ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

Recognition of The Zonal Soil Types of the Forest-Steppe On the Landsat TM Images Using the Logistic Regression Method

Pavel Ukrainskiy*, Alla Zemlyakova, Edgar Terekhin, Olga Marinina, and Zhanna Buryak.

Belgorod State National Research University, 85, Pobedy St., Belgorod, 308015, Russia.

ABSTRACT

The work is dedicated to identifying the soil cover of the forest-steppe. A territory of alternating gray forest soils and black soils, which is located in the center of the Belgorod Region and which is mainly an arable land, has been surveyed. Through the example of recognizing a soil type, a possibility of applying quantitative methods for remote sensing of the qualitative characteristics of the soil has been studied. The logistic regression method is proposed for this purpose. To test the method, a Landsat 5 TM image was used, which served as a source of data about the spectral reflectance of the soils. The image was processed in the ENVI software. The fields with bare soil were highlighted on it. For them, random sample points were generated and the reflectance values were derived from these points. The belonging of these sample points to a certain soil type was determined using the soil maps of the farms located on the territory under survey. The R programming language was used for the statistical data processing. Based on it, a logistic regression model was created. It represents an equation of the connection between the reflectance in the third, fourth and fifth bands of Landsat TM and a certain type of soil (black soil or gray forest soil). This model allows calculating a log odds ratio (logit), which may be then directly re-calculated into probability values for the presence of a certain type of soil. The model is characterized by a high predictive ability (McFadden's pseudo-R2 is 0.94). The advantage of such models is the possibility of showing a gradual transition between soils (using the logit maps or probability maps). They also allow setting distinct quantitative criteria for separating soil types on a satellite image.

Keywords: chernozems, digital soil mapping, gray forest soil, Landsat, logistic regression, soil recognition

**Corresponding author*

INTRODUCTION

The forest-steppe zone of the East European Plain is a region, where the remote sensing studies of the soil cover are currently important and required. On the one hand, it is due to an intensive agricultural development of the territory, on the other hand - due to a relatively high variety of the soil cover. It determines the need for a regular updating of the soil maps and for increasing their accuracy. In addition, the large areas of tilled land are favorable for the remote sensing of the soil. The made the greater part of the soils accessible for a direct aerospace surveillance [1].

The primary task of the soil mapping is the detection of soil boundaries and identification of soils. First of all, this task is solved for the upper levels of the soil classification – the type and sub-type [2]. Quite a few works are dedicated to the recognition of the types and sub-types of the forest-steppe soils [3-5]. Initially, a visual identification was used. It was substituted by an automated identification (classification with learning and autonomous classification) [2, 6]. However, these methods do not use distinct quantitative criteria for the separation of soils. There are attempts to introduce such criteria on the basis of the spectral indices method5. But the most optimal way for this is creating regression models, which link the reflectance and the soil characteristics. A lot of models for determining the quantitative characteristics of the steppe and forest-steppe soils have been developed, for example, for determining the humus content [7-10] and the particle size distribution [5, 11]. But in the domestic practice of identification, no qualitative characteristics of the soil have been determined with the help of regression models so far. This is mainly due to the fact that the necessary methods are not widely practiced yet. Such methods include the logistic regression [12]. In Russia, this method is used predominantly in the works relating to the economics and medicine. But abroad this method has found its application also in the studies of the landscape [13-15] and soils [16-18].

So, the presented study is dedicated to the problem of the remote sensing of the qualitative characteristics of the soil by quantitative methods. A solution of the problem is demonstrated through an example of recognizing zonal soil types of the forest-steppe zone on Landsat 5 TM space images: black soils and gray forest soils. We set a task of creating a logistic regression model allowing their identification.

MATERIALS AND METHODS

As a territory for the study, an area was selected in the centre of the Belgorod Region, on the boundary of the Belgorod District and Shebekino District (Figure 1). It has the shape of a wedge narrowing towards the South of the total area of 23,760ha.

The length of the studied area in the north-eastern direction is 25km. The maximum length from the West to the East is 13km. In the West, the studied territory is limited by the Seversky Donets River, and in the South - by the Nezhegol River. The eastern boundary runs along the edge of the wooded areas stretching along the right bank of the Koren River. The northern boundary runs along the Belgorod–Blizhnyaya Igumenka–Sevryukovo line.

The major part of the studied territory is tilled land. Due to this, its soil cover may be observed on space images in the periods, when it is not covered by agricultural vegetation. Here, most widespread are two zonal soil types of the forest steppe – the black soils and gray forest soil [19]. In the historical past the gray forest soils were covered by deciduous forests. It was a large woodland between the Razumnaya, Koren, Korocha and Hezhegol rivers. By now, a considerable part of these forests has been cut clear. Their boundary shifted towards the East in the direction of the Koren River [20]. Therefore, the major part of the gray forest soils is now a tilled land, and their boundary with the black soils may be seen on space images.

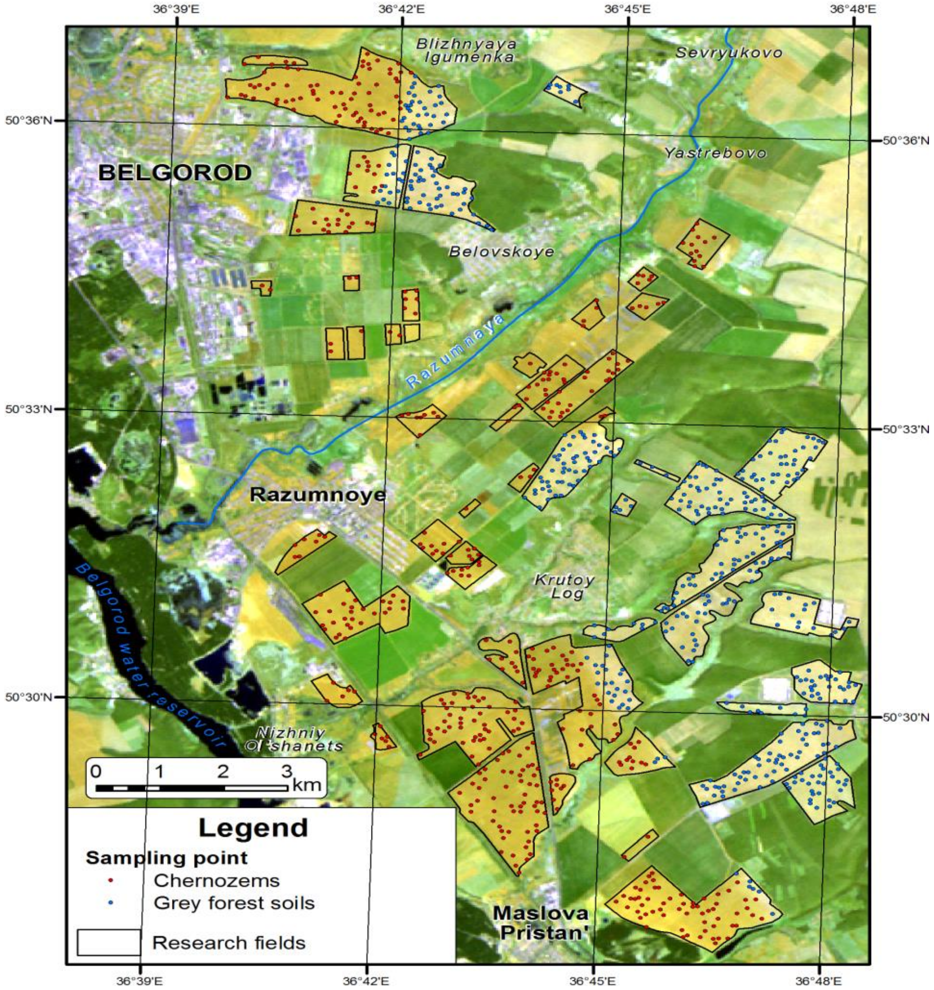


Figure 1: The studied territory (the map is superimposed onto a Landsat 5 TM image in the combination of bands 7:5:3)

To solve the preset task, a satellite image from Landsat 5 TM of May 6, 2007 (path/row 177/025) was used (Figure 1). For the period of 12 days preceding the shooting, there were no precipitations on the territory (according to the Belgorod meteorological station). It is an evidence of the air-dry condition of the soil on the day of shooting. The image was radiometrically calibrated to the reflectance value using the ENVI 4.8 software.

After that, the reflectance values were extracted using ArcGIS 10.1. It was done as follows. During a visual analysis of the image, cropless fields of the total area of 3,260ha were singled out and vectorized (i.e. a mask of bare soil was created) Then, within the bare soil mask, random sample points were generated. Their number was set equal for the both soil types (500 for each). The correspondence of these sample points to the soil types was verified using the soil maps of the farms located on the territory under survey. Using a zonal statistics tool, the reflectance values were extracted in these points from each band. Such approach to extracting data from images is used in many works relating to the application of the logistic regression for the soil and landscape interpretation [14, 16-18].

The programming language R, version 3.2.2, was used for the statistical data processing. Beside the R basic packages, a number of additional packages was used: car, ggplot2, pscl [21-24]. The logistic regression model was created using the glm() function from the stats package (included into the basic configuration of R). The logit conversion is used to enable modeling of the qualitative characteristics by quantitative methods. It consists in calculating the natural logarithm of the odds ratio (logit). The logistic regression model is as follows [12]:

$$l = \ln(P1/P2) = \ln(P1/(1-P1)) = C0 + C1\times X1 + \ldots + Cn\times Xn, \quad (1)$$

where l – logit, P1 – the probability of the object's belonging to Class 1 (in our case, it is the gray forest soil), P2 – the probability of the object's belonging to Class 2 (in our case, it is the black soil), C0, C1, Cn – regression coefficient, X1, Xn – independent variables (predictors).

The logits are re-calculated into the probability value using the formula:

$$P1 = (\exp l)/(1+\exp l), \quad (2)$$

the keys to Equation 2 are given in the description of Equation 1.

The independent variables in the currently developed model are the reflectance values in various bands of the Landsat image. The bands for independent variables were selected to meet the following requirements: the maximum possible number of independent variables, absence of redundant variables, statistical significance of the model coefficients, maximum predictive ability of the model.

The check for redundant variables was performed by calculating the variance inflation factor (VIF) [25]. The calculation was done using the vif() function from the additional car package. As a redundancy criteria for variables, the VIF value exceeding 5 was adopted. The predictive ability of logistic regression models is evaluated by the value of pseudo-R2. It takes the value of 0 to 1, and its interpretation is similar to R2 for linear regression models. But, as other ways of calculation are used, it is assumed to be designated with the prefix "pseudo" [26]. The most widely used is the calculation of the pseudo-R2 according to McFadden (R2MF) [27]. To determine this indicator, the pR2() function from the additional package of pscl was used. All graphs in this work were built using the basic graphic package of R (graphics) and, also, an additional package of ggplot2.

The work was completed by creating in the ENVI program of a map of the logarithm of the odds ratio for the gray forest zone (logit), which was subsequently transformed into a probability map (using formula 2). For this purpose we used a tool of mathematical operations with patterns, into which the developed regression model was incorporated.

RESULTS AND DISCUSSION

For separating zonal soils of the forest steppe on the space images, a logistic regression model was created (formula 3). It was designed for application with the Landsat TM imagery. Potentially, it may be used also for the Landsat ETM+ and Landsat OLI imagery, which have bands similar in their spectral range. The reflectance in the bands of the Landsat TM sensor, expressed in percentage, are used as the independent variables (predictors) of the model.

Out of all possible variants of the model, the three-predictor model proved to be the most efficient. The models with six, five and four predictors had either statistically insignificant coefficients, or redundant predictors. While models with two or one predictor had a lower R2MF value compared with the three-component models, i.e. they produced a poorer description of the modeled object.

Among the three-component models, the highest R2MF value of 0.94 had the model based on the third, fourth and fifth bands of the Landsat TM sensor. This model (Formula 3) describing the dependence of the spectral reflectance of the soil on its type looks as follows:

$$l = -32{,}64 + 5{,}50\times B3 + 0{,}45\times B4 - 0{,}88\times B5, \quad 3)$$

where l – logit of odds ratio for the gray forest soil, B3 – reflectance in percentage in the third band of the TM sensor, B4 – reflectance in percentage in the fourth band of the TM sensor, B5 – reflectance in percentage in the fifth band of the TM sensor.

The characteristics of the model coefficients (3) are given in Table. All coefficients are statistically significant (p<0,05) and are not redundant (VIF<5)

Table: Characteristics of the model coefficients

Coefficients	Estimate	Standard error	z-score	p-value	VIF
Intercept	-32,64	4,27	-7,64	2,26*10-14	-
B3	5,50	0,84	6,57	5,01*10-11	3,01
B4	0,45	0,20	2,28	0,02	1,15
B5	-0,88	0,24	-3,64	0,0003	2,80

The created model allows building maps of the logit of odds ratio for the gray forest soil. An example of such map is shown in Figure 2.

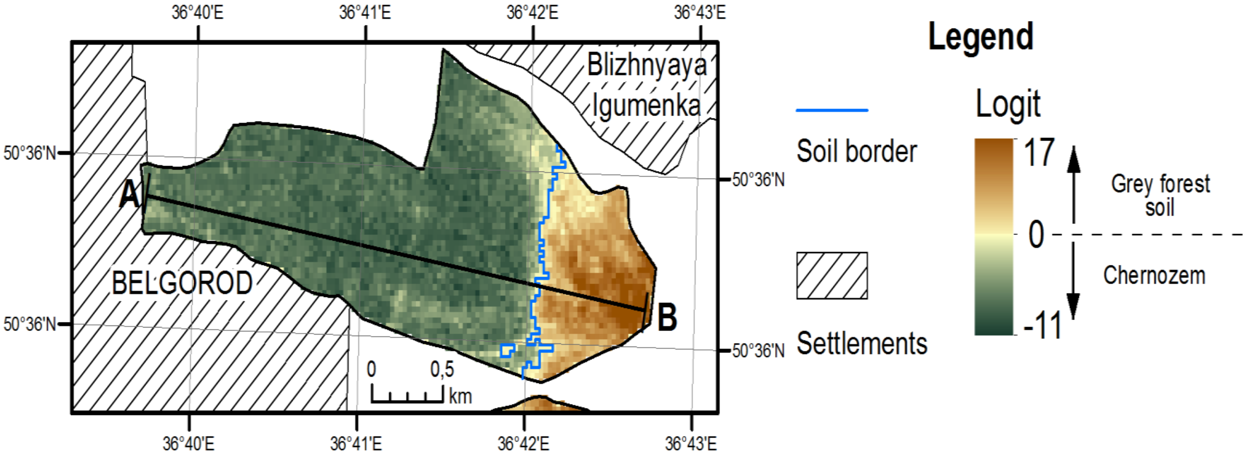
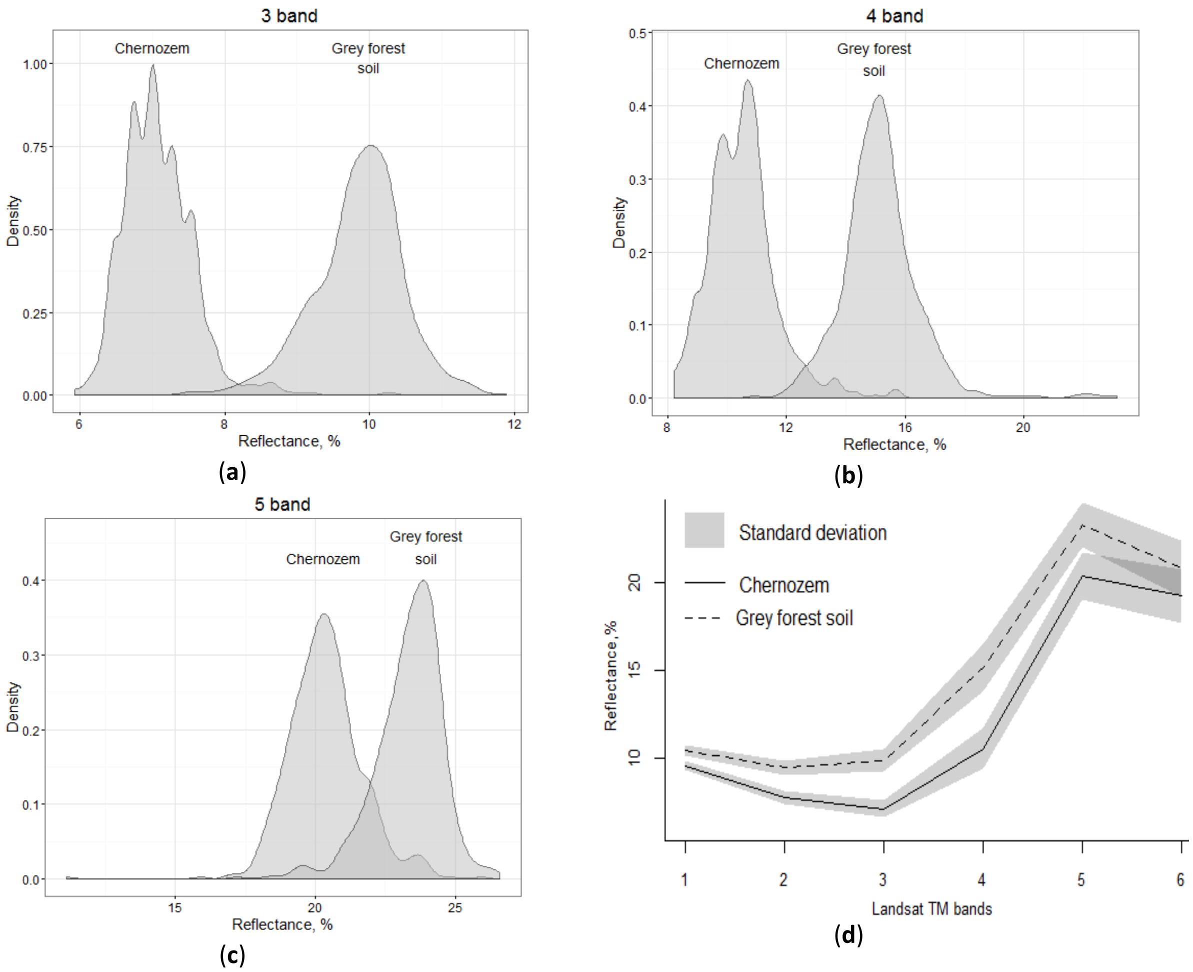


Figure 2: Map of values of the logarithm of odds ratio (logit) for gray forest soil

For the illustrative purposes, a field is specially selected with the soil boundary running across it. The gray forest soil is characterized by positive values of the logit, while the black soils are characterized by negative values. The soils boundary runs over the pixels with the zero value of the logit. On both sides of the boundary there runs a narrow strip of the logit values close to zero (shown in yellow in Figure 3). This is a zone of a gradual transition from one soil to another.

Of all the predictors of the created model, the B3 variable, which characterizes the reflectance in the third band of Landsat TM (i.e. in the red part of the spectrum), has the highest absolute value of the coefficient. Then, in decreasing order, follow the absolute values of the coefficients relating to the B4 and B5 variables (Table). It coincides well with the degree of separability in this bands of the samples of the black soil and gray forest soil. From the diagrams of the kernel density estimation of these samples (Figure 3a-c) it is evident that the smallest overlapping is in the third band, while the largest one is in the fifth band. When comparing the spectral reflectance curves of the black soil and gray forest curve it is visible that in the third, fourth and fifth bands the distance between the curves is the largest. And it decreases with the transition from the third band to the fifth band (Figure 3d).



(a)

(b)

(c)

(d)

Figure 3: The difference in the spectral reflectance of the black soil and gray forest soil in the Landsat TM (A) bands, the fourth band (Б), the fifth band (В) of Landsat TM, and across the whole spectrum (Г): in the third band (a); in the fourth band (b); in the fifth band (c); and across the whole spectrum (d)

The fact, that the reflection in the third band exceeds many times the contribution of the fourth and fifth band, has its scientific grounds. From the graph of the spectral reflectance of the soils it is evident that the red part of the spectrum accounts for the maximum solar radiation absorbed by the studied soils (Figure 3d). This range is characterized by the most intensive absorption of the light by the humus [1]. And it is the difference in the humus content that accounts for the difference in the spectral reflectance of the black soils and gray forest soils [1]. And here both the humus content and the ratio of the humic acids and fulvic acids in the humus play their role [28].

The possibility to render the gradual change in the soils is an important merit of the logistic regression method. The actual soil boundaries are indistinct. The distinct boundaries of the plots on the traditional soil maps is a forced simplification due to the problem of rendering a gradual transition between soils. On the space images this gradual change is rendered if it is visible to the interpreter. But the manual vectorizing (like the automated classification) creates distinct boundaries. While the application of logistic regression models allows resolving this problem.

When using logistic regression methods, a question arises which indicator is best to operate for separating the soils – logit or probability? For the purpose of comparing these indicators, we made a profile crossing the soils boundary (Line AB in Figure 2). The change in the value of the logit and probability along the profile line in shown in Figure 4.

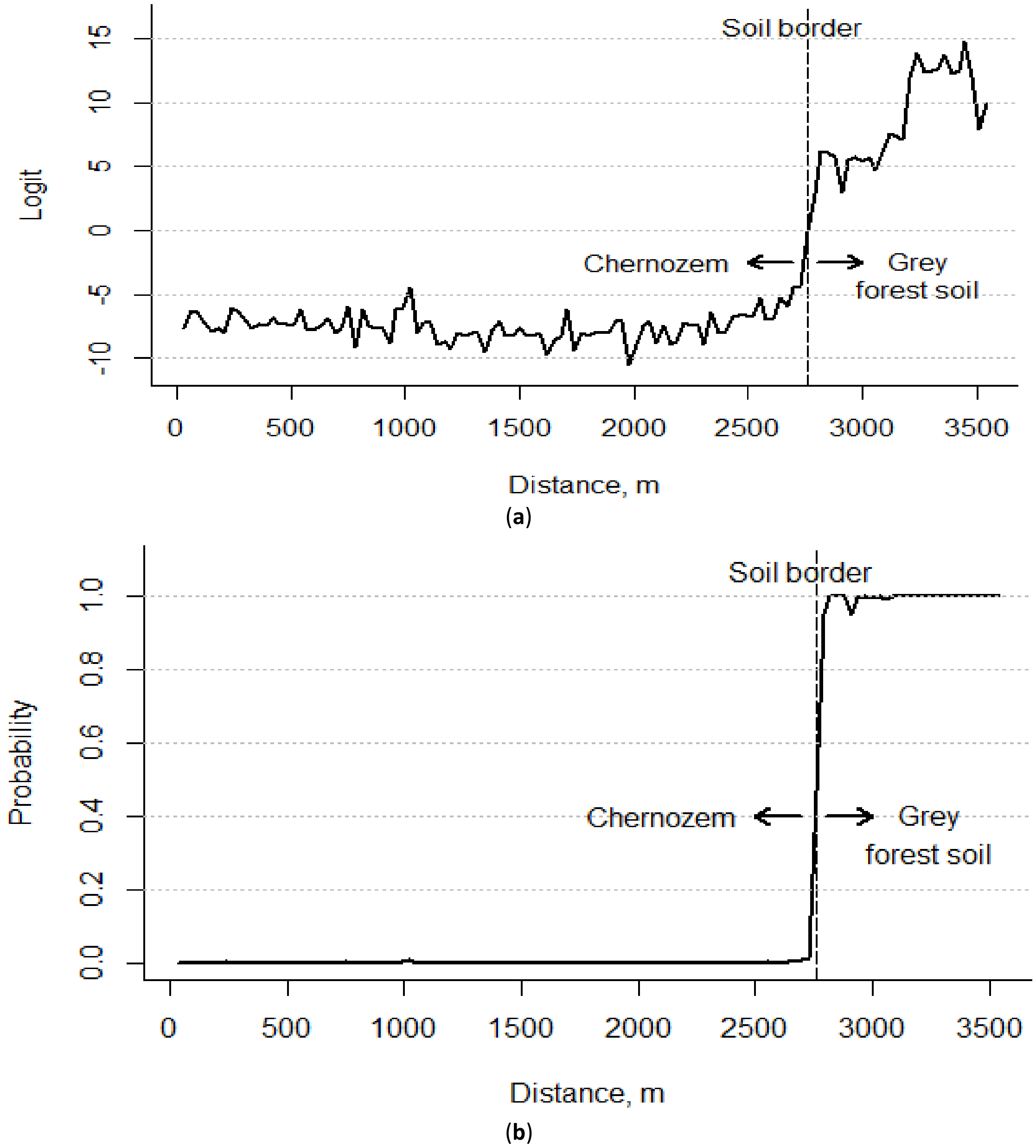


Figure 4: The values of the logit (a) and detection probability (b) for the gray forest soil along the profile AB

The logit is unusual for comprehension as compared with the probability. But it is quite convenient for separating the modeled objects. The principal difference between the logit and probability is the boundaries of the range of assumed values. The logit values change from zero to infinity. While the probability varies from 0 to 1. Correspondingly, the probabilities are better for the purpose of comparison. Such comparison may be required in case of creating a series of logistic regression models built for different soils on the basis of different predictors. In this case the image pixels are assigned to the soil with the greater detection probability.

CONCLUSION

The conducted study showed that the logistic regression allows solving the task of the remote determination of the qualitative characteristics of the soil by quantitative methods. The application of this method may be useful for the geosciences, in particular, in the thematic mapping of the soils. The logistic regression model obtained in the course of the study allows a highly efficient separation of zonal soil types of the forest-steppe zone (black soils and gray forest soils) on the Landsat TM images. From the point of view of the mapping tasks, an important advantage of such models is the possibility of showing on the map the gradual transition between soils. Prom the point of view of the practice of the automated soil interpretation, the advantage of the logistic regression models is the possibility of introducing distinct numeric criteria for the soil separation and to eliminate, as much as possible, the subjective factor from the interpretation process.

ACKNOWLEDGMENT

Results were obtained in the framework of the state assignment of the Ministry of Education and Science of the Russian Federation No 5.78.2014/K.

REFERENCES

[1] Kravtsova V I. Space methods of soil research. Aspekt Press: Moscow, Russia, 2005. 190 p. (in Russian)

[2] Simakova M S. From visual aerial photo interpretation and field soil survey to automated decoding and soil mapping by satellite imagery. Byulleten' instituta im. V.V. Dokuchaeva. 2014; 74:3–19. Available online: https://yadi.sk/i/kf1UkWMnfhjrW. Date accessed: 24/06/2016. (in Russian)

[3] Ovechkin S V, Savin I Yu. Interpretation of satellite images on ecological and soil characteristics of forest-steppe zone the Central Russian Upland. Aerospace methods in soil science and its use in agriculture. Nauka: Moscow, Russia; 1990. 103–108. (in Russian)

[4] Mazikov V M. Distantsionnaya indikatsiya svoistv pochv i pochvennogo pokrova. Dr. geogr. sci. thesis, Institute of Geography, RAS, Moscow, 2001. 222 p. (in Russian)

[5] Ukrainskii P A. Assessment of agricultural land by fertility indicators for remote monitoring of land (on the example of the Belgorod region), cand. geogr. Sci. thesis, VGPU, Voronezh, 2011. Available online: http://93.88.136.1/files/08_10_2011/ref_ukr.pdf. Date accessed: 24/06/2016. (in Russian)

[6] Savin I Yu, Simakova M S. Satellite technologies for soil inventorying and monitoring in Russia, Sovremennye problemy distantsionnogo zondirovaniya Zemli iz kosmosa. 2012; 9:104–115. Available online: http://d33.infospace.ru/d33_conf/sb2012t5/104-115.pdf. Date accessed: 24/06/2016. (in Russian)

[7] Shatokhin A V. Remote indication of humus content in the soils of the forest-steppe and steppe zones of Ukraine. Agrokhimiya. 1998; 6:21–25. (in Russian)

[8] Shatokhin A V, Achasov A B. Use of Modern Technologies for Mapping the Soil Cover of the Northern Donets Steppe). Pochvovedenie. 2005; 7:790–798. (in Russian)

[9] Terekhov A G, Kauazov A M. Method of estimation of content of humus in the arable lands of northern Kazakhstan on the basis of satellite data. Sovremennye problemy distantsionnogo zondirovaniya Zemli iz kosmosa. 2007; 4:358–364. Available online: http://d33.infospace.ru/d33_conf/vol2/358-364.pdf. Date accessed: 24/06/2016. (in Russian)

[10] Ukrainskii P A, Narozhnyaya A G, Gagina I S. On the possibility of modeling relationship between humus content and soil reflectance based on data tradition agrochemical surveys and multispectral satellite imagery Landsat 8 OLI. The Agrarian Scientific Journal. 2015; 12:29–32. Available online: http://www.sgau.ru/files/pages/846/14513759090.pdf#page=29. (in Russian)

[11] Ukrainskii P A, Chepelev O A. Studing of soil texture at Pooskolye according to space imagery decoding. Izvestiya Samarskogo nauchnogo tsentra Rossiiskoi akademii nauk. 2011; 13:1225–1229. Available online: http://www.ssc.smr.ru/media/journals/izvestia/2011/2011_1_1225_1229.pdf. Date accessed: 24/06/2016. (in Russian)

[12] Hosmer D W, Lemeshow S. Applied Logistic Regression. 2nd ed.; John Wiley & Sons: New York, USA, 2000. 383 p. DOI: 10.1002/0471722146.ch1.

[13] Bricklemyer R S, Lawrence R L, Miller P R. Documenting no-till and conventional till practices using Landsat ETM+ imagery and logistic regression. J. Soil Water Conserv. 2002; 57:267–271.

[14] Pradhan B, Lee S, Manso S, Buchroithner M, Jamaluddin N, Khujaimah Z. Utilization of optical remote sensing data and geographic information system tools for regional landslide hazard analysis by using binomial logistic regression model. J. Appl. Remote Sens. 2008; 2:1–11. DOI: 10.1117/1.3026536.

[15] Debella-Gilo M, Etzelmüller B. Spatial prediction of soil classes using digital terrain analysis and multinomial logistic regression modeling integrated in GIS: Examples from Vestfold County, Norway. Catena. 2009; 77:8–18. DOI: 10.1016/j.catena.2008.12.001.

[16] Campling P, Gobin A, Feyen J. Logistic modeling to spatially predict the probability of soil drainage classes. Soil Sci. Soc. Am. J. 2002; 66:1390–1401. DOI: 10.2136/sssaj2002.1390.

[17] Kempen B, Brus D J, Heuvelink G B M, Stoorvogel J J. Updating the 1:50000 Dutch soil map using legacy soil data: A multinomial logistic regression approach. Geoderma. 2009; 151:311–326. DOI: 10.1016/j.geoderma.2009.04.023.

[18] Jafari A, Finke P A, Vande W J, Ayoubi S, Khademi H. Spatial prediction of USDA-great soil groups in the arid Zarand region, Iran: comparing logistic regression approaches to predict diagnostic horizons and soil types. Eur. J. Soil Sci. 2012; 6:284–298. DOI: 10.1111/j.1365-2389.2012.01425.x.

[19] Solovichenko V D, Tyutyunov S I. The soil cover of the Belgorod region and his rational use, Belgorod: Izd-vo “Otchii krai”; 2013. 371 p. (in Russian)

[20] Chendev Yu G. Agrotechnogenic transformation of dark gray soil in the central forest-steppe zone during the last 200 years. Eurasian Soil Science. 1997; 30:5–15.

[21] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing: Vienna, Austria, 2016. Available online: https://www.R-project.org. Date accessed: 24/06/2016.

[22] Fox J, Weisberg S. An {R} Companion to Applied Regression, 3rd ed. Sage: Thousand Oaks, USA, 2011. Available online: http://socserv.socsci.mcmaster.ca/jfox/Books/Companion. Date accessed: 24/06/2016.

[23] Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York: New York, USA, 2009. DOI: 10.1111/j.1467-985X.2010.00676_9.x.

[24] Jackman S. pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory, Stanford University. Department of Political Science, Stanford University: Stanford, USA, 2015. Available online: http://pscl.stanford.edu. Date accessed: 24/06/2016.

[25] O’brien R M. A caution regarding rules of thumb for variance inflation factors. Qual. Quant, 2007; 41:673–690. DOI: 10.1007/s11135-006-9018-6.

[26] Veall M R, Zimmermann K F. Pseudo-R2 Measures for Some Common Limited Dependent Variable Models. J. Econ. Surv. 1996; 10:241–259. DOI: 10.1111/j.1467-6419.1996.tb00013.x.

[27] McFadden D. Conditional logit analysis of qualitative choice behavior. Frontiers in econometrics; Academic Press: New York, USA, 1973. 105-142.

[28] Karavanova E I. Optical properties of soils and their nature [monograph online]. Moscow: Izd-vo Mosk. un-ta, Russia, 2003. 153 p. http://soil.msu.ru/attachments/article/1366/