# 基于搜索和新闻的原油价格预测

## 8.1 基于搜索指数的油价预测模型

ARMA 模型（Auto-Regressive and Moving Average Model）是研究时间序列的重要方法由自回归模型简称AR模型与滑动平均模型简称MA模型为基础“混合”构成，是时间序列模型的一种，是由美国统计学波克斯(Box)和金肯(Jenkins)在20世纪70年代提出，在预测领域有非常广泛的应用。

### 8.1.1 ARIMA油价预测模型

我们以2004年3月-2012年6月WTI月度原油价格序列数据建立ARIMA（8,1,0）模型来预测2012年7月至11月期间内原油价格。模型方程如下：



模型估计结果如表所示。

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **模型統計資料** | | | | | | | | | | | | | |
| 模型 | 預測變數數目 | 模型適合度統計資料 | | | | | | | | Ljung-Box Q(18) | | | 離群值數目 |
| 平穩 R 平方 | R 平方 | RMSE | MAPE | MAE | MaxAPE | MaxAE | 標準化 BIC | 統計資料 | DF | 顯著性 |
| 油价-模型\_1 | 1 | .078 | .982 | 2.967 | 3.091 | 2.200 | 17.916 | 13.938 | 2.315 | 7.916 | 10 | .637 | 0 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ARIMA 模型參數** | | | | | | | | |
|  | | | | | 估計 | SE | T | 顯著性 |
| 油价-模型\_1 | 油价 | 無轉換 | 常數 | | 4.711 | 11.648 | .404 | .686 |
| AR | 落後 1 | .187 | .048 | 3.884 | .000 |
| 落後 2 | .014 | .049 | .296 | .767 |
| 落後 3 | .011 | .049 | .222 | .824 |
| 落後 4 | .036 | .049 | .730 | .466 |
| 落後 5 | .016 | .049 | .318 | .750 |
| 落後 6 | .027 | .049 | .559 | .577 |
| 落後 7 | .004 | .049 | .078 | .938 |
| 落後 8 | .170 | .048 | 3.515 | .000 |
| 差異 | | 1 |  |  |  |
| 时间序列 | 無轉換 | 分子 | 落後 0 | .000 | .000 | -.395 | .693 |

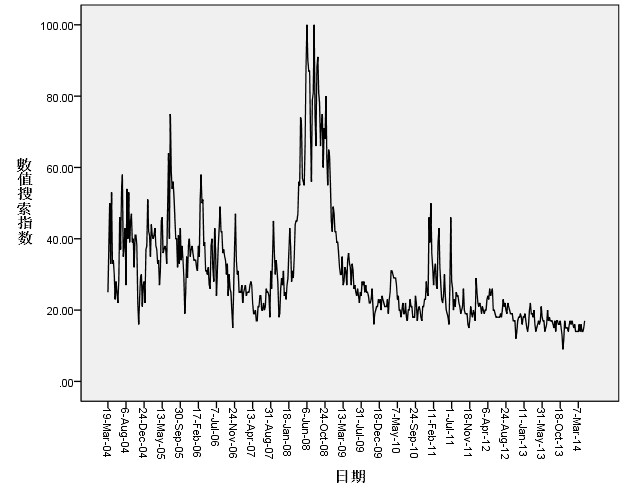
对样本内原油价格月度数据进行预测，得到样本区间内即2004年3月-2012年6月的月度结果，如图4.9所示，从样本内预测精度的指标上看，样本容量共436，该模型样本内的预测误差是2.97%左右，可以看出时间序列模型样本期内拟合效果较好。

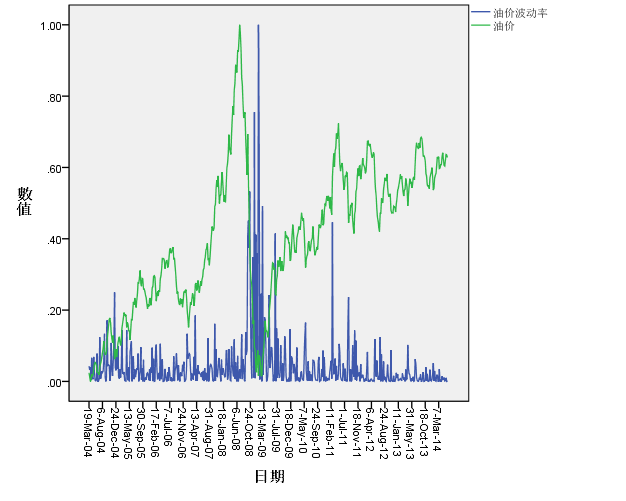
为检验该模型的预测效果，对样本外的原油价格进行预测，预测结果如下表所示。从预测精度的指标上看，样本外的预测误差为6.70%。预测精度显示无论样本内拟合情况还是样本外的预测情况，该回归模型的预测精度均在93%以上，预测效果较好。

### 8.1.2加入搜索指数的ARIMA油价预测模型

1）搜索数据准备

为了准确反映互联网中对于原油价格的关注度，采用人工范围划定的方式选取了若干关键词，包括‘crude oil’, ‘WTI’, ‘crude oil price’, ‘NYMEX oil price’,以及‘NYMEX crude oil’。时间与油价的选取时间相同，为2004年3月-2012年11月。经过统计分析，发现各个关键词得到的搜索指数序列较为相似，序列间的线性相关系数较高。其中，‘crude oil’对应的搜索量最大，且包含更多的具体信息，因此选取‘crude oil’作为最能代表互联网搜索关注度的关键词。

下图为搜索指数以及原油价格波动的变化趋势，可以看到，搜索指数序列与原油价格波动序列具有一定的相似性，在2008年9月全球金融危机期间，油价变动值达到了一个高点，而期间的搜索指数也出现了明显的升高，表现了类似的趋势。



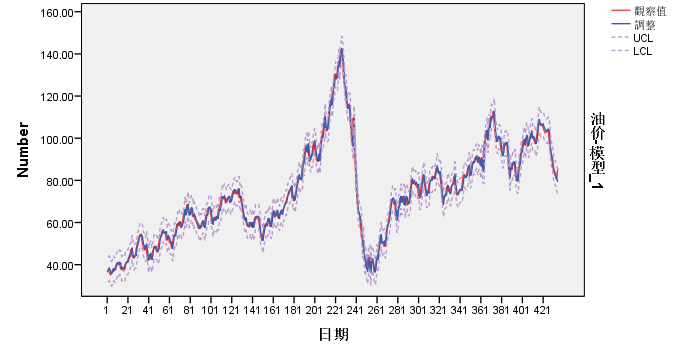
采用加入了搜索指数的ARIMA(8,1,0)模型来预测2012年7月至11月期间内原油价格。其模型方程为：



对样本内原油价格月度数据进行预测，得到样本区间内即2004年3月-2012年6月的月度结果，如图4.9所示，从样本内预测精度的指标上看，样本容量共436，该模型样本内的预测误差是2.95%左右，可以看出时间序列模型样本期内拟合效果较好。

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **模型統計資料** | | | | | | | | | | | | | |
| 模型 | 預測變數數目 | 模型適合度統計資料 | | | | | | | | Ljung-Box Q(18) | | | 離群值數目 |
| 平穩 R 平方 | R 平方 | RMSE | MAPE | MAE | MaxAPE | MaxAE | 標準化 BIC | 統計資料 | DF | 顯著性 |
| 油价-模型\_1 | 1 | .090 | .983 | 2.947 | 3.115 | 2.210 | 18.007 | 13.775 | 2.301 | 6.671 | 10 | .756 | 0 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ARIMA 模型參數** | | | | | | | | |
|  | | | | | 估計 | SE | T | 顯著性 |
| 油价-模型\_1 | 油价 | 無轉換 | 常數 | | 1.369 | .550 | 2.490 | .013 |
| AR | 落後 1 | .169 | .048 | 3.507 | .001 |
| 落後 2 | .012 | .049 | .253 | .801 |
| 落後 3 | .006 | .049 | .131 | .895 |
| 落後 4 | .026 | .049 | .543 | .588 |
| 落後 5 | .004 | .049 | .075 | .940 |
| 落後 6 | .018 | .049 | .359 | .720 |
| 落後 7 | -.004 | .049 | -.091 | .928 |
| 落後 8 | .161 | .048 | 3.328 | .001 |
| 差異 | | 1 |  |  |  |
| 搜索指数 | 無轉換 | 分子 | 落後 0 | -.027 | .011 | -2.526 | .012 |



2）模型效果

为检验该模型的预测效果，对样本外的原油价格进行预测，并与不加入搜索指数的模型对比，预测结果如下表所示。

表4.21: 原油价格序列的ARIMA模型估计结果-样本外

|  |  |  |
| --- | --- | --- |
|  | 模型1 | 模型2 |
| *MAE* | 5.80 | 4.95 |
| *MAPE* | 6.16 | 5.36 |
| *RMSE* | 6.70 | 5.73 |

从预测精度的指标上看，样本外的预测误差为5.73%。预测精度显示无论样本内拟合情况还是样本外的预测情况，该回归模型的预测精度均在94%以。与不加入搜索指数的模型相比样本外的预测精度提升了近1%，并且模型平稳的R方从0.73提升到0.90，说明了加入搜索的模型能更好刻画原油价格序列，搜索数据对于预测原油价格具有一定作用。

## 8.2基于情感分析的石油价格走势预测

### 8.2.1 方法概述

1）情感分析

现有的经济学研究已经表明，从股票价格预测的角度来考虑，充分考虑投资者情绪的预测模型才能更准确的预测股票价格的走势，才能正确的描述市场规律。

情感分析一种度量文本内容情感倾向性的方法。通过情感分析我可以们获得文本 内容的情感，但是同时能够知道市场参与者是怎样处理新闻和反馈的。

在金融市场领域，情感分析主要应用于股票市场预测。金融文本挖掘方法主要基于词典的方式。这种基于词典的方式通过在给定词典中出现的预先定义的正向词和负向词的频率，在结合一些情感指标的计算规则得到最终到情感分值，然后基于这些情 感序列再做相应的分析活着预测。情感指标的计算规则主要有Tetlock-Negative方法，Net—Optimism方法。情感分析的公式中，S(A)表示一篇新闻的情感，而表示一篇文章中的所包涵的词的个数， 表示一篇文章中所包涵的负向词的个数，表示一篇文章中的所包涵的 正向词的个数。

Demers和Vega(2010)提出了一种不同的计算文本情绪的方法，这种方法主要 度量文本中正向词和负向词的差异，如下面公式所示，



本节采用的情绪计算方法如下，该方法也是Net-Optimism方法的一种变体，并且 这种方法在股票价格预测领域中使用的非常广泛，其情绪计算得分计算如下，



相对于Net-Optimism方法，计算时改变了分子，使得文本中正向词 和负向词的差异刻画的更加准确。

2） 预测模型

在没有其他额外的变量下，基于历史石油价格序列预测步长为h的情况下的石油价格走势的公式可以表述为如下：



其中表示在t时刻的石油价格走势预测值，m表示滞后期，h是步长。特别的，预示着石油价格将要下降（），相反预示着石油价格将要上升（）。引入新闻情感和预测滞后期l，则预测模型可以推导为如下：



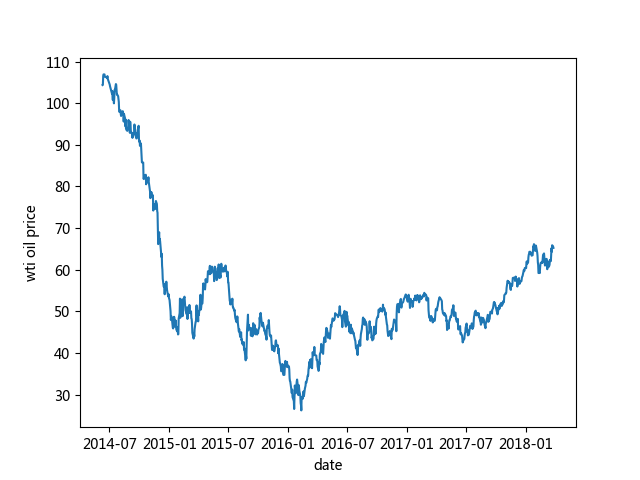
该模型同时引入线性预测模型和非线性预测模型来 研究新闻情感对于石油价格走势的预测能力，在对未来石油价格走势预测时，不只考 虑历史价格序列，同时引入历史新闻情绪序列辅助石油价格走势的预测。

### 8.2.2石油价格与新闻数据采集

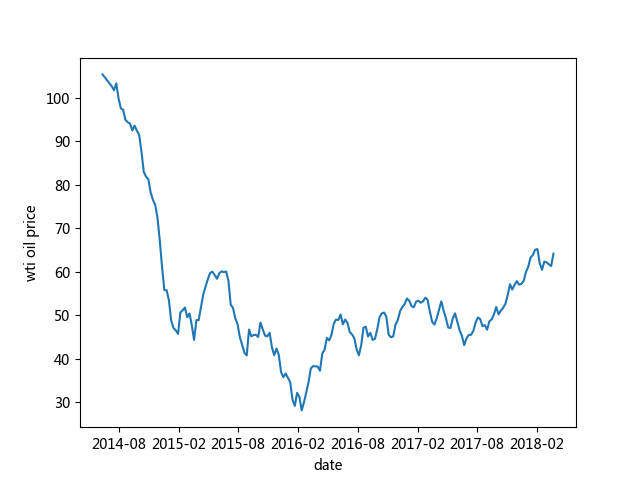
1）石油价格

选择了西德克萨斯原油期货价格(WTI)作为本节的研究对象，美国西得克萨斯中质原油一直 被众多投资者视为国际能源市场的基准价，同时许多国内外的学者也把WTI原油价 格作为研究对象

从美国能源署(EIA)获得的从2014年6月10号到2018年3月21号，总共193周的WTI原油价格数据和相应时期的日度价格，如图所示。



WTI原油周度价格相对于日度价格来说，噪音少，周度走势更加明确，而日度价格影响因素众多，而新闻中情感往往是针对于一些影响石油价格幅 度较大的事件的报道和分析，因此使用WTI原油周度价格应该更加具有实际意义。周度价格如下图所示。



实验中，本节把从2014年6月10号到2017年6月16号的数据(共154个样本， 80％)作为训练集，剩下时间的数据(共39个样本，20％)作为测试集。

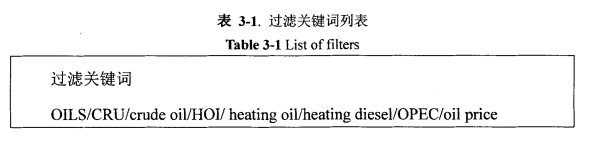
2）新闻数据获取

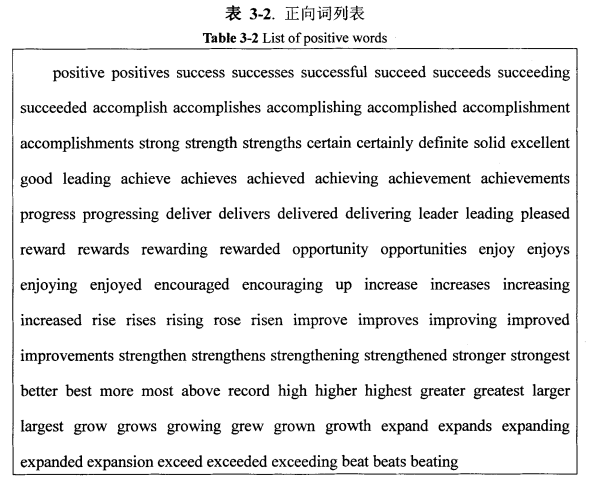
采用了网页爬虫技术获取海量新闻数据。程序的输入主要有两个，一个是带爬 的路透社石油专题模块的网址，另外一个是新闻爬取的时间跨度，这个时间跨度主要 控制爬取新闻的开始日期和截止日期，同时程序的一些默认参数主要包括请求网页超时时间，最大尝试连接等。

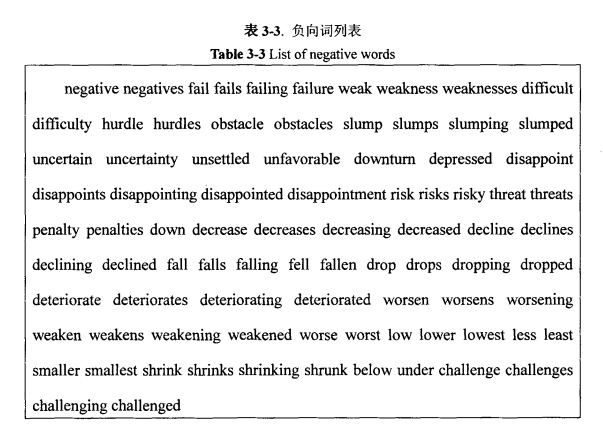
### 8.2.3新闻数据情感分析

从路透英国官方网站获得了从2014年6月10号到2018年3月21号（总共943天）石油板块的新闻数据。虽然是石油板块的新闻但是仍然存在一些与石油无关的信息，对于分析石油新闻情感来说是噪声。因此采用Wex等人（2013）的文献自定义的关键词列表（如下表所示），对新闻标题进行过滤，保留与石油高度相关的新闻，总计7785条，平均每天8.25条新闻。

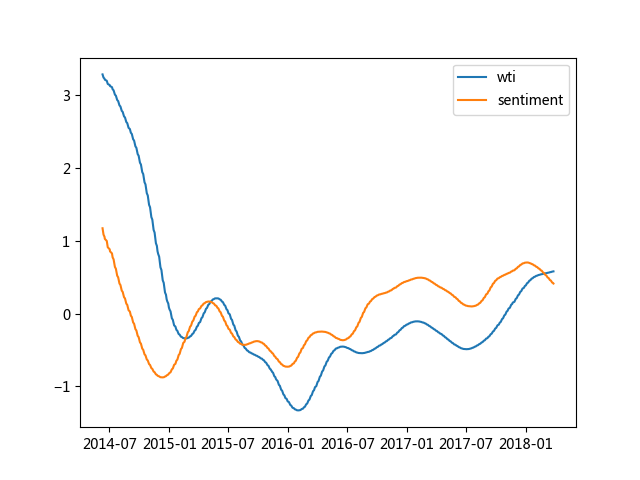
对过滤后的日度新闻合成为周度新闻，采用斯坦福分词工具对合并语料进行处理。基于文献中的Henry’s词典，包括正向词列表（共103个词）和负向词列表（共85个词）进行提取，计算情感得分。

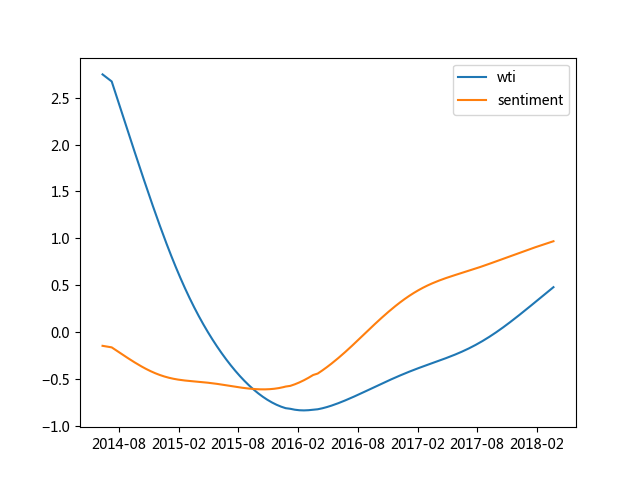






将日度、周度的新闻情感值和WTI油价进行标准化和HP滤波去掉噪声后，得到的趋势序列如下图所示。从图中可以看出，新闻情感和WTI油价的趋势有明显的相关性。对日度的新闻情感和WTI油价进行格兰杰检验，也得到了新闻情感和石油价格在滞后期为1，2和3的时候(p-value<0．01)，具有明显的格兰杰因果关系。这表明新闻情感相对于石油价格来说有重要的参照作用，新闻情感的变化早于石油价格的变化，并且在短时间内新闻情感对石油价格具有预测能力。





### 8.2.4实验设计

选取了分类准确率(PCC)作为模型好坏的评价指标。



其中M表示测试集的数量，和分别表示在t时刻石油价格走势 的预测值和真实值，特别的，(或者)表示预测石油价格将会上升(或者 下降)，同理。

为了研究新闻情感对于石油价格走势的预测能力，分别引入总共两类预测模型来进行实验研究，分别是基准模型(标记为Type-I)和加入情感信息的预测模型(标记为Type-Ⅱ)。对于基准模型，如式(4．18)所示，其输入只包括历史石油价格序列， 而对于本节新提出的模型，如式(4—19)所示，新闻情感被当做一个重要的预测变量引入到石油价格走势预测模型中。因此，考虑到上文提到的分类模型(KNN、LR、RF、GBDT和SVM)，总共有10个预测模型，分别是KNN-I、LR-I、RF-I、GBDT-I和SVM-I作为基准模型，而KNN-Ⅱ、LR-Ⅱ、RF-Ⅱ、GBDT-Ⅱ和SVM-Ⅱ作为新提出的预测模型。 关于模型的参数设置问题，针对于Type-I和Type-Ⅱ模型，所有的分类模型采用的相同的参数设定标准。

### 8.2.5预测结果分析

|  |  |  |  |
| --- | --- | --- | --- |
| 模型 | PCC | | |
| Type-I | Type-Ⅱ | 提升率 |
| KNN | 57.14% | 72.73% | 15.59% |
| LR | 57.14% | 64.71% | 7.57% |
| RF | 66.67% | 75.00% | 8.33% |
| GBDT | 64.29% | 71.43% | 7.14% |
| SVM | 47.62% | 65.00% | 17.38% |

十个预测模型的预测结果在上表中。结果表明，所有的Type-II模型的预测结果优于Type-I模型，也就是说新闻情感对于石油价格变化的预测具有重要的影响， 即所有加入新闻情感的模型预测准确率要明显高于只使用历史石油价格预测的模型。 特别地，所有Type-Ⅱ模型的预测表现显著地优于相对应的Type-I模型。Type-II模 型的平均分类准确率为67.12％，远高于基准模型的平均分类准确率57.90％。

同时，在所有的基础模型，即没有加入新闻情感的预测模型中，随机森林模型准确率最高为66.67%，其次是GBDT模型为64.29%。在引入新闻情感辅助石油价格走势预测中后，每一个分类模型都有了准确率的提升，其中支持向量机模型提升的幅度最大，达17.38％，KNN模型准确率提升了15.59％，随机森林模型提升了8.33％，逻辑回归模型提升了7.57％，GBDT模型提升了7.14%。结果表明无论对于线性预测模型还是非线性预测模型，新闻情感的引入都提高了预测准确率，而且从提升幅度来讲，非线性模型效果提升幅度远高于线性模型的提升幅度。