AUSTRIAN JOURNAL OF STATISTICS
Volume 42 (2013), Number 2, ??–??

Statistical Indicators for the Analysis of Digitalized Brain Tumor Images

Matthias Templ[1,2], Semaguel Aklan[1], Peter Filzmoser[1],
Matthias Preusser[3], Johannes A. Hainfellner[3]

[1]Vienna University of Technology, Austria
[2] Statistics Austria
[3] Department of Medicine I, Medical University of Vienna

Abstract: In this contribution, indicators for computer-based analysis and assessment of tumor cell proliferation in human brain tumors are developed. The methods are applied on (two) samples of digitized human brain tumor tissue sections immunostained with an antibody against the Ki67 epitope. The Ki67 immunostaining highlights cells undergoing cell division and is thus a surrogate marker for tumor growth.

The challenges are related to the enormous size of the images ("*big data*") analyzed, some of them are larger than 100 GB. Thus, efficient methods to extract relevant information have to be applied.

Before starting with the statistical analysis, the digitized images are pre-processed to extractthe highlighted cells. Then the distribution of Ki67 immunostaining patterns is analyzed. Starting with a bivariate kernel density estimation, the proposed indicators are used to evaluate and compare the resulting density estimates. Moreover, the spatial distribution of clusters of Ki67-labeled tumor cells is of particular interest.

The results allow to evaluate and compare images or sectors of the images. This evaluation and comparisons of samples or sectors could turn out to be useful in practice since it allows for a pre-selection of interesting sectors and samples. Thus, the time-consuming part of manual inspection of the huge images could be reduced.

Zusammenfassung: Dieser Beitrag befasst sich mit der Entwicklung von Indikatoren aufbauend auf computer-basierten Analysemethoden und der Evaluierung der Verteilung von Tumorzellen. Anhand von (zwei) Proben von digitalisierten Tumorgeweben (immungefärbt mit einem Antikörper gegen das Ki-67-Epitop) des menschlichen Gehirns. Die Ki67 Immunfärbung markiert Zellen und deren Zellteilung und ist damit ein Ersatz-Marker für das Tumorwachstum.

Die Herausforderungen bestehen vor allem in der enormen Größe der Bilder ("*Big Data*"), einige davon sind größer als 100 GB. Um relevante Information aus diesen Daten zu gewinnen, müssen sehr effiziente Methoden entwickelt werden.

Zuerst werden die digitalisierten Bilder – vor der statistischen Analyse – bearbeitet. Danach wird die Verteilung der Ki67 Immunfärbung analysiert. Be-

ginnend mit einer bivariaten Kerndichteschätzung werden die vorgeschlagenen Indikatoren auf die resultierenden Dichteschätzungen angewandt. Darüber hinaus ist die räumliche Verteilung von Clustern von Ki67 markierten Tumorzellen von besonderem Interesse.

Bilder oder Sektoren der Bilder können somit verglichen werden. Diese Auswertung und Vergleiche von Proben oder Sektoren könnten sich zukuenftig in der Praxis als nützlich erweisen, da sie eine Vorauswahl der interessantesten Sektoren und Proben ermöglichen. Der manuele Aufwand zur Inspektion der großen Datensätze könnte damit verringert werden.

Keywords: Brain Tumor Image Analysis, Density Estimation, Indicators.

1 Introduction

In Austria the second most common cause of death is cancer. The chance to get a brain tumor is about 0.5 % for men and about 0.4 % for women (Statistik Austria, 2011).

For the histomorphological analysis the typing and grading of tumors play an important role. The cell proliferation is used as an index for the growth behavior of tumors. It is an important factor for the estimation of the projected survival time of a patient, particularly in intracranial ependymomas, a certain subtype of primary brain tumors occurring mostly with children and young adults. As yet the determination of the proliferation index is usually performed manually. Moreover, images on brain tumors are evaluated by extensive visual analysis of images of about 10-150 GB size with distance between pixel of 40 nanometer. To see special structures, a zoom of about 40 is necessary, which makes an analysis of all areas of a sample impossible. The aim is to provide indicators to first highlight interesting images or regions in images for further visual analysis.

The goal of this contribution is to investigate the usefulness of indicators for the computer-based analysis and assessment of tumor cell proliferation in human brain tumors. The aim is to define indicators that ensure an objective assessment and highlights interesting areas in images for further manual visual inspection. Exemplary, two samples of human brain tumors are used in this study and results from the smaller sample are presented.

The paper is structured in the following manner: Section 2 gives detailed information about the material and data. For the reason of readability (especially for readers with minor background in neurology), the problem is shown in detail. The proposed indicators are shown in Section 3. The indicators are then practically applied on one sample and the results are presented and compared in Section 4. Section 5 concludes and gives an outlook to further developments.

2 Material

For illustration, Figure 1 shows an infratentroial ependymoma in a pediatric case, detected by magnetic resonance imaging (source: CERN, 2011).

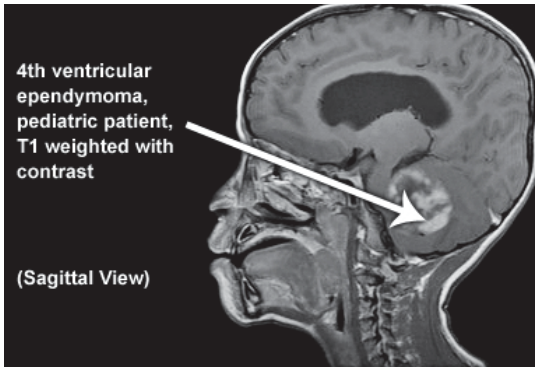

Figure 1: MRI-sagittal view: infratentorial ependymoma.

After neurosurgical resection, such an infratentorial ependymoma is specially for histological evaluation. More precisely, after fixing the resulting tumor tissue blocks with formalin and embedding those with paraffin, sections are cut at a thickness of 3-5 μm. Then a heat-induced epitope retrieval in 0.01 M (molar mass) citrate buffer (pH 6.0) is conducted with the slides for 30 minutes in a microwave oven at 600 W. After the incubation of the sections with a monoclonal mouse anti-Ki67 antibody at a dilution of 1:50 for 25 minutes, the detection process of the immunoreactivity using the ChemMate kit (Dako) and diaminobenzidine as chromogen is performed. The Ki67 immunohistochemistry is conducted according to the standard operating procedure of the laboratory of the Institute of Neurology, Medical University of Vienna (Preusser et al., 2008).

After this process the tissue sections are digitized with a digital pathological scanner, a NanoZoomer from the Hamamatsu Corporation (Hamamatsu, 2012). The resulting virtualized slides are available in the NDPI (NanoZoomer Digital Pathology Image) format. From this digitized data the information of interest is filtered. Hereby, the difference in colors for the Ki67 labeled cell nuclei (brown) and unlabeled cell nuclei (blue) is used for filtering. The slides have an average resolution of 100.000×100.000 pixels and thus processing of such an image with a color depth of 24 bit would need a memory consumption of about 30 GB. For some slides this goes beyond 100 GB. To process the data, the slides are divided into 5000×5000 pixel blocks, later on defined as *sectors*.

For illustration of the largeness of the images, a (rather small) image is selected and from this image a 40-times magnified area is shown. Figure 2 shows an original scanned and digitized tumor tissue slide, whereby the grid, which splits the sample, is also added. The brown points represent the cell nuclei of the proliferating cells. The arrow in the bottom of the image shows the location where the zoomed picture (40-times magnified) in Figure 3 is extracted. It is easy to see that it is rather time-consuming to zoom-in in only few areas to evaluate the structure of the tumor and cells.

For further processing, the image is split to essentially 2 gray-value images with a depth of 8 bit, whereas one of the grey images includes most information on brown cell nuclei and one corresponds to blue cell nuclei (for details, see Walser, 2011). By fixing a threshold, the objects are separated whereas this threshold is chosen carefully (Walser, 2011, Otsu, 1979). For both images the thresholds are applied and the result is a binary matrix indicating the Ki67 labeled cell nuclei.

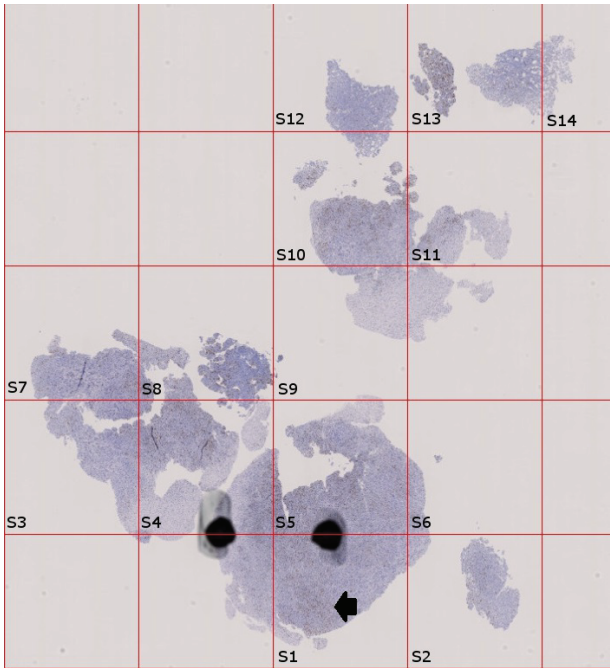

Figure 2: Digitalized image visualized with the NanoZoomer from Hamamatsu. Sectors are already built and used for analysis in Section 4.

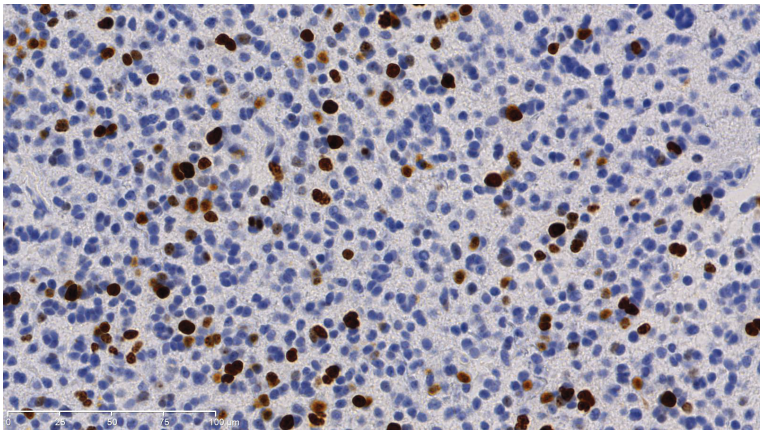

Figure 3: 40-times magnified zoom of a small area from Figure 2.

The materials used in this study consist of two samples of 78 specimens of intracranial ependymomas, which had been collected from a group of scientists from the General Hospital Vienna (AKH) primarily for diagnostic and research purposes including the assessment of the Ki67 index, a histopathological biomarker is used to determine the tumor cell proliferation. The 78 tissues had a size greater than one microscopic field at a magnification of x400 (Preusser et al., 2008).

3 Indicators for the Evaluation of the Images/Tumors

3.1 Conventional Calculation of the Cell Proliferation

The conventional determination of the tumor cell proliferation index is carried out in the following way. The anti-Ki67 immunoreactive tissue section is scanned at a low magnification and the area with the highest density of immunolabelled tumor cell nuclei is determined. This area is also called "hot-spot". Then a total of 500 tumor cell nuclei are evaluated within the hot-spot area and through manual counting. The fraction of the labeled cell nuclei per 500 tumor cell nuclei is calculated and is expressed as a percentage. The counting of 500 cells per case yields good results since it takes two minutes by an experienced person.

However, as already mentioned, the whole process of manual scanning is very time-consuming and therefore not workable.

The aim is therefore to find indicators that help to find regions of interest and to evaluate the images or regions. With that help, neurologists can evaluate the tumor in less time.

In the previous section, the process of receiving a digitized image was described. The aim is now to extract meaningful information out of a digitized image.

3.2 Preliminary Considerations

In several studies (Wählby, Sintorn, Erlandsson, Borgefors, and Bengtsson, 2004) the watershed algorithm (Beucher and Lantuéjoul, 1979, Roerdink and Meijster, 2000) has been the basis to find areas of high density. It consists of placing a *water* source in each regional minimum (area with many pixels) to flood the relief from sources. The rise will stop when each seeded catchment basin in the gradient magnitude image meets another seeded catchment basin. For more information, see Wählby et al. (2004).

Summarize and extract information on a grid:
Another idea is to lay a certain grid over the image and to calculate in each area an average or total or mean of ones (remember, we have already extracted a binary matrix). This now contains a summary of the areas and this information on the grid is of lower resolution. Thus, standard statistical methods may then be suitable to apply.

Bivariate kernel density estimation:
To enhance this idea one can think of a sliding window over the grid, which is already a kind of kernel density estimation. In our contribution we focus on bivariate kernel density estimation using Gaussian kernels with an optimal bandwidth (compare Figure 4). Usually, the optimal bandwidth is given by $1.06\hat{\sigma}n^{1/5}$. We plug in a robust estimator for σ based on the interquartile range, which is well-known to be consistent and unbiased if the data follow a normal distribution. The mathematical basics of kernel density estimation are not touched in this contribution, but we refer interested readers to Härdle, Müller, Sperlich, and Werwatz (2004) or Scott (1992) for details on that topic.

3.3 Upper-To-Total Ratio (*UTR*) and the *giniUTR* Indicator

The density estimation yields for every grid point an estimated density value. Let $\boldsymbol{y} = (y_1, \ldots, y_n)^t$ be the vector containing the density estimates on n grid points. On regions where no information is given, the resulting density estimates are not zero but very close to zero. Therefore a baseline – a minimum level, to split regions with no information from the rest – is defined and only values larger than this baseline are considered for the analysis. For example, for Figure 4 the baseline should be chosen at a height which allows to separate the black area from the rest. This is also visible in Figure 2 where regions with no information have to be intuitively separated from the rest. The level of the baseline is denoted by $\alpha_1 \in \mathbb{R}_+$.

The UTR is then defined as

$$UTR_j = \frac{\sum_{l=1}^{n} \mathbb{I}(y_l \geq \alpha_j)}{\sum_{l=1}^{n} \mathbb{I}(y_l \geq \alpha_1)} \in [0, 1]\,, \tag{1}$$

with

$$\mathbb{I}(y_l \geq \alpha_j) = \begin{cases} 1, & \text{for } y_l \geq \alpha_j \\ 0, & \text{else } , \end{cases}$$

where $\alpha_1 \leq \alpha_j \in \mathbb{R}_+$.

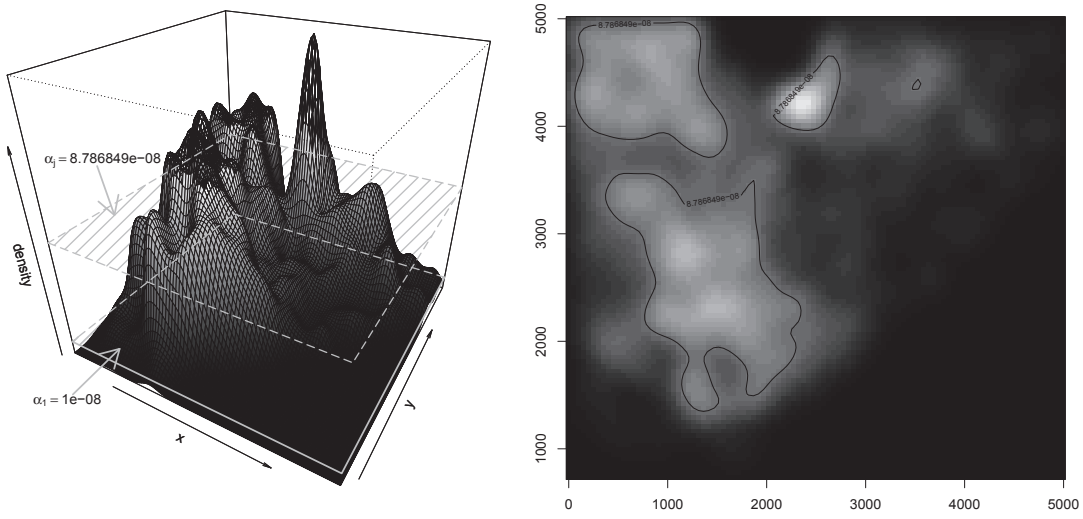

Figure 4: Bivariate kernel density estimates. In the left plot, a 3D presentation of the values of the bivariate density estimate is presented. In the right plot, the same result is presented in a two dimensional representation. In both plots, the level cut at $\alpha_j = 8.786849 \cdot 10^{-8}$ along the z-axis is shown.

Figure 4 shows the result of a two-dimensional kernel density estimate of an image. The level cut $\alpha_j = 8.786849 \cdot 10^{-8}$ is visible in both plots, in the left as a level cut in the z-axis, and in the right plot marked as the border (black line) between lightgrey and darkgrey. The UTR is just the ratio of lightgrey marked pixels to the sum of darkgrey and lightgrey marked pixels. The area in black indicates density values below the baseline $\alpha_1 = 10^{-8}$.

Later on, the UTR is calculated for m level cuts $\alpha_1, \ldots, \alpha_m$. More precisely, a vector of 50 equidistant level cuts $\alpha_j \in [bl, ul]$ is defined, where bl is the baseline. The upper limit ul is the maximum value of the density estimations of all regarded sectors and is defined as

$$ul = \max_{p=1,\ldots,P}(\boldsymbol{y}^p)\,, \tag{2}$$

where the vector $\boldsymbol{y}^p$ contains all density estimations of sector S_p, and P is the number of sectors. The definition of one common threshold for all sectors allows for comparisons between sectors.

The Gini and the *giniUTR*:
For a sorted data vector $\boldsymbol{x} = (x_1, \ldots, x_n)^t$ (from smallest to largest values) the Gini Index (Gini, 1912) is given by

$$gini(\boldsymbol{x}) = \frac{\sum_{i=1}^{n}(2i - n - 1)x_i}{n^2\mu} \in [0, 1]\,, \tag{3}$$

where i describes the rank order number and $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $x_i > 0$. If the Gini Index takes the value 0 then a perfect equality is given and every data value is equal. The higher the inequality, the higher the Gini Index. More details on interpretation of the Gini Index (over the Lorenz curve (Lorenz, 1905)) can be found in basic statistical textbooks.

The following proposed indicator compares the upper-to-total-ratio (UTR) of the data with the UTR of the uniform distribution for m level cuts. Using the definition of the UTR in Equation (1) and the definition of the Gini Index in Equation (3), the proposed indicator *giniUTR* is given as follows

$$giniUTR = gini(UTR) = \frac{\sum_{i=1}^{n}(2i - m - 1)UTR_i}{m^2\mu}\,, \tag{4}$$

where $UTR_i \in (0, 1]$ and $\mu = \frac{1}{m}\sum_{i=1}^{m} UTR_i$.

3.4 Further Comparison and Indicators Using Upper-to-Total-Ratios

The previous indicator compares the density of the uniform distribution to the empirical one (even if the comparison between sectors is in main focus). Having the data uniformly distributed is somehow an extreme case. Kernel density estimates that would follow a bivariate normal distribution is another extreme case. This is covered by the $cpUTR$ indicator, i.e. the UTR's of all sectors again are considered and compared with the UTR of the normal distribution. Therefore, the aim is to compare the UTR's of the empirical data with the UTR's of the theoretical normal distribution for a set of level cuts along the z-axis.

The UTR of the bivariate normal distribution with the parameters $\rho = 0$, $\mu_1 = \mu_2 = 0$ and $\sigma_1 = \sigma_2 = \sigma$ can also be calculated analytically. Note that other choices of σ can be used but here we concentrate on a symmetric distribution in two dimensions, i.e. on circles instead of ellipses. This is a natural choice when no prior information about the bivariate distribution is given.

Given two random variables $(X, Y) \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ and if one defines

$$\tilde{x} = \frac{x - \mu_1}{\sigma_1} \qquad \text{and} \qquad \tilde{y} = \frac{y - \mu_2}{\sigma_2}\,,$$

then it is well-known that the bivariate standard normal probability density function is given by

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}(\tilde{x}^2 - 2\rho\tilde{x}\tilde{y} + \tilde{y}^2)\right)\,. \tag{5}$$

The contours of the bivariate normal distribution describe in this case circles with center $(0, 0)$ and radius $r_i = \sqrt{-2\sigma^2 \log(2\pi\sigma^2\alpha_i)}$. For several level cuts $\alpha_i, i = 1, \ldots, m$, the UTR_i can be estimated as the ratio of the area of the circle at the level cut α_i in relation to the area of the circle at the first level α_1. Therefore, the UTR at a level of i (for which $\alpha_i > \alpha_1$ holds) of the bivariate normal distribution with the parameters $\rho = 0$, $\mu_1 = \mu_2 = 0$ and $\sigma_1 = \sigma_2 = \sigma$ is defined as

$$nUTR_i = \frac{r_i^2}{r_1^2} \in [0, 1]\,. \tag{6}$$

The increase and decrease of $nUTR$ can also be expressed via an approximation, which is derived in the following.

Derivation of the *nUTR* in the $N(0, 0, \sigma^2, \sigma^2, 0)$ case:
As already mentioned before, we expect a circular bivariate normal distribution where the means of the bivariate normal can be anywhere. For simplicity we set the mean to the origin $(0, 0)$ since we are only interested in the UTR's that are not dependent on the location of the theoretical normal distribution.

In the following, the aim is to find a numerical approximation for the derivation of the $nUTR$, i.e. to be able to calculate the $nUTR$ for each level cut a_i.

At first the equidistant level cuts α_i with $i = 1, \ldots, m$, where $0 < a = \alpha_1 < \alpha_2 < \cdots < \alpha_m = b < 2\pi\sigma^2$, are defined as a function k through

$$\begin{aligned} k(i) &= a + (i-1)\frac{b-a}{m-1} \\ &= a + (i-1)h_\alpha\,. \end{aligned} \tag{7}$$

Considering $r_i = \sqrt{-2\sigma^2 \log(2\pi\sigma^2\alpha_i)}$ and (6) yields the following definition of the $nUTR$

$$\begin{aligned} nUTR_i \;&=\; \frac{r_i^2}{r_1^2} = \frac{-2\sigma^2 \log(2\pi\sigma^2\alpha_i)}{-2\sigma^2 \log(2\pi\sigma^2\alpha_1)} \\ &\overset{(7)}{=} \frac{-2\sigma^2 \log(2\pi\sigma^2 k(i))}{-2\sigma^2 \log(2\pi\sigma^2 k(1))} = \frac{\log(2\pi\sigma^2 k(i))}{\log(2\pi\sigma^2 k(1))} \\ &:= V(k(i))\,. \end{aligned} \tag{8}$$

Now it is possible to calculate the derivative of $V(k(i))$ and hence the behavior of the $nUTR$ can be analyzed. Using the chain rule

$$\frac{\partial V(k(i))}{\partial i} = \frac{\partial V}{\partial k} \cdot \frac{\partial k}{\partial i}\,,$$

the derivative of V is

$$\frac{\partial V(k(i))}{\partial i} = \frac{1}{2\sigma^2\pi \log(2\pi\sigma^2 k(1)) \cdot k(i)} \cdot \frac{\partial k}{\partial i}\,, \tag{9}$$

where

$$\frac{\partial k}{\partial i} = h_\alpha = \frac{b-a}{m-1} > 0\,, \tag{10}$$

and thus

$$\begin{aligned}\frac{\partial V(k(i))}{\partial i} &= \frac{h_\alpha}{-r_1^2\pi}\frac{1}{k(i)} \\ &= -\frac{h_\alpha}{r_1^2\pi(a+(i-1)h_\alpha)} < 0 \quad \forall i \in \mathbb{N}\,.\end{aligned}$$

If $i \in \mathbb{N}$ increases by one unit, then k increases with the factor h_α and hence the derivative of V changes according to the above derivations. Knowing these properties, it is easily possible to calculate the $nUTR$ of a normal distribution for each level cut analytically.

The choice of the parameter a is essential, and here we choose the lower limit $a = bl$ – the baseline that separates the region with no information (see Section 3.3).

3.4.1 The Scaled *UTR* Indicator

Since the maximum of the normal distribution is rather high in comparison to empirical density estimates in sectors, a different approach to compare the densities has been developed. We tried different approaches to define level cuts, like equal level cuts for each sector as done in the section before, equal level cuts for each sector except the normal density (which was scaled down), etc. We obtained the best results with level cuts $\boldsymbol{\alpha}$ that are defined for each sector and for the bivariate normal distribution individually. In each sector, the 50 level cuts are made from the baseline up to the maximum of the density values of the corresponding sector and for the maximum of the theoretical normal density.

The $scaledUTR$ is defined in a sector S_p, $p \in \{1, \ldots, P\}$, as

$$scaledUTR_j^p = \frac{\sum_{l=1}^n \mathbb{I}(y_l^p \geq \alpha_j^p)}{\sum_{l=1}^n \mathbb{I}(y_l^p \geq \alpha_1^p)} \in [0,1] \tag{11}$$

with

$$\mathbb{I}(y_l^p \geq \alpha_j^p) = \begin{cases} 1, & \text{for } y_l^p \geq \alpha_j^p \\ 0, & \text{else } . \end{cases}$$

Here, y_l^p are the n density estimates in sector S_p, and α_j^p refers to the j-th level cut in sector S_p. The upper limit of the level cuts in the p-th sector is limited by the maximum of the density estimates in this sector, $\mathbb{R}_+ \ni \alpha_1^p \leq \alpha_j^p \leq \alpha_m^p = \max_l(y_l^p)$ (with $j \in \{1, \ldots, m\}$). For the bivariate normal density, the m level cuts are made from the baseline to the top.

3.4.2 The *gini-cpUTR* Indicator

In the following an indicator is defined for the comparison of the UTRs between the data and the bivariate normal distribution. Here, the absolute value of the difference of both Gini Indices is calculated:

$$gini\text{-}cpUTR = |gini(scaledUTR \text{ of emp. data}) - gini(nUTR \text{ of normal distr.})| \,. \quad (12)$$

3.5 The Indicator *NGroups*

The following indicator is supposed to give a view into the behavior of corresponding clusters within the density estimation for each level cut α_i, with $i = 1, \ldots, m$, and within a predetermined evaluation area (*eval.area*). Note that the level cuts now are again to be chosen equal for each sector (as done in Section 3.3).

The evaluation area is the largest coherent area containing the majority of the density estimates over the baseline. The purpose is to determine the number of groupings/clusters of the density values within such a defined evaluation area for several level cuts. Figure 4 shows a simple situation, where the evaluation area is equal to the non-black grey area and within this grey area there are four clusters (marked by the black contour lines and region diplayed in lightgrey).

The indicator $NGroups$ is defined as the total number of the clusters within the evaluation area (*eval.area*) at a level cut $\alpha_i \in \{1, \ldots, m\}$:

$$NGroups_i = \sum \mathbb{I}(C_j \in eval.area) \,, \quad (13)$$

with

$$\mathbb{I}(C_j \in eval.area) = \begin{cases} 1, & \text{for } C_j \in eval.area, \\ 0, & \text{else} \ , \end{cases} \quad (14)$$

where C_j denotes the j-th cluster.

The calculation of $NGroups$ for several level cuts $\alpha_1, \ldots, \alpha_m \in \mathbb{R}$ is of further interest and is also needed for the calculation of the last proposed indicator $modCHI$ in Equation (17). $NGroups$ is a vector of length m, i.e. one value for each level cut.

The challenges were to implement functionality for receiving the largest contour line considering several conditions, the coordinates of the requested evaluation area, the baseline, the level cut area and assigning the density values to the groups.

3.6 Modified Calinski Harabasz Index (*modCHI*)

The last proposed indicator is used to evaluate the spatial distribution of clusters, i.e. the separation of clusters is considered. The aim is to use an indicator that evaluates the density and the separation between clusters. One choice is to use the Calinski Harabasz Index to measure the dissimilarity between clusters over the dissimilarity within clusters (Calinski and Harabasz, 1974, Maulik and Bandyopadhyay, 2002 and Schlittgen, 2009).

Slight modifications of the Calinski Harabasz Index yields the definition of this last indicator. Instead of the usual between cluster sum of squares and the within cluster sum of squares that is used in the definition of the original version of the Calinski Harabasz

Index, here the absolute distances and pairwise distances between clusters, respectively, are used. These modifications lead to the following definitions:

$$BSA = \sum_{j=1}^{K} \sum_{l=j+1}^{K} |\overline{\boldsymbol{x}}_l - \overline{\boldsymbol{x}}_j| \tag{15}$$

and

$$WSA = \sum_{j=1}^{K} \sum_{i=1|x_i \in C_j}^{N_j} |x_{ij} - \overline{\boldsymbol{x}}_j| \,. \tag{16}$$

Here, x_i refers to the coordinates of a grid points in the picture, x_{ij} is a grid point in cluster C_j, N_j is the number of grid points in cluster C_j, K is the number of clusters ($NGroups$) within this area, and $\overline{\boldsymbol{x}}_j$ is the center of the j-th cluster C_j. The overall number of grid points in the considered evaluation area is $N = \sum_j N_j$. The indicator $modCHI$ is obtained by setting BSA and WSA in relation:

$$modCHI = \frac{BSA}{WSA} \frac{N-K}{K-1} \,. \tag{17}$$

Since the $modCHI$ is a dissimilarity measure, the larger the values for the $modCHI$, the better the separation of the clusters within the considered area.

The indicator $modCHI$ is also used to be a vector of length of m, since it is calculated for m level cuts $\boldsymbol{\alpha} \in \mathbb{R}^m$ within the predetermined evaluation area. The level cuts are again to be chosen equal for each sector.

4 Results

The original digitized slides of about 100.000 $\times$ 100.000 pixels are split into sectors of about 5000×5000 pixels. In the following we present results of one of the smallest digitized slides for illustration. From that data we already presented the Ki67 labeled cell nuclei in Figure 2, and from Sector 1, we have already shown the density estimation in Figure 4. Note that for the two-dimensional kernel estimates with grid size of 100, an optimal bandwidth (Venables and Ripley, 1994) is used for evaluation. All values above the baseline (grey and lightgrey area, see, e.g., Figure 4) for each sector are used in the calculations.

4.1 *UTR* and *giniUTR*

As already mentioned, 50 equidistant level cuts $\alpha_j \in [bl, ul]$ are used, see Equation (2).

In the following, the UTR's for all 14 sectors of Sample 1 are estimated (according to Equation (1)). Figure 5 presents the UTR's for the Sectors 1–5 (left), 6–10 (right) and 11–14 (bottom) on a logarithmic scale of the level-cuts for better comparisons. The UTR has to be 1 for the uniform distribution (grey lines in the figures).

The UTR's for the Sectors S_1, S_4 and S_5 seem to behave similar, nearly identical with a fast decrease to 0, whereas the Sectors S_{12} and S_{14} but also S_2 show a clear deviation.

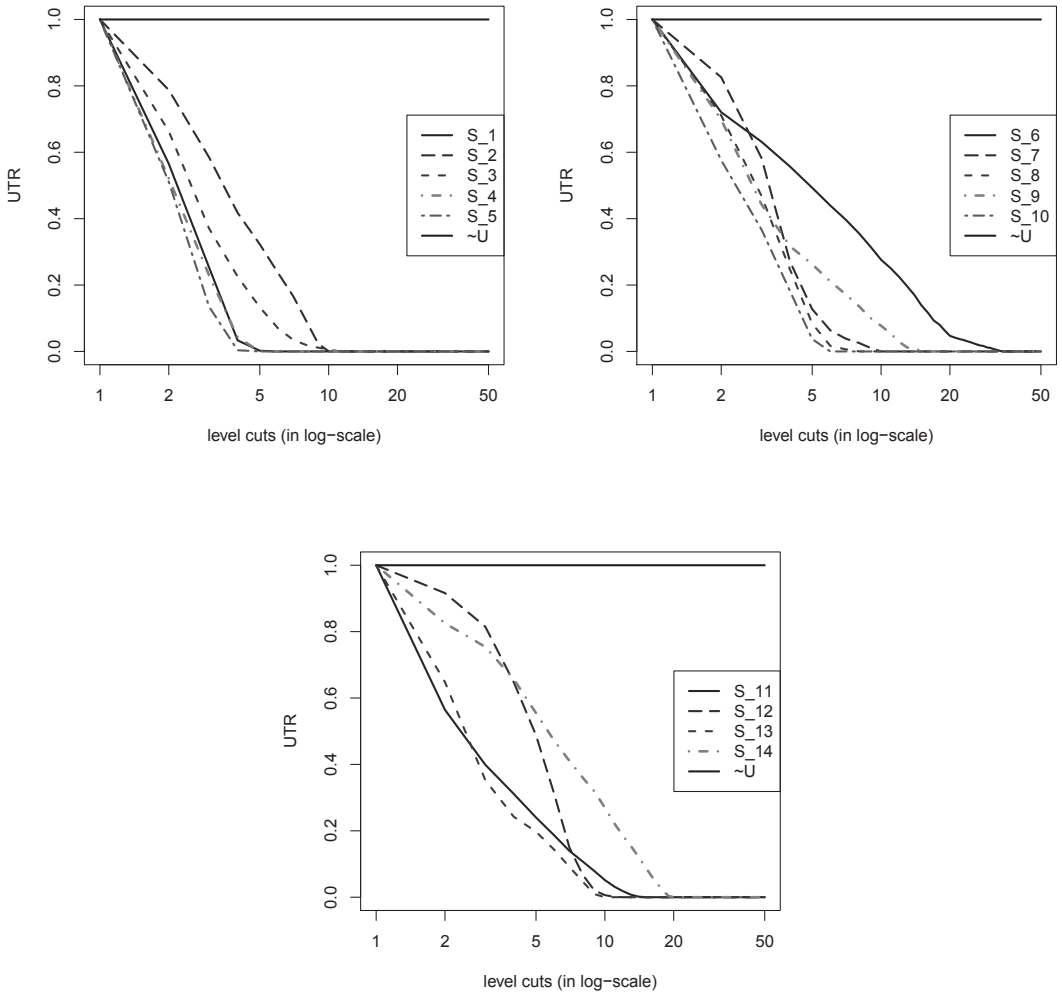

Figure 5: Upper-to-total ratio (UTR) for Sectors S_1 to S_{14}.

In addition, it is obvious that the UTR of Sector S_6 falls slower than in other sectors. The Sector S_{14} has also a slower decrease than the other sectors. Generally, the comparison of the UTR of all sectors of Sample 1 shows that the Sectors S_2, S_6, S_{12} and S_{14} have a clear deviation in the behavior compared to the remaining ones.

Table 1 reports the inequality based on the *giniUTR* indicator (in percentages) of all sectors.

Naturally, the obtained results show a high inequality to the uniform distribution. Sector S_5 shows the largest inequality with 96.08 % and the smallest inequality is in Sector S_6 with 73.35 %.

4.2 Results for the Indicators *scaledUTR* and *gini-cpUTR*

The following parameters for the arithmetic mean and standard deviation for the bivariate normal distribution have been selected: $\mu_1 = \mu_2 = 0$ and $\sigma_1^2 = \sigma_2^2 = 10$, and correlation $\rho = 0$.

Table 1: $giniUTR$ of all sectors of Sample 1.

Sector	1	2	3	4	5	6	7
$giniUTR$ in %	95.45	89.73	92.54	95.52	96.08	73.35	92.39
Sector	8	9	10	11	12	13	14
$giniUTR$ in %	93.50	87.47	94.29	88.43	89.32	91.35	79.72

The next step is to define the level cuts $\boldsymbol{\alpha}$ for each sector. This is necessary since the aim here is on the comparison with the behavior of the normal distribution. The level cuts are now within the interval $[bl, \max(N(0, 0, 10, 10, 0))]$ and hence $\alpha_j \in [10^{-8}, 0.0159]$, with $j = 1, \ldots, 50$. Again 50 equidistant level cuts are considered.

The calculation of the UTR for the bivariate normal distribution is given in Equation (8). The following figures show the scaled UTR's of all sectors of Sample 1 including the UTR for the bivariate normal distribution.

From Figure 6 it is visible that the scaled UTR of Sector S_{12} of Sample 1 shows the most deviant behavior of the $nUTR$ and of the other sectors, whereas Sector S_6 has the most similar behavior of the UTR to the $nUTR$. Sectors S_9 and S_{11} show also an UTR near to the UTR of the normal distribution.

In the following Table 2 the results for the comparison of the scaled UTRs of the sectors with the $nUTR$ of the bivariate normal distribution are presented.

Table 2: Results for $gini\text{-}cpUTR$ for Sample 1.

Sector	1	2	3	4	5	6	7
$gini\text{-}cpUTR$ in %	1.39	10.20	8.21	7.44	6.71	7.36	2.74
Sector	8	9	10	11	12	13	14
$gini\text{-}cpUTR$ in %	1.19	5.45	8.47	2.81	6.94	3.63	8.56

The obtained results show that the maximum absolute difference in behavior of the scaled UTR's is between the Sector S_2 and the normal distribution and is 10.2 %. The minimum difference is between S_8 and normal distribution and S_1 and normal distribution. In average, the densities in the sectors are more similar to the normal distribution (black thick solid line) than to the univariate distribution (grey solid line). Anyhow, the comparison of sectors or samples is the main intention here.

4.3 Results for the Indicator *NGroups*

The following two indicators deal with the spatial distribution of possible clusters within the density estimations of all sectors. Therefore 50 equidistant level cuts $\alpha_j \in [bl, ul]$, with $j = 1, \ldots, 50$ are considered again. The level of the baseline (bl) is chosen as usual (see e.g. Section 3.3) and the upper limit (ul) is again defined as the maximum value of the density estimations of all regarded sectors (see Equation 2). Figure 7 shows results for the indicator $NGroups$ for all sectors of Sample 1.

From Figure 7 it is apparent that all sectors start with one cluster at the level of baseline and all end with one cluster at their last level. The largest number of clusters is in Sector

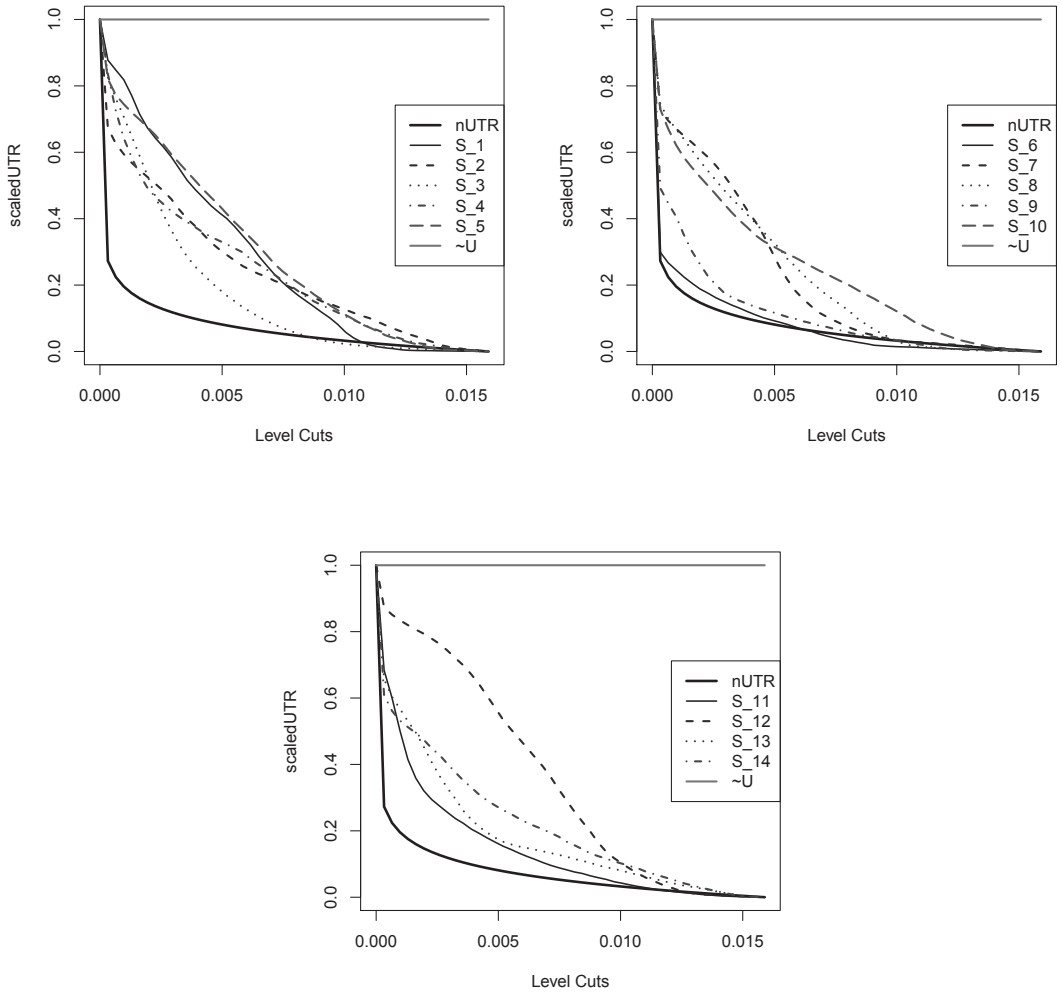

Figure 6: Scaled UTRs for Sectors S_1 to S_{14} vs. $nUTR$.

S_7 at the 4-th level cut and has the value 9. Sector S_6 is the sector with the highest density peak, followed by Sector S_{14}.

4.4 Results for the Indicator *modCHI*

After computing the number of clusters per level in the previous subsection, now the separation of these clusters is of further interest. The following plots represent the clusters of the density estimation for each sector at the level where the sectors have a maximum of clusters.

Equations (15), (16) and (17) are applied to all sectors considering all 50 levels. The larger the value for the $modCHI$, the better the separation of the groups within the evaluation area.

The obtained results for Sample 1 are shown in Figure 9.

The maximum value for $modCHI$ is obtained for Sector S_{11} at the 13-th level being 36.43. $modCHI = 0$ means that there is only one cluster within the evaluation area at this

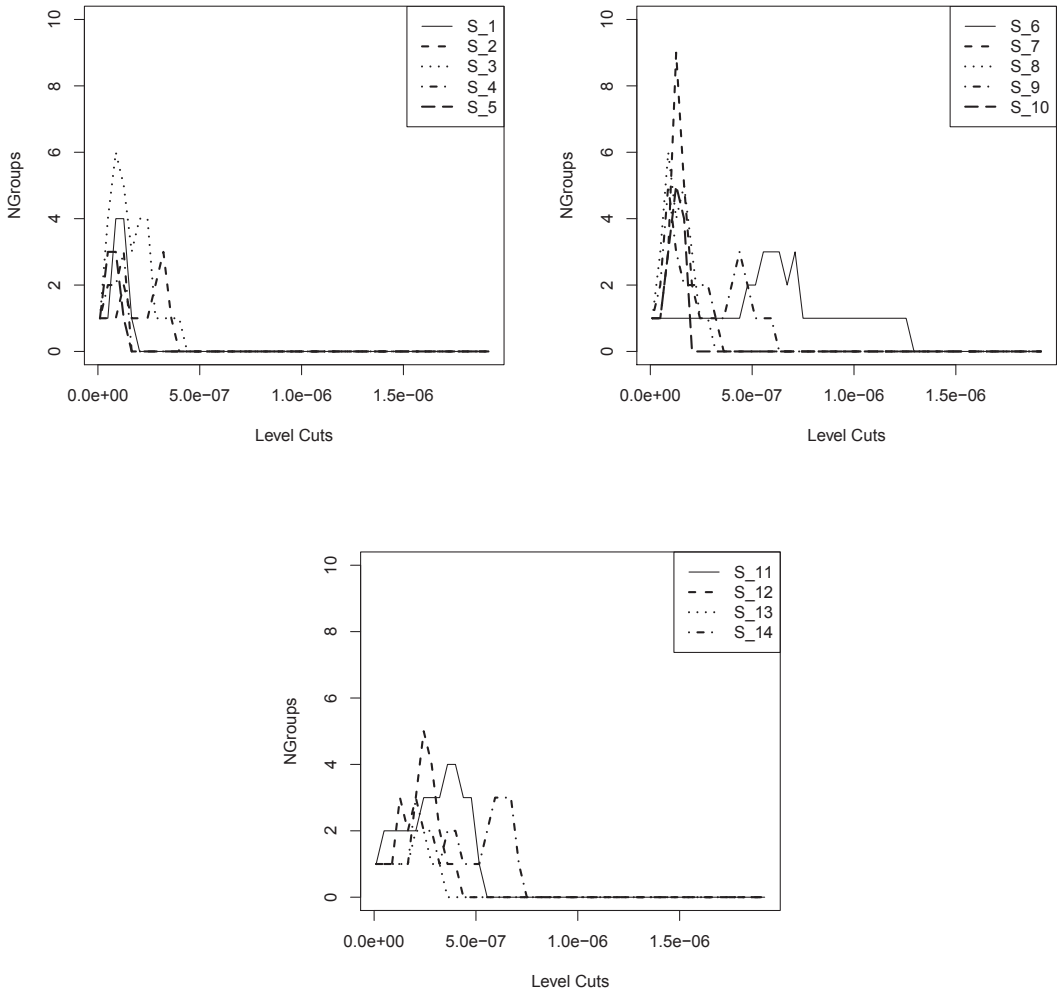

Figure 7: Indicator $NGroups$ for Sectors S_1 to S_{14}.

level, whereas a missing value is obtained if there is not any cluster within the evaluation area and thus the curves in Figure 9 stop at the last level containing a cluster.

5 Summary and Outlook

This contribution deals with the basic research of statistical support on the issue of computer-based assessment of pathological characteristics of brain tumors and it has been performed on request of a team of scientists from the Institute of Neurology, Medical University of Vienna.

Since manual inspection of the digitized tissues results is very time-consuming, the aim was the definition of some indicators that ensure an objective assessment. Two digitized human brain tumor tissue-sections were analyzed in detail, results from the second (larger) sample are omitted to keep within the limit of pages. However, results on the second sample are published in a master thesis (Aklan, 2012).

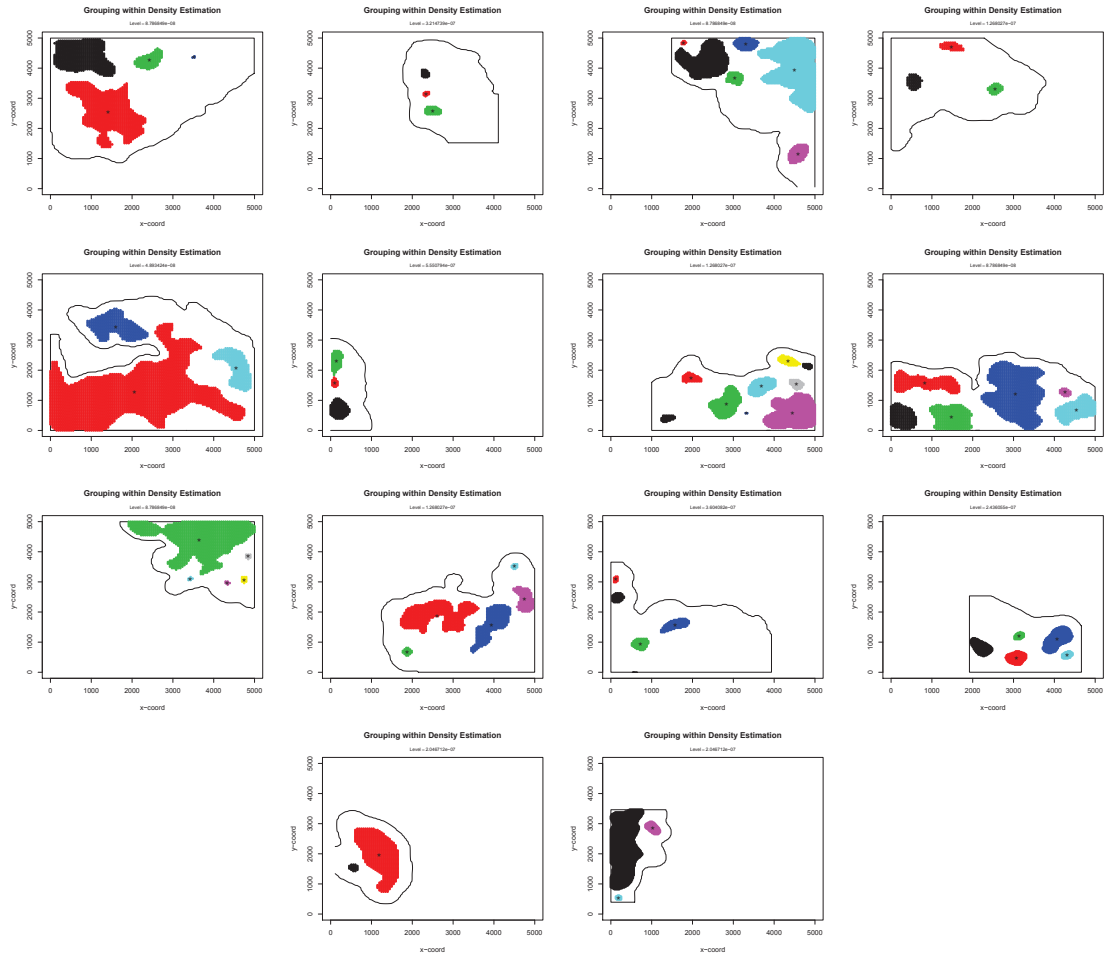

Figure 8: Clusters within Sectors S_1 – S_{14}. The cluster centers are marked with a star ($*$).

At first the scanned and digitized brain tumor samples underwent a process of segmentation in several parts and the determination of the marked cell nuclei during the pre-processing. This splitting into sectors was also necessary because the digitized slides had an average resolution of the size of 100.000 $\times$ 100.000 pixels and the processing of such an image with a color depth of 24 bit would need a memory consumption of 30 GB just to read in the data. The images could even have larger sizes, and this would lead to a huge memory consumption, i.e. 100 GB or larger. For this reason but mainly to compare different sectors in an image, the slides were divided into sectors of size 5000 $\times$ 5000 pixels.

To extract the most important information out of the data, bivariate kernel density estimations have been applied and evaluated. Different indicators were defined for different needs. The upper-to-total ratios (UTR, $nUTR$ and related Gini's) measure and compare the inequality to the bivariate uniform and normal distribution for each sector. They account for the form/profile and the height of the density. For the scaled version ($scaledUTR$), the level cuts are not cut equidistantly but individually. This allows to concentrate on the distribution of the samples and the very large impact of few small

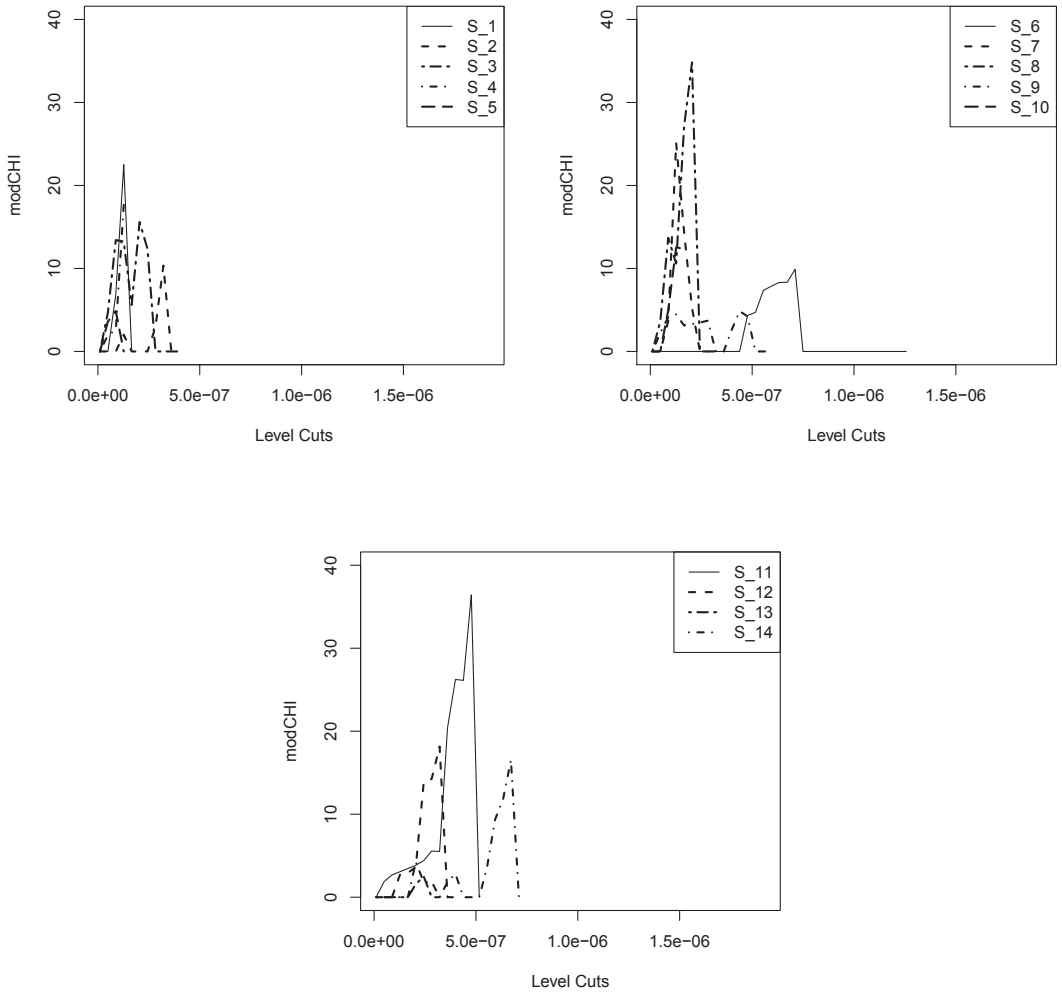

Figure 9: "Separation" of clusters within the evaluation area of $S_1 - S_{14}$.

high-density regions is reduced. The form/profile of the surface is in focus. The indicator $NGroups$ concentrates on the number of peaks at certain level cuts. Therefore, this indicator should give an impression if and how many small high-density regions are present in the sample/sector. Another concept is related to the spatial distribution for each level cut within a predetermined coherent evaluation area. This gives an impression about the separation of the peaks of the distributions in a sector/sample.

Differences between sectors (shown in this contribution) of digitalized brain tumor samples can be made visible by the proposed set of indicators and further manual analysis might only be necessary for those sectors/samples with abnormally high or low indicator values. Automated image analysis may offer objective and time-efficient assessment of immunostained slides in the clinical setting.

For small tissues like the one used, some problems still have to be solved. Especially, further investigations may consider the non-rectangular form of data in the sectors of smaller tissue samples. Even for larger samples, the minor parts of several sectors may not be fully covered by data values (blue and brown pixels). This requires further research

on this topic. It is also reasonable to apply other statistical techniques, e.g. by splitting the sectors again into parts and to apply a log-linear model to estimate the counts of cell nuclei in the defined areas. However, the main problem – dealing with *Big Data* – still exists.

References

Aklan, S. (2012). *Development and implementation of statistical indicators for the assessment of brain tumors*. Unpublished master's thesis, Vienna University of Technology.

Beucher, S., and Lantuéjoul, C. (1979). Use of watersheds in contour detection. In *Realtime edge and motion detection/estimation.*

Calinski, R., and Harabasz, J. (1974). A Dendrite Methode For Cluster Analysis. *Communications in Statistics*, *3*, 1–27.

CERN, F. (2011). *Pediatric Ependymoma Images.* Website. (Available online at: `http://www.cern-foundation.org/Content.aspx?id=608`)

Gini, C. (1912). Variabilità e mutabilità: contributo allo studio delle distribuzioni e delle relazioni statistiche. *Studi Economico-Giuridici della R. Università di Cagliari*, *3*, 3–159.

Hamamatsu. (2012, March). *Virtual microscopy / nanozoomer.* Website. (Available online at: `http://sales.hamamatsu.com/en/products/system-division/virtual-microscopy.php`)

Härdle, W., Müller, M., Sperlich, S., and Werwatz, A. (2004). *Nonparametric and Semiparametric Models*. Berlin New York: Springer-Verlag.

Lorenz, M. O. (1905). Methods of measuring the concentration of wealth. *Publications of the American Statistical Association, Volume 9, Number 70, p. 209-219*, *9*, 209-219.

Maulik, U., and Bandyopadhyay, S. (2002). Performance Evaluation of Some Clustering Algorithms and Validity Indices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*, 1650–1654.

Otsu, N. (1979, January). A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, man and Cybernetics*, *9*(1), 62–66.

Preusser, M., Heinzl, H., Gelpi, E., Höftberger, R., Fischer, I., Pipp, I., et al. (2008). Ki67 index in intracranial ependymoma: a promising histopathological candidate biomarker. *Histopathology*, *53*, 39–47.

Roerdink, J., and Meijster, A. (2000). The watershed transform: Definitions, algorithms and parallelization strategies. *FUNDINF: Fundamenta Informatica*, *41*.

Schlittgen, R. (2009). *Multivariate Statistik*. München: Oldenbourg Wissenschaftsverlag München.

Scott, D. (1992). *Multivariate Density Estimation*. New York, Chichester, Brisbane, Toronto, Singapure: Wiley-Interscience, John Wiley & Sons INC.

Statistik Austria. (2011, Oktober). Website. (Available online at: `http://www.statistik.at/web\_de/statistiken/gesundheit/krebserkrankungen/gehirn\_zentralnervensystem/index.html`)

Venables, W., and Ripley, B. (1994). *Modern Applied Statistics with S-Plus*. New York and Berlin and Heidelberg: Springer.

Wählby, C., Sintorn, I.-M., Erlandsson, F., Borgefors, G., and Bengtsson, E. (2004, July). Combining Intensity, Edge and Shape Information for 2D and 3D Segmentation of cell nuclei in tissue sections. *Journal of Microscopy*, *215*, 67–76.

Walser, A. (2011). *Automatisierte Auswertung der Zellproliferation in menschlichen Gehirntumoren*. Unpublished master's thesis, Vienna University of Technology.

Authors' addresses:
Matthias Templ[1,2], Semaguel Aklan[1], Peter Filzmoser[1], Matthias Preusser[3], Johannes A. Hainfellner[3,4]

[1] Department of Statistics and Probability Theory
Vienna University of Technology
Wiedner Hauptstrasse 8-10
A-1040 Vienna
E-Mail: Templ@tuwien.ac.at and P.Filzmoser@tuwien.ac.at

[2] Methods Unit
Statistics Austria
Guglgasse 13
A-1110 Vienna
E-Mail: Matthias.Templ@statistik.gv.at

[3] Department of Medicine I & Comprehensive Cancer Center Vienna
Medical University of Vienna
Währinger Gürtel 18-20
A-1090 Vienna, Austria
E-Mail: Matthias.Preusser@meduniwien.ac.at

[4] Institute of Neurology
Medical University of Vienna
AKH 4J, Währinger Gürtel 18-20
A-1090 Vienna, Austria
E-Mail: Johannes.Hainfellner@meduniwien.ac.at