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：大数据背景下投资者关注对“一带一路”概念股预测能力的实证检验[1]

学　　校　山东财经大学

参赛队员　王冬冬　吴寒　史光燕

学科专业　统　计　学

指导教师　田　金　方

院　　系　统　计　学　院

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

随着“ 一带一路 ”规划的陆续出台和实施，“ 一带一路 ”主题直接成为股市炒作的热点，“ 一带一路 ”概念股应运而生。本文首先攫取该热点话题，选取“ 一带一路 ”概念股的有关股票，并得到沪深两市中“ 一带一路 ”概念股的相关市场表现。其次在大数据背景下借助百度指数这一网络搜索量作为投资者关注度的代理指标，探究投资者关注与概念股市场表现的相关关系。最后结合 5 折交叉验证技术，利用线性回归、回归树、神经网络、随机森林和支持向量机等五种模型对沪深两市中“ 一带一路 ”概念股的市场表现：平均成交量和收益率进行预测。通过模型拟合度和稳定性的比较，得到针对不同数据集的相对最优模型。并基于预测结果，对个人投资者和政府提供建设性建议。

本文借助大数据理论和数据挖掘方法，首先从国泰安 CSMAR 系列研究数据库中获取了沪深两市中 61 只股票从 2014 年 6 月 6 日至 2015 年 5 月 22 日共 236 个交易日的收盘价和成交量作为“ 一带一路 ”概念股收益率预测的基础数据。同时引入网络大数据——百度指数作为投资者关注的代理指标，增强了对“ 一带一路 ”概念股预测的效果。

关键词：“ 一带一路 ”概念股；成交量；收益率；投资者关注

一、引言

（一）研究背景

继沪港通之后，A 股市场近期又掀起了一波新的主题炒作—“ 一带一路 ”。在这种炒作氛围下，基建、涉外工程、港口等概念股持续受到资金关注，此前遭到冷遇的中国交建 、中国铁建、中材国际等这些传统基建股自 2014 年 10 月份起相继出现 20%以上的涨幅，多只与“ 一带一路 ”概念相关的股票出现连续涨停。

所谓的“ 一带一路 ”，即“ 丝绸之路经济带 ”和“ 21 世纪海上丝绸之路 ”。2013 年 9 月和 10 月由中国国家主席习近平分别提出建设“ 新丝绸之路经济带 ”和“ 21 世纪海上丝绸之路 ”的战略构想，两者并称为一带一路（One Belt And One Road，简称“ OBAOR ”；或 One Belt One Road，简称“ OBOR ”）。随着“ 一带一路 ”规划的陆续出台和实施，“ 一带一路 ”主题直接成为股市炒作的热点，“ 一带一路 ”

概念股应运而生。民生证券 、申银万国 、兴业证券等多家券商相继发布研究报告，推荐该主题的投资机会。 如民生证券认为，“ 一带一路 ”战略将是我国未来十年的重大政策红利；而兴业证券更是认为，“ 一带一路 ”是堪比 “ 加入 WTO ”性质的惠普式机会、系统性机会，受益于“ 一带一路 ”大战略的系统行情将可能延续 3 到 5 年。下表详细阐述了“ 一带一路 ”的投资主线，如表 1 所示：

表 1 “ 一带一路 ”投资主线

分类	行业/省份	逻辑	相关企业
按行业	基建（铁路、港口）	基建投资机会	新疆城建、西部建设等
	油气管网及设备	中国是中亚油气资源最大的购买国 ;中俄能源合作	广汇能源、金州管道等
	旅游	带动西部旅游业发展	西安旅游等
	电网设备类	对外扩张	特变电工等
	交通物流	公路、铁路、航运等联通	永贵电器等
	农产品	西部农业开发	新农开发等
	商业贸易	进出口贸易增加	友好集团等
	金融业	区域经济整合增长	宏源证券等
按区域	新 疆	建设“ 核心区 ”	北新路桥等
	山 西	提出“ 新起点 ”	建设机械等
	甘 肃	打造“ 黄金段 ”	上峰水泥等
	宁 夏	提出“ 战略支点 ”；与阿拉伯国家“ 捆绑 ”	宁夏建材等
	东南沿海	北部湾城市群、广东、福建、海南、浙江等省港口	港口型公司、远洋运输公司

资料来源：21 世纪经济报道

“ 一带一路 ”的持续关注催生了股市的新投资热点——“ 一带一路 ”概念股。这一现象无疑佐证了投资者关注对股票交易的影响。因此如何看待“ 一带一路 ”概念股的未来走势，投资者关注在其中扮演的角色又是怎样的正是本文的研究重点。

在大数据背景下，借助互联网科技和信息技术的高速发展，搜索引擎正成为股市投资者获取信息的重要途径。投资者通过网络搜索引擎可以获取大量与投资

相关并且契合自己投资兴趣的投资信息，借助分析信息来获取更高的利润。与此同时搜索引擎通过记录投资者的搜索数据，反映其关注的焦点和强度，从而映射投资者的投资行为和倾向。基于以上事实，本文将搜索引擎数据作为衡量投资者关注的代理指标，检验投资者关注对“一带一路”概念股的影响。

（二）研究综述

针对投资者关注及其对股票市场影响的研究，国内外学者做了大量的研究工作，也有了一定的文献基础。Barber 和 Odean（2001）通过研究发现，个人投资者的关注度在一定程度上决定了他们是否购买股票。Da,Engelberg 和 Gao(2011)首次提出网络搜索数据的可利用价值，他们使用谷歌趋势获得关键词的搜索数量作为投资者关注度的衡量指标，研究了投资者关注度对股票价格的短期影响。Thomas Dimp 和 Stephan Jank(2011)同样用谷歌趋势作为代理指标，研究了投资者关注度与股票指数的双向因果关系。Mah,Larch and Peter (2011)也用谷歌趋势作为代理指标，研究了投资者关注度与股市的关系，结论为搜索量的增加促进较高的收益率和交易量，并增强股票市场的流动性。Vlastakis and Markellos(2012)同样利用谷歌趋势获得搜索量，证明了投资者关注对股市交易量存在显著正相关。

国内在此方面的研究也在近几年来逐渐展开。宋双杰、曹阵、杨坤(2011)最早利用谷歌趋势提供的搜索量数据构建了投资者关注的直接衡量指标，系统解释了 IPO 市场上存在的三种异象。研究结果表明，投资者关注对资产价格有直接的影响。俞庆进、张兵(2012)首次使用了百度搜索量作为投资者关注的代理指标，他们实证检验了百度指数与创业板股票市场表现的相关性，验证了投资者关注对股价的正向影响作用，并发现非交易日的投资者关注度对下一交易日股票集合竞价时的价格有显著影响。王镇、郝刚（2013）将百度指数作为投资者关注度的度量指标，研究了投资者关注度对股票收益率的影响。研究发现，当期投资者的关注度促进股票收益率提高，前期投资者关注度与收益率有负向影响。杨欣，吕本富（2014）年利用百度指数构建了衡量投资者对突发事件关注度的指标，研究了突发事件、投资者关注与股市波动之间的关系，研究证实了突发事件关注度对股市波动的良好解释能力。

依据以上理论和实证探讨，本文拟探究在“一带一路”这一政策背景下，投资者关注与“一带一路”概念股之间的影响机制，并建立相对最优的模型实现对“一带一路”概念股成交量和收益率的预测。文章的具体结构框架如下图 1 所示：

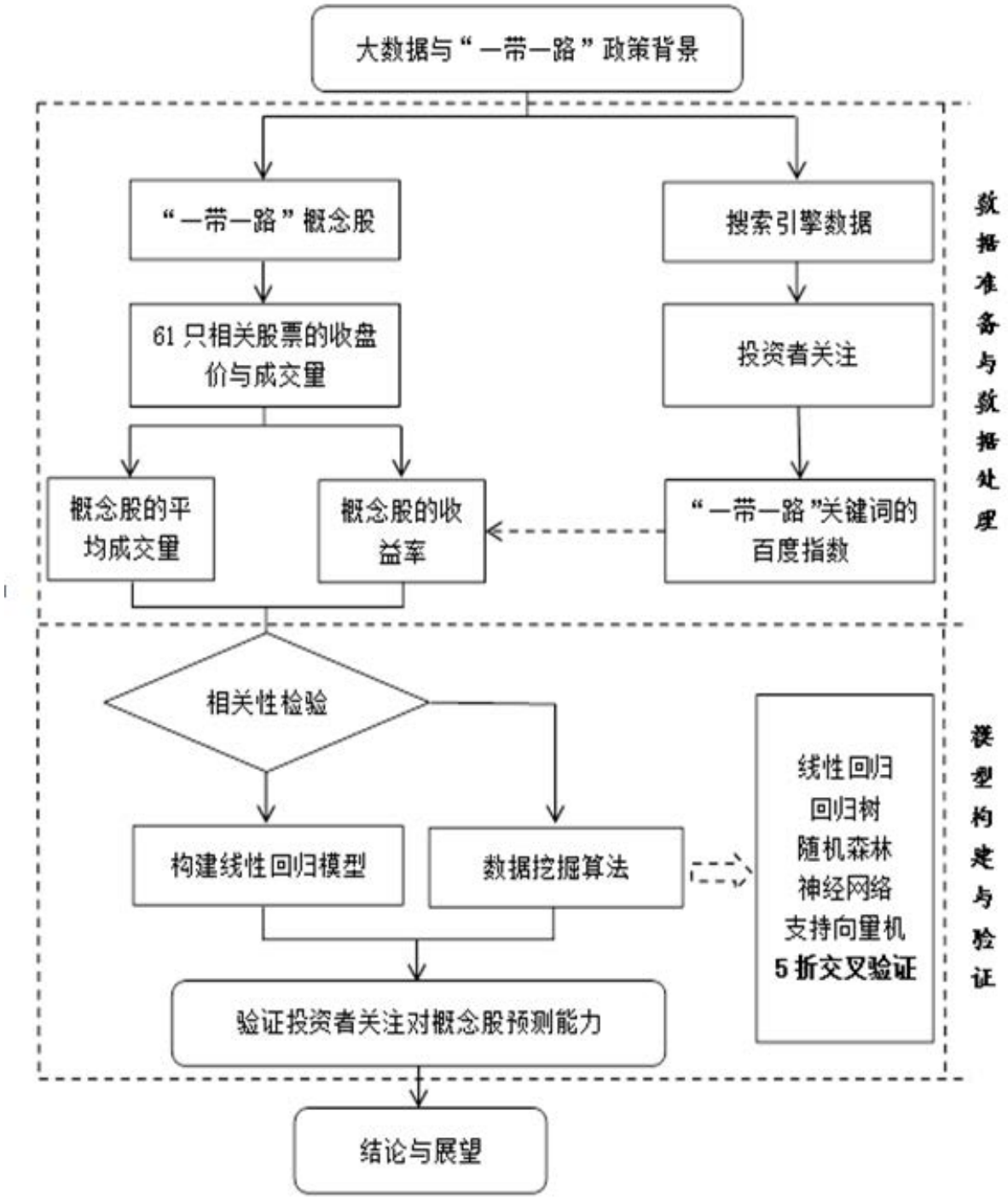

图 1　建模的结构框架

二、变量描述及数据准备

（一）研究对象

由于股市概念强大的广告效应，一只自身或许没多大吸引力的股票，一旦被纳入某个概念中，会引起投资者的广泛关注，从而概念股为研究投资者关注提供

了很好的素材。本文将前景较好、受到广泛关注的“一带一路”概念股作为研究对象。据分析，四行业将从“一带一路”中获利：第一是基建类，“一带一路”沿途国家基础建设差，建筑、高铁、电力设备均受益，建筑首当其冲；第二是能源建设，中国与中亚油气合作利好油气钻采、服务等；第三是交通运输；第四是旅游业，丝路建设带动相关区域旅游业发展。因此本文选取了与四行业相关的沪深两市的 61 只股票，如下表 2 所示。

表 2　四行业的“一带一路”概念股相关股票

细分方向	股票简称及代码
基建工程	中铁二局（600528.SH）中国交建（601800.SH）中国中冶（601618.SH）、中国电建（601669.SH）中国铁建（601186.SH）北新路桥（002307.SZ）、中工国际（002051.SZ）包钢股份（600010.SH）天山股份（000877.SZ）、祁连山（600720.SH）　宁夏建材（600449.SH）达刚路机（300103.SZ）、新疆城建（600545.SH）西部建设（002302.SZ）上峰水泥（000672.SZ）、*ST 建机（600984.SH）青松建化（600425.SH）北方国际（000065.SZ）、安阳钢铁（600569.SH）成都路桥（002628.SZ）福建水泥（600802.SH）、四川路桥（600039.SH）西部建设（002302.SZ）重庆路桥（600106.SH）、四川成渝（601107.SH）北方创业（600967.SH）博实股份（002698.SZ）、海南瑞泽(002596.SZ）　柳工（000528.SZ）　　青海华鼎（600243.SH）、青龙管业(002457.SZ）　浙富控股（002266.SZ）
交通运输	连云港（601008.SH）　南方航空（600029.SH）厦门港务（000905.SZ）、盐田港（000088.SZ）　日照港（600017.SH）　锦州港（600190.SH）、北部湾港（000582.SZ）大连港（601880.SH）福建高速（600033.SH）、宁波海运（600798.SH）天津港（600717.SH）五洲交通（600368.SH）、营口港（600317.SH）　招商轮船（601872.SH）重庆港九（600279.SH）、深赤湾 A（000022.SZ）
能源建设	渝开发（000514.SZ）　特变电工（600089.SH）广汇能源（600256.SH）、银星能源（000862.SZ）许继电气（000400.SZ）新疆浩源（002700.SZ）、神火股份（000933.SZ）
旅游业及其他	西安旅游（000610.SZ）曲江文旅（600706.SH）西安饮食（000721.SZ）、小商品城（600415.SH）渤海租赁（000415.SZ）兰石重装（603169.SH）、象屿股份（600057.SH）

资料来源：南方财富网（www.southmoney.com）

鉴于百度指数中收录“一带一路”关键词较晚，为保持一致本文选取了从 2014 年 6 月 6 日至 2015 年 5 月 22 日共 236 个交易日的 61 只股票每日的收盘价和成交量。数据来源国泰安 CSMAR 系列研究数据库。

（二）变量描述

1．“一带一路”概念股的交易量和收益率

首先鉴于我国上海证券交易所和深圳证券交易所市场的不同，本文将“一带

一路”概念股涉及的 61 只股票区分为沪深两市，并分别进行指标构建和分析。具体的区分如下表 3 所示,

表 3　沪深两市的“一带一路”概念股股票

	股票简称	个数
上海证券交易所	*ST 建机、安阳钢铁、包钢股份、北方创业、大连港、福建高速、福建水泥、广汇能源、锦州港、兰石重装、连云港、南方航空、宁波海运、宁夏建材、祁连山、青海华鼎、青松建化、曲江文旅、日照港、四川成渝、四川路桥、特变电工、天津港、五洲交通、象屿股份、小商品城、新疆城建、营口港、招商轮船、中国电建、中国交建、中国铁建、中国中冶、中铁二局、重庆港九、重庆路桥	36
深圳证券交易所	北部湾港、北方国际、北新路桥、博实股份、渤海租赁、成都路桥、达刚路机、海南瑞泽、柳工、青龙管业、厦门港务、上峰水泥、深赤湾 A、神火股份、天山股份、西安旅游、西安饮食、西部建设、新疆浩源、许继电气、盐田港、银星能源、渝开发、浙富控股、中工国际	25

据此可以求得“一带一路”概念股的每日平均成交量，即

沪市“一带一路”概念股平均成交量为 $HSP_t = \sum_i^n V_{it} P_{it} / n \qquad i = 1, 2, ..., 36$

深市“一带一路”概念股平均成交量为 $SSV_t = \sum_i^m V_{it} / m \qquad i = 1, 2, ..., 25$

其次本文采用成交量加权平均价的方法计算“一带一路”概念股的每日价格，公式为：

沪市“一带一路”概念股的平均价格为 $HSP_t = \sum_i^n V_{it} P_{it} / \sum_i^n V_{it} \qquad i = 1, 2, ..., 36$

深市“一带一路”概念股的平均价格为 $SSP_t = \sum_i^m V_{it} P_{it} / \sum_i^m V_{it} \qquad i = 1, 2, ..., 25$

其中，P 表示股票的收盘价，V 表示股票的成交量，n 表示沪市中股票数，m 表示深市中股票数。

由此可以分别得到沪深两市中的“一带一路”概念股的平均价格，如下图 2 所示：

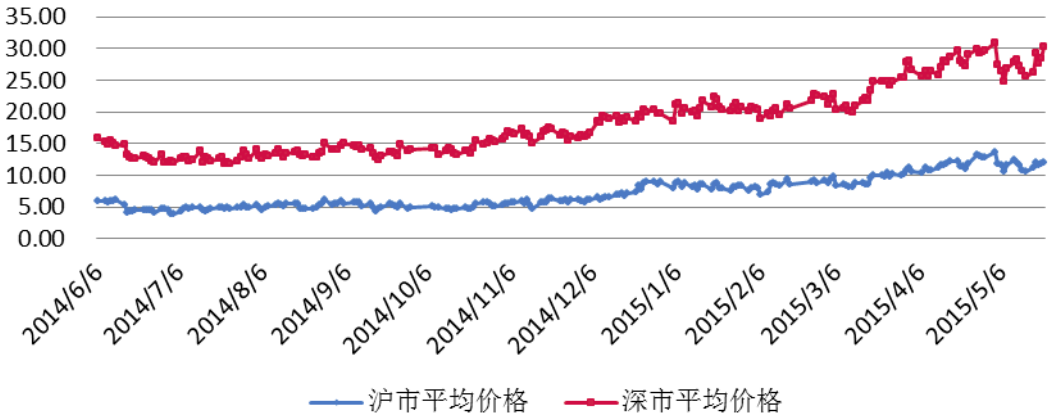

图 2　沪深两市中的“一带一路”概念股的平均价格

根据两市的平均价格，可以得到两市“一带一路”概念股的收益率，公式如下：

沪市“一带一路”概念股的收益率为 $HSR_t = \ln(HSP_t / HSP_{t-1})$

深市“一带一路”概念股的收益率为 $SSR_t = \ln(SSP_t / SSP_{t-1})$。

2．投资者关注度

本文采用“一带一路”关键词的百度指数来衡量投资者对“一带一路”概念股的关注度，检验投资者关注度对“一带一路”相关股票市场收益率的影响。百度指数是以百度搜索引擎的搜索量为数据基础，以关键词为统计对象，计算出各关键词在在百度网页上的搜索频数并加权平均获得的。百度搜索引擎是全球最大的中文搜素引擎，是我国用户最多的搜索引擎，所以选择百度指数作为一个衡量关注度的指标有较强的代表性。百度指数分为整体趋势、PC 趋势和移动趋势，整体趋势是 PC 趋势和移动趋势的加总。本文选取的为整体趋势。

本文搜集了自 2014 年 6 月 6 日到 2015 年 5 月 23 日“一带一路”关键词每日的百度指数（BI），为与股票价格数据相对应，仅保留交易日的搜索数据，共计 236 条记录。具体趋势如下图 3 所示：

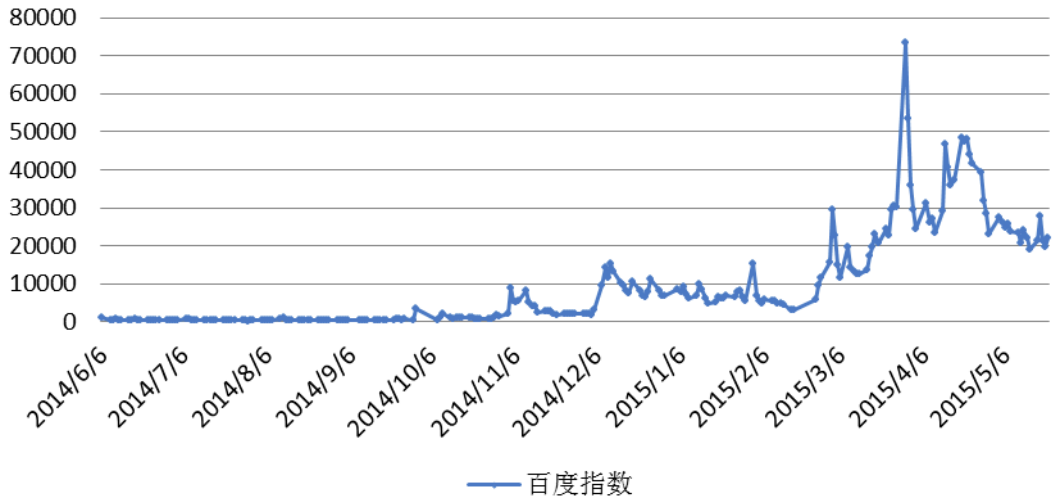

图 3 “一带一路”关键词每日的百度指数

3．其他变量

为加强模型的预测效果，本文添加了上证综指、深证成指和美元兑人民币汇率等变量。其中上证综指（SHI）和深证成指（SSI）代表了大盘走势，“一带一路”概念股的变动必然受到大盘波动的影响。大盘是整体股市活跃度和资金注入的反应。如果大盘一路上涨，那么概念股就一定会有所涨幅，只不过幅度的大小而已。如果大盘低迷，那么即使概念股会有所上涨，但是到达一定幅度和时间后也要补跌。

鉴于“一带一路”战略不仅仅局限在国内，更强调国际贸易，因此有必要增加汇率变量作为预测“一带一路”概念股收益率的变量。人民币升值使得国际热钱流入国内，有利于股市的上涨。

（三）数据预处理

通过上述指标构建，可以得到“一带一路”概念股的市场表现变量：平均价格（HSP、SSP）、平均成交量（HSV、SSV）以及收益率（HSR、SSR）；投资者关注度变量：百度指数（BI）；控制变量：上证综指（SHI），深证成指（SSI）和汇率（ER）。为消除数据之间的不同量级，对平均价格、平均成交量、百度指数、上证综指、深证成指和汇率进行对数化处理。

对整理后的数据进行简单描述统计，如下表4所示。

表 4　变量的简单描述统计

	数目	最小值	最大值	平均数	标准差
HSV	236	16.16	19.42	18.0007	0.70712
SSV	236	15.41	17.79	16.6666	0.53877
HSP	236	1.38	2.62	1.9177	0.33089
SSP	236	1.95	2.91	2.3788	0.24691
HSR	236	-0.25	0.21	0.0028	0.06551
SSR	236	-0.17	0.16	0.0026	0.05337
BI	236	5.55	11.21	7.9534	1.72402
SHI	236	7.61	8.45	7.9378	0.25144
SSI	236	8.88	9.68	9.1846	0.24122
ER	236	1.81	1.84	1.8226	0.0072

由上述描述统计可以看出：各变量的规模和量级基本近似，为下文模型的构建与应用奠定了基础。

三、投资者关注与概念股市场表现的相关性检验

通过对指标的构建和数据的预处理，本节的目的在于验证投资者关注的代理变量—百度指数能否作为预测“一带一路”概念股市场表现的显著指标：成交量和收益率，同时探索其中的影响机制。

（一）平稳性检验

为避免构建线性回归模型的时候出现伪回归的问题，需要对以上10个指标进行平稳性检验，通过检验得到所有指标ADF检验的结果都在5%的显著水平下拒绝存在单位根的原假设，即认为上述时间序列平稳，可以进行线性回归模型构建。

（二）投资者关注与概念股市场表现的相关性分析

首先分别计算投资者关注度与沪市和深市中“一带一路”概念股市场表现的Pearson 相关系数来简单探究投资者关注度与“一带一路”概念股市场表现的联系，具体结果如下表所示：

表 5　投资者关注与沪市概念股市场表现的 Pearson 相关系数

	BI	HSP	SHI	HSR	ER	HSV
BI	1	.904**	.934**	.060	.423**	.769**
HSP	.904**	1	.966**	.130*	.497**	.640**
SHI	.934**	.966**	1	.044	.431**	.744**
HSR	.060	.130*	.044	1	-.023	.000
ER	.423**	.497**	.431**	-.023	1	-.062
HSV	.769**	.640**	.744**	.000	-.062	1

注：**表示该相关系数在1%的显著水平下显著，*表示其在5%的显著水平下显著

表 6　投资者关注与深市概念股市场表现的 Pearson 相关系数

	BI	ER	SSP	SSI	SSR	SSV
BI	1	.423**	.947**	.925**	.046	.751**
ER	.423**	1	.393**	.473**	-.038	.034
SSP	.947**	.393**	1	.946**	.153*	.727**
SSI	.925**	.473**	.946**	1	.066	.739**
SSR	.046	-.038	.153*	.066	1	.060
SSV	.751**	.034	.727**	.739**	.060	1

注：**表示该相关系数在1%的显著水平下显著,*表示其在5%的显著水平下显著

由上述两表可以看出：两市中投资者关注度的代理指标—“一带一路”关键词的百度指数（BI）与“一带一路”概念股的平均成交量具有较强的相关关系，但是与概念股的收益率指标的相关关系较弱，其中沪市中为0.06，深市中为0.046，即相关关系并不显著。综上，初步认为投资者关注度影响“一带一路”概念股的平均成交量，但无法直接影响概念股的收益率。为进一步探究其关系，拟通过建立回归模型的方式验证其结论，即分别建立两市中“一带一路”概念股平均成交量和收益率与投资者关注度的一元回归方程，结果如下表所示：

表 7　投资者关注与概念股市场表现的回归结果

	成交量指标		收益率指标	
被解释变量	沪市（HSV）	深市（SSV）	沪市（HSR）	深市（SSR）
Intercept	15.4931**	14.7995**	-0.0152	-0.0087
BI	0.3253**	0.2348**	0.0023	0.0014
R^2	0.5909	0.5643	0.0036	0.0021
F-statistic	338.01**	303.04**	0.8349	0.4877

注：**表示在1%的显著水平下显著，*表示在5%的显著水平下显著

如上表所示，投资者关注度的代理指标—“一带一路”关键词的百度指数能显著解释沪深两市中“一带一路”概念股的平均成交量，但无法有效解释其收益率。其中在投资者关注度与成交量的线性模型中，两方程的F检验以及系数显著性检验均在1%的显著水平下显著，同时拟合优度R^2分别为0.5909和0.5643，说明投资者关注度指标对“一带一路”概念股成交量指标的解释能力较强，能较好预测该指标的走势。但在分析投资者关注度与收益率指标时，回归方程的显著性和系数的显著性检验均未通过，即认为无法通过该关键词的百度指数解释“一带一路”概念股的收益率趋势。

综上，我们认为投资者关注度的代理指标—“一带一路”关键词的百度指数与“一带一路”概念股的市场表现有着密切联系：投资者者关注度与概念股的平均成交量具有较强的线性相关关系，即投资者的关注强度能够直接影响概念股成

交量的变化；但投资者关注度与概念股的收益率之间没有形成显著的影响模式，需要通过其他方式进行研究。针对该问题本文拟结合 5 折交叉验证技术，采用线性回归、回归树、随机森林、支持向量机和神经网络等 5 种模型进行预测，并对各种模型的预测结果进行对比研究。

四、预测模型构建

（一）两市“一带一路”概念股平均成交量预测

通过上一章节对投资者关注度代理变量与“一带一路”概念股平均成交量和收益率等市场表现的相关性分析，对于概念股平均成交量的预测可采用线性回归的模式，即将两市概念股的成交量指标作为被解释变量，“一带一路”关键词的百度指数作为解释变量，而上证综指或深证成指、汇率作为控制变量建立多元回归模型，分析结果如下：

表 8　投资者关注与概念股平均成交量的回归结果

被解释变量	沪市（HSV）	深市（SSV）
Intercept	95.3057**	58.9655**
BI	0.2684**	0.1346**
SHI	0.9633**	——
SSI	——	1.1793**
ER	-47.7817**	-29.7383**
R^2	0.7880	0.6999
F-statistic	287.52**	180.3288**

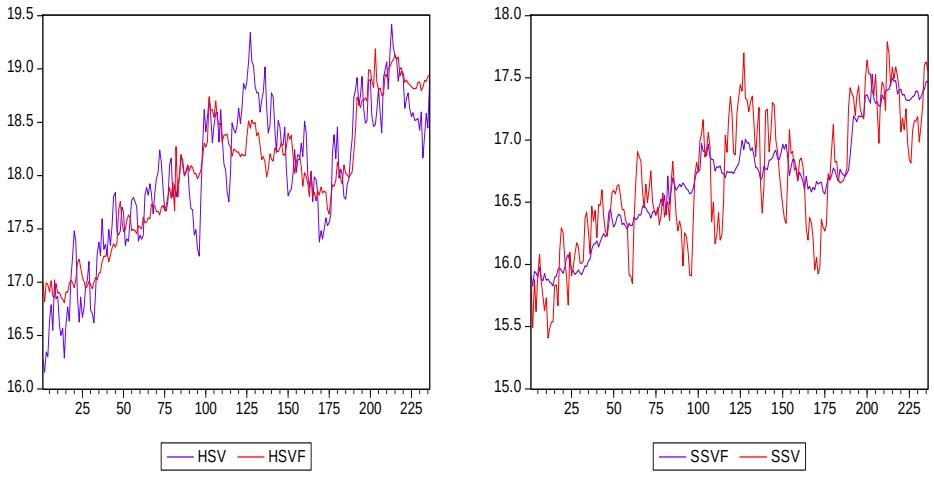

图 4　概念股平均成交量的预测结果

由表 8 和图 4 可以看出：通过分别构建两市中“一带一路”概念股的平均成交量和投资者关注度代理指标之间的多元线性回归模型，均能显著解释概念股成

交量指标的变动，预测结果与实际值的走势都基本一致，适合对成交量指标进行预测。同时根据回归方程可以得到：“一带一路”关键词的百度指数每变动 1%，沪市中概念股成交量变动 0.27%，深市中概念股平均成交量变动 0.13%。因此认为投资者关注的代理指标对沪深两市概念股平均成交量的预测能力均较强。

（二）两市“一带一路”概念股收益率预测

针对投资者关注度与概念股的收益率之间没有形成显著的线性影响模式，本文拟结合 5 折交叉验证技术，采用线性回归、回归树、随机森林、支持向量机和神经网络等 5 种模型进行预测，并对各种模型的预测结果进行对比研究，增强了其结果的可靠性。

为对比研究 5 种算法之间的优劣，首先定义均方误差 MSE：$MSE=\sum(y-\hat{y})^2$ 其中 y 为实际值，$\hat{y}$ 为预测值，通过比较 MSE 的大小来判定模型的稳定性，即 MSE 的值越小，模型的稳定性越好。其次定义参数 NMSE：$NMSE=\frac{MSE_i}{MSE}$，即将线性回归模型作为基准模型，构建其他模型均方误差 MSE 与线性回归模型 MSE 的比值，作为衡量模型拟合度的指标，NMSE 小于 1 表明模型的拟合度优于线性回归模型，且值越小，模型的拟合度越好。

模型构建时要充分考虑影响“一带一路”概念股收益率的因素，因此本文将概念股的平均价格、上证综指或深证成指、汇率等作为输入变量，概念股收益率作为输出变量，同时对比加入投资者关注度指标前后的模型，进而研究投资者关注度对“一带一路”概念股收益率的预测能力。

依据以上思路，分别对沪深两市中“一带一路”概念股的收益率进行预测，通过设计交叉验证随机建立 5 个训练集和测试集，同时每个数据集分别建立线性回归、回归树、随机森林、神经网络、支持向量机等 5 种模型，利用 R 软件对模型进行模拟，得到每个模型的训练集和测试集的平均均方误差（MSE）和平均 NMSE 来评估模型的稳定性和拟合度。沪市“一带一路”概念股收益率预测的均方误差（MSE）和 NMSE 如下表 9 所示：

表 9　沪市中概念股收益率预测的 MSE 和 NMSE（未加百度指数）

	训练集		测试集	
	MSE	NMSE	MSE	NMSE
线性回归	0.2913	1.0000	0.2811	1.0000
回归树	0.0053	0.0464	0.0054	0.0424
随机森林	0.0242	0.0324	0.0189	0.0537
神经网络	0.0145	0.0546	0.0076	0.0731
支持向量机	0.0325	0.0535	0.0562	0.0386

表 10　沪市中概念股收益率预测的 MSE 和 NMSE（加入百度指数）

	训练集		测试集	
	MSE	NMSE	MSE	NMSE
线性回归	0.2923	1.0000	0.2820	1.0000
回归树	0.0044	0.0150	3.370e-03	0.0133
随机森林	0.0045	0.0133	7.532e-03	0.0155
神经网络	0.0038	0.0155	3.532e-03	0.0272
支持向量机	0.0138	0.0272	0.0193	0.0129

由表 9 和表 10 可以看出：

（1）从整体的训练集和测试集拟合的效果来看，回归树、随机森林、神经网络和支持向量机四种模型均比线性回归模型表现优异，说明四种算法能更好地描绘和预测“一带一路”概念股收益率的变化。

（2）从投资者关注度代理指标进入模型前后的效果来看，除线性回归模型不显著外，其他四个模型在投资者关注度代理指标加入后不论是 MSE 还是 NMSE 都有了较为明显的降低，说明对概念股收益率的预测更为精准和稳定，这也证实了投资者关注度代理指标—“一带一路”关键词百度指数对“一带一路”概念股收益率的预测能力。

（3）从模型之间的对比来看，由于测试集更能体现对输出变量的预测能力，因此我们主要分析不同模型在测试集上的 MSE 和 NMSE，其中回归树模型和神经网络的平均 MSE 数值相对最小，说明两者在模型稳定性上表现较好；回归树和支持向量机模型的平均 NMSE 数值最小，说明两者对输出变量的拟合效果更好。综合平均 MSE 数值和平均 NMSE 数值，认为回归树模型对于沪市中“一带一路”概念股的收益率预测更好。

同理可以得到深市“一带一路”概念股收益率预测的平均均方误差（MSE）和平均 NMSE 如下表 11 所示：

表 11　深市中概念股收益率预测的 MSE 和 NMSE（未加百度指数）

	训练集		测试集	
	MSE	NMSE	MSE	NMSE
线性回归	0.3056	1.0000	0.3190	1.0000
回归树	0.0346	0.0894	0.0725	0.1190
随机森林	0.0124	0.0356	0.0642	0.0489
神经网络	0.0186	0.0349	0.0128	0.0396
支持向量机	0.0086	0.0234	0.0105	0.0375

表 12　深市中概念股收益率预测的 MSE 和 NMSE（加入百度指数）

	训练集		测试集	
	平均 MSE	平均 NMSE	MSE	NMSE
线性回归	0.3453	1.0000	0.3525	1.0000
回归树	0.0132	0.0656	0.0432	0.0632
随机森林	0.0032	0.0245	0.0234	0.0345
神经网络	0.0046	0.0144	4.324e-03	0.0224
支持向量机	0.0023	0.0121	2.342e-03	0.0219

由表 11 和表 12，通过对深市“一带一路”概念股的分析，同样可以得到：回归树、随机森林、神经网络和支持向量机等模型表现比线性回归模型更佳，同时加入投资者关注度指标后模型的预测精度和稳定性也得到了较大提高。

另外通过分析测试集中的平均 MSE 和平均 NMSE，支持向量机模型的 MSE 数据相对最小，同时其 NMSE 数值也相对最小，其次为神经网络模型。所以说明支持向量机模型对于沪市中“一带一路”概念股的收益率预测更好。支持向量机模型在解决小样本、非线性模式识别中体现了较大的优势，能够根据有限的样本信息在模型中特定训练样本的学习精度并寻求最佳的结果。

五、结论建议

（一）研究结论

本文首先对投资者关注度与股票市场的关系方面的研究进行了回顾，得出结论：投资者关注度促进股票交易，进而影响股票价格。然后基于投资者关注—“一带一路”关键词的百度指数，探究其与“一带一路”概念股市场表现（平均成交量和收益率）的相关关系，并采用线性回归模型对“一带一路”概念股平均成交量进行了拟合和预测，采用回归模型、支持向量机、BP 神经网络、随机森林和回归树 5 种模型对“一带一路”概念股的收益率进行了拟合和预测，得到结论如下：

（1）通过选用投资者关注的代理指标—“一带一路”关键词的百度指数，实证探究了投资者关注对“一带一路”概念股市场表现的影响，证实投资者关注的强度对“一带一路”概念股具有较强的相关作用。

（2）通过建立线性回归模型分析投资者关注度对“一带一路”概念股平均成交量的预测能力，证实投资者关注的代理指标对沪深两市概念股平均成交量的预测能力均较强，关键词的百度指数每变动 1%，沪市中概念股成交量变动 0.27%，深市中概念股平均成交量变动 0.13%。

（3）通过结合 5 折交叉验证技术，采用回归模型、支持向量机、BP 神经网络、随机森林和回归树 5 种模型对“ 一带一路 ”概念股的收益率进行了拟合和预测。同时计算各个模型训练集和测试集的平均 MSE 和平均 NMSE，得到在沪市“ 一带一路 ”概念股的收益率预测中回归树模型表现最佳；在深市“ 一带一路 ”概念股的收益率预测中支持向量机模型表现最佳。综上较好地结合投资者关注度实现了对“ 一带一路 ”概念股收益率的预测。

（二）策略建议

本文引入百度指数衡量投资者关注度并预测“ 一带一路 ”概念股的平均成交量和收益率具有重要的意义。网络搜索指数可以反应人们的关注重点，衡量投资者关注度，并用来做关注度与股市关系研究，丰富了理论研究。另外，这一研究还对投资者、股票发行人、监管部门具有一定的指导意义。

理论意义在于，借助对“ 一带一路 ”概念股的研究，实现对投资者关注度的衡量由间接衡量到直接衡量的过渡，有效地使用搜索引擎产生的海量数据，更精确的获得投资者的关注度。网络搜索指数量化了投资者关注度，促进了网络搜索数据在金融领域的应用。同时验证了我国概念股变动与投资者关注的关联性，丰富了投资者关注和概念股的相关研究，并为后来者的研究提供了参考。

对个人投资者而言，投资者关注度对股市影响的研究为他们在股市的投资决策提供了参考。我国股市中，个人投资者数量非常多，但他们的投资行为不够成熟，他们由于有限关注及有限的信息处理能力，经常受媒体、炒作的影响，盲目跟风，不能合理进行决策。而本文投资者关注度与股市变动的关系研究，可以帮助投资者合理有效地投资并管理风险，从而获取利润。

对政府监管部门，我国证券市场起步较晚，市场还不成熟，在监管上有一些缺陷。通过使用网络搜索数据可以了解投资者的关注情况，从而加强投资者关注度较高的股票的监管，提高透明度，规范市场，促进我国证券市场的稳定。

对于上市公司而言，投资者关注度对股市影响的研究可以知道他们进行最小化融资成本。由于个体有限关注，投资者关注度是稀缺资源，上市公司为了获得更高收益就要提高自身关注度，抓住投资者的注意力。通过的宣传吸引投资者的关注，在股票市场中实现最小化的融资成本。另外，上市公司可以通过有效的管理投资者关系，披露信息是选取最恰当的时间，在发布“ 利好 ”消息时能够获取最大的收益，当发布“ 利空 ”消息时达到最小损失。

（三）创新与展望

本文的主要创新之处在于：

1. 以“一带一路”概念股为研究重心，结合我国“政策市”和个人投资者盲目跟风的现状，实证探究了投资者关注与“一带一路”概念股之间的相关性，并实现了对概念股市场表现的预测。

2. 根据我国搜索引擎市场的现状，选用“一带一路”关键词的百度指数作为投资者对“一带一路”概念股的关注度，降低了关键词选取的不确定性，增强了指标的代表性。

3. 借助线性回归、回归树、支持向量机、神经网络和随机森林等多个模型对概念股的收益率进行预测，增强了预测结果的可信度和稳定性。

本文通过实证分析验证了投资者关注与概念股市场表现的相关关系，同时证实了其在概念股预测方面的优良特性。在之后的研究中可以进一步完善投资者关注的代理指标，借助大数据时代的诸多网络平台，例如微博、股吧等，利用文本挖掘等数据挖掘技术获取数据，从而实现对概念股，以及普通个股和大盘的实时精确的预测，为个人投资者投资风险的减低和股市的成熟稳定提供保障。

参考文献

[1] Hsieh, J., Walkling, R.. The history and performance of concept stocks[J]. Journal of Banking & Finance, 2006: 2433–2469.

[2]舒志斌，兰宜生.我国高科技概念股市场定价的实证分析[J].汕头大学学报，2001.

[3]赖以容.中国概念股研究［D］.复旦大学，2010.

[4]Brad M.Barber,Terrance Odean.Boys Will Be Boys: Gender,Overconfidence, and Common Stock Investment[J].The Quarterly Journal of Economics,2001.

[5]Da,Z,J.Engelberg,P.Gao.In search of attention[J].The Journal of Finance,2011.

[6]Thomas Dimp,Stephan Jank.Can internet search queries help to predict stock[J].Centre for Economics Working Paper,2011.

[7] Bank M., M. Larch, and G. Peter. Google search volume and its influence on liquidity and returns of German stocks [J]. Financial Markets and Portfolio Management, 2011: 1-26.

[8]Nikolaos Vlastakis, Raphael N.Markellos. Information Demand and Stock Market Volatility[J].Journal of Banking and Finance, 2012.

[9]宋双杰，曹阵，杨坤.投资者关注与 IPO 异象———来自网络搜索量的经验证据[J].经济研究，2011.

[10]俞庆进,张兵.投资者有限关注与股票收益———以百度指数作为关注度的一项实证研究[J].金融研究，2012.

[11]王镇,郝刚.投资者关注度对股票收益率的影响———基于百度指数指标[J].新疆财经，2013.

[12]董倩，孙娜娜，李伟. 基于网络搜索数据的房地产价格预测[J].统计研究，2014.