Journal of Environmental Statistics

September 2018, Volume 8, Issue 2. http://www.jenvstat.org

# Rater Classification by Means of Set-theoretic Methods Applied to Forestry Data

**Dietrich Stoyan** **Arne Pommerening** **Andreas Wünsche**

**Abstract**

We consider a situation where $r$ raters select subsets from a set of $n$ items by marking them by '0' or '1', as in classification problems, approval voting and in general subset voting. The number $r$ of raters is small in comparison to the number $n$ of items. We intend to classify the raters, to understand their behavior and to go beyond the possibilities of classical statistical methods such as Fleiss' kappa, cluster analysis or latent class analysis.

We use a non-parametric set-theoretic approach, which is natural for the given dichotomous setting. We recommend the determination of a set-theoretic mean, the Vorob'ev expectation, to play a role similar to the classical mean of a sample. In particular, we use distances of the raters' subsets from the mean as characteristics of the individual raters.

Furthermore, we introduce a new measure of conformity of a given rater with all others, characterizing the extent to which the rater deviates from the whole group of raters.

We demonstrate the use of these methods in a case study, where the raters are forest managers and the items are trees in a forest thinning experiment. Our aim is to contribute to an understanding of the psychological processes involved, when forest managers mark trees for forest operations.

Keywords: binary measurements, conformity of raters, marking of forest trees, expectation of random set, random set, rater classification in forestry.

## 1. Introduction

Many problems in environmental, psychological and multivariate statistics (Cichetti, 1994; Hallgreen, 2012; Sheskin, 1997) are related to the following situation:

There is a group of $r$ raters or voters and $n$ subjects, items or candidates. (In the following we speak about 'raters' and 'items'.) Every rater evaluates every item by marking approvals by '1' and disapprovals by '0', i.e., the evaluation or measurement is dichotomous. The array of $r \times n$ marks '0' and '1' is considered as data and analyzed in various ways by well-known statistical methods. In the present paper $r$ is relatively small and $n$ comparatively large, i.e., a relatively small number of raters rates a large number of items.

Several standard statistical analysis methods aim to make inference of the behavior of raters. We give four examples.

– *Cochran's Q test* considers the question whether there are significant differences in rater activity (Sheskin, 1997). If the null-hypothesis that all raters assign mark 1 with equal frequencies is rejected by the chi-square test, it can be assumed that at least two of the $r$ raters have significantly different rating activities. These raters can be identified by pairwise comparisons using the McNemar test.
– *Fleiss' kappa* is a numerical statistical characteristic which helps to characterize rater agreement in the whole rater group (Fleiss et al., 2003). It has the character of a correlation coefficient: $\kappa = 1$ means complete agreement and $\kappa = -1$ disagreement; the case $\kappa = 0$ is related to rating by chance. By its nature, $\kappa$ does not provide information on the individual behavior of single raters (Stoyan et al., 2018).
– *Cluster analysis* (see e.g., Everitt et al., 2011) applied in the present setting yields clusters or classes of raters with similar behavior. Raters with close inter-rater distances (measured for example by the Manhattan metric, which is suitable for dichotomous data), are arranged in clusters. This approach in the context of the rating problem was used already by Schouten (1982) and Stoyan et al. (2018).

None of these methods directly helps to understand the behavior of individual raters, i.e., to understand the formation of classes, clusters or groups obtained by classification methods. This understanding is not an easy problem, in particular if only the dichotomous data are available without any additional information from covariates.

In contrast, *latent class analysis, LCA* (see Collins and Lanza, 2010, and Uebersax and Grove, 1993) allows an individual characterization of raters, however, in our case in an indirect way only: because the number of raters $r$ is small in comparison to the number of items LCA cannot be used to form groups of raters. Nevertheless, it can be applied to form two groups of items that consist of the items with 'true' marks '1' and '0', respectively, and then to determine the diagnostic power of raters. Statistical indicators for this purpose are the so-called *specificity* and *sensitivity* (Uebersax and Grove, 1993), where specificity is the probability of a negative rating given a negative case and sensitivity is the probability of a positive rating given a positive case. However, the task of interpreting these indicators remains.

The present paper suggests non-parametric solutions to the problem of explaining the individual behavior of single raters. Adapted to the dichotomous nature of the data we use set-theoretic ideas and try to exhaust the information given in this way. Indeed, every rater $i$ forms a subset $X_i$ of the set $E$ of all items, simply the set of items marked by '1'. This approach is inspired by the idea of subset voting (Regenwetter et al., 2006). In subset voting, stochastic random-set models exist, e.g., the 'size-independent model of approval voting' (Regenwetter et al., 2006). However, the present paper uses a non-parametric approach.

We assume that the $X_i$ are realizations of a finite discrete random set $X$. We recommend the use of a random-set mean, the empirical Vorob'ev expectation $\overline{X}$, as a result of the rating process, produced by the 'wisdom of rater crowd'. It supports the determination of distances of single raters from $\overline{X}$. Finally, we introduce a measure for the conformity of single raters with the whole group of all raters.

Our paper is organized as follows. Section 2 presents the set-theoretic methods, namely the Vorob'ev expectation $\overline{X}$, the distances $\delta_i$ of the raters from this expectation, and the conformity numbers $c_i$. Then, in order to motivate the new set-theoretic approach we consider a data example from environmental statistics, namely from forest management. It is described in detail in Section 3. In

this example, $r = 15$ forest managers evaluate $n = 387$ trees. This includes the application of the classical statistical methods mentioned above and some discussion of their results. Then the new summary characteristics are presented for the forest data. In the last Section 4 we then discuss the benefit of the application of the set-theoretic methods.

# 2. Set-theoretic statistics

## 2.1. Description of sets

There is a set $E$ of $n$ elements, $\{1, 2, \ldots, n\}$, in rating terminology the set of all items. We consider subsets of $E$, which we call 'sets.'

The standard descriptor of a set $A$ is its indicator function $\mathbf{1}_A(\cdot)$ defined by

$$\mathbf{1}_A(j) = \begin{cases} 1 & \text{for } j \in A \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

The number of elements of a set $A$ is denoted by $|A|$,

$$|A| = \sum_{j=1}^{n} \mathbf{1}_A(j).$$

*Distances* between sets are defined in a set-theoretic sense, based on the notion of symmetric difference. For two subsets $A$ and $B$ of a set $E$ the symmetric difference $A\Delta B$ is

$$A\Delta B = A \setminus B \cup B \setminus A,$$

where $\setminus$ is the set-theoretic minus. The distance between $A$ and $B$ is $\delta(A, B) = |A\Delta B|$. It can be computed as

$$\delta(A, B) = \sum_{j=1}^{n} |\mathbf{1}_A(j) - \mathbf{1}_B(j)|, \tag{2}$$

which shows that the distance is just the distance corresponding to the $L^1$-norm (or sum norm or Manhattan norm) in $\mathbb{R}^n$.

## 2.2. The Vorob'ev expectation

Let $X$ be a finite discrete random subset of $E$. (The general theory of random sets is presented in detail in Molchanov (2005, 2017), while for this paper a naive understanding suffices; we use some of the ideas of the theory.)

An important summary characteristic of a random set $X$ is its *coverage function* $p_X(\cdot)$ defined by

$$p_X(j) = \mathbf{E}(\mathbf{1}_X(j)) = \mathbf{P}(j \in X) \text{ for } j \in E. \tag{3}$$

Using the coverage function so-called *t-th quantiles* $\{p_X \geq t\}$ are defined by

$$\{p_X \geq t\} = \{j \in E : p_X(j) \geq t\} \text{ for } t \geq 0, \tag{4}$$

which are deterministic sets.

The *Vorob'ev expectation* of $X$, $\mathbf{E}_V(X)$, is defined as the set $\{p_X \geq t\}$ for $t \in [0,1]$ which is determined by the equation

$$\mathbf{E}(|X|) = |\{p_X \geq t\}|,$$

if this equation has a solution, or, in general, from the condition

$$|\{p_X \geq s\}| \leq \mathbf{E}(|X|) \leq |\{p_X \geq t\}| \text{ for all } s > t. \tag{5}$$

This definition is the same as Definition 2.1 in Molchanov (2005), p. 177, with $\mu(\cdot)$ replaced by $|\cdot|$, and Definition 2.2.3 in Molchanov (2017), p. 283. The discrete case used here was also considered in the original paper by Vorob'ev (1984).

We see that $\mathbf{E}_V(X)$ is the $t$-th quantile of $X$ having minimum element number equal or larger than $\mathbf{E}(|X|)$.

The deeper sense behind the definition of Vorob'ev expectation is the following minimization property: For each finite set $M$ with $|M| \leq \mathbf{E}_V(X)$ it holds

$$\mathbf{E}|X \Delta \mathbf{E}_V(X)| \leq \mathbf{E}|(X \Delta M)|,$$

see Molchanov (2005), p. 177, and Molchanov (2017), p. 284.

## 2.3. Statistical analysis

*General*

Now we turn to samples of finite random sets and present statistical tools for their analysis.

We assume that the random sets $X_1, X_2, \ldots, X_n$ are given and have the same distribution as a prototype $X$. Our aim is to estimate the Vorob'ev expectation $\mathbf{E}_V(X)$ and then to analyse the relationship of the $X_i$, which are assumed to be not empty, to the empirical version of $\mathbf{E}_V(X)$.

We use the notation

$$s_i = |X_i| \text{ for } i = 1, 2, \ldots, r$$

and

$$\overline{s} = \frac{1}{r} \sum_{i=1}^{r} s_i.$$

Clearly, $\overline{s}$ serves as an estimator of $\mathbf{E}(|X|)$. We estimate the coverage function $p_X(j)$ by

$$\hat{p}_X(j) = \frac{n_j}{r} \text{ for } j = 1, 2, \ldots, n, \tag{6}$$

where $n_j$ is the number of sets containing element $j$, i.e.

$$n_j = \sum_{i=1}^{r} \mathbf{1}_{X_i}(j) \text{ for } j = 1, 2, \ldots, n.$$

We consider below the coverage function as a function of a real variable (which we also denote by $j$), which is constant in the intervals between integers.

Note that the $s_i$ and $n_j$ are the margins of the two-way table of 0's and 1's corresponding to the rating problem.

Finally, $\overline{X}$ is the empirical Vorob'ev expectation, given by (5) with $p_X(j)$ replaced by $\hat{p}_X(j)$ and $\mathbf{E}(|X|)$ by $\overline{s}$.

When the Vorob'ev mean is determined, the distances $\delta_i$ of the sets $X_i$ from $\overline{X}$ in the sense of equation (2) can be used as characteristics of the degree of conformity of the raters with the whole collective of raters.

*Ranking of items*

The presentation of coverage function and Vorob'ev mean becomes clearer if we rank the items $j$ according to the numbers $n_j$, where items with larger $n_j$ are ranked lower than those with smaller $n_j$, while the ranking of items with equal $n_j$ does not matter. Thus after ranking the $n_j$ form a decreasing sequence.

This ranking is used in the following for numbering the items. It implies that the empirical coverage function $\hat{p}_X(j)$ is decreasing in $j$. Furthermore, the empirical Vorob'ev expectation $\overline{X}$ has the form $\{1, 2, \ldots, m\}$ with $m$ obtained as

$$m = \max\{j : \hat{p}_X(j) = \hat{p}_X(\overline{s})\}. \tag{7}$$

This can be interpreted as follows: $\overline{s}$ is a positive real number between 1 and $n$ and lies in an interval with integer end points where $\hat{p}_X(\cdot)$ is constant. The right end point $m$ of this interval is also the end point of the empirical Vorob'ev mean. We remark that the number $m$ leads to another way of understanding $\overline{X}$. By definition of the empirical coverage function there is an integer $L$ with

$$\hat{p}_X(m) = L/r.$$

Thus $\overline{X}$ is the set of all items with at least $L$ 1-marks, with

$$L = r\hat{p}_X(m). \tag{8}$$

*Conformity numbers $c_i$*

The conformity of rating of rater $i$ with the other raters can be characterized also by a numerical characteristic, which we call *conformity number* $c_i$. It is defined by

$$c_i = \frac{1}{s_i} \sum_{j=1}^{n} \mathbf{1}_{X_i}(j) \cdot n_j \text{ for } i = 1, 2, \ldots, r. \tag{9}$$

That means: $c_i$ is the mean of the numbers $n_j$ of items chosen by rater $i$.

A large value of $c_i$ means that rater $i$ has a 'tendency to conform' with the general rating tendency of all $r$ raters, since most of the items marked by this rater are items frequently chosen by others as well.

The conformity numbers offer information from the interior of the two-way table which is not included in the margins. However, they have a weakness: active raters with large $s_i$ necessarily mark also items that are not frequently marked. This reduces the values of the corresponding $c_i$, which may lead to a biased impression. Therefore, we recommend the use of the *relative conformity numbers* $r_i$ defined by

$$r_i = c_i/C_i \text{ for } i = 1, 2, \ldots, r, \tag{10}$$

with

$$C_i = \frac{1}{s_i} \sum_{j=1}^{s_i} n_j \text{ for } i = 1, 2, \ldots, r. \tag{11}$$

The quantity $C_i$ is the conformity number of an idealized rater who marks $s_i$ items as rater $i$ does, but he chooses the $s_i$ items with the largest numbers $n_j$ in the list of all items. The ratio $r_i$ aims to compensate for the size bias of the $c_i$. The $r_i$ are positive numbers smaller than 1. A large value of $r_i$ indicates a high degree of conformity of rater $i$ with the whole group of raters.

# 3. Forestry example

## 3.1. Data description

We consider a situation where $r$ forest managers classify $n$ trees either as trees to be maintained or as trees to be removed. They assign tree marks, either '0' or '1', where mark '1' means 'remove'. In our data example there are $r = 15$ forest managers, which we call in the following 'raters', and $n = 387$ trees.

The aim of such tree-marking experiments is to study the personal strategies of forest managers and their agreement in the tree-selection procedure, which can provide psychological insights into the foresters' thinking, see Vítková et al. (2016). As part of standard forest management practice, thinnings are regularly applied to forest stands in order to reduce tree density while trees naturally increase in size. The selection of trees for thinnings and of those to leave behind is not trivial and requires detailed ecological and silvicultural skills that are provided in forestry training courses. The long-term development of forest stands and their ability to provide anticipated goods and services largely depends on decisions made by trained forest managers. Yet, even the most skilled forest managers continue to be human beings with a unique set of personal preferences, experiences and flaws that influence their decision making. Considering how important marking decisions are for the development of managed forest ecosystems, it is crucial to understand the selection behavior of forestry staff. The analysis of human selection behavior is likely to lead to valuable psychological insights. As part of this it is possible to understand interactions within and between groups of forest managers better and why certain forest managers nearly always select the same trees while others make completely different choices. This understanding can prove helpful for preparing goal-oriented training courses, which take lessons learnt from this research into account.

The matter also plays an important role in computer-based forest models which simulate the development of forests over time. In such models, it has commonly been assumed that forest managers and machine operators mark trees according to theoretical rules published in forestry textbooks and according to best practice. Previous studies, however, have cast doubt on this assumption (Zucchini and von Gadow, 1995; Füldner et al., 1996; Spinelli et al., 2016; Pommerening et al., 2018). The authors of these studies have found only little agreement in the marking behavior among forestry professionals. While Spinelli et al. (2016) speculate that different practical experience in tree marking is a possible explanation, we assume that also education and individual personality play a role. This was confirmed by Vítková et al. (2016) who report a tree marking experiment in Ireland involving raters with different experience and education. They required the raters to perform the marking twice in the same experimental forest, once before and once after training in a new marking technique. It turned out that experts were unwilling to adopt the new marking

method and the training led to confusion and decreasing agreement in this group. In contrast, novices responded well to the training and the agreement in this group was significantly higher than among the experts. Pommerening et al. (2018) came to similar conclusions when analysing 36 marking experiments from all over Britain. The authors also found that more complex forest structure appears to facilitate the decision process.

Before the experiment referred to in this paper started, the raters were informed and coached using thinning instructions, which implied a clear thinning strategy involving the retention of dominant, good quality trees, so-called frame trees, and the removal of frame-tree competitors.

For the statistical analysis, we ordered the raters with respect to the number of trees they marked with '1'; rater #1 is thus the rater with most '1' marks. This simplifies the presentation of our results.

The experiment analysed in this paper deliberately included forest managers and other persons with quite different backgrounds and therefore different marking psychology. Most of the forest managers were from the UK. It was a problem for the majority of participants that they were asked to use a thinning strategy that conflicted with their experience. The strategy they usually use is low thinning, otherwise known as 'thinning from below', where trees are removed mainly from the lower canopy layer and from among the smaller diameter trees. But in the experiment they were asked to apply crown thinning, also referred to as 'thinning from above', where trees are removed that are part of the upper canopy layer in order to favour the best trees of the main canopy by removing their direct competitors. Forest manager #1 was an employee of the Forestry Commission, #2 and #6 were experienced employees of British forest management firms and #8 was a forest engineer of Bangor University. In contrast, persons #3, #4, #5, #7, #10, and #12 were inexperienced employees or students. The others had a varied background, some were working for forestry firms, others were self-employed or private persons with an interest in forestry or were private persons. Finally, #9 was the organizer and trainer of the workshop.

It is clear that while marking a particular tree in one part of the forest a forest manager can hardly recall all other trees of the forest he has already visited. There are complicated spatial correlations between the marks of trees close together, where the aforementioned frame trees play an important role. There could, for example, be agreement in the selection of frame trees but differences in which specific trees to thin around them. Or, there could be differences in the frame tree selection with consequences for many more trees even at larger distances. Since we concentrate on data analytic methods and statistical mean values, we ignore these dependencies here.

The forestry data studied in this paper can be downloaded from `www.pommerening.org`. We have used data from research plot 7 in Coed y Brenin (North Wales, UK) surveyed in 2006. The data were originally collected as part of a training workshop for forest managers.

### 3.2. Classical data analysis

In order to convince the reader that a detailed statistical analysis of the forestry data really makes sense, first simple classical data analysis methods were used to check whether or not there is agreement between the raters. We used for this purpose the methods mentioned in the Introduction.

Fleiss' kappa is very small, $\kappa = 0.102$, which according to Landis and Koch (1977) means that there is only 'slight agreement'; Table 4 in Stoyan et al. (2018) interprets this value as between 'slight' and 'fair agreement'. Cochran's $Q$ test yields a clear rejection of the null hypothesis of equality of the probabilities of marking with '1' for all raters, i.e. the hypothesis of equal rater activity.

Pairwise comparisons using the McNemar test show that, for example, even the differences between forest managers #1 and 2, #3 and 4, #11 and 12 and #14 and 15 are significant.

These results are plausible since the numbers $s_i$ of trees marked by raters $i$ are quite different, see Fig. 1. This figure, termed rater histogram in Pommerening et al. (2018), shows for each of the 15 raters the numbers $s_i/387$ of trees marked with '1'. For example, raters #1 and #15 assigning 184 and 37 1-marks, respectively, have extreme positions in the histogram.

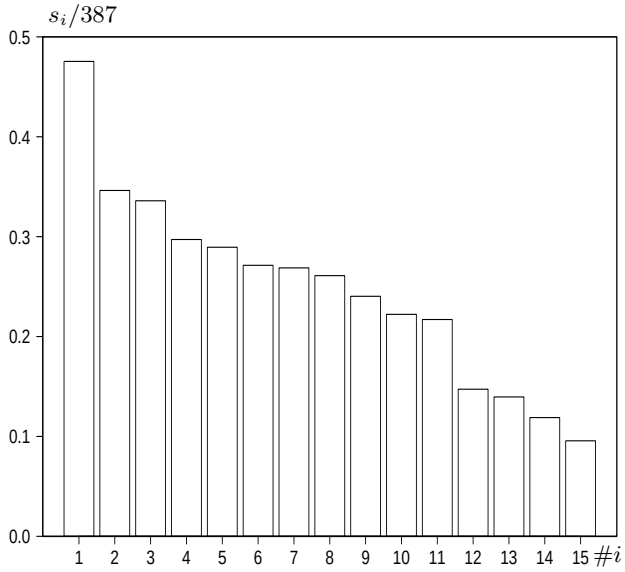


Figure 1: The marking activities of the 15 forest managers shown by the proportions $s_i/387$ of marked trees, where $s_i$ is the number of trees marked by forest manager $\#i$.

Also the passive marking frequency of the trees shows great variability, see Fig. 2. This marking histogram (Pommerening et al., 2018) shows the empirical distribution of the proportions of trees getting mark '1'. There are 30 trees without any '1' mark, which is shown by the bar at 0, and no tree has received the theoretically possible maximum of 15 '1' marks; also 14 '1' marks were not received. The zero class looks like a case where all forest managers are in perfect agreement, however, this is only a kind of pseudo- or passive agreement. The situation of trees that are marked by none of the forest managers is not unusual, see the discussion in Zuchini and von Gadow (1995).

Following Schouten (1982) and Stoyan et al. (2018) we used cluster analysis to find subgroups of raters with similar behavior. Fig. 3 shows a dendrogram resulting from cluster analysis with the Ward algorithm and Manhattan distance. The 15 variables used are the rater-related sequences of 387 0's and 1's.

As expected, if one chooses a four-cluster solution (like in Fig. 3), forest manager #1 forms a single cluster probably simply because of the large number of trees marked. The other clusters are difficult to explain. In the large #4, ..., #15 cluster, there are mainly forest managers with low marking activity.

Finally we applied LCA using the TAM::tamaan function in the R-package TAM (Kiefer et al., 2016; R Development Core Team, 2016) to the data in order to form two classes of trees, which we identify with the trees which should be marked by '1' and '0'. The corresponding latent class probabilities are 0.355 (for '1') and 0.645 (for '0'), respectively. The model with two classes was

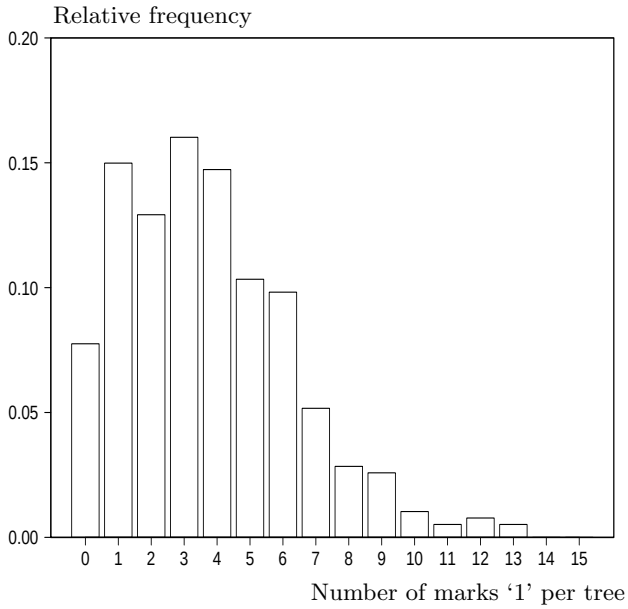


Figure 2: The passive marking frequencies, i.e., the frequencies of '1' marks assigned to the trees.

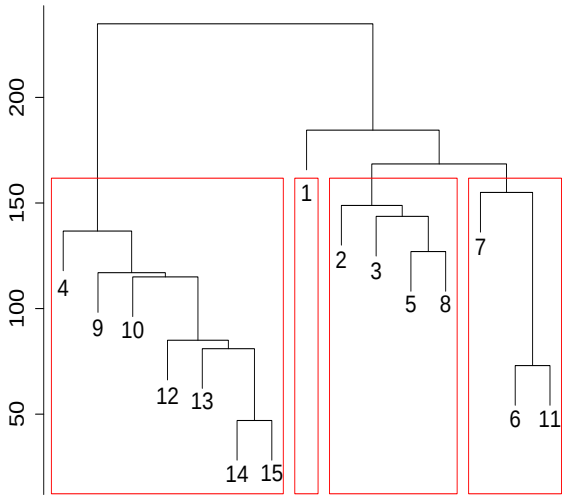


Figure 3: Cluster analysis dendrogram for the 15 forest managers with four distinctive clusters highlighted in red.

preferred over a latent class model with three classes due to the smaller Bayesian information criterion (BIC) value.

Fig. 4 shows the estimated probabilities of trees belonging to class '1' in dependence on the frequency of their '1' marks. For example, for a tree with three '1' marks the probability has quartiles of 0.027 and 0.138. Since the trees are ordered with respect to the frequency of being marked by the forest managers, there is a tendency of monotonous increase. The degenerated boxplot at frequency 7 suggests interpreting trees marked more often than 7 times as trees of class 1, while rarely marked trees (not marked at all or only once) may be interpreted as trees of class 0.

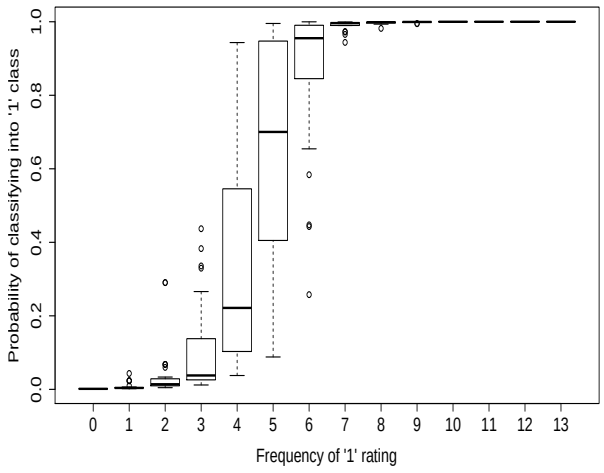


Figure 4: Boxplots of the estimated probabilities of classification as '1' for the 387 trees in dependence on the frequency of '1' rating.

We observe a great variability of the probabilities of trees marked by four and five forest managers. In a next step the forest managers were characterized by the values of specificity and sensitivity shown in Fig. 5.

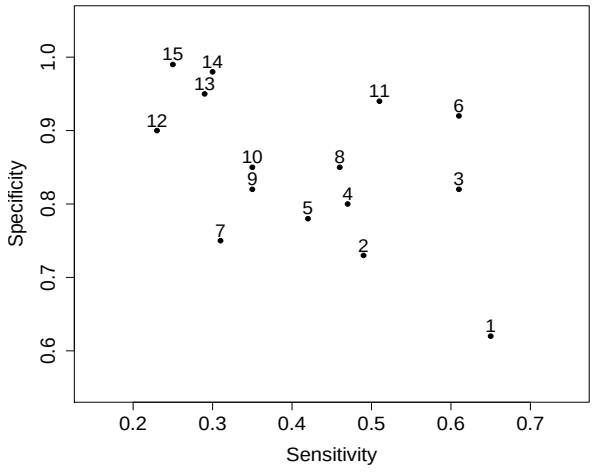


Figure 5: Scatterplot of specificity and sensitivity of the 15 forest managers as obtained by LCA.

We see the (not surprising) tendency that active raters (raters that often assigned mark '1') mark those trees frequently with '1' which are often considered by LCA as '1' trees, and that non-active raters mark those trees frequently with '0' which are considered as '0' trees by LCA. As Fig. 5 shows, rater #1 has an extreme position in the scatterplot, which may be explained by his rating activity. Raters #3, #6 and #11 may be considered 'normal' raters, since they both show large sensitivities and specificities. Raters #12 to #15 form a subgroup of raters of small sensitivity, mainly because

of their small rating activities. Finally, all other raters form a group with medium sensitivities and specificities. If one determines the probability of correct classification using sensitivity and specificity, the largest value of 0.81 is obtained for rater #6, followed by #11 with 0.79, while the smallest value is 0.59 for rater #7. That may mean that this rater has a rating behavior quite different from that of the whole group.

### 3.3. Set-theoretic results for the forestry data

Fig. 6 shows the coverage function $\hat{p}_X(j)$ for the forest data. It is a decreasing function of tree number $j$. We use it for the determination of the Vorob'ev mean $\overline{X}$. The mean number of marked trees per forest manager is $\overline{s} = 96.13$. This number belongs to the interval $[91, 130]$, where $\hat{p}_X(j)$ takes the constant value $1/3$; by the way, this interval belongs to trees marked by five forest managers. By definition, the Vorob'ev mean $\overline{X}$ is the set $\{1, 2, \ldots, 130\}$.

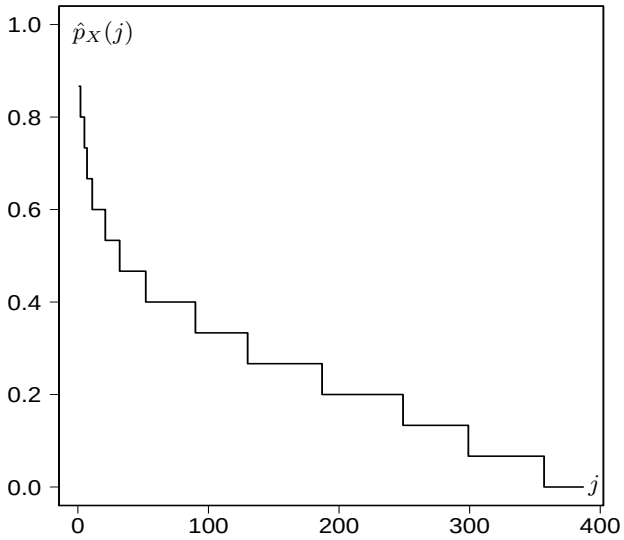


Figure 6: The empirical coverage function for the forestry data. For a given tree number $j$ it gives the proportion of the numbers of forest managers who marked tree $j$. The maximum number of marks is 13 and the minimum is 0, it is $\hat{p}_X(1) = 13/15$ and $\hat{p}_X(358) = 0$. As a result of ordering the trees $\hat{p}_X(j)$ is a decreasing function.

In other words, $\overline{X}$ is the set of all trees marked by five or more forest managers. This set may be interpreted as the set of trees the 15 raters want to determine as '1' trees.

By the way, also LCA leads to such a set. One could take a tree number close to 96.13 trees with a probability of '1' classification larger than some limit. (The probability is 0.883 for the 96th and 97th tree when ordering the trees according to the probabilities.) This set would include some trees marked only by four forest managers, and some trees marked by six forest managers would not belong to this set. The differences between the two sets result from differences in weighting the raters. While in the set-theoretic approach all raters are considered equal, in the LCA approach raters considered 'normal' (e.g. raters #6 and #11) have higher weights.

Fig. 7 shows a scatterplot of the pairs $(s_i, c_i)$ (number of trees marked by forest manager $\#i$, conformity number of forest manager $\#i$) for the forest data. Additionally, in red (+) are shown 15 pairs $(s_i, c_i)$ for simulated data obtained from 15 sequences of 387 randomly marked trees. (For

every pair of forester #$i$ and tree ($j$) the probability of 0.248 for choosing '1' was taken, and the selection process was completely random. 0.248 is the proportion of '1' marks in the data.)

There are big differences between the two data sets. The points of the simulated data are close together and show, as expected, no structure. By contrast, for the forest data we clearly see a distinctive structure, with a high degree of negative correlation between the $s_i$ and $c_i$, i.e., active raters tend to be less conform than passive ones.

Table 1: The mark numbers $s_i$, the conformity numbers $c_i$, the relative conformity numbers $r_i$ and the distances $\delta_i$ from the empirical Vorob'ev expectation $\overline{X}$ for the 15 raters.

| $i$ | $s_i$ | $c_i$ | $r_i$ | $\delta_i$ |
|---|---|---|---|---|
| 1 | 184 | 4.83 | 0.81 | 128 |
| 2 | 134 | 5.07 | 0.77 | 124 |
| 3 | 130 | 5.52 | 0.82 | 102 |
| 4 | 115 | 5.43 | 0.79 | 115 |
| 5 | 112 | 5.37 | 0.77 | 114 |
| 6 | 105 | 6.05 | 0.85 | 77 |
| 7 | 104 | 4.88 | 0.69 | 136 |
| 8 | 101 | 5.63 | 0.79 | 101 |
| 9 | 93 | 5.44 | 0.74 | 123 |
| 10 | 86 | 5.57 | 0.74 | 114 |
| 11 | 84 | 6.29 | 0.83 | 88 |
| 12 | 57 | 5.95 | 0.72 | 117 |
| 13 | 54 | 6.41 | 0.76 | 104 |
| 14 | 46 | 7.07 | 0.81 | 102 |
| 15 | 37 | 6.95 | 0.76 | 107 |

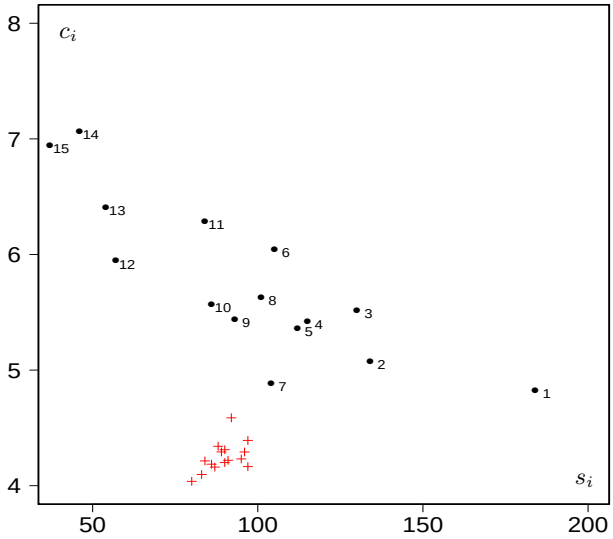


Figure 7: Scatter plot of the $(s_i, c_i)$, (number of trees marked by forest manager $i$, average mark number of the trees marked by forest manager $i$ = conformity number).

Fig. 5 and 7 show some similarity. Indeed, the $s_i$ are positively correlated with the sensitivities and the $c_i$ with the specificities. But note how simple the determination of $s_i$ and $c_i$ is compared

to that of sensitivities and specificities!

Perhaps the relative conformity numbers $r_i$ in Table 1 show the structure of the rater group in a still clearer way. We now see that rater #1 is not only very active but also rather conform, he has one of the largest relative conformity numbers. The most conform rater in terms of the relative conformity numbers is rater #6, one of the older, experienced raters. The most non-conform rater is #7, a young student. The trainer of the experiment, rater #9, belongs to the raters with low conformity.

Finally, the set-theoretic distances $\delta_i$ between the $X_i$ of the forest managers and the empirical Vorob'ev expectation are given in Table 1. It is interesting that raters #6 and #7 also here assume extreme positions, #6 has the minimum distance and #7 the maximum. (By the way, a refined analysis of the data revealed that #6 shows a somewhat special behavior: he has the maximum number of marks '1' under the first 50 trees.) Raters with small $s_i$ tend to have small distances and rater #11 with his very small distance of 88 conforms well in terms of $r_i$.

*Interpretation of the results*

1) #1 clearly constitutes an outlier as indicated by the extreme number $s_1$ of marked trees. Nevertheless, as $r_1$ shows he conforms well with the whole group of raters.

2) #6 and #11 show a kind of medium marking activity (the corresponding $s_i$ are close to $\overline{s}$) and a high degree of conformity with the whole group of raters. Because of the small distances from the Vorob'ev expectation they may serve as representatives of the whole group of forest managers.

3) The other forest managers show an average behavior.

4) The organizer and trainer of the experiment, forest manager #9, does not take a central place in Fig. 7 and has a considerable distance $\delta_9$ from the Vorob'ev expectation. Rater #7 followed strictly the rules given by #9 and thus became a marginal figure within the collective of raters. Obviously, more training is necessary to teach traditional British forest managers how best to apply thinnings from above.

## 4. Discussion

This paper shows that set-theoretic methods indeed yield information on the individual rater behavior. The conformity numbers $c_i$ and the relative conformity nubers $r_i$ for the raters $i$ characterize the relation of the rating of rater $i$ in comparison to that of all raters. And the Vorob'ev mean helps to understand more subtle differences in the rating behavior of individual raters.

The results of both the set-theoretic approach and of LCA show similarity in important points. Both methods suggest that raters #6 and #11 occupy the positions of extreme conformists and label rater #7 as non-conformist. The set-theoretic approach explains the role of rater #1 better; he is not only characterized by his high marking activity but also shows a rather conform behavior. Rater #3 does not play a particular role in the set-theoretic approach in contrast to LCA. Raters #12 to #15 have similar results with both approaches. Finally, both approaches classify the raters not named here as 'medium'. It is interesting that raters #6 and #11, which have largest sensitivity and specificity in LCA, are the raters with the smallest distance from the Vorob'ev mean, i.e. the average rater. Perhaps two advantages of the set-theoretic approach in comparison to LCA are its theoretically simpler approach and its greater suitability.

Clearly, the methods presented in this paper can be generalized in such a way that the raters do not

assign marks of type '1' or '0', but provide scores instead. Then the item sets can be interpreted as fuzzy sets (if the scores are between 0 and 1) and the analysis can follow the pattern suggested in this paper.

The Vorob'ev expectation may be interpreted as a result of a wisdom of crowd approach of psychology (Surowiecky, 2004). In this approach, judgements of multiple experts (raters) are aggregated to obtain results which are closer to a ground truth than single judgements. If estimates of real numbers have to be aggregated, a simple aggregation method involves averages of individual experts' judgements. Other more sophisticated methods were designed for aggregating rankings (Steyvers et al., 2009; Lee et al., 2014) and solutions of combinatorial problems (Yi et al., 2012). Similarly, we have aggregated sets to mean sets, generalizing the simple averaging approach. However, it must be noted that the whole crowd may go astray, as it seems a bit with our forestry example, where the majority of the group tended not to follow the thinning rules given by the organizer of the experiment.

Finally, we remark that the empirical Vorob'ev expectation $\overline{X}$ may have also an additional application. It may represent a set of selected items (in the case of our example: trees), as it contains the most frequently used items in a number suitable to the investigation purpose. Some examples are:

- Trees finally determined to fell,
- Patients considered ill,
- Candidates elected.

If in such situations the number of items to select is not prescribed a priori, $\overline{X}$ may be a plausible solution.

At the end of a tree-marking experiment the question may arise which trees to finally remove. This decision may be made on a 'democratic' basis, using the data of the experiment: just take the empirical Vorob'ev mean $\overline{X}$ as the set of trees to fell. This decision will be plausible for every rater, since $\overline{X}$ consists of all trees marked by at least $L$ raters, where $L$ is given by equation (8). (In the case of the forestry experiment discussed in the present paper it is $L = 5$.)

One may proceed similarly in the case of a medical rating experiment as, for example, in Landis and Koch (1977a, b) and Stoyan et al. (2018). There the raters are doctors and the items patients, and giving the mark '1' to a patient may mean that the doctor thinks that he has some disease of interest. While the rating experiment also here primarily aims at studying the agreement of doctors in making decisions and to find groups of similar behavior, the question may also arise which of the patients to finally consider as ill. The Vorob'ev mean is then a plausible solution: the set all patients classified by at least $L$ doctors as ill is considered as really ill.

The Vorob'ev mean can be also used in the election of a council, in a modification of the classical approval voting system. In this system, $r$ voters can vote for as many candidates as they want from a set of $n$ candidates and the $k$ candidates with most votes are finally elected. Here $k$ is a number fixed before the election and known to all voters, i.e., the size of the council. This voting system was suggested by Brams and Fishburn (1978), and is nowadays often applied, see also Brams and Taylor (1996) and Rothe (2015).

One may soften this system by not fixing the size of the council a priori, but by allowing it to be determined by the voters as well. This means: one proceeds as in usual approval voting, but the number of members of the council is $L$ and the set of candidates elected is the empirical Vorob'ev mean. If it does not happen that two candidates have the same number of votes, $L$ equals simply $\overline{s}$, the mean number of candidates per ballot, rounded up. And the Vorob'ev mean is the set of

the $L$ candidates with most votes. (Perhaps rounding up may be replaced by rounding, up or down.) Voting systems for councils of variable numbers of members are discussed in Faliszewski et al. (2018), but the system discussed just here is not mentioned there.

## Acknowledgements

The authors thank Michael D. Lee for the hint on 'wisdom of the crowd'. Michel Regenwetter kindly pointed us towards the general theory of subset voting and provided helpful comments on an earlier version of this paper whilst Anthony Marley recommended the application of methods of LCA, in which we enjoyed the help of Alexander Robitzsch. Jörg Rothe discussed voting systems with us.

The authors wish to thank the Welsh European Funding Office and the Forestry Commission Wales for the financial support for the Tyfiant Coed project as part of which the data used in this study were collected. Jens Haufe, Owen Davies and Gareth Johnson helped with the data collection and with organising the training workshop.

This study also supports the COST action FP1206 'EuMixFor'.

## References

Brams S, Fishburn P (1978). "Approval Voting." *American Political Science Review*, **72**, 831–847.

Brams S, Taylor A (1996). "Fair Division: From Cace-Cutting to Dispute resolution." Cambridge University Press, Cambridge. **72**, 831–847.

Collins L, Lanza ST (2010). "Latent Class and Latent Transition Analysis." J. Wiley & Sons, Hoboken, New Jersey.

Everitt BS, Landau S, Leese M, Stahl D (2011). "Cluster Analysis." J. Wiley & Sons, London.

Faliszewski P, Slinko A, Talmon N (2018). "The complexity of multiwinner voting rules with variable number of winners."https://arxiv.org/pdf/1711.06641.pdf

Fleiss JL, Levin B, Paik MC (2003). "Statistical Methods for Rates and Proportions." J. Wiley & Sons, New York.

Füldner K, Sattler S, Zucchini W, v. Gadow K (1996). "Modellierung personenabhängiger Auswahlwahrscheinlichkeiten bei der Durchforstung" [Modelling of Person-specific Tree Selection Probabilities in a Thinning]. *Allgemeine Forst- und Jagdzeitung*, **167**, 159–162.

Hallgren KA (2012). "Computing Inter-rater Reliability for Observational Data: An Overview and Tutorial." *Tutor. Quant. Methods Psychol.*, **8**, 23–34.

Kiefer T, Robitzsch A, Wu M (2016). "TAM: Test Analysis Modules." R package version 1.995-0. URL https://CRAN.R-project.org/package=TAM.

Landis JR, Koch GG (1977a). "The Measurement of Observer Agreement for Categorical Data." *Biometrics*, **33**, 159–174.

Landis JR, Koch GG (1977b). "An Application of Hierarchical Kappa-type Statistics in the Assessment of Majority Agreement among Multiple Observers." *Biometrics*, **33**, 363–374.

Molchanov IS (2005). "Theory of Random Sets." Springer, London.

Molchanov IS (2017). "Theory of Random Sets. Second Edition." Springer, London.

Pommerening A, Pallarés Ramos C, Kędziora W, Haufe J, Stoyan D (2018). "Rating Experiments in Forestry: How Much Agreement Is There in Tree Marking?" *Plos One*, **13**, e0194747.

R Development Core Team (2018). "R: A Language and Environment for Statistical Computing." Vienna, Austria.

Regenwetter M, Grofman B, Marley AAJ, Tsetlin IM (2006). "Behavioral Social Choice. Probabilistic Models, Statistical Inference, and Applications." Cambridge University Press, Cambridge.

Rothe J (ed.) (2015). "Economics and Computation. An Introduction to Algorithmic Game Theory, Computational Social Choice, and Fair Division." Springer-Verlag, Heidelberg.

Schouten HJA (1982). "Measuring Pairwise Interobserver Agreement when all Subjects are Judged by the Same Observers." *Statistica Neerlandica*, **36**, 45–61.

Sheskin DJ (1997). "Handbook of Parametric and Nonparametric Statistical Procedures." CRC Press, Boca Raton.

Spinelli R, Magagnotti N, Pari L, Soucy M (2016). "Comparing Tree Selection as Performed by Different Professional Figures." *Forest Science*, **62**, 213–219.

Steyvers M, Lee MD, Miller B, Hemmer P (2009). "The Wisdom of Crowds in the Recollection of Order Information." In: J. Lafferty & C. Williams (Eds.): "Advances in Neural Information Processing Systems". MIT Press, Cambridge, MA, vol 23, 1785–1793.

Stoyan D, Stoyan, H (1994). "Fractals, Random Shapes and Point Fields." J. Wiley & Sons, Chichester.

Stoyan D, Pommerening A, Hummel M, Kopp-Schneider A (2018). "Multiple-rater Kappas for Binary Data: Models and Interpretation." *Biometrical Journal*, **22**, 22–33.

Surowiecki J (2004). "The Wisdom of Crowds." Random House, New York.

Uebersax JS, Grove WM (1993). "A Latent Trait Finite Mixture Model for the Analysis of Rating Agreement." *Biometrics*, **49**, 823–835.

Vítková L, Ní Dhubhàin AN, Pommerening A (2016). "Agreement in Tree Marking: What is the Uncertainty of Human Tree Selection in Selective Forest Management?" *Forest Science*, **62**, 288–296.

Vorob'ev OYu (1984). "Mean-value Modelling." (Russ.), Nauka, Moscow.

Yi KM, Steyvers M, Lee MD, Dry MJ (2012). "The Wisdom of the Crowd in Combinatorial Problems." *Cognitive Science*, **36**, 452–470.

Zucchini W, v. Gadow K (1995). "Two Indices of Agreement among Foresters Selecting Trees for Thinning." *Forest & Landscape Research*, **1**, 199–206.

**Affiliation:**

Dietrich Stoyan and Andreas Wünsche, Institut für Stochastik, TU Bergakademie Freiberg, 09596 Freiberg, Germany.
Arne Pommerening, Swedish University of Agricultural Sciences, Department of Forest Ecology and Management, Umeå, Sweden.
e-mail: stoyan@math.tu-freiberg.de

Journal of Environmental Statistics
Volume 8, Issue 2
September 2018

http://www.jenvstat.org
Submitted: 2017-12-28
Accepted: 2018-07-04