Application of Clustered Resampling Methods in Assessing Accuracy of Cross-Validated Estimators of Cross-Classification Probabilities of Nearest-Neighbor Classifiers

Yuri K. Belyaev

Research Report
Centre of Biostochastics

Swedish University of
Agricultural Sciences

Report 2005:1
ISSN 1651-8543

Application of Clustered Resampling Methods in Assessing Accuracy of Cross-Validated Estimators of Cross-Classification Probabilities of Nearest-Neighbor Classifiers

Yuri K. Belyaev

Centre of Biostochastics
Swedish University of Agricultural Sciences
SE-901 83 Umeå, Sweden

Abstract

In several applications objects of interest can be investigated only if they are observed from a long distance. Remotely sensed raster data are collected by sophisticated sensors. These remotely sensed data are transformed into digital images showing special properties of the observed objects by differently colored pixels. Transformation of elements of remotely sensed data into colored pixels can be realized by classifiers. The case with a finite number of classes (colors) is considered. In geomatics such classes can be different species of trees, e.g. spruce, pine and broad-leaved. In this paper the nearest neighbor (NN-) classifiers are considered. The NN-classifiers are widely used in the analysis of statistical data in many areas of research. The most important their characteristics are the cross-classification (CC-) probabilities. The cross-validated (CV-) estimators of the CC-probabilities can be obtained using training sets. The accuracy of NN-classifiers is characterized by the distributions of the deviations of the CV-estimators from the true values of CC-probabilities, which depend on the training sets. Resampling from NN-clusters of connected NN-points can be used to obtain consistent estimators for the distributions of the deviations in the case of $1NN$-classifiers. Two numerical experiments illustrate the suggested methods of resampling.

Key words: remotely sensed data, nearest-neighbor classifiers, probability of misclassification, estimation, cross-validation, point processes, clusters, resampling.

2000 AMS subject classification: 62G09, 62H30, 62G20

1 Introduction

In several applications objects of interest can be investigated only if they are observed from a distance and large raster data are collected by sophisticated sensors. These data have to be transformed into digital images showing special properties of the observed objects by differently colored pixels. This situation is typical in geomatics where satellite sensors collect huge remotely sensed data sets by scanning the surface of areas. Also field data have to be collected.

Let $\mathcal{A}_0$ be an area of interest where it is necessary to create a digital map which shows its land cover. In order to be more specific we assume that $\mathcal{A}_0$ is covered by forest and it is of interest to create a map that shows where the different species of trees grow, e.g. pine, spruce, and broad-leaved trees. Field data integrated with remotely sensed data are essentially used to create these maps. Usually the maps are discretely colored and each color represent a species. The whole process of creating discretely colored maps is rather complex.

In this paper we consider methods of statistical classification which are an essential part in the process of creating the maps. Also we show how it is possible to assess the accuracy of the classification methods. The considered methods are rather general and they can be applied in the creation digital images in many fields of research.

In the example with forest, the field data contain particular information related to trees growing in plots placed in $\mathcal{A}_0$ and in a larger area $\mathcal{A}_1$ which includes $\mathcal{A}_0$. It is essential that $\mathcal{A}_1$ together with $\mathcal{A}_0$ has been simultaneously scanned by the same satellite sensor.

For each plot rather rich field information has been collected and it can be used to find the proportions of these three species, i.e the proportions of pine, spruce and broad leaved trees which grow in any part of the plot. It is also possible to calculate the density of wood in the plot measured in m^3/ha and some other characteristics of forest. A description of the field data used in this study is given in Holmström et al. (2001) and Holmström and Fransson (2003).

The positions of 771 circular plots in the Remningtorp area (in Västergötland, Sweden) are shown in Fig. 1, where distances are given in meters. The inventory was made during the winter 1997/1998 and 1998/1999. It is typical that the plots placed in $\mathcal{A}_0$ constitute only a small part of its surface. Therefore, it is necessary to use the remotely sensed data which is related to the whole area $\mathcal{A}_0$.

The remotely sensed data are collected by a satellite sensor which registers energies reflected from small elements of the area's surface and the data are

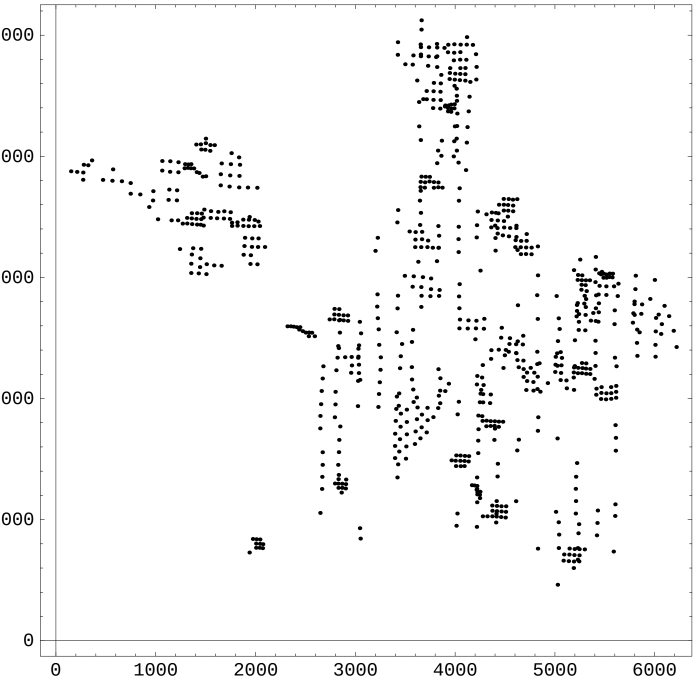

Figure 1: Positions of plots in the Remningtorp area, Västergötland, Sweden.

saved for future needs. These remotely sensed data can be used in the creation of digital maps by applying special software. The remotely sensed data contain many uncertainties caused by restricted resolution of the sensors, the state of atmosphere, noise in sensors and etc. That is why we have to consider the remotely sensed data as observations of random statistical data.

Computer intensive methods such as cross-validation and bootstrap are shown to be efficient tools in the analysis of statistical data, Davison and Hinkley (1997). The nearest neighbor $NN-$classifiers are of wide usage. There are several books and papers devoted to evaluation of the classifiers' characteristics, e.g. Efron and Tibshirani (1993, 1997), Holst and Arle (2001), Steele and Patterson (2000). Mostly only general questions are discussed there and they are not directly connected to the analysis of remote sensed data as we do in this paper. However the accuracy of NN-classifiers are not enough investigated. In the paper we apply some new methods to evaluate the accuracy of NN-classifiers.

The paper is organized as follows. In Section 2 we consider the creation of training sets. From reference data, e.g. the field data in geomatics, for a small subset of remotely sensed data, it is possible correctly to find out which classes will have the most of related pixels in the digital images of the observed objects. The definition of classes is an important part in creating training sets. Section 3 is devoted to the NN-classifiers and their CC-probabilities. The central part of the paper is Section 4 where special resampling methods for the evaluation of the accuracy of CV-estimators of CC-probabilities are suggested. Results of two numerical experiments supporting the suggested resampling methods are given in Section 5. We end this paper with discussion in Section 6.

2 Classes and training sets

The remotely sensed data are raster data sets. They have to be transformed to sets of picture elements (pixels) of created digital maps. Each $\{i, j\}$-element of the raster data contains coordinates x_i, y_j showing its position in the surface, and the reflected from this part energies in several bands of light $e_{ij1}, ..., e_{ijb_0}$. The remotely sensed data related to $\mathcal{A}_0$ are saved as a huge list of elements $\{x_i, y_j, e_{ij1}, ..., e_{ijb_0}\}$, $i, j \in \{1, 2, ...\}$. A classifier is a function which supplies the elements of the raster data by numbers which code classes which can be used to color pixels in the created digital maps.

In our example with forest one can define such classes as pure (100%) pine forest, pure (100%) spruce forest, pure (100%) broad leaves forest, and several classes of mixed forest, e.g. with more than 80% of spruce, 0% of pine, and less than 20% of broad leaves and etc.

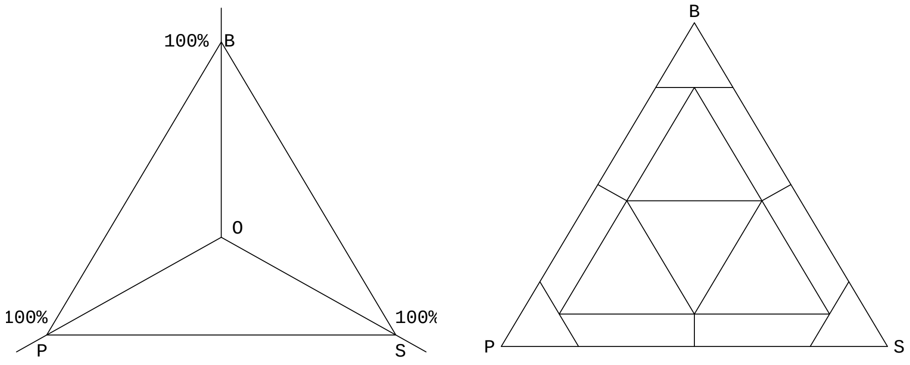

Figure 2: Simplex $\mathcal{S}_3$ in $\mathbf{R}^3$ is shown (left) for all possible mixtures of pines broad leaved and spruces in plots. $\mathcal{S}_3$ is shown in the plane (right) as being divided to 13 subclasses.

All possible values of mixtures can be shown by points in the triangular simplex in three-dimensional space $\mathbb{S}_3 = \{\{z_1, z_2, z_3\} : 0 \leq z_i \leq 100,\ z_1 + z_2 + z_3 = 100\}$, see Fig. 2 (left), where the points $P = \{100, 0, 0\}$, $S = \{0, 100, 0\}$, $B = \{0, 0, 100\}$ are pure pine, spruce and broad leaved classes, respectively. The triangular simplex can be divided into several classes by introducing restrictions on mixtures. In Fig. 2 (right) an example of a division of $\mathbb{S}_3$ into many classes is shown. The vertices P, S, B are for pure pine, spruce and broad leaves parts of forests. The edges PB, BS, PS of $\mathbb{S}_3$ are for mixtures of pine and broad leaves, broad leaves and spruce, pine and spruce parts of forest, respectively.

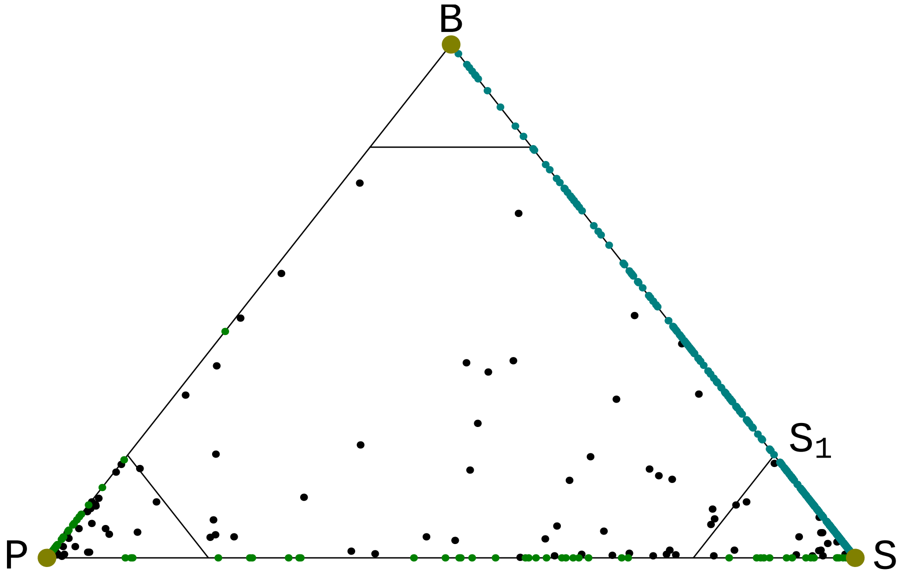

Figure 3: Each point in the simplex $\mathbb{S}_3$ has coordinates pP, pS, pB which are percentages of pines, spruces and broad leaved species of trees in the plot. The plots were placed in the Remningtorp area (1999).

In Fig. 3 each point in the simplex $\mathbb{S}_3$ corresponds to a plot in the Remningtorp area. There are 9 plots with 100% of broad leaves trees 156 plots with 100% of spruce trees and 29 plots with 100% of pine trees. The points in the three triangles near vertices P, B, and S correspond to plots which contain more than 80% of pine, spruce and broad-leaved trees, respectively. The points on the edges corresponds to the plots with forest containing only two of three species, i.e. pines and broad leaves, broad leaves and spruces, spruces and pines. The above mentioned mixture of two species with more than 80% of spruce and less than 20% of broad leaves trees corresponds to the interval

SS_1 on the edge SB in Fig. 3. There are also plots with presence of pines, spruces and broad leaved trees simultaneously. They are shown by the points inside $\mathcal{S}_3$.

If we consider a division of each plot to several small subplots which are squares with sides corresponding to resolution of satellite sensor, say with side length 2 or even 1 meter then the distribution of points in $\mathcal{S}_3$ will be essentially different compared to the distribution in Fig. 3. In this case we will have a lot of points interior and on edges of $\mathcal{S}_3$. Most of the subplots will be shown as points placed at vertices P, S and B of $\mathcal{S}_3$. The other fraction of subplots related to the mixtures of two species will be shown as points on three edges of $\mathcal{S}_3$ and a small fraction of subplots will correspond to points in the inner part of $\mathcal{S}_3$. The distributions of points on the edges and inside $\mathcal{S}_3$ will be approximated by some limit distributions. Then one can hope to use only 7 classes: three for P, S, and B for not mixed forest, 3 for forest with mixtures of two species PS, SB and SP, and PSB for subplots with presence of pines, spruces and broad leaved trees simultaneously.

In our example this division to small subplots was not possible to realize because we have only the remote sensed data with the low resolution $20m \times 20m$. This seems as the main reason that the distributions of points interior and edges of $\mathcal{S}_3$ in Fig. 3 are essentially differ from the more uniform distributions in the case of division plots to small surface elements.

If a set of classes is defined then it is possible to consider many classifiers which are transformations from the raster data into the set of classes. These transformations can be defined on single elements or on groups of elements of raster data. From an infinite number of possible classifiers (transformations into the set of classes) it is necessary to select an optimal one in some sense or at least we have to have the possibility to compare any two classifiers to select the "better" one of them. In order to be able to do that it is necessary to have a so-called training set. The field data can be used in order to obtain the training set related to a selected set of classes. We suppose that the coordinates $\{x_i, y_j\}$ of the raster data correctly correspond to small elements of $\mathcal{A}_1$. Otherwise it is necessary to correct the coordinates. If the resolution of censors is rather high then each circular plot can be included in the union of several small parts of surface corresponding to elements of the remotely sensed raster data. Many such parts occurred to be inside each plot, see Fig. 4 (left). These parts are shown as dark grey squares.

In the case of low resolution as in Fig. 4 (right) we identify the recalculated for each plot field data with one square. In the case of high resolution we consider the squares inside the plots which represent elements of the raster

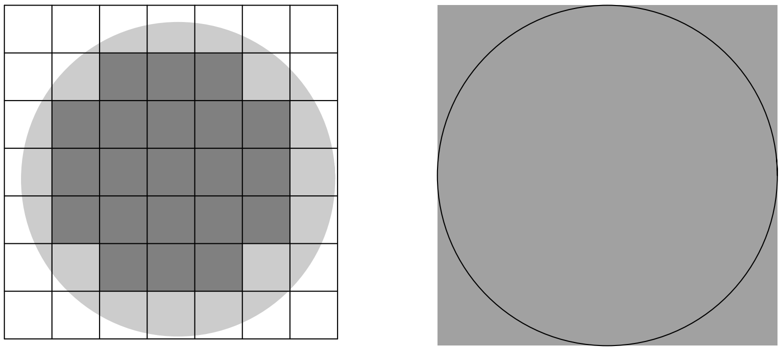

Figure 4: Possible mutual positions of square surface elements and circular plots. High resolution (left) and low resolution (right).

data. Some of the squares are placed inside crowns of the same trees others can contain parts of crowns related to different trees. In the last case if the species of these trees are different we have to consider mixtures of two or tree species simultaneously. Then the corresponding points will be placed on the edges of the simplex $\mathcal{S}_3$ or inside it. Points inside $\mathcal{S}_3$ correspond to squares containing parts of crowns simultaneously belonging to three different species. The boundaries of the visible crowns are fractal. It would be interesting to find out the limit distribution of points on the edges and inside $\mathcal{S}_3$ when the size of squares decreases, i.e. the sensor's resolution is growing.

Let $\mathcal{P}_{i_1j_1}, ..., \mathcal{P}_{i_mj_m}$ be such small parts of the plots placed in $\mathcal{A}_1$. For each such part we can use the field data in order to calculate percentages of $z_P(i_hj_h), z_S(i_hj_h)$, $z_B(i_hj_h)$ of pines, spruces, and broad leaved trees which grow on $\mathcal{P}_{i_hj_h}$, $h = 1, ..., m$. Let $c(\cdot)$ be a function, defined on the simplex $\mathcal{S}_3$, which values are classes of forest.

Then the value $c(z_P(i_hj_h), z_S(i_hj_h), z_B(i_hj_h)) = c_h \in \mathcal{K}_0 = \{1, ..., k_0\}$ is the true class of forest which grows on $\mathcal{P}_{i_hj_h}$. Let $\mathbf{x}_h = \{e_{i_hj_h1}, ..., e_{i_hj_hb_0}\}$ be the vector of registered energies reflected from $\mathcal{P}_{i_hj_h}$, $h = 1, ..., m$. The 3-tuple $\{h, \mathbf{x}_h, c_h\}$ is considered as the hth element of the training set $\mathbb{T}^0 = \{\{1, \mathbf{x}_1, c_1\}, ..., \{m, \mathbf{x}_m, c_m\}\}$.

In the example with circular plots of radius $10m$ in the Remningtorp area the remotely sensed data were registered by the SPOT4 satellite sensor with low resolution (pixel size $20m \times 20m$) which corresponds Fig. 4 (right). Fig. 5 shows that in the considered example elements of the remotely sensed data do not correspond exactly to the positions of the plots (shown as black squares).

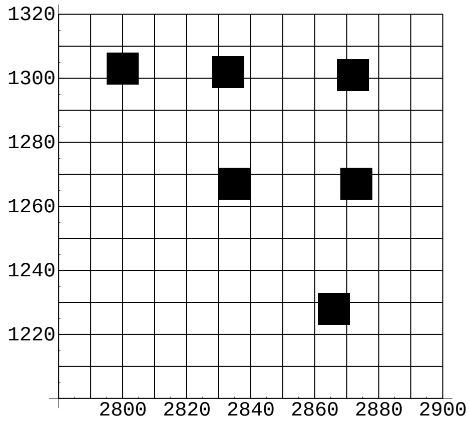

Figure 5: Positions of 6 plots in Remningtorp area with respect to a grid on $(2780, 2900) \times (1200, 1320)$. Plots are approximated by black squares. Each square has size $20m \times 20m$.

The vectors $\{\mathbf{x}_h\}$ of reflected energies have to be recalculated to the plots. It was done by Holmström and Fransson (2003) using cubic convolution (e.g., Niblack (1986)). In our example we will only consider the two bands $3(r)$ and $4(swir)$ related to the reflected light of low frequencies. In Table 1 a fragment of the training set, related to the plots in Remningtorp area, is given.

Table 1. Fragment of the training set

No of data	wood	% of trees			registered energies	
element	m^3/ha	P	S	B	band $3(r)$	band $4(swir)$
42	319.9	100	0	0	102.8	86.9
53	168.6	100	0	0	106.3	98.2
68	98.6	100	0	0	99.1	91.8
...	...	...	...	...	...	...
14	314.1	0	100	0	124.8	83.2
16	536.5	0	100	0	100.1	59.5
17	27.1	0	100	0	143.8	88.4
...	...	...	...	...	...	...
8	58.	0	0	100	124.7	80.9
57	67.1	0	0	100	149.9	116.5
58	131.9	0	0	100	157.4	116.7
...	...	...	...	...	...	...

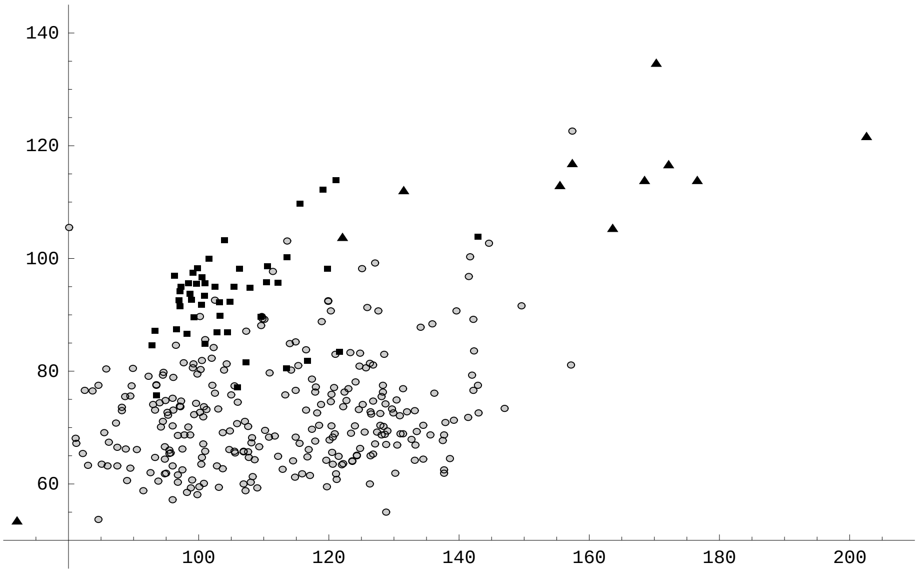

Figure 6: The points which coordinates are energies registered by the SPOT 4 (1999) in the two bands 3(r)(x-axis) and 4th (swir) (y-axis) reflected from the field plots with (not less than 80%) of pines (■), spruce (○) or broad leaved (▲) trees with total density of wood not less than $100m^3/ha$.

In Table 1 P, S and B correspond to percentages of pine, spruce, and broad leaved forest on data elements (plots) numbered as $1, 2, \ldots$. The last two columns contain values of the reflected energies in bands $3(r)$ and $4(swir)$. We denote by $PB20$ a forest with more than 80% of pines, less than 20% of broad leaved trees, spruces being absent, by $BS20$ a forest with more than 80% of broad leaved trees, less than 20% of spruces and pines being absent, and by $SB20$ a forest with more than 80% of spruces less than 20% of broad leaved, and pines being absent.

In Fig. 6 the x-axis corresponds to values of energies in band $3(r)$, and the y-axis to values of energies in band $4(swir)$. Small black squares correspond to the elements of training set belonging to plots with forest $PB20$, small gray circles correspond to plots with forest $SB20$ and small black triangles correspond to plots with forest $BS20$. We can see that there are three clusters related to these three classes.

The variability of the reflected energies in different bands of light is a serious source of possible misclassification when remotely sensed data is used. The definition of a set of classes and the construction of a training set are essential preliminary steps in selecting an appropriate classifier. After that any classifier is defined as a specified transformation from a set of raster data (from the set of all possible values of energies in several bands of light) into the set of classes. The training set can also be used to estimate the accuracy characteristics of classifiers. It is also possible to use the training set to compare two or more suggested classifiers in order to select the one which will be used to create a discretely colored digital map of an area $\mathcal{A}_0$. Classification is an important area of statistical inference.

3 Nearest neighbor classifiers and their characteristics

Suppose that the *remotely sensed data* related to an area $\mathcal{A}$ are written as the following raster array

$$\mathbb{D}_{RS}(\mathcal{A}) = \{\{i, j, \mathbf{e}(i,j)\} : \mathcal{A}_{ij} \subset \mathcal{A}_1\}, \tag{1}$$

where $\{\mathcal{A}_{ij}\}$ are disjoint surface elements and $\mathbf{e}(i,j) = \{e_1(i,j), ..., e_{b_0}(i,j)\}$ are the reflected energies in b_0 bands of light from $\mathcal{A}_{ij}$ registered by a satellite sensor. The indices i, j code the coordinates (longitude and latitude) of $\mathcal{A}_{ij}$. In the theoretical model we approximate $\mathcal{A}_{ij}$ by squares which sides are equal to the sensor's resolution. In order to avoid the necessity to use all components of the registered energy $\mathbf{e}(i,j)$ one can apply a vector function $\mathbf{f}_{RS}(\cdot)$ which transforms the value $\mathbf{e}(i,j)$ into $\mathbf{x}(i,j) = \mathbf{f}_{RS}(\mathbf{e}(i,j)) \in \mathfrak{X} = (\mathbf{x})$. $\mathfrak{X}$ is called a *feature space.* If only the reflected energies in bands $b_1, b_2, ..., b_r$, $r < b_0$ are used then $\mathbf{f}_{RS}((e_1, ..., e_{b_0})) = \{e_{b_1}, ..., e_{b_r}\}$, $r < b_0$.

Let $\mathcal{A}_{i_1 j_1}, ..., \mathcal{A}_{i_m j_m}$ be surface elements which are parts of the field plots placed in $\mathcal{A}_1$. For each $\mathcal{A}_{i_h j_h}$ one can use the field data in order to find the exact (true) class $c(i_h, j_h) \in \mathcal{K}_0 = \{1, 2, ..., k_0\}$ of the part of forest which grows on $\mathcal{A}_{i_h j_h}$. The set of pairs

$$\{\{\mathbf{x}(i_h, j_h), c(i_h, j_h)\},\ h = 1, ..., m\} \tag{2}$$

is called a *training set,* and the points $\mathbf{x}(i_h, j_h) \in \mathfrak{X}$ are called training points. Let m_k be the number of all training points which have true class $k \in \mathcal{K}_0$. We can also consider (2) as the union of subsets

$$\mathbb{T}^0_{m_k} = \{\{\mathbf{x}(i_h, j_h), c(i_h, j_h)\} : c(i_h, j_h) = k,\ h = 1, ..., m_k,\ \mathbf{x}(i_h, j_h) \in \mathfrak{X}\},$$

i.e. the training set is the union of the subsets

$$\mathbb{T}^0_{\mathbf{m}} = \bigcup_{k=1}^{k_0} \mathbb{T}^0_{m_k}, \quad \mathbf{m} = \{m_1, ..., m_{k_0}\}, \quad m. = \sum_{k=1}^{k_0} m_k. \tag{3}$$

Let $\mathfrak{X}^0_{\mathbf{m}}$ be the set of all training points $\{\mathbf{x}(i_h, j_h)\}_{1\le h\le m}$. We do not know the true classes $c(i,j)$ for $\mathbf{x}(i,j) \in \mathfrak{X} \smallsetminus \mathfrak{X}^0_{\mathbf{m}}$ which are not completely placed inside the plots. A classifier has to supply each such point $\mathbf{x}(i,j)$ with a class which we denote by $c^\bullet(i,j)$. If $c^\bullet(i,j) \neq c(i,j)$ then the classifier has made an error. It is desirable to find classifiers which do not do too many misclassifications. In the general case a classifier is a function $f^\bullet(\mathbf{x}, \mathbb{T}^0_{\mathbf{m}})$ defined on the feature space $\mathfrak{X} = (\mathbf{x})$ which values are numbers in $\mathcal{K}_0 = \{1, ..., k_0\}$. Usually classifiers essentially depend on the training sets.

We will consider some special classifiers. We suppose that $\mathfrak{X} \subseteq \mathbb{R}^d$, $d \ge 1$. We supply the feature space $\mathfrak{X}$ with a distance function $d(\cdot,\cdot)$. For each point $\mathbf{x} \in \mathfrak{X}$ we can find the k nearest neighbor $(NN\text{-})$points in $\mathfrak{X}^0_{\mathbf{m}}$, i.e.
$\mathbf{x}_{(1)} = \mathbf{x}(i_{h_1}, j_{h_1})$, $d(\mathbf{x}, \mathbf{x}_{(1)}) = \min_{1\le h\le m} d(\mathbf{x}, \mathbf{x}(i_h, j_h))$, $\mathbf{x}_{(2)} = \mathbf{x}(i_{h_2}, j_{h_2})$,
$d(\mathbf{x}, \mathbf{x}_{(2)}) = \min_{1\le h\le m, h\neq h_1} d(\mathbf{x}, \mathbf{x}(i_h, j_h)), ..., \mathbf{x}_{(k)} = \mathbf{x}(i_{h_k}, j_{h_k})$,
$d(\mathbf{x}, \mathbf{x}_{(k)}) = \min_{1\le h\le m, h\neq h_l, l<k} d(\mathbf{x}, \mathbf{x}(i_h, j_h))$, $\mathbf{x}(i_h, j_h) \in \mathfrak{X}^0_{\mathbf{m}}$. The true classes $c(l) = c(i_{h_l}, j_{h_l})$, $l = 1, ..., k$, are known from the field data. The sequence of true classes of the k NN-points $\{\mathbf{x}_{(1)}, ..., \mathbf{x}_{(k)}\}$ to $\mathbf{x}$ is denoted by $\mathbf{c}^k_1(\mathbf{x}) = \{c(1), ..., c(k)\}$.

Definition 1. *The classifier $f^\bullet_{1NN}(\mathbf{x}, \mathbb{T}^0_{\mathbf{m}}) = c(1)$ is called the nearest neighbor classifier or shortly $1NN$-classifier.*

The $1NN$-classifier supplies any point $\mathbf{x} \in \mathfrak{X}$ with the class $c(1) = c(i_{h_1}, j_{h_1})$ of the NN-point $\mathbf{x}(i_{h_1}, j_{h_1})$. Let $\mathcal{K}^k_0 = \mathcal{K}_0 \times ... \times \mathcal{K}_0 = (\{c_1, ..., c_k\})$ be the set of all sequences of length k with $c_l \in \mathcal{K}_0$. Let $g_k(\cdot) : \mathcal{K}^k_0 \to \mathcal{K}_0$ be a function defined on $\mathcal{K}^k_0$ which values are numbers of classes in $\mathcal{K}_0$. For example, if $k = 3$ and $\mathcal{K}_0 \in \{1, 2, 3\}$ one can define

$$g_3(c_1, c_2, c_3) = \begin{cases} c_1, & \text{if } c_1, c_2, c_3 \text{ are different,} \\ c, & \text{if at least 2 of } c_{(1)}, c_{(2)}, c_{(3)} \text{ equal } c. \end{cases} \tag{4}$$

Definition 2. *The classifier $f^\bullet_{kNN}(\mathbf{x}, \mathbb{T}^0_{\mathbf{m}}) = g_k(\mathbf{c}^k_1(\mathbf{x}))$ is called the kNN-classifier.*

Note that the numbers of misclassifications essentially depend on the selection of the feature space $\mathfrak{X}$ and the distance $d(\cdot,\cdot)$. We consider the vectors $\mathbf{x}(i,j) = \mathbf{f}_{RS}(\mathbf{e}(i,j))$ as values of the random variables $\mathbf{X}(i,j)$. The distributions of these r.v.s essentially depend on the types of trees which grow on $\mathcal{A}_{ij} \subset \mathcal{A}$. We introduce the following basic assumption.

Assumption A1. *For all classes of forest $c \in \mathcal{K}_0$ the feature values $\mathbf{x}(i,j)$ given $c(i,j) = c \in \mathcal{K}_0$, $\mathcal{A}_{ij} \subset \mathcal{A}_1$ are values of independent random variables with continuous probability distributions $\mathrm{P}_c[\cdot]$ defined on the Borel σ-algebra $\mathfrak{B}(\mathfrak{X})$, and $\mathrm{P}_{c_1}[\cdot] \not\equiv \mathrm{P}_{c_2}[\cdot]$ if $c_1 \neq c_2$.*

Suppose that energies reflected from disjoint parts of trees growing in the surface elements are independent r.v.s which distributions depend on the type of trees, e.g. pine, spruce and broad leaved trees. Then the registered energies within each band are also independent r.v.s. The distributions of these energies are defined by the sizes $z_P(i,j), z_S(i,j), z_B(i,j)$ of the elements' parts occupied by pines, spruce and broad leaved trees.

The set of classes $\mathcal{K}_0$ is defined by a division of the simplex $\mathcal{S}_3$ into disjoint subsets. We can consider assumption **A1** as a reasonable approximation if all classes $c \in \mathcal{K}_0$ are defined by sufficiently small subsets of $\mathcal{S}_3$ because then the positions of the points $\mathbf{z}(i,j) = \{z_P(i,j), z_S(i,j), z_B(i,j)\}$ within these small subsets are approximately identically distributed. If a satellite sensor has a high resolution then, hopefully, **A1** approximately holds and we need to consider only 7 classes P, S, B, PS, PB, BS, PSB defined in the previous section. In this case we have no need to consider the division of $\mathcal{S}_3$ into many classes which are small subsets. In general if there are classes where the field data show that the points $\mathbf{z}(i,j)$ are not identically distributed then it is necessary to divide these classes in two or more smaller subclasses where the points $\mathbf{z}(i,j)$ can be considered as identically distributed.

The following more technical assumption can be used in the justification of the below suggested methods.

Assumption A2. *For any class $c \in \mathcal{K}_0$ the probability distribution P_c has continuous positive density $p_c(\mathbf{x})$.*

It follows from **A2** that there are no ties, i.e. $\mathbf{x}(i',j') \neq \mathbf{x}(i'',j'')$ if $\{i',j'\} \neq \{i'',j''\}$. Note also that we do not assume that $\mathrm{P}_c[\cdot]$, $c \in \mathcal{K}_0$, are known distributions. If $\mathbf{x}(i,j) \in \mathfrak{X} \setminus \mathfrak{X}^0_{\mathbf{m}}$, $\mathcal{A}_{ij} \subset \mathcal{A}$, then we do not know the true value $c(i,j)$ and we do not assume that $\mathbf{x}(i,j) \in \mathfrak{X} \setminus \mathfrak{X}^0_{\mathbf{m}}$ are values of independent r.v.s. We use capital letters for r.v.s, e.g. $\mathbf{X}(i,j) = \mathbf{x}(i,j)$ means that $\mathbf{x}(i,j)$ is an observed value of the r.v. $\mathbf{X}(i,j)$, $c(1) = c$ means that the true class $c(1)$ of the NN-point to $\mathbf{x} \in \mathfrak{X}$, is $c \in \mathcal{K}_0$. Suppose that the unknown true class of the forest on $\mathcal{A}_{ij} \subset \mathcal{A}_1$ is $r \in \mathcal{K}_0$, $\mathbf{x}(i,j) \in \mathfrak{X} \setminus \mathfrak{X}^0_{\mathbf{m}}$, and s is the class of the forest on $\mathcal{A}_{ij}$ suggested by a classifier $f^\bullet(\cdot, \mathbb{T}^0_{\mathbf{m}})$, i.e. $f^\bullet(\mathbf{x}(i,j), \mathbb{T}^0_{\mathbf{m}}) = s \in \mathcal{K}_0$. Formally, we can write the probability of this event as

$$p_{rs}(f^\bullet, \mathbb{T}^0_{\mathbf{m}}) = \mathrm{P}_r[f^\bullet(\mathbf{X}(i,j), \mathbb{T}^0_{\mathbf{m}}) = s \mid \mathbb{T}^0_{\mathbf{m}}]. \tag{5}$$

Let $f^{\bullet}_{kNN}$ be a kNN-classifier with a function $g_k(\cdot)$ on $\mathcal{K}_0^k$. Then we can write (5) as follows

$$p_{rs}(f^{\bullet}_{kNN}, \mathbb{T}^0_{\mathbf{m}}) = \mathrm{P}_r[g_k(\mathbf{c}_1^k(\mathbf{X}(i,j))) = s \mid \mathbb{T}^0_{\mathbf{m}}], \tag{6}$$

where we consider the training set $\mathbb{T}^0_{\mathbf{m}}$ as fixed and a r.v. $\mathbf{X}(i,j)$ has the distribution $\mathrm{P}_r[\,\cdot\,]$. The matrix $\mathbb{P}(f^{\bullet}, \mathbb{T}^0_{\mathbf{m}}) = (p_{rs}(f^{\bullet}, \mathbb{T}^0_{\mathbf{m}}))_{r,s\in\mathcal{K}_0}$ is called the *confusion matrix* of a classifier $f^{\bullet}$ given a training set $\mathbb{T}^0_{\mathbf{m}}$. Henceforth, we write in short $p_{rs}(\mathbb{T}^0_{\mathbf{m}})$ and $\mathbb{P}(\mathbb{T}^0_{\mathbf{m}})$ instead of $p_{rs}(f^{\bullet}, \mathbb{T}^0_{\mathbf{m}})$ and $\mathbb{P}(f^{\bullet}, \mathbb{T}^0_{\mathbf{m}})$.

The confusion matrix is the most important characteristic of each classifier. The probabilities $p_{rs}(\mathbb{T}^0_{\mathbf{m}})$, $r, s \in \mathcal{K}_0$, are unknown and it is very difficult to find out how accurate we may estimate of their values. If probabilities (5) or (6) would be known then they could be used in assessing the accuracy of the created digital maps of the area $\mathcal{A}_0$, Belyaev (2000, 2003a). If accurate estimators of the probabilities $p_{rs}(f^{\bullet}_a, \mathbb{T}^0_{\mathbf{m}})$ and $p_{rs}(f^{\bullet}_b, \mathbb{T}^0_{\mathbf{m}})$, $r, s \in \mathcal{K}_0$, would be found then they could be used to select the better one of the two classifiers $f^{\bullet}_a, f^{\bullet}_b$.

4 CV-estimators of confusion matrices of NN-classifiers

In this section we consider possible methods of assessing the accuracy of classifiers based on distances between points in $\mathbb{T}^0_{\mathbf{m}}$ corresponding to a training set. The main difficulty is bound with the presence of dependencies between values of classifiers related to nearly placed training points in $\mathfrak{X}^0_{\mathbf{m}}$. Here we give somewhat heuristical justification of the below suggested resampling methods. In the next Section 5 we test these resampling methods in two numerical experiments. A more rigorous justification will take many more pages and is worth to be published in separate papers.

We will use a common cross-validation CV-method in order to obtain values of estimators for $\mathbb{P}(\mathbb{T}^0_{\mathbf{m}}) = (p_{rs}(\mathbb{T}^0_{\mathbf{m}}))_{r,s\in\mathcal{K}_0}$. The CV-method can be described as follows. Let $\mathbf{x}_{ri} \in \mathfrak{X}^0_{\mathbf{m}}$ and the true class is r, i.e. the element $\{\mathbf{x}_{ri}, r\} \in \mathbb{T}^0_{\mathbf{m}}$. We exclude this element from $\mathbb{T}^0_{\mathbf{m}}$ and consider $\mathbb{T}^0_{\mathbf{m}(r,i)} = \mathbb{T}^0_{\mathbf{m}} \setminus \{\mathbf{x}_{ri}, r\}$ as the reduced training set. The classifier based on $\mathbb{T}^0_{\mathbf{m}(r,i)}$ "recognizes" the point $\mathbf{x}_{ri}$ as having the class s, if $f^{\bullet}(\mathbf{x}_{ri}, \mathbb{T}^0_{\mathbf{m}(r,i)}) = s$. If $s = r$ then the classification is correct otherwise it is erroneous. Similarly we can apply the classifier $f^{\bullet}(\cdot)$ to all points in $\mathfrak{X}^0_{\mathbf{m}}$. For each pair of classes $r, s \in \mathcal{K}_0$ we can find the following frequencies

$$\hat{p}^{CV}_{rs}(\mathbb{T}^0_{\mathbf{m}}) = \frac{1}{m_r}\sum_{i=1}^{m_r} \mathrm{I}(f^{\bullet}(\mathbf{x}_{ri}, \mathbb{T}^0_{\mathbf{m}(r,i)}) = s), \tag{7}$$

where the sum is taken over all m_r elements $\{\mathbf{x}_{ri}, r\}$. The frequencies $\hat{p}_{rs}^{CV}(\mathbb{T}_{\mathbf{m}}^0)$ are called the *CV-estimates* of the CC-probabilities (5).

From (7) we see that the values $\hat{p}_{rs}^{CV}(\mathbb{T}_{\mathbf{m}}^0)$ can essentially depend on the original training set $\mathbb{T}_{\mathbf{m}}^0$. We have to consider $\mathbb{T}_{\mathbf{m}}^0$ as a value of a random training set $\mathbb{T}_{\mathbf{m}}$. Is it possible to find the distribution of the deviation $\hat{p}_{rs}^{CV}(\mathbb{T}_{\mathbf{m}}^0)$ from the true unknown value $p_{rs}(\mathbb{T}_{\mathbf{m}}^0)$? If we consider the training set $\mathbb{T}_{\mathbf{m}}$ as random and $\mathbb{T}_{\mathbf{m}}^0$ as an observed value of $\mathbb{T}_{\mathbf{m}}$ then we are interested to know the cumulative distribution function (c.d.f.) $F_{rs\mathbf{m}}(z) = \mathrm{P}[\hat{p}_{rs}^{CV}(\mathbb{T}_{\mathbf{m}}) - p_{rs}(\mathbb{T}_{\mathbf{m}}) \leq z]$ of the deviations. Can we estimate $F_{rs\mathbf{m}}(\cdot)$ consistently if all $m_r \to \infty$, $r \in \mathcal{K}_0$, i.e. is there an estimator $\hat{F}_{rs\mathbf{m}}(\cdot)$ which converges to $F_{rs\mathbf{m}}(\cdot)$ in probability if all m_r tend to infinity? This is a complex problem and we will try to solve it.

For a short time we introduce the following temporary assumption. Suppose that we know all probability distributions $\mathrm{P}_r[\,]$, $r \in \mathcal{K}_0$. Then, for any given $\mathbb{T}_{\mathbf{m}}$, we can simulate any number n of points $\mathbf{x}_r(i) \in \mathfrak{X}$, $i = 1, ..., n$ which are values of i.i.d. r.v.s with probability distribution $\mathrm{P}_r[\cdot]$ and true class r. Then by the law of large numbers (LLN) the exact values of the CC-probabilities (5) for any given value of $\mathbb{T}_{\mathbf{m}}$ would be found as the following limit

$$p_{rs}(\mathbb{T}_{\mathbf{m}}) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}(f^{\bullet}(\mathbf{x}_r(i), \mathbb{T}_{\mathbf{m}}) = s), \quad r, s \in \mathcal{K}_0. \tag{8}$$

We stress that the exact values of the CC-probabilities are depend on the training sets. Under the temporary assumption we could also simulate any number of independent copies $\mathbb{T}_{\mathbf{m}}^1, ..., \mathbb{T}_{\mathbf{m}}^n, ...$ of training sets which are values of random sets identically distributed as the original training set $\mathbb{T}_{\mathbf{m}}^0$. For each training set $\mathbb{T}_{\mathbf{m}}^j$ similarly as in (7) and (8) we could be able to find the values of the CV-estimators $\hat{p}_{rs}^{CV}(\mathbb{T}_{\mathbf{m}}^j)$ and their exact values $p_{rs}(\mathbb{T}_{\mathbf{m}}^j)$, and thus write the following list of deviations

$$\{\hat{p}_{rs}^{CV}(\mathbb{T}_{\mathbf{m}}^j) - p_{rs}(\mathbb{T}_{\mathbf{m}}^j)\}_{1\leq j\leq n}, \tag{9}$$

where n is a large number, say $n \geq 2000$. The larger n the better.

From (9) as $n \to \infty$ we find the c.d.f.s of interest

$$F_{rs\mathbf{m}}(z) = \lim_{n\to\infty} F_{rs\mathbf{m}n}(z), \quad r, s \in \mathcal{K}_0, \tag{10}$$

where $F_{rs\mathbf{m}n}(z) = \frac{1}{n}\sum_{j=1}^{n} \mathrm{I}(\hat{p}_{rs}^{CV}(\mathbb{T}_{\mathbf{m}}^j) - p_{rs}(\mathbb{T}_{\mathbf{m}}^j) \leq z)$. If we know $F_{rs\mathbf{m}}(z)$ or close to it $F_{rs\mathbf{m}n}(z)$ with a large n then we can say how typical the (unknown!) deviation $\hat{p}_{rs}^{CV}(\mathbb{T}_{\mathbf{m}}^0) - p_{rs}(\mathbb{T}_{\mathbf{m}}^0)$ can be if the original training set is $\mathbb{T}_{\mathbf{m}}^0$. The c.d.f.s $F_{rs\mathbf{m}n}(\cdot)$ have been obtained by simulation in two artificial numerical experiments (see Section 5). Here, we consider $\mathbb{T}_{\mathbf{m}}$ as a r.v. and

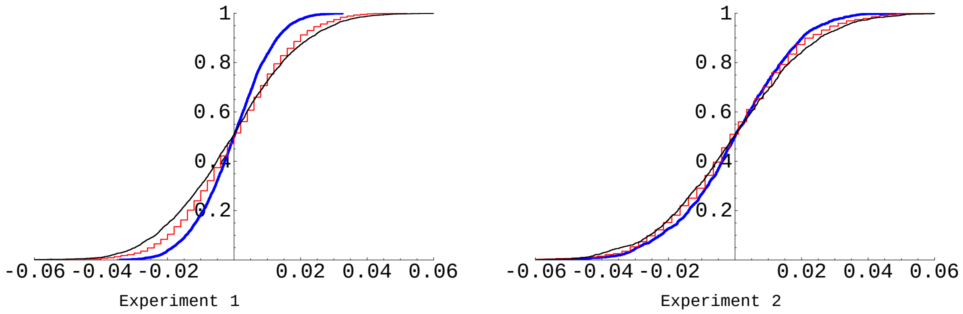

Figure 7: Three distribution functions of deviations $\hat{p}_{11}^{CV}(\mathbb{T}_{\mathbf{m}}) - p_{11}(\mathbb{T}_{\mathbf{m}})$, $\hat{p}_{11}^{CV}(\mathbb{T}_{\mathbf{m}}) - \operatorname{med}\hat{p}_{11}^{CV}(\mathbb{T}_{\mathbf{m}})$ and $p_{11}(\mathbb{T}_{\mathbf{m}}) - \operatorname{med} p_{11}(\mathbb{T}_{\mathbf{m}})$ are shown by thin, stepwise and thick lines, respectively.

then $\hat{p}_{rs}^{CV}(\mathbb{T}_{\mathbf{m}})$, $p_{rs}(\mathbb{T}_{\mathbf{m}})$ are also r.v.s. We denote the medians of the c.d.f.s related to these r.v.s $\hat{p}_{rs}^{CV}(\mathbb{T}_{\mathbf{m}})$ and $p_{rs}(\mathbb{T}_{\mathbf{m}})$ by $\operatorname{med}\hat{p}_{rs}^{CV}(\mathbb{T}_{\mathbf{m}})$ and $\operatorname{med} p_{rs}(\mathbb{T}_{\mathbf{m}})$, respectively.

In Fig. 7 the c.d.f.s of deviations $\hat{p}_{rs}^{CV}(\mathbb{T}_{\mathbf{m}})-p_{rs}(\mathbb{T}_{\mathbf{m}})$, $\hat{p}_{rs}^{CV}(\mathbb{T}_{\mathbf{m}})-\operatorname{med}\hat{p}_{rs}^{CV}(\mathbb{T}_{\mathbf{m}})$, and $p_{rs}(\mathbb{T}_{\mathbf{m}}) - \operatorname{med} p_{rs}(\mathbb{T}_{\mathbf{m}})$, $r = s = 1$, are shown by thin black, stepwise, and thick black lines, respectively. It shows that the typical deviations $\hat{p}_{rs}^{CV}(\mathbb{T}_{\mathbf{m}})$ from $p_{rs}(\mathbb{T}_{\mathbf{m}})$ often can be larger than the typical deviations of $\hat{p}_{rs}^{CV}(\mathbb{T}_{\mathbf{m}})$ and $p_{rs}(\mathbb{T}_{\mathbf{m}})$ from their medians. If one knows $F_{rs\mathbf{m}n}(\cdot)$, which is close to $F_{rs\mathbf{m}}(\cdot)$, then it is possible to obtain confidence intervals for the CV-estimators $\hat{p}_{rs}^{CV}(\mathbb{T}_{\mathbf{m}}^{0})$, $r, s \in \mathcal{K}_0$. Estimates $\hat{F}_{rs\mathbf{m}}(\cdot)$ of $F_{rs\mathbf{m}}(\cdot)$ can be used in assessing the numbers of correctly and erroneously classified pixels in a preliminary version of the digital image (map) of interest. If we could estimate consistently $F_{rs\mathbf{m}}(z)$ by using for each $\mathbf{m}$ only the original training set $\mathbb{T}_{\mathbf{m}}^{0}$ as all $m_k \to \infty$ then we could drop the temporary assumption, that we know all $\mathrm{P}_r[\,]$, $r \in \mathcal{K}_0$, which is absolutely unrealistic and which we have used here only to simplify the understanding of this complex problem. We can exclude this temporary assumption by using instead the theory of resampling methods from non-homogeneous data which has been developed and applied in a series of papers: Belyaev (1996, 2003b), Belyaev and Sjöstedt-de Luna (2000), Ekström and Belyaev (2001), Ekström and Sjöstedt-de Luna (2004).

In order to justify appropriate resampling methods we consider the original training set $\mathbb{T}_{\mathbf{m}}^{0}$ as a realisation of a marked point process, where the marks are the classes $k \in \mathcal{K}_0$. Let us consider vector-valued r.v.s $\mathbf{M} = \{M_1, ..., M_{k_0}\}$ which components are independent and which have the Poisson distributions with mean values $m_1, ..., m_{k_0}$. Let $\mathbb{T}_{M_r} = \{\{\mathbf{X}_{ri}, r\} : \ i = 1, ..., M_r\}$, where $\{\mathbf{X}_{ri}\}$ are i.i.d. $\mathrm{P}_r[\ ]$-distributed r.v.s. The resulting superposition $\mathbb{T}_{\mathbf{M}} = \cup_{r=1}^{k_0} \mathbb{T}_{M_r}$ is the marked Poisson point process with independent discretely distributed marks and $\mathfrak{X}_{\mathbf{M}} = \{\mathbf{X}_{ri} : \{\mathbf{X}_{ri}, r\} \in \mathbb{T}_{\mathbf{M}},\ i = 1, ..., M_r,\ r \in \mathcal{K}_0\}$. We can approximate the probabilities of events which can occur in the training set $\mathbb{T}^0_{\mathbf{m}}$ by using the distribution of $\mathbb{T}_{\mathbf{M}}$ if all m_k, $k \in \mathcal{K}_0$ are sufficiently large.

Now we will consider the training set $\mathbb{T}^0_{\mathbf{m}}$ as a realisation of $\mathbb{T}_{\mathbf{M}}$. Here, we still assume that we know all $\mathrm{P}_r[\ ]$, $r \in \mathcal{K}_0$. The appropriate realisation of the random set $\mathbb{T}_{\mathbf{M}}$ can be obtained by adding to $\mathbb{T}_{\mathbf{m}}$ the i.i.d. pairs $\{\mathbf{X}_{ki}, k\}$, $i = m_k+1, ..., M_k$, if $M_k > m_k$, $k = 1, ..., k_0$ or by excluding from $\mathbb{T}_{\mathbf{M}}$ the randomly chosen $m_k - M_k$ pairs if $m_k > M_k$, $k = 1, ..., k_0$. Suppose that $\underline{\lim}_{m.\to\infty} \frac{m_k}{m.} > 0$ for all $k \in \mathcal{K}_0$, $m. = m_1 + \cdots + m_{k_0}$. In short, we denote this asymptotic $\mathbf{m} \rightrightarrows \infty$. If $\mathbf{m} \rightrightarrows \infty$ then $\mid M_k - m_k \mid = O_p(m_k^{-1/2})$, $k = 1, ..., k_0$. Therefore, the probabilities, of many similarly defined on realisations of $\mathbb{T}_{\mathbf{M}}$ and $\mathbb{T}_{\mathbf{m}}$ events of interest, are asymptotically equivalent as $\mathbf{m} \rightrightarrows \infty$. Conditionally $\mathbb{T}_{\mathbf{M}} = \mathbb{T}_{\mathbf{m}}$ if $\mathbf{M} = \mathbf{m}$.

Recall that $p_k(\mathbf{x})$ is the probability density introduced in Assumption **A2**. The marked Poisson point process corresponding to $\mathbb{T}_{\mathbf{M}}$ we denote by $\mathfrak{P}_{\mathbf{M}}(\lambda_{\mathbf{m}}(\cdot))$, where the intensity measure$\lambda_{\mathbf{m}}(\cdot) = \sum_{k=1}^{k_0} m_k p_k(\cdot)$. We write $\mathfrak{P}(\lambda_{\mathbf{m}}(\cdot))$ if we consider points of $\mathfrak{P}_{\mathbf{M}}(\lambda_{\mathbf{m}}(\cdot))$ without marks. Hence, we can approximately consider $\mathfrak{X}^0_{\mathbf{m}}$ as a realisation of the marked non-homogeneous Poisson point process $\mathfrak{P}_{\mathbf{M}}(\lambda_{\mathbf{m}}(\cdot))$ with the first moment intensity measure $\lambda_{\mathbf{m}}(\mathbf{x}) = \sum_{k=1}^{k_0} m_k p_k(\mathbf{x})$, $\mathbf{x} \in \mathfrak{X} \subseteq \mathbb{R}^d$, and marks are independently placed at $\mathbf{x} \in \mathfrak{X}^0_{\mathbf{m}}$ with the discrete distribution with the probabilities

$$q_k(\mathbf{x}) = \frac{m_k p_k(\mathbf{x})}{\sum_{k'=1}^{k_0} m_{k'} p_{k'}(\mathbf{x})}, \quad k = 1, ..., k_0. \tag{11}$$

Let $\mathbf{X}_r$ be a r.v. with the distribution $\mathrm{P}_r[\cdot]$ and recall that $\mathbb{T}^0_{\mathbf{m}(r,i)} = \mathbb{T}^0_{\mathbf{m}} \setminus \{\mathbf{x}_{ri}, r\}$. Then

$$p_{rs}(\mathbb{T}^0_{\mathbf{m}}) = \mathrm{E}_r[\mathrm{I}(f^\bullet(\mathbf{X}_r, \mathbb{T}_{\mathbf{M}}) = s) \mid \mathbb{T}_{\mathbf{M}} = \mathbb{T}^0_{\mathbf{m}}],$$

and

$$p_{rs}(\mathbb{T}^0_{\mathbf{m}(r,i)}) = \mathrm{E}_r[\mathrm{I}(f^\bullet(\mathbf{X}_r, \mathbb{T}_{\mathbf{M}(r,i)}) = s) \mid \mathbb{T}_{\mathbf{M}(r,i)} = \mathbb{T}^0_{\mathbf{m}(r,i)}]. \tag{12}$$

From (12) we can write the deviation of $\hat{p}_{rs}^{CV}(\mathbb{T}_{\mathbf{m}}^0)$ from $p_{rs}(\mathbb{T}_{\mathbf{m}}^0)$ as follows

$$\begin{aligned}
&\hat{p}_{rs}^{CV}(\mathbb{T}_{\mathbf{m}}^0) - p_{rs}(\mathbb{T}_{\mathbf{m}}^0) \\
&= \frac{1}{m_r}\sum_{i=1}^{m_r}\Big(\mathrm{I}(f^{\bullet}(\mathbf{x}_{ri}, \mathbb{T}_{\mathbf{m}(r,i)}^0) = s) - p_{rs}(\mathbb{T}_{\mathbf{m}(r,i)}^0)\Big) \\
&\quad - \frac{1}{m_r}\sum_{i=1}^{m_r}(p_{rs}(\mathbb{T}_{\mathbf{m}}^0) - p_{rs}(\mathbb{T}_{\mathbf{m}(r,i)}^0)). \qquad (13)
\end{aligned}$$

The bias of the CV-estimator $\hat{p}_{rs}^{CV}(\mathbb{T}_{\mathbf{m}}^0)$ is equal to the last sum in (13) and we call it the CV-bias. For kNN-classifiers the bias has order $O(m_r^{-1})$, i.e. the CV-estimators are asymptotically unbiased. This fact can be easily proved in the case of $1NN$-classifier with the feature space $\mathfrak{X} \subseteq \mathbb{R}^d$, $d \geq 1$.

The terms in the first sum (13) are values of r.v.s with zero means. But these r.v.s. are dependent and, therefore, it is not correct to use the ordinary resampling methods for independent r.v.s. For example, if a kNN-classifier is used then the i-th term in the CV-estimator (7) is defined by the set of k NN-points in $\mathbb{T}_{\mathbf{m}(r,i)}^0$ which are nearest to $\mathbf{x}_{ri}$, $i = 1, ..., m_r$. Some of these sets with k NN-points can have nonempty intersections, i.e. they contain the same points for different $\mathbf{x}_{ri_1}$ and $\mathbf{x}_{ri_2}$, $i_1 \neq i_2$, therefore, these terms in (7) have to be considered as values of dependent r.v.s.

Let $\mathcal{B}(\mathbf{x}_0, r_0) = \{\mathbf{x} : \mathbf{x} \in \mathfrak{X} \subseteq \mathbb{R}^d,\ d(\mathbf{x}_0, \mathbf{x}) \leq r_0\}$, $d \geq 1$. By $\mathbf{x}_{(k)ri}$ we denote the training point which is the k-th closest to $\mathbf{x}_{ri}$. The set of all balls $\{\mathcal{B}(\mathbf{x}_{ri}, d(\mathbf{x}_{ri}, \mathbf{x}_{(k)ri})) : \mathbf{x}_{ri} \in \mathfrak{X}_{\mathbf{m}}^0,\ i = 1, ..., m_r,\ r \in \mathcal{K}_0\}$ is a cover of $\mathfrak{X}_{\mathbf{m}}^0$ and in short we call it the kNN-scale. In Fig. 9 (left) a part of a $3NN$-scale is shown. If the balls related to terms in (7) are disjoint then from assumption **A1** it follows that the terms can be considered as values of independent r.v.s. The idea, to use growing blocks for raster spatially dependent data, has been suggested in Hall (1985) and it was elaborated to the case of sums of raster m-dependent non-identically distributed r.v.s in Belyaev (1996), Ekström and Belyaev (2001).

The case with kNN-classifiers ($k > 1$) essentially differs from the cases with blocks based on raster data. In the case of the $1NN$-classifier we can handle the difficulty with the dependency between terms in (7) by using resampling from specially created clusters of points in the set $\mathfrak{X}_{\mathbf{m}}$. Henceforth, we will consider only the case of the $1NN$-classifier, $f_{1NN}^{\bullet}(\cdot)$. In the general case of kNN-classifiers it is necessary to use resampling from sums of terms in (7) related to a set of growing special blocks which are subsets of $\mathfrak{X}_{\mathbf{m}}$. The idea, to use resampling from growing blocks of terms in (7) in order to estimate the

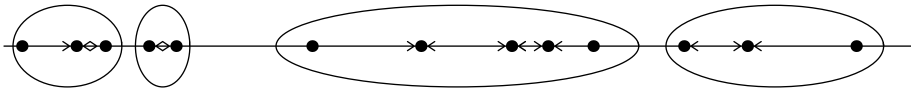

Figure 8: Four NN-clusters, $\mathfrak{X}_{\mathbf{m}} \subset \mathbb{R}^1$. Arrows show the nearest neighbor points.

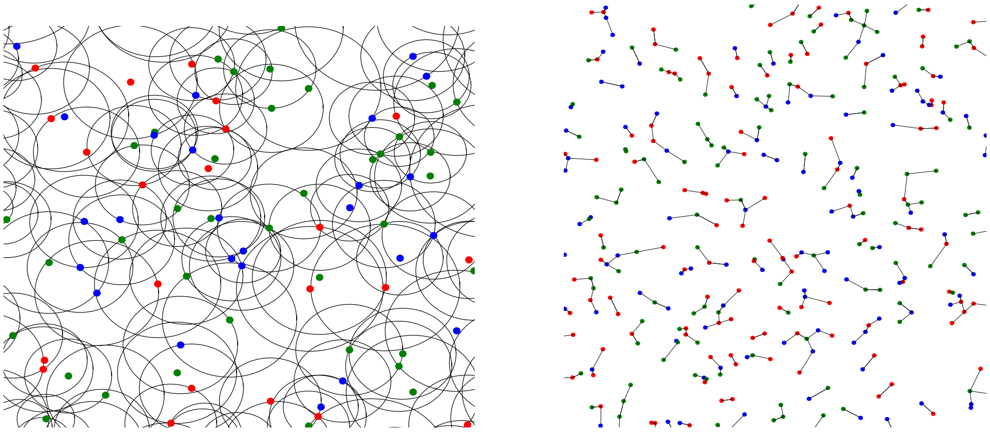

Figure 9: In the left part of the figure the $3NN$-scale of circles, related to $3NN$-classifiers defined in (4), is shown. In the right part of the figure NN-clusters are shown. They are formed by connected training points in a realisation of $\mathfrak{X}_{\mathbf{m}} \subset \mathbb{R}^2$.

c.d.f. of the deviations the CV-estimators from the true CC-probabilities has to be elaborated. It will be done in a separate paper. Here, we concentrate on the case with $1NN$-classifiers.

Definition 3. *Two points $\mathbf{x}', \mathbf{x}'' \in \mathfrak{X}_{\mathbf{m}}$ are called connected if at least one of them is the NN-point to the other, i.e. one or both of the two following relations holds $d(\mathbf{x}', \mathbf{x}'') = \min_{\mathbf{x} \in \mathfrak{X}_{\mathbf{m}} \setminus \mathbf{x}'} d(\mathbf{x}', \mathbf{x})$ or $d(\mathbf{x}', \mathbf{x}'') = \min_{\mathbf{x} \in \mathfrak{X}_{\mathbf{m}} \setminus \mathbf{x}''} d(\mathbf{x}, \mathbf{x}'')$. A set $\mathcal{C}$ is called NN-cluster if in any disjoint subsets $\mathcal{C}', \mathcal{C}'', \mathcal{C} = \mathcal{C}' \cup \mathcal{C}''$ there are points $\mathbf{x}' \in \mathcal{C}'$, $\mathbf{x}'' \in \mathcal{C}''$ which are connected, and if any point $\mathbf{x} \in \mathfrak{X}_{\mathbf{m}} \setminus \mathcal{C}$, i.e. $\mathbf{x}$ is not in $\mathcal{C}$, then $\mathbf{x}$ has no connection with points which are in $\mathcal{C}$.*

Typical NN-clusters are shown in the following Fig. 8, $\mathfrak{X}^0_{\mathbf{m}} \subseteq \mathbb{R}^1$, and in Fig. 9 (right), $\mathfrak{X}^0_{\mathbf{m}} \subseteq \mathbb{R}^2$. Note that the NN-clusters can be considered as random graphs without cycles.

Let $\{\mathcal{C}_h\}_{1\le h\le n_{CL}}$ be the division of the set $\mathfrak{X}_{\mathbf{m}}$ of the training points without marks into $n_{CL} = n_{CL}(\mathfrak{X}^0_{\mathbf{m}})$ NN-clusters, $\mathcal{C}_{h_1} \cap \mathcal{C}_{h_2} = \emptyset$, if $h_1 \neq h_2$, $\cup_{h=1}^{n_{CL}} \mathcal{C}_h = \mathfrak{X}^0_{\mathbf{m}}$. Here, the numbering of NN-clusters is arbitrary and it does not depend on shapes and positions of NN-clusters. From assumption **A1** it follows that all points in $\mathfrak{X}^0_{\mathbf{m}}$ have different components of their coordinates. By $M(\mathcal{C}_h)$ we denote the random number of points in the NN-cluster $\mathcal{C}_h$. We can number these points as follows. Note that there are only two points in any $\mathcal{C}_h$ which are nearest each other. We give them numbers 1 and 2 and denote them by $\mathbf{X}^j_h = \{X^j_{h1}, ..., X^j_{hd}\}$, $j = 1, 2$, $X^1_{h1} < X^2_{h1}$. If there are other points in $\mathcal{C}_h$ then they obtain numbers $3, ..., M(\mathcal{C}_h)$ due to their distances from $\mathbf{X}^1_h$, i.e. $d(\mathbf{X}^1_h, \mathbf{X}^2_h) < d(\mathbf{X}^1_h, \mathbf{X}^3_h) < \cdots < d(\mathbf{X}^1_h, \mathbf{X}^{M(\mathcal{C}_h)}_h)$. We call $\mathbf{X}^1_h$ as the root of $\mathcal{C}_h$. We can use notation $\mathcal{C}_h(\mathbf{X}^1_h)$ and $\mathcal{C}_h(\mathfrak{X}_{\mathbf{m}})$ if we want to stress that $\mathbf{X}^1_h$ is the root point of $\mathcal{C}_h$, and that $\mathcal{C}_h = \mathcal{C}_h(\mathfrak{X}_{\mathbf{m}}) \subseteq \mathfrak{X}_{\mathbf{m}}$.

Let us write $\{j_1, j_2\}$ if $\mathbf{X}^{j_2}_h$ is the nearest neighbor point to the point $\mathbf{X}^{j_1}_h$. The list $\mathcal{L}(\mathcal{C}_h) = \{\{1, 2\}, \{2, 1\}, \{3, j_3\}, ..., \{M(\mathcal{C}_h), j_{M(\mathcal{C}_h))}\}\}$ contains all $M(\mathcal{C}_h)$ such pair of points' numbers. The lists $\mathcal{L}(\mathcal{C}_1), ..., \mathcal{L}(\mathcal{C}_{n_{CL}})$ are invariant w.r.t. the synchronic extensions or contractions of all coordinates of points in $\mathfrak{X}_{\mathbf{m}}$ and the numbers $M(\mathcal{C}_h)$, $h = 1, ..., n_{CL}$ are also invariant. If $\mathbf{m} \rightrightarrows \infty$ then the linear sizes of the NN-clusters $\mathcal{C}_h$, $h = 1, ..., n_{CL}$, tend to zero. In a neighborhood of size $O(m_{\cdot}^{-1/d})$ around any point $\mathbf{x}_0$ with $p_{\cdot}(\mathbf{x}_0) = \sum_{k=1}^{k_0} p_k(\mathbf{x}_0) > 0$ the training points can be considered as a realisation of the homogeneous Point process $\mathfrak{P}(\lambda(\mathbf{x}_0))$ as $\mathbf{m} \rightrightarrows \infty$.

Due to the invariancy mentioned above, the distribution, of the number of points in a randomly chosen NN-cluster, can be asymptotically approximated by the distribution of a randomly chosen NN-cluster generated by the homogeneous Point process $\mathfrak{P}(\lambda(\mathbf{x}_0))$ with intensity $\lambda(\mathbf{x}_0) \equiv 1$. Let $\mathfrak{P}(1)$ be the homogeneous Poisson point process on the line $\mathbb{R}^1$. Then the mean number of points in a randomly taken NN-cluster $\mathcal{C}$ is $\mathrm{E}[M(\mathcal{C})] = 3$. If $\mathfrak{P}(1)$ is defined on the plane $\mathbb{R}^2$ then $\mathrm{E}[M(\mathcal{C})] = 2(6\pi + 2\sqrt{3})/(3\pi) = 5.102...$ Here, E[ ] relates to the Palm distributions, see Daley and Vere-Johens (1988). It is possible to show that the variance of $M(\mathcal{C})$ is finite if $\mathfrak{P}(1)$ is defined in $\mathbb{R}^d$, $d \ge 1$.

Henceforth, we consider $\mathbb{T}_{\mathbf{m}}$ as a random training set. Let the $1NN$-classifiers $f^{\bullet}_{1NN}(\cdot, \mathbb{T}_{\mathbf{m}(r,i)})$, $i = 1, ..., m_r$, are used and $p_{rs}(\mathbb{T}_{\mathbf{m}})$, $r, s \in \mathcal{K}_0$, are related cross-classification probabilities. If $\mathbf{X}_{ri'}$ is connected with its NN-point $\mathbf{X}_{si''}$ then both $\mathbf{X}_{ri'}$ and $\mathbf{X}_{si''}$ belong to the same NN-cluster and $f^{\bullet}_{1NN}(\mathbf{X}_{ri'}, \mathbb{T}_{\mathbf{m}(r,i')}) = s$.

We prove that the CV-estimator $\hat{p}^{CV}_{rs}$ has very small bias of order $O_p(m_r^{-1})$, $\mathbf{m} \rightrightarrows \infty$. Let $\mathfrak{X}_{\mathbf{m}}$ corresponds to the random set $\mathbb{T}_{\mathbf{m}}$. For each training point

$\mathbf{X}_{sj} \in \mathfrak{X}_{\mathbf{m}}$, $s \in \mathcal{K}_0$, we define the following open subset

$$\mathcal{U}(\mathbf{X}_{sj}, \mathfrak{X}_{\mathbf{m}}) = \{\mathbf{x} : \, d(\mathbf{x}, \mathbf{X}_{sj}) < \min_{\mathbf{X}_{s'j'} \in \mathfrak{X}_{\mathbf{m}} \setminus \mathbf{X}_{sj}} d(\mathbf{x}, \mathbf{X}_{s'j'})\}.$$

Note that these subsets are disjoint. If a point $\mathbf{x} \in \mathcal{U}(\mathbf{X}_{sj}, \mathfrak{X}_{\mathbf{m}})$ then the distance from $\mathbf{x}$ to the point $\mathbf{X}_{sj}$ is less than the distance from $\mathbf{x}$ to any other point $\mathbf{X}_{s'j'} \in \mathfrak{X}_{\mathbf{m}}$. We call $\mathcal{U}(\mathbf{X}_{sj}, \mathfrak{X}_{\mathbf{m}})$ the Voronoi cell related to $\mathbf{X}_{sj}$ (Okabe et al. (1992)).

Let $\mathbf{X}_r$ be a r.v. with the distribution $\mathrm{P}_r[\;]$ independent of the random training set $\mathbb{T}_{\mathbf{m}}$. The set $\mathfrak{X}_{\mathbf{m}(r,i)} = \mathfrak{X}_{\mathbf{m}} \setminus \mathbf{X}_{ri}$ is also random, and we have

$$p_{rs}(\mathbb{T}_{\mathbf{m}}) = \sum_{j=1}^{m_s} \mathrm{P}_r[\mathbf{X}_r \in \mathcal{U}(\mathbf{X}_{sj}, \mathfrak{X}_{\mathbf{m}})], \tag{14}$$

$$p_{rs}(\mathbb{T}_{\mathbf{m}(r,i)}) = \sum_{j=1}^{m_s} \mathrm{P}_r[\mathbf{X}_r \in \mathcal{U}(\mathbf{X}_{sj}, \mathfrak{X}_{\mathbf{m}(r,i)})]. \tag{15}$$

The difference of the indicators $\mathrm{I}(\mathbf{X}_r \in \mathcal{U}(\mathbf{X}_{sj}, \mathfrak{X}_{\mathbf{m}(r,i)})) - \mathrm{I}(\mathbf{X}_r \in \mathcal{U}(\mathbf{X}_{sj}, \mathfrak{X}_{\mathbf{m}}))$ can be non-zero almost surely only if $\mathbf{X}_r \in \mathcal{U}(\mathbf{X}_{ri}, \mathfrak{X}_{\mathbf{m}})$. Hence, it follows that

$$\mid p_{rs}(\mathbb{T}_{\mathbf{m}}) - p_{rs}(\mathbb{T}_{\mathbf{m}(r,i)}) \mid \leq \mathrm{P}_r[\mathbf{X}_r \in \mathcal{U}(\mathbf{X}_{ri}, \mathfrak{X}_{\mathbf{m}})]. \tag{16}$$

We can write

$$\mathrm{E}_r[\hat{p}_{rs}^{CV}(\mathbb{T}_{\mathbf{m}})] = \frac{1}{m_r} \sum_{i=1}^{m_r} \sum_{j=1}^{m_s} \mathrm{P}_r[\mathbf{X}_r \in \mathcal{U}(\mathbf{X}_{sj}, \mathfrak{X}_{\mathbf{m}(r,i)})]. \tag{17}$$

From (13) - (17) we can estimate the CV-bias of the CV-estimator as follows

$$\mid \mathrm{E}_r[\hat{p}_{rs}^{CV}(\mathbb{T}_{\mathbf{m}})] - p_{rs}(\mathbb{T}_{\mathbf{m}}) \mid = \left| \frac{1}{m_r} \sum_{i=1}^{m_r} (p_{rs}(\mathbb{T}_{\mathbf{m}(r,i)}) - p_{rs}(\mathbb{T}_{\mathbf{m}})) \right|$$

$$\leq \;\; \frac{1}{m_r} \sum_{i=1}^{m_r} \mathrm{P}_r[\mathbf{X}_r \in \mathcal{U}(\mathbf{X}_{ri}, \mathfrak{X}_{\mathbf{m}})] \leq \frac{1}{m_r}. \tag{18}$$

From (18) we have that the bias of $\hat{p}_{rs}^{CV}(\mathbb{T}_{\mathbf{m}})$ is negligibly small and the random parts of the deviations $\hat{p}_{rs}^{CV}(\mathbb{T}_{\mathbf{m}}) - p_{rs}(\mathbb{T}_{\mathbf{m}})$ characterize the accuracy of the CV-estimators as $\mathbf{m} \rightrightarrows \infty$.

Let $\mathcal{C}_1, ..., \mathcal{C}_{n_{CL}(r)}$ be the list of all NN-clusters containing one or more training points with mark r, $n_{CL}(r) = n_{CL}(r, \mathfrak{X}_{\mathbf{m}})$. Each NN-cluster $\mathcal{C}_h$ contains $M_{rsh} > 0$ training points $\mathbf{X}_{ri}$ from class r which have been recognized as belonging to class s by the $1NN$-classifiers with the training sets $\mathfrak{X}_{\mathbf{m}(r,i)}$, $i = 1, ..., m_r$. We have

$$M_{rs}(\mathbb{T}_{\mathbf{m}}) = \sum_{h=1}^{n_{CL}(r)} M_{rsh}, \quad M_{r\cdot}(\mathbb{T}_{\mathbf{m}}) = \sum_{s=1}^{k_0} M_{rs}(\mathbb{T}_{\mathbf{m}}) = \sum_{s=1}^{k_0} \sum_{h=1}^{n_{CL}(r)} M_{rsh}, \quad (19)$$

where

$$\begin{aligned} M_{rsh} &= \sum_{i\colon \mathbf{X}_{ri} \in \mathcal{C}_h} \mathrm{I}(f^{\bullet}_{1NN}(\mathbf{X}_{ri}, \mathbb{T}_{\mathbf{m}(r,i)}) = s) \\ &= \sum_{i:\{\mathbf{X}_{ri}, \mathbf{X}_{sj}\} \in \mathcal{C}_h} \mathrm{I}(\mathbf{X}_{ri} \in \mathcal{U}(\mathbf{x}_{sj}, \mathfrak{X}_{\mathbf{m}(r,i)})). \end{aligned} \quad (20)$$

From (19) and (20) we obtain that the CV-estimator can be written as follows

$$\hat{p}^{CV}_{rs}(\mathbb{T}_{\mathbf{m}}) = \frac{M_{rs}(\mathbb{T}_{\mathbf{m}})}{M_{r\cdot}(\mathbb{T}_{\mathbf{m}})} = \sum_{h=1}^{n_{CL}(r)} M_{rsh} \Bigg/ \sum_{h=1}^{n_{CL}(r)} M_{r\cdot h}, \quad (21)$$

where

$$M_{r\cdot h} = \sum_{s=1}^{k_0} M_{rsh}. \quad (22)$$

We will now try to investigate the distribution of the normed deviations

$$\sqrt{n_{CL}(r)}(\hat{p}^{CV}_{rs}(\mathbb{T}_{\mathbf{m}}) - p_{rs}(\mathbb{T}_{\mathbf{m}})) \quad \text{as} \quad \mathbf{m} \rightrightarrows \infty. \quad (23)$$

We need some properties of NN-clusters. The justification of them will be easier if instead of $\mathbb{T}_{\mathbf{m}}$ we will consider $\mathbb{T}_{\mathbf{M}}$, generated, as it has been described above, by the marked Poisson point process. The most of NN-clusters $\mathcal{C}_{h'}(\mathfrak{X}_{\mathbf{M}})$ in $\mathfrak{X}_{\mathbf{M}}$ are identical with the NN-clusters $\mathcal{C}_h(\mathfrak{X}_{\mathbf{m}})$ in $\mathfrak{X}_{\mathbf{m}}$. Here, we have that

$$\sum_h \sum_{h'} \mathrm{I}(\mathcal{C}_h(\mathfrak{X}_{\mathbf{m}}) = \mathcal{C}_{h'}(\mathfrak{X}_{\mathbf{M}})) = n_{CL}(\mathfrak{X}_{\mathbf{m}}) + O_p(\sqrt{m.}) = n_{CL}(\mathfrak{X}_{\mathbf{M}}) + O_p(\sqrt{m.}), \quad (24)$$

and $n_{CL}(\mathfrak{X}_{\mathbf{M}}) = m./\mu(d) + o_p(\sqrt{m.})$, $\mathbf{m} \rightrightarrows \infty$. Here, $\mu(d)$ is the mean number of points in a randomly chosen NN-cluster generated by the homogeneous Poisson point process $\mathfrak{P}(1)$ in $\mathbb{R}^d$. Hence, the c.d.f. of deviations

$$\sqrt{n_{CL}(r)}(\hat{p}_{rs}^{CV}(\mathbb{T}_{\mathbf{M}}) - p_{rs}(\mathbb{T}_{\mathbf{M}})) \tag{25}$$

approach in probability with the c.d.f. of deviations given in (23) as $\mathbf{m} \rightrightarrows \infty$.

Let $\tilde{M}_{rsh}$ be the random number of points in $\mathcal{C}_h(\mathfrak{X}_{\mathbf{M}})$ with mark r which have been recognised as belonging to class s, $r, s \in \mathcal{K}_0$ by the $1NN$-classifiers with the training sets $\mathfrak{X}_{\mathbf{M}} \smallsetminus \mathbf{X}_{ri}$. If $\mathbb{T}_{\mathbf{M}} = \mathbb{T}_{\mathbf{m}}$ and $\mathfrak{X}_{\mathbf{M}} = \mathfrak{X}_{\mathbf{m}}$ then $\tilde{M}_{rsh} = M_{rsh}$. In order to simplify the notations below we drop tilda over $M_{rsh}, M_{r\cdot h}, ...$, and instead of $\tilde{M}_{rsh}, \tilde{M}_{r\cdot h}, ...$ we will write $M_{rsh}, M_{r\cdot h}, ...$. Then as in (21) we can write

$$\hat{p}_{rs}^{CV}(\mathbb{T}_{\mathbf{M}}) = \frac{\sum_{h=1}^{n_{CL}(r)} M_{rsh}}{\sum_{h=1}^{n_{CL}(r)} M_{r\cdot h}}, \tag{26}$$

where $M_{r\cdot h} = \sum_{s=1}^{k_0} M_{rsh}$, $n_{CL}(r) = n_{CL}(r, \mathfrak{X}_{\mathbf{M}})$. Recall that the number of points in each NN-cluster is not less than 2.

The r.v.s M_{rsh}, $h = 1, ..., n_{CL}(r, \mathfrak{X}_{\mathbf{M}})$ can be considered as independent r.v.s. It is easy to prove it in the case $\mathfrak{X} \subseteq \mathbb{R}^1$. We consider four sequentially placed training points in $\mathfrak{X}_{\mathbf{M}}$ without marks $X_{i+1} < X_{i+2} < X_{i+3} < X_{i+4}$. If $X_{i+2} - X_{i+1} < X_{i+3} - X_{i+2}$ and $X_{i+4} - X_{i+3} < X_{i+3} - X_{i+2}$ then $\{X_{i+1}, X_{i+2}\}$ and $\{X_{i+3}, X_{i+4}\}$ belong to different NN-clusters. From the absence of dependency in the Poisson point processes the point X_{i+4} will be a point of regeneration and positions of all points X_{i+4+j}, $j \geq 1$, will be independent of the "past" before X_{i+4}. It follows that the number of points in the NN-cluster started from X_{i+3} is independent from contents all NN-clusters in the "past". Note that the linear sizes of NN-clusters can be dependent. From assumptions **A1** and **A2** we conclude that all r.v.s $M_{rs}(\mathcal{C}_h(\mathfrak{X}_{\mathbf{M}}))$, are independent.

In the general case let $\mathfrak{X}_{\mathbf{M}} \subseteq \mathbb{R}^d$, $d > 1$, and let $\mathcal{C}_h$ be any NN-cluster. Together with the points in $\mathcal{C}_h$ we consider all pairs of connected points which are not in $\mathcal{C}_h$ and one point in each such pair is the nearest to a point in $\mathcal{C}_h$. Let set $\mathcal{S}_h$ be the union of all points in $\mathcal{C}_h$ and all such pairs of connected points. From the properties of the Poisson point processes it follows that all events defined on $\mathfrak{X}_{\mathbf{M}} \smallsetminus \mathcal{S}_h$ and $\mathcal{C}_h$ are independent. From assumptions **A1** and **A2** and the property that all NN-clusters contain not less than two points it follows then that all events related to marks on $\mathcal{C}_h$ and on $\mathfrak{X}_{\mathbf{M}} \smallsetminus \mathcal{C}_h$ are also independent. Therefore, the r.v.s M_{rsh}, $h = 1, ..., n_{CL}(r) = n_{CL}(r, \mathfrak{X}_{\mathbf{M}})$, are independent.

Let $\mu_{rs}(\mathbf{m}) = \mathrm{E}[M_{rsh}]$ be the mean value of the number of points, in the randomly placed NN-cluster $\mathcal{C}_h = \mathcal{C}_h(\mathbf{X}_h^1)$, with mark r which has been classified by the $1NN$-classifier as having mark s. We will consider the centered r.v.s $M^0_{rsh} = M_{rsh} - \mu_{rs}(\mathbf{m})$, $M^0_{r\cdot h} = \sum_{s=1}^{k_0} M_{rsh} - \mu_{r\cdot}(\mathbf{m})$, $\mu_{r\cdot}(\mathbf{m}) = \sum_{s=1}^{k_0} \mu_{rs}(\mathbf{m})$, and the following sums $M^0_{rs\cdot} = \sum_{h=1}^{n_{CL}} M^0_{rsh}$, $M^0_{r\cdot\cdot} = \sum_{s=1}^{k_0} M^0_{rs\cdot}$. We can rewrite (26) as follows

$$\hat{p}^{CV}_{rs}(\mathbb{T}_{\mathbf{M}}) = \frac{\mu_{rs}(\mathbf{m}) + (1/n_{CL}(r))M^0_{rs\cdot}}{\mu_{r\cdot}(\mathbf{m}) + (1/n_{CL}(r))M^0_{r\cdot\cdot}}. \tag{27}$$

All random root points $\mathbf{X}^1_1, ..., \mathbf{X}^1_{n_{CL}}$ are considered as a random subset generated by the Poisson marked process $\mathfrak{P}_{\mathbf{M}}(\lambda(\mathbf{m}))$. The numbers of marks M_{rsh} and $M_{r\cdot h}$ in $\mathcal{C}_h(\mathbf{X}^1_h)$, $h = 1, .., n_{CL}$ are conditionally independent r.v.s given their root points $\mathbf{X}'_h$. The distributions of these numbers depend on the positions of the root points. The distributions of the numbers $M_{\cdot\cdot h}$ of all points in $\mathcal{C}_h(\mathbf{X}^1_h)$, $M_{\cdot\cdot h} = \sum_{r=1}^{k_0}\sum_{s=1}^{k_0} M_{rsh}$, are asymptotically the same as $\mathbf{m} \rightrightarrows \infty$.

These properties give us heuristical suggestion to investigate asymptotic properties of the considered complex probability model of the training data $\mathbb{T}_{\mathbf{M}}$ by applying some techniques from the renewal theory. We will consider the additive process $M_{\cdot\cdot 1} + \cdots + M_{\cdot\cdot h}$ with independent r.v.s $\{M_{\cdot\cdot h}\}_{h\geq 1}$ and the time parameter h. Then n_{CL} is the stopping time when $M_{\cdot\cdot 1} + \cdots + M_{\cdot\cdot n_{CL}}$ is the number of all points in $\mathfrak{X}_{\mathbf{M}}$. The asymptotic distribution of each $M_{\cdot\cdot h}$ will be the same as if the NN-cluster $\mathcal{C}_h(\mathbf{X}^1_h)$ was generated by the Poisson point process $\mathfrak{P}(1)$ on $\mathbb{R}^d$. Then it follows that the variances $\mathrm{E}[(M^0_{rs}(\mathcal{C}_h))^2]$ are uniformly bounded because $M_{rsh} \leq M_{\cdot\cdot h}$. We can also use independency of M_{rsh_1}, M_{rsh_2} and $\mathrm{I}(n_{CL} > h_2), h_1 < h_2$. Then we can obtain that

$$\bar{M}^0_{rs\cdot} = \frac{1}{n_{CL}(r)} M^0_{rs\cdot} = O_p(m_r^{-\frac{1}{2}}) \quad \text{and} \quad \bar{M}^0_{r\cdot\cdot} = \frac{1}{n_{CL}(r)} M^0_{r\cdot\cdot} = O_p(m_r^{-\frac{1}{2}}). \tag{28}$$

From (27) and (28) it follows that $\hat{p}^{CV}_{rs}(\mathbb{T}_{\mathbf{M}})$ is a consistent estimator for $p^{CV}_{rs}(\mathbb{T}_{\mathbf{M}})$,

$$\hat{p}^{CV}_{rs}(\mathbb{T}_{\mathbf{M}}) - p^{CV}_{rs}(\mathbb{T}_{\mathbf{M}}) \xrightarrow{p} 0, \quad \mathbf{m} \rightrightarrows \infty. \tag{29}$$

From (27) we have

$$\begin{aligned} & \mathrm{E}\left[\hat{p}^{CV}_{rs}(\mathbb{T}_{\mathbf{M}}) - \frac{\mu_{rs}(\mathbf{m})}{\mu_{r\cdot}(\mathbf{m})}\right] \\ = \quad & \mathrm{E}\left[\frac{\mu_{r\cdot}(\mathbf{m})\bar{M}^0_{rs\cdot} - \mu_{rs}(\mathbf{m})\bar{M}^0_{r\cdot\cdot}}{\mu_{r\cdot}(\mathbf{m})(\mu_{r\cdot}(\mathbf{m}) + \bar{M}^0_{r\cdot\cdot})} - \frac{\mu_{r\cdot}(\mathbf{m})\bar{M}^0_{rs\cdot} - \mu_{rs}(\mathbf{m})\bar{M}^0_{r\cdot\cdot}}{\mu^2_{r\cdot}(\mathbf{m})}\right] \\ = \quad & \mathrm{E}\left[\frac{\bar{M}^0_{r\cdot\cdot}(\mu_{rs}(\mathbf{m})\bar{M}^0_{r\cdot\cdot} - \mu_{r\cdot}(\mathbf{m})\bar{M}^0_{rs\cdot})}{\mu^2_{r\cdot}(\mathbf{m})(\mu_{r\cdot}(\mathbf{m}) + \bar{M}^0_{r\cdot\cdot})}\right]. \end{aligned} \tag{30}$$

By applying the Schwartz inequality to (30) we obtain

$$\begin{aligned}
&\left|\mathrm{E}\left[\hat{p}_{rs}^{CV}(\mathbb{T}_{\mathbf{M}}) - \frac{\mu_{rs}(\mathbf{m})}{\mu_{r\cdot}(\mathbf{m})}\right]\right| \\
\leq\ & \left(\mathrm{E}\left[\frac{(\bar{M}^0_{r\cdot\cdot})^2}{\mu^2_{r\cdot}(\mathbf{m})}\right]\mathrm{E}\left[\left(\frac{\mu_{r\cdot}(\mathbf{m})\bar{M}^0_{rs\cdot} - \mu_{rs}(\mathbf{m})\bar{M}^0_{r\cdot\cdot}}{\mu_{r\cdot}(\mathbf{m})(\mu_{r\cdot}(\mathbf{m}) + \bar{M}^0_{r\cdot\cdot})}\right)^2\right]\right)^{1/2}. \qquad (31)
\end{aligned}$$

Note that

$$\left|\frac{\mu_{r\cdot}(\mathbf{m})\bar{M}^0_{rs\cdot} - \mu_{rs}(\mathbf{m})\bar{M}^0_{r\cdot\cdot}}{\mu_{r\cdot}(\mathbf{m})(\mu_{r\cdot}(\mathbf{m}) + \bar{M}^0_{r\cdot\cdot})}\right| \leq 1$$

and it tends to 0 in probability as $\mathbf{m} \rightrightarrows \infty$. Hence, the expectation of second factor in (31) also tends to zero by the Lebesque dominated convergence theorem. By applying the mentioned above technique of the renewal theory we can obtain that $\mathrm{E}[(\bar{M}^0_{r\cdot\cdot})^2] = O(m_r^{-1})$ as $\mathbf{m} \rightrightarrows \infty$. Hence, we have that

$$\sqrt{m_r}\,\mathrm{E}\left[\hat{p}_{rs}^{CV}(\mathbb{T}_{\mathbf{M}}) - \frac{\mu_{rs}(\mathbf{m})}{\mu_{r\cdot}(\mathbf{m})}\right] \to 0, \quad \mathbf{m} \rightrightarrows \infty. \qquad (32)$$

From (28) we obtain the following asymptotic expansion of (27)

$$\hat{p}_{rs}^{CV}(\mathbb{T}_{\mathbf{M}}) = \frac{\mu_{rs}(\mathbf{m})}{\mu_{r\cdot}(\mathbf{m})} + \frac{1}{\mu_{r\cdot}(\mathbf{m})}\left(\bar{M}^0_{rs\cdot} - \bar{M}^0_{r\cdot\cdot}\frac{\mu_{rs}(\mathbf{m})}{\mu_{r\cdot}(\mathbf{m})}\right) + O_p(m_r^{-1}). \qquad (33)$$

We have that $\bar{M}_{r\cdot} = \mu_r(\mathbf{m}) + o_p(1)$ and $n_{CL}(r) = \frac{m_r}{\mu_{r\cdot}(\mathbf{m})}(1 + o_p(1))$, as $\mathbf{m} \rightrightarrows \infty$.

From (18), (24)-(27), (32) and (33) we can write the normed deviations of interest in the following asymptotic form

$$\begin{aligned}
&\sqrt{n_{CL}(r)}\ (\hat{p}_{rs}^{CV}(\mathbb{T}_{\mathbf{M}}) - p_{rs}(\mathbb{T}_{\mathbf{M}})) \\
=\ & \sqrt{n_{CL}(r)}\Bigg(\frac{1}{\mu_{r\cdot}(\mathbf{m})}\left(\bar{M}^0_{rs\cdot} - \bar{M}^0_{r\cdot\cdot}\frac{\mu_{rs}(\mathbf{m})}{\mu_{r\cdot}(\mathbf{m})}\right) \\
& - \left(\mathrm{E}[\hat{p}_{rs}^{CV}(\mathbb{T}_{\mathbf{M}})] - \frac{\mu_{rs}(\mathbf{m})}{\mu_{r\cdot}(\mathbf{m})}\right) + \left(\mathrm{E}[\hat{p}_{rs}^{CV}(\mathbb{T}_{\mathbf{M}})] - p_{rs}(\mathbb{T}_{\mathbf{M}})\right)\Bigg) \\
=\ & \sum_{h=1}^{n_{CL}(r)} \frac{1}{\sqrt{n_{CL}(r)}}\left(\frac{1}{\bar{M}_{r\cdot}}(M^0_{rsh} - M^0_{r\cdot h}\hat{p}_{rs}^{CV}(\mathbb{T}_{\mathbf{M}}))\right) + O_p(m_\cdot^{-1/2}) \\
=\ & \sum_{h=1}^{m_\cdot/\mu(d) - f(m)} \frac{M^0_{rsh} - M^0_{r\cdot h}\frac{\mu_{rs}(\mathbf{m})}{\mu_{r\cdot}(\mathbf{m})}}{\sqrt{m_\cdot/\mu(d)}\ \mu_{r\cdot}(\mathbf{m})} + O_p(m_\cdot^{-1/2}), \qquad (34)
\end{aligned}$$

where $f(\cdot)$ is a function, $f(m)/\sqrt{m.} \to \infty$ and $f(m.)/m. \to 0$, $\mathbf{m} \rightrightarrows \infty$.

All terms in the last sum are independent and asymptotically uniformly small in probability as $\mathbf{m} \rightrightarrows \infty$. The numbers of points $M_{\cdot\cdot h}$ in $\mathcal{C}_h(\mathbf{X}_h^1)$, $h = 1, ..., n_{CL}$, have uniformly integrable second moments and their distributions can be asymptotically considered as if the r.v.s $M_{\cdot\cdot h}$ correspond to NN-clusters generated by the homogeneous Poisson point process $\mathfrak{P}(1)$. The Lindeberg assumption can be obtained by using the inequalities $(M^0_{rsh})^2 \le M^2_{\cdot\cdot h}$, $(M^0_{r\cdot h})^2 \le M^2_{\cdot\cdot h}$, $h = 1, ..., n_{CL}$. The deviation from the case of triangular array of independent r.v.s, caused by the presence of random root points $\mathbf{X}_h^1$, $h = 1, ..., n_{CL}$, is non-essential and the consistency of resampling method as $\mathbf{m} \rightrightarrows \infty$ follows from a slight extension of Corollary 3 in Belyaev (2003b). These assertions are given here without detailed proofs and they have to be considered by the reader as heuristical arguments which lead us to estimating the distribution of the normed deviations by resampling the terms in the last sum in (34). Note that the last two sums in (34) are approaching in probability as $\mathbf{m} \rightrightarrows \infty$. Hence, we can simulate resampling from the terms $\frac{1}{\sqrt{n_{CL}(r)\bar{M}_{r\cdot}}}\,(M^0_{rsh} - M^0_{r\cdot h}\hat{p}^{CV}_{rs}(\mathbb{T}^0_{\mathbf{m}}))$ in the sum preceding the last sum in (34). Recall that we consider the training set $\mathbb{T}^0_{\mathbf{m}}$ as a value of $\mathbb{T}_{\mathbf{M}}$, i.e. as if the event $\mathbb{T}_{\mathbf{M}} = \mathbb{T}^0_{\mathbf{m}}$ was occurred.

Let $\mathcal{C}^0_1, ..., \mathcal{C}^0_{n^0}$ be the NN-clusters containing training points in $\mathfrak{X}^0_{\mathbf{m}}$, which have mark $r \in \mathcal{K}_0$, $n^0 = n_{CL}(r, \mathfrak{X}^0_{\mathbf{m}})$. The resampling method we want to use can be described as follows. Let m_{rsh} and $m_{r\cdot h}$ be values of r.v.s M_{rsh} and $M_{r\cdot h}$, $h = 1, ..., n^0$, respectively, when $\mathbb{T}_{\mathbf{M}} = \mathbb{T}^0_{\mathbf{m}}$. Let $\{j^{\star b}_{1n^0}, ..., j^{\star b}_{n^0n^0}\}$, $b = 1, ..., B$, be values of B independently simulated realisations of i.i.d. random variables $J^{\star b}_{hn^0}$, $h = 1, ..., n^0$, uniformly distributed on $\{1, ..., n^0\}$. For each b we calculate n^0 values $n^{\star b}_{hn^0}$ of r.v.s $N^{\star b}_{hn^0} = \sum_{i=1}^{n^0} \mathrm{I}(J^{\star b}_{in^0} = h)$. Note that $\sum_{h=1}^{n^0}(n^{\star b}_{hn^0} - 1) = 0$, $\mathrm{E}[N^{\star b}_{hn^0}] = 1$, $\mathrm{E}[(N^{\star b}_{hn^0} - 1)^2] = 1 - 1/n^0$, and $\mathrm{E}[(N^{\star b}_{h_1n^0} - 1)(N^{\star b}_{h_2n^0} - 1)] = -1/n^0$. Then we have

$$\sum_{h=1}^{n^0}(n^{\star b}_{hn^0} - 1)m_{rs\cdot}(\mathbf{m}) = \sum_{h=1}^{n^0}(n^{\star b}_{hn^0} - 1)m_{r\cdot\cdot}(\mathbf{m}) = 0. \tag{35}$$

Let $\mathcal{C}_{j^{\star b}_{1n^0}}, ..., \mathcal{C}_{j^{\star b}_{n^0n^0}}$ be the bth resampled copy of NN-clusters. The related copy of the CV-estimator $p^{CV\star b}_{rs}(\mathbb{T}^0_{\mathbf{m}})$ is defined as follows

$$\hat{p}^{CV\star b}_{rs}(\mathbb{T}^0_{\mathbf{m}}) = \frac{\sum_{h=1}^{n^0} m_{rsj^{\star b}_{hn^0}}}{\sum_{h=1}^{n^0} m_{r\cdot j^{\star b}_{hn^0}}}. \tag{36}$$

From (35) and (36) we have

$$\hat{p}_{rs}^{CV\star b}(\mathbb{T}^0_{\mathbf{m}}) = \frac{\bar{m}_{rs\cdot}(\mathbf{m}) + \bar{m}^{0\star b}_{rs\cdot}}{\bar{m}_{r\cdot\cdot}(\mathbf{m}) + \bar{m}^{0\star b}_{r\cdot\cdot}}, \tag{37}$$

where $\bar{m}_{rs\cdot}(\mathbf{m}) = \frac{1}{n^0}\sum_{h=1}^{n^0} m_{rsh}$, $\bar{m}_{r\cdot\cdot}(\mathbf{m}) = \sum_{s=1}^{k_0} \bar{m}_{rs\cdot}(\mathbf{m})$, and

$$\bar{m}^{0\star b}_{rs\cdot} = \frac{1}{n^0}\sum_{h=1}^{n^0}(n^{\star b}_{hn^0} - 1)m_{rsh} \xrightarrow{p} 0,$$

$$\bar{m}^{0\star b}_{r\cdot\cdot} = \frac{1}{n^0}\sum_{h=1}^{n^0}(n^{\star b}_{hn^0} - 1)m_{r\cdot h} \xrightarrow{p} 0,$$

as $\mathbf{m} \rightrightarrows \infty$. The asymptotic expansion of (37) can be written as follows

$$\begin{aligned}
&\hat{p}_{rs}^{CV\star B}(\mathbb{T}^0_{\mathbf{m}}) \\
&= \frac{\bar{m}_{rs\cdot}(\mathbf{m})}{\bar{m}_{r\cdot\cdot}(\mathbf{m})} + \frac{1}{\bar{m}_{r\cdot\cdot}(\mathbf{m})}\left(\bar{m}^{0\star b}_{rs\cdot} - \bar{m}^{0\star b}_{rs\cdot}\frac{\bar{m}_{rs\cdot}(\mathbf{m})}{\bar{m}_{r\cdot\cdot}(\mathbf{m})}\right) + O_p(m_{\cdot}^{-1}) \\
&= \hat{p}_{rs}^{CV}(\mathbb{T}^0_{\mathbf{m}}) + \frac{1}{n^0}\sum_{h=1}^{n^0}(n^{\star b}_{hn^0} - 1)\frac{1}{\bar{m}_{r\cdot\cdot}(\mathbf{m})}\left(m_{rsh} - m_{r\cdot h}\hat{p}_{rs}^{CV}(\mathbb{T}^0_{\mathbf{m}})\right) \\
&\quad + O_p(m_r^{-1}), \quad \mathbf{m} \rightrightarrows \infty.
\end{aligned} \tag{38}$$

Hence, from (38) we obtain that

$$\begin{aligned}
&\sqrt{n^0}(\hat{p}_{rs}^{CV\star}(\mathbb{T}^0_{\mathbf{m}}) - \hat{p}_{rs}^{CV}(\mathbb{T}^0_{\mathbf{m}})) \\
&= \sqrt{n^0}\sum_{h=1}^{n^0}(n^{\star b}_{hn^0} - 1)\frac{1}{\bar{m}_{r\cdot\cdot}(\mathbf{m})}(m_{rsh} - m_{r\cdot h}\hat{p}_{rs}^{CV}(\mathbb{T}^0_{\mathbf{m}})) \\
&\quad + O_p(m_{\cdot}^{-1/2}), \quad \mathbf{m} \rightrightarrows \infty.
\end{aligned} \tag{39}$$

The sum in (39) corresponds, asymptotically as $\mathbf{m} \rightrightarrows \infty$, to resampling terms from the sums in (34). By the Central Resampling Theorem, Belyaev (2003b, Corollary 3), we obtain that asymptotically the d.f.s of independently simulated deviations

$$\sqrt{n^0}(\hat{p}_{rs}^{CV\star b}(\mathbb{T}^0_{\mathbf{m}}) - \hat{p}_{rs}^{CV}(\mathbb{T}^0_{\mathbf{m}})), \quad b = 1, ..., B, \tag{40}$$

approach in probability the d.f.s of deviations (34) as $\mathbf{m} \rightrightarrows \infty$. This result gives solution of consistent estimation accuracy of CV-estimators of cross-classification probabilities $p_{rs}^{CV}(\mathbb{T}_{\mathbf{m}}^0)$ when $1NN$-classifiers are used. We have used only the original training set in order to discover the deviations which one can meet if other training sets would occur at her/his disposal.

We finish this section with a verbal description of the suggested clustered resampling method. Suppose that it was decided to use a feature space $\mathfrak{X} \subseteq \mathbb{R}^d$ with a distance function $d(\cdot, \cdot)$ and there is a training set $\mathbb{T}_{\mathbf{m}}^0$. Suppose that the $1NN$-classifiers were used in order to obtain the CV-estimators $\hat{p}_{rs}^{CV}(\mathbb{T}_{\mathbf{m}}^0)$ of the CC-probabilities $p_{rs}(\mathbb{T}_{\mathbf{m}}^0)$. We have to find numbers $m_{rsh}, m_{r\cdot h}$, $r, s \in \mathcal{K}_0$ of cross-classified numbers

$$m_{rsh} = \sum_{i:\mathbf{x}_{ri} \in \mathcal{C}_h} \mathrm{I}(f_{1NN}^{\bullet}(\mathbf{x}_{ri}, \mathbb{T}_{\mathbf{m}(r,i)}^0) = s), \quad m_{r\cdot h} = \sum_{s=1}^{k_0} m_{rsh}, \tag{41}$$

where $\mathcal{C}_h, h = 1, ..., n_{CL}(r)$ are the all NN-clusters in $\mathbb{T}_{\mathbf{m}}^0$ which contain points with mark r. Then we simulate $B \gg 1$ resampling copies of CV-estimators (36) and find differences

$$\hat{p}_{rs}^{CV\star b}(\mathbb{T}_{\mathbf{m}}^0) - \hat{p}_{rs}^{CV}(\mathbb{T}_{\mathbf{m}}^0), \quad r, s \in \mathcal{K}_0, \ \ b = 1, ..., B. \tag{42}$$

Note that we can simultaneously calculate values (42) for several pairs $r_j, s_j, j = 1, 2, ..., k_0$. The obtained lists of differences (42) mimic deviations of the CV-estimators from true CC-probabilities if other, independent and identically distributed as $\mathbb{T}_{\mathbf{m}}^0$, training sets would be used. If B pairs of differences

$$\{\hat{p}_{rs_1}^{CV\star b}(\mathbb{T}_{\mathbf{m}}^0) - \hat{p}_{rs_1}^{CV}(\mathbb{T}_{\mathbf{m}}^0), \hat{p}_{rs_2}^{CV\star b}(\mathbb{T}_{\mathbf{m}}^0) - \hat{p}_{rs_2}^{CV}(\mathbb{T}_{\mathbf{m}}^0)\}_{1 \leq b \leq B}$$

were simulated then this list can be used to estimate the correlation between deviations

$$\hat{p}_{rs_1}^{CV}(\mathbb{T}_{\mathbf{m}}) - p_{rs_1}(\mathbb{T}_{\mathbf{m}}) \quad \text{and} \quad \hat{p}_{rs_2}^{CV}(\mathbb{T}_{\mathbf{m}}) - p_{rs_2}(\mathbb{T}_{\mathbf{m}}),$$

where we consider training set $\mathbb{T}_{\mathbf{m}}$ as random. For justification of this fact it is possible to use a multidimensional variant of the Central Limit Resampling Theorem, Belyaev (2004).

The d.f. based on deviations (42) gives us information how typical deviations can be if other identically distributed with $\mathbb{T}_{\mathbf{m}}^0$ training sets would be used. It is worth noting that our result is asymptotic. For a given $\mathbf{m}$ it is possible to have such training sets that give via resampling rather good estimates of accuracy, but it is also possible to meet training sets that give much less accurate estimates of accuracy. The probabilities to meet such "bad" training sets are vanishing as $\mathbf{m} \rightrightarrows \infty$. It is still an open problem to find out how large the numbers $m_1, ..., m_{k_0}$ have to be in order to be sure to have sufficiently accurate estimators of the true d.f.s of deviations CV-estimators from theoretical (unknown) values.

We want to emphasize that in (42) we have only used the original training set $\mathbb{T}^0_{\mathbf{m}}$. We are not able to simulate independent copies of training sets which have the same distributions as the original training set because we do not know the true distributions $\mathrm{P}_r[\,]$, $r \in \mathcal{K}_0$. The suggested resampling from NN-clusters helps to handle with this difficulty. In the next section we consider two numerical examples.

5 Numerical experiments

The asymptotic analysis, developed in Section 4 for $1NN$-classifiers, shows that it is possible to obtain consistent estimates of the accuracy of the CV-estimators $\hat{p}^{CV}_{rs}(\mathbb{T}^0_{\mathbf{m}})$ by using resampling from NN-clusters. Here, we illustrate how the suggested type of resampling works. We use simulated data because then we can see how the estimated typical distribution of deviations of the CV-estimators $\hat{p}^{CV}_{rs}(\mathbb{T}^0_{\mathbf{m}})$ deviates from the true typical distribution of deviations.

In order to avoid large computing time and the complexity of needed programs the one-dimensional Euclidean feature space $\mathfrak{X} = \mathbb{R}^1$ and the set with three classes $\mathcal{K}_0 = \{1, 2, 3\}$ had been considered. Here, we present selected results of two experiments.

In Experiment 1, one original training set $\mathbb{T}^0_{\mathbf{m}_1}$ and a series of independent training sets $\mathbb{T}^s_{\mathbf{m}_1}$, $\mathbf{m}_1 = \{m_{11}, m_{12}, m_{13}\}$, $s = 1, ..., B_1$, $B_1 = 3300$, $s = 1, ..., B_1$, $B_1 = 3300$, were simulated with $m_{11} = 500$, $m_{12} = 400$, and $m_{13} = 300$ which are numbers of simulated real-valued independent normally distributed r.v.s belonging to classes 1, 2 and 3. Their mean values are $\mu_{11} = 1$, $\mu_{12} = 5$, $\mu_{13} = 9$ and variances $\sigma^2_{11} = 3$, $\sigma^2_{12} = 2$, $\sigma^2_{13} = 1$, respectively.

Similarly in Experiment 2, one original training set $\mathbb{T}^0_{\mathbf{m}_2}$ and a series of independent training sets $\mathbb{T}^s_{\mathbf{m}_2}$, with $\mathbf{m}_2 = \{m_{21}, m_{22}, m_{23}\}$, $s = 1, ..., B_2$, $B_2 = 2300$, $m_{21} = 400$, $m_{22} = 600$, and $m_{23} = 300$ were simulated. The parameters of the normal distributions were $\mu_{21} = 1$, $\mu_{22} = 5$, $\mu_{23} = 2.75$, and $\sigma^2_{21} = 0.125$, $\sigma^2_{22} = 9$, $\sigma^2_{23} = 0.25$.

The probability densities of the related normal distributions $\mathcal{N}(\mu_{ij}, \sigma^2_{ij})$, $i = 1, 2$, $j = 1, 2, 3$, are shown in Fig. 10. For each simulated training set $\mathbb{T}^s_{\mathbf{m}_i}$, $i = 1, 2$, we simulated $N = 10000$ values of i.i.d. $\mathcal{N}(\mu_{ij}, \sigma^2_{ij})$-distributed r.v.s,

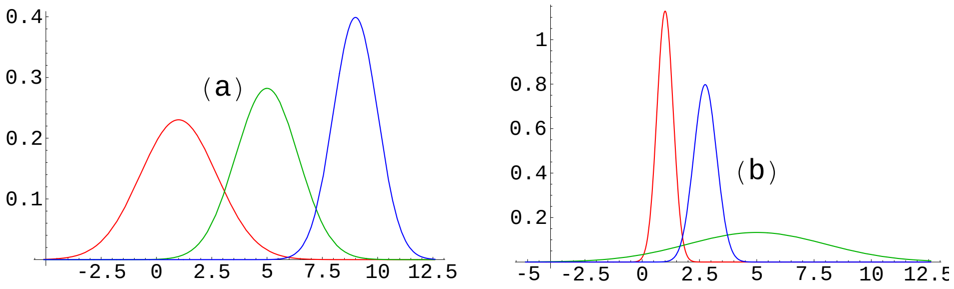

Figure 10: Probability densities of $\mathcal{N}(1,3)$, $\mathcal{N}(5,2)$, $\mathcal{N}(9,1)$ in Experiment 1 (left) and $\mathcal{N}(1,0.125)$, $\mathcal{N}(5,9)$, $\mathcal{N}(2.75,0.25)$ in Experiment 2 (right).

$i=1,2$, $j=1,2,3$. The $1NN$-classifier was applied to each of these values and the frequencies $\hat{p}_{rs}(\mathbb{T}^s_{\mathbf{m}_i})$ of cross-classifications were calculated.

The differences $\hat{p}_{rs}(\mathbb{T}^s_{\mathbf{m}_i}) - p_{rs}(\mathbb{T}^s_{\mathbf{m}_i})$ decrease to zero as N increases to infinity. $N = 10000$ is sufficiently large and we used the frequencies $\hat{p}_{rs}(\mathbb{T}^s_{\mathbf{m}_i})$ as the true values of $p_{rs}(\mathbb{T}^s_{\mathbf{m}_i})$, $r,s \in \mathcal{K}_0$. For each $\mathbb{T}^s_{\mathbf{m}_i}$ also the CV-estimates have been computed. We have $M_1 = 3300$ and $M_2 = 2300$ pairs of the true values and the CV-estimates for the confusion matrices $\mathbb{P}(\mathbb{T}^s_{\mathbf{m}_i}) = (p_{rs}(\mathbb{T}^s_{\mathbf{m}_i}))$ in the two experiments. Two pairs of the true values and the CV-estimates of confusion matrices related to two simulated training set are shown in Table 2.

Table 2. True confusion matrices and their CV-estimators

Experiment 1	
True	CV-estimator
$\begin{pmatrix} 0.8661 & 0.1339 & 0.0000 \\ 0.1650 & 0.7751 & 0.0599 \\ 0.0000 & 0.0721 & 0.9279 \end{pmatrix}$	$\begin{pmatrix} 0.8860 & 0.1140 & 0.0000 \\ 0.1600 & 0.7875 & 0.0525 \\ 0.0000 & 0.0667 & 0.9333 \end{pmatrix}$
$\begin{pmatrix} 0.8536 & 0.1461 & 0.0003 \\ 0.1678 & 0.7746 & 0.0576 \\ 0.0003 & 0.0923 & 0.9074 \end{pmatrix}$	$\begin{pmatrix} 0.8640 & 0.1360 & 0.0000 \\ 0.1750 & 0.7475 & 0.0775 \\ 0.0033 & 0.1033 & 0.8933 \end{pmatrix}$
Experiment 2	
True	CV-estimator
$\begin{pmatrix} 0.8720 & 0.1196 & 0.0084 \\ 0.0830 & 0.7756 & 0.1414 \\ 0.0233 & 0.2896 & 0.6871 \end{pmatrix}$	$\begin{pmatrix} 0.8650 & 0.1275 & 0.0075 \\ 0.0783 & 0.7817 & 0.1400 \\ 0.0067 & 0.3000 & 0.6933 \end{pmatrix}$
$\begin{pmatrix} 0.8655 & 0.1167 & 0.0178 \\ 0.0816 & 0.7761 & 0.1423 \\ 0.0247 & 0.3078 & 0.6675 \end{pmatrix}$	$\begin{pmatrix} 0.8475 & 0.1325 & 0.0200 \\ 0.0917 & 0.7683 & 0.1400 \\ 0.0333 & 0.2500 & 0.7167 \end{pmatrix}$

We observe that the true values and the CV-estimators of confusion matrices are essentially depending on the related training sets $\mathbb{T}^s_{\mathbf{m}_i}$. We compute the lists of deviations $\{\hat{p}^{CV}_{rs}(\mathbb{T}^s_{\mathbf{m}_i}) - \hat{p}_{rs}(\mathbb{T}^s_{\mathbf{m}_i})\}_{1\leq s\leq B_i}$, $i = 1, 2$, and the list of deviations $\{\hat{p}^{CV\star b}_{rs}(\mathbb{T}^0_{\mathbf{m}_i}) - \hat{p}^{CV}_{rs}(\mathbb{T}^0_{\mathbf{m}_i})\}_{1\leq s\leq B_i}$ by using the suggested resampling from the list of NN-clusters $\mathcal{C}^0_1, ..., \mathcal{C}^0_{n^0_{CL}(r)}$ related to the original training set, $\mathbb{T}^0_{\mathbf{m}_i}$, $i = 1, 2$. Besides that we have computed the list of deviations $\{\hat{p}^{CV*b}_{rs}(\mathbb{T}^0_{\mathbf{m}_i}) - \hat{p}^{CV}_{rs}(\mathbb{T}^0_{\mathbf{m}_i})\}_{1\leq s\leq B_i}$, $i = 1, 2$, $B_1 = B_2 = 3000$, where $\hat{p}^{CV*b}_{rs}(\mathbb{T}^0_{\mathbf{m}_i})$ is obtained as the bth copy of resampled terms in (7), i.e.

$$\hat{p}^{CV*b}_{rs}(\mathbb{T}^0_{\mathbf{m}_i}) - \hat{p}^{CV}_{rs}(\mathbb{T}^0_{\mathbf{m}_i}) = \frac{1}{m_r}\sum_{h=1}^{m_r}(n^\star_{hm_r} - 1)\mathrm{I}(f^\bullet_{1NN}(\mathbf{x}_{rh}, \mathbb{T}^0_{\mathbf{m}_i(r,h)}) = s). \quad (43)$$

Then the following three d.f.s were found

$$F^i_{Trs}(z) = \frac{1}{M_i}\sum_{s=1}^{M_i}\mathrm{I}(\hat{p}^{CV}_{rs}(\mathbb{T}^s_{\mathbf{m}_i}) - p_{rs}(\mathbb{T}^s_{\mathbf{m}_i}) \leq x), \quad (44)$$

$$\hat{F}^i_{Crs}(z, \mathbb{T}^0_{\mathbf{m}_i}) = \frac{1}{B_i}\sum_{b=1}^{B_i}\mathrm{I}(\hat{p}^{CV\star b}_{rs}(\mathbb{T}^0_{\mathbf{m}_i}) - p^{CV}_{rs}(\mathbb{T}^0_{\mathbf{m}_i}) \leq z), \quad (45)$$

$$\hat{F}^i_{Srs}(z, \mathbb{T}^0_{\mathbf{m}_i}) = \frac{1}{B_i}\sum_{b=1}^{B_i}\mathrm{I}(\hat{p}^{CV*b}_{rs}(\mathbb{T}^0_{\mathbf{m}_i}) - p^{CV}_{rs}(\mathbb{T}^0_{\mathbf{m}_i}) \leq z). \quad (46)$$

Fig. 11 illustrates results of Experiment 1. These three d.f.s, corresponding to a simulated original training set $\mathbb{T}^0_{\mathbf{m}_1}$, are shown in Fig. 11. The line of $\hat{F}^1_{Crs}(z)$ related to resamplings from NN-clusters is tagged by arrows "$\leftarrow CR$", and "$CR \rightarrow$ ". The line of $\hat{F}^1_{Srs}(z)$ related to resamplings from single training points is tagged by arrows "$\leftarrow SR$", and "$SR \rightarrow$ ". The d.f. of interest $F^1_{Trs}(z)$ is shown by a line without tags. We see that the resampling from NN-clusters in this example gives a very good estimate of $F^1_{Trs}(z)$.

Resampling from the list of all terms in (7) gives the d.f. $\hat{F}^1_{Srs}(z, \mathbb{T}^0_{\mathbf{m}_i})$ which has worse accuracy, i.e. it underestimates the probability to have larger deviation of the CV-estimator from the true cross-classification probability. One can say that $\hat{F}^1_{S_{rs}}(z, \mathbb{T}^0_{\mathbf{m}_i})$ is "over-optimistic". Our extensive simulations show that resampling from the list of all terms in (7) in the most of cases gives the "over-optimistic" estimates (46) for the distributions of true deviations (44).

It is necessary to know that the accuracy of the approximation essentially depends on the original training set. If the numbers $m_1, ..., m_{k_0}$ are not sufficiently large then the estimators (45) and especially (46) can essentially deviate

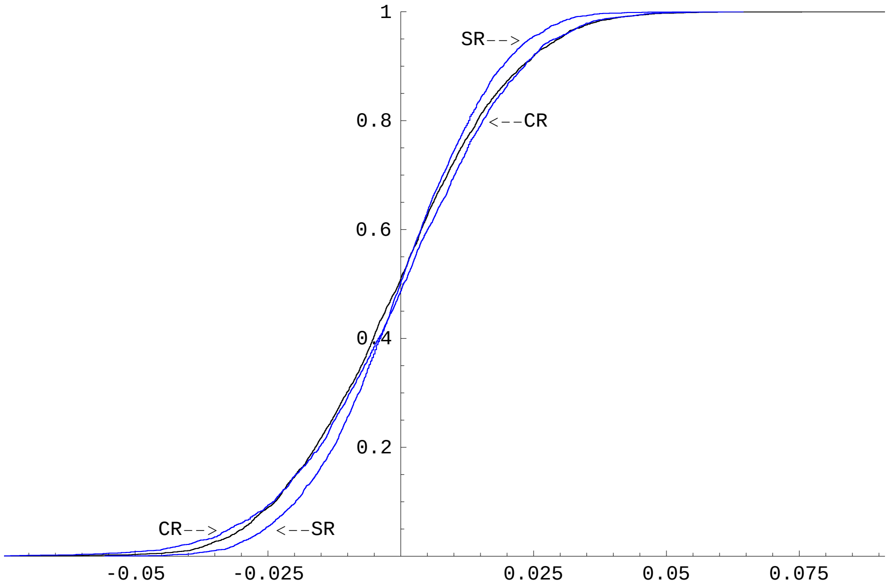

Figure 11: True distribution functions of $\hat{p}_{rs}^{CV}(\mathbb{T}_{\mathbf{m}_1})-p_{11}(\mathbb{T}_{\mathbf{m}})$ and its estimates $\hat{p}_{11}^{\star CV}(\mathbb{T}_{\mathbf{m}_i})-\hat{p}_{11}^{CV}(\mathbb{T}_{\mathbf{m}_i})$, and $\hat{p}_{11}^{*CV}(\mathbb{T}_{\mathbf{m}_i})-\hat{p}_{11}^{CV}(\mathbb{T}_{\mathbf{m}_i})$, obtained in Experiment 1, by resampling from the related sets of $1NN$-clusters (shown by "$\leftarrow CR$", "$CR \rightarrow$") and by resampling from terms in (7) (shown by "$\leftarrow SR$", "$SR \rightarrow$").

from the true distribution of interest (44). Fig. 12 and 13 illustrate that the accuracy of the estimators (45) and (46) essentially depends on the training sets. Most of the estimates (46) are over-optimistic.

In Fig. 13 (Experiment 2) you can see that both estimates (45) and (46) have a stepwise shape on the left tails and they are not accurate. This character of estimates caused by the presence of too small number of those training points in class 3 which have been classified as belonging to class 1. There were only a few such points in Experiment 2.

We conclude that the results of extensive simulations correspond the theory considered in Section 4. The resampling from NN-clusters can be used if there are many cross-classified training points in all classes. The question of how large the numbers $m_1, ..., m_{k_0}$ of training points should be has to be considered as a separate complex problem. We can only say that several hundreds of training points in each class usually can give satisfactory estimators to the true distributions of deviations $\hat{p}_{rs}^{CV}(\mathbb{T}_{\mathbf{m}}^0) - p_{rs}(\mathbb{T}_{\mathbf{m}}^0)$. We also stress that the usage of a thoroughly prepared training set is an essential and necessary part for obtaining objective results.

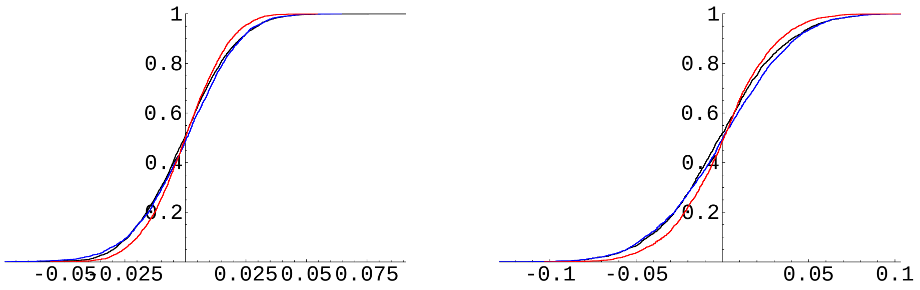

Figure 12: True distribution functions of $\hat{p}_{11}^{CV}(\mathbb{T}_{\mathbf{m}_i})-p_{11}(\mathbb{T}_{\mathbf{m}_i})$ and $\hat{p}_{33}^{CV}(\mathbb{T}_{\mathbf{m}_i})-p_{33}(\mathbb{T}_{\mathbf{m}_i})$ in Experiment 1 ($i=1$, left) and in Experiment 2 ($i=2$, right), respectively, and their estimates obtained by resamplings from the related sets of NN-clusters and "over-optimistic" estimates obtained by resampling from single terms (43) (gray lines).

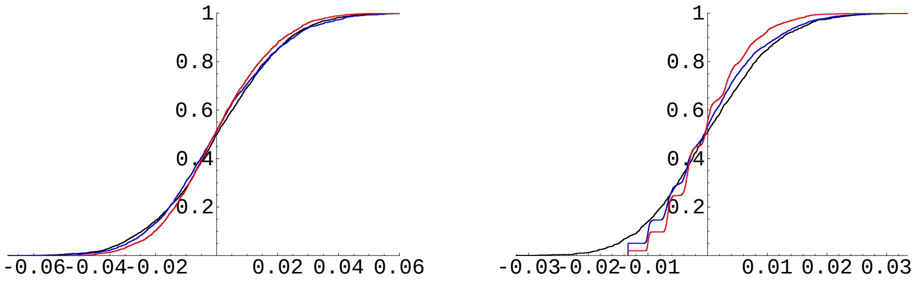

Figure 13: True distribution functions of $\hat{p}_{32}^{CV}(\mathbb{T}_{\mathbf{m}_i})-p_{32}(\mathbb{T}_{\mathbf{m}_i})$ and $\hat{p}_{31}^{CV}(\mathbb{T}_{\mathbf{m}_i})-p_{31}(\mathbb{T}_{\mathbf{m}_i})$ in Experiment 1 ($i=1$, left) and in Experiment 2 ($i=2$, right), respectively, and their estimates obtained by resamplings from the related sets of NN-clusters and from single terms in (43).

All illustrating calculations and figures have been obtained by using a series of programs developed by the author of this paper. The programs are based on the language of *Mathematica* 5.1.

6 Discussion

The NN-classifiers can be used in many applications. The suggested methods of consistent estimation of the classifiers' accuracy show the crucial importance of training sets. In the case with remotely sensed data the process of creating training sets is very labor-intensive if satellite sensors have a high resolution, e.g. if the pixel size is $1m \times 1m$. In this case it is necessary to have the additional information on the shape of individual tree crowns in the plots and the more exact coordinates of their positions have to be known as well. The creation of training sets can be realized only if some appropriate computer-intensive methods will be developed. In the case when the remotely sensed data obtained by high resolution sensors it is necessary to investigate the influence of fractal properties of crown boundaries and the porosity of crowns when the reflection from ground grass will be an essential part in the registered energies of the light reflected from tree crowns.

Application of NN-classifiers in the analysis of biomedical raster data in non-invasive medicine is potentially useful. Such type of data are obtained and transformed into digital images during the analysis of patients by various methods of tomography, Korn et al. (1996). Sophisticated sensors in the optical tomography (OCT) are developed to register the reflected lazer light during the observation of the eye's retina, Jaffe and Caprioli (2004). The obtained digital images are essential for correct diagnosis. Methods of creating training sets in the biomedical applications have to be developed. The creation of training sets in the biomedical application substantially differs from the approach considered in Section 2 due to the absence of data similar to the field data. Thoroughly collected biomedical data bases and anatomical knowledge can be used to create training sets. If elaborated training sets will be developed then the NN-classifiers can be efficiently used to create digital images which will obviously show different types of tissue, e.g. the different types of tumors, and it will help to make more precise diagnoses.

Acknowledgements

Importance of developing classification methods in analysis of remotely sensed data has become obvious to author during participation in the Swedish research program "Remotely Sensing for Environment" (RESE). This work was supported by the Swedish National Space Board (Dnr. 10403) and by the Grant of the Swedish Research Council (No. 621-2002-3795)

The author is thankful to Prof. B. Ranneby, Drs. M. Ekström, S. Sjöstedt-de Luna, M. Nilsson and J. Yu for discussion problems and useful comments related to the nearest neighbor classifiers.

References

Belyaev, Yu.K. (1996) Central limit resampling theorems for m-dependent heterogeneous random variables. *Research Report 1996-5, Department of Mathematical Statistics, Umeå University, Sweden.*

Belyaev, Yu.K. (2000) On the accuracy of discretely colored maps created by classifying remotely sensed data. *Research Report 2000-10, Department of Forest Resource Management and Geomatics, Swedish University of Agricultural Sciences, Sweden.*

Belyaev, Yu.K. (2003a) On the accuracy of classifiers and corresponding digital discretely colored images.*Theory Probability and Math. Statist., American Mathematical Society*, **66**, 15-25.

Belyaev, Yu.K. (2003b) Necessary and sufficient conditions for consistency of resampling. *Research Report 2003-1, Centre of Biostochastics, Swedish University of Agricultural Sciences, Sweden.*

Belyaev, Yu.K. (2004) Application of resampling methods to linear heteroscedastic regression with vector responses. *Research Report 2004-2, Department of Mathematical Statistics, Umeå University, Sweden.*

Belyaev, Yu.K. and Sjöstedt-de Luna, S. (2000) Weakly approaching sequences of random distributions. *Journal of Applied Probability,* **37**, No 3, 807-822.

Daley, D.J. and Vere-Jones (1988) *An Introduction to The Theory of Point Process.* Springer-Verlag, N.Y.

Davison, A.C. and Hinkley, D.V. (1997) *Bootstrap Methods and Their Applications.* Cambridge University Press.

Efron, B. and Tibshirani, R.J. (1993) *An Introduction to the Bootstrap.* Chapman and Hall, New York.

Efron, B. and Tibshirani, R.J. (1997) Improvement on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, **92,** 548-560.

Ekström, M. and Belyaev, Yu.K. (2001) On the estimation of the distribution of sample means based on non-stationary spacial data. *Working Paper 2001:89, Department of Forest Resource Management and Geomatics, Swedish University of Agricultural Sciences, Sweden.*

Ekström, M. and Sjöstedt-de Luna, S. (2004) Subsampling methods to estimate the variance of sample means based on nonstationary spatial data with varying expected values. *Journal of the American Statistical Association.* **99**, No 465, 82-95.

Hall, P. (1985) Resampling a coverage pattern. *Stochast. Process. and Appl.*, **20**, 231-246.

Holmström, H. and Fransson, J.E.S. (2003) Combining remotely sensed optical and radar data in kNN-estimation of forest variables. *Forest Science.* **49(3)**, 409-418.

Holmström, H., Nilsson, M. and Ståhl, G. (2001) Simultaneous estimations of forest parameters using aerial photograph interpreted data and the k nearest neighbor method. *Scandinavian Journal of Forest Research.* **16**, 67-78.

Holst, M. and Arle, A. (2001) Nearest neighbor classification with dependent training sequences. *The Ann. of Math. Statist.* **29**, No 5, 1424-1442.

Jaffe, G.J. and Caprioli, J. (2004) Optical coherence tomography to detect and manage retinal disease and glaucoma. *American Journal of Ophthalmology,* **137,** No. 1, 156-169.

Korn, F., Sidiropoulos, N., Falontos, C., Siegel, E. and Protopapas, M.D. (1996) Fast nearest neighbor search in medical image databases. *Proceedings of the 22nd Very Large Data Bases Conference, Bombay, India,* 1-12.

Niblack, W. (1986) *An Introduction to Digital Image Processing.* Prentice-Hall International, London.

Okabe, A., Boots, B.N. and Suqihara, K. (1992) *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams.* Wiley, New York.

Steele, B.M. and Patterson, D.A. (2000) Ideal bootstrap estimation of expected prediction error for k-nearest neighbor classifiers i application for classification and error assessment. *Statistical and Computing*, **10**, 349-355.