国重点景区动态流量监测和服务系统可以将游客的手机信号数据进行分类。为了深入挖掘旅游实时客流、游客来源地、驻留时间的特点及之间的关系，本文利用手机信号频数，生成的数据集进行分析。

（一）动态客流的空间差异性分析

由于 5 月属于旅游旺季，游客人数集中，本节选取山东省 53 个主要景点作为研究对象，采集了 2015 年 5 月份每天的数据进行整理，生成数据集。为了观测山东省 17 地市主要景区客流量总体分布情况，本节利用 GEODA10.2 软件基于 2015 年 5 月份主要景点的实时客流量做出其空间四分位图，从图 2 可以看出，各地区 2015 年 5 月份的客流量存在很大差异，其范围处于 20440-935500 之间，由此将 17 地级市分为四个类型，其中，第一类地区为旅游业发达地区，月客流量最大（243400-935500），分别是青岛市、烟台市、泰安市、济宁和枣庄市；第二类地区为旅游业发展地区，月客流量处于 175700-215600 之间，分别是济南市、威海市、潍坊市和德州市；第三类地区为旅游业欠发达地区，分别是滨州市、东营市、淄博市和临沂市；第四类地区为旅游业欠发展地区，其月度客流量最少，分别是聊城市、莱芜市、日照市和菏泽市。

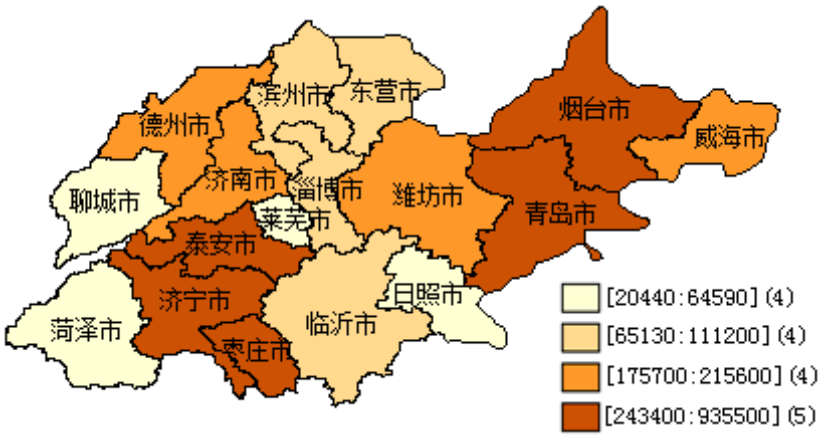

图 2　2015 年 5 月山东省 17 地市主要景区实时客流量的空间四分位图

（二）动态客流的波动性分析

为了便于分析，本节从三大类别中随机挑选出一个具有代表性的景区进行分析，分别是趵突泉景区（AAAAA）、刘公岛景区（AAAA）、龙悦湖旅游渡假区（AAA）。分析三个典型景区在年内、周内及日内客流量波动情况及特征，旨在对山东省主要景区的季节波动、工作日及周末客流量的特征探究。

1.景区客流量季节特征分析

使用“全国重点景区动态流量监测和服务系统”中 2014 年 1 月 1 号至 2014 年 12 月 31 日山东省典型景点的实时客流量数据，并会汇总整理得到月度数据，作出 2014 年典型景区客流量变化趋势图。

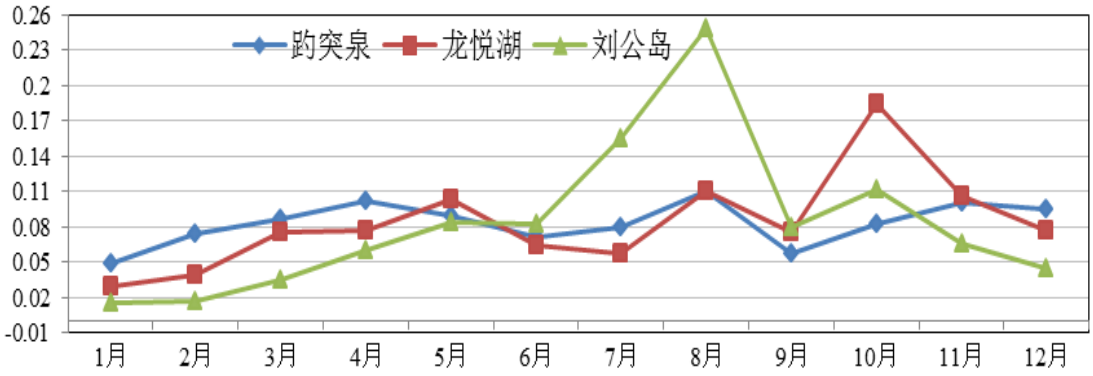

图 3　2014 年典型景区客流量平均变化趋势图

图 3 显示，由于受气候、自然等多种因素的影响，旅游景区客流呈现出明显的季节波动。三个景区在 8 月和 10 月月客流量都出现高峰值，是景区的旺季。其中，刘公岛景区和趵突泉景区的客流量占全年比值在 8 月达到最高，分别为 24.9%和 11.1%，这主要是在学生暑期时段，说明学生群体占游客很大比例；龙悦湖旅游渡假区的高峰值出现在 10 月，比例高达 18.43%，这主要是受到黄金周

的影响，使得客流量有大幅提升。三个景区在冬季（12 月、1 月、2 月）出现低谷期，这一时段是景区的淡季，这与处在温带季风气候的自然景观受冬季气温、风向和天气的影响有直接联系；另外，综合全年来看，龙悦湖旅游度假区的波动频率最大，这主要是因为其作为 AAA 级景区客流量基数相对较小，所以每月占全年客流量这一比例对月客流量的波动反映较之于另外两个景区较为敏感。

2.景区周内客流量变动分析

为避免节日效应造成的景区客流量大规模的上涨，本节采集了 2015 年 5 月 4 日-2015 年 5 月 31 日典型景区每天的客流量，为了探索典型景区客流量在一周内的波动特征，我们分别计算三个景区平均每天的客流量比重，如图 4 所示。三个典型景区的客流量在周一均较低，周一至周五五天工作日内均没有明显起伏波动，从周五开始，客流量开始呈上升趋势，在周末两天达到高峰，其中周六的客流量到达最高值，周日有下降趋势。这说明景区周末吸引了大量短途客人，显示了景区的假日型特征。

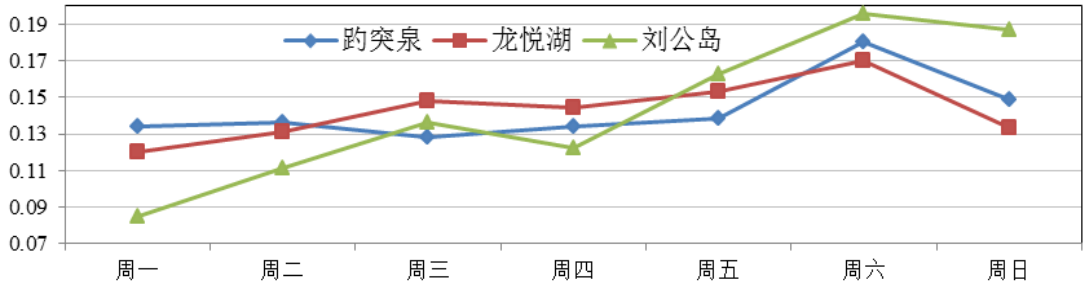

图 4　典型景区平均周客流量变化趋势图

3.景区日高峰期时段分析

为探讨不同季节对景区日高峰期时段的影响，我们选取了趵突泉景区在 3 月、6 月、9 月和 12 月这 4 个月客流量在每个时点上的分别的平均值，绘制波动曲线如下：

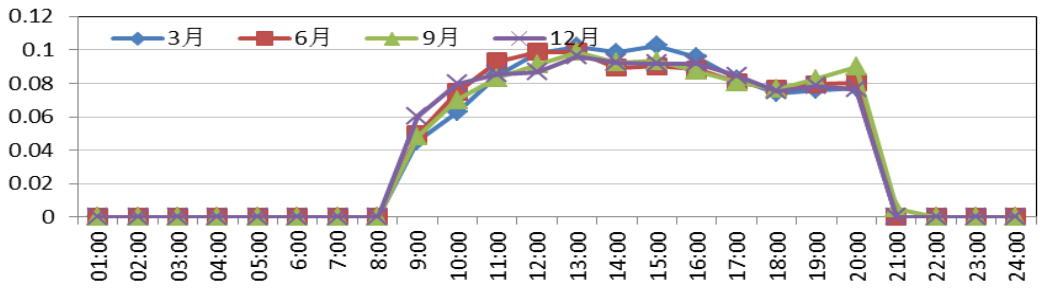

图 5　趵突泉景区在不同季节的日客流量平均波动

从图 5 我们可以看出，4 条曲线几乎是重合的，即趵突泉景区在 4 个月份日客流量波动变化情况是相似的，说明这四个月份的变动对景区日客流量波动没有

显著影响。所以在随后的分析中，我们直接收集趵突泉景区、刘公岛景区和龙悦湖旅游渡假区这三个典型景区 2015 年 5 月的客流量数据进行分析，对一天之内每时客流量与当日客流量之比进行统计与平均，三个景区的天内变动曲线如图 6 所示。

图 6 显示，三个景区在 21 时-次日 7 时客流量均为 0，说明三个景区主要以短期游玩的旅客为主，暂无驻留超过 24 小时的游客群体。一天之内景区饱和度最高的高峰时段在 11 时-13 时，15 时-17 时；13 时-14 时出现低谷；新增客流量高峰期在营业时间开始的前两个小时；营业结束前一小时之内客流量下降明显。

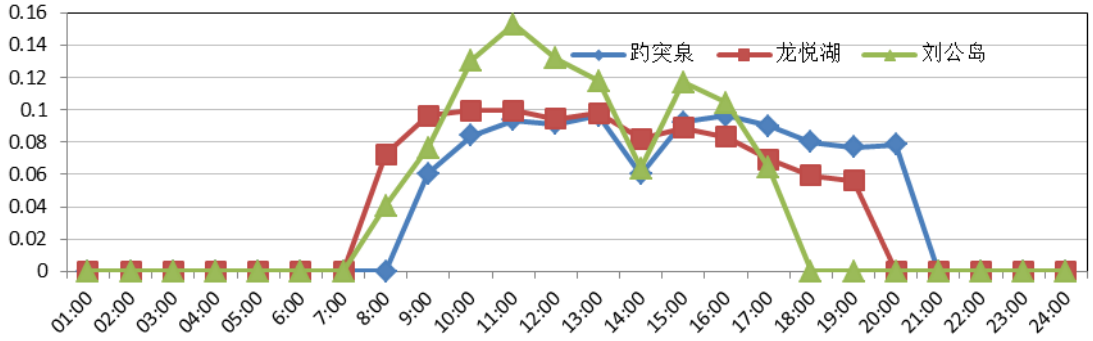

图 6　典型景区平均日客流量变动趋势图

（三）动态客流的来源地分析

本节选取山东省 53 个主要景点作为研究对象，采集了 2015 年 5 月份每天的数据进行汇总整理，得到山东省主要景点动态客流的客源地数据，利用 GEODA10.2 软件做出空间百分位图进行分析。百分位图能够显示客流量人数较多的来源地区和较少的省市区。如图 7 所示，空间百分位图依据客源地人数将全国 31 个省市区分为 6 大类。其中，山东省主要景点的客源 80%以上都来源于山东省；其次是相邻的省份，河北省，河南省和江苏省，但其比例远远小于本身客源的比例，处于 1%-10%之间；第三大类是离山东省距离相对较小并且经济水平较高的地区，主要分布在第二大类省份的周围，比如北京市、浙江省、广东省等地区；第四-第六类地区距离山东省较远，客流人数较少，不足总人数的 0.5%，主要分布在偏远的西部地区，比如西藏，广西、云南等地区。

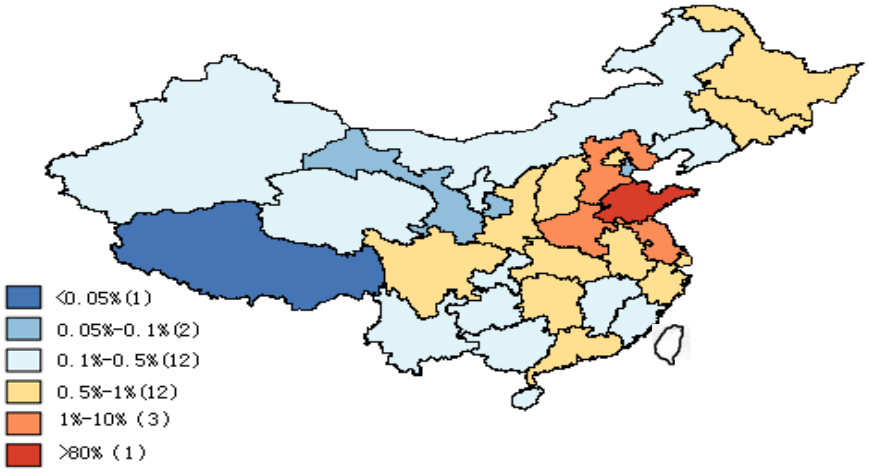

图 7　2015 年 5 月山东省主要景区实时客流量客源地的空间百分位图

可以看出, 山东省客源主要来源于当地, 客源量以山东省为中心向外围扩散, 随着距离的增加逐渐减少。省际客流量差异的原因首先地理位置造成的距离，还有可能和经济发展水平相关。

（四）动态客流的驻留时间分析

本节选取山东省 53 个主要景点作为研究对象，采集了 2015 年 5 月 31 日当天不同景点的驻留时间加以析，如表 2 分别列举了 AAAAA 景区、AAAA 景区、AAA 景区实时客流量的驻留时间，其中驻留时间在 0-2 小时的实时客流量最大，其次是 2-4 小时，随着统计驻留时间的增加，客流量渐少。较少的游客的驻留时间达到 24 小时以上, 特别是三大级别的景区驻留时间在 48 小时以上的人数为 0, 说明几乎没有游客驻留时间超过 48 小时。另外，通过比较三大级别景区的驻留时间可以看出，相同的驻留时间内级别越高的景区内的客流量越大，这也说明景区级别是影响人们选择旅游的因素之一。

表 2　不同景区客流量的驻留时间表

驻留时间 景点级别	0-2 小时	2-4 小时	4-8 小时	8-24 小时	24-48 小时	48 小时以上
AAAAA 景区	5766	3230	1488	1338	317	0
AAAA 景区	3394	1984	812	535	50	0
AAA 景区	1147	1257	377	377	39	0

五、基于机器学习算法模型的实证分析

大数据是无目的的数据搜集和挖掘, 我们要做的关键工作是在挖掘的海量原始数据去寻找有意义有价值的信息。在数据分析的过程中，我们假设实时客流量与各个景区的类别、等级、门票价格与当地的经济发展情况或政府支出情况等变

量有关联，对客流量影响因素进行了深入挖掘以便探究。在旅游研究中，BP 神经网络已广泛用于对国内游客量进行预测分析。本文对山东省重点景区实时客流量进行建模分析，选取算法模型中的 BP 神经网络方法和支持向量机方法进行建模，并采用线性回归建模方法与算法模型预测精度进行对比分析。

（一）模型的选择依据

人工神经网络是数据挖掘常用方法之一，该方法可以处理连续型和类别型数据，对数据进行分类和预测。20 世纪 80 年代中期，得益于神经网络的使用在全国乃至全世界范围内的发展，国内外学术界渐渐掀起了研究神经网络的狂潮。近年来，卢金秋（2006）[1]在原有基础上对 BP 神经网络进行改进，对照标准 BP 神经网络算法，通过对实际税收数据的测试运行，验证了改进算法在训练性能上的优越性；戴丹（2006）[2]利用 BP 神经网络模型预测股票市场未来 2 周的收盘价中期变化趋势，提高了预测精度；牛忠远（2006）[3]基于物流需求的时间序列统计数据，利用人工神经网络多步预测和滚动预测方法对我国物流需求进行了科学预测；张丽娟（2013）[4]采用基于神经网络算法的内模控制进行发动机排气噪声的消除，采用有源控制方法消除排气噪声，对汽车排气噪声有源控制效果有一定程度上的改善；王晓军（2014）[5]提出了将数据挖掘与 BP 神经网络以及 RBF 神经网络相结合的动态线性建模方法，有利于煤气化系统的数据预测和控制和工厂中控制系统的优化；马保忠、陈传明（2015）[6]利用 BP 神经网络模型对时间序列模型进行修正，从而对黄金期货价格进行研究和预测。

基于统计学习理论的支持向量机算法具有理论完备、全局优化、适应性强的优点，是机器学习的一种新方法和研究新热点[7]。近年来，支持向量机在回归方面有很好的表现[8][9]。Smola（2004）[10]对支持向量机的回归问题进行综述探究，总结了目前训练支持向量机的算法，包括二次规划部分及如何处理大规模数据的算法，并针对标准支持向量算法进行修正及改善；针对大规模回归问题，Mangasarian 等（2000,2002）[11][12]通过线性规划解决大规模支持向量回归的问题，并且效果显著；GAO（2003）[13][14]提出了基于平均场方法及平均域方法的支持向量机回归算法。本文选取上述两个算法模型对山东省重点景区实时客流量进行建模分析。

（二）模型的介绍

1.BP 神经网络模型

BP 神经网络(Back Propagation)是机器学习中一种经典的学习方法，通过对带有标记的训练进行学习，建立从样本属性到目标概念的非线性函数关系。可以证明，当满足一定条件时，BP 神经网络可以以任意精度逼近任意函数，预测能

力较强。BP 神经网络，包括误差反传误差反向传播算法的学习过程，由信息的正向传播和误差的反向传播两个过程组成。由图 8 可知，BP 神经网络是一个三层的网络：

输入层（Input layer）：输入层各神经元负责接收来自外界的输入信息，并传递给中间层各神经元；隐藏层（Hidden Layer）：中间层是内部信息处理层，负责信息变换, 根据信息变化能力的需求, 中间层可以设计为单隐层或者多隐层结构；最后一个隐层传递到输出层各神经元的信息，经进一步处理后，完成一次学习的正向传播处理过程；输出层（Output Layer）：顾名思义，输出层向外界输出信息处理结果。

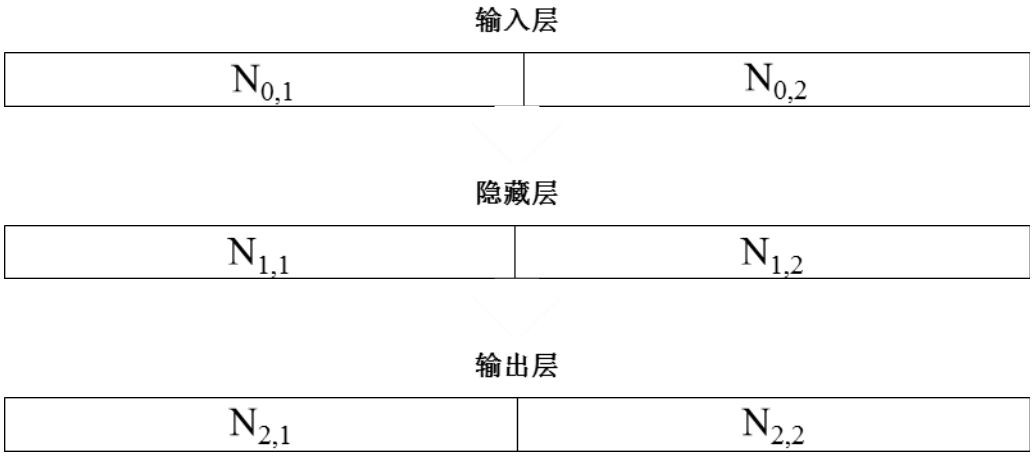

图 8　BP 网络神经结构图

当实际输出与期望输出不符时, 进入误差的反向传播阶段。误差通过输出层, 按误差梯度下降的方式修正各层权值，向隐藏层、输入层逐层反传。周而复始的信息正向传播和误差反向传播过程，是各层权值不断调整的过程，也是神经网络学习训练的过程，此过程一直进行到网络输出的误差减少到可以接受的程度，或者预先设定的学习次数为止。

2.支持向量机方法

支持向量机（SVM）是一种线性和非线性数据的分类方法，它使用非线性映射将原始数据映射到高维空间，在该空间内搜索最佳分离超平面。在线性可分的情况下，存在这样的超平面把空间中的类分开，并且该超平面与类的距离最大即最大边缘超平面，它等价于求解约束的凸二次最优化问题。在线性不可分的情

况下，可以允许个别样本分类错误，但需要借助非线性映射把原输入数据变换到高维空间，在高维空间中构造线性决策函数来实现原空间中的非线性决策函数，巧妙地解决了维数问题，并保证了有较好的推广能力，而且算法复杂度与样本维数无关。目前，SVM 算法在模式识别、回归估计、概率密度函数估计等方面都有应用，且算法在效率与精度上已经超过传统的学习算法或与之不相上下。为适应训练样本集的非线性，传统的拟合方法通常是在线性方程后面加高阶项。此法诚然有效，但由此增加的可调参数未免增加了过拟合的风险。支持向量回归算法采用核函数解决这一矛盾。用核函数代替线性方程中的线性项可以使原来的线性算法“ 非线性化 ”，即能做非线性回归。与此同时，引进核函数达到了“ 升维 ”的目的，而增加的可调参数通过拟合依然能控制。

（三）指标的选择和数据处理

1.指标选择

本文利用“ 全国重点景区动态流量监测和服务系统 ”取得了研究所需的原始数据及记录，从中选取了 2013 年 9 月至 2015 年 3 月 53 个景区的游客数量(people)作为模型的因变量，并对这 1007 条数据进行了数据清洗和分析整理，保证了数据使用的有效性、可靠性和可测度性。

根据以往旅游需求印象因素的研究，旅游需求的影响因素一般包含以下几个方面[15][16][17]：景区所在地的经济发展水平、旅游价格、政策支持、景区设施等因素，故我们选取了 53 个景区所在地公共性财政支出（pay）的月度数据作为度量当地公共事业发展的指标；选取各个景区的门票价格（price）作为度量旅游价格的指标；选取景区所在城市是否列入山东省启动的“ 一圈一带 ”区域发展战略城市(plan)作为度量政策支持的指标；选取景区质量等级划分的级别(class)作为度量景区设施等因素的指标。

2.数据处理

为了消除不同变量的单位对模型分析造成的影响，我们对数值型数据统一进行了无量纲化处理，对标准化后的数据进行模型的拟合分析。

（四）模型的拟合与分析

1.模型的拟合

(1) BP 神经网络模型构造

在构建 BP 神经网络模型的过程中，我们使用 nnet 函数进行分析，在所选取的样本中随机选取训练集和测试集，其中，训练集用于构建模型，我们选取 70%

的比例，测试集用于评估模型的预测能力，我们选取的比例为 30%。输入层中，people 作为因变量，price、plan、class 和 pay 作为自变量输入。根据经验分析，我们构建包含 5 个节点的单个隐藏层的神经网络。为保证结构收敛于局部最优值，我们选择参数权重递减值为 0.01，最大迭代次数为 2000。本文选取 NMSE 指标作为指标参数，其中 NMSE 的计算公式如公式（1）示：

$$NMSE = \frac{mean((y - \hat{y})^2)}{mean((\bar{y} - y)^2)} \quad (1)$$

其中，NMSE 值越小说明模型预测精度越好。测算结果如表 3 示：

表 3　BP 神经网络 NMSE

指标	NMSE
训练集	0.4128
测试集	0.3989

(2) 支持向量机模型构造

在构建支持向量机的模型中选用了 svm 函数进行分析。同样，在构建样本的过程中，随机抽取了 70%作为训练集，用于拟合模型；30%作为测试集，用于评估模型的拟合能力。在拟合支持向量机回归的阶段，我们选择“ eps-regression ”类型进行回归拟合，并使用核函数“ radial ”，惩罚因子取 10，对模型进行拟合和评估。判断模型拟合效果的参数我们依然选用 NMSE 指标，测算结果如表 4 所示：

表 4　支持向量机 NMSE

指标	NMSE
训练集	0.5191
测试集	0.5001

(3) 线性回归模型

为对比机器算法模型与传统算法模型拟合效果，我们对选取线性回归模型对变量进行分析，模型形式如下公式（2）：

$$y_t = c + x_{1t} + x_{2t} + x_{3t} + x_{4t} \quad t=1, 2, 3, \cdots, n \quad (2)$$

其中，y_t 表示实时客流量，x_{1t} 代表景区价格，x_{2t} 表示景区的级别，x_{3t} 表示景区所在城市是否为一圈一带规划城市，x_{4t} 表示景区所在城市的公共财政预算支出。通过分析我们得出如下表 5 所示回归结果：

表 5　线性回归结果

变量名	Estimate	Std. Error	t value	Pr(>|t|)	F
Intercept	-0.010877	0.032826	-0.331	0.740484	23.75
price	0.129930	0.033462	3.883	0.000113	p-value<
class	0.265133	0.032973	8.041	3.82e-15	2.2e-16
Plan	0.107539	0.035660	3.016	0.002657	
pay	0.003028	0.034974	0.087	0.931032	

其中，整个回归方程和多数变量均通过显著性检验。为了与上述算法模型拟合效果进行同度量分析，我们对线性回归模型也用了 70%的数据作为训练集拟合模型，30%的数据作为测试集验证模型，其训练集和测试集的拟合效果如表 6 示：

表 6　线性回归 NMSE

指标	NMSE
训练集	0.8802
测试集	0.8835

2.模型的分析

通过三种算法模型对变量的拟合，我们得到三种模型的预测精度对比，如下：

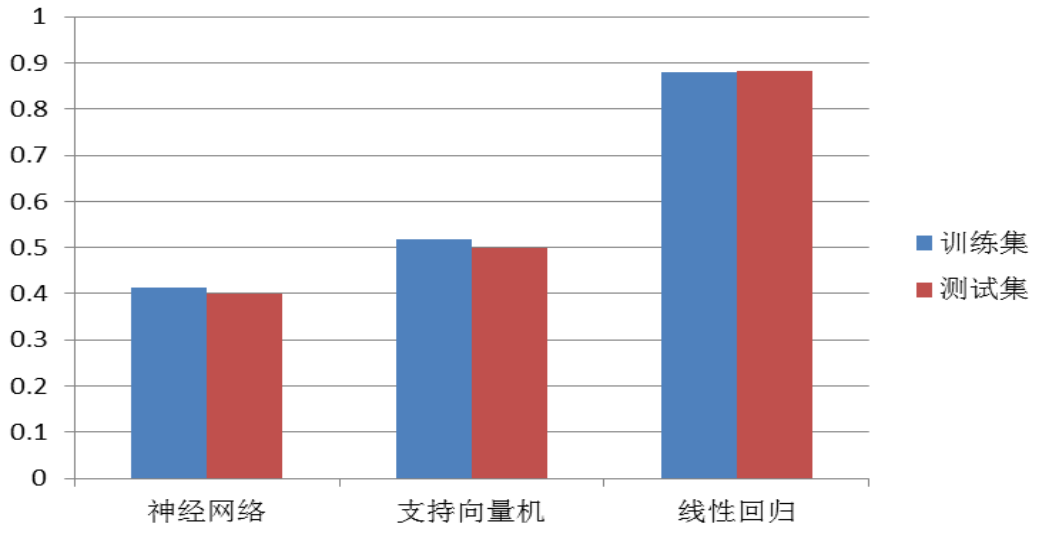

图 9　三种模型 NMSE

如图 9 所示，与传统建模方法相比，神经网络和支持向量机的预测性能要明显优于传统计量分析方法。其原因可能是传统的预测方法需要对数据存在一定的基本假设，而一些假设在实际获得的数据或者是变量中很难满足，以至于影响模型的预测精度。机器算法模型对数据本身的要求不高，在做回归和预测分析时也不需要数据的正态性、独立性等假设条件，适用性强，在模型拟合和预测效果上可能更加精确。

六、应用及展望

（一）研究结论及建议

1.研究结论

本文首先运用 EDA 统计方法对“ 全国重点景区动态流量监测和服务系统 ”获取的数据进行客流量的空间差异性分析、客流量波动性分析、客流的来源地以及驻留时间分析。我们得出以下结论：

第一，以 2015 年 5 月份天度客源数据分析得出，山东省 17 地级市客流量存在很大的空间差异性，可以分为旅游业发达地区、旅游业发展地区、旅游业欠发达地区和旅游业欠发展地区四大类。

第二，以趵突泉景区（AAAAA）、刘公岛景区（AAAA）、龙悦湖旅游渡假区（AAA）三个典型景区为例，得出景区的客流量有明显季节趋势和假日性特征，且景区一天内游客量高峰时段集中，因此可根据动态客流波动引导景区游客错时消费；根据对不同级别的景区驻留时间分析，不同驻留时间段中景区级别越高的人数越多，则说明景区级别也是影响客流量的因素之一。

第三，使用 2015 年 5 月最新日度客源数据，根据来源地人数差异在空间百分位图上将全国 31 个省市区分为 6 大类。其中景区的客流来源以山东省为主，并向外围扩散，随着距离的增加逐渐减少。省外客源的不足与距离有关，也可能与景区自身建设与影响力有关；

第四, 运用 BP 神经网络和支持向量机算法模型以及传统线性回归计量模型，测算影响景区实时客流量的各因素指标，对客流量进行了回归建模。其中 BP 神经网络的拟合效果最好。与之相比，传统的线性回归方法对这种实时数据的拟合效果差强人意。

2.政策建议

根据上述研究结果，本文针对山东省各地市扩大当地旅游业发展，提高客流量提供如下政策建议：

(1)加大政府扶持力度。政府可对旅游业欠发达地区以及经济欠发展地区的进行扶植，加大政策性资金投入和项目引进。其中，山东省于 2014 年 8 月实施“ 一圈一带 ”城市规划政策，对“ 一圈一带 ”规划城市提供了强有力的政策扶持和旅游项目带动，无疑为当地旅游业发展注入了一支“ 强心剂 ”。

(2)提高政府财政支出中公共设施支出部分。当地政府应加大基础设施建设力度，加强城市硬件设施建设，并对当地现有的特色自然景观和人文景观进行有效的开发和保护，促使当地旅游业繁荣发展。另外，可对加大对当地景区的宣传力度，扩大景区在省外的影响力。

(3)加强景区自身基础设施建设。可在景区附近引进投资，提高经济型餐厅、宾馆数量，延长游客驻留时间，同时提高对省外或长距离客源的吸引力。也可根据景区自身优势，开发吸引游客的游玩和观赏的新看点，丰富景区内观赏类型。

(4)引导景区错时消费。景区可以根据自身的旅游旺季和淡季制定不同的门票价格，鼓励游客在淡季来景区游玩，以保证景区淡季客源。在工作日可对门票价格或游玩场所的价格进行折减。在高峰期设立提示牌，引导游客避开高峰期进入景区游玩。此外，也可在新增客流量高峰期增开售票窗口，引导游客分流快捷取票。

（二）不足及展望

1.变量选取。模型的自变量是各景区月度实时客流量，在选取影响自变量变动的指标因素中，只搜集到了一些政府支出、政策支持以及景区自身因素的指标，而未能搜集反映景区当地经济发展、服务业发展等指标的相关月度资料。

2.数据使用。本文使用的数据是从“ 全国重点景区动态流量监测和服务系统 ”上获取的的实时各时点上的日度数据，我们为了建立影响景区客流量因素模型，使用了天度数据汇总下的月度数据，没有充分利用大数据每日时点更新数据。如果能获取影响景区客流量的一些天度数据，可能模型结果会更有针对性和实际意义。

3.模型使用。本文使用传统建模方法和算法模型进行对比分析，但算法模型只使用了具有代表性的支持向量机和 BP 神经网络建模，可以使用一些其他的算法模型建模，分析比较预测精度。

参考文献

[1]卢金秋.数据挖掘中的人工神经网络算法及应用研究[D].浙江工业学,2006.

[2]戴丹.BP 神经网络用于市场预测的研究[D].武汉理工大学,2006.

[3]牛忠远.我国物流需求预测的神经网络模型和实证分析研究[D].浙江大学,2006.

[4]张丽娟.基于神经网络算法的车辆噪声有源控制研究[D].上海工程技术大学,2014.

[5]王晓军.神经网络建模方法及数据挖掘在煤造气过程中的应用[D].北京交通大学,2014.

[6]马保忠,陈传明. BP 神经网络模型在黄金期货价格预测中的实证分析[J].企业导报,2015,11:19-20.

[7]常甜甜.支持向量机学习算法若干问题的研究[D].西安电子科技大学,2010.

[8]V.N.Japnik,S.Golowich, A. Smola. Support vector method for function approximation, regression estimation,and signal processing [J] ,In Advances in Neural Information Processing Systems 9, Cambridge,MA,MIT Press, 1997, 281-287.

[9]H.Drucker, C.J.C.Burges, L.Kaufman,et al. Support vector regression machines [J] , Advances in Neural Information Processing Systems,9,Cambridge,MA,MIT Press,1997,155-161.

[10]A.J.Smola, B.Scholkopf. A tutorial on support vector regression [J] ,Stat.Comput.2004,14:199-222.

[11] O. L. Mangasarian, D. R. Musicant. Robust linear and support vector regression [J] . IEEE Trans. Pattern Analysis Mach. Intell,2000,22: 950-955.

[12]O. L. Mangasarian, D.R. Musicant. Large scale Kernel regression via linear programming [J] , Mach. Learn, 2002,46:255-269.

[13] J.B. Gao, S. R. Gunn, C.J. Harris. SVM regression through variational methods and its sequential implementation [J] , Neurocomputing,2003, 55:151-167.

[14] J. B. Gao, S. R. Gunn, C. J. Harris. Mean field method for the support vector machine regression [J] ,Neurocomputer,2003,50:391-405.

[15] Law R. A neural network model to forecast Japanese demand fortravel to Hong Kong [J]. Tourism Management, 1999, 20:89—97.

[16] Qu H —L, Lam S. A travel demand model for Mainland Chinesetourists to Hong Kong [J]. Tourism Management, 1997, 18(8):593—597.

[17]张玉娟, 赵定涛. 中国入境旅游需求影响因素分析 [J]. 经济理论与经济管理, 2008, (5):51-55.