Land use and climate control the spatial distribution of soil types in the grasslands of Inner Mongolia

F.K. Barthold \textsuperscript{a,\*}, M. Wiesmeier \textsuperscript{b}, L. Breuer \textsuperscript{a}, H.-G. Frede \textsuperscript{a}, J. Wu \textsuperscript{a,c}, F.B. Blank \textsuperscript{a}

\textsuperscript{a} Research Centre for BioSystems, Land Use and Nutrition (IFZ), Institute for Landscape Ecology and Resources Management, Justus-Liebig-University Gießen, Heinrich-Buff-Ring 26, 35392 Gießen, Germany

\textsuperscript{b} Lehrstuhl für Bodenkunde, Department für Ökologie und Ökosystemmanagement, Wissenschaftszentrum für Ernährung, Landnutzung und Umwelt, Technische Universität München, Emil-Ramann-Straße 2, 85350 Freising-Weihenstephan, Germany

\textsuperscript{c} Cold and Arid Regions Environmental and Engineering Research Institute, Chinese Academy of Sciences, West Dongang Road 320, 730000 Lanzhou, China

\textbf{ABSTRACT}

The spatial distribution of soil types is controlled by a set of environmental factors such as climate, organisms, parent material and topography as well as time and space. A change of these factors will lead to a change in the spatial distribution of soil types. In this study, we use a digital soil mapping approach to improve our knowledge about major soil type distributing factors in the steppe regions of Inner Mongolia (China) which currently undergo tremendous environmental change, e.g. climate and land use change. We use Random Forests in an effort to map Reference Soil Groups according to the World Reference Base for Soil Resources (WRB) in the Xilin River catchment. We benefit from the superior prediction capabilities of RF and additional interpretive results in order to identify the major environmental factors that control spatial patterns of soil types. The nine WRB soil groups that were identified and spatially predicted for the study area are Arenosol, Calcisol, Cambisol, Chernozem, Cryosol, Gleysol, Kastanozems, Phaeozem and Regosol.

Model and prediction performances of the RF model are high with an Out-of-Bag error of 51.6\% for the model and a misclassification error for the predicted map of 28.9\%. The main controlling factors of soil type distribution are land use, a set of topographic variables, geology and climate. However, land use and climate are of major importance and topography and geology are of minor importance. The visualizations of the predictions, the variable importance measures as result of RF and the comparisons of these with the spatial distribution of the environmental factors delivered additional, quantitative information of these controlling factors and revealed that intensively grazed areas are subjected to soil degradation. However, most of the area is still governed by natural soil forming processes which are driven by climate, topography and geology. Most importantly though, our study revealed that a shift towards warmer temperatures and lower precipitation regimes will lead to a change in the spatial distribution of RGs towards steppe soils that store less carbon, i.e. a decrease of spatial extent of Phaeozems and an increase of spatial extent of Chernozems and Kastanozems.

\textsuperscript{*} Corresponding author. Present address: Institute of Earth and Environmental Sciences, University of Potsdam, Karl-Liebknecht-Str. 24-25, 14476 Potsdam, Germany. Tel.: +49 331 977 2463; fax: +49 331 977 2068. E-mail address: f.barthold@gmail.com (F.K. Barthold).

\textcopyright 2012 Elsevier Ltd. All rights reserved.

1. Introduction

Soils provide important ecosystem services such as carbon sequestration, water purification, provision of ground for humans to live on and to secure food production among many others (Millenium Ecosystem Assessment, 2005). Changes in soils, i.e. their properties, will consequently have an impact on these services. Soils naturally develop and change through time and space. And environmental factors such as climate, organisms, relief and parent material are also important controls on soil development and distribution (Jenny, 1941; McBratney et al., 2003). A change of any one of these controlling factors may induce a change of soil, i.e. its properties. However, the extents of this change and if this change poses a threat to soils depends on the factors that change and their impact on the soil at a specific location.

Steppe ecosystem, of global importance since they cover about one fifth of the terrestrial land surface (Allard et al., 2007) and storage of large amounts of carbon (C) and nitrogen (N), are
especially vulnerable ecosystems to changes. They are currently undergoing tremendous environmental changes, especially in China, where the quick economic rise causes rapid development of the country and induces large land use changes. The Inner Mongolian Grasslands (China) belong to the regions that are currently undergoing tremendous environmental changes due to the vast economic developments. These changes include altered and intensified grazing regimes as well as increased agricultural production (Akiyama and Kawamura, 2007; Butterbach-Bahl et al., 2011). One of the largest ecologic problems related to the quick development in the East Asian Steppe in this context is the grazing-induced degradation of the sensitive grassland soils that is associated with wind driven soil erosion (Hoffmann et al., 2008; Kaiser, 2004; Wiesmeier et al., 2009, 2012). And an increase of grazing intensity also leads to soil degradation, e.g. it decreases soil aggregation and the amount of fresh litter-like particulate organic matter (Kölbl et al., 2011) while at the same time animal trampling leads to a decrease of infiltration rates, saturated hydraulic and air conductivities (Reszkowska et al., 2011). Grazing in combination with degradation of vegetation patterns enhance soil structure disruption and result in a texture driven wettability of the soil surfaces (Kölbl et al., 2011). The ability to retain water in the topsoil also declines with increasing grazing pressure on short (Schneider et al., 2008) and long time scales (Zhao et al., 2011b). Spatial heterogeneity of soil moisture was shown to be reduced with increased grazing intensity and is mainly controlled by soil properties and vegetation patterns (Zhao et al., 2011a). However, these findings apply on smaller spatial scales, e.g. at the plot scale and there exists an urgent need to find out if the effects of land use changes on soil spatial patterns are visible at larger scales.

In addition to the environmental changes that are due to steppe management, the grasslands in Inner Mongolia are threatened by climate change (Butterbach-Bahl et al., 2011). Smit and Cai (1996) and Polley et al. (2000) predicted an increase of temperature and a decrease in precipitation, particularly in the summer months, which is expected to lead to a reduced primary production since water is the limiting factor in this semi-arid environment (Christensen et al., 2004; Xiao et al., 1995; Xue, 1996). This in turn is predicted to induce, for example, changes in C and N storage as a result of reduced input of organic matter. Xiao et al. (1995) showed that the amount of C and N that can be stored in steppe soils depends on precipitation and temperature. First estimates of C and N storage and the storage capacity depending on precipitation and temperature in the Central Inner Mongolian Steppe indicate that storage is close to an optimum due to the current precipitation and temperature regime (Butterbach-Bahl et al., 2011), Wiesmeier et al. (2010) provide a detailed study of C, N and S storage and spatial patterns in the grasslands of Inner Mongolia, however, the impact of climate is not considered directly and a more detailed study on the impacts of climate change on carbon sequestration and soil nutrient storage is lacking.

There exists an urgent need to improve our knowledge of the relationships between environmental factors and the distribution of soil types to understand how environmental change will affect soils and to provide recommendations for a sustainable land management.

There already exist many studies that explore the relationships between environmental factors and spatial soil patterns using digital soil mapping (see a summary of them in McBratney et al., 2003; Grunwald, 2009) but only a limited amount of them are focused on steppe areas (Florinsky et al., 2002; Wiesmeier et al., 2010).

Hence, the objectives of this study are (a) to identify the main factors that control the spatial distribution of soil types at the catchment scale in the Inner Mongolian Steppe, (b) to assess the relationships between spatial patterns of soil types and environmental factors and (c) to provide a soil map for decision makers in a region that is threatened by soil degradation.

2. Material and methods

2.1. Study area

The study area is a sub catchment of the 10,000 km² large Xilin River Basin (43°24’ to 44°40’ N and 115°20’ to 117°13’ E) which is located in the Xilingol League in central Inner Mongolia A.P., China (Fig. 1a). The sub catchment (Fig. 1b) is defined by the contributing area to a point located near the city of Xilinhot and comprises about 3600 km². The elevation ranges between 1010 m and 1609 m. The Xilin River is an endorheic river system. It is characterized by a semi-arid continental climate with cold, dry winters and warm, relatively wet summers. The mean annual precipitation is 350 mm but is highly variable in space and time due to prevailing convective weather conditions. Chen (1988) reported annual ranges of 150 and 500 mm, with 60–80% falling between June and August. The mean annual temperature is 2 °C, with a January average of −23 °C and a July average of 18 °C (Chen, 1988). Due to its central location, the study area represents average climate conditions of Inner Mongolia where a west to east trending gradient with increasing precipitation rates and temperature characterizes the climate.

Inner Mongolia belongs to the Eurasian steppe. Its primary natural resources are grasslands which cover more than 70% of the area (870.000 km²) (Kobayashi et al., 1994). The Xilingol grasslands are referred to as the highest quality natural pasture in China (Chen et al., 2008). Livestock farming is the major industry (Chen et al., 2008). The Xilin River Basin represents the characteristic features of the grasslands in Inner Mongolia. Fig. 1b shows that 72.5% of our study area is covered with steppe. The steppe is characterized by grassland species, dominated by Leymus chinensis and Stipa grandis. Cleistogenes squarroso and Artemisia frigida dominate in degraded areas of the steppe regions (Tong et al., 2004). In higher elevated areas in the east and northeast of the basin, the vegetation changes from steppe to mountain meadows. The dominant species of the meadow grassland include Bromus inermis, Agrostis gigantean, Carex pediformis, Stipa baikalenis and Calamagrostis epigeos. Marshland dominates in the vicinity of the river and its tributaries. The dominant species are Phragmites australis, Carex appendiculata, Iris lactea var. chinensis and Hippuris vulgaris. Most of these grassland areas are used for agricultural purposes, mainly grazing of sheep and to a lower proportion of cattle and goats. Even sparsely vegetated areas, e.g. the paleo sand dunes that stretch through the centre of the sub catchment in a northwest–southeast direction and branching out northwards (Fig. 1b), are used for grazing. The sand dunes are populated with dense vegetation, particularly Ulmus pumila and other tree genii, i.e. Betulus spp., Malus spp., Prunus spp. and Populus spp. on north to northwest facing slopes or in depressions, while south to southeast facing slopes have generally sparser shrubs and grassland vegetation. A substantial and considerably increasing part of the area is also used for cultivation of crops such as maize, wheat and rapeseed (Guo et al., 2004) despite unfavourable climate conditions and high risk for wind erosion.

The geological setting of the study area is quite diverse and spans a large range of rocks (Fig. 1c, Table 1). The oldest systems which form the basis of the area originate from the Palaeozoic and comprise a variety of igneous rocks (Granites, Granodiorites and Diorites), sediments and metamorphic rocks (Schists). Parts of them outcrop on the northwestern boundary of the catchment and in the east, as well as in a strip that stretches in a northeast–southwest direction through the centre of the catchment. Some of
these oldest areas are capped by volcanic rocks from the Mesozoic and together they form low mountains in the present-day landscape (Chen, 1988; Tong et al., 2004). They also mark a borderline between an old sedimentation plain consisting of shales in the west and a large area of Quaternary deposits in the east. The Quaternary deposits are distributed along the river and take up most of the area in the east. Landforms such as sandy lands (Chen, 1988; Tong et al., 2004) and hills characterize their present-day landscape. Almost equally spatially dominant are Pleistocene basalts covering the southwest of the study area and forming lava table lands and high plateaus (Chen, 1988; Tong et al., 2004). A more detailed description of rock types is given in Table 1.

2.2. Field sampling

We used a design-based, stratified sampling plan (Brus and deGrujter, 1997) along the lines of McKenzie and Ryan (1999) with land use and topography as stratifying variables. Based on a Landsat TM7 image from August 17th, 2005 seven land use classes were delineated (Fig. 1b). These classes were further stratified using the topographic wetness index (TWI). The TWI is a secondary terrain attribute (Wilson and Gallant, 2000) that is calculated as

\[ \text{TWI} = \log(A / \tan \alpha) \]  

Fig. 1. The maps show (a) the location of the Xilin River Basin (·) in Inner Mongolia, China, (b) hillshade of the Xilin River sub catchment derived from the SRTM digital elevation model. Land use, locations of the sampling sites and the river are superimposed and (c) geology of the study area. Land use was classified based on a Landsat TM7 image from August 17th, 2005. The 1:200,000 geological map of the Inner Mongolian Bureau of geology (1973) was modified and information was lumped into 8 new geological map units based on formation processes and age.
a See map in Fig. 1c.

where A is the specific catchment area and α is the slope. The TWI is a topographic variable that indicates soil moisture conditions. Large values are usually found in lower positions of the watershed and in hollows and indicate an increased likelihood of saturated conditions (Minasny and McBratney, 2007). We classified the TWI into three classes based on an equal area basis. From each of the resulting 18 distinct environments we randomly selected 8 replicates. With sampling at one additional site in a steppe class we resulted in 145 samples. Fig. 1b presents the locations of the result of the lab analyses were used to guide the soil classification. At the remaining 115 sampling sites, a reduced number of soil properties were collected with the Puerckhauer auger down to 1 m depth, i.e. depth of soil horizons, field texture, colour (according to Munsell colour charts), organic matter content, carbonate content, BD, and proportion of rock fragments were determined using a Puerckhauer auger. Soils were classified based on these field measurements.

2.3. Laboratory analyses

Samples from the 30 soil profiles were bagged and shipped to the Soil Lab of the Technical University of Munich in Germany where they were analysed for soil texture [%], BD [g cm$^{-3}$], pH [−] and SOC [mg g$^{-1}$].

In order to determine soil texture, fine earth material (<2 mm) was oxidized with H$_2$O$_2$ to remove organic material. The remaining material was dispersed with Na$_2$P$_2$O$_7$ and shaken for at least 16 h, followed by wet sieving to isolate sand fractions of 2000 to 630 μm, 630 to 200 μm and 200 to 63 μm. To determine silt and clay fractions, approximately 3 g of the <63 μm fraction was suspended in deionized water with Na$_4$P$_2$O$_7$ and an ultrasonication was
conducted for 180 s with 75 J ml⁻¹. Afterwards the distribution of 63 to 20 μm, 20 to 63 μm, 63 to 2 μm and <2 μm fractions were obtained by measuring the X-ray absorption of the soil–water suspension during sedimentation of the soil particles with a Micromeritics Sedigraph 5100 (Micromeritics, Norcross, US).

Bulk density (BD) was quantified with the mass of the oven-dry soil (105 °C) divided by the core volume. Soil pH values were measured in 0.01 M CaCl₂ at a soil/solution ratio of 1-to-2.5 at room temperature.

Concentrations of SOC were determined in duplicate by dry combustion on a Vario Max CNS elemental analyser (Elementar, Hanau, Germany). The measured C concentrations of the samples that were free of carbonate represent the SOC concentration. Samples that contained CaCO₃ were heated to 500 °C to remove organic carbon and the concentration of inorganic C of the residual material was determined by dry combustion. The content of inorganic C was subtracted from the C concentration of the untreated material and represents the SOC content.

All samples were analysed in duplicate and the mean value and coefficient of variation were calculated. Measurements in the texture analyses with a coefficient of variation >5% and all other measurements with a coefficient of variation >1% were repeated.

2.4. Environmental covariates

The 90 m DEM presented in Fig. 1b was derived from the NASA Shuttle Radar Topography Mission (SRTM) 90 m dataset (Tile 60_04, Data Version 4.1) (Jarvis et al., 2008) which was obtained from the CGIAR Consortium for Spatial Information (http://srtm.csi.cgiar.org). The seamless, near-global coverage DEM was available for download in geographic coordinate system, WGS84 datum, with a resolution of 3 arc seconds that had been preprocessed by Reuter et al. (2007). We reprojected the DEM in Universal Transverse Mercator (UTM) coordinates, Zone 50, to obtain a resolution of 90 m. The DEM was then used to calculate a set of 12 primary and secondary terrain attributes as defined by Wilson and Gallant (2000). Table 2 provides a list of these attributes as well as their definition and pedologic significance. The computation of the Contributing Area (CA) as an indicator for soil moisture was based on a Monte–Carlo approach to derive natural flow pattern on noisy DEM (Behrens et al., 2008). Within each step in the Monte–Carlo approach a new random error was added to each cell of the DEM. After removing sinks and pits from each DEM the D8 single-flow algorithm (Jenson and Domingue, 1988) was applied to derive CA. This process was repeated 500 times and the resulting datasets were averaged for each pixel. This results in much more smooth and realistic representations of potential soil wetness compared to single applications of the D8 algorithm as well as multiple-flow algorithms on the original cleaned DEM (Behrens et al., 2008).

A Landsat TM image with a spatial resolution of 30 m (UTM WGS84 projection) from 17th August 2005 was available for land use classification using Idrisi Kilimanjaro (version 14.02, Clark Labs, Worcester MA, USA). In a first step, supervised classification was carried out. Training sites derived from ground truth data and additional expert knowledge were used to classify the catchment into four main ecological units, i.e. (1) sand dunes, (2) mountain meadow, (3) marshland and (4) steppe. In a second step, unsupervised classification was implemented to refine clarity between vegetation types and land management. Therefore, each of the ecological units was further classified by cluster analysis using bands 1 to 5 and 7. The spectral information from these bands was forced into three to five classes, applying expert knowledge about land use management and vegetation present in each area in order to indicate increments of vegetation density in each ecological unit. Cloud cover was masked out of the classification. The resulting classification provided 11 land use types. The validity of these 11 units was tested by visual inspection and refined by further ground surveys. For this study, the resulting classification was lumped into 7 main units that deliver a simplified map of the land use of the region. Fig. 1b presents the seven land use classes: arable land, bare soil, marshland and water, mountain meadow, sand dunes, steppe and water.

Soils are a weathering product of the underlying bedrock. Knowledge of the geology is therefore valuable information when mapping soil types. A geological map of the scale 1:200,000 was provided from the Inner Mongolian Bureau of Geology (Inner Mongolia Bureau of Geology (Team 7), 1973). The map was digitized and georeferenced. The geology of the region is quite

---

Table 2

<table>
<thead>
<tr>
<th>Primary terrain attributes</th>
<th>Abbreviation</th>
<th>Definition</th>
<th>Significance for pedology</th>
<th>Calculation method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elevation</td>
<td>Elev</td>
<td>Height above sea level</td>
<td>Climate, vegetation, potential energy</td>
<td>D8 algorithm (O’Callaghan and Mark, 1984)</td>
</tr>
<tr>
<td>Slope</td>
<td>Slope</td>
<td>Rate of change in elevation direction</td>
<td>Velocity of surface and subsurface flow, geomorphology, soil water content</td>
<td>D8 algorithm</td>
</tr>
<tr>
<td>Aspect</td>
<td>Aspect</td>
<td>Direction of steepest downslope direction</td>
<td>Solar insolation, evapotranspiration, species distribution and abundance</td>
<td>D8 algorithm</td>
</tr>
<tr>
<td>Profile curvature</td>
<td>Profcurv</td>
<td>Rate of aspect change along a flow line</td>
<td>Flow acceleration, erosion, deposition rate, geomorphology</td>
<td>D8 algorithm</td>
</tr>
<tr>
<td>Plan curvature</td>
<td>Plancurv</td>
<td>Rate of aspect change along a contour</td>
<td>Converging/diverging flow, soil water content soil characteristics</td>
<td>D8 algorithm</td>
</tr>
<tr>
<td>Mean curvature</td>
<td>Meancurv</td>
<td>Combination of plan and profile curvature</td>
<td>Both of plan and profile curvature</td>
<td>D8 algorithm</td>
</tr>
<tr>
<td>Secondary terrain attributes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total upslope length</td>
<td>Tlen</td>
<td>Length of entire upslope flowpath to join to a point</td>
<td>Erosion rates, sediment yield, time of concentration</td>
<td>Dinf algorithm (Tarboton, 1997)</td>
</tr>
<tr>
<td>Longest upslope length</td>
<td>Plen</td>
<td>Length of flowpath from the furthest cell that drains to each cell</td>
<td>Flow acceleration, erosion rates</td>
<td>Dinf algorithm</td>
</tr>
<tr>
<td>Contributing area</td>
<td>Ca</td>
<td>Area above a cell that drains into that cell</td>
<td>Runoff volume</td>
<td>Monte Carlo</td>
</tr>
<tr>
<td>Topographic wetness index</td>
<td>Twi</td>
<td>Log (ca/tan slope)/(ca/22.13)ᵃ * (sin slope/0.0896)ᵇ</td>
<td>Soil moisture conditions</td>
<td>Monte Carlo</td>
</tr>
<tr>
<td>Transport capacity index</td>
<td>Tci</td>
<td>ca * tan slope</td>
<td>Net erosion and deposition rates</td>
<td>Monte Carlo</td>
</tr>
<tr>
<td>Stream power index</td>
<td>Spi</td>
<td>Measures erosive power of flowing water</td>
<td>Monte Carlo</td>
<td></td>
</tr>
</tbody>
</table>
complex. The original map exhibits 29 map units from a broad range of geologic periods. Detailed information of the original map units is supplied in Table 1. For use in our digital soil-landscape model, we reduced the number of dimensions of the map based on geological expert knowledge. We lumped information from the map based on formation processes and age and defined nine new geological map units which are presented in Fig. 1c.

Climate influences weathering processes in the soil. Spatial representations of precipitation and temperature were used as climate variables (Fig. 2a and b, respectively). These were available from The Climate Source Inc. (Corvallis, OR, USA) (http://climatesource.com) with a spatial resolution of 1.5 km. Both variables are annual estimates averaged over the period 1961–1990 and computed with the Parameter-elevation Regressions on Independent Slopes Model (PRISM) (Daly et al., 1994).

2.5. Statistical modelling

We chose Random Forests (RF) (Breiman, 2001) as statistical model. RF is an ensemble method that was developed as an extension from CART (Breiman et al., 1984). CART is a non-parametric data mining technique that uses recursive partitioning of the dataset to explore relationships between a response variable and predictor variables, and to predict data. For this purpose, the dataset of the response variable is split in a tree like manner into successively smaller, increasingly homogenous subsets. The splits are based on that value of the predictor variable that decreases node impurity the most. Their most advantageous feature is that they often give a clear picture of the structure of the data, i.e. about the nature of the relationships between response variable and predictor variables (Prasad et al., 2006). However, they are associated with a high sensitivity to the selection of the dataset with respect to the resulting tree structure (Prasad et al., 2006) and with overfitting of the data.

RF was developed to improve the prediction performances of the model. In RF, the model building process is the same as in CART with the difference that many trees are built, i.e. a forest of models. For each tree, only a subset of the predictor variables is used. The number of predictor variables is a user-defined parameter and the permutation of the variables between the trees occurs randomly. This permutation process is implemented in the R "randomForest" routine. The random selection of predictor variables, among which the best split is searched in each tree, increases diversity of the forest and hence, decreases correlation of individual trees. This prevents overfitting (Breiman, 2001).

Also, each tree is built from a bootstrap sample of the input data set which allows for robust error estimation with the remaining test set, the so called Out-Of-Bag (OOB) sample. The OOB sample is used to calculate the OOB error for every tree and then generalize over all classifiers. Each tree is built to maximum size without pruning. The result of RF is one single prediction that is inferred from the ensemble of predictions by aggregation over all trees.

Fig. 2. The maps show a) mean annual precipitation (MAP, mm/yr) and b) mean annual temperature (MAT, °C/yr), period 1961–1990, as obtained from The Climate Source Inc. (Corvallis, OR) in the Xilin river catchment.
achieved with 16 predictors (Fig. 3). The OOB errors varied between dataset and found that the lowest OOB error (39.3%) could be the lowest OOB error. We applied this function to the original data set. The software package offers a tuning function based on majority vote. That means that the most popular class of the forest (\(\text{randomForest}\) package V4.5-30 (Liaw and Wiener, 2002) of \(R\)) (1) mean decrease in accuracy and (2) mean decrease of Gini Index. The first is the importance of a variable with regard to its prediction accuracy where the difference in prediction accuracy is calculated before and after permuting a variable. The second refers to the improvement of the splitting criterion “Gini Index” which measures the reduction in class impurity from partitioning the data set (Myles et al., 2004). We applied RF to get a robust spatial prediction of RSGs in our study area.

We excluded two thirds of the 145 RSG data points as training set in order to have an independent validation data set. The exclusion was based on stratified sampling so that the ratio of 2:1 for each RSG in both, the training and the validation set, was assured. This was done based on the assumption that the RSGs reflect the environmental properties and to ensure that both datasets have the same statistical properties. The RF model was then built with the training set. The RF routine again separated our input training set into a bootstrap sample and an OOB sample, repeating this procedure for each tree. Model accuracy was evaluated with the OOB error estimate which is based on majority vote from all trees. The resulting model was then used to spatially predict RSGs. Since all input variables were available as 90 m resolution rasters, the prediction resulted in a 90 m resolution map with an RSG predicted for each pixel. We then evaluated the accuracy of the map with our previously selected validation set in order to obtain an independent error estimate. The misclassification error rate which is the ratio between misclassified observations and total observations in the validation set was used as error estimate.

There are two user-defined parameters in the “randomForest” package: (1) the number of predictors to be used in each tree-building process \((m_{\text{try}})\) and (2) the number of trees to be built in the forest \((n_{\text{tree}})\). Each of these parameters can be varied by the user to improve the model performance. Before running the model replications we assessed the best \(m_{\text{try}}\) and \(n_{\text{tree}}\) settings for the original data set. The software package offers a tuning function which can be used to find the optimal \(m_{\text{try}}\) settings with respect to the lowest OOB error. We applied this function to the original dataset and found that the lowest OOB error (39.3%) could be achieved with 16 predictors (Fig. 3). The OOB errors varied between 0.392 for \(m_{\text{try}} = 16\) and 0.441 for \(m_{\text{try}} = 2\). Grimm et al. (2008) used a similar approach to tune for best \(m_{\text{try}}\) settings. In their study, the range of obtained OOB errors is in the same magnitude as in our approach, namely, most changes occurring at the second decimal place. In correspondence with findings of Grimm et al. (2008) and other studies (Diaz-Uriarte and de Andres, 2006) that addressed parameter optimization in RF, we conclude that for our dataset the default setting of \(m_{\text{try}}\), which is the square root of the total number of predictors, is a good choice. Parameter optimization for the \(n_{\text{tree}}\) setting resulted in a value of 1000. This is a recommended value to achieve more stable results in estimating variable importance (Diaz-Uriarte and de Andres, 2006).

We repeated the model building procedure 10, 100 and 1000 times to assess the sensitivity of the model and the prediction to the selection of the training set in terms of OOB error and independent error for prediction accuracy. The results of the sensitivity analysis, i.e. the misclassification errors of all the map’s prediction accuracies, are plotted in Fig. 4. The boxplots indicate that with a higher number of model replications a larger range of misclassification errors is obtained. In total, the error rates range between 28.9 and 59.6%. However, the boxplots show that the interquartile ranges of the error distributions decrease with increasing number of model runs. The confidence intervals (CI) that are represented as notches in Fig. 4 also become narrower. The median error for 10 model runs is 45.19 ± 3.40 (95% CI), for 100 model runs it is 30.71 ± 2.00 (95% CI), and for 1000 model runs it is 26.15 ± 1.50 (95% CI).
44.23 ± 1.04 (95% C.I.) and for 1000 model runs it is 44.23 ± 0.30 (95% C.I.). We selected the training set with the lowest prediction error for further analysis.

Confusion matrices summarize the final classifications of the model and of the spatial prediction. Plots of the variable importance measures and spatial visualizations of the RSGs were consulted to help interpreting which predictor variables drive the spatial distributions of the soil types.

3. Results

3.1. Field and lab observations

We identified seven RSGs among the 30 soil profiles with detailed description based on the classification of the WRB (IUSS Working Group WRB, 2007): Arenosol, Calcisol, Cambisol, Chernozem, Cryosol, Gleysol and Phaeozem. Table 3 lists their differences in depth of A horizons, texture, BD, pH and SOC content. Kastanozems and Regosols were not among the 30 soil profiles. They were among the 115 sites that were classified based on field observations. Laboratory results therefore do not exist for these two RSGs (Table 3). Most common are the typical steppe soils Phaeozem (47 observations), Chernozem (29) and Kastanozem (10). They differ in their thicknesses and colour of humic horizons and amount and depth of secondary carbonates. The differences in these properties are induced by different moisture and temperature regimes. Arenosols are also quite frequent in the study area (34 observations) where they were often found in the sand dunes. Gleysols are also a common soil type within the study area (14 observations) and can be found in groundwater dominated areas. Less frequent are Cambisols (4 observations), Calci sols (3), Regosols (3) and Cryosols (1).

3.2. Random Forests

3.2.1. Model performance

We conducted RF analysis with WRB RSG as response variable and 16 environmental covariates, achieving an OOB error rate of 51.61%. The confusion matrix of the OOB predictions in Table 4 shows that the prediction accuracy is high for RSGs that have high frequencies, e.g. Arenosols, Chernozems, Gleysols and Phaeozems, and low for the RSGs that are less frequent, e.g. Calcisols, Cambisols, Kastanozems, Cryosols and Regosols. Most of the incorrectly classified RSGs ended up in classes that are similar in terms of pedogenesis, e.g. Kastanozems are predicted as Chernozems, the Cryosol is predicted as Gleysol, some Phaeozems as Chernozems and some Chernozems as Phaeozems.

3.2.2. Variable importance

There are two measures that quantify the importance of the variables: the mean decrease in accuracy and the mean decrease in the Gini Index (Fig. 5). In both cases, land use is the variable that increases model accuracy and node impurity most efficiently. In terms of increasing model accuracy it is followed by a list of 4 topographic variables, namely CA, elevation, stream power index (SPI) and TWI, and the geology that show values above 1. Mean annual precipitation (MAP) and temperature (MAT) follow in importance forming a group of two parameters with values between 0.5 and 1.

In terms of decreasing node impurity, there is a list of topographic variables that have almost equal importance values. These are elevation, TWI, CA, SPI, plan curvature, slope, transport capacity index (TCI) and aspect. Among them are also the geology, MAP and MAT with values above 3. Since it is not possible to quantify the predictor variables we used spatial visualizations of the predictions.

3.2.3. Model prediction

Fig. 6 presents the predicted RSGs based on the RF model. Table 3 presents the area coverages of the predicted RSGs. Phaeozems are by far the most extensive soils, with 2121.7 km² covering 58.6% of the area. They occur mainly in more elevated areas throughout the whole area. Arenosols and Chernozems are almost equally extensive (Table 3). Arenosols are mainly developed in sand dune areas (Figs. 1b and 6). But they are also predicted in areas classified as bare soil. Chernozems mainly occur in close vicinity with Gleysols and Phaeozems. They appear to fill the spatial gaps between Phaeozems and Gleysols.

Gleysols and Kastanozems cover with 4.9 and 3.3%, respectively, appreciable parts of the area. Gleysols are mostly developed along the river network and in groundwater dominated areas which are indicated by marshland vegetation such as P. australis, C. appendiculata, I. lacteal var. chinensis and H. vulgaris (personal field observations).

Larger areas of Kastanozems are developed in lower elevated areas, dominantly in the west of the catchment. Many very small patches occur on the basaltic plateau in the southwestern part of the catchment, fewer small patches occur north of the river. Often times they abut Chernozems.

Less distributed are Calci sols, Cambisols, Cryosols and Regosols. Very small patches of Calcisols and Regosols occur in areas underlain by sediments in the west of the catchment, northwards and southwards of the river. In these areas they occur in close vicinity to each other. Some Calcisols are also developed in areas underlain by shales in the north of the catchment. The distribution of Cambisols is restricted to areas which are located in the west and only occur on land that is used for agricultural purposes. The few sites that are mapped as Cryosols are linked with wetter regions in the vicinity of the river network in higher elevated areas. They only occur in the wetland area located in the eastern part of the catchment. We performed validation analysis using an independent validation set to estimate the accuracy of the predictions.

Table 3

| RSG          | n  | Ah horizon [cm] | Sand content [%] | Silt content [%] | Clay content [%] | SOCa [g cm⁻³] | BDa [g cm⁻³] | pH⁺ | Area [%]
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Arenosol</td>
<td>9</td>
<td>18 ± 17</td>
<td>86 ± 9</td>
<td>9 ± 6</td>
<td>5 ± 3</td>
<td>13.8 ± 18.5</td>
<td>1.48 ± 0.35</td>
<td>6.4 ± 0.2</td>
<td>543.1</td>
</tr>
<tr>
<td>Calcisol</td>
<td>3</td>
<td>15 ± 18</td>
<td>50 ± 29</td>
<td>31 ± 21</td>
<td>19 ± 9</td>
<td>32.4 ± 23.5</td>
<td>1.14 ± 0.28</td>
<td>7.0 ± 0.2</td>
<td>10.2</td>
</tr>
<tr>
<td>Cambisol</td>
<td>2</td>
<td>23 ± 4</td>
<td>74 ± 9</td>
<td>16 ± 6</td>
<td>9 ± 3</td>
<td>12.8 ± 3.8</td>
<td>1.34 ± 0.09</td>
<td>6.5 ± 1.3</td>
<td>9.5</td>
</tr>
<tr>
<td>Chernozem</td>
<td>7</td>
<td>59 ± 14</td>
<td>60 ± 19</td>
<td>25 ± 13</td>
<td>14 ± 6</td>
<td>15.2 ± 7.0</td>
<td>1.33 ± 0.16</td>
<td>6.6 ± 0.7</td>
<td>616.6</td>
</tr>
<tr>
<td>Cryosol</td>
<td>1</td>
<td>20</td>
<td>87</td>
<td>8</td>
<td>5</td>
<td>33.6</td>
<td>1.06</td>
<td>6.3</td>
<td>2.2</td>
</tr>
<tr>
<td>Gleysol</td>
<td>1</td>
<td>25</td>
<td>92</td>
<td>5</td>
<td>3</td>
<td>44.7</td>
<td>1.09</td>
<td>7.4</td>
<td>179.2</td>
</tr>
<tr>
<td>Kastanozem</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Phaeozem</td>
<td>7</td>
<td>62 ± 35</td>
<td>50 ± 17</td>
<td>32 ± 11</td>
<td>18 ± 6</td>
<td>33.3 ± 15.9</td>
<td>1.15 ± 0.18</td>
<td>5.9 ± 0.5</td>
<td>2121.7</td>
</tr>
<tr>
<td>Regosol</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>19.3</td>
</tr>
</tbody>
</table>

¹ A horizon.
3.2.4. Validation

The application of an independent, previously selected validation set allows us to assess the model’s ability of regionalization. From the 52 validation samples, 37 were classified correctly (71.1%). This results in an overall independent misclassification error rate of 28.9%. Table 5 presents the confusion matrix of the predicted and the observed classes based on the validation set. It shows a similar classification pattern as the model. Spatially extensive RSGs have high prediction accuracy, e.g. Arensoul, Chernozem, Gleysol and Phaeozem. Even the prediction performance of the lesser extensive RSGs is high, e.g. Cambisoul, Kastanozem and Regosol. Only the map accuracy of the Calcisoul, a lesser spatial extensive soil, is low (Table 5). Since there is only one site classified as Cryosol in the original data set, Cryosols were predicted by the model but no conclusion can be drawn on the map’s accuracy in predicting this RSG because it was not present in the validation data set.

4. Discussion

4.1. Random Forests performance

The superior capability of RF is its prediction performance (Prasad et al., 2006). Our results support this assessment with high prediction accuracies in model performance and regionalization performance. The OOB error of our model is 51.61%. This is much lower than in the study of Grimm et al. (2008) where soil organic carbon concentrations and stocks were predicted using RF. They achieved an OOB error rate ranging between 75% and 94% in five soil depths. The RF model performances in other studies, e.g. in ecology, are slightly higher than our results. Peters et al. (2008a) achieved an OOB error of 23.9% for their model of groundwater dependent vegetation patterns. In another study, where they modelled wetland vegetation distribution (Peters et al., 2008b), they achieved an OOB error that ranged between 19% and 45%.

We used an independent validation set to evaluate the model’s regionalization capabilities. The test revealed a low misclassification error of 28.9%. This is in the same error range as obtained in the study of Wiesmeier et al. (2010). There exist no other studies in soil science which used RF as a statistical method and validation to test the accuracy of the predictions. However, other studies in ecology that applied RF for spatial predictions of vegetation patterns (Peters et al., 2008a) achieved less satisfying results with a misclassification error of 80.2%.

Our confidence in the model and the map accuracy is even more enhanced by the summary of the classifications which is presented in the confusion matrices (Tables 4 and 5). In both cases, many of the samples that were voted into the wrong RSG class indicate taxonomic similarity with even that class. However,
misclassification is also partly due to some RSGs being underrepresented in the data set. An increased number of these classes is likely to reduce the misclassification error.

4.2. Soil–environment relationships

The strong RF model performance delivers evidence that the spatial distributions of soil types are controlled by land use, topographic variables such as catchment area (CA), elevation (ELEV), stream power index (SPI) and topographic wetness index (TWI), and geology.

A quantitative assessment of these relationships was possible by benefitting from RFs strong advantages in prediction performance and the RF model was used to spatially predict the RSGs in our study area. The spatial visualizations of the prediction helped to interpret the variable importance measures and assess the soil–environment relationships that drive spatial distribution of soil types. The soil map shows some striking distribution patterns that allow interpretations about the soil–environment relationships. The interpretation is supported by the hillshade which is superimposed on the soil map (Fig. 6) and by comparison with the land use (Fig. 1b), geologic (Fig. 1c), MAP and MAT (Fig. 2a and b) maps.

Phaeozems are the most extensive RSG in the region. They are the typical soils of the long grass steppe which accommodate soils with a thick, humus rich topsoil, high base saturation, which is still less than in Chernozems and Kastanozems, and that may contain secondary carbonates. These features indicate a certain combination of climate conditions and transport processes. The occurrence of the Phaeozems is linked with higher elevations as the hillshade of the DEM indicates (Fig. 6). In the Xilin river catchment, higher elevations are associated with higher precipitation regimes, colder temperatures and an increased gradient in transport energy. This is also supported by the spatial distribution of MAT and MAP in Fig. 2a and b, respectively. These are the preconditions for the natural development of Phaeozems in semi-arid regions. The large spatial extent that the Phaeozems cover in the area indicates that these natural soil forming processes are still dominating soil development.

Calcisols are soils that accumulate substantial amounts of lime, usually originating from calcareous bedrock material. We identified hypocalcic, leptic petric and luvic hypocalcic petric horizons. Regosols are weakly developed soils that do not classify for any of the other reference soil groups. They usually indicate eroded areas in arid and semi-arid areas. One site in the vicinity of the river was classified as Cryosol. Cryosols usually occur in permafrost regions and accommodate a frozen soil layer. At this site, the soil was frozen at about

---

![Fig. 6. Predicted RSGs based on the Random Forest model (OOB error = 51.61%, misclassification error = 28.9%).](image-url)
0.9 m depth. Since the time of sampling was at the beginning of July it is questionable if we would have identified this site as Cryosol also at the end of summer, when thawing of the soil would have proceeded. Less extensive but strikingly distributed are the relatively shallow Arenosols that are characterized by a sandy texture and low SOC concentrations. They mainly occur in the sand dunes (Fig. 1b). In this environment, they indicate initial soil development. But they are also predicted on bare soil. Bare soil is a land use unit that represents areas of intensively grazed steppe. This may imply that intensively grazed areas already account for the degradation of soils. This is in line with the findings of Liu et al. (2008) who found that human induced land use changes lead to soil degradation by decreasing the fine earth fraction.

Gleysols feature a soil layer that develops a characteristic colour pattern of grey soil with orange and red mottles that is evoked through subsequent wetting and drying by a rising and falling groundwater table. The spatial pattern of the Gleysol is similarly remarkable. Their spatial distribution clearly reflects the structures of the river network indicating intact natural soil development in this environment that is mainly driven by topography. A high proportion of the soil profiles were classified as Chernozems that are also dark, humus rich soils. They usually accommodate moderate to high amounts of secondary carbonates below the mollic horizon. The major feature of the spatial distribution of the Chernozems is their location. They seem to fill the spatial gaps between the Phaeozems and the Gleysols. Their spatial distribution seems to be linked with topographic variables that control soil moisture conditions, e.g. TWI. One explanation for their spatial pattern may be that the groundwater contains dissolved carbonates which show capillary ascent. The closer a location is situated to the river the more carbonate ascends in the capillaries. These are the sites that we identify as Chernozems. The concentration of dissolved carbonates in the groundwater may decrease with increasing distance to the river and with them also the amount and depth of secondary carbonates in the soil which we then classify as Phaeozems. In addition, Chernozems are less extensively distributed in higher elevations in the east of the catchment where precipitation is higher. This is in line with their definition in the WRB where more humid climate conditions define the separation towards Phaeozems.

In contrast, Kastanozems have a thinner black, humus-rich horizon but therefore feature high amounts of secondary carbonates compared to Chernozems and Phaeozems. They are the typical soils of the shortgrass steppe. The spatial distribution of the Kastanozems is also clearly linked with the climate. A comparison with the spatial pattern of the MAP and MAT indicates that they dominate in areas that are associated with lower precipitation and higher temperatures in the west of the catchment (Fig. 2a and b). The west of the catchment is also characterized by high TWI values which indicate high soil moisture conditions. The relationship between these specific site and climate conditions with the spatial distribution of the Kastanozems is concordant with their definition in the WRB. Kastanozems are defined to populate areas with drier and warmer climate conditions and wetter site conditions in which a reduced input and an increased mineralization of organic materials limits the amount of soil organic matter in the upper soil, a feature that separates the Kastanozems from the Chernozems by definition. The occurrences of the Cryosols can also be related to topography and climate. They mainly occur in the vicinity of the river, in higher elevated areas that exhibit lower MATs, e.g. near the origin of the river in the eastern part of the area.

Cambisols, Cambisols and Regosols are far less extensive. Still, they show a distinct distribution pattern by mainly occurring in the western part of the area. They seem to be restricted to a certain combination of the climate with higher temperatures and lower precipitation and the underlying geology, volcanic rocks in this case. 

5. Conclusions

The RF approach that we applied to map RSGs allowed deeper insight into the current soil forming processes of a grassland area. We could benefit from the prediction capabilities of RF which served as a sound basis for interpretation of the soil/environment relationships that we were able to establish between soil types and environmental properties. Our results imply that land use and climate change in the area towards intensively grazed areas and warmer and drier climate conditions will have an impact on the spatial distribution of RSGs in the area. A change towards an increasing extent of intensively grazed land will lead to degradation of soils. Soils will e.g. lose their capability of storing carbon and provision of land that can be used for agricultural purposes. But more importantly, a climate change towards warmer and drier climate conditions may lead to a transition to soil types with reduced carbon sequestration due to an unfavourable proportion of addition of organic material and weathering rate. However, a detailed investigation on the relationship of C storage and climate is necessary to confirm these results.

Acknowledgements

We are grateful to the German Science Foundation (DFG) for their generous funding of FG 536. We also thank the Institute of Botany for providing existing soil information of the study area, T. Behrens for helpful suggestions and support during GIS analysis and F. Gerschlauer for substantial support with the preparation of the PRISM data. We particularly appreciate the field efforts provided by T. Lenhauwsky (Ingenieurbüro Feldwisch, Bergisch- Gladbach, Germany). Technical and logistical support from the Inner Mongolian Grassland Ecosystem Research Station and the administrative support from Yuandi Zhu are deeply acknowledged.