DOI: 10.13758/j.cnki.tr.2019.05.027

基于传递函数的土壤数据库缺失数据的填补研究①

韩光中[1]，杨银华[2]，吴 彬[2]，李山泉[3*]

(1 内江师范学院地理与资源科学学院，四川内江 641112；2 内江市东兴区气象局，四川内江 641100；
3 邢台学院资源与环境学院，河北邢台 054001)

摘 要：数据缺失在土壤调查研究中是一个非常普遍的现象，处理不当一定程度上会影响研究结果的可靠性。土壤转换函数(pedotransfer functions，PTFs)是简单、快速、大批量填补土壤数据库缺失信息的有效手段。但目前分析和厘定我国土壤数据库缺失数据特征的研究较少，针对土壤数据库缺失数据的填补方法也亟待规范。本文对我国第二次土壤普查数据库进行分析，探讨该数据库的数据缺失特征，并对数据缺失严重的土壤属性进行预测，以期为今后的土壤数据库缺失数据填补工作提供参考。总体来看，质地(砂粒、粉粒和黏粒含量)、pH、有机质、全氮、全磷、全钾是土壤普查中最基础的调查项目，这些土壤属性信息的完整性最好。有效磷、速效钾和阳离子交换量数据有一定的缺失。碱解氮、容重、砾石含量、各种类型氧化铁数据缺失严重。在填补缺失数据时，建议首先考虑模型的稳定性，尽量使用那些相对稳定且数据完整性好的土壤属性来预测缺失数据。我国第二次土壤普查数据库基本都缺少空间属性信息，在填补缺失数据时最好采用简单而相对稳定的回归模型。利用回归分析得到的土壤传递函数可以较好地实现容重、碱解氮和部分阳离子交换量缺失数据的填补工作。尽管如此，由于部分土壤属性信息有一定的时效性，应用传递函数时要注意数据源的历史背景。

关键词：土壤数据库；数据缺失；传递函数；数据填补

中图分类号：S159.2 文献标识码：A

数据缺失在土壤调查研究中是一个非常普遍的现象，处理不当一定程度上会影响研究结果的可靠性。20 世纪 80 年代后期，土壤传递函数(pedotransfer functions，PTFs)的出现为简单、快速、大批量填补土壤数据库缺失信息提供了有效的手段。作为土壤属性推绎模型(soil attribute inference models)的典型代表，土壤传递函数的核心功能就是利用已知的土壤信息预测和估算那些缺失或难以直接观测或观测成本高昂的土壤属性[1-2]。

目前，国内已有的土壤传递函数多集中应用于预测土壤容重[3-6]、土壤水力特性[7-9]等。分析和厘定我国土壤数据库数据缺失特征的研究较少，针对土壤数据库缺失数据的填补方法也亟待规范。本文对我国第二次土壤普查数据库进行分析，探讨该数据库的数据缺失特征，并对数据缺失严重的土壤属性进行预测，以期为今后的土壤数据库缺失数据填补工作提供参考。

1 材料与方法

1.1 数据来源

数据来自全国第二次土壤普查成果资料[10]，覆盖了我国主要的土壤类型。第二次土壤普查成果资料通过与中国土壤系统分类的近似参比[11-14]，转化成土壤系统分类体系。通过对数据进行质量筛选，剔除部分异常值，现有完整的 8 731 条数据代表了我国主要的土壤类型。异常值剔除条件为：砂粒、粉粒和黏粒三者百分含量大于 106% 或小于 94%[15]。

基于中国土壤系统分类的土纲检索[11]，将上述数据大致分选为 12 个土纲，灰土和火山灰土土纲本研究未涉及。这是因为第二次土壤普查中的发生学土种均不属于灰土土纲，也就是说我国尚未发现属于中国土壤系统分类体系的灰土剖面。火山灰土土纲只有极少量的数据，这可能主要是因为我国火山灰土的分布面积极少，因此本研究也不再涉及。土壤粒径大小

①基金项目：四川省科技计划项目(2018JY0527)、四川省教育厅重点项目(17ZA0223)和内江师范学院成果转化重大培育项目(17CZ03)资助。
* 通讯作者(li-9yuyan@163.com)
作者简介：韩光中(1981—)，男，山东费县人，博士，副教授，主要从事土壤发生与土壤退化研究。E-mail: hanguangzhong@163.com

的分级采用 1978 年制定的中国土壤质地粒级分类标准：砂粒(1 ~ 0.05 mm)、粉粒(0.05 ~ 0.001 mm)和黏粒(<0.001 mm)。

1.2　研究方法

对全国第二次土壤普查数据进行统计，计算土壤属性信息的完整度，公式如下：

$$\text{土壤属性信息的完整度} = x_i/N \times 100\% \quad (1)$$

式中：x_i：含有土壤某一属性的土壤样本个数；N：所有土壤样本个数。

再使用 IBM Statistics SPSS 20.0，利用回归分析，得到各土壤属性传递函数。利用 R^2 对各传递函数的预测精度进行评价。

2　结果

2.1　土壤属性信息的完整性

根据土壤属性信息的完整度将我国第二次土壤普查数据样本分为 4 个等级(表 1)。从土壤属性信息的完整性上看，质地(砂粒、粉粒和黏粒含量)、pH、有机质(SOM)、全氮(TN)、全磷(TP)、全钾(TK)数据完整度大于 80%，是土壤普查中最基础的调查项目。这些土壤属性信息的完整性好，不会影响到土壤数据库的利用。质地、TP 和 TK 在土壤中相对稳定，没有外部干扰的情况下，短期内一般很少发生重大变化，可以用作土壤属性预测的基础值。有效磷(AP)和速效钾(AK)数据完整度介于 50% ~ 80%，有一定的缺失，碱解氮(AN)数据缺失比较严重。它们是农业上最常用的土壤速效肥力指标，这些属性信息的缺失会明显限制对已有数据库的开发利用。而且这些土壤属性短期内容易发生变化，尤其是在人为干扰或自然侵蚀等状态下，需要通过传递函数演绎出当时的数据作为历史参考数据，供以后的研究对照。阳离子交换量(CEC)数据完整度介于 50% ~ 80%，有一定的缺失，总体而言可以满足数据库的需要。CEC 相对稳定，没有外部干扰的情况下短期内也一般很少发生较大变化，如果研究条件许可，可以先补测数据；同时考虑到 CEC 的测试方法比较复杂，当大规模补测数据不现实时，可以通过传递函数对缺失数据进行填补。容重是土壤的一个重要物理性质，可用作计算土壤持水力和导水性[16-17]，也可用作土壤属性数据的换算，如有机碳库计算[18]中将质量分数换算为体积分数等，因此是土壤学很多模型中的一个必要参数。由于一些土壤中植物根系和砾石较多，很难或无法通过环刀法采样来测定土壤容重；此外，系统获取大量的土壤容重数据也是一项费时费力，甚至不切实际的工作[19]，这造成我国大多数土壤数据库缺失或部分缺失土壤容重数据，利用传递函数预测土壤容重具有重要意义。氧化铁是中国土壤系统分类富铁土、水耕人为土等土壤类型划分中的一个非常重要的指标，在以往的研究中重视不够，测试方法也比较复杂，数据缺失严重。砾石和 $CaCO_3$ 含量的数据信息完整度均不足 15%。考虑到砾石和 $CaCO_3$ 在土壤中不是普遍存在，这一数据不能真实反映砾石和 $CaCO_3$ 含量的缺失程度。但在剖面描述中土壤明显存在大量砾石的粗石土、粗骨土(土壤发生分类)等，剖面数据中并没有统计砾石含量。砾石含量的缺失对计算碳库和评估农业机械运行不利。土壤砾石含量数据一旦缺失，只能在未来的研究中补测数据，今后要注意评估砾石含量对土壤容重、有机碳库估算等方面的影响作用。$CaCO_3$ 多存在于干旱、半干旱区土壤或石灰性母质的土壤中，本研究也不再涉及其填补工作。

表 1　我国第二次土壤普查土壤属性信息的完整度

Table 1　Integrity of soil property information in National Second Soil Survey

级别	数据信息完整度	土壤属性
1	>80%	质地(砂粒、粉粒和黏粒含量)、有机质(SOM)、全氮(TN)、全磷(TP)、全钾(TK)、pH
2	50% ~ 80%	有效磷(AP)、速效钾(AK)、阳离子交换量(CEC)
3	20% ~ 50%	碱解氮(AN)
4	<20%	容重(BD)、砾石含量、游离铁(Fe_d)、全铁(Fe_t)、$CaCO_3$ 含量

2.2　土壤容重的预测

大量研究表明，土壤容重与 SOM 和土壤质地关系密切[20-24]，也与土壤深度[25-26]、土壤类型[27-28]、土地利用和植被[29]有一定的关系。研究还表明，基于土壤系统分类的数据分组可以改进模型的预测精度[30-31]。国内已有大量土壤容重传递函数的研究，如韩光中等[6]基于我国现有土壤数据库，利用 SPSS 采用逐步回归方法确定了我国主要土壤类型的容重传递函数。基于土壤基本理化性质的土壤容重预测工作目前已经比较完善(表 2)，可以很好地实现土壤容重缺失数据的填补。

2.3　土壤速效养分的预测

AN、AP 和 AK 是农业上最常用土壤速效肥力指标，在全国第二次土壤普查中，它们还是土壤肥力分

表 2 不同土壤类型的最优容重传递函数[3,6]
Table 2 Optimal PTFs of BD for different types of soils

土纲	土壤容重传递函数
有机土	lnBD = 0.373–0.0028SOM
人为土	lnBD = 0.407–0.019SOM+0.028(lnSOM)^2+0.001clay
铁铝土	BD = 0.186×1.541/[1.541SOM+0.186(1–SOM)]
干旱土	lnBD = 0.277–0.0019depth
盐成土	lnBD = 0.407–0.0069SOM
潜育土	lnBD = 0.215–0.0025SOM+0.0017depth
均腐土	lnBD = 0.341–0.054SOM+0.0006depth
富铁土	lnBD = 0.283–0.0039SOM–0.040TN+0.0022depth
淋溶土	BD = 0.197×1.506/[1.506SOM+0.197(1–SOM)]
雏形土	BD = 0.156×1.538/[1.538SOM+0.156(1–SOM)]
新成土	BD = 0.154×1.529/[1.529SOM+0.154(1–SOM)]
变性土	lnBD = 0.436–0.0103SOM+0.0006depth
所有土纲	lnBD = 0.4345–0.0356SOM$^{0.5}$–0.0007SOM–0.0215TN+0.0001Clay

注：BD：土壤容重(g/cm^3)；SOM：土壤有机质(淋溶土、雏形土和新成土模型中单位为 g/g，其他模型中单位为 g/kg)；clay：黏粒(%)；depth：土壤深度(cm)。

级的主要依据[32]。本研究利用 SPSS 采用逐步回归方法确定了土壤速效养分的传递函数(表 3)。从预测结果上看，AN 的传递函数预测精度比较高(R^2 为 0.754)，可以用来填补缺失数据；而 AP 和 AK 传递函数的预测精度很低。考虑到 AK 和 AP 数据缺失度不大，建议尽量避免使用本研究提议的传递函数来填补缺失数据。因为土壤速效养分容易变化，其观测结果有一定的时效性，所以在应用传递函数预测 AN 时要注意数据源的历史背景。

2.4 CEC 的预测

土壤 CEC 是土壤的基本特性和主要肥力影响因素之一，直接反映土壤保蓄、供应和缓冲阳离子养分的能力，同时影响其他土壤理化性质。土壤 CEC 常被作为土壤资源质量的评价指标和土壤施肥、改良的主要依据[33-34]。考虑到土壤 CEC 的测试方法比较复杂，大规模补测数据不现实，利用土壤其他属性来预测土壤 CEC 具有重要意义。

从表 4 中可以看出，有机土、潜育土和变性土

表 3 土壤速效养分的最优传递函数
Table 3 Optimal PTFs of soil available nutrients

样本个数	速效养分传递函数	R^2
2 241	AN = 55.52+62.90TN–5.92pH	0.754
3 896	AP = 1.534+10.25TP	0.360
4 115	AK = –17.44+1.01TK+ 12.85pH–0.45sand–17.44	0.119

注：AN：碱解氮(mg/kg)；AP：有效磷(mg/kg)；AK：速效钾(mg/kg)；sand：砂粒(%)；TN：全氮(g/kg)；TP：全磷(g/kg)；TK：全钾(g/kg)。

表 4 不同土壤类型的最优 CEC 传递函数
Table 4 Optimal PTFs of CECs for different types of soils

样本数	土纲	CEC 传递函数	R^2
14	有机土	CEC = 15.392+0.214SOM	0.915
880	(水耕)人为土	CEC = 5.204+0.23clay+0.164SOM	0.236
94	(旱耕)人为土	CEC = 1.815+0.42clay	0.538
32	铁铝土	CEC = 5.306+0.127SOM+0.005TK	0.467
150	干旱土	CEC = 45.32+0.295SOM–4.22pH	0.151
155	盐成土	CEC = 1.257+0.283TK+0.18clay+0.06sand	0.439
89	潜育土	CEC = 55.93+0.073SOM–5.14pH	0.699
550	均腐土	CEC = 27.15+0.221SOM+0.22clay–2.26pH	0.477
60	富铁土	CEC = 15.93+2.76TP–0.14slit	0.267
620	淋溶土	CEC = 9.89+0.27clay+1.72TN	0.256
1 741	雏形土	CEC = 6.96+3.22TN+0.26clay	0.344
440	新成土	CEC = 3.574+0.40clay+0.181SOM	0.528
74	变性土	CEC = 35.66+14.49TN–0.38slit	0.723

注：CEC：土壤阳离子交换量(cmol/kg)；SOM：土壤有机质(g/kg)；clay：黏粒(%)；slit：粉粒(%)；sand：砂粒(%)；TN：全氮(g/kg)；TK：全钾(g/kg)。

CEC 传递函数的预测精度最高(R^2> 0.669)，旱耕人为土和新成土 CEC 传递函数的预测精度较高(R^2> 0.528)，传递函数在这些土纲可以较好地实现 CEC 缺失数据的填补工作。均腐土、盐成土和铁铝土 CEC 传递函数的预测精度中等，水耕人为土、干旱土、富铁土、淋溶土和雏形土 CEC 传递函数的预测精度很低，这些土纲不适合利用本研究提议的 CEC 传递函数来填补缺失数据。

2.5　氧化铁的预测

土壤中的氧化铁是由硅酸盐类矿物在地表特定水热条件下，经过风化作用形成的，它的形态和各种性质易随成土环境的改变而改变，所以它们既是成土过程的产物，也是成土条件的反映[35]。在土壤发生、分类与土壤物理、化学性质以及土壤肥力方面，氧化铁的研究都是极其重要的，如在中国土壤系统分类中是划分水耕人为土、富铁土、铁铝土和淋溶土的一个重要指标[11]。通常将连二亚硫酸钠–柠檬酸钠–重碳酸钠溶液(DCB)提取的铁称为游离铁(Fe_d)，包含了土壤中所有形态的铁氧化物。尽管如此，土壤 Fe_d 的测试方法比较复杂，大批量测试费时费力，成本高昂，这造成我国大多数土壤数据库严重缺失土壤 Fe_d 数据。传递函数可以较好地实现水耕人为土、铁铝土、富铁土和淋溶土 Fe_d 的预测(表 5)。尽管如此，土壤数据库中 Fe_t 的数据也缺失严重，所以大批量填补土壤 Fe_d 缺失数据仍然非常困难。但在一些研究中，如果对样品测试精度要求不太高，可以利用土壤已有数据来估算土壤 Fe_d 含量。

表 5　土壤氧化铁的最优传递函数
Table 5　Optimal PTFs of soil Fe_d

样本数	土纲	Fe_d 传递函数	R^2
152	水耕人为土	$Fe_d = 0.870Fe_t–0.893TK+1.20$	0.790
32	铁铝土	$Fe_d = 0.793Fe_t+0.525TK–18.01$	0.875
90	富铁土	$Fe_d = 0.676Fe_t+0.026clay–11.13$	0.683
113	淋溶土	$Fe_d = 0.665Fe_t–7.06pH+26.46$	0.621
55	雏形土	$Fe_d = 0.464Fe_t+0.27clay–4.64$	0.365
474	所有土纲	$Fe_d = 0.692Fe_t–0.348TK–2.64pH+13.09$	0.660

注：clay：黏粒(%)；Fe_t：全铁(g/kg)；TK：全钾(g/kg)。

3　讨论

本研究的建模数据集主要来自于全国第二次土壤普查数据库，具有一定的时效特征，这可能会影响到传递函数在其他研究中的应用。在填补缺失数据时，本研究首先考虑模型的稳定性。在保证精度的前提下，尽可能地少选参数，使用那些相对稳定且数据完整性好的土壤属性，如质地、pH、SOM、TN、TP 和 TK 等，来预测缺失数据。我国第二次土壤普查数据库基本都缺少空间属性信息，在填补缺失数据时采用回归模型，未考虑地形、气候、母质等因素。一方面是因为经验模型在应用时预测精度相对较稳定，另一方面是因为地形、气候、母质等因素的数据量大，完整获取较困难。这些因素应在今后的深入研究中加以考虑。

需要特别指出的是，将中国土壤发生分类体系转为中国土壤系统分类体系，是一个非常复杂的过程。即使在土纲级别的转换仍只是大致的，一般一个发生学土纲对应 1 ~ 3 个系统分类土纲[36]。这一转换的不精确性会影响到传递函数的预测精度。不管土壤发生分类还是土壤系统分类，同土纲内的土壤仍复杂多变，影响土壤属性的非土壤因素多而复杂，这也会影响到传递函数的预测精度。此外，土壤传递函数的预测精度和稳定性与建模样本数量也有密切关系。有机土和铁铝土的建模样本数比较少，在应用土壤传递函数填补缺失数据时一定要特别小心。

本研究中，AK 和 AP 传递函数的预测精度偏低。可能是因为提议模型没有反映影响土壤吸附或固定的因素，如 Fe_d、$CaCO_3$、酶活性、根际分泌物和黏土矿物类型等指标在数据库中多是缺失或无法定量的。类似的，水耕人为土、干旱土、富铁土、淋溶土和雏形土 CEC 传递函数和雏形土 Fe_d 传递函数的预测精度也很低。这可能同样是因为影响因素比较复杂，在以后的研究中需要重视。

4　结论

质地(砂粒、粉粒和黏粒含量)、pH、SOM、TN、TP、TK 是土壤普查中最基础的调查项目，这些土壤属性信息的完整性最好。AP、AK 和 CEC 数据有一定的缺失。AN、容重、砾石含量和各种类型氧化铁数据缺失严重。在填补缺失数据时，建议首先考虑模型的稳定性，尽量使用那些相对稳定且数据完整性好的土壤属性来预测缺失数据。我国第二次土壤普查数据库基本都缺少空间属性信息，在填补缺失数据时最好采用简单而相对稳定的回归模型。利用回归分析得到的土壤传递函数可以较好地实现容重、AN 和部分 CEC 缺失数据的填补工作。尽管如此，由于部分土壤属性信息有一定的时效性，应用传递函数时要注意数据源的历史背景。

参考文献：

[1]　Wöstena J H M, Pachepsky Y A, Rawls W J. Pedotransfer functions: Bridging gap between available basic soil data

and missing soil hydraulic characteristics[J]. Journal of Hydrology, 2001, 251: 42–49
[2] Mcbratney A B, Minasny B, Cattle S. et al. From pedotransfer functions to soil inference systems[J]. Geoderma, 2002, 109: 41–73
[3] Han G Z, Zhang G L, Gong Z T, et al. Pedotransfer functions for estimating soil bulk density in China[J]. Soil Science, 2012, 177(3): 158–164
[4] 刘继红, 兰传宾, 陈杰. 区域土壤容重转换函数构建与预测结果评价——以河南省封丘县为例[J]. 土壤通报, 2013, 44(1): 77–82
[5] 王巧利, 林剑辉, 许彦峰. 基于 BP 神经网络的土壤容重预测模型[J]. 中国农学通报, 2014, 30(24): 237–245
[6] 韩光中, 王德彩, 谢贤健. 中国主要土壤类型的土壤容重传递函数研究[J]. 土壤学报, 2016, 53(1): 93–102
[7] 廖凯华, 徐绍辉, 程桂福, 等. 基于不同 PTFS 的流域尺度土壤持水特性空间变异性分析[J]. 土壤学报, 2010, 47(1): 33–41
[8] 王改改, 张玉龙. 土壤传递函数模型的研究进展[J]. 干旱地区农业研究, 2012, 30(1): 99–103
[9] 邹刚华, 李勇, 李裕元, 等. 亚热带小流域稻田土壤饱和导水率传递函数构建[J]. 土壤通报, 2013, 44(2): 302–307
[10] 中国土壤普查办公室. 中国土种志(1–6)[M]. 北京: 中国农业出版社, 1995
[11] 中国科学院南京土壤研究所土壤系统分类课题组&中国土壤系统分类课题研究协作组. 中国土壤系统分类检索[M]. 3 版. 合肥: 中国科学技术大学出版社, 2001
[12] 龚子同, 张甘霖, 陈志诚, 等. 以中国土壤系统分类为基础的土壤参比[J]. 土壤通报, 2002, 33(1): 1–5
[13] 杨国祥, 史学正, 于东升, 等. 基于 WebGIS 的中国土壤参比查询系统研究[J]. 土壤学报, 2007, 44(1): 1–6
[14] 李德成, 张甘霖, 龚子同. 我国砂姜黑土土种的系统分类归属研究[J]. 土壤, 2011, 43(4): 623–629
[15] Heuscher S A, Brandt C C, Jardine P M. Using soil physical and chemical properties to estimate bulk density[J]. Soil Science Society of America Journal, 2005, 69: 1–7
[16] Lenaviciute N. Predicting soil bulk and particle densities by pedotransfer functions from existing soil data in Lithuania[J]. Geografijos metrastis, 2000, 33: 317–328
[17] Marco A, Marcello D. SOILPAR 2. 00: Software to estimate soil hydrological parameters and functions[J]. European Journal of Agronomy, 2003, 18: 373–377
[18] Yu D S, Shi X Z, Wang H J, et al. Regional patterns of soil organic carbon stocks in China[J]. Journal of Environmental Management, 2007, 85(3): 680–689
[19] Benites V M, Machado P L O A, Fidalgo E C C, et al. Pedotransfer functions for estimating soil bulk density from existing soil survey reports in Brazil[J]. Geoderma, 2007, 139: 90–97
[20] Post W M, Kwon K C. Soil carbon sequestration and land-use change: Processes and potential[J]. Global Change Biology, 2000, 6: 317–327
[21] Tremblay S, Ouimet R, Houle D. Prediction of organic carbon content in upland forest soils of Quebec, Canada[J]. Canadian Journal of Forest Research-revue Canadienne De RechercheForestiere, 2002, 32: 1–12
[22] Kaur R, Kumar S, Gurung H P. A pedo-transfer function (PTF) for estimating soil bulk density from basic soil data and its comparison with existing PTFs[J]. Australian Journal of Soil Research, 2002, 40: 847–857
[23] Prévost M. Predicting soil properties from organic matter content following mechanical site preparation of forest soils[J]. Soil Science Society of America Journal, 2004, 68: 943–949
[24] Périé C, Ouimet R. Organic carbon, organic matter and bulk density relationships in boreal forest soils[J]. Canadian Journal of Soil Science, 2008, 88: 315–325
[25] Huntington T G, Johnson C E, Johnson A H, et al. Carbon, organic matter and bulk density relationships in a forested Spodosol[J]. Soil Science, 1989, 148: 380–386
[26] Leonavičiutė N. Predicting soil bulk and particle densities by pedotransfer functions from existing soil data in Lithuania[J]. Geografijos metraštis, 2000, 33: 317–330
[27] Alexander E B. Bulk densities of California soils in relation to other soil properties[J]. Soil Science Society of America Journal, 1980, 44: 689–692
[28] Salifu K F, Meyer W L, Murchison H G. Estimating soil bulk density from organic matter content, pH, silt and clay[J]. Journal of Tropical Forest Science, 1999, 15: 112–120
[29] Harrison A F, Bocock K L. Estimation of soil bulk-density from loss-on-ignition values[J]. Journal of Applied Ecology, 1981, 8: 919–927
[30] Manrique L A, Jones C A. Bulk density of soils in relation to soil physical and chemical properties[J]. Soil Science Society of America Journal, 1991, 55: 476–481
[31] Wösten J H M, Pachepsky Y A, Rawls W J. Pedotransfer functions: Bridging the gap between available basic soil data and missing soil hydraulic characteristics[J]. Journal of Hydrologic Engineering, 2001, 251: 123–150
[32] 全国土壤普查办公室. 中国土壤[M]. 北京: 中国农业出版社, 1998
[33] 黄昌勇. 土壤学[M]. 北京: 中国农业出版社, 2000
[34] 李学垣. 土壤化学[M]. 北京: 高等教育出版社, 2001
[35] Schwertmann U. The effect of pedogenic environments on iron oxide minerals[J]. Advances in Soil Science, 1985, 1: 171–200
[36] 赵其国, 史学正, 等. 土壤资源概论[M]. 北京: 科学出版社, 2007

Missing Data Imputation Approach for Soil Database Based on Pedotransfer Functions

HAN Guangzhong[1], YANG Yinhua[2], WU Bin[2], LI Shanquan[3*]

(1 *College of Geography and Resources Science*, *Neijiang Normal University*, *Neijiang*, *Sichuan* 641112, *China*; 2 *Dongxing Meteorological Bureau*, *Neijiang*, *Sichuan* 641100, *China*; 3 *College of Resources and Environment*, *Xingtai University*, *Xingtai*, *Hebei* 054001, *China*)

Abstract: Data missing is a common problem in soil survey and related researches. When this problem proposed, the common solution in most studies was to neglect it or remove records that have missing data due to the lack of the understanding of the importance of data missing. Obviously, this solution could not to satisfy the needs of practical studies. The application of pedotransfer functions (PTFs) provides a broad prospect for the interpolation of missing data of soil database in a simple, rapid and batch processing way. At present, few studies were carried to analyze or interpolate the missing data of soil database in China. More importantly, the method to interpolate the missing data of soil database needs to be standardized. In this study, the characteristics of missing data in Chinese Soil Database from the Second National Soil Survey were analyzed, and the interpolations of serious missing date of soil properties were tried in order to provide knowledge for future researches. Results showed, the data of soil texture (sand, silt and clay contents), pH, organic matter, total nitrogen, total phosphorus and total potassium were most complete for they are the basic survey factors in soil survey. The data of available phosphorus, available potassium and cation exchange capacity had a certain miss. The data of alkaline nitrogen, bulk density and iron oxides were missed seriously. Considering that the stability of prediction model is essential, so soil properties with complete data would be used with top priority in the interpolation of data missing. The existing Chinese Soil Database is in short of spatial attribution data, so it is better to use regression model than use spatial interpolation in the interpolation of missing data of soil database. In this study, PTFs from regression analysis could meet the requirement of data interpolation of bulk density, alkaline nitrogen and partial cation exchange capacity. Besides, some soil properties such as available potassium could be time limited, so the historical background of data sources should be considered in the application of PTFs.

Key words: Soil database; Data missing; Pedotransfer functions; Data imputation