The influence of resolution on the predictive power of spatial heterogeneity measures as biomarkers of liver fibrosis

Peer-reviewed author version

CLAES, Jari; AGTEN, Annelies; Blazquez-Moreno, Alfonso; Crabbe, Marjolein; Tuefferd, Marianne; Goehlmann, Hinrich; GEYS, Helena; Peng, Cheng-Yuan; NEYENS, Thomas & FAES, Christel (2024) The influence of resolution on the predictive power of spatial heterogeneity measures as biomarkers of liver fibrosis. In: COMPUTERS IN BIOLOGY AND MEDICINE, 171 (Art N° 108231).

DOI: 10.1016/j.compbiomed.2024.108231
Handle: http://hdl.handle.net/1942/42892

The influence of resolution on the predictive power of spatial heterogeneity measures as biomarkers of liver fibrosis

Jari Claes [1], Annelies Agten[1], Alfonso Blázquez-Moreno[2], Marjolein Crabbe[2], Marianne Tuefferd[3], Hinrich Goehlmann[3], Helena Geys[2], Cheng-Yuan Peng[4], Thomas Neyens[1,5], and Christel Faes[1]

[1]Data Science Institute, UHasselt - Hasselt University, Agoralaan 1, BE 3590 Diepenbeek, Belgium.

[2]Discovery Statistics, Global Development, Janssen Research and Development, Turnhoutseweg 30, 2340 Beerse, Belgium.

[3]Translational Biomarkers, Infectious Diseases, Janssen Research and Development, Turnhoutseweg 30, 2340 Beerse, Belgium.

[4]China Medical University Hospital, Taichung, Taiwan.

[5]L-BioStat, KU Leuven, Kapucijnenvoer 35, 3000 Leuven, Belgium.

Abstract

Spatial heterogeneity of cells in liver biopsies can be used as biomarker for disease severity of patients. This heterogeneity can be quantified by non-parametric statistics of point pattern data, which make use of an aggregation of the point locations. The method and scale of aggregation are usually chosen ad hoc, despite values of the aforementioned statistics being heavily dependent on them. Moreover, in the context of measuring heterogeneity, increasing spatial resolution will not endlessly provide more accuracy. The question then becomes how changes in resolution influence heterogeneity indicators, and subsequently how they influence their predictive abilities. In this paper, cell level data of liver biopsy tissue taken from chronic Hepatitis B patients is used to analyze this issue. Firstly, Morisita-Horn indices, Shannon indices and Getis-Ord statistics were evaluated as heterogeneity indicators of different types of cells, using multiple resolutions. Secondly, the effect of resolution on the predictive performance of the indices in an ordinal regression model was investigated, as well as their importance in the model. A simulation study was subsequently performed to validate the aforementioned methods. In general, for specific heterogeneity indicators, a downward trend in predictive performance could be observed. While for local measures of heterogeneity a smaller grid-size is outperforming, global measures have a better performance with medium-sized grids. In addition, the use of both local and global measures of heterogeneity is recommended to improve the predictive performance.

1 Introduction

In the study of disease progression, examination of the tissue has been an essential tool. Disease staging and resistance to treatment are often linked not only to the presence or abundance of certain cells in the tissue microenvironment, but also to the dynamics and heterogeneity between certain cells of different types and purposes [1, 2]. In recent years, digital methods have gained a foothold in pathology. In particular, digital pathology has proven to be essential within the context of the tumor microenvironment [3]. Technological developments have allowed for fast and accurate microscopic imaging, as well as for the subsequent application of artificial intelligence (AI) techniques designed to identify cells within the environment. Given the digitized information, further AI and statistical techniques can be applied to quantify the aforementioned heterogeneity between cells in more detail. Specifically, techniques stemming from point pattern analysis as a branch of spatial statistics are relevant.

In point pattern analysis, in order to calculate non-parametric statistics, point locations often need to be aggregated. Examples of these statistics include but are not limited to the Morisita-Horn and the Shannon index, statistics that are frequently used in ecology [4, 5] and more recently histology [6, 7]. The method of aggregation, however, is not standardized but generally chosen by the author ad hoc. Methodological differences can potentially lead to differences in the values of the calculated statistics and therefore differences in conclusion.

The issue of determining an optimal scale has already been studied in spatial point pattern data, specifically in the context of crime patterns [8]. The paper builds on the goodness-of-fit-for-multiple-resolutions method designed by Constanza [9] and calculates Andresen's S Index [10], which indicates similarity of two point patterns, over iteratively decreasing grid cell sizes. The conclusion from the paper reads that the best scale should be a balanced and data context dependent trade-off between the benefits, being the accuracy, and the pitfalls, being the possible noise and data collection issues, of using increased resolutions.

Although the method and conclusions provided by the paper are generally applicable, the issues for point pattern data concerned with measuring heterogeneity can be slightly different. Increasing spatial resolution will not endlessly provide more accuracy, as limiting each grid cell to only a few data points might not yield as much information about heterogeneity between species in the sample. In the most extreme case, limiting grid cells to either being empty or containing one single data point will yield no such information within any grid cell at all. The question then becomes how changes in grid choice influence heterogeneity indicators derived from this grid aggregation, and subsequently how they influence predictive qualities of these indicators. Considering heterogeneity indicators have recently been increasingly used as biomarkers in histological studies [11, 12, 13, 14, 15, 16], studying the influence of grid size is essential to safeguard the accuracy of conclusions drawn from these statistical analyses. The aim of this paper is to analyze this issue by evaluating different heterogeneity indicators as well as their influence in models attempting to predict disease staging under different choices of aggregation.

This paper builds on the concepts and application from Agten and colleagues [17]. They investigated both spatial summary statistics and spatial models to summarize the cellular micro-environment of liver biopsies, to be used as biomarker for the patient's stage of liver fibrosis. In this paper, the spatial summary statistics are further investigated, and in particular the dependency on the aggregation scale. Using a penalized ordinal regression model, allowing to take into account the censored nature of the fibrosis score, we will investigate the sensitivity of the grid choice on the predictive behavior of liver fibrosis. The next section summarizes the data and different spatial summary statistics that are investigated. Both a simulation study and data application are used to investigate the impact on the grid-choice.

2 Methods

2.1 Data

The data contains 110 liver biopsies that were collected from chronic hepatitis B patients, consisting of 100 formalin-fixed and paraffin-embedded core needle biopsies delivered by Avaden Biosciences and 10 fresh frozen liver biopsies from China Medical University in Taiwan. Further details concerning the datasets can be found in the Appendix. Ethics commission numbers

Table 1: Range of fibrosis stages defined by METAVIR scores found in the liver data. 15 of 110 scores are not defined as one stage, but rather as an indefinite score ranging between two values.

Fibrosis score	**0**	**0-1**	**1**	**1-2**	**2**	**2-3**	**3**	**3-4**	**4**
Number of samples	37	4	23	6	13	4	17	1	5

for the samples are found in the Appendix. Collection was done fully adhering to IRB-approved protocols. Written informed consent was provided by all participants. Fibrosis stage assessment of the liver was done with a METAVIR score [18]. This is summarized in Table 1. Note that some inconclusive scores were given by the pathologist. 15 of the 110 METAVIR scores were provided by the pathologist with two neighboring values instead of one, indicating that neither scores could be definitively chosen over the other.

Immunofluorescent staining was applied to the biopsies. As biopsies were pooled from different sources, multiple scanners were used for scanning the biopsies, with more details found in the Appendix. All were scanned at 20X magnification. CellProfiler and HALO were used for cell segmentation, identifying each cell, their two-dimensional (x and y) coordinates and the cell type according to the aforementioned stains. The two-dimensional window of interest was chosen to be the rectangular window drawn around the most extreme x and y coordinates. Each cell was identified as one of three cell types: HBsAg-negative hepatocyte, HBsAg-positive hepatocyte and immune cell. To classify the cells into these cell types, antibodies against HBsAg and CD45 were applied, as well as a nuclear dye. Cells that showed a level of fluorescence related to CD45 that was higher or lower than a certain threshold were considered respectively immune cells or hepatocytes. Within this latter class, cells that showed a level of fluorescence related to HBsAg higher or lower than a certain threshold were considered respectively HBsAg positive and HBsAg negative hepatocytes. Thresholds of both CD45 and HBsAg were determined using intensity distributions found in control samples.

2.2 Spatial summary statistics

Spatial measures are needed to quantify the spatial micro-environment, often referred to as the spatial heterogeneity amongst different types of biological cells. Biological cells are hereafter referred to as points. The arrangement of said points then form a spatial point pattern. Different summaries of spatial statistics start by discretizing the spatial point pattern data by dividing the window into grid cells of the same size. The number of points corresponding to each cell type in any given grid cell is assumed to be distributed by a Poisson distribution. Certain biopsies were shaped such that the surrounding window of interest contained a substantial area not of importance to the study. To combat this issue, empty grid cells were subsequently discarded. The size of each grid cell was chosen for each sample separately so that the mean amount of points in the non-empty grid cells approaches a pre-specified value chosen as 5, 10, 15, 20, 25, 50, 100, 200, 500 or 1000. To choose the size that best approaches this value, a simple algorithm was constructed, defined as Algorithm 1 in the Appendix.

In the following formulations, i denotes the index of a grid cell, c_i^k denotes the number of points of type k in grid cell i, p_i^k denotes the ratio of c_i^k to the total number of points in grid cell i, and n^k denotes the total number of points of type k.

2.2.1 Morisita-Horn index

The Morisita-Horn index (MHI) [19] is a measure calculating the similarity - or differently put, heterogeneity - of two cell types across the whole grid. The formula of the index is defined as

$$M = \frac{2\sum_i \frac{c_i^k}{n^k}\frac{c_i^l}{n^l}}{\sum_i(\frac{c_i^k}{n^k})^2 + \sum_i(\frac{c_i^l}{n^l})^2}.$$

This index is thus calculated for the entirety of the sample and ranges from 0 to 1, indicating complete dissimilarity or homogeneity at value 0, and complete similarity or heterogeneity at value 1. Specifically, in the extreme case where some grid cells only contain points of a certain type, and the other grid cells only contain points of the other type, this would point to homogeneity and would result in a value of 0. In the extreme case, then, where all grid cells contain an equal portion of points of either type, this would point to heterogeneity and would result in a value of 1. Since this is a pairwise statistic, this was calculated for each pairwise combination of the three cell types studied in this paper.

2.2.2 Shannon diversity index

The Shannon diversity index (SDI), which is analogous to the Shannon entropy in information theory [20], is an index measuring diversity per grid cell i and is formulated as follows:

$$S_i = -\sum_l^3 p_i^l \log(p_i^l).$$

Whereas the Morisita-Horn index does pairwise comparisons over the entirety of the sample, the Shannon index compares all point types at once but only for one grid cell at a time. The Shannon index ranges from 0 to $\log(l)$ with l indicating the amount of distinct point types, which in the context of this paper results in $\log(3) \approx 1.1$. Similar to the Morisita-Horn index, low values indicate homogeneity and high values indicate heterogeneity. Specifically, in the extreme case a grid only contains points of one type, this would point to homogeneity and would result in a value of 0. In the extreme case, then, where a grid cell contains an equal portion of points across all types, this would point to heterogeneity and would result in a value of $\log(3)$. Both the mean and variance of Shannon indices across all grid cells are used to summarize the information about the entirety of the sample.

Note that both the Morisita-Horn index and the summary measures of Shannon diversity index give a quantitative measure of cell type heterogeneity found in the entirety of the sample, which is why they are referred to as global spatial summary statistics or henceforth global covariates for short.

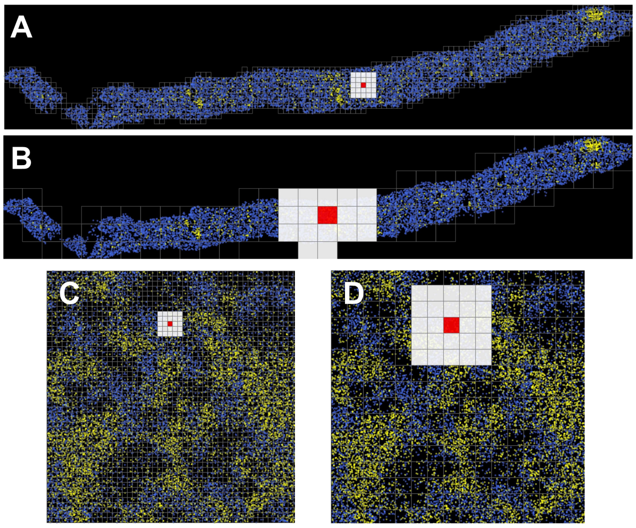

Figure 1: Visualization of different grid choices and corresponding neighborhood structures. (A) and (B) show a sample of the real data with only two cell types shown, with a grid overlay for an average of 20 and 200 point cells per grid cell respectively. (C) and (D) show a simulated sample from simulation type 4 with only two cell types shown, with a grid overlay for an average of 20 and 200 point cells per grid cell respectively.

2.2.3 Getis-Ord hotspot colocalization

The Getis-Ord hotspot analysis [21] is often used in ecology and more recently in histology. It is designed to detect significant clusters or hotspots in the distributions of types. In addition to measuring the significant spots of types individually, within the context of constructing indicators for heterogeneity between types, our aim is to combine significant spots of different types and to measure their frequency. In its algorithm, for any point cell type, in each grid cell a z-score is calculated by comparing the number of points in the grid cells and its neighboring cells with the expected number of cells. The expected number is based on the average number of cells per grid cell. A significantly high and low z-score were defined respectively as a hotspot or coldspot. Formally, in grid cell i, the z-score for point type k is defined as:

$$z = \frac{\sum_{j=1}^{m} w_{i,j} c_j^k - \overline{c}^k \sum_{j=1}^{m} w_{i,j}}{SU},$$

with normalizing factors

$$S = \sqrt{\frac{\sum_{j=1}^{m} (c_j^k)^2}{N} - (\overline{c}^k)^2},$$

$$U = \sqrt{\frac{N \sum_{j=1}^{m} w_{i,j}^2 - (\sum_{j=1}^{m} w_{i,j})^2}{N-1}},$$

where $w_{i,j}$ is the binary indicator whether grid cell i and j are neighbors, $\overline{c}^k$ is the mean of the points in the grid cells and N is the total amount of grid cells.

Different summary measures of the hotspot analysis are calculated. First, for each cell type separately, the proportion of significant hotspots (and coldspots) across all the grid cells is calculated. Secondly, the colocalization of two cell types is summarizes as the proportion of grid cells that showed significance for both types. This results in different combinations of significant spots (hotspots of both types, coldspots of both types or hotspot of one and coldspot of the other). Both a second order and third order neighborhood structure were considered. A second order neighborhood structure considers horizontally, vertically and diagonally adjacent cells as neighbors, and subsequently adds adjacent cells of these neighbors to the neighborhood. A third order neighborhood structure then adds a subsequent adjacency layer.

All statistics derived from the Getis-Ord statistics summarizes local information to a sample wide level, which is why they are referred to as local spatial summary statistics or henceforth local covariates for short.

2.3 Penalized ordinal regression

All spatial and nonspatial (i.e., cell type as percentage of total number of cells) statistics were subsequently used as predictors in an ordinal regression model. The probabilities $\pi_{im} = P(Y_i = m)$ that a biopsy i ($i = 1, \ldots, N$) is classified as from a patient with fibrosis stage m ($m = 0, \ldots, 4$) is modeled. A forward cumulative probability logit model was used. The associated cumulative probabilities of a score less than m is equal to $P(Y_i \leq m) = \sum_{l=1}^{m} \pi_{il}$, and is modeled as function of the P covariate(s) x_i as

$$\text{logit}(P(Y_i \leq m|x_i)) = \log\left(\frac{P(Y_i \leq m|x_i)}{P(Y_i > m|x_i)}\right) = \alpha_m + x_i^T \beta$$

where α_m are level-specific intercepts and β is the common slope corresponding to the P covariates. A Least Absolute Shrinkage and Selection Operator (LASSO) type penalization was added by adding a term directly related to the magnitude of covariate weights to the likelihood function to be minimized, inhibiting covariate influence. The formula to be minimized function is then

$$M = -\frac{l}{N} + \lambda \sum_{j=1}^{P} |\beta_j|$$

where l is the loglikelihood of the cumulative probability logit model

$$l = \sum_{i=1}^{N} \sum_{m=0}^{4} I(Y_i = m) \log(\pi_{im})$$

and λ is a tuning parameter. Fibrosis stages with two neighboring values were accounted for by including both possible values once into the model, but with only half the importance given in the likelihood calculations.

The OrdinalNet package [22] was used in R. The OrdinalNet package uses a coordinate descent algorithm, providing an initial output of a selection of models ranging from high to low λ and thus low to high covariate usage, where for each model the loglikelihood, Akaike information criterion (AIC) and Bayesian information criterion (BIC) were provided. Due to experiencing excessive parameter shrinking yielding non-informative models when choosing to minimize the loglikelihood, only AIC and BIC minimized models were considered in the following analyses.

The performance of different model subtypes were investigated: a model including all covariates (M1), a model with only local covariates (M2), with only global and nonspatial covariates (M3) and a model with only local and global covariates (M4).

2.4 Simulated data

To validate the proposed methods and illustrate the influence of grid cell size on the measured variables and model outcome, a simulation study was done. Four datasets were created with 100 patients per dataset. Similar to the data, per patient, a disease stage score was assigned as a categorical outcome ranging from 0 to 4 (i.e. five categories). A point pattern with three different types of cells was then simulated for each patient. Whereas the three different types of cells in the original data include HBsAg-positive, HBsAg-negative hepatocytes and immune cells, the three types of cells in the simulated data are more simply named type 1, type 2 and type 3. The settings with which this point pattern was simulated were dependent on the disease stage assigned to the patient, therefore creating a link between features found in the point pattern and the disease stage. Specifically, type 1 and type 2 cells had different levels of clustering depending on the disease stage of the patient. The techniques detailed above were then used on these datasets in addition to the original dataset. The complete simulation outline and pseudocode can be found in the Appendix (Algorithm 2).

Figure 1 shows two grids of different cell sizes on top of a sample of the real data as well as a sample of the simulated data of the fourth type. Grid choices shown are 20 and 200 average points per grid cell. Visualization of the neighborhood was done by implementing a second order neighborhood structure. Note the more accurate outlining of the sample shape in the case of a smaller grid size for the real data, as well as the difference in neighborhood definition due to grid cell sizes.

3 Results

3.1 Data analysis

Statistics concerning amount of point cells and amount of grid cells used to discretize the real data per setting are shown in Figure 2. The sample with the least amount of point cells contained 4459 cells, the one with the highest amount of point cells contained 71195 cells.

The predictive quality of our fitted OrdinalNet models was measured by a leave-one-out-cross-validation (LOOCV) process for each combination of model type and grid choice separately. The absolute difference between model predicted fibrosis score and observed fibrosis score being only one or below (hereafter referred to as lag one) was considered the primary measure of predictive quality, rather than an exact prediction. This was chosen due to the aforementioned ambiguity of scoring liver fibrosis. Ratios of number

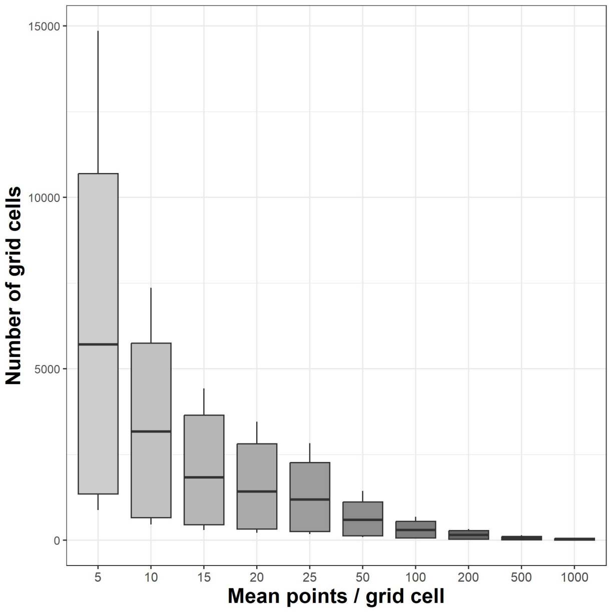

Figure 2: Boxplots for the number of grid cells per sample per chosen grid size.

of samples with lag one prediction and total number of samples are shown in Figure 3, expressed in percentages for all models fitted on the real data and all grid choices. Scores that were deemed ambiguous by the pathologist and thus had two neighboring values were included twice weighted 50% in the calculation of the predictive error, once using the lower and once using the higher score, e.g. an observed score of 0.5 and predicted score of 1 would result in both an absolute difference of zero and of one).

The model with the highest predictive quality is the AIC minimized model with second order neighborhood structure where all covariates were included with a grid choice of an estimated average of 200 point cells per grid cell. AIC minimized models and models with second order neighborhood structure tend to have a higher quality prediction in most covariate inclusion scenarios and grid choices. The best model can predict 84% of the samples correctly with one lag step. In order to understand how the grid size impacts each of the spatial summary measures, we look at the coefficient of variation, defined as the ratio of the standard deviation to the mean, between covariate values for different grid sizes. Figure 4 shows the coefficient of variation for each of the 110 samples. This coefficient would therefore indicate per sample how much the value of a given covariate changes when the chosen grid cell size changes. Nonspatial covariates are left out since grid size could

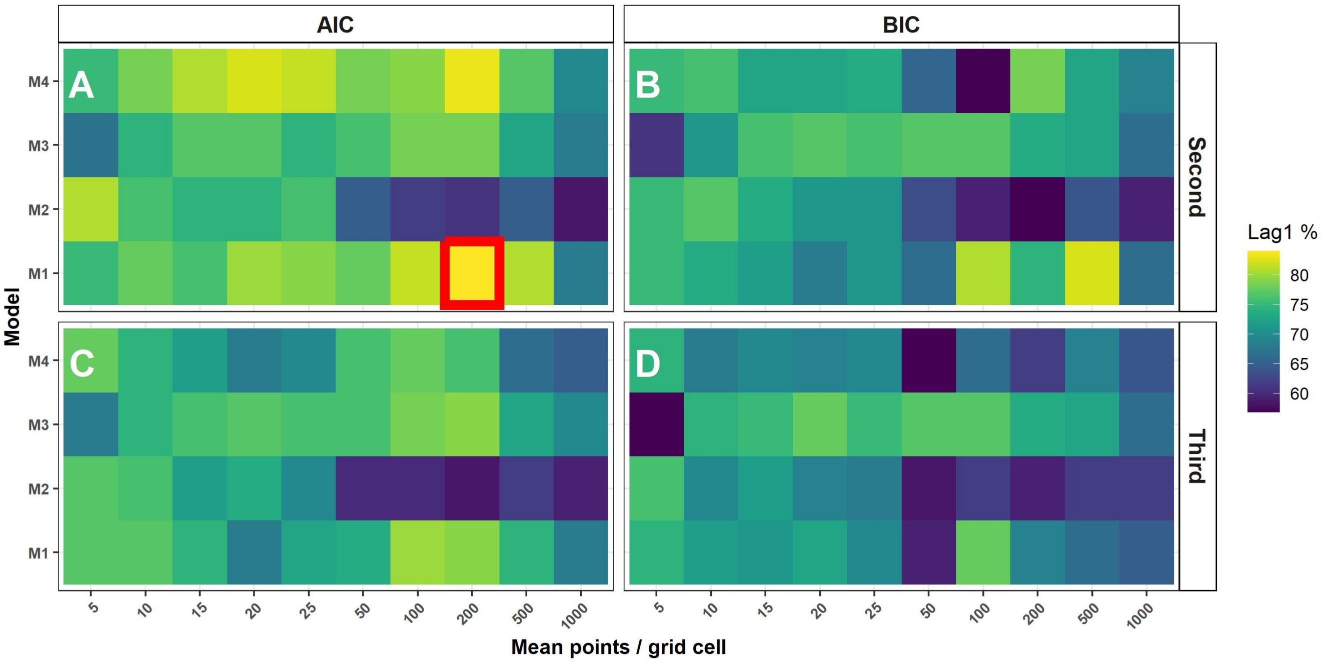

Figure 3: LOOCV lag 1 results (absolute difference between predicted and observed score being one or less) for different model constructions using real data. Plot A and B show results for all models that were respectively AIC and BIC minimized with a second order neighborhood structure. Plot C and D show these for a third order neighborhood structure. M1, M2, M3 and M4 refer to models where all, local, global & nonspatial and local & global covariates were included respectively. The x axis refers to the choice of mean point cells per grid cell as indicator of grid choice. The highest percentage was indicated using a red border.

not affect them. Consistently across samples, the global covariates have a small coefficient of variation between grid sizes. Most local covariates have larger coefficients of variation, some consistently and others inconsistently across samples. This explains why the choice of the grid can influence the predictive performance when using these summary measures as covariates in a prediction model.

Figure 5 further details the predictive performance and the variable importance results of the real data using the AIC minimized models with a second order neighborhood structure. To indicate the importance of a covariate, the absolute value of the coefficient corresponding to said covariate in the OrdinalNet model using all samples was used, since all covariates were standardized prior to their inclusion in the model. The local covariate in-

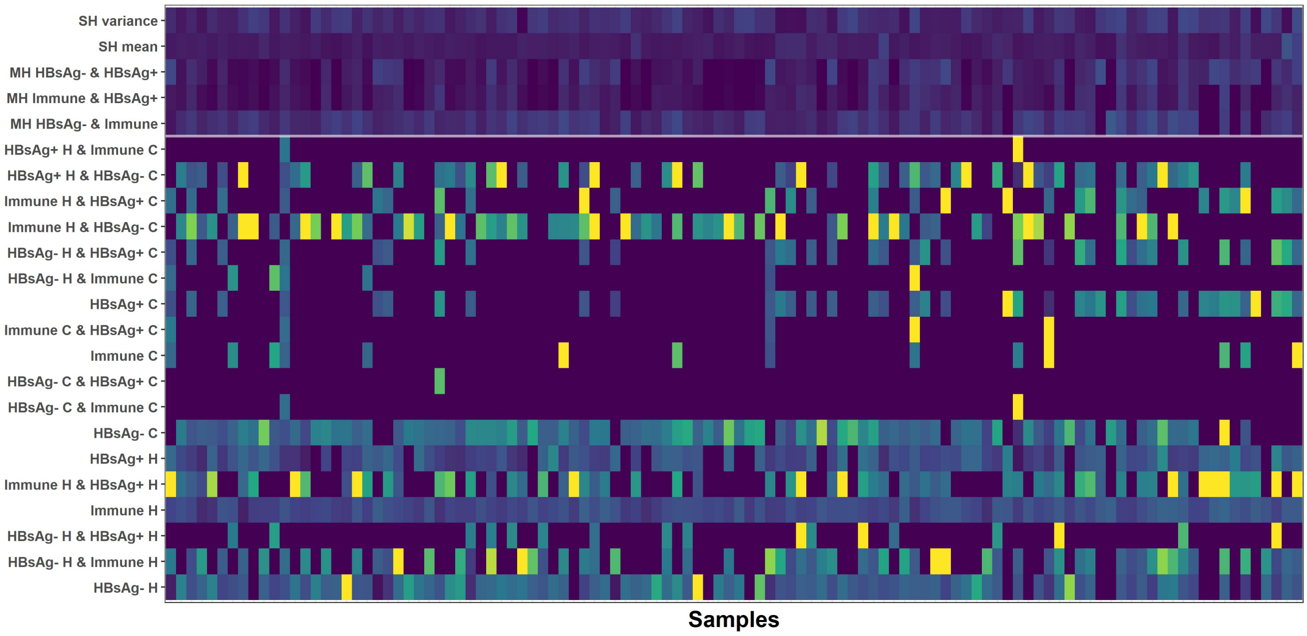

Figure 4: Coefficients of variation between covariate values for different grid sizes per sample. "SH" refers to Shannon index, "MH" refers to Morisita-Horn, "H" and "C" refer to Getis-Ord hotspots and coldspots respectively, and any combination refers to a colocalization of these spots. Coefficient values range from low in dark blue to high in yellow. Samples are sorted from highest to lowest total number of data points.

cluded model has a predictive quality drop-off from grid choice 25 to 50 point cells per grid cell. This drop-off is not as present in the three other covariate inclusion choices. In all models where local covariates are included, local covariates whose coefficients are largest in smaller grid choices such as the Getis-Ord HBsAg-positive coldspots percentage have smaller coefficients in larger grid choices, with the pivotal grid choice coinciding with the aforementioned drop-off point. In M1 and M4, global and if present nonspatial covariate weights become larger for grids choices 20 point cells and above per grid cell.

3.2 Simulation study

Figure 6 shows the performance and the variable importance results of the data collected using the first simulation type for the AIC minimized models with a second order neighborhood structure. Both models where global and nonspatial covariates are simultaneously included have near perfect lag one performance for grid choice below 500 and 1000 point cells per grid cell. The local and global covariates model mirrors this result with slightly worse performing grid choices of 5 and 200 point cells per grid cell. Performance of the local covariates model peaks at grid choice 25 point cells per grid cell and rapidly decays beyond 50. The MHI of type 1 and 2 is almost consistently the most important variable in the models where it was included, with a drop-off in importance for the largest grid choices. The model with only local covariates for most grid choices retains many covariates; most importantly Getis-Ord coldspots of type 1 and a colocalization of type 1 coldspots with type three hotspots. Their importance dwindles in models where other aforementioned important covariates are included. The performance and the variable importance results of the data collected using the second, third and fourth simulation type for the AIC minimized models with a second order neighborhood structure are shown in the Appendix. Some notable differences can be observed. Generally, models using the data of the second simulation perform worse than the other simulations, and the third and fourth simulation show near perfect prediction for models M1, M3 and M4. For all simulations but the fourth, variable importance is highest for the covariate concerning the Morisita-Horn of type 1 and 2, in models where this covariate was included. In contrast, for the fourth simulation, variable importance is much higher for the covariate concerning the percentage of type 1 cells, in models where this covariate was included. Many of the same phenomena found in the results of the first simulation are also present in the results of the remaining simulations. Models generally show a drop in quality beyond grid choice 25 point cells per grid cell. Grid choices 500 and 1000 yield very low percentages for all models, and in the M2 model, all grid choices beyond 25 yield these percentages as well.

4 Discussion

4.1 Real data

Firstly, the 110 real data samples have highly inconsistent shapes and sizes, which is indicated by the highly variable number of total points per sample. As a result, indicated in Figure 2, even within the same choice of grid cell setting, the largest sample can have close to 20 times more grid cells to take into account than the smallest one. Among all models applied to the real data, the best performing is the AIC minimized model using a second order neighborhood structure with a grid cell choice of an average 200 data points

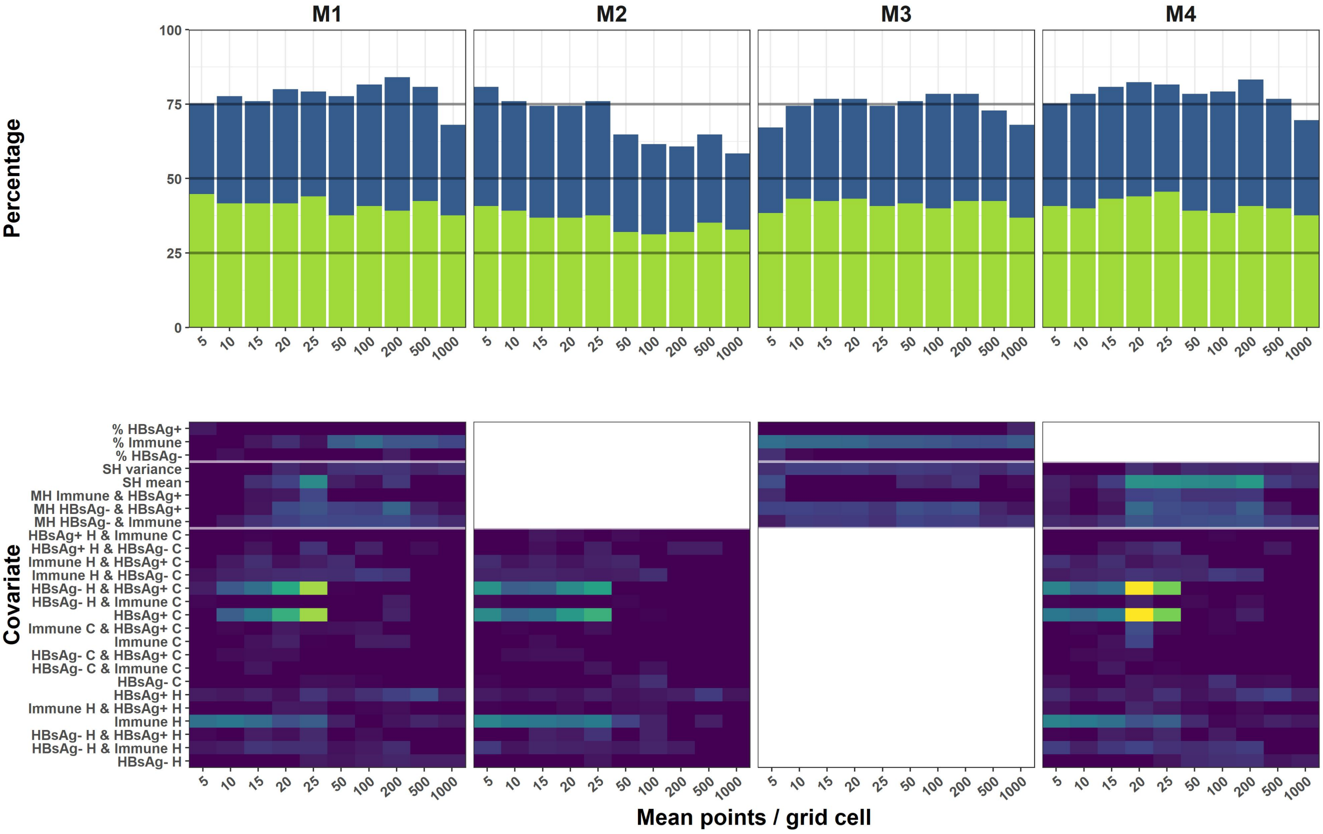

Figure 5: Performance and variable importance of liver data. Columns indicate the results from the model where respectively all (M1), only local (M2), global and nonspatial (M3), and local and global covariates (M4) were included. "SH" refers to Shannon index, "MH" refers to Morisita-Horn, "H" and "C" refer to Getis-Ord hotspots and coldspots respectively, and any combination refers to a colocalization of these spots. Row one shows LOOCV results, with lag zero % in green and lag one % cumulatively in green and blue.

per grid cell. This is an intriguing result, since this grid cell choice no longer provides as much local information as smaller grid cell choices, as can be seen from the predictive quality of any M2 model. Note that in most cases, AIC minimized models performed better than their BIC minimized counterparts. Models using covariates with second order neighborhood structure performed better in comparison to their counterparts using a third order neighborhood structure. Due to this, the summarizing figures are only based on the AIC minimized model with second order neighborhood structure. Looking only at

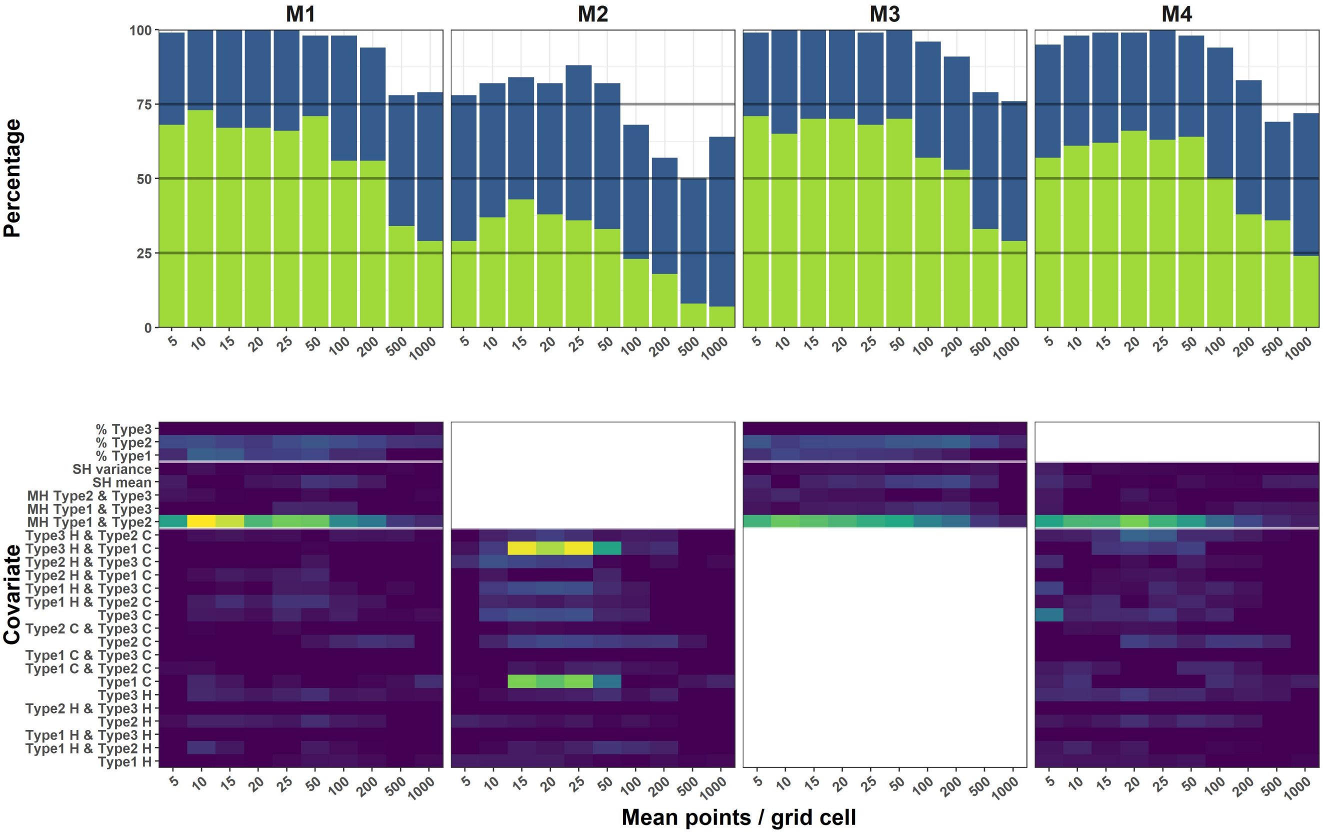

Figure 6: Performance and variable importance of simulation 1. Fibrosis score dependent on b, the maximum distance in either dimension that a parent point of type 2 that is colocalized with type 1 is simulated from a parent point of type 1. Columns indicate the results from the model where respectively all (M1), only local (M2), global and nonspatial (M3), and local and global covariates (M4) were included. "SH" refers to Shannon index, "MH" refers to Morisita-Horn, "H" and "C" refer to Getis-Ord hotspots and coldspots respectively, and any combination refers to a colocalization of these spots. Row one shows LOOCV results, with lag zero % in green and lag one % cumulatively in green and blue.

grid size influence, then, in Figure 5, there is no conclusive worse performing grid size as expected, with the exception of the decay of predictive quality in the local covariate model as grid size increases. What is noticeable, however, is the clear change in variable importance due to grid size. For smaller grid sizes, local covariates – in particular Getis-Ord hotspots of HBsAg negative cells, hotspots of HBsAg negative cells that collocate with coldspots of HBsAg positive cells and immune hotspots – carry the model, whereas in larger grids their influence dies out and gets taken over by the global and nonspatial covariates -– in particular the Morisita-Horn measures of HBsAg negative and positive and HBsAg negative and immune, as well as the percentage of immune cells relative to the total amount of cells – for larger grid sizes. Local covariates are most influenced by grid changes, as could be seen in Figure 4. Many coefficients of variation between values in different grid sizes for local covariates were quite large in comparison to those for global covariates. Those that had zero coefficients of variation had a value of zero in each grid choice. Therefore, these provided no information regardless. On the other hand, many of the large coefficients of variation were only inflated by having a small mean due to only having one grid size where the covariate value was nonzero. Nonetheless, local covariates, especially ones related to colocalization of different types, are most vulnerable to grid changes. Nonetheless, their volatility is likely the cause of their diminished influence in the models for larger grid sizes. By extension, the relatively higher consistency among global covariates is the reason they gain importance when local covariates can no longer provide predictive information.

4.2 Simulated data

Figure 6 as well as Figure S1, S2 and S3 (found in the Appendix) also showcase grid size influence, but for the simulated data. Despite slight differences, the local covariates models in all simulation types point to the same general rule as the real data sets: smaller grid sizes provide better local information than larger grid sizes. The diminished influence of local covariates in larger grid size models confirms this. In other model types, however, it is clear that the type of simulation controls the variable importance. In the first three simulation types, in the models where global covariates are included, the MHI of type 1 and 2 explicitly dominates any other included covariate and loses influence in the two to three largest grid sizes. In the fourth simulation type, where included, the proportion of type 2 cells dominates other covariates. The conclusions regarding the respectively waning and growing influence of local and other covariates as grid size increases previously drawn from the variable importance plot for the real data can not be drawn from these simulations.

The method outlined in this paper and in the paper of Agten et al. [17], with the previously described algorithm, always guarantee that each grid cell has on average approximately the chosen number of data points (e.g. 20 number of points per grid cell). . Any possible change in tissue (e.g. tissue size difference, different tissue type such as tumors, tissue from different organs or organisms, temporal differences in the same tissue such as hypertrophy and hyperplasia) might affect the number of grid cells analyzed, but performance of heterogeneity indicators will be invariant due to the approximately invariant average number of data points per grid cell. Therefore, we argue that the method used in Agten et al. and this paper would be well-suited in contexts of other types of tissues.

4.3 Limitations

Firstly, in this study grid choice was defined by the average point cells per grid cell. The simple algorithm used to find the best approximate length of the side of a grid cell that would yield a specified average point cells per grid cell in a given sample was imperfect as can be seen in Figure 2. Thusly, inconsistencies, as small as they might seem, could have influenced subsequent calculations. Secondly, all data that was simulated, was simulated in a unit square window. This is in stark contrast to the shapes of the actual data samples, which due to the nature of biopsies were often stretched, thin and irregular. Biases stemming from these shapes such as having the grid size choice influence the edge structure are therefore not present in the simulated data. Thirdly, the four simulation settings heavily influenced a single covariate as most important for prediction each time, as was mentioned in the discussion. This was never a local covariate. This could be interpreted to mean that the use of local covariates is generally ineffective, though it is more likely a shortcoming in the construction of these simulations. The local covariate importance found in smaller grid sizes using the real data supports the latter interpretation. Fourthly, as previously mentioned, the primary response variable of interest, namely the fibrosis score of the liver tissue, suffered from a level of ambiguity. Some fibrosis scores were unable to be marked as one value and were therefore chosen by the pathologist to lie between two neighboring values. Moreover, one study showed that, despite inter-observer agreement between the assessment of fibrosis score in chronic viral hepatitis being shown to be rather good (Cohen's kappa equal to 0.59), inexperience and lack of consensus reading might cause agreement to dwindle [23]. As scoring of the entire cohort was done by multiple different pathologists, potential subjectivity in scoring as well as uncertainty to define a definite score could have potentially influenced the analyses in this paper. Finally, the predictive value of the heterogeneity measures was determined by a LASSO type model. One study showed that variable selection using LASSO for data with low signal to noise ratio (which is influenced by among others number of data points and number of predictors, and can be assumed to be low in our context due to the high number of data points relative to the number of predictors), outperforms some other selection methods [24]. However, other studies have been more critical of LASSO as a variable selection technique [25]. More advanced methods might be better suited to avoid inaccurate assignment of variable importance in future studies.

5 Conclusion

In both real and simulated data, the model subtype where only local covariates were included had a mostly downward trend in predictive ability as grid cell size increased. This trend was also reflected in the real data in the dwindling importance of the local covariates due to a lack of information at larger grid sizes. Even models containing all variables, only global and nonspatial covariates, and only local and global covariates had a dropoff in predictive quality in the largest choice(s) of grid size. While for local measures a smaller grid-size is outperforming, global measures have a better performance with medium-sized grids. Thus, across all data and models, trends are not unambiguously apparent, and there is no 'one-size-fits-all' solution. This study aimed to find spatial resolutions that would be optimal on which to calculate heterogeneity indices that in turn would be successfully predictive of disease staging. However, across all data and models, one such resolution cannot be determined in this context. Furthermore, trends in predictive power of these heterogeneity measures are not unambiguously apparent and consistent. Despite this limitation, it is clearly shown that local spatial measures of heterogeneity perform better with more detailed resolutions, while global covariates are more stable with mild resolutions. As such, there is no 'one-size-fits-all' solution, and conclusions depend on the measure of heterogeneity used. This should be taken into account in future studies investigating heterogeneity as a biomarker in histological samples.

References

[1] P. A. Kenny, G. Y. Lee, M. J. Bissell, Targeting the tumor microenvironment, Frontiers in bioscience: a journal and virtual library 12 (2007) 3468. doi:10.2741/2327.

[2] L. Cassetta, J. W. Pollard, Targeting macrophages: therapeutic approaches in cancer, Nature reviews Drug discovery 17 (12) (2018) 887–904. doi:10.1038/nrd.2018.169.

[3] R. Natrajan, H. Sailem, F. K. Mardakheh, M. Arias Garcia, C. J. Tape, M. Dowsett, C. Bakal, Y. Yuan, Microenvironmental heterogeneity parallels breast cancer progression: a histology–genomic integration analysis, PLoS medicine 13 (2) (2016) e1001961. doi:10.1371/journal.pmed.1001961.

[4] Z. Kikvidze, M. Ohsawa, Richness of colchic vegetation: comparison between refugia of south-western and east asia, BMC ecology 1 (2001) 1–10. doi:10.1186/1472-6785-1-6.

[5] L. M. Gomiero, F. M. Braga, Ichthyofauna diversity in a protected area in the state of são paulo, southeastern brazil, Brazilian Journal of Biology 66 (2006) 75–83. doi:10.1590/s1519-69842006000100010.

[6] C. C. Maley, K. Koelble, R. Natrajan, A. Aktipis, Y. Yuan, An ecological measure of immune-cancer colocalization as a prognostic factor for breast cancer, Breast Cancer Research 17 (1) (2015) 1–13. doi:10.1186/s13058-015-0638-4.

[7] D. Zhang, Y. Dong, Y. Zhang, X. Su, T. Chen, Y. Zhang, B. Wu, G. Xu, Spatial distribution and correlation of adipocytes and mast cells in superficial fascia in rats, Histochemistry and Cell Biology 152 (2019) 439–451. doi:10.1007/s00418-019-01812-5.

[8] N. Malleson, W. Steenbeek, M. A. Andresen, Identifying the appropriate spatial resolution for the analysis of crime patterns, PloS one 14 (6) (2019) e0218324. doi:10.1371/journal.pone.0218324.

[9] R. Costanza, Model goodness of fit: a multiple resolution procedure, Ecological modelling 47 (3-4) (1989) 199–215. doi:10.1016/0304-3800(89)90001-x.

[10] M. A. Andresen, Testing for similarity in area-based spatial patterns: A nonparametric monte carlo approach, Applied Geography 29 (3) (2009) 333–345. doi:10.1016/j.apgeog.2008.12.004.

[11] E. Azzalini, R. Barbazza, G. Stanta, G. Giorda, L. Bortot, M. Bartoletti, F. Puglisi, V. Canzonieri, S. Bonin, Histological patterns and intra-tumor heterogeneity as prognostication tools in high grade serous ovarian cancers, Gynecologic Oncology 163 (3) (2021) 498–505. doi:10.1016/j.ygyno.2021.09.012.

[12] Q. Chen, M. Cai, X. Fan, W. Liu, G. Fang, S. Yao, Y. Xu, Q. Li, Y. Zhao, K. Zhao, et al., An artificial intelligence-based ecological index

for prognostic evaluation of colorectal cancer, BMC cancer 23 (1) (2023) 763. doi:10.1186/s12885-023-11289-0.

[13] Y. R. Chung, H. J. Kim, Y. A. Kim, M. S. Chang, K.-T. Hwang, S. Y. Park, Diversity index as a novel prognostic factor in breast cancer, Oncotarget 8 (57) (2017) 97114. doi:10.18632/oncotarget.21371.

[14] S. Nawaz, A. Heindl, K. Koelble, Y. Yuan, Beyond immune density: critical role of spatial heterogeneity in estrogen receptor-negative breast cancer, Modern Pathology 28 (6) (2015) 766–777. doi:10.1038/modpathol.2015.37.

[15] I. P. Nearchou, D. A. Soutar, H. Ueno, D. J. Harrison, O. Arandjelovic, P. D. Caie, A comparison of methods for studying the tumor microenvironment's spatial heterogeneity in digital pathology specimens, Journal of Pathology Informatics 12 (1) (2021) 6. doi:10.4103/jpi.jpi$_2$6$_2$0.

[16] F. Sobhani, S. Muralidhar, A. Hamidinekoo, A. H. Hall, L. M. King, J. R. Marks, C. Maley, H. M. Horlings, E. S. Hwang, Y. Yuan, Spatial interplay of tissue hypoxia and t-cell regulation in ductal carcinoma in situ, npj Breast Cancer 8 (1) (2022) 105. doi:10.1038/s41523-022-00419-9.

[17] A. Agten, A. Blázquez-Moreno, M. Crabbe, M. Tuefferd, H. Goehlmann, H. Geys, C.-Y. Peng, J. Claes, T. Neyens, C. Faes, Measures of spatial heterogeneity in the liver tissue micro-environment as predictive factors for fibrosis score, Computers in Biology and Medicine (2023) 107382doi:10.1016/j.compbiomed.2023.107382.

[18] Z. D. Goodman, Grading and staging systems for inflammation and fibrosis in chronic liver diseases, Journal of hepatology 47 (4) (2007) 598–607. doi:10.1016/j.jhep.2007.07.006.

[19] H. S. Horn, Measurement of" overlap" in comparative ecological studies, The American Naturalist 100 (914) (1966) 419–424. doi:10.1086/282436.

[20] C. E. Shannon, A mathematical theory of communication, The Bell system technical journal 27 (3) (1948) 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x.

[21] A. Getis, J. K. Ord, The analysis of spatial association by use of distance statistics, Geographical analysis 24 (3) (1992) 189–206. doi:10.1111/j.1538-4632.1992.tb00261.x.

[22] M. J. Wurm, P. J. Rathouz, B. M. Hanlon, Regularized ordinal regression and the ordinalnet r package, Journal of Statistical Software 99 (6) (2021). doi:10.18637/jss.v099.i06.

[23] M.-C. Rousselet, S. Michalak, F. Dupré, A. Croué, P. Bedossa, J.-P. Saint-André, P. Calès, Sources of variability in histological scoring of chronic viral hepatitis, Hepatology 41 (2) (2005) 257–264. doi:10.1002/hep.20535.

[24] T. Hastie, R. Tibshirani, R. Tibshirani, Best subset, forward stepwise or lasso? analysis and recommendations based on extensive comparisons, Statistical Science 35 (4) (2020). doi:10.1214/19-sts733.

[25] L. Freijeiro-González, M. Febrero-Bande, W. González-Manteiga, A critical review of lasso and its derivatives for variable selection under dependence among covariates, International Statistical Review 90 (1) (2021) 118–145. doi:10.1111/insr.12469.