2021 年高教社杯全国大学生数学建模竞赛

承　诺　书

我们仔细阅读了《全国大学生数学建模竞赛章程》和《全国大学生数学建模竞赛参赛规则》（以下简称“竞赛章程和参赛规则”，可从 http://www.mcm.edu.cn 下载）。

我们完全清楚，在竞赛开始后参赛队员不能以任何方式，包括电话、电子邮件、“贴吧”、QQ 群、微信群等，与队外的任何人（包括指导教师）交流、讨论与赛题有关的问题；无论主动参与讨论还是被动接收讨论信息都是严重违反竞赛纪律的行为。

我们完全清楚，在竞赛中必须合法合规地使用文献资料和软件工具，不能有任何侵犯知识产权的行为。否则我们将失去评奖资格，并可能受到严肃处理。

我们以中国大学生名誉和诚信郑重承诺，严格遵守竞赛章程和参赛规则，以保证竞赛的公正、公平性。如有违反竞赛章程和参赛规则的行为，我们将受到严肃处理。

我们授权全国大学生数学建模竞赛组委会，可将我们的论文以任何形式进行公开展示（包括进行网上公示，在书籍、期刊和其他媒体进行正式或非正式发表等）。

我们参赛选择的题号（从 A/B/C/D/E 中选择一项填写）：B

我们的报名参赛队号（12 位数字全国统一编号）：B22460

参赛学校（完整的学校全称，不含院系名）：西安电子科技大学

参赛队员 (打印并签名) ：1. 张凤

2. 夏雨烟

3. 唐祥

指导教师或指导教师组负责人 (打印并签名)：

（指导教师签名意味着对参赛队的行为和论文的真实性负责）

日期：2022 年 5 月 3 日

（请勿改动此页内容和格式。此承诺书打印签名后作为纸质论文的封面，注意电子版论文中不得出现此页。以上内容请仔细核对，如填写错误，论文可能被取消评奖资格。）

赛区评阅编号：
（由赛区填写）

全国评阅编号：
（全国组委会填写）

2021 年高教社杯全国大学生数学建模竞赛

编 号 专 用 页

赛区评阅记录（可供赛区评阅时使用）：

评阅人						
备注						

送全国评阅统一编号:
（赛区组委会填写）

（请勿改动此页内容和格式。此编号专用页仅供赛区和全国评阅使用，参赛队打印后装订到纸质论文的第二页上。注意电子版论文中不得出现此页。）

基于人工智能的卷烟销售数据分析与预测研究

摘要

本文针对一家烟草公司于 2019-2021 年的卷烟销售数据进行研究分析并进行了对未来一年销量的预测。

对于问题一：我们从不同销售环境对品牌的销售状况这一问题入手，建立**多元线性回归模型**来刻画客户与品牌之间的关系。统计分析了十个客户在 2021 年于不同销售环境下的四种具有代表性香烟的销量，利用最小二乘法与 SPSS **得到销量的 47.1%是由于市场类型以及品牌决定的**。

由于市场类型的显著性过高达到 72.4%,说明市场的差异对销量的影响不是那么大。考虑到品牌的销量可能会受到不同销售地区的影响，于是对模型进行了改进，改进后的模型求得系数最大的显著性水平降至 50.1%，这说明了**销量与卷烟品牌和市场类型之间的相关性**。

对于问题二：我们首先统计了 2019-2021 年城市与乡村不同终端卷烟销量并得出其变化规律。随后，延用问题一中的多元线性回归模型分析造成此变化规律的主要因素，得到**卷烟的销量与销售的时间，销售终端类型以及销售市场类型有很强的相关性，且普通终端的销量在绝大部分情况下会比现代终端占优势的结论**。

为预测未来 1 年该分公司每个月的销量趋势，使用人工智能方法，建立了 **BP 神经网络模型**，利用 2019 年与 2020 年的销量对模型进行训练，后预测 2021 年的销量并与真实销量进行比对，发现预测效果较好但仍不准确。

为达到更好的预测效果，我们对模型进行了改进，决定采用 **RBF 神经网络**。通过训练得到更为准确的乡村城镇及不同终端的销量预测以及销量总和预测。预测结果的**相对误差限平均为 9%,相对误差上限为 58.98%**，与 BP 神经网络模型的相对误差相比，误差的范围缩小，精确度提高。随后，利用该模型对客户 0811170 进行 1 年的未来情况的预测。

对于问题三：我们分别对四个价位段卷烟 2019-2021 年销量进行了统计，分析了其主要特征及之间的数量关系。考虑到问题二中的 RBF 神经网络预测效果较好，延用此模型对四个价位段销量前三的卷烟品牌进行未来一年的销量预测。

对于问题四：根据上述问题中的模型与结果，针对性的对就卷烟的月销量变化趋势给出建议：**可以把更多精力放在农村普通终端的建设营销上，城市中可以以普通终端为主，现代终端为辅且 1 月份在近几年来始终是销量的最高潮，公司可以抓住该时机进行宣传。**

最后对本文所建立的模型进行了讨论和分析，综合评价模型。

关键词：多元回归线性模型　BP 神经网络　RBF 神经网络

一、 问题重述

1.1 问题背景

烟草行业在国民经济体系中占有重要地位，它的存在和发展对于满足人们正常消费需求和增加国家财政收入都具有重要的意义和作用。烟草行业有别于其他行业的市场竞争，由国家烟草专卖局和地方政府对卷烟的产量和价格起到双重监管和控制。近几年来，各烟草公司围绕客户、品牌、销售战略开展了多项工作，深入推进政府主导式改革，力图提高卷烟质量、增加销量、创造上档次品牌等。然而目前依然存在很多问题，比如卷烟产品之间的差异性不够明显、产品特色不够突出、产品结构有待进一步优化、卷烟品牌较多但重点品牌规模偏小、优势品牌竞争力和影响力不够等多种问题。

1.2 问题提出

现有一家烟草公司，收集了 2019-2021 年的部分客户相关销售数据（见附件 1）某一分公司客户的相关销售数据（附件 2）以及总公司的多指标销售情况（附件 3），其中客户、品牌、销售三种数据的结构与解释见数据结构类型表（附 4）。请你们的团队基于这些附件数据进行分析和建模，完成以下任务：

（1）请根据附件 1 和附件 2 中的数据，可选择不同角度进行分析，建立适当的数学模型来刻画客户与品牌之间的关系。诸如但不仅仅限于如下问题：不同经营业态的客户对哪种商品名称的销量最大？不同销售环境对品牌的销售状况如何？现代终端销售最好的前 10 种品牌分别是什么？经营业态、市场类型、终端类型三种客户中哪些销量数据波动最大和最小？数据平稳的客户群体有何共同特征？

（2）使用人工智能方法，建立相应的数学模型，讨论当前分公司的月销量变化规律，分析影响销量波动的主要因素，并预测未来 1 年该分公司每个月的销量趋势。应用你们的模型，请分析某一特定客户的销售规律以及对其进行 1 年的未来情况的预测。（1 箱=250 条=2500 盒）

（3）卷烟在价位上大体可分为四个段位：低端 100-150 元/条、中端 200-250 元/条、中高端 250-500 元/条、高端 500-1000 元/条，这四个价位段的销量占比有何特点？其中，真龙（起源）、硬盒芙蓉王、84 软盒玉溪是价位在 200-250 元/条的三种主要销售商品，在销量上各有怎样的特征？三者有什么样的数量关系？是否存在竞争和替代？未来一年的趋势是否会发生变化？其他价位段的主销商品是哪三种？在销量上有什么关系？未来一年内的销售占比有何变化？

（4）请根据你们的模型与结果，给该公司的领导们写一封总结信，就卷烟的月销量变化趋势给出你们的见解。

二、 问题分析

2.1 问题一的分析

问题一要求我们刻画品牌与客户之间的关系。品牌的种类、客户的类型以及两者之间的影响因素太多，本文着重关注客户所处不同销售环境对品牌的销售状

况的影响。本文假设市场销售环境的不同对销量的影响是一个常数的变化；卷烟的不同品牌亦是如此，即不同品牌之间烟的销量差异是一个常数。结合线性回归方程求解，以此得到卷烟品牌和市场类型之间的相关性。

2.3 问题二的分析

问题二要求我们讨论分公司的月销量变化规律及影响因素并预测未来 1 年该分公司每个月的销量趋势。本文统计了该分公司三年内的月销量并作为得到销量变化趋势，且根据销量产出的不同环境进行因素分析。并以相关因素和特定用户往年销量作为依据对相关模型进行训练，最后得到预测数据。

2.2 问题三的分析

问题三要求我们讨论卷烟的四个价位段的销售特征及未来一年的预测结果。本题首先要把百种卷烟品牌根据价位进行分类并统计销量分析其特征，随后还要统计各价位段的销量前三的卷烟品牌及其销量并分析特征。得到数据之后，对相关的模型进行训练并得到预测结果。

2.4 问题四的分析

问题四要求我们模型与结果，就卷烟的月销量变化趋势写一封总结信。在统计大量数据与趋势预测以及模型的结果上，针对性的对整个市场环境、销售环境以及卷烟品牌的选取上进行总结并给出见解。

三、 模型的假设

3.1 假设市场环境以及卷烟的品牌对销量的影响是线性的

自变量和因变量之间存在多元线性关系，因变量 y 能够被 x_1，$x_2 \ldots x_k$ 完全地线性解释。若为非线性，那么曲线斜率是一直变化的，没有意义，预测出的因变量会非常不准确。

3.2 假设市场环境与卷烟的品牌不存在线性关系

若市场环境与卷烟品牌之间存在线性关系，回归系数将会无法求解。

3.3 假设月份、市场类型以及终端类型对销量的影响是线性的

自变量和因变量之间存在多元线性关系，因变量 y 能够被 x_1，$x_2 \ldots x_k$ 完全地线性解释。若为非线性，那么曲线斜率是一直变化的，没有意义，预测出的因变量会非常不准确。

3.4 假设月份、市场类型以及终端类型不存在线性关系

若月份、市场类型以及终端类型之间存在线性关系，回归系数将会无法求解。

3.5 市场类型以及终端类型为 BP 神经网络输入层的数据输入端进行训练

烟卷的销量受时间、市场类型，终端类型这三类因素的影响比较大。

3.6 卷烟的销量只受时间、市场类型和终端类型三类因素的影响

四、 符号说明

符号	含义
y	销量
x_1	市场环境
x_2	卷烟品牌
x_3	月份
x_4	终端类型
a_0	常数项
a_1、a_2、a_3	回归系数
R^2	可决系数
Y	期望输出

五、 模型的建立与求解

5.1 问题一的模型建立与求解

5.1.1 模型建立

由上述的假定可知市场环境以及卷烟的品牌对销量的影响是线性的，即市场销售环境的不同对销量的影响是一个常数的变化；卷烟的不同品牌亦是如此，即不同品牌之间烟的销量差异是一个常数。其中烟的品牌，销售市场环境这两因素之间是没有相互作用，由此可以建立二元线性回归模型：

$$y = a_0 + a_1 x_2 + a_2 x_2 \tag{1}$$

式(1)在市场类型x_1以及卷烟品牌x_2（我们选择了各个价位段最具代表性的卷烟品牌）中由如下定义：

$$x_1\begin{cases}1，代表乡村\\2，代表城市\end{cases}，\quad x_2\begin{cases}1，代表红塔山\\3，代表玉溪\\5，代表芙蓉王\\7，代表南京\end{cases}$$

5.1.2 模型求解

要估计二元回归模型中的a_0、a_1、a_2，最常用的方法就是最小二乘法。设根据给定的一组样本数据Y_0、Y_1、Y_2，采用最小二乘法估计得到的样本回归模型为

$$y=\hat{a}_0+\hat{a}_1x_1+\hat{a}_2x_2 \tag{2}$$

并记为模型 1。则回归系数估计量$\hat{a}_0$、$\hat{a}_1$、$\hat{a}_2$应该使残差平方和

$$\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2=\sum_{i=1}^{n}(Y_i-\hat{a}_0-\hat{a}_1x_1-\hat{a}_2x_2)^2 \tag{3}$$

达到最小。

根据极值存在的必要条件，应该有

$$\begin{cases}-2\sum(Y_i-\hat{a}_0-\hat{a}_1x_1-\hat{a}_2x_2)=0\\-2\sum(Y_i-\hat{a}_0-\hat{a}_1x_1-\hat{a}_2x_2)x_1=0\\-2\sum(Y_i-\hat{a}_0-\hat{a}_1x_1-\hat{a}_2x_2)x_2=0\end{cases} \tag{4}$$

而根据我们的假设，x_1、x_2之间不存在线性关系，那么由式(4)正规方程组可以解出$\hat{a}_0$、$\hat{a}_1$、$\hat{a}_2$

$$\begin{cases}\hat{a}_0=\overline{Y}-\hat{a}_1\overline{X_1}-\hat{a}_2\overline{X_2}\\[2ex]\hat{a}_1=\dfrac{(\sum y_ix_{1i})(\sum x_{2i}^{\ 2})-(\sum y_ix_{2i})(\sum x_{1i}x_{2i})}{(\sum x_{1i}^{\ 2})(\sum x_{2i}^{\ 2})-(\sum x_{1i}x_{2i})^2}\\[2ex]\hat{a}_2=\dfrac{(\sum y_ix_{2i})(\sum x_{1i}^{\ 2})-(\sum y_ix_{1i})(\sum x_{1i}x_{2i})}{(\sum x_{1i}^{\ 2})(\sum x_{2i}^{\ 2})-(\sum x_{1i}x_{2i})^2}\end{cases} \tag{5}$$

其中，$x_i=X_i-\overline{X}$、$y_i=Y_i-\overline{Y}$、$\overline{X}=\dfrac{1}{n}\sum x_i$、$\overline{Y}=\dfrac{1}{n}\sum y_i$。

实际操作中，我们统计了十个客户在 2021 年于不同销售环境下的四种具有代表性香烟的销量，如下图：

客户	类型	红塔山	南京	玉溪	芙蓉王
0812184	城市	147	92	303	239
0815895	城市	144	80	299	235
0813360	城市	143	80	287	232
0814521	城市	140	83	284	215
7233396	城市	144	82	283	211
城市总和		718	417	1456	1132
0812772	乡村	131	57	260	197
0811170	乡村	132	39	270	212
0813474	乡村	131	76	277	214
0811854	乡村	131	37	275	204
0301029	乡村	131	48	281	223
乡村总和		1363	257	656	1050

图 1　节选十个客户在 2021 年的四种卷烟品牌销量

之后利用 SPSS 进行回归系数以及检验统计量 R^2 的求解，如下表：

表 1　检验统计量 R^2 的求解

模型摘要				
模型	R	R^2	调整后 R^2	标准估算的错误
1	0.686	0.471	0.260	375.98743

预测变量：市场类型 x_1，卷烟品牌 x_2；因变量：销量 y

在多元回归分析中，需要说明各个解释变量的相对重要性，或者比较被解释变量对各个解释变量的敏感性。然而，回归系数与变量原有的计量单位有直接关系，计量单位不同，彼此不能直接比较。为此，需要引入标准化系数 *Beta*，用 $\hat{\beta}_j^*$ 表示。

对于二元线性回归模型，可以按下列公式直接计算 *Beta* 系数：

$$\hat{\beta}_1^* = \hat{\beta}_1 \sqrt{\frac{\sum x_{1i}^2}{\sum y_i^2}}, \hat{\beta}_2^* = \hat{\beta}_2 \sqrt{\frac{\sum x_{2i}^2}{\sum y_i^2}} \tag{6}$$

同样利用 SPSS 软件分析得到各类系数：

表 2　系数的求解

系数						
	模型	未标准化系数 B	标准化系数 $Beta$	显著性检验值 t	显著性	共线性统计容差
1	常量	1226.150	/	2.539	0.052	/
	品牌	-123.475	-0.676	-2.077	0.092	1.000
	市场类型	99.250	0.121	0.373	0.724	1.000

R^2 为 0.471 表明销量的 47.1%是由于市场类型以及品牌决定的，系数中卷烟品牌标准化系数为负，表明销量会随卷烟品牌的价格上涨呈现负相关地减少。

5.1.3 模型改进

由于市场类型的显著性过高达到 72.4%，说明市场的差异对销量的影响不是那么大。考虑到品牌的销量可能会受到不同销售地区的影响，我们添加一个交互项：$x_1 * x_2$。则新模型为：

$$y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_1 x_2 \tag{7}$$

并记为模型 2，同样利用多元线性回归分析以及上面的数据得：

表 3　改进模型检验统计量 R^2 的求解

模型摘要				
模型	R	R^2	调整后 R^2	标准估算的错误
2	0.766	0.587	0.277	371.63561

预测变量:市场类型 x_1，卷烟品牌 x_2；因变量：销量 y

表 4　改进模型系数的求解

系数						
	模型	未标准化系数 B	标准化系数 $Beta$	显著性检验值 t	显著性	共线性统计容差
2	常量	1971.650	/	2.315	0.082	/
	品牌	-309.850	-1.695	-1.667	0.171	1.000
	市场类型	-397.750	-0.487	-0.739	0.501	0.238
	交互项	124.250	1.235	1.057	0.350	0.076

由上表可见，R^2 提高至 58.7%，且系数最大的显著性水平降至 50.1%更加说明了销量与卷烟品牌和市场类型之间的相关性。

5.1.4 小结

我们通过分析客户所处的销售环境以及卷烟品牌来刻画客户与品牌之间的

关系，可见销售环境以及卷烟品牌对卷烟的销售至关重要。可猜测其包含的两大主要因素为人口因素和经济因素。人口因素包括该区域总人数和男女人数之比。通常，总人数和订购量两者之间呈正比。一般男性越多，烟的市场需求越大。经济因素主要是居民人均可支配收入，卷烟销售与居民人均可支配收入呈正相关，在经济发展较好的城市，居民对生活质量的要求越来越高，对档次较高的卷烟需求量逐渐变大，而经济发展相对滞缓的农村，居民更倾向于综合性价比高的卷烟。

5.2 问题二的模型建立与求解

5.2.1 分公司月销量变换规律及其因素

我们通过统计 2019-2021 年分公司的月销量，得到图 2（2019 与 2020 年图见附录 A 图 1、2）。分析图表可知，三年间乡村现代终端的销量一直处于低水平，只有一月份的销量突破 100 箱；普通终端销量远大于其现代终端销量。城市普通终端销量略大于现代终端销量，二者几乎持平。而乡村普通终端与城市两终端销量很相近。

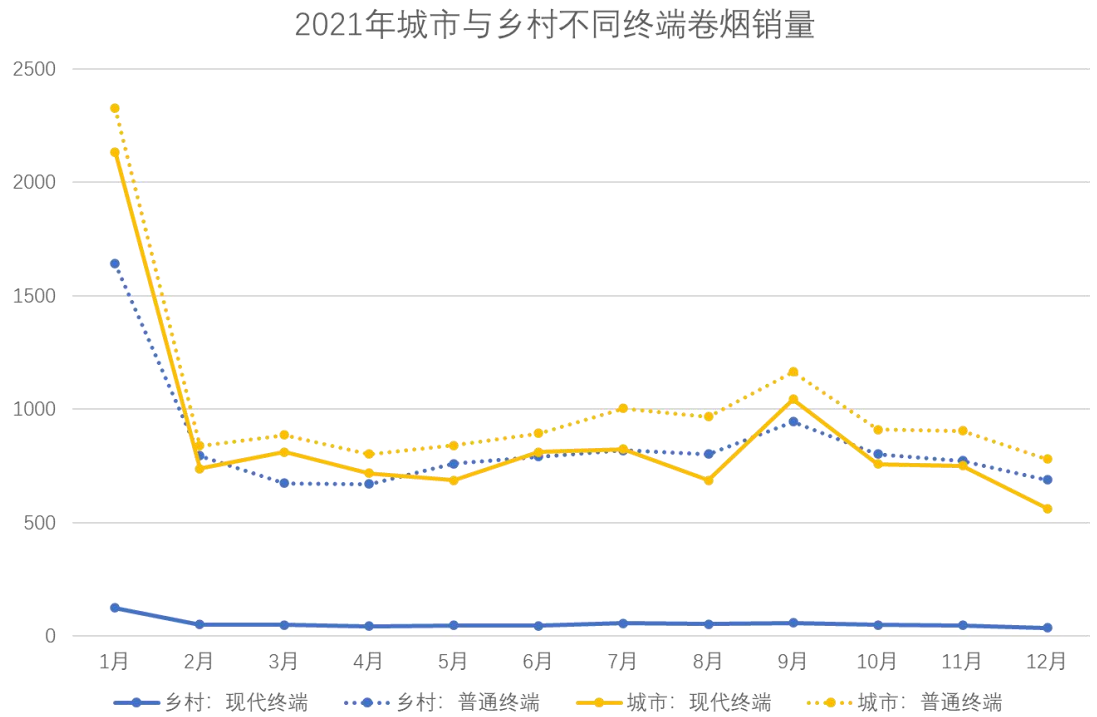


图 2　2021 年城市与乡村不同终端卷烟销量

主要因素的分析，由于问题一中利用改进二元线性回归模型的说明效果较好，我们在此问题中继续沿用。根据假定，月份、市场类型以及终端类型对销量的影响是线性的，且这三个因素之间没有线性关系。我们将此模型即为 3，同时，月份 x_3、市场类型 x_1 与销售终端 x_4 有如下定义：

$$x_1{:}1-12\text{月份}；\ x_1\begin{cases}0，\text{代表乡村}\\1，\text{代表城市}\end{cases}；\ x_4\begin{cases}0，\text{代表普通终端}\\1，\text{代表现代终端}\end{cases}$$

同样地，我们利用 SPSS 进行分析，得到 2019 年可决系数 R^2 为 0.413，2020 年为 0.442，2021 年为 0.521。同时也到了其他相关系数及其显著值。（2019 与 2020 年图见附录 A 表 1、2）

表 5　改进模型对 2021 年卷烟销量系数的求解

模型		未标准化系数 B	标准化系数 $Beta$	显著性检验值 t	显著性
3	常量	908.150	/	7.190	0.000
	月份	-34.031	-0.240	-2.380	0.022
	市场类型	500.604	0.512	5.070	0.000
	销售终端	-470.372	-0.481	-4.764	0.000

以上三年的多元线性回归分析都可以从理论上说明：卷烟的销量与销售的时间（月份），销售终端类型，销售市场类型有很强的相关性（从极低的显著性水平可知），且和月份大致呈负相关，即一年之初销量最大，随着一年过去，每月的销量呈现递减趋势。市场类型变量的系数为正说明：城镇相较乡村卷烟销售量有明显的更多的大小关系。三个表格中的销售终端的系数显著性水平均为 0，且值为负数，说明普通终端的销量在绝大部分情况下会比现代终端占优势。

5.2.1 模型建立

为拟合出未来 1 年该分公司各月销量趋势曲线，我们使用 BF 神经网络模型。

5.2.1.1 模型准备

图 3 是一个三层 BP 神经网络的结构示意图。流程为：一开始先将数据输入，然后经过正向传播，经过隐含层来到最终的输出层，然后通过输出层的计算与判断与真实值的之间的误差是否过大，如果过大则将信号进行反向的传播，通过这些误差的数据去修改网络的权值和与权值对应的阈值。然后如此反复循环，直到他超过了它的最大传输次数或者实验结果符合预期设定的标准，结束循环。

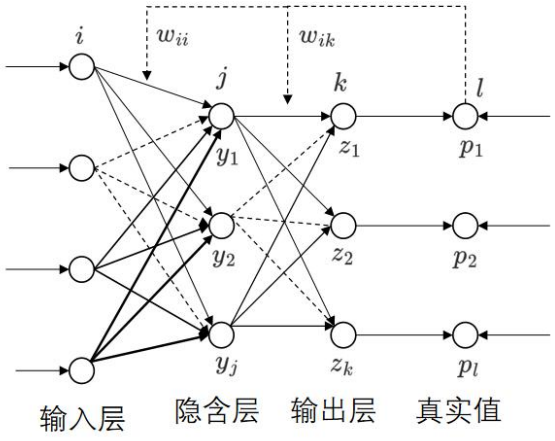


图 3　三层 BP 神经网络的结构图

5.2.1.2 模型流程

传统 BP 神经网络有如下的简易流程(以下公式中 $i=1,2\ldots n$；$j=1,2\ldots n$ ）。

• 将网络初始化。

首先，设置它的激励函数为 $g(x)$， $g(x)$ 为 Sigmoid 函数，表达式为:

$$g(x)=\frac{1}{1+e^{-x}} \tag{8}$$

• 隐含层的输出:

一般取三层 BP 网络，有几层隐含层则 n 为几，输入层到达隐含层的偏位置为 a_j，取输入层到达隐含层的权值为 ω_{ij}，隐藏层的输入为 H_j：

$$H_j=g(\sum_{j=1}^{n}\omega_{ij}x_i+a_j) \tag{9}$$

• 输出层的输出:

输出层的输出为 O， b 为隐含层到输出偏置，表达式为：

$$O_k=\sum_{j=1}^{i}H_j\omega_{jk}+b_k \tag{10}$$

• 计算误差:

一般来说，误差的计算公式如下:

$$E=\frac{1}{2}\sum_{k=1}^{m}(Y_k-O_k)^2 \tag{11}$$

将 $Y-O=e$ 记作，则 E 可以将其表达为:

$$E=\frac{1}{2}\sum_{k=1}^{m}e_k{}^2 \tag{12}$$

• 最后将权值重新更新下，公式如下:

$$\begin{cases}\omega_{ij}=\omega_{ij}+\eta H_j(1-H_j)x_i\sum_{k=1}^{m}\omega_{jk}e_k\\ \omega_{jk}=\omega_{jk}+\eta H_j e_k\end{cases} \tag{13}$$

• 偏置的更新

$$\begin{cases}a_j=a_j+\eta H_j(1-H_j)\sum_{k=1}^{m}\omega_{jk}e_k\\ b_k=b_k+\eta e_k\end{cases} \tag{14}$$

隐含层传递到输出层的偏置为:

$$\frac{\partial E}{\partial b_k}=(Y_k-O_k)(-\frac{\partial O_k}{\partial b_k})=-e_k \tag{15}$$

然后更新后结果为：

$$b_k=b_k+\eta e_k \tag{16}$$

输入层传递到隐含层的偏置为:

$$\frac{\partial E}{\partial a_j}=\frac{\partial E}{\partial H_j}\cdot\frac{\partial H_j}{\partial a_j} \tag{17}$$

更新后的结果为:

$$a_k=a_k+\eta H_j(1-H_j)\sum_{k=1}^{m}\omega_{jk}e_k \tag{18}$$

5.2.2 模型求解

由前面对本分公司的 3 年每月卷烟销量的分析知：烟卷的销量受时间、市场类型，终端类型这三类因素的影响比较大，故选取时间，市场类型以及终端类型为 BP 神经网络输入层的数据输入端进行训练，销量作为唯一的输出端口。且根据假设，卷烟的销量只受这三类因素的影响。

表 6　BP 神经网络网络信息

网络信息			
输入层	协变量	1	Zscore(月份)
		2	Zscore(市场类型)
		3	Zscore(终端类型)
	单元数		3
	协变量的重新标度方法		标准化
隐藏层	隐藏层数		1
	隐藏层 1 中的单元数		1
	激活函数		双曲正切
输出层	因变量	1	销量
	单元数		1
	标度因变量的重新标度方法		标准化
	激活函数		恒等式
	误差函数		平方和

搜集 2019 年 1 月至 2021 年 12 月共 36 个月每个月的不同市场类型，不同销售终端的销量，其中用前 24 个月的 96 个数据进行训练，2021 年的 48 个数据用于检验模型预测的准确度。数据整理并对输入的元素归一化后有如下表格（截取部分）。

月份	市场类型	终端类型	销量	Z月份	Z市场类型	Z终端类型	MLP_PredictedValue
1	0	0	1449.7308000000000	-1.69191	-.99739	-.99739	962.2488805
1	0	1	158.3232000000000	-1.69191	-.99739	.99739	159.2002484
1	1	0	1779.2918000000000	-1.69191	.99739	-.99739	1006.964186
1	1	1	2212.5926000000000	-1.69191	.99739	.99739	919.6407938
2	0	0	561.2032000000000	-1.61992	-.99739	-.99739	961.2184196
2	0	1	58.2892000000000	-1.61992	-.99739	.99739	154.6101539
2	1	0	635.8020000000000	-1.61992	.99739	-.99739	1006.934312
2	1	1	773.8144000000000	-1.61992	.99739	.99739	917.7369122
3	0	0	743.4584000000000	-1.54792	-.99739	-.99739	960.1659148
3	0	1	72.2392000000000	-1.54792	-.99739	.99739	150.0767956
3	1	0	821.5406000000000	-1.54792	.99739	-.99739	1006.903741
3	1	1	951.4480000000000	-1.54792	.99739	.99739	915.7958221
4	0	0	889.3840000000000	-1.47592	-.99739	-.99739	959.0909418
4	0	1	81.1488000000000	-1.47592	-.99739	.99739	145.6003178
4	1	0	878.0448000000000	-1.47592	.99739	-.99739	1006.872456
4	1	1	1008.9484000000000	-1.47592	.99739	.99739	913.8169563
5	0	0	874.0138000000000	-1.40393	-.99739	-.99739	957.9930700
5	0	1	75.4080000000000	-1.40393	-.99739	.99739	141.1808312
5	1	0	853.1316000000000	-1.40393	.99739	-.99739	1006.840442

图 4　2019-2021 年份销量

借助 SPSS 数据分析工具，经过不断地调整中间隐藏层的神经元结点数以及训练的方向等优化操作，最终可以得到预测的结果：

表 7　BP 神经网络训练集

个案处理摘要			
		N	百分比
样本	训练	144	75.0%
	检验	48	25.0%
有效		192	100.0%
总计		192	

表 8　BP 神经网络预测结果

模型摘要		
训练	平方和误差	30.165
	相对误差	0.422
	使用的中止规则　误差在 1 个连续步骤中没有减小	
检验	平方误差	6.597
	相对误差	0.186

5.2.3 检验分析

依据上面的模型分析数据可知该模型的相对误差高达为 18.6%，处于较好的预测态势，预测与实际的对比图如下：

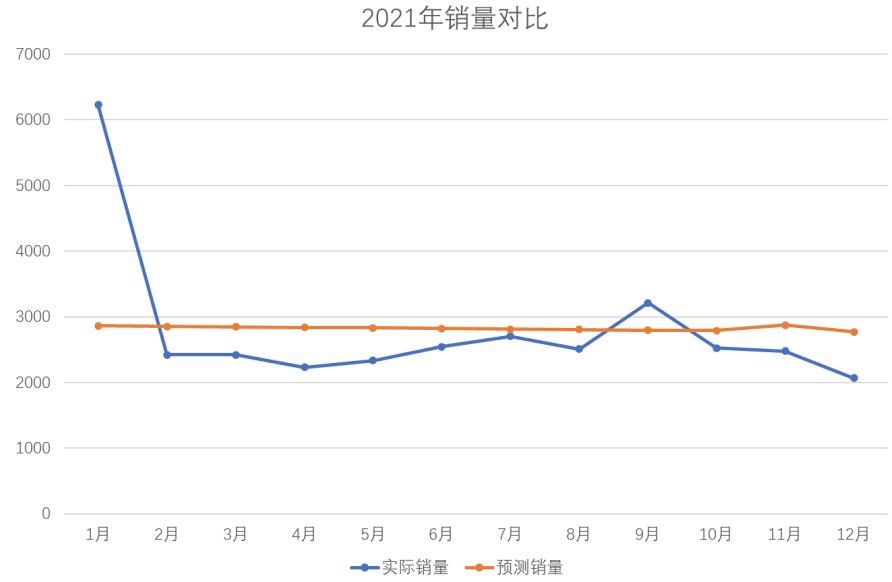


图 5　2021 年实际销量与模型预测销量比对

5.2.4 问题结论

故运用该预测模型预测 2022 年该分公司的月销售情况可得如下图：

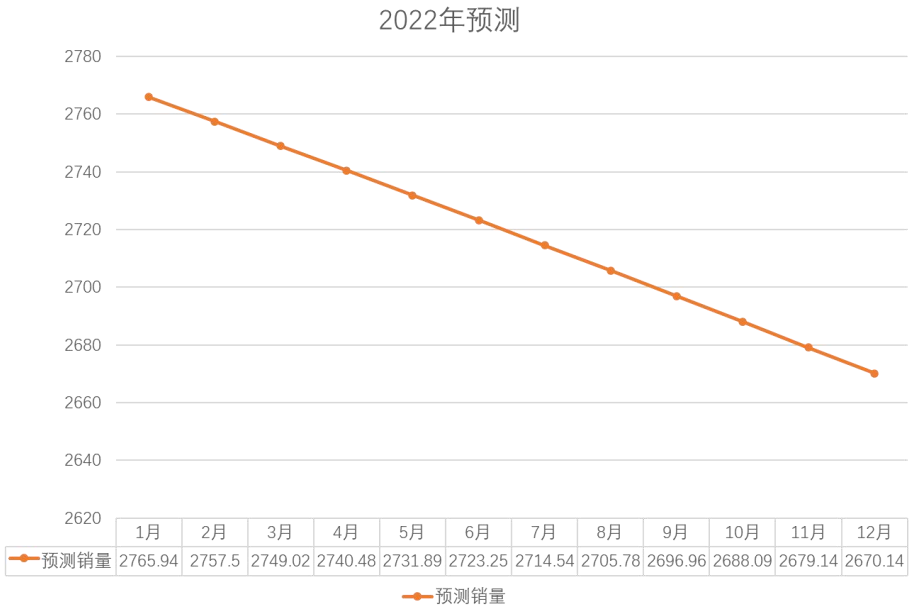


图 6　2022 年分公司每个月的预测销量

5.2.5 模型改进

尽管 BP 神经网络算法具有组成成分不复杂，逼近性良好，训练的网络层数等可以依据训练的不同状况不断调整，由于其算法的本质和网络体系结构，不避免地存在以下的缺点：

1）计算权值容易出现局部极小化，导致 BP 神经网络无法实现应有的作用，网络训练不成功；

2）由于 BP 神经网络的算法是梯度下降法，在某些情况下，算法训练时会产生停滞，使得收敛效率十分低下，大大延长了 BP 神经网络的收敛时间；

3）由于 BP 神经网络的隐含层数不是确定的，可以是一层也可以是多层，建立 BP 神经网络模型时，没有明确的标准指导选择多少隐含层，这就导致 BP 神经网络的结构差异较大，只能根据研究人员凭借经验一一测试，在一些实际问题之中，实用性降低。

相较于 BP 神经网络，RBF 神经网络能够更加快速的完成函数逼近，拥有更加快速的收敛速度。RBF 神经网络具有一般性，从理论上讲，只要 RBF 神经网络的隐含层中神经单元节点的个数足够多，它就可以逼近任意函数。故试着使用 RBF 神经网络模型进行优化。

5.2.5.1 模型准备

RBF 网络最基本的结构由输入层 m，隐含层 m_2，输出层 m_g 构成。从 m 到 m_2 是非线性的传递信息，训练速度慢； m 到 m 是线性的传递信息，训练速度快，层与层之间通过径向基函数实现信息传递。其中，输入层的输入向量的个数等于隐含层神经单元的节点个数。径向基函数节点构成了隐含层的变换函数，高斯核

函数是常用径向基函数，形式如下:

$$R_i(x) = \exp[-\frac{\|x - c_i\|}{2\sigma_i^{\ 2}}] \qquad i = 1,2,3\ldots m \tag{19}$$

其中 x 是 n 维输入向量； c_i 是第 i 个径向基函数的中心点的参数，两个向量具有一样的维度； σ_i 是输入向量到径向基函数中心点的距离， m 是神经单元节点的个数； $\|x - c_i\|$ 为径向基函数的自变量，表示 x 与 c_i 之间的间隔； $R_i(x)$ 有且只有一个最大值存在于 c_i 处，随着 $\|x - c_i\|$ 的变大， $R_i(x)$ 快速的减小为零。

5.2.5.2 模型流程

输入层是神经网络与外界交互的窗口，由大量的信息源节点构成，输入层除了数据传输，没有任何其他作用；隐函数层有且仅有的一层，通过径向基函数将输入样本从输入层映射到隐含层上。隐含层的神经单元节点数与实际解决的问题有关，数目等于输入的样本数据的总数;径向基函数是一种局部响应函数。输出层是线性的，作用是响应输入层的输入样本数据。

RBF 网络的基本思想是：将定义构成神经网络空间的基本单元为径向基函数，大量的携带径向基函数的神经单元节点构成了隐含层，无监督学习确定径向基函数的中心参数，通过输入向量和中心点的间距建立映射关系，将输入向量投影到隐含层空间上。而隐含层到输出层的信息传递是线性的，RBF 神经网络的输出就是隐含层到输出层权值之和。尽管，从整体上看，RBF 神经网络的输入层到输出层的传递是非线性的，但是从局部上看，隐含层到输出层的输出权值是线性的，所以，对隐含层到输出层的线性关系建立方程组，从而就可以快速的求解隐含层到输出层的输出权值，完成神经网络训练，集大程度的规避了局部收敛的问题。

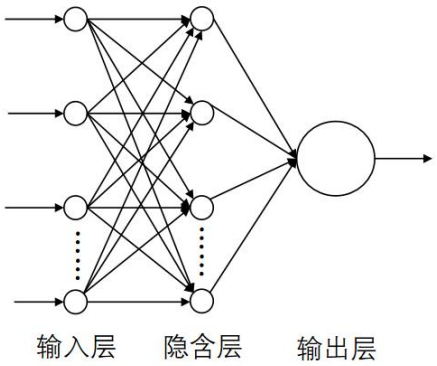


图 7　RBF 网络结构图

5.2.6 模型结果

应用搜集所得的 2019 年 1 月到 2021 年 12 月每个月的总销量以及按照除月份外的 2 个因素分的 4 类的对应销量分别作为 RBF 神经网络的输入数据端口，一次性预测 2020 年 1 月到 2022 年 12 月份的每月总销量以及这 4 类每月的销量请况。通过一定的训练方法以及参数的调整最终达到最佳效果的预测模型。

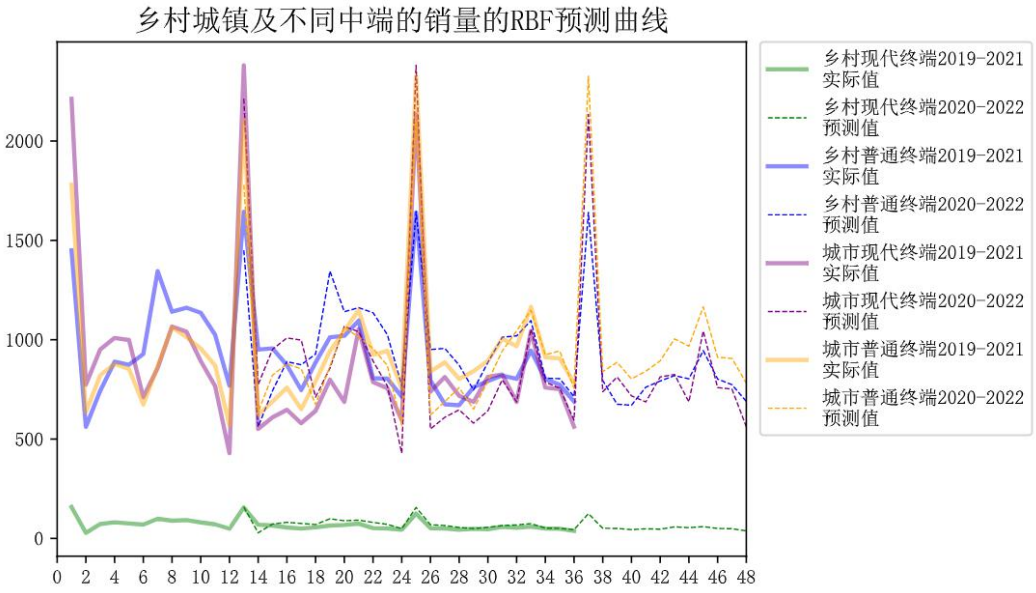


图 8　乡村城镇及不同终端的销量 RBF 预测曲线

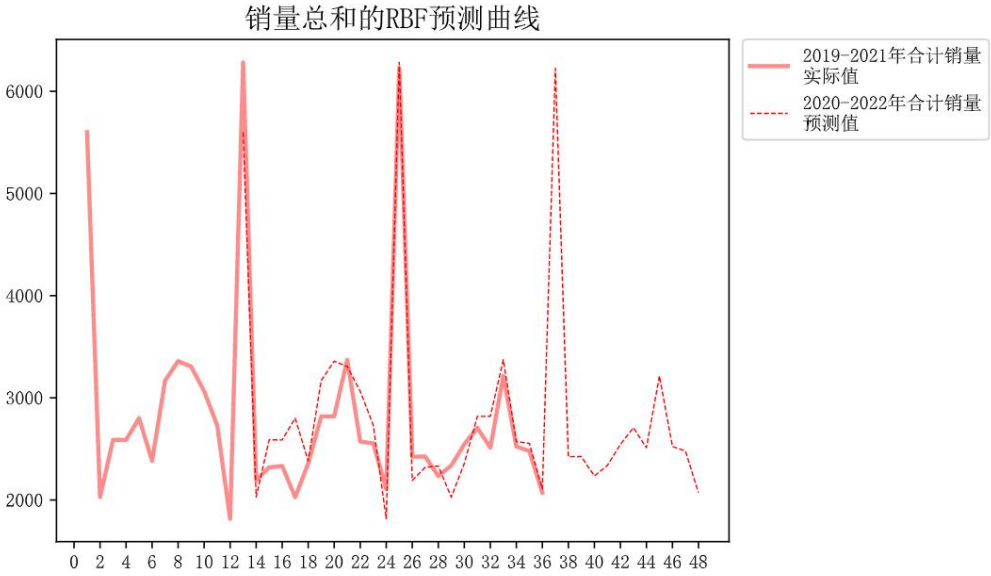


图 9　销量总和 RBF 预测曲线

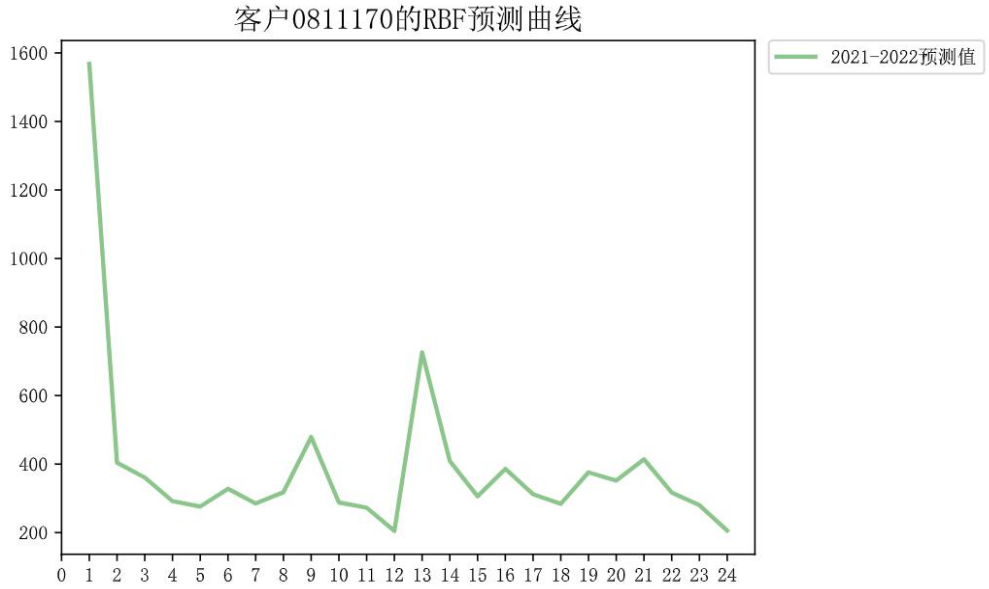
相对误差限平均为 9%, 相对误差上限为 58. 98%。与 BP 神经网络模型的相对误差相比，误差的范围缩小，精确度提高。故可认为该模型适合用于分析和预测卷烟的月销售请况。则最终采用该模型对该分公司未来一年的月销售数据预测。

表 9　RBF 预测 2022 年分公司卷烟销量

月份	1	2	3	4	5	6	7	8	9	10	11	12
销量	6225.885	2424.079	2424.375	2235.017	0338.006	2544.856	2707.038	2511.731	3213.828	2523.208	2479.943	2070.989

从上图的预测结果（第 36 个月到第 48 个月的销量数据）可看出未来一年的月销售大体趋势：1 月为整年卷烟销售的高峰期，直到 4 月份卷烟的销售都呈现销量减少的态势，从 5 月 8 月，卷烟的销售呈现略微地上升趋势，之后又开始销量递减。

将该模型用在客户 0811170 的月销量分析上可以获得该用户在 2022 年的卷烟月销售情况，预测结果如下（从第 12 个月到第 24 个月的销量数据）



月份	1	2	3	4	5	6	7	8	9	10	11	12
销量/箱	726	409	306	386	312	284	376	352	414	317	280	206

图 10　客户 0811170 的 RBF 预测曲线

5.3 问题三的模型建立与分析

5.3.1 四个卷烟价位段的销售占比及其特点

通过统计 2019-2021 年度四个价位段的卷烟销售情况，得到了其销售占比。

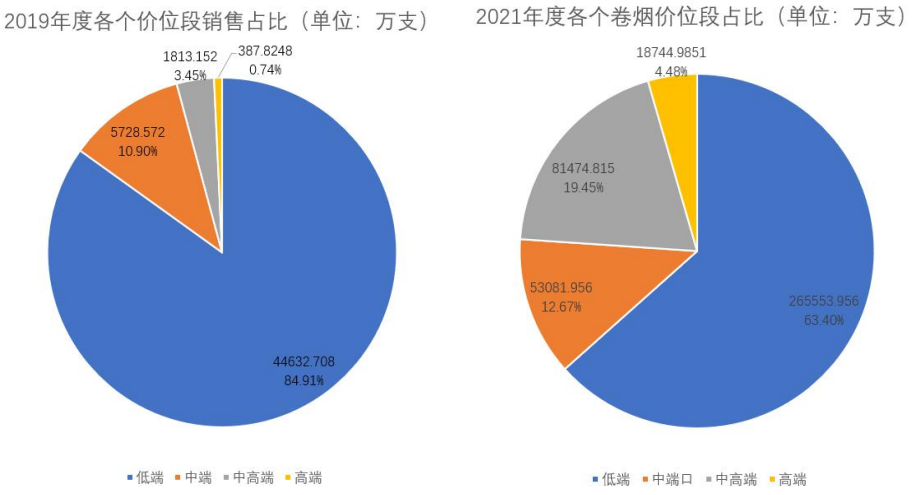


图 11　四个价位段卷烟市场销售占比

同时通过比较两个年份各价位段卷烟的供需关系，得到了如下图示：

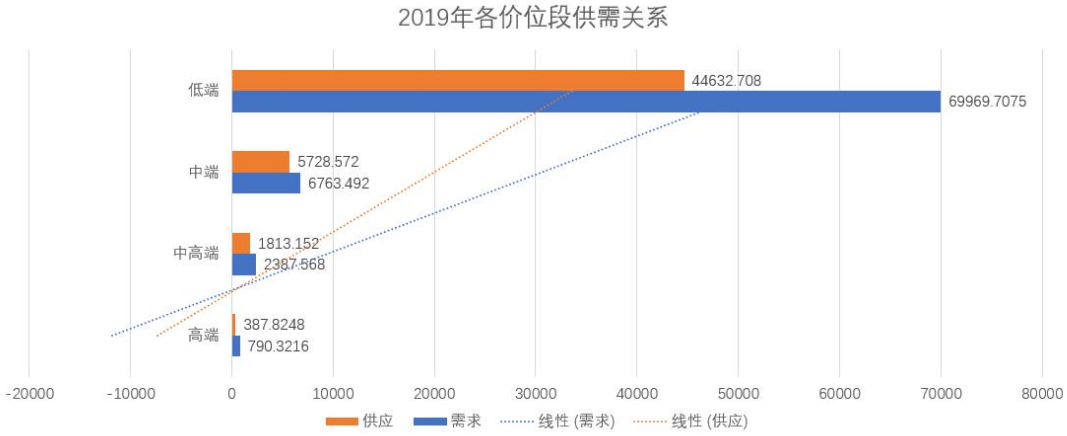


图 12　2019 年各价位段卷烟供需关系

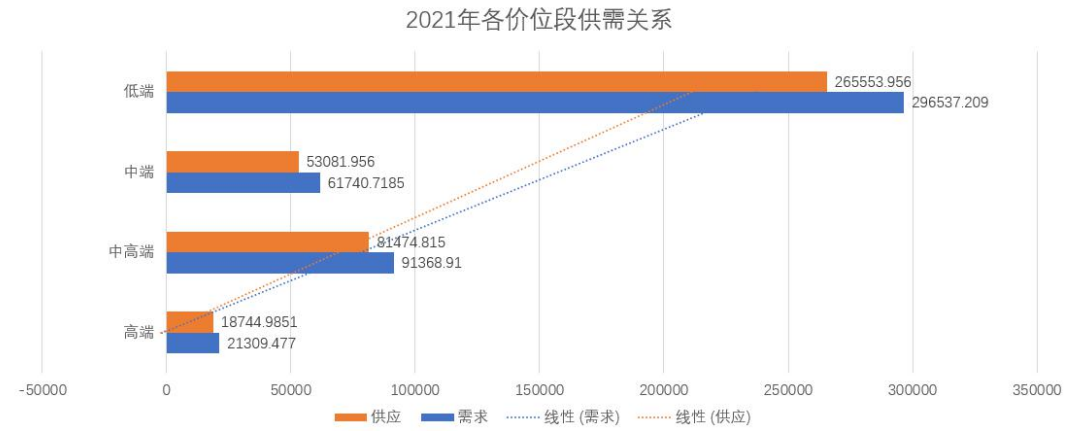


图 13　2021 年各价位段卷烟供需关系

由图可得出以下特征：

（1）高端及中高端类卷烟供求。2021 年全国高端卷烟销售比重为 4.48%，中高端卷烟为 19.45%，两者之和为 23.93%，与 2019 年可比口径的原一类卷烟销售比重 4.19%相比，增加了 19.74 个百分点。这两个段位的基本涵盖了零售价从 250-1000 元/条八个价格梯次，一个价类内部的价格差异过大。而因计税价格改变，实际上烟草商业对系统外销售价格没有变化的大约增加了 1 个百分点。目前高端及中高端卷烟销售比重在 23~24%水平基本上是与市场需求相适应的。同时我们还统计了它们的供需关系，发现中高档卷烟供给一般可能略小于需求，但基本上可以由市场自动调节。

（2）中端卷烟供求。2021 年中端卷烟销售比重为 12.67%，梯次和占比较小。这是由于中端卷烟的价格梯次过少，这与其作为消费主导价位卷烟的显示状况不相适应。目前的市场供求关系与 2019 年情况没有发生大的变化。

（3）低端卷烟供求。2019 年低端卷烟比重为 84.91%，占据了绝大部分的市场。2021 年低端卷烟销售比重为 63.40%，与 2019 年的同口径比较减少了 21.51 个百分点。这类烟规格较多，零售价基本保持在 15 元/包以下，价格梯度不大，但是价类区间较多，而销量越来越少，队行业效益增长的贡献度也越来越小。且此价格段供不应求的情况最为明显。

（4）纵向比对可知，四个价位段在两年间均存在或大或小的供不应求。

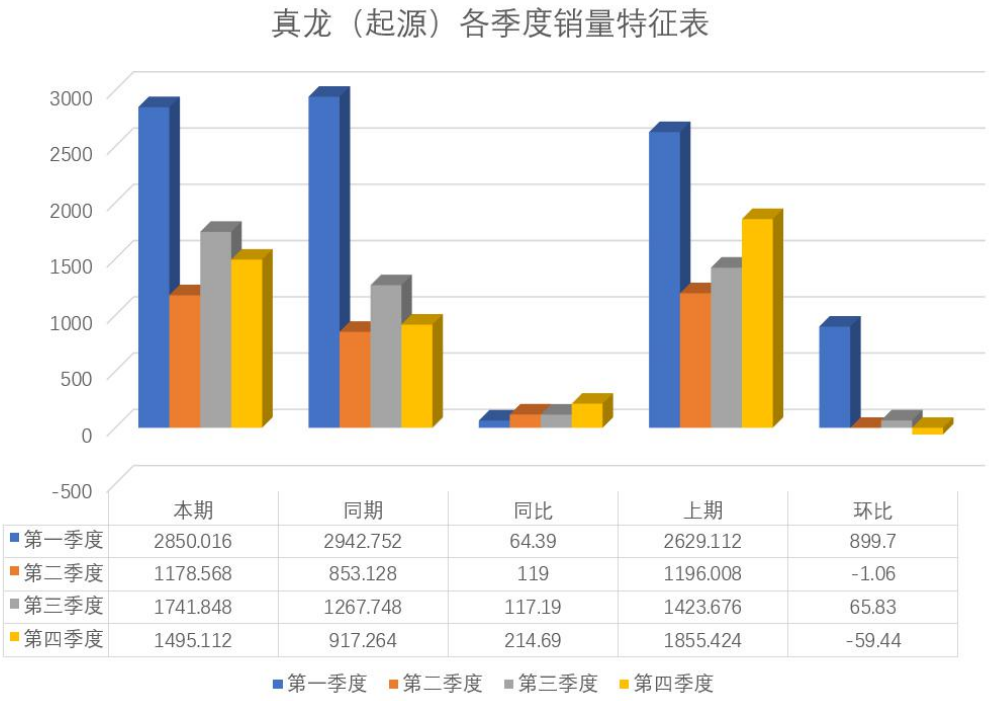
通过以上分析可知，在卷烟价类梯次化管理工作中，由于现行的卷烟价类划分均衡性不够，不仅影响卷烟价格“同档同价同差率”工作的开展，也对行业经济运行平稳发展、宏观决策和分析产生一定影响。

5.3.2 中端价位主要销售产品

5.3.2.1 真龙（起源）、硬盒芙蓉王、84 软盒玉溪各自特点

我们通过统计，分别得出了真龙（起源、）硬盒芙蓉王、84 软盒玉溪三种中端卷烟品牌在 2021 年四个季度销量的本期、同期、同比、上期以及环比的五个维度特征图进行自身特点分析。

对于真龙（起源），可见第一季度的销量最高，这与时间密不可分。根据中国传统文化习俗，逢年过节，年轻人更偏好于买烟送给有抽烟习惯的长辈，若时间间隔中恰逢节假日，卷烟的销售数量都比上期有明显的增长。其中涨势最高的是第一季度，1-3 月份正值元旦春节，涨势第二的第三季度中也有中元节和中秋节。对比同期，各个季度的销售趋势大体不变，只在销售数量上有小幅增加，同比增长率保持上升趋势，增长速度快且平稳。环比更能说明趋势问题，第一季度销量的大幅增加导致其环比远超过其他三个季度。



	本期	同期	同比	上期	环比
第一季度	2850.016	2942.752	64.39	2629.112	899.7
第二季度	1178.568	853.128	119	1196.008	-1.06
第三季度	1741.848	1267.748	117.19	1423.676	65.83
第四季度	1495.112	917.264	214.69	1855.424	-59.44

图 14　真龙（起源）各季度销量

硬盒芙蓉王与真龙（起源）于各个季度所显示的特征极为相似。不同的是，硬盒芙蓉王在前两个季度都呈现不同程度的同比下降，是三者最低增长率。

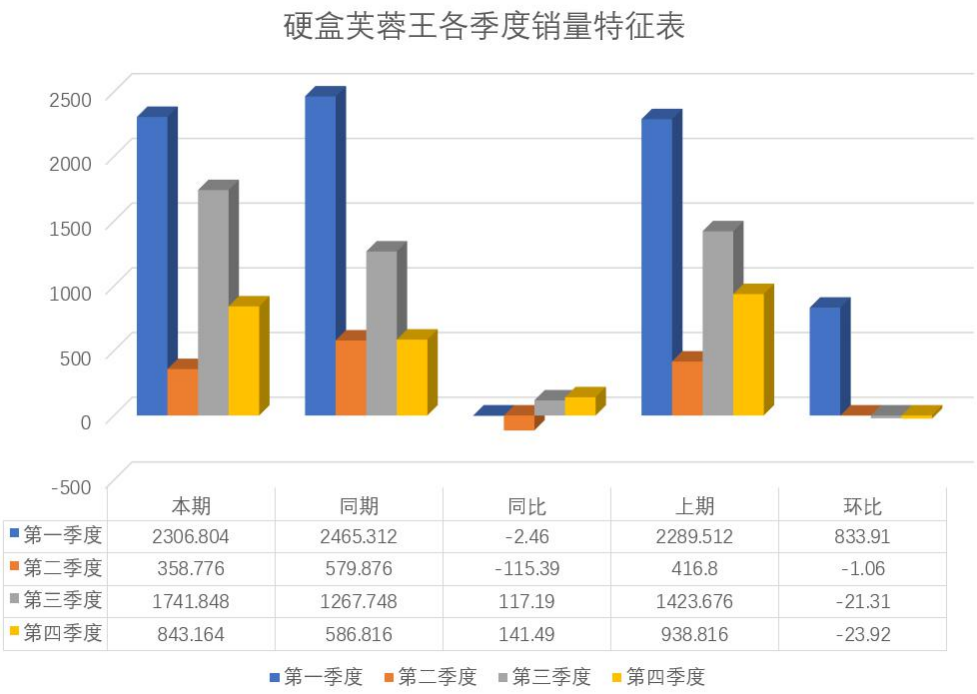


图 15　硬盒芙蓉王各季度销量

84 软盒玉溪也具有上述两个品牌的特征。值得注意的是，除了第一季度，在其他三个季度其销量都较为平稳，没有出现巨大落差。

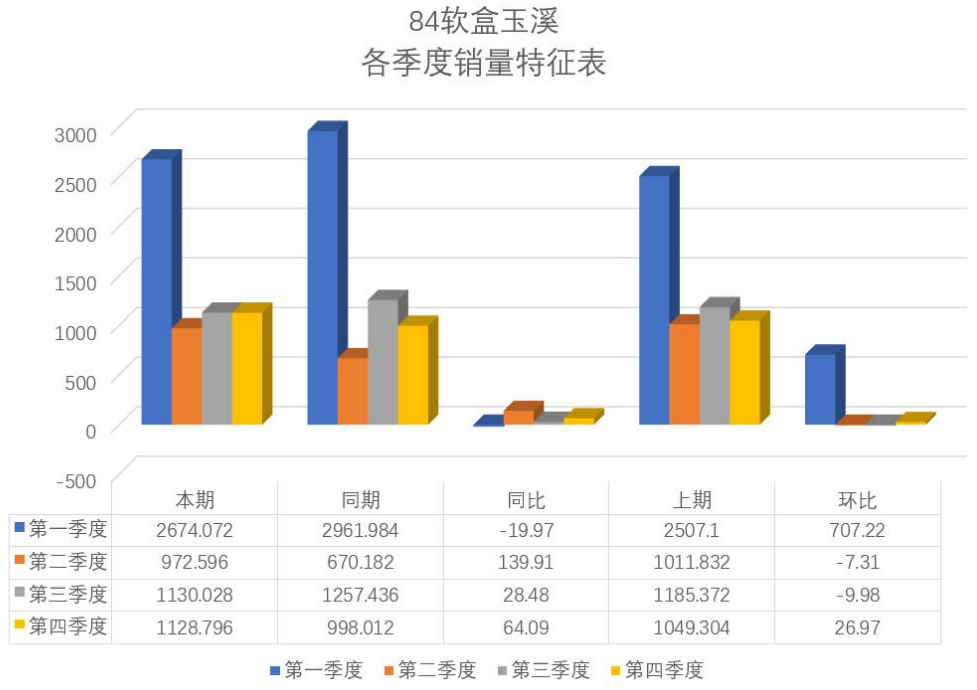


图 16　84 软盒玉溪各季度销量

5.3.2.2 三者数量关系

随后，我纵向们对比分析了三种卷烟品牌在 2021 年的本期销量。三者都在一月份有着 1500 万支以上的销量，随后在二月份销量急剧下降并趋于平缓。综合来看，真龙（起源）是中端价位的最畅销卷烟品牌，硬盒芙蓉王相对于其他两者销量最差。

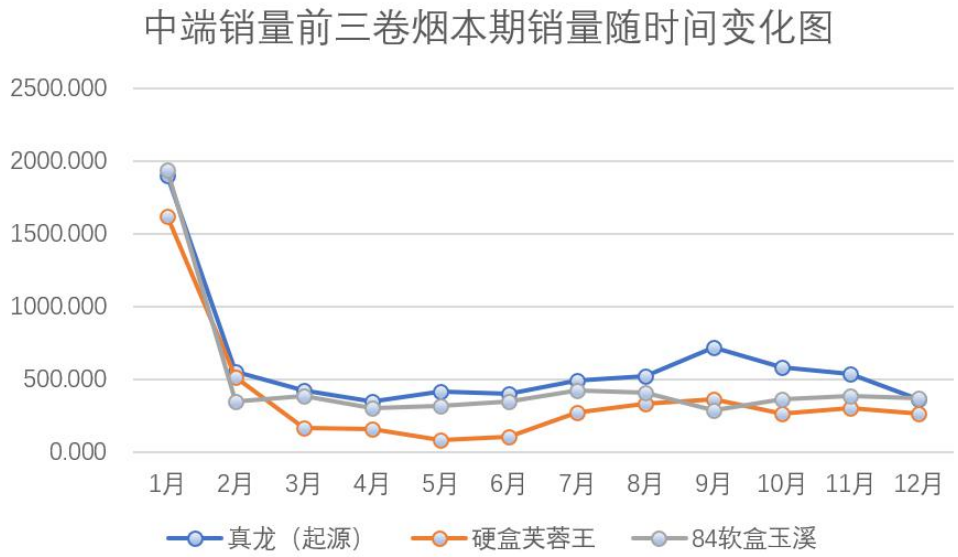


图 17　中端销量前三卷烟年销量变化

5.3.2.3 竞争与替代

我们认为三者之间存在竞争，主要从以下几个方面分析：

• 从产品品类上来看：“真龙”品牌产品特征为烟草本香飘然而入，浓而不烈，芬芳怡然，烟气醇和，热血却不冲动，豪迈而不嘈杂，洗尽铅华，回归本质；“芙蓉王”品牌是属于“中间香型”品类的代表，产品特征为整体平衡性好，醇和雅致，层次丰富；而“玉溪”品牌则属于“清香型”品类的代表，产品特征为香味清新的感受比较强烈，香味飘逸，幽雅而愉快，远扬而留暂，突出而清新。这也表明了“真龙”、“芙蓉王”和“玉溪”这三大品牌长期以来一直拥有着各自相对稳定的目标消费群体，双方都在各自的品类优势领域里面深耕细作。

• 从产品传播方式上看：“真龙”品牌运用龙这一意向，或缘于物华天宝、稀世天成，或系于王者风范、高贵典雅，或出于内心景仰、推崇备至，“真龙（真男儿）”三者皆有，浑然一体、玉成其美，无愧是中式卷烟的集大成者。“芙蓉王”品牌的传播方式是“乱和精”。“乱”是较难看出清晰的逻辑路径，既有公益性质的，也有体育性质，更有娱乐性质等等；虽说传播遍布全中国，但却缺乏一条清晰的逻辑路径，进而导致了整合传播的效用很难全盘发挥出来。“精”则是指“芙蓉学子”大型公益活动，为“芙蓉王”品牌树立了高度的社会责任形象，也让“芙蓉王”品牌的“传递价值，成就你我”得到了最充分的演绎。而“玉溪”品牌的传播方式是“专注”，只专注于自己最擅长的领域，只专注于与品牌文化高度吻合的领域。正如，“围棋之道”，于经纬交织木上演绎“天地万物环环相围而生”之理，于微中见大，常被古人谓之“善围者，可以王天下。”“玉溪”

品牌的德文化与“围棋运动”的完美结合，使得了“玉溪”品牌的德文化在传播中得到了高端消费者的广泛认同，也赢得了高端消费者在情感上的深度共鸣。

• 从产品价位上来看：“真龙”品牌产品线下共生产 27 款产品，主要覆盖了从 55 元/条至 1000 元/条的价位区间，品牌内部差距十分大。其中最畅销的中端子品牌起源占其价格段销量的九成。“芙蓉王”品牌产品线主要覆盖了从 250 元/条至 500 元/条的数款产品规格。产品规格部分为零售价 250 元/条的“硬盒芙蓉王”、280 元/条的“芙蓉王（硬细支）”、350 元/条的“芙蓉王（硬蓝新版）”、500 元/条的“芙蓉王（硬领航）”。分析具体数据可知，“硬盒芙蓉王”和“芙蓉王（硬蓝新版）”是“芙蓉王”品牌销量最大的两款主力产品规格，尤其是“硬盒芙蓉王”这款产品规格甚至占到了“芙蓉王”品牌整体销量的七成左右。“玉溪”品牌产品线主要覆盖了从 200 元/条至 1000 元/条的九款产品规格。其中，“硬盒玉溪（和谐）”和“84 软盒玉溪”是“玉溪”品牌销量最大的两款主力产品规格，尤其是“84 软盒玉溪”这款产品规格甚至占到了“玉溪”品牌整体销量的七成左右。由此可见，三大品牌在零售价 200 元/条-250 元/条即中端这一价位区间竞争最为激烈，尤其是这三款各自品牌的绝对主力产品规格更是在同价位区间展开最为激烈的较量。

对于三者是否存在替代，单从价格而言是完全可以替代的，但具体落实到品牌文化、地域风俗乃至个人习惯，其实每个产品都是不可替代的。

5.3.2.4 未来一年的趋势

由于第二问中 RBF 模型预测的结果较好，此处不再建立新的模型，继续用 RBF 进行预测。

图中采用 2021 年的各月份本期销量作为源数据，来预测 2022 年的各月份销量。可见预测的未来一年销量趋势跟 2021 年大相径庭，真龙（起源）除 5、8 月份，84 软盒玉溪除 3、6 月份销量可能较去年有所下降外，其他月份都近似等同；硬盒芙蓉王销量几乎稳定不变。

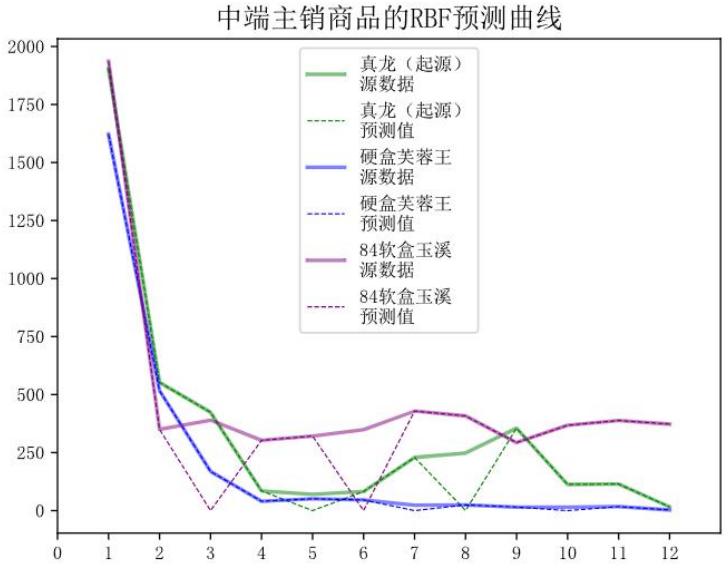


图 18　中端销量前三卷烟未来一年销量预测

5.3.3 高端价位主要销售产品

5.3.3.1 中华（双中支）、软盒中华、真龙（海韵细支）数量关系

高端价位最畅销的三种卷烟品牌分别是：中华（双中支）、软盒中华以及真龙（海韵细支）。可见软盒中华销量略超过其他两个品牌，而中华（双中支）和真龙（海韵细支）在销量上差别不大。

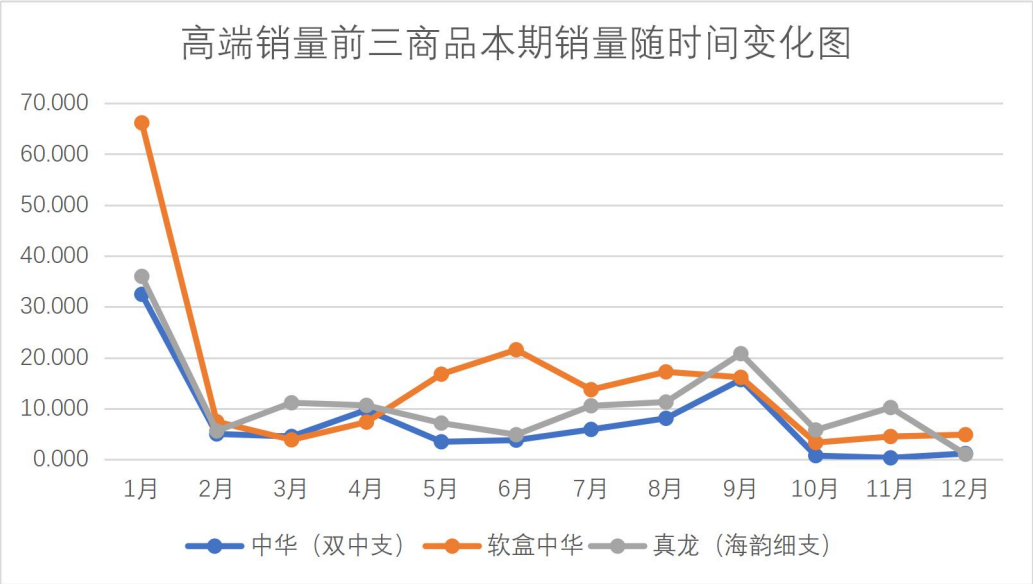


图 19　高端销量前三卷烟年销量变化

5.3.3.2 未来一年的趋势

由 RBF 预测曲线可知道，未来一年销量与 2021 年大致相同。且对比预测数据可知，软盒中华仍会占到高端主销产品的 5 成，而中华（双中支）只占 2 成。

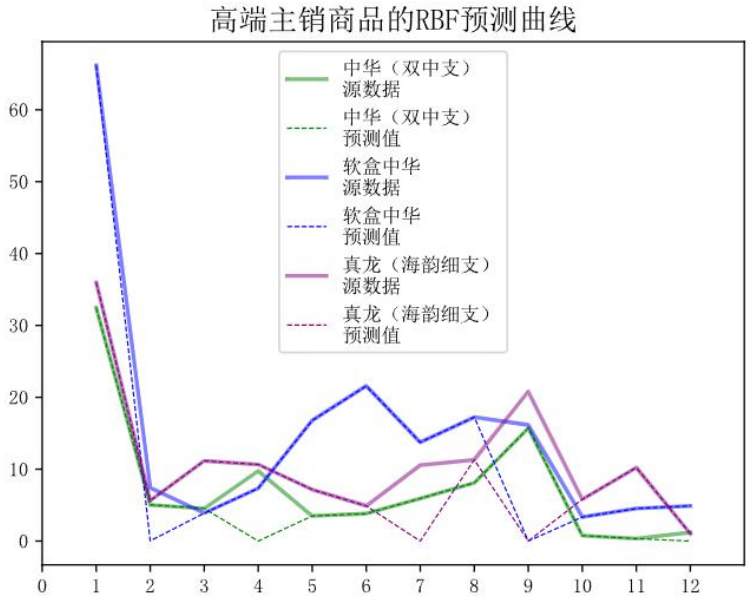


图 20　高端销量前三卷烟未来一年销量预测

5.3.4 中高端价位主要销售产品

5.3.4.1 真龙（海韵）、玉溪（软尚善）、硬盒中华数量关系

中高端价位最畅销的三种卷烟品牌分别是：真龙（海韵）、玉溪（软尚善）和硬盒中华。可见真龙（海韵）销量远远超过其他两个品牌，而中华（双中支）和真龙（海韵细支）在销量上差别不大且十分平稳。

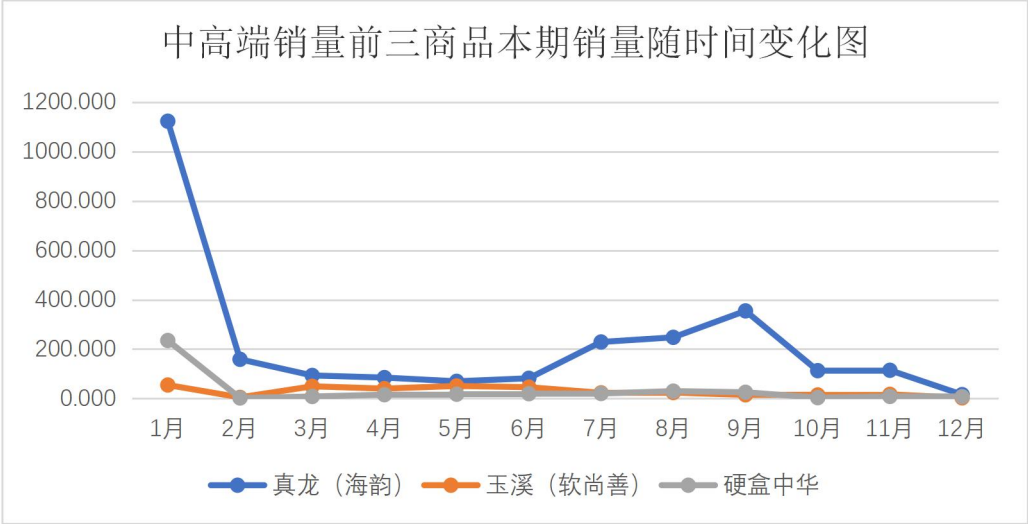


图 21　中高端销量前三卷烟年销量变化

5.3.4.2 未来一年的趋势

由 RBF 预测曲线可知道，未来一年销量与 2021 年大致相同。且对比预测数据可知，真龙（海韵）销量将会在 5、7 月份有所下降，但仍占到中高端主销产品的 7 成。

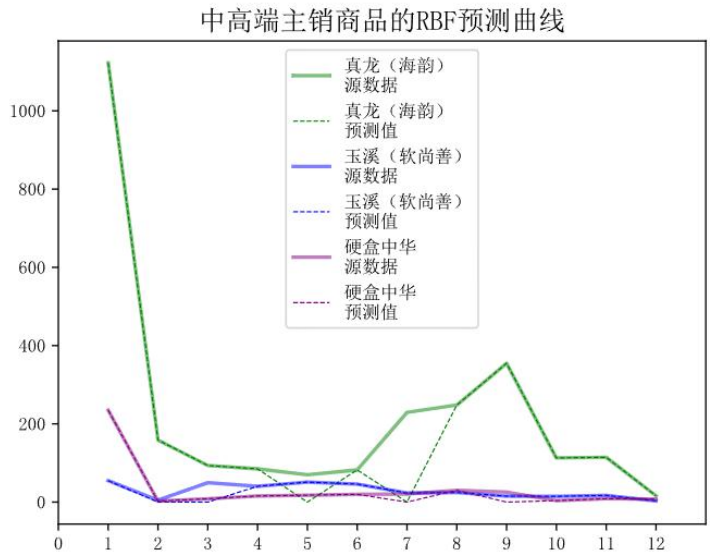


图 22　中高端销量前三卷烟未来一年销量预测

5.3.5 低端价位主要销售产品

5.3.5.1 黄金叶（金满堂）、真龙（软祥云）、红塔山（硬经典 100）数量关系

低端价位最畅销的三种卷烟品牌分别是：黄金叶（金满堂）、真龙（软祥云）以及红塔山（硬经典 100）。可见在整体上真龙（软祥云）在销量上领先于其他两个品牌。黄金叶（金满堂）总体销量一直处于最低，而红塔山（硬经典 100）各个月份之间的销量呈现跳跃式变化，不稳定。

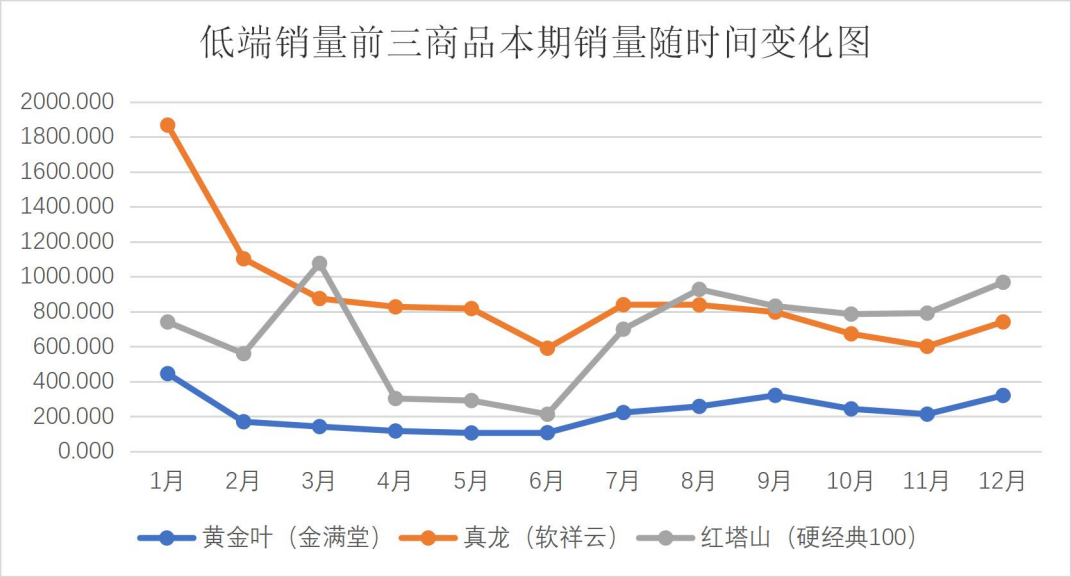


图 23　低端销量前三卷烟年销量变化

5.3.5.2 未来一年的趋势

低端卷烟预测结果同其他几个价位段有较大出入。真空（软祥云）的销量在 1 月份快要接近 0，这与事实不符，且三个产品在 6 月份销量出现了不同程度的下降。

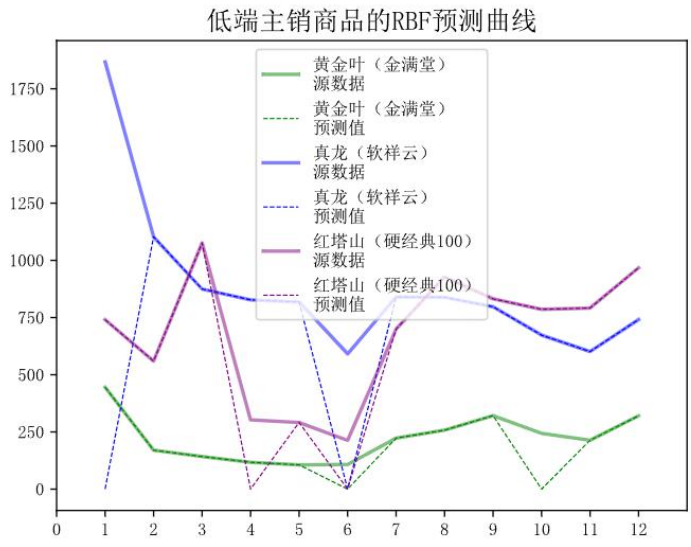


图 24　低端销量前三卷烟未来一年销量预测

5.4 问题四的分析总结

图 25 为该分公司的月销量的品牌宽度均值数据，从以上数据不难看出这三年里，城市的卷烟品牌宽度都比乡村的大。也可以发现分公司的月销售数据也可以发现现代终端的销量几乎都低于普通终端的销量，这在一定程度上说明公司可以把更多精力放在农村普通终端的建设营销上，城市中可以以普通终端为主，现代终端为辅，综合其各自优势，达到最好销售效果。

	2019年		2020年		2021年	
	乡村	城市	乡村	城市	乡村	城市
1月	25.32685	48.419	39.44277	40.50578	37.98607	59.97759
2月	16.82783	31.33499	25.08894	36.74207	24.43134	39.37927
3月	18.5949	35.33715	25.24125	37.60864	18.69334	39.83543
4月	22.63013	38.70737	25.46722	38.68589	18.39041	40.914
5月	24.60052	42.17198	27.74961	42.48983	18.25511	38.21366
6月	25.11515	35.97083	26.87493	40.93148	20.02735	41.92404
7月	31.9171	48.75478	31.79969	46.57928	26.04875	45.56051
8月	31.46217	48.61641	32.63081	51.90119	24.90547	44.30996
9月	29.18022	43.90505	30.7113	48.48308	24.6781	45.76263
10月	32.11771	50.29449	28.1958	46.03336	23.9723	41.44743
11月	32.43321	49.96886	26.70005	41.39174	24.76825	40.20469
12月	21.9746	27.48125	23.53987	36.50957	20.28729	34.00576

图 25　乡村与城市 2019-2021 卷烟品牌销售宽度

同时根据统计分析及模型预测结果可知，1 月份在近几年来始终是销量的最高潮，公司可以抓住该时机，进行大规模宣传。结合时代背景，由于新冠疫情的影响，可以发现 3 年以来的总销量都所有下滑的趋势。在此大背景下，全民经济收入受到一定影响，居家抗疫期间对烟的需求量减少，且由于交际活动的限制，疫情期间的吸烟量也减少。此外，近年来全民戒烟健体意识的不断加强，对卷烟的销量也有一定的冲击。对于此情况，可以创新营销模式，丰富服务内容，营销充分利用大数据优势，积极开展“线上+线下”相结合的形式。

最后根据卷烟价类划分来看，其标准与行业发展和经济社会发展相适应、相匹配、相吻合。在现行的标准中，在零售价 200～500 元/条即中端、中高端区间，市场集中度已经处于非常高的水平，该价区无疑是一类烟价类中市场规模最大的基础价区。在卷烟消费不断升级的大背景下，我们不妨设想，将一类烟价类划分标准整体上移，从而带动整个类别的变化，跟上行业发展和经济社会发展的步伐。理顺卷烟价格区间，为经济运行分析创造更为有利的条件。现行卷烟价类划分标准中，高端烟价格空间和梯次范围过大，如果能使各价类卷烟的价格跨度划分趋于均衡，在此前提下深入分析卷烟价格，有利于“同档同价同差率”工作的顺利推进，也可以为经济运行分析创造更为有利的条件。

为行业生产经营管理决策提供辅助。从统计学角度分析，现行卷烟价类划分标准已经不能客观反映各类卷烟的真实发展趋势，比如低端烟的销量比重和利润比重在缩小，但又不能因为其比例缩小就忽略不计，因此需要设计一套更突出可比性、更科学的新标准，以有效弥补不足，对卷烟生产经营管理决策起到更好的指导作用。

六、 模型评价与推广

6.1 多元线性回归模型

分析探究影响销量的因素时，采用了多元线性回归模型主要对市场类型和销售终端的影响进行了理论化的探究，最初是只把两个变量认作是毫不相关的两者，所得结果显示二者只可影响销量的 47.1%，最大显著性达到 70%多，进而猜想二者对销量的影响不是毫不相互影响的，故而在线性回归时加入市场类型与销售终端的乘积作为一个交互项变量。分析结果显示自变量对销量的决定程度提高至 58.7%，系数的最大显著性水平降低至 50%，表明改进后的模型更加从理论层面说明了这两个因素的影响力。故而，在分析多因素对因变量的影响大小时，线性回归时最简单的关系，但在建模时要考虑到变量不是孤立的，可能对因变量有交互的影响，从而更加有利于主成分的分析。

6.2 BP 神经网络与 RBF 神经网络模型

在预测月销量时，最开始使用的 BP 神经网络算法模型有着良好的逼近性能，且在神经元训练过程中可以自由调控训练的方向等状况，但是在预测结果的相对误差上分析还是有些不足的，在训练预测时，BP 神经网络的平均相对误差达到 18.6%。为了追求一个更“全局化”，预测效果更佳的结果，采用 RBF 神经网络模型，RBF 神经网络模型以其更加快速地逼近任意函数，以及相对没有 BP 神经网络那么局部的范围，预测相对误差缩小到 9%。故可认为 RBF 神经网络模型在某些情况下（尤其是所给样例较少，BP 神经网络的训练会因局限性受影响的情况）会有有更高效的预测效果。

参考文献

[1]赵梦娜.基于 SVM 和 BP 神经网络的量化策略研究[D].大连理工大学,2021.DOI:10.26991/d.cnki.gdllu.2021.001540.

[2]王明哲.基于 RBF 神经网络的重力固体潮信号的建模与预测[D].昆明理工大学,2019.DOI:10.27200/d.cnki.gkmlu.2019.000775.

[3]张岭,孙雨燕.基于表征客户对卷烟品牌群落的分析[J].中外企业家,2014(23):198-199.

[4]邹雯.基于改进随机森林的卷烟订购量预测的研究与应用[D].南昌大学,2020.DOI:10.27232/d.cnki.gnchu.2020.002475.

[5]李晓亮,闫晓雯,马晓敏,曹越.基于目标消费者分析的卷烟品牌培育策略研究[J].中国管理信息化,2020,23(22):142-143.

附录 A　问题二涉及图表

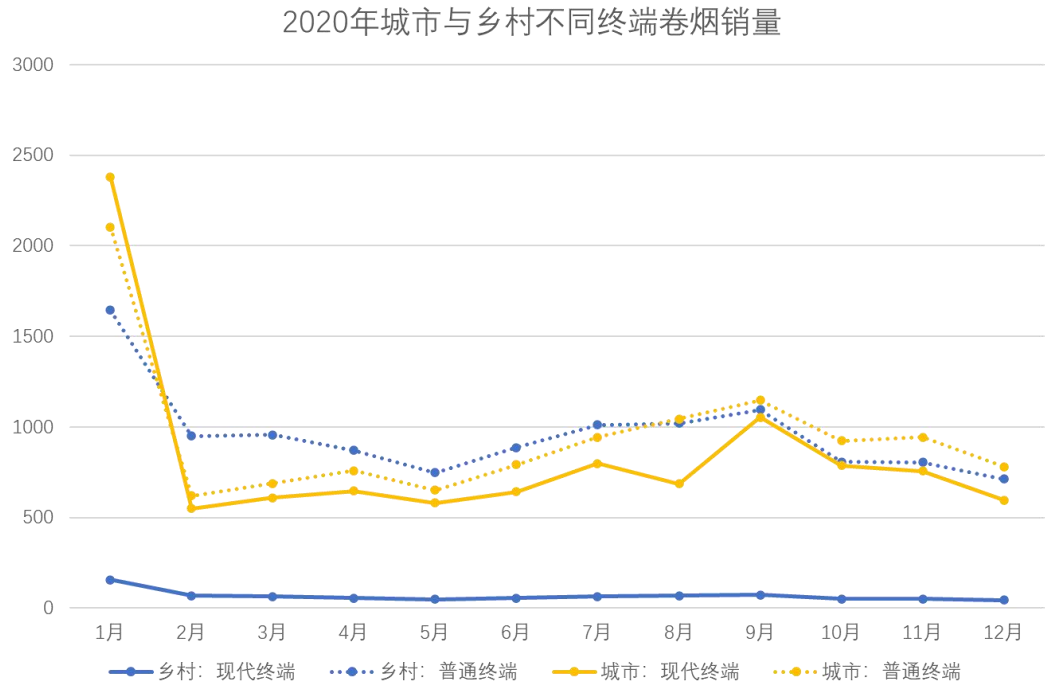


图 1　2020 年城市与乡村不同终端卷烟销量

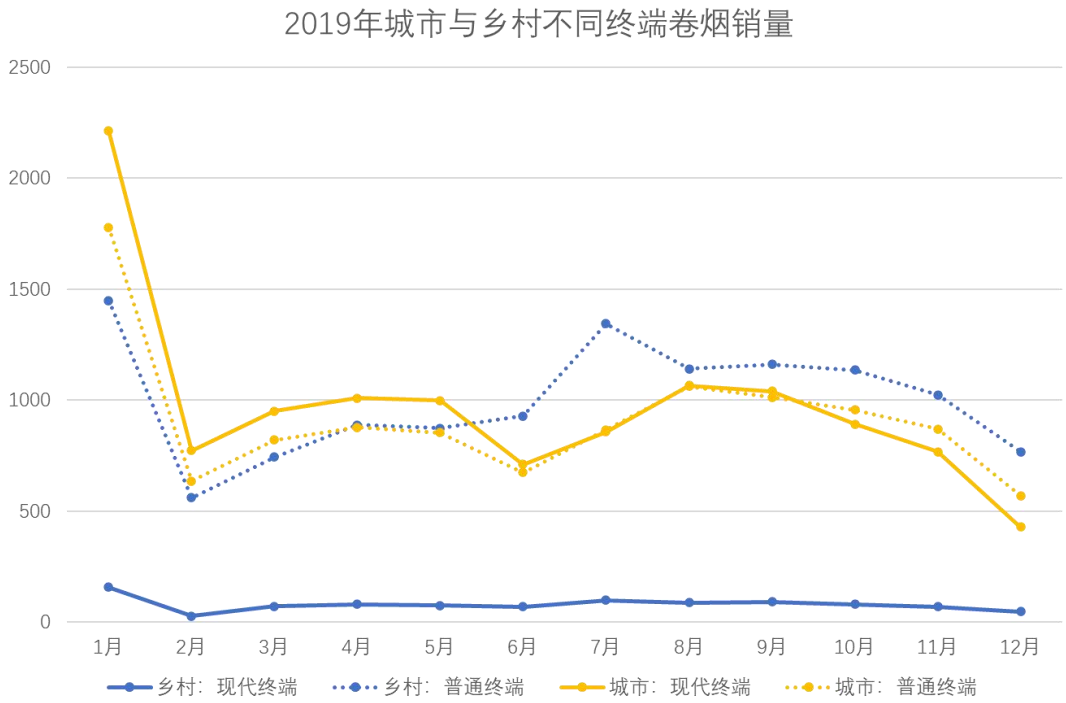


图 2　2019 年城市与乡村不同终端卷烟销量

表 1　改进模型对 2020 年卷烟销量系数的求解

	模型	未标准化系数 *B*	标准化系数 *Beta*	显著性检验值 *t*	显著性
3	常量	949.737	/	6.648	0.000
	月份	-28.760	-0.200	-1.778	0.082
	市场类型	381.130	0.384	3.413	0.001
	销售终端	-499.517	-0.504	-4.473	0.000

表 2　改进模型对 2019 年卷烟销量系数的求解

	模型	未标准化系数 *B*	标准化系数 *Beta*	显著性检验值 *t*	显著性
3	常量	911.771	/	6.491	0.000
	月份	-23.812	-0.173	-1.497	0.142
	市场类型	402.979	0.424	3.670	0.001
	销售终端	-429.105	-0.451	-3.970	0.000

附录 B　RBF 乡村城镇不同终端销量及总和预测——python 源程序

```
# -*- coding: utf0-8 -*-

import numpy as np
import matplotlib as mpl
import matplotlib.pyplot as plt
from scipy.linalg import norm, pinv
import math

np.random.seed(20)

class RBF:
    """
    RBF 神经网络
    """
    def __init__(self, input_dim, num_centers, out_dim):
        """
        本身变量 输入维度 中间层数量 输出维度
        """
        self.input_dim = input_dim #多少个输入
        self.num_centers = num_centers #中间神经元数
```

```
        self.out_dim = out_dim #多少个输出
        self.beta = 8 #扩展常数
        self.centers = [np.random.uniform(-1, 1, input_dim) for i in range(num_cen
ters)] #中间点
        self.w = np.random.random((self.num_centers, self.out_dim)) #中间层与输出之
间的关系 是一个输入 size*输出 size 的矩阵

    def _basisfunc(self, c, d):
        return np.exp(-self.beta * norm(c - d) ** 2)

    def _calcAct(self, X): #激活函数
        G = np.zeros((X.shape[0], self.num_centers), dtype = np.float)
        for  ci, c in enumerate(self.centers):
            for xi, x in enumerate(X):
                G[xi, ci] = self._basisfunc(c, x) #中心点 样本
        return G


    def train(self, X, Y): #训练
        rnd_idx = np.random.permutation(X.shape[0])[:self.num_centers] #随机选择中
心点
        self.centers = [X[i, :] for i in rnd_idx]

        #计算 RBF 的激活函数的值
        G = self._calcAct(X) #_calcAct 是激活函数

        self.W = np.dot(pinv(G), Y) #点乘


    def predict(self, X): #预测
        G = self._calcAct(X)
        Y = np.dot(G, self.W)
        return Y

def read_data(file):
    #从文件中读取数据并返回数组
    with open(f'{file}', mode='r', encoding='utf-8') as f:
        name = f.readline()
        X = []
        while f:
            s = f.readline().strip()
            if not s:
                break
            X.append(float(s))
```

```
    return name, X

def plusAyear(old):
    new = old.replace('2019-2021', '2020-2022')
    return new

 #构造数据 画图
#file = input() #数据文件
n = 36 #总的数据个数
nn = 48
title_group = ['乡村城镇及不同中端的销量的RBF预测曲线', '销量总和的RBF预测曲线'] #图的名称
x = np.linspace(1, n, n).reshape(n, 1) #横坐标
xx = np.linspace(1, nn, nn).reshape(nn, 1) #完整横坐标
num_centers = 200 #越大 学习能力越强 模拟效果越好
rbf = RBF(1, num_centers, 1)

file1 = '1.txt' #乡 现 2019-2021
file2 = '2.txt' #乡 普 2019-2021
file3 = '3.txt' #城 现 2019-2021
file4 = '4.txt' #城 普 2019-2021
file5 = '5.txt' #乡 现 2020-2021
file6 = '6.txt' #乡 普 2020-2021
file7 = '7.txt' #城 现 2020-2021
file8 = '8.txt' #城 普 2020-2021
file9 = '9.txt' #2019-2021总和
file10 = '10.txt' #2020-2021总和

#print(f'\n---*** {i} {graphname} ***--\n')

#源数据
name1, y1 = read_data(file1)
name2, y2 = read_data(file2)
name3, y3 = read_data(file3)
name4, y4 = read_data(file4)
name5, y5 = read_data(file5)
name6, y6 = read_data(file6)
name7, y7 = read_data(file7)
name8, y8 = read_data(file8)
name9, y9 = read_data(file9)
name10, y10 = read_data(file10)

#预测
rbf.train(x, y1)
```

```
z1 = rbf.predict(x) #乡 现 2020-2022
rbf.train(x, y2)
z2 = rbf.predict(x) #乡 普 2020-2022
rbf.train(x, y3)
z3 = rbf.predict(x) #城 现 2020-2022
rbf.train(x, y4)
z4 = rbf.predict(x) #城 普 2020-2022
rbf.train(x, y9)
z9 = rbf.predict(x) #合计 2020-2022

#绘图
plt.rcParams['font.sans-serif']=['SIMSUN'] #用来正常显示中文标签
plt.xlim(1, n+1)
plt.xticks(np.arange(0, nn+1, 2))

plt.plot(xx[:36], y1, 'k-', label=f'{name1}实际值', color = 'green', alpha = 0.4
5, linewidth = 2.2)
plt.plot(xx[12:], z1, 'k-', label=f'{plusAyear(name1)}预测值', color = 'green', li
newidth = 0.7, linestyle = '--')
plt.plot(xx[:36], y2, 'k-', label=f'{name2}实际值', color = 'blue', alpha = 0.4
5, linewidth = 2.2)
plt.plot(xx[12:], z2, 'k-', label=f'{plusAyear(name2)}预测值', color = 'blue', lin
ewidth = 0.7, linestyle = '--')
plt.plot(xx[:36], y3, 'k-', label=f'{name3}实际值', color = 'purple', alpha = 0.4
5, linewidth = 2.2)
plt.plot(xx[12:], z3, 'k-', label=f'{plusAyear(name3)}预测值', color = 'purple', l
inewidth = 0.7, linestyle = '--')
plt.plot(xx[:36], y4, 'k-', label=f'{name4}实际值', color = 'orange', alpha = 0.4
5, linewidth = 2.2)
plt.plot(xx[12:], z4, 'k-', label=f'{plusAyear(name4)}预测值', color = 'orange', l
inewidth = 0.7, linestyle = '--')

plt.title(title_group[0], fontsize = 15, fontweight = 'semibold')
plt.legend(loc = 'upper left', ncol = 1, bbox_to_anchor=(1.02, 1), borderaxespa
d = 0)
plt.savefig(f'{title_group[0]}_centers={num_centers}.svg',bbox_inches= 'tight')
plt.cla()

plt.xticks(np.arange(0, nn+1, 2))
plt.plot(xx[:36], y9, 'k-', label=f'{name9}实际值', color = 'red', alpha = 0.45, l
inewidth = 2.2)
plt.plot(xx[12:], z9, 'k-', label=f'{plusAyear(name9)}预测值', color = 'red', line
width = 0.7, linestyle = '--')
```

```
plt.title(title_group[1], fontsize = 15, fontweight = 'semibold')
plt.legend(loc = 'upper left', bbox_to_anchor=(1.02, 1), borderaxespad = 0)
plt.savefig(f'{title_group[1]}_centers={num_centers}.svg',bbox_inches='tight')

#计算相对误差
ex1 = []
ex2 = []
ex3 = []
ex4 = []
ex9 = []
for i in range(24):
    e1 = math.fabs(y5[i]-z1[i])/y5[i]
    ex1.append(e1)
    e2 = math.fabs(y6[i]-z2[i])/y6[i]
    ex2.append(e2)
    e3 = math.fabs(y7[i]-z3[i])/y7[i]
    ex3.append(e3)
    e4 = math.fabs(y8[i]-z4[i])/y8[i]
    ex4.append(e4)
    e9 = math.fabs(y10[i]-z9[i])/y10[i]
    ex9.append(e9)

def print_ex(pre, ex, name):
    print(f"\n---------------------\n{name}\n")
    print('预测值：\n')
    for i in pre:
        print(i, end='\n')
    print('\n')
    print('相对误差：\n')
    for i in ex:
        print(i, end='\n')
    print('\n')
    print('相对误差上下限：',min(ex), max(ex),'\n')
    print('相对误差平均值：',sum(ex)/n,'\n')
    print('\n----------------------\n')

print_ex(z1, ex1, '乡村现代 2020-2022 销量：')
print_ex(z2, ex2, '乡村普通 2020-2022 销量：')
print_ex(z3, ex3, '城镇现代 2020-2022 销量：')
print_ex(z4, ex4, '城镇普通 2020-2022 销量：')
print_ex(z9, ex9, '合计 2020-2022 销量：')
```

附录 C　客户 0811170 的 RBF 模型预测——python 源程序

```
# -*- coding: utf-8 -*-

import numpy as np
import matplotlib as mpl
import matplotlib.pyplot as plt
from scipy.linalg import norm, pinv
import math

np.random.seed(20)

class RBF:
    """
    RBF 神经网络
    """
    def __init__(self, input_dim, num_centers, out_dim):
        """
        本身变量 输入维度 中间层数量 输出维度
        """
        self.input_dim = input_dim #多少个输入
        self.num_centers = num_centers #中间神经元数
        self.out_dim = out_dim #多少个输出
        self.beta = 8 #扩展常数
        self.centers = [np.random.uniform(-1, 1, input_dim) for i in range(num_cen
ters)] #中间点
        self.w = np.random.random((self.num_centers, self.out_dim)) #中间层与输出之
间的关系 是一个输入 size*输出 size 的矩阵

    def _basisfunc(self, c, d):
        return np.exp(-self.beta * norm(c - d) ** 2)

    def _calcAct(self, X): #激活函数
        G = np.zeros((X.shape[0], self.num_centers), dtype = np.float)
        for  ci, c in enumerate(self.centers):
            for xi, x in enumerate(X):
                G[xi, ci] = self._basisfunc(c, x) #中心点 样本
        return G


    def train(self, X, Y): #训练
        rnd_idx = np.random.permutation(X.shape[0])[:self.num_centers] #随机选择中
心点
        self.centers = [X[i, :] for i in rnd_idx]
```

```
        #计算 RBF 的激活函数的值
        G = self._calcAct(X) #_calcAct 是激活函数

        self.W = np.dot(pinv(G), Y) #点乘


    def predict(self, X): #预测
        G = self._calcAct(X)
        Y = np.dot(G, self.W)
        return Y

def read_data(file):
    #从文件中读取数据并返回数组
    with open(f'{file}', mode='r', encoding='utf-8') as f:
        s = f.readline().split()
        X = [float(x) for x in s]
    return X

 #构造数据 画图
#file = input() #数据文件
n = 24 #一年的数据

title_group = ['客户 0811170 的 RBF 预测曲线'] #图的名称
x = np.linspace(1, n, n).reshape(n, 1) #横坐标
num_centers = 100 #越大 学习能力越强 模拟效果越好
rbf = RBF(1, num_centers, 1)

file1 = 'kh2020-2021.txt'

#print(f'\n---*** {i} {graphname} ***--\n')

#源数据
y1 = read_data(file1) #2020-2021

#预测
rbf.train(x, y1)
z1 = rbf.predict(x)

#绘图
plt.rcParams['font.sans-serif']=['SIMSUN'] #用来正常显示中文标签
plt.xlim(1, n+1)
plt.xticks(np.arange(0, n+1, 1))
```

```
plt.plot(x, z1, 'k-', label='2021-2022 预测值', color = 'green', alpha = 0.45, line
width = 2.2)

plt.title(title_group[0], fontsize = 15, fontweight = 'semibold')
plt.legend(loc = 'upper left', ncol = 1, bbox_to_anchor=(1.02, 1), borderaxespa
d = 0)
plt.savefig(f'{title_group[0]}_centers={num_centers}.svg',bbox_inches= 'tight')
```

附录 D　各价位卷烟销量前三的销量 RBF 预测——python 源程序

```
# -*- coding: utf-8 -*-

import numpy as np
import matplotlib as mpl
import matplotlib.pyplot as plt
from scipy.linalg import norm, pinv

np.random.seed(20)

class RBF:
    """
    RBF 神经网络
    """
    def __init__(self, input_dim, num_centers, out_dim):
        """
        本身变量 输入维度 中间层数量 输出维度
        """
        self.input_dim = input_dim #多少个输入
        self.num_centers = num_centers #中间神经元数
        self.out_dim = out_dim #多少个输出
        self.beta = 8 #扩展常数
        self.centers = [np.random.uniform(-1, 1, input_dim) for i in range(num_cen
ters)] #中间点
        self.w = np.random.random((self.num_centers, self.out_dim)) #中间层与输出之
间的关系 是一个输入 size*输出 size 的矩阵

    def _basisfunc(self, c, d):
        return np.exp(-self.beta * norm(c - d) ** 2)

    def _calcAct(self, X): #激活函数
        G = np.zeros((X.shape[0], self.num_centers), dtype = np.float)
        for  ci, c in enumerate(self.centers):
            for xi, x in enumerate(X):
                G[xi, ci] = self._basisfunc(c, x) #中心点 样本
```

```
        return G


    def train(self, X, Y): #训练
        rnd_idx = np.random.permutation(X.shape[0])[:self.num_centers] #随机选择中心点
        self.centers = [X[i, :] for i in rnd_idx]

        #计算 RBF 的激活函数的值
        G = self._calcAct(X) #_calcAct 是激活函数

        self.W = np.dot(pinv(G), Y) #点乘


    def predict(self, X): #预测
        G = self._calcAct(X)
        Y = np.dot(G, self.W)
        return Y

def read_data(file):
    #从文件中读取数据并返回数组
    with open(f'{file}', mode='r', encoding='utf-8') as f:
        name = f.readline()
        s = f.readline().split()
        X = [float(xi) for xi in s]
    return name, X

 #构造数据 画图
#file = input() #数据文件
n = 12 #一年的数据

title_group = ['高端主销商品的 RBF 预测曲线', '中高端主销商品的 RBF 预测曲线',
         '中端主销商品的 RBF 预测曲线', '低端主销商品的 RBF 预测曲线', '各价位段主销商品的 RBF 预测曲线']
x = np.linspace(1, n, n).reshape(n, 1) #横坐标
num_centers = 10 #越大 学习能力越强 模拟效果越好
rbf = RBF(1, num_centers, 1)
j = 0
for i in range(1, 13, 3):

    file1 = str(i)+'.txt'
    file2 = str(i+1)+'.txt'
    file3 = str(i+2)+'.txt'
    graphname = title_group[j] #图的名称
```

```
    j+=1
    print(f'\n---*** {i} {graphname} ***--\n')

    #源数据
    name1, y1 = read_data(file1)
    name2, y2 = read_data(file2)
    name3, y3 = read_data(file3)

    #预测
    rbf.train(x, y1)
    z1 = rbf.predict(x)
    rbf.train(x, y2)
    z2 = rbf.predict(x)
    rbf.train(x, y3)
    z3 = rbf.predict(x)

    #绘图
    plt.rcParams['font.sans-serif']=['SIMSUN'] #用来正常显示中文标签
    plt.xlim(1, n+1)
    plt.xticks(np.arange(0, n+1, 1))

    plt.plot(x, y1, 'k-', label=f'{name1}源数据', color = 'green', alpha = 0.5, li
newidth = 2)
    plt.plot(x, z1, 'k-', label=f'{name1}预测值', color = 'green', linewidth = 0.
7, linestyle = '--')
    plt.plot(x, y2, 'k-', label=f'{name2}源数据', color = 'blue', alpha = 0.5, lin
ewidth = 2)
    plt.plot(x, z2, 'k-', label=f'{name2}预测值', color = 'blue', linewidth = 0.
7, linestyle = '--')
    plt.plot(x, y3, 'k-', label=f'{name3}源数据', color = 'purple', alpha = 0.5, l
inewidth = 2)
    plt.plot(x, z3, 'k-', label=f'{name3}预测值', color = 'purple', linewidth = 0.
7, linestyle = '--')

    plt.title(graphname, fontsize = 15, fontweight = 'semibold')
    plt.legend(loc = 'upper center')
    plt.savefig(f'{graphname}.svg')
    plt.cla()
```