基于近红外光谱技术的茶鲜叶海拔高度判别模型建立

王胜鹏[1] 郑鹏程[1] 龚自明[1]
张正竹[2] 滕 靖[1] 王雪萍[1] 卢素芳[1]

1.湖北省农业科学院果树茶叶研究所,武汉 430064;
2.安徽农业大学茶树生物学与资源利用国家重点实验室,合肥 230036

摘要 以不同海拔高度的茶鲜叶为研究对象,扫描获取其近红外光谱(NIRS)并筛选特征光谱区间后,分别应用逐步多元线性回归法(SMLR)、主成分回归法(PCR)和联合区间偏最小二乘法(Si-PLS)建立茶鲜叶海拔高度预测模型。结果表明,在5 542.41～6 888.48 cm^{-1}区间内,对原始光谱进行一阶导数+3点Norris平滑预处理后,建立的SMLR模型预测集相关系数和预测均方差分别为0.800 5和0.486;在4 929.16～6 965.62 cm^{-1}区间内,当主成分数为3时,对原始光谱进行一阶导数+3点Norris平滑预处理后,建立的PCR模型预测集相关系数和预测均方差分别为0.803 6和0.472;当将光谱划分为18个子区间、因子数为13时,选用[5 8 11 17]4个子区间建立的Si-PLS模型预测集相关系数和预测均方差分别为0.944 3和0.295。经比较,Si-PLS模型预测结果最佳。

关键词 茶鲜叶;海拔高度;近红外光谱;多元线性回归法;主成分回归法;联合区间偏最小二乘法
中图分类号 O 657.33 **文献标识码** A **文章编号** 1000-2421(2018)01-0089-06

茶鲜叶质量是成品茶品质的基础,只有应用高质量的茶鲜叶才会加工出高品质的成品茶。茶鲜叶质量除与自身遗传特性有关外,还与茶产地生态环境等因素密切相关,而海拔高度就是其中一个非常重要的因素[1]。一般来说,高海拔地区茶鲜叶质量要优于低海拔地区茶鲜叶,高海拔地区茶鲜叶收购价格也要高于低海拔地区茶鲜叶。由于存在着较大的利润空间,部分茶农采摘低海拔地区的茶鲜叶冒充高海拔地区茶鲜叶,并以较高的价格出售给茶叶加工厂,而收购人员很难凭自身经验判别不同海拔高度茶鲜叶,且存在着较大的主观性。因此,建立一种科学便捷的茶鲜叶海拔高度判别方法十分必要。

近红外光谱(near infrared spectroscopy,NIRS)是指波长介于可见光区与中红外区间的电磁波,波长范围为0.8～2.5 μm。NIRS是一种快速、无损、绿色的分析方法,分析的样品不需要进行任何预处理,具有简便、快速特点,目前已经广泛应用于农业、石化、纺织业和医药等行业[2-4]。国内外很多学者已经将近红外光谱技术应用于茶行业之中。在定量研究方面,已经实现了对茶叶内含成分如含水量、茶多酚、咖啡碱[5]以及抗氧化能力进行快速测定,茶叶等级的精确定级和茶鲜叶质量评价[6]等方面。在定性研究方面,主要集中在对茶叶的种类进行鉴定、判别[7],以及茶叶真伪的鉴定和茶叶原产地溯源[6]等方面。目前,将近红外光谱技术结合多种化学计量学算法应用于茶鲜叶海拔高度的判别方面还鲜有报道。

本研究以采自湖北省恩施土家族苗族自治州恩施市不同海拔高度的茶鲜叶为研究对象,扫描其近红外光谱后,分别应用逐步多元线性回归法(stepwise multiple linear regression,SMLR)[8]、主成分回归法(principal component regression,PCR)[9]和联合区间偏最小二乘法(synergy interval partial least squares,Si-PLS)[9-10]建立不同海拔高度茶鲜叶判别模型,并对其准确性进行验证,为茶叶收购提供一种快速、科学的茶鲜叶海拔高度判别方法。

收稿日期:2017-05-31
基金项目:国家自然科学基金项目(31400586);国家现代茶产业技术体系建设专项(CARS-23)
王胜鹏,博士,副研究员. 研究方向:茶叶加工/茶叶品质快速无损检测. E-mail:wwsspp0426@163.com
通信作者:龚自明,研究员. 研究方向:茶叶加工. E-mail:ziminggong@163.com

1 材料与方法

1.1 试验材料

茶鲜叶样品共120个，其中0 m＜海拔高度＜450 m的鲜叶、450 m≤海拔高度＜800 m的鲜叶、800 m≤海拔高度≤1100 m的鲜叶各40个，均采自湖北省恩施土家族苗族自治州恩施市。采摘标准为单芽、一芽一叶和一芽二叶。采摘时间为2017年4月10日至4月22日。将样品按照3∶1划分为2个集合，其中校正集样品90个，验证集样品30个，分别用于建立校正集模型和预测集模型(不同海拔高度的茶鲜叶设定值分别为1.00、2.00和3.00)。

1.2 试验方法

1)近红外光谱采集。采用美国赛默飞世尔Antaris Ⅱ型傅里叶变换近红外光谱仪；光谱扫描范围4 000～10 000 cm^{-1}；分辨率8 cm^{-1}；InGaAa检测器。仪器开机预热1 h状态稳定后再扫描光谱。每个样品扫描3条光谱，每条光谱扫描64次，取3条光谱的平均值作为该样品的最终光谱值。

2)光谱预处理及模型建立。将每条光谱转化为1 557对数据点，分别应用TQ Analyst 9.4.45软件、OPUS 7.0软件和Matlab 7.0软件对数据点进行预处理，筛选出最佳预处理方法。分别利用SMLR、PCR和Si-PLS 3种方法建立不同海拔高度茶鲜叶近红外光谱判别模型，结果用校正集相关系数(correlation coefficient of cross validation，R_c)、预测集相关系数(correlation coefficient of prediction，R_v)、交互验证均方根方差(root mean square error of cross validation，RMSECV)、预测均方差(root mean square error of prediction，RMSEP)表示。

2 结果与分析

2.1 SMLR预测模型建立

应用TQ Analyst 9.4.45软件建立不同海拔高度茶鲜叶样品SMLR模型，光谱预处理方法分别为原始光谱、一阶导数＋7点卷积平滑和一阶导数＋3点Norris平滑，筛选最佳光谱区间后将样品光谱与赋值间进行联立，结果见表1。

表1 不同预处理方法SMLR建模结果

Table 1 Results of SMLR calibration and prediction models with three different pretreatment methods

预处理方法 Pretreatment methods	光谱区间/cm^{-1} Spectral regions	校正集 Calibration set		验证集 Prediction set	
		R_c	RMSECV	R_v	RMSEP
原始光谱 Original spectra	5 503.84～6 892.34	0.761 6	0.529	0.684 7	0.716
一阶导数＋7点卷积平滑 First derivative＋7 points Savitzky-Golay filter	6 892.34～9 268.21	0.812 2	0.476	0.771 5	0.506
一阶导数＋3点Norris平滑 First derivative＋3 points Norris filter	5 542.41～6 888.48	0.821 6	0.465	0.800 5	0.486

从表1可知，在5 503.84～6 892.34 cm^{-1}区间内，应用原始光谱建立的不同海拔高度茶鲜叶样品预测模型，校正集R_c和RMSECV分别为0.761 6和0.529；当用30个验证集样品对模型进行验证时，验证集R_v和RMSEP分别为0.684 7和0.716。在6 892.34～9 268.21 cm^{-1}区间内，对原始光谱进行一阶导数＋7点卷积平滑预处理后建立不同海拔高度茶鲜叶样品预测模型，校正集R_c和RMSECV分别为0.812 2和0.476；当用30个验证集样品对该模型进行验证时，验证集R_v和RMSEP分别为0.771 5和0.506。在5 542.41～6 888.48 cm^{-1}区间内，对原始光谱进行一阶导数＋3点Norris平滑预处理后建立不同海拔高度茶鲜叶样品预测模型，校正集R_c和交互验RMSECV分别为0.821 6和0.465；当用30个验证集样品对该模型进行验证时，验证集R_v和RMSEP分别为0.800 5和0.486。可见，3种预处理方法建立的模型中，以一阶导数＋3点Norris平滑建立的模型预测结果最佳，筛选的特征光谱区间如图1。从图1可以看出，筛选的光谱区间为5 542.41～6 888.48 cm^{-1}，位于鲜叶光谱的长波区域，光谱信息丰富，而且光谱基本上避开了H_2O的－OH

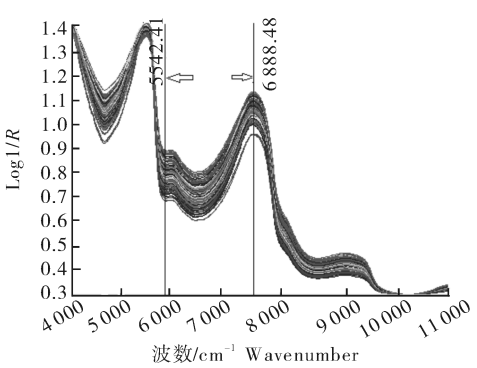

图1 SMLR筛选的特征光谱区间

Fig.1 Optimal spectral regions selected by SMLR method

吸收峰的信号干扰，模型的预测效果也较好。

2.2　PCR 预测模型建立

PCR 是借助主成分分析，先对样品光谱矩阵进行有效分解，然后选取其中的主成分进行多元线性回归分析。在主成分回归中，确定参与回归的最佳主成分数最为重要。如果选取的主因子太少，将会丢失原始光谱较多的有用信息，模型拟合不充分。如果选取的主因子太多，则会将大量噪声信息带入模型之中，导致模型的预测误差增大。本研究应用 TQ Analyst 9.4.45 软件建立不同海拔高度茶鲜叶样品 PCR 模型，光谱预处理方法分别为原始光谱、一阶导数＋7 点卷积平滑和一阶导数＋3 点 Norris 平滑，筛选建模最佳光谱区间和主成分数，然后将样品光谱与赋值间进行联立，结果如表 2 所示。

从表 2 可知，在 4 944.59～5 970.53 cm^{-1}区间内、主成分数为 6 时，应用原始光谱建立不同海拔高度茶鲜叶样品预测模型，校正集 R_c和 RMSECV 分别为 0.633 3 和 0.632；当用 30 个验证集样品对该模型进行验证时，验证集 R_v和 RMSEP 分别为0.582 5 和 0.876。在 4 709.32～5 017.87 cm^{-1}区间内、主成分数为 8 时，对原始光谱进行一阶导数＋7 点卷积平滑预处理后建立不同海拔高度茶鲜叶样品预测模型，校正集 R_c和 RMSECV 分别为 0.658 3 和0.615；当用 30 个验证集样品对该模型进行验证时，验证集 R_v和 RMSEP 分别为 0.632 7 和 0.775。在4 929.16～6 965.62 cm^{-1}区间内、主成分数为3 时，对原始光谱进行一阶导数＋3 点 Norris 平滑预处理后建立不同海拔高度茶鲜叶样品预测模型，校正集 R_c和 RMSECV分别为 0.825 1 和 0.461；当用 30 个验证集样品对该模型进行验证时，验证集 R_v和 RMSEP 分别为 0.803 6 和 0.472。可见，3 种预处理方法建立的模型中以一阶导数＋3 点 Norris 平滑建立的模型预测结果最佳，筛选的特征光谱区间和主成分数如图 2 和图 3。

从图 2 可以看出，筛选的光谱区间为 4 929.16～6 965.62 cm^{-1}，位于鲜叶光谱的长波区域，光谱信息丰富，光谱区间包括了一部分H_2O的－OH吸收峰的信号，但建立的模型预测效果也较好，这可能是由于反映鲜叶海拔高度的光谱信息有可能与水吸收峰存在部分重叠的原因。从图 3 可以看出，PC1 代表了 86.876％光谱信息，PC2 代表了 9.678％光谱信息，PC3 代表了 1.216％光谱信息，前 3 个主成分累计贡献了 97.77％的光谱信息，已经可以完全代表原光谱信息进行建立模型。因此，建立的最佳回归模型选取了 3 个主成分数。

表 2　不同预处理方法 PCR 建模结果

Table 2　Results of PCR calibration and prediction models with three different pretreatment methods

预处理方法 Pretreatment methods	主成分数 PCs	光谱区间/cm^{-1} Spectral regions	校正集 Calibration set		验证集 Prediction set	
			R_c	RMSECV	R_v	RMSECV
原始光谱 Original spectra	6	4 944.59～5 970.53	0.633 3	0.632	0.582 5	0.876
一阶导数＋7 点卷积平滑 First derivative＋7 points Savitzky-Golay filter	8	4 709.32～5 017.87	0.658 3	0.615	0.632 7	0.775
一阶导数＋3 点 Norris 平滑 First derivative＋3 points Norris filter	3	4 929.16～6 965.62	0.825 1	0.461	0.803 6	0.472

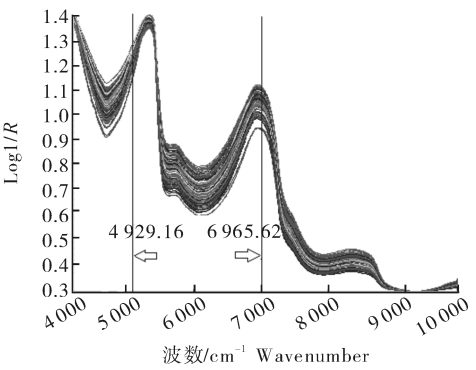

图 2　应用 PCR 筛选的特征光谱区间

Fig.2　Optimal spectral regions selected by PCR method

2.3　Si-PLS 预测模型建立

1)光谱预处理方法比较。本研究比较了多元散射校正、一阶导数和平滑等 3 种光谱预处理方法，均采用偏最小二乘法(PLS)建立定量模型，根据模型预测效果确定最佳光谱预处理方法，结果见表 3。从表 3 可以看出，应用预处理后的鲜叶光谱建立的 PLS 模型预测能力得到明显提高，最佳光谱预处理方法为平滑，此时校正集 R_c为 0.802 1，RMSECV 为 1.215，预测集 R_v为 0.783 5，RMSEP 为 1.081。

2)特征光谱筛选。全光谱数据信息量巨大，建模时容易引入过多的无用数据，模型预测精度降低，需要对光谱区间进行筛选。而联合区间偏最小二乘法(Si-PLS)将全光谱划分为 10、11、12、13…25 个光谱子区间，联合其中2、3和4个子区间建立局部模

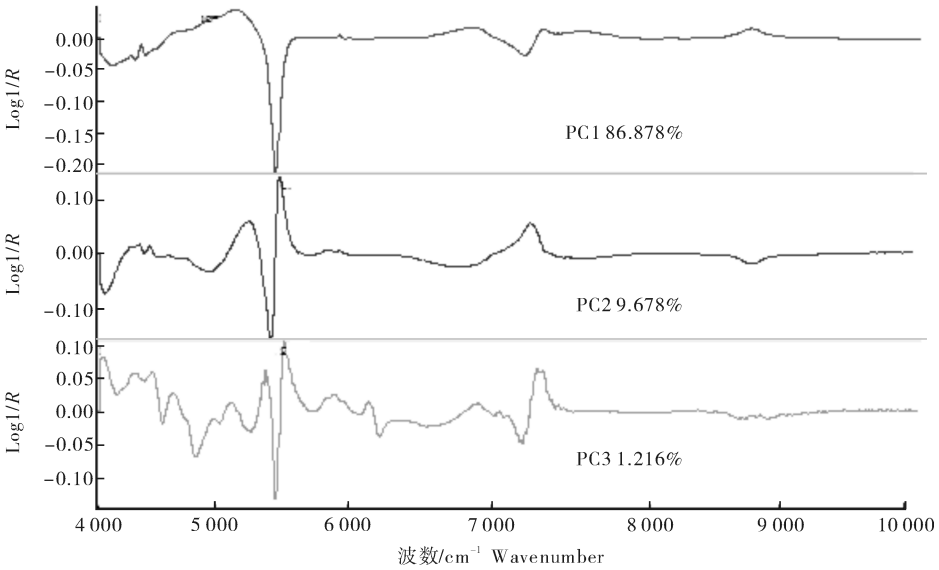

图3 PC1～PC3主成分光谱信息图

Fig.3 The information of PC1-PC3

表3 不同预处理方法PLS建模结果

Table 3 Results of PLS calibration and prediction models with three different pretreatment methods

预处理方法 Pretreatment methods	校正集 Calibration set		验证集 Prediction set	
	R_c	RMSECV	R_v	RMSEP
原始光谱 Original spectra	0.635 3	2.032	0.575 9	2.223
多元散射校正 MSC	0.694 5	1.953	0.642 1	1.988
一阶导数 First derivative	0.765 4	1.585	0.714 9	1.743
平滑 Mean	0.802 1	1.215	0.783 5	1.081

型，对比各个模型的RMSECV值，RMSECV值越小模型精度越高，最小的RMSECV值对应的光谱区间即为筛选的最优子区间。从表4可以看出，当将全光谱划分为18个子区间、因子数为13时，选用[5 8 11 17]4个子区间建立的不同海拔高度茶鲜叶样品Si-PLS模型结果最佳(RMSECV为0.225)，参与建模的数据占全部光谱数据比例为22.22%，显著降低了建模光谱数据量，简化了模型，提高了稳健性。

从表4中还可以看出，在将全光谱划分为10～25子区间建立模型时，最佳模型全部为应用4个光谱子区间建立的，并没有出现应用2个或3个光谱子区间得到的最佳模型，说明建立模型时光谱信息越充分，模型预测效果越好，这进一步确认了优选光谱区间的必要性(图4)。从表4还可以得出，在建立模型的过程中，当将全光谱划分为不同子区间时，光谱子区间的选择也有所不同，这可能是由于茶鲜叶近红外光谱信息丰富，光谱信息间存在着交叉和重叠等现象，不同的光谱子区间可能代表相同的信息，或者代表多少有效光谱信息。在5 000 cm^{-1}和7 000 cm^{-1}附近为H_2O的－OH的近红外光谱吸收峰，从图4可以看出，筛选的光谱子区间主要集中在光谱信息丰富的长波区域，而且避开了水羟基吸收峰的干扰，这样建立的模型才可能更加稳健。

3)预测模型建立。应用上述筛选的光谱子区间建立不同海拔高度茶鲜叶近红外光谱判别模型，结果如图5所示。从图5可以看出，应用Si-PLS方法建立的不同海拔高度茶鲜叶校正集模型R_c和RMSECV分别为0.962 5和0.225；当用30个验证集样品对模型进行检验时，所得验证集模型R_v和RMSEP分别为0.944 3和0.295；可见，模型稳健性具有高稳健性，具有很好的预测效果，可用于茶鲜叶海拔高度的判别。

表 4　不同海拔高度茶鲜叶 Si-PLS 建模结果

Table 4　Results of Si-PLS calibration models for fresh tea leaves at different altitudes with selected spectral regions

区间数 Interval numbers	因子数 Factors	光谱区间 Selected intervals	数据比例/% Data ratio	交互验证均方根方差 RMSECV
10	13	[6 8 9 10]	40.00	0.363
11	13	[7 8 9 10]	36.36	0.374
12	12	[6 8 11 12]	33.33	0.367
13	14	[8 10 11 12]	30.77	0.381
14	14	[8 9 12 14]	28.57	0.374
15	11	[8 10 14 15]	26.67	0.367
16	12	[910 14 16]	25.00	0.356
17	14	[5 7 10 16]	23.53	0.364
18	13	[5 8 11 17]	22.22	0.225
19	10	[11 12 17 19]	21.05	0.371
20	12	[11 12 18 20]	20.00	0.362
21	14	[9 12 17 20]	19.05	0.354
22	15	[6 8 9 20]	18.18	0.369
23	15	[3 13 15 18]	17.39	0.350
24	12	[713 22 24]	16.67	0.354
25	14	[11 14 18 20]	16.00	0.344

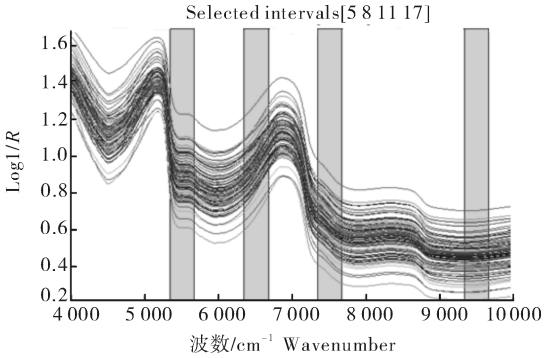

图 4　Si-PLS 方法筛选的最佳建模光谱子区间

Fig.4　Optimal spectral regions selected by Si-PLS method

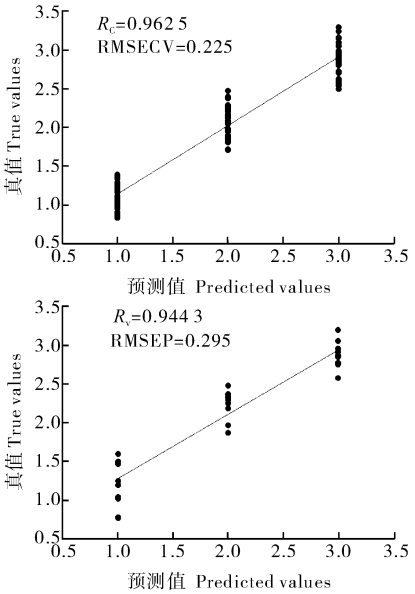

图 5　茶鲜叶近红外光谱预测模型的直值与预测值

Fig.5　True values versus predicted values by Si-PLS in calibration and prediction sets

3　讨　论

本研究利用傅里叶变换近红外光谱技术，在对茶鲜叶光谱进行预处理剔除部分噪声和筛选特征光谱区间后，分别结合逐步多元线性回归方法、主成分回归方法和联合区间偏最小二乘法建立了茶鲜叶海拔高度的判别模型。综合比较建立的 3 种近红外光谱，预测结果最佳的是应用 Si-PLS 建立的模型（R_v=0.944 3，RMSEP=0.295），其次为应用主成分回归方法建立的模型（R_v=0.803 6，RMSEP=0.472），最差为应用逐步多元回归方法建立的模型（R_v=0.800 5，RMSEP=0.486）。这可能是由于 Si-PLS筛选的建模光谱信息最为充分，不仅大大降低了建模光谱数据量（仅占全部光谱数据量的 22.22%），而且还剔除了大量噪声信息，因此，建立的模型预测结果最佳。而主成分回归模型在筛选特征光谱区间的基础上，还对光谱信息进行了有效压缩，但是建模用光谱信息占全部信息的比例为 33.94%，可能夹杂了部分噪声信息，影响了模型预测结果。而逐步多元线性回归模型存在着共线性问题，导致模型计算误差变大，因此，模型预测结果最差。本研究表明应用傅里叶变换近红外光谱技术结合联合区间偏最小二乘法快速、无损预测茶鲜叶样品海拔高度是可行的，有利于鲜叶收购过程中公平交易的实现。

参　考　文　献

[1]　童启庆.茶树栽培学[M].3 版.北京：中国农业出版社，2000.

[2] 陆婉珍.现代近红外光谱分析技术[M].2版.北京:中国石化出版社,2007.
[3] 付苗苗,刘梅英,牛智有.基于高光谱图像技术的配合饲料主要营养成分的检测方法[J].华中农业大学学报,2017,36(2):123-129.
[4] 丁驰竹,谭佐军.基于光纤探头的洋葱近红外光谱检测的数值模拟[J].华中农业大学学报,2007,36(5):110-116.
[5] YAN S H.Evaluation of the composition and sensory properties of tea using near infrared spectroscopy and principal component analysis[J].J Near Infrared Spec,2005,13(6):313-325.
[6] WANG S P,ZHANG Z Z,NING J M,et al.Back propagation-artificial neural network model for prediction of the quality of tea shoots through selection of relevant near infrared spectral data via synergy interval partial least squares[J].Analytical letters,2013,46:184-195.
[7] REN G X,WANG S P,NING J M,et al.Quantitative analysis and geographical traceability of black tea using Fourier transform near-infrared spectroscopy (FT-NIRS)[J].Food research international,2013,53:822-826.
[8] 沃尔夫冈·哈德勒,利奥波德·西马.应用多元统计分析[M].3版.北京:北京大学出版社,2014.
[9] 盖钧镒.试验统计方法[M].北京:中国农业出版社,2001.
[10] NORGAARD L,SAUDLAND A,WAGNER J,et al.Interval partial least squares regression (iPLS):a comparative chemometric study with an example from near-infrared spectroscopy[J].Appl Spectrosc,2000,54:413-419.

Establishment of discrimination model for different elevation fresh tea leaves based on near infrared spectroscopy

WANG Shengpeng[1] ZHENG Pengcheng[1] GONG Ziming[1] ZHANG Zhengzhu[2]
TENG Jing[1] WANG Xueping[1] LU Sufang[1]

1.*Institute of Fruit and Tea, Hubei Academy of Agricultural Science, Wuhan* 430064, *China*;
2.*State Key Laboratory of Tea Plant Biology and Utilization, Anhui Agricultural University, Hefei* 230036, *China*

Abstract There is a certain relationship between the quality of fresh tea leaves and the elevation of growing, but at present, it is no effective method to discriminate the elevation of fresh leaves picked. In this study, fresh tea leaves of different elevation were used as the research objects, after near infrared spectroscopy scanned and the characteristic spectral interval selected, the prediction models of elevation of fresh tea leaves were established by stepwise multiple linear regression (SMLR), principal component regression (PCR) and synergy interval partial least squares (Si-PLS). The results showed that, the correlation coefficient and root mean square error of prediction set was respectively 0.800 5 and 0.486 by SMLR method, which used the spectroscopy in the range of 5 542.41-6 888.48 cm^{-1} and the first derivative +3-point Norris smoothing pretreatment; the correlation coefficient and root mean square error of prediction set was respectively 0.803 6 and 0.472 by PCR method, which used the spectroscopy in the range of 4 929.16-6 965.62 cm^{-1} and the first derivative + 3-point Norris smoothing pretreatment; the correlation coefficient and root mean square error of prediction set was respectively 0.944 3 and 0.295 by Si-PLS method, which contained 18 spectral intervals combined with [5 8 11 17] of four subintervals and 13 factors. By comparison, the Si-PLS model has the best prediction results. It was preliminary realized to discriminate the elevation of fresh tea leaf samples rapidly and nondestructively by using NIRS-Si-PLS method.

Keywords fresh tea leaves; elevation; near infrared spectroscopy; stepwise multiple linear regression; principal component regression; synergy interval partial least squares

(责任编辑:陆文昌)