INFORMATION VALUES OF ABSOLUTE ELEVATION AND ELEVATION DIFFERENCE FOR ILLUSTRATION OF THERMAL BELT

INFORMATIVNI VREDNOSTI NADMORSKE VIŠINE IN VIŠINSKE RAZLIKE ZA PONAZORITEV TERMALNEGA PASU

Rok Ciglič

BOJAN ERHARTIČ

Vineyards near Jeruzalem.
Vinogradi pri Jeruzalemu.

Information values of absolute elevation and elevation difference for illustration of thermal belt

DOI: 10.3986/AGS50201
UDC: 913(497.4):531.717.3
551.581:531.717.3(497.4)
COBISS: 1.01

ABSTRACT: This paper estimates the information gain and the information gain ratio, which are usually used in machine-learning processes, to assess which data layer – absolute elevation or elevation difference – better reflects the topoclimatic characteristics (especially the thermal belt). Both attributes are compared based on their information value in explaining the locations of vineyards, which depend largely on the thermal belt. The analysis is performed on 9,000 cells covering various winegrowing districts. In general, elevation difference proves to be a better attribute, but certain differences can be observed between individual areas, especially between the continental and submediterranean parts of Slovenia.

KEYWORDS: geography, information gain, gain ratio, thermal belt, elevation difference, absolute elevation, Slovenia

The article was submitted for publication on February 2, 2010.

ADDRESS:
Rok Ciglič
Anton Melik Geographical Institute
Scientific Research Centre of the Slovenian Academy of Sciences and Arts
Gosposka ulica 31, SI – 1000 Ljubljana, Slovenia
E-mail: rok.ciglic@zrc-sazu.si

Contents

1 Introduction

Defining natural geographical divisions using geographical information systems and quantitative methods requires precise spatial data that contain as much information as possible. For areas for which the desired data are unavailable, other data must be used. Due to the correlation between temperature and absolute elevation (Bailey 1996, 68; Pezzi, Ferrari in Corazza 2008, 452), a digital elevation model can be used to illustrate temperatures. However, in doing so attention must be paid to the lowest, concave areas and the thermal belt lying above them, which make it impossible to draw a simple linear connection between absolute elevation and specific climatic characteristics (e.g., temperature), especially when conducting research on a larger scale.

Elevation difference complements the information on absolute elevation and makes it possible to approximately assess the limits between the lake of cold air and the thermal belt because these are already roughly known (Gams 1996; Žiberna 1999; Ogrin 2007). If a narrower complete area with a uniform bottom is studied, there are no differences between these two attributes. However, when comparing large number of separate concave forms at various absolute elevations, it is presumed that the information on elevation difference is more relevant (Figure 1).

Based on vineyard locations, which reflect the climatic phenomena described above (Ogrin 2007), it was determined which attribute (i.e., absolute elevation or elevation difference) provides a greater quantity of information. This quantity was established using the information gain and the gain ratio. These two measures are generally used in machine learning methods and increasingly more often in environmental sciences (Džeroski 2002). Information gain as an estimate of an attribute's importance was introduced by Hunt et al. in 1966 (Kononenko 1994, 171). In this study, several Slovenian winegrowing areas were selected in order to determine whether there are differences between the explanatory power for vineyard locations with elevation difference and the explanatory power for vineyard locations with absolute elevation, and at what scale this difference is most pronounced.

2 Climate and relief

In Slovenia, relief factors are of above-average importance to the climate. The roughness of the relief is important especially for local and micro-climatic conditions. Relief influences the topoclimatic conditions (those involving modification of climatic conditions due to relief) primarily in terms of the elevation structure, aspect, slope, and the type of surface, in which concavity plays an especially important role (Ogrin 2000).

Concave relief forms that can keep the air close to the basin floor from being mixed by the ambient flow once a stable stratification has developed (Whiteman et al. 2004, 1232) provide favorable relief conditions for the formation of temperature inversion and a lake of cold air. Its height depends on the elevation of the relief that surrounds a depression (usually between 50 and 200 m). Above the lake of cold air lies a thermal belt, which is warmer than the adjacent lower or higher areas (Ogrin 2000). This is why the minimum temperature at hill stations is 2 °C higher and the average temperature 1 °C than at stations located

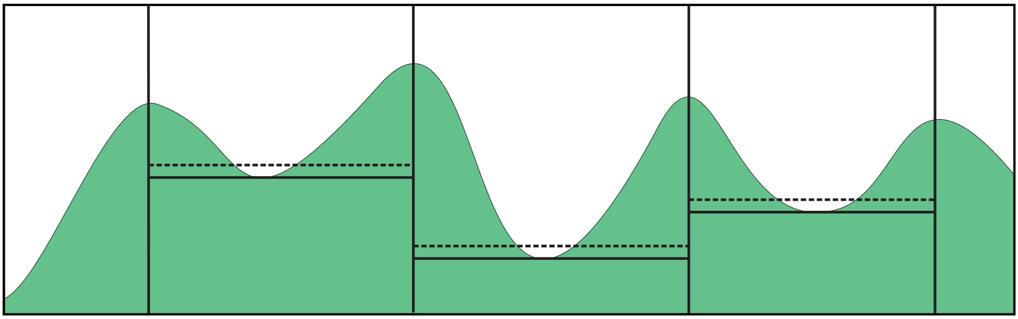


Figure 1: Basin bottom elevations (continuous line) and thermal belt lower limits (dashed line) differ between individual concave areas, whereas the elevation difference (and thus also the difference between both lines) is about the same.

at the same absolute elevation in basins (Gams 1996). The boundaries of the thermal belts, which also occur along the coast, are different in various areas of Slovenia (Ogrin 2000; 2007; Žiberna 1999). During the winter lake of cold air can persist several days (Vrhovec 1991, 91).

2.1 Determining the thermal belt

In some regions it is impossible to determine the thermal belt through climatic measurements, but it can be determined through profile measurements of minimum temperatures in individual weather situations, and through detailed cartography of habitats of temperature-sensitive crops (Ogrin 2007).

Using data from meteorological stations, the reference ceiling for all of Slovenia can be determined. The average minimum temperatures indicate the thermal belt limit at a elevation difference of approximately 500 m, and the average annual temperature indicates it at 200 to 250 m (Gams 1996; Žiberna 1999; Ogrin 2007). Vineyards indicate a thermal belt lower limit at 15 to 30 m above the basin bottom and an upper limit at an elevation of approximately 450 to 550 m above sea-level (Žiberna 1992, cited in Ogrin 2007).

Žiberna (1992) and Ogrin (2007) delimited the thermal belt by the minimum and maximum elevation of vineyards, in which the minimum elevation is also the average elevation of the lake of cold air in the warm half of the year or at the beginning of vine growth. Ogrin drew attention to the fact that vineyard locations are not always rational. The thermal belt determined in this way mainly reflects climatic changes at the beginning of vine growth (i.e., at the end of April and in May), when frost risk is greatest. In summer, the vertical range of the thermal belt is larger, and in winter it is smaller. This is a dynamic phenomenon, which also depends on weather conditions (Ogrin 2007). There is thus no fixed limit, but the influence of the frequent formation of the lakes of cold air in Slovenia can be observed in the annual averages of minimum and average temperatures. The only exception to this rule is the coastal plains (Gams 1996).

The density of vineyard areas decreases with absolute elevation (Hrvatin and Perko 2003, 43). In addition to absolute elevation, elevation difference is also connected with the share of vineyard areas. The intensity of coincidence between vineyard areas and elevation difference varies among vineyard regions (Žiberna 1992). This raises the question of which factor (i.e., absolute elevation or elevation difference) is more connected with vineyards within a broader area because vineyard regions have different absolute elevations.

2.2 Problems in determining the limits of the thermal belt and lake of cold air based on vineyards

There are a few limits to the method of determining the range of the thermal belt based on vineyards (Ogrin 2007). It is primarily useful in regions with an old winegrowing tradition and in regions with more rough relief (Ogrin 2007). However, it must be taken into account that, despite their favorable climate, not all the regions are used for winegrowing because the geological and soil conditions may be less favorable (Žiberna 1992). With statistical methods has been proved that several natural factors influence vineyards location (Watkins 1997; Hrvatin, Perko and Petek 2006).

3 Preparing the data for analysis

Vineyards are a cultivation type that is especially associated with the thermal belt. Due to this feature, vineyard locations were selected in order to determine which data better show topoclimatic properties: absolute elevation or elevation difference.

This paper primarily focuses on areas in which it is simpler to determine the thermal belt and the lake of cold air. This means it focuses on areas with more rough relief within winegrowing regions that have traditionally been engaged in winegrowing.

In order to select several different areas, the classification into winegrowing regions (i.e., the Primorska, Posavje, Podravje) and districts was used (Natek 1998, 209). Three districts were selected from each region, totaling nine districts (Table 1, Figure 4). In this way, the importance of elevation difference at the national (winegrowing) level can be compared with that at the level of an individual region. It is also possible to compare the continental and submediterranean parts of Slovenia in this regard.

Table 1: Selected areas.

Primorje winegrowing region districts	Posavje winegrowing region districts	Podravje winegrowing region districts
Goriška Brda	Šmarje-Virštanj	Radgona-Kapela
Kras	Dolenjska	Maribor
Koper	Bela krajina	Ljutomer-Ormož

A small sample area was selected in each district that opens up to and connects to the bottom of a major concave area (e.g., a part of polje). The only exception was Bela krajina, where part of a broad slope above a karst plain was selected.

The range of an individual sample area within a winegrowing district, from which the cells were captured, was determined using ArcGIS 9.3 software (Figure 2). Using a set of commands and according to the digital elevation model (DEM) individual drainage basins were determined that simultaneously encompass complete relief unit with a uniform bottom (the range of a valley or polje from its bottom to rim).

The *FILL* command was used to fill all depressions with a depth less than 3.2 m, which is the same as the average error in the DEM (Digitalni modeli višin 2007). This prevented the runoff area from being determined separately and too high above the actual bottom due to a potential error in the DEM or due to an actual minor depression.

At their lowest points, the selected basins' absolute elevations are the same as those at the bottom of the nearest major concave form or plain (Figure 4). The elevation difference was determined by subtracting the absolute elevation at the bottom of the nearest major complete concave form from the absolute elevation of individual cells.

The bottoms of the concave forms were determined based on concavity and inclination. The concave relief was determined by comparing the original and smoothed DEMs, a method already used by Podobnikar and Šprajc (2007). Cells with a slope inclination below 2° were classified as plain (Perko 2001). The entire procedure is presented in Figure 3. A DEM with a 25 m resolution was used in the study (Hrvatin and Perko 2005; GURS 2009).

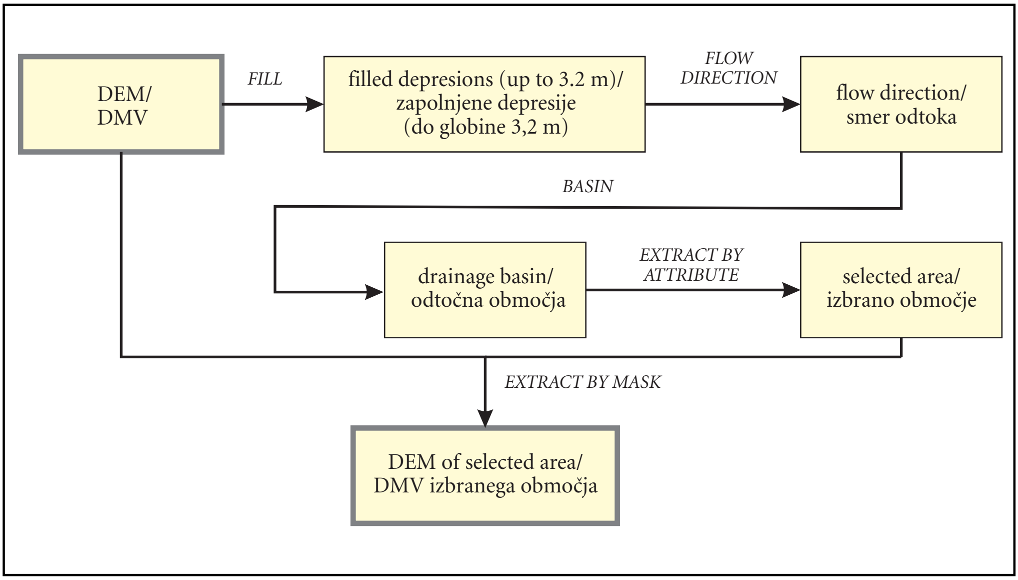


Figure 2: Determining the range of individual area within a winegrowing district.

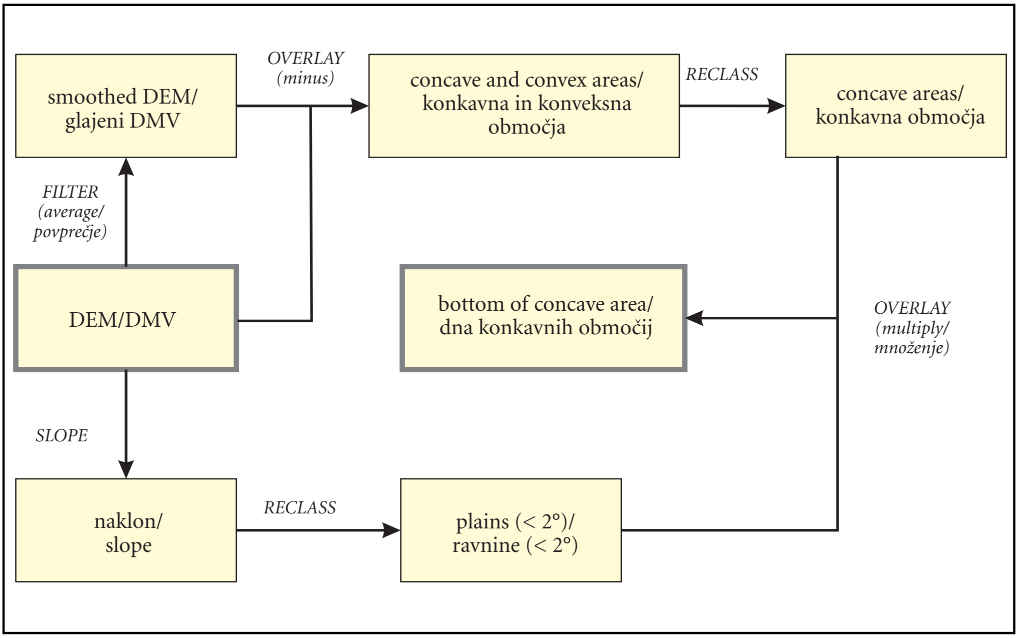


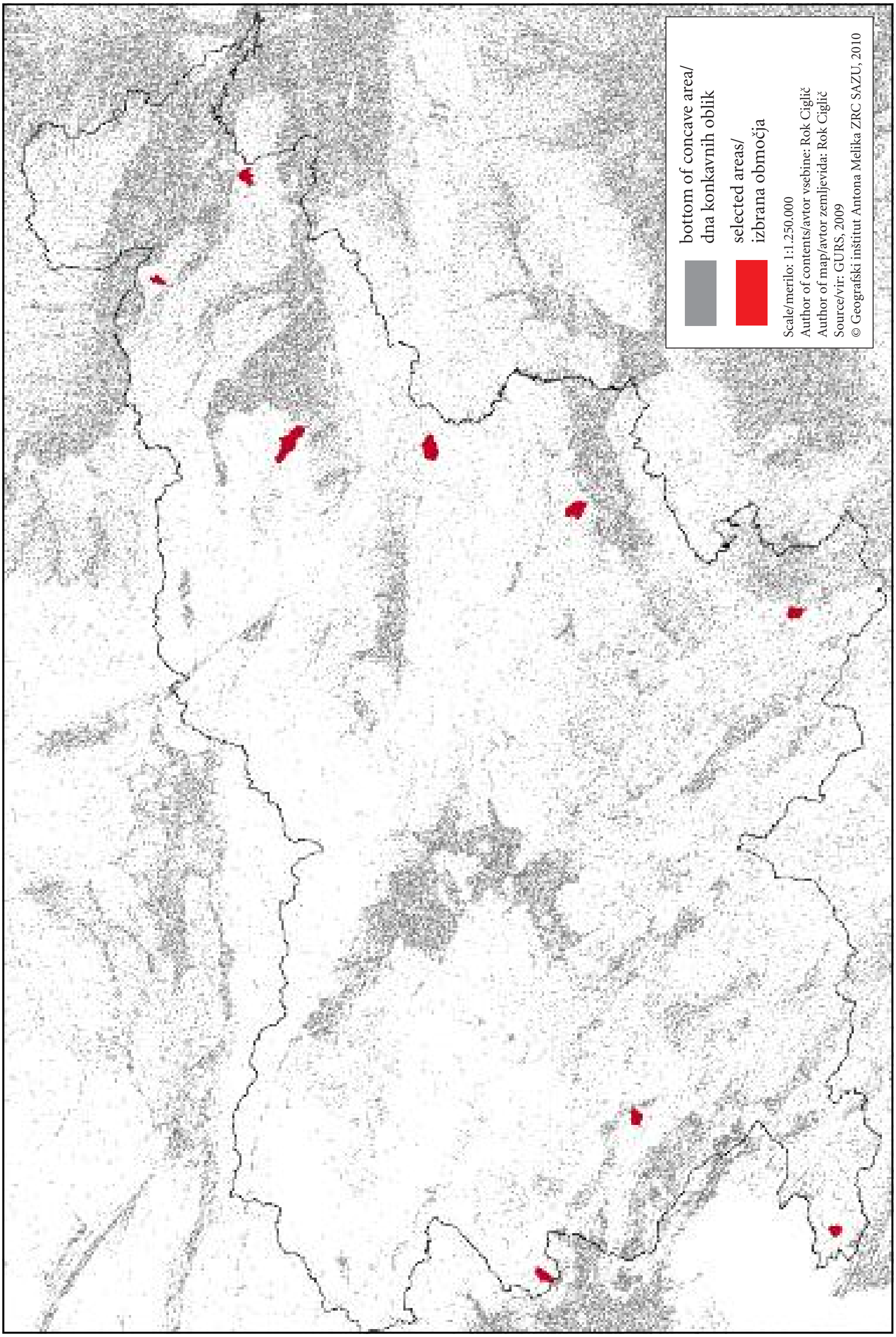
Figure 3: Determining the bottoms of concave forms using a DEM.

Five hundred cells with vineyards and 500 cells outside the vineyards were then selected in each sample area (the example of the Ljutomer-Ormož Hills is presented in Figure 5), which means that 1,000 cells were selected for each winegrowing district, thus 3,000 cells for each winegrowing region, or 9,000 altogether. The ratio between vineyard areas and the cells without vineyards is rarely the same – this is only the case in the Goriška brda (Table 2). In our case, we were thus forced to randomly select 500 cells with vineyards and 500 cells without vineyards. In this way, this characteristic of the region (i.e., the share of vineyards) was distorted, but this was the only way in which the general characteristics (absolute elevations) of vineyards within individual areas were preserved and at the same time suitably compared among one another. If the number of cells had been selected according to the shares of the area of individual regions, this would have led to incorrect results because some regions with a larger share of vineyards would have increased the influence of its vineyards' absolute elevation and elevation difference. The similar solution was also used by Saito, Nakayama, and Matsuyama (2009) in comparing various samples of data layers showing the presence of landslides.

Table 2: Total number of cells and share of vineyards.

Winegrowing district	Total number of cells	Share of vineyards (%)
Goriška brda	9,770	52.1
Kras	9,370	27.3
Koper	6,836	11.9
Šmarje-Virštanj	17,960	3.9
Dolenjska	14,774	10.7
Bela krajina	8,530	9.3
Radgona-Kapela	5,396	19.0
Maribor	26,568	6.1
Ljutomer-Ormož	10,463	27.0
(Total)	109,667	15.5

Figure 4: Locations of the bottoms of concave form and selected areas. ➤



bottom of concave area/
dna konkavnih oblik
selected areas/
izbrana območja
Scale/merilo: 1:1.250.000
Author of contents/avtor vsebine: Rok Ciglič
Author of map/avtor zemljevida: Rok Ciglič
Source/vir: GURS, 2009
© Geografski inštitut Antona Melika ZRC SAZU, 2010

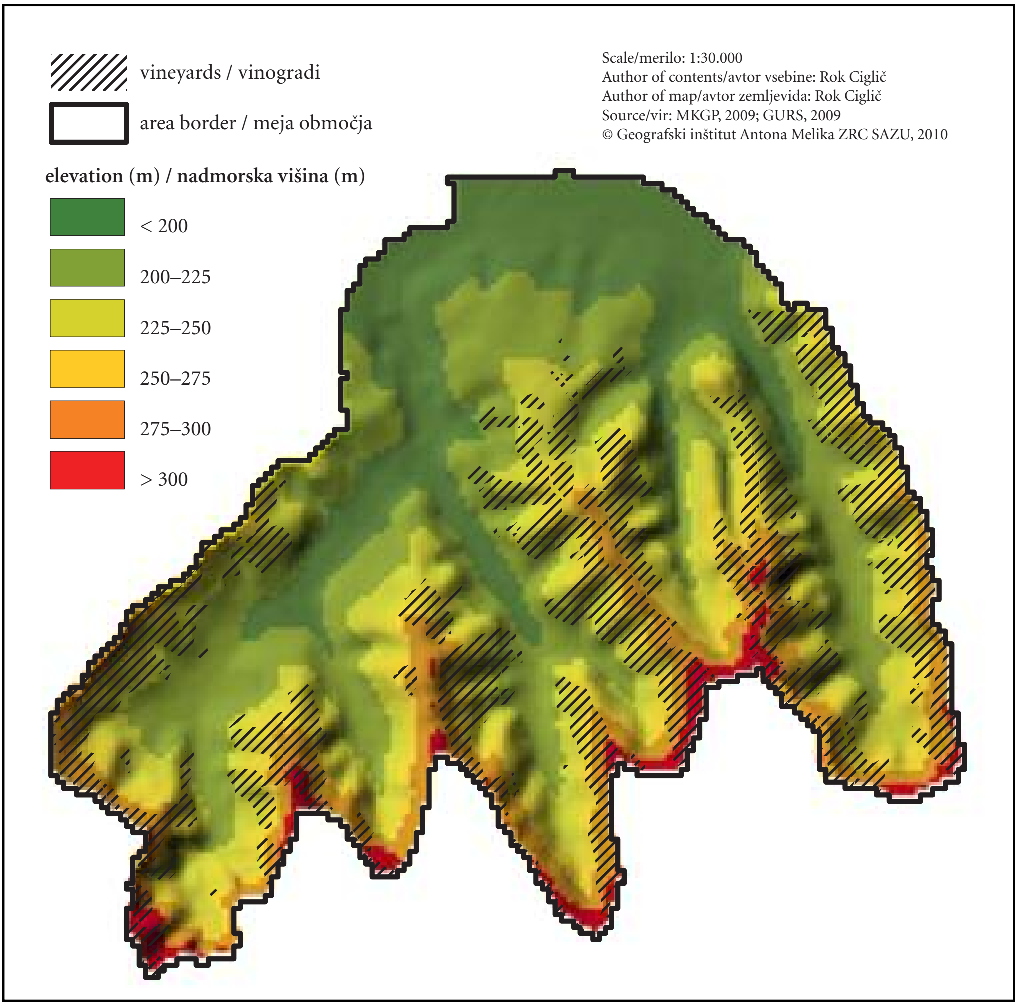

vineyards / vinogradi
area border / meja območja
elevation (m) / nadmorska višina (m)
< 200
200–225
225–250
250–275
275–300
> 300
Scale/merilo: 1:30.000
Author of contents/avtor vsebine: Rok Ciglič
Author of map/avtor zemljevida: Rok Ciglič
Source/vir: MKGP, 2009; GURS, 2009
© Geografski inštitut Antona Melika ZRC SAZU, 2010

Figure 5: Example of an area selected in order to select cells (Ljutomer-Ormož Hills).

According to the borders of the winegrowing areas presented in *Geografski atlas Slovenije* (Natek 1998), and the information on land use provided by the Ministry of Agriculture, Forestry, and Food (Raba tal 2009), vineyards account for approximately 3% of the total area of winegrowing regions.

This paper compares several areas, the majority of which have different ranges of values with regard to elevation difference and absolute elevations. Because this influences the value of the information gain, it is necessary to make an additional calculation of the information gain ratio, which can reduce this influence and further support the result (Tables 3 and 4).

Table 3: Value range of the 9,000 cells used in the study.

Winegrowing area	Absolute elevation (m)	Value range (m)	Elevation difference (m)	Value range (m)
Primorje (3,000 cells)	45–426	381	0–332	332
Posavje (3,000 cells)	141–568	427	0–427	427
Podravje (3,000 cells)	179–847	668	0–582	582
Total	45–847	802	0–582	582

Table 4: Range of attribute values.

	Absolute elevation (m)	Value range (m)	Elevation difference (m)	Value range (m)
Goriška Brda	57–260	203	0–203	203
Kras	94–429	335	0–335	335
Koper	44–275	231	0–231	231
Primorje winegrowing region	44–429	385	0–335	335
Šmarje-Virštanj	200–432	232	0–232	232
Dolenjska	171–469	298	0–298	298
Bela Krajina	141–588	447	0–447	447
Posavje winegrowing region	141–588	447	0–447	477
Radgona-Kapela	222–341	119	0–119	119
Maribor	265–849	584	0–584	584
Ljutomer-Ormož	179–337	158	0–158	158
Podravje winegrowing region	179–849	670	0–584	584
All regions together	44–849	805	0–584	584

4 Methods for calculating attribute significance

The information gain and gain ratio are measures of the significance of an attribute (Kononenko 2005), a variable, or data layer (this paper uses the term 'attribute'). These two measures are used together with others primarily to direct and control the hypothesis search in machine learning algorithms. In doing this search, the basic task of such an algorithm is to assess the significance of an attribute for a given learning problem (Kononenko 2005).

This method is thus used for determining, on the basis of the values of an existing attribute (e.g., in the land-use category), which of the remaining attributes (e.g., absolute elevation or bedrock type) best explains these values. On the one hand, there is the target or predicted attribute, and on the other several explanatory attributes. For example, information theory approach was used for determination of the most suitable classification of rivers according to the pollution of sediments (Kraft, Einax and Kowalik 2004).

The measures for determining attribute significance described above are based on the information value. This is the value required to determine the outcome of an event. It is defined as a minus binary logarithm of the event probability (Shannon and Weaver 1949; Kononenko 2005) and expressed in bytes.

Information value: $I(X_i) = -\log_2 P(X_i)$

The anticipated average information value (Witten and Frank 2005) required to find out which of the incompatible outcomes X_i ($i = 1 \dots n$, $\sum_\mathrm{i} P(X_i) = 1$ byte) took place is referred to as entropy of an event (Kononenko 2005, 174), expressed with the following equation:

Event entropy: $H(X) = -\sum_\mathrm{i} P(X_i) \log_2 P(X_i)$

in which X_i denotes the event and $P(X_i)$ denotes the probability of event X_i.

4.1 Calculating the information gain and gain ratio

Prior to explaining the information gain and gain ratio calculation in greater detail, it should be mentioned that the term »class« and not »value« is used for the value/quantity of the attribute predicted. In the analytical part of this paper, the predicted attribute is the vineyard location, which contains two classes: 0 and 1, or 'there is a vineyard' and 'there is no vineyard.' The term 'value' is used for all the other attributes that explain the predicted one. In our case, both attributes (i.e., the elevation difference and absolute elevation) have values ranging from zero to several hundred meters.

Attribute information gain or Gain(*A*) is defined as the attribute's contributive information. It is calculated by subtracting the conditional entropy of a class at a given attribute value ($H_{R|A}$) from the class entropy (H_R; Kononenko 2005):

$$\text{Gain}(A) = H_R - H_{R|A}$$

in which $\text{Gain}(A) \geq 0$, and max. $\text{Gain}(A) = H_R$.

The information gain ratio or GainR(*A*) eliminates the deficiency of the information gain. According to information gain the attribute's quality increases with the number of possible values. The overestimation of multivalue attributes is eliminated by normalizing the information gain through the entropy of the attribute's value (Kononenko 2005):

$$\text{GainR}(A) = \text{Gain}(A) \mathbin{/} H_A$$

At the same time, GainR(*A*) sometimes favourise attributes with low intrinsic information, which is why both methods of calculation must be taken into account in the final estimation (Witten and Frank 2005).

Table 5 presents the data and equations required to calculate the information gain and gain ratio (taken from Kononenko 2005).

Table 5. Data, ratios, and equations required to calculate information gain and gain ratio (Kononenko 2005).

Data:

n – number of training examples (number of all units or cells)
n_k – number of training examples from class r_k (number of units in class k of predicted attribute r)
n_j – number of training examples with a j value of a given attribute A_i (number of units with value j of explanatory attribute A_j)
n_{kj} – number of training examples from class r_k with a j value of a given attribute A_i (number of units in the class of the predicted attribute with one of the explanatory attribute values)

Individual data ratios:

$p_{kj} = n_{kj}/n$
$p_k = n_k/n$
$p_j = n_j/n$
$p_{k|j} = p_{kj}/p_j = n_{kj}/n_j$

Equations:

$H_R = -\Sigma_k p_k \log p_k$: class entropy (entropy of the predicted attribute)
$H_A = -\Sigma_j p_j \log p_j$: entropy of the attribute (entropy of the explanatory attribute)
$H_{R|A} = H_{RA} - H_A = -\Sigma_j p_j \Sigma_k p_{k|j} \log p_{k|j}$ ($H_{RA} \geq H_{R|A}$): conditional class entropy at a given attribute value
$H_{RA} = -\Sigma_k \Sigma_j p_{kj} \log p_{kj}$: entropy of the product of class-value attribute events

Some algorithms used for classifying and sorting into groups can only process descriptive variables. In such cases, it is thus necessary to make the numeric variables mathematically discrete or divide them into intervals (Witten and Frank 2005), which must also be done in calculating the information gain of the numeric attributes.

There are several discretization methods (Witten and Frank 2005); the one used in estimating the significance of the numeric attribute in the software we used – Weka 3.5.8. (Hall et al. 2009) – is based on entropy and takes into account the principle of minimum description length (MDL). By dividing units into intervals, the type of division is sought that renders the groups as »clean« as possible in terms of the classes they contain (Witten and Frank 2005).

In this procedure, it is first established which division of the numeric values into two groups according to the attribute selected provides the biggest information gain. When the division point is set, based on which the units are classified, the procedure is repeated for values above and below the division point selected (i.e., the process continues in both directions). Without a set criterion, this process can continue until all of the units are divided in such a way that the unit groups are clean. This does not make sense because the findings would be overly adapted to the training sample and it would be more difficult to generalize the result; therefore, criteria for halting the discretization process must be introduced. It is thus verified for each division whether the information gain (of this division) is sufficiently large given the number of units N, the number of classes k, the entropy of units E, the entropy of units in each subinterval (following the division) E_1 and E_2, and the number of classes in each subinterval (following the division) k_1 and k_2 (Fayyad and Irani 1993; Witten and Frank 2005):

$$gain > \frac{\log_2(N-1)}{N} + \frac{\log_2(3^k-2)-kE+k_1E_1+k_2E_2}{N}$$

To illustrate the significance of the values of the information gain calculated, the values of the information gain (Figure 6) was calculated for various explanatory attributes A_x by taking into account an invented predicted attribute C (with two possible classes: 0 and 1; both classes have equal number of instances).

Table 6: Values of attributes C and A_x.

C	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}
0	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
0	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No
0	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No
0	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	No
0	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No
1	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No	No
1	Yes	Yes	Yes	Yes	No	No	No	No	No	No	No
1	Yes	Yes	Yes	No	No	No	No	No	No	No	No
1	Yes	Yes	No	No	No	No	No	No	No	No	No
1	Yes	No	No	No	No	No	No	No	No	No	No
Probability of event »Yes« (%)	100	90	80	70	60	50	40	30	20	10	0

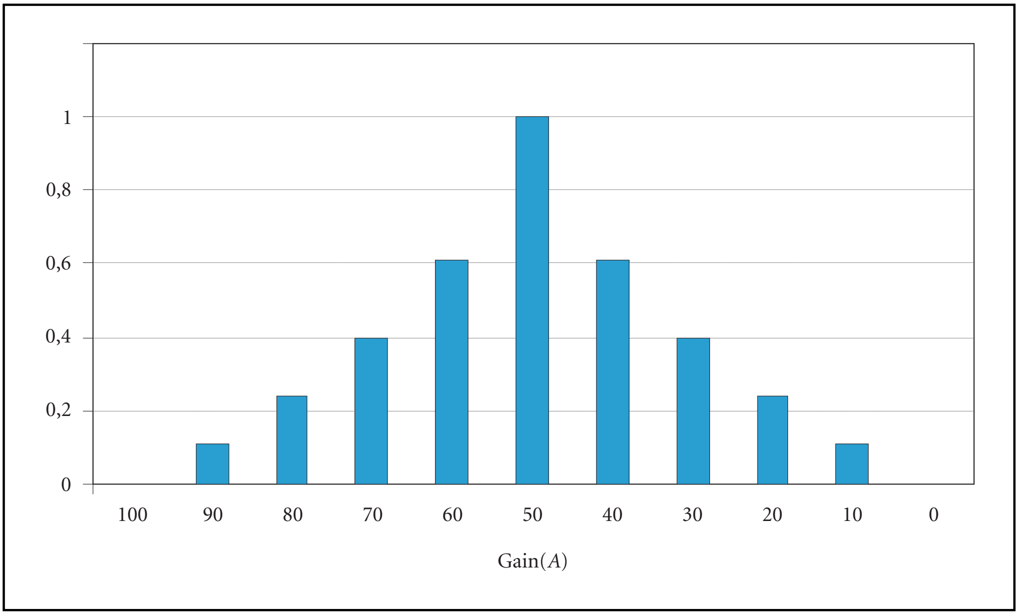


Figure 6: Gain(A) for the explanatory attributes. Gain(A) is expressed in bytes and the probability of event »Yes« ($p(A)$) is expressed in percentages.

These attributes are binary, like attribute *C*. In this, it must be noted that the values of attributes *A* and *C* for individual units match perfectly, which means that 0-values match the Yes-values, and 1-values match the No-values to the greatest possible extent (Table 6). Figure 6 shows that the information gain is the biggest where the match between the predicted and explanatory attributes is complete and falls towards the point at which the explanatory attribute has only one value and thus cannot provide any information.

5 Calculating information gain and gain ratio of absolute elevation and elevation difference according to vineyard locations

Using Weka software the information gain and gain ratio were calculated for the explanatory attributes *absolute elevation* and *elevation difference* according to the predicted attribute *vineyards*. 500 cells with vineyards and 500 cells outside vineyards were taken from each vineyard district. Then the data from individual districts were combined by winegrowing regions, and finally the data for all regions were also combined. In this way, the information gain and gain ratio were calculated for individual winegrowing regions and the total Slovenian winegrowing area. In addition, measures were also calculated for the continental Slovenian region (i.e., Posavje and Podravje together). Through this it can be established whether there are differences in results when comparing locations within individual winegrowing regions and all the locations in the total winegrowing area.

In calculating both of these measures, the 10-fold cross-validation method was used (Kirkby and Frank 2010). This means that the units were divided into ten groups or 'folds' and the information gain and ratio were calculated ten times.

The information gains (Table 7) and gain ratios (Table 8) calculated show elevation difference proves to be a more significant attribute nearly in all cases because higher values mean greater quantities of information.

Table 7: Information gain values. Higher values are in bold.

Calculated based on:	Absolute elevation	Elevation difference
Posavje winegrowing region cells	0.169 (±0.006)	**0.251 (±0.005)**
Podravje winegrowing region cells	0.309 (±0.004)	**0.404 (±0.004)**
Primorje winegrowing region cells	**0.016 (±0.003)**	0.012 (±0.002)
Continental Slovenia cells	0.212 (±0.003)	**0.283 (±0.004)**
Total	0.104 (±0.002)	**0.127 (±0.002)**

Table 8: Information gain ratio values. Higher values are in bold.

Calculated based on:	Absolute elevation	Elevation difference
Posavje winegrowing region cells	0.079 (±0.011)	**0.117 (±0.006)**
Podravje winegrowing region cells	0.116 (±0.001)	**0.172 (±0.009)**
Primorje winegrowing region cells	0.032 (±0.003)	**0.034 (±0.010)**
Continental Slovenia cells	0.079 (±0.006)	**0.118 (±0.006)**
Total	0.042 (±0.001)	**0.067 (±0.001)**

By comparing the calculated values with the information gain values presented in Figure 6, the values of the information gain in the analysis described can be better understood. The elevation difference in the Podravje winegrowing region has the highest information value (0.404 byte). On the other hand, values around 0.1 byte have extremely low information value; they are at the level of a binary attribute, whose values are 90% identical and do not provide a great deal of information.

Vineyard locations can be better described by elevation difference than absolute elevation, although values generally vary by winegrowing area. By comparing the results, it can be seen that in continental Slovenia vineyards are found at more specific locations or at specific elevations above the basin or polje bottoms as there is a greater importance of elevation difference. There is also evident higher discrepancy between the significance of elevation difference and absolute elevation, which could also result from the fact that the thermal belt and lake of cold air are more pronounced in continental Slovenia. The same applies to absolute elevation, although it has little information significance everywhere.

In the Primorje winegrowing region, the significance of elevation difference and absolute elevation is the smallest among all areas; however, with regard to the information gain, absolute elevation is even slightly more significant, which is an exception in this analysis. Both measures prove that the thermal belt may be less pronounced in this region or that the formation of the lake of cold air is rare, which provides generally better climatic conditions even at lower elevation difference. Good physical-geographical conditions enable some regions to largely specialize in winegrowing and allocate most of their land to vineyards (e.g., Goriška brda). Together with some other areas (e.g., the Koper, Bilje-Vrtojba, and Vipava hills), Goriška brda belong to the areas with the most favorable natural conditions, which is also reflected in the high concentration and terracing of vineyards (Ažman Momirski and Kladnik 2009).

With regard to the broadest area (i.e., all the winegrowing regions together), elevation difference and absolute elevation explain the vineyard location (and consequently primarily the topoclimatic conditions) to a smaller extent than for individual winegrowing regions or the total winegrowing area of the continental Slovenia. The extremely low information value of both layers equalizes the significance of both layers in small-scale analyses. This result is contrary to our expectations because one would expect the significance of elevation difference in explaining vineyard locations to increase with the increased number of cells from various areas (at different absolute elevations). There are several possible reasons for this: (1) the vineyards are actually not equally divided across all slopes, which can also be a result of the fact that the thermal belt or lake of cold air boundaries are not that uniform in Slovenia (as already mentioned in the introduction), (2) the cell sample was not sufficiently adequate and it would be prudent to repeat the analysis, or (3) other factors may influence vineyard locations. This last possibility was not of particular interest in this paper because the main goal was to only compare the data on elevation difference and absolute elevation. In the introduction we mentioned that relief roughness is important especially for local and micro-climatic conditions (Ogrin 2006, 126), which is supported by these results.

6 Conclusion

With the aid of a suitable database – which contains data on the predicted attribute and several explanatory attributes or data layers – this method may prove useful when faced with the dilemma of what data layers to use for analysis. This method can be used to establish which data provide the most information in explaining a specific feature. A further advantage of this method is that it makes it possible to compare nominal and numeric attributes.

This paper uses vineyard locations as indicators of topoclimatic features or, more precisely, the thermal belt, and data on absolute elevation and elevation difference as explanatory attributes. The main goal was thus to establish which attribute can be used to better explain or describe the local topoclimatic conditions. Due to geographically extremely diverse areas, and their indirect selection through vineyards, the explanatory power (or information value) of both attributes is very low; nonetheless, a comparison of both attributes makes it possible to draw certain conclusions about the features of vineyard locations and subsequently thermal belt. It was confirmed that elevation difference is usually more significant for explaining the vineyard locations and thus indicated that the thermal belt really exists. Among the conclusions, it can be highlighted that elevation difference plays a more important role with regard to vineyard locations in Slovenia's inland winegrowing regions than in its submediterranean regions; this could lead to the conclusion that the formation of the lake of cold air and thermal belt is more pronounced in continental area. According to the results obtained by comparing all of the cells (low values of both measures), it can also be concluded that, at a small scale, the locations of vineyards and thus the thermal belt cannot be as successfully demonstrated with elevation difference and absolute elevation as at a larger scale. It was expected that by taking into account all of the cells (of all winegrowing areas together) elevation difference would be even more significant. Results may also be the result of the fact that vineyard locations – and thus perhaps also thermal belt boundaries – actually vary so evident by areas, or the fact that the cell sample was inappropriate. It would be prudent to repeat the analysis on a larger sample by including more areas. It is also possible to include more attributes and to perform the analysis using the RELIEF method (Kononenko 1994), which is used for evaluating several interconnected attributes.

7 References

Ažman Momirski, L., Kladnik, D., 2009. Terraced landscapes in Slovenia. Acta geographica Slovenica 49-1. Ljubljana. DOI: 10.3986/AGS49101

Bailey, R. G. 1998: Ecosystem geography. New York.

Digitalni modeli višin, 2007. Internet: http://prostor.gov.si/vstop/fileadmin/struktura/DMV.doc (12. 10. 2009).

Džeroski, S. 2002. Environmental sciences. Handbook of data mining and knowledge discovery. Oxford.

Fayyad, U. M., Irani, K. B. 1993: Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. Proceedings of the Thirteenth international joint conference on artificial intelligence 2. San Mateo.

Gams, I. 1996. Termalni pas v Sloveniji. Geografski vestnik 68. Ljubljana.

Geodetska uprava Republike Slovenije. Public information of Slovenia/Javne informacije Slovenije. Digital elevation model 25/ Digitalni mode višin 25. Ljubljana. 2009.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H. 2009: The WEKA data mining software: an update. SIGKDD Explorations 11-1. Washington. DOI: 10.1145/1656274.1656278

Kononenko, I. 1994: Estimating Attributes: Analysis and Extensions of RELIEF. Machine Learning: ECML-94. Catania.

Kononenko, I. 2005: Strojno učenje. Ljubljana.

Kraft, J., Einax, J. W., Kowalik, C. 2004: Information theory for evaluating environmental classification systems. Analytical and Bioanalytical Chemistry 380-3. Berlin. DOI: 10.1007/s00216-004-2769-9

Hrvatin, M., Perko, D., 2003. Surface roughness and land use in Slovenia. Acta geographica Slovenica 43-2. Ljubljana. DOI: 10.3986/AGS43202

Hrvatin, M., Perko, D., 2005. Differences between 100-meter and 25-meter digital elevation models according to types of relief in Slovenia. Acta geographica Slovenica 45-1. Ljubljana. DOI: 10.3986/AGS45101

Hrvatin, M., Perko, D., Petek, F. 2006. Land use in selected erosion-risk areas of Tertiary low hills in Slovenia. Acta geographica Slovenica 46-1. Ljubljana. DOI: 10.3986/AGS46103

Kirkby, R., Frank, E. 2010: WEKA Explorer User Guide for Version 3-4. URL: http://garr.dl.sourceforge.net/project/weka/documentation/3.4.x/ExplorerGuide-3.4.16.pdf (21. 4. 2010).

Natek, M., 1998. Vinorodna območja. Geografski atlas Slovenije. Ljubljana.

Ogrin, D. 2000: Nekatere topoklimatske značilnosti razporejanja temperature zraka in burje v razgibanem reliefu Slovenije. Dela 15. Ljubljana.

Ogrin, D. 2007: Uporabnost kartiranja vinogradov kot metode za ugotavljanje prostorskih značilnosti termalnega pasu. Dela 28. Ljubljana.

Perko, D. 2001: Analiza površja Slovenije s stometrskim digitalnim modelom reliefa. Geografija Slovenije 3. Ljubljana.

Pezzi, G., Ferrari, C., Corazza, M. 2008: The Altitudinal Limit of Beech Woods in the Northern Apennines (Italy). Its Spatial Pattern and Some Thermal Inferences. Folia Geobotanica 43-4. Prague. DOI: 10.1007/s12224-008-9025-6

Podobnikar, T., Šprajc, I., 2007: Spatial analyses and Maya cultural landscape. Informatica 2007. Havana. Internet: http://www.ipf.tuwien.ac.at/publications/2007/podobnikar_cuba_2007.pdf (16. 5. 2009).

Land use/ Raba tal (September 2008). Internet: http://rkg.gov.si/GERK/ (1. 5. 2009).

Saito, H., Nakayama, D., Matsuyama, H. 2009: Comparison of landslide susceptibility based on a decision-tree model and actual landslide occurrence: The Akaishi Mountains, Japan. Geomorphology 109, 3-4. New York. DOI: 10.1016/j.geomorph.2009.02.026

Shannon, C. E., Weaver, W. 1949: The mathematical theory of communication. Illinois.

Vrhovec, T. 1991: A cold air lake formation in a basin – a simulation with a mesoscale numerical model. Meteorology and Atmospheric Physics 46-1/2. Vienna, New York. DOI: 10.1007/BF01026626

Watkins, R. L. 1997: Vineyard site suitability in Eastern California. GeoJournal 43-3. Dordrecht.

Whiteman, C. D., Haiden, T., Pospichal, B., Eisenbach, S., Steinacker, R. 2004: Minimum Temperatures, Diurnal Temperature Ranges, and Temperature Inversions in Limestone Sinkholes of Different Sizes and Shapes. Journal of applied meteorology 43-8. Boston. DOI:10.1175/1520-0450(2004)043<1224:MTDTRA>2.0.CO;2

Witten, I. H., Frank, E. 2005: Data mining. Practitcal machine learning tools and techniques. Amsterdam.

Žiberna, I. 1992: Vpliv klime na lego in razširjenost vinogradov na primeru Srednjih Slovenskih goric. Geografski zbornik 32. Ljubljana.

Žiberna, I. 1999: Temperaturna inverzija v hriboviti Sloveniji. Dela 13. Ljubljana.

Informativni vrednosti nadmorske višine in višinske razlike za ponazoritev termalnega pasu

DOI: 10.3986/AGS50201
UDK: 913(497.4):531.717.3
551.581:531.717.3(497.4)
COBISS: 1.01

IZVLEČEK: V prispevku smo uporabili izračun informacijskega prispevka ter razmerja informacijskega prispevka, ki se navadno uporabljata v procesih strojnega učenja, z namenom, da ocenimo, kateri podatkovni sloj bolje odraža topoklimatske značilnosti (predvsem termalni pas) – nadmorska višina ali višinska razlika. Oba atributa smo primerjali na podlagi njune informacijske vrednosti pri pojasnjevanju lokacije vinogradov, ki so zelo navezani na termalni pas. Analizo smo opravili na podlagi 9000 celic, ki smo jih zajeli z območij različnih vinorodnih okolišev. Kot boljši atribut se je večinoma izkazala višinska razlika, so pa opazne razlike med posameznimi območji, še posebej med celinskim in submediteranskim delom Slovenije.

KLJUČNE BESEDE: geografija, informacijski prispevek, razmerje informacijskega prispevka, termalni pas, višinska razlika, nadmorska višina, Slovenija

Uredništvo je prejelo prispevek 2. februarja 2010.

NASLOV:
Rok Ciglič
Geografski inštitut Antona Melika
Znanstvenoraziskovalni center Slovenske akademije znanosti in umetnosti
Gosposka ulica 13, SI – 1000 Ljubljana, Slovenija
E-pošta: rok.ciglic@zrc-sazu.si

Vsebina

1 Uvod

Za izdelavo naravnogeografske členitve s pomočjo geografskih informacijskih sistemov ter kvantitativnih metod potrebujemo prostorske podatke, ki so natančni in vsebujejo čim več informacij. Na območjih, za katere nimamo želenih podatkov, se moramo opreti na druge podatke. Zaradi povezanosti temperature in nadmorske višine (Bailey 1996, 68; Pezzi, Ferrari in Corazza 2008, 452) se lahko za ponazoritev temperatur uporablja digitalni model višin. Pri tem pa je treba biti pozoren na najnižje, konkavne dele ter nad njimi ležeči termalni pas, ki onemogočajo preprosto linearno povezavo nadmorske višine in nekaterih značilnosti podnebja (na primer temperature), predvsem pri preučevanju v večjem merilu.

Dopolnilo podatku o absolutni nadmorski višini je višinska razlika, s katero je mogoče približno določiti, kje so meje pojava jezera hladnega zraka ter termalnega pasu, saj so te okvirno že poznane (Gams 1996; Žiberna 1999; Ogrin 2007). Če raziskujemo ožje, zaključeno območje z enotnim dnom, razlike med atributoma ni; pri primerjavi več ločenih konkavnih oblik na različnih nadmorskih višinah, pa predpostavljamo, da pride bolj do izraza podatek o višinski razliki (slika 1).

Slika 1: Med posameznimi konkavnimi območji so nadmorske višine dna kotlin (neprekinjena črta) in spodnje meje termalnega pasu (prekinjena črta) različne, višinska razlika – in s tem tudi razlika med obema črtama – pa je približno enaka.

Glej angleški del prispevka.

Na podlagi lokacije vinogradov, ki odražajo omenjene klimatske pojave (Ogrin 2007), smo preverili, kateri atribut posreduje večjo količino informacije – višinska razlika ali absolutna nadmorska višina. Količino informacije smo ugotovili s pomočjo informacijskega prispevka in razmerja informacijskega prispevka. Omenjeni meri se uporabljata predvsem pri metodah strojnega učenja in vse pogosteje tudi na področju okoljskih znanosti (Džeroski 2002). Informacijski prispevek je kot oceno pomembnosti atributa predlagal Hunt leta 1966 (Kononenko 1994). V raziskavi smo izbrali več vinorodnih območij Slovenije ter poskušali ugotoviti, ali je moč pojasnjevanja lokacij vinogradov z višinsko razliko v primerjavi z močjo pojasnjevanja lokacij vinogradov z absolutno višino različna ter pri kakšnem prostorskem obsegu pride ta razlika najbolj do izraza.

2 Podnebje in relief

Reliefni dejavniki v Sloveniji so za podnebje nadpovprečnega pomena. Reliefna razčlenjenost je pomembna predvsem za lokalne in mikroklimatske razmere. Na topoklimatske razmere (tiste, kjer gre za modifikacijo klimatskih razmer zaradi reliefa) relief vpliva predvsem z višinsko strukturo, ekspozicijo in naklonom pobočij ter tipom površja, kjer je pomembna predvsem konkavnost (Ogrin 2000).

Konkavne oblike reliefa, ki lahko zadržujejo zrak pri dnu ter ne omogočajo mešanja z okoliškim zrakom (Whiteman s sodelavci 2004, 1232), nudijo ugodne reliefne pogoje za nastanek inverzije in jezera hladnega zraka. Višina je odvisna od višine reliefa, ki zapira depresijo, običajno pa je med 50 in 200 m. Nad jezerom hladnega zraka leži termalni pas, ki je toplejši od nižje in višje ležečih predelov (Ogrin 2000). Postaje na vzpetinah imajo zato za 2 °C višje minimalno in za 1 °C povprečno temperaturo kot pa enako visoko ležeče postaje s kotlinsko lego (Gams 1996). Meje termalnih pasov, ki se pojavljajo tudi ob morju, so v različnih območjih Slovenije različne (Ogrin 2000; Ogrin 2007; Žiberna 1999). V zimskem času se jezera hladnega zraka lahko obdržijo več dni (Vrhovec 1991, 91).

2.1 Določanje termalnega pasu

Določanje termalnega pasu s pomočjo klimatskih meritev je v posameznih pokrajinah nemogoče, mogoče pa je s profilnimi meritvami minimalnih temperatur ob posameznih vremenskih situacijah in s podrobnim kartiranjem rastišč toplotno zahtevnih kultur (Ogrin 2007).

S podatki meteoroloških postaj lahko določimo orientacijsko zgornjo mejo za celotno Slovenijo. Povprečne minimalne temperature nakazujejo mejo termalnega pasu pri višinski razliki okoli 500 m, povprečne letne temperature pa pri 200 do 250 m (Gams 1996; Žiberna 1999; Ogrin 2007). Vinogradi nakazujejo spodnjo mejo termalnega pasu med 15 in 30 m nad dnom dolin in zgornjo približno 450 do 550 m nadmorske višine (Žiberna 1992 po: Ogrin 2007).

Žiberna (1992) in Ogrin (2007) sta omejila termalni pas s spodnjo in zgornjo višinsko mejo vinogradov, spodnja meja pa je hkrati tudi povprečna višina jezera hladnega zraka v topli polovici leta oziroma v začetku rasti vinske trte. Pri tem je Ogrin opozoril, da lokacije vinogradov niso vedno racionalne. Na ta način določen termalni pas kaže predvsem na podnebne razmere v začetku rasti vinske trte, to je konec aprila in v maju, ko je ogroženost trte zaradi pozebe največja. Poleti je vertikalni obseg termalnega pasu večji, pozimi manjši. Gre za dinamičen pojav, ki je odvisen tudi od vremenskih razmer (Ogrin 2007). Stalna meja torej ne obstaja, je pa vpliv pogostega pojava jezera hladnega zraka v Sloveniji opazen na letnih povprečjih minimalnih in povprečnih temperatur. Izjema so le priobalne ravnice (Gams 1996).

Gostota vinogradniških površin z nadmorsko višino pada (Hrvatin in Perko 2003). Z deležem vinogradniških površin se poleg nadmorske višine povezuje tudi višinska razlika. Intenzivnost sovpadanja pojava vinogradniških površin in višinske razlike je med vinogradniškimi območji različna (Žiberna 1992). To poraja vprašanje, katera od obeh višin je bolj povezana z vinogradi, če obravnavamo širše območje, saj so nadmorske višine vinogradniških območij različne.

2.2 Težave pri določanju meja termalnega pasu in jezera hladnega zraka na podlagi vinogradov

Metoda za določanje obsega termalnega pasu na podlagi vinogradov ima nekaj omejitev (Ogrin 2007). Uporabna je predvsem v pokrajinah z dolgo vinogradniško tradicijo in v bolj reliefno izoblikovanih pokrajinah (Ogrin 2007). Pri tem moramo upoštevati, da za vinogradništvo niso izkoriščena vsa območja, ki so ugodna s klimatskega vidika, saj so lahko geološke in pedološke razmere ter družbenogeografske razmere manj ugodne (Žiberna 1992).

Na lokacije vinogradov vplivajo tudi drugi naravnogeografski dejavniki, kar so s statističnimi metodami dokazal tudi Watkins (1997 ter Hrvatin, Perko in Petek 2006).

3 Priprava podatkov za analizo

Vinogradi spadajo med tiste kulturne rastline, ki so bolj navezane na termalni pas. Zaradi te lastnosti smo izbrali lokacije vinogradov za ugotavljanje, kateri podatek bolje prikazuje topoklimatske značilnosti – višinska razlika ali nadmorska višina.

V tem prispevku smo se osredotočili predvsem na območja, kjer je določanje termalnega pasu in jezera hladnega zraka bolj enostavno. To pomeni reliefno bolj izoblikovana območja znotraj vinorodnih okolišev, ki so območja tradicionalnega vinogradništva.

Za izbor več različnih območij smo se oprli na razdelitev na vinorodne rajone (primorski, posavski in podravski) in okoliše (Natek 1998). Izbrali smo tri okoliše iz vsakega rajona, skupaj torej devet območij (preglednica 1, slika 4). Tako lahko primerjamo pomen višinske razlike na državni (vinorodni) ravni in ravni posameznega rajona. Možna je tudi primerjava med celinskim in submediteranskim delom Slovenije.

Preglednica 1: Izbrana območja.

primorski vinorodni rajon	posavski vinorodni rajon	podravski vinorodni rajon
Briški vinorodni okoliš	Šmarsko-Virštanjski vinorodni okoliš	vinorodni okoliš Radgonsko-Kapelske gorice
Kraški vinorodni okoliš	Dolenjski vinorodni okoliš	Mariborski vinorodni okoliš
Koprski vinorodni okoliš	Belokranjski vinorodni okoliš	Ljutomersko-Ormoške gorice

V vsakem okolišu smo izbrali manjše zaključeno vzorčno območje, ki se hkrati na eni strani odpira in dotika dna večjega konkavnega območja (na primer del kraškega polja). Izjema je le Bela krajina, kjer smo izbrali del širšega pobočja nad kraškim ravnikom.

Določanje obsega posameznega vzorčnega območja znotraj vinorodnega okoliša, s katerega smo zajeli celice, smo opravili s programom ArcGIS 9.3 (slika 2). Na podlagi digitalnega modela višin (DMV) smo z več ukazi določili posamezna odtočna območja, ki hkrati obsegajo zaključene reliefne enote z enotnim dom (obseg doline ali kraškega polja od dna do obronka).

Z ukazom *FILL* smo zapolnili vse depresije, ki so globlje manj kot 3,2 m, kolikor znaša tudi povprečna napaka DMV (Digitalni modeli višin 2007). S tem smo preprečili, da se zaradi morebitne napake v DMV-ju ali pa tudi dejanske manjše depresije odtočno območje ne bi določilo ločeno in previsoko nad dejanskim dnom.

Slika 2: Določanje obsega posameznega območja znotraj vinorodnega okoliša.
Glej angleški del prispevka.

Izbrana območja imajo na najnižji točki nadmorsko višino enako kot dno najbližje večje konkavne oblike oziroma uravnave (slika 4). Višinsko razliko smo določili tako, da smo od višine posamezne celice odšteli višino dna najbližje večje zaokrožene konkavne oblike.

Dna konkavnih oblik smo določili na podlagi konkavnosti in naklona. Konkavni relief smo določili po metodi primerjave originalnega in zglajenega DMV-ja, ki sta jo uporabila Podobnikar in Šprajc (2007). Za ravnino smo določili tiste celice, ki imajo naklon manjši od 2° (Perko 2001). Celoten postopek prikazuje slika 3. V raziskavi smo uporabili DMV z ločljivostjo 25 metrov (Hrvatin in Perko 2005; GURS 2009).

Slika 3: Določanje dnov konkavnih oblik s pomočjo DMV.
Glej angleški del prispevka.

Slika 4: Lokacije dnov konkavnih oblik ter izbranih območij.
Glej angleški del prispevka.

Na vsakem vzorčnem območju (primer Ljutomersko-Ormoških goric je na sliki 5) smo nato izbrali po 500 celic z vinogradi in 500 celic izven vinogradov, ker pomeni, da smo za vsak vinorodni okoliš izbrali 1000 celic ter za vinorodni rajon 3000 celic. Skupaj smo torej izbrali 9000 celic. Razmerje površja z vinogradi in celic brez vinogradov je le redko v enakem razmerju, dejansko le v Goriških Brdih (preglednica 2). V našem primeru smo bili zaradi tega prisiljeni naključno izbrati 500 celic z vinogradi in 500 celic brez vinogradov. Tako smo popačili lastnost pokrajine (delež vinogradov), vendar smo lahko le na ta način obdržali splošne značilnosti (višine) vinogradov znotraj posameznih območij in jih hkrati ustrezno primerjali med seboj. Če bi izbrali število celic glede na deleže površine posameznih območij, bi prišlo do napačnih rezultatov, saj bi nekatere območja z večjim deležem vinogradov povečala vpliv nadmorske višine in višinske razlike lastnih vinogradov. Enak pristop so uporabili tudi Saito, Nakayama in Matsuyama (2009), ko so primerjali različne vzorce podatkovnih slojev, ki so prikazovali prisotnost plazov.

Slika 5: Primer izbranega območja za izbor celic (Ljutomersko-Ormoške gorice).
Glej angleški del prispevka.

Preglednica 2: Število vseh celic in delež vinogradov.

	število vseh celic	delež vinogradov (%)
Briški vinorodni okoliš	9.770	52,1
Kraški vinorodni okoliš	9.370	27,3
Koprski vinorodni okoliš	6.836	11,9
Šmarsko-Virštanjski vinorodni okoliš	17.960	3,9
Dolenjski vinorodni okoliš	14.774	10,7
Belokranjski vinorodni okoliš	8.530	9,3
vinorodni okoliš Radgonsko-Kapelske gorice	5.396	19,0
Mariborski vinorodni okoliš	26.568	6,1
Ljutomersko-Ormoške gorice	10.463	27,0
skupaj	109.667	15,5

Glede na meje vinorodnih območij, prikazanih v Geografskem atlasu Slovenije (Natek 1998), in podatkih o rabi tal Ministrstva za kmetijstvo, gozdarstvo in prehrano (Raba tal 2009) na vinorodnih območjih vinogradi obsegajo približno 3 % površja.

V prispevku smo primerjali več območij, ki imajo v večini različen razpon vrednosti pri višinskih razlikah in tudi pri nadmorskih višinah. Ker to vpliva na vrednost informacijskega prispevka, je dodaten izračun razmerja informacijskega prispevka, s katerim lahko ta vpliv zmanjšamo in dodatno podpremo rezultat, nujen (preglednici 3 in 4).

Preglednica 3: Razpon vrednosti atributov.

	nadmorska višina (m)	razpon vrednosti (m)	višinska razlika (m)	razpon vrednosti (m)
Briški vinorodni okoliš	57–260	203	0–203	203
Kraški vinorodni okoliš	94–429	335	0–335	335
Koprski vinorodni okoliš	44–275	231	0–231	231
primorski vinorodni rajon	44–429	385	0–335	335
Šmarsko-Virštanjski vinorodni okoliš	200–432	232	0–232	232
Dolenjski vinorodni okoliš	171–469	298	0–298	298
Belokranjski vinorodni okoliš	141–588	447	0–447	447
posavski vinorodni rajon	141–588	447	0–447	477
vinorodni okoliš Radgonsko-Kapelske gorice	222–341	119	0–119	119
Mariborski vinorodni okoliš	265–849	584	0–584	584
Ljutomersko-Ormoške gorice	179–337	158	0–158	158
podravski vinorodni rajon	179–849	670	0–584	584
vsi rajoni skupaj	44–849	805	0–584	584

Preglednica 4: Razpon vrednosti 9000 celic, uporabljenih v raziskavi.

	nadmorska višina (m)	razpon vrednosti (m)	višinska razlika (m)	razpon vrednosti (m)
primorski vinorodni rajon (3000 celic)	45–426	381	0–332	332
posavski vinorodni rajon (3000 celic)	141–568	427	0–427	427
podravski vinorodni rajon (3000 celic)	179–847	668	0–582	582
vsi rajoni skupaj	45–847	802	0–582	582

4 Metode za izračun pomembnosti atributa

Informacijski prispevek (*information gain*) in razmerje informacijskega prispevka (*gain ratio*) sta meri za pomembnost atributa (Kononenko 2005) ali spremenljivke oziroma podatkovnega sloja (v prispevku uporabljamo izraz ‘atribut’). Ti meri se skupaj z drugimi uporabljata predvsem za usmerjanje in nadzor iskanja hipoteze v algoritmih strojnega učenja. Pri iskanju je osnovna naloga takega algoritma oceniti pomembnost atributa za dani učni problem (Kononenko 2005).

Torej, metoda se uporablja, ko želimo glede na vrednosti nekega obstoječega atributa (na primer kategorije rabe tal) ugotoviti, kateri izmed ostalih atributov (na primer nadmorska višina, tip kamnine) najbolje pojasnjuje te vrednosti. Na eni strani imamo ciljni oziroma napovedani atribut, na drugi pa več pojasnjevalnih atributov. Pristopa informacijske teorije so se tako na primer poslužili pri določanju najbolj ustrezne klasifikacije rek glede na onesnaženost sedimentov (Kraft, Einax in Kowalik 2004).

Omenjeni meri za ugotavljanje pomembnosti atributa temeljita na količini informacije. To je količina, ki je potrebna, da izvemo kakšen je izid nekega dogodka. Definirana je kot minus dvojiški logaritem verjetnosti dogodka (Shannon in Weaver 1949; Kononenko 2005) in jo izražamo bitih.

Količina informacije: $I(X_i) = -\log_2 P(X_i)$

Povprečni pričakovani količini informacije (*average information value*, Witten in Frank 2005), da izvemo, kateri izmed nezdružljivih izidov X_i ($i = 1 \dots n$, $\sum_{\mathrm{i}} P(X_i) = 1$ bit) se je zgodil, pravimo entropija dogodka (Kononenko 2005, 174) in ima naslednjo enačbo:

Entropija dogodka: $H(X) = -\sum_{\mathrm{i}} P(X_i) \log_2 P(X_i)$

Pri tem je X_i – dogodek, $P(X_i)$ pa je verjetnost dogodka X_i.

4.1 Izračun informacijskega prispevka in razmerja informacijskega prispevka

Pred podrobnejšo razlago izračuna informacijskega prispevka in razmerja informacijskega prispevka naj omenimo, da uporabljamo za vrednost/količino atributa, ki ga napovedujemo, izraz *razred* in ne *vrednost.* V analitičnem delu tega prispevka je napovedani atribut lokacija vinogradov, ki ima dva razreda: 0 in 1 oziroma 'vinograd je' in 'vinograda ni'. Za ostale atribute, s katerimi pojasnjujemo napovedanega, uporabljamo izraz *vrednost.* Torej v našem primeru imata obe višini vrednosti od 0 do nekaj 100 m.

Informacijski prispevek atributa – Gain(A) je definiran kot prispevna informacija atributa. Izračuna se tako, da od entropije razredov (H_R) odštejemo pogojno entropijo razreda pri dani vrednosti atributa ($H_{R|A}$) (Kononenko 2005):

$$\mathrm{Gain}(A)= H_R - H_{R|A}$$

Pri tem velja, da je informacijski prispevek atributa Gain(A) večji ali enak 0 in da je maksimalna vrednost informacijskega prispevka atributa Gain(A) enaka entropiji razredov (H_R).

Razmerje informacijskega prispevka – GainR(*A*) odpravlja pomanjkljivost informacijskega prispevka. Pri slednjem namreč kvaliteta atributa s številom različnih vrednosti raste. Precenjevanje večvrednostnih atributov se odpravi z normalizacijo informacijskega prispevka z entropijo vrednosti atributa (Kononenko 2005):

$$\mathrm{GainR}(A) = \mathrm{Gain}(A) \,/\, H_A$$

GainR(A) lahko precenjuje atribute z nižjo informacijsko vrednostjo, zato moramo pri končni oceni dejansko upoštevati oba načina izračuna (Witten in Frank 2005).

V preglednici 5 so navedeni podatki in enačbe, ki jih potrebujemo za izračun informacijskega prispevka ter razmerja informacijskega prispevka (po Kononenku 2005).

Preglednica 5: Podatki, razmerja ter enačbe, ki jih potrebujemo za izračun informacijskega prispevka ter razmerja informacijskega prispevka (Kononenko 2005).

Podatki:
n – število učnih primerov (število vseh enot ali celic)
n_k – število učnih primerov iz razreda r_k (število enot v razredu k napovedanega atributa r)
n_j – število učnih primerov z j-to vrednostjo danega atributa A_i (število enot z vrednostjo j pojasnjevalnega atributa A_i),
n_{kj} – število učnih primerov iz razreda r_k in z j-to vrednostjo danega atributa A_i (število enot v razredu napovedanega atributa z eno izmed vrednosti pojasnjevalnega atributa)

Razmerja posameznih podatkov:
$p_{kj} = n_{kj} / n$
$p_k = n_k / n$
$p_j = n_j / n$
$p_{k|j} = p_{kj} / p_j = n_{kj} / n_j$

Enačbe:
$H_R = -\sum_k p_k \log p_k$ – entropija razredov (entropija napovedanega atributa)
$H_A = -\sum_j p_j \log p_j$ – entropija vrednosti atributa (entropija vrednosti pojasnjevalnega atributa)
$H_{R|A} = H_{RA} - H_A = -\sum_j p_j \sum_k p_{k|j} \log p_{k|j} (H_{RA} \geq H_{R|A})$ – pogojna entropija razreda pri dani vrednosti atributa
$H_{RA} = -\sum_k \sum_j p_{kj} \log p_{kj}$ – entropija produkta dogodkov razred-vrednost atributa

Nekateri algoritmi uvrščanja in razvrščanja v skupine lahko obravnavajo le opisne spremenljivke. Zato je treba v takih primerih številske spremenljivke diskretizirati oziroma razdeliti na intervale (Witten in Frank 2005), kar je treba narediti tudi pri računanju informacijskega prispevka številskih atributov.

Metod diskretizacije je več (Witten in Frank 2005), tista, ki je uporabljena pri ocenjevanju pomembnosti številskega atributa v tej analizi uporabljenega programa Weka 3.5.8. (Hall s sodelavci 2009), temelji na entropiji ter upošteva načelo najkrajšega opisa (MDL – *minimum description length*). Pri tem, ko razdeljujemo enote na intervale, iščemo takšno razdelitev, ki naredi skupine glede na vsebnost razredov čim bolj žčiste’ (Witten in Frank 2005).

Pri tem postopku se najprej ugotovi, katera delitev številskih vrednosti v dve skupini glede na izbran atribut zagotavlja največji informacijski prispevek. Ko je meja, na podlagi katere se enote razvrstijo, določena, se postopek ponovi za vrednosti nad in pod izbrano mejo – torej proces se nadaljuje v obe smeri. Brez določenega kriterija se ta proces lahko nadaljuje dokler ne razdeli čisto vseh enot tako, da so skupine enot čiste. Ker to ni smiselno, saj bi se ugotovitve preveč prilagajale učnemu vzorcu in bi bilo rezultat teže posplošiti, so vpeljana določila za ustavitev procesa diskretizacije. Zato se pri vsaki delitvi preveri, ali je informacijski prispevek (te delitve) dovolj velik glede na število enot *N*, število razredov *k*, entropije enot *E*, entropije enot v vsakem podintervalu (po delitvi) E_1 in E_2 in števila razredov v vsakem podintervalu (po delitvi) k_1 in k_2 (Fayyad in Irani 1993; Witten in Frank 2005):

$$gain > \frac{\log_2(N-1)}{N} + \frac{\log_2(3^k-2) - kE + k_1E_1 + k_2E_2}{N}$$

Za ponazoritev, kaj pomenijo vrednosti izračunanega informacijskega prispevka, smo glede na izmišljeni napovedani atribut *C* (z dvema možnima razredoma – 0 in 1, v katerih je enakovredno število enot) izračunali vrednosti informacijskega prispevka (slika 6) za različne pojasnjevalne atribute A_x. Ti atributi so, tako kot *C*, dvojiški. Pri tem še dodajamo, da se vrednosti atributov *A* in *C* za posamezne enote ujemajo idealno, kar pomeni, da se vrednosti 0 čimbolj ujemajo z vrednostmi ‘da’, vrednosti 1 pa z vrednostmi ‘ne’ (preglednica 6). Iz slike 6 je razvidno, da je informacijski prispevek največji tam, kjer je ujemanje napovedanega in pojasnjevalnega atributa popolno in pada proti točki, kjer ima pojasnjevalni atribut le eno vrednost in zato ne more podati nobene informacije.

Preglednica 6: Vrednosti atributov *C* in A_x.

C	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}
0	da	da	da	da	da	da	da	da	da	da	ne
0	da	da	da	da	da	da	da	da	da	ne	ne
0	da	da	da	da	da	da	da	da	ne	ne	ne
0	da	da	da	da	da	da	da	ne	ne	ne	ne
0	da	da	da	da	da	da	ne	ne	ne	ne	ne
1	da	da	da	da	da	ne	ne	ne	ne	ne	ne
1	da	da	da	da	ne	ne	ne	ne	ne	ne	ne
1	da	da	da	ne	ne	ne	ne	ne	ne	ne	ne
1	da	da	ne	ne	ne	ne	ne	ne	ne	ne	ne
1	da	ne	ne	ne	ne	ne	ne	ne	ne	ne	ne
verjetnost dogodka »da« v odstotkih	100	90	80	70	60	50	40	30	20	10	0

Slika 6: Informacijski prispevek (Gain(*A*)) za pojasnjevalne atribute. Gain(*A*) je podan v bitih, verjetnost dogodka ‘da’ (*p*(*A*)) pa v odstotkih. Glej angleški del prispevka.

5 Izračun informacijskega prispevka in razmerja informacijskega prispevka nadmorske višine in višinske razlike glede na lokacije vinogradov

S programom Weka smo izračunali informacijski prispevek in razmerje informacijskega prispevka za pojasnjevalna atributa *nadmorska višina* in *višinska razlika* glede na napovedani atribut *vinogradi.* Z vsakega vinorodnega okoliša smo vzeli 500 celic z vinogradi in 500 celic izven vinogradov. Nato smo podatke iz posameznih okolišev združili po rajonih, na koncu pa smo skupaj združili tudi podatke za vse rajone. Na ta način smo lahko izračunali informacijski prispevek in razmerje informacijskega prispevka za območja posameznih rajonov in za območje vinorodne Slovenije. Dodatno smo izračunali ti meri tudi za območje rajonov celinske Slovenije (Posavski in Podravski rajon skupaj). S tem lahko ugotovimo, ali pride ob primerjavi območij znotraj posameznih rajonov in vseh območij na celotnem vinorodnem območju do različnih rezultatov.

Pri izračunu obeh mer smo uporabili način razdelitve enot v 10 skupin (*10-fold cross-validation*; Kirkby in Frank 2010). To pomeni, da smo enote razdelili v 10 skupin (*folds*) ter 10 krat izračunali informacijski prispevek in njegovo razmerje.

Izračunani *informacijskimi prispevki* (preglednica 7) ter *razmerja informacijskega prispevka* (preglednica 8) kažejo, da se višinska razlika skoraj povsod izkazuje kot pomembnejši atribut, saj višje vrednosti pomenijo večjo količino informacije.

Preglednica 7: Vrednosti informacijskega prispevka. Višje vrednosti so napisane v krepkem tisku.

izračun na podlagi:	nadmorska višina	višinska razlika
celic posavskega rajona	0,169 (±0,006)	**0,251 (±0,005)**
celic podravskega rajona	0,309 (±0,004)	**0,404 (±0,004)**
celic primorskega rajona	**0,016 (±0,003)**	0,012 (±0,002)
celic rajonov celinske Slovenije	0,212 (±0,003)	**0,283 (±0,004)**
celic vseh rajonov skupaj	0,104 (±0,002)	**0,127 (±0,002)**

Preglednica 8: Vrednosti razmerja informacijskega prispevka. Višje vrednosti so napisane v krepkem tisku.

izračun na podlagi:	nadmorska višina	višinska razlika
celic posavskega rajona	0,079 (±0,011)	**0,117 (±0,006)**
celic podravskega rajona	0,116 (±0,001)	**0,172 (±0,009)**
celic primorskega rajona	0,032 (±0,003)	**0,034 (±0,010)**
celic rajonov celinske Slovenija	0,079 (±0,006)	**0,118 (±0,006)**
celic vseh rajonov skupaj	0,042 (±0,001)	**0,067 (±0,001)**

Če primerjamo izračunane vrednosti z vrednostmi informacijskega prispevka na sliki 6, lahko bolje razumemo vrednosti informacijskega prispevka v naši analizi. Največjo informativno vrednost (0,404 bita) ima višinska razlika v podravskem rajonu. Na drugi strani imajo vrednosti okoli 0,1 bita izjemno malo informativno vrednost; te so na ravni dvojiškega atributa, ki ima 90 % vrednosti identičnih in ne podajo veliko informacij.

Z višinsko razliko se da lokacije vinogradov v večini rajonov bolje opisati kot pa z nadmorsko višino. Vrednosti se od rajona do rajona sicer razlikujejo. Ob primerjavi rezultatov vidimo, da so na območju rajonov celinske Slovenije vinogradi na bolj specifičnih legah – na določenih višinskih razlikah oziroma višinah nad dnom kotlin ali polij, saj je pomen višinske razlike bolj izrazit. Opazen pa je tudi večji razkorak med pomenom višinske razlike in dejanske nadmorske višine, kar bi lahko bila tudi posledica dejstva, da sta termalni pas in pojav jezera hladnega zraka na območju celinske Slovenije bolj izrazita. Enaka ugotovitev velja sicer tudi za nadmorsko višino, vendar ima ta povsod manjši informativni pomen.

V Primorskem vinorodnem rajonu je pomen obeh višin med vsemi rajoni najmanjši, pri informacijskem prispevku pa je celo nadmorska višina rahlo pomembnejša, kar je edini primer v tej analizi. Obe meri dokazujeta, da je termalni pas morda tam manj izrazit oziroma je tudi pojav jezera hladnega zraka redek in so posledično na splošno boljši klimatski pogoji tudi na območjih z manjšo višinsko razliko. Dobre fizičnogeografske razmere omogočajo, da so nekatere pokrajine izrazito usmerjene v vinogradništvo in večino površin namenjajo vinogradom – na primer Goriška brda. Ta spadajo, skupaj z nekaterimi drugimi območji (Koprska brda, Biljensko-Vrtojbenski griči, Vipavska brda) med območja z najbolj ugodnimi naravnimi razmerami, kar se odraža tudi v veliki zgoščenosti in terasiranosti vinogradniških površin (Ažman Momirski in Kladnik 2009).

Za najširše območje, območje vseh rajonov skupaj, obe višini pojasnjujeta lego vinogradov (in s tem predvsem topoklimatske razmere) v manjši meri, kot pa za posamezne rajone ali celotno območje rajonov celinske Slovenije skupaj. Zelo majhna informativna vrednost obeh slojev izenačuje pomen obeh slojev pri analiziranju v manjših merilih. Ta rezultat je v nasprotju s pričakovanji, saj bi pričakovali, da z večanjem števila celic z različnih območij (ki so na različnih nadmorskih višinah) tudi pomen višinske razlike pri razlagi lokacij vinogradov postaja večji. Tu dopuscamo nekaj možnih vzrokov; prvi je, da vinogradi dejansko niso enako porazdeljeni po vseh pobočjih, kar je lahko posledica tudi dejstva, da meje termalnega pasu ali pa jezera hladnega zraka niso tako enotne za območje Slovenije, na kar smo že uvodoma opozorili; druga možnost je ta, da vzorec celic ni bil dovolj ustrezen in bi analizo bilo smotrno ponoviti; tretja možnost pa je ta, da na lokacije vinogradov močneje vpliva še kakšen drug dejavnik. Zadnja možnost nas v prispevku ni toliko zanimala, saj smo želeli primerjati zgolj podatke o višinski razliki in nadmorski višini med seboj. K temu pa dodajamo, da smo že uvodoma pa smo omenili, da je reliefna razčlenjenost pomembna predvsem za lokalne, mikroklimatske razmere (Ogrin 2006, 126), kar ti rezultati tudi nakazujejo.

6 Sklep

Metoda se lahko ob primerni podatkovni bazi – taki, kjer imamo podatke za napovedani atribut in več pojasnjevalnih atributov oziroma podatkovnih slojev – izkaže kot koristna, kadar smo v dilemi, katere podatkovne sloje uporabiti za analizo. Z metodo lahko namreč izvemo, kateri podatki nam prinašajo največ informacij pri razlagi nekega pojava. Prednost metode je tudi ta, da omogoča primerjavo nominalnih in številskih atributov.

V prispevku smo uporabili lokacije vinogradov kot pokazatelje topoklimatskih značilnosti, točneje termalnega pasu, podatke o nadmorski višini in višinski razliki pa kot pojasnjevalne atribute. Dejansko smo torej ugotavljali, s katerim atributom lahko bolje pojasnimo oziroma opišemo krajevne topoklimatske razmere. Zaradi geografsko zelo raznolikih območij in posrednega določanja z vinogradi, je pojasnjevalna moč (informacijska vrednost) obeh atributov sicer zelo nizka, a kljub temu primerjava obeh atributov omogoča nekatera sklepanja o značilnosti lokacij vinogradov in s tem termalnega pasu. Potrdili smo, da je višinska razlika navadno pomembnejša za razlago lokacij vinogradov in s tem nakazali, da torej termalni pas obstaja. Med sklepi lahko posebej izpostavimo dejstvo, da ima višinska razlika v notranjosti vinorodne Slovenije večji pomen pri legi vinogradov kot pa na submediteranski strani, s čimer bi lahko sklepali tudi na bolj izrazit pojav jezera hladnega zraka ter termalnega pasu. Glede na rezultate ob primerjavi vseh celic (majhna vrednosti obeh mer) lahko tudi zaključimo, da v manjšem merilu lokacij vinogradov in s tem termalnega pasu ne moremo tako uspešno prikazati z nadmorsko višino ali pa višinsko razliko, kot v večjem merilu. Nasprotno smo pričakovali, da bo višinska razlika z upoštevanjem vseh celic (vseh vinorodnih rajonov skupaj) prišla še bolj do izraza. Rezultat je lahko posledica dejstva, da so lokacije vinogradov med območji dejansko tako različne in s tem morda tudi meje termalnega pasu, ali pa dejstvo, da vzorec celic ni bil ustrezen. Analizo bi bilo smotrno ponoviti z večjim vzorcem in pa z vključitvijo več območij. Možna je tudi vključitev več atributov ter analiza z metodo RELIEF (Kononenko 1994), ki se uporablja za vrednotenje več atributov, ki so med seboj tudi povezani.

7 Literatura

Glej angleški del prispevka.