Journal of Geochemical Exploration 197 (2019) 37–47

Contents lists available at ScienceDirect

Journal of Geochemical Exploration

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

Spatial distributions and the identification of ore-related anomalies of Cu across the boundary area of China and Mongolia

Mi Tian, Xueqiu Wang*, Lanshi Nie, Hanliang Liu, Wei Wang, Taotao Yan

Key Laboratory of Geochemical Exploration, Institute of Geophysical and Geochemical Exploration (IGGE), CAGS, Langfang 065000, China
UNESCO International Centre on Global-scale Geochemistry (ICGG), Langfang 065000, China

ARTICLE INFO

Keywords:
Stream sediment
Copper
Random forest regression
Boundary areas of China and Mongolia
Ore-related anomalies

ABSTRACT

The 1:1, 000, 000 geochemical mapping across the boundary area of China and Mongolia has acquired high-quality basic geochemical data, yielding more than 9000 stream sediment samples and over 400 rock samples. Each stream sediment and rock sample site was assigned to the regional tectonic unit, geological background and geomorphic landscape maps, respectively, resulting in spatial data for three categorical variables. The contents and spatial distributions of Cu in stream sediment and rock samples across the boundary area of China and Mongolia were studied. The stream sediment geochemical data set was centered logratio transformed (clr) to avoid the closure effect. Random forest regression (RFR) was applied to predict Cu geochemical background for each sample with tectonic unit, geological background, and geomorphic landscape as characteristic variables, then the residuals were used to map geochemical anomalies. The study shows that Cu contents in stream sediments and rocks are high in the west and low in the east of the study area, and the median Cu varies with different tectonic units, geological backgrounds, and geomorphic landscapes. The background values of Cu are strongly spatial heterogeneous. The distributions of predicted Cu background values by random forest regression algorithm are similar to the distribution characteristics of Cu in stream sediments and rocks. The influence of tectonic units on the spatial variations of Cu geochemical background was greatest, followed by geological background, the influence of geomorphic landscape was smallest. The geochemical anomalies drawn by the residuals produced in random forest regression are in good agreement with the known deposits, indicating that the predicted geochemical background of Cu is reasonable and accurate, and has certain theoretical and practical significance.

1. Introduction

The area across the boundary of China and Mongolia has excellent geological conditions for mineralization. There are Oyu Tolgoi ultra-large porphyry copper-gold deposits, Tsagaan Suvarga large porphyry copper-molybdenum deposits, Asgarte large silver deposits, Talimetz large gold deposits, Chabou and bayanur large lead-zinc-silver deposits and other small-medium sized deposits (Kirwin et al., 2005; Wainwright et al., 2017; Hou et al., 2010). Copper is one of the most widely used metals. At present, more than 600 million tons of copper have been proved in the world, and copper is widely used in the fields of electricity, light industry, machinery manufacturing, construction industry and national defense industry. As one of the most important metallogenic belts of porphyry copper in the world, the study area has huge resource potential, making it one of the hot spots in international geoscientific research and exploration (Li et al., 2015; Nie et al., 2004).

The 1:1, 000, 000 geochemical mapping project across the boundary area of China and Mongolia began in 2008, and was implemented by China Geological Survey, Mongolia Ministry of Geology and Mineral Resources and Mongolia Geological Survey Center. By 2017, more than 9000 stream sediment samples and 400 rock samples from about 1.3 million km^2 have been collected, and the high quality geochemical data of 68 elements, organic carbon and total carbon have been obtained (Nie et al., 2013). The geochemical data processing and the identification of ore-related geochemical anomalies of Cu are helpful to delineate the metallogenic target area of Cu in the study area to guide further prospecting work.

The decoupling of anomalies from background is an important step in geochemical data processing. Traditionally, a geochemical anomaly can be defined as a concentration of an element that is greater than a

* Corresponding author at: Key Laboratory of Geochemical Exploration, Institute of Geophysical and Geochemical Exploration (IGGE), CAGS, Langfang 065000, China.
E-mail address: wangxueqiu@igge.cn (X. Wang).

https://doi.org/10.1016/j.gexplo.2018.11.010
Received 22 May 2018; Received in revised form 14 November 2018; Accepted 20 November 2018
Available online 22 November 2018
0375-6742/

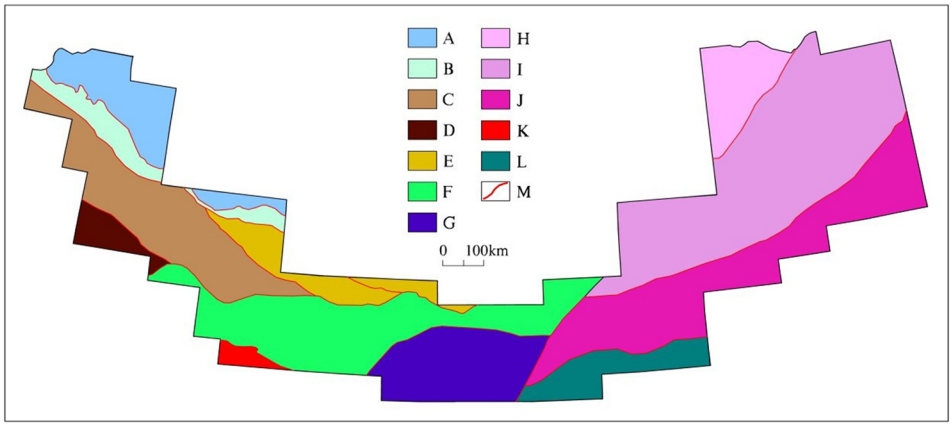

Fig. 1. Geological units across the boundary area of China and Mongolia (Modified according to Li et al., 2015).
A - Altai Tectonic Belt; B - Altai Southern Arc Basin System; C - East and West Junggar Arc Basin System; D - Junggar Block; E - Gobi Altai Arc Basin System; F - Beishan-Gobi Tianshan Arc Basin System; G - Bayinmaodao-Yagan-Baruun Tsohio Tectonic Belt; H - Ereen Davaa - Erguna Microblock; I - Baruun Urt-Hutag Uul-Dongwuqi-Aershan Arc Basin; J - Sulinheer- Mandula-Horingolor Arc Basin System; K - Tarim Block; L - North China Block; M - Fault.

threshold value. The threshold value was usually identified by various statistical methods, such as the probability plots method, 85% cumulative frequency method, median + 2mad method, mean + nSDEV, QQ graph method and fractal method (Hawkes and Webb, 1962; Tukey, 1977; Cheng et al., 1994, Cheng, 2007; Zhang et al., 2008; Zuo et al., 2009; Daya, 2015). However, the applications of these methods were based on the assumption that the geochemical data are normally, log-normally or fractal distributed and moreover the data are homogeneous, obviously these methods were unsuitable for most areas. The lithological units and geographic landscapes of most study area are often complex and thus the geochemical data are obviously heterogeneous. It is demonstrated that backgrounds may change from area to area within a region and between regions (Reimann and Garrett, 2005).

The boundary area of China and Mongolia is such a very typical area, with multiple tectonic units and complex geological backgrounds, moreover, the geomorphic landscapes are diverse, which results in significant elemental natural spatial variations (Tomurtogoo and Badarch, 1998; Tomurtogoo, 2005). Therefore, geochemical data are strongly heterogeneous and most traditional methods are no longer applicable. Researchers proposed that geochemical background values can be fitted through the relationships of metallogenic element and some rock-forming elements. For example, Hao et al. (2014) applied linear regression method to calculate the background values of metallogenic element based on the relationships between metal element concentrations and oxide concentrations (in particular SiO_2). The residual errors, namely the differences between the measured and the predicted values, were used to map the geochemical anomalies. Although this approach takes into account the spatial variations of the geochemical backgrounds, the relationships of elements are generally complex and non-linear in geological events (Lark, 1999).

Random forest (Breiman, 2001) is a highly accurate, adaptable and interpretable machine learning method, which uses an ensemble of decision trees, and is capable of both classification and regression. Based on the non-linear relationships between the dependent variable and the independent variables, it can give the predicted values of the dependent variable and relative importance of each independent variable. It is gaining popularity for use in various fields: for example geological mapping (Cracknell et al., 2014), digital soil mapping (e.g. Henderson et al., 2005; Wiesmeier et al., 2009) and mineral prospectivity mapping (e.g. Carranza and Laborte, 2015; Harris and Grunsky, 2015).

In this paper, the contents and overall distributions of Cu in stream sediment and rock samples across the boundary area of China and Mongolia were studied. The concentrations and spatial distributions of Cu in different geological units, geological backgrounds and geomorphic landscapes were discussed. And then the random forest regression was applied to predict the geochemical background values of Cu for every sample sites, with the abovementioned three factors as characteristic variables, considering about the natural spatial variations of elements. The residual errors were used to identify geochemical anomalies related to mineralization, and then the impacts of these factors on the Cu natural spatial variations were given. The study of the influence factors of elemental natural spatial variations is essential for accurately determining the elemental geochemical backgrounds, then identifying Cu geochemical anomalies related to mineralization across the boundary area of China and Mongolia.

2. The study area

The study area is located at the boundary areas of China and Mongolia, covering the Altai Mountains, south of the Mongolia Plateau and west of the Great Khingan, extending about 50–100 km from each side of the border. The length of borderline between China and Mongolia is up to 4673 km. Geographical coordinates of the study area are between East longitude 86°–120° and North latitude 41°–50°, with a total area of about 1.3 million km^2 (Nie et al., 2013).

The study area is made of two primary tectonic units, the Altai-Xingmeng orogenic system of the central Asian tectonic belt with a large area (areas including A–J, Fig. 1) and Tarim-North China block with a small area (K and L, Fig. 1). The Altai-Xingmeng orogenic system are subdivided into 10 secondary tectonic units, including Altai Tectonic Belt (A), Altai Southern Arc Basin System (B), East and west Junggar Arc Basin System (C), Junggar Block (D), Gobi Altai Arc Basin System (E), Beishan-Gobi Tianshan Arc Basin System (F), Bayinmaodao-Yagan-Baruun Tsohio Tectonic Belt (G), Ereen Davaa- Erguna Microblock (H), Baruun Urt-Hutag Uul-Dongwuqi- Aershan Arc Basin (I) and Sulinheer- Mandula-Horingolor Arc Basin System (J). Tarim-north China block are subdivided into 2 secondary tectonic units, including Tarim Block (K) and North China Block (L) (Liu et al., 2018; Li et al., 2015).

The geological background is complex in the boundary area of China and Mongolia, distributed from Proterozoic to Quaternary (A-N, Fig. 2). Proterozoic, Cambrian, Silurian, Devonian and Andesite (A, B, D, E, L, Fig. 2) are mainly located in the north-western part with a small area. Permian, Jurassic and Tertiary (G, H, J, Fig. 2) are mainly distributed in the eastern part of the area and that is sporadically distributed in the west. Cretaceous (I, Fig. 2) is distributed in the middle-east of the study area with a large area. Carboniferous (F, Fig. 2) is distributed sporadically in the study area with a small area. Ordovician and Granite (C, M, Fig. 2) are distributed evenly in the study area. Quaternary (K, Fig. 2) is distributed over a large range across the area. In addition, basalt (N, Fig. 2) is distributed in the north-eastern part of the study area with a small area (Fu et al., 2016; Tang et al., 2016).

The geomorphic landscapes here are mainly grassland (B, E, F, G, I, K, Fig. 3), forest-grass (C, D, Fig. 3) and semi-desert area (A, H, J, Fig. 3), where the northwest of the study area is dominated by forest grass area, the eastern part is the grassland area, and the middle and southwest regions are occupied by large semi-desert regions.

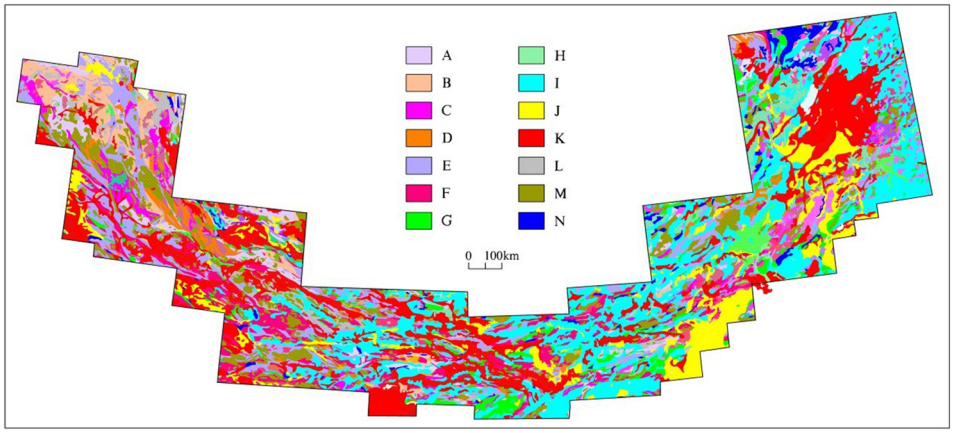

Fig. 2. Geological backgrounds across the boundary area of China and Mongolia
A - Proterozoic; B - Cambrian; C - Ordovician; D - Silurian; E - Devonian; F - Carboniferous; G - Permian; H - Jurassic; I - Cretaceous; J - Tertiary; K - Quaternary; L - Andesite; M - Granite; N - Basalt.

3. Materials and methods

3.1. Sampling and sample preparation

The samples were collected by China and Mongolia cooperation team. The stream sediment sampling density was one sample per 100 km^2. Sampling sites distributed as evenly as possible throughout the survey area and each sample site was at a position, which can control the maximum area of a sampling cell. Active stream sediment samples were collected at the mouths of the streams inside each sampling cell. The sample site was at the lowermost point of the largest drainage catchment in each sampling cell. Composite samples were collected in a range of 50 m (generally 3–4 sites). Rock samples were collected in the area where bedrocks outcrop. GPS units were used to record the coordinates. A total of 9045 stream sediment and 469 rock sampling sites were evenly distributed in the region (Fig. 4). The sampling weight was about 500–1000 g. Duplicate samples were distributed at the same site but different locations (at least 2 m apart), and their quantity was 5% of the total samples (Wang et al., 2007; Tian et al., 2018).

Stream sediment samples were prepared before sending to laboratory for analysis. The procedures were as follows: drying (not directly under the sun)—crushing (to prevent the grains clustering into lumps)—sieving (discarding the portion over or lower the × mesh)—grinding (Grind samples to 200 mesh in agate or pure-aluminium-porcelain mill)—splitting and weighing (depending on the requirement of analysis)—bottling (polypropylene or plastic bottle)—storing (Store the rest of samples in storage room) (Wang et al., 2007, 2011, 2016). The rock samples were coarsely crushed to about 1 cm in the jaw crusher, and then finely grinded to 200 mesh in the agate or pure-aluminium-porcelain mill, and the following steps were the same as stream sediment samples.

3.2. Analysis and quality control

Sample pretreatment method was as follows: 0.25 g sample was dissolved by HF + HNO_3 + $HClO_4$ + aqua regia, then put 25 ml solution in 5% aqua regia, pipette 1 ml clear solution, and then dilute to 10 ml with 2% HNO_3. The samples were analyzed in the laboratory of Institute of Geophysical and Geochemical Exploration (IGGE), Chinese Academy of Geological Sciences. 68 elements (Ag, As, Au, B, Ba, Be, Bi, Br, Cd, Cl, Co, Cr, Cs, Cu, F, Ga, Ge, Hf, Hg, I, In, Li, Mn, Mo, N, Nb, Ni, P, Pb, Rb, S, Sb, Sc, Se, Sn, Sr, Ta, Te, Th, Ti, Tl, U, V, W, Zn, Zr, Y, La, Ce, Pr, Nd, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu, SiO_2, Al_2O_3, Fe_2O_3, MgO, CaO, Na_2O, K_2O), OrgC (organic carbon) and TC (total carbon) were analyzed. Concentrations of Cu were analyzed by ICP-MS, the detection limit was 1 mg/kg. Certificate samples were used, and the accuracy of analyses was better than 5% for major elements and better than 10% for trace elements. Duplicate samples with random sampling 5% of total number of samples in secret code were analyzed, and the relative errors of repeated sample analyses were less than 10%.

3.3. Centered logratio transformation

Since a stream sediment geochemical data set is a closed number system, with regard to the closure effect, centered logratio

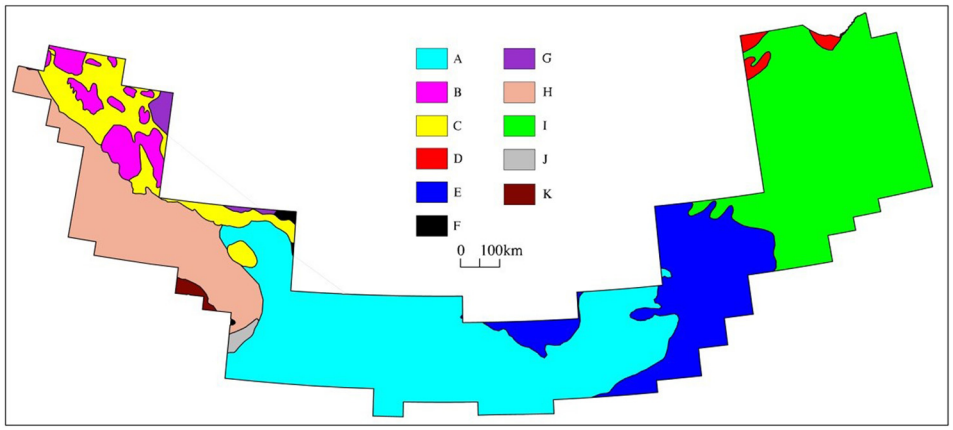

Fig. 3. Geomorphic landscapes across the boundary area of China and Mongolia.
A - Altai alpine meadow and tundra; B - Altai montane forest and forest steppe; C - Altai steppe and semi-desert; D - Daurian forest steppe; E - Eastern Gobi desert steppe; F - Gobi lakes valley desert steppe; G - Great lakes basin desert steppe; H - Junggar basin semi-desert; I - Mongolian-Manchurian grassland; J - Taklimakan desert; K - Tian Shan montane steppe and meadows.

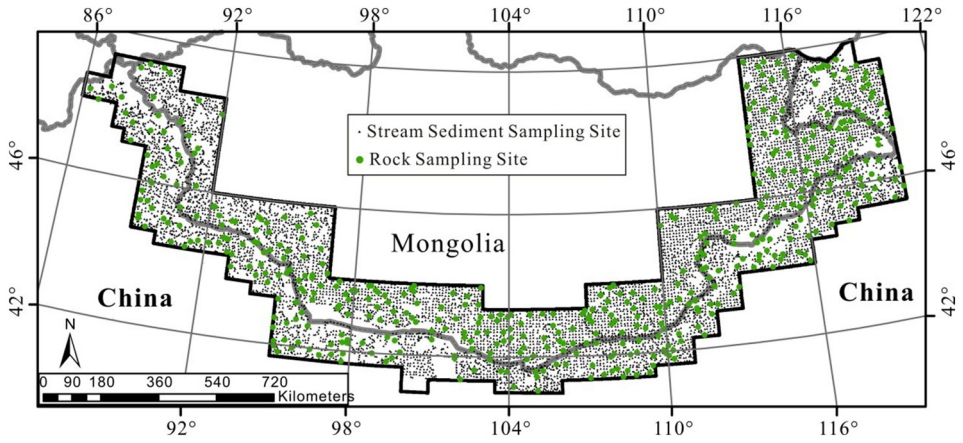

Fig. 4. Locations of samples.

Table 1
Statistical parameters of Cu analytical results (mg/kg) in stream sediment and rock samples.*

Sample type	N	DL	Min	25%	75%	85%	Max	GM	GSD	MAD
Stream sediment	9045	1	1.00	15.29	26.10	29.90	675.75	20.20	13.29	5.30
Rock	469	1	3.88	15.51	25.90	29.51	108.00	20.50	9.98	5.20

* N = number of samples; DL = detection limit; GM = geometric mean; GSD = geometric standard deviation; MAD = median absolute deviaiton.

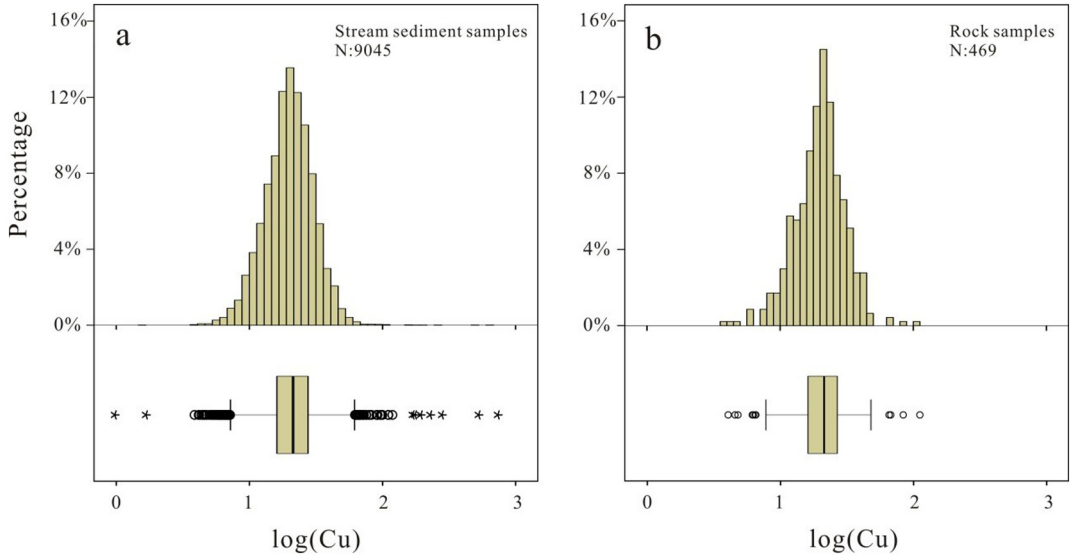

Fig. 5. Histograms and box plots of Cu in stream sediments and rocks.

transformation (clr) was proposed to open compositional data and transform the data from simplex space to Euclidean space (Aitchison, 1986; Egozcue et al., 2003; Tolosana-Delgado, 2008).

Suppose there are p dimension random vectors of compositional data $x = (x_1, x_2, \cdots x_p)$. The steps of centered logratio transformations are as follows: calculate the geometric means of all variables, and then the variable is divided by geometric mean and log-transformed, as the following equation shows:

$$y_i = \ln \frac{x_1}{\sqrt[p]{\prod_{i=1}^{p} x_i}}, \cdots \ln \frac{x_p}{\sqrt[p]{\prod_{i=1}^{p} x_i}} \quad (1)$$

where x is the compositional variable, p is the number of compositional variables, and the denominator is the geometric mean. Correspondingly the random vectors x are transformed into $Y = (y^1, y^2, \cdots y^p)$, which can be assigned in p dimension real space R^p. Hence clr is one-to-one correspondence of X_p and R^p. In consequence, data treated with clr got rid of the constraint of compositional data and came into the degree of freedom of R^p, in which the negative skewness of covariance matrix disappeared.

3.4. Random forest regression

Random forest regression (RFR) was proposed by Leo Breiman and Adele Cutler in 2001, which was an integrated learning algorithm based on decision tree. This algorithm is capable of dealing with continuous and discrete attributes simultaneously. RFR has many advantages, such as high efficiency and anti-noise (Meng et al., 2016; Liu et al., 2015). This method can prevent overfitting without checking the interaction and nonlinearity of variables. The random forest algorithm is insensitive to the outliers, and thus is stable in the case of much random disturbance (Zhang et al., 2017; Sreenivas et al., 2014; Hao et al.,

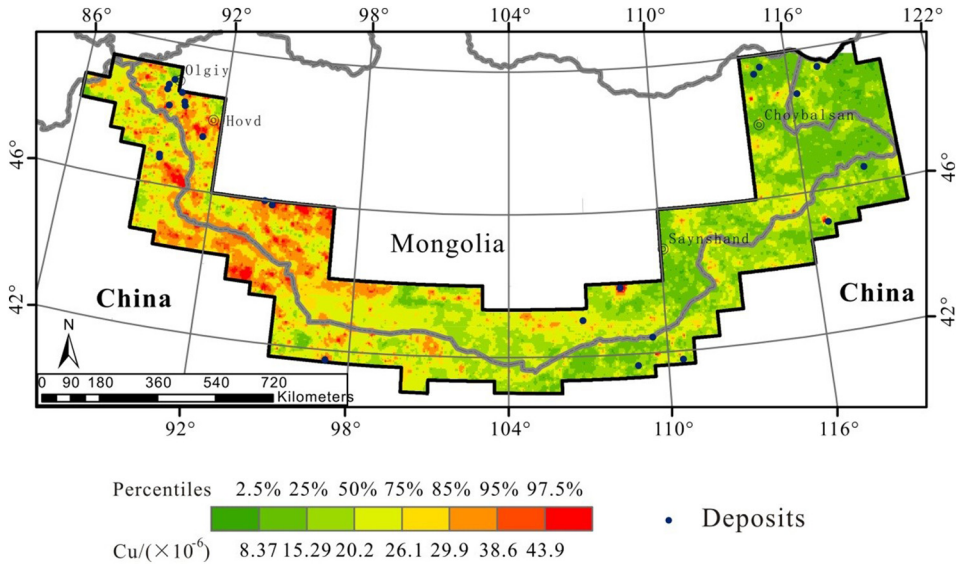

Fig. 6. Geochemical map of Cu in stream sediments.

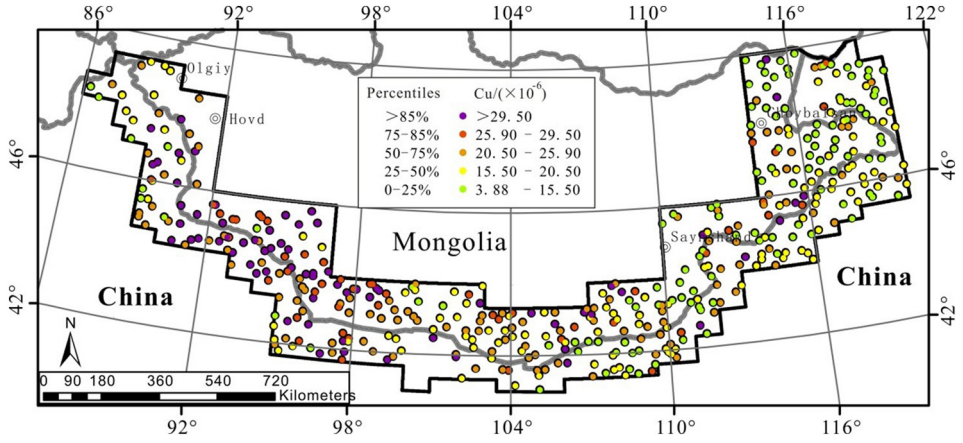

Fig. 7. Geochemical map of Cu in rocks.

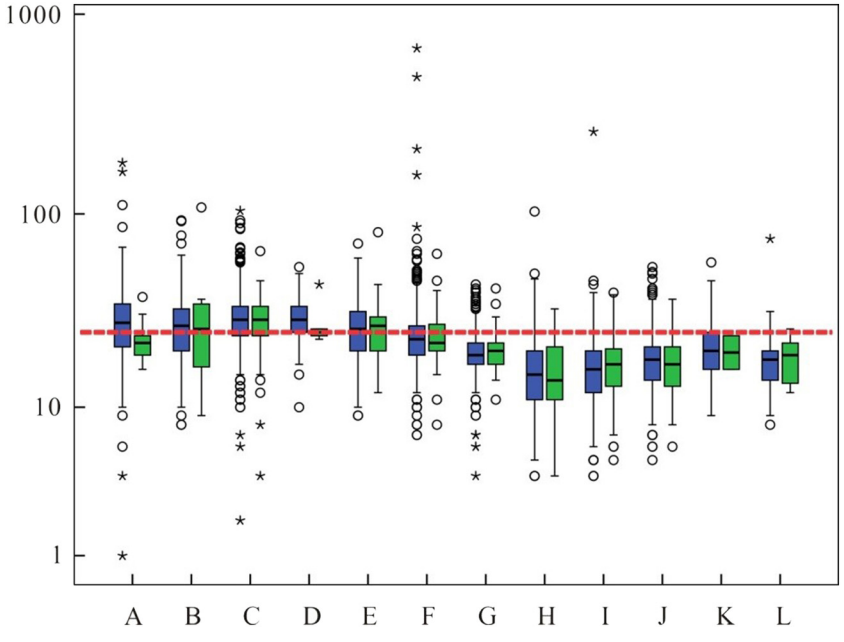

Fig. 8. Boxplots for Cu concentrations (mg/kg) in stream sediment (blue) and rock (green) samples per tectonic unit (Fig. 1). The crustal abundance in China (26 mg/kg, Chi, 2007) is denoted by the red dashed line. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Statistical parameters for Cu concentrations (mg/kg) in stream sediment and rock samples per tectonic unit (Fig. 1).

Tectonic unit	Sample type	N	Min	25%	75%	85%	Max	GM	GSD	MAD
A	Stream sediment	617	0.95	21.43	35.43	39.91	180.52	27.64	14.15	6.61
	Rock	13	16.02	19.08	24.56	24.66	37.74	22.45	5.92	2.35
B	Stream sediment	408	7.56	20.50	33.01	36.47	93.40	26.87	10.90	6.17
	Rock	15	8.72	16.45	35.70	35.70	108.00	25.64	24.18	9.19
C	Stream sediment	1074	1.62	23.58	33.93	37.94	103.80	28.60	9.76	5.24
	Rock	63	3.88	24.30	33.97	37.30	65.14	29.39	9.68	4.84
D	Stream sediment	87	10.29	24.62	34.24	39.66	54.07	28.55	8.24	4.16
	Rock	5	23.35	23.72	34.72	25.70	43.73	24.53	9.64	1.17
E	Stream sediment	472	8.90	20.10	31.68	34.70	71.00	25.50	9.25	5.75
	Rock	29	11.80	19.75	30.85	32.30	80.90	27.40	12.48	6.10
F	Stream sediment	1268	6.57	19.40	26.69	29.73	675.75	22.60	24.58	3.60
	Rock	75	8.24	20.30	27.97	30.20	63.26	22.50	8.32	3.07
G	Stream sediment	757	4.05	16.57	22.30	24.80	43.78	18.90	5.73	2.90
	Rock	37	10.90	16.85	22.26	23.19	41.60	20.00	5.98	2.50
H	Stream sediment	612	3.73	10.91	20.05	22.90	103.08	14.89	7.56	4.32
	Rock	30	4.37	10.42	21.41	23.84	33.46	14.06	8.07	4.97
I	Stream sediment	2551	4.00	12.30	20.00	22.40	259.30	15.80	7.51	3.80
	Rock	135	4.60	12.90	20.54	22.40	40.40	16.58	6.43	3.92
J	Stream sediment	1019	5.20	14.00	21.40	23.90	54.40	17.80	6.08	3.70
	Rock	54	5.90	12.93	21.15	23.80	36.70	16.95	5.78	4.00
K	Stream sediment	31	9.08	15.60	24.65	28.44	56.55	20.33	9.62	4.62
	Rock	2	15.81	**	**	**	23.66	19.74	**	**
L	Stream sediment	149	8.50	14.45	20.15	22.10	74.60	17.60	6.53	2.90
	Rock	11	11.50	13.10	22.90	22.90	26.10	18.90	5.05	4.40

Asterisks represent there are no parameters.

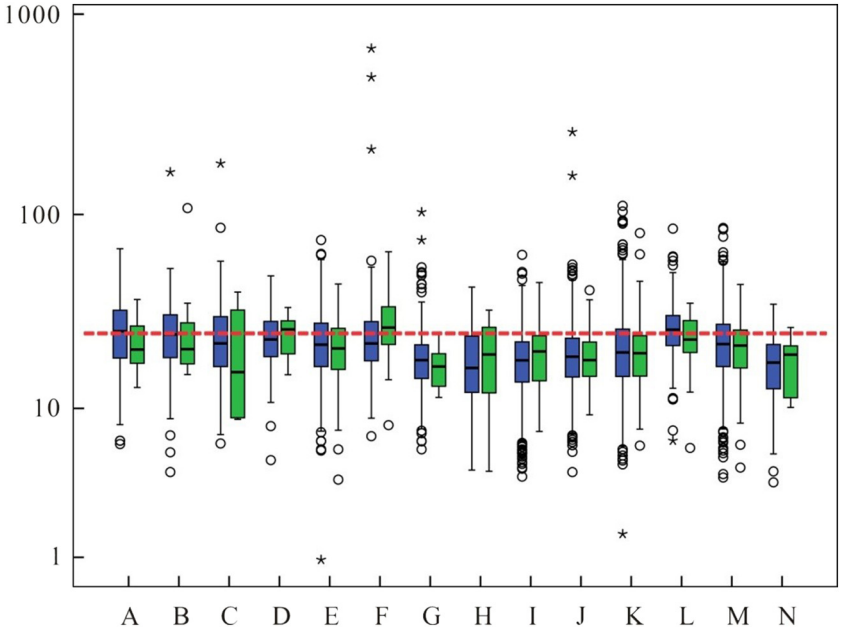

Fig. 9. Boxplots for Cu concentrations (mg/kg) in stream sediment (blue) and rock (green) samples per geological background (Fig. 2). The crustal abundance in China (26 mg/kg, Chi and Yan, 2007) is denoted by the red dashed line. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

2016).

3.4.1. Algorithm principle

The random forest algorithm is composed of a set of regression decision trees$\{h(x,\theta_t), t = 1,2,\cdots T\}$. Where, θ_t is a random variable that obeys independent distribution, x is independent variable, T represents the number of decision trees. The mean of each decision tree $\{h(x,\theta_t)\}$ is taken as the result of regression prediction according to the idea of integrated learning:

$$\overline{h}(x) = \frac{1}{T}\sum_{t=1}^{T}\{h(x,\theta_t)\} \tag{2}$$

where $h(x,\theta_t)$ are the outputs based on x and θ.

RFR introduces the idea of bagging (Breiman, 1996) and random subspace (Ho, 1998) in order to overcome the problem of low accuracy and overfitting of the decision tree model.

3.4.1.1. Bagging. Take multiple training samples from the original samples with replacement randomly, and the size of a set of training samples is equal to the original sample size. Then construct regression decision subtree T for each set of training samples. Finally, the average value of each tree is taken as the final prediction result.

Suppose S is the original sample, N is the sample size of S, the probability of each sample in S being not extracted is $\left(1-\frac{1}{N}\right)^N$.

When $N \to \infty$:

$$\left(1-\frac{1}{N}\right)^N \approx \frac{1}{e} \approx 0.368 \tag{3}$$

Eq. (3) indicates that about 36.8% samples were not extracted at

Table 3
Statistical parameters for Cu concentrations (mg/kg) in stream sediment and rock samples per geological background (Fig. 2).

Geological background	Sample type	N	Min	25%	75%	85%	Max	GM	GSD	MAD
A	Stream sediment	320	6.33	18.60	32.89	33.96	67.50	25.65	15.81	7.09
	Rock	11	13.00	17.40	27.90	29.78	37.30	20.60	7.28	4.15
B	Stream sediment	254	4.32	18.66	31.19	25.35	163.47	24.67	9.03	7.21
	Rock	8	15.23	16.72	32.40	82.70	108.00	20.74	33.60	3.30
C	Stream sediment	198	6.38	16.80	30.60	26.25	180.52	22.19	12.53	5.19
	Rock	4	8.72	8.81	37.17	**	40.69	17.85	15.90	8.95
D	Stream sediment	201	5.10	18.91	28.99	30.42	49.20	23.26	10.24	6.46
	Rock	8	15.20	19.00	29.21	32.37	33.97	26.23	6.42	4.29
E	Stream sediment	861	0.95	16.78	28.24	24.83	74.63	21.85	9.60	4.96
	Rock	47	3.88	16.02	26.68	32.68	44.74	20.90	9.59	5.48
F	Stream sediment	602	7.02	18.00	28.76	35.52	675.75	22.16	11.50	5.94
	Rock	30	8.10	21.50	34.25	37.05	65.14	26.80	10.64	5.70
G	Stream sediment	304	5.90	14.52	21.82	31.98	103.08	18.14	11.30	5.90
	Rock	13	11.50	12.75	19.60	20.50	24.90	16.80	4.12	3.20
H	Stream sediment	180	4.44	12.20	24.20	37.19	43.13	16.50	16.41	8.41
	Rock	14	4.37	11.51	27.38	29.87	32.90	19.46	8.99	7.36
I	Stream sediment	2079	4.05	13.90	22.60	32.62	62.85	18.10	10.63	5.91
	Rock	106	7.47	14.06	24.51	29.18	45.51	20.15	7.52	5.64
J	Stream sediment	785	4.32	14.77	23.60	32.50	259.30	18.90	35.10	5.13
	Rock	40	9.24	14.80	22.73	27.43	41.60	18.15	7.59	3.48
K	Stream sediment	1918	1.62	14.87	26.30	38.38	110.90	19.90	13.51	7.45
	Rock	95	6.20	14.70	24.50	28.72	80.90	19.73	11.25	4.77
L	Stream sediment	203	6.62	21.60	30.90	33.57	85.20	26.10	8.53	6.34
	Rock	18	6.01	19.68	29.75	32.85	35.70	23.22	7.74	4.79
M	Stream sediment	863	4.00	16.80	27.88	24.80	85.88	22.00	7.58	5.83
	Rock	60	4.60	16.50	26.48	30.60	44.50	21.63	8.42	5.13
N	Stream sediment	277	3.73	12.77	21.95	28.50	35.30	17.62	8.35	5.96
	Rock	15	10.14	10.94	22.00	23.20	26.82	19.40	5.85	3.62

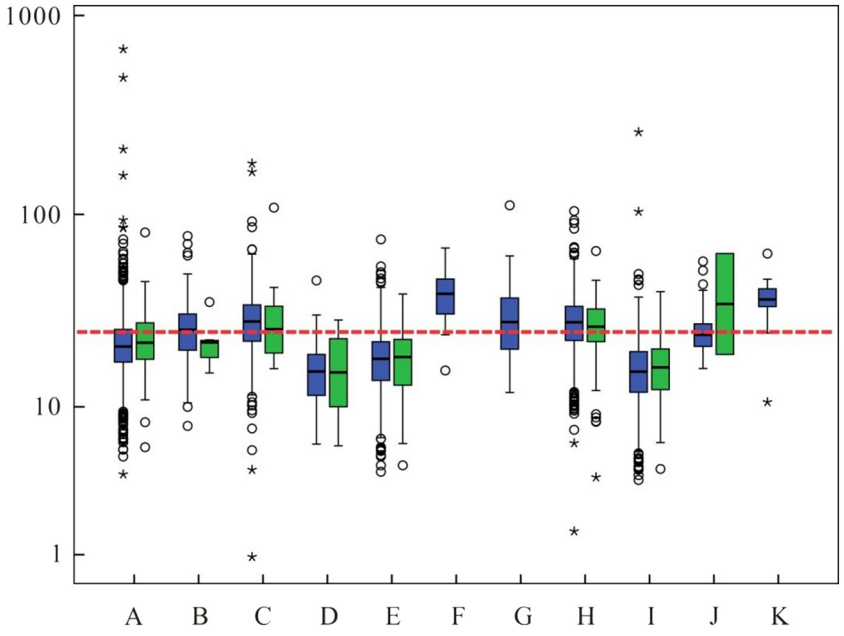

Fig. 10. Boxplots for Cu concentrations (mg/kg) in stream sediment (blue) and rock (green) samples per geomorphic landscape (Fig. 3). The crustal abundance in China (26 mg/kg, Chi, 2007) is denoted by the red dashed line. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

each time, called OOB (out of bag). The bagging idea can not only use randomization to build more regression decision subtrees, but also guarantee the independence between subtrees.

3.4.1.2. Random subspace. In the process of constructing the regression decision subtrees, each split node randomly selects the feature subspace from the total feature space. The subspace is taken as the candidate feature set of nodes, and the optimal feature is selected to be divided. This method ensures that the nodes between trees or in the same tree are different from each other, and it also guarantees the independence and diversity of the tree, and thus enhancing the randomness of RFR node splitting.

3.4.2. Algorithm steps

(1) Generate a subset of samples randomly using bagging idea.
(2) Select f features to carry out node split using the idea of random subspace, and construct a single regression decision subtree.
(3) Repeat step 1 and 2, build T regression decision trees. Each tree grows freely without pruning, then form a forest.
(4) The average of predicted values of T decision subtrees is taken as the final result.

Random forest analysis was implemented in R3.4.2 (R Development Core Team, 2007).

Table 4
Statistical parameters for Cu concentrations (mg/kg) in stream sediment and rock samples per geomorphic landscape (Fig. 3).

Geomorphic landscape	Sample type	N	Min	25%	75%	85%	Max	GM	GSD	MAD
A	Stream sediment	2945	4.05	17.40	25.81	29.10	675.75	21.00	17.28	4.04
	Rock	165	5.90	18.00	27.94	30.16	80.90	22.00	8.65	4.50
B	Stream sediment	230	7.82	20.08	31.12	36.94	77.59	25.70	10.06	5.60
	Rock	9	15.23	17.21	22.67	29.23	35.72	22.06	6.07	3.32
C	Stream sediment	628	0.95	22.35	34.57	39.36	180.52	28.37	13.30	6.14
	Rock	18	16.05	18.95	34.83	38.45	108.00	25.88	21.48	7.77
D	Stream sediment	80	6.15	11.44	19.10	23.19	46.11	15.51	6.71	3.68
	Rock	8	6.01	9.15	23.47	27.17	28.88	15.68	8.20	7.13
E	Stream sediment	1292	4.20	13.90	22.28	25.00	74.60	18.10	6.87	4.20
	Rock	69	4.60	13.00	23.25	26.70	39.30	18.50	7.87	5.10
F	Stream sediment	21	15.70	28.35	46.95	52.12	67.50	39.40	12.68	7.80
	Rock	**	**	**	**	**	**	**	**	**
G	Stream sediment	78	11.95	20.16	37.67	42.26	110.90	28.16	15.21	8.52
	Rock	**	**	**	**	**	**	**	**	**
H	Stream sediment	1112	1.62	22.59	33.96	38.45	103.80	28.10	10.05	5.70
	Rock	60	3.88	22.22	33.35	35.87	65.14	26.65	10.66	5.15
I	Stream sediment	2609	3.73	12.00	19.75	22.30	259.30	15.47	7.69	3.83
	Rock	137	4.37	12.34	20.45	22.00	40.40	16.30	6.07	3.40
J	Stream sediment	32	16.09	20.98	27.98	33.80	57.70	24.18	9.97	3.14
	Rock	2	19.09	**	**	**	63.26	41.18	**	**
K	Stream sediment	18	10.60	32.81	41.82	44.66	63.19	36.85	10.66	4.70
	Rock	**	**	**	**	**	**	**	**	**

Asterisks represent there are no parameters.

Table 5
Statistical parameters for clrCu in stream sediments.

	Min	25%	75%	85%	Max	GM	GSD	MAD
ClrCu	−2.94	−0.65	−0.23	−0.12	3.82	1.31	0.18	0.12

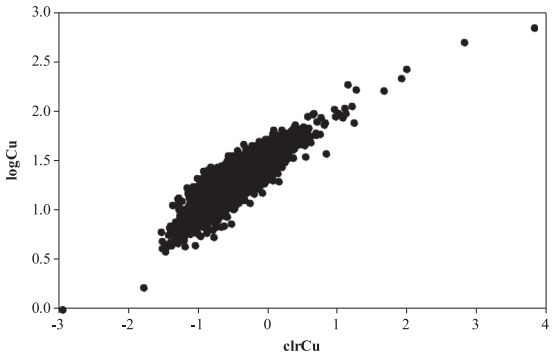

Fig. 11. Scatterplot of clrCu and logCu.

4. Results

The concentration of Cu in rock samples varies from 3.88 mg/kg to 108.00 mg/kg, with a median value of 20.50 mg/kg (Table 1), and with its log-transformed values showing a fairly symmetrical bell-shaped distribution (Fig. 5b). The concentration of Cu in stream sediment samples varies from 1.00 mg/kg to 675.75 mg/kg, with a median value of 20.20 mg/kg, thus a very similar value to that observed for the rock samples (Table 1), and with its log-transformed values also showing a fairly symmetrical bell-shaped distribution (Fig. 5a). The geometric standard deviation of Cu in stream sediment and rock samples are 13.29 mg/kg and 9.98 mg/kg, respectively, and median absolute deviation of Cu is 5.30 mg/kg and 5.20 mg/kg, respectively, which indicates that the spatial variations of Cu in stream sediment are more significant than that in rocks.

The spatial distributions of Cu across the boundary area of China and Mongolia are shown on the geochemical maps (Figs. 6 and 7). The maps were produced from gridded data with the interpolation method of inverse distance weighting (IDW), which were processed by Arcgis 10.2. The geochemical patterns of Cu concentrations in the stream sediment samples and in rock samples are fairly similar (Figs. 6 and 7), both showing a general trend of increasing concentrations westwards.

4.1. Cu concentrations per tectonic unit

The median value for stream sediment samples of the western part of the study area are above the Earth crustal abundance in China (26 mg/kg, Chi, 2007, red line in Fig. 8), namely for the tectonic units (Fig. 1) A, B, C, D (Table 2). The highest median concentrations are observed for the East and West Junggar Arc Basin System (C, 28.60 mg/kg, Table 2) and for the Junggar Block (D, 28.55 mg/kg, Table 2). The lowest median concentrations are observed for the Ereen Davaa- Erguna Microblock (H, 14.89 mg/kg, Table 2). In each tectonic unit, the concentrations of Cu observed for stream sediment samples closely follow the concentrations for rock samples (Fig. 8).

4.2. Cu distributions per geological background

The median value for stream sediment samples in most stratums of the study area are lower than the Earth crustal abundance in China (26 mg/kg, Chi, 2007, red line in Fig. 9), namely for the geological backgrounds (Fig. 2) C, D, E, F, G, H, I, J, K, M, N (Table 3). The highest median concentrations are observed for the Andesite (L, 26.10 mg/kg, Table 3). The lowest median concentrations are observed for the Jurassic (H, 16.50 mg/kg, Table 3). In each geological background, the concentrations of Cu observed for stream sediment samples closely follow the concentrations for rock samples (Fig. 9).

4.3. Cu concentrations per geomorphic landscape

Cu contents of stream sediment samples per geomorphic landscape are significantly different. The median value for stream sediment samples of the western part of the study area are above the Earth crustal abundance in China (26 mg/kg, Chi, 2007, red line in Fig. 10), namely for the geomorphic landscapes (Fig. 3) B, C, F, G, H, K (Table 4). The highest median concentrations are observed for the Gobi lakes valley desert steppe (F, 39.40 mg/kg, Table 4). The lowest median concentrations are observed for the Daurian forest steppe (D, 15.51 mg/kg, Table 4) and Mongolian-Manchurian grassland (I, 15.47 mg/kg, Table 4). In each geomorphic landscape, the concentrations of Cu

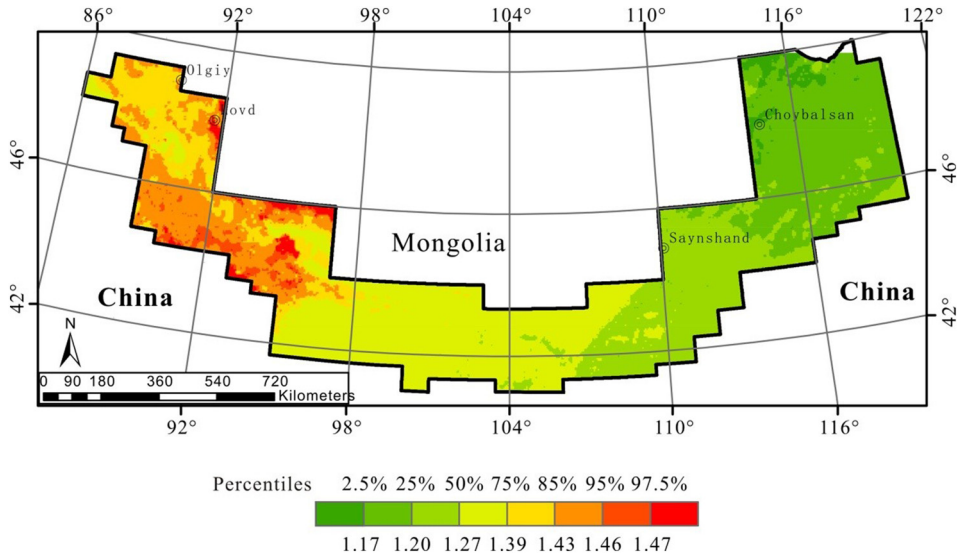

Fig. 12. Cu predicted background values by RFR.

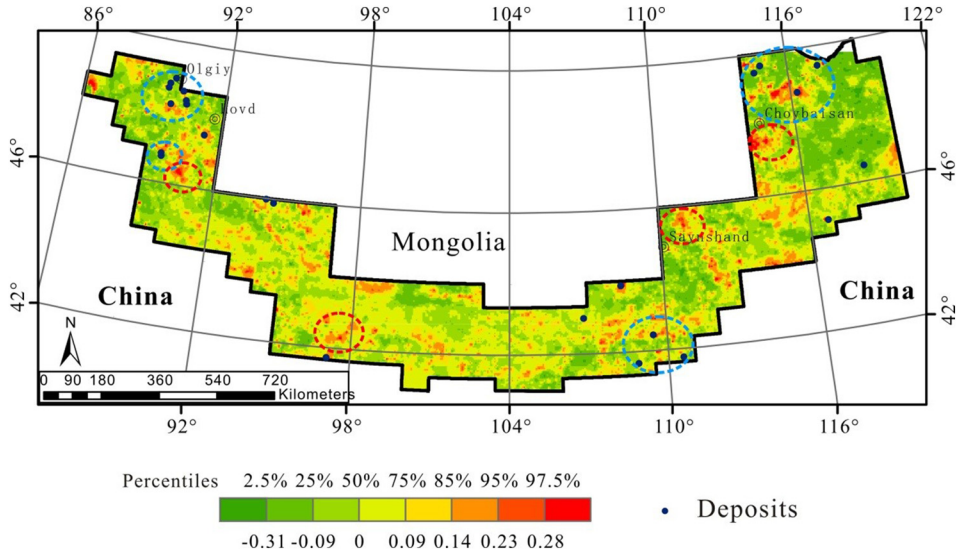

Fig. 13. Anomalies identified by RFR (blue dots represent the known deposits, blue dashed lines show the obvious discrepancies of the two methods, and red dashed lines show the strong anomalies which may be potential prospecting targets). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

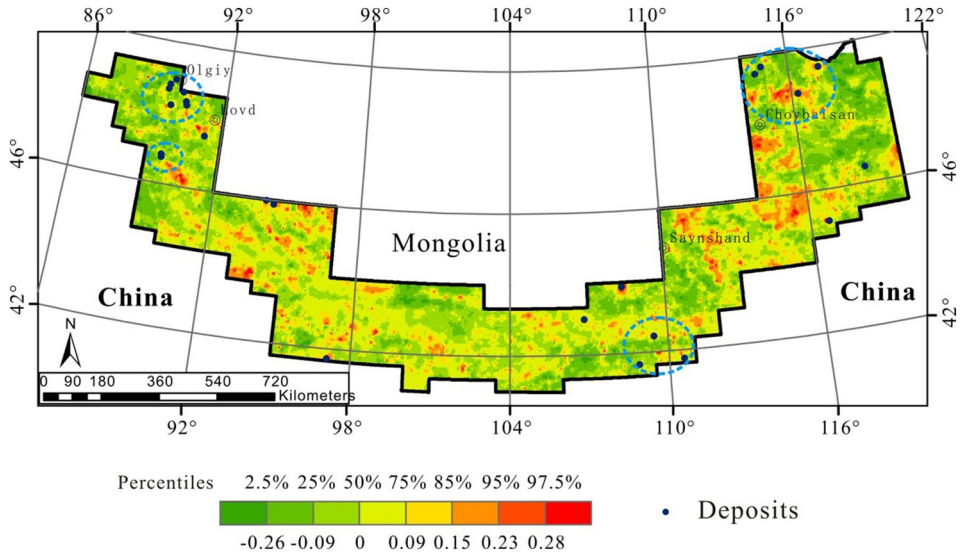

Fig. 14. Anomalies identified by linear regression (blue dots represent the known deposits, and blue dashed lines show the obvious discrepancies of the two methods). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

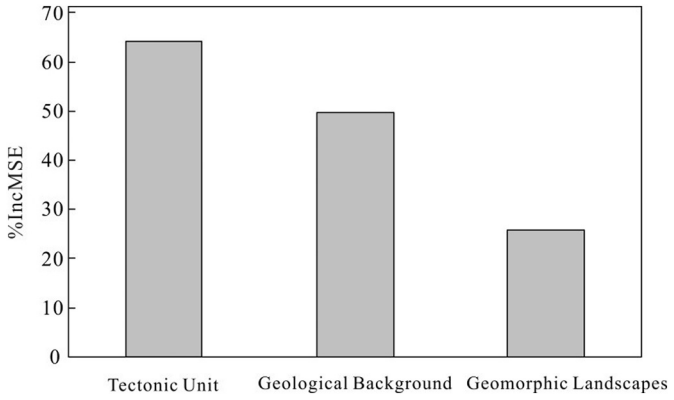

Fig. 15. Importance of variables.

observed for stream sediment samples closely follow the concentrations for rock samples (Fig. 10).

4.4. The result of RFR

The tectonic unit, geological background and geomorphic landscape are set as three characteristic variables, and the contents of Cu in stream sediments are taken as dependent variable to conduct random forest regression analysis. The non-numerical characteristic variables are converted into discrete numerical features. These are attributed such that each subarea is represented by the Cu median concentration observed for the respective stream sediment samples.

All the analyzed element data were centered logratio transformed to open the data and avoid the influence of closure effect. And in order to be more intuitive, the data were converted back to the original unit in geochemical maps.

The statistics of the clrCu are shown in Table 5 and the XY scatterplot of clrCu and logCu is shown in Fig. 11. The background values predicted by RFR (Fig. 12) are similar to that of stream sediments and rocks, with high values in the west and relatively low in the east.

The residuals of RFR and linear regression were used to represent the geochemical anomalies. The geochemical anomalies identified by RFR (Fig. 13) are more evenly distributed in the study area and coincide better with known deposits (blue circled areas in Fig. 13) than the anomalies estimated by linear regression (blue circled areas in Fig. 14). There were no corresponding anomalies in many known mineral deposits by traditional linear regression method.

5. Discussion

The efficiency of geochemical survey can be expressed as EI = TP/(TP + FN) according to Yilmaz et al. (2017), where TP represents the samples containing known mineralization are identified as anomalies, and FN represents that there is no anomalies identified despite the presence of known mineralization. The EI of linear regression method was 50%, and EI of random forest regression was about 66.7%. The efficiency of geochemical survey had been greatly improved by using RFR method, which indicates that the RFR method could identify the anomalies related to mineralization more accurately.

The random forest algorithm can provide the degree of importance of the characteristic variables. As shown in the Fig. 15, the tectonic units influence greatly on the natural spatial variations of the Cu geochemical backgrounds, followed by the geological backgrounds, and the influence of the geomorphic landscapes is the lowest.

Traditionally, raw or logratio geochemical data were used to identify geochemical anomalies, just as Fig. 6 shows, large-area anomalies (orange and red areas) are identified in the western part but nearly no obvious anomalies in the east. Actually, the result of RFR indicates that Cu background values show a general trend of increasing concentrations westward with strong spatial variations. Therefore, most of the anomalies in the western part are false anomalies which resulted from the high background values of this area, and some true anomalies in low background value area are ignored mistakenly. Thus the natural background discrepancies should be eliminated in order to identify ore-related anomalies.

Although linear regression method can eliminate the influence of elemental natural spatial variation by comparison (Fig. 14), this method considers the linear relations between elements and is highly susceptible to outliers that are unavoidable in geochemical data set, and thus the result is unsatisfactory. The random forest regression method takes into account the natural spatial variations of elements and the non-linear relationships between variables, gives the predicted values of element contents by random sampling, and eliminates the influence of natural spatial variations of elements to identify ore-related anomalies (Fig. 13), which improves the efficiency of geochemical survey greatly. The RFR method also delineates some strong continuous anomalies (red circled areas in Fig. 13), which may be potential prospecting targets and worthy of further study.

6. Conclusions

(1) The spatial distributions of Cu in the stream sediment and rock samples across the boundary area of China and Mongolia are fairly similar, both showing a general trend of increasing concentrations westwards. The median Cu varies with different tectonic units, geological backgrounds and geomorphic landscapes.
(2) Geochemical data set is a closed number system, which should be transformed to open the data and avoid the influence of closure effect.
(3) The geochemical data consist of outliers, the relationships between geochemical variables are non-linear, moreover, the geochemical backgrounds are natural spatial variation, and thus traditional geochemical data processing methods are inapplicable.
(4) The geochemical anomalies identified by random forest regression (RFR) coincide better with the known deposits, the RFR is a suitable method to process geochemical data to produce the elemental geochemical backgrounds and identify ore-related anomalies.
(5) Tectonic unit influences the natural spatial variations of stream sediment geochemical backgrounds of Cu most, followed by the geological background and then the geomorphic landscape.

Acknowledgements

This study was funded by National Key R&D Program of Deep-penetrating Geochemistry (2016YFC0600600) and International Science Cooperation Program of *Mapping Chemical Earth*: The Belt and Road Geochemical Mapping Project (121201108000150005). We thank the two anonymous reviewers for their professional comments which greatly improve the quality of this manuscript.

References

Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Chapman&Hall.
Breiman, L., 1996. Bagging predictors. Mach. Learn. 24 (2), 123–140.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Carranza, E.J.M., Laborte, A.G., 2015. Random forest predictive modeling of mineral prospectivity with small number of prospects and data with missing values in Abra (Philippines). Comput. Geosci. 74, 60–70.
Cheng, Q., 2007. Mapping singularities with stream sediment geochemical data for prediction of undiscovered mineral deposits in Gejiu, Yunnan Province, China. Ore Geol. Rev. 32 (1–2), 314–324.
Cheng, Q., Agterberg, F.P., Ballantyne, S.B., 1994. The separation of geochemical anomalies from background by fractal methods. J. Geochem. Explor. 51 (2), 109–130.
Chi, Q.H., 2007. Applied Geochemical Element Abundance Data Manual. Geological Publishing House, Beijing.
Cracknell, M.J., Reading, A.M., Mcneill, A.W., 2014. Mapping geology and volcanic-hosted massive sulfide alteration in the Hellyer-Mt Charter region, Tasmania, using

Random Forests™ and Self-organising maps. J. Geol. Soc. Aust. 61 (2), 287–304.
Daya, A.A., 2015. Comparative study of C–A, C–P, and N–S fractal methods for separating geochemical anomalies from background: a case study of Kamoshgaran region, northwest of Iran. J. Geochem. Explor. 150, 52–63.
Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barcelo-Vidal, C., 2003. Isometric logratio transformations for compositional data analysis. Math. Geol. 35 (3), 279–300.
Fu, C., Li, J.J., Tang, W.L., Zhang, F., Ren, J.P., Dang, Z.C., Orolmaa, D., Tumurtogoo, O., 2016. The division and correlation of strata in the mid-western part of Sino-Mongolian border area. Geol. Bull. China 35 (4), 503–518.
Hao, L., Zhao, X., Zhao, Y., et al., 2014. Determination of the geochemical background and anomalies in areas with variable lithologies. J. Geochem. Explor. 139 (1), 177–182.
Hao, L.B., Tian, M., Zhao, X.Y., Zhao, Y.Y., Lu, J.L., Bai, R.J., 2016. Spatial distribution and sources of trace elements in surface soils, Changchun, China: insights from stochastic models and geostatistical analyses. Geoderma 273, 54–63.
Harris, J.R., Grunsky, E.C., 2015. Predictive Lithological Mapping of Canada's North Using Random Forest Classification Applied to Geophysical and Geochemical Data. Pergamon Press. Inc.
Hawkes, H.E., Webb, J.S., 1962. Geochemistry in Mineral Exploration. Harper & Row.
Henderson, B.L., Bui, E.N., Moran, C.J., Simon, D.A.P., 2005. Australia-wide predictions of soil properties using decision trees. Geoderma 124, 383–398.
Ho, T.K., 1998. The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. 20 (8), 832–844.
Hou, W.R., Nie, F.J., Jiang, S.H., et al., 2010. The geology and ore-forming mechanism of the Tsagaan Suvarga large-size Cu-Mo porphyry deposit in Mongolia. Acta Geosci. Sin. 31 (3), 307–320.
Kirwin, D.J., Forster, C.N., Kavalieris, L., et al., 2005. The Oyu Tolgoi copper-gold porphyry deposits, South Gobi, Mongolia. In: Seltmann, R., Gerel, O., Kirwin, D.J. (Eds.), Geodynamics and Metallogeny of Mongolia With a Special Emphasis on Copper and Gold Deposits. IAGOD Guidebook Series. 11, Londonpp. 5–12.
Lark, R.M., 1999. Soil–landform relationships at within-field scales: an investigation using continuous classification. Geoderma 92, 141–165.
Li, J.J., Zhang, F., Ren, J.P., et al., 2015. Tectonic units in China-Mongolia border area and their fundamental characteristics. Geol. Bull. China 34 (4), 636–662.
Liu, C., Chan, Y., Kazmi, S.H.A., et al., 2015. Financial fraud detection model: based on random forest. Int. J. Econ. Financ. 7 (7), 178–188.
Liu, H.L., Nie, L.S., Wang, X.Q., et al., 2018. Regional geochemistry of lithium in the Altay area across the boundary of China and Mongolia. Geoscience 2018 (3), 493–499.
Meng, H.D., Xiao-Qing, J.I., Xiao, Y.L., et al., 2016. Research on random forest classification algorithm based on sensitivity degree in Hadoop environment. J. Inner Mongolia Univ. Sci. Technol. 2016 (3), 297–301.
Nie, F.J., Jiang, S.H., Zhang, Y., et al., 2004. Geological features and origin of porphyry copper deposits in China-Mongolia border region and its neighboring areas. Mineral Deposits 2004 (2), 176–189.
Nie, L. S., Wang, X. Q., Chen, Z., et al. 2013. Study of 1:100 000 Geochemical Mapping in the Border Area of China and Mongolia. Acta Geologica Sinica. 2013(1), 214–215.
Reimann, C., Garrett, R.G., 2005. Geochemical background—concept and reality. Sci. Total Environ. 350 (1), 12–27.
Sreenivas, K., Sujatha, G., Sudhir, K., et al., 2014. Spatial assessment of soil organic carbon density through random forests based imputation. J. Indian Soc. Remote Sens. 42 (3), 577–587.
Tang, W.L., Li, J.J., Fu, C., Dang, Z.C., Orolmaa, D., Tomurtogoo, O., Delgersaikhan, A., 2016. The division and correlation of strata in the mid-eastern part of Sino-Mongolian border area. Geol. Bull. China 35 (4), 488–502.
Team, R.R.D.C., 2007. A language and environment for statistical computing. Computing 1, 12–21.
Tian, M., Wang, X., Nie, L., Zhang, C., 2018. Recognition of geochemical anomalies based on geographically weighted regression: a case study across the boundary areas of China and Mongolia. J. Geochem. Explor. 190, 381–389.
Tolosana-Delgado, R., 2008. Compositional data analysis in a nutshell. www.sediment.uni-goettingen.de/staff/tolosana/extra/CoDaNutshell.pdf.
Tomurtogoo, O., 2005. Tectonics and structural evolution of Mongolia. Geodynamic and metallogeny of Mongolia with a special emphasis on copper and gold deposits. In: Seltmann, R., Gerel, O., Kirwin, D.J. (Eds.), Geodynamics and Metallogeny of Mongolia with a Special Emphasis on Copper and Gold Deposits. IAGOD Guidebook Series 11. Londonpp. 5–12.
Tomurtogoo, O., Badarch, G., 1998. Stratigraphy of Mongolia. Geological map of Mongolia (at the Scale 1:1000000) and its explanatory note. Mineral Resources Authority of Mongolia and Mongolian Academy of Sciences (MRAM-dMAS). Ulaanbaatar, 4–20.
Tukey, J.W., 1977. Exploratory Data Analysis.
Wainwright, A.J., Tosdal, R.M., Lewis, P.D., et al., 2017. Exhumation and Preservation of Porphyry Cu-Au deposits at Oyu Tolgoi, South Gobi Region, Mongolia. Econ. Geol. 112 (3), 591–601.
Wang, X.Q., Chi, Q.H., Liu, H.Y., Nie, L.S., Zhang, B.M., 2007. Wide-spaced sampling for delineation of geochemical provinces in desert terrains, northwestern China. Geochem.: Explor., Environ., Anal. 7 (2), 153–161.
Wang, X.Q., Xu, S.F., Zhang, B.M., Zhao, S.D., 2011. Deep-penetrating geochemistry for sandstone-type uranium deposits in the Turpan–Hami basin, north-western China. Appl. Geochem. 26 (12), 2238–2246.
Wang, X.Q., Zhang, B.M., Lin, X., Xu, S.F., Yao, W.S., Ye, R., 2016. Geochemical challenges of diverse regolith-covered terrains for mineral exploration in China. Ore Geol. Rev. 73, 417–431.
Wiesmeier, M., Steffens, M., Kölbl, A., Kögel-Knabner, I., 2009. Degradation and small-scale spatial homogenization of topsoils in intensively grazed steppes of Northern China. Soil Tillage Res. 104, 299–310.
Yilmaz, H., Cohen, D., Sonmez, F.N., 2017. Comparison between the effectiveness of regional bleg and − 80# stream sediment geochemistry in detection of precious and base metal mineral deposits in western Turkey. J. Geochem. Explor. 181, 69–80.
Zhang, C.S., Fay, D., McGrath, D., Grennan, E., Carton, O.T., 2008. Statistical analyses of geochemical variables in soils of Ireland. Geoderma 146 (1–2), 378–390.
Zhang, H., Wu, P., Yin, A., et al., 2017. Prediction of soil organic carbon in an intensively managed reclamation zone of eastern China: a comparison of multiple linear regressions and the random forest model. Sci. Total Environ. 592, 704–713.
Zuo, R., Cheng, Q., Xia, Q., 2009. Application of fractal models to characterization of vertical distribution of geochemical element concentration. J. Geochem. Explor. 102 (1), 37–43.