sciendo
DOI 10.2478/jazcas-2021-0055

A ROBUST APPROACH TO VARIATION IN CARPATHIAN RUSYN: RESAMPLING-BASED METHODS FOR SMALL DATA SETS

MOULAY ZAIDAN LAHJOUJI-SEPPÄLÄ – ACHIM RABUS
Slavonic Institute of the Albert-Ludwig-University of Freiburg, Freiburg, Germany

LAHJOUJI-SEPPÄLÄ, Moulay Zaidan – RABUS, Achim: A robust approach to variation in Carpathian Rusyn: Resampling-based methods for small data sets. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 603 – 617.

Abstract: Quantitative, corpus based research on spontaneous spoken Carpathian Rusyn language can cause several data-related problems: Speakers are using ambivalent forms in different quantities, resulting in a biased data set – while a stricter data-cleaning process would lead to a large scale data loss. On top of that, polytomous categorical dependent variables are hard to analyze due to methodological limitations. This paper provides several approaches to face unbalanced and biased data sets containing variation of conjugational forms of the verb *maty* 'to have' and *(po-)znaty* 'to know' in Carpathian Rusyn language. Using resampling based methods like Cross-Validation, Bootstrapping and Random Forests, we provide a strategy for circumventing possible methodological pitfalls and gaining the most information from our precious data, without trying to p-hack the results. Calculating the predictive power of several sociolinguistic factors on linguistic variation, we can make valid statements about the (sociolinguistic) status of Rusyn and the stability of the old dialect continuum of Rusyn varieties.

Keywords: oral corpora, border effects, language variation, spoken language corpus, robust statistics, Carpathian Rusyn

1 INTRODUCTION

As the size of empirical data and the number of bigger corpora rose as steadily as processing power of computers, complex statistical methods have obtained more and more approval in the field of linguistics. This trend also applies to dialectology and sociolinguistics – subfields with a greater focus on variation in spoken language. Compared to statistical methods applied to written language data, spoken language data can evoke several data related problems. As oral corpora are often unequally smaller than written corpora, results and the application of statistical tests have to be treated with special caution. Working with smaller data sets, outliers as well as autocorrelations between independent variables can pose the risk of causing a higher effect on the result of estimations or elaborated statistical tests than in larger, balanced data sets.

In this paper we discuss statistical methods from a sociolinguistic point of view. By analyzing a specific case of linguistic variation in Carpathian Rusyn, we

problematize the use of statistical methods by taking the rather complicated nature of spoken-language based data into account. We propose to analyze small and unevenly distributed datasets with resampling-based and robust methods, rather than reducing the complexity of the analysis or the data set for the sake of high significance levels. The aim is to avoid false positive or negative results by assessing statistics based on estimations, rather than absolute values.

The methods discuss are applied to verbal inflection in Carpathian Rusyn. The verbs *maty*, *znaty*, *poznaty*$_{3Ps.Sg.Pres.}$: 'to have' and 'to know' are analyzed with respect to the sociolinguistic embedding of the variation within the states Carpathian Rusyn is spoken, i.e.e., Poland, Slovakia, and Ukraine. The aim is to analyze which sociolinguistic factors with high influence on the outcome of the variation can be detected and whether so called "border effects" ([1], [2]) can be observed. The first section is dedicated to giving a short overview of the specific situation of Carpathian Rusyn, the background of the dataset, and the motivation of the analysis.

In the second section, the resampling methods cross validation and bootstrapping are applied to a multinomial logistic regression model, resulting in robust estimations of regression coefficients.

In the third section, we approach the variable importance via categorization with the decision-tree-bases methods Random Forest and Conditional Forest.

Since categorical dependent (and independent) variables are common in (socio-) linguistics and small, unevenly distributed data samples are more the rule than the exception when analyzing minority language data, our approaches are applicable beyond the Rusyn test case. For analyses, the open source software R-studio [3] is used.[1]

2 LINGUISTIC DATA AND METHODOLOGY

2.1 Variation in Carpathian Rusyn

Rusyn is a Slavic minority language mainly spoken in the Carpathian area, with the highest population of speakers in Transcarpathian Ukraine, Eastern Slovakia and Poland. Within the continuum of Northern Slavic languages, Rusyn is located right on the border between East- and West Slavic.

While Ukrainian is the linguistically closest language to the Rusyn varieties, their linguistic status and the national recognition of Rusyns as minorities is disputed. Some scholars claim that the Rusyn varieties are to be considered dialects of the Ukrainian language [4], others argue in favor of a separate linguistic and cultural identity of the speakers of Rusyn ([5], [6]). From a structural viewpoint, there are certain similarities with Ukrainian, e.g., with respect to common sound changes on the one hand, e.g., East Slavic polnoglasie, such as in *molodŷj* 'young' or the

[1] The R-script used for this work can be found via: https://bwsyncandshare.kit.edu/s/bGyJiGfHYkZHBa2 (please download the .html file and open with browser).

rendering of Common Slavic jat' as /i/ such as in *bilŷj* 'white'. On the other hand, certain properties make the Carpathian Rusyn varieties similar to the adjacent West Slavic languages, i.e., Polish and Slovak (for instance the use of clitic pronouns or the past tense formation using forms of the auxiliary verb 'to be' ([7], [8]). Resulting from the ambivalent status of Rusyn within the different European states, the situation is complex and dynamic. The current state of Rusyn can be researched using the online Corpus of Spoken Rusyn.[2]

In his grammar "The Rusyn Language" Stefan M. Pugh [9] describes the Prešov standardized variety of Carpathian Rusyn, from time to time with respect to other Rusyn non-standard variations (Slovak Rusyn, Lemko and Subcarpathian Rusyn). An interesting case of verbal variation is described within the conjugation classes "E(1) A(J): Conjugation Ia" and "E(2) AJ Proper" [9, p. 117–120]: The original stem marking A(J)[3] only appears in imperatives and in the non-past tense forms. As examples the verbs *čitaty* 'to read' and *maty* 'to have' are given, where the only form including the stem mark (A)J would be $čitaty_{3Ps.Pl.Pres.}$ (*čitajut'*) and $maty_{3Ps.Pl.Pres.}$ (*majut'*). The more one progressed to the east of the Rusyn dialect continuum, the more common a full A(J) conjugation would be evident (*mam* < *maju*, *mat'* < *maje*).

However, Pugh states that the A(J) forms within conjugations of this class were limited to the 3rd person plural, except the verbs *maty*, *znaty* and *poznaty*, where the A(J) forms can also be found in 3rd person singular forms. This leads to three competing forms of *maty*, *znaty*, $poznaty_{3Ps.Sg.Pres.}$:

ma, maje, mat'; zna, znaje, znat'; pozna, poznaje, poznat'.

The dataset we analyze this variation on contains 284 utterances of the above mentioned forms, by 56 speakers. The data has been obtained via query search in the Corpus of Spoken Rusyn. Corpus results can be downloaded and imported into the software R-Studio. In this case, the data set has been manually checked and cleaned[4] before the import. Besides the language samples (also available as anonymized audio recordings), the corpus also features speaker metadata (age, gender, living place, citizenship, GPS-locations).

Another variable (dialect area) has been added manually to our dataset. This variable reflects the affiliation of the villages to isoglosses that were the result of traditional dialectological research [10], before the current state borders had been established.[5] In this way, we can compare whether the traditional dialectal areas or the current states (and their respective roofing standard languages) have a stronger

[2] Accessible via www.russinisch.uni-freiburg.de/corpus (26.08.2021).

[3] Read as vowel "a" + J.

[4] We intentionally did not remove multiple utterances of the verbs by the same speaker as long as they were not within-sentence repetitions.

[5] This only applies to Rusyn data from Eastern Slovakia and Transcarpathia. The traditional Lemko dialect constellations have been torn apart by the violent resettlements of Lemko Rusyns (Akcja Wisła).

influence on the variable of interest. Statistics can also reveal differences between older and younger speakers in the sense of an apparent time study [11].

2.2 Methodological background

In order to analyze the relative importance of certain factors that might predict the outcome variable, i.e., the realization of the verb forms, we want to compare the usage of the variants of a linguistic phenomenon between several (sociolinguistic) subgroups within our data set. To do so, the variation has to be quantified. Most commonly, frequencies (of e.g. uttered word forms or specific grammatical constructions) are calculated and further on compared between several subgroups.[6] While performing statistical tests on a small data set of spoken language, quite a few problems can occur that might affect the quality of our results. Generally, researchers have a wider range of options when dealing with numerical outcome variables. Parametric statistics allows for making profound guesses about the population based on a certain underlying distribution of data, then comparing the variance within the data set with the natural distribution to assess statistical effects. However, when dealing with categorical language data there are quite a few more methodological limitations and tripping hazards.

A common statistical test in (socio-)linguistics, which is very similar to the test we are applying to our data below, is binomial logistic regression. Binomial distributions are traditionally described with a "success" "not success" scenarios like e.g. flipping coins, where each toss is independent from the latter and the probability for each side showing when it lands is equally probable (50/50 chance). The observations, derived from a random sample taken from a population are analyzed with the underlying assumption of a binomial distribution, similar to numerical data as weight and size are assumed to be distributed normally. Deviations from the distribution within the underlying population can be explained to a certain degree by factors determined within the regression formula. Besides the fact that a study design with bivariate variables can lead to (more or less necessary) oversimplification[7] of linguistic variation, the assumption of a natural 50/50 chance between two forms can as well be a bad starting point.[8]

[6] Subgroups can be defined in many possible ways and by multiple conditions. A group does not necessarily consist of many individual speakers; it could also be defined as a set of all the utterances of individuals. In our case, subgroups could consist of e.g. all female speakers. In between factor relations can be taken into account by defining and cross testing subgroups by multiple conditions (e.g. gender, age group, living place).

[7] Simplification of the variables can be an advantage form the methodological point of view because statistics involving polytomous dependent variables are disproportionately more computationally intensive and harder to analyze. The calculation of *n* baseline models can evoke to prohibitively high level of manual work and can pose the risk of bad model fitting.

[8] In researching e.g. the use of *L1* and *L2* forms, the chance of which form could be uttered may vary between individuals and groups due to random factors like weather, sympathy, geographically ambivalent perceptions of language and other factors that are hard to grasp statistically.

The most problematic property of our data set – typical for spoken and written linguistic corpora – is that many speakers utter several ambivalent forms, ranging from one utterance up to as many as twelve utterances in very different proportions. It is impossible to exclude the speakers, as we would not only have to work with an even smaller data set, but we would also willingly ignore existing variation and therefore making our whole study obsolete. For this reason, we decided to keep as much data within our sample as possible, even if this leads to some individuals dominating the data set and even though the assumption of independence between single data points is violated.

2.3 Resampling methods I: Cross validation and non-parametric bootstrapping

Taking the threefold nature of our dependent variable and several sociolinguistic factors into account, we conduct a multinomial logistic regression analysis with the formula *verb_form ~ Variety + Area + Age + Gender*. We used the function "multinom()" from the package "nnet". This function as part of the "nnet" package has several advantages such as the usual "lme4"-alike [12] regression output content and that there is no need to reshape the data set to long format. However, it does not provide p-values or t-statistics. The significance levels in Table 1 and 2 are provided by the function stargazer() of the R-package "stargazer" [13], that has been used to print the tables in HTML-format. It is important to note that this multinomial regression works by setting a baseline category and comparing two regressions side by side automatically. In our case, the set baseline of the dependent variable is the finite verb *mat* 'has'. The multinomial regression model predicts the *logit* of the two other verb forms with respect to the baseline model. As *verb_form* consists of three categories, the formula of the basic multinomial regression translates to:

$$\ln\left(\frac{P(\text{Verb\_Form} = \text{ma})}{P(\text{Verb\_Form} = \text{mat})}\right) = b_{10} + b_{11}\,(\text{Variety=Slo}) + b_{12}\,(\text{Variety=Tra}) + b_{13}\text{Age} + b_{14}(\text{Gender=m}) + b_{15}(\text{Area=1}) + b_{16}(\text{Area=2}) \cdots + \epsilon$$

$$\ln\left(\frac{P(\text{Verb\_Form} = \text{mae})}{P(\text{Verb\_Form} = \text{mat})}\right) = b_{20} + b_{21}\,(\text{Variety=Slo}) + b_{22}\,(\text{Variety=Tra}) + b_{23}\text{Age} + b_{24}(\text{Gender=m}) + b_{25}(\text{Area=1}) + b_{26}(\text{Area=2}) \cdots + \epsilon$$

Here, *P* are the odds, b_i are the regression coefficients of the respective factors and ϵ is the error term. Baselines are also set for the factors.

Trusting this *naïve* model, some model coefficients (**Tab. 1**) seem to be (highly) significant. However, we cannot solely rely on the meaningfulness of these values (even if they seem likely), not taking the violation of assumptions and the data related bias into account.

As described in **2.2**, several speakers produced different realizations of the response variable. The multinomial model cannot consider the individuals as a random factor.[9] In other words: the assumption of independence between data points is violated.

However, one can use this naïve approach in order to find a good model fit (the combination of predictive factors with the highest explanatory power) for further processing. The quality of the model fit (meaning how much of the effect can be explained by our predictors) can be assessed by comparing the Akaike Information Criterion or testing the model and measuring the accuracy.

A common approach is to split the dataset into a training and a test set in order to test the predictive power of the model on new data. For unbalanced small data sets, this approach is problematic as it worsens the model quality as some of the multiple utterances by the same speakers might be used for training and testing at the same time, moreover we would lose a greater part of our data during this process.

Alternatively, one can assess the accuracy of the model via *K-fold Cross Validation* ([15], [16], [17]). *CV* allows to split the data into *k* subsets and then compare each subset with all the other subsets. The *CV error rate is the* average *error rate* of the aggregated subset-based regression model. Doing so, the accuracy of the model can be predicted precisely without losing valuable data. The above mentioned formula has been chosen on the basis of the best *CV* accuracy rate. A formula including interaction effects between variety and age reached approximately the same accuracy rate and is therefore mentioned in the results (**Fig. 2** and **Fig. 3**). However, due to data related bias and violation of assumptions, the accuracy of the naïve regression model is merely 63%. That means the error term of the regression formulas has a predictive power of 37%.

Correlations between independent variables can have a strong influence on the outcome of the model. As shown in **Fig. 2**, the regression model with interaction effects seems to perform better (AIC) than the basic model. Nevertheless, the estimations of the regression are unreliable. The coefficient of the factor Transcarpathian variety (VarietyTRA) has become negative, even though there is no sound reason to assume a negative effect. This behavior can be explained by confounding [18]. A correlation between age and the Transcarpathian samples leads to the effect, that with the inclusion of the interaction variable, the VarietyTra:Age has not only an effect on the dependent variable, but also on the independent variable *variety*.

Hinneburg et al. [19] problematize the analyses of small datasets with a categorical dependent variable. Among other approaches, the authors show that non-parametric bootstrap can provide robust estimations of the statistics that help to avoid false assumptions about the underlying linguistic mechanisms. Fox [20] explains the principles

[9] Multinomial Logistic Mixed-Effects Regressions could potentially account for the individual variation of utterances. However, R-packages that are able to perform Mixed-Effects Regression Models for multinomial data reliably are rare. It is possible to perform several types of Multinomial Logistic Mixed-Effects Regression with the R-package "mclogit" [14] but in our case the algorithm did not converge.

behind bootstrapping regression models in R, which is that bootstrap allows estimating the distribution of regression statistics without making a priori assumptions about the distribution within the population. Therefore, data set is resampled *n*-times and the regression is calculated for each subset of the samples:

> "The essential idea of the non-parametric bootstrap is as follows: We proceed to draw a sample of size n from among the elements of S, sampling with replacement. […] The key bootstrap analogy is therefore as follows: The population is to the sample as the sample is to the bootstrap samples" [20, p. 1–2].

In this manner, not only the bias within the dataset, but also the effects of dependencies between several observations (several utterances of the same speakers) are reduced. We bootstrapped the regression model using the "boot()" function of the R-package "boot" [21].

As shown in **Tab. 2**, the median 1000-fold[10] bootstrapped regression coefficients as well as their significance levels are in the most cases less extreme than in the naïve model (**Tab. 1**).

The bootstrap process allows checking the distribution of the bootstrapped coefficients. After 1000-fold bootstrap, most coefficients seem to be normally distributed (cf. **Fig. 1**), with some of the distributions showing rather large spikes, skewness or broadly distributed minimum/maximum values. A straightforward way to calculate confidence intervals in R is by using the boot.ci() function of the package "boot" 2021 [21] or the boot_ci() function of the package "sjstats" [22]. If the distribution of bootstrapped coefficients contains larger spikes or extreme limits, boot_ci() will provide unrealistically large confidence limits for all variables. This is due to the methods being either entirely based on t-distribution or sample quantiles and the distributions are expected to be normal (no spikes, no skewness, no extreme limits). Boot.ci() provides the possibility to calculate bias-corrected and accelerated (*BCa*) confidence intervals, that seek to take skewness and bias within the distribution of coefficients into account. *BCa-CI* provide a far more realistic picture of the bootstrapped confidence intervals. **Fig. 2** displays the 95% *BCa* confidence limits of the multinomial logistic regression (without interactions), the black dots indicating the original, non-bootstrapped coefficients. Despite the fact that some *CI* are very large, the results show a more robust and less biased estimation of the coefficients. In some cases (Variety, Area), the *CI* indicates that the factors potentially have an even larger effect on the category of the dependent variable, than the median values in **Tab. 2** suggest.

[10] Meaning that the data set has been subsampled and the statistics have been calculated 1000 times.

2.4 Resampling methods II: Random Forest

When it comes to analyzing data with categorical outcome variables in R, CART[11]-based methods [23] provide a useful alternative to logistic regression models. The *bagging*[12] approach of Random Forests [24] is similar to the aggregated bootstrapping-approaches from above, but the underlying mechanisms behind CART differs from logistic regressions. Comparing CART-based models to the multinomial regression analysis (or vice versa), another equally valid perspective can be obtained. The alternative perspective can help to create a clearer picture of the calculated statistics and can, in case of a very unbalanced data set, help to verify or falsify results. Using the R-packages "randomForest" [25] and "party" ([26], [27], [28]), a robust estimation of variable importance is assessed easily, without the need to implement manual bootstrapping to the R-script. As Random Forests are even considered to be robust against presence of in-between variable interactions [29], they provide an additional corrective to the regression analysis. Without going into too much detail, we want to address a few tripping hazards that can occur while assessing the predictive power of factors via CART-based Forests.

The principle behind *decision trees* is rather straightforward. A "tree" is grown by deciding on several occasions (nodes) which factor is the most important for splitting the data between the categories of the dependent variable. Like the aggregated bootstrapped coefficients, *Random Forests* provide a robust estimate of several parameters that indicate the predictive power of factors, by combining the predictions of *n* numbers of trees, which are again based on random subsets of the data set. The data that is left out within each of the *n*-trees is used for assessing the overall accuracy of the model (OOB (out of bag)-error rate). In contrast to the regression models, *RF* algorithms use a random set of possible factors for each of these splits. It is important to check whether it is necessary to adjust the numbers of those factors. Within the formula, which is very similar to the regression formula above, the argument "mtry" indicates the amount of factors considered for each split. If "mtry" is set high, the choices between factors are less random and pose a higher risk of bias. If "mtry" is set low, the choice between factors is smaller, which may lead to a larger OOB-error rate. The OOB-error rate for the *RF* model (ntree = 10000, mtry = 3) *verb_form ~ Variety + Area + Age + Gender* was 24.7%, meaning the accuracy of the model is 75.3%.

The variable importance can be displayed with the help of the function "varImpPlot" (**Fig. 3**, left graph). The ranking of the variable importance of our analysis proves the point of Strobl et al. [29], that the mean *decrease Gini* and *mean decrease accuracy* indexes tend to be biased towards continuous independent variables (or in other cases towards variables with many categories). As shown in

[11] Classification and Regression Tree, also known as Decision Tree.

[12] Bagging is short for bootstrap aggregating.

Fig. 2, age has no predictive power in the regression models. The reason behind this error is that numerical variables like *age* can be split into various fractions, leading to considerably more options to split the decision tree branches compared to categorical variables with a very limited amount of possible splits.

Following Strobl et al. [29], the better approach for data sets with mixed (categorical and numerical) predictors is to use Conditional Forests via the function "cforest()". Conditional Forests [28], while being more computationally intensive, perform multiple significance tests at each splitting point of the trees. These significant tests (permutation tests, conceptually similar to the cross validation technique mentioned above) take several covariates of the variables into account, performing multiple significance tests on all possible combinations of predictors and covariates in the data set, preserving possible covariance structure of e.g. *variety* and *age*. As shown on the right-hand side of **Fig. 3**, the highest ranked (and therefore most important) factor is *variety*.

2.5 Interpretation

Our analysis shows that the predominating factor determining the verb forms *maty*, *znaty*, *poznaty*$_{3Ps.Sg.Pres.}$ is the factor variety, distinguishing between Transcarpathian, Lemko or Slovak Rusyn. While the old (formerly border-transgressing dialectal Areas haven't been ranked as unimportant (Area1), it seems that, at least in most cases, variety has the strongest effect. Comparing the coefficients (*ma*, *maje* vs. *mat*) of the regression model in between the varieties, Transcarpathian has by far the most homogeneous distribution of verb forms (the dominating form *maje* is congruent to the Standard Ukrainian form). Following the hypothesis of Border Effects [2] and the model of Auer and Hinskens [1, p. 17], the different embedding of Rusyn, brought about by the respective state (i.e. Poland, Slovakia and Ukraine), leads to convergence between non-standard varieties and their respective *dachsprache* and divergence within old dialectal continua. Considering the fact, that Rusyn is acknowledged as minority language in Slovakia and Poland, the more heterogeneous use of the verb forms within these varieties, including a strong use of verb forms differing from the respective umbrella languages, might not be accidental. Whereas the codified standard of Rusyn is taught in schools in Rusyn villages in Slovakia as well as in the Institute of Rusyn Language in Culture at *Prešov University*[13], the speakers of Rusyn are tending to be more confident about their language and identity [30].

3 CONCLUSION

Making correct statistical assumptions about inferences of sociolinguistic factors in spoken language data, especially dealing with a polytomous categorical variable of interest is unequally more difficult and error-prone than when dealing with parametric/

[13] https://www.unipo.sk/cjknm/hlavne-sekcie/urjk/o-institute/ (18.03.2021).

continuous data. Meeting all assumptions of the regression models and providing a balanced, unbiased data set is theoretically possible, but practically very unlikely to achieve without a prohibitively high amount of data manipulation or oversimplification of the variables of interest. The robust statistical methods suggested in this paper provide a broader perspective on the linguistic mechanisms behind the variation in spoken language, without 1. oversimplification of the data set, 2. without restricting the regression models to a binary outcome variable or just few predictors, and 3. without p-hacking. Even though the results of robust approaches are sometimes unspectacular, reporting robust estimations will reveal realistic tendencies and often significant results, instead of p-values with an unrealistically high level of significance (**Fig. 1**). By comparing several methodological approaches such as multinomial logistic regressions and Random (or Conditional) Forests, indistinct results can be re-evaluated from different points of view. As for our specific case, several statistical methods helped to uncover the underlying sociolinguistic factors behind variation within the inflectional system of verbs in Rusyn. The modern states where Rusyn is spoken have a stronger impact variation than the historical dialect areas or sociolinguistic factors such as age and gender.

It would be desirable to conduct further statistical analysis taking random factors into account as well as special factors such as the distance of the geographical location of the living place of speakers to the center of dialect areas or state borders.

ACKNOWLEDGEMENTS

The research has been funded by the German Research Foundation *DFG* under the project Number RA 2212/2-2 *"Rusyn as a minority language across state borders: a quantitative perspective"*.

Multinom. Log. Reg.: Verb Forms ~ Factors without & with Interaction Effects				
	Dependent variable:			
	ma	maje	ma	maje
	(1)	(2)	(3)	(4)
VarietySLO	-3.476***	-3.017***	2.529	-0.947
	(0.471)	(0.421)	(2.395)	(1.393)
VarietyTRA	10.800***	14.146***	**-13.792****	18.253***
	(0.355)	(0.355)	(0.001)	(0.001)
Age	0.006	0.023	0.024	0.050*
	(0.018)	(0.018)	(0.025)	(0.026)
Genderm	-1.079	-0.125	-0.876	0.163
	(0.733)	(0.724)	(0.732)	(0.740)
Area1	8.480***	11.786***	-1.262	14.462***
	(0.437)	(0.437)	(0.905)	(0.908)

Multinom. Log. Reg.: Verb Forms ~ Factors without & with Interaction Effects				
	Dependent variable:			
	ma	maje	ma	maje
	(1)	(2)	(3)	(4)
Area2	-1.155**	-0.658	-10.001***	2.845**
	(0.471)	(0.422)	(1.818)	(1.194)
VarietySLO:Age			0.034	-0.087**
			(0.056)	(0.039)
VarietyTRA:Age			0.427***	-0.056***
			(0.008)	(0.008)
Constant (mat)	3.270***	1.477	2.261*	-0.153
	(1.091)	(1.100)	(1.292)	(1.344)
Note:	*p<0.1; **p<0.05; ***p<0.01			

Tab. 1. Result table of *naïve* Multinomial Logistic Regressions model

Bootstrap Multinom. Log. Reg. Coeff. Median Values: Verb Forms ~ Factors without & with Interaction Effects R = 1000				
	Dependent variable:			
	ma	maje	ma	maje
	(1)	(2)	(3)	(4)
VarietySLO	-3.097	-2.267	0.993	0.713
	(2.857)	(1.619)	(59.64)	(34.74)
VarietyTRA	7.538**	12.564***	**-14.579**	19.052
	(3.139)	(2.739)	(24.089)	(34.87)
Age	0.007	0.026	0.0239	0.0507
	(0.020)	(0.019)	(1.087)	(1.087)
Genderm	-1.106	-0.142	0.955	0.088
	(1.551)	(1.523)	(5.237)	(5.228)
Area1	7.08	11.95***	-2.025	17.688
	(5.597)	(4.430)	(27.216)	(48.321)
Area2	-1.844	-1.608	-10.17	1.924
	(2.462)	(1.745)	(53.36)	(16.127)
VarietySLO:Age			0.041	-0.1
			(1.162)	(1.22)
VarietyTRA:Age			0.437	-0.058
			(1.68)	(1.81)

Bootstrap Multinom. Log. Reg. Coeff. Median Values: Verb Forms ~ Factors without & with Interaction Effects R = 1000				
	Dependent variable:			
	ma	maje	ma	maje
	(1)	(2)	(3)	(4)
Constant	1.41	1.4081796	2.193	-0.287
	(2.488)	(2.452)	(23.76)	(23.759)
Note:	*p<0.1; **p<0.05; ***p<0.01			

Tab. 2. Result table of *bootstrap* Multinomial Logistic Regressions models

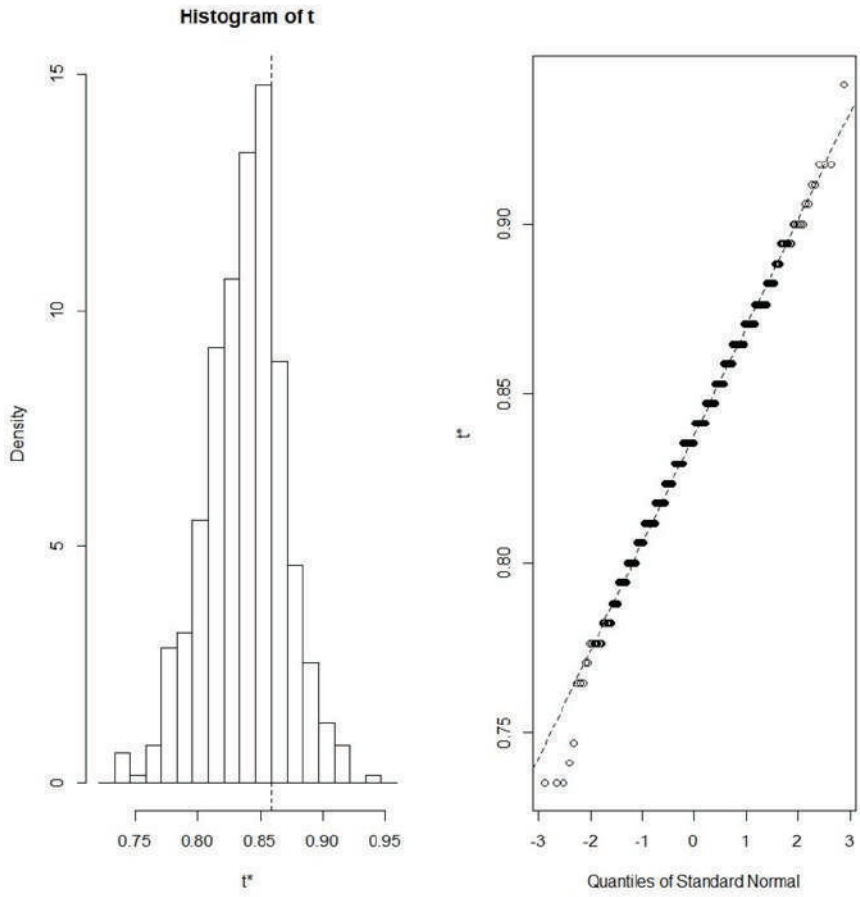

Fig. 1. Normal-like distributed bootstrap cofficients (t)

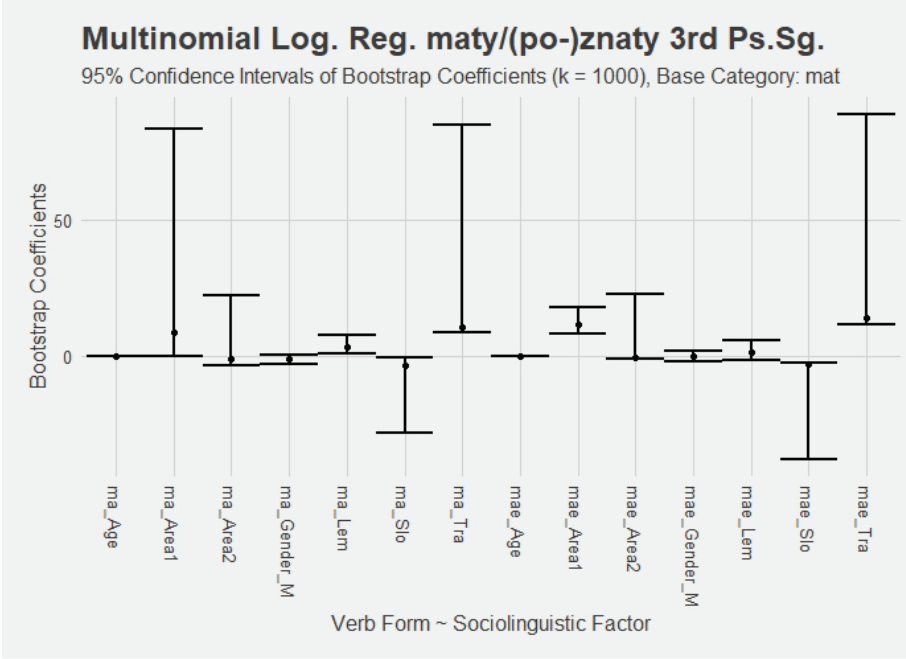

Fig. 2. Bootstrap Confidence Intervals (95%)

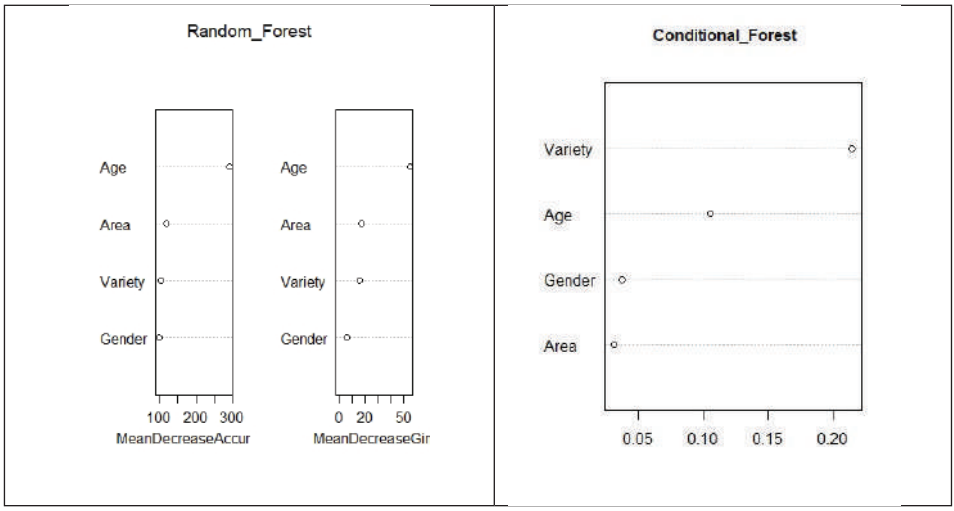

Fig. 3. Variable importance of Random Forest and Conditional Forest

References

[1] Auer, P., and Hinskens, F. (1996). Convergence and Divergence of Dialects in Europe. In Sociolinguistica (10).

[2] Woolhiser, C. (2005). Political borders and dialect divergence/convergence in Europe. P. Auer, F. Hinskens and P. Kerswill (eds.). Dialect change: Convergence and divergence in european languages. Cambridge, pages 236–262.

[3] RStudio Team. (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA. Accessible at: http://www.rstudio.com/.

[4] Magocsi, P. R. (2015). With Their Backs to the Mountains: A History of Carpathian Rus' and Carpatho-Rusyns. Budapest.

[5] H. A. Skrypnyk (ed.). (2013). Ukrajinci-Rusyny: Etnolinhvistyčni ta etnokul'turni procesy v istoryčnomu rozvytku. Kyjiv.

[6] Boudovskaia, E. E. (2006). The morphology of Transcarpathian Ukrainian dialects. Los Angeles.

[7] Rabus, A. (2019). Vergangenheitsbildung in gesprochenen karpatorussinischen Varietäten: Quantitativ-statistische Perspektiven. Die Welt der Slaven 69(1), pages 15–33.

[8] Plishkova, A. (2009). Language and national identity: Rusyns south of Carpathians. Translated by Patricia A. Krafcik. With a bio-bibliographic introduction by Paul Robert Magocsi. New York (Classics of Carpatho-Rusyn scholarship, 14).

[9] Pugh, S. M. (2009). The Rusyn language: A grammar of the literary standard of Slovakia with reference to Lemko and Subcarpathian Rusyn. München (Languages of the World/ Materials, 476).

[10] Pan'kevyč, I. (1938). Ukrajins'ki hovory Pidkarpats'koji Rusy i sumežnych oblastej. Z pryložennjam 5 dialektolohičnych map. Častyna I. Zvučnja i morfolohija. Praha.

[11] Chambers, J. K. (2002). Patterns of Variation including Change. The Handbook of Language Variation and Change, pages 358–361.

[12] Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), pages 1–48. DOI 10.18637/jss. v067.i0.

[13] Hlavac, M. (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.2. Accessible at: https://CRAN.R-project.org/package=stargazer.

[14] Elff, M. (2020). mclogit: Multinomial Logit Models, with or without Random Effects or Overdispersion. R package version 0.8.5.1. Accessible at: https://CRAN.R-project.org/ package=mclogit.

[15] Mosteller, F., and Tukey, J. W. (1968). Data analysis, including statistics. In Handbook of Social Psychology. Addison-Wesley, Reading, MA.

[16] Efron, B. (1982). The Jackknife, the Bootstrap and Other Resampling Plans. CBSM38, SIAM, Philadelphia, Penn.

[17] Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. J. Am. Stat. Assoc., 78, pages 316–331.

[18] VanderWeele, T. J., and Shpitser, I. (2013). On the definition of a confounder. Annals of Statistics. 41(1), pages 196–220.

[19] Hinneburg, A., Mannila, H., Kaislaniemi, S., Nevalainen T., and Raumolin-Brunberg, H. (2006). How to Handle Small Samples: Bootstrap and Bayesian Methods in the Analysis of Linguistic Change. Literary and Linguistic Computing 22(2), pages 137–150.

[20] Fox, J. (2002). Bootstrapping Regression Models Appendix to An R and S-PLUS Companion to Applied Regression.
[21] Canty, A., and Ripley, B. (2021). boot. Bootstrap R (S-Plus) Functions. R package version 1.3-25. Accessible at: https://cran.r-project.org/web/packages/boot/boot.pdf.
[22] Lüdecke, D. (2020). _sjstats: Statistical Functions for Regression Models (Version 0.18.0). Accessible at: https://CRAN.R-project.org/package=sjstats.
[23] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). Classification and regression trees.
[24] Breiman L. (2001). Random forests. Machine Learning, 45(1), pages 5–32.
[25] Liaw A., and Wiener, M. (2002). Classification and Regression by randomForest. R News 2(3), pages 18–22.
[26] Hothorn, T., Buehlmann, P., Dudoit, S., Molinaro, A., and Van Der Laan, M. (2006). Survival Ensembles. Biostatistics, 7(3), pages 355–373.
[27] Strobl, C., Boulesteix, A., Zeileis, A., and Hothorn, T. (2007). Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. BMC Bioinformatics, 8(25). Accessible at: http://www.biomedcentral.com/1471-2105/8/25.
[28] Strobl, C., Boulesteix, A., Kneib, T., Augustin, T. and Zeileis, A. (2008). Conditional Variable Importance for Random Forests. BMC Bioinformatics, 9(307). Accessible at: http://www.biomedcentral.com/1471-2105/9/307.
[29] Strobl, C., Hothorn, T. and Zeileis, A. (2009). Party on! A New, Conditional Variable Importance Measure for Random Forests Available in the party Package. Department of Statistics: Technical Reports, No. 50.
[30] Schimon, A., and Rabus, A. (2016). Wahrnehmungsdialektologische Untersuchungen zum Russinischen in Zakarpattja am Beispiel der Region Chust. Zeitschrift für Slawistik 61(3), pages 401–432.