METHODS/METODE

ASSESSING THE IMPACT OF INPUT DATA INCONGRUITY IN SELECTED QUANTITATIVE METHODS FOR MODELLING NATURAL LANDSCAPE TYPOLOGIES

PREVERJANJE VPLIVA NESMISELNOSTI VHODNIH PODATKOV PRI IZBRANIH KVANTITATIVNIH METODAH ZA MODELIRANJE NARAVNOPOKRAJINSKIH TIPIZACIJ

AUTHOR/AVTOR
dr. Rok Ciglič
Research Centre of the Slovenian Academy of Sciences and Arts, Anton Melik Geographical Institute, Gosposka ulica 13, SI – 1000 Ljubljana, Slovenia
rok.ciglic@zrc-sazu.si

DOI: 10.3986/GV90107
UDC/UDK: 911.5:528:004(497.4)
COBISS: 1.01

ABSTRACT
Assessing the impact of input data incongruity in selected quantitative methods for modelling natural landscape typologies
With supervised classification methods, we can determine classification rules for landscape types of existing landscape typologies. In this article, we analyse whether supervised classification methods could also define adequate rules for landscape types determination in the case of poorly designed typologies. We tried to model two Slovenian intentionally distorted natural landscape typologies. We noted that due to the incongruity of the distorted typologies, decision tree methods were not capable of forming rules for determination of landscape types. Although we did manage to create modelled distorted typologies with minimum distance to means method, maximum likelihood method, and k-nearest neighbours method, they matched the basic distorted typology only slightly.

KEY WORDS
geography, geographic information systems, models, landscape classification, Slovenia

IZVLEČEK

Preverjanje vpliva nesmiselnosti vhodnih podatkov pri izbranih kvantitativnih metodah za modeliranje naravnopokrajinskih tipizacij

Z metodami nadzorovane klasifikacije lahko za obstoječe naravnopokrajinske tipizacije določimo klasifikacijska pravila za posamezne pokrajinske tipe. V prispevku razpravljamo, ali bi tudi v primeru zelo slabo zasnovanih tipizacij z metodami nadzorovane klasifikacije lahko izdelali dovolj natančna pravila za določanje pokrajinskih tipov. Poskusili smo modelirati dve namenoma popačeni naravnopokrajinski tipizaciji Slovenije. Opazili smo, da zaradi nesmiselnosti popačenih tipizacij metode odločitvenih dreves sploh niso bile sposobne izdelati pravil za določanje pokrajinskih tipov. Z metodami najmanjše razdalje, največje verjetnosti in k *najbližjih sosedov pa smo sicer uspeli izdelati modelirane popačene tipizacije, a so se te z osnovno popačeno tipizacijo le malo ujemale.*

KLJUČNE BESEDE

geografija, geografski informacijski sistemi, modeli, pokrajinska klasifikacija, Slovenija

The article was submitted for publication on February 1, 2018.
Uredništvo je prispevek prejelo 1. februarja 2018.

1 Introduction

1.1 Theoretical background of the research study

The determination of natural landscape types has a long tradition in Slovenia (Melik 1946; Perko 1998; Špes et al. 2002; Perko, Hrvatin and Ciglič 2015; Perko and Zorn 2016; see also Perko, Hrvatin, and Ciglič 2017), as well as globally (i.e., Olson et al. 2001; Mücher et al. 2010). During the analysis of the landscapes, the authors pointed out the numerous challenges present in the process of landscape classification (Gams 1986; Mücher et al. 2003; Bailey 2004). One of the challenges is also the absence of a general agreement on the perception of the term *landscape*. Udo de Haes and Klijn (1994) pointed out that an ecosystem may be defined as an abstract notion or as an actually recognisable object. In order for the units to actually be identified, Bailey (1996, 4) states that ecosystems as geographic landscape units include all natural features and may thus be identified and delimited with boundaries. On the contrary, Gams (1978, 15) states that »*any region delimited by a line on the map is an unnatural and artificial formation that serves only as a means for determining differences.*«

Confirmation of the existence of natural landscape types is significant, since different computer algorithms, which enable the modelling of previously formed typologies, are available. We can observe different cases of assessment of landscape classifications (Breskvar Žaucer and Marušič 2006; Strand 2011). The analyses (Ciglič 2012; 2014; Kokalj and Oštir 2013; Ciglič and Perko 2015) of two Slovenian natural landscape typologies (Perko 1998; Špes et al. 2002) have shown that both typologies, although they were formulated with a manual determination of boundaries, are of sufficient quality and can be modelled with quantitative methods.

Here, the question arises whether in the case of poorly designed or random typologies supervised classification methods, which are often used in landscape analysis (e.g. decision trees, minimum distance to means method, maximum likelihood method, and *k*-nearest neighbours method), could even result in the formation of models. We came across a study on the effectiveness of methods in geographical analyses performed on simulated data (Belbin and McDonald 1993). Each method has its own specific advantages and disadvantages, and each yields different results.

The article aims to determine the effectiveness of methods for modelling typologies, even if the typologies in question were created without taking into account their natural landscape background, i.e., completely at random. With an experiment on the case of Slovenia, we will check how certain methods perform in the case of random, in some instances also incongruous natural landscape typologies. In this manner, we will be able to assess the impact of methods on the result of the modelling. For the purpose of this experiment, two original natural landscape typologies were distorted – the first according to Perko (1998) and the second according to Špes et al. (2002).

1.2 Terminology

In the article, longer terms are used to differentiate among categories of typologies:

- **original (natural landscape) typology** – this term refers to the two original typologies; these are the typologies of Slovenia according to Perko (1998) and Špes et al. (2002),
- **(basic) distorted (natural landscape) typology** – this term refers to intentionally distorted original natural landscape typologies that were used to model the incongruous typologies of Slovenia,
- **modelled distorted (natural landscape) typology** – this term designates any typology that was created by the supervised classification method based on the basic distorted natural landscape typology.

2 Methodology

2.1 Original natural landscape typologies

The article is based on the analysis of natural landscape typologies by Perko (1998) and Špes et al. (2002).

Perko determined 9 types:

- Alpine mountains (type code is 1.1),
- Alpine hills (1.2),
- Alpine plains (1.3),
- Pannonian low hills (2.1),
- Pannonian plains (2.2),
- Dinaric plateaus (3.1),
- Dinaric lowlands (3.2),
- Mediterranean low hills (4.1),
- Mediterranean plateaus (4.2).

Špes et al. (2002) determined 13 types:

- mountains (type code is 1),
- wide river valleys in mountains, hills and in karst areas (2),
- high karst plateaus and hills in carbonate rocks (3),
- hills in non-carbonate rocks (4),
- inter-mountain basins (5),
- low hills in the inner part of Slovenia (6),
- the plains and wide valleys in the area of low hills of the inner part of Slovenia (7),
- poljes (8),
- the low karst of the regions of Notranjska and Dolenjska (9),
- the low karst of the region of Bela Krajina (10),
- Kras and Podgorje karst (11),
- the low hills in the Primorska region of Slovenia (12),
- wide valleys and coastal plains in the Primorska region of Slovenia (13).

In the interest of clarity, the natural landscape typology by Perko (1998) was named TIPI9, while the ecological landscape typology by Špes et al. (2002) was named TIPI13.

2.2 Preparation of distorted natural landscape typologies

First, we rasterized both original natural landscape typologies. The resolution of the raster layer was 200 m (506,450 cells). This amount of cells can still be processed by usual computer equipment. Than we randomly redistributed both original typologies and thereby also the selection of random learning cells across the space (Figure 1). With both, we preserved all the types in the same ratio and with the same number of cells, as the cells were just randomly redistributed. In this way, we acquired a distorted (random or incongruous) typology. The redistribution was performed with SPSS software. We expected the distorted typologies to not reach the same scores and modelling success rates as the original typologies.

The original TIPI9 typology corresponded with the distorted typology in 15.1%, which is equal to the calculation of the theoretical agreement. The original TIPI13 typology corresponded with its distorted typology in 13.1%, which is close to the calculation of the theoretical agreement (13.4%). Random agreement between the original and the distorted typology is proof of both typologies being appropriately distorted.

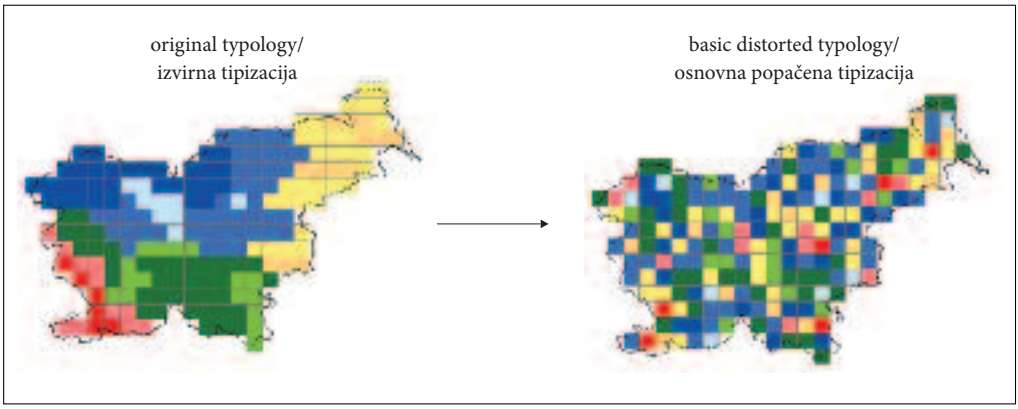


Figure 1: Outline of the formation of the distorted typology (right) with a random redistribution of the original typology cells (left).

2.3 Data layers for modelling

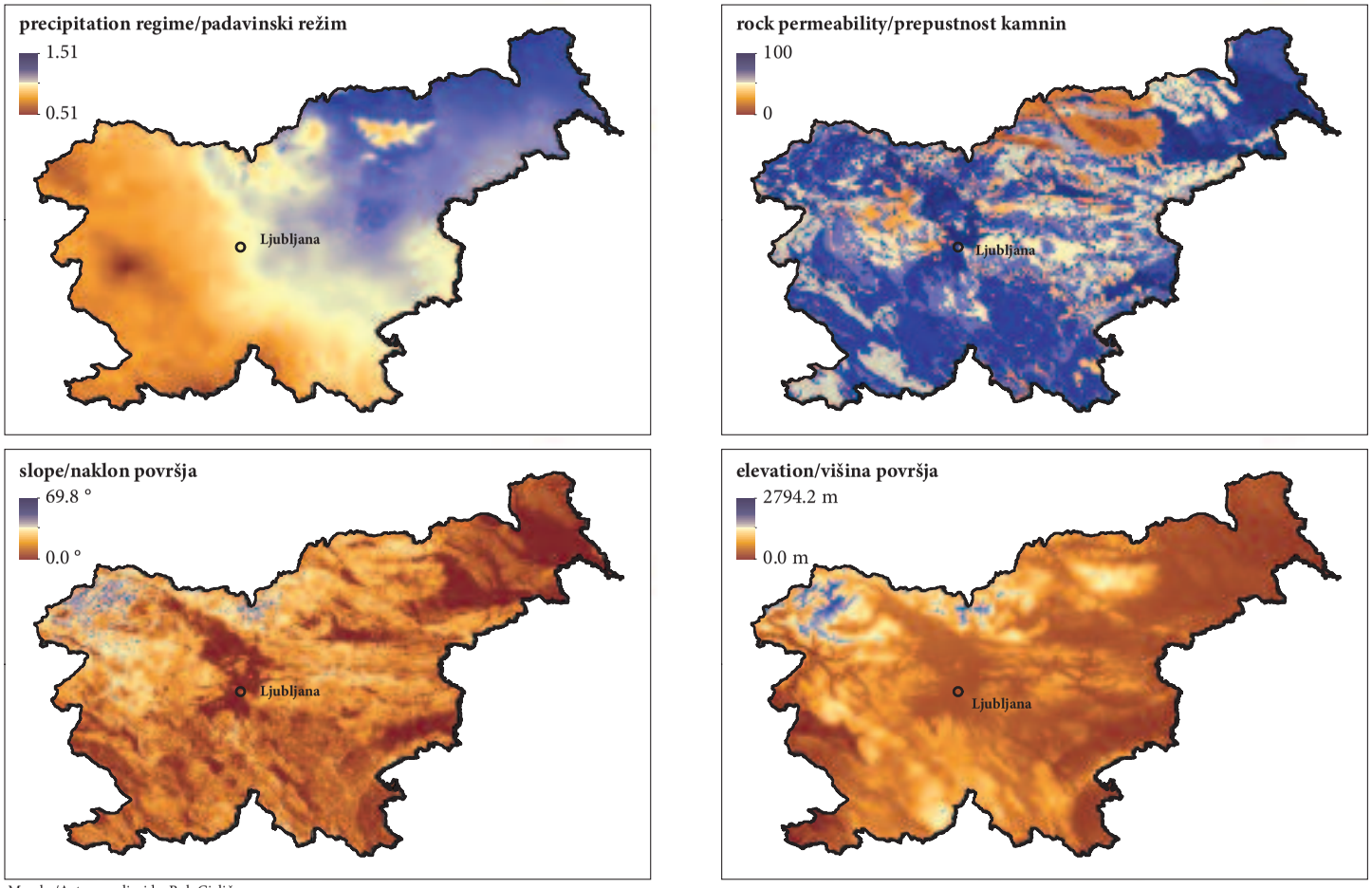
For the purposes of modelling we selected four data layers, which had proved to be adequate in previous studies, having already been conducive to relatively successful original typology modelling (e.g. Ciglič and Perko 2015). These data layers are: elevation (original unit: meter), slope (degree), precipitation regime (ratio), and rock permeability (level of permeability) (Figure 2). All of the data layers were prepared at a resolution of 200 m and standardised by linear transformation between minimum and maximum values to a scale of 0–100.

2.4 Modelling and evaluation of distorted typologies

For the modelling of distorted typologies we tested out four types of the decision-tree method, the minimum distance to means method, the maximum likelihood method, and the *k*-nearest neighbours method (Table 1). The models were created based on a random sample, which encompassed 2,000 training cells from each type. This means that 18,000 cells (3.6% of the total number of cells) were used in the modelling of Perko's (1998) distorted classification, and 26,000 cells (5.1% of all cells) were used with the distorted classification by Špes et al. (2002).

We evaluated the modelled distorted typologies by calculating the correlation between the modelled distorted typology and the explanatory data layers, as well as the correlation between the basic distorted typology and the modelled distorted typology, taking into account all (and not only the learning) cells. If the modelling based on a congruous baseline typology was successful, the correlation between the modelled distorted typology and the explanatory data layers should be as high as possible. Also the correlation between the modelled distorted typology and the basic distorted typology should be as high as possible. We determined correlations between the typology and the explanatory variables by using information gain, the gain ratio, and the eta^2 coefficient. Correlations between typologies were determined using the kappa coefficient and the Cramer coefficient of correlation. The modelled distorted typologies were also verified with a discriminant analysis.

Figure 2: Selected data layers for modelling. ➤ *p. 120*



Map by/Avtor zemljevida: Rok Ciglič
Sources/Viri: Litostratigrafska karta ... 2007; Zemljevid tipov kamnin 2012; Zemljevid povprečnih ... 2010; Digitalni model ... 2010

0 25 50 75 100 km

Table 1: Supervised classification methods used in the study. Modelling was performed using SPSS and Idrisi/Terrset software.

Method	Settings	Method description (source)
decision tree, SPSS version with the Gini coefficient measure	10 layers, 100 units in internal nodes, 50 units in external nodes, minimal improvement of the Gini coefficient: 0.0001, pruning SE = 1	Lin, Noe and He 2006
decision tree, Idrisi/Terrset version with the gain ratio measure	pruning (nodes with less than 1% of cells within the type)	Eastman 2015
decision tree, Idrisi/Terrset version with the Gini coefficient measure	pruning (nodes with less than 1% of cells within the type)	Eastman 2015
decision tree, Idrisi/Terrset version with the gain ratio measure	pruning (nodes with less than 1% of cells within the type)	Eastman 2015
minimum distance to means	distance type is not additionally standardised, maximum distance is not limited	McCoy 2005; Eastman 2015
maximum likelihood	the same a priori probability for each type, the minimum likelihood for classification is 0	Richards 1986; Eastman 2015
k-nearest neighbours	the number of *k*-neighbours is 30, the highest permitted value of cells from individual categories was 2000 training cells	Kononenko and Kukar 2007; Eastman 2015

3 Results and discussion

We noted that due to the incongruity of the basic distorted typology, decision tree methods did not construct rules at all; namely, the computer programme didn't find suitable classification rules and the algorithm stopped. With the minimum distance to means method, the maximum likelihood method, and the *k*-nearest neighbours method, however, we were able to design rules for the distorted TIPI9 typology (Figure 3, Table 2), as well as the distorted TIPI13 typology (Figure 4, Table 3). Moreover, it should be noted that when modelling the TIPI13 typology by the minimum distance to means method, only 12 of the 13 types were recognised.

Figure 3: Modelled distorted TIPI9 typologies following successful supervised classification.
► *p. 122*
Figure 4: Modelled distorted TIPI13 typologies following successful supervised classification.
► *p. 123*

basic distorted typology TIPI9/
osnovna popačena tipizacija TIPI9
modelirana popačena tipizacija TIPI9
po metodi največje verjetnosti/modelled distorted
typology TIPI9 created by the maximum
likelihood method
modelirana popačena tipizacija TIPI9 po
metodi najmanjše razdalje/modelled distorted typology
TIPI9 created by the minimum
distance to means method
modelirana popačena tipizacija TIPI9 po metodi
k najbližjih sosedov/modelled distorted typology TIPI9
created by the k–nearest
neighbours method
Type/tip
1.1
1.2
1.3
2.1
2.2
3.1
3.2
4.1
4.2
Content and map by/Avtor vsebine in zemljevida: Rok Ciglič
© 2018, Anton Melik Geographical Institute/Geografski inštitut Antona Melika ZRC SAZU
0 25 50 75 100 km

basic distorted typology TIPI13/
osnovna popačena tipizacija TIPI13
modelirana popačena tipizacija TIPI13
po metodi največje verjetnosti/modelled distorted
typology TIPI13 created by the maximum
likelihood method
modelirana popačena tipizacija TIPI13 po
metodi najmanjše razdalje/modelled distorted typology
TIPI13 created by the minimum
distance to means method
modelirana popačena tipizacija TIPI13 po metodi
k najbližjih sosedov/modelled distorted typology TIPI13
created by the k–nearest
neighbours method
Type/tip 1 2 3 4 5 6 7 8 9 10 11 12 13
Content and map by/Avtor vsebine in zemljevida: Rok Ciglič
© 2018, Anton Melik Geographical Institute/Geografski inštitut Antona Melika ZRC SAZU
0 25 50 75 100 km

Table 2: Share of types according to the basic distorted TIPI9 typology and according to individual modelled distorted TIPI9 typologies.

Type code	Share of area (%)			
	Basic distorted TIPI9 typology	Modelled distorted TIPI9 typologies		
		k-nearest neighbours	Maximum likelihood	Minimum distance to means
1.1	15.1	14.9	12.6	24.5
1.2	23.0	14.3	4.1	3.4
1.3	4.0	12.0	3.2	0.6
2.1	14.8	11.1	12.5	3.6
2.2	6.4	10.4	38.1	20.3
3.1	18.8	9.9	3.5	6.3
3.2	9.4	9.2	6.4	<0.1
4.1	5.2	9.4	13.6	27.9
4.2	3.3	8.8	6.0	13.4
Total	100.0	100.0	100.0	100.0

Table 3: Share of types according to the basic distorted TIPI13 typology and according to individual modelled distorted TIPI13 typologies.

Type code	Share of area (%)			
	Basic distorted TIPI13 typology	Modelled distorted TIPI13 typologies		
		k-nearest neighbours	Maximum likelihood	Minimum distance to means
1	8.3	12.1	3.7	3.6
2	3.6	11.8	4.6	0.3
3	17.7	9.8	9.0	19.5
4	22.4	8.8	13.5	28.5
5	6.2	8.5	22.2	2.0
6	16.2	7.6	12.4	/
7	8.8	7.4	9.7	0.4
8	4.4	6.2	2.3	6.7
9	5.4	5.7	5.7	<0.1
10	1.6	6.2	3.4	5.8
11	2.5	6.1	0.1	<0.1
12	1.7	4.6	4.4	6.6
13	1.2	5.1	9.0	26.6
Total	100.0	100.0	100.0	100.0

3.1 Analysis of TIPI9 results

A review of the correlation between the modelled distorted typologies and the data layers (Table 4) showed that, according to several evaluation criteria, the two modelled distorted typologies created by the minimum distance to means method and the maximum likelihood method have considerably higher correlation rates than the basic distorted typology and the modelled distorted typology created by the *k*-nearest neighbours algorithm. A higher correlation rate means that the types are formed in such a way as to better fit the numeric values of data layers. The share of appropriately classified types following verification by discriminant analysis also showed that the aforementioned methods yielded the more congruous models (Table 5).

Table 4: Evaluation of TIPI9 typologies according to their correlation with individual data layers.

	Typology	Information gain (bit)	Information gain ratio	Eta^2 coefficient
Slope	basic distorted typology	0.00	0.00	0.00
	modelled distorted typology created by the *k*-nearest neighbours method	0.04	0.01	0.00
	modelled distorted typology created by the maximum likelihood method	0.55	0.10	0.50
	modelled distorted typology created by the minimum distance to means method	0.55	0.11	0.43
Precipitation regime	basic distorted typology	0.00	0.00	0.00
	modelled distorted typology created by the *k*-nearest neighbours method	0.09	0.01	0.01
	modelled distorted typology created by the maximum likelihood method	0.71	0.09	0.54
	modelled distorted typology created by the minimum distance to means method	0.87	0.11	0.65
Rock permeability	basic distorted typology	0.00	0.00	0.00
	modelled distorted typology created by the *k*-nearest neighbours method	0.05	0.01	0.00
	modelled distorted typology created by the maximum likelihood method	0.65	0.20	0.35
	modelled distorted typology created by the minimum distance to means method	1.17	0.35	0.73
Elevation	basic distorted typology	0.00	0.00	0.00
	modelled distorted typology created by the *k*-nearest neighbours method	0.06	0.01	0.01
	modelled distorted typology created by the maximum likelihood method	0.61	0.11	0.47
	modelled distorted typology created by the minimum distance to means method	0.56	0.10	0.41

Table 5: Evaluation of TIPI9 typologies based on all the data layers simultaneously using discriminant analysis.

Typology	Share (%) of cells that are classified correctly according to the discriminant analysis
Basic distorted typology	7.5
Modelled distorted typology created by the *k*-nearest neighbours method	14.2
Modelled distorted typology created by the maximum likelihood method	67.1
Modelled distorted typology created by the minimum distance to means method	72.5

When evaluating the modelled distorted typologies, we also had to ask ourselves how successfully we approached the baseline, i.e., the basic distorted typology. For this purpose, we compared the modelled distorted typologies with the basic distorted typology, calculated the kappa coefficient and the Cramer coefficient (Table 6), and subsequently also compared the number of cells that were classified the same within the basic distorted typology and an individual modelled distorted typology (Table 7).

Table 6: Kappa coefficients and Cramer coefficients for the comparison of individual modelled distorted typologies with the basic distorted TIPI9 typology.

Modelled distorted typology	Cramer coefficient	Kappa coefficient (%)
Modelled distorted typology created by the *k*-nearest neighbours method	0.006 (stat. sig, p=0,001)	0.04 (stat. sig, p=0,001)
Modelled distorted typology created by the maximum likelihood method	0.004 (not stat. sig,)	0.0 (not stat. sig,)
Modelled distorted typology created by the minimum distance to means method	0.004 (not stat. sig,)	0.0 (not stat. sig,)

Table 7: Correspondence of modelled distorted typologies with the basic distorted TIPI9 typology.

Modelled distorted typology	Share of all cells (%)	Share of training cells (%)
Modelled distorted typology created by the *k*-nearest neighbours method	12.2	21.1
Modelled distorted typology created by the maximum likelihood method	9.4	11.8
Modelled distorted typology created by the minimum distance to means method	9.4	11.7

We found agreement with the basic distorted typology to be extremely low, which means that the modelled distorted typologies differ considerably from what they should be similar to. The Cramer coefficient, the kappa coefficient, and the share of identically classified cells were low in all cases; they were

lowest with those modelled distorted typologies that were created by the maximum likelihood method and the minimum distance to means method. This means that methods have a way of »imposing« their own structure on the model; the lower the correspondence, the more the structure of a certain method is imposed. Some higher values in the evaluation of modelled distorted typologies based on data layers are actually due to methods classifying cells according to certain rules, while basically imposing their own structure with regard to data layers and without taking into account the basic distorted typology.

A lesser consideration of the basic distorted typology and the imposition of the inherent structure of an individual method are a consequence of the distorted TIPI9 typology being completely random and non-objective, and therefore not suitable for adequate modelling.

3.2 Analysis of results TIPI13

In the evaluation regarding the connection to data layers (Table 8), the modelled distorted typologies TIPT13 by using the method of minimum distance to means and the method of maximum likelihood

Table 8: Evaluating typologies TIPI13 on the basis of connection with data layers.

	Typology	Information gain (bit)	Information gain ratio	Eta^2 coefficient
Slope	Basic distorted typology	0.00	0.00	0.00
	Modelled distorted typology created by the *k*-nearest neighbours method	0.05	0.01	0.00
	Modelled distorted typology created by the maximum likelihood method	0.62	0.12	0.42
	Modelled distorted typology created by the minimum distance to means method (12 types!)	0.55	0.11	0.43
Precipitation regime	Basic distorted typology	0.00	0.00	0.00
	Modelled distorted typology created by the *k*-nearest neighbours method	0.09	0.01	0.00
	Modelled distorted typology created by the maximum likelihood method	0.69	0.09	0.41
	Modelled distorted typology created by the minimum distance to means method (12 types!)	0.80	0.10	0.59
Rock permeability	Basic distorted typology	0.00	0.00	0.00
	Modelled distorted typology created by the *k*-nearest neighbours method	0.05	0.02	0.01
	Modelled distorted typology created by the maximum likelihood method	0.77	0.23	0.30
	Modelled distorted typology created by the minimum distance to means method (12 types!)	1.40	0.43	0.83
Elevation	Basic distorted typology	0.00	0.00	0.00
	Modelled distorted typology created by the *k*-nearest neighbours method	0.07	0.01	0.49
	Modelled distorted typology created by the maximum likelihood method	0.71	0.12	0.33
	Modelled distorted typology created by the minimum distance to means method (12 types!)	0.56	0.10	0.47

are evaluated better than the modelled distorted typology using the method *k*-nearest neighbours, and the basic distorted typology. Since the minimum distance to means method also »enforces« its structure, the evaluation scores are relatively good; however (as will be seen in the second part of the article), this is why the model does not correspond to the basic distorted TIPI13 typology. The share of adequately classified cells after performing the discriminant analysis has shown that the model according to the method of minimum distance to means proved to be the most adequate (Table 9).

Table 9: Evaluating typologies TIPI13 on the basis of all data layers simultaneously using discriminant analysis.

Typology	Share (%) of cells that are classified correctly according to the discriminant analysis
Basic distorted typology	3.7
Modelled distorted typology created by the *k*-nearest neighbours method	7.4
Modelled distorted typology created by the maximum likelihood method	50.2
Modelled distorted typology created by the minimum distance to means method (12 types!)	76.4

In the modelling of the distorted TIPI13 typology, we also noted slightly higher scores for modelled typologies using certain methods (in particular, the method of minimum distance to means). Since we considered this to be a similar feature as with TIPI9, we continued the analysis also in this case and calculated how effectively the modelled distorted typologies connect to the basic distorted TIPI13 typology (Tables 10 and 11).

Table 10: Kappa coefficient and Cramer's coefficient for the comparison of modelled distorted typologies with the basic distorted TIPI13 typology ([a] we manually changed one cell's type from 8 to 6, in order for all types to be represented and to be able to calculate the kappa coefficient; [b] statistically significant at p<0.001; [c] statistically not significant).

Modelled distorted typology	Cramer coefficient	Kappa coefficient (%)
Modelled distorted typology created by the *k*-nearest neighbours method	0.009[b]	0.5[b]
Modelled distorted typology created by the maximum likelihood method	0.005[c]	0.0[c]
Modelled distorted typology created by the minimum distance to means method (12 types!)[a]	0.005[c]	0.1[c]

According to the number of types, the agreement of the modelled and the original distorted typology is, as expected, even smaller than in the case of the analysis of TIPI9. The Cramer's coefficient, the kappa coefficient, and the share of equally classified cells are very low. In comparison to all the cells, the share of equally classified cells is very low in all modelled typologies (app. 10%) and is mostly a result of a random agreement. Some higher values of connection to data layers are actually the consequence

Table 11: Agreement of modelled distorted typologies with basic distorted TIPI13 typology.

Modelled distorted typology	Share of all cells (%)	Share of training cells (%)
Modelled distorted typology created by the *k*-nearest neighbours method	9.0	17.0
Modelled distorted typology created by the maximum likelihood method	10.0	8.5
Modelled distorted typology created by the minimum distance to means method (12 types!)	11.2	8.1

of the fact that the methods classify the cells relatively well according to certain rules, but at the same time impose their own structure.

A smaller consideration of the basic distorted original typology and imposing own structure of an individual method are a consequence of the fact that the distorted TIPI13 typology is actually not proper and does not enable adequate modelling.

3.3 Comparison of results with previous research studies

By comparing the results of modelling the distorted typologies to the results, acquired with modelling the original (undistorted!) typologies, it can be seen that the agreements in modelling the originals were substantially higher. In modelling TIPI9 on the basis of a 3.6% learning sample, four explanatory data layers, and different classification methods, all the tools were able to determine the rules, with the agreement between the original typology and its model being between 51% (method of minimum distance to means) and 75% (method of *k*-nearest neighbours) (Ciglič 2014; Ciglič and Perko 2015). In modelling TIPI9 on the basis of a 1.0% learning sample, six explanatory data layers, and the method of random forests, a 94% agreement was achieved (Ciglič et al. 2017). In modelling TIPI13 on the basis of a 5.1% learning sample, four explanatory data layers, and different classification methods, the rules were set by all the methods, with the agreement between the original typology and its model being between 47% (method of minimum distance to means) and 69% (method of *k*-nearest neighbours) (Ciglič 2014). In modelling TIPI13 on the basis of a 100.0% learning sample, different combinations of data layers, and the decision tree method, the agreement between the original typology and its model ranged from 71.2% to 79.5% (Ciglič 2012). All the cases show a much higher agreement than the modelling of distorted classifications. All the tools formed the classification rules, meaning that the original classifications TIPI9 and TIPI13 are not random and are designed with a high degree of objectivity, since they can be confirmed by different quantitative models.

The modelling of distorted typologies has shown that some methods (different decision tree algorithms) cannot prepare classification rules if input data is incongruous or illogical. Some methods (minimum distance to means, maximum likelihood, *k*-nearest neighbours) indeed managed to form a modelled distorted typology; however, it was entirely different than its baseline.

This means that by using the supervised classification methods we can model features (in this case landscape types) for which the data of sufficient quality is available. If not, the rule »*garbage in, garbage out*« is applied, meaning that bad data (input into a model) cannot lead to congruous or useful results (output from a model). Due to these findings, caution is advised when using classification methods, making sure that results are checked in different ways.

4 Conclusion

The article addresses the capability of prediction of certain supervised classification methods. We tried to assess the results deriving from different supervised classification methods, if they are used for modelling a random or incongruously based classification. For this purpose, we distorted two natural landscape typologies of Slovenia and tried to determine logical classification rules for them by using different methods.

We noted that certain methods (decision tree method) cannot confirm random or incongruous classifications even to the lowest extent, since the algorithm is not capable of finding logical classification rules and clearly determined types. Certain methods are indeed capable of modelling a classification, but only to the extent where the entire agreement between the original and the model can be attributed merely to a coincidence. It is important to emphasise that certain methods, which we checked (foremost the method of minimum distance to means), produce visually relatively logical natural landscape types, however, these come as a result of »enforcing« the method's own structure, which produces a typology entirely different than the original. Due to the mentioned findings, it is suggested the results be checked in several ways. Both compositions of modelling distorted typologies (TIPI9 and TIPI13) led us to similar conclusions. Distorted typologies can objectively be regarded as inadequate, since:

- by using certain methods, we were not able to produce a model due to the incongruity of distorted typology,
- produced modelled distorted typologies did not comply with the basic distorted typology despite some relatively positive scores from the perspective of correlation with the explanatory data layers.

Checking the modelling of distorted classifications can be helpful for studying the characteristics of geoinformation tools or for familiarising with their operation, since the user is given the information about which supervised classification method enforces its own structure more and which method fits the learning data better (even though the data is bad or distorted). This way a user with a representative database can use the methods that can fit the learning data better. Otherwise he should try out different methods and compare results.

Acknowledgements: The author acknowledges financial support from the Slovenian Research Agency (project no. L1-7542: Advancement of computationally intensive methods for efficient modern general-purpose statistical analysis and inference; program no. P6-0101: Geography of Slovenia).

5 References

Bailey, R. 1996: Ecosystem Geography. New York.

Bailey, R. 2004: Identifying ecoregion boundaries. Environmental Management 34-S1. DOI: https://doi.org/10.1007/s00267-003-0163-6

Belbin, L., McDonald, C. 1993: Comparing three classification strategies for use in ecology. Journal of Vegetation Science 4-3. DOI: https://doi.org/10.2307/3235592

Breskvar Žaucer, L., Marušič, J. 2006: Analiza krajinskih tipov z uporabo umetnih nevronskih mrež. Geodetski vestnik 50-2.

Ciglič, R. 2012: Evaluation of digital data layers for establishing natural landscape types in Slovenia. Geopolitics, History and International Relations 4-2.

Ciglič, R. 2014: Analiza naravnih pokrajinskih tipov Slovenije z GIS-om. Geografija Slovenije 28. Ljubljana.

Ciglič, R., Perko, D. 2015: Modelling as a method for evaluating natural landscape typology: The case of Slovenia. Landscape Analysis and Planning. Cham. DOI: https://doi.org/10.1007/978-3-319-13527-4_4

Ciglič, R., Perko, D., Hrvatin, M., Štrumbelj, E. 2017: Modeling and evaluating older landscape classifications with modern quantitative methods. From Pattern and Process to People and Action. Ghent.

Digitalni model višin 25. Geodetska uprava Republike Slovenije. Ljubljana, 2010.
Eastman, J. R. 2015: Terrset Manual. Worcester.
Gams, I. 1978: Kvantitativna prirodnogeografska regionalizacija Slovenije. Ljubljana.
Gams, I. 1986: Za kvantitativno razmejitev med pojmi gričevje, hribovje in gorovje. Geografski vestnik 58.
Kokalj, Ž., Oštir., K. 2013: Vrednotenje pokrajinskoekoloških tipov Slovenije v luči pokrovnosti, izdelane s klasifikacijo satelitskih posnetkov Landsat. Prostor, kraj, čas 1. Ljubljana.
Kononenko, I., Kukar, M. 2007: Machine Learning and Data Mining: Introduction to Principles and Algorithms. Chichester.
Lin, N., Noe, D., He., X. 2006: Tree-based methods and their applications. Springer Handbook of Engineering Statistics. London.
Litostratigrafska karta Slovenije. Geološki zavod Slovenije. Ljubljana, 2007.
McCoy, R. M. 2005: Field Methods in Remote Sensing. New York.
Melik, A. 1946: Prirodnogospodarska sestava Slovenije. Geografski vestnik 18.
Mücher, C. A., Bunce, R. G. H., Jongman, R. H. G., Klijn, J. A., Koomen, A. J. M., Metzger, M. J., Wascher, D. M. 2003: Identification and Characterisation of Environments and Landscapes in Europe. Alterra-Rapport 832. Wageningen.
Mücher, C. A., Klijn, J. A., Wascher, D. M., Schaminée, J. H. J. 2010: A new European landscape classification (LANMAP): A transparent, flexible and user-oriented methodology to distinguish landscapes. Ecological Indicators 10-1. DOI: https://doi.org/10.1016/j.ecolind.2009.03.018
Olson, D. M., Dinerstein, E., Wikramanayake, E. D., Burgess, N. D., Powell, G. V. N., Underwood, E. C., D'Amico, J. A., Itoua, I., Strand, H. E., Morrison, J. C., Loucks, C. J., Allnut, T. F., Ricketts, T. H., Kura, Y., Lamoreux, J. F., Wettengel, W. W., Hedao, P., Kassem, K. R. 2001: Terrestrial ecoregions of the World: A new map of life on earth. BioScience 51-11. DOI: https://doi.org/10.1641/0006-3568(2001)051[0933:TEOTWA]2.0.CO;2
Perko, D. 1998: The regionalization of Slovenia. Geografski zbornik 38.
Perko, D., Hrvatin, M., Ciglič, R. 2015: A methodology for natural landscape typification of Slovenia. Acta geographica Slovenica 55-2. DOI: https://doi.org/10.3986/AGS.1938
Perko, D., Hrvatin, M., Ciglič, R. 2017: Determination of landscape hotspots of Slovenia. Acta geographica Slovenica 57-1. DOI: https://doi.org/10.3986/AGS.4618
Perko, D., Zorn, M. 2016: Sedemdeset let raziskovanj na Geografskem inštitutu Antona Melika ZRC SAZU. Geografski vestnik 88-2. DOI: https://doi.org/10.3986/GV88207
Richards, J. A. 1986: Multispectral Transformations of Image Data. Remote Sensing Digital Image Analysis: An Introduction. Berlin.
Špes, M., Cigale, D., Lampič, B., Natek, K., Plut, D., Smrekar, A. 2002: Študija ranljivosti okolja: metodologija in aplikacija. Geographica Slovenica 35, 1-2. Ljubljana.
Strand, G.-H. 2011: Uncertainty in classification and delineation of landscapes: A Probabilistic approach to landscape modeling. Environmental Modelling and Software 26-9. DOI: https://doi.org/10.1016/j.envsoft.2011.03.005
Udo de Haes, H. A., Klijn, F. 1994: Environmental policy and ecosystem classification. Ecosystem Classification for Environmental Management. Dordrecht.
Zemljevidi povprečnih mesečnih in letnih padavin 1971–2000. Agencija Republike Slovenije za okolje. Ljubljana, 2010.
Zemljevid tipov kamnin. Geografski inštitut Antona Melika ZRC SAZU, različica 9.12. Ljubljana, 2012.

PREVERJANJE VPLIVA NESMISELNOSTI VHODNIH PODATKOV PRI IZBRANIH KVANTITATIVNIH METODAH ZA MODELIRANJE NARAVNOPOKRAJINSKIH TIPIZACIJ

1 Uvod

1.1 Teoretično ozadje raziskave

Določanje naravnopokrajinskih tipov ima v Sloveniji dolgo tradicijo (Melik 1946; Perko 1998; Špes in sodelavci 2002; Perko, Hrvatin in Ciglič 2015; 2017; Perko in Zorn 2016), prav tako tudi drugod po svetu (na primer Olson in sodelavci 2001; Mücher in sodelavci 2010). Ob analizah pokrajine so avtorji opozarjali na številne izzive, ki so prisotni pri procesu delitve pokrajine na posamezne enote (Gams 1986; Mücher in sodelavci 2003; Bailey 2004). Eden izmed izzivov je tudi odsotnost splošnega dogovora o dojemanju pokrajine. Udo de Haes in Klijn (1994) sta izpostavila, da je lahko ekosistem definiran kot abstrakten pojem ali pa kot dejansko prepoznaven objekt. Da so enote dejansko lahko prepoznane, trdi Bailey (1996, 4), ki pravi, da ekosistemi kot geografske enote pokrajine vključujejo vse naravne pojave in so lahko prepoznani in zamejeni z mejami. Gams (1978, 15) pa nasprotno trdi, da »… *je vsaka regija z omejeno črto na karti nenaravna, umetna tvorba in rabi samo kot sredstvo ugotavljanja razlik* …«.

Zato je tudi preverjanje obstoja naravnopokrajinskih tipov ena izmed pomembnih prvin, saj so že dalj časa na voljo različni računalniški algoritmi, ki omogočajo modeliranje že izdelanih tipizacij. Tako lahko zasledimo različne primere tovrstnega preverjanja pokrajinskih klasifikacij (Breskvar Žaucer in Marušič 2006; Strand 2011). Analize (Ciglič 2012; 2014; Kokalj in Oštir 2013; Ciglič in Perko 2015), ki obravnavajo dve slovenski naravnopokrajinski tipizaciji (Perko 1998; Špes in sodelavci 2002), so pokazale, da sta obe, kljub temu, da sta bili narejeni z ročnim določanjem meja, dovolj kakovostni in ju je mogoče modelirati s kvantitativnimi metodami.

Na tem mestu se sprašujemo, ali bi v primeru zelo slabo zasnovanih oziroma naključnih tipizacij metode nadzorovane klasifikacije, ki so pogosto uporabljene v pokrajinskih analizah (na primer odločitvena drevesa, metoda največje verjetnosti, metoda *k* najbližjih sosedov, metoda najmanjše razdalje), sploh lahko izdelale modele. Zasledili smo namreč preizkus, kako se obnesejo metode v geografskih analizah, ki jih avtorji opravijo na simuliranih podatkih (Belbin in McDonald 1993). Vsaka metoda ima določene prednosti in slabosti, ponudijo pa tudi različne rezultate.

V prispevku želimo preveriti, kako se obnesejo metode za modeliranje tipizacij, tudi če so te narejene povsem brez upoštevanja kakršnega koli naravnogeografskega ozadja, torej povsem naključno. S poskusom smo na primeru Slovenije preverili, kako se metode obnašajo v primeru naključnih, lahko bi dejali tudi nesmiselnih naravnopokrajinskih tipizacij. Tako bomo lahko ocenili, kakšen je vpliv samih metod na rezultat modeliranja. Za ta preizkus smo popačili dve izvirni naravnopokrajinski tipizaciji – po Perku (1998) ter Špesovi in sodelavcih (2002).

1.2 Terminologija

V prispevku uporabljamo daljše izraze za ločevanje med vrstami tipizacij:

- **izvirna (naravnopokrajinska) tipizacija** – s tem izrazom označujemo izvirni tipizaciji; to sta tipizaciji Slovenije po Perku (1998) ter po Špesovi in sodelavcih (2002),
- **(osnovna) popačena (naravnopokrajinska) tipizacija** – s tem izrazom označujemo namerno popačeni izvirni naravnopokrajinski tipizaciji, ki sta služili za modeliranje nesmiselnih tipizacij Slovenije,
- **modelirana popačena (naravnopokrajinska) tipizacija** – s tem izrazom označujemo vsako tipizacijo, ki je bila narejena z metodo nadzorovane klasifikacije na podlagi osnovne popačene izvirne tipizacije.

2 Metodologija

2.1 Izvirni naravnopokrajinski tipizaciji

Prispevek smo zasnovali na analizi naravnopokrajinskih tipizacij po Perku (1998) ter Špesovi in sodelavcih (2002). Perko je določil 9 tipov:
- alpska gorovja tip (oznaka tipa je 1.1),
- alpska hribovja (1.2),
- alpske ravnine (1.3),
- panonska gričevja (2.1),
- panonske ravnine (2.2),
- dinarske planote (3.1),
- dinarska podolja in ravniki (3.2),
- sredozemska gričevja (4.1),
- sredozemske planote (4.2).

Špesova in sodelavci (2002) pa so določili 13 tipov:
- visokogorski svet (oznaka tipa je 1),
- širše rečne doline v visokogorju, hribovju in na krasu (2),
- visoke kraške planote in hribovja v karbonatnih kamninah (3),
- hribovja v pretežno nekarbonatnih kamninah (4),
- medgorske kotline (5),
- gričevje v notranjem delu Slovenije (6),
- ravnine in širše doline v gričevju notranjega dela Slovenije (7),
- kraška polja in podolja (8),
- nizki kras Notranjske in Dolenjske (9),
- nizki kras Bele krajine (10),
- Kras in Podgorski kras (11),
- gričevje v primorskem delu Slovenije (12),
- širše doline in obalne ravnice v primorskem delu Slovenije (13).

Zaradi preglednosti smo naravnopokrajinsko tipizacijo po Perku (1998) poimenovali TIPI9, pokrajinskoekološko tipizacijo po Špesovi in sodelavcih (2002) pa TIPI13.

2.2 Priprava popačenih naravnopokrajinskih tipizacij

Najprej smo izvirni tipizaciji rasterizirali. Ločljivost rastrskega sloja je bila 200 m, kar predstavlja 506.450 celic in še omogoča analize z običajno računalniško opremo. Nato smo obe izvirni tipizaciji in s tem tudi nabor naključnih učnih celic naključno prerazporedili po prostoru (slika 1). Pri obeh smo ohranili vse tipe v enakem razmerju in z enakim številom celic, celice smo le naključno prerazporedili. Na ta način smo dobili osnovno popačeno (naključno ali nesmiselno) tipizacijo. Prerazporeditev je bila narejena s pomočjo programa SPSS. Za popačeni tipizaciji pričakujemo, da ne bosta dosegli takšnih ocen in uspešnosti modeliranja kot pa izvirni tipizaciji.

Izvirna tipizacija TIPI9 se je s popačeno različico ujemala v 15,1 %, kar je enako izračunu teoretičnega ujemanja. Izvirna tipizacija TIPI13 se je s svojo popačeno različico ujemala v 13,1 %, kar je blizu izračunu teoretičnega ujemanja (ta je 13,4 %). Naključno ujemanje med izvirno in osnovno popačeno tipizacijo je dokaz, da smo tipizaciji ustrezno popačili.

Slika 1: Shema izdelave osnovne popačene tipizacije (desno) z naključno prerazporeditvijo celic izvirne tipizacije (levo).
Glej angleški del prispevka.

2.3 Podatkovni sloji za modeliranje

Za modeliranje smo izbrali štiri podatkovne sloje, ki so se v preteklih raziskavah pokazali za ustrezne in so z njimi izvirne tipizacije že razmeroma uspešno modelirali (na primer Ciglič in Perko 2015). Ti podatkovni sloji so: nadmorska višina (osnovna enota: meter), naklon (stopinja), padavinski režim (padavinsko razmerje) in prepustnost kamnin (stopnja prepustnosti) (slika 2). Vsi podatkovni sloji so bili pripravljeni v ločljivosti 200 m ter standardizirani z linearno transformacijo med minimalno in maksimalno vrednostjo na vrednostno lestvico 0–100.

Slika 2: Izbrani podatkovni sloji za modeliranje.
Glej angleški del prispevka.

2.4 Modeliranje in vrednotenje popačenih tipizacij

Za modeliranje popačenih tipizacij smo preizkusili štiri vrste metod odločitvenih dreves, metodo najmanjše razdalje, metodo največje verjetnosti in metodo *k* najbližjih sosedov (preglednica 1). Modele smo izdelali na naključnem vzorcu, v katerem smo zajeli po 2000 učnih celic iz vsakega tipa. To pomeni, da smo pri Perkovi (1998) popačeni klasifikaciji za izdelavo modela uporabili 18.000 celic (3,6 % vseh celic), pri popačeni klasifikaciji Špesove in sodelavcev (2002) pa 26.000 celic (5,1 %).

Preglednica 1: Metode nadzorovane klasifikacije, ki smo jih uporabili v raziskavi. Modeliranje smo izvedli v programih SPSS in Idrisi/Terrset.

metoda	nastavitve	opis metode (vir)
odločitveno drevo, različica SPSS z mero Ginijev koeficient	deset ravni, 100 enot v notranjih vozliščih, 50 enot v zunanjih vozliščih, minimalno izboljšanje Ginijevega koeficienta: 0,0001, obrezovanje/pruning SE = 1)	Lin, Noe in He 2006
odločitveno drevo, različica Idrisi/Terrset z mero razmerje informacijskega prispevka	obrezovanje (vozlišča z manj kot 1 % celic v tipu)	Eastman 2015
odločitveno drevo, različica Idrisi/Terrset z mero Ginijev koeficient	obrezovanje (vozlišča z manj kot 1 % celic v tipu)	Eastman 2015
odločitveno drevo, različica Idrisi/Terrset z mero informacijski prispevek	obrezovanje (vozlišča z manj kot 1 % celic v tipu)	Eastman 2015
metoda najmanjše razdalje	tip razdalje ni dodatno standardiziran, najdaljša razdalja ni omejena	McCoy 2005; Eastman 2015
metoda največje verjetnosti	enake apriorne verjetnosti za vsak tip, minimalna verjetnost za klasifikacijo je 0	Richards 1986; Eastman 2015
metoda *k* najbližjih sosedov	število sosedov *k* je 30, najvišje dovoljeno število celic iz posamezne kategorije je bilo 2000 učnih celic	Kononenko in Kukar 2007; Eastman 2015

Dobljene modelirane popačene tipizacije smo ovrednotili tako, da smo na podlagi vseh celic (ne le učnih celic!) izračunali povezanost med modelirano popačeno tipizacijo in pojasnjevalnimi podatkovnimi sloji ter povezanost med modelirano popačeno tipizacijo in osnovno popačeno tipizacijo. Ob uspešnem modeliranju na podlagi smiselne izhodiščne tipizacije, bi morala biti povezanost med modelirano popačeno tipizacijo in pojasnjevalnimi podatkovnimi sloji ter med modelirano popačeno tipizacijo in osnovno popačeno tipizacijo čim večja. Povezanosti med tipizacijo in pojasnjevalnimi spremenljivkami smo ugotavljali z informacijskim prispevkom, razmerjem informacijskega prispevka ter koeficientom eta^2. Povezanost med tipizacijami pa smo ugotavljali s koeficientom kappa in Cramerjevim koeficientom povezanosti. Modelirane popačene tipizacije smo preverili tudi z diskriminančno analizo.

3 Rezultati in diskusija

Opazili smo, da zaradi nesmiselnosti osnovne popačene tipizacije metode z odločitvenimi drevesi sploh niso uspele izdelati pravil; računalniški program namreč ni našel ustreznih klasifikacijskih pravil in algoritem se je ustavil. Z metodami najmanjše razdalje, največje verjetnosti in *k* najbližjih sosedov pa smo pravila uspeli izdelati za popačeno tipizacijo TIPI9 (slika 3, preglednica 2) in tudi za popačeno tipizacijo TIPI13 (slika 4, preglednica 3). Tu je treba še opozoriti, da je bilo pri modeliranju tipizacije TIPI13 po metodi najmanjše razdalje prepoznanih samo 12 od 13 tipov!

Slika 3: Modelirane popačene tipizacije TIPI9 po uspelih metodah nadzorovane klasifikacije.
Glej angleški del prispevka.

Slika 4: Modelirane popačene tipizacije TIPI13 po uspelih metodah nadzorovane klasifikacije.
Glej angleški del prispevka.

Preglednica 2: Delež tipov po osnovni popačeni tipizaciji TIPI9 in po posameznih modeliranih popačenih tipizacijah TIPI9.

oznaka tipa	delež površja (%)			
	popačena tipizacija	modelirane popačene tipizacije		
		k najbližjih sosedov	največja verjetnost	najmanjša razdalja
1.1	15,1	14,9	12,6	24,5
1.2	23,0	14,3	4,1	3,4
1.3	4,0	12,0	3,2	0,6
2.1	14,8	11,1	12,5	3,6
2.2	6,4	10,4	38,1	20,3
3.1	18,8	9,9	3,5	6,3
3.2	9,4	9,2	6,4	<0,1
4.1	5,2	9,4	13,6	27,9
4.2	3,3	8,8	6,0	13,4
skupaj	100,0	100,0	100,0	100,0

Preglednica 3: Delež tipov po osnovni popačeni tipizaciji TIPI13 in po posameznih modeliranih popačenih tipizacijah.

oznaka tipa	delež površja (%)			
	popačena tipizacija	modelirana popačene tipizacije		
		k najbližjih sosedov	največja verjetnost	najmanjša razdalja
1	8,3	12,1	3,7	3,6
2	3,6	11,8	4,6	0,3
3	17,7	9,8	9,0	19,5
4	22,4	8,8	13,5	28,5
5	6,2	8,5	22,2	2,0
6	16,2	7,6	12,4	/
7	8,8	7,4	9,7	0,4
8	4,4	6,2	2,3	6,7
9	5,4	5,7	5,7	<0,1
10	1,6	6,2	3,4	5,8
11	2,5	6,1	0,1	<0,1
12	1,7	4,6	4,4	6,6
13	1,2	5,1	9,0	26,6
skupaj	100,0	100,0	100,0	100,0

3.1 Analiza rezultatov TIPI9

Pri pregledu povezanosti modeliranih popačenih tipizacij s podatkovnimi sloji (preglednica 4), smo opazili, da modelirani popačeni tipizaciji po metodi najmanjše razdalje in metodi največje verjetnosti dosegata precej višje povezanosti po več načinih vrednotenja kot pa osnovna popačena tipizacija in modelirana popačena tipizacija po metodi *k* najbližjih sosedov. Višja povezanost pomeni, da so tipi oblikovani tako, da bolje ustrezajo številskim vrednostim podatkovnih slojev, in se zato s podatkovnimi sloji bolje povezujejo. Delež ustrezno klasificiranih po preverjanju z diskriminančno analizo je prav tako pokazal, da sta omenjeni metodi naredili bolj smiselna modela (preglednica 5).

Pri vrednotenju modeliranih popačenih tipizacij se moramo tudi vprašati, kako uspešno smo se približali izhodišču, torej osnovni popačeni tipizaciji. Zato smo primerjali modelirane popačene tipizacije in osnovno popačeno tipizacijo ter izračunali koeficient kappa in Cramerjev koeficient (preglednica 6), nato pa smo primerjali tudi, koliko celic je enako klasificiranih v osnovni popačeni tipizaciji in posamezni modelirani popačeni tipizaciji (preglednica 7).

Ugotovili smo, da je ujemanje z osnovno popačeno tipizacijo izredno majhno, kar pomeni, da se modelirane popačene tipizacije precej razlikujejo od tistega, čemur bi morale biti podobne. Cramerjev koeficient, koeficient kappa in delež enako klasificiranih celic so povsod zelo nizki, najnižji pri modeliranih popačenih tipizacijah po metodah največje verjetnosti in najmanjše razdalje. To pomeni, da metode »vsilijo« svojo strukturo; manjše kot je ujemanje, bolj vsiljena je določena struktura metode.

Nekatere višje vrednosti vrednotenja modeliranih popačenih tipizacij na podlagi podatkovnih slojev so dejansko posledica tega, da metode klasificirajo celice po določenih pravilih, a dejansko vsilijo svojo strukturo glede na podatkovne sloje in ne upoštevajo izhodiščne popačene tipizacije.

Manjše upoštevanje osnovne popačene tipizacije in vsiljevanje lastne strukture posamezne metode, sta posledici dejstva, da je popačena tipizacija TIPI9 dejansko povsem naključna, neobjektivna in zato ne omogoča zadovoljive stopnje modeliranja.

Preglednica 4: Vrednotenje tipizacij TIPI9 na podlagi povezanosti s posameznimi podatkovnimi sloji.

	tipizacija	informacijski prispevek (bit)	razmerje informacijskega prispevka	koeficient eta^2
naklon	osnovna popačena tipizacija	0,00	0,00	0,00
	modelirana popačena tipizacija z metodo *k* najbližjih sosedov	0,04	0,01	0,00
	modelirana popačena tipizacija z metodo največje verjetnosti	0,55	0,10	0,50
	modelirana popačena tipizacija z metodo najmanjše razdalje	0,55	0,11	0,43
padavinski režim	osnovna popačena tipizacija	0,00	0,00	0,00
	modelirana popačena tipizacija z metodo *k* najbližjih sosedov	0,09	0,01	0,01
	modelirana popačena tipizacija z metodo največje verjetnosti	0,71	0,09	0,54
	modelirana popačena tipizacija z metodo najmanjše razdalje	0,87	0,11	0,65
prepustnost kamnin	osnovna popačena tipizacija	0,00	0,00	0,00
	modelirana popačena tipizacija z metodo *k* najbližjih sosedov	0,05	0,01	0,00
	modelirana popačena tipizacija z metodo največje verjetnosti	0,65	0,20	0,35
	modelirana popačena tipizacija z metodo najmanjše razdalje	1,17	0,35	0,73
nadmorska višina	osnovna popačena tipizacija	0,00	0,00	0,00
	modelirana popačena tipizacija z metodo *k* najbližjih sosedov	0,06	0,01	0,01
	modelirana popačena tipizacija z metodo največje verjetnosti	0,61	0,11	0,47
	modelirana popačena tipizacija z metodo najmanjše razdalje	0,56	0,10	0,41

Preglednica 5: Vrednotenje tipizacij TIPI9 na podlagi vseh podatkovnih slojev hkrati z diskriminančno analizo.

tipizacija	delež pravilno klasificiranih celic po diskriminančni analizi (%)
osnovna popačena tipizacija	7,5
modelirana popačena tipizacija z metodo *k* najbližjih sosedov	14,2
modelirana popačena tipizacija z metodo največje verjetnosti	67,1
modelirana popačena tipizacija z metodo najmanjše razdalje	72,5

Preglednica 6: Izračunani koeficienti kappa in Cramerjevi koeficienti za primerjavo posameznih modeliranih popačenih tipizacij z osnovno popačeno tipizacijo TIPI9.

modelirana popačena tipizacija	Cramerjev koeficient	koeficient kappa (%)
modelirana popačena tipizacija z metodo *k* najbližjih sosedov	0,006 (stat. značilen p=0,001)	0,04 (stat. značilen p=0,001)
modelirana popačena tipizacija z metodo največje verjetnosti	0,004 (ni stat. značilno)	0,0 (ni stat. značilno)
modelirana popačena tipizacija z metodo najmanjše razdalje	0,004 (ni stat. značilno)	0,0 (ni stat. značilno)

Preglednica 7: Ujemanje modeliranih popačenih tipizacij z osnovno popačeno tipizacijo TIPI9.

modelirana popačena tipizacija	delež enako klasificiranih celic od vseh celic (%)	delež enako klasificiranih učnih celic (%)
modelirana popačena tipizacija z metodo *k* najbližjih sosedov	12,2	21,1
modelirana popačena tipizacija z metodo največje verjetnosti	9,4	11,8
modelirana popačena tipizacija z metodo najmanjše razdalje	9,4	11,7

3.2 Analiza rezultatov TIPI13

Pri vrednotenju glede na povezanost s podatkovnimi sloji (preglednica 8) sta modelirani popačeni tipizaciji po metodi najmanjše razdalje in po metodi največje verjetnosti ocenjeni bolje kot modelirana popačena tipizacija po metodi *k* najbližjih sosedov ter osnovna popačena tipizacija. Ker metoda najmanjše razdalje tudi »vsili« svojo strukturo, so ocene relativno dobre, a (kot bomo videli v nadaljevanju) se model zato ne ujema z osnovno popačeno tipizacijo TIPI13. Delež ustrezno klasificiranih celic po preverjanju z diskriminančno analizo je pokazal, da se je model po metodi najmanjših razdalj prav tako izkazal kot najbolj ustrezen (preglednica 9).

Tudi pri modeliranju popačene tipizacije TIPI13 smo za modelirane tipizacije po nekaterih metodah (predvsem po metodi najmanjše razdalje) dobili nekoliko višje ocene. Ker sklepamo, da gre za podoben pojav kot pri TIPI9, smo tudi tukaj nadaljevali analizo in izračunali, kako dobro se modelirane popačene tipizacije povezujejo z osnovno popačeno tipizacijo TIPI13 (preglednici 10 in 11).

Ujemanje modelirane in izvirne popačene tipizacije je glede na število tipov pričakovano še manjše kot v primeru analize TIPI9. Cramerjev koeficient, koeficient kappa in delež enako klasificiranih celic so zelo nizki. Delež enako klasificiranih celic glede na vse celice je zelo majhen pri vseh modeliranih tipizacijah (okrog 10 %) in je predvsem rezultat naključnega ujemanja. Nekatere višje vrednosti povezanosti s podatkovnimi sloji so dejansko posledica dejstva, da metode klasificirajo celice po določenih pravilih relativno dobro, a vsilijo svojo strukturo.

Manjše upoštevanje izhodiščne osnovne popačene tipizacije in vsiljevanje lastne strukture posamezne metode sta posledici tega, da je popačena tipizacija TIPI13 dejansko slaba oziroma neobjektivna ter ne omogoča ustreznega modeliranja.

Preglednica 8: Vrednotenje tipizacij TIPI13 na podlagi povezanosti s podatkovnimi sloji.

	tipizacija	informacijski prispevek (bit)	razmerje informacijskega prispevka	koeficient eta^2
naklon	osnovna popačena tipizacija	0,00	0,00	0,00
	modelirana popačena tipizacija z metodo *k* najbližjih sosedov	0,05	0,01	0,00
	modelirana popačena tipizacija z metodo največje verjetnosti	0,62	0,12	0,42
	modelirana popačena tipizacija z metodo najmanjše razdalje (12 skupin!)	0,55	0,11	0,43
padavinski režim	osnovna popačena tipizacija	0,00	0,00	0,00
	modelirana popačena tipizacija z metodo *k* najbližjih sosedov	0,09	0,01	0,00
	modelirana popačena tipizacija z metodo največje verjetnosti	0,69	0,09	0,41
	modelirana popačena tipizacija z metodo najmanjše razdalje (12 skupin!)	0,80	0,10	0,59
prepustnost kamnin	osnovna popačena tipizacija	0,00	0,00	0,00
	modelirana popačena tipizacija z metodo *k* najbližjih sosedov	0,05	0,02	0,01
	modelirana popačena tipizacija z metodo največje verjetnosti	0,77	0,23	0,30
	modelirana popačena tipizacija z metodo najmanjše razdalje (12 skupin!)	1,40	0,43	0,83
nadmorska višina	osnovna popačena tipizacija	0,00	0,00	0,00
	modelirana popačena tipizacija z metodo *k* najbližjih sosedov	0,07	0,01	0,49
	modelirana popačena tipizacija z metodo največje verjetnosti	0,71	0,12	0,33
	modelirana popačena tipizacija z metodo najmanjše razdalje (12 skupin!)	0,56	0,10	0,47

Preglednica 9: Vrednotenje tipizacij TIPI13 na podlagi vseh podatkovnih slojev hkrati z diskriminančno analizo.

tipizacija	delež pravilno klasificiranih celic po diskriminančni analizi (%)
osnovna popačena tipizacija	3,7
modelirana popačena tipizacija z metodo *k* najbližjih sosedov	7,4
modelirana popačena tipizacija z metodo največje verjetnosti	50,2
modelirana popačena tipizacija z metodo najmanjše razdalje (12 skupin!)	76,4

Preglednica 10: Koeficient kappa in Cramerjev koeficient za primerjavo posameznih modeliranih popačenih tipizacij s popačeno tipizacijo TIPI13 ([a] ročno smo eni celici spremenili tip iz 8 v 6 zato, da so bili zastopani vsi tipi in je bilo mogoče izračunati koeficient kappa; [b] statistično značilno pri p<0,001; [c] ni statistično značilno).

modelirana popačena tipizacija	Cramerjev koeficient	koeficient kappa (%)
modelirana popačena tipizacija z metodo *k* najbližjih sosedov	0,009[b]	0,5[b]
modelirana popačena tipizacija z metodo največje verjetnosti	0,005[c]	0,0[c]
modelirana popačena tipizacija z metodo najmanjše razdalje (12 skupin!)[a]	0,005[c]	0,1[c]

Preglednica 11: Ujemanje modeliranih popačenih tipizacij z osnovno popačeno tipizacijo TIPI13.

modelirana popačena tipizacija	delež enako klasificiranih celic od vseh celic (%)	delež enako klasificiranih učnih celic (%)
modelirana popačena tipizacija z metodo *k* najbližjih sosedov	9,0	17,0
modelirana popačena tipizacija z metodo največje verjetnosti	10,0	8,5
modelirana popačena tipizacija z metodo najmanjše razdalje (12 skupin!)	11,2	8,1

3.3 Primerjava rezultatov s predhodnimi raziskavami

Ob primerjavi rezultatov modeliranj popačenih tipizacij z rezultati modeliranj, ki so bili dobljeni z modeliranjem izvirnih (nepopačenih!) tipizacij, lahko ugotovimo, da so bila ujemanja pri modeliranju izvirnikov precej večja. Pri modeliranju TIPI9 na temelju 3,6 % učnega vzorca, štirih pojasnjevalnih podatkovnih slojev in različnih klasifikacijskih metod so prav vsa orodja uspela določiti pravila, ujemanje med izvirno tipizacijo in njenim modelom pa je bilo od 51 % (metoda najmanjše razdalje) do 75 % (metoda *k* najbližjih sosedov) (Ciglič 2014; Ciglič in Perko 2015). Pri modeliranju TIPI9 na temelju 1,0 % učnega vzorca, šestih pojasnjevalnih podatkovnih slojev in metode naključnih gozdov je bilo doseženo ujemanje 94 % (Ciglič in sodelavci 2017). Pri modeliranju TIPI13 na temelju 5,1 % učnega vzorca, štirih podatkovnih slojev in različnih klasifikacijskih metod so pravila uspela določiti prav vse metode, ujemanje med izvirno tipizacijo in njenim modelom pa je bilo od 47 % (metoda najmanjše razdalje) do 69 % (metoda *k* najbližjih sosedov) (Ciglič 2014). Pri modeliranju TIPI13 na temelju 100,0 % učnega vzorca, različnih kombinacij podatkovnih slojev in metode odločitvenega drevesa je bilo ujemanje med izvirno tipizacijo in njenim modelom od 71,2 % do 79,5 % (Ciglič 2012). V vseh primerih gre torej za precej višje ujemanje, kot pri modeliranju popačenih klasifikacij. Klasifikacijska pravila so izdelala prav vsa orodja, kar pomeni, da izvirni klasifikaciji TIPI9 in TIPI13 nista naključni in sta narejeni s precejšnjo mero objektivnosti, saj ju lahko potrdimo z različnimi kvantitativnimi modeli.

Modeliranje popačenih tipizacij je pokazalo, da nekatere metode (razni algoritmi odločitvenih dreves) ne morejo pripraviti klasifikacijskih pravil, če so vhodni podatki nesmiselni oziroma nelogični.

Nekatere metode (najmanjša razdalja, največja verjetnost, *k* najbližjih sosedov) so sicer uspele pripraviti modelirano popačeno tipizacijo, a je bila ta povsem drugačna od svojega izvirnika.

To pomeni, da lahko z metodami nadzorovane klasifikacije modeliramo pojave (v tem primeru pokrajinske tipe), za katere imamo dovolj kakovostne podatke. V nasprotnem primeru velja pravilo »*garbage in, garbage out*«, kar pomeni, da s slabimi podatki (vnos v model) ne moremo pričakovati smiselnih oziroma uporabnih rezultatov (izvoz iz modela). Zaradi omenjenih ugotovitev je treba biti pri rabi klasifikacijskih metod previden in je treba rezultate preveriti na različne načine.

4 Sklep

Prispevek obravnava sposobnost napovedovanja nekaterih metod nadzorovane klasifikacije. Želeli smo preveriti, kakšne rezultate podajo različne metode nadzorovane klasifikacijske, če z njimi modeliramo naključno oziroma nerazumsko zasnovano klasifikacijo. Za ta namen smo popačili dve naravnopokrajinski tipizaciji Slovenije ter skušali zanju določiti smiselna klasifikacijska pravila z različnimi metodami.

Ugotovili smo, da nekatere metode (odločitvena drevesa) naključnih oziroma nerazumskih klasifikacij ne morejo potrditi niti v najmanjši meri, saj algoritem ni sposoben najti logičnih klasifikacijskih pravil in čistih skupin (tipov). Nekatere metode sicer uspejo modelirati klasifikacijo, a le v tolikšni meri, da lahko celotno ujemanje med izvirnikom in modelom pripišemo zgolj naključju. Pomembno je opozoriti, da nekatere metode, ki smo jih preverjali (predvsem metoda najmanjše razdalje), proizvedejo vizualno precej smiselne naravnopokrajinske tipe, a so ti posledica »vsiljevanja« lastne strukture same metode, ki poda povsem drugačno tipizacijo od izvirnika. Zaradi omenjenih ugotovitev je priporočljivo rezultate preveriti na več načinov. V obeh sklopih modeliranja popačenih tipizacij (TIPI9 in TIPI13) smo prišli do podobnih ugotovitev. Popačeni tipizaciji lahko objektivno označimo kot neustrezni, saj:

- po nekaterih metodah nismo uspeli izdelati modela zaradi nesmiselnosti popačene tipizacije,
- se izdelane modelirane popačene tipizacije, kljub nekaterim relativno dobrim ocenam z vidika povezovanja s pojasnjevalnimi podatkovnimi sloji, niso ujemale z osnovno popačeno tipizacijo.

Za preučevanje lastnosti geoinformacijskih orodij oziroma spoznavanje njihovega delovanja, je preverjanje modeliranja popačenih klasifikacij lahko v pomoč, saj uporabnik dobi informacijo o tem, katera metoda nadzorovane klasifikacije bolj vsiljuje svojo strukturo in katera metoda se bolj prilagodi učnim podatkom (čeprav slabim oziroma popačenim). Tako lahko uporabnik, ki meni, da ima reprezentativno podatkovno bazo, uporabi metode, ki se bolj prilagodijo učnim podatkom, v nasprotnem primeru pa mora preizkusiti različne metode in primerjati rezultate.

Zahvala: Raziskavo je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije v okviru raziskovalnega projekta »Napredek računsko intenzivnih metod za učinkovito sodobno splošnonamensko statistično analizo in sklepanje« (L1-7542) in raziskovalnega programa »Geografija Slovenije (P6-0101)«.

5 Viri in literatura

Glej angleški del prispevka.