RESEARCH PAPER

SPATIAL VARIATION OF FEATURE DENSITY IN MULTISCALE TOPOGRAPHIC DATA

Timofey E. Samsonov[1*], Olga P. Yakimova[2], Daniil A. Potemkin[2] and Olga A. Guseva[3]

[1] Lomonosov Moscow State University, Faculty of Geography, Leninskiye Gory 1, 119234, Moscow, Russia

[2] Demidov Yaroslavl State University, Faculty of Mathematics, Souyznaya str. 144, 150008, Yaroslavl, Russia

[3] Demidov Yaroslavl State University, Faculty of Biology and Ecology, Matrosova lane 9, 150057, Yaroslavl, Russia

***Corresponding author:** tsamsonov@geogr.msu.ru

Received: August 17th, 2022 / Accepted: February 15th, 2023 / Published: March 31st, 2023

https://DOI-10.24057/2071-9388-2022-127

ABSTRACT. Digital topographic maps are created in a series of scales from large to small, and the underlying spatial data is commonly organized as a multiscale database consisting of several levels of detail (LoDs). Spatial density of features (or spatial objects) in such database varies both between LoDs (coarser levels are less densely populated with features) and within each LoD (feature density changes over the area). While the former type of density variation is caused by generalization, the latter one is mainly conditioned by geographic location and its properties, such as landscape complexity or fraction of urban areas. Since topographic database LoDs are derived using different data sources and generalization techniques, there is a need for a method that can help with automated evaluation of resulting feature density in terms of its appropriateness for the specified location and level of detail. This paper provides such method by uncovering dependencies between the location properties and the density of spatial data in multiscale topographic database. Changes in feature density are modeled as a function of spatial (landscape complexity and terrain ruggedness) and non-spatial (land cover types ratio) measures estimated via independent data sources. Resulting model predicts how much higher or lower is the expected spatial density of features over the area in comparison to the average density for the LoD. This information can be used further to assess the fitness of the data to the desired level of detail of the topographic map.

KEYWORDS: level of detail; topographic data; spatial databases; multiscale mapping

CITATION: Samsonov T. E., Yakimova O. P., Potemkin D. A. and Guseva O. A. (2023). Spatial variation of feature density in multiscale topographic data. Geography, Environment, Sustainability, 1(16), 86-102
https://DOI-10.24057/2071-9388-2022-127

ACKNOWLEDGEMENTS: The part of the study on vector features and land cover analysis was funded by Russian Foundation for Basic Research (RFBR) according to the research project 18-07-01459-a. The work of Timofey Samsonov on density of terrain features has been funded by Russian Science Foundation (RSF) grant No 19-77-00071.

Conflict of interests: The authors reported no potential conflict of interest.

INTRODUCTION

Spatial data for topographic mapping are commonly derived at multiple levels of detail (LoDs) which comprise a multiscale topographic database (Jones and Abraham 1986; Kilpeläinen 2000). Official standards for topographic map compilation are essentially definitions of level of detail which correspond to a specific map scale (Military Topographic Service 1978, 1980, 1985). The standards prescribe multiple rules for selection and generalization of features (spatial objects), as well as precision of their representation. Hence, LoD cannot be easily defined as one number such as scale, and definitions vary significantly. Meng and Forberg (2007) describe LoD as an arbitrary milestone in scale-space continuum which corresponds to a certain degree of generalization. Lemmens (2011) understands LoD as a combination of resolution and the amount of spatial, temporal and semantic detail. Ruas and Bianchin (2002) conceptualize LoD of a spatial database as a combination of the conceptual schema of the data, the semantic resolution, the geometric resolution, the geometric precision, and the granularity. In many cases an LoD can be effectively defined as a specific combination of elements which have a particular size or granularity. This approach is used in 3D city modeling (Kolbe, Gröger, and Plümer 2005; Biljecki, Ledoux, and Stoter 2016), where each level of detail is defined by a specific set of building elements. Samsonov (2022) identified the typical granularity of terrain features selected for small-scale cartographic relief presentation, which is 5-6 mm at mapping scale.

Since rules for LoD derivation can be quite sophisticated, the differences and inconsistencies in LoD can be inferred using the machine learning methods (Touya and Brando-Escobar 2013). In a pursuit of a universal and effective approach to LoD estimation raster analysis methods are developed as well. In particular, a detail resolution method by Cheng et al. (2017) is based on calculation of a rasterized line coalescence. The similar approach can be applied to describe the legibility of individual spatial features (Cheng, Liu, and Zhang 2021).

The notion of LoD can be also traced through the literature on cartographic generalization (or generalization of spatial data). In particular, formalized LoD-based representations are widespread in surface (especially TIN-

based) modeling where the precision of the resulting LoD is defined by simple metric criteria such as vertical error (de Floriani, Marzano, and Puppo 1996). The similar criteria usually expressed in terms of distances, areas or point densities are used in geometric simplification of lines (Douglas and Peucker 1973; Visvalingam and Whyatt 1993; Li and Openshaw 1992), polygons (Buchin et al. 2016; Haunert and Wolff 2010) or in point selection (Töpfer and Pillewizer 1966). Since the LoD is itself a complex notion, its reduction during generalization most probably should be expressed as a combination of multiple characteristics. Such approach was tested by Samsonov and Yakimova (2020), where the authors achieved a similar change in level of detail by joint alteration of Modified Hausdorff Distance and the number of line bends.

Despite a steady interest in the detail-related issues in geographical information science, such investigations remain quite rare. At the same time, one of the most critical requirements for spatial data used in analysis or mapping is the appropriate level of detail or degree of generalization. Heterogeneous natural conditions produce different spatial patterns, some of which can be characterized as a complex interplay of different land cover types and underlying surface (Phillips 1999). Specifically, mountainous areas are characterized by complex terrain, and therefore tend to require denser representation of relief (Imhof 1982). Economically developed, especially urban areas, are characterized by complex configuration of spatial elements (Batty 2013). Topographic maps respond to this by denser patterns of spatial features. From a cartographic point of view, the question arises how much this density should variate over the area, and is there any way to determine if the data underlying the topographic map is appropriate for the selected level of detail and location. To date, no formalized methods have been developed for this purpose. To bring the problem closer to solution we developed a new approach which considers the *relative feature density* – i.e., how much denser the features are over the selected area in relation to the average density for the whole LoD. This property is modeled as a function of location properties which are expressed in a number of spatial (landscape complexity and terrain ruggedness) and non-spatial (ratio between various land cover types) measures estimated via independent detailed data sources.

The rest of the paper is organized into five sections. In the Materials and Methods section, we introduce the notion of relative feature density and then conceptualize our approach to model it as a function of location properties. Experimental work subsection sheds the light on topographic and land cover data used for the case study, as well as their preprocessing needed to construct the desired model. In the Results section we demonstrate how our approach can be effectively used to improve the prediction of LoD based on feature density. Limitations of the approach are settled in the Discussion section. Finally, the main insights gained during the research are summarized in the Conclusion.

MATERIALS AND METHODS

General formulation

The aim of the developed method is to model the relationships that exist between the geographic location and density of features in topographic data. For the sake of brevity, we will use the term density to name the feature density, unless other type of density is explicitly defined. The flowchart of the method in general form is represented in Fig. 1.

We start from some abstract density measure d calculated for each training data fragment (Fig. 1a). Since the expected value of d varies with LoD over the same area, we divide it by the mean $\overline{d}$ for that LoD to obtain the relative density $\hat{d}$(Fig. 1b):

$$\hat{d} = d / \overline{d}$$

Relative density shows how much denser is the LoD fragment in relation to the average density over the whole LoD. Relative density is then modeled as a function of location properties:

$$g\left(\hat{d}_i\right) = \sum\nolimits_j \beta_{ij} f\left(l_j\right) + \varepsilon$$

where $\hat{d}_i$ is i-th relative density measure, l_j is the value of j-th location property, ε is a free term, g and f are linearizing transformation functions specific to the pair of $\hat{d}_i$ and l_j, and β_{ij} are the coefficients. *While building the model, we expect that spatial distribution of the relative density is similar for all LoDs.* Therefore, relative densities are merged into one sample (Fig. 1c) and then used in model fitting with location properties extracted for the same areas (Fig. 1d).

For any new topographic data fragment with known values of l_j the relative density $\hat{d}_i$ can be predicted as $\hat{d}_i = g^{-1}\left(\sum_j \beta_{ij}(l_j) + \varepsilon\right)$ (Fig. 1e). Let's assume that the model is $\ln\left(\hat{d}+1\right) = 0.74\ln\left(l_i+1\right) + 0.2$ and $l_1 = 2$ for the new data. Then the predicted feature density will be $\hat{d}_1 = e^{0.74\ln 3 + 0.2} - 1 \approx 1.753$ times higher at that location than $\overline{d}_1$ (the mean for whole LoD).

Having the actual density d_i for the new data fragment and its predicted relative value $\hat{d}_i$, we can obtain a *normalized density* (Fig. 1f):

$$\tilde{d}_1 = d_i / \hat{d}_i$$

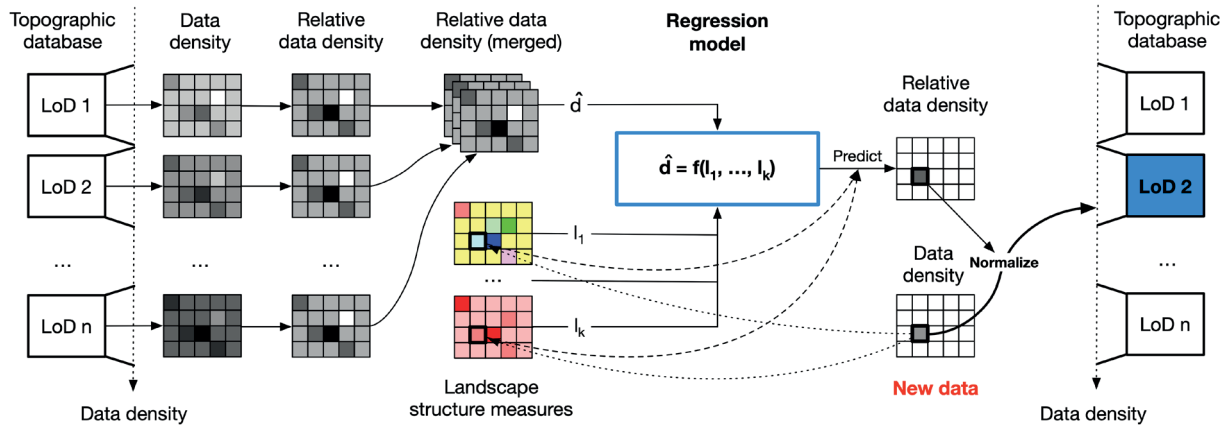

Fig. 1. The flowchart of the method

If the specific implementation of the model (2) is effective, then $\tilde{d}_i$ should be equal to the $\bar{d}_i$ of the desired LoD for well-prepared data. It is expected that for the real-world cases there will be some difference between these values. If the densities (or their transformed values) are distributed normally, then *Z*-score can be used to measure the difference in a statistical way:

$$z_i = \frac{\tilde{d}_i - \bar{d}_i}{s_i}$$

where s_i is a standard deviation of *i*-th density measure.

Finally, it is important not only to calculate the difference, but also to assess its ability to differentiate LoDs effectively. It is expected that for any fragment of k-th LoD its Z-score calculated against its own mean and standard deviation is smaller than a Z-score calculated against the mean and standard deviation of any other LoD (Fig. 1g):

$$\left|z_i^k\right| < \left|z_i^m\right|,\ m \neq k$$

This hypothesis is the main objective tested in the experimental part of our work.

Specific implementation

For this study points, lines and intersections were selected as features which densities are modeled. The corresponding density measures are calculated as follows:

• *Point density* (d_p). Each spatial data feature is converted to the point features. For linear and polygonal features their vertices are extracted. The total number of resulting points is divided by the area covered by the data.

• *Line density* (d_L). The total length of all linear features and the total perimeter of polygonal features are summed and then divided by the area covered by the data.

• *Intersection density* (d_I). An overlay of all linear features and borders of polygonal features is computed. Resulting geometry is set to be point geometry – it means that all intersections between linear and polygonal layers are derived. The number of intersections is divided by the area covered by the data.

While point and line density characterize the total abundance of spatial data that cover the area, intersection density encodes the complexity of topological relations between the features in the database: more intersections indicate more complex pattern of the features.

Location properties act as density predictors. Two groups of measures were considered for this purpose:

• *Non-spatial measures* describe general properties of location and do not account for the shape and spatial pattern of geographic objects that cover the area. Specifically, we use the ratios occupied by different land cover classes such as water, forest, urban and others.

• *Spatial measures* characterize the location through geometry and shape of geographic features that cover the area. For this we used the landscape complexity measures (joint entropy, contagion index, fractal dimension) and terrain ruggedness indices described below.

Joint entropy describes the overall complexity of the landscape pattern (Nowosad and Stepinski 2019):

$$joinet = -\sum_{i=1}^{K}\sum_{j=1}^{K} p_{ij}\log_2 p_{ij}$$

where p_{ij} is probability that *i*-th and *j*-th class are observed in neighboring raster cells, and *K* is the total number of land cover classes.

Contagion index is calculated in a similar way (Riitters et al. 1996):

$$contag = 1 + \frac{\sum_{i=1}^{K}\sum_{j=1}^{K} p_{ij}\ln p_{ij}}{2\ln K}\ \#$$

but accounts for the number of classes. *contag* describes the probability of two random cells belonging to the same class.

Perimeter-Area Fractal Dimension measures the complexity of landscape patches shape and is calculated as (Burrough 1981):

$$pacfac = \frac{2}{\beta}$$

where β is the slope of the regression of landscape patch area A_i against the patch perimeter P_i for all *n* patches in the landscape:

$$\sum_{i=1}^{n} \ln A_i = a + \beta \sum_{i=1}^{n} \ln P_i$$

The value of *pafrac=1* if patches are simple (squares, circles) and for irregular shapes with high fractal dimension.

Terrain ruggedness index is essentially a vertical distance between the central cell of a moving window and its surrounding cells calculated in a raster digital elevation model (Riley, De Gloria, and Elliot 1999):

$$tri = \sqrt{\sum_{i=-1}^{1}\sum_{j=-1}^{1}\left(z_{ij} - z_{00}\right)^2}$$

where z_{00} is a central cell of the floating window.

Data preparation

Experimental evaluation of the method was performed on multiscale topographic database with 3 levels of detail corresponding to 1:200 000, 1:500 000 and 1:1 000 000 mapping scales (referred further as 200, 500 and 1000 LoDs). The database represents layers of digital Russian topographic maps of the corresponding scales which were compiled by The Federal Service for State Registrations, Cadaster and Cartography (Rosreestr) using the generalization of larger-scale maps. Each LoD is represented in Esri geodatabase storage format and contains 47/47/40 layers for 200/500/1000 LoD respectively. The layers in each LoD are grouped into eight feature datasets inside each geodatabase: administrative (3/2/2 layers), economy (7/9/6 layers), geodesy (3/3/2 layers), hydrography (11/11/10 layers), relief (5/5/5 layers), settlements (6/5/4 layers), transport (6/6/6 layers) and vegetation/ground (6/6/5 layers). The number of layers slightly differ between LoDs because some types of objects are removed or added between scales.

33 sample fragments centered on settlements located in different geographic conditions were extracted from each level of detail, resulting in 99 data fragments in total. Each fragment was clipped by 100×100 km rectangle, and then projected into Lambert Azimuthal Equal Area Projection with corresponding center. This projection was selected because it allows each fragment to cover the similar area. The ratio between different land cover types is also correct, while other distortions are negligible within the extent of each fragment. The map of sample fragments' locations is represented in Fig. 2. Samples were divided into training and testing groups, which is explained later in the Experimental work section. The possible difference in feature density between fragments can be judged from Fig. 3. It can be clearly seen that highly urbanized Moscow fragment is characterized by significantly higher feature density at each LoD.

To describe the location, we used external data sources derived independently of topographic data, with better detail and generated without cartographic generalization. The main data source is Copernicus Global Land Cover (CGLC) (Buchhorn et al. 2020), which is a recent high-quality 100 m resolution global raster dataset obtained

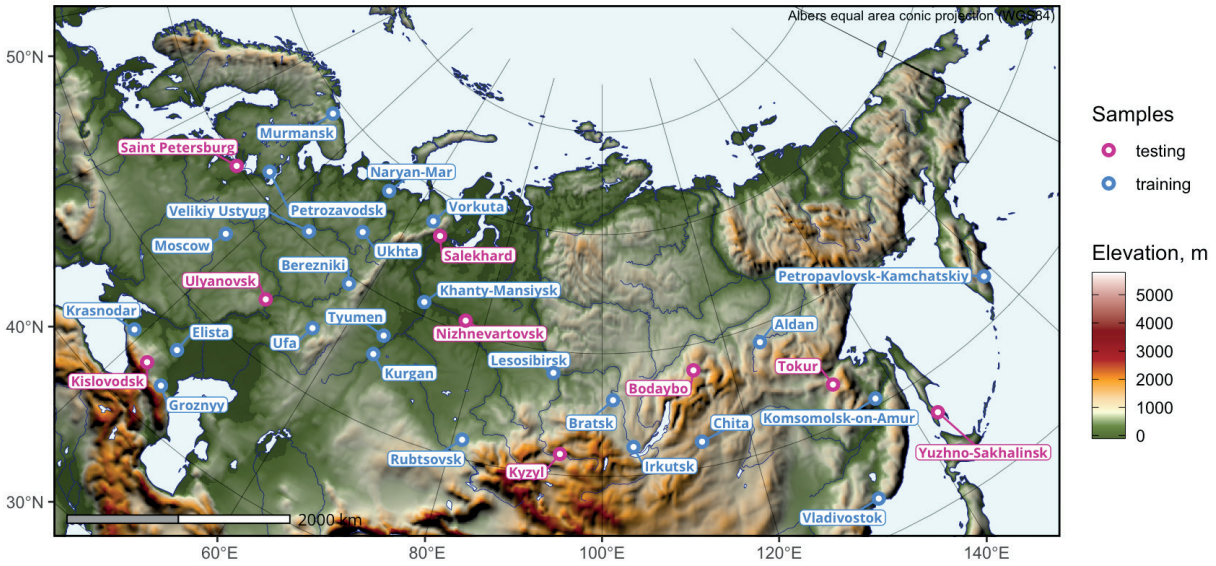

Fig. 2. Locations of sample fragments

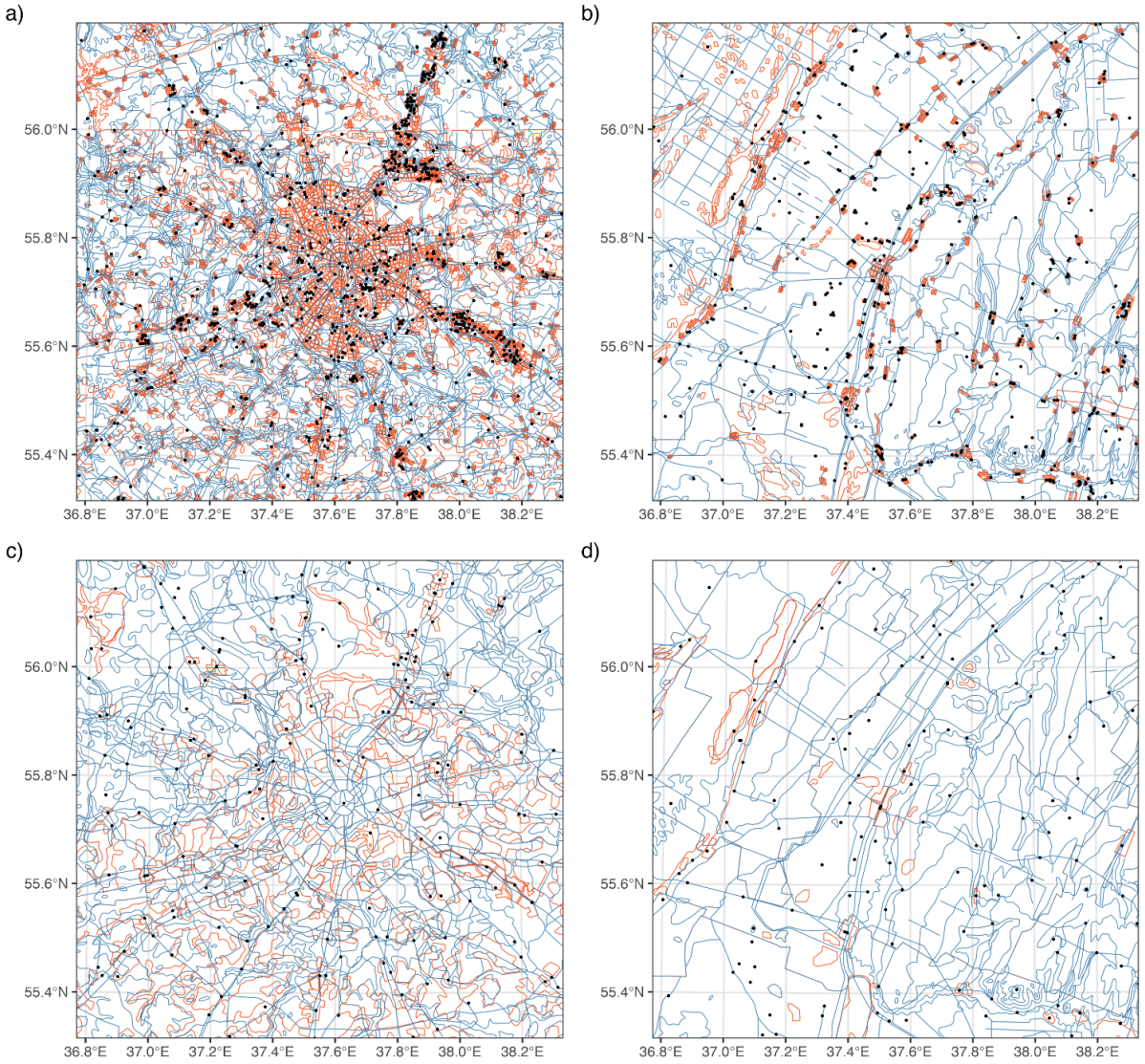

Fig. 3. Example renderings of 500 and 1000 database LoDs for two fragments: (a) Moscow, 500 LoD; (b) Rubtsovsk, 500 LoD, (c) Moscow, 1000 LoD; (d) Rubtsovsk, 1000 LoD. Point, linear and polygon features are shown by black, blue and red color correspondingly. Each fragment is 100 per 100 km. Lambert Azimuthal Equal Area Projection

by classification of satellite imagery. From this dataset 33 fragments were extracted which cover the same area and have the same projection as samples from the database. All land cover extracts are represented in Appendix Fig. A.1.–A.3. with colors according to the official CGLC legend in Appendix Table A.1. which includes 23 classes. These figures illustrate the variety of landscapes selected for the case study, which corresponds to initial selection of fragments. In particular, there are fragments with few (e.g. Aldan, Tokur) and large (e.g. Irkutsk, Tyumen, Saint Petersburg) number of land cover classes. The fragments can also be characterized by small (e.g. Komsomolsk-on-

Amur, Naryan-Mar) or large (e.g. Elista, Rubtsovsk) size of land cover patches. There is also a clear distinction between mostly natural (e.g. Lesosibirsk, Ukhta) and urban/agricultural (e.g. Ufa, Ulyanovsk) landscapes.

Before calculating the location property measures CGLC data was reclassified to a smaller number of classes. This preprocessing step was performed to avoid redundancy in the data in cases where land cover is significantly more detailed than topographic in terms of classes. In particular, all closed forest classes (111–116) and open forest classes (121–126) were merged in just two classes: *cforest* (closed forest) and *oforest* (open forest), which corresponds to differentiation applied in topographic databases under the study. Similarly, open sea (20) and permanent water bodies (80) were merged into one *water class*. Additionally, no input (0), bare (60), snow (70) and moss (100) classes were excluded from the case study, since they occupy a negligible area on most of the fragments. Resulting 8-class land cover classification in addition to open forest, closed forest and water, included the initial classes *shrubs* (20), *grass* (30), *crops* (40), *urban* (50), and *wetland* (90).

To estimate terrain ruggedness, we used GMTED_2010 global raster digital elevation model with resolution 7.5 arc seconds (approximately 140 m on 50°N latitude).

Since 100 m land cover does not adequately reproduce linear objects, we additionally used independent data on drainage and road density. Drainage network data was obtained from newest MERIT Hydro-Vector database (Lin et al. 2021) which was derived globally at 3 arc second resolution (approximately 60 m on 50°N latitude). Thus, land cover, terrain and drainage density data sources have ~100 m spatial resolution. Additionally, road density was averaged from GRIP database Meijer et al. (2018) with 5 arc minute resolution (approximately 6 km on 50°N latitude), which was refined using the official Russian road length statistics for municipalities (Russian Federal State Statistics Service 2021).

Experimental work

For modeling purposes 24 fragments (73% of total number) were used as training subset, and 9 fragments (27% of total number) were used as test subset, which is close to the commonly accepted 70/30 ratio. These subsets were formed to have the comparable variety of landscape types in both of them.

For each sample fragment $d_{P'}$ d_L and d_I were calculated as variables to be modeled. For training subset, the mean value of each measure was calculated, and relative densities $\hat{d}_{P'}$ $\hat{d}_L$ and $\hat{d}_I$ were obtained by division of raw values on corresponding means.

Based on our CGLC reclassification, we calculated the ratio of each fragment's area occupied by each land cover class. *contag*, *pafrac* and *joinent* measures were calculated based on CGLC data. *tri* measure was calculated from GMTED2010 extracts. *roads* and *rivers* variables are the densities of road and drainage network correspondingly aggregated from GRIP and MERIT Hydro-Vector datasets for each database sample as a total length divided by fragment area (10 000 km^2).

Since the initial set of calculated predictors consists of 14 variables, we continued with principal component analysis (PCA) of their values aimed at reduction of dimensions to a smaller number. Before applying PCA the variables were log-transformed, centered at zero and scaled to unit variance. Log transformation was performed in a form $In\left(l_j+1\right)$, where l_j is the value of location property. The term +1 is used to exclude possibility of transforming $l_j=0$ into-∞.

After performing the PCA, we fitted the following model to each relative density measure:

$$In\left(\hat{d}_i+1\right)=\sum\nolimits_j \beta_{ij} PC_j+\varepsilon_i$$

Z-scores for all obtained results were calculated and assessed for their effectiveness in differentiating the results. Extraction and counting of database objects for density estimation was performed using MapAnalyzer QGIS plugin written in Python programming language (Yakimova et al. 2021). All remaining analytical calculations, including PCA, building the regression model and performing the statistical tests were programmed in *R* language, including the l*andscapemeterics R* package by Hesselbarth et al. (2019).

RESULTS

Raw density analysis

We begin our analysis of the training subset from visualization of probability density estimates for raw density variables (Fig. 4). Each variable in each scale has lognormal distribution, which is supplemented by all p-values of Shapiro-Wilk test for log-transformed variables being greater than 0.05. This allows us to apply Z-scores in further analysis.

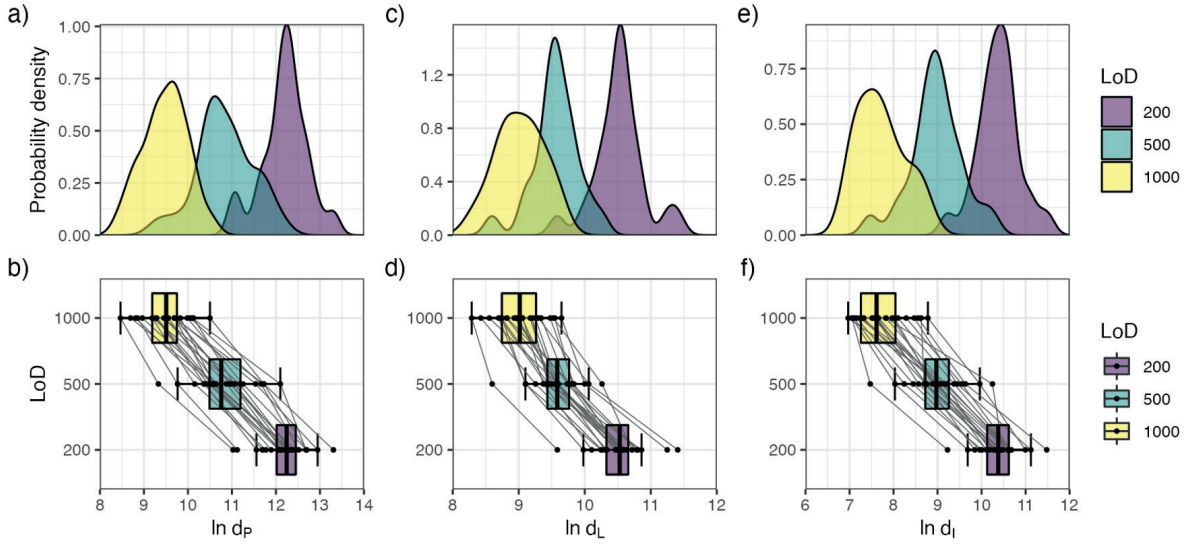

Fig. 4. Distributions of feature density across LoDs: (a-b) point density, (c-d) line density, (e-f) intersection density

As can be seen from lines connecting LoDs of each fragment in Fig. 4, raw densities decrease systematically when the level of detail is decreased correspondingly. This observation is also true for the median values across all fragments. However, distributions overlap significantly. To test whether each of the training fragments can be unmistakably attributed to its level of detail using the density measures, we calculate *Z*-scores and then compare them across all levels. We have the mean and standard deviation of *i*-th density measure at *j*-th level of detail $d_{ij} = d_i(L_j)$. Then for each combination $d_{ikm} = d_i(L_k, F_m)$ of *i*-th density measure calculated at *k*-th level of detail and *m*-th fragment ($3_i x 3_k x 24_m = 216$ combinations in total) *Z*-scores are derived against distribution of each d_{ij}. Expectation is that absolute value of *Z*-score should be minimal when *k=j*, i.e., when *Z*-score of a density measure value is calculated against the mean and standard deviation of its own LoD. If this is true, then raw density measures can be used to assess the suitability of the data fragment for inclusion into the LoD.

Results of *Z*-scores calculation are presented in Fig. 5a, where the rows correspond to d_{km} and the columns correspond to d_{ij}. All values are grouped in *3x3* matrices which correpond to one group of calculations. *Z*-scores for density measures against their own distributions are located on the antidiagonal of each matrix. Therefore, if results meet expectations, these scores should have smallest absolute values in each row of the matrix (colored with blue color).

Fig. 5. Z-scores for density characteristics (training fragments): (a) feature density; (b) normalized feature density. Each row is a unique combination of fragment and level of detail. Each column is a unique combination of feature density measure and level of detail

Fig. 5a indicates that for fragments out of there are cases where expectation is not realized. In particular, for 17% of all d_{ikm} combinations (37 out of 216) the smallest absolute value of Z-score is obtained against another LoD. It means that each 6-th fragment is statistically closer to another LoD in terms of feature density due to it being significantly lower or greater than the mean for its own LoD. For example, Rubtsovsk is located in plain and dry mostly homogeneous sparsely populated landscape, and its database LoDs 200 and 500 are more similar in terms of feature density to the average LoD fragment of 500 and 1000 LoD correspondingly.

Such significant variability in feature density poses a question, whether it is determined by the properties of its geographic location and can be predicted. To answer this question, we apply the method presented in the current paper.

Principle component analysis

PCA results of the derived models are summarized in Fig. 6, Fig. 7, and Fig. 8. According to Fig. 6 the first two PCs explain 53.3% of the total variance, while most of the variance (96%) is explained by the first eight PCs, which were selected for further analysis.

The quality of each PC and variable contributions can be assessed from Fig. 7. In particular, PC1 is formed mainly by *joinent* and *contag* landscape metrics, as well as both types of *forest*, which most probably describes a variability in forest patterns. PC2 is dominated by *pafrac, shrubs, crops* and *roads*. PC3 is formed by *tri, cforest, grass* and *wetland* predictors. PC4 is strongly contributed by *rivers* and *tri*. PC5 is dominated by *tri, grass* and *water*, while PC6 is formed mainly by *pafrac, contag* and *water* variables. PC7 is comprised mainly by *oforest, urban, crops, roads* and *rivers* predictors. Finally, PC8 has significant contributions from *pafrac, contag, tri, shrubs, wetland, crops, roads* and *rivers*. Every variable contributed significantly to at least one PC, while most of them contribute significantly to 2-3 PCs.

It is notable that, despite the expectation, the fraction of urban areas does not contribute as significant as other location properties to any specific PC. However, it is the only location property which contributes equally to the most important PCs 1-2, which results in significant contribution overall analyzed further.

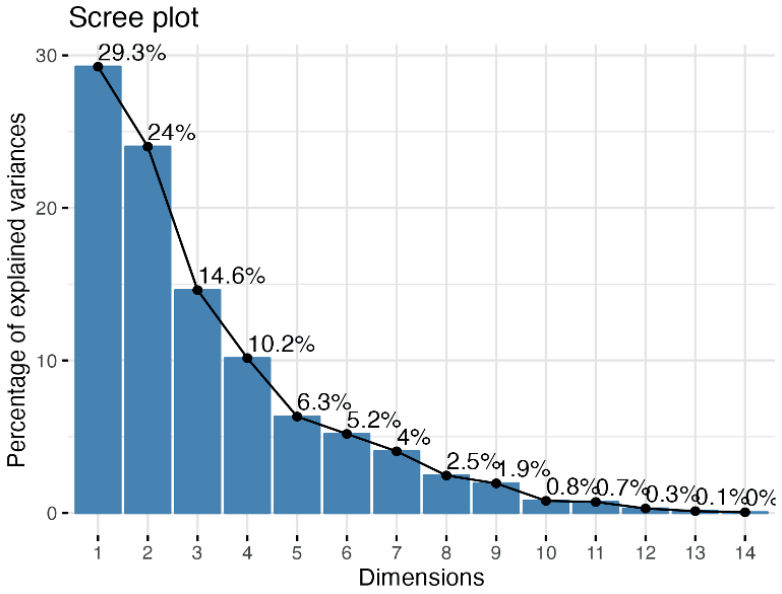

Fig. 6. Fraction of total variance explained by each principal component (scree plot)

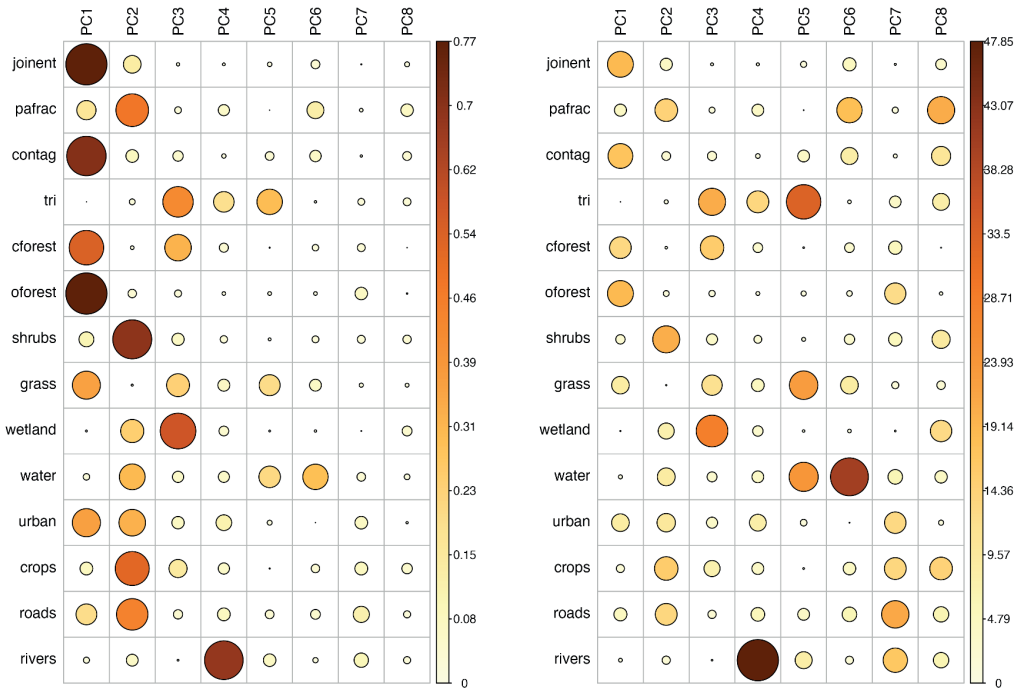

Fig. 7. Principal component analysis summary: (left) quality of representation, *cos*²; (right) variable contributions

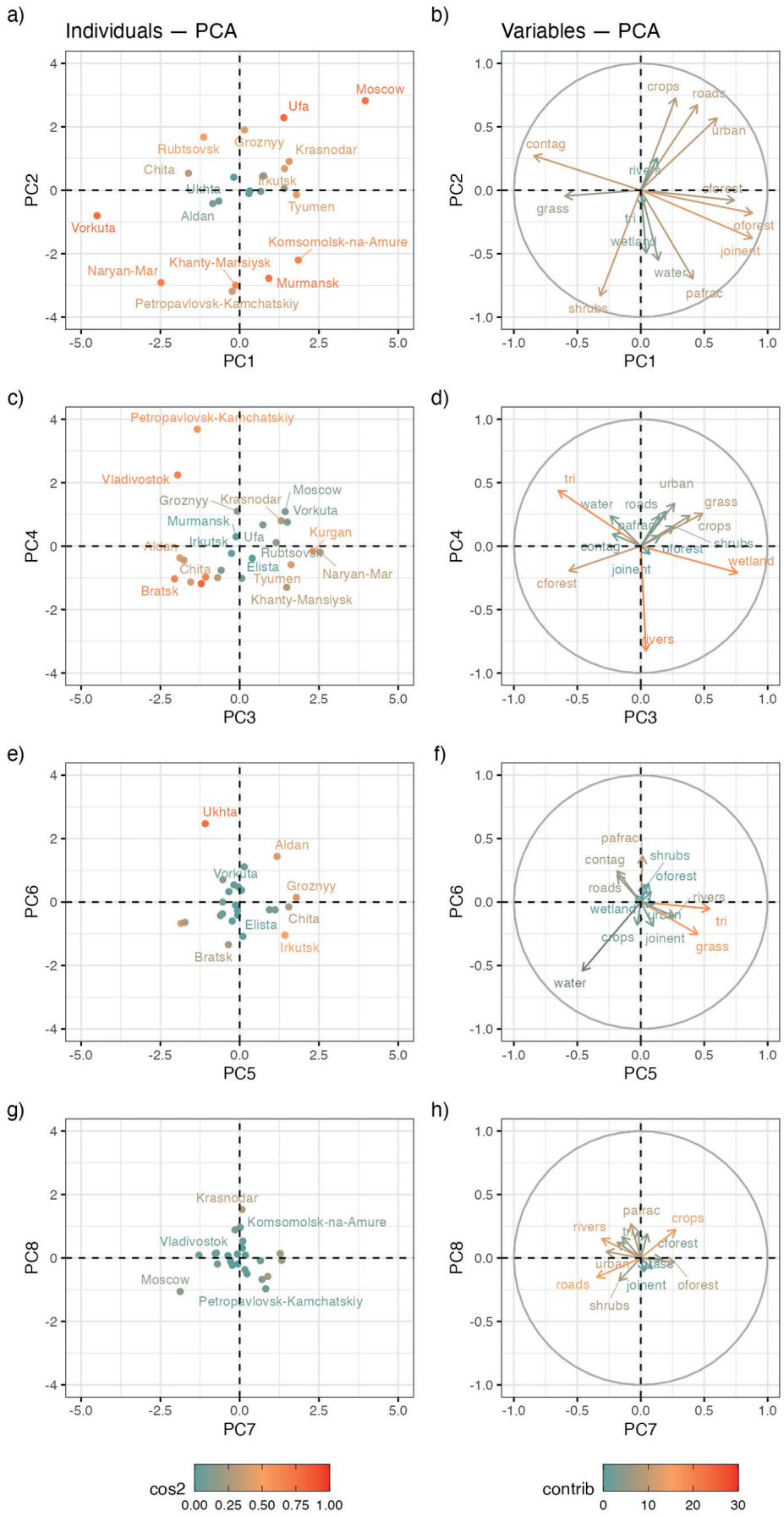

Fig. 8. Principal component analysis details: (a,c,e,g) graph of individuals; (b,d,f,h) graph of variables

Fig. 8 exposes the details of PCA in terms of how the fragments and variable contributions are distributed along first eight PC dimensions. From Fig. 8a we see clearly that the PCs 1–2 differentiate urban/agricultural areas (upper right quarter) and mostly natural ones with lots of forest (lower left quarter). Fig. 8b reveals that fragments are arranged from mountains (left part) to planes (right part) and from less (upper part) to more (lower part) developed river network by PCs 3–4. On Fig. 8c we see mainly differentiation on the ratio of large water bodies by PCs

5-6 with near-water fragments located in lower left quarter of the plot. Finally, the contribution of PCs 7-8 on Fig. 8d on fragment differentiation is less distinct. The graphs of variables represented in Fig. 8b,d,f,h supplement our conclusions with contribution vectors of each PC.

Normalized feature density

The summary of the model for each relative density measure is presented in Table 1. As we see, PCs 1-5 and 7 contribute significantly to regression on $\widehat{d}_P$ and $\widehat{d}_{L'}$ while only PCs 1, 4 and 7 exhibit significant influence on $\widehat{d}_I$. PCs 6 and 8 have neglegible effect on the result in all regressions. These relations between principal components and modeled relative density measures can be assessed visually from Fig. 9 and Fig. 10. In particular, trend lines for PCs 6 and 8 have less explicit slope, which indicates a weak correlation with relative features density.

The derived regression model allows calculation of normalized densities $\tilde{d}_i = d_i / \widehat{d}_{i'}$ which account for landscape heterogeneity and reduce densities to their expected values for corresponding LoD. Distributions of normalized densities can be assessed from Fig. 11. Comparing to raw densities in Fig. 4 it can be seen that normalized distributions have higher kurtosis and intersect less than for raw densities, which means that they their populations can potentially be separated more easily based on Z-scores.

To check this hypothesis, we calculated Z-scores of normalized values against the mean and standard deviation of raw values. Results are presented in Fig. 5b. Now only for 4.6% of all d_{ikm} (10 out of 216) the smallest absolute value

Table 1. Summary of relative feature density regression models

PC***	Points ($\ln\widehat{d}_P$)			Lines ($\ln\widehat{d}_L$)			Intersections ($\ln\widehat{d}_I$)		
	Beta	95% CI**	*p-value	Beta	95% CI	p-value	Beta	95% CI	p-value
Intercept	0.66	0.63, 0.70	•••	0.67	0.64, 0.69	•••	0.66	0.62, 0.71	•••
PC1	0.02	0.00, 0.04	•	0.02	0.01, 0.03	••	0.04	0.02, 0.06	•••
PC2	–0.08	–0.10, –0.06	•••	–0.04	–0.05, –0.02	•••	–0.02	–0.05, 0.00	°
PC3	–0.04	–0.06, –0.01	••	–0.03	–0.05, –0.02	•••	0.01	–0.02, 0.05	°
PC4	0.08	0.05, 0.11	•••	0.04	0.02, 0.06	•••	0.09	0.06, 0.13	•••
PC5	0.05	0.02, 0.09	••	0.07	0.05, 0.10	•••	0.03	–0.02, 0.08	°
PC6	0.01	–0.04, 0.05	°	0.01	–0.02, 0.04	°	0.04	–0.02, 0.09	°
PC7	–0.13	–0.18, –0.08	•••	–0.07	–0.10, –0.03	•••	–0.13	–0.19, –0.06	•••
PC8	0.00	–0.06, 0.06	°	–0.02	–0.07, 0.02	°	–0.01	–0.09, 0.07	°

* p-values: ••• < 0.001, •• < 0.01, • < 0.05, ° ≥ 0.05
** CI – confidence interval, *** PC – principal component

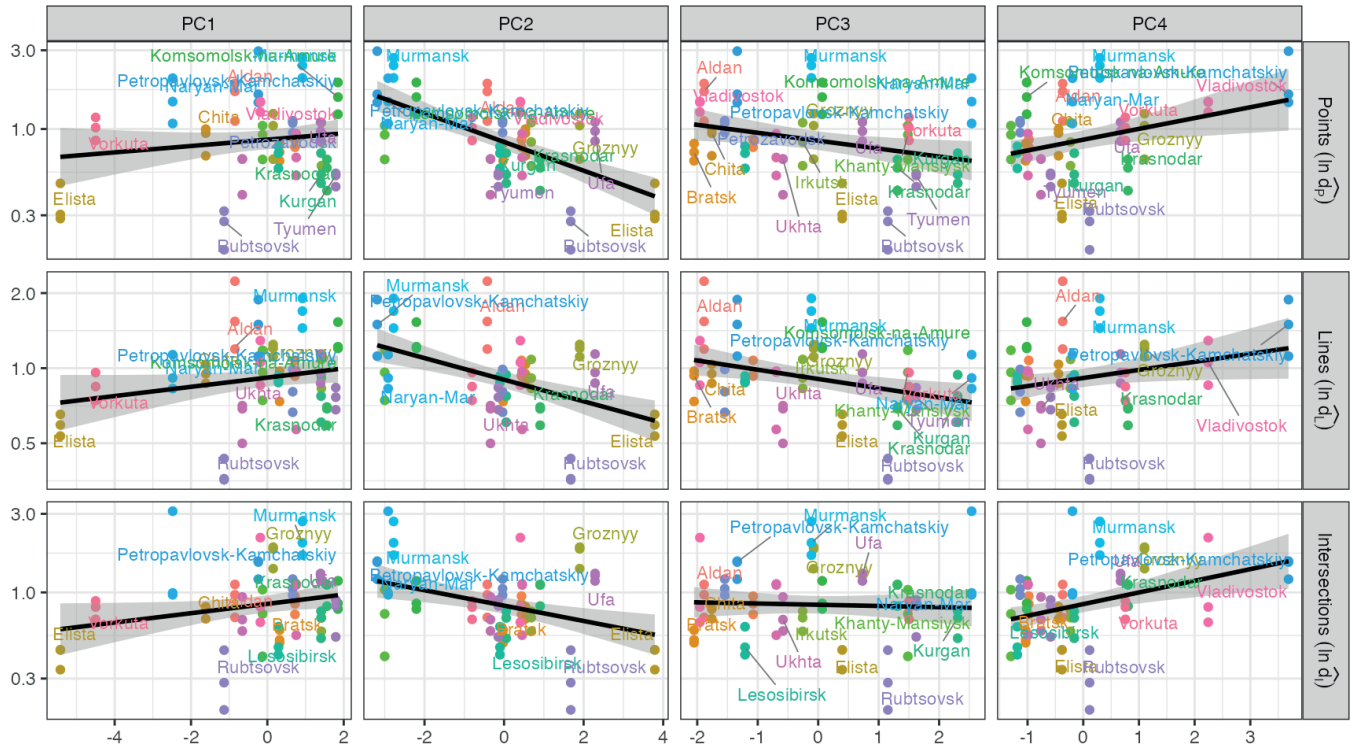

Fig. 9. Relationships of relative feature density with principal components 1-4

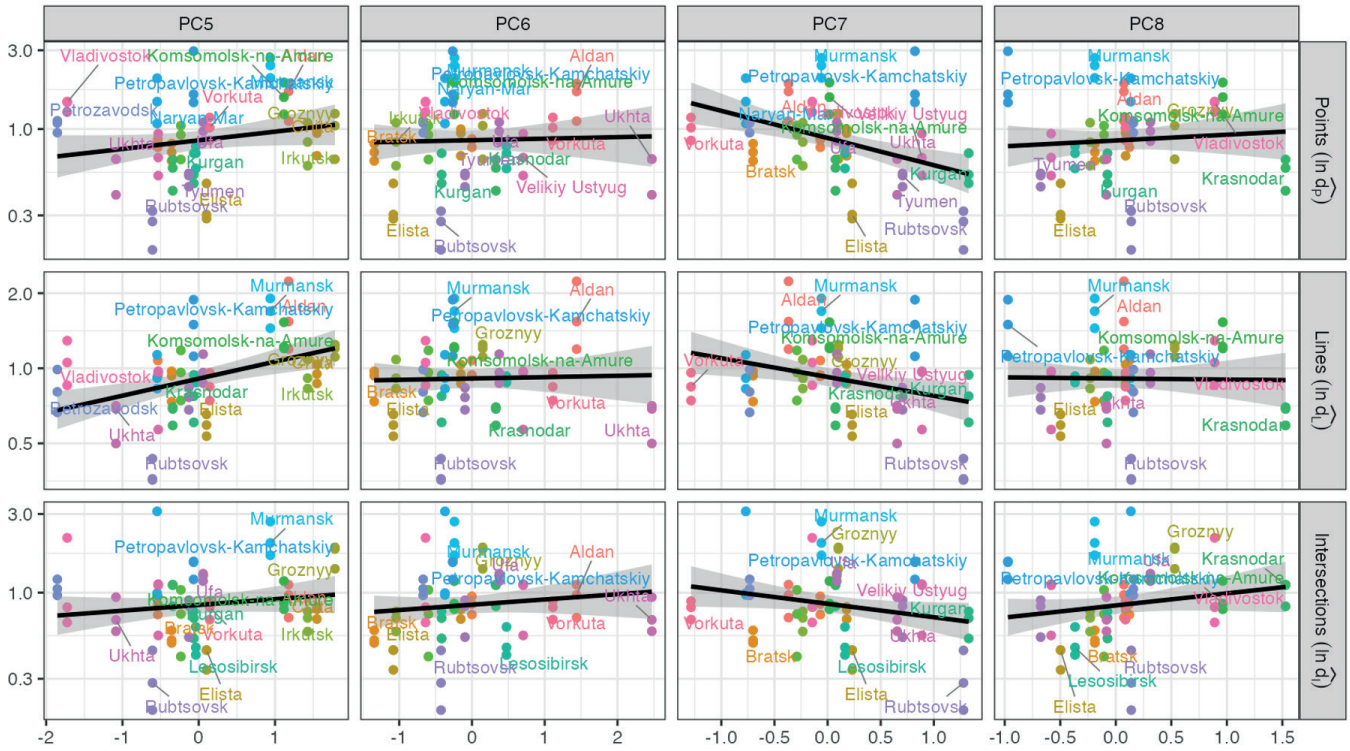

Fig. 10. Relationships of relative feature density with principal components 5-8

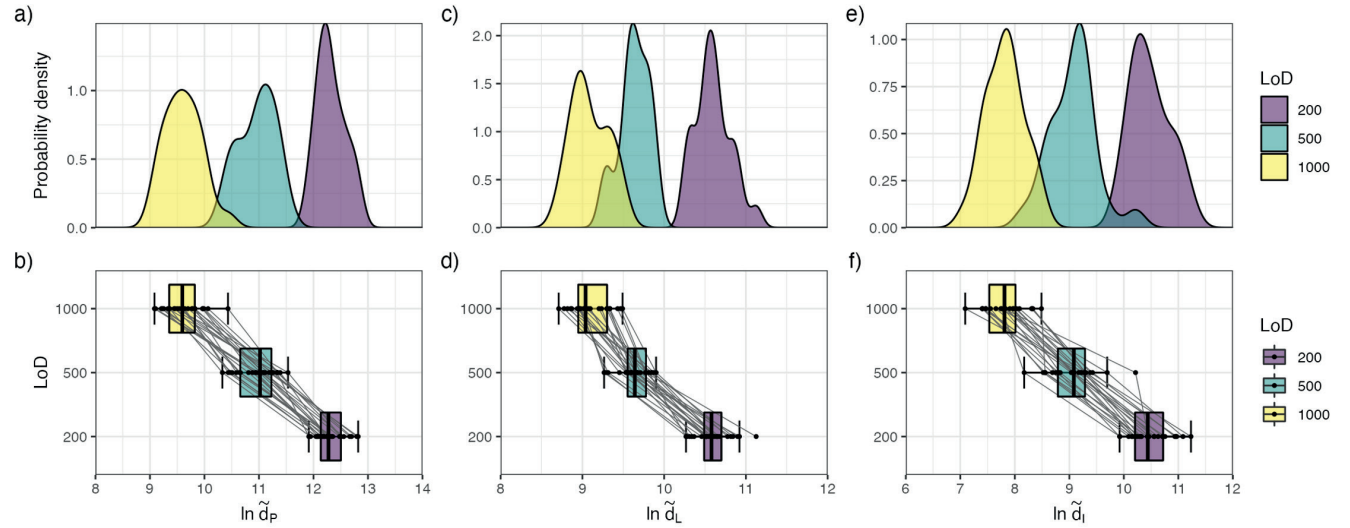

Fig. 11. Distributions of normalized feature density across scales: (a-b) point density; (c-d) line density; (e-f) intersection density

of Z-score is obtained against another LoD. This result means that the error of identifying the most appropriate LoD is significantly reduced after processing with our method, and approximately 95% of training dataset can be recognized correctly.

The robustness of the developed approach was assessed by using the derived PCA transform and regression model to predict relative density for 9 testing fragments ($3_i x 3_k x 9_m = 81$ combinations in total). As can be seen from Fig. 12, our approach improves the error of LoD recognition from 12.3% (10 out of 81) to 2.5% (2 out of 81), which is a significant gain in precision. As in the case of the training dataset, more than 95% of the total samples were related to the true LoD after density normalization.

DISCUSSION

The developed model has several potential applications. First, if a topographic data misses metadata about its level of detail or the most appropriate scale of visualization, the model can be used to predict the most proper level of detail of the data. Second, it can be used as a preliminary means of judging if two topographic databases covering different areas have similar level of detail and can be essentially merged into one database. Third, it is useful to evaluate the results of collaborative generalization during which multiple areas are generalized using different algorithms (Touya, Duchêne, and Ruas 2010). A ratio between feature densities which naturally reflects the difference in landscape structures, is an indicator of good generalization. It must be stressed, however, that all these potential applications need practical assessment of their effectiveness.

There are several limitations of the presented results which need to be overcome in future improvements. The current study is based on the experimental sample which now consists of 33 database fragments. The modest sample size and the fact that fragments do not cover all possible varieties of landscapes may potentially limit the effectiveness of the derived PCs and regression model. In

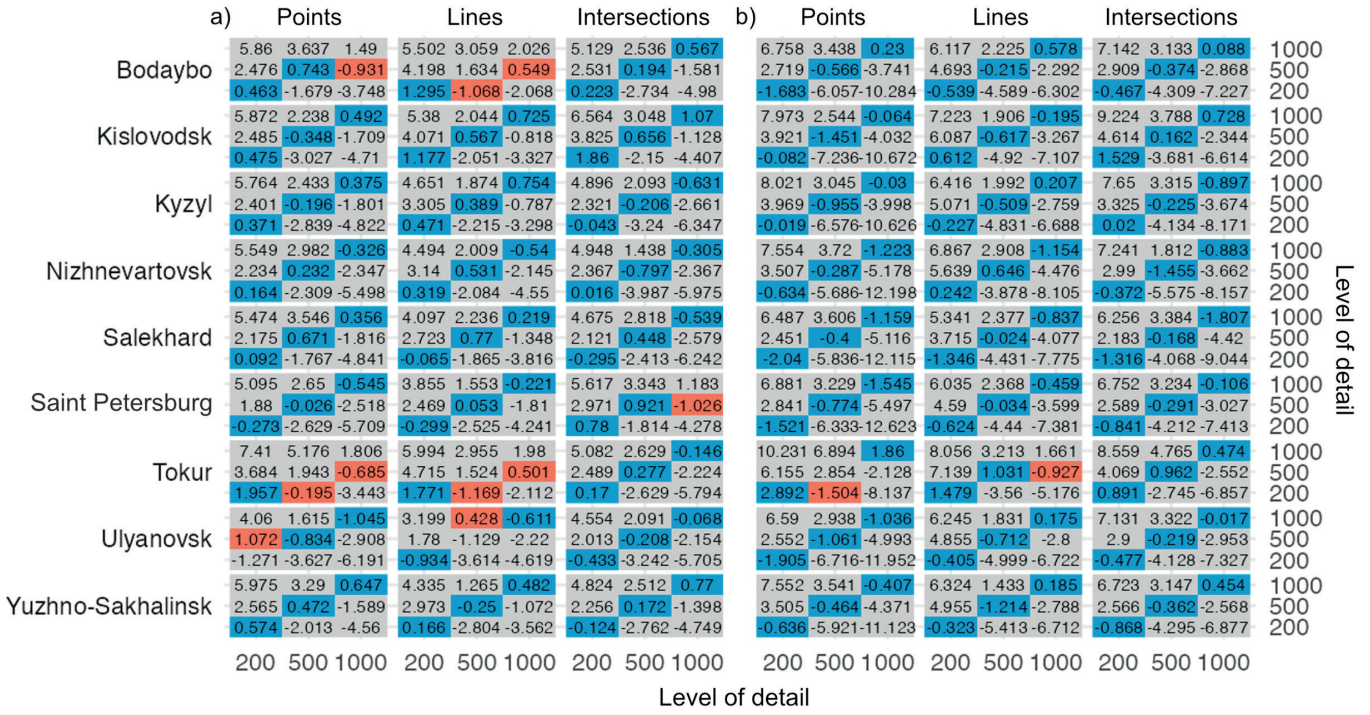

Fig. 12. Z-scores for density characteristics (testing fragments): (a) feature density; (b) normalized feature density. Each row is a unique combination of fragment and level of detail. Each column is a unique combination of feature density measure and level of detail

particular, expected (mean) value of each density measure will change if the sample is extended to include additional fragments. A reasonable way to improve the current result is to subdivide the total area covered by a spatial database product by the grid of similar fragments, and perform model construction and testing based on the total amount of fragments. Since the area of Russia is approximately $1.7 \times 10^7 km^2$, and the area of each fragment is $10^4 km^2$, this will result in approximately 1700 fragments. This amount can be increased multiple times by extracting smaller portions of data (i.e. 50x50 km). Such a large population can be randomly sampled to construct more reliable model for relative feature density. However, this experiment will be substantially more demanding in terms of computing time.

Additional shortcoming is that current study is tied to specific spatial data product and to only three levels of detail. Since each product and LoD is derived using a specific data selection and generalization rules, the effectiveness of the model should be assessed before its application. The same issue is true for specific data sources used for prediction. In particular, these data sources are not suitable for density assessment of detailed (large scale) spatial data due to their resolution. Also, each of the data sources has its own errors and limited precision, which inevitably affects the quality of results. However, assessment of error propagation is out of the scope of the current paper.

A more convoluted problem is hidden in the assumption that the relative feature density is similar for each fragment independent from LoD. However, as we've seen in the paper, the database fragment which is 2 times denser than the average in one LoD, can be less different from the average in another LoD. We cannot integrate such information directly in our model, since LoD is unknown in our case. In addition, it should be stressed that only density-based characteristics of LoD are considered in the current study. Surely, other LoD descriptors such as feature granularity must be investigated in a similar way.

Finally, more sophisticated methods of model building for LoD recognition such as deep learning (DL) can be used instead of a simple linear regression with PCA. This may potentially improve the reliability and robustness of the derived model. Application of DL is impractical on current sample, but can be feasible if thousands of data extracts comprise the sample.

CONCLUSIONS

Digital topographic maps are commonly represented as multiscale spatial databases with several levels of detail (LoDs) corresponding to map scales. A sequence of LoDs covering the same area is characterized by monotonous change of the detail properties such as granularity and density of features (spatial objects): the smaller is the desired map scale, the less detailed is the LoD. However, the properties of one LoD may differ significantly between locations. For example, elevation contours in a rugged terrain are naturally more complex than over the gently sloped areas. The spatially varying shape of individual objects and terrain surface as well as their arrangement in specific patterns results in heterogeneity of each LoD. Hence, the measures which describe the level of detail may vary significantly. This complicates the integration of topographic data fragments covering different areas into one database. There is a need for an automated method that will help to assess feature density of the fragments and their correspondence to the desired level of detail.

In this paper we approached this problem by statistical learning from multiscale topographic database constructed through manual generalization. For each of the three LoDs we analyze how much the density of points, lines and intersections in each fragment is higher or lower than the average LoD density. This ratio, called relative density, is then modeled as a function of location properties expressed in a number of spatial and non-spatial measures. Principal components derived from measures showed significant contributions both from landscape complexity measures, and from ratios of different land cover classes. Urban and natural, mountainous and flat, with less and more developed river network areas were effectively differentiated. Results showed that normalization of raw feature density on its predicted relative value provides the

measure close to the mean value of the desired LoD. Using normalized values, the error of LoD recognition is reduced several times. This proves the hypothesis that location properties can be used as effective predictors of relative feature density in topographic data.

Results obtained in this paper can be used in spatial data integration and generalization workflows which involve spatial partitions – either in collaborative generalization or merging the data obtained independently for different areas. A more representative and large data sample and different machine learning methods can be used in future investigations to improve the predictive power of the proposed approach.

REFERENCES

Biljecki F., Ledoux H., and Stoter J. (2016). An Improved LOD Specification for 3d Building Models. Computers, Environment and Urban Systems, 59, 25–37, DOI: 10.1016/j.compenvurbsys.2016.04.005.

Buchhorn M., Lesiv M., Tsendbazar N.-E., Herold M., Bertels L., and Smets B. (2020). Copernicus Global Land Cover Layers Collection 2. Remote Sensing, 12 (6), 1044, DOI: 10.3390/rs12061044.

Buchin K., Meulemans W., Van Renssen A., and Speckmann B. (2016). Area-Preserving Simplification and Schematization of Polygonal Subdivisions. ACM Transactions on Spatial Algorithms and Systems, 2(1), 1-36, DOI: 10.1145/2818373.

Burrough P.A. (1981). Fractal Dimensions of Landscapes and Other Environmental Data. Nature, 294 (5838), 240-42, DOI: 10.1038/294240a0.

Cheng X., Liu Z., and Zhang Q. (2021). MSLF: Multi-Scale Legibility Function to Estimate the Legible Scale of Individual Line Features. Cartography and Geographic Information Science, 48(2), 151-68, DOI: 10.1080/15230406.2020.1857307.

Cheng X., Wu H., Ai T., and Yang M. (2017). Detail Resolution: A New Model to Describe Level of Detail Information of Vector Line Data. In Advances in Geographic Information Science, edited by Chenghu Zhou, Fenzhen Su, Francis Harvey, and Jun Xu, Singapore: Springer Singapore, 167-77, DOI: 10.1007/978-981-10-4424-3_12.

De Floriani L., Marzano M., and Puppo E. (1996). Multiresolution Models for Topographic Surface Description. The Visual Computer, 12 (7), 317-45, DOI: 10.1007/s003710050068.

Douglas D.H. and Peucker T.K. (1973). Algorithms for the Reduction of the Number of Points Required to Represent a Digitized Line or Its Caricature. The Canadian Cartographer, 10(2), 112-22, DOI: 10.3138/FM57-6770-U75U-7727.

Haunert J.-H. and Wolff A. (2010). Area Aggregation in Map Generalisation by Mixed-Integer Programming. International Journal of Geographical Information Science, 24(12), 1871-97, DOI: 10.1080/13658810903401008.

Hesselbarth M. H. K., Sciaini M., With K.A., Wiegand K, and Nowosad J. (2019). Landscapemetrics : An Open-Source R Tool to Calculate Landscape Metrics. Ecography, 42 (10), 1648-57, DOI: 10.1111/ecog.04617.

Imhof E. (1982). Cartographic Relief Presentation, Berlin: Walter der Gruyter, 416.

Jones C.B. and Abraham I.M. (1986). Design Considerations for a Scale-Independent Database. In Proceedings of Second International Symposium on Spatial Data Handling, Seattle, 384-98.

Kilpeläinen T. (2000). Maintenance of Multiple Representation Databases for Topographic Data. The Cartographic Journal, 37(2), 101-7, DOI: 10.1179/0008704.37.2.p101.

Kolbe T.H., Gröger G., and Plümer L. (2005). CityGML: Interoperable Access to 3d City Models. In: Geo-Information for Disaster Management, edited by Peter van Oosterom, Siyka Zlatanova, and Elfriede M. Fendel, Berlin, Heidelberg: Springer Berlin Heidelberg, 883-99, DOI: 10.1007/3-540-27468-5_63.

Lemmens M. (2011). Quality of Geo-Information. In: Geo-Information, Dordrecht: Springer Netherlands, 211-27, DOI: 10.1007/978-94-007-1667-4_11.

Li Z. and Openshaw S. (1992). Algorithms for Automated Line Generalization Based on a Natural Principle of Objective Generalization. International Journal of Geographical Information Systems, 6(5), 373-89, DOI: 10.1080/02693799208901921.

Lin P., Pan M., Wood E. F., Yamazaki D., and Allen G. H. (2021). A New Vector-Based Global River Network Dataset Accounting for Variable Drainage Density. Scientific Data, 8(1), 28, DOI: 10.1038/s41597-021-00819-9.

Meijer J.R., Huijbregts M.A.J., Schotten K.C.G.J., and Schipper A.M. (2018). Global Patterns of Current and Future Road Infrastructure. Environmental Research Letters, 13(6), 064006, DOI: 10.1088/1748-9326/aabd42.

Meng L. and Forberg A. (2007). 3D Building Generalisation. In: Generalisation of Geographic Information: Cartographic Modelling and Applications, Elsevier, 211-31, DOI: 10.1016/B978-008045374-3/50013-2.

Military Topographic Service. (1978). Guide to cartographic and cartographic works. Part 1. Drafting and preparation for publication of 1:25 000, 1:50 000, and 1:100 000 scale topographic maps [in Russian]. Moscow: Editorial; Publishing Department of Military Topographic Service.

Military Topographic Service. (1980). Guide to cartographic and cartographic works. Part 2. Drafting and preparation for publication of 1:200 000 and 1:500 000 topographic maps [in Russian]. Moscow: Editorial; Publishing Department of Military Topographic Service.

Military Topographic Service. (1985). Guide to cartographic and cartographic works. Part 3. Drafting and preparation for publication of 1:1 000 000 scale topographic maps [in Russian]. Moscow: Editorial; Publishing Department of Military Topographic Service.

Nowosad J. and Stepinski T. F. (2019). Information Theory as a Consistent Framework for Quantification and Classification of Landscape Patterns. Landscape Ecology, 34(9), 2091-101, DOI: 10.1007/s10980-019-00830-x.

Phillips J. (1999) Earth surface systems : complexity, order and scale – Malden, Mass: Blackwell Publishers.

Riitters K.H., O'Neill R.V., Wickham J.D., and Jones K.B. (1996). A Note on Contagion Indices for Landscape Analysis. Landscape Ecology, 11(4), 197-202, DOI: 10.1007/BF02071810.

Riley S.J., De Gloria S.D., and Elliot R. (1999). A Terrain Ruggedness That Quantifies Topographic Heterogeneity. Intermountain Journal of Science, 5 (1-4), 23-27.

Ruas A., and Bianchin A. (2002). Echelle Et Niveau de Detail. In: Generalisation et Representation Miltiple, edited by Anne Ruas, Paris: Hermes Lavoisier, 25-44.

Russian Federal State Statistics Service. (2022). Database of Municipal Indicators. Available at: https://www.gks.ru/free_doc/new_site/bd_munst/munst.htm. [Accessed 10 June 2022].

Samsonov T. (2022). Granularity of Digital Elevation Model and Optimal Level of Detail in Small-Scale Cartographic Relief Presentation. Remote Sensing, 14(5), 1270, DOI: 10.3390/rs14051270.

Samsonov T. and Yakimova O. (2020). Regression Modeling of Reduction in Spatial Accuracy and Detail for Multiple Geometric Line Simplification Procedures. International Journal of Cartography, 6(1), 47-70, DOI: 10.1080/23729333.2019.1615745.

Töpfer F. and Pillewizer W. (1966). The Principles of Selection. The Cartographic Journal, 3(1), 10-16, DOI: 10.1179/caj.1966.3.1.10.

Touya G. and Brando-Escobar C. (2013). Detecting Level-of-Detail Inconsistencies in Volunteered Geographic Information Data Sets. Cartographica: The International Journal for Geographic Information and Geovisualization, 48(2), 134-43, DOI: 10.3138/carto.48.2.1836.

Touya G., Duchêne C., and Ruas A.. (2010). Collaborative Generalisation: Formalisation of Generalisation Knowledge to Orchestrate Different Cartographic Generalisation Processes. In: Theories and Methods of Spatio-Temporal Reasoning in Geographic Space, Berlin, Heidelberg: Springer Berlin Heidelberg, 264-78, DOI: 10.1007/978-3-642-15300-6_19.

Visvalingam M. and Whyatt J.D. (1993). Line Generalisation by Repeated Elimination of Points. The Cartographic Journal, 30(1), 46-51, DOI: 10.1179/caj.1993.30.1.46.

Yakimova O., Samsonov T., Potemkin D., Usmanova E. (2021). QGIS processing tool for spatial data detail assessment. InterCarto. InterGIS, 27(2), 260-279, DOI: 10.35595/2414-9179-2021-2-27-268-279.

Appendix

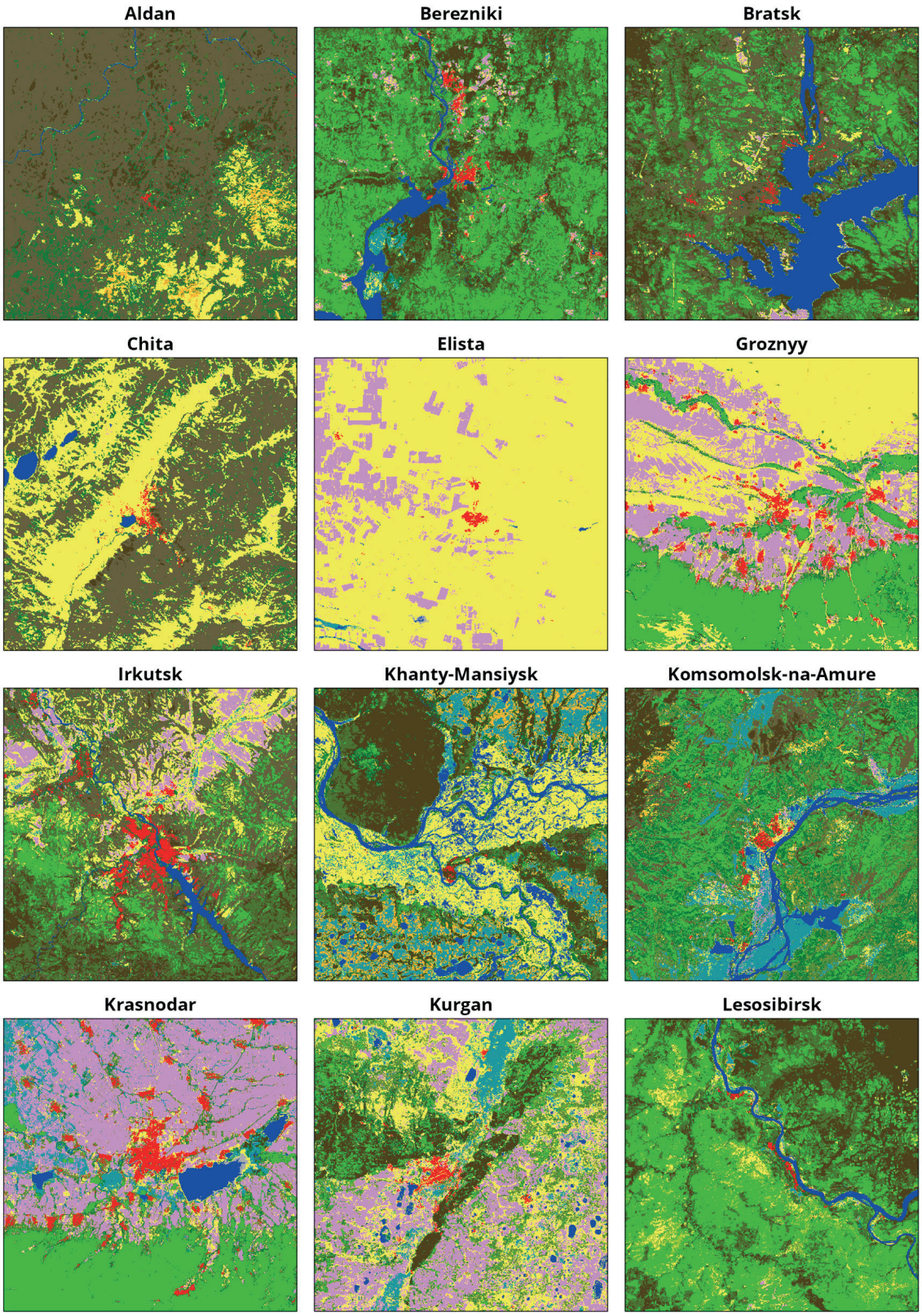

Fig. A.1. Land cover extracts for training fragments 1-12

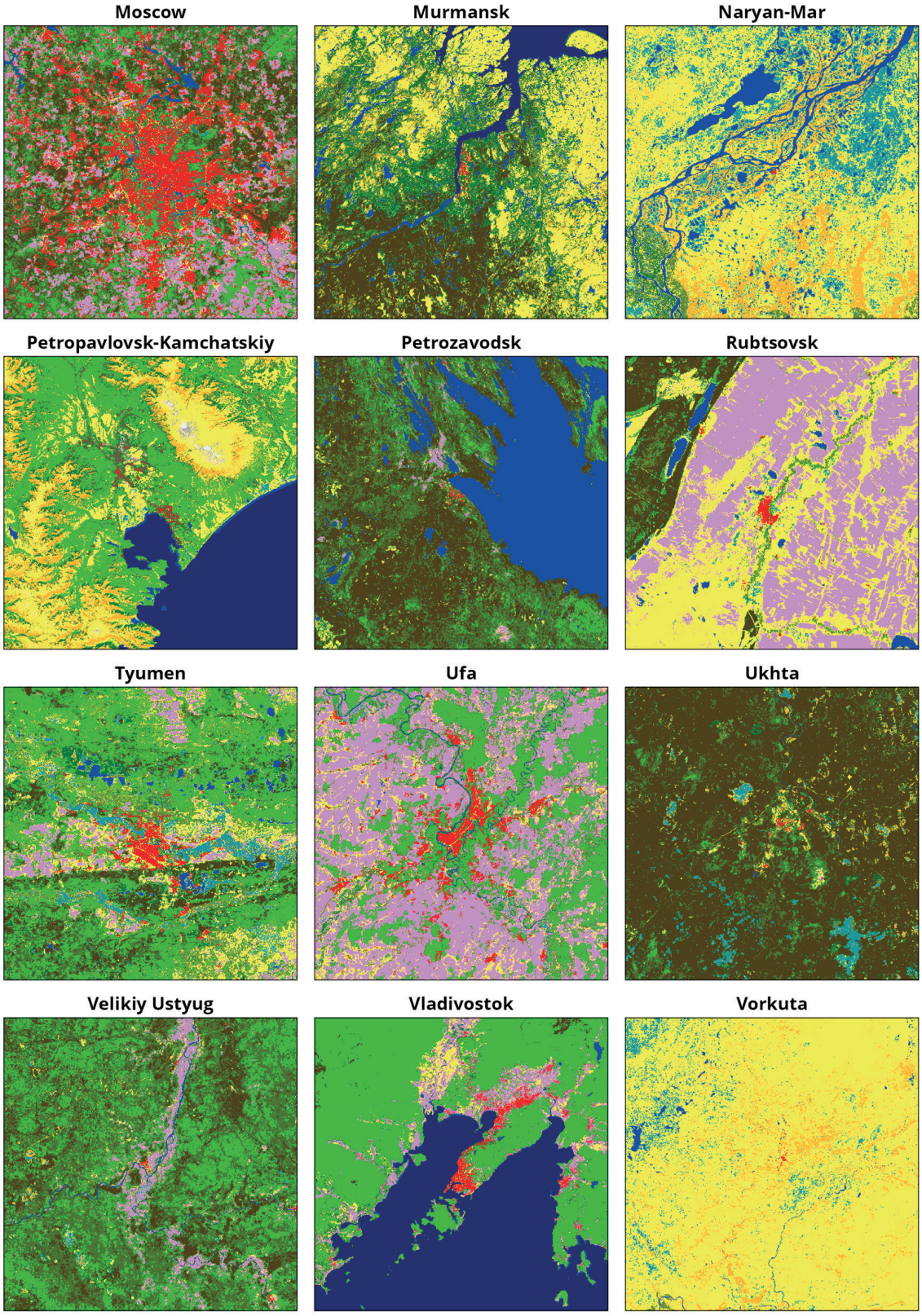

Fig. A.2. Land cover extracts for training fragments 13-24

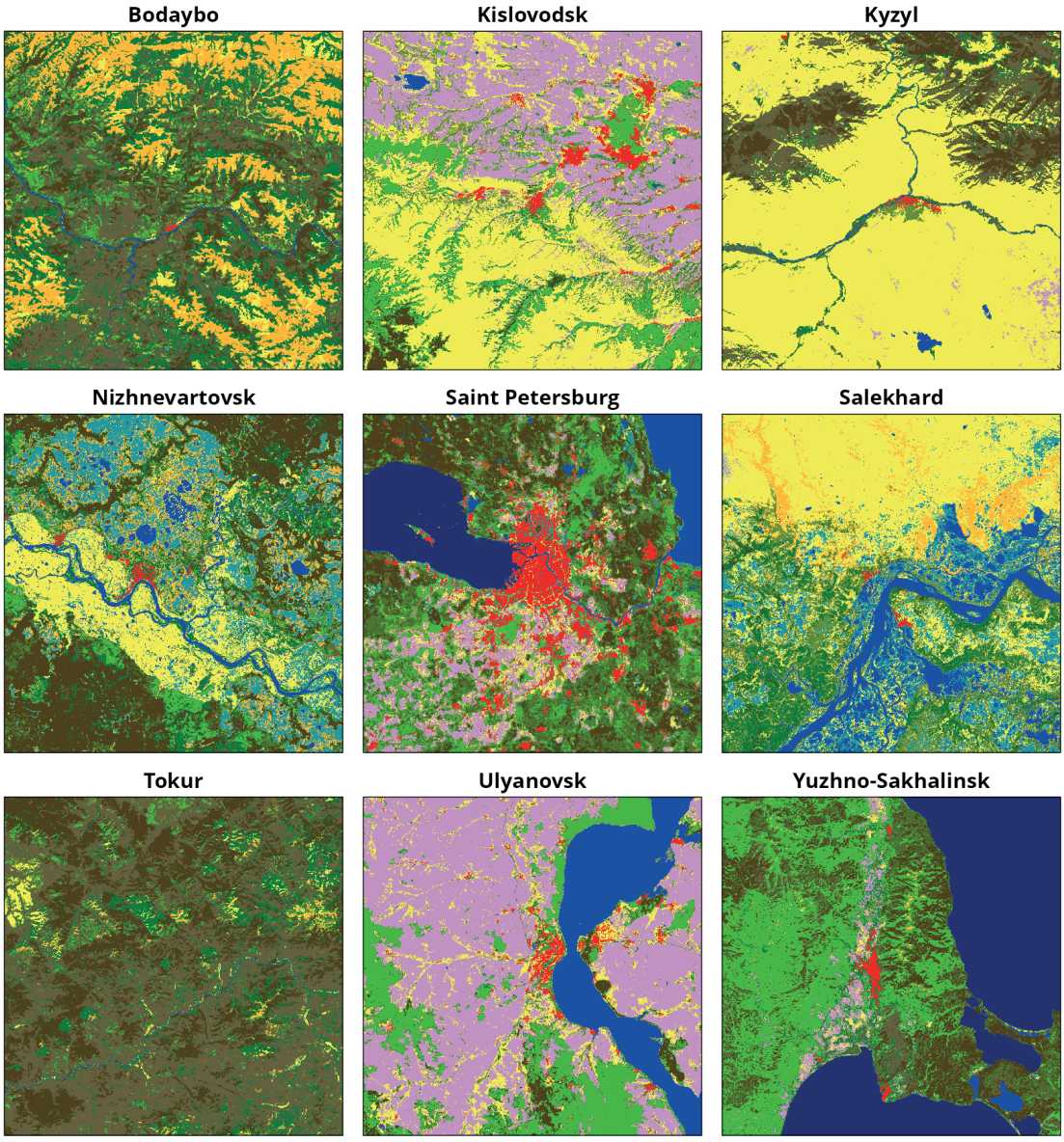

Fig. A.3. Land cover extracts for testing fragments

Table A.1. Copernicus Global Land Cover classes and colors

Type	Color	Description
0		No input data available
20		Shrubs
30		Herbaceous vegetation
40		Cultivated and managed vegetation/agriculture (cropland)
50		Urban / built up
60		Bare / sparse vegetation
70		Snow and Ice
80		Permanent water bodies
90		Herbaceous wetland
100		Moss and lichen
111		Closed forest, evergreen needle leaf
112		Closed forest, evergreen, broad leaf
113		Closed forest, deciduous needle leaf
114		Closed forest, deciduous broad leaf
115		Closed forest, mixed
116		Closed forest, unknown
121		Open forest, evergreen needle leaf
122		Open forest, evergreen broad leaf
123		Open forest, deciduous needle leaf
124		Open forest, deciduous broad leaf
125		Open forest, mixed
126		Open forest, unknown
200		Open sea