DOI: 10.15439/2021F87
ISSN 2300-5963 ACSIS, Vol. 26

Alternatives for greedy discrete subsampling: various approaches including cluster subsampling of COVID-19 data with no response variable

Lubomír Štěpánek
Department of Statistics and Probability
Faculty of Informatics and Statistics
University of Economics
nám. W. Churchilla 4, 130 67 Prague, Czech Republic
lubomir.stepanek@vse.cz
&
Institute of Biophysics and Informatics
First Faculty of Medicine
Charles University
Salmovská 1, Prague, Czech Republic
lubomir.stepanek@lf1.cuni.cz

Filip Habarta
Department of Statistics and Probability
Faculty of Informatics and Statistics
University of Economics
nám. W. Churchilla 4, 130 67 Prague, Czech Republic
filip.habarta@vse.cz

Ivana Malá
Department of Statistics and Probability
Faculty of Informatics and Statistics
University of Economics
nám. W. Churchilla 4, 130 67 Prague, Czech Republic
malai@vse.cz

Luboš Marek
Department of Statistics and Probability
Faculty of Informatics and Statistics
University of Economics
nám. W. Churchilla 4, 130 67 Prague, Czech Republic
marek@vse.cz

***Abstract*—An exhaustive selection of all possible combinations of $n = 400$ from $N = 698$ observations of the COVID-19 dataset was used as a benchmark. Building a random set of subsamples and choosing the one that minimized an averaged sum of squares of each variable's category frequency returned similar results as a "forward" subselection reducing the dataset one-by-one observation by the same metric's permanent lowering. That works similarly as k-means clustering (with a random clusters' number) over the original dataset's observations and choosing a subsample from each cluster proportionally to its size. However, the approaches differ significantly in asymptotic time complexity.**

I. Introduction

SUBSAMPLING is a method that reduces a size of a dataset by selecting a subset from the original dataset. However, in many areas, including biomedicine and many others, we often face a kind of opposite problem, i. e. we obtain a sample of only insufficient size and would need to enlarge its size. That can be done, e. g., by one of the resampling methods such as bootstrapping or others, or we need to use various inference methods to estimate properties of the entire population that our dataset comes from.

While such a data size reduction could not sound meaningful for the first impression, there are various situations where subsampling makes sense or is even necessary.

Usually, we can distinguish between two kinds of subsampling. Firstly, when we do the subsampling, we cannot even in theory collect all possible observations of an entire population. Or, secondly, we can gain all possible observations or, furthermore, we have already got them, but for some reason, we have to reduce the number of observations that will be utilized.

A typical example of the first subsampling kind is one of the large fields of statistics, called *sampling*, where subsampling as a method of choice deals with an idea of an entire population and its parameters but is limited to an option of gaining data of only a (small) subset coming from the population. Then, regardless of whether the population is more or less virtual, getting the sample that belongs to the population is still a problem fulfilling the subsampling definition.

The motivations for the subsampling could also be different and usually arise from any impossibility to utilize the entire original dataset, as may be true for the latter family of the subsampling problems. Thus, the rank of those motivations varies from the lack of (computational) power to analyze the entire original dataset to the lack of economic sources, making it impossible to collect all values for each observation of the original sample, e. g. populating a new (important) variable is considered to enrich the original dataset but can be done only for a limited number of observations (of the sub-selected dataset).

As a motivation for our study, using online surveys, we collected an original dataset of patients suffering from COVID-19 and undergoing anti-COVID-19 vaccination. To study a time development of COVID-19 antibodies after the vaccination, it is necessary to check the blood levels of the patients' antibodies from time to time. However, no matter how helpful would be the checking of antibodies for each patient, our financial sources were limited (and the antibody kits for laboratory serology tests are relatively expensive), so we had to select a subset of patients from the original dataset, no greater than a maximal number of laboratory tests funded by our financial sources. Furthermore, since the subsample can be done in many ways, we wanted to keep all categories of all categorical variables well balanced, i. e., to keep their frequencies in the final subsample equal or at least near-equal.

All the motivations share the demand on the quality by which the subsampling is done. As is naturally feasible, we usually want to avoid the "garbage in, garbage out", also known as the GIGO paradigm, which means that we cannot expect great outputs whenever the inputs are of low quality. The same logic applies to subsampling if followed by whatever kind of another analysis uses the subsample as an input. Thus, the authors suggest replacing the "garbage in, garbage out" paradigm more positively with "great in, great out".

However, regardless of the primary motivation why do subsampling, there is always a demand to keep the data homogeneity in the sub-selected sample, corresponding to the original data. More technically spoken, assuming the dataset contains only categorical variables, the homogeneity means that all categories of all categorical variables are near-equally represented in the final subsample.

In case there is a response variable included in the dataset, a popular and well-established method called propensity scoring (or propensity matching) is usually performed to identify the "best" subset of a given size that harmonizes effect sizes of individual explanatory variables [1].

Nevertheless, when a response variable is missing in the data because e. g. is planned to measure its values rather only for observations in the subsample than for the entire original sample, the logistic regression model behind the propensity scoring could not be built at all (since the response variable is not available). In such a case, the methodology that could be used for subsampling differ from naive approaches such as random sampling, even-odd sampling [2], to more intuitive, rather manual than automated sampling based on matching the observations so that they are balanced in pairs (or larger groups than pairs) [3]. In other words, when a response variable, commonly participating as a key part of the subsampling quality checking, is not available in the dataset, it could be "substituted" by a metric that might control for the quality of the subsampling process.

To check how balanced the subsample is, some metrics could be used [4]. They usually assume that numerical variables – if any – were prior transformed to categorical ones following more or less complex categorization rule. There are several commonly used metrics describing the rate of the categorical variables' levels balance in a final sample [5] such as entropy, mutability, Gini impurity, Simpson index, Shannon-Wiener index, and other diversity metrics. A sum of squares of categories' frequencies also becomes very popular; it is somewhat similar to Shannon entropy but is scaled, so it cannot be greater than 1.0 at maximum.

Based on the metric choice, the lower (or, the higher) is the metric's value; the better balanced is the subsample. Thus, for example, considering Shannon entropy or sum of squares of categories' frequencies, a lower value means better balancing the subsample; i. e. the frequencies of the categories in the subsample are equal or at least near-equal.

In fact, the subsampling itself is a discrete optimization task since the selection of a final subsample from the original sample may be made using a finite number of ways, but some of them are better than others, taking into account there is a given metric, checking the subsampling quality (categorical variables' levels well balancing) that is about to be minimized.

In this study, we selected a subpopulation ($n = 400$) from a COVID-19 dataset (or original size $N = 698$) with a missing response variable, which was up to be collected later. Whereas the response variable was not available, there were 18 more (explanatory) variables of interest. First, numerical variables were categorized. The quality of subpopulation selecting was measured using a sum of squares of each variable's category frequency and averaged over all variables. Minimizing the metric reflects the demand for keeping all the variables' categories numerically balanced, i. e. of similar sizes. Several subset-selecting strategies were applied. Besides a single random subsampling, an exhaustive method selecting all possible combinations of $n = 400$ observations from initial $N = 698$ observations was performed, choosing the subsample that grand totally minimized the metric. Similarly, a "forward" subselection, reducing the original dataset by one observation per each step, permanently lowering the metric, was done. A repeated random subsampling enabled to model a prior distribution of the metric and helped estimate its empirical minimum, determining one given subsample. Finally, k-means clustering (with a random number of clusters) of the original dataset's observations and choosing for a subsample from each cluster, proportionally to its size, and also based on a joint occurrence of each pair in one cluster, also lowered the metric compared to the random subsampling.

The aim of this study is to demonstrate that all the approaches except for a single random subsample offer a valid alternative to exhaustive sampling grant-totally minimizing the chosen metric.

II. Proposed research methodology

There are overall research methodology and the formal description of the dataset, the metric chosen for controlling the quality of the subsampling, and the proposed methods of the subsampling discussed in the following subsections.

A. Formal description of a dataset used for subsampling

The original dataset consists of N rows containing one observation per row and k categorical variable in columns.

The subsampling task means selecting a subset of n rows and k columns, where $n < N$. Thus, the sampling is applied on rows, not on columns.

For each $i \in \{1, 2, 3, \ldots, k\}$, the variable i contains exactly n_i categories and the frequencies of the category j is $n_{i,j}$. We can easily show that for the original dataset, and the subsample is

$$\sum_{j=1}^{n_i} n_{i,j} = N \qquad \text{and} \qquad \sum_{j=1}^{n_i} n_{i,j} = n,$$

respectively, so the sum of frequencies of a given variable i's categories is equal to N in the original dataset and is equal to n in the subsample, respectively, and based on the context.

B. A metric for controlling the quality of the subsampling

The Shannon entropy is defined as

$$H_i = -\sum_{j=1}^{n_i} p_{i,j} \log p_{i,j}$$

where $p_{i,j}$ is a probability of a category j for each $j \in \{1, 2, 3, \ldots, n_i\}$ in a sample of n_i categories of a variable i. We can easily prove by Jensen's inequality the upper bound of the entropy H_i defined by such formula is dependent on the probabilities $p_{i,j}$. Also, the formula may struggle with zero probabilities, i. e. when $\exists j \in \{1, 2, 3, \ldots, n_i\}$ such that $p_{i,j} = 0$ since the term $\log p_{i,j}$ is not defined for $p_{i,j} = 0$.

To overcome these difficulties, we rather used a sum of squares of each variable's category frequency. Using the finite samples, probabilities are only estimated by their frequencies, therefore we will replace the probability $p_{i,j}$ by its unbiased estimate $\pi_{i,j} = \frac{n_{i,j}}{n} = \hat{p}_{i,j}$, where $n_{i,j}$ is a number of occurrence of category j of variable i in the sample of size n. The sum of squares of the variable i's category frequencies then follows as

$$S_i = \sum_{j=1}^{n_i} \pi_{i,j}^2 = \sum_{j=1}^{n_i} \left(\frac{n_{i,j}}{n}\right)^2 = \sum_{j=1}^{n_i} \hat{p}_{i,j}^2. \tag{1}$$

Finally, when there is more than one variable, i. e. $i \in \{1, 2, 3, \ldots, k\}$ then in order to take into account for each variable's sum of squares given by formula (1), we can calculate the average value $\bar{S}$ of the sums of squares for individual variables, so

$$\bar{S} = \frac{1}{k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\frac{n_{i,j}}{n}\right)^2. \tag{2}$$

Let us derive the lower and upper bound of the sum of squares for the variable i.

(i) Firstly, let us consider one of the two possible extreme scenarios – the sample is populated by only one category. More technically, let us assume that $\exists j^* \in \{1, 2, 3, \ldots, n_i\}$ so that $n_{i,j^*} = n_i$. Then, $\forall j \in \{1, 2, 3, \ldots, n_i\} \setminus j^*$ is $n_{i,j} = 0$ and, eventually, $\frac{n_{i,j^*}}{n} = 1$ and $\frac{n_{i,j}}{n} = 0$.

The sum of squares S_i then follows the term

$$\begin{aligned} S_i &= \sum_{j=1}^{n_i} \left(\frac{n_{i,j}}{n}\right)^2 = \\ &= \left(\frac{n_{i,j^*}}{n}\right)^2 + \sum_{j \in \{1,2,\ldots,n_i\} \setminus j^*} \left(\frac{n_{i,j}}{n}\right)^2 = \\ &= 1^2 + (n_i - 1) \cdot 0^2 = \\ &= 1. \end{aligned}$$

Thus, we derived the maximum value of the sum of squares S_i for the variable i is equal to 1.

(ii) Now suppose the other extreme scenario – all categories are equally populated in the sample and no one of the categories occurred more than once. So, in other words,

$$\frac{n_{i,1}}{n} = \frac{n_{i,2}}{n} = \cdots = \frac{n_{i,n_i}}{n} = \frac{1}{n}$$

and also $n_i = n$.

The sum of squares S_i then follows as

$$\begin{aligned} S_i &= \sum_{j=1}^{n_i} \left(\frac{n_{i,j}}{n}\right)^2 = \sum_{j=1}^{n_i} \left(\frac{1}{n}\right)^2 = \\ &= \sum_{j=1}^{n} \left(\frac{1}{n}\right)^2 = \\ &= n \cdot \left(\frac{1}{n}\right)^2 = \\ &= \frac{1}{n}. \end{aligned}$$

So, we derived the minimum value of the sum of squares S_i for the variable i is equal to $\frac{1}{n}$, where n is the size of a sample containing only categories of the variable i.

Concluding this up, we derived that for each variable i and its sample size n is the sum of squares S_i of the variable's category frequencies lower then or equal to 1 and greater than or equal to $\frac{1}{n}$, more formally $\frac{1}{n} \leq S_i \leq 1$.

Going back to the idea of a well-balanced subsample, all category frequencies of all variables in the subsample should be of (near) equal sizes. That is a situation very close to scenario (ii) with balanced frequencies $\frac{1}{n}$ above – on the other hand, the frequencies in scenario (i) are imbalanced. Assuming this, the sum of squares S_i of the variable's category frequencies in the well-balanced subsample should be as low as possible and should approach the $\frac{1}{n}$. Finally, if all the variable would minimize their sums of squares, then also the average value $\bar{S}$ of all the sums of squares should be minimal.

Keeping the subsample well balanced, i. e. ensuring the categories of all the variables in the subsample are of (near) equal frequencies, means lowering the average value $\bar{S}$ of the sums of squares as much as possible. In theory, the minimal possible value of the average value $\bar{S}$ of the sums of squares is

$$\bar{S} = \frac{1}{k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\frac{n_{i,j}}{n}\right)^2 \geq \frac{1}{k} \sum_{i=1}^{k} \frac{1}{n} = \frac{1}{k} \cdot \frac{k}{n} = \frac{1}{n}.$$

In practise, assuming the categories are well balanced for each variable, i. e. for each $i \in \{1, 2, 3, \dots, k\}$ is $n_{i,1} \approx n_{i,2} \approx \dots \approx n_{i,n_i}$, then $\sum_{j=1}^{n_i} n_{i,j} = n \approx n_i \cdot n_{i,j}$ and so $\frac{n_{i,j}}{n} \approx \frac{n/n_i}{n} \approx \frac{1}{n_i}$, we can expect rather

$$\bar{S} = \frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\frac{n_{i,j}}{n}\right)^2 \gtrsim \frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\frac{1}{n_i}\right)^2 \approx$$
$$\approx \frac{1}{k}\sum_{i=1}^{k} n_i\left(\frac{1}{n_i}\right)^2 \approx \frac{1}{k}\sum_{i=1}^{k}\frac{n_i}{n_i^2} \approx \frac{1}{k}\sum_{i=1}^{k}\frac{1}{n_i}.$$

Eventually, what worth to be mentioned, is a comparison of each variable's sum of squares S_i given by formula (1) and Gini impurity. Using still the same mathematical notation, then Gini impurity is defined as

$$G_i = 1 - \sum_{j=1}^{n_i}\pi_{i,j}^2 = 1 - \sum_{j=1}^{n_i}\left(\frac{n_{i,j}}{n}\right)^2 = 1 - \sum_{j=1}^{n_i}\hat{p}_{i,j}^2,$$

which is obviously equal to $S_i = 1 - G_i$. That being written, using the Gini impurity G_i in this study instead of the sum of squares S_i would return exactly the same results (as far as the sign of Gini impurity is opposite than the one of the sum of squares and shifted by 1.0).

C. Single random subsampling without replacement

The term of random subsampling without replacement means that each observation of the original dataset has only one chance to be selected in the subsample.

If we subsample the original dataset of size N to a dataset of size n only once, there are in theory $\binom{N}{n}$ options how to do the random subsampling. Assuming one of the subsamples[1] minimizing the averaged sums of squares $\bar{S}$, the probability of randomly hitting such a subsample is about $\frac{1}{\binom{N}{n}} \simeq 0$ for large $N > n$.

An expected value of the averaged sums of squares $\bar{S}$, calculated using the obtained subsample, is in between the expected value of the worst-case scenario, 1, and the best-case scenario, $\frac{1}{n}$, so $\frac{1}{n} \leq \mathbb{E}(\bar{S}) \leq 1$.

The asymptotic time complexity is easy to derive, $\Theta(1)$, assuming the random subset generating costs 1 unit of complexity time.

D. Repeated random subsampling without replacement

Similarly to the previous approach, here we repeat the random subsampling $m > 1$ times.

The repetition of the random subsampling enables us to estimate an expected value $\hat{\mathbb{E}}(\bar{S})$ of the averaged sums of squares $\bar{S}$ and standard deviation $\sqrt{\hat{\text{var}}(\bar{S})}$, using the values of m obtained subsamples. Assuming the Ljapunov's version of the central limit theorem, the derived variable $\frac{\bar{S}-\hat{\mathbb{E}}(\bar{S})}{\sqrt{\hat{\text{var}}(\bar{S})}}$ follows standard normal distribution, formally $\frac{\bar{S}-\hat{\mathbb{E}}(\bar{S})}{\sqrt{\hat{\text{var}}(\bar{S})}} \sim \mathcal{N}(0, 1^2)$. That helps us to estimate the minimum value of the averaged sums of squares $\bar{S}$ following way. Supposing there $\frac{\bar{S}-\hat{\mathbb{E}}(\bar{S})}{\sqrt{\hat{\text{var}}(\bar{S})}} \sim \mathcal{N}(0, 1^2)$ holds, we know that

$$P\left(\frac{\bar{S}-\hat{\mathbb{E}}(\bar{S})}{\sqrt{\hat{\text{var}}(\bar{S})}} \leq u_{0.01}\right) = 0.01,$$

where $u_{0.01}$ is a 0.01-th quantile of the standard normal distribution. Continuing in the derivations, we get

$$P\left(\bar{S} \leq \hat{\mathbb{E}}(\bar{S}) - |u_{0.01}|\sqrt{\hat{\text{var}}(\bar{S})}\right) = 0.01, \tag{3}$$

so approximately, the minimum value of $\bar{S}$ is very likely close to the term of $\hat{\mathbb{E}}(\bar{S}) - |u_{0.01}|\sqrt{\hat{\text{var}}(\bar{S})}$. Utilizing this piece of information, we can not only estimate the minimum value of the averaged sums of squares $\bar{S}$, but can also highlight the subsample approaching this minimum value (surely it is the subsample with minimal value – somewhat close to the subtraction $\hat{\mathbb{E}}(\bar{S}) - |u_{0.01}|\sqrt{\hat{\text{var}}(\bar{S})}$ from the positive direction – of the averaged sums of squares $\bar{S}$ in the set of all m generated subsamples).

The asymptotic time complexity of the (m times) repeated random subsampling without replacement is $\Theta(m)$, again assuming the random subset generating costs 1 unit of complexity time. The pseudocode of the repeated random subsampling process is in Algorithm 1.

E. Exhaustive subsampling

The method of exhaustive subsampling is based on greedy generating all possible subsamples of size n from the original dataset of size $N > n$.

In theory, there are $\binom{N}{n} = \frac{n!}{k!(n-k)!}$ ways how a subsample of size n could be sampled from the dataset of size N. It implies there is also $\binom{N}{n}$ values of the averaged sums of squares $\bar{S}$ (one value per each subsample), but the values are not necessarily different.

Regardless of that, this approach enables to convenient pick the subsample with a minimum possible value of the averaged sums of squares $\bar{S}$ (no other subsample could practically have the value of the averaged sums of squares $\bar{S}$ lower).

However, there is an obvious trade-off between the possibility to reach the practical minimum of the value of the averaged sums of squares $\bar{S}$ and asymptotic time complexity, which is enormous, $\Theta\left(\binom{N}{n}\right) = \Theta\left(\frac{n!}{k!(n-k)!}\right)$, assuming the random subset generating costs 1 unit of complexity time.

F. Subsampling by forwarding step-by-step size reduction of the original dataset

The logic of the step-by-step size reduction of the original dataset by permanent lowering a value of the averaged sums of squares $\bar{S}$ is based on random selection of such an observation that its removing from the original dataset tends to decrease (or at least not increase) a value of the averaged sums of squares $\bar{S}$. Thus, we also call this approach as *one-by-one* observation's sample reduction of also as *row-by-row* observation's sample reduction. Let's define a size of the dataset after τ steps, i. e. after removing of τ observations, as $n(\tau)$,

[1] Theoretically, there could be more than one subsample with the same but minimal value of the metric of the averaged sums of squares $\bar{S}$.

Algorithm 1: Repeated random subsampling without replacement and estimating of the minimum value of the averaged sums of squares $\bar{S}$, together with highlighting of the subsample minimizing the averaged sums of squares $\bar{S}$

```
Data: an original dataset of size $N$ containing $k$ variables
Result: a set of $m$ random subsamples of size $n < N$, an estimate of the minimum value of the averaged sums of squares $\bar{S}$ and highlighting of the subsample minimizing the averaged sums of squares $\bar{S}$
$N$   // size of the original dataset ;
$n$   // size of the subsample;
$m$   // number of repetitions of;
      // subsampling;
$\mathcal{S}$   // a tuple of subsamples of size $n$;
$\mathcal{A}$   // a tuple of averaged sums of squares $\bar{S}$;
for $\ell = 1 : m$ do
    generate a random subsample $\int$ of size $n$ without replacement from the original dataset of size $N$ and calculate its averaged sums of squares $\bar{S}$ ;
    $\mathcal{S} = \{\mathcal{S}, \int\}$;
    $\mathcal{A} = \{\mathcal{A}, \bar{S}\}$;
end
find the minimum of $\mathcal{A}$ and a corresponding subsample with $\bar{S} = \min\{\mathcal{A}\}$ ;
calculate an estimate $\hat{\mathbb{E}}(\bar{S})$ and $\hat{\text{var}}(\bar{S})$ ;
calculate the estimated minimum of $\bar{S}$ as $\hat{\mathbb{E}}(\bar{S}) - |u_{0.01}|\sqrt{\hat{\text{var}}(\bar{S})}$ ;
compare $\min\{\mathcal{A}\}$ and $\hat{\mathbb{E}}(\bar{S}) - |u_{0.01}|\sqrt{\hat{\text{var}}(\bar{S})}$ ;
```

and the averaged sums of squares $\bar{S}$ after τ steps as $\bar{S}(\tau)$. Evidently, $n(0) = N$, $n(1) = N - 1$, $n(2) = N - 2$, $\dots$, $n(N - n) = N - (N - n) = n$. Analogously, we demand on $\bar{S}(\tau + 1) \leq \bar{S}(\tau)$ for each $\tau \in \{0, 1, 2, \dots, N - n - 1\}$.

It is easy to demonstrate that $\bar{S}(N - n) \leq \bar{S}(0)$, i. e. the averaged sums of squares $\bar{S}$ after $N - n$ steps (when dataset size is n) is lower than or equal to the value of the averaged sums of squares $\bar{S}$ in the beginning. Assuming the initial original dataset is not well balanced, then $\bar{S}(N - n) < \bar{S}(0)$ or even $\bar{S}(N - n) \ll \bar{S}(0)$. Based on the fact the selection of one observation per each step is random (until it leads to decreasing of the averaged sums of squares $\bar{S}$ value), a deterministic value of $\bar{S}(N - n)$ is not possible to calculate.

Let us suppose the random selection of the observation tending to reduce the averaged sums of squares $\bar{S}(\tau)$ in the $(\tau + 1)$-th step (so that $\bar{S}(\tau + 1) \leq \bar{S}(\tau)$), when the dataset contains exactly $N - \tau$ observations, would take averagely about $(N - \tau)/2$ samplings. Then the average asymptotic time complexity [6] of the row-by-row reduction of the original dataset by permanent lowering a value of the averaged sums of squares $\bar{S}$ is $\Theta(\bullet)$, so that

$$\begin{aligned}
\Theta(\bullet) &= \Theta\left(\sum_{\tau=0}^{N-n-1} (N - \tau)/2\right) = \\
&= \Theta\left(\frac{1}{2}\sum_{\tau=0}^{N-n-1} (N - \tau)\right) = \\
&= \Theta\left(\frac{1}{2}\left(\sum_{\tau=0}^{N-n-1} N - \sum_{\tau=0}^{N-n-1} \tau\right)\right) = \\
&= \Theta\left(\frac{1}{2}\left((N - n)N - \frac{(N - n - 1)(N - n)}{2}\right)\right) = \\
&= \Theta\left(\frac{1}{2}\left(\frac{(N - n)(N + n + 1)}{2}\right)\right) = \\
&= \Theta\left((N - n)(N + n + 1)\right) \approx \\
&\approx \Theta(N^2).
\end{aligned}$$

The pseudocode of the subsampling by row-by-row reduction of the original dataset is in Algorithm 2.

Algorithm 2: Subsampling by row-by-row reduction of the original dataset, decreasing the value of the averaged sums of squares $\bar{S}$ per each step

```
Data: an original dataset of size $N$ containing $k$ variables
Result: a subsample minimizing the averaged sums of squares $\bar{S}$
$N$     // size of the original dataset;
$n$     // size of the subsample;
$n_t$   // current size of the dataset;
$\bar{S}$     // current averaged sums of squares;
$n_t = N$;
while $n_t > n$ do
    while $\bar{S}$ after removing the random observation $\geq \bar{S}$ do
        pick another random observation from the current dataset of size $n_t$ (# of observations)
    end
    remove the picked observation from the dataset;
    $n_t = n_t - 1$;
    update $\bar{S}$;
end
use the subsample of size $n$;
```

G. Subsampling using clustering

An idea behind the subsampling using unsupervised learning of clustering kind is to utilize the fact that observations within each cluster are similar enough, while observations between each cluster are different enough. Thus, when we require subsamples with well-balanced category frequencies for each variable, we should consider observations from different clusters when creating the final subsample. Thus, a big question is *how* to pick the observations from different clusters to ensure the final subsample of a given size is well balanced.

The paper's authors suggest several ideas on how to use clusters for subsampling and, particularly, how to draw the observations from existing clusters when the final subsample is constructed.

Firstly, regardless of the fact the observations are picked randomly or following some pattern from d clusters of sizes $|c_1|$, $|c_2|$, …, $|c_d|$, a number of observations picked from the cluster $\delta \in \{1, 2, \ldots, d\}$ should be proportional to its size, $|c_\delta|$.

Let us assume that a number of category frequencies of a variable i that are greater than zero is η_i in a given cluster δ. A total count of categories of a variable i is, following the previous notation, n_i, and a mean frequency for average category is about $\frac{|c_\delta|}{n_i}$. As we can see, the mean frequency is proportional to the cluster size $|c_\delta|$. In other words, the larger is the cluster (the larger is $|c_\delta|$), more categories would get non-zero frequency. Assuming the count of the variable i's categories with non-zero frequency is η_i in a given cluster δ, the η_i is proportional to $|c_\delta|$, $\eta_i \propto |c_\delta|$, and those frequencies are roughly similar, i. e. $n_{i,j} \approx \frac{n}{\eta_i} \approx \frac{|c_\delta|}{\eta_i}$, we can derive

$$\begin{aligned} S_i &= \sum_{j=1}^{n_i} \left(\frac{n_{i,j}}{n}\right)^2 \propto \sum_{j=1}^{\eta_i} \left(\frac{|c_\delta|/\eta_i}{|c_\delta|}\right)^2 \propto \\ S_i &\propto \sum_{j=1}^{\eta_i} \frac{1}{\eta_i^2} \propto \sum_{j=1}^{\eta_i} \frac{1}{|c_\delta|^2} \propto \\ &\propto \sum_{j=1}^{|c_\delta|} \frac{1}{|c_\delta|^2} \propto |c_\delta| \cdot \frac{1}{|c_\delta|^2} \propto \\ &\propto \frac{1}{|c_\delta|}, \end{aligned} \tag{4}$$

that supports our suggestion to draw observations from the clusters proportionally to their sizes[2], i. e. the larger the cluster is, the more observations should be picked from the cluster towards the final subsample to minimize the sum of squares S_i.

Secondly, we also propose an experimental approach that requires another ongoing research. Considering the (not necessarily k-means) clustering is repeated m times, with a random number of clusters in each of m iterations, we can construct a symmetric square matrix T of dimensions $N \times N$, that for the p-th row and the q-th column describes a number of times that the p-th observation of the original dataset was together in the same cluster with the q-th observation of the original dataset. The matrix T follows a form of

$$T = \begin{pmatrix} t_{1,1} & t_{1,2} & \cdots & t_{1,N} \\ t_{2,1} & t_{2,2} & \cdots & t_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ t_{N,1} & t_{N,2} & \cdots & t_{N,N} \end{pmatrix}, \tag{5}$$

where $t_{p,q}$ stands for a number of times both the p-th observation and q-th observation of the original dataset were together in the same cluster.

Once we want to construct a subsample of size n from the original dataset of size N, we demand on keeping all variables' each category frequency balanced with other frequencies, so the final subsample should include all categories of all variables with (near) similar frequencies, if possible. Drawing such observations that were many times together in the same clusters within the multiple clustering procedure would result in the final subsample containing too many similar observations, which would reduce the native variability of the variables.

Consequently, the final subsample should be constructed using observations that are mutually non-similar. The way of constructing such a subsample could be to pick two original observations with a minimum value of $t_{p,q}$ and then add to the subsample one by one new observation (until the subsample size is sufficient) such that each new one (the q-th) has the minimum value of

$$\sum_{\forall p \in \{\text{observations in subsample}\}} t_{p,q},$$

so that

$$q = \operatorname{argmin}_{q \in \{1,2,\ldots,N\}} \sum_{\forall p \in \{\text{observations in subsample}\}} t_{p,q}, \tag{6}$$

that minimizes a chance of getting a subsample with too much similar observations. While this approach may look as completely deterministic, it contains a part that is based on randomness, namely the clustering part.

Adopting the time complexity of the m times repeated k-means clustering for small number of clusters is $\Theta(m \cdot N \cdot k)$ [7] and for the T matrix construction (5), the ongoing part using the formula (6) takes averagely $\Theta(n^2)$ complexity time units.

Whereas the clustering algorithm itself could vary (it is not necessary to apply only k-means algorithm), it is worth to be mentioned that – since the variables in the original dataset are categorical (or transformed into categorical ones) – the Gower distance was chosen for the clustering as it can handle categorical variables well within the clustering [8].

The pseudocode of the subsampling by clustering the original dataset is in Algorithm 3.

III. Results

We used COVID-19 survey data of our provenience for the application of the proposed methods. The original dataset contains $N = 698$ rows corresponding to observations and $k = 18$ columns related to variables. Since the dataset is of

[2] The proportional equation (4) might be confusingly understood as to pick a maximum of observations (towards the final subsample) from the larger cluster since this would lead to the minimization of the sum of squares S_i for the given variable. However, such a subsample would be constructed using almost only one of the clusters – the largest one – and thus, tends to include very similar observations, which could break the demand of well-balanced category frequencies over all variables.

Algorithm 3: Subsampling by clustering the original dataset using the matrix T of mutual occurrences in the same clusters as in (5)

```
Data: an original dataset of size $N$ containing $k$
      variables and $T$ matrix of mutual occurrences in
      the same clusters as in (5)
Result: a subsample minimizing the averaged sums of
        squares $\bar{S}$
$n$      // size of the subsample;
$n_t$    // current size of the dataset;
$T$      // matrix of mutual occurrences in;
         // the same clusters;
$\bar{S}$      // current averaged sums of squares;
$\mathcal{S}$      // current subsample;
populate the subsample $\mathcal{S}$ by the first two
  observations corresponding to row and column
  indices of minimum of $T$;
$n_t = 2$;
while $n_t < n$ do
    pick the $q$-th observation such that
        $q = \operatorname{argmin}_{q \in \{1,2,\ldots,N\} \setminus \mathcal{S}} \sum_{\forall p \in \mathcal{S}} t_{p,q}$,
    where $t_{p,q}$ is the value of $p$-th row and $q$-th
    column of the matrix $T$ ;
    $n_t = n_t + 1$;
    update $\bar{S}$;
end
use the subsample $\mathcal{S}$ of size $n$;
```

a questionnaire form including questions with the close format, the vast majority of the variables are categorical. A few of the numerical variables were categorized following experts' suggestions or natural logic, e. g. age was categorized into intervals of lengths ten years, starting and ending at an age divisible by a number 10, etc. Applying this approach, there are only categorical variables in the original dataset before the subsampling. The reason why the response variable, i. e. the serology levels of COVID-19 antibodies, is missing in the original dataset is that patients involved in the study were planned to undergo relatively expensive serology tests; thus, the original size ($N = 698$) had to be reduced significantly ($n = 400$) to keep the costs of the serology testing manageable.

The task was to get a subsample of $n = 400$ rows from the original dataset, containing the original number of $k = 18$ variables.

All the computations were performed using R programming language and environment [9]. There are more numerical applications of R language to various fields in [10]–[15].

We applied all the methods mentioned above to do the subsampling and compare the results using the metric controlling the quality of the subsampling in between the methods.

The metric of the subsampling quality, depicting particularly how well the category frequencies of all the variables are balanced, is the averaged sums of squared $\bar{S}$ as defined in (2).

Besides the single random subsampling without replacement, we started with the repeated random subsampling without replacement. Repeating the random subsampling multiple times ($m = 100$) enables modeling the prior distribution of the averaged sums of squared $\bar{S}$, and was also used for estimation of the minimum value of averaged sums of squares $\bar{S}$ using the formula (3).

Histogram of the prior distribution of the averaged sums of squared $\bar{S}$ is in figure 1. The minimum value of averaged sums of squares $\bar{S}$ was estimated following the (3) to be equal $\hat{\bar{S}} \doteq 0.247$.

The next method, subsampling by forwarding one-be-one reduction of the original dataset, was also performed $m = 100$ times. Histogram of the prior distribution of the averaged sums of squared $\bar{S}$ is in figure 2. The minimum value of averaged sums of squares $\bar{S}$ for the one-be-one size reduction that was obtained is equal to $\hat{\bar{S}} \doteq 0.242$.

The subsampling by clustering the original dataset was performed $m = 100$ times, too. The final subsample was designed using the T matrix (5) and creating the subsample from scratch using the logic of formula (6). Histogram of the prior distribution of the averaged sums of squared $\bar{S}$ is in figure 3. The minimum value of averaged sums of squares $\bar{S}$ for the one-be-one size reduction that was obtained is equal to $\hat{\bar{S}} \doteq 0.245$.

As we can see, all the three applied methods return similar accuracy considering the minimization of the averaged sums of squares. The formal comparison of statistical differences in between mean values of the averaged sums of squares $\bar{S}$ for the repeated ($m = 100$) random subsampling without replacement, repeated ($m = 100$) subsampling by forwarding one-be-one reduction of the original dataset, and the repeated ($m = 100$) subsampling by clustering the original dataset could be performed using one-way analysis of variance (ANOVA). However, considering the figures 1, 2 and 3, the practical differences are minimal. What practically differs is the asymptotic time complexity of the mentioned techniques, as discussed before.

IV. Conclusion

Subsampling may be an important task when the original dataset is larger than required. If there is a response variable available in the dataset, then the methodology used for the subsampling is well established; the popular propensity scoring is used to extract the subsample from the original data that harmonize size effects of all predictors using logistic regression model.

When the response variable from some reason or another is missing, e. g. is planned to be collected later, the methodology of the subsampling is not so straightforward. Many various

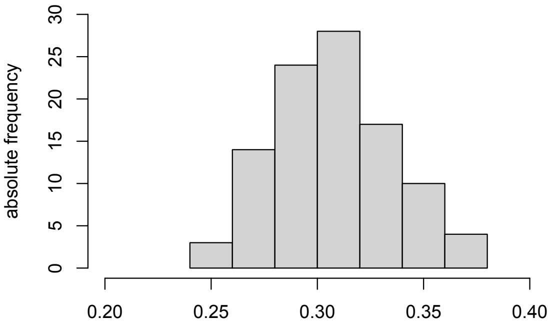

Fig. 1. Histogram of the prior distribution of the averaged sums of squares $\bar{S}$ calculated for the repeated ($m = 100$) random subsampling without replacement.

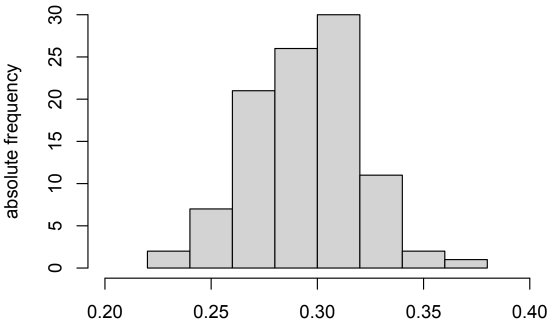

Fig. 2. Histogram of the prior distribution of the averaged sums of squares $\bar{S}$ calculated for the repeated ($m = 100$) subsampling by forwarding one-be-one reduction of the original dataset.

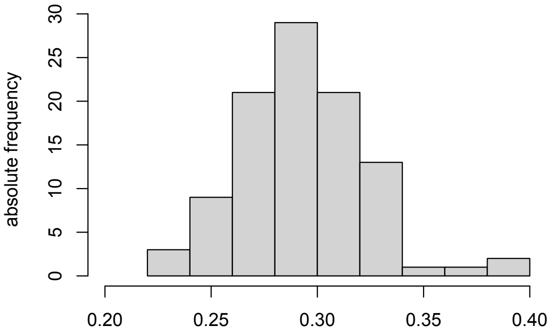

Fig. 3. Histogram of the prior distribution of the averaged sums of squares $\bar{S}$ calculated for the repeated ($m = 100$) subsampling by clustering the original dataset.

methods of low significance are used, based on different approaches – from totally random subsampling to manually matched pairs of observations with balanced all variables' category frequencies.

In this study, we proposed one metric – the averaged sums of squares – enabling to control a quality of the subsampling, including the fact the metric is in theory scaled to an interval not dependent on entry data, as was proven. Furthermore, we compared several methods; some of them are novel and proposed by this paper.

While the repeated random subsampling without replacement is relatively fast method, it can reach the minimum of the averaged sums of squares only approximately. The subsampling using one-by-one reduction of the original sample is a bit slower than the random multiple subsampling, but still feasibly applicable; it can approach the minimum of the averaged sums of squares only approximately, too. The exhaustive subsampling as the only one method can numerically calculate the exact value of the minimum of the averaged sums of squares; however, its executing time is enormously high. Finally, the subsampling by clustering is an innovative method that is relatively fast if implemented using standard algorithms and maturated computational environments, and furthermore, it offers a way to keep control over the mutual occurrences of each two observations from the same clusters, when the final subsample is constructed. Even the subsampling by clustering approached the minimum of the averaged sums of squares relatively closely.

All the proposed methods, i. e. repeated random subsampling without replacement, subsampling using one-by-one reduction of the original dataset and subsampling by clustering seem to be valid alternatives to exhaustive subsampling.

V. Acknowledgement

This paper is supported by the grant OP VVV IGA/A, CZ.02.2.69/0.0/0.0/19_073/0016936 with no. 18/2021, which has been provided by the Internal Grant Agency of the Prague University of Economics and Business.

References

[1] Peter C. Austin. "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies". In: *Multivariate Behavioral Research* 46.3 (May 2011), pp. 399–424. DOI: 10.1080/00273171.2011.568786. URL: https://doi.org/10.1080/00273171.2011.568786.

[2] Santhosh Pathical and Gursel Serpen. "Comparison of subsampling techniques for random subspace ensembles". In: *2010 International Conference on Machine Learning and Cybernetics*. IEEE, July 2010. DOI: 10.1109/icmlc.2010.5581032. URL: https://doi.org/10.1109/icmlc.2010.5581032.

[3] Elizabeth A. Stuart. "Matching Methods for Causal Inference: A Review and a Look Forward". In: *Statistical Science* 25.1 (Feb. 2010). DOI: 10.1214/09-sts313. URL: https://doi.org/10.1214/09-sts313.

[4] Sarda Sahney, Michael J. Benton, and Paul A. Ferry. "Links between global taxonomic diversity, ecological diversity and the expansion of vertebrates on land". In: *Biology Letters* 6.4 (Jan. 2010), pp. 544–547. DOI: 10.1098/rsbl.2009.1024. URL: https://doi.org/10.1098/rsbl.2009.1024.

[5] David MacKay. *Information theory, inference, and learning algorithms*. Cambridge, UK New York: Cambridge University Press, 2003. ISBN: 0-521-64298-1.

[6] Lubomír Štěpánek, Filip Habarta, Ivana Malá, et al. "Analysis of asymptotic time complexity of an assumption-free alternative to the log-rank test". In: *Proceedings of the 2020 Federated Conference on Computer Science and Information Systems*. IEEE, Sept. 2020. DOI: 10.15439/2020f198. URL: https://doi.org/10.15439/2020f198.

[7] Malay K. Pakhira. "A Linear Time-Complexity k-Means Algorithm Using Cluster Shifting". In: *2014 International Conference on Computational Intelligence and Communication Networks*. IEEE, Nov. 2014. DOI: 10.1109/cicn.2014.220. URL: https://doi.org/10.1109/cicn.2014.220.

[8] J. C. Gower. "A General Coefficient of Similarity and Some of Its Properties". In: *Biometrics* 27.4 (Dec. 1971), p. 857. DOI: 10.2307/2528823. URL: https://doi.org/10.2307/2528823.

[9] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2017. URL: https://www.R-project.org/.

[10] Lubomír Štěpánek, Pavel Kasal, and Jan Měšťák. "Evaluation of facial attractiveness for purposes of plastic surgery using machine-learning methods and image analysis". In: *2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom)*. IEEE, Sept. 2018. DOI: 10.1109/healthcom.2018.8531195. URL: https://doi.org/10.1109/healthcom.2018.8531195.

[11] Lubomír Štěpánek, Pavel Kasal, and Jan Měšťák. "Machine-learning at the service of plastic surgery: a case study evaluating facial attractiveness and emotions using R language". In: *Proceedings of the 2019 Federated Conference on Computer Science and Information Systems*. IEEE, Sept. 2019. DOI: 10.15439/2019f264. URL: https://doi.org/10.15439/2019f264.

[12] Lubomír Štěpánek, Pavel Kasal, and Jan Měšťák. "Machine-Learning and R in Plastic Surgery – Evaluation of Facial Attractiveness and Classification of Facial Emotions". In: *Advances in Intelligent Systems and Computing*. Springer International Publishing, Sept. 2019, pp. 243–252. DOI: 10.1007/978-3-030-30604-5_22. URL: https://doi.org/10.1007/978-3-030-30604-5_22.

[13] Lubomír Štěpánek, Pavel Kasal, and Jan Měšťák. "Machine-learning at the service of plastic surgery: a case study evaluating facial attractiveness and emotions using R language". In: *Proceedings of the 2019 Federated Conference on Computer Science and Information Systems*. IEEE, Sept. 2019. DOI: 10.15439/2019f264. URL: https://doi.org/10.15439/2019f264.

[14] Lubomír Štěpánek, Pavel Kasal, and Jan Měšťák. "Evaluation of Facial Attractiveness after Undergoing Rhinoplasty Using Tree-based and Regression Methods". In: *2019 E-Health and Bioengineering Conference (EHB)*. IEEE, Nov. 2019. DOI: 10.1109/ehb47216.2019.8969932. URL: https://doi.org/10.1109/ehb47216.2019.8969932.

[15] Lubomír Štěpánek, Filip Habarta, Ivana Malá, et al. "A Machine-learning Approach to Survival Time-event Predicting: Initial Analyses using Stomach Cancer Data". In: *2020 International Conference on e-Health and Bioengineering (EHB)*. IEEE, Oct. 2020. DOI: 10.1109/ehb50910.2020.9280301. URL: https://doi.org/10.1109/ehb50910.2020.9280301.