基于随机森林回归的制丝过程参数影响权重分析

刘继辉[1]，许磊[2]，马晓龙[1]，李达[1]，林鸿佳[1]，杨洋[1]，杨晶津[1]，李兴绪[2]，王慧[*1]

1. 红云红河烟草(集团)有限责任公司，昆明市红锦路 367 号　650231

2. 云南财经大学云南省经济社会大数据研究院，昆明市龙泉路 237 号　650221

摘要：为提高制丝工艺质量评价中参数赋权分析的科学性和客观性，选取“云烟”某一类规格卷烟制丝过程全批次数据的稳态数据样本，通过 Pearson 相关性矩阵筛选各工序出口含水率的解释变量，然后利用随机森林回归进行建模分析，采用拟合优度和 5 折交叉验证的测试集标准化均方误差分别验证模型的拟合效果和外推预测性能，最终根据 OOB 均方误差的平均递减值进行解释变量影响权重的测度和关键参数的筛选。结果表明：① 综合 Pearson 相关性矩阵和设备控制原理，筛选得到 37 个解释变量；②制丝过程 5 个工序随机森林回归模型的拟合优度均大于 0.9、五折交叉验证测试集的标准化均方误差均小于 1，表明模型的拟合效果和外推预测性能较好；③根据解释变量影响权重的测度分析，筛选得到 18 个关键参数；④本研究基于全样本数据建立的制丝过程关键参数筛选和赋权方法，可为制丝关键质量特性精准控制和工艺质量评价提供参考。

关键词：制丝过程；稳态数据样本；Pearson 相关性；随机森林回归；权重分析

中图分类号：TS452　　**文献标志码：**A

Influence Weight Analysis of Parameter in Primary Processing based on Random Forests for Regression

LIU Jihui[1], XU Lei[2], MA Xiaolong[1], LI Da[1], Lin Hongjia[1], YAN Yang[1], YAN Jingjin[1], LI Xingxu[2], and WANG Hui[*1]

1. HongyunHonghe Tobacco (Group) Co., Ltd., Kunming 650231, China

2. Yunnan University of Finance and EconomicsYunnan Economy &Society Bigdata Research Institute, Kunming 650221, China

Abstract: In order to improve the scientificity and objectivity of parameter weighting analysis of tobacco primary quality evaluation in primary processing, Pearson correlation matrix was used to screen explanatory variables of moisture content in output for process step respectively based on the whole batch steady state samples of the primary processing of a certain kind of “YunYan”, after that models were established by random forest for regression, then goodness of fit and NMSE of test set by 5-fold cross-validation were used to verify the fitting and extrapolation forecasting of models. Finally, the influence weight of explanatory variables were calculated and key parameters were screened according to average decline value of OOB mean square error. The results showed that: 1) Thirty-seven explanatory variables of moisture content in output for process step were screened by Pearson correlation matrix and

收稿日期：2016–05–06　　修回日期：2016–07–06
基金项目：红云红河烟草（集团）有限责任公司科技项目“基于云平台的工艺质量智能管控研究及应用”（HYHH2016GY04）
作者简介：刘继辉（1983—），硕士，工程师，主要从事卷烟工艺研究。 E-mail：bobbyrna@sina.cn ；*通讯作者：王慧，E-mail：gywh01@163.com
引文格式：刘继辉，许磊，马晓龙，等. 基于随机森林回归的制丝过程参数影响权重分析[J].烟草科技，2016，49（）：（LIU Jihui, XU Lei, MA Xiaolong, et al. Influence weight Analysis of parameter in primary processing based on random forests for regression [J].Tobacco Science & Technology, 2016, 49（）:）
DOI：10.16135/j.issn1002–0861.2016.0229

control principle of equipment. 2) Random forest regression model of five process steps in primary processing has good performance in fitting and extrapolation forecasting. 3) Eighteen key parameters were screened according to the influence weight of explanatory variables. 4) A method of weighting and screening key parameters based on full sample size data in primary processing was established, which will offer reference for precise control of critical to quality and evaluation of tobacco primary quality.**Keywords:** Primary processing; The steady state sample; Pearson correlation; Random forests for regression; Weight analysis; Key parameters

制丝过程是凸显卷烟感官风格、稳定产品质量、降低原料消耗的重要环节。卷烟产品多点加工布局下，地域气候、工艺布局、装备水平差异较大，如何建立一套科学的制丝过程工艺质量评价方法，确保产品质量稳定一致显得尤为重要。制丝加工设备参数繁多，且内部存在大量交互效应；此外制丝加工流程较长，上游工序的质量指标都直接或间接影响下游工序乃至最终产品的质量控制，所以关键参数的筛选及其权重的测度是建立科学评价方法的重要环节。目前，参数权重的测度方法有：①主观赋权法。主要根据经验判断以及最终目标来设定权重，能较好地反映决策者对评价目标的主观意向和偏好，但无法克服主观随意性较大等问题，包括专家评分法、层次分析法等。如刘晓龙等[1]运用专家评分法分析卷烟制造过程特性参数与需求参数之间的关联程度；史艳霞等[2]通过因果矩阵打分表筛选产品的关键质量特性并确定指标的重要程度；张新锋等[3]运用网络分析法 ANP 和比较判断矩阵计算制丝关键工序对制丝质量的影响权重。②客观赋权法。应用统计分析方法充分挖掘样本中蕴涵的数据信息，可有效降低主观因素的影响，包括主成分分析法、最大熵技术法等。如张慧筠等[4]运用主成分分析法筛选出化学成分因子、感官质量因子、烟气成分因子共 3 个主成分对 15 个牌号的卷烟质量进行评价；张天栋等[5]运用熵值法对消费者可分辨的香气特征等 7 项感官指标赋权。③主客观综合集成赋权法，即将主观赋权法、客观赋权法结合使用。如刘馨[6]选取生产组织、工艺质量以及物耗成本 3 个方面的指标进行精益生产管理评价赋权研究。客观赋权法中，熵值赋权法是根据解释变量所提供信息量的大小来确定权重，未将目标变量纳入建模分析；主成分赋权法是以多元回归分析为基础，考虑了解释变量对目标变量的影响，但与机器学习组合算法相比较，多元回归模型预测的稳健性不高[7]。随机森林[8]（Random Forest，RF）是由 Breiman 于 2001 年提出的一种以决策树（CART）为基础的机器学习组合算法，相比上述客观赋权方法，具有更好的噪声容忍度和外推预测性[9]。近年来，由于处理

高频实时数据的巨大优势，随机森林已在经济[10]、金融[11]、生物[12]、烟草[13]等领域得到广泛应用。本研究中选择随机森林开展制丝过程关键参数的筛选及其权重的测度研究。

目前，随着制丝工序工艺装备和信息技术的成熟应用，制丝过程关键质量特性的过程控制水平得到了明显提升，但是地域、气候、环境温湿度和蒸汽质量的差异均对含水率控制影响较大。所以选取制造执行系统采集的正常生产批次数据，在工艺参数严格执行生产技术标准、特殊过程（加香、加料和掺配）复合工艺要求、过程控制能力测评显示制丝过程处于稳定受控状态的前提下，以各工序出口含水率作为研究变量，通过参数筛选及赋权研究，为制丝工序物料含水率的精准控制提供理论依据。

1 数据处理

1.1 数据样本

2015 年 1 月 1 日至 12 月 30 日昆明卷烟厂“云烟”某一类规格卷烟的全批次制丝过程数据，共计 219 批，包含松散回潮、加料、叶丝干燥及冷却、加香等工序的工艺参数、设备参数和出口含水率，数据采集频次为每 6 s 采集 1 次。

1.2 数据预处理

对制造执行系统（Manufacturing Execution System，简称 MES）采集的原始数据进行稳态数据筛选，流程见图 1。首先剔除停机断料批次数据，停机断料批次的判定规则为生产过程中某一工序入口流量降至 0 kg/h 且持续时间超过 90 s 的批次。然后参考表 1 的截取规则筛选稳态数据，制丝过程中部分参数的设定值为一常量的数据，如加料比例、切丝宽度等，在筛选过程中一并剔除。此外，制丝滚筒类设备混合加工的特性造成参数和出口含水率的实时数据无法一一对应，根据制丝正常生产过程标识物（纯白卷烟纸）工序停留时间的实测值，按 8 min 的时间间隔进行数据分组并计算稳态数据的均值，最终形成稳态数据样本。

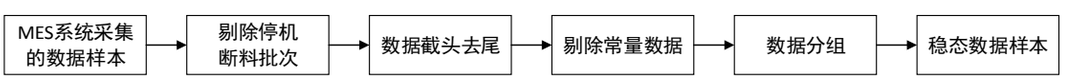


图 1 稳态数据筛选流程

Fig.1 The steady state data screening process

表 1 稳态数据截取规则①

Tab.1 Steady state data capture rules

关键质量特性		开始条件	结束条件
物料流量 Q		Q>100 kg/h，且时间延时 3 min	Q<100 kg/h，且时间前移 3 min
出口含水率 $X_{出}$	低含水率物料（≤15%）	$X_{出}$>8 %，且时间延时 3 min	$X_{出}$<8 %，且时间前移 3 min
	中高含水率物料（>15%，且≤23%）	$X_{出}$>12 %，且时间延时 3 min	$X_{出}$<12 %，且时间前移 3 min
	高含水率物料（>23%）	$X_{出}$>20 %，且时间延时 3 min	$X_{出}$<20 %，且时间前移 3 min
出口温度		与出口含水率同步	与出口含水率同步
热风温度、筒壁温度等其他参数		与工序出口含水率起始点同步	与工序入口物料流量结束点同步

注：①参考《中式卷烟制丝生产线创新效果评价测试大纲》制定。

通过对原始数据进行以上 3 个步骤的预处理后，最终形成了各工序的稳态数据样本。其中，松散回潮工序的样本量为 1 627 个，二次回潮工序的样本量为 1 796 个，加料工序的样本量为 1 794 个，叶丝干燥及冷却工序的样本量为 2 286 个，加香工序的样本量为 2 283 个。

2 研究方法

2.1 相关性分析

为研究稳态数据样本内各工序参数的交互效应，选择 Pearson 相关系数对稳态数据样本内各工序的参数进行相关性分析。利用数据样本计算相关系数 r 时具有一定的随机性，需要检验相关系数的显著性。采用 SPSS 统计软件内的 t 检验[14]推断各参数之间的 Pearson 相关系数及其检验的 P 值。针对稳态数据样本中高度显著相关（相关系数绝对值大于 0.9 且检验 P 值小于 0.05）的参数，结合设备控制原理剔除跟随变量，剩余参数为该工序统计建模的解释变量。

2.2 随机森林回归分析

2.2.1 随机森林简介

随机森林有回归和分类两种方法，当研究变量为连续变量时，采用随机森林回归（Random Forest for Regression）进行分析；当研究变量为分类变量时，采用随机森林分类（Random Forest for Classification）进行分析。本研究中的稳态数据样本为连续变量，属于回归方法的范畴。随机森林回归建模的主要思路是，首先由原始数据集 D 生成随机向量序列$\theta_i(i=1,2,\ldots k)$，然后采用 Bootstrap 从 D 中有放回地随机抽取 k 个子样本集，记为$D_i(i=1,2,\ldots k)$；其次，对每个子样本集D_i分别构建制丝过程该工序出口含水率的决策树模型$\{h(X,\theta_i)\}$，并假定子样本集$\{\theta_k\}$独立同分布；最后，由多个决策树组合$\{h_1(X),h_2(X),\ldots,h_k(X)\}$构成随机森林回归模型。模型的预测结果是以上 k 个决策树$\{h(X,\theta_i)\}$回归结果的平均值。其数学定义[15]如下所示：

$$H(x)=\frac{1}{k}\sum_{i=1}^{k}h_i(x) \qquad (1)$$

其中：$H(x)$表示随机森林回归模型的预测值，$h_i(x)$表示第 i 个决策树模型。

2.2.2 OOB 估计及重要性度量

原始数据集 D 采用 Bootstrap 抽样后每个样本未被抽中的概率为$(1-1/N)^N$，其中 N 为原始数据 D 的样本量。当样本量 N 较大时，$(1-1/N)^N$将收敛于$1/\mathrm{e}$，约为 0.368，表明原始样本集中约有 36.8%的“袋外数据”（Out of Bag，OOB）可能不在子样本集中，其可作为评价随机森林及决策树预测性能的测试数据集。由袋外数据进行评价的方法称为 OOB 估计（Out of Bag Estimation）[16]，当随机森林中的决策树足够多时，OOB 估计具有无偏性[13]。

当模型中某个解释变量被随机改变（噪声扰动）时，可将改变前后对模型的影响程度作为度量该变量的相对重要性。在随机森林分类的分析中，通常以 OOB 估计所得的预测准确率作为相对重要性的评价指标，通过推断出原始数据与加入噪声扰动后的 OOB 准确率之差度量变量的重要性[13]。OOB 准确率之差越大，该变量的相对重要性越高[17]。而对于随机森林回归分析，亦可采用 OOB 估计的均方误差平均递减值来评价解释变量对回归模型的重要性程度，均方误差（Mean Squares Error，MSE）的数学定义如下：

$$\mathrm{MSE}=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \qquad (2)$$

其中：$\hat{y}_i$表示第 i 个观测值y_i的预测值。

2.2.3 模型评价

建模分析中常用拟合优度R^2来衡量模型拟合的优劣。为了更客观地评价模型的预测性能，通常采用交叉验证（Cross Validation）[18-19]估计模型的预测误差，即把一部分数据作为训练集来构建模型，另一部分作为测试集来推断模型预测的误差。本研究中采用五折交叉验证方法来评价随机森林回归分析模型预测结果的可靠性，评价指标是标准化均方误差（Normalized Mean Squares Error，NMSE），其数学定义为：

$$\mathrm{NMSE}=\frac{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2}=\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2} \qquad (3)$$

其中：$\hat{y}_i$表示第 i 个观测值y_i的预测值，$\overline{y}$表示样本平均值。

如果 NMSE≥1，表明直接用均值预测的效果要优于模型预测，说明所拟合的回归模型不具有预测性。对于训练集而言，NMSE 等于$1-R^2$；对于测试集而言，NMSE 与测试集的R^2没有直接关系，NMSE 越小，说明模型外推预测性能越好。[9]

2.3 参数权重的测度

参照综合评价中的回归分析法[20]，将多元回归方程中的标准化回归系数作为解释变量对研究变量影响程度的度量，然后通过归一化处理得到相应解释变量的影响权重。因此，本研究中运用随机森林进行赋权[21]，将度量参数x_i（i=1,2,…m）相对重要性的均方误差（MSE）平均递减值定义为g_i，那么将g_i的绝对值归一化后，可得参数x_i的影响权重为：

$$w_i=\frac{|g_i|}{\sum_{i=1}^{m}|g_i|} \qquad (4)$$

由公式（4）推断出各工序参数对出口含水率的影响权重，按照从大到小进行排序，将影响权重累积达到 80 %以上的参数定义为关键参数。

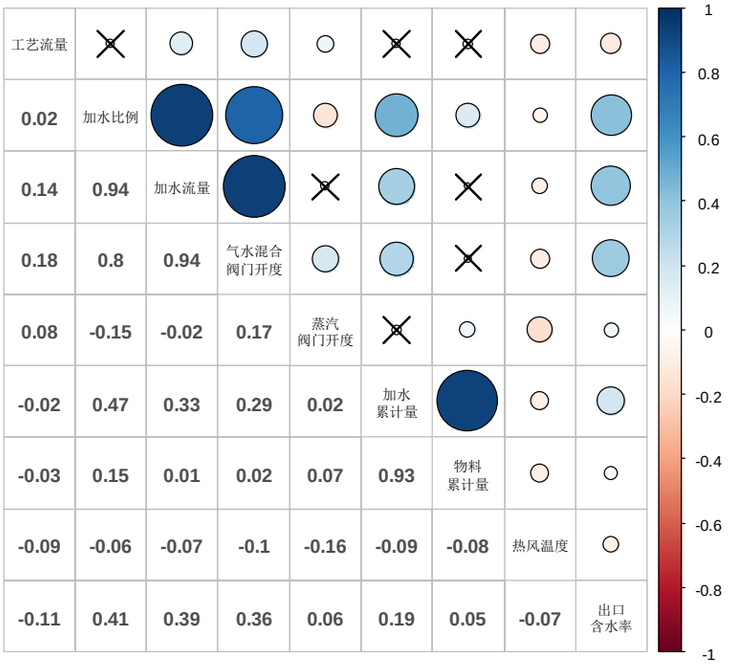
3 结果与分析

3.1 松散回潮工序参数影响权重分析

3.1.1 稳态数据的相关性

松散回潮工序稳态数据的相关性分析见图 2。出口含水率与各参数的相关性 t 检验的 P 值都小于 0.05，说明在 5 %的显著性水平下，各参数与出口含水率均存在显著相关关系。此

外，加水比例与加水流量、加水流量与气水混合阀门开度、加水累计量与物料累计量 3 组参数高度显著相关。综合稳态数据的相关性分析结论及松散回潮设备的控制原理，剔除加水流量和加水累计量。最终，松散回潮工序统计建模的解释变量确定为：工艺流量、加水比例、蒸汽阀门开度、物料累计量、热风温度、气水混合阀门开度。



注：矩阵图左下数字部分表示参数间的相关系数值；右上部分图示表示参数间的相关性及 *t* 检验的 *P* 值，其中：“×”表示 *P* 值大于 0.05，“〇”的形状和颜色表示参数间相关性的大小。下同。

图 2 松散回潮工序参数的相关系数矩阵图

Fig.2 Parameter correlation coefficient matrix in process of loosening and conditioning

3.1.2 随机森林回归分析

运用随机森林回归模型对松散回潮工序的参数进行统计建模，模型的拟合优度为 0.90，表明该模型拟合效果较好。采用公式（3）计算五折交叉验证的测试集 NMSE 为 0.51，说明该模型外推预测性能较好。

松散回潮 6 个解释变量对出口含水率的影响程度见表 2。由表 2 可以看出，松散回潮工序的关键参数是加水比例、物料累计量和气水混合阀门开度，影响权重分别是 33.74%、31.31%和 16.29%。

表 2 松散回潮工序解释变量的 MSE 平均递减值及影响权重

Tab.2 The average decline value of MES and influence of variable weights in process of loosening and conditioning

变量	MSE 平均递减值①	权重/%
加水比例	0.152 75	33.74
物料累计量	0.141 71	31.31
气水混合阀门开度	0.073 73	16.29
蒸汽阀门开度	0.036 75	8.12
工艺流量	0.029 47	6.51
热风温度	0.018 25	4.03

注：①MSE 平均递减值指以该变量为解释变量所造成的均方误差的平均递减量；数值越大，说明变量越重要[9]。下同。

3.2 二次回潮工序参数影响权重分析

3.2.1 稳态数据的相关性

二次回潮工序稳态数据的相关性分析见图 3。除蒸汽阀门开度外，二次回潮工序出口含水率与其他参数的相关性 t 检验的 P 值都小于 0.05。说明在 5%的显著性水平下，除蒸汽阀门开度外，其余参数均与出口含水率存在显著相关关系。二次回潮工序统计建模的解释变量确定为：工艺流量、入口含水率、蒸汽阀门开度、物料累计量、热风温度。

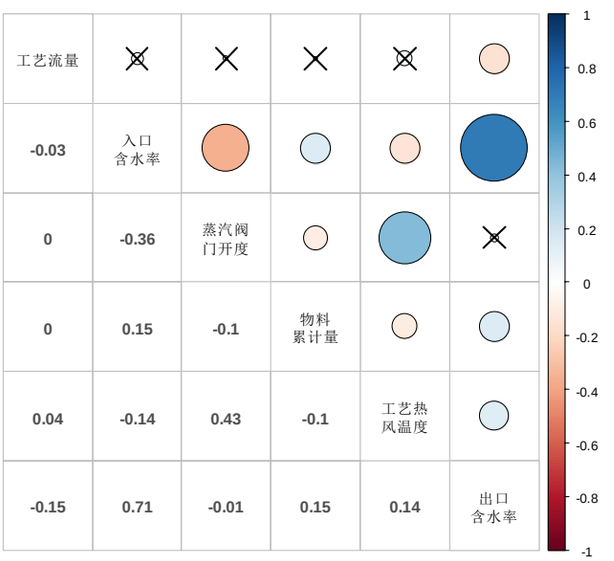


图 3 二次回潮工序参数的相关系数矩阵图

Fig. 3 Parameter correlation coefficient matrix in process of secondary loosening and conditioning

3.2.2 随机森林回归分析

运用随机森林回归模型对二次回潮工序的参数进行统计建模，模型的拟合优度为 0.93，表明该模型拟合效果较优。采用公式（3）计算五折交叉验证的测试集 NMSE 为 0.34，说明该模型外推预测性能较优。

二次回潮工序 5 个解释变量对出口含水率的的影响程度见表 3。由表 3 可以看出，二次

回潮工序的关键参数是入口含水率、热风温度和物料累计量，影响权重分别是 68.86%、10.54%和 7.40%。

表 3 二次回潮工序解释变量的 MSE 平均递减值及影响权重

Tab.3 The average decline value of MES and influence of variable weights in process of secondary loosening and conditioning

变量	MSE 平均递减值	影响权重/%
入口含水率	0.259 90	68.86
热风温度	0.039 79	10.54
物料累计量	0.027 94	7.40
工艺流量	0.025 51	6.76
蒸汽阀门开度	0.024 27	6.43

3.3 加料工序参数影响权重分析

3.3.1 稳态数据的相关性

加料工序稳态数据的相关性分析见图 4。除工艺流量、加料累计量、物料累计量和料液温度外，加料工序出口含水率与其他参数的相关性 t 检验的 P 值均小于 0.05。说明在 5%的显著性水平下，出口含水率

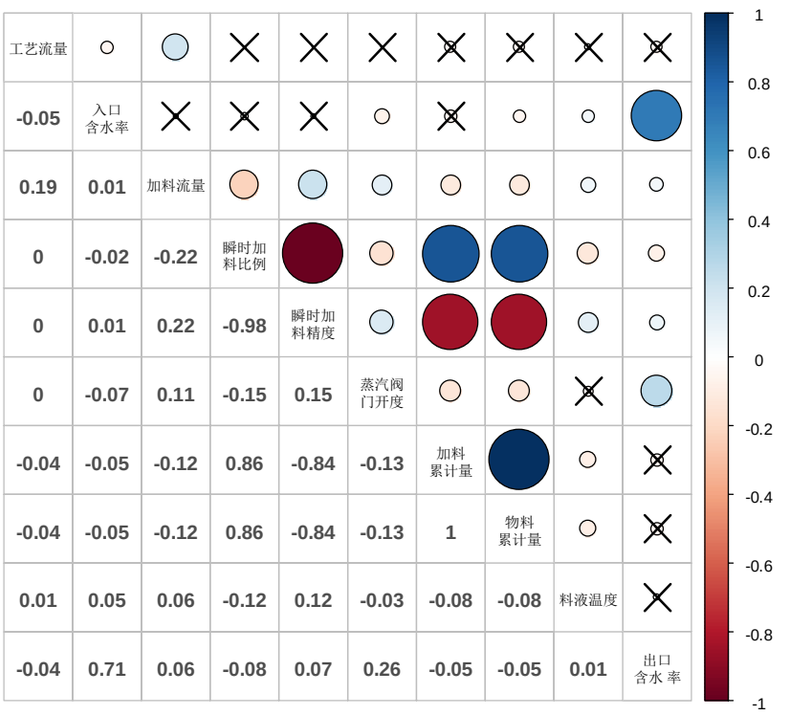


图 4 加料工序参数的相关系数矩阵图

Fig. 4 Parameter correlation coefficient matrix in process of casing

与入口含水率、加料流量、瞬时加料比例、瞬时加料精度和蒸汽阀门开度等参数存在显著相关关系。此外，瞬时加料比例与瞬时加料精度、加料累计量与物料累计量 2 组参数高度显著

相关。综合稳态数据的相关性分析结论及加料设备的控制原理，剔除瞬时加料精度和加料累计量。最终，加料工序统计建模的解释变量为：工艺流量、入口含水率、加料流量、瞬时加料比例、蒸汽阀门开度、物料累计量、料液温度。

3.3.2 随机森林回归分析

运用随机森林回归模型对加料工序的参数进行统计建模，模型的拟合优度为 0.96，表明模型拟合效果较优。采用公式（3）计算五折交叉验证的测试集 NMSE 为 0.20，表明该模型外推预测性能较优。

加料工序 7 个解释变量对出口含水率的的影响程度见表 4。由表 4 可以看出，加料工序的关键参数是入口含水率和蒸汽阀门开度，影响权重分别是 59.66%、23.73%。

表 4 加料工序解释变量的 MSE 平均递减值及影响权重

Tab.4 The average decline value of MES and influence of variable weights in process of casing

变量	MSE 平均递减值	权重/%
入口含水率	0.151 04	59.66
蒸汽阀门开度	0.060 07	23.73
料液温度	0.010 59	4.18
瞬时加料比例	0.010 44	4.12
工艺流量	0.008 48	3.35
物料累计量	0.008 48	3.35
加料流量	0.004 06	1.60

3.4 叶丝干燥及冷却工序参数影响权重分析

3.4.1 稳态数据的相关性

叶丝干燥及冷却工序稳态数据的相关性分析见图 5。除工艺流量、膨胀单元蒸汽体积流量、膨胀单元蒸汽质量流量、排潮阀门开度、Ⅰ区筒壁温度、Ⅱ区筒壁温度和热风温度外，叶丝干燥工序出口含水率与其他参数的相关性 t 检验的 P 值都小于 0.05。说明在 5%的显著性水平下，干燥出口含水率与切叶丝含水率、膨胀单元蒸汽阀门开度、物料累计量、筒壁二区阀门开度、循环风阀门开度、循环风蒸汽阀门开度、负压和工艺气速度等参数存在显著的相关关系。此外，筒壁Ⅱ区蒸汽阀门开度与Ⅰ区筒壁温度、筒壁Ⅱ区蒸汽阀门开度与Ⅱ区筒壁温度、Ⅰ区筒壁温度与Ⅱ区筒壁温度 3 组参数高度显著相关。综合稳态数据的相关性分析结论及薄板烘丝机的控制原理，剔除筒壁Ⅱ区蒸汽阀门开度。最终，叶丝干燥及冷却工序统计建模的解释变量确定为：切叶丝含水率、工艺流量、膨胀单元蒸汽阀门开度、物料累计量、膨胀单元蒸汽体积流量、膨胀单元蒸汽质量流量、排潮阀门开度、循环风阀门开度、循环风蒸汽阀门开度、负压、工艺气速度、Ⅰ区筒壁温度、Ⅱ区筒壁温度、热风温度。

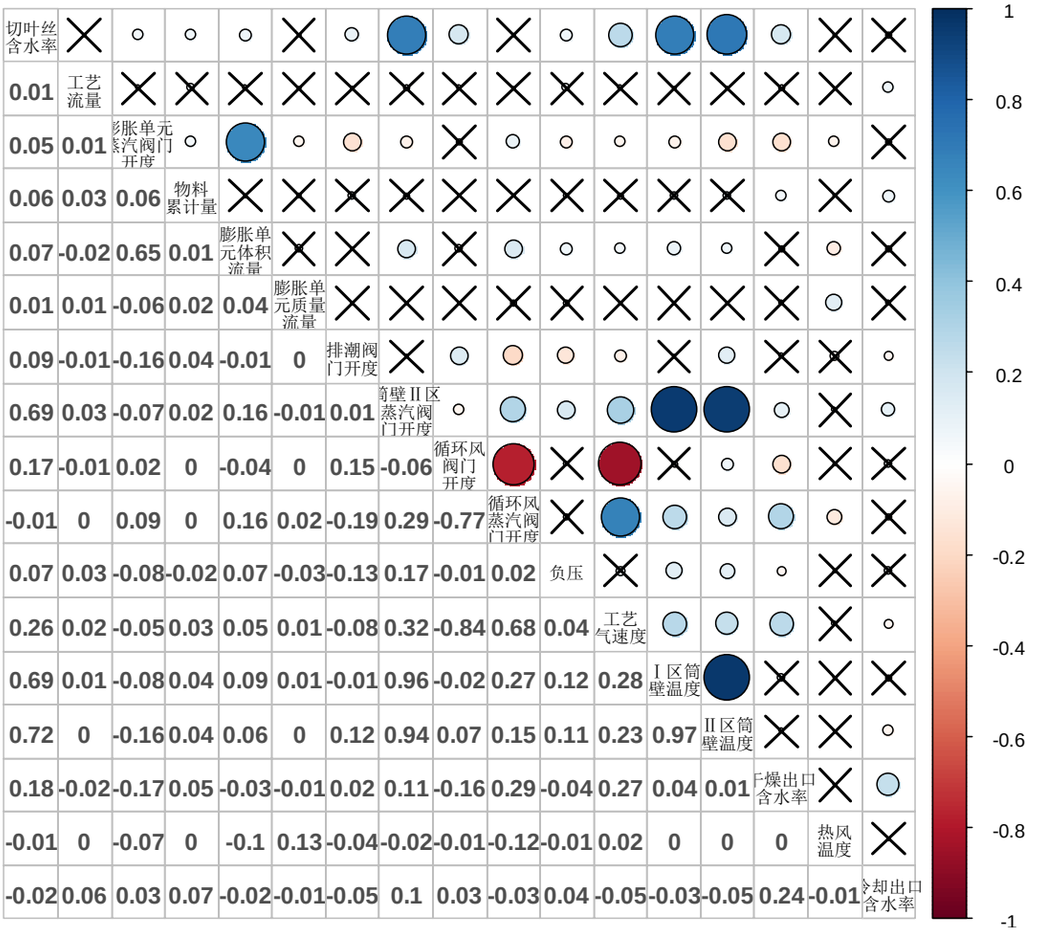


图 5 叶丝干燥及冷却工序参数的相关系数矩阵图

Fig. 5 Parameter correlation coefficient matrix in process of cut tobacco drying and cooling

3.4.2 随机森林回归分析

运用随机森林回归模型对叶丝干燥及冷却工序的参数进行统计建模，模型的拟合优度为 0.95，表明该模型拟合效果较优。采用公式（3）计算五折交叉验证的测试集 NMSE 为 0.29，说明该模型外推预测性能较优。

叶丝干燥及冷却工序 14 个解释变量对叶丝干燥出口含水率的影响程度见表 5。由表 5 可以看出，叶丝干燥及冷却工序的关键参数是循环风蒸汽阀门开度、排潮阀门开度、循环风阀门开度、Ⅰ区筒壁温度、Ⅱ区筒壁温度和工艺气速度，影响权重分别是 21.87%、18.12% 、15.35%、10.60%、8.87%和 8.85%。

表 5 叶丝干燥及冷却工序解释变量的 MSE 平均递减值及影响权重

Tab.5 The average decline value of MES and influence of variable weights in process of cut tobacco drying and cooling

变量	MSE 平均递减值	权重/%
循环风蒸汽阀门开度	0.024 64	21.87
排潮阀门开度	0.020 41	18.12
循环风阀门开度	0.017 29	15.35
Ⅰ区筒壁温度	0.011 94	10.60
Ⅱ区筒壁温度	0.009 99	8.87
工艺气速度	0.009 97	8.85
切叶丝含水率	0.008 72	7.74
膨胀单元蒸汽阀门开度	0.007 18	6.37
负压	0.000 96	0.85
膨胀单元蒸汽体积流量	0.000 69	0.61
热风温度	0.000 57	0.51
物料累计量	0.000 20	0.17
工艺流量	-5.82E-05	0.05
膨胀单元蒸汽质量流量	-3.48E-05	0.03

3.5 加香工序参数影响权重分析

3.5.1 稳态数据的相关性

加香工序稳态数据的相关性分析见图 6。除加香累计量和物料累计量外，加香工序出口含水率与其他参数的相关性 t 检验的 P 值均小于 0.05。说明在 5%的显著性水平下，加香出口含水率与冷却出口含水率、工艺流量、加香流量、瞬时加香比例和瞬时加香精度等参数存在显著相关关系。此外，加香出口含水率和冷却出口含水率、工艺流量和加香流量、加香累计量和物料累计量 3 组参数高度显著相关。综合稳态数据的相关性分析结论及加香机的控制原理，剔除加香流量和加香累计量。最终，加香工序统计建模的解释变量为：冷却出口含水率、工艺流量、瞬时加香比例、瞬时加香精度、物料累计量。

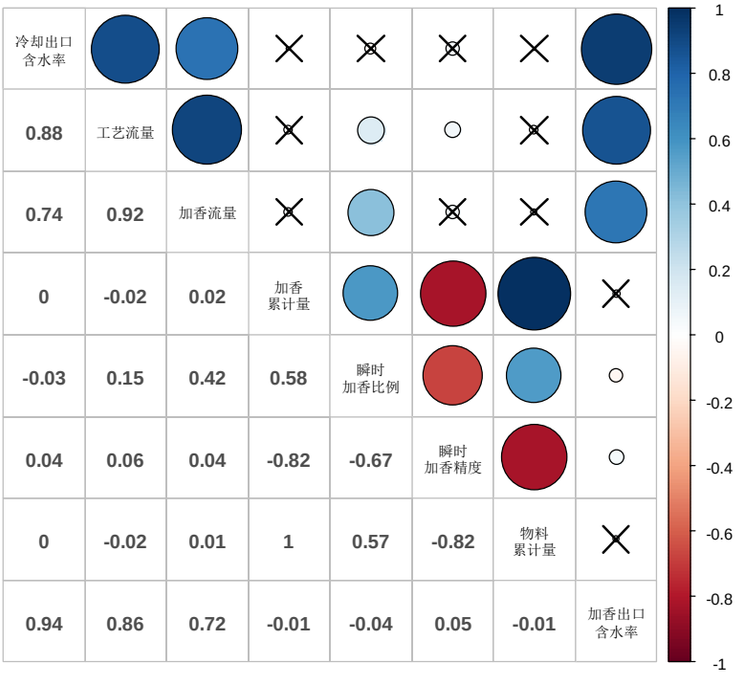


图 6 加香工序参数的相关系数矩阵图

Fig. 6 Parameter correlation coefficient matrix in process of flavoring

3.5.2 随机森林回归分析

运用随机森林回归模型对加香工序的参数进行统计建模，模型的拟合优度为 0.95，表明该模型拟合效果较优。采用公式（3）计算五折交叉验证的测试集 NMSE 为 0.51，表明该模型外推预测性能较好。

加香工序 5 个解释变量对出口含水率的的影响程度见表 6。由表 6 可以看出，加香工序的关键参数是工艺流量、冷却出口含水率、瞬时加香精度和瞬时加香比例，影响权重分别是 22.23%、22.12%、20.72%和 18.35%。

表 6 加香工序解释变量的 MSE 平均递减值及影响权重

Tab.6 The average decline value of MES and influence of variable weights in process of flavoring

变量	MSE 平均递减值	影响权重/%
工艺流量	0.255 25	22.23
冷却出口含水率	0.254 03	22.12
瞬时加香精度	0.237 86	20.72
瞬时加香比例	0.210 72	18.35
物料累计量	0.190 39	16.58

4 结论与讨论

① 综合 Pearson 相关性矩阵和设备控制原理，筛选得到 37 个解释变量；②制丝过程 5 个工序随机森林回归模型的拟合优度均大于 0.9、五折交叉验证测试集的标准化均方误差均小于 1，表明模型的拟合效果和外推预测性能较好；③根据 OOB 均方误差的平均递减值实

现了解释变量影响权重的测度分析，筛选得到 18 个影响制丝过程出口含水率的关键参数；④本研究基于全样本数据建立的制丝过程关键参数筛选和赋权方法，可为制丝关键质量特性精准控制和工艺质量评价提供参考。

参考文献

[1] 刘晓龙. 卷烟制造过程关键质量特性识别及实证研究[D]. 郑州: 郑州大学, 2013.
LIU Xiaolong. Research on key quality characteristics recognition in the cigarette machining process[D]. Zhengzhou: Zhengzhou University, 2013.

[2] 史艳霞. 基于 DMAIC 方法的泰山(AF)卷烟质量改进研究[D]. 济南: 山东大学, 2014.
SHI Yanxia. Study on the quality improvement of Taishan (AF) based on the DMAIC methods[D]. Ji'nan: Shandong University, 2014.

[3] 张新锋. 基于 ANP 的卷烟制丝质量评价方法[J]. 郑州轻工业学院学报(自然科学版), 2015, 30(3–4): 34–38.
ZHANG Xinfeng. The evaluation method of tobacco primary quality based on ANP[J]. Journal of Zhengzhou University of Light Industry (Natural Science), 2015, 30(3–4): 34–38.

[4] 张慧筠, 王玉胜, 陈玉筠. 主成分分析法在卷烟质量评价中的应用[J]. 广东化工, 2011, 38(5): 216–217, 223.
ZHANG Huijun, WANG Yusheng, CHEN Yujun. Application of principal component analysis in evaluation of cigarette quality[J]. Guangdong Chemical Industry, 2011, 38(5): 216–217, 223.

[5] 张天栋, 杨建云, 朱东来, 等. 熵值法在卷烟消费者感官质量评价中的应用[J]. 西南农业学报, 2014, 27(2): 823–828.
ZHANG Tiandong, YANG Jianyun, ZHU Donglai, et al. Entropy value method application in cigarette consumer sensory quality evaluation[J]. Southwest China Journal of Agricultural Sciences, 2014, 27(2): 823–828.

[6] 刘馨. C 卷烟厂精益生产管理体系建设研究[D]. 长沙: 中南大学, 2014.
LIU Xin. Research on construction of lean production management system in C cigarette factory[D]. Changsha: Central South University, 2014.

[7] 明均仁, 肖凯. 基于 R 语言的面向需水预测的随机森林方法[J]. 统计与决策, 2012(9): 81–83.
MING Junren, XIAO Kai. Random forest method on water demand prediction based on R language[J]. Statistics & Decision, 2012(9): 81–83.

[8] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1): 5–32.

[9] 吴喜之. 复杂数据统计方法: 基于 R 的应用[M]. 北京: 中国人民大学出版社, 2012.
WU Xizhi. The statistical methods for complex data: Application based on R[M]. Beijing: China Renmin University Press,2012. （自译）

[10] 崔俊富, 苗建军, 陈金伟. 基于随机森林方法的中国经济增长动力研究[J]. 经济与管理研究, 2015, 36(3): 3–7.
CUI Junfu, MIAO Jianjun, CHEN Jinwei. Research on the impetus of china's economic growth based on random forest[J]. Research on Economics and Management, 2015, 36(3): 3–7.

[11] 方匡南, 吴见彬, 朱建平, 等. 信贷信息不对称下的信用卡信用风险研究[J]. 经济研究, 2010(S): 97–107.
FANG Kuangnan, WU Jianbin, ZHU Jianping, et al. Forecasting of credit card credit risk under asymmetric information based on nonparametric random forests[J]. Economic Research Journal, 2010(S): 97–107.

[12] 李贞子, 张涛, 武晓岩, 等. 随机森林回归分析及在代谢调控关系研究中的应用[J]. 中国卫生统计, 2012, 29(2): 158–163.
LI Zhenzi, ZHANG Tao, WU Xiaoyan, et al. Methodology of regression by random forest and its application on metabolomics[J]. Chinese Journal of Health Statistics, 2012, 29(2): 158–163.

[13] 秦玉华, 宫会丽, 宋楠, 等. 改进随机森林的波长选择用于烟叶近红外稳健校正模型的建立[J]. 烟草科技, 2014(6): 64–67, 72.
QIN Yuhua, GONG Huili, SONG Nan, et al. Wavelength selection based on modified random forest for establishing robust near infrared calibration model of tobacco[J]. Tobacco Science & Technology, 2014(6): 64–67, 72.

[14] 张文彤, 邝春伟. SPSS 统计分析基础教程[M]. 2 版. 北京: 高等教育出版社, 2011.
ZHANG Wentong, KUANG Chunwei. The basic course of SPSS statistical analysis[M].2th ed. Beijing: Higher Education Press, 2011. （自译）

[15] 方匡南, 吴见彬, 谢邦昌. 基于随机森林的保险客户利润贡献度研究[J]. 数理统计与管理, 2014, 33(6): 1122–1131.
FANG Kuangnan, WU Jianbin, SHIA Benchang. Measurement of customer profitability of insurance company in china based on random forest[J]. Journal of Applied Statistics and Management, 2014, 33(6): 1122–1131.

[16] Breiman L. Out-of-bag estimation[EB/OL]. [2010–06–30]. http//stat.berkeley.edu/pub/users/breiman/OOB estimation.ps.

[17] 秦玉华, 丁香乾, 宫会丽. 高维特征选择方法在近红外光谱分类中的应用[J]. 红外与激光工程, 2013, 42(5): 1355–1359.
QIN Yuhua, DING Xiangqian, GONG Huili. High dimensional feature selection in near infrared spectroscopy classification[J]. Infrared and Laser Engineering, 2013, 42(5): 1355–1359.

[18] 杨柳, 王钰. 泛化误差的各种交叉验证估计方法综述[J]. 计算机应用研究, 2015, 32(5): 1287–1290, 1297.
YANG Liu, WANG Yu. Survey for various cross-validation estimators of generalization error[J]. Application Research of Computers, 2015, 32(5): 1287–1290, 1297.

[19] 吴喜之. 统计学: 从数据到结论[M]. 3 版. 北京: 中国统计出版社, 2009.
WU Xizhi. Statistics: From data to Conclusion [M]. 3th ed. Beijing: China Statistics Press, 2009. （自译）

[20] 吕远, 王大洋, 赵方剑. 中高渗水驱油藏单井控制可采储量影响因素定量研究[J]. 中国石油和化工标准与质量, 2014(7): 139–140.
LV Yuan, WANG Dayang, ZHAO Fangjian. Quantitative study on influencing factors of single well controlled recoverable reserves in Mid-High Permeability water drive reservoir[J]. China Petroleum and Chemical Standard and Quality, 2014(7): 139–140. （自译）

[21] 王瑛, 王娜, 肖薇. 基于随机森林赋权和改进 ELECTRE-Ⅲ方法的科技奖励评价研究[J]. 湖南大学学报(自然科学版), 2015, 42(3): 140–144.
WANG Ying, WANG Na, XIAO Wei. Research on the evaluation of science and technology award based on random forest and improved ELECTRE-Ⅲ[J]. Journal of Hunan University (Natural Sciences), 2015, 42(3): 140–144.