

非洲猪瘟舆情信息对我国猪肉价格波动影响的研究

肖莉*, 卢佳贤#, 郑丹霓, 蔡易彤, 朱文慧, 钟黛曼

华南农业大学, 广东 广州

收稿日期: 2021年12月14日; 录用日期: 2022年1月14日; 发布日期: 2022年1月25日

摘要

非洲猪瘟自2018年8月爆发以来, 对我国生猪产业和人民生活带来了重大的影响。为了实现猪瘟舆情的量化以及研究非洲猪瘟舆情与猪价的传导关系, 以及猪价空间分布情况。本文利用从LDA主题模型, 支持向量机、随机森林、梯度下降等机器学习构造情感分类器, TF-IDF计算特征词权重加权计算出非洲猪瘟舆情指数, 并通过VAR模型研究舆情指数与猪价的传导关系, 最后再利用空间自相关分析方法分析猪价的省域特征, 并用聚类分析加以验证。本文丰富了非洲猪瘟舆情分析的研究方法, 并首次将舆情的文本处理方法应用于猪瘟舆情分析, 为国内相关企业和政府对于生猪的生产调控和猪肉定价等提供了参考意见。

关键词

舆情指数, 文本挖掘, VAR模型, 空间自相关, 聚类分析

Study on the Influence of Public Opinion Information of the African Swine Fever on the Fluctuation of Pork Prices in China

Li Xiao*, Jiaxian Lu#, Danni Zheng, Yitong Cai, Wenhui Zhu, Daiman Zhong

South China Agricultural University, Guangzhou Guangdong

Received: Dec. 14th, 2021; accepted: Jan. 14th, 2022; published: Jan. 25th, 2022

*通讯作者。

#第一作者。

文章引用: 肖莉, 卢佳贤, 郑丹霓, 蔡易彤, 朱文慧, 钟黛曼. 非洲猪瘟舆情信息对我国猪肉价格波动影响的研究[J]. 数据挖掘, 2022, 12(1): 49-72. DOI: 10.12677/hjdm.2022.121007

Abstract

Since the outbreak of the African Swine Fever in August 2018, it has had a significant impact on my country's pig industry and people's lives. In order to quantify the public opinion of swine fever and study the conduction relationship between the public opinion of African swine fever and the prices of pigs, as well as the spatial distribution of pig prices. This paper uses machine learning from the LDA topic model, support vector machine, random forest, gradient descent to construct sentiment classifier, uses TF-IDF to calculate the weight of feature words to calculate the African swine fever public opinion index, and uses the VAR model to study the conduction relationship between the public opinion index and pig prices. Finally, the spatial autocorrelation analysis method is used to analyze the provincial characteristics of pig prices, and cluster analysis is used to verify it. This article enriches the research methods of African swine fever public opinion analysis, and for the first time the text processing method of public opinion to be applied to the public opinion analysis of swine fever, and provides reference opinions for domestic-related enterprises and governments on pig production regulation and pork pricing.

Keywords

Public Opinion Index, Text Mining, VAR Model, Spatial Autocorrelation, Cluster Analysis

Copyright © 2022 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

1.1. 研究背景

非洲猪瘟是影响生猪产业的主要传染病之一,我国深受其影响[1] [2]。非洲猪瘟[3]自2018年8月2日传入我国后,对我国的养猪行业造成了巨大冲击,使我国猪价产生频繁、剧烈的波动,对国民的生活产生巨大影响。在经过政府的防控后,非洲猪瘟传播速度放缓,疫情总体可控,而在2020年3月3日,非洲猪瘟疫情再次出现,而且非洲猪瘟疫苗的研发离生产上市还需要至少2年以上的时间[4],这表示非洲猪瘟疫情在我国仍未完全被控制住,我国的非洲猪瘟疫情防控任务依旧任重道远。突发事件的冲击会使得猪肉价格波动产生杠杆效应,导致猪价波动呈现非对称性、集簇性、持续性与记忆性特征[5],研究非洲猪瘟疫情如何影响猪价的波动,分析猪价的空间依赖性,这对有效应对非洲猪瘟疫情,稳住猪肉市场定价具有重大的参考价值和现实意义。

1.2. 研究现状

目前,我国对非洲猪瘟舆情指数的建立方法[6] [7]仅有利用百度指数或谷歌趋势,通过对搜索量进行加权合成,这类方法虽然获取数据简单方便,但没有充分利用非结构化数据,对文本的主题和情感进行进一步分析,存在忽略文本信息的缺点。非洲猪瘟舆情的数据挖掘关键是文本数据处理[8]。长期以来,国内外都在积极探索如何有效、准确和快速地处理文本数据,从运用较简单的文本情感词典[9] [10] [11] [12],到运用机器学习[13]-[18],以及近一两年崛起的深度学习[19] [20] [21] [22]。同时与国外相比,我

国在文本挖掘相关研究上起步稍晚，主要原因是中文文本的处理难度远大于英文文本且中文语义较为复杂[23]。总体来说，文本数据处理主要是依靠情感词典与机器学习结合，准确率和处理效率都不高[24][25]，无法满足准确了解猪瘟舆情信息的需要。

在猪价的空间传导方面，国外学者主要关注于分析国内外市场价格的空间传导[26]，在国内，谭莹[27]等发现我国猪价波动在空间上有省际间的溢出传导效应，王刚毅[28]等得出我国各省区间的生猪价格具有跨区域的空间溢出效应。

综上，在舆情分析领域，已有很多学者提出建立舆情指数的可行方法，但对于突发疫病如非洲猪瘟的舆情指数的研究还很少，且缺少对舆情的文本信息的数据挖掘。另外，对于猪肉价格在空间上的分布，已有部分学者提出相关的结论，但有关研究还是较少，且未与非洲猪瘟等动物疫病进行联系。

1.3. 研究目的

针对“非洲猪瘟疫情”焦点话题和舆情新闻的文本主题和情感倾向难以分类的问题，本文提出了一种结合 LDA 主题模型挖掘和机器学习文本情感分析的非洲猪瘟舆情指数建立方法，并将其作为非洲猪瘟舆情信息的测度指标，构建 VAR 模型分析不同时间非洲猪瘟舆情信息对猪肉价格的影响。另一方面，本文运用莫兰指数探究猪价的空间自相关性，并且对各省份的猪价进行聚类，将聚类结果与空间自相关的结果进行相互验证，从而寻找猪价在空间上的影响因素，解决由于忽略空间自相关而带来的猪价预测问题，有利于政府和企业根据非洲猪瘟舆情信息变化对猪价变化做出对应的决策。

1.4. 研究创新

现有舆情指数的测度方法多为利用舆情量进行简单加权合成，忽略了舆情文本中的有效信息，目前已有文献的非洲猪瘟舆情指数的度量方法也是如此。因此，本文提出利用特征词向量权重、文本主题分类和情感倾向分类对舆情信息进行提取，构建非洲猪瘟舆情指数，有效地利用了舆情中的文本信息，更加客观地对非洲猪瘟舆情进行度量。

另外，现有的猪价分析大多分析其在某地区或某时段的走势特征，很少将其与所在空间联系起来。因此，本文提出空间分析方法引入我国猪价的空间分布特征分析，研究我国猪价在空间上的自相关性与空间聚集特征。

2. 研究方法

2.1. 研究思路

通过爬取各大网络平台上关于非洲猪瘟的网络舆情数据，文本信息包括标题，时间，正文等，并进行数据预处理，包括去重、中文分词、去停用词等，将处理后的数据储存为语料库。

结合 LDA 主题模型、机器学习情感分类算法和特征词权重对非洲猪瘟舆情文本进行数据挖掘，建立非洲猪瘟舆情指数。

从布瑞克中国农业大数据和中国国家统计局获取到去皮带骨猪肉的价格数据，结合非洲猪瘟舆情指数建立 VAR 模型，通过 Grange 检验、脉冲响应和方差分解研究两者的时变特征和动态影响关系。

对猪价数据进行空间自相关检验，并结合聚类分析验证其结果。

技术路线图见图 1 所示。

技术路线主要可以分为对舆情文本数据的处理和舆情指数的综合计算，舆情指数与猪价的传导分析以及猪价在省域范围的空间相关性分析三部分。

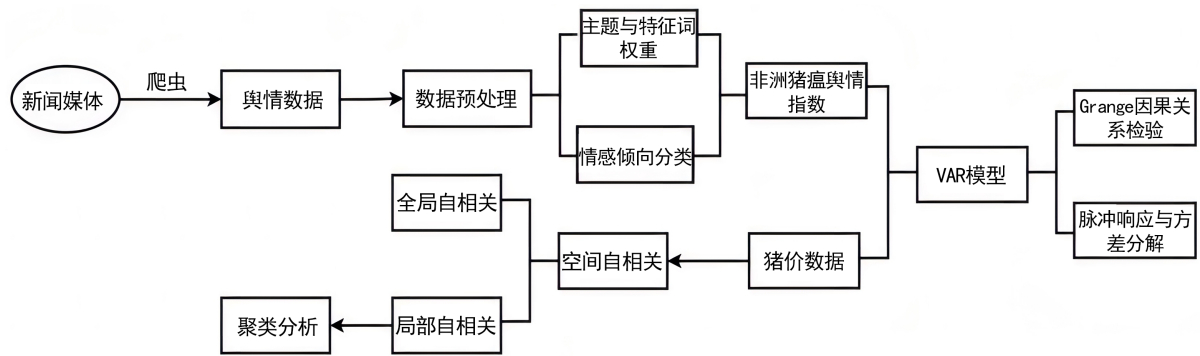


Figure 1. Technology road-map

图 1. 技术路线图

2.2. 数据挖掘方法

2.2.1. 主题挖掘

隐含狄利克雷分布(LDA)是由 Blei 等[29]提出的一种文档主题生成模型。模型假设每个主题是各个词的频率分布,文档表示为一系列主题的多项式分布。

本文基于 LDA 主题分类模型,利用 TF-IDF 算法计算特征词权重值 W_d 。TF-IDF 是基于文档频率的统计方法,基本思想为,一个词语能视为该文档的特征词,必定可以划分该文档和其他文档,即认为该词语在该文档中匹配次数较高,但在其他文本中匹配次数较低。其优点是能够过滤掉无关的词,同时保留文本的重要特征词。计算公式如下:

$$(\text{TF-IDF})_i = \text{TF}_i * \text{IDF}_i \quad (1)$$

其中 $(\text{TF-IDF})_i$ 值越高,说明该特征词对文本的重要性越大。

在主题模型领域,模型性能的度量常用困惑度(perplexity)作为指标。模型性能与困惑度成反比,困惑度越小,则语言模型越有效。其计算方法如下:

$$\text{perplexity}(W) = \exp\left(-\frac{\sum_m \ln p(W_m)}{\sum_m N_m}\right) \quad (2)$$

其中, W 为测试集, W_m 为测试集中可观测到的单词, N_m 为测试集的单词数。为了比较概率主题模型的性能,需要绘制对应的困惑度图来对概率主题模型的性能进行评价。

2.2.2. 情感分类器

为了分析非洲猪瘟舆情情感倾向,本文选择了表 1 所示的 8 种机器学习情感分析算法。通过对比各个分类算法的测试结果,选出最优的情感分类器并通过情感分类器对待分类的文本进行情感倾向分类。

Table 1. Principles of sentiment classification algorithm

表 1. 情感分类算法原理

情感分类算法	原理
朴素贝叶斯	基于贝叶斯定理,利用类别和特征项的联合概率来估计文档的类别概率。
支持向量机	在于找到一个分离超平面 H ,使数据能够正确划分并且不同类别间距离最大化。
随机森林	是决策树的组合,通过对原始数据集中的训练数据进行训练生成决策树,多数决策树的决策结果构成最终的随机森林决策结果。

Continued

AdaBoost	通过降低每个个体学习器的分类误差, 增加效果好的个体学习器的权重, 进而得到最终的集成学习器。
多层感知机	基于梯度下降法不断更新参数使损失函数减小
随机最速下降法	能够自动逃离鞍点, 自动逃离比较差的局部最优解
梯度提升	每一次建立模型时, 都是在之前建立模型损失函数的梯度下降方向, 即通过优化损失函数(loss function)来生成这些模型
LightGBM	是一个梯度 boosting 框架, 使用基于学习算法的决策树

2.3. 舆情指数的建立

定义文档 d 的特征权重 W_d 为该文档所有特征值的权重和, 计算公式:

$$W_d = \sum_{i=1}^{N_d} (\text{TF-IDF})_i, i = 1, 2, \dots, N_d \quad (3)$$

其中, N_d 为文档 d 的特征词数。

定义主题权重 t_d 为该类主题中文档数量占全体文档数量的比例, 计算公式:

$$t_d = \begin{cases} \frac{n_1}{\sum_{i=1}^m n_i} & \text{topic} = 1 \\ \frac{n_2}{\sum_{i=1}^m n_i} & \text{topic} = 2 \\ \vdots & \\ \frac{n_m}{\sum_{i=1}^m n_i} & \text{topic} = m \end{cases} \quad (4)$$

其中, m 为主题类别个数, n_i 为第 i 个主题中文档数量。

定义文档的情感值, 公式如下:

$$E_d = \begin{cases} -1 & \text{情感类别为负面} \\ 0 & \text{情感类别为中性} \\ 1 & \text{情感类别为正面} \end{cases} \quad (5)$$

定义文档 d 的舆情指数为该文档的情感值 E_d 、特征权重 W_d 、主题权重 t_d 的乘积, 用 $Index_d$ 来表示, 计算公式如下:

$$Index_d = E_d * W_d * t_d \quad (6)$$

2.4. VAR 模型的建立

VAR 模型通过对模型中的内生变量的滞后性进行回归, 从而预测和分析全部内生变量的动态关系, 其一般形式为:

$$Y_t = A_1 Y_{t-1} + \dots + A_p Y_{t-p} + B_r X_{t-r} + \varepsilon_t \quad (7)$$

其中 $t \in [1, 2, 3, \dots, n]$, Y_t 表示 k 维度内生变量向量; X_t 表示 d 维度外生变量向量; p 为滞后阶数。

本文将非洲猪瘟舆情指数作为模型的外生变量建立 VAR 模型。通过 VAR 模型衡量非洲猪瘟舆情信息对猪价的影响。

2.5. 空间自相关分析

考虑到猪价可能存在空间相关性，省份与省份之间的猪价可能互相影响，为了验证是否存在这种空间相关性，本文采用莫兰指数进行分析。

2.5.1. 空间权重矩阵

定义邻接权重矩阵，公式如下：

$$w_{ij} = \begin{cases} 1, & i \text{ 与 } j \text{ 空间邻接} \\ 0, & i \text{ 与 } j \text{ 空间不邻接} \end{cases} \quad (8)$$

其中， w_{ij} 为区域 i 与区域 j 之间的距离。

由于相邻省份之间的猪价的相互影响可能更大，当超过关键距离时，影响权重可能会出现快速下降。故本文定义地理距离矩阵，将“反距离”和“固定距离范围”二者结合，在距离范围指定的距离范围内的相邻要素具有相等的权重，距离范围外的权重为距离取倒数。

2.5.2. 莫兰指数

莫兰指数分为全局莫兰指数和局部莫兰指数。定义全局莫兰指数为如下公式：

$$I = \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{S^2 \sum_{i=1}^n \sum_{j=1}^n w_{ij}} \quad (9)$$

其中 x_i 和 x_j 分别代表空间要素(猪价) x 在空间单元(省级行政区) i 和 j 的值， \bar{x} 为猪价 x 的均值，

$S^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$ 为样本方差， w_{ij} 为空间权重矩阵的元素， n 为省级行政区的数量(本文为 30 个)。全局莫兰指数是空间相关性的度量，用于判定空间相互之间是否存在相关关系。

某个省份附近的空间集聚情况可用局部莫兰指数表示，定义局部莫兰指数为如下公式：

$$I_i = \frac{(x_i - \bar{x}) \sum_{j=1}^n w_{ij} (x_j - \bar{x})}{S^2} \quad (10)$$

I_i 的取值介于 -1 和 1 之间，当 I_i 小于 0 时，表示为不同属性之间聚集，即高值与低值集聚； I_i 大于 0 时，表示为相同属性之间聚集，即高值与高值集聚，低值与低值集聚；如果 I_i 接近 0，则不存在空间相关性。

2.6. 聚类分析

聚类分析是研究样本分类的分析方法。本文基于 K-means 聚类算法，将各省级行政区的猪价之间的价差作为评估猪价相似程度的评价指标，按照价差，将各省级行政区划分为 k 个地区集合，集合内价差越小，则集合之间的相似性越大。若省份之间的猪价存在空间自相关关系，则省份与省份之间的猪价会互相影响，从而呈现相近的变化趋势和价格，因此在聚类分析中具有空间自相关性的省份会归属于同一个类别[30]。

若样本被划分为 k 个类别，设为 (C_1, C_2, \dots, C_k) ，则误差平方和可以表示为：

$$E = \sum_{i=1}^k \sum_{x \in C_i} \|x - \mu_i\|^2 \quad (11)$$

其中， μ_i 表示为集合 C_i 的质心，为 C_i 的均值向量，表示为：

$$\mu_i = \frac{1}{N_i} \sum_{x \in C_i} x \quad (12)$$

其中, N_i 表示为 C_i 的样本数。在聚类过程中, E 表示为所有样本子集的总误差平方和。对于不同样本集得到的不同分类结果, 集合 C_i 的质心 μ_i 和样本子集也会发生改变, 从而得到不同的 E 值, 在 E 达到最小时, 聚类的效果最好。

3. 实证分析

3.1. 数据采集及预处理

3.1.1. 舆情数据来源

采集各大网站平台上 2016 年 1 月至 2020 年 8 月关于“非洲猪瘟”相关网络舆情数据, 共计 47,541 条。对“非洲猪瘟”文本数据的来源进行统计分析, 得到文本数据来源分布图, 如图 2 所示。由图 2 可知, 新闻媒体、搜索引擎是“非洲猪瘟”舆情信息的主要来源。

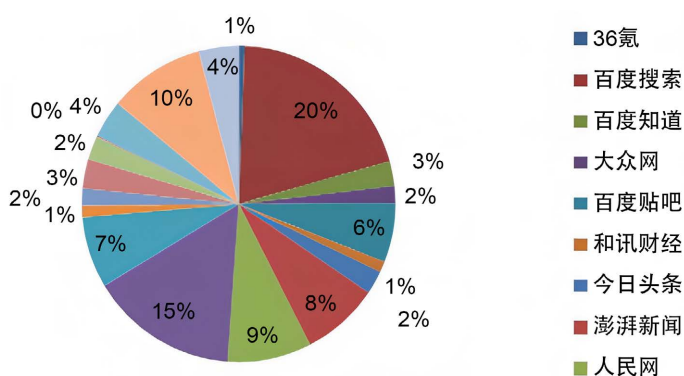


Figure 2. Data sources

图 2. 数据来源

3.1.2. 猪瘟数据来源

猪价数据来自布瑞克中国农业大数据和中国国家统计局年鉴, 选取省级行政区 2018 年 6 月到 2020 年 7 月的猪价月度数据和 2016 年 1 月 1 日至 2018 年 9 月 4 日的去皮带骨猪肉全国周度数据进行分析, 没有考虑其中, 除西藏、台湾、香港和澳门猪价数据缺失, 其他 30 个省级行政区。

3.1.3. 数据预处理

新闻数据预处理包括缺失值处理(剔除新闻时间或正文这类关键字段缺失的新闻), 去重, 对文本数据进行文本分词, 停用词过滤。通过预处理, 可以减少无效信息的干扰, 保留文本的有效信息, 提高文本处理的效率和精确度, 最终得到有效数据 28,473 条。之后, 从时间维度对每月的舆情数据量进行统计, 得到每月舆情量变化如图 3 所示。

由图 3 可知, 从 2016 年 2 月至 2018 年 7 月, 我国未发现非洲猪瘟病例, 此时舆情量较低。

直到 2018 年 8 月 3 日, 沈阳发现中国首例非洲猪瘟病例, 相关舆情总量迅速上升达到峰值。由于政府扑杀病猪、补贴养殖户、稳定物价等有力措施, 舆情量于 2018 年 10 月跌落至较低水平, 此后一直略有波动。

2020 年年初, 河南和四川排查出多起有关违法违规调运引发的非洲猪瘟病例。自此, 舆情量再度飙升。

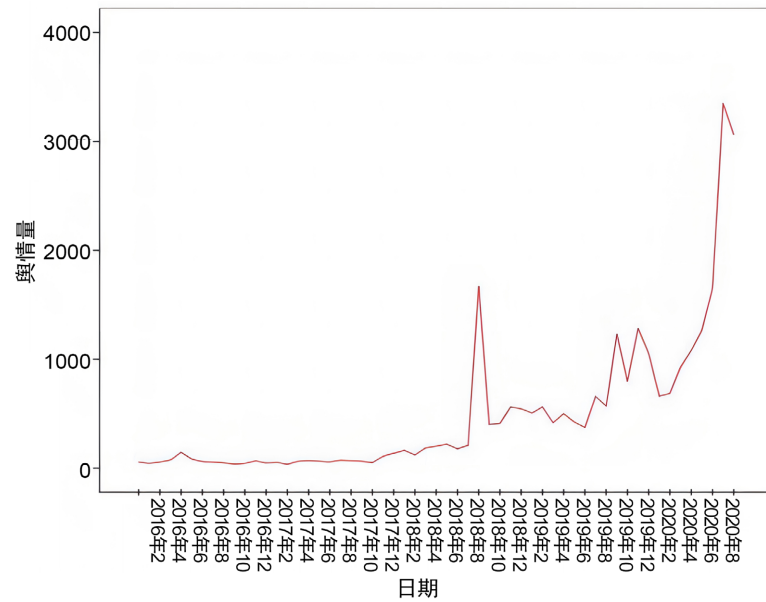


Figure 3. Monthly changes in public opinion
图 3. 每月舆情变化量

3.2. 非洲猪瘟舆情指数的建立

3.2.1. 主题挖掘

① LDA 主题模型

对非洲猪瘟舆情数据应用 LDA 主题模型建模，并计算模型困惑度如图 4 所示，其纵坐标即困惑度的负自然对数值。

由图 4 可知，当主题数设置为 5 时，困惑度最低，模型效果最佳。应用对应的 LDA 主题模型进行训练，得到每个模型内的主题词及对应权重，生成最终的主题分类结果。

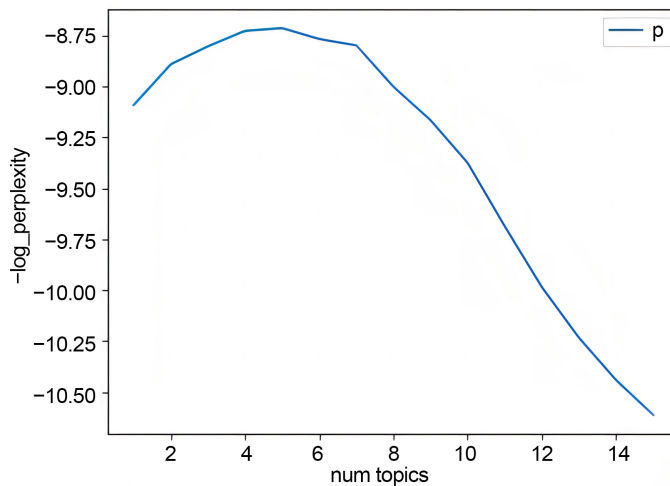


Figure 4. Perplexity graph
图 4. 困惑度图

计算五个主题对应的主题权重值如下：

Continued

情感分类	支持向量机			随机最速下降法		
	精准率	召回率	F 值	精准率	召回率	F 值
正面	0.70	0.08	0.15	0.27	0.11	0.15
中性	0.70	0.97	0.81	0.81	0.76	0.78
负面	0.56	0.27	0.37	0.43	0.66	0.52
准确率		0.686			0.645	
情感分类	随机森林			梯度提升		
	精准率	召回率	F 值	精准率	召回率	F 值
正面	0.62	0.12	0.21	0.49	0.22	0.30
中性	0.73	0.91	0.81	0.76	0.91	0.83
负面	0.58	0.47	0.52	0.63	0.50	0.56
准确率		0.701			0.719	
情感分类	AdaBoost			LightGBM		
	精准率	召回率	F 值	精准率	召回率	F 值
正面	0.32	0.20	0.25	0.37	0.11	0.17
中性	0.71	0.92	0.80	0.77	0.90	0.83
负面	0.53	0.21	0.30	0.57	0.54	0.55
准确率		0.657			0.709	

精准率是所有被预测为正的样本中实际为正的样本的比例，召回率是实际为正的样本中被预测为正样本的改了，而 F 值则是描述精准率和召回率的关系，即 $2 * \text{精准率} * \text{召回率} / (\text{精准率} + \text{召回率})$ 。最后得出的准确率即为预测正确的结果占总样本的百分比。

由表 3 的测试结果可知，梯度提升分类算法准确率最高，达到 0.719，同时由于负面情感被误判的代价比较高，在兼顾精确率和召回率下，梯度提升分类算法的负面情感所算得的 F 值较其他分类算法高，故认为梯度提升分类算法为最佳的情感分类器。接着，用训练好的分类算法对非洲猪瘟舆情文本数据进行情感倾向性分类，得到的分类结果如图 10 所示。

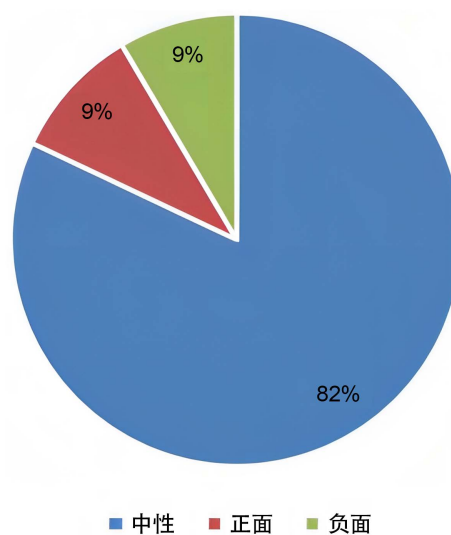


Figure 10. Proportion of public opinion
图 10. 舆情占比

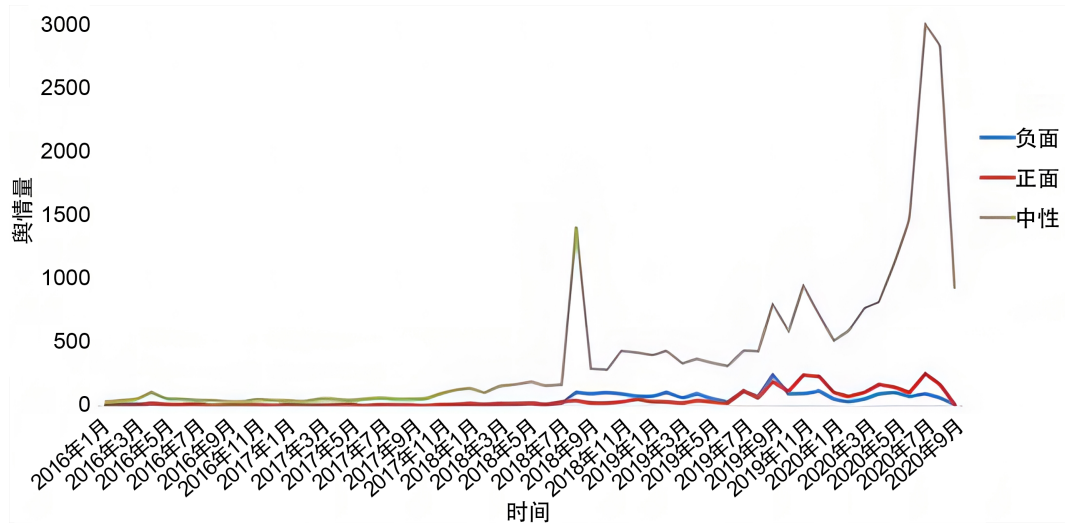


Figure 11. Changes in public opinion

图 11. 舆情变化

由图 10 的情感分布来看, 全网有关非洲猪瘟信息的情绪以中性为主, 主要内容为非洲猪瘟引起的肉价变动, 政府采取的相应措施及社会各界对于疫情状况的理性分析。正面舆情和负面舆情数量相当, 各占据总量的 9%。负面舆情主要为非洲猪瘟传播引起的担忧和恐慌, 养殖户对于生意亏损的焦虑不满, 市民对于非洲猪瘟过后肉价上涨的抱怨。正面言论则主要与政府针对非洲猪瘟采取的得力措施及非洲猪瘟得到控制、猪肉价格逐步恢复正常的良好情况有关。

根据非洲猪瘟舆情文本数据的情感倾向分类结果, 统计每月的舆情情感变化如图 11 所示, 由于中性数据量较多, 不易观察正负面的结果, 绘制了每月正负面的舆情变化图, 如图 12 所示。

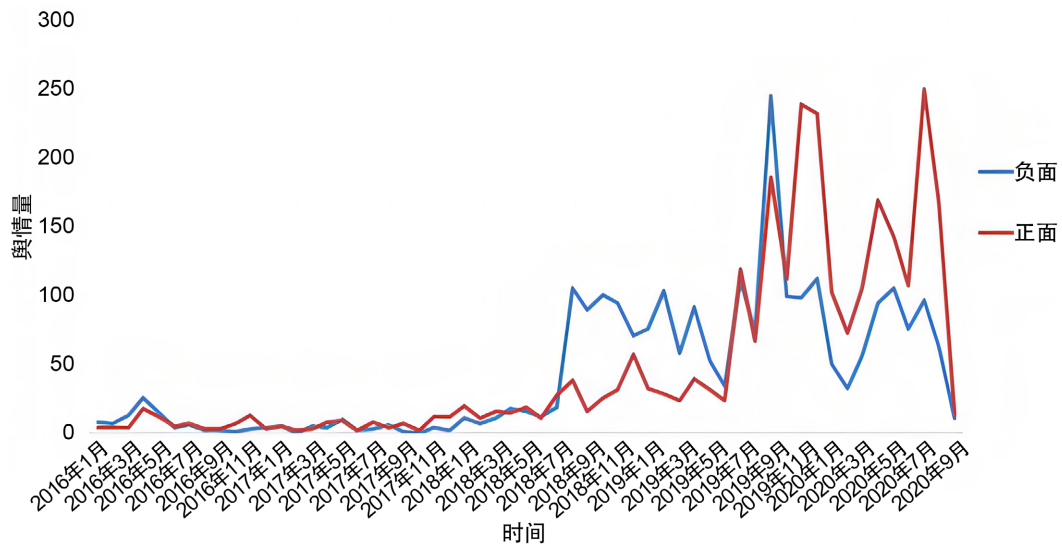


Figure 12. Positive and negative public opinion changes

图 12. 正负面舆情变化

由图 11, 图 12 可知, 从 2016 年 1 月至 2018 年 7 月, 我国尚未出现非洲猪瘟病例。网上有关猪瘟的舆情, 主要为普及相关知识的中性言论。

2018年8月,中国出现首例非洲猪瘟病例。从2018年7月至2019年4月,中性舆情达到峰值后略有下降,但一直维持在较高水平。负面舆情和正面舆情数量上升,且负面舆情多于正面舆情。这说明非洲猪瘟引起了社会各界的广泛关注,其对人民生活造成了较大困扰。

从2019年1月至2019年3月,中性舆情仍然维持在较高水平,负面舆情略有下降,这说明政府补贴养殖户、调节肉价等一系列举措起到了一定成效。

2019年7月以后,由于非洲猪瘟再度接连发生,加上猪瘟对于生猪存栏量和肉价的后续影响逐渐显现,舆情总体呈上升趋势,言论情绪以中性为主,正面舆情多于负面舆情。

3.2.3. 舆情指数建立

为了更好地观察各月非洲猪瘟疫情舆情指数变化趋势,本文根据舆情指数计算公式各个文档的舆情指数,将每个文档的舆情指数求绝对值再求代数和得到每月舆情总数,绘制2016年1月至2020年8月各月非洲猪瘟舆情指数折线图如图13所示,直观反映每月舆情指数及其变化趋势。

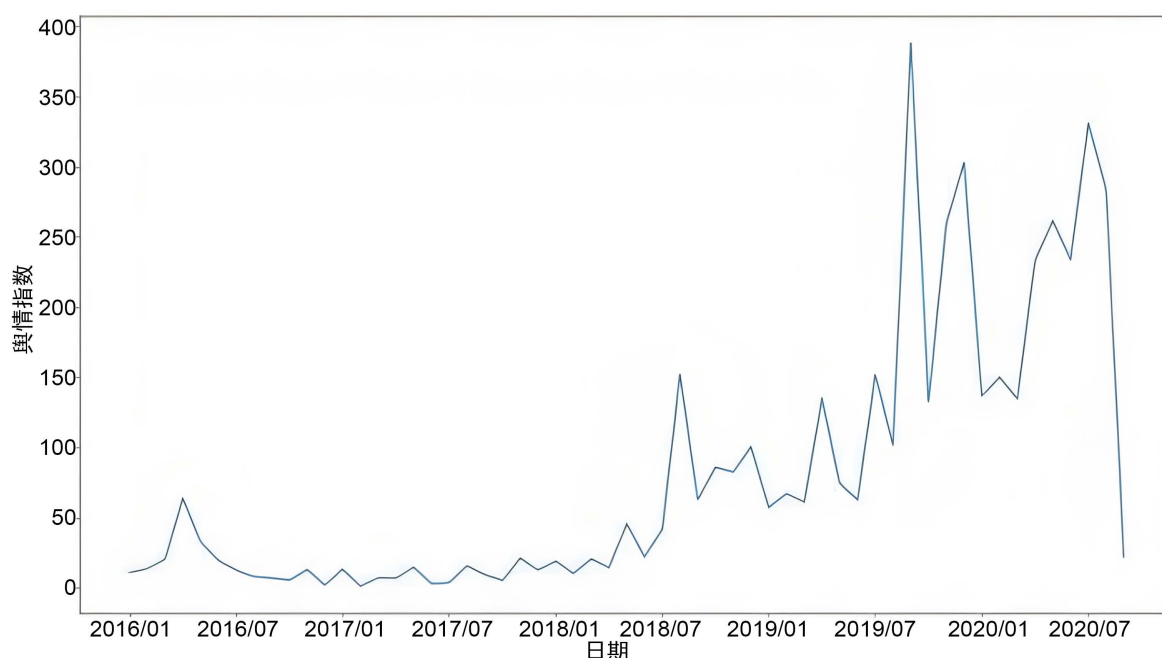


Figure 13. Line chart of monthly public opinion index

图 13. 每月舆情指数折线图

由图13可知,从2016年1月开始,非洲猪瘟疫情在各媒体平台开始有一定的搜索量,舆情指数较低且趋于平稳;在2018年8月,非洲猪瘟在国内爆发,各搜索引擎舆情指数暴涨,波动较大,并在2018年8月达到局部高峰,表明非洲猪瘟在此时是社会关注热点问题,在之后一年时间内,非洲猪瘟一直有较高的热度;2019年9月,非洲猪瘟疫情舆情指数飙升,在调查过程中我们发现,在2019年9月非洲猪瘟在国内大面积爆发,全国多省生猪供应不足,猪价飙涨,猪瘟疫情成为全民讨论热点,这与舆情指数折线图所反映的结果一致。

3.3. 基于VAR模型的猪价与舆情指数传导研究

3.3.1. 猪肉价格数据处理

本文对猪肉价格用线性插值方法进行插值补缺数据处理,得到图14。

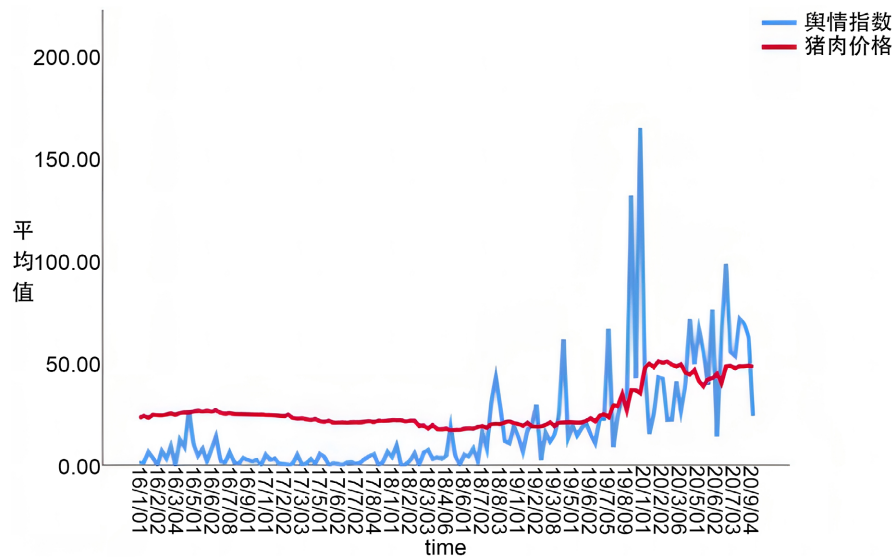


Figure 14. Line chart of pork prices and public opinion index
图 14. 猪肉价格与舆情指数折线图

从图 14 可以看出，猪肉价格与舆情指数的变化趋势较为接近，峰值的位置存在一定的滞后关联。为了准确地衡量这种影响的大小和方向，本文通过 VAR 模型进行分析。

3.3.2. 模型建立

① ADF 平稳性检验

由表 4，猪肉价格原始数据不平稳，一阶差分后，猪肉价格和舆情指数均平稳，故猪价和舆情指数同阶单整。

Table 4. Stationarity test results
表 4. 平稳性检验结果

变量	ADF 统计量	临界值			P 值	平稳性
		1%	5%	10%		
猪肉价格 Y	-1.7373	-3.9963	-3.4284	-3.1376	0.7317	不平稳
舆情指数 X	-8.8796	-3.9961	-3.4283	-3.1375	0.0000	平稳
ΔY	-6.8982	-2.5745	-1.8421	-1.6528	0.0000	平稳
ΔX	-9.9602	-2.5748	-1.8422	-1.6158	0.0000	平稳

注： Δ 表示一阶差分后的变量。

② 滞后阶数选择

Table 5. Lag order and each criterion information
表 5. 滞后阶数与各准则信息

Lag	LogL	LR	FPE	AIC	SC	HQ
0	-1321.41	NA	267.0071	11.26303	11.29247	11.2749
1	-1239.75	161.2214	137.8823	10.60215	10.69048	10.63776
2	-1228.52	21.98163	129.6555	10.54062	10.68784	10.59997

Continued

3	-1214.43	27.35191	118.9831	10.4547	10.6608	10.53779
4	-1203.24	21.52428	111.9244	10.3935	10.65849	10.50033
5	-1197.15	11.60504	109.9603	10.37574	10.69961	10.50631
6	-1191.85	10.01818	108.7579	10.36465	10.74742	10.51896
7	-1171.25	38.57362	94.43898	10.22336	10.66501	10.40141
8	-1151.86	35.96632	82.86163	10.09242	10.59296*	10.29421*
9	-1145.19	12.26117*	81.01636*	10.06970*	10.62912	10.29523

注：*号标注为该准则下最优数值。

AIC 又称赤池信息量准则，它可以权衡所估计模型的复杂度和此模型拟合数据的优良性；HQ 统计量构造思想如 AIC 类似，也是依照自变量个数施加“惩罚”，只是力度不同；SC 又称施瓦兹准则，是通过比较不同分布滞后模型的拟合优度来确定合适的滞后期长度；LR 又称似然比准则，是反映灵敏度和特异度的综合指标；FPE 准则即最小最终误报误差准则。以上准则都能在一定角度对模型的优劣进行评价。

由表 5 可得，滞后阶数为 9 时，LR、FPE、AIC 达到最优；滞后阶数为 8 时，SC 和 HQ 达到最优，因此最终选定 VAR(9)模型作为拟合模型。

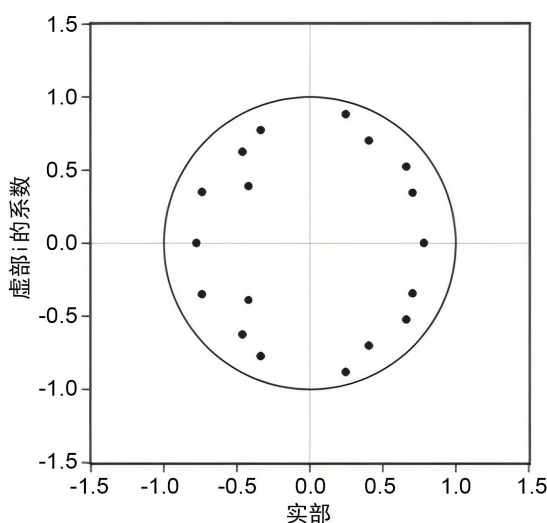


Figure 15. AR characteristic polynomial inverse root graph of VAR(9)

图 15. VAR(9)的 AR 特征多项式逆根图

由图 15 的结果，可以看出各点均在单位圆内，故 VAR(9)模型是平稳的。

3.3.3. Granger 因果关系检验

Table 6. Granger causality test results

表 6. Granger 因果关系检验结果

H_0	Chi-sq	Df	Prob	Result
X 不是 Y 的 Granger 原因	22.84736	9	0.0065	拒绝 H_0
Y 不是 X 的 Granger 原因	85.04049	9	0.0000	拒绝 H_0

由表 6 中 Granger 因果关系检验结果可得,猪肉价格 Y 和舆情指数 X 互为 Granger 原因,说明非洲猪瘟舆情信息会对猪价产生冲击,同时猪价波动又会引起网络舆情的增长,影响非洲猪瘟舆情指数,它们之间存在着传导关系。

3.3.4. 脉冲响应分析

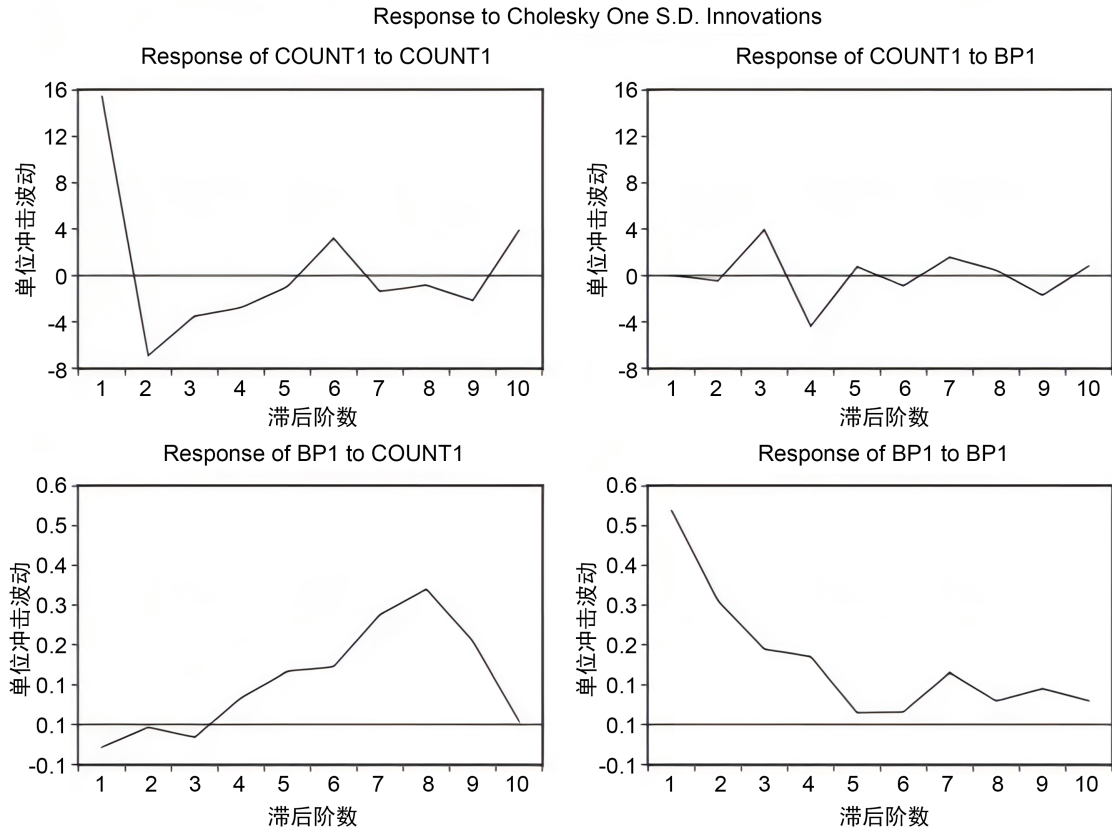


Figure 16. Impulse response results

图 16. 脉冲响应结果

由图 16 的脉冲响应结果,当猪肉价格受到非洲猪瘟舆情的一单位标准差的冲击时,先是在滞后 3、4 期时达到最大冲击,而后呈现波动下降的趋势,这说明随着非洲猪瘟舆情对猪价产生了一定地冲击,但这种冲击随着非洲猪瘟舆情的控制而逐渐减弱。

从另一个方面看,当猪肉价格对非洲猪瘟舆情的一单位标准差的冲击时,先是逐步上升直到滞后 8 期达到峰值,而后快速下降,这说明猪价价格的波动先是导致非洲猪瘟舆情的长期逐步增长,而后在政府对猪价采取调控后,对舆情的冲击开始快速减弱。

3.3.5. 方差分解分析

本文采用 CholeskyDecomposition 法研究非洲猪瘟舆情指数对猪肉价格冲击的贡献度。

由图 17 可知,猪肉价格在第 1、2 期时仅受自身影响,在第 3 期时受到非洲猪瘟舆情影响,冲击的影响逐渐增大并在第 4 期趋于稳定,这说明舆情的快速增加,使得公众对猪肉产生一定恐慌情绪,市场需求随之受到一定影响而下降,而非洲猪瘟疫情又导致猪肉的出栏量不足,猪肉价格随之逐步增长,非洲猪瘟舆情在一定程度上加剧了猪肉价格的波动。

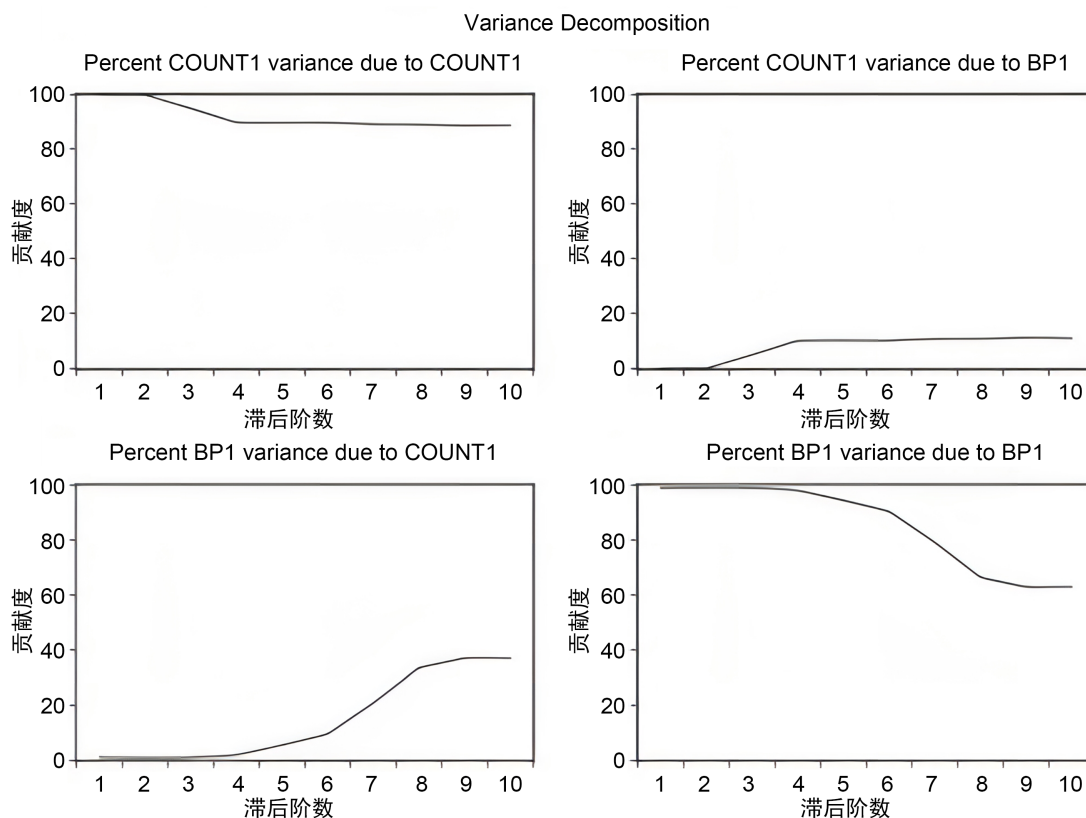


Figure 17. Variance decomposition results
图 17. 方差分解结果

3.4. 猪价的空间相关性分析

3.4.1. 全局自相关

本文利用地理邻接和地理距离理论构建两类空间权重矩阵，计算猪价的莫兰指数，见表 7。

Table 7. Moran index based on pig prices of 2 spatial weight matrices from June 2018 to July 2020

表 7. 2018 年 6 月至 2020 年 7 月基于 2 种空间权重矩阵猪价的莫兰指数

时间	邻接权重矩阵		地理距离矩阵	
	Moran's	P 值	Moran's	P 值
2020 年 7 月	0.218224	0.021529	0.136239	0.000914
2020 年 6 月	0.246812	0.011425	0.167627	0.0001
2020 年 5 月	0.215296	0.025387	0.182928	0.00003
2020 年 4 月	0.199286	0.035402	0.175088	0.000055
2020 年 3 月	0.375094	0.000288	0.231631	0
2020 年 2 月	0.554598	0	0.308624	0
2020 年 1 月	0.465327	0.000011	0.26857	0
2019 年 12 月	0.325592	0.001487	0.224644	0.000001
2019 年 11 月	0.265814	0.007823	0.141642	0.000801
2019 年 10 月	0.213677	0.025626	0.093706	0.013567

Continued

2019年9月	0.151371	0.093333	0.078304	0.029377
2019年8月	0.090024	0.224444	0.036967	0.149007
2019年7月	0.187286	0.047461	0.093132	0.014446
2019年6月	0.09064	0.238966	0.005599	0.42458
2019年5月	0.150852	0.08842	0.004652	0.443586
2019年4月	0.254875	0.008553	0.058882	0.069986
2019年3月	0.29031	0.003619	0.104184	0.007759
2019年2月	0.334083	0.001027	0.140802	0.000805
2019年1月	0.361753	0.000402	0.184619	0.000027
2018年12月	0.337916	0.000932	0.176018	0.000059
2018年11月	0.316364	0.001551	0.148246	0.000421
2018年10月	0.325498	0.000598	0.106754	0.004499
2018年9月	0.270586	0.003917	0.050959	0.087608
2018年8月	0.157958	0.087088	0.017522	0.32082
2018年7月	0.118275	0.175246	-0.01225	0.671702
2018年6月	0.103775	0.221529	-0.00631	0.592231

由表 7 可知, 基于邻接权重矩阵计算的莫兰指数显示, 猪价在空间上具有正相关性, 在显著性水平为 0.05 的检验水准下, 除部分月份如 2018 年 6 月至 8 月等的检验 P 值没有通过检验, 接受了不存在空间相关性的原假设, 其余月份的莫兰指数的检验 P 值均通过了检验。

基于地理距离矩阵计算的莫兰指数总体上为呈正值, 且绝大部分月份的莫兰指数的检验 P 值都通过了检验。

因此, 我国猪价在空间上存在高值与高值集聚、低值与低值集聚的分布, 且空间依赖性可能具有动态性或仅存在于局部省级行政区, 且由图 18 所示, 两个空间权重矩阵计算的莫兰指数都不平稳, 这表明猪价之间的空间相关性相对不稳定。

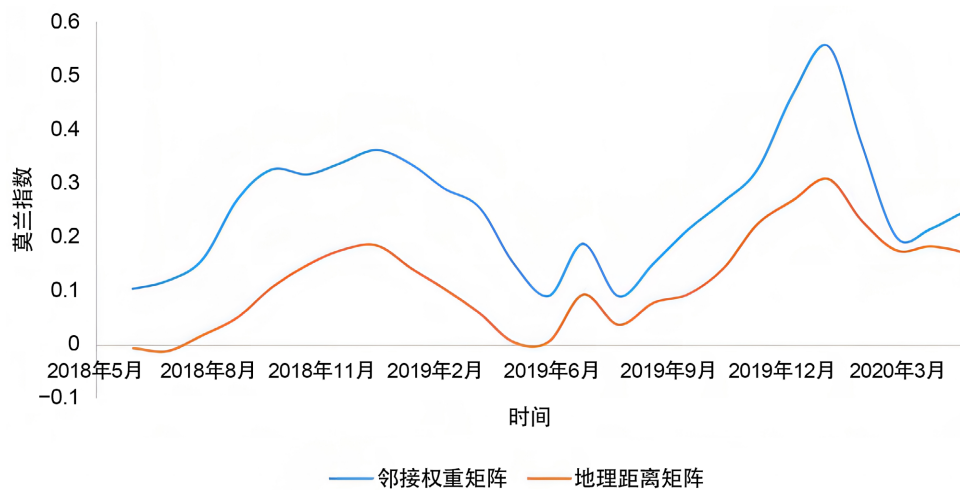


Figure 18. Pig prices moran index from June 2018 to July 2020

图 18. 2018 年 6 月至 2020 年 7 月猪价莫兰指数

3.4.2. 局部自相关

为了更清晰地看出各省级行政区的空间相关性，本文选取了猪价有显著空间自相关的月份，进行局部自相关分析。

从个别省份来看，其类型变化分别为：

1) 江苏：2018年9月到10月从高-高型变成低-高型，说明江苏的猪价受到邻省猪价的抑制。2019年3月变为高-高型，猪价上升，逐渐与相邻猪价高水平省份齐平。

2) 内蒙古：猪价类型变化较大，在2018年9月到10月从高-低型变为低-低型，说明此时内蒙古猪价对相邻省份有抑制作用，后在11月又变为高-低型，说明内蒙古自身猪价是较低的，但逐渐超过相邻省级行政区成为猪价高水平地区，在2019年1月变为低-低型，后在2020年5月从低-低型变为高-低型，而在6月又变为低-低型，猪价下降，又变为与猪价低水平地区齐平。

3) 东三省：黑龙江、吉林和辽宁在2018年9月到10月都从低-高型变为低-低型，自身猪价逐渐降低，对相邻区域猪价有抑制作用。

4) 四川：2019年7月从高-高型变为低-低型，自身猪价和相邻省份由高水平变为低水平，说明此时四川省与相邻省份的猪价关系密切，且呈正相关关系，后在10月从低-低型变为高-低型，四川省的猪价自身较低，但逐渐超越相邻省份成为猪价高水平地区，此时四川省对相邻省份的猪价无带动作用。

从总体上来看，猪价的高-高集聚型主要集中在东部地区和中部地区，可以看出这些省份的猪价都持较高水平，东北地区的猪价主要呈低-低型，猪价持较低水平，西部地区的猪价类型较不稳定，受相邻省份的影响较大。

本文选取了2019年11月2020年1月这两个有代表性意义，能描述这种集中效应的猪价图，见图19、图20。由图19和图20可以看出猪价在各省水平相差不大，但东部和中部地区的猪价较高，西部地区猪价会偏低。

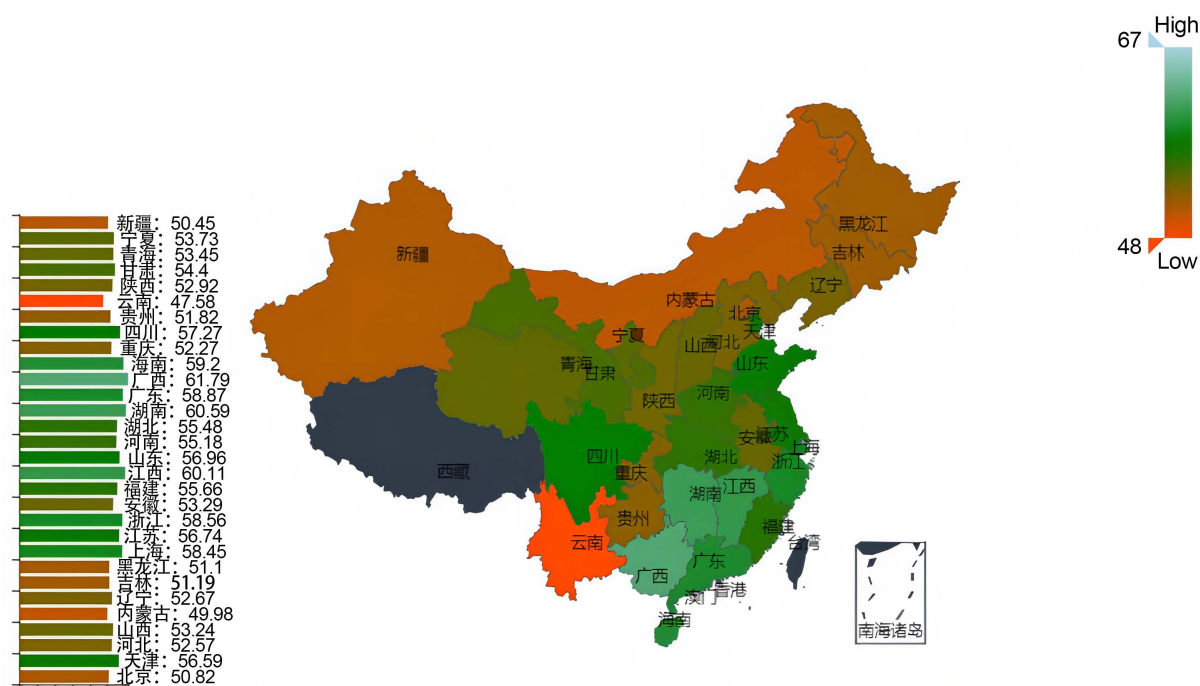


Figure 19. Local autocorrelation of pig prices in 30 provinces in November 2019

图 19. 2019 年 11 月 30 个省的猪价局部自相关情况



Figure 20. Local autocorrelation of pig prices in 30 provinces in January 2020
 图 20. 2020 年 1 月 30 个省的猪价局部自相关情况

3.5. 聚类分析

当省份之间的猪价存在空间自相关关系时，各个省份的猪价变化趋势会受到周围省份的影响，此时存在空间自相关性的省份的猪价变化趋势会由于其相互影响而相近，则可以以猪价的时间变化数据建立指标体系，进行聚类分析，寻找具有相同类别的省份集合。当省份之间存在自相关关系时，这些省份会由于变化趋势上和价格上的相同而聚集为同一个类别，根据聚类结果对空间自相关分析结论进行验证。

为保证结论的可靠性，本文的聚类分析选取空间自相关分析中应用的各个省份 2018 年 6 月到 2020 年 7 月的猪价数据，以每个月度的猪价数据作为聚类指标，对各个省份进行聚类，对聚类的结果以地图的形式可视化。

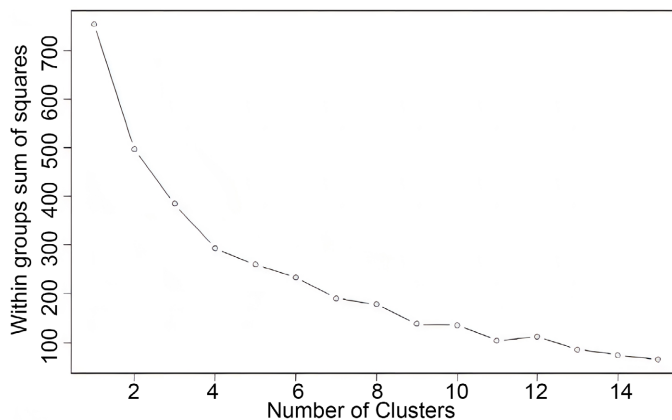


Figure 21. Results of the elbow method
 图 21. 手肘法结果图

在进行聚类分析之前,利用手肘法和 Calinsky 法对聚类应取的类别个数进行判断,由图 21 可知,手肘法函数图像的拐点在 4 和 7 取到,则聚类应该取定的类别数为 4 或者 7。由图 22 可知,Calinskicriterion 在 2 和 4 取到极小值,而在 7 取到最大值,聚类应该取定的类别数为 2 和 4 且不能选择 7。而结合实际情况分析,当聚类个数取定为 2 时,虽然聚类的结果较好,但是由于类别个数太少,将无法对省份之间的空间自相关关系进行分析,聚类为 2 的结果没有分析的意义。因此,取定聚类个数为 4 进行聚类分析。

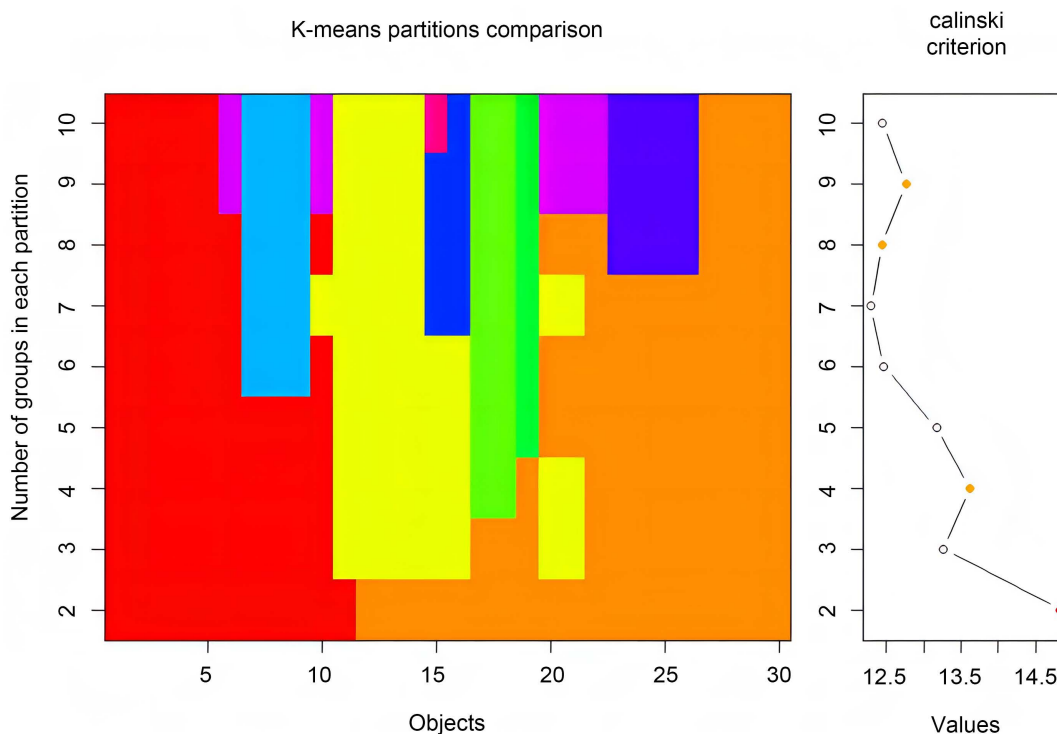


Figure 22. Calinsky criterion method result graph

图 22. Calinsky 准则法结果图

由图 23 的聚类结果可以看出,聚类结果将省份大致切分为东部地区,东北部地区,西部地区和中部地区四个部分。

结合图 19 和图 20 的结果,这四个部分的分布特征为:

- 1) 东部地区类别对应的省份在空间自相关分析中呈现了高 - 高型空间自相关关系,这些地区的猪价较为接近且比较高;
- 2) 中部地区类别对应的省份在空间自相关分析中也呈现了高 - 高型空间自相关关系,这些地区的猪价较为接近,但没有东部地区高;
- 3) 东北部地区类别对应的省份在空间自相关分析中呈现了低 - 低型空间自相关关系,这些地区的猪价较为接近且比其他省份要低;
- 4) 西部地区分类仅有两个省份:新疆和内蒙古,这两个省份的猪价类型不稳定,这两个省份的猪价变动幅度比较大。

从图 23 可以注意到,归属于中部地区类别的省份越接近东部地区类别,则它们的价格相比起远离东部地区类别的省份会更高一点,这说明了东部地区在猪价的相互影响作用中起到了主导因素,即东部地区在空间上对猪价的影响相比起中部地区更大一点。



Figure 23. Pig prices clustering result map
图 23. 猪价聚类结果地图

4. 研究结论与对策建议

4.1. 研究结论

非洲猪瘟舆情文本资料中富含公众以及媒体关注的主题焦点、情感倾向、关注热词，因此衡量非洲猪瘟舆情指数时，不能仅从舆情量进行简单加权计算，还要结合舆情的文本信息，提取其中的有效信息并进行量化，结合舆情的主题、情感、关键词等进行综合评价。

猪肉价格受非洲猪瘟舆情信息的冲击和影响，并且在非洲猪瘟舆情指数发生变动后的不同时期对猪肉价格的冲击大小存在差异，即非洲猪瘟舆情指数滞后影响于猪肉价格，这种冲击再滞后 3、4 期时强度最大。因此，通过对舆情指数进行监控预警，可以指导生猪企业对猪肉进行合理定价，以及政府相关部门对猪价的宏观调控等。

非洲猪瘟舆情指数同样也会受到猪肉价格变动的冲击和影响，并且舆情指数在猪价变动后的一定时期内这种冲击逐渐增强，持续时间较长，达到峰值后开始快速减弱。因此，可以从舆情管理的视角，对网络舆情进行监测与预警。

猪肉价格在我国存在全局空间相关性与局部空间相关性，即一个省份或地区的猪价会与其相邻或距离较近的省份或地区具有空间上的关联，一个省份或地区的猪价降低或增长会对相邻省份起到抑制或带动作用，并且会形成猪价上的高-高、高-低、低-低和低-高聚集区域，但这种聚集区域并不是恒定的，随着时间的推移聚集区域也会有一定的转移。

我国 30 个省份的猪价空间聚集大致可划分为四块区域，分布为猪价水平较高、呈现高-高聚集且猪价易于影响其他地区的东部地区，猪价水平中等偏高、呈现高-高聚集且猪价易受东部区域影响的中部地区，猪价偏低、且呈现低-低聚集的东北地区和猪价类型不稳定易受外部影响的西部地区，这与我国的四大经济区域划分较为接近，说明了各地区猪价之间存在着普遍的空间经济联系。

4.2. 对策建议

结合本文的研究结论,为更好地对非洲猪瘟舆情变化情况进行反映,保证猪肉产业稳定发展,提出以下对策建议:

生猪企业和政府部门应该加强对非洲猪瘟等突发重大事件的监测,基于互联网大数据的背景下,建立非洲猪瘟舆情预警系统,设置非洲猪瘟舆情预警线,当网络舆情超过一定阈值后,指导生猪企业和政府部门及时对猪瘟舆情做出反应,对猪价进行合理的定价与调控,从而稳定猪肉行业市场,以降低非洲猪瘟冲击下的经济损失。

从舆情管理的视角,政府应做好非洲猪瘟疫情的防控与宣传,普及相关疫情知识,为民众答疑解惑,及时辟谣,对网络舆情进行有序引导,避免突发重大疫情导致的网络负面舆情爆发,从而影响人民生活与猪肉市场的供求平衡,降低非洲猪瘟舆情带来的负面影响。

政府要协调好各个经济区域的猪价水平,避免猪价的不稳定波动和区域间价格的传导产生的不良影响。如对于经济不发达且猪价不稳定的西部地区要加强猪肉价格的调控,防止不稳定的猪价对于人民的生产生活造成不良影响;对于经济发达且猪价偏高的东部地区进行猪价的合理定价,防止过高的价格水平影响人民的生活质量;对于易受东部地区价格影响的中部地区,要及时关注其猪价水平,防止其受影响而产生较大波动;对于猪价持续偏低的东北地区,政府要关注生猪养殖户的生产经营情况,防止“肉贱伤农”。

基金项目

国家社会科学面上项目“中国农业高质量发展的统计测度与实现路径研究”(批准号:21BTJ057)。

参考文献

- [1] 胡浩,戈阳.非洲猪瘟疫情对我国生猪生产与市场的影响[J].中国畜牧杂志,2020,56(1):168-172.
- [2] 梁兴群,夏庆利.非洲猪瘟疫情对我国生猪产业的影响[J].饲料与畜牧,2019(6):58-63.
- [3] 赵静.探讨非洲猪瘟对我国养猪业的影响与防控[J].农家参谋,2020(23):63.
- [4] 肖和良.浅析中国非洲猪瘟疫苗研制最新进展[J].中国畜牧业,2019(21):53-54.
- [5] 苗珊珊.突发事件信息冲击对猪肉价格波动的影响[J].管理评论,2018,30(9):246-255.
- [6] 马名慧,邵喜武.非洲猪瘟疫情下我国生猪产业价格传导机制研究——基于VAR模型的实证分析[J].价格月刊,2020(3):7-14.
- [7] 段琮琮,刘灵芝.非洲猪瘟疫情影响下我国畜禽产品价格波动的动态关系研究——基于舆情管理视角[J].农业现代化研究,2020,41(4):678-686.
- [8] 杨霞,黄陈英.文本挖掘综述[J].科技信息,2009(33):82+99.
- [9] 王勇,吕学强,姬连春,肖诗斌.基于极性词典的中文微博情感分类[J].计算机应用与软件,2014,31(1):34-37+126.
- [10] 张成功,刘培玉,朱振方,方明.一种基于极性词典的情感分析方法[J].山东大学学报(理学版),2012,47(3):47-50.
- [11] Turney, P.D. and Littman, M.L. (2003) Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Transactions on Information Systems*, **21**, 315-346. <https://doi.org/10.1145/944012.944013>
- [12] Yang, A.M., Lin, J.H., Zhou, Y.M., et al. (2012) Research on Building a Chinese Sentiment Lexicon Based on SO-PMI. *Applied Mechanics & Materials*, **263-266**, 1688-1693. <https://doi.org/10.4028/www.scientific.net/AMM.263-266.1688>
- [13] 李婷婷,姬东鸿.基于SVM和CRF多特征组合的微博情感分析[J].计算机应用研究,2015,32(4):978-981.
- [14] 李晓东.隐朴素贝叶斯在情感分类中的应用研究[D]:[硕士学位论文].衡阳:南华大学,2019.
- [15] 郑志伟,邱佳玲,阳庆玲,龚晓春,郭山清,贾忠伟,郝春.随机森林对文本情感分析的应用与R软件实现[J].现代预防医学,2018,45(8):1345-1348+1353.

- [16] Alqaryouti, O., Siyam, N., Monem, A.A. and Shaalan, K. (2019) Aspect-Based Sentiment Analysis Using Smart Government Review Data. *Applied Computing and Informatics*, 1-20. <https://doi.org/10.1016/j.aci.2019.11.003>
- [17] Soumya, S. and Pramod, K.V. (2020) Sentiment Analysis of Malayalam Tweets Using Machine Learning Techniques. *ICT Express*, **6**, 300-305.
- [18] Dey, L., Chakraborty, S., Biswas, A., Bose, B. and Tiwari, S. (2016) Sentiment Analysis of Review Datasets Using Naïve Bayes' and K-NN Classifier. *International Journal of Information Engineering and Electronic Business (IJIEEB)*, **8**, 54-62. <https://doi.org/10.5815/ijieeb.2016.04.07>
- [19] 金志刚, 胡博宏, 张瑞. 基于深度学习的多维特征微博情感分析[J]. 中南大学学报(自然科学版), 2018, 49(5): 1135-1140.
- [20] 王宏生, 金相宇. 基于深度学习的中文电商评论情感分析[J]. 信息通信, 2018(3): 51-53.
- [21] 何炎祥, 孙松涛, 牛菲菲, 李飞. 用于微博情感分析的一种情感语义增强的深度学习模型[J]. 计算机学报, 2017, 40(4): 773-790.
- [22] 吴鹏, 刘恒旺, 沈思. 基于深度学习和 OCC 情感规则的网络舆情情感识别研究[J]. 情报学报, 2017, 36(9): 972-980.
- [23] 石凤贵. 中文文本分词及其可视化技术研究[J]. 现代计算机, 2020(12): 131-138+148.
- [24] 洪巍, 李敏. 文本情感分析方法研究综述[J]. 计算机工程与科学, 2019, 41(4): 750-757.
- [25] Xu, J.P., Tang, W.Y., Zhang, Y. and Wang, F.J. (2020) A Dynamic Dissemination Model for Recurring Online Public Opinion. *Nonlinear Dynamics*, **99**, 1269-1293. <https://doi.org/10.1007/s11071-019-05353-3>
- [26] Čechura, L. and Šobrová, L. (2008) The Price Transmission in Pork Meat Agri-Food Chain. *Agricultural Economics*, **54**, 77-84. <https://doi.org/10.17221/272-AGRICECON>
- [27] 谭莹, 周建军, 何勤英. 我国猪肉价格波动的省际空间传导研究[J]. 价格理论与实践, 2017(5): 65-68.
- [28] 王刚毅, 王孝华, 李洪妹. 中国生猪价格空间溢出效应研究——基于同步系数矩阵的空间计量分析[J]. 农业现代化研究, 2018, 39(1): 105-112.
- [29] Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003) Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, **3**, 993-1022.
- [30] 杨俊闯, 赵超. K-Means 聚类算法研究综述[J]. 计算机工程与应用, 2019, 55(23): 7-14+63.