工程与管理科学
Engineering and Management Science
第 7 卷◆第 12 期◆版本 1.0◆2025 年
文章类型：论文|刊号（ISSN）：2705-0637(P) / 2705-0645(O)

基于 SARIMAX 和 Apriori 的菜品销量预测与关联性分析

何岩
Virginia Epsiscopal School Virginia Lynchburg 24503

DOI:10.12238/ems.v7i12.16415

[摘 要] 为了降本增效以让餐厅获得更多的收益，刻画客户的膳食特征并对菜品的销量预测来筛选合适的菜品对于餐饮企业有着重大的意义。基于某餐厅的销售订单数据，提出基于 SARIMAX（季节自回归移动移动平均模型）的特定菜品销量预测方法，建立了基于 Apriori 算法的不同菜品的关联性分析。通过使用某餐厅六个月的订单数据，训练模型并预测了其中三个最大众的菜：红烧肉，黄韭小炒和米饭的销量。最终通过分析预测销量和实际销量的均方根误差（RMSE）、均方误差（MSE）得出训练模型的预测度。模型预测红烧肉，黄韭小炒，米饭三个菜品的销量的均方根误差分别为 4.12，3.29，62.70；三个菜品的均方误差分别为 16.99，10.80，3930.7。同时，基于 Apriori 算法，对销售订单的菜品建立关联规则，得到了 20 个高度相关的菜品关联关系，便于商家制定合适的套餐吸引顾客购买。

[关键词] SARIMAX；时间序列；菜品预测；Apriori

1.研究背景

近年来，餐饮行业竞争日益激烈，食材成本、人力成本及租金等运营开支持续攀升，使得企业利润空间不断压缩。正如 Ganjar A.所说，人们对于食物的重视性日益增长[1]。在此背景下，餐饮企业需通过精细化管理优化运营效率，以提升市场竞争力。与此同时，随着信息技术的快速发展，餐饮行业的信息化系统（如前厅点餐系统、后厨管理系统、供应链管理平台等）已广泛普及，根据丁铛所说，这为大数据分析提供了坚实的数据基础[2]。这些系统能够实时记录菜品销量、顾客点餐偏好、库存消耗、采购成本等关键数据，为企业优化经营策略提供了丰富的信息支持。

数据分析技术的引入，使得餐饮企业能够从海量运营数据中挖掘有价值的规律，从而制定更科学的决策。例如，利用时间序列分析（如 ARIMA、SARIMAX 模型）。时间序列模型在短期数据预测中应用广泛，其预测方法主要可分为回归模型和机器学习方法两类[3]。回归模型中的自回归（AR）模型、移动平均（MA）模型以及自回归滑动平均（ARMA）模型均要求时间序列具有平稳性，即序列的均值与方差为常数，且任意两个时点之间的协方差与时间无关[4]。可以精准预测未来菜品销量，帮助餐厅优化食材采购计划，减少库存浪费；而关联规则挖掘（如 Apriori 算法）则能分析不同菜品之间的消费关联性，为套餐设计、促销策略提供依据[5]。此外，小波变换[6]等信号处理方法可用于分析销售数据的周期性波动，进一步优化餐厅的运营节奏。

餐饮业一直面临着如何去精准预测顾客需求以达到减少食材浪费同时保持顾客新鲜感的这么一个‘世纪难题’餐厅厨师固定的菜单已经难以满足现在顾客的高要求，所以我们可以将菜单内不同销量的各个菜品，例如‘明星菜品’或‘高销量菜品’结合顾客的口味偏向一起代入到 SARIMAX 中去计算，从而得到分析出某系菜销量高的原因，比如口味，性价比等。这些高销量就可以成为当季的明星进行重点宣传。此外通过分析和预测菜品销量，可以结合后厨系统，对食材采购进行优化[7]，减少不必的浪费。随着大数据和人工智能技术的快速发展，数据驱动的运营模式正在逐步取代经验驱动的管理模式，我们可以使用算法模型将菜单的优化分成三个模块（1）需求预测模块，预测不同季节气候的菜品需求；（2）套餐优化模块，探索菜品之间的关联性生成高关联度的餐品套餐；（3）成本优化模块，以利润最大损耗最小为优化目标节省食材成本开支，优化菜单结构。

本研究基于某餐厅六个月的真实订单数据，采用 SARIMAX（季节性自回归移动平均模型）对最受欢迎的菜品（如麻婆豆腐、红烧肉、番茄炒蛋）进行销量预测，并通过 RMSE、MSE 等指标验证模型的准确性。同时，利用 Apriori 算法挖掘不同菜品之间的关联规则，识别出 20 组高频共现菜品组合，为餐厅制定套餐策略提供数据支持[8]。这些分析方法不仅能够帮助餐厅降低采购成本、减少食材浪费，还能通过精准营销（如推荐套餐、动态定价）提升客单价，最终实现降本增效的目标。

综上所述，大数据分析技术的应用，使得餐饮企业能够从传统的经验驱动决策转向数据驱动决策，从而在激烈的市场竞争中占据优势。本研究旨在探索数据分析方法在餐饮管理中的实际应用，为企业提供可落地的技术方案，助力其实现精细化运营和可持续发展。

2.菜品销量预测

2.1 背景介绍

本研究使用的数据来自国内某连锁中式家常菜餐厅，数据记录时间为 2022 年 9 月至 2023 年 1 月，共计 6 个月的运营数据。该餐厅主要提供大众化的中式家常菜品，菜单包含 89 个菜品项目，涵盖热菜、凉菜、主食和饮料等类别（如表 1 所示），其中包括白斩鸡、素烧鹅等经典家常菜品。

表 1 菜品销量图

菜品编号	菜品名称	单位	均价	数量	金额	实际金额
000	米饭	份	2.00	648.00	1296.00	1296.00
203	白斩鸡	份	13.00	31.00	403.00	403.00
208	素烧鹅	份	8.00	8.00	64.00	64.00
210	韩式泡菜	份	6.00	10.00	45.00	45.00
			…			
			…			
			…			
929	芬达	听	4.00	1.00	4.00	4.00
930	红牛	瓶	7.00	4.00	28.00	28.00

文章类型：论文|刊号（ISSN）：2705-0637(P) / 2705-0645(O)

数据记录了餐厅每日所有菜品的销售情况，形成完整的日销售记录。需要特别说明的是，在数据采集期的最后一个月（次年 1 月），由于春节假期影响，餐厅有将近 10 天的歇业期，这段时间没有产生销售数据，所以我们剔除这段时间的空白数据。

该餐厅作为连锁经营的中式家常菜馆，其菜品结构和销售模式在同类餐饮企业中具有典型性和代表性。完整的销售记录为后续的菜品销量预测和关联分析提供了可靠的数据基础，较为庞大且完善的数据为后面的数据分析提供优质的基础。

2.2 数据分析

为了使模型更加贴合现实中的餐厅以达到最具有参考价值的值，本研究使用餐厅 9 月份到 1 月份的销售额，具体研究并分析了三道菜品，分别为红烧肉，黄韭小炒以及米饭。

首先分析缺失值。数据的缺失主要包括缺失和记录中某段信息的缺失，而两者都会导致数据挖掘建模分析的结果不准确。对数据进行分析发现除了已知的在 1 月份由于春节假期，餐厅没有营业导致没有营业额，在 8 月 11 日有一处不明的数据缺失。为了防止缺失值对数据挖掘模型造成不良的影响，此处使用时间序列差值法对缺失值进行处理。此外，为了使数据更加稳定平滑以减少数据方面对数据挖掘模型精准度的影响，在此对三道菜的数据做平稳化处理。处理结果如图 1，2，3（分别为红烧肉，黄韭小炒以及米饭）

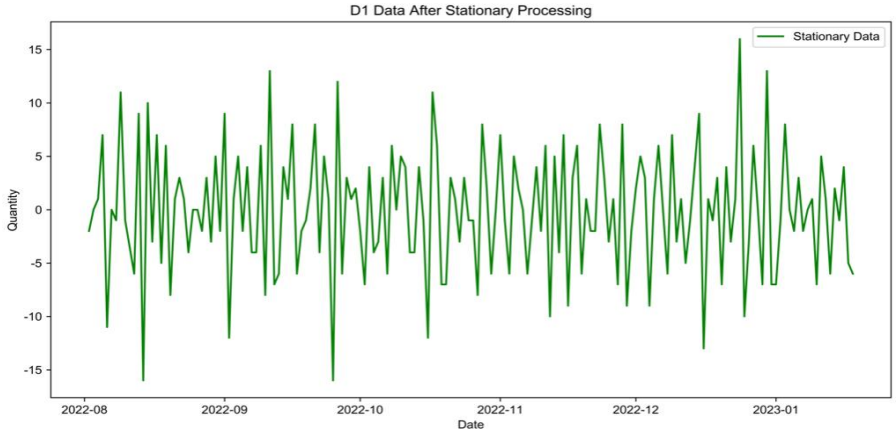


图 1　红烧肉的稳定处理后的数据

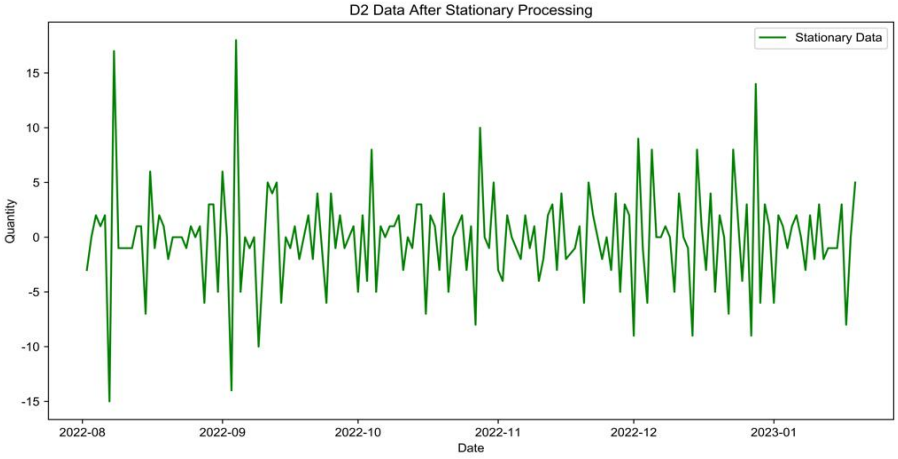


图 2　黄韭小炒的稳定处理后的数据

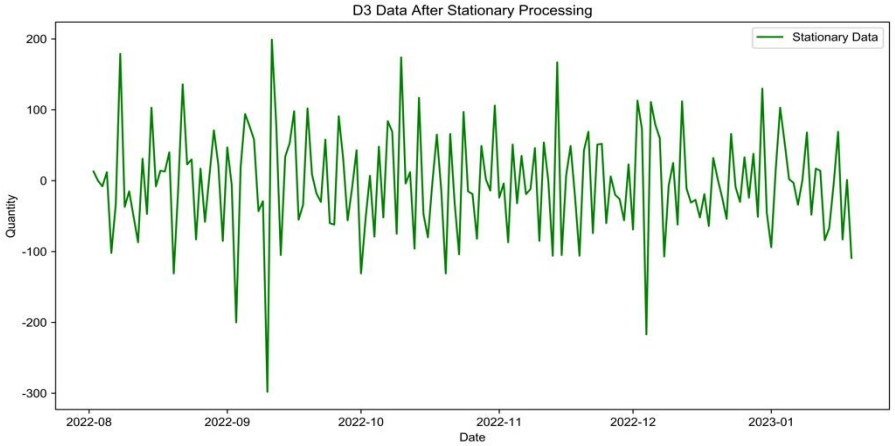
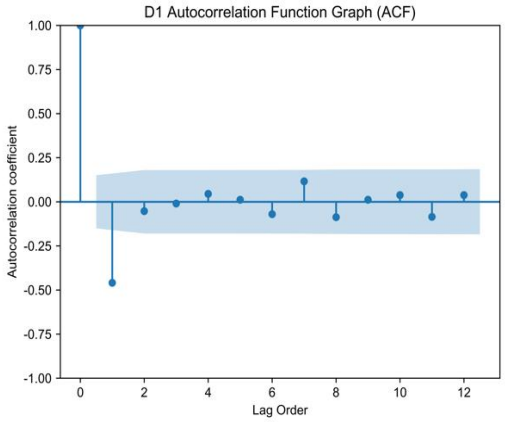
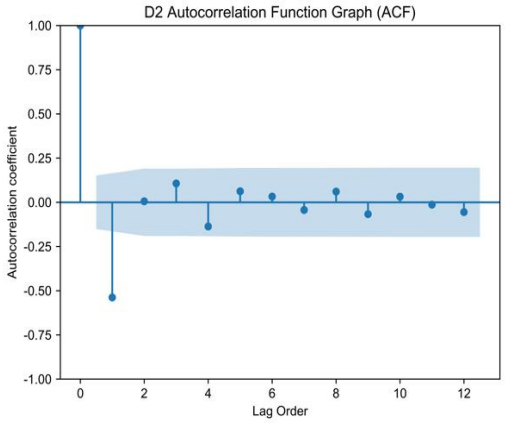


图 3　米饭的稳定处理后的数据

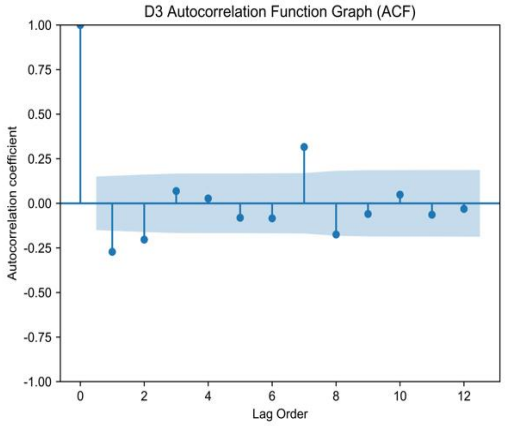
处理完缺失值并平稳化处理之后的数据也并不能直接用于训练模型，在训练之前应先对其进行白噪声检验，以排查数据是否为白噪声。如果数据被判定为白噪声，则说明时间序列数据具有纯随机性，意味着它没有可预测的模式或趋势，完全由随机波动组成因此不适合进行数据淘金。白噪声检验结果显示，三种菜品的 Ljung-Box 检验 p 值均小于 0.05，拒绝白噪声假设，说明数据具有可预测性。拒绝白噪声假设之后，使用 ACF（自相关函数图）和 PACF（偏自相关函数图）分析数据的自相关结构并且可以帮助选择合适的 ARIMA 模型参数。如图可视。



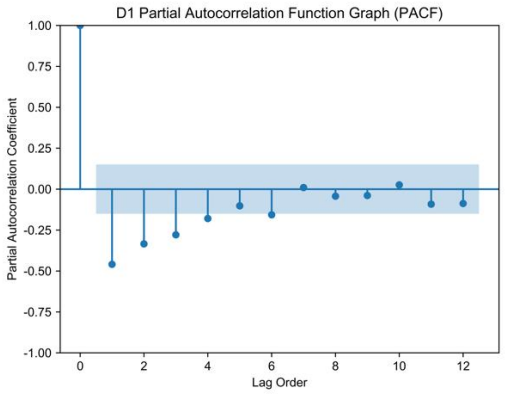
图表 4　红烧肉的自相关图



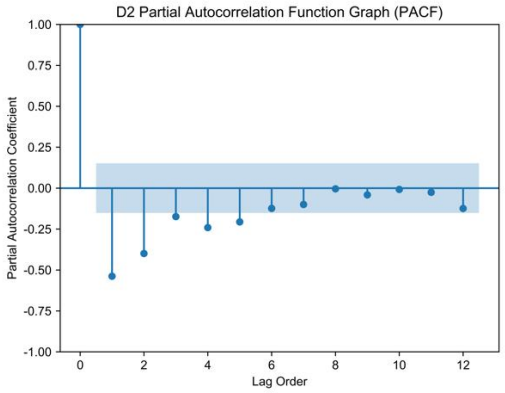
图表 5　黄韭小炒的自相关图



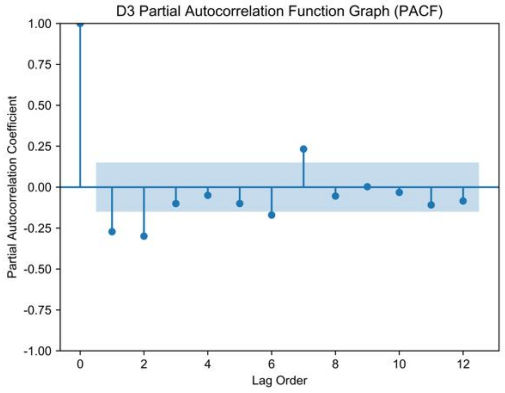
图表 6　米饭的自相关图



图表 7　红烧肉偏自相关图



图表 8　黄韭小炒偏自相关图



图表 9　米饭偏自相关图

ACF 图展示一个时间序列与其过去各滞后期之间的整体相关性，而 PACF 图则显示剔除中间滞后影响后，当前值与某一滞后期之间的“净相关性”。通过这两个图，我们可以判断序列是否平稳，并据此确定 ARIMA 模型中 AR（自回归）和 MA（移动平均）部分的阶数，从而为时间序列建模提供直观依据。自相关图（ACF）和偏自相关图（PACF）是分析时间序列数据的常用工具，用于识数据中的规律和结构，辅助建模和预测。

自相关图（ACF）的作用：

1. 判断序列自身在不同时间间隔下的相关性，高自相关值表示滞后值与当前值关系密切。

2. 帮助确定 ARIMA 模型中移动平均(MA)部分的阶数(Q)，若自相关系数快速降至零，通常说明 MA 阶数较低。

3.检验序列是否随机：若所有自相关系数接近零，序列可能为随机波动，不易建模。

4.识别季节性：在季节性滞后（如 12 或 4）处出现高峰，表明存在季节性模式。

偏自相关图（PACF）的作用：

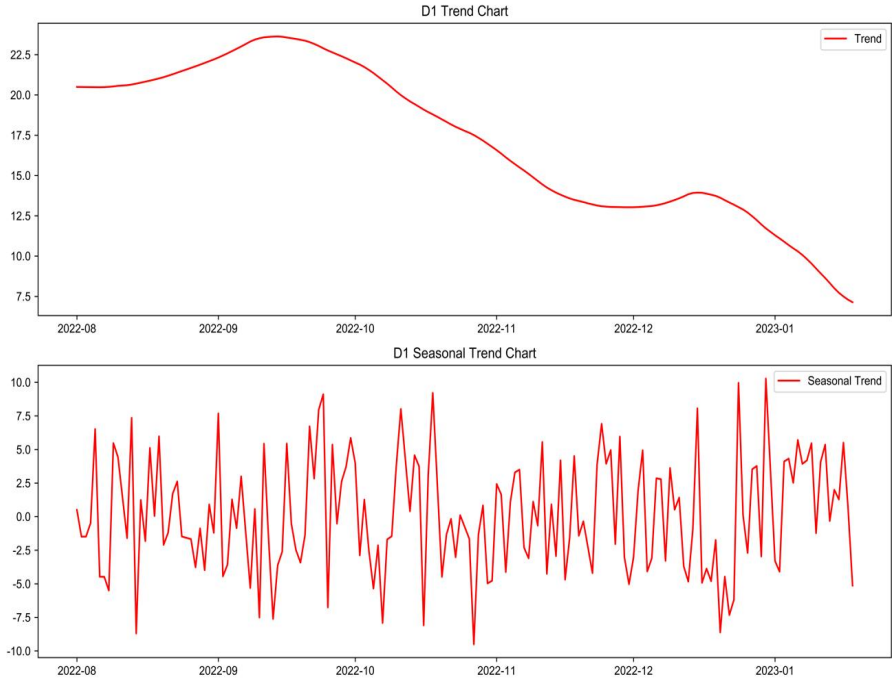
1.分析两个时间点之间的直接相关性，排除其他滞后项的影响。

2.帮助确定 ARIMA 模型中自回归（AR）部分的阶数（P），若偏自相关系数突然降为零，则 AR 阶数可据此确定。

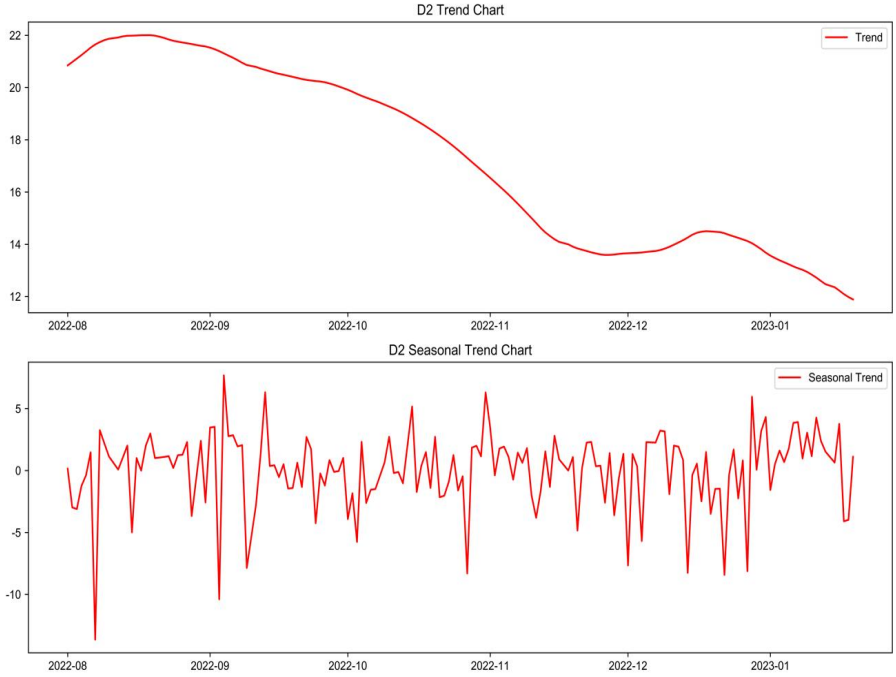
3.检验模型效果：若建模后 PACF 仍有明显相关性，说明模型可能遗漏重要信息。

4.辅助识别非平稳性，提示是否需要进行差分等处理使序列平稳。

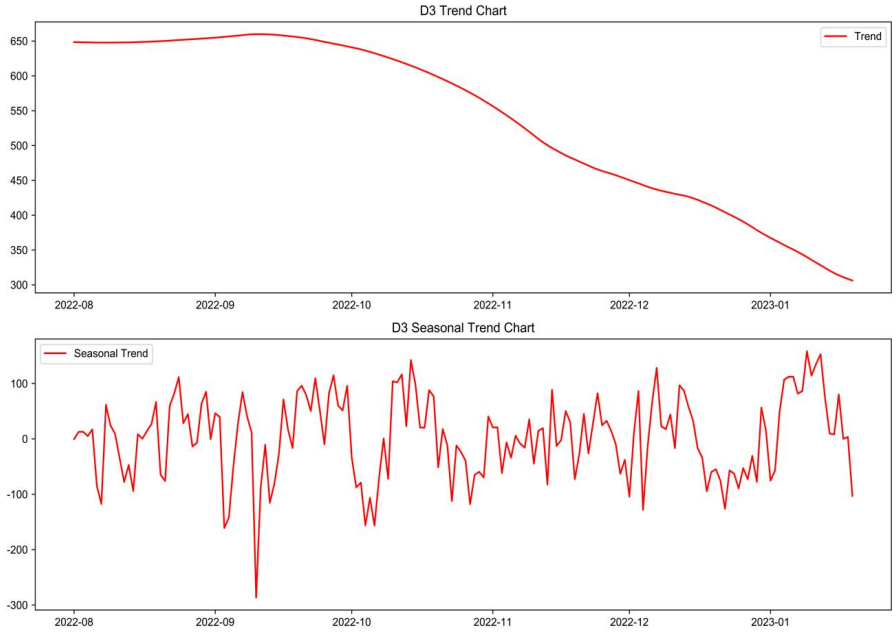
ACF 和 PACF 图看出三道菜品的时间序列可能具有季节性（如 ACF/PACF 图中 lag=7 处出现异常）。于是引用小波变换（公式如下）提取时间序列的趋势并且去噪。发现这三个时间序列缺失含有季节性（详情请看图 10，11，12），于是使用 SARIMAX（季节性差分自回归滑动平均外生模型）进行模型训练。



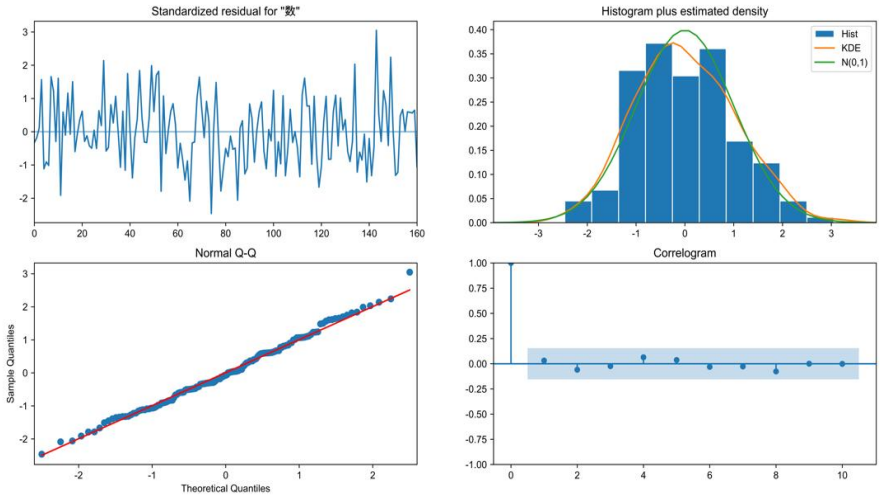
图表 10　红烧肉的趋势和季节性趋势图



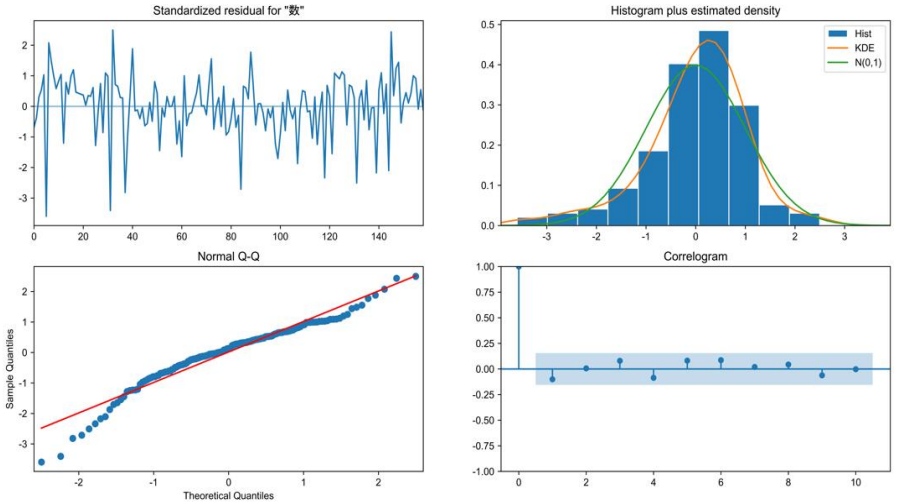
图表 11　黄韭小炒的趋势和季节性趋势图



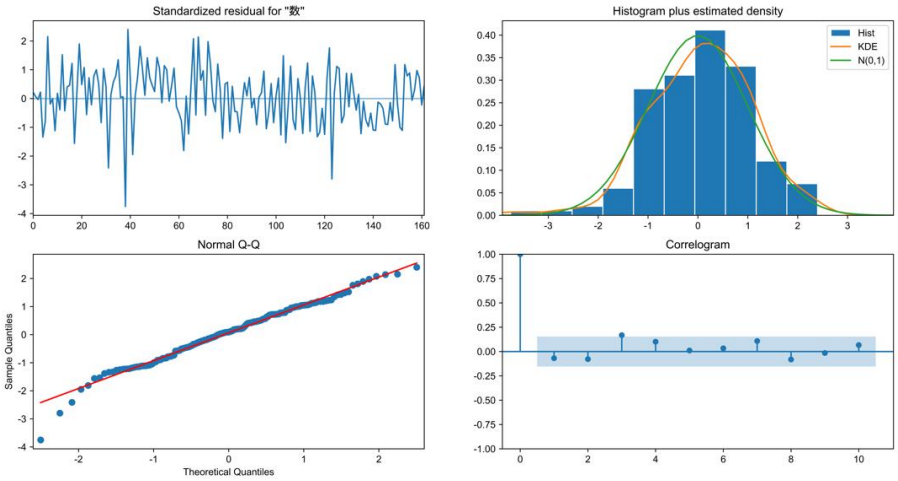
图表 12　米饭的趋势和季节性趋势图



图表 14　红烧肉 SARIMAX 表



图表 15　黄韭小炒 SARIMAX 表



图表 16　米饭 SARIMAX 表

2.3 模型构建

SARIMA（季节性自回归积分移动平均模型）是在 ARIMA 模型基础上加入了季节性成分的扩展模型，适用于具有明显周期性波动的时间序列数据，如月度销售额或气象数据。其模型表示为 SARIMA（p，d，q）(P，D，Q，s)，其中（P，D，Q）分别表示季节性部分的自回归、差分和移动平均的阶数，s 代表季节周期（例如 12 个月）。与 ARIMA 相比，SARIMA 能够同时捕捉数据中的非季节变化和季节模式，因此在处理季节性数据时通常拟合效果更好，预测也更准确。为找到最优模型，通常采用 BIC（贝叶斯信息准则）等指标寻找总阶数（包括季节与非季节部分）最小的 SARIMA 模型，以在预测精度和模型简洁性之间取得平衡。

采用季节性自回归移动平均模型（SARIMAX），其一般形式为：设置参数优化范围如下：p，q∈[0，2]，P，Q∈[0，1]，d=1，D=1（由 ADF 检验确定），遍历最佳 AIC 值。采用网格搜索法遍历 216 种参数组合，以 AIC 准则作为评价标准：AIC = 2k - 2ln（L），其中 k 为参数个数，L 为似然函数值。最终米饭的最优参数为 SARIMAX（1，1，1）（1，0，1）7，AIC= 1894.27；黄韭小炒为（0，1，1），AIC= 833.674；红烧肉为（0，1，1）（1，0，1）7，AIC=942.701。将数据按 80：20 比例划分为训练集和测试集。

2.4 结果分析

菜品名称	RMSE	MAE	MSE	MAPE（%）	最佳模型参数
红烧肉	4.12	3.27	16.99	79.00	（0，1，1）x（1，0，1，7）
黄韭小炒	3.29	2.38	10.80	24.56	（0，1，1）
米饭	62.70	54.09	3930.7	17.87	（1，1，1）x（1，0，1，7）

从预测结果可以看出来在绝对误差方面，米饭的 RMSE 值（62.70）显著高于其他两种菜品，说明米饭销量的绝对波动较大，但模型在相对预测精度上仍具备较强适用性。当转换为相对误差 MAPE 时，米饭的表现（17.87%）反而优于红烧肉（79.01%）。并且黄韭小炒表现出最佳的预测稳定性，其 RMSE（3.29）和 MAE（2.38）均处于较低水平，MAPE（24.56%）也相对合理，说明该菜品的销售模式最具有规律性。

3. 菜品关联分析

3.1 背景介绍

现代菜品种类繁多，顾客往往会因此而变得疲于选择，且顾客并不会因为选择丰富购买更多的商品。繁杂的选购过程往往会给顾客带来疲惫的用餐体验。对于某些菜品来说顾客可能会选择同时购买，比如已经点了较为油腻的肉的话，顾客往往会选择清淡的蔬菜来中和口感。如果说餐厅能够抓出菜品之前微妙的关联的话，则可以在点单系统上做出很多改进，使得顾客的点单体验和效率得到极大的提升，如把几道相关度高的菜作为套餐售卖以及在选择某样菜品之后自动推荐出其他相关度高的菜品以供顾客选择。因此，为了使利益最大化，找到各个菜品之间的关联对于商户来说极为重要。顾客在某餐馆消费的每一单数据，可以通过算法找到其中的规律。关联度高的菜品可以绑定从而推出新套餐，既可以帮助疲于选择的顾客做出最佳选择，又可以热销菜品带动其他菜品的销售额从而提高餐厅总体的销量，如果在绑定套餐上做出优惠调整，也会让顾客比起单点更倾向于去选择套餐。

3.2 数据探索与预处理

数据样本包括 2024 年 8 月 1 号，8 月 15 号及 8 月 30 号全天的所有订单。通过对所有订单进行检索，发现米饭出现的次数过高，考虑到可能会掩盖其他菜品之间的潜在关联于是把米饭从数据中剔除。此外考虑到数据都是由各类菜名组成的，不存在量级之间的差距，并且检索数据之后未发现缺失值，所以不对数据做其他处理。数据的精准采集在本节起到关键作用。数据探索与预处理是模型成功的基石俗话说‘垃圾进，垃圾出’原始数据往往是混乱的，不完整且不一致的，直接用于建模往往会导致最后呈现混乱的状况。

（1）首先需要整合多项数据，去收集店内的菜品的历史销量，买菜的成本，顾客的评价，以及各个时间段的订单销量。

（2）数据的探索，在预处理之前需要了解数据的全貌和特点，做到数据可视化和重点的摘要，对数值进行描述性统计：均值、四分位数、最大值、最小值标准差、计数，从而做到快速发现数据的异常值、数值范围。

（3）缺失值的分析，将数据的缺失部分做成条形图记录缺失比例。

（4）异常值的检测，将由于系统错误产生的负销量去除。

（5）时间序列的分析，各个时间段、气候的分段数据统计，计算数据的周期性，完善每个部分数据的完整性。

数据探索与预处理是一个相当繁琐的过程，但是它确确实实的能直接决定一个数据模型的上限，一个数据丰富全面的数据库，既是使用最基础的数字模型去计算也会获得相当不错的效果。

3.3 模型构建

Apriori 数据挖掘总体流程如下：（1）收集数据。（2）数据探索和预处理。通过查看数据特征，缺失值处理，异常

值处理等方式对数据进行预处理，以让模型分析更加准确。（3）分析和建模。通过 Apriori 进行关联规则分析并且创建模型。（4）结果反馈。模型由三个基本部分构成：输入、算法处理和输出。输入包括建模样本数据及参数设置（最小支持度和最小置信度）。算法处理部分使用 Apriori 关联规则算法对数据进行分析。输出则是算法处理后的关联规则结果。具体实现过程如下：首先设定最小支持度和最小置信度参数，并输入样本数据；随后运行 Apriori 算法进行分析。若没有规则满足所设参数条件，则需调整参数重新计算；否则，直接输出最终生成的关联规则。Apriori 核心算法指标如下：

（1）支持度（support）：项集 A、B 同时发生的概率称为关联规则的支持度（也称相对支持度）

$$Support(A \Rightarrow B) = P(A \cup B)$$

（2）置信度（confidence）：项集 A 发生，则项集 B 发生的概率为关联规则的置信度

$$Confidence(A \Rightarrow B) = P(B|A)$$

（3）最小支持度和最小置信度

最小支持度是用户或专家定义的衡量支持度的一个阈值，表示项目集在统计意义上的最重要性；最小置信度是用户或专家定义的衡量置信度的一个阈值，表示关联规则的最低可靠性。同时满足最小支持度阈值和最小置信度阈值的规则称作强规则。

3.4 结果反馈

根据 Apriori 实现的关联规则中大多包含米饭，参考价值不大，故而剔除关于米饭的关联关系。在所有生成的关联规则中选择了关联性最强的前十组，对其的分析如下：

1.{'酸辣土豆丝'}=>{'水波蛋'} 支持度约为 1.28%,置信度约为 17.35%。说明同时选择酸辣土豆丝和水波蛋这两种菜品的概率达 17.35%，而这种情况发生的可能性约为 1.28%。

2.{'红烧大排'}=>{'水波蛋'} 支持度约为 1.13%,置信度约为 15.00%。说明同时选择红烧大排和水波蛋这两种菜品的概率达 15.00%，而这种情况发生的可能性约为 1.13%。

3.{'番茄炒蛋'}=>{'白斩鸡'} 支持度约为 1.13%,置信度约为 14.56%。说明同时选择番茄炒蛋和白斩鸡这两种菜品的概率达 14.56%，而这种情况发生的可能性约为 1.13%。

4.{'酸辣土豆丝'}=>{'番茄炒蛋'} 支持度约为 1.05%,置信度约为 14.29%。说明同时选择酸辣土豆丝和番茄炒蛋这两种菜品的概率达 14.29%，而这种情况发生的可能性约为 1.05%。

5.{'肉饼蒸蛋'}=>{'青韭豆芽'} 支持度约为 1.05%,置信度约为 15.22%。说明同时选择肉饼蒸蛋和青韭豆芽这两种菜品的概率达 15.22%，而这种情况发生的可能性约为 1.05%。

6.{'肉饼蒸鲞'}=>{'水波蛋'} 支持度约为 0.90%,置信度约为 19.35%。说明同时选择肉饼蒸鲞和水波蛋这两种菜品的概率达 19.35%，而这种情况发生的可能性约为 0.90%。

7.{'酸菜鱼'}=>{'水波蛋'} 支持度约为 0.90%,置信度约为 15.79%。说明同时选择酸菜鱼和水波蛋这两种菜品的概率达 15.79%，而这种情况发生的可能性约为 0.90%。

8.{'豆腐皮肉包子'}=>{'番茄炒蛋'} 支持度约为 0.90%，置信度约为 15.00%。说明同时选择豆腐皮肉包子和番茄炒蛋这两种菜品的概率达 15.00%,而这种情况发生的可能性约为 0.90%。

9.{'红烧大排'}=>{'快茭菜小炒'} 支持度约为 0.83%，置信度约为 11.00%。说明同时选择红烧大排和快茭菜小炒这两种菜品的概率达 11.00%，而这种情况发生的可能性约为 0.83%。

10.{'水波蛋'}=>{'白斩鸡'} 支持度约为 0.83%，置信度约为 8.03%。说明同时选择水波蛋和白斩鸡这两种菜品的概率达 8.03%，而这种情况发生的可能性约为 0.83%。

其中置信度最高的是 规则 6（肉饼蒸鲞 => 水波蛋，19.35%),说明顾客在选择肉饼蒸鲞时,很可能也会点水波蛋。而最有代表性的依然是 酸辣土豆丝与水波蛋的组合（规则 1，17.35%），支持度和置信度都比较突出。在这十条规则中，水波蛋与番茄炒蛋之所以与其他菜品呈现出较高的关联性，主要源于它们在饮食习惯、口味接受度和营养结构上的普适性。水波蛋口感清淡能够与酸辣、荤菜或清淡类菜品搭配，并且能通过优质蛋白质弥补营养不足，再加上价格低廉、常被作为附加选择。而番茄炒蛋则因其酸甜适口的口味以及营养均衡的特性，几乎人人接受，往往成为点餐时的“必点菜”。这两道菜分别扮演着“百搭配角”和“大众主角”的角色，因此在不同的组合中频繁出现，形成了与多种菜品的高度关联性。

4.结论

通过整合 Apriori 关联规则算法和 SARIMAX 时间序列模型，构建了一个双维度的餐饮数据分析框架，为餐饮企业的精细化运营提供了可靠的方法论支持。可以有效的帮助餐饮行业预测菜品销量，降低采购成本，优化产品套餐及菜单，用大数据和算法能极大程度的帮助餐厅灵活应对市场变化，持续性的经营增长并且对决策提供了使用方法论。

尽管数据模型在实施过程中面临数据质量‘技术能力’成本和组织文化等一系列复杂且未知的挑战，但随着科技的发展和行业的经验积累，这些困难都会随之解决，未来，数据模型将更加智能化，个性化和可持续化，为餐饮行业带来更多建议和机遇。

[参考文献]

[1]Alfian，Ganjar，Jongtae Rhee，Hyejung Ahn，Jaeho Lee，Umar Farooq，Muhammad Fazal Ijaz，and M.Alex Syaekhoni.“Integration of RFID，Wireless Sensor Networks，and Data Mining in an E-Pedigree Food Traceability System.”Journal of Food Engineering 212（November 2017）：65－75.

[2]丁铛，张志飞，苗夺谦，等.基于消费者行为的点餐推荐算法[J].计算机科学，2017，44（S2）：46-50.

[3]任守纲，张景旭，顾兴健，等.时间序列特征提取方法研究综述[J].小型微型计算机系统,2021,42(2):271-278.

[4]辛曼玉.基于 arima-rbf 神经网络的沿海港口吞吐量预测研究[J].武汉理工大学学报（交通科学与工程版），2014，38（1）：241-244.

[5]吴斌，肖刚，陆佳炜.基于关联规则挖掘领域的 Apriori 算法的优化研究[J].计算机工程与科学，2009，31（6）：116-118.

[6]叶昭呈，李梦诗，季天瑶.基于连续小波变换和深度学习的不可靠观测系统的供热管网渗漏识别[J/OL].工业安全与环保，1-8[2025-07-17].

[7]朱廷杰，王鹏举，孙卫强.基于群体决策特征的中式自选餐厅菜品销量预测模型[J].计算机应用研究，2022，39（06）：1731-1736.DOI：10.19734/j.issn.1001-3695.2021.11.0622.

[8]钱雪忠,孔芳.关联规则挖掘中对 Apriori 算法的研究[J].计算机工程与应用，2008（17）：138-140.