Catena 126 (2015) 98–104

Contents lists available at ScienceDirect

Catena

journal homepage: www.elsevier.com/locate/catena

# Soil distribution modeling using inductive learning in the eastern part of permafrost regions in Qinghai–Xizang (Tibetan) Plateau

Wangping Li [a,b], Lin Zhao [a,*], Xiaodong Wu [a,*], Shijie Wang [c], Yu Sheng [a], ChienLu Ping [d], Yonghua Zhao [a], Hongbing Fang [a], Wei Shi [a]

[a] Cryosphere Research Station on Qinghai–Xizang Plateau, State Key Laboratory of Cryospheric Sciences, Cold and Arid Regions Environmental and Engineering Research Institute, Chinese Academy of Sciences, Lanzhou, Gansu 730000, China
[b] University of Chinese Academy of Sciences, Beijing 100049, China
[c] Faculty of Geomatics, LanZhou Jiaotong University, Lanzhou, Gansu 730000, China
[d] University of Alaska Fairbanks Agricultural and Forestry Experiment Station Palmer, AK 99645, United States

ARTICLE INFO

*Article history:*
Received 20 December 2013
Received in revised form 25 September 2014
Accepted 27 October 2014
Available online 22 November 2014

*Keywords:*
Soil–landscape model
Decision tree
SoLIM
Permafrost
Soil classification

ABSTRACT

Soil–landscape models serve as a basis for understanding the relationships between soils and landscape, and such models also allow us to study soil distribution, and classification. To understand the distribution of soils and landform relationships in permafrost regions, we developed a soil–landscape model by using See5.0 decision tree software in the Wenquan region of Xinghai County, in the eastern part of permafrost regions of Qinghai–Xizang (Tibetan) Plateau (QTP). The model was based on soil pedon data from 2009. Nine environmental factors closely related to permafrost-affected-soil formation were selected as variables for the model: land surface temperature for warm season and cool season, elevation, slope gradient, slope aspect, planform and profile curvatures, wetness index and NDVI. A 5-fold cross-validation method was applied to verify the effectiveness of the soil–landscape model. The model results were consistent with field observations, and the slope was the most strongly correlated factor with soil type of the nine environmental variables. The soils in the modeled area are mainly Ustic Cambosols and Ustic Isohumisols, which cover about 60.0% and 25.6% of the total area, respectively. Gelic cambosols occur mainly in permafrost region, while Ustic isohumosols occur in the transition region between permafrost and seasonally frozen ground. Further studies are required to utilize the soil–landscape model to predict the spatial distribution of soil types over the QTP.



## 1. Introduction

Soil is a fundamental natural resource that is for the biophysical and biogeochemical functioning of the planet (Scull et al., 2003). Soil–landscape models serve as a basis for understanding the relationships between soil and the environmental conditions of a landscape, and such models also allow us to study soil distribution, and classification (Qi and Zhu, 2003).

Many researchers have explored methods for constructing soil–landscape models based on soil maps or soil samples, such as multiple linear regression (Sun et al., 2008b), similarity modeling (Zhu, 1997), neural network (Behrens et al., 2005; Chagas et al., 2013; Zhu, 2000), and support vector machines (SVM) (Shi et al., 2011). These methods all involve implicit classification rules that are difficult to express clearly and cannot be modified or improved. Recently, the See5.0 decision tree algorithm was applied to the study of soil classification (Qi and Zhu, 2011; Taghizadeh-Mehrjardi et al., 2014). This algorithm like other tree-based algorithms, offers a number of advantages (Lacoste et al., 2011). First, it can handle both continuous and discrete data (Bou Kheir et al., 2010). Second, it is independent of the distribution of sampling sites (Zhou et al., 2004), and the classification rules mined by the algorithm are readily interpretable (Qi and Zhu, 2003). Third, it can generate higher accuracy than many efforts on the soil of the region (Schmidt et al., 2008). Other advantages of the See5.0 algorithm include winnow and order attributes according to their importance (Quinlan, 2001).

The Qinghai–Xizang (Tibetan) Plateau (QTP) is the largest geomorphological unit of the Eurasian continent, and it is the largest low-latitude permafrost region in the world. Permafrost-affected soils are strongly affected by thawing and freezing, and their characteristics can vary (Baumann et al., 2014; Chen et al., 2005). Little is known about the relationships between soil, environmental conditions and soil distribution on the QTP. Considering the extensive permafrost in the region, it is extremely important to understand the relationships between soil and environmental conditions, particularly for those areas that currently lack data and are unpopulated (Wu et al., 2012).

Furthermore, the QTP is sensitive to climate change (Yashiro, et al., 2010); therefore, there is good reason to focus research efforts on the soil of the region. Soil–landscape models can help us to understand

* Corresponding authors.
*E-mail addresses:* linzhao@lzb.ac.cn (L. Zhao), wxd565@163.com (X. Wu).

http://dx.doi.org/10.1016/j.catena.2014.10.025
0341-8162/

the effects of permafrost distribution and development on the soil types and to promote the study of permafrost distribution, degradation processes, mechanisms and preventive treatments. The purpose of this study is to use soil profiles from field investigations in the permafrost regions of the eastern QTP and employ software to explore and evaluate statistically meaningful relationships involving soil distribution patterns and landform features of the eastern QTP. The results will offer some insight into the terrestrial ecosystems response to environmental changes, as well as provide datasets for further ecological studies of this important area.

## 2. Study area

The study area is in Wenquan District, and it was selected because it is a typical transitional area between permafrost and seasonally frozen ground in the northeastern part of the QTP (Fig. 1**a**). The region is a vast plateau with elevations of 3405–5294 m asl. Two high and steep mountain ranges, the Ela and Jiangluling mountains, cross the study area in a north-west–southeast pattern. The Qing-Kang Road traverses the area in the northeast–southwest direction. Additionally, there are two basins, the Wenquan Basin and the Kuhai Basin, which have lower elevations and flat terrain (Fig. 1**b**).

The annual mean temperature is about −3.2 °C, and the annual precipitation is between 500 and 600 mm in the study area (Chou et al., 2009). The annual potential evaporation is about 1264 mm (Zhang et al., 2010).

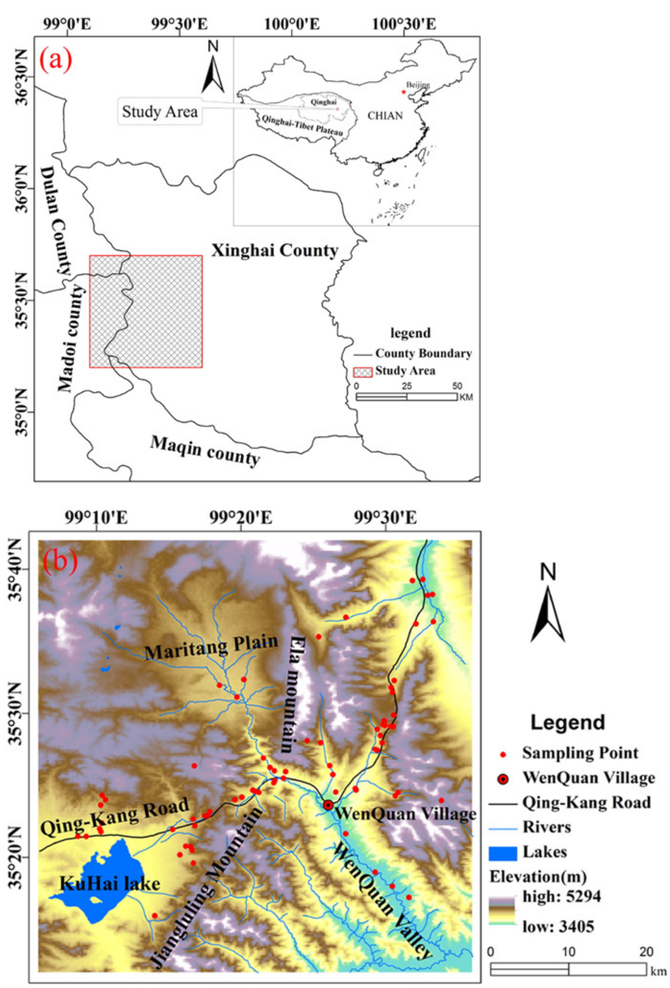


**Fig. 1.** Map of the study area (a) and sampling plot locations (b).

According to a field investigation from 2009, the study region is primarily a grassland ecosystem, with the main vegetation communities being alpine steppe, meadow, wet meadow, and smaller areas of alpine shrub. Alpine steppe is the most prevalent type of community in the high plains, with the main plant species being *Stipa purpurea*. Alpine meadow, and alpine wet meadow are mainly distributed in the low mountains and water-filled depression areas, with the main plant species being *Kobresia pygmaea*, *Kobresia humilis*, *Carex tris-tachya*, and *Kobresia tibetica*. Alpine shrubs are distributed mainly in the northern slopes of the low elevation mountains, with the main plant species being *Salix oritrepha*, *Potentilla fruticosa*, and *Caragana jubata*. The parent material for the soil in the study region consists primarily of carbonates rocks, clastic rocks, volcanic rocks, metamorphic rocks, glacial outwash and moraine (Li et al., 2004; Wang et al., 2004).

## 3. Data and methods

### *3.1. Remote sensing and field investigation data*

The basic data used to build our model include a 30 m resolution Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) Global Digital Elevation Model (GDEM) provided by the International Scientific Data Service Platform, 250 m resolution average 16 d NDVIs over a 10-year period (2000–2009) provided by NASA, and 1 km resolution seasonal mean surface temperature for 6 years (2003–2008), also provided by NASA.

Fieldwork was undertaken from September 12 to October 20, 2009. A total of 73 soil pits were excavated with a mechanical digger (Fig. 1**b**). Each sample was documented by recording details about the soil profile, soil properties (color, texture, structure, etc.), topography information (slope and aspect), and the dominant vegetation species and cover. The soil samples were subjected to laboratory analysis. According to the "Chinese Soil Taxonomic Classification" (Gong, 1999), the soils are classified as Ustic cambosols, Aquic cambosols, Gelic cambosols, Permagelic gleyosols, and Ustic isohumosols (suborder). Representative physical and chemical properties of five soils are shown in the Table 1. The design of sampling points complied with the following principles: the sampling points were distributed evenly along the Qing-Kang Road. Adjustments were made to account for altitude, slope, aspect and the general distributed information for permafrost. Furthermore, we had to adjust the sampling points slightly according to the observed type of vegetation, the soil conditions and the field accessibility of the sites. To some extent, this sampling scheme, in combination with of the researchers, was able to compensate for the uneven distribution and poor representativeness of the samples.

### *3.2. The decision tree See5.0 software*

Decision trees constitute a nonparametric method for the analysis of hierarchical relationships. This is one of the most efficient forms of inductive learning (Schmidt et al., 2008). Continuous subsets are separated by using different predictive variables, and they are divided until each subset is homogeneous. To ensure that the identified variables are the best choices, they are selected based on the concept of entropy from information theory (Quinlan, 1986). Some of the most famous algorithms include ID3 and its derivatives, C4.5 and C5.0 (Quinlan, 1993, 2001). C4.5 uses the information gain rate to choose decision attributes, and it increases the discretization of continuous attributes, the processing of unknown attributes, the production of rules, and other functions on the basis of ID3. C5.0 uses a boosting method based on C4.5, which makes the calculations faster and improves the accuracy of the rules (Wen et al., 2007). See5.0 software was developed by the Rulequest company on the basis of the C5.0 algorithm. It allows cross validation and pruning of decision trees (http://www.rulequest.com/see5-win.html#DATA).