南 开 大 学

本 科 生 毕 业 论 文（设 计）

中文题目：基于 ARRB 模型的中国汽车销量的预测

外文题目：Accurate Estimation of China Auto Sales via ARRB Model

学　　号：1310140

姓　　名：梁思琪

年　　级：2013 级

专　　业：统计学

系　　别：统计系

学　　院：数学科学学院

指导教师：邹长亮

完成日期：2016 年 05 月

关于南开大学本科生毕业论文（设计）的声明

本人郑重声明：所呈交的学位论文，是本人在指导教师指导下，进行研究工作所取得的成果。除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人创作的、已公开发表或没有公开发表的作品内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。本学位论文原创性声明的法律责任由本人承担。

学位论文作者签名：

年　　月　　日

本人声明：该学位论文是本人指导学生完成的研究成果，已经审阅过论文的全部内容，并能够保证题目、关键词、摘要部分中英文内容的一致性和准确性。

学位论文指导教师签名：

年　　月

摘 要

汽车产业作为我国经济的支柱产业，准确预测汽车销量，无论是对于政策制定者还是对于汽车厂商，都具有十分重要的意义。受隐马尔可夫模型启发，本文提出一个采用公开的在线搜索数据对汽车销量进行预测的模型—ARRB（AutoRegression with Rearranged Baidu search data）模型。统计意义上，该模型是以经错位处理后的百度搜索项为外生变量的自回归模型。在进行参数估计时，ARRB 模型采用长度为 26 个月的滚动窗口在模型训练期间捕获人们搜索模式和时间序列趋势的改变，同时还利用l_1正则化实现了变量选择。ARRB 模型不仅考虑了汽车销售的季节性，捕捉了人们在线搜索行为随着时间推移的变化，同时也通过采用关键词错位技术，消除了汽车销量和搜索数据间的不同步性，进一步提升了预测的准确性。实验结果表明，虽然仅使用百度指数提供的低质量公开搜索数据作为输入，ARRB 模型仍优于目前所有用于汽车预测的经典模型。ARRB 模型具有灵活性、稳健性、可扩展性和自我修正性等优点，可用于在多重时空分辨率下对其他社会事件进行实时跟踪。

关键词：汽车销量预测；隐马尔可夫模型；自回归模型；百度指数；l_1正则；关键词错位技术

Abstract

Nowadays, automobile industry has become the pillar industry of the national economy. It is of great significance for the policy makers and the automobile manufacturers to forecast auto sales accurately. In this paper, motivated by hidden Markov model, we propose a model termed ARRB (AutoRegression with Rearranged Baidu search data) to predict the sales volume of cars using the public online search data. Statistically, this model is an autoregressive model with rearranged Baidu search data as exogenous variables. All the parameters are dynamically trained every month with a 26-month rolling window to capture the changes of time series during the training period and the behavior patterns of people. ARRB also uses l_1 regularization to automatically select the most relevant information. ARRB not only takes the seasonality of auto sales and the changes in people's search patterns over time into consideration, but also eliminates the asynchronization between the search data and auto sales using rearrangement technology. Although ARRB model only takes the inferior-quality public search data provided by the Baidu index as input, it still has a better performance than the present state-of-the-art models used for automotive prediction. The ARRB model is flexible, robust, scalable, and self-correcting at the same time. Therefore, it is a potentially powerful tool, which can be used for real-time tracking of other social events at multiple temporal and spatial resolutions.

Keywords: auto sales prediction; hidden Markov; auto regressive model; Baidu index; l_1 regularization; keyword rearrangement technology

目 录

一、引言

近年来，我国国民经济不断发展，人民生活水平也在不断提高，在这种时代背景下，我国的汽车产业逐渐成为四大支柱产业之一。根据国际汽车制造商协会发布的数据,我国 2016 年汽车销量达 2802.8 万辆,已经连续八年获得全球第一并再次创造历史新高。随着汽车产业的不断发展,其在我国国民经济中的地位也日渐深入人心,因此如何准确的预测汽车销量对我国经济发展有着十分重要意义。

目前用于汽车销量预测方面的方法大致分为定性预测和定量预测两种。其中定性预测法主要包括主观概率预测和专家预测等。如：门峰等针对我国汽车产业的发展方向进行了研究,从汽车市场、组织结构、新能源汽车和行业管理等方面进行了分析并对未来汽车产业的发展趋势和特点进行了预测[1]。定量预测法包括灰色系统预测法[2]，多元回归分析预测法[3]，时间序列预测法[4]和 BP 神经网络预测法[5]等。如：杨月英等人运用灰色时间序列对我国未来两年的汽车销量进行了预测，该方法只是利用历年汽车销量数据进行建模，并没有考虑到市场中其它随机因素的影响[2]；王旭天等人根据月度数据同时具有长期趋势效应，季节效应和随机波动的特点，构建了 SARIMA 模型并用于汽车销量预测[6]；李响等人提出了一种 RBF 神经网络与 ARMA 与模型相结合的混合模型，并将之用于汽车销量的预测，其中销量数据中的线性部分利用 ARMA 模型来进行拟合，非线性部分通过 RBF 神经网络来进行逼近[7]。这些预测方法大多都只依赖于汽车销量的历史数据，因此都具有较大的延迟性。同时，以上预测方法往往也具有较大的预测颗粒度，因此一般用于汽车销量的年度预测。

互联网技术的快速发展使得我国的互联网普及率逐年增加。根据中国互联网络信息中心（CNNIC）于今年颁布的第 39 次《中国互联网络发展状况统计报告》，截至 2016 年 12 月，我国网民数量已经达到 7.31 亿，普及率高达 53.2%，其中

搜索引擎的用户数量更是达到了 6.02 亿，使用率为 82.4%，图 1.1 展示了近年来搜索引擎用户规模及使用率的变化趋势。作为中国网民使用率最高的网络工具之一，搜索引擎记录了不计其数的网络搜索数据。中国互联网络信息中心（CNNIC）在社区进行的搜索营销调查显示，有 77%的消费者会在购买商品之前使用搜索引擎搜索相关的商品信息。随着研究的深入，网络搜索数据的重要性逐渐得到重视，近期一些研究进展也表明，很多经济、社会行为都与搜索行为存在着很高的相关关系。近些年来，越来越多的学者开始利用网络搜索数据开展相关的学术研究，如：2009 年，谷歌流感趋势（GFT）--一个数字疾病检测系统，采用选定的谷歌搜索数据来估计当前流感的疫情趋势，被认为是展示大数据如何改变传统统计预测分析的很好的例子[8]。受此启发， S. C. Kou 等人于 2015 年提出了 ARGO 模型，将季节性流感的信息与搜索信息权重的动态调整相结合，使得该模型在预测精度、灵活性和稳健性等方面有了极大的提升[9]。

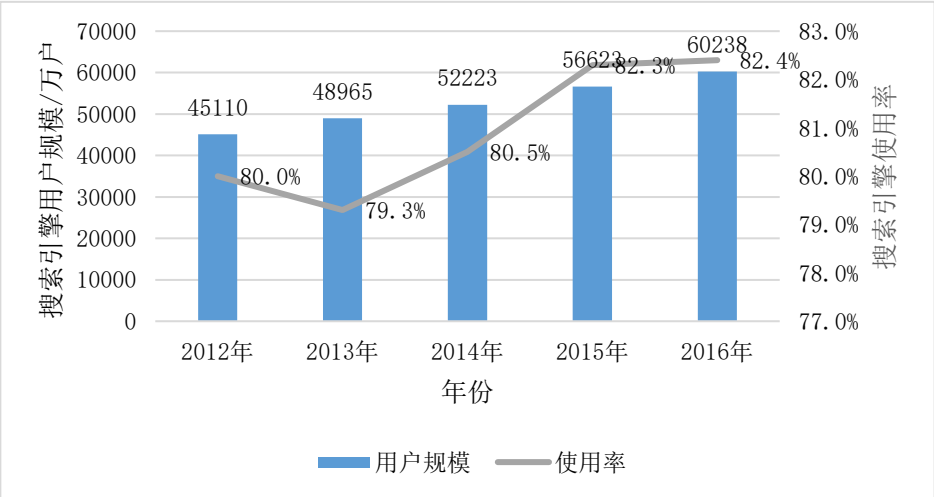

图 1.1 2012-2016 年搜索引擎用户规模及使用率

与此同时，基于网络搜索数据的汽车销量预测模型也得到了发展。如：2011 年，袁庆玉等人就建立了网络关键词搜索数据与汽车销量理论框架，并且在此基础上利用选取的关键词得到关键词合成指数，然后通过回归模型探究了不同价格区间的汽车销量与相应合成指数之间的关系，并进行了预测试验[10]；2014 年，

崔东佳等人利用综合赋权和错位逐步合成法来合成百度搜索关键词，从而得到相应网络搜索指数，然后通过建立回归模型，对网络搜索指数和品牌汽车销量进行协整分析和 Granger 因果检验[11]；2015 年，王炼等人验证了网络搜索与汽车销量和市场份额之间均存在显著的正相关关系[12]。

受此启发，本文基于自回归模型提出 ARRB（AutoRegression with Rearranged Baidu search data）模型。ARRB 模型相对于目前存在的基于网络搜索数据的汽车销量预测模型做出了以下改进：目前基于网络搜索数据的汽车销量预测模型大多为静态模型，不能根据最新数据来演变模型。ARRB 模型实现了静态模型向动态模型的转变，当网络上的新信息可用时，可以动态地合并新信息；其次，现存的方法大多将多个查询词聚合成一个单一变量，不能很好地捕获人们的网络搜索行为的变化，而 ARRB 模型将查询词作为自回归模型的外生变量，全面捕捉网民的搜索模式；另外，ARRB 模型可以自动选择最有用的百度搜索词，在进行预测的同时进行变量选择；最后 ARRB 模型采用关键词错位技术，进一步增强关键词搜索行为与汽车销量之间的相关性，从而可以得到更加准确的预测结果。

本文按照下面的模式展开：第 2 节为文献综述，介绍了目前用于汽车销量预测的几种现有模型；第 3 节介绍了在构建 ARRB 模型时所需要的预备知识。在 4 节中，我们详细介绍了基于错位百度搜索指数的自回归模型（ARRB）的结构框架；第 5 节为实证分析，即利用现实数据进行模拟实验，并将 ARRB 模型与其他几种经典预测模型的预测结果进行分析比较；在第 6 节中，我们对全文进行总结并得出相应结论。

二、文献综述

在汽车销量预测方面，专家学者主要从定性和定量这两个角度进行研究。这里，我们主要介绍汽车销量定量预测方面的研究成果。

2012 年，杨月英等人基于 2000-2010 年的中国汽车年销量的历史数据，通过建立 GM(1,1)模型对中国汽车年销量进行了预测[2]。精度检验结果显示，2000-2010 年预测值的相对误差的绝对值均不超过 20%，小误差概率 P=1，后验差比值 C=0.1438，表明用灰色模型对中国汽车年销量进行预测可以取得令人满意的结果。但是，由于灰色模型只考虑了历史数据的影响，忽略了市场中的随机因素，导致预测结果在 2008 年（金融危机影响），2009 年（补涨行情出现）和 2011 年（美国债务危机影响）出现了较大的偏差。

赵颖等人在 2014 年提出运用主成分回归模型对我国汽车年销量进行预测。他们首先通过格兰杰因果关系检验和灰色关联分析选取了影响我国汽车销量增长的若干影响因素，如人均 GDP、私人汽车拥有量、汽车市场平均价格、燃料价格指数等，然后建立了以这些影响因素 2001-2010 年的年度数据为自变量，以相应年份中国汽车年销量为因变量的主成分回归模型[3]。其中，第一主成分的累积贡献率就达到了 85%。但是由于自变量的选取过程融入了过多的人为因素，可能会导致无关影响因素的乱入或主要影响因素的遗漏等现象，因此预测结果没有充足的说服力。

2016 年，王旭天等人对我国 2004 年 1 月-2015 年 1 月的汽车月度销量数据进行了研究分析，通过构建 SARIMA 模型对全国汽车月销量进行了预测[4]。预测结果表明，除 2014 年 12 月预测偏差较大以外，其余月份的预测偏差均控制在 3%以内。同时预测值的平均绝对百分比误差为 MAPE=2.36，预测效果较为满意。但是 SARIMA 模型只是依据汽车销量的历史数据进行建模，有着较为严重的滞后性，忽略了其他社会因素的影响，导致了部分月份预测偏差较大。

同年，王旭天等人基于 BP 神经网络，将 1998 年-2010 年人均 GDP、钢材产量、城镇居民人均可支配收入等六项经济指标的年度数据作为输入值，汽车年销量作为输出值，构建了包含一个隐含层，4 个隐含层节点的 BP 神经网络，并将

训练好的神经网络用于汽车年销量的预测。预测结果显示，预测结果的平均相对误差为 2.34%[5]。但是由于神经网络存在不能解释自己的推理过程和推理依据，训练速度慢，训练失败的可能性较大等缺点，因此在汽车销量预测方面应用较少。

以上方法均是将统计学中的经典预测模型应用于汽车销量预测。在基于网络搜索数据进行预测方面，袁庆玉等人于 2011 年首次提出基于网络搜索对汽车销量进行预测。该方法在人工选出基准关键词后使用 Google 等工具进行相关词推荐，从而得到网络搜索关键词集合，再采用综合赋权法对搜索数据进行合成，最后通过对合成指数和汽车销量进行一元线性回归来得到最终模型。最终预测结果显示，对不同价格区间的汽车销量月度数据该模型预测值的平均绝对误差百分数均不超过 4%[11]。但是该模型将多个搜索关键词聚合成一个单一变量，当人们的互联网搜索行为发生变化时，不能很好地进行调整。

为此，本文提出基于百度指数的 ARRB 模型，该模型不仅克服了以上模型所存在的问题，大幅度提升了预测的精度，同时也具有很好的灵活性和可扩展性，使得 ARRB 模型有着广阔的应用范围。

三、预备知识

（一）隐马尔可夫模型（HMM）

隐马尔可夫模型是一种统计模型，它假定被建模的系统为一个伴随有隐含状态的马尔科夫过程。一个隐马尔科夫模型可以作为最简单的动态贝叶斯网络。隐马尔科夫模型最早是由 L. E. Baum 和他的同事们在一系列统计著作中提出的[13-17]。

下图显示了一个实例化的 HMM 的一般结构。每个椭圆形代表一个随机变量。随机变量$x(t)$是 t 时刻的隐含状态，随机变量$y(t)$是在 t 时刻可观测到的状

态。图中的箭头表示条件依赖关系。

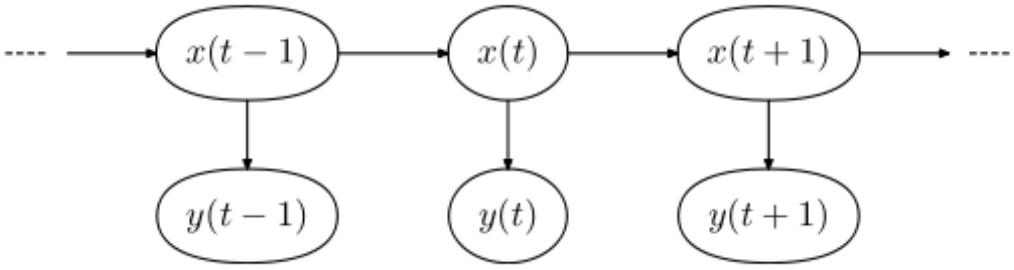

图 3.1 隐马尔可夫模型一般结构

从图 3.1 中可以看出，当给定隐含变量x所有时刻的值时，它在 t 时刻的条件概率分布只取决于上一时刻的取值，即$x(t-1)$，而$t-2$及以前的时刻的取值则对其没有影响，这就是所谓的马尔可夫性。类似的，可观测变量$y(t)$的值仅取决于同时刻隐含变量$x(t)$的值。隐马尔可夫模型有如下三个重要假设：

假设 1：马尔可夫假设

$$P[x(t)|x(t-1),\cdots,x(1)] = P[x(t)|x(t-1)] \tag{3.1}$$

假设 2：不动性假设

$$P[x(i+1)|x(i)] = P[x(j+1)|x(j)], \forall i,j \tag{3.2}$$

假设 3：输出独立性假设

$$P[y(1),\cdots,y(T)|x(1),\cdots,x(T)] = \prod P[y(t)|x(t)] \tag{3.3}$$

在标准隐马尔可夫模型中，隐含变量的状态空间是离散的，而可观测变量本身可以是离散的或连续的（通常服从高斯分布）。隐马尔可夫模型的参数通常有两种类型，转移概率和输出概率。其中转移概率控制在给定 t-1 时刻下的隐含状态时，t 时刻隐含状态的选取。也就是说，假设隐含状态空间有 N 种可能的取值，在 t-1 时刻隐含状态可能处在 N 种可能状态中的任何一种状态，则在 t 时刻，便有从该状态转移到 N 种可能状态的共计N^2种转移概率。注意，从任何给定状态进行跃迁的转移概率集合的总和必须为 1。因此，这个$N \times N$的转移概率矩阵是马尔可夫矩阵。由于当其他转移概率确定时，剩下的那一种转移概率也能被确定，因此，共有$N(N-1)$个转移参数。

此外，对于N种可能状态的每一种状态来说，当给定某时刻隐含变量的状态时，都有一个由可观测变量的分布所决定的输出概率的集合。集合的大小由观测变量的性质所决定。例如，如果可观测变量是离散的且有 M 种可能的取值，那么对所有隐含状态共有$N(M-1)$种输出参数。另一方面，如果可观测变量是一个服从多元高斯分布的 M 维向量，则会有 M 个参数控制均值以及$\frac{M(M+1)}{2}$个参数控制协方差矩阵，因而共有$\mathrm{N}\left(\mathrm{M}+\frac{M(M+1)}{2}\right)=\frac{NM(M+3)}{2}=O(NM^2)$个输出参数。

（二）Lasso 算法

在统计和机器学习中，Lasso 是通过变量选择和正则化来提升预测精度和增强模型的可解释性的一种回归分析方法。它最初是在 1996 年由 Robert Tibshirani 在 Leo Breiman 的“Nonnegative Garrote”的启发下提出的[18,19]。

考虑一个有着 N 个样本的集合，每个样本都有 p 个协变量和一个响应变量。令y_i表示响应变量，$x_i \coloneqq (x_1, x_2, \cdots, x_p)^T$表示第 i 个样本的协变量。Lasso 的目标即为解决下面的问题：

$$\min_{\beta_0,\beta}\left\{\frac{1}{N}\sum_{i=1}^{N}(y_i-\beta_0-x_i^T\beta)^2 \ \ subject\ \ to\ \ \sum_{j=1}^{p}|\beta_j|\le t\right\} \tag{3.4}$$

这里，t 是一个预先设定的自由参数，用来调节正则量。令X为协变量矩阵，则$X_{ij}=(x_i)_j$，且x_i^T为X的第 i 行，上式也可写为：

$$\min_{\beta_0,\beta}\left\{\frac{1}{N}\|y-\beta_0-X\beta\|_2^2\right\} \ \ subject\ \ to\ \ \|\beta\|_1\le t \tag{3.5}$$

其中$\|Z\|_p=\left(\sum_{i=1}^{N}|Z_j|^p\right)^{1/p}$为$\ell^p$范数。

由于$\hat{\beta}_0=\bar{y}-\bar{x}^T\beta$，因此

$$y_i-\hat{\beta}_0-x_i^T\beta=y_i-(\bar{y}-\bar{x}^T\beta)-x_i^T\beta=(y_i-\bar{y})-(x_i-\bar{x})^T\beta \tag{3.6}$$

变量通常都经过中心化处理。另外，为使最终结果不受量纲影响协变量通常

都是经过标准化处理的$\left(\sum_{i=1}^{N} {x_{ij}}^2 = 1\right)$。

Lasso 的拉格朗日形式为：

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{N} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\} \tag{3.7}$$

其中参数λ用来控制回归复杂度的调整，随着λ的增大，对线性模型的惩罚程度也就越大。

（三）滚动窗口分析

在模型参数的训练中，我们用到了时间序列的滚动窗口（Rolling-Window Analysis）分析，下面简要介绍滚动窗口分析的具体步骤。

1.选择滚动窗口的长度 m。

2.选择一个预测范围 h ，该值取决于数据的具体应用及数据的周期。

3.划分数据集。如果连续滚动窗口之间的增量数是 1 期，则将整个数据集划分为 N = T－M + 1 个分区。第一个滚动窗口包含 1 到 m 期的观测数据，第二个滚动窗口包含 2 到 m+1 期的观测数据，以此类推。图 3.4 展示了具体的分区方法。

4.对每个滚动窗口的子样本：

①对每个模型进行估计。

②对未来 h 期的值进行预测。

③计算每次预测的误差，即$e_{nj} = y_{m-h+n+j-}\hat{y}_{nj}$。其中，$e_{nj}$是由滚动窗口 n 训练出的模型所得到的未来 j 期预测值的预测误差；y是响应变量；$\hat{y}_{nj}$是由滚动窗口 n 训练出的模型得出的 j 期预测。

5.计算预测均方误差（RMSEs），即比较不同模型间的 RMSEs，其中 RMSEs 值最小的模型的预测效果最好。

$$\mathrm{RMSE}_j = [\left(\frac{1}{\mathrm{N}}\right)\sum_{n=1}^{N} {e_{nj}}^2]^{\frac{1}{2}} \quad for\ j = 1,\cdots,h \tag{3.8}$$

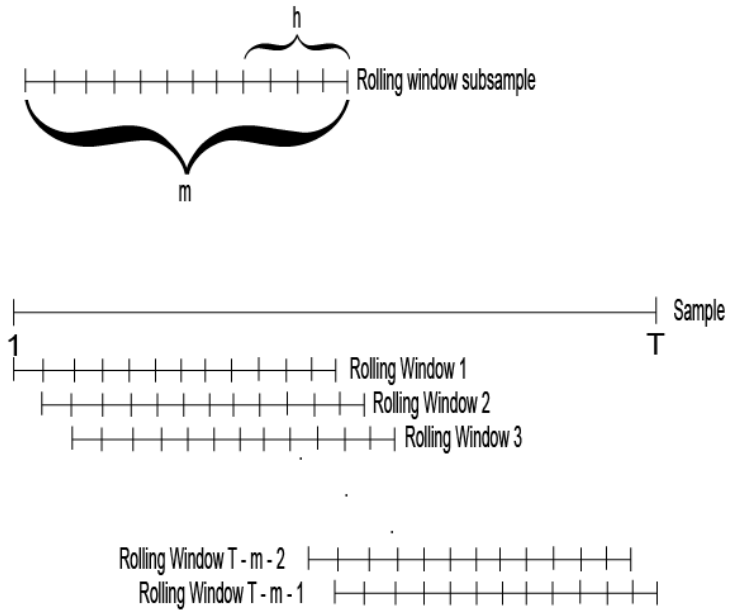

图 3.4 滚动窗口分区方法

（四）Stationary Bootstrap 方法

Stationary bootstrap 方法是由 Dimitris N. Politis 与 Joseph P. Romano 于 1991 年提出的一种针对平稳时间序列的再抽样方法[20]。该方法保持了通过再抽样得到的伪时间序列的平稳性。

假设$\{X_n, n\epsilon\mathbb{Z}\}$是一个弱相关的平稳时间序列，μ为序列$\{X_n, n\epsilon\mathbb{Z}\}$的总体联合分布的一个参数。给定数据$X_1,\cdots,X_N$，我们的目的是基于某些估计量$T_N = T_N(X_1,\cdots,X_N)$对μ做出推断。特别的，我们想要建立μ的置信区间或估计估计量T_N的标准误差。通常，我们需要知道T_N的分布$R_N = R_N(X_1,\cdots,X_N;\mu)$，而 stationary bootstrap 方法就可以达到这个目的。令$B_{i,b} = \{X_i, X_{i+1},\cdots,X_{i+b-1}\}$为从$X_i$开始包含$b$个观测的块。当$j > N$时，$X_j$等于$X_i$，其中$\mathrm{i} = j(modN)$，且$X_0 = X_N$。令$p$为$[0,1]$之间一个固定的数。令$L_1, L_2,\cdots$为$i.i.d.$，服从几何分布且独立于$X_1,\cdots,X_N$的一组随机变量，因此$\{L_i = m\}$的概率为$(1-\mathrm{p})^{m-1}p$，$m = 1,2,\cdots$。令$I_1, I_2,\cdots$为

$i.i.d.$，服从$\{1,\cdots,N\}$离散均匀分布且独立于X_i和L_i的一组随机变量。则伪时间序列$X_1^*,\cdots,X_N^*$按如下方式产生。首先产生块的随机长度，则伪时间序列$X_1^*,\cdots,X_N^*$前L_1个观测由第一个块$B_{I_1,L_1}=\{X_{I_1},\cdots,X_{I_1+L_1-1}\}$所确定，以此类推，接下来的$L_2$个观测由第二个样本块$B_{I_2,L_2}=\{X_{I_2},\cdots,X_{I_2+L_2-1}\}$确定，直到产生了伪时间序列的 N 个观测。

一旦$X_1^*,\cdots,X_N^*$被产生，我们就可以计算该伪时间序列的$T_N(X_1^*,\cdots,X_N^*)$或$R_N(X_1^*,\cdots,X_N^*;T_N)$。用同样的方法通过重复抽样产生大量的伪时间序列，则真正的分布$R_N(X_1,\cdots,X_N;\mu)$就可以通过这些为时序列的经验分布来逼近。

四、ARRB 模型构建

我们的模型受隐马尔可夫模型所启发，通过隐马尔可夫模型将因果序列融合在一起。

经过 Logit 变换的汽车销量序列$\{y_t\}$是我们所感兴趣的隐含时间序列。在对该序列进行 ADF 检验后发现，该序列存在单位根，为平稳序列，因此我们建立一个滞后为 N 的自回归模型，即向量集合$\{y_{(t-N+1):t}\}_{t\geq N}$是一个马尔科夫链。同时，假设在 t 时刻，某相关百度搜索词条 i 的搜索指数仅与某一时刻人们的购买行为相关，而与其他因素如媒体宣传等无关，即经过处理的的某相关词条 i 的搜索指数$x_{i,t-\tau_i}$仅与未来某时刻的汽车销售量y_t相关，其中τ_i为常量（这个结论是由直觉得出的，在生活中，人们通常会在买车之前在网络上搜索心仪车辆的相关信息）。

隐马尔可夫模型结构如下：

$$\begin{array}{ccccc} y_{[1:N]} & \rightarrow & y_{[2:(N+1)]} & \rightarrow\cdots\rightarrow & y_{[(T-N+1):T]} \\ \downarrow & & \downarrow & & \downarrow \\ x_{i,N-\tau_i} & & x_{i,N-\tau_i+1} & & x_{i,T-\tau_i} \end{array} \tag{4.1}$$

为了使表达更为简洁，我们假设下文中的$X_t=(x_{1,t},x_{2,t},\ldots,x_{K,t})$已经过了错

位处理。则 ARRB 模型的前提假设如下：

1. $y_t = \mu_y + \sum_{j=1}^{N} \alpha_j y_{t-j} + \varepsilon_t \qquad \varepsilon_t \overset{iid}{\sim} N(0, \sigma^2)$

2. $X_t | y_t \sim N_k(\mu_x + y_t\beta, Q)$

3. 以y_t为条件时，X_t与$\{y_l, x_l : l \neq t\}$独立

其中$\beta = (\beta_1, \beta_2, \cdots, \beta_K)^T, \mu = \left(\mu_{x_1}, \mu_{x_2}, \cdots, \mu_{x_K}\right)^T$, 且 Q 为协方差阵。则预测分布为$\mathrm{f}\left(y_t \middle| y_{1:(t-1)}, X_{1:t}\right)$是以$y_{(\mathrm{t-N}):(t-1)}$和$X_t$的线性组合为均值，有着常数方差的正态分布。由这个观测，我们得到了最终的 ARRB 模型。

定理 3.1 $\mathrm{f}\left(y_t \middle| y_{1:(t-1)}, x_{1:t}\right)$

$$\sim N\left(\left(\frac{1}{\sigma^2} + \beta^T Q^{-1}\beta\right)^{-1}\left(\frac{\mu_y + \alpha^T y_{(t-N+1):t}}{\sigma^2} + \beta^T Q^{-1}(X_{t+1} - \mu_x)\right), \left(\frac{1}{\sigma^2} + \beta^T Q^{-1}\beta\right)^{-1}\right)$$

证明：由$X_t | y_t \sim N_k(\mu_x + y_t\beta, Q)$可知，

$$X_t = \mu_x + y_t\beta + e_t \qquad e_t \sim N_k(0, Q) \tag{4.2}$$

上式可写为

$$y_{t+1}\beta = -\mu_x + X_{t+1} + \epsilon_t \quad \epsilon_t \sim N_k(0, Q) \tag{4.3}$$

则

$$y_{t+1}\beta | X_{1:(t+1)} \sim N(X_{t+1} - \mu_x, Q) \tag{4.4}$$

$$\therefore \beta^T Q^{-1}\beta y_{t+1} | X_{1:(t+1)} = \beta^T Q^{-1}(X_{t+1} - \mu_x) + \beta^T Q^{-1}\epsilon_t | X_{1:(t+1)}$$

$$\sim N(\beta^T Q^{-1}(X_{t+1} - \mu_x), \beta^T Q^{-1}\beta) \tag{4.5}$$

又有

$$\frac{1}{\sigma^2} y_{t+1} \left| y_{1:t} = \left[\frac{1}{\sigma^2}\left(\mu_y + \alpha^T y_{(t-N+1):t}\right) + \frac{1}{\sigma^2}\varepsilon_t\right] \right| X_{t+1} \sim N\left(\frac{\mu_y + \alpha^T y_{(t-N+1):t}}{\sigma^2}, \frac{1}{\sigma^2}\right) \tag{4.6}$$

则由式(4.5)和式(4.6)可知，$\left(\beta^T Q^{-1}\beta + \frac{1}{\sigma^2}\right) y_{t+1} \left| y_{1:t}, X_{1:(t+1)}\right.$所服从的分布为，

$$\left(\beta^T Q^{-1}\beta + \frac{1}{\sigma^2}\right) y_{t+1} | y_{1:t}, X_{1:(t+1)}$$

$$\sim N\left(\beta^T Q^{-1}(X_{t+1} - \mu_x) + \frac{\mu_y + \alpha^T y_{(t-N+1):t}}{\sigma^2}, \beta^T Q^{-1}\beta + \frac{1}{\sigma^2}\right) \tag{4.7}$$

从而有

$$f(y_t|y_{1:(t-1)},x_{1:t}) \sim N\left(\begin{matrix}(\frac{1}{\sigma^2}+\\ \beta^TQ^{-1}\beta)^{-1}\end{matrix}\left(\frac{\mu_y+\alpha^Ty_{(t-N+1):t}}{\sigma^2}+\beta^TQ^{-1}(X_{t+1}-\mu_x)\right),\left(\frac{1}{\sigma^2}+\beta^TQ^{-1}\beta)^{-1}\right)\right) \quad (4.8)\blacksquare$$

令$y_t = logit(s_t)$为将 t 时刻归一化汽车销量s_t进行 logit 变换后的结果，$x_{i,t}$为 t 时刻搜索词条 i 的搜索频数（错位排列后）经 log 变换后的结果。ARRB 模型如下：

$$y_t = \mu_y + \sum_{j=1}^{N}\alpha_j y_{t-j} + \sum_{i=1}^{K}\beta_i x_{i,t} + \varepsilon_t \qquad \varepsilon_t \overset{iid}{\sim} N(0,\sigma^2) \qquad (4.9)$$

x_t可看做是时间序列$\{y_t\}$的外生变量。

五、实证分析

（一）网络搜索关键词的选取

网络搜索关键词的加入可以避免在进行样本预测时受到过多的前瞻性信息的影响，如何进行关键词的选取是本项研究的基础及关键所在，下面将介绍关键词选取的详细步骤。

注意到在 Google 所发布的数据挖掘工具中，Google Correlate 能够很好地完成关键词选取的工作。当我们把汽车销量的历史数据输入到 Google Correlate 中，并且限定搜索区域（中国）以及相应的搜索时间范围时，就能得到在限定地区与时间范围内与所输入的数据相关性最高的若干搜索关键词。遗憾的是，百度指数并未推出类似功能的工具。但是考虑到人们在不同搜索引擎上的搜索行为的相似性，因此我们仍然利用 Google Correlate 来选取关键词。

由于 Google 于 2014 年 5 月在中国大陆被封锁，因此我们仅导入 2007 年 1 月到 2014 年 4 月的中国大陆汽车销量数据，限定搜索区域为中国，限定搜索时

间范围为 2007 年 1 月到 2014 年 4 月，从而得到与导入数据相关性最高的 100 个关键词。在这些关键词中，往往存在一些干扰项，如：福昕，淘宝网等与汽车无关的关键词。去除这些干扰项后，我们便得到第一个关键词集合。同时考虑到网络搜索行为与汽车销售之间可能存在不同步性，我们分别将销售数据提前 1 到 6 个月，再次利用 Google Correlate 进行关键词选取，从而分别得到 6 组关键词，去除其中的重复项和干扰项后与之前不考虑滞后时得到的关键词共同构成一个关键词集合，最终的关键词集合由 55 个关键词组成。图 5.1 展示了我们所选出关键词集合中的部分关键词。

变速箱油	刹车油	飞歌导航	后视镜	机油
节气门	路畅导航	人保车险	一嗨租车	广汽
胎压监测	提车	倒车影像	波箱油	车钥匙
轮毂	地图导航	多功能方向盘	腾讯汽车	平安车险
……	……	……	……	……

图 5.1 部分相关关键词

（二）数据来源

1、网络搜索数据

本文所采用的网络搜索数据来源于百度指数--由百度发布的一个以数据为基础的数据分享平台。百度指数基于百度网民在百度搜索引擎上的搜索行为，计算出某个关键词的搜索规模（如图 5.2）。我们选取了关键词集合中的全部关键词从 2007 年 1 月到 2016 年 6 月的百度指数数据，为了能更好地与来自搜狐汽车的大陆汽车月度销量数据结合，我们需要将这些日度搜索数据进行合并，得到相关关键词的月度搜索数据。

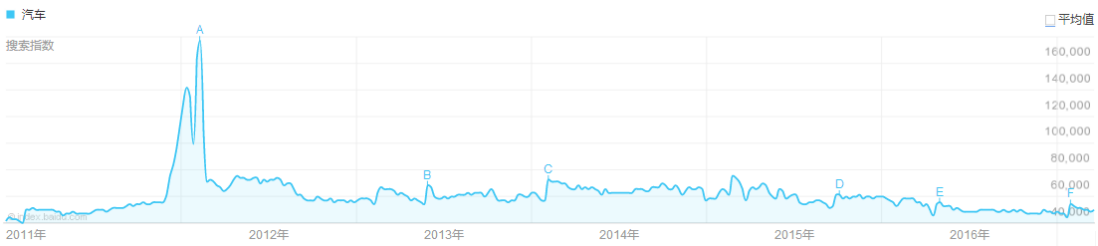

图 5.2 关键词“汽车”的搜索指数趋势图

2、中国汽车销量数据

本文所采用的中国汽车销量数据来源于搜狐汽车网站，该网站提供国内汽车销量的月度数据。我们选取了 2007 年 1 月至 2016 年 6 月共计 114 个月的汽车销量数据。

（三）数据处理

1. 将百度指数日度数据改为月度数据

由于汽车销量数据是月度数据，而百度指数数据是日度数据，为了进行进一步地研究分析，我们需要将日度百度指数数据按月合并，从而得到相关关键词的月度搜索数据。

2. 数据错位

显然，相关关键词的网络搜索行为与购车行为并不一定是时时对应的关系。对某些关键词来说，人们可能会在购车前几个月进行搜索，如：人们通常会在买车前几个月在搜索引擎上搜索相关的汽车网站。因此本文采用关键词错位技术对原始搜索指数进行错位，进一步提高网络搜索数据与汽车销量之间的相关性。我们先分别将搜索数据的日期提前 1 到 6 个月，从而得到六组数据，再分别计算出汽车销量数据与这六组数据的 Pearson 相关系数，然后选取每个关键词搜索数据与汽车销量相关性最高时所对应的提前期。如图 5.3 所示，若

某关键词的最佳错位期为三个月，则该关键词的 2014 年 1 月、2 月和 3 月份的搜索指数应分别对应 2014 年 4 月、5 月和 6 月的中国汽车销量，以此类推。

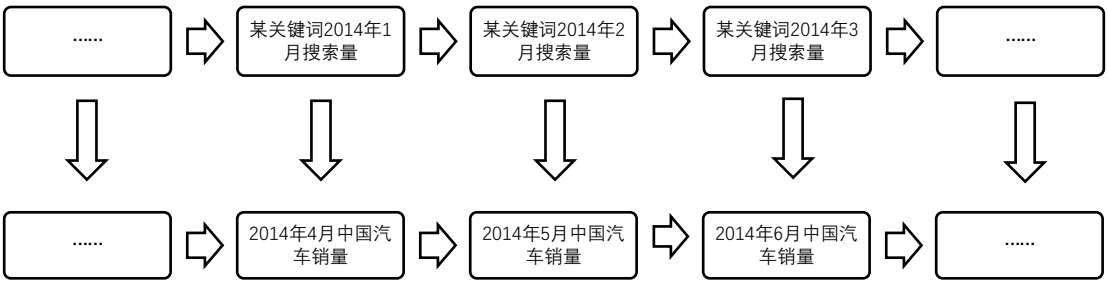

图 5.3 关键词提前期 3 个月的示例

最后，根据上述方法将关键词集合中所有关键词进行错位。表 5.1 展示了部分相关关键词的最佳提前期。

表 5.1 部分相关关键词及其最佳提前期

关键词	最佳提前期	关键词	最佳提前期
变速箱油	3	节气门	6
刹车油	3	人保车险	1
飞歌导航	6	一嗨租车	2
后视镜	1	广汽	1
机油	1	胎压监测	3

3. 规范化处理

为使变量更规范，首先需对原始的销量数据和搜索数据进行预处理。对于销量数据，我们首先用除以最大值的方法将销量数据归一化，从而使得销量数据转化到(0,1]上，再使用 logit 函数使其转化到$\mathbb{R}$上，从而得到y_t。

对于使用关键词错位技术处理过后的搜索数据，我们利用 log 函数使百度指数数据从$\mathbb{N}$转化到$\mathbb{R}$上，得到$X_t = (x_{1,t}, x_{2,t}, \dots, x_{K,t})$。其中$x_{i,t}(i = 1,2, \dots, K)$为 t 时刻搜索词条 i 的搜索频数经 log 变换后的结果。之所以选择选择 log 函数，是

因为百度搜索频数在顶峰附近通常有指数增长的速率。另外，由于百度指数的取值为在$[0,+\infty)$上的整数，因此，我们在 log 变换之前给每个数据都加上δ=0.5，以防止$\log 0$的出现。

（四）参数估计

经过多次模拟试验，我们最终选择移动窗口长度 N=26（月）去捕获汽车销售的季节性趋势，百度指数相关关键词的数目 K=55。由于独立变量的数目多于观测变量的数目，即该问题为“Large p, small n”问题，不能用传统的极大似然估计（最小二乘）法去估计参数，因此我们通过添加惩罚项来解决这个问题。通常而言，我们有三种惩罚项：l_1惩罚项，l_2惩罚项和l_1惩罚项与l_2惩罚项的线性组合。所有参数都是通过长度为 26 个月的滚动窗口进行动态训练得到的。

在给定的月份，我们的目标就是寻找参数$\mu_y, \boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_{26})$和$\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_{55})$使得下面的目标函数达到最小：

$$\sum_t (y_t - \mu_y - \sum_{j=1}^{26} \alpha_j y_{t-j} + \sum_{i=1}^{55} \beta_i x_{i,t})^2 + \lambda_\alpha \left|\left|\alpha\right|\right|_1 \qquad (5.1)$$

其中，λ_α，λ_β，η_α，η_β是超参数。在[10]中，为了模型的简洁性和稀疏性考虑，以及在经过交叉验证实验后，选择$\eta_\alpha = \eta_\beta = 0$且$\lambda_\alpha = \lambda_\beta$，因为此时模型有着最好的表现。

因此同样的，最终我们的 ARRB 模型的优化参数$\mu_y, \boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_N)$和$\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_K)$可以通过求解下面的问题得到：

$$\underset{\mu_y, \alpha, \beta}{\operatorname{argmin}} \sum_t (y_t - \mu_y - \sum_{j=1}^{26} \alpha_j y_{t-j} + \sum_{i=1}^{55} \beta_i x_{i,t})^2 + \lambda(||\alpha||_1 + ||\beta||_1) \qquad (5.2)$$

由于进行了数据错位，因此我们将从 2007 年 7 月到 2014 年 1 月的历史数据作为训练集来训练模型。图 5.4 是根据训练结果画出的系数热力图，从图中可以看出模型中的哪些变量较为显著，因此 ARRB 模型在进行预测的同时可以做出变

量选择：

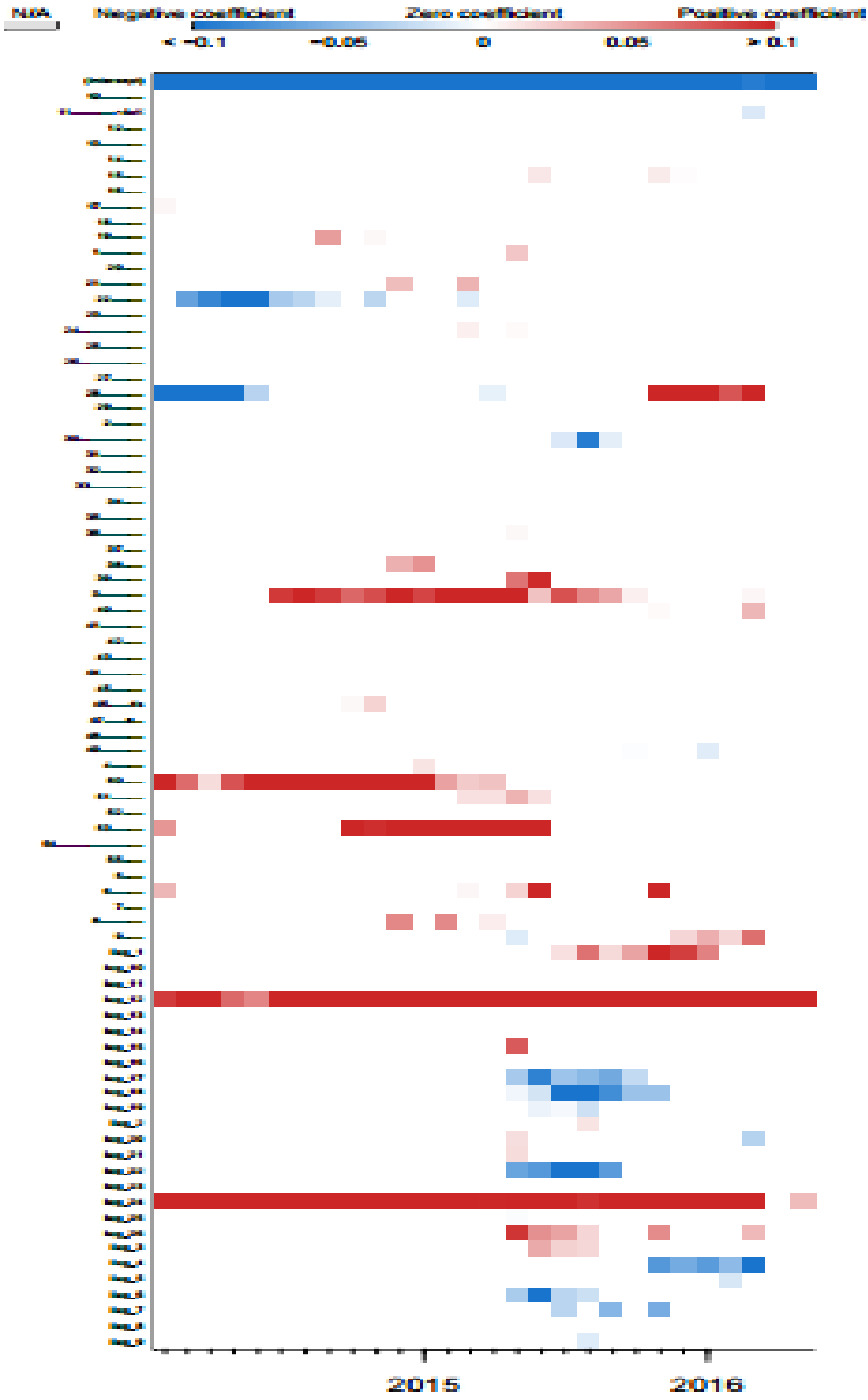

图 5.4 模型系数热力图

（五）模型预测

利用 2007 年 7 月到 2014 年 1 月的历史数据，对 2014 年 2 月到 2016 年 6 月的中国汽车销量进行预测。为了说明 ARRB 模型的优越性，我们将同时使用以下几种经典预测模型进行预测，最后将这几种模型的预测结果展示在一张图中,并对这些模型的预测精度进行分析。

SARIMA 模型：全称为季节性差分自回归移动平均模型，是用来处理同时存在趋势性和季节性的非平稳时间序列数据的一种时间序列模型。

GM(1,1)模型：用将原始数据构造成规律性较强的生成序列的手段来寻求现实现象的变动规律的一种预测方法。

BP 神经网络：一种运用最速下降法，通过反向传播算法来不断调整网络的权值和阈值，使网络的误差平方和最小的神经网络模型。

Naïve Method：也叫天真预测法，一种以前一期的观测值作为这一期的观测值的预测方法。

各种方法预测结果以及预测误差如图 5.5 所示。从图中可以直观的看出，ARRB 模型的预测误差要明显小于其他几种预测方法。

图 5.5 各模型预测结果

（六）预测精度分析

为了进一步比较上述几种方法的预测精度，我们分别用所要预测的汽车销量估计$\hat{s}$的 RMSE，MAE，MAPE 来进行预测精度分析。它们的定义分别如下：

$$\mathrm{RMSE}(\hat{s}_t, s_t) = [(1/n)\sum_{t=1}^{n}(\hat{s}_t - s_t)^2]^{1/2} \quad (5.3)$$

$$\mathrm{MAE}(\hat{s}_t, s_t) = (1/n)\sum_{t=1}^{n}|\hat{s}_t - s_t| \quad (5.4)$$

$$MAPE(\hat{s}_t, s_t) = (1/n)\sum_{t=1}^{n}|\hat{s}_t - s_t|/s_t \quad (5.5)$$

估计$\hat{s}_t$和目标对象s_t之间的相关性定义为它们的样本相关系数。

根据上述定义，我们分别计算出了各个方法在不同时间段内相应的精度指标。其中表中的 RMSE，MAE，MAPE 是其他给定方法相对于天真预测法的误差比，绝对误差在天真预测法的括号中给出（每种评价指标的最优结果被加粗强调）：

表 5.2 各模型预测精度分析表

	2014-02/2016-05	2014-02-01/2014-12-01	2015-01-01/2015-12-01	2016-01-01/2016-05-01
RMSE				
ARRB	**0.516**	**0.722**	**0.687**	**0.195**
SARIMA	0.680	0.787	0.863	0.462
GM	1.257	1.564	1.490	0.939
BPNet	2.351	5.075	1.681	1.329
Naive	1.000 (317238.249)	1.000 (190742.816)	1.000 (294995.668)	1.000 (511700.172)
MAE				
ARRB	**0.504**	**0.579**	**0.686**	**0.208**
SARIMA	0.820	0.991	0.944	0.544
GM	1.411	1.829	1.535	0.971
BPNet	2.563	5.775	1.682	1.418
Naive	1.000 (237088.889)	1.000 (145320.000)	1.000 (242775.000)	1.000 (406980.000)
MAPE				
ARRB	**0.481**	**0.556**	**0.644**	**0.208**
SARIMA	0.833	1.032	0.952	0.540
GM	1.531	2.022	1.667	1.015
BPNet	2.550	6.148	1.558	1.256
Naive	1.000 (0.142)	1.000 (0.087)	1.000 (0.146)	1.000 (0.242)
Correlation				
ARRB	**0.852**	0.762	**0.884**	**0.942**
SARIMA	0.824	**0.872**	0.838	0.885
GM	0.416	0.650	0.430	-0.228
BPNet	0.291	0.375	0.365	-0.534
Naive	0.417	0.444	0.565	-0.114

从表 5.2 中可以看出，ARRB 模型在各个评价指标下的表现均优于其他模型。

接下来考察 ARRB 模型相对于其他模型的效率如何。ARRB 模型估计$\hat{s}^{(1)}$对给定模型估计$\hat{s}^{(2)}$的相对效率定义为：

$$e\left(\hat{s}^{(1)},\hat{s}^{(2)}\right) = MSE_{true}^{(2)}/MSE_{true}^{(1)} \tag{5.6}$$

其中$MSE_{true}^{(i)} = \mathbb{E}[(\hat{s}_t - s_t)^2]$

它可以被如下估计：

$$\hat{e}\left(\hat{s}^{(1)},\hat{s}^{(2)}\right) = MSE_{obs}^{(2)}/MSE_{obs}^{(1)} \tag{5.7}$$

其中$MSE_{obs}^{(i)} = 1/n\sum_{t=1}^{n}(\hat{s}_t - s_t)^2$

$\hat{e}\left(\hat{s}^{(1)},\hat{s}^{(2)}\right)$的 95%置信区间可以由时间序列的 stationary bootstrap 方法构造，其中使用平均长度为 26（对应 26 个月）的随机块(随机块的长度服从几何分布)产生残差的再生时间序列。我们得到$\log\{e\left(\hat{s}^{(1)},\hat{s}^{(2)}\right)\}$基本的 bootstrap 置信区间，然后用指数运算恢复数据原有的量纲。非参数的 bootstrap 置信区间考虑到了误差的自相关和互相关，对随机块的平均长度不敏感。表 4.3 给出了 ARRB 模型分别对 SARIMA 模型，GM(1,1)模型，BP 神经网络模型和天真预测法的相对效率的点估计以及 95%置信区间。

表 5.3 相对效率的置信区间

	point estimate	basic 95% CI lower bond	basic 95% CI upper bond	normal 95% CI lower bond	normal 95% CI upper bond
SARIMA	1.91	1.28	2.77	1.32	2.74
GM	6.78	4.32	10.35	4.36	10.32
BPNet	22.54	9.08	58.68	10.12	50.87
Naive	3.99	2.17	7.18	2.34	6.83

表 5.3 中的结果表明，ARRB 模型相对于其他模型的效率的点估计均大于 1 且置信区间的下边界也均大于 1,所以 ARRB 模型的效率是高于其他所有方法的。

六、结论

受隐马尔科夫链启发，本文提出了利用公开搜索数据进行汽车销量预测的 ARRB 模型。首先我们对数据进行预处理，采用关键词错位技术，本质上，ARRB 模型是以百度搜索指数为外生变量的自回归模型。在进行参数估计时，我们采用了时间序列的滚动窗口法，多次实验分析后，选用了长度为 26 个月的滚动窗口。同时由于独立变量的数目多于观测变量的数目，不能用传统的极大似然估计（最小二乘）法去估计参数，因此我们运用 LASSO 算法，在参数估计的同时进行了变量选择，剔除冗余变量。在本文的实证分析部分，我们将 ARRB 模型与其他经典模型的预测结果进行对比分析，最终结果表明，ARRB 模型在预测精度方面表现明显优于其他方法。相较于其他目前用于汽车预测的模型，ARRB 模型不仅具有灵活稳健的优点，同时还具备可扩展性和自我修正性，使得 ARRB 模型有着非常广泛的适用范围，可以应用于许多其他社会事件的实时跟踪。

参考文献

[1] 门峰，王今．中国汽车产业发展趋势预测［J］．汽车工业研究，2011，(2)：2-5.

[2] 杨月英，马萍．基于灰色时间序列预测中国汽车销量[J]．湖州职业技术学院学报，2012，10(1):5-7.

[3] 赵颖．基于回归分析的我国汽车销量预测模型研究[D]．华中师范大学，2014.

[4] 郭顺生，王磊，黄琨．基于时间序列模型预测汽车销量研究[J]．机械工程师，2013(5):8-10.

[5] 王旭天．基于 BP 神经网络的我国汽车销量预测分析[D]．东华大学，2016.

[6] 王旭天，李政远．基于 SARIMA 的我国汽车销量预测分析［J］．中国市场．2016，(1)：71-74

[7] 李响，宗群，童玲．汽车销售混合预测方法研究[J]．天津大学学报(社会科学版)，2006，8(3):175-178.

[8] Helft M (November 11, 2008) Google uses searches to track flu's spread. NY Times. Available at www.nytimes.com/2008/11/12/technology/internet/12flu.html?_r=0#. Accessed July 11, 2015.

[9] Yang S, Santillana M, Kou S C. Accurate estimation of influenza epidemics using Google search data via ARGO[J]. Proceedings of the National Academy of Sciences of the United States of America, 2015, 112(47):14473-8.

[10] 袁庆玉，彭赓，刘颖,等．基于网络关键词搜索数据的汽车销量预测研究[J]．管理学家:学术版，2011(1):12-24.

[11] 崔东佳．大数据时代背景下的品牌汽车销量预测的实证研究[D]．河南大学，2014.

[12] 王炼，宁一鉴，贾建民．基于网络搜索的销量与市场份额预测:来自中国汽车市场的证据[J]．管理工程学报，2015，29(4):56-64.

[13] Baum, L. E.; Petrie, T. (1966). "Statistical Inference for Probabilistic Functions of Finite State Markov Chains". The Annals of Mathematical Statistics. 37 (6): 1554–1563.doi:10.1214/aoms/1177699147. Retrieved 28 November 2011.

[14] Jump up^ Baum, L. E.; Eagon, J. A. (1967). "An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology". Bulletin of the American Mathematical Society. 73 (3): 360. doi:10.1090/S0002-9904-1967-11751-8.Zbl 0157.11101.

[15] Jump up^ Baum, L. E.; Sell, G. R. (1968). "Growth transformations for functions on manifolds". Pacific Journal of Mathematics. 27 (2): 211–227.doi:10.2140/pjm.1968.27.211. Retrieved 28 November 2011.

[16] Jump up^ Baum, L. E.; Petrie, T.; Soules, G.; Weiss, N. (1970). "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains". The Annals of Mathematical Statistics. 41:
164. doi:10.1214/aoms/1177697196. JSTOR 2239727. MR MR287613.Zbl 0188.49603.

[17] Jump up^ Baum, L.E. (1972). "An Inequality and Associated Maximization Technique in Statistical Estimation of Probabilistic Functions of a Markov Process". Inequalities. 3: 1–8.

[18] Tibshirani, Robert. 1996. “Regression Shrinkage and Selection via the lasso”. Journal of the Royal Statistical Society. Series B (methodological) 58 (1). Wiley: 267–88. http://www.jstor.org/stable/2346178.

[19] Breiman, Leo. 1995. “Better Subset Regression Using the Nonnegative Garrote”. Technometrics 37 (4). Taylor & Francis, Ltd.: 373–84. doi:10.2307/1269730.

[20] Politis DN, Romano JP (1994) The stationary bootstrap. J Am Stat Assoc 89(428): 1303–1313.

致　谢

随着毕业设计的完成，我的大学生活也将结束。大学四年学习生涯，让我有了质的蜕变，专业技能、个人处世能力、人生目标和信仰的加强提高与具体化 ,但每一次的突破自我成长都是心酸的,因此我要感谢在我成长过程中支持、关心、帮助过我的人,也要感谢培育我的母校，南开大学。 诚挚的感谢我的论文指导老师邹长亮老师。他在忙碌的教学工作中挤出时 间来审查、修改我的论文。还有教过我的所有老师们，你们严谨细致、一丝不 苟的作风一直感染着我,；他们循循善诱的教导和不拘一格的思路给予我无尽的 启迪。在将来的博士学习，工作中，我将会将各位老师作为的我榜样，一直继 续努力下去。 再感谢四年中陪伴在我身边的同学、朋友、感谢他们为我提出的可贵的建 议和意见，有了他们的支持、鼓励和帮助，我才能充实的度过了四年的学习生活。

梁思琪 2017 年 05 月