Associate Professor Daniela MANEA, PhD
E-mail: daniela.todose@csie.ase.ro
Professor Emilia TITAN, PhD
E-mail: emilia.titan@csie.ase.ro
Associate Professor Cristina BOBOC, PhD
E-mail: cristina.boboc@csie.ase.ro
The Bucharest Academy of Economic Studies
Institute of National Economy
Student Andra ANOAICA
E-mail:andra.anoaica@gmail.com

LOGISTIC REGRESSION IN MODELLING SOME SUSTAINABLE DEVELOPMENT PHENOMENA

Abstract. *Technological innovations of the last decade have led to a real explosion of data and a practically unlimited capacity to create and to store them, remodelling day to day life.*
This paper analyses theoretical models for qualitative variables used in sustainable development. More exact and detailed information on natural resources are vital to the state, as well as to environmental agencies and to the private sector. The type of forest vegetation is one of the basic characteristics that are recorded and analysed in order to maintain the ecological balance.
Generally, the type of forest vegetation is either recorded directly by the agents, or by tele-detection. Both techniques are costly both in financial and time terms or even impossible to do. Predictive models offer an alternative to obtain this data. Although linear regression models are used on a wide scale by biologists and ecologists, these models are inadequate when the dependent variable is qualitative.. Logit models are a natural complement to regression models, where the endogenous variable is a qualitative variable, a situation that may be obtained or not, or a category of a classification. The popularity of logit models is explained by the multivariate nature of the models and the easiness with which they can be interpreted.

Key words*: physician's migration, health professionals, Romania.*

JEL Classification: C44, Q23, Q56

1. Introduction

Technological innovations of the last decade have led to a real explosion of data and a practically unlimited capacity to create and to store them, remodelling day to day life. There is a real movement in the direction of quantifying activities of the academic, scientifically, industrial, governmental and NGO fields.

The main objective of this paper is to describe and to apply statistical models suitable for treating qualitative variables used in sustainable development. Sometimes direct techniques for measuring variables used in sustainable development have proven to be costly and too time consuming. Even more, decision makers may consider useful to have available information regarding the inventories of nearby terrains that are not under direct control in which case it is usually economically or legally impossible to collect inventory data. Predictive models offer an alternative method of obtaining this data.

Linear regression is the popular statistical method used on a wide scale by biologists andecologists. The necessary calculations are elementary in case of simple linear regressions. However, these methods are inadequate in the case in which the dependent variable is qualitative.

The logit models represent a natural complement to linear regression models, where the endogenous term is not a continuous variable, but a state that may be obtained or not, or a category from a certain classification. The popularity of the logit models comes from the multivariate nature of methods and the easiness of interpretation. The logit models have the quality of being stochastic and to admit decisional variables. These decisional variables constitute the determinist part of the utility function, which is used to calculate the probability to make a choice from a series of available alternatives.

The empirical study described in this paper responds to an issue regarding sustainable development. Ecologists are facing more and more issues regarding estimations of an extended geographical area or of a long period of time. These problems vary from understanding the spatial distribution of species, in order to detect invading species in time, to the way in which local forest ecology interacts with fire models to determine continental carbon fluxes. By applying the logit model on our database, we intend to predict the chances of a certain type of species to develop in a certain geographical area, depending on the cartographic characteristics of the studied area.

Since our study has been conducted on a database with a large number of explanatory variables, and the expected answer is a categorical type, we will focus our attention on the multinomial logit model. The multinomial logit model, allows a superior number of possibilities for the dependent variable.

The first part of the paper is a presentation of the main concepts and notions used in the modelling of qualitative variables and multinomial logit models. The second part of the paper is an empirical study on the affiliation of some individuals to a certain type of trees depending on its geographical characteristics of its location.

2. Literature review

The study of the models describing qualitative variables started from the years 1940-1950. Biology, followed by psychology and sociology has been the first application areas. Recently these models started to be applied on economic data, where the development of qualitative models has taken two main directions.

The first direction is building of individual behavioural models based on economic theory. This approach has led to a better understanding of certain models, the logit model being developed by **McFadden**[1]. A second approach is one in which exogenous variables have been introduced to explain values that were supposed to be qualitative. The main role of these models is explanatory.

The multinomial logit model has been introduced at the end of the 60s by McFadden (1968) and Theil (1969). Boskin (1974) however, and Schmidt and Strauss (1975) have applied it in choosing a profession, in very different implementations.

Boskin has used the multinomial logit model to explain the (economic) behaviour in which choosing a profession is determined by monetary criteria which, on one side is represented by the costs of education necessary to access a certain profession and on the other side, the salary that individual hopes to achieve. The education costs and salaries vary from one profession to another. One such model can be used to predict, for example, the impact of decreasing the cost of education when choosing a profession.

Schmidt and Strauss, have used the multinomial logit model as an instrument for studying a discriminate analysis, with the purpose of highlighting distinctive features of each professional category. For example, mentioning the low percentage of women in management positions, they question if the gender of the individual has truly been a discriminatory criteria for belonging to a certain category, taking into consideration real differences regarding educational level and professional experience between men and women. In order to do this, they have tried to explain, in their model the "professional class" by using variables such as gender and education level.

The multinomial logit model has been extended in two directions: GEV models (Generalized Extreme-Value Logit Models), introduced by **McFadden** (1978)[2]out of which we mention the Logit Nested model and the random coefficient logit models (Mixed Multinomial Logit), developed in the90s by **Revelt and Train**

[1] D. McFadden (1968) : "Specification Tests for the Multinomial Logit Model", Econometrica, 52, (5), 1219-40

[2]D. McFadden (1978) : "Quantitative Methods for Analyzing Travel Behavior of Individuals: Some Recent Developments", BEHAVIOURAL TRAVEL MODELLING, 279-318, Croom Helm London: London, 1978.

(1998)[3]. The first development has been a generalization of the law followed by the residuals of the multinomial logit model. The second follows the same law but sets the parameters of the model in a random manner. The two models are a response to the critiques on the multinomial logit model, about satisfying a hypothesis of independence between the offered alternatives, many times unrealistic. This hypothesis is known in Anglo-Saxon literature as IIA (Independence from Irrelevant Alternatives) which is roughly translated as "independence depending on unstocked alternatives". The idea of this hypothesis is that the logit model doesn't consider the closeness between the different answers of the model. It has been conceived in a way that an individual has to choose between two alternatives independently from other choices offered. A particular fact that brings critique to this model is that introducing a new element in the set of possible choices does not re-evaluate the percentages taken into consideration during the process of decision-making between the two alternatives. The nested logit model take into account this problem and make the individual decision dependent on common criteria of several alternatives that are close by nature and the specific criteria of each alternative offered.

3. The logit model

Logical regression or the logit model is a probability model, developed by the statistician D. R. Cox in 1958. Logical regression measures the relation between dependent categorical variables and one or more independent variables, which are generally, but not mandatory, continuous, by estimating the probabilities of events associated to categorical variables[4]. The logistic regression can be seen as a special case of generalized linear models.

However, there are important differences between logistic regression and linear regression, mainly in assumed hypothesis set and also in the general form of regression function. In order to highlight these differences we will illustrate a model with a dichotomous answer as a linear regression.

The logit models are part of the nonlinear logit-probit which is based on a hypothesis that the probability law is known that can be normal – in case of probit models or the logistic in case of the logit model. The logit and probit model are used in order to model dependent variables which are, as type, probabilities, percentages, specific weights or variables of binary type.

The general formula of a logit-probit model is:

[3]D. Revelt, K. Train (1998) : "Customer-Specific Taste Parameters and Mixed Logit: Households' Choice of Electricity Supplier" Economics Working Papers E00-274

[4] C. Hurlin (2015), curs online "Econométrie des Variables Qualitatives" http://www.univ-orleans.fr/deg/masters/ESA/CH/Qualitatif_Chapitre1.pdf

$$y = \int_{-\infty}^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_nx_n} f_Y(y)dy + \varepsilon,$$

where

$f_Y(y)$represents the probability density of a depend variable.

As it can be noticed integral precedent defines the repartition function of the dependent variable Y, which means that the model obtained the following format:

$$y = F_Y(\beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_nx_n) + \varepsilon.$$

If the probability law is normal, the previous model is called probit, but if the probability law is logistic, the model is of logit type.

We assume that we have:

-a sample of N observations,

- y_i, i=1,..,N a dichotomous endogenous variable, which is coded conventionally as

$$y_i = \begin{cases} 1 \text{ in case that the event happens} \\ 0 \text{ otherwise} \end{cases}$$

- x_i= (x_i^1… x_i^k), i=1,…,N, the characteristics of the exogenous variables

In this case, the linear model is as follows:

$$y_i = x_i\beta + \varepsilon_i, i=1,\ldots,N$$

where $\beta = (\beta_1 \ldots \beta_K)' \in R^K$ is a vector for K unknown parameters and where ε_i are assumed to be independently distributed. Knowing that the endogenous variable y_i can only have the values 0 or 1, the linear specification implies the fact that the perturbations ε_i can only have two values conditioned by the vector x_i, in this way ε_i admits only a discreet law that do not verify the normality hypothesis of the residuals. Since we are assuming that residualsε_i have a means equal to 0, the probability p_i associated with event y_i=1 is determined in a unique manner. Therefore the expected value of residual scan be written as follows:

$$E(\varepsilon_i) = p_i(1 - x_i\beta) - (1 - p_i)x_i\beta = p_i - x_i\beta = 0 \Rightarrow$$
$$p_i = x_i\beta = Prob(y_i = 1)$$

The quantity $x_i\beta$ corresponds to a probablility and has to fulfil certain conditions, such as being included in the closed interval [0,1], a condition that cannot be fulfilled.

Another issue is the presence of heteroscedasticity. It could be observed that the variance-covariance matrix of the residuals is dependent on the individual, more exactly, depends on the characteristics associated with the exogenous variables x_i:

$$V(\varepsilon_i) = x_i\beta(1 - x_i\beta) i\ 1, \ldots,N$$

Therefore, the linear specification proves to be inadequate for using in case of a qualitative dependent variable

Logit models have as an endogenous variable the probability of an event happening, conditioned by exogenous variables. We consider the following model:

$$p_i = Prob(y_i = 1|x_i) = F(x_i\beta) i=1,...,N$$

where F(.) is the repartition function of the logistic law, $\forall w \in R$:

$$F(w) = \frac{e^w}{1+e^w} = \frac{1}{1+e^{-w}} = \Lambda(w)$$

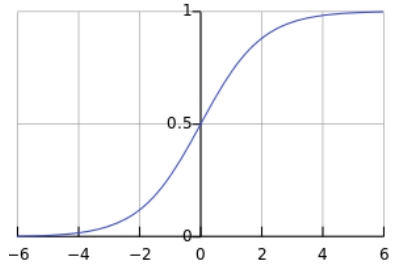

Figure 1. Logistic regression function Λ(w)

The logit model defines the probability associated to the event $y_i = 1$, as a value of the placement of the logical law considered in point $x_i\beta$:

$$p_i = \Lambda(x_i\beta) = \frac{1}{1+e^{-x_i\beta}}, \; i = 1,.., N$$

In order to give a correct interpretation of a logit model it is necessary a good understanding of the theoretical notion as **odds ratio**. Most often the probability is seen as a way to quantify chances for an event to take place: 0 meaning the event will not take place and 1 means the event happens. There are also other ways of representing chances for an event to happen, one of which being the odds ratio.[5]

Used by gambling houses, the odds ratio of an event is the ratio between the probability of an event taking place and the probability of not happening. An odds ratio of 4, for example, means we expect 4 times more for positive events than non-events. An odds ratio of ⅕ indicates a fifth of positive events compared to non-events. If p is the probability of an event and O is the odds ratio of the event, then

$$0 = \frac{p}{1-p}, \; p = \frac{0}{1+0}$$

[5]SAS online support "The Logistic Procedure"
http://support.sas.com/documentation

The odds ratio less than 1 corresponds to probabilities less than 0.5, and an odd ratio higher than 1, corresponds to a probability higher than 0.5. Same as in probabilities' case, the odds ratio have a lower limit equal to 0, however unlike probabilities, there is no upper limit. This characteristic makes the odds ratio to be useful in multiple comparisons, which makes it useful for measuring the relationship between two dichotomous variables.

A particular case of logit models is the multinomial logit model. These models have endogenous variables that may take more than two modalities.

We consider the multinomial model, in which the observed qualitative dependent variable for the individual i where i=1,..,N, noted y_i may take m_i+1 modalities, where j=1,.., m_i are assumed to be mutually exclusive for every individual i:

$$\sum_{j=0}^{m_i+1} Prob(y_i = j) = 1, i = 1, \dots, N$$

The probability associated with each answer can be defined by:

$$Prob(y_i=j) = F_{ij}(x,\beta),\ i=1,...N,\ j=0,1,...\ m_i$$

Starting from this definition we can make the following 4 remarks:

1. The number of modalities m_i taken by the endogenous variable y_i may depend on the individual: $m_z \neq m_k$.
2. The repartition function F_{ij} (x, β) corresponds with the probability that the individual i may choose modality j depending on the explanatory variables x and the parameters vector β. This function might be dependent on the individuals (indicator i), but also in on modalities (indicator j)
3. In the multinomial model, the probability associated with modality m_i+1 modalities (event usually coded with 0) is not necessary to be specified, since it can be calculated starting from m_i:
4. $$Prob(y_i=j) = 1-\sum_{j=1}^{mi} F_{ij}(x,\beta), i = 1, \dots, N$$

We define $\sum_{i=1}^{N}(m_i + 1)$ as binary variables y_{ij} so that:

$$y_{ij} = \begin{cases} 1\ if\ y_i = j \\ 0\ if\ y_i \neq j \end{cases} \text{ i=1,...,N}j=0,1,...\ m_i$$

The estimation of the multinomial logit model parameters can be made by using:

1. The maximum likelihood method
2. The GMM moment method, simulated moments
3. Non-parametrical and semi-parametrical methods

The interpretation of logistic regression coefficients is not as intuitive as in the case of interpreting usual coefficients with the help of linear regression. Unlike a linear regression where a positive coefficient of 0.5 indicates a rise with a unit of the exogenous term, for a logistic regression it indicates the fact that a rise with a unit of the exogenous term will lead to a 0.5 rise in the chance logarithm. This fact occurs due to the existence of a non-linear relationship between probability and exogenous term.
In order to easily make an interpretation of these coefficients odds ratios are used. An odd ratio of 1.5 shows that for a unit rise of the independent variable we will have a raise of 50% of the chances to realize a dependent variable. A value of 0,5of odds ratio indicates that for a unit rise of the independent variable we will have a decrease of 50% of the chances to realize a dependent variable.
In the case of linear regression the coefficient R^2 is used as a measure of the power of the model as the proportion from the total variation explained by the model. Cox and Snell propose the following coefficient:

$$R^2 = 1 - \left\{\frac{L(0)}{L(\widehat{\beta})}\right\}^{\frac{2}{n}}$$

where L(0) is the probability of the model that has only the constant, and $L(\widehat{\beta})$ is the probability of the specified model, and n is the size of the sample. The value of R^2 is less than one,with the maximum value:

$$R^2_{max} = 1 - \{L(0)\}^{\frac{2}{n}}$$

Nagelkerke(1992)[6] proposes the following adjusted coefficient, which can have a max value of 1:

$$\widetilde{R^2}^2 = \frac{R^2}{R^2_{max}}$$

Same as AIC statistics, R^2 and $\widetilde{R^2}$ are most useful in comparing models, a superior value indicating a better model.

4. Use of Logit model in sustainable development areas

The case study of this paper responds to some sustainable development issues. The ecologists are facing more and more issues that imply in a way or another, predictions on large geographical areas or over a long period of time. By applying the multinominal logit model on our database, we intend to predict the chances for a certain type of species to develop in a certain area, depending on cartographic characteristics of the studied area.

[6] Nagelkerke, Nico J. D. (1992): "Maximum Likelihood Estimation of Functional Relationships", Pays-Bas. Lecture Notes in Statistics 69. ISBN 0-387-97721-X.

We have a database available that has the characteristics of seven species of trees. The data has been gathered by the Forestier Service of USA, USFS – Forest Inventory and Analysis. Study zones include 4 natural areas situated in Roosevelt National Forest, from the north of the Colorado state: Rawth (29628 hectare), Neota (3904 hectare), Comanche Peak (27389 hectare) şi Cache la Poudre (3817 hectare). These natural areas contain forest terrain that has been subjected to very few direct human interventions. As a result, the current composition of vegetation inside these areas is a result of natural processes, than an active forest management.

Every observation represents an area of 30 m^2. The seven types of trees considered are: spruce, pine simple, yellow pine, willow, aspen, subalpine spruce and Douglas fir. The data set contain 15120 observations that include both characteristics of every species, as well as the tree itself.

The 7 types of trees are coded as follows: 1. spruce, 2. Simple pine, 3. Yellow pine, 4. willow, 5. Douglas fir, 7. Subalpin spruce.

Cartographies characteristics of the trees in the database set are:

- *Altitude – altitude measured in meters*
- *Aspect – aspect measured in degrees compared to the azimuth*
- *Slope – slope measured in degrees*
- *Dist_horiz_hidro – horizontal distance measured in meters until the next water surface*
- *Dist_vert_hidro – vertical distance measured in meters until the closest water surface*
- *Dist_horiz_road – distance measured in meters until the closest road*
- *Shadow_9am (index between 0 and 255) – index of the shadow at 9 o clock 9, summer solstice*
- *Shadow_12 (index between 0 and 255) – index of shadow at noon, summer solstice*
- *Shadow_3pm (index between 0 and 255) – index of shadow at 3 o clock, summer solstice*
- *Dist_horiz_fire – horizontal distance measured in meters between the points where wildfires start*

Natural areas of the observation are:

1 - Rawah natural area
2 - Neota natural area
3 - Comanche Peak natural area
4 - Cache la Poudre natural area

Before starting the analysis we will introduce two new variables in the database. We replace the 4 binary columns of the natural area variable with a single column with values from 1 to 4 corresponding to each natural area included in the database. The same treatment will be received by the other 40 binary columns

corresponding to the types of soil. The two new variables will be:Natural_aria (4 binary colums, 0-absence,1-presence) – natural area of which they belong to Soil_type (40 de coloanebinare, 0-absence, 1-presence)- type of soil

Decisional trees

Applying logistic regression directly on the database will lead to 6 regressions that compare one of the selected species as a way of reference for the other types of trees. The probabilities that one of the species to be more probable depending on the characteristics are calculated consequently, as a ratio to a single species. The results would be more interesting if a smaller number of trees were compared to the reference model.

In order to facilitate data analysis using the multinomial logit model, we will divide the species in two groups. Because we know the classes beforehand, the seven species, we have the case of a supervised classification. So we will turn to an instrument designed to solve this kind of issue, decisional trees. Since the results are categorical types, the tree will be a classification, the alternative being a regression tree.

The coding of the decisional tree is R, and the variables taken into consideration for building the tree are all the initial variables. R uses the algorythm CART. CART segments a set of data creating binary subtrees. A predictor is preferred to another predictor depending on the value of the entropy.

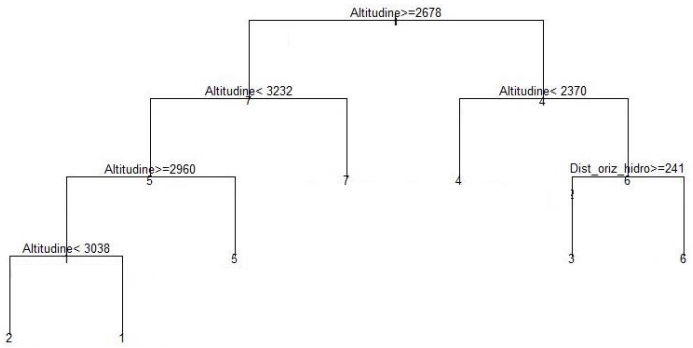

Figure 2. Decisional tree builts for grouping the seven species

We have obtained a grouping of the 7 species in two. In the first group we have the species 4- willow, 3- yellow pine, 6 douglas fir, and in the second we have 1-spruce, 2-simple pine, 5-aspen, 7-subalpine spruce.

We will do the logistic regression for each of the two groups. We will choose as a modality for the first group species number 3 – yellow pine and for the second group, species number 3 – simple pine.

Analysis of the first group

By using STEPWISE method and SAS procedure PROC logistic (annex 1), we will introduce explanatory variables all the cartographic characteristics we have, less the natural area, because we want to use the above mentioned model also for trees that don't necessarily belong to one of the natural areas.: For the Soil_type variable, we will use as reference modality, 10-Bullwark – Catamount family.

By introducing the Soil_type variable in the model we have a quasi-complete separation of data, so that some maximum likelihood estimators cannot be calculated or have too big values, as the SAS output states. Therefore, the **Soil_type** variable is eliminated from the analysis.

Table 1. Chance report for the value of the Soil_type variable

Odds Ratio Estimates				
Effect	**Species**	**Point Estimate**	**95% Wald**	
Soil_type 2 vs 10	1	<0.001	<0.001	>999.999
Soil_type 2 vs 10	5	11.432	3.322	39.342
Soil_type 2 vs 10	7	67.144	<0.001	>999.999

For the multinomial logit modelling we are using the SAS generalized logit model. The numerical optimization method for obtaining maximum likelihood estimates is Newton-Raphson. The maximum likelihood algorithm of the model converges, so that it eliminates the issue of the quasi-complete separation of data. If we take the generalized adjusted determination coefficientR^2into consideration, 78% of the variation is explained by our model.

Table 2. Testing global nullity of the first group parameters

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	11366.7628	24	<.0001
Score	7246.6899	24	<.0001
Wald	3368.2650	24	<.0001

All the three tests asymptotically equivalent, Wald test, maximum likelihood test and score test are indicators of the estimator's significance.

Table 3. Maximum likelihood estimates of the first group

Analysis of Maximum Likelihood Estimates						
Parameter	**Species**	**DF**	**Estimate**	**Standard**	**Wald**	**Pr >χ^2**
Intercept	**1**	1	-20.8689	0.9706	462.3072	<.0001
Intercept	**5**	1	14.1627	0.9503	222.1146	<.0001

Intercept	**7**	1	-80.7762	2.1638	1393.6471	<.0001
Altitude	**1**	1	0.00938	0.000282	1108.4809	<.0001
Altitude	**5**	1	-0.0059	0.000303	380.2957	<.0001
Altitude	**7**	1	0.028	0.000646	1875.1462	<.0001
Aspect	**1**	1	0.000151	0.000455	0.11	0.7402
Aspect	**5**	1	0.00358	0.000495	52.2619	<.0001
Aspect	**7**	1	0.001	0.00068	2.1807	0.1397
Slope	**1**	1	-0.029	0.00696	17.3564	<.0001
Slope	**5**	1	0.0322	0.00605	28.2907	<.0001
Slope	**7**	1	-0.0215	0.00998	4.6255	0.0315
Dist_hidro	**1**	1	-0.00253	0.000175	207.915	<.0001
Dist_hidro	**5**	1	-0.00034	0.000189	3.1579	0.0756
Dist_hidro	**7**	1	-0.00392	0.00023	290.5769	<.0001
Dist_horiz_road	**1**	1	-0.00012	0.000026	22.8764	<.0001
Dist_horiz_road	**5**	1	-0.00028	0.000034	65.1949	<.0001
Dist_horiz_road	**7**	1	-0.0001	0.000041	5.8677	0.0154
Shadow_12	**1**	1	-0.032	0.00315	103.5308	<.0001
Shadow_12	**5**	1	0.0248	0.00277	80.1853	<.0001
Shadow_12	**7**	1	-0.0254	0.00433	34.2735	<.0001
Shadow_3pm	**1**	1	0.00659	0.00165	15.9931	<.0001
Shadow_3pm	**5**	1	-0.0196	0.00154	161.9856	<.0001
Shadow_3pm	**7**	1	-0.00976	0.00246	15.6918	<.0001
Dist_horiz_fire	**1**	1	0.00006	0.000031	3.6927	0.0547
Dist_horiz_fire	**5**	1	-0.00038	0.000034	130.6522	<.0001
Dist_horiz_fire	**7**	1	0.000162	0.00005	10.2899	0.0013

The procedure will estimate three regressions as compared to the reference model. Almost all coefficients are significant at 5% level.

Table 4. Odds ratios estimates of the first group

Odds Ratios			
Effect	**Species**	**Unit**	**Estimate**
Altitude	**1**	10	1.098
Altitude	**5**	10	0.943
Altitude	**7**	10	1.323
Aspect	**1**	1	1
Aspect	**5**	1	1.004
Aspect	**7**	1	1.001
Slope	**1**	1	0.971
Slope	**5**	1	1.033
Slope	**7**	1	0.979
Dist_hidro	**1**	10	0.975
Dist_hidro	**5**	10	0.997
Dist_hidro	**7**	10	0.962
Dist_horiz_road	**1**	10	0.999
Dist_oriz_road	**5**	10	0.997
Dist_oriz_road	**7**	10	0.999
Shadow_12	**1**	1	0.968
Shadow_12	**5**	1	1.025
Shadow_12	**7**	1	0.975
Shadow_3pm	**1**	1	1.007
Shadow_3pm	**5**	1	0.981
Shodow_3pm	**7**	1	0.99
Dist_horiz_fire	**1**	10	1.001
Dist_horiz_fire	**5**	10	0.996
Dist_horiz_fire	**7**	10	1.002

A coefficient greater than one indicates a rise in the chance ratios for the event associated with the reference modality to be less probable than the one corresponding to the modality in which the comparison is done. We will interpret the odds ratio only for the values that have significant parameters. The values of the odds ratio are expressed in units measures specified in the logistic procedure syntax.

Therefore an increase of 10 meters in altitude will lead to a decrease of 9,8% of the chances for the tree to be a spruce, an increase of 32 % of the chances that the tree is a subalpine spruce and a decrease of 5,7 % of the chances that the tree is an aspen. The big difference regarding subalpine aspen is not surprising, this being a species that grows at high altitudes.

The odds ratio corresponding to the Slope variable indicates the fact that an increase in the degree of the slope will lead to a 3% increase in the chances for the

tree to be an aspen and a decrease of respectively 3 %. and 2 % for the tree to be a spruce, or a subalpine spruce.
The odds report corresponding to the Dist_hidro variable indicate the fact that an increase in the distance to a water source with 10 m decreases the probability that the respective tree species to be spruce with 2,5 % and with 3,8% to be a subalpine spruce. The coefficient associated to the aspen modality is insignificant in this case.
The odds ratio for the Shadow_3pm andDist_horiz_fire variables is close to the value of 1 and therefore we cannot interpret the differences registered between the reference modality, simple pine and compared modality.

Analysis of the second group

For the second group we have used the same SAS syntax and the same explicative variables :Altitude, Aspect, Slope, Dist_hidro, Dist_horiz_road, Shadow_9am, Shadow_12, Shadow_3pm, Dist_horiz_fire.
In this group we have 6480 observations and asin previouscase, for the multinominal logit modeling we have used the generalised logit model in SAS. The method of numerical optimization for obtaining the maximum likelihood estimates is Newton-Raphson. The maximum likelihood algorithm of the model converges, so that it eliminates the issue of the quasi-complete separation of data.
If we take the generalized adjusted determination coefficient R^2into consideration, 58% of the variation is explained by our model. All the three tests asymptotically equivalent, Wald test, maximum likelihood test and score test are indicators of the estimators' significance.

Table 4. Testing global nullity of the second group parameters

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	4711.6879	16	<.0001
Score	3555.0699	16	<.0001
Wald	2174.9667	16	<.0001

Table 5. Maximum likelihood estimates of the second group

Analysis of Maximum Likelihood Estimates						
Parameter	**Specie**	**DF**	**Estimate**	**Standard**	**Wald**	**Pr>ChiSq**
Intercept	**4**	1	19.4759	2.9668	43.0939	<.0001
Intercept	**6**	1	-20.7561	2.8603	52.6604	<.0001
Altitude	**4**	1	-0.0122	0.000391	982.1735	<.0001
Altitude	**6**	1	0.000974	0.000261	13.9269	0.0002
Slope	**4**	1	-0.055	0.0167	10.8504	0.001
Slope	**6**	1	0.067	0.0163	16.8316	<.0001
Dist_hidro	**4**	1	-0.00238	0.000284	70.4431	<.0001
Dist_hidro	**6**	1	-0.00246	0.000231	113.0065	<.0001
Dist_oriz_road	**4**	1	0.00101	0.000093	118.4871	<.0001
Dist_oriz_road	**6**	1	0.00025	0.00007	12.7112	0.0004
Shadow_9am	**4**	1	0.0145	0.0173	0.7044	0.4013
Shadow_9am	**6**	1	0.1716	0.0174	97.5467	<.0001
Shadow_12	**4**	1	0.0361	0.014	6.6537	0.0099
Shadow_12	**6**	1	-0.1857	0.0143	168.2014	<.0001
Shadow_3pm	**4**	1	-0.0204	0.014	2.1205	0.1453
Shadow_3pm	**6**	1	0.1562	0.0144	118.3323	<.0001
Dist_horiz_fire	**4**	1	0.000961	0.00009	114.0508	<.0001
Dist_horiz_fire	**6**	1	0.000357	0.000071	24.9611	<.0001

We observe that the shadow indicator for the three periods of the day is not significant.

Table 6. Odds ratio estimates of the second group

Odds Ratios			
Effect	**Species**	**Unit**	**Estimate**
Altitude	4	10	0.885
Altitude	6	10	1.01
Slope	4	1	0.946
Slope	6	1	1.069
Dist_hidro	4	10	0.976
Dist_hidro	6	10	0.976
Dist_horiz_road	4	10	1.01
Dist_horiz_road	6	10	1.003
Shadow_9am	4	1	1.015
Shadow_9am	6	1	1.187
Shadow_12	4	1	1.037
Shadow_12	6	1	0.831
Shadow_3pm	4	1	0.98
Shadow_3pm	6	1	1.169
Dist_horiz_fire	4	10	1.01
Dist_horiz_fire	6	10	1.004

An increase of 10 meters in altitude leads to an 18 % decrease in the chances for the tree to be willow, an increase of only 1 % for the respective tree to be a douglas fir. The willow is a species that grows at a lower altitude than a yellow pine.
The odds ratio corresponding to Slope variable indicates the fact that a rise in slope will lead to a 5,6 % decrease in the chances of the tree to be a willow, but it will increase with 7 % the chances for the tree to be a douglas fir.
The odds ratio corresponding to the Dist_hidro variable indicates the fact that an increase of the distance to a water source with 10 m decreases the probability that the tree species is a douglas fir or a willow with 2,4 %.
As for the odds ratio corresponding to the shadow index corresponding to the 3 hours taken into consideration we have obtained the following results. An increase with an unit of the 9 o clock shadow index increases the probability with 18% that the tree to be douglas fir, at 12 o clock decreases with 17%, and at 3 o clock this increases with 16%. An unit increase of the noon shadow index increases with 3,7% the probability that the respective tree is a willow.
The odds ratio for the Dist_horiz_fire and Dist_horiz_road variables are close to the value of 1, and an interpretation of it does not bring any extra information to the actual study. The parameter estimators for the variablesShadow_9am and Shadow_3pm are not significant for the willow, consequently it would be wrong to interpret the corresponding odds ratio.

Conclusions

Through this study we tried to solve a problem of sustainable development through statistical tools. By using a a database with 15120 observations, we have tried to predict which of the seven species of trees are more likely to be on a certain area of $30m^2$ depending on its characteristics.

The technique taken into consideration is multinomial logical regression. Although multinomial logical regression it doesn't offer a direct classification of the trees species as discriminant analysis, no condition of normality is needed. This constraint is not fulfilled by variables such as Dist_horiz_fire (the horizontal distance to the points of where wildfiresstart), Dist_horiz_road (horizontal distance to the road) or Aspect.

In order to facilitate the analysis we preferred to divide the species in two categories using a decisional tree using the CART algorithm. The first group contains the spruce, simple pine, aspen and subalpine spruce, and the second group contains the yellow pine, willow and douglas fir. We have chosen for both groups as reference models the simple pine and the yellow pine.

As expected, a modification in altitude determines a considerable change of the odds for a certain tree to be subalpine spruce in case of an altitude increase and a decrease for the simple spruce and aspen. In the second group an increase in altitude leads to the decreasing of chances that the tree is a willow and an insignificant increase to be a douglas fir, the douglas fir and the yellow pine being at relatively similar altitudes.

Another characteristic that offers important information regarding the type of tree that would be most probable to occur with the modification of the value is the slope measured in degrees. The trees on a terrain with a higher slope are the douglas fir and the aspen, and on the opposite pole are the spruce, will owand the subalpine spruce.

Dist_hidro indicates the species that are nearest to water sources : willow and douglas fir, spruce and subalpine spruce.

Significant variables, that do not modify the odds ratio very much are: Dist_roadand Dist_horiz_fire. In this case the odds ratio are very close to one.

The discriminant analysis is seen as an alternative to logical regression. Even if in theory the discriminant analysis wields better results than logistic regression by offering a direct classification of individuals on categories, in our case this analysis couldn't be applied because the initial hypothesis of normality are not verified. Therefore logistic regression is preferred because is a more flexible and easily applied to the data in real life, which sometimes are far from the theoretical ones.

REFERENCES

[1] **Hill T., LewickiP. (2007),** ***Statistics: Methods and Applications***; StatSoft, Tulsa, OK;
[2] **Hurlin C. (2015),** ***Econométrie des Variables Qualitatives***, http://www.univ-orleans.fr/deg/masters/ESA/CH/Qualitatif_Chapitre1.pdf;
[3] **McFadden D. (1968),** ***Specification Tests for the Multinomial Logit Model****;* Econometrica, 52, (5), 1219-40;
[4] **McFadden D. (1978),** ***Quantitative Methods for Analyzing Travel Behaviour of Individuals: Some Recent Developments****:* Behavioural travel modelling, 279-318, *Croom Helm London*: London, 1978;
[5] **Nagelkerke D., Nico J.D. (1992),** ***Maximum Likelihood Estimation of Functional Relationships****;* Pays-Bas. Lecture Notes in Statistics 69. ISBN 0-387-97721-X;
[6]**Press S, Wilson S. (1978),** ***Choosing Between Logistic Regression and Discriminant Analysis****;* Journal of the American Statistical Association, Vol. 73, No. 364., pp. 699-705 ;
[7] **Revelt D., Train K. (1998),** ***Customer-Specific Taste Parameters and Mixed Logit: Households' Choice of Electricity Supplier;*** Economics Working Papers, E00-274.