APPLICATION OF DECISION-TREE ON WELLNESS TOURISM VISITORS' HEALTH BEHAVIOUR CHANGES

DÖNTÉSI FA ALKALMAZÁSA WELLNESS TURISZTIKAI UTAZÁSON RÉSZTVEVŐK EGÉSZSÉGMAGATARTÁS-VÁLTOZÁSÁNAK ALAKULÁSÁRA

Orsolya Savella [1] *– Péter Dudás*[2]

[1,] Tourism and Catering Department, John von Neumann University, Hungary
[2] Economic, Finance and Management Department, John von Neumann University, Hungary

Keywords:
health tourism
health behaviour
scene of health education
research directions
Kulcsszavak:
egészségturizmus
egészségmagatartás
egészségnevelési színtér
kutatási irányok

Abstract
In the case of social science research, the link between observation and theory can cause difficulties for researchers. Direct observation and measurement are rarely possible. Because the processes are constantly changing, it is possible to observe the phenomena once. Some of the models of mathematical, statistical methods are made in uncertain circumstances, which take into account the more important factors, but also the probable variables can be expressed. The technique used to make such a model is the decision tree method, which is a graphical model used in decision making, in which several choices are available and their outputs are uncertain. The study illustrates the use of this model, focusing on the health behavior of individuals involved in health tourism travel.

Összefoglalás
A társadalomtudományi kutatások estén a megfigyelés és az elmélet közötti kapcsolat megteremtése nehézséget okozhat a kutatók számára. Közvetlen megfigyelésre és mérésre ritkán van lehetőség. Mivel a folyamatok állandóan változó körülmények között mennek végbe, a jelenségek egyszeri megfigyelése lehetséges. A matematikai, statisztikai módszerekkel alkotott modellek egy része bizonytalan körülmények között készül, melyek a lényegesebb tényezőket veszik figyelembe, ugyanakkor a valószínűsíthető változók is kifejezésre juttathatók. Ilyen modellt alkotó technika a döntési fák módszere, amely egy olyan döntéshozatalban használt grafikus modell, melyben több választási lehetőség is rendelkezésre áll, és a kimeneteik bizonytalanok. A tanulmány ennek a modellnek az alkalmazását mutatja be, egészségturisztikai utazáson résztvevő egyének egészségmagatartására irányulva.

---

[1] Contact author. Tel.: +36 56 510 300 / 5538
   E-mail address: savella.orsolya@gk.uni-neumann.hu

1. Introduction

One of the relevant factors in the manifestation of health behavior, the strength of the intention, and the self-efficacy of the individual which includes the ability, skills, or effort required to implement it [1].

Knowledge-based behavioral intentions have also had an impact on the potential of health tourism to improve health behaviors (physical activity, physically passive leisure spending habits, mental health, nutrition). The results obtained to confirm that the informal knowledge, the intention to change the way of life are positive for health. Our research focused on the impact of gaining knowledge on behavioral intent. However, research did not explore how the acquisition or lack of knowledge affects the health behavior of in wellness tourists. In this article, we show how the qualitative variables that can be considered significant change in the current behavior, in addition to the observations that can be described with essentially quantitative variables.

2. Methods

The basic function of decision trees is pick up a complex decision problem to some minor problems. The method is a tool of decision-making that helps outline the full picture of the decision-making situation [2].

Several methods of making decision trees are known. Most algorithms can produce binary trees. The advantage of using it is to recognize irrelevant variables. This also helps to understand the problem. Because it reveals which variables are important and which are not. By this decisions are come more secure.

2.1. CHAID-based decision tree

CHAID (Chi-squared Automatic Interaction Detector) is a preferred segmentation technique because it is alternative to conventional cluster analysis, which is basically capable of grouping quantifiable observations. It is a multivariate recursive grading method that was originally about categorical variables [3]. With the development of the algorithm, it becomes suitable for treating continuous variables in case of either dependent or independent variables [4].

The decision trees are popular because the relationship between the assigned dependent variable, and the explanatory variables can be seen in a visual form. In an easily understandable tree structure variables can be easily interpreted. The advantage of the method, there is no restriction on the scale and their distribution of the variables, and by the extension of the algorithm both continuous and category dependent and independent variables can be handled.

The problem with the practical application of the scales is the level of measurement of variable. The proliferation of statistical programs also raises the question of whether and to what extent measurement methods can influence applicable methods. Using the programs speeds up the calculations, but the researcher must keep in mind what data he is working on.

In case of metric variables several researchers have confirmed either ordinal or interval type of data can be used [5].

The main steps of the algorithm forming the CHAID model:
- to select the dependent variable,
- for all explanatory variables, to merge the categories that are least different from each other in reference to the dependent variable,
- to segment the division of observations by categories of explanatory variable / after the division sub-databases are the next level of the tree structure /,
- the algorithm continues to merge categories and allocate cases until it reaches the predefined stop-criteria

The illustration of what is describes are presented in the field of tourism. Tourism is one of the most important sectors of the economy. There is a need for scientific research on tourism, and it is also a fact that tourism is a complex system that affects almost everyone.

We used the SPSS program for data processing and analysis.

3. Results

In a previous study we measured the influence of knowledge acquaintance on behavior intention [6]. In this study we examined the current health behavior after the trip. The respondents were previously classified into two large groups based on different characteristics. In the first group are those who have shown an attitude for behavior changing. They have been marked with the "active" indicator. In the second group are those who were not influenced by the travel experience. This is the "indifferent" category. We assumed that the questions of behavioral intention, current behavior and the traditional socio-demographic criteria could be the basis of segmentation. The CHAID-based method was expected to refine these groups. It is a question of what significant variables can be used to.

The sample included 332 people who took part in a health tourism trip in 2017. The conclusions drawn about the sample are guidelines and can be used for further research. Characteristics:

- Gender: female 75%; male 25%.
- Age: 21-35 years 46%; 36-55 years 43%; 56-74 years 11%.
- Took part in health tourism trip in 2017: once 55.4%; twice 30.1%; three times: 14.5%.
- Primary motivation: having rest 75%; health preservation, - improvement 14%; curing 9%; knowledge acquaintance 2%.

Based on the calculations, the groups were well separated, but the hypothesis was only partially verified.

Respondents were categorized on the strengths of the available database. The database contains 332 cases on six categories of nominal variables. The highlighted variables were selected from a preliminary research. These are the following:

1: age (under 20; 21-35; 36-55; 56-74; over 75)
2: health tourism trips in 2017 (1; 2; 3 or more)
3: spend more leisure time
4: conscious nutrition
5: less harmful habit
6: more conscious knowledge about disease prevention, health preservation

(Optional categories for 3-6 variables: it was also true before the trip; it is true since the trip; partly true; it was true temporarily; it isn't true neither since the trip.)

Target variable is the behavior after the trip, on the basis of which the respondent can be considered active. Based on preliminary research, this variable has two categories of aggregates: active and indifferent.

In this study $\alpha = 0.05$ level was used recommended by the program for construction of the CHAID model.

With regard to the stop criteria, the automatic setting was used offered by the program. After the adoption of the basic parameters, the model is shown in Figure 1.

Number of cases are shown according to given subgroup with their percentage distribution. In small tables, the estimated value is the category that represented by more than 50% of the cases. This is indicated by background coloring.

Level 0 of the tree shows the distribution of the entire database by category of target variable. The right column shows the element numbers for each category, and the center column shows the percentage distribution. If only this result is known, it can be stated that two-thirds of the respondents were active after the trip, one-third remained indifferent. The question: it is really possible to say whether one-third of the interviewees are really indifferent or there is some hidden information that can be used to reduce this ratio.

For this the algorithm performs the possible mergers for each explanatory variable firstly, and the program selects the one with the strongest effect.

In the study, this is the variable that describes the attitude of spending leisure time actively. This means that the categories of this variable have the greatest impact on the evolution of the groups. The initial database was distributed along the categories of this variable. This is the first level of the tree structure. As a result of the subdivision, the three sub-databases (1-3) show the distribution of the categories.
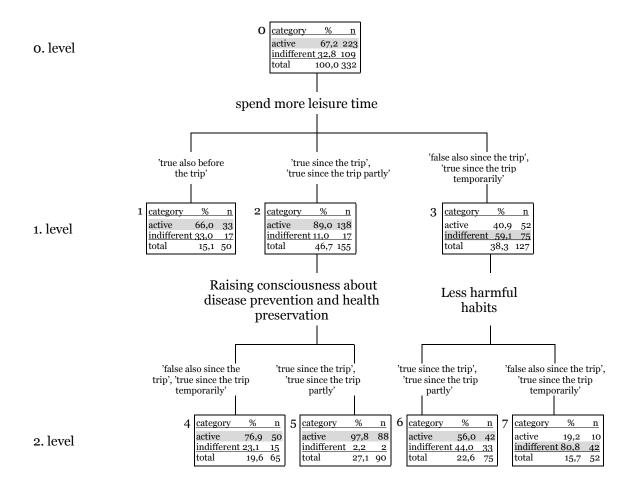
0. level

Node 0
category	%	n
active	67,2	223
indifferent	32,8	109
total	100,0	332

spend more leisure time

'true also before the trip'

'true since the trip', 'true since the trip partly'

'false also since the trip', 'true since the trip temporarily'

1. level

Node 1
category	%	n
active	66,0	33
indifferent	33,0	17
total	15,1	50

Node 2
category	%	n
active	89,0	138
indifferent	11,0	17
total	46,7	155

Node 3
category	%	n
active	40,9	52
indifferent	59,1	75
total	38,3	127

Raising consciousness about disease prevention and health preservation

Less harmful habits

'false also since the trip', 'true since the trip temporarily'

'true since the trip', 'true since the trip partly'

'true since the trip', 'true since the trip partly'

'false also since the trip', 'true since the trip temporarily'

2. level

Node 4
category	%	n
active	76,9	50
indifferent	23,1	15
total	19,6	65

Node 5
category	%	n
active	97,8	88
indifferent	2,2	2
total	27,1	90

Node 6
category	%	n
active	56,0	42
indifferent	44,0	33
total	22,6	75

Node 7
category	%	n
active	19,2	10
indifferent	80,8	42
total	15,7	52

Figure 1 CHAID model in database
Source: own editing

Knowledge of belonging to the criteria category reduces uncertainty. If everyone who was considered indifferent in preliminary research and would be excluded, we would exclude that 34 people who actually spend their free time also before their journey (17 people) or since the trip (17 people). It should be noted that in the preliminary grouping, 52 of the active people were in the set of indifferent category.

This means that 25.9 percent of the respondents ([52 + 17 + 17] / 332) would only make a wrong decision on the basis of the preliminary research. The initial 32.8 percent passivity was reduced by knowing the post-travel activity.

Continuing with the merge-splitting algorithm, on the level 2 has two explanatory variables with the same effect. The two sub-databases were divided in different ways. Those who spend their leisure time actively also before and after the trip are divided by the "more conscious acquisition of disease prevention and health care" by 40% -60% (4, 5 subsets). Those who were not active before the trip and still stayed there, or had little or no interest in them, were affected at least by the abandonment of harmful habits. This factor decomposed the respondents by 60% -40% (6, 7 subsets).

After separating the two sub-databases, the decision uncertainty dropped to 23.2 percent. At this level, the original data matrix consists of four disjoint partial databases.

In summary, the relationship of all explanatory variables is significant with the result variable, so they are included in the model (Table 1). (The significance study was done with the Bonferroni adjustment. This is commonly used when testing multiple hypotheses at the same time. In our case, the different and simultaneous hypotheses are independent hypotheses examined at different levels of the tree structure.)

Table 1 Tree table - Representation of the decision tree

Pile	'active'		'indifferent'		Predicted value	Siga
	N	%	N	%		
0	223	67,2	109	32,8	active	0,000
1	33	66,0	17	33,0	active	0,000
2	138	89,0	17	11,0	active	0,000
3	52	40,9	75	59,1	indifferent	0,000
4	50	76,9	15	23,1	active	0,001
5	88	97,8	2	2,2	active	0,001
6	42	56,0	33	44,0	active	0,000
7	10	19,2	42	80,8	indifferent	0,000

a: Bonferroni adjusted
Source: own editing

It should be mentioned that a more conscious diet is also a significant variable. This is level 3. The scope of the study can't be communicated.

Table 2 Change in unsteadiness

Level	'indifferent' (unsteadiness %)	Changes of unsteadiness (%)	
		To the previous level	To the zero level
0.	32,8	–	–
1.	25,9	– 6,9	– 6,9
2.	23,2	– 2,7	– 9,6
3.	22,0	– 1,2	– 10,8

Source: own editing

The change of uncertainty measures how another variable improves the uncertainty of the explanation (Table 2). In this case, the number of indifferent grades at each level is reduced by the percentage of the previous level.

In the table, the "indifferent" column shows how increasing levels (more and more explanatory variables) can result in more accurate ratings. The reduction in the error rate compared to the previous level is declining, but to a lesser extent. The last column shows a monotonous reduction in error after each step compared to the initial state.

Table 3 Model summary

"Letter"	Node		'active'		Response %	Index %
	case	%	case	%		
5	90	27,1%	88	39,5%	97,8%	145,7%
4	65	19,6%	50	22,4%	76,9%	116,1%
2	50	15,1%	33	14,8%	66,0%	98,3%
6	75	22,6%	42	18,8%	56,0%	83,4%
7	52	15,7%	10	4,5%	19,2%	28,6%

Source: own editing

Table 3 provides information on cases in the final "letters", which show the final classification (Table 3). The "Node" column shows the number of cases in the letter and their

proportion to all respondents (332). The "active" column shows the number of cases that fall into the target category in each letter, and their proportion to the total number of active.

The „response" column in each letter shows the 'active' ratio (%) for the number of cases in the "letter". This is a rate among index (%), response (%,) and expected root activity (67.2%). This would be 100% in every letter if the algorithm did not find the causes of activity in the independent variables. The higher the index (%), the greater effect of the rule in the "letter".

Table 4 shows the verification of the model. It can see how many times the match accuracy was, so how many times did the actual cases belong to each category and how many times the error was in the forecast (the rows contain the actual data, the columns the forecasts). The better the model, the greater proportion of observations are on the main diagonal. These are the right decisions. The total accuracy is 76.8%, so the risk of misclassification is 23.2%.

Table 4 Results of classification

Observed classification	Predicted classification		Right decision (%)
	active	*indifferent*	
active	213	10	95,5%
indifferent	67	42	38,5%
Overall percentage	84,3%	15,7%	76,8%

a: with level 3 not shown in Figure 1
Source: own editing

4. Summary

In this paper, after analyzing the available data, results are the following: research shows applied mathematical and statistical methods that are suitable for exploring the structure of data. In uncertain cases, it is advisable to base decisions on multiple sides. The advantage of the decision tree-based process is that over only a part-by-one forecast, complete picture of the composition can be made about each segment that is crucial in planning marketing campaigns.

This presented method is based on rules that increases the graph that defines the data grouping by level. In the decision tree, the rules also define the prediction for the given condition, which is based on the simple majority principle after formulating the rule. Decision trees can be built up efficiently also for large data sets.

"Don't make a decision without knowledge." (Bernoulli)

Acknowledgement

Bibliographical References

[1] Nagy, L. –Barabás, K. (2011): Az egészségműveltség és egészségmagatartás diagnosztikus mérésének lehetőségei. In: Csapó, B. – Zsolnay, A. (Ed.): Kognitív és affektív fejlődési folyamatok diagnosztikus értékelésének lehetőségei az iskola kezdő szakaszában. Budapest, Nemzeti Tankönyvkiadó, pp. 173-224.
[2] Winston, W. L. (2003): Operations Research - Applications and Algorithms. Duxbury Press, pp. 641-645.
[3] Kass, G. (1980): An exploratory technique for investigating large quantities of categorical data. In: *Applied Statistics*, 29. (2) pp. 119–127.
[4] Hámori, G. (2001): A CHAID alapú döntési fák jellemzői. In: *Statisztikai Szemle*, 79. (8) pp. 703-710.
[5] Borgatta, E.F. – Bohrnstedt, G.W. (1980): Level of measurement – once over again. In: *Sociological Methods and Research*, 9 (2) pp. 147-160.
[6] Savella, O. – Horváth, H. A. – Csizmásné Tóth, J. – Dudás, P. (2017): Az egészségturisztikai utazások során szerzett ismeretek hatása az egészséggel kapcsolatos viselkedési szándék kialakulására. In: Gradus, 4 (2) pp. 91-97.