Statistics and Application 统计学与应用, 2016, 5(2), 196-202
Published Online June 2016 in Hans. http://www.hanspub.org/journal/sa
http://dx.doi.org/10.12677/sa.2016.52019

Hans汉斯

The Statistical Modeling and Analysis of Electricity Business Big Data Network Index

Fei Wang, Hui Mei, Ruili Zhang, Chengping Gong, Xiong Xie, Liyun Su*

College of Mathematics and Statistics, Chongqing University of Technology, Chongqing
Email: *1093464745@qq.com

Received: Jun. 9th, 2016; accepted: Jun. 24th, 2016; published: Jun. 30th, 2016

Abstract

This article first briefly describes the present situation of the big e-commerce data age, and chooses Baidu and Ali index to explain its concept. Based on the Ali index of smartphone industry, we analyze the data from January 1, 2016 to March 3, 2016. Using SAS software, the stationarity and randomness of the sequence are researched. Then we carry on model fitting and forecast of two stationary white noise sequences to conclude the model expression and calculate the forecast error. The result shows that the conclusion can provide a certain reference value for the purchasers and suppliers of 1688.

Keywords

Electricity Business Big Data, Ali Index, Time Series Analysis, Baidu Index

电商大数据网络指数统计建模与分析

王　飞，梅　辉，张瑞丽，龚铖萍，谢　熊，苏理云*

重庆理工大学数学与统计学院，重庆
Email: *1093464745@qq.com

收稿日期：2016年6月9日；录用日期：2016年6月24日；发布日期：2016年6月30日

摘　要

本文首先简述电商大数据时代的现状，选取百度指数和阿里指数，对其概念进行说明。基于智能手机行

文章引用: 王飞, 梅辉, 张瑞丽, 龚铖萍, 谢熊, 苏理云. 电商大数据网络指数统计建模与分析[J]. 统计学与应用, 2016, 5(2): 196-202. http://dx.doi.org/10.12677/sa.2016.52019

业的阿里指数分析了自2016年1月1日至2016年3月3日的数据，运用SAS软件编程，研究了序列的平稳性和随机性，并对得到的两个平稳非白噪声序列进行模型拟合和预测，最后得出模型表达式并计算了预测误差。结果发现，所得出的结论能为1688采购商和1688供应商提供一定的参考价值。

关键词

电商大数据，阿里指数，时间序列分析，百度指数

1. 引言

随着社会的不断发展，人民生活消费水平的显著提高，科学技术的进步使网络成为了生活中不可或缺的因素。因此，电子商务迎来了大好的发展前景，经过近二十年的迅猛发展，已逐渐满足了绝大多数人的生活需求。

孙易冰，赵子东等参照官方 CPI 的制度方法，研究出一种基于网络爬虫技术的价格指数计算模型。并把该模型试算值与官方数据进行比较，同时对原始数据的特征挖掘，发现该模型具有时效性强和灵敏度高的特点[1]。南豪峰从社会研究方法论这一视角审视，认为目前大数据的研究主要集中于哲学基础和范式转变方面，缺乏设计、信度和效度、伦理等方面的研究[2]。张国发，邵树琴利用协整理论对淘宝搜索指数与成交指数进行了协整性研究，建立了误差修正模型，分析了成交指数变动的长期趋势和短期波动[3]。赵萱基于淘宝指数的比较观察，分析了中国智能手机市场，研究发现得出了中国智能手机市场上销量最大的是中档次手机的结论[4]。

近年来，随着大数据时代的来临，人们面对纷繁复杂的数据，已经辨别不出真正的有价值的数据，不能从中获得准确的信息。而互联网搜索引擎所提供的相关数据(如阿里指数和百度指数)则能把握人们对某事物的关注情况，准确反映了电子商务这一领域的情况[5] [6]。若基于这些网络指数，对电商大数据进行统计建模分析，将能间接摸清消费者的心理活动购买行为。基于此，本文将利用阿里指数所提供的智能手机数据作为本文的研究对象。

2. 电商大数据时代的现状

2.1. 电商时代发展简述

1996 年前后，电子商务开始在中国起步。在 21 世纪初期，中国的电子商务迎来了第一次发展高潮，以当当网、卓越网为第一梯队，京东商城亦在中关村发迹。2010~2011 年，团购网站也是如雨后春笋搬发展起来。自 2013 年开始，在中国互联网巨头中，以百度和阿里为代表开启了并购的风潮。2015 年，根据全球十大电商市场数据盘点，中国的电商市场遥遥领先，阿里巴巴进军曼谷，口碑网投 100 亿推“全面开店”计划等等都说明了目前我国的电商发展还处于黄金时期。

2.2. 大数据时代电商面临的挑战

大数据时代的来临，对电子商务既是极大机遇，也是一系列挑战。

在数据集非常庞大的情况下，如何处理这些数据是基础。首先必须要对数据进行筛选和清洗，以便得到有价值和相关联的数据。然后才能进行客观实际的全面分析，摸清消费者的喜好。最后立足于这些喜好，设计并生产出迎合市场的产品。只有以消费者需求为中心，顺应市场，就能够真正实现销量与利润的双增长，最终促使电商得到更好的发展。目前一些企业已开始重视并开发大数据。

3. 网络指数的选取与概念说明

3.1. 网络指数的选取

网络指数有很多种，由于百度和阿里巴巴作为我国互联网巨头，而且电子商务的运营方式基本相同，数据采集和指数定义也相似，因此笔者只选择其中最具有代表性的百度指数和阿里指数作为研究对象。

百度指数，即百度搜索次数。如某词汇的百度指数为 10 万，那么就意味着网民通过百度搜索该词汇 10 万次。

阿里指数比较真实的反映了电子商务对数据的挖掘与分析，通过这些数据可以预测顾客的需求以及预测某一物品未来的需求程度。再大一点来说，可以反映电子商务的发展状况。与此相同，网络指数是对生活中各种现象的分析，是对社会民生的观察分析，可以预测人们生活状况和社会发展状况，对国家的各种政策的制定有重大的影响。

3.2. 两种网络指数的概念说明

3.2.1. 百度指数

百度指数是以全球最权威的中文检索数据为基础，根据搜索量和媒体检索量进行过滤和加权，通过科学、标准的运算，并且以直观的图形界面展现，帮助用户最大化的获取有价值信息。

百度指数能综合反映关键词在过去 1 天用户对它的关注和媒体对它的关注的一个参考值。任意关键词的百度指数都是该关键词在比较期的数值除以该关键词在基期的数值。比较期的数值和基期的数值是通过当天的用户搜索量和百度新闻中过去 30 天相关的新闻数量相比得来。

3.2.2. 阿里指数

阿里指数是阿里巴巴为了解电子商务平台市场动向而研发的数据分析平台，2012 年 11 月 26 日，阿里指数正式上线。它是由阿里巴巴网站每日运营的 5 项基本数据统计计算得出，包括每天网站浏览量、每天浏览的人次、每天新增供求产品数、新增公司数和产品数这 5 项。

阿里指数主要包括如下三项：

1) 淘宝采购指数：根据在淘宝市场(淘宝集市 + 天猫)里所在行业的成交量计算而成的一个综合数值，指数越高表示在淘宝市场的采购量越多。

2) 1688 采购指数：根据在 1688 市场里所在行业的搜索频繁程度计算而成的一个综合数值，指数越高表示在 1688 市场的采购量越多。

3) 1688 供应指数：根据在 1688 市场里所在行业已上网供应产品数计算而成的一个综合数值，指数越高表示 1688 市场的供应产品越多。

4. 实证分析

4.1. 数据来源

在阿里指数官网上搜集并整理了智能手机行业 2016 年 1 月 1 日至 2016 年 3 月 3 日阿里指数里的淘宝采购指数、1688 采购指数和 1688 供应指数[7]。

4.2. 统计建模分析

三组指数的数据是随时间变化的，因此笔者利用 SAS 软件编程[8]对数据进行时间序列分析。

4.2.1. 绘制时序图

图 1 中最上面红色表示的是淘宝采购指数的时间序列，绿色表示的是 1688 采购指数的时间序列，蓝

色表示的是 1688 供应指数的时间序列。

4.2.2. 平稳性检验

为了判断序列是否平稳，除了需要考察时序图的性质，还必须对自相关图进行检验。下面笔者着重对淘宝采购指数的时间序列的平稳性进行分析。

图 2 横轴表示自相关系数，纵轴表示延迟阶数。从图 2 中发现序列的自相关系数衰减到零的速度非常缓慢，在比较长的延迟时期里，自相关系数一直为正，而后才开始变为负数。在自相关图中显出了一定的三角对称性，这是具有单调趋势的非平稳序列的一种典型的自相关图形式。这和图 1 中红色的曲线显示的显著的单调性是一致的，说明该指数的序列是非平稳序列。用同样的方法可以得出 1688 采购指数和 1688 采购指数的序列是平稳时间序列，从图 1 中也可以直观的看出，序列基本上在一个数值上随机波动。现将数据的简单描述汇总至表 1。

4.2.3. 纯随机性检验

纯随机检验又称白噪声检验，此处显著性水平 α 取 0.05。检验结果显示，无论是 6 阶，12 阶还是 24 阶，检验统计量的 P 值都非常小(<0.0001)，所以可以以相当大的把握(置信水平 > 99.99%)断定 1688 采购指数和 1688 供应指数的时间序列为非白噪声序列(图 3、图 4)。

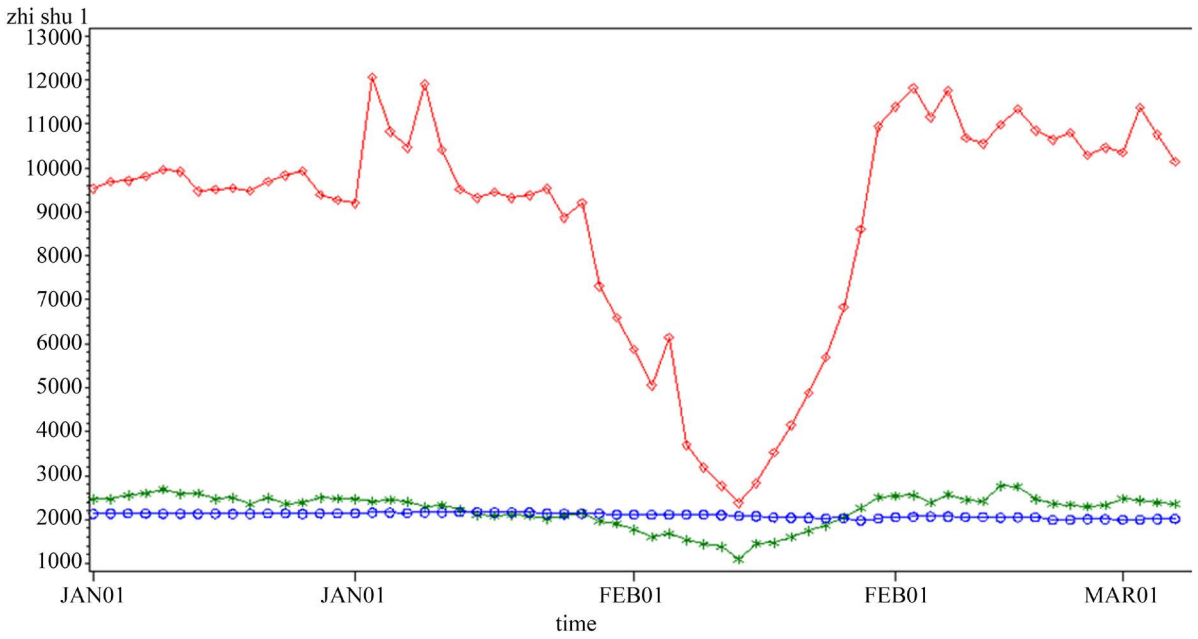

Figure 1. Time series figure
图 1. 时序图

Table 1. Data analysis
表 1. 数据分析

指数	均值	方差	平稳性
淘宝采购指数	8892.3	2560.4	非平稳
1688 采购指数	2226.0	378.1	平稳
1688 供应指数	2101.7	55.8	平稳

4.3. 模型拟合和预测

4.3.1. 模型拟合

1688 采购指数模型定阶输出结果见表 2，所以该模型为 AR(1)模型时拟合最好。同时 1688 供应指数模型也同样为 AR(1)模型时拟合最好。

经过检验，所有的 P 值均小于 0.05，可以认为所有参数均显著。

Autocorrelations

Lag	Covariance	Correlation	−1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	6555887	1.00000	*********************	0
1	6169901	0.94112	********************	0.125988
2	5718315	0.87224	******************	0.209740
3	5124822	0.78171	*****************	0.261043
4	4370863	0.66671	************	0.295876
5	3599198	0.54900	**********	0.318832
6	2776106	0.42345	********	0.333500
7	1978944	0.30186	******	0.341928
8	1115884	0.17021	***	0.346132
9	293400	0.04475	*	0.347458
10	−490219	−0.07478	*	0.347549
11	−1088528	−0.16604	***	0.347805
12	−1560783	−0.23807	*****	0.349060
13	−2016545	−0.30759	******	0.351628
14	−2312229	−0.35270	*******	0.355874
15	−2635749	−0.40204	********	0.361380
16	−2831338	−0.43188	*********	0.368411
17	−2988228	−0.45581	*********	0.376361
18	−3149928	−0.48047	**********	0.385024
19	−3084530	−0.47050	**********	0.394426
20	−2981654	−0.45481	**********	0.403236
21	−2811990	−0.42893	*********	0.411298
22	−2623398	−0.40016	**********	0.418338
23	−2377561	−0.36266	**********	0.424370
24	−2115048	−0.32262	**********	0.429262

"." marks two standard errors

Figure 2. Autocorrelations of Taobao procure index
图 2. 淘宝采购指数自相关图

Autocorrelation Check for White Noise

To Lag	Chi-Square	DF	Pr> ChiSq	Autocorrelations					
6	226.15	6	<0.0001	0.938	0.873	0.790	0.688	0.587	0.486
12	244.80	12	<0.0001	0.381	0.267	0.154	0.068	−0.010	−0.082
18	284.20	18	<0.0001	−0.136	−0.188	−0.248	−0.298	−0.334	−0.364
24	385.90	24	<0.0001	−0.395	−0.414	−0.425	−0.431	−0.419	−0.398

Figure 3. Autocorrelations of check for white noise of 1688 procure index
图 3. 1688 采购指数白噪声检验

Autocorrelation Check for White Noise

To Lag	Chi-Square	DF	Pr> ChiSq	Autocorrelations					
6	269.15	6	<0.0001	0.948	0.893	0.838	0.777	0.715	0.649
12	381.09	12	<0.0001	0.586	0.527	0.493	0.466	0.453	0.430
18	431.36	18	<0.0001	0.402	0.373	0.336	0.296	0.241	0.179
24	439.37	24	<0.0001	0.090	0.024	−0.037	−0.096	−0.149	−0.192

Figure 4. Autocorrelations of check for white noise of 1688 supply index
图 4. 1688 供应指数白噪声检验

Table 2. Model fitting of 1688 procure index
表 2. 1688 采购指数模型拟合

Lags	MA 0	MA 1	MA 2	MA 3	MA 4	MA 5
AR 0	11.80232	11.76387	11.6846	11.56017	11.4251	11.27005
AR 1	9.708679	9.774373	9.825117	9.852718	9.912487	9.955944
AR 2	0.768793	9.795009	9.84508	9.894549	9.901933	9.940308
AR 3	9.799704	9.833512	9.899254	9.950976	9.962964	10.00535
AR 4	9.808772	9.873049	9.934122	9.989545	10.01521	10.05808
AR 5	9.857694	9.918336	9.977967	9.995878	10.05778	10.07904

Table 3. Model error analysis
表 3. 模型误差分析

1	日期	3.4	3.5	3.6	3.7	3.8
2	真实值	2340	2342	2176	2317	2333
3	预测值	2366	2369	2272	2375	2378
4	相对误差	0.011	0.012	0.044	0.025	0.019
5	真实值	2044	2049	2054	2074	2043
6	预测值	2024	2053	2025	2083	2058
7	相对误差	0.009	0.002	0.014	0.004	0.007

由此得到 1688 采购指数 AR(1)模型为

$$x_t = 2438.1 + 0.95641x_{t-1} + \varepsilon_t \quad (1)$$

1688 供应指数 AR(1)模型为

$$x_t = 2133.7 + 0.99746x_{t-1} + \varepsilon_t \quad (2)$$

4.3.2. 模型预测

通过软件计算得到两个指数 AR(1)模型 2016 年 3 月 3 日之后 5 天的指数预测值和真实值比较如表 3 所示。2~4 行为 1688 采购指数模型，5~7 行为 1688 供应指数模型。

模型拟合预测的相对误差 = |预测值 − 真实值|/真实值，从表 3 可知，相对误差均较小，说明模型拟合不错，预测效果比较好。

5. 结语

首先，淘宝采购指数的序列是非平稳时间序列，而 1688 采购指数和 1688 供应指数是平稳时间序列，进一步分析发现这两个还是非白噪声序列，说明还存在可提取有价值的信息，能够进行模型拟合和预测。

其次，笔者进行了模型拟合和未来五天的预测，得到了较好的预测效果。

最后，根据本文研究所得的结论和预测的结果，对 1688 采购商和 1688 供应商具有一定的参考价值。比如说淘宝采购商可以在淘宝采购指数较低时进行商品采购，这时采购人数一般较少。

当然，本文的研究还有很多不完善的地方。首先，本文所选取的样本量不是太多；其次虽研究了阿里指数和百度指数，但对智能手机只单纯考虑了阿里指数的数据，未能将两者有机地结合起来。这些不足对未来该领域的研究也许提供了一定的方向。

基金项目

重庆市教育委员会人文社会科学研究一般项目(15SKG136)，全国统计科学研究项目(2014LY069)，重庆理工大学科研立项重点项目(KLA15004)。

参考文献 (References)

[1] 孙易冰, 赵子东, 刘洪波. 一种基于网络爬虫技术的价格指数计算模型[J]. 统计研究, 2014, 31(10): 74-80.

[2] 南豪峰. 大数据在社会研究中的应用现状[J]. 江汉大学学报(社会科学版), 2015(5): 12-17.

[3] 张国发, 邵树琴. 淘宝搜索指数与成交指数的协整分析——以羽绒服为例[J]. 产业经济, 2014(3x): 293-293.

[4] 赵萱. 基于淘宝指数的一项中国智能手机市场的经验研究[J]. 法制与社会, 2013(7): 191-194.

[5] 陈正坤. 浅析淘宝指数[J]. 电子世界, 2012(5): 3-6.

[6] 徐贵登, 游国斌, 游天嘉, 等. 宁德市县域电子商务发展分析[J]. 淮海工学院学报, 2015(11): 88-91.

[7] http://alizs.taobao.com/?spm=0.0.0.0.BfKKDW

[8] 王燕. 应用时间序列分析(第四版)[M]. 北京: 中国人民出版社, 2015.