**DOI: 10.13758/j.cnki.tr.2020.02.025**

肖云飞, 高小红, 李冠稳. 土壤有机质可见光–近红外光谱预测样本优化选择. 土壤, 2020, 52(2): 404–413

# 土壤有机质可见光–近红外光谱预测样本优化选择①

肖云飞，高小红*，李冠稳

(青海师范大学地理科学学院，青海省自然地理与环境过程重点实验室，青藏高原地表过程与生态保育教育部重点实验室，西宁 810008)

**摘 要：**土壤有机质可见光-近红外光谱预测中建模样本的优化选择对提高有机质模型估算精度具有重要作用。本文以湟水流域土壤有机质为例，采用基于土壤单一属性信息考虑的建模样本选择方法：浓度梯度法、Kennard-Stone(KS)方法，以及基于土壤多种信息考虑的建模样本选择方法：Rank-KS(RKS)法、土壤类型结合浓度梯度法以及土壤类型结合 KS 法。通过偏最小二乘回归建模，探索可见光–近红外光谱预测青海湟水流域有机质的最优样本集。结果表明：不同级别样本数的最佳建模样本选择方法不同，整体表现为基于土壤多种信息挑选的建模样本集的模型精度相比土壤单一信息均较高，特别是 KS 方法结合土壤类型后的建模样本集模型精度明显提高且在样本数较少时更为明显。土壤类型可以优化建模样本选择方法提高模型预测精度。在保证固定验证样本模型预测精度的情况下，土壤类型参与建模样本的选择可以有效减少建模样本数，进而降低了建模成本。

**关键词：**土壤有机质；可见光-近红外光谱；土壤类型；建模样本构建；湟水流域

**中图分类号：**S151.9 **文献标志码：**A

## Optimal Selection of Calibration Sample Sets for Predicting Soil Organic Matter Contents from Visible and Near Infrared Reflection Spectrum

XIAO Yunfei, GAO Xiaohong*, LI Guanwen

(*College of Geographical Sciences*, *Key Laboratory of Physical Geography and Environmental Process in Qinghai Province*, *Qinghai Normal University*, *Xining* 810008, *China*)

**Abstract:** Selecting optimal samples for calibration sets in the visible and near infrared reflectance spectrum prediction of soil organic matter is very important to improve the prediction accuracy of SOM contents. In this paper, the Huangshui river basin in Qinghai Province was selected to screen the optimal sample set method by partial least squares regression model for SOM prediction from visible-near infrared reflectance spectrum. Sample selection methods only considered single soil attribute information such as, concentration gradient method, kennard – stone (KS) method, and sample selection methods based on a variety of soil information including Rank – KS (RKS) method, soil type combined with concentration gradient method and soil type combined with KS. The results showed that under different sample number levels, the optimal sample selection method were obviously different, and the model accuracies of the calibration sample set with multiple soil information, especially the precision of calibration sample set model from KS method combined with soil type with low number of samples, were higher than those of the calibration sample set with single soil information. Adding soil type in soil sample selection can improve the accuracy of model prediction. Under the condition of fixed validation samples and model prediction accuracy, adding soil type into the calibration sample selection can effectively reduce the calibration sample numbers and the prediction cost.

**Key words:** SOM; Visible and near-infrared reflection spectrum; Soil type; Modeling sample construction; the Huangshui river basin

土壤有机质是土壤的重要组成成分,是植物主要营养来源之一，是土壤肥力的重要指标[1]。近年来可见光–近红外光谱技术以高效性、实时性、成本低的特点在土壤理化性质中得到了快速发展[2]。为提高可见光–近红外光谱对土壤有机质的预测精度，国内外学者从土壤粒径[3]、光谱预处理[4]、特征光谱波段选

①基金项目：国家自然科学基金项目(41550003)和青海省科技厅自然科学基金项目(2016-ZJ-907)资助。

* 通讯作者(xiaohonggao226@163.com)

作者简介：肖云飞(1992—)，女，甘肃兰州人，硕士研究生，主要从事遥感应用与地理空间数据分析研究。E-mail: 1328056300@qq.com

择[5]、建模方法[6]等方面分别研究了对有机质预测精度的影响，期待找到最佳的土壤有机质可见光–近红外光谱预测方法。近来有学者从建模样本选择优化角度出发[7-8]，研究最佳建模样本集，以提高模型精度。

实际情况中，由于土壤样本的采集受多种原因影响，如交通可通达性、采样区域的范围大小、经费的不足等未能全面地考虑成土母质、理化性质、地理空间位置等，使采样不具有代表性、样本空间分布的不均匀等情况，影响建模样本的选择，进而影响模型的预测能力。目前常用的建模样本选择方法有浓度梯度法、Kennard-Stone(KS)方法、Rank-KS(RKS)方法。KS 方法以土壤光谱差异性作为建模样本与验证样本选择的依据，当样本之间的光谱差异较小时，选择出的建模样本就不具有代表性；浓度梯度法主要考虑了土壤有机质的含量，但未考虑土壤的光谱特性、地理位置及其他理化性质对建模样本选择的影响；RKS 方法虽考虑了土壤有机质含量与土壤光谱特性，但也未考虑土壤的其他理化性质。所以该方法也使得挑选的建模样本因未考虑其他因素的影响而缺少代表性，且 RKS 方法在利用有机质含量对样本分类时没有一个具体的标准，可能会造成建模样本的挑选不均匀。

陈奕云等[7]分别采用 KS、RKS、SPXY(Sample set Partitioning based on joint X-Y distance)3 种方法挑选不同比例的建模样本来预测固定验证样本的精度，研究表明 KS 方法无法提高模型预测精度，SPXY 方法用 50% 总建模样本数就能达到建模预测精度，RKS 方法在保证建模预测精度时可以减少 70% 的建模样本。刘艳芳等[8]利用土地利用类型结合土壤理化信息、光谱信息挑选建模样本集，研究表明具有多种土壤信息结合的方法选择的建模样本更具有代表性，可以有效地提高模型预测精度。邬登巍和张甘霖[9]研究表明母质和土地利用类型的差异会显著影响异地模型的适应性，一个地区建立的估算模型不可随便用于母质和土地利用类型不同的其他地区。Liu 等[10]利用含有不同土地利用类型的建模样本的模型很好地预测了单一土地利用类型有机质含量。刘伟等[11]提出将光谱信息和理化信息结合构成的 RKS 方法能够明显地提高二甲亚砜溶液浓度的预测精度。以上研究表明加入土壤多种信息的建模样本选择方法，可以构建更具有代表性的建模集，从而提高模型预测精度。

土壤光谱反射率是土壤的众多理化性质的综合反映，理化性质不同，光谱反射率也就不同。有机质含量相同的不同土壤类型的光谱也可能不尽相同[12-14]。本文将土壤类型加入建模样本的选择方法中，结合浓度梯度法、KS 方法，构成 5 种建模样本选择方法，对比不同建模样本选择方法的模型精度，研究土壤类型对建模样本选择的影响，寻找最佳的建模样本构建方法，以及在固定验证样本情况下模型达到一定预测精度至少所需的建模样本数，为今后湟水流域有机质预测提供较好的建模集构建方法，同时为湟水流域野外采样提供数据支持。

## 1　材料与方法

### 1.1　研究区概况

湟水是黄河上游最大的一级支流，发源于青海省海晏县境内，青海境内全长 336 km。湟水流域位于青海东部地区，地处 36°02′ ~ 37°28′N，100°42′ ~ 103°04′E(图 1)，是黄土高原向青藏高原的过渡地带。流域内地形比较复杂，内有河谷盆地、丘陵和中高山地，海拔 1 655 ~ 4 860 m，流域西宽东窄，西高东低，为高原干旱、半干旱气候，气温由西向东逐渐升高[15]。流域内主要土壤类型为灰钙土、栗钙土、黑钙土、灰褐土、高山草甸土、山地草甸土等，主要农作物有油菜、马铃薯、春小麦、玉米、青稞、燕麦等，是青海省主要的农业区。

### 1.2　样品采集与分析

研究所使用的土壤样本为 2015 年、2016 年采集的 418 个土壤样品，采样点空间分布如图 1，采样时间选择在农作物收割结束的 10 月至 11 月初，共计 34 d，为了与野外光谱采集时间一致，土壤采样在天气晴朗的 11:00—15:30 间。采集土壤样本时考虑土壤类型、可到达性、耕地类型、采样地未翻晒等因素，相对均匀地分布在整个流域内。土壤样本采集方式在相对平坦的地方采用“梅花型”方法采集，坡耕地采用“S”型方法采集，采集土壤表层 0 ~ 20 cm 的表土，去除植物根系和石粒，搅拌均匀装入密封袋，使用 GPS 实时记录采样点坐标和高程。土壤样本在实验室避光条件下自然风干过 100 目筛，用于土壤反射光谱的测量和有机质的实验室分析测试。有机质含量的测定采用重铬酸钾外加热法。有机质含量特征统计结果见表 1。

变异系数用来衡量土壤特征的空间变异强度，变异系数 0 ~ 10% 为小变异，10% ~ 100% 为中等变异，大于 100% 为高度变异[16]。从表中可以看出不同土壤类型的有机质含量都属于中等变异，且变异大小为高山草甸土<灰钙土<灰褐土<黑钙土<山地草甸土<栗钙土<总体变异系数，高山草甸土与山地草甸土的土壤有机质含量较高，平均值分别达到了

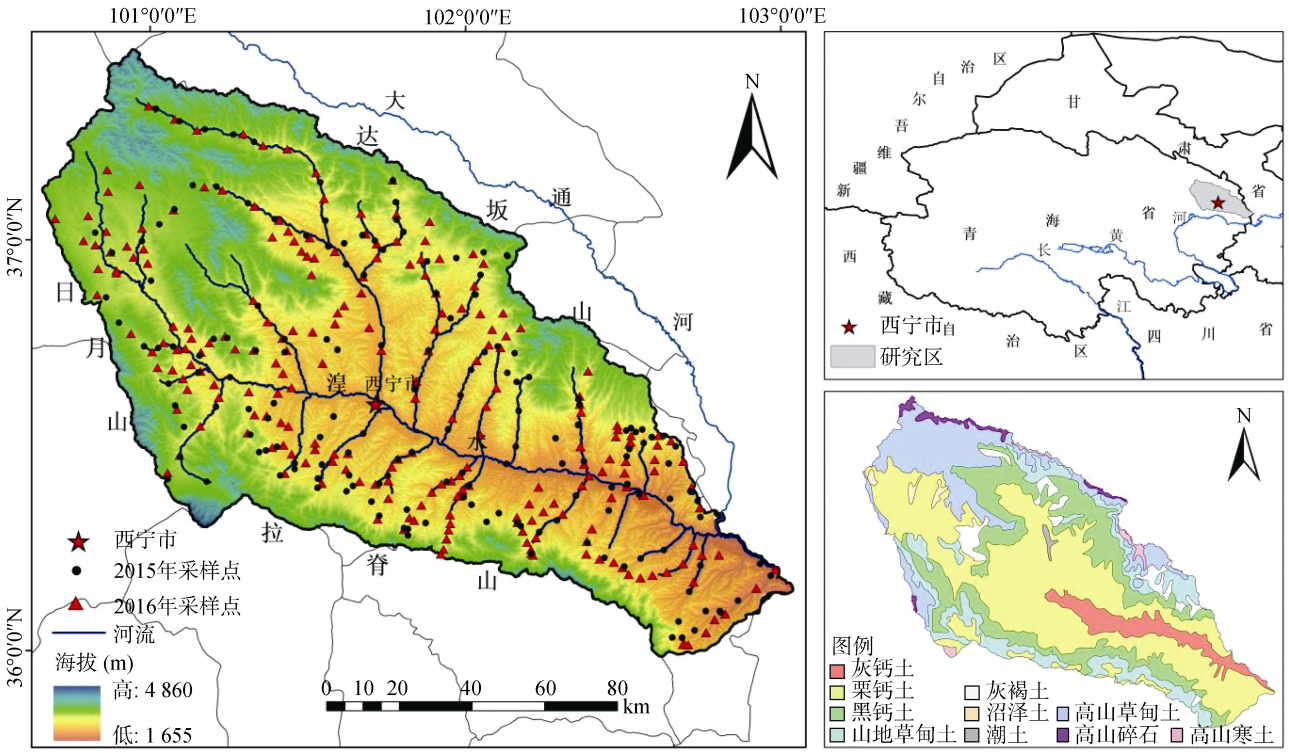


**图 1　研究区采样点空间分布及土壤类型**

Fig. 1　Spatial distribution of soil types and sampling sites in study area

**表 1　土壤有机质含量特征统计**

Table 1　Descriptive statistics of soil organic matter contents for different soil types

| 土壤类型 | 样本数 | 最大值(g/kg) | 最小值(g/kg) | 平均值(g/kg) | 标准差(g/kg) | 变异系数% |
|---|---|---|---|---|---|---|
| 灰钙土 | 23 | 26.30 | 8.96 | 16.96 | 4.38 | 25.85 |
| 栗钙土 | 191 | 144.41 | 4.86 | 27.73 | 20.10 | 72.50 |
| 黑钙土 | 129 | 131.29 | 6.14 | 31.74 | 16.11 | 50.77 |
| 灰褐土 | 19 | 69.82 | 19.94 | 37.58 | 14.52 | 38.65 |
| 山地草甸土 | 27 | 150.44 | 19.03 | 67.47 | 38.94 | 57.71 |
| 高山草甸土 | 11 | 98.97 | 59.30 | 91.53 | 23.61 | 25.80 |
| 总计 | 400 | 150.44 | 4.86 | 33.30 | 24.57 | 73.76 |

91.53、67.47 g/kg，灰钙土有机质含量最低，平均值为 16.96 g/kg。

### 1.3　室内光谱的测量及预处理

采用美国 ASD Field Spec 4 地物光谱仪在暗室内测量光谱，波长范围为 250～2 500 nm，重采样间隔为 1 nm，输出波段 2 151 条。将土壤装入直径为 12 cm、高为 2 cm 的黑色玻璃器皿中。用 75 W 卤素灯作为光源，光源入射角为 30°，距土壤样本表面中心 30 cm。光谱探头距土壤样本中心 15 cm 处垂直向下，土壤样本每旋转 90°测 5 条光谱，共 20 条光谱，光谱测量中每测 5 个土壤样本白板定标一次。检查光谱中是否有异常光谱，剔除异常光谱计算平均值作为土壤的原始光谱。利用主成分分析方法剔除光谱异常值 18 条，剩余 400 条光谱进行后续研究。

在光谱测量中，由于光谱测量环境、测量仪器等因素的影响，光谱中会出现噪声，噪声的存在会影响光谱信息的表达、分析及模型的精度[17]。光谱的微分变换可以减少噪声和背景的影响，放大光谱特征的差异[18]。本文中首先去除噪声比较大的 350～399 nm、2 401～2 500 nm 波段。光谱预处理方法选择 Savitzky-Golay(SG)加一阶微分变换。原始光谱及一阶微分光谱反射率如图 2 所示。

### 1.4　建模集样本选择方法

本文考虑土壤有机质含量、光谱特征和土壤类型构建了 5 种建模样本优化选择方法。浓度梯度法是一种基于理化性质的建模样本选择方法，将土壤有机质含量按大小顺序排列，并依顺序每隔一个样本抽取两个样本为建模样本，剩余为验证样本；土壤类型结合

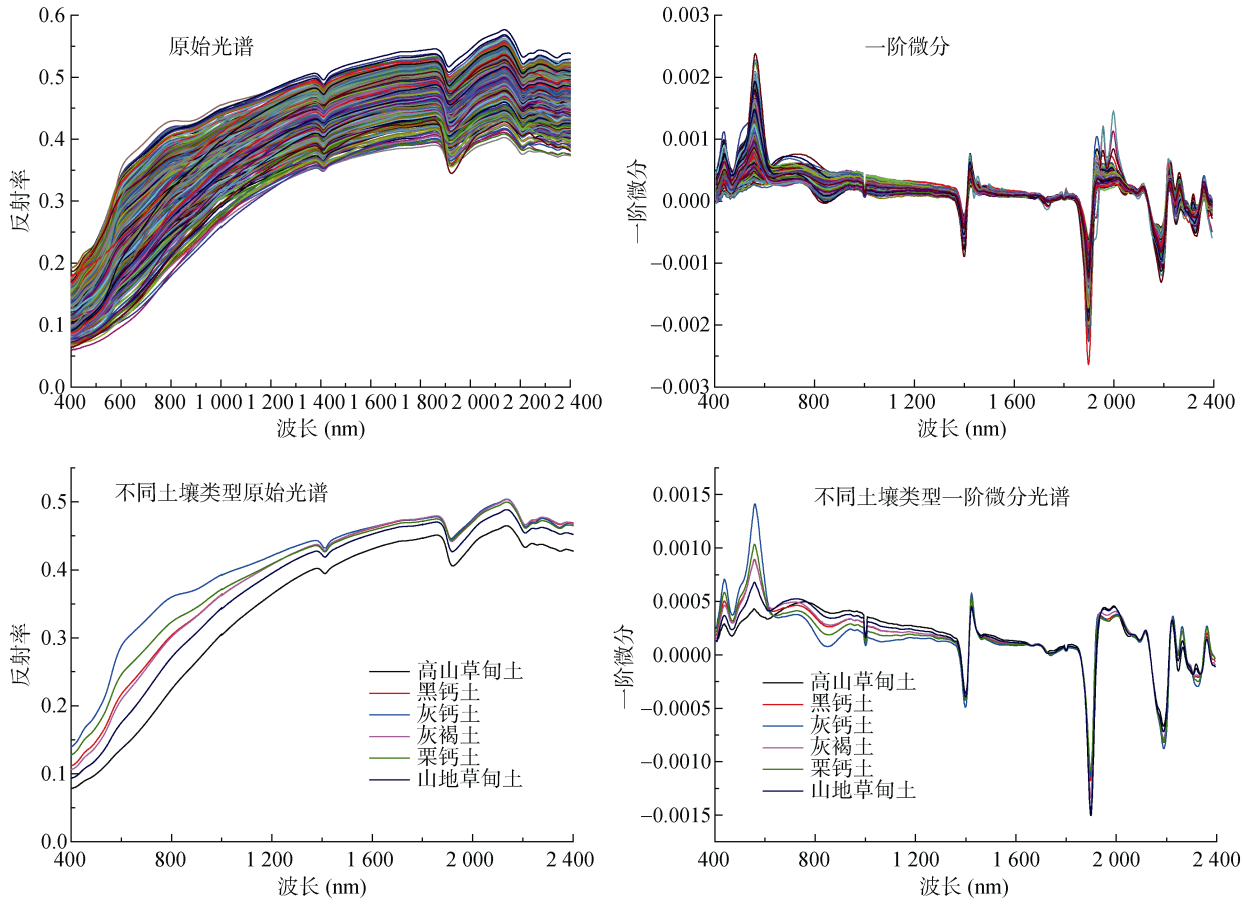


**图 2　土壤原始光谱反射率及一阶微分变换**

Fig. 2　Soil original spectral reflectance and first-order differential spectra of different soil types

浓度梯度法原理是首先将土壤按照类型分开，每一种土壤类型按土壤有机质含量大小排序，按顺序每隔一个样本抽取两个样本，将每种土壤类型所抽取的样本合为一个整体作为建模集的样本；KS 方法根据光谱主成分空间的欧氏距离选择样本，先寻找全体样本空间中欧氏距离最远的两个样本，归入建模集。再依次计算全体样本中每个剩余样本到建模集样本的距离，选取每个剩余样本的最短距离，将这些剩余样本最短距离中的最长距离所对应的样本选入建模集。重复上一个步骤，直至建模集中样本的数量和所需建模集样本数量一致；土壤类型结合 KS 方法原理是先将土壤按类型分开，每种土壤类型中按 KS 方法挑选出一定数目的样本作为建模样本；RKS 法是一种既考虑土壤样本的理化性质又考虑其光谱性质的建模样本挑选方法，首先按样本有机质的含量将样本分为多份，每一份中再按 KS 方法选择一定数目的样本作为建模样本[11]。

### 1.5　偏最小二乘回归(PLSR)建模与验证

偏最小二乘回归(PLSR)是由 Wold 和 Albano 等在 1989 年提出的，模型同时实现了多元回归、主成分分析、变量之间相关分析的新型多元统计分析方法。通过因子分析实现了光谱数据的降维，同时也除去了干扰组分和干扰因素的影响，消除了自变量间多重共线性，很好地解决了样本数少于变量数的问题[19]。模型采用留一法交叉验证方法。模型精度评价指标采用决定系数($R^2$)、均方根误差(RMSE)、相对分析误差(RPD)，当 $R^2$、RMSE 越小，RPD 越大，模型的预测效果越好，Chang 等[20]认为当 RPD≥2 时模型有较好的预测效果，当 1.4≤RPD<2 时模型有粗略估算能力，当 RPD<1.4 时模型不具备估算能力。

### 1.6　不同级别样点数的建模样本选择方法

为了研究不同级别样点数的最佳建模样本选择方法以及加入土壤类型对建模样本选择方法的优化效果，将剔除异常值后的 400 个土壤样本点按其经纬度坐标导入到研究区地理空间中，在满足样点相对均匀分布在整个研究区内且包含不同土壤类型的条件下，计算样本点之间的空间距离(ArcGIS 软件中完成)。在样本数分别为 400、350、300、250 时，土壤样点空间分布图上计算相同土壤类型样点的空间距离，考虑每一类土壤类型的样点占总样点的百分比，

依据经验样点间距离分别不小于 0.12、0.22、0.32、0.44 km 对样点进行删除，按土壤样点间隔数 50 剔除样本数，最后得到 350、300、250、200 的样点分布图。在样本数分别为 200、150 时，土壤采样点空间分布图上不考虑土壤类型计算邻近点之间的距离，删除距离较近两采样点中的一个样点(删除样点为土壤类型样本数较多的样点)，按土壤样点间隔数 50 剔除样本数，得到 150、100 的样本分布图。最终样本分成 7 个级别，即样本数为 400、350、300、250、200、150、100，并分别应用浓度梯度方法、KS 法、RKS 法、土壤类型结合浓度梯度方法、土壤类型结合 KS 方法，按建模样本与验证样本比为 2∶1 确定每一级别样本的建模样本与验证样本。不同级别样点的空间位置如图 3 所示。

### 1.7 固定验证样本下不同建模样本数的建模样本选择方法

将 400 个土壤样本考虑空间位置、土壤类型、有机质含量挑选出 1/5 样本(80 个)作为验证样本，剩余的 4/5 的样本(320 个)作为总建模样本。浓度梯度方法、KS 方法、RKS 方法按占样本数的 90%、80%、70%、60%、50%、40%、30%、20%、10% 分别挑选出建模样本。土壤类型结合浓度梯度方法、KS 方法按土壤类型将总建模样本分开，每种土壤类型的建模样本占该土壤类型样本数的 90%、80%、70%、60%、50%、40%、30%、20%、10% 挑选出来作为一个整体作为总建模样本。验证样本地理空间分布见图 3H。不同建模样本特征统计见表 2。

## 2 结果

### 2.1 不同级别样本数的建模样本选择方法模型精度对比及变化趋势

不同级别样本数的 5 种建模样本选择方法模型精度结果如图 4。KS 方法在样本数为 400、350 时，建模 $R^2$ 与验证 $R^2$ 差值较小，模型具有较好的预测能力(RPD>2)；且在样本数为 400 时 RPD 值最大(RPD=2.459)，模型具有最佳预测能力；当样本数小于 300 时，建模 $R^2$ 不断在增大，而验证 $R^2$ 不断减小，差值不断增大，模型有过拟合现象，预测能力较差。浓度梯度法在不同级别样本数的模型精度变化不大，表现为在样本数分别为 400、350、250、200 时，RPD>2，模型具有较好的预测能力；且在样本数为 200 时 RPD 为 2.579，模型具有最佳预测能力；在样本数分别为 300、150、100 时，RPD<2，模型具有粗略预测能力；样本数在 150 时建模 $R^2$ 为 0.896，验证 $R^2$ 为 0.707，差值变大，模型不稳定。RKS 方法在样本数分别为 400、350、300、200、150 时，RPD>2，模型较稳定，具有较好的预测能力；在样本分别为 250、100 时，RPD<2，模型具有粗略预测能力，且样本数为 100 时建模 $R^2$ 为 0.909，验证 $R^2$ 为 0.667，模型不稳定。土壤类型结合浓度梯度法挑选的建模样本的模型整体 RPD>2，具有较好的预测能力；当样本数为 300 时 RPD 为 3.237，模型具有极好的预测能力。土壤类型结合 KS 方法，样本数为 300 时，RPD 为 1.910，模型具有粗略预测能力；样本数为 200 时，RPD 为 3.01，模型具有极好的预测能力；其他样本数时，RPD>2，具有较好的预测能力。

当样本数为 350、400 时，不同的建模样本选择方法的模型 RPD 值都大于 2，模型预测精度差值不大，且在样本数为 400 时，不同建模样本选择方法的建模 $R^2$、验证 $R^2$、RPD 值都较接近，说明对本研究区来说样本数为 400 是最佳的样本选择。当样本数小于 350 时，不同建模样本选择方法的建模 $R^2$ 值差值不大，但验证 $R^2$ 的差值变大，RPD 值差值也变大。当样本数为 150、100 时，验证 $R^2$、RPD 值都较小，说明样本数少于 150 时，模型只具有粗略的预测能力。

对比模型预测精度发现加入土壤类型的浓度梯度法和 KS 方法挑选的建模样本集的模型相比其他建模样本选择方法挑选的建模样本集的模型验证 $R^2$ 和 RPD 值相对较大。且在样本数较少时更加明显，说明加入土壤类型可以很好地优化建模样本选择方法满足样本点较少达到很好的预测效果。

### 2.2 固定验证样本下建模样本选择方法精度对比

在固定验证样本的情况下，建模样本的不同选择方法精度对比如图 5。对于任何一种建模样本选择方法挑选的子建模集，在建模样本≥50% 时，建模 $R^2$>0.82，验证 $R^2$>0.75，RPD>2，且差值较小，模型较稳定。说明不考虑建模样本选择方法情况下，只需要 50% (160) 的建模样本数就能保证模型具备好的预测精度。当建模样本小于总建模样本的 50% 时，不同建模样本选择方法的模型预测能力的差异变大，首先 RKS 方法的模型精度下降，说明 RKS 方法的建模样本数量在下降到一定比例时，容易丢失对模型预测精度有显著贡献的建模样本。加入土壤类型的浓度梯度法和 KS 方法在建模样本数只有总建模样本数的 20% 时，RPD≥2，模型具有好的预测精度，很大程度减少了建模成本。

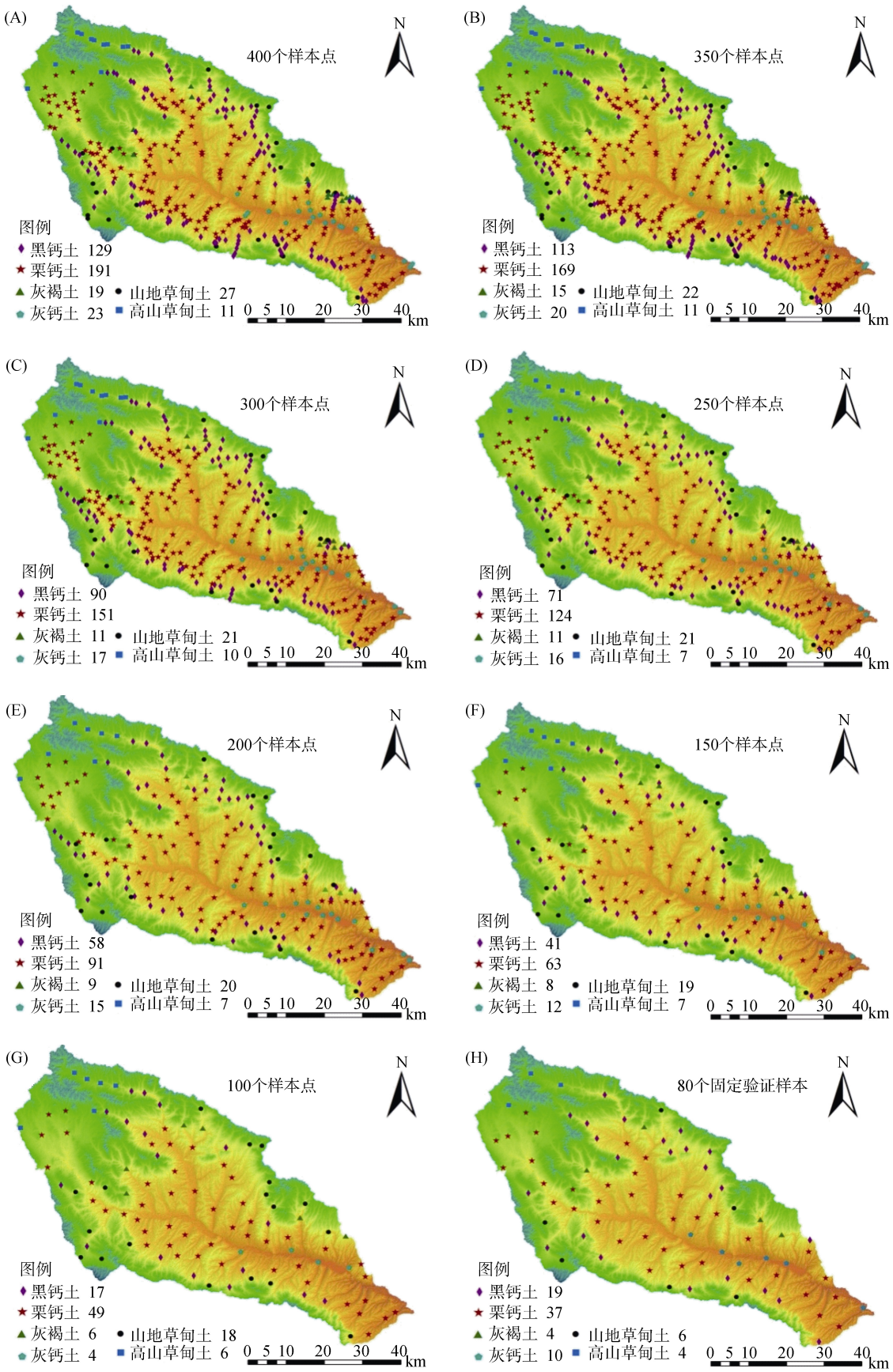


**图 3　不同级别样本空间分布图**

Fig. 3　Spatial distribution of sample sites at different levels

**表 2 固定验证样本下的不同建模样本特征统计**

Table 2 Descriptive statistics of different calibration samples under fixed verification samples

| 建模样本选择方法 | 百分比(%) | 建模样本数 | | | | | | | 有机质含量 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 灰钙土 | 栗钙土 | 黑钙土 | 灰褐土 | 山地草甸土 | 高山草甸土 | 总计 | 最大值(g/kg) | 最小值(g/kg) | 平均值(g/kg) |
| 固定验证样本 | | 6 | 39 | 27 | 2 | 3 | 3 | 80 | 133.56 | 8.569 | 33.550 |
| 浓度梯度法 | 90 | 15 | 138 | 91 | 15 | 22 | 7 | 288 | 150.44 | 4.86 | 33.27 |
| | 80 | 14 | 120 | 80 | 15 | 19 | 8 | 256 | 150.44 | 4.86 | 33.23 |
| | 70 | 12 | 106 | 69 | 13 | 17 | 7 | 224 | 150.44 | 4.86 | 33.27 |
| | 60 | 9 | 95 | 60 | 10 | 13 | 5 | 192 | 150.44 | 4.86 | 33.26 |
| | 50 | 5 | 77 | 53 | 9 | 12 | 4 | 160 | 148.74 | 4.86 | 32.99 |
| | 40 | 8 | 56 | 43 | 7 | 11 | 3 | 128 | 148.74 | 6.14 | 33.22 |
| | 30 | 6 | 40 | 38 | 3 | 8 | 1 | 96 | 144.41 | 6.14 | 33.09 |
| | 20 | 3 | 32 | 22 | 2 | 5 | 0 | 64 | 144.41 | 8.26 | 33.28 |
| | 10 | 2 | 14 | 11 | 2 | 2 | 1 | 32 | 124.27 | 9.27 | 32.98 |
| KS | 90 | 14 | 139 | 89 | 16 | 22 | 8 | 288 | 150.44 | 4.86 | 33.43 |
| | 80 | 13 | 119 | 83 | 13 | 21 | 7 | 256 | 150.44 | 4.86 | 34.39 |
| | 70 | 12 | 98 | 75 | 12 | 21 | 6 | 224 | 150.44 | 4.86 | 34.97 |
| | 60 | 10 | 85 | 65 | 11 | 16 | 5 | 192 | 150.44 | 4.86 | 35.75 |
| | 50 | 10 | 67 | 52 | 10 | 16 | 5 | 160 | 150.44 | 4.86 | 37.40 |
| | 40 | 8 | 56 | 40 | 6 | 14 | 4 | 128 | 150.44 | 4.86 | 38.13 |
| | 30 | 4 | 43 | 32 | 4 | 10 | 3 | 96 | 150.44 | 4.86 | 40.13 |
| | 20 | 2 | 26 | 22 | 2 | 10 | 2 | 64 | 150.44 | 4.86 | 43.25 |
| | 10 | 2 | 13 | 10 | 2 | 5 | 0 | 32 | 150.44 | 4.86 | 41.84 |
| RKS | 90 | 15 | 136 | 91 | 17 | 23 | 6 | 288 | 150.44 | 4.86 | 33.54 |
| | 80 | 14 | 121 | 80 | 15 | 21 | 5 | 256 | 150.44 | 4.86 | 33.58 |
| | 70 | 10 | 106 | 72 | 12 | 19 | 5 | 224 | 150.44 | 4.86 | 34.16 |
| | 60 | 10 | 89 | 62 | 9 | 17 | 5 | 192 | 150.44 | 4.86 | 34.64 |
| | 50 | 8 | 77 | 48 | 8 | 14 | 5 | 160 | 150.44 | 4.86 | 34.87 |
| | 40 | 7 | 62 | 38 | 6 | 12 | 3 | 128 | 150.44 | 4.86 | 35.44 |
| | 30 | 6 | 49 | 27 | 5 | 8 | 1 | 96 | 150.44 | 4.86 | 34.60 |
| | 20 | 5 | 31 | 20 | 3 | 4 | 1 | 64 | 150.44 | 4.86 | 35.11 |
| | 10 | 3 | 14 | 10 | 1 | 3 | 1 | 32 | 150.44 | 4.86 | 34.17 |
| 土壤类型结合浓度梯度法 | 90 | 15 | 137 | 92 | 15 | 22 | 7 | 288 | 150.44 | 4.86 | 33.35 |
| | 80 | 13 | 122 | 82 | 14 | 19 | 6 | 256 | 150.44 | 4.86 | 33.50 |
| | 70 | 12 | 106 | 71 | 12 | 17 | 6 | 224 | 150.44 | 4.86 | 33.28 |
| | 60 | 10 | 91 | 61 | 10 | 15 | 5 | 192 | 150.44 | 4.86 | 33.70 |
| | 50 | 8 | 76 | 51 | 9 | 12 | 4 | 160 | 150.44 | 4.86 | 33.66 |
| | 40 | 7 | 61 | 41 | 7 | 9 | 3 | 128 | 144.41 | 8.26 | 32.55 |
| | 30 | 5 | 46 | 31 | 5 | 7 | 2 | 96 | 138.56 | 9.27 | 33.38 |
| | 20 | 4 | 30 | 20 | 3 | 5 | 2 | 64 | 138.56 | 9.27 | 32.29 |
| | 10 | 2 | 15 | 10 | 2 | 2 | 1 | 32 | 89.07 | 10.17 | 31.67 |
| 土壤类型结合KS方法 | 90 | 15 | 137 | 92 | 15 | 22 | 7 | 288 | 150.44 | 4.86 | 33.57 |
| | 80 | 13 | 122 | 82 | 14 | 19 | 6 | 256 | 150.44 | 4.86 | 34.11 |
| | 70 | 12 | 106 | 71 | 12 | 17 | 6 | 224 | 150.44 | 4.86 | 34.76 |
| | 60 | 10 | 91 | 61 | 10 | 15 | 5 | 192 | 150.44 | 4.86 | 35.70 |
| | 50 | 8 | 76 | 51 | 9 | 12 | 4 | 160 | 150.44 | 4.86 | 34.83 |
| | 40 | 7 | 61 | 41 | 7 | 9 | 3 | 128 | 150.44 | 4.86 | 34.53 |
| | 30 | 5 | 46 | 31 | 5 | 7 | 2 | 96 | 150.44 | 4.86 | 37.47 |
| | 20 | 4 | 30 | 20 | 3 | 5 | 2 | 64 | 150.44 | 4.86 | 40.11 |
| | 10 | 2 | 15 | 10 | 2 | 2 | 1 | 32 | 150.44 | 8.96 | 33.21 |

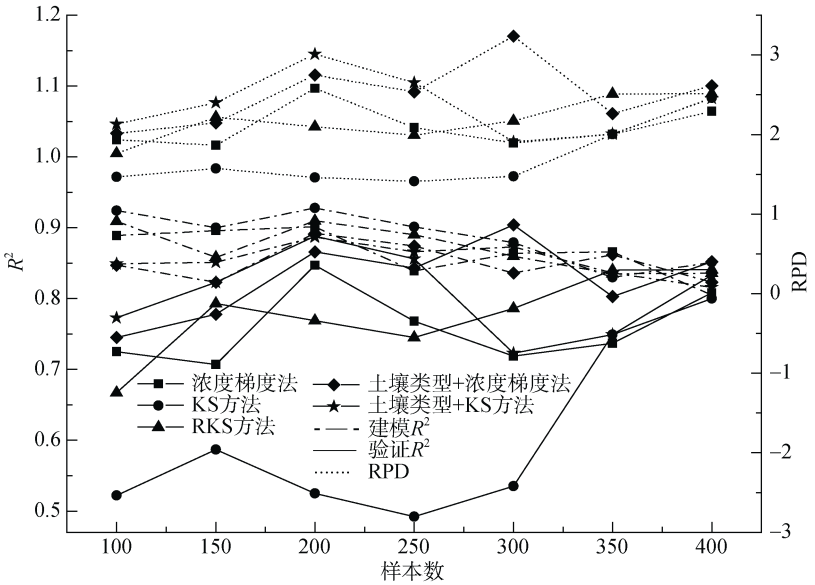


**图 4　不同级别采样点建模样本选择方法模型精度**

Fig. 4　Models accuracies for different calibration samples at different samples levels

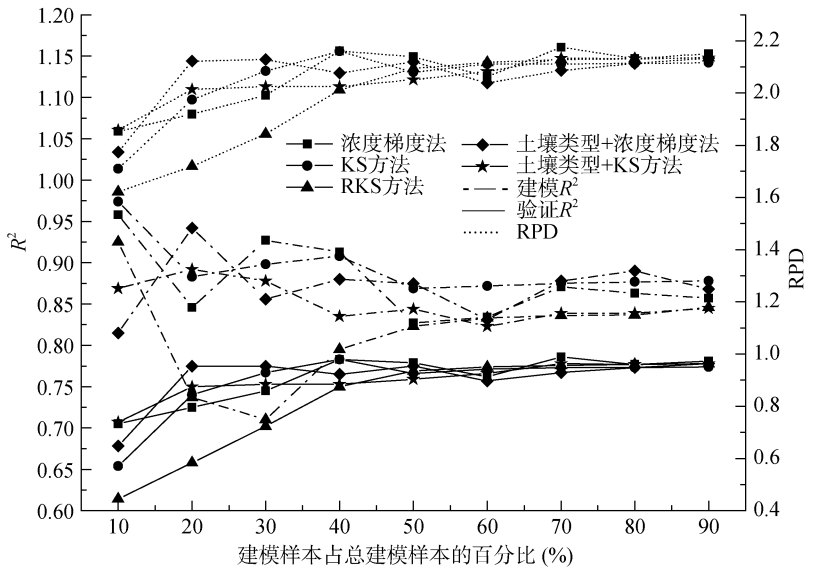


**图 5　固定验证样本下不同建模样本选择方法精度**

Fig. 5　Models accuracies for different calibration samples under fixed validation samples

## 3　讨论

不同类型土壤的反射光谱曲线存在着一定的规律性及基本一致的变化趋势，但土壤可见光-近红外光谱只能反映土壤一部分理化性质的差异，不能很好地揭示不同类型土壤的类间差异[21]。加入土壤类型有效地控制了地理空间位置、环境背景、其他理化性质对光谱的影响，使得建模样本选择更具有代表性。

本文研究与刘艳芳等[8]在江汉平原的研究结论是一致的，即加入土地利用类型可以构建更具有代表性的建模样本集。在该结论的基础上利用湟水流域 400 个土壤样本，对比不同样本数级别下加入土壤类型对建模样本选择方法的优化效果表明：当样本数大于 350 时，土壤类型的加入对建模样本选择方法的优化效果并不显著；而当样本数小于 300 时，土壤类型对建模样本选择方法的优化效果则较为显著。

KS 方法在本文有机质预测中表现为：当样本数在 400、350 时模型预测精度较好，模型相对稳定；但当样本数小于 350，模型精度较差，模型较不稳定。土壤类型对 KS 方法的优化结果表现为：在样本点为 400、350 时，土壤类型结合 KS 方法相比 KS 方法的模型预测精度提高并不明显；但当样本数小于 350

时，土壤类型结合 KS 方法相比 KS 方法模型精度明显提高，模型变稳定。出现上述情况的原因是：在不断剔除样本数的过程中可能缩小了有机质含量范围使光谱差异逐渐变小，或当样本理化性质含量低或含量范围较窄时，采样点间光谱差异较小，导致 KS 方法所挑选的建模样本不具代表性[10]；而加入土壤类型可以弥补 KS 方法的不足，使建模样本集更具有代表性。进而表明当样本较多时，在保证模型预测精度的情况下，为考虑建模时间成本，可以只选择 KS 方法挑选建模样本；当样本较少的时候，为保证模型的预测精度，加入土壤类型可以优化 KS 方法，挑选出更具有代表性的建模样本。

RKS 方法考虑了土壤的光谱信息与理化性质，相比 KS 方法也有效地提高了模型的精度，但整体提高效果没有土壤类型结合 KS 方法好，可能因为 RKS 方法是将样本按有机质含量排序划分等级，再进行 KS 方法挑选建模样本集，不同等级的划分没有指定的标准，随机性比较大。

土壤类型结合浓度梯度法相比浓度梯度法模型精度提高不明显，可能是由于有机质含量的考虑进而也间接考虑了光谱的原因。有研究证明有机质含量高低在可见光-近红外波段对光谱反射率有较大的影响，有机质越低，反射率越高，反之亦然[22]。

固定验证样本(考虑了验证样本的空间分布位置、土壤类型及有机质含量)，减少建模样本数量，研究在保证预测精度情况下建模的最小成本。表 2 中建模样本占总建模样本的不同比例下，KS 方法的建模样本的有机质含量平均值与固定验证样本的有机质含量平均值差值相对较大，而其他建模样本选择方法的建模样本有机质含量平均值与固定验证样本的有机质含量平均值接近，且不同建模样本选择方法的建模样本有机质含量的范围变化不大，固定验证样本的有机质含量都在建模样本有机质范围以内，但模型预测精度不同。说明土壤光谱是土壤众多性质的综合反映，只考虑单一因素的建模样本选择方法不能很好地挑选出具有代表性的建模样本集，且在样本数量较少时更为显著，加入土壤类型可以有效地优化建模样本选择方法，提高模型精度和减少建模样本数。

陈奕云等[7]在固定验证样本的情况下，KS 方法建模样本数仅占总建模样本数的 70% 就能很好地保证模型的预测精度。本文中 KS 方法建模样本数仅占总建模样本数的 30% 就能很好地保证模型的预测精度，KS 方法在保证模型精度下选择的建模样本占总建模样本的比例更小，可能是因为样本数不同和固定样本选择考虑的因素不同，或因地区之间的土壤类型及有机质差异所造成的。以后的研究中在选择验证样本时可以考虑加入更多的因素，如耕地类型、地形、土壤质地等因素，增加验证样本数量，使验证样本相对于本研究区域更具有代表性，为以后该区域野外采样方案提供参考意见，减少野外采样成本。

不同样本数的不同建模样本选择方法的模型精度对比以及在固定验证样本下不同建模样本选择方法达到一定的预测精度所需的最少的建模样本对比表明：具有多种土壤要素考虑的建模集更具有代表性，土壤类型对建模样本选择方法的优化具有可行性与必要性。

## 4 结论

本文通过对不同建模样本选择方法的模型精度对比，比较了不同建模样本选择方法的构建对模型精度的影响。结果表明浓度梯度法和 KS 方法所选的建模样本集所建立的模型预测能力较差。加入土壤类型后使所选择的建模样本更有代表性，模型精度得到提高。不同级别的样本数下最佳建模样本选择方法不同，但整体表现为土壤多种信息结合的建模样本选择方法模型精度较高。在固定验证样本下不同建模样本选择方法预测模型精度对比表明，浓度梯度法、KS 方法及 RKS 方法 3 种方法建模样本数至少要分别达到总建模样本数的 40%、30%、40% 时，才能保证模型精度较好。土壤类型结合浓度梯度法与土壤类型结合 KS 方法在建模样本数占总建模样本数的 20% 时，就能保证很好的建模精度，有效地减少了建模样本数，减少了建模成本。

## 参考文献：

[1] 龙军, 张黎明, 毛艳玲, 等. 福建省不同耕地土壤和土地利用类型对“碳源/汇”的贡献差异研究[J]. 土壤学报, 2013, 50(4): 664–674.

[2] 蒙继华, 吴炳方, 杜鑫, 等. 遥感在精准农业中的应用进展及展望[J]. 国土资源遥感, 2011(3): 1.

[3] 李冠稳, 高小红, 杨灵玉, 等. 不同粒径土壤有机质含量可见光-近红外光谱估算研究——以湟水流域为例[J]. 土壤通报, 2017, 48(6): 1360–1370.

[4] 陈奕云, 赵瑞瑛, 齐天赐, 等. 结合光谱变换和 Kennard-Stone 算法的水稻土全氮光谱估算模型校正集构建策略研究[J]. 光谱学与光谱分析, 2017, 37(7): 2133–2139.

[5] 杨梅花, 赵小敏. 基于可见-近红外光谱变量选择的土壤全氮含量估测研究[J]. 中国农业科学, 2014, 47(12): 2374–2383.

[6] 郭斗斗, 黄绍敏, 张水清, 等. 多种潮土有机质高光谱预测模型的对比分析[J]. 农业工程学报, 2014, 30(21): 192–200.
[7] 陈奕云, 齐天赐, 黄颖菁, 等. 土壤有机质含量可见光-近红外光谱反演模型校正集优选方法[J]. 农业工程学报, 2017, 33(6): 107–113.
[8] 刘艳芳, 卢延年, 郭龙, 等. 基于地类分层的土壤有机质光谱反演校正样本集的构建[J]. 土壤学报, 2016, 53(2): 332–341.
[9] 邬登巍, 张甘霖. 母质与土地利用类型对土壤光谱反演模型的影响[J]. 土壤, 2016, 48(1): 173–179.
[10] Liu Y L, Jiang Q H, Fei T, et al. Transferability of a visible and near-Infrared model for soil organic matter estimation in riparian landscapes[J]. Remote Sensing, 2014, 6(5): 4305–4322.
[11] 刘伟, 赵众, 袁洪福, 等. 光谱多元分析校正集和验证集样本分布优选方法研究[J]. 光谱学与光谱分析, 2014, 34(4): 947–951.
[12] 纪文君, 史舟, 周清, 等. 几种不同类型土壤的 VIS-NIR 光谱特性及有机质响应波段[J]. 红外与毫米波学报, 2012, 31(3): 277–282.
[13] 赵小敏, 杨梅花. 江西省红壤地区主要土壤类型的高光谱特性研究[J]. 土壤学报, 2018, 55(1): 31–42.
[14] 张颖帝, 张佳宝, 李晓鹏. 基于高光谱的砂姜黑土含水量反演研究[J]. 土壤, 2017, 49(3): 630–634.
[15] 邱玉利, 周建中, 马林. 湟水流域地表水资源特征[J]. 水资源与水工程学报, 2017, 18(6): 98–102.
[16] 雷志栋, 杨诗秀, 谢森传. 土壤水动力学[M]. 北京: 清华大学出版社, 1988.
[17] 黄明祥, 王珂, 史舟, 等. 土壤高光谱噪声过滤评价研究[J]. 光谱学与光谱分析, 2009, 29(3): 722–725.
[18] 周倩倩, 丁建丽, 唐梦迎, 等. 干旱区典型绿洲土壤有机质的反演及影响因素研究[J]. 土壤学报, 2018, 55(2): 313–324.
[19] 于雷, 洪永胜, 耿雷, 等. 基于偏最小二乘回归的土壤有机质含量高光谱估算[J]. 农业工程学报, 2015, 31(14): 103–109.
[20] Chang C W, Laird D, Mausbach M J. Near infrared reflectance spectroscopy-principal components regression analyses of soil properties[J]. Soil Science Society of America Journal, 2001, 65(2): 480–490.
[21] 史舟, 王乾龙, 彭杰, 等. 中国主要土壤高光谱反射特性分类与有机质光谱预测模型[J]. 中国科学: 地球科学, 44(5): 978–988.
[22] 彭杰, 周清, 张杨珠, 等. 有机质对土壤光谱特性的影响研究[J]. 土壤学报, 2013, 50(3): 517–524.