Image and Data Integration Using Spatial Point Processes

Kasthuri Kannan, PhD
Associate Professor
Translational Molecular Pathology
UT MD Anderson Cancer Center

Who thinks mathematics is awesome?

Banach-Tarski paradox

It is a theorem in mathematics!!

Are we living in the world of matrix?

Contents

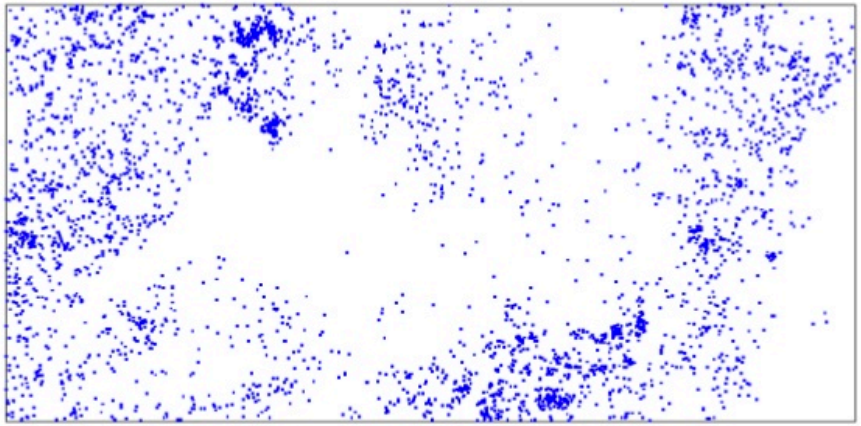
- Introduction to Spatial Point Processes
  - Point process methods, CSR
  - Conditional intensity
  - Gibbs, Strauss and Geyer models
- Omics Data Integration
  - Marked point process
  - Histopathology image segmentation
  - Power of spatial point process modeling
- An Application of Spatial Point Process
  - Case study: Glioblastoma multiforme
  - K and L functions
  - Contact inhibition of locomotion

Introduction to Spatial Point Processes

Motivation

A collection of points in space, R^d, $d >= 2$. The arrangement of points is called a *pattern*.

Example: tropical rainforest



Example : the milky way galaxy

1. Are the points uniformly spread over the survey region?
2. Does the density of points depend on an explanatory variable?
3. Are the points randomly scattered?
4. Is there evidence of clustering?
5. Is the spatial pattern consistent with my scientific hypothesis?
6. How reliable is my statistical analysis?

Adrian Baddeley, Ege Rubak, Rolf Turner. Spatial Point Patterns. Chapman & Hall/CRC

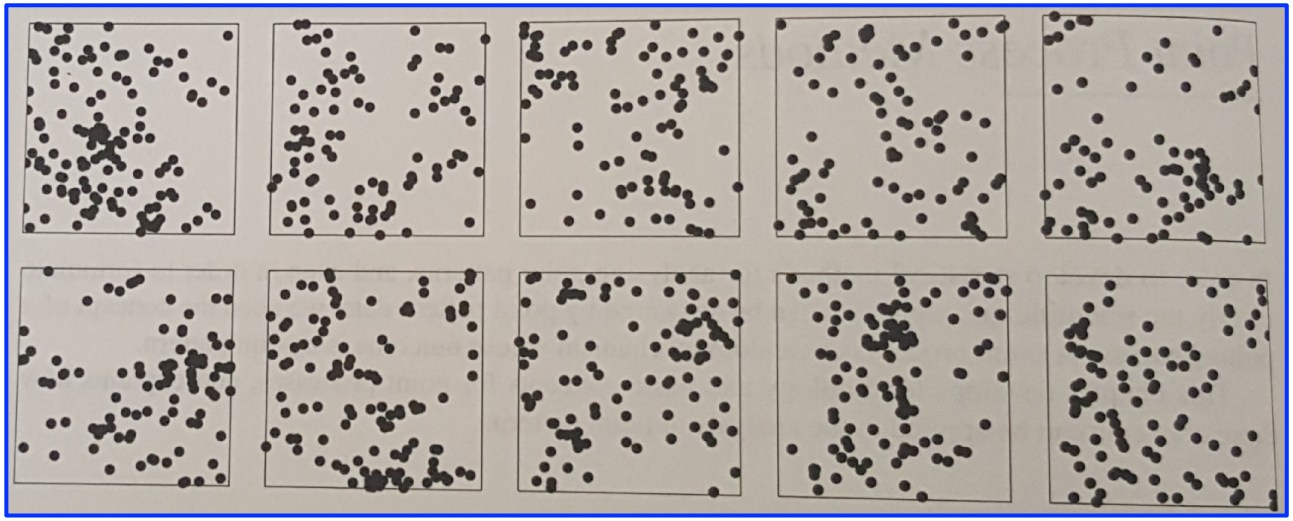
Point Process

Not about points themselves

How the points were generated on the first place?

- Random variable
  - Numerical quantity that is observed/measured
  - “Inherently variable” – measurement errors, sampling variability etc.

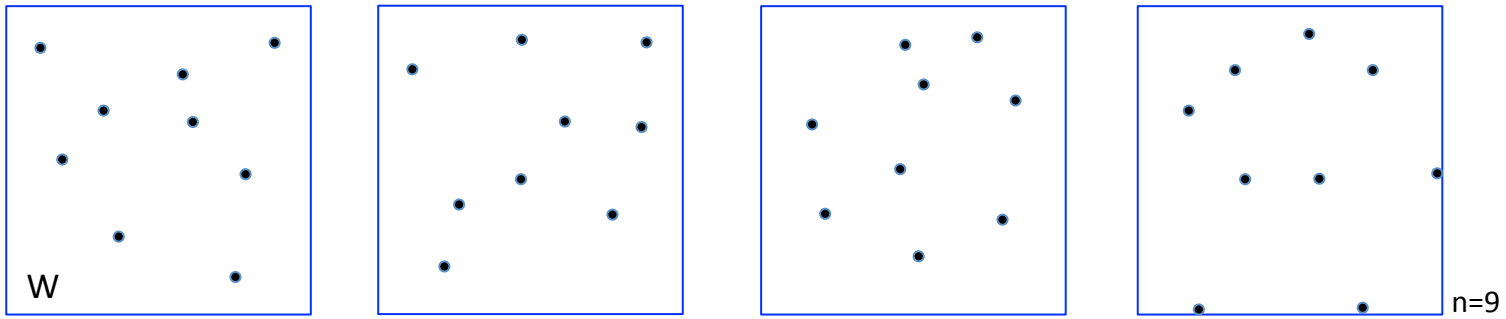
- Point process
  - A random mechanism whose outcome is a point pattern
  - Same process could lead to different point patterns (realizations)



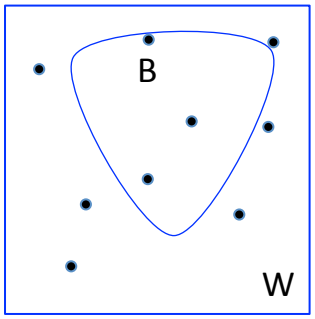
Adrian Baddeley, Ege Rubak, Rolf Turner. Spatial Point Patterns. Chapman & Hall/CRC

The Binomial Process

- Number of points, *n*, is constant and are spread in *independent* random locations:
  $X = \{x_i : i = 1,2,3,\ldots,n\}$
- The points are *uniformly* spread over a region W



For a given region B in W, we need to determine the probability distribution, $n(X \cap B)$



We can model X_i as success/heads if it belongs to B, otherwise failure/tails

$$\mathrm{P}\{n(X \cap B) = k\} = \binom{n}{k} p^k (1-p)^{n-k}$$

where

$$p = \frac{area(B)}{area(W)} \text{ and } k = 1,2,3,\ldots,n$$

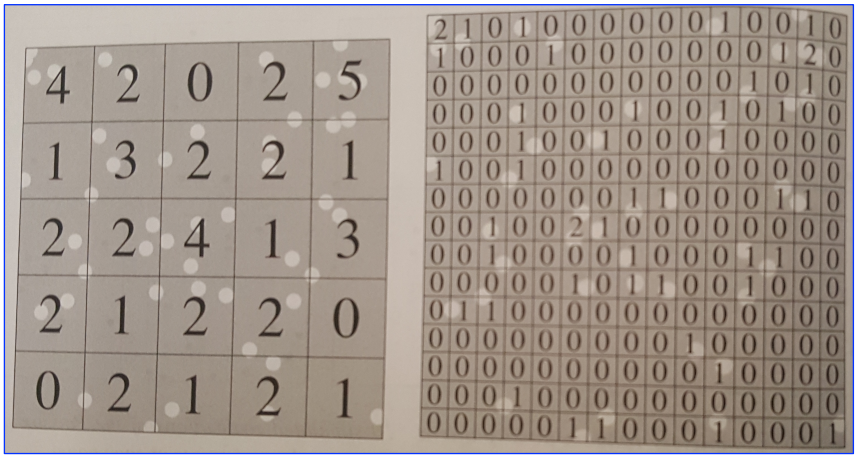
Complete Spatial Randomness (CSR)

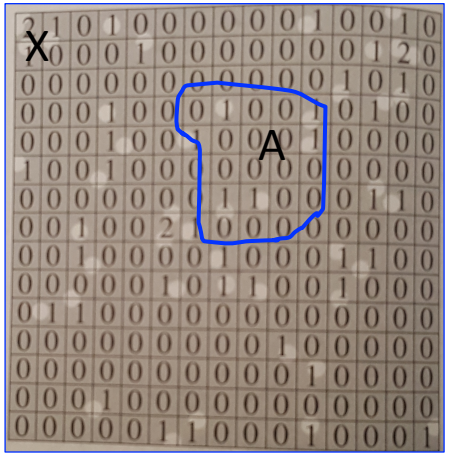
- *Homogeneity*: the points have no preference for any spatial location
- *Independence*: information about the outcome in one region of space has no influence on the outcome in other regions of space
- Binomial process does not have the property of spatial independence.
  - Given *k* points in a region and total *n* points, *n-k* points lie in the complement of the region
- CSR is a realistic model for several phenomena, radioactive decay, rare events, extreme events etc.
- Null model
- What is the probability distribution of CSR?

Homogeneous Poisson Process

Homogeneity: $\mathrm{E}[n(X \cap A)] = \lambda \ \mathrm{area}(A)$

Independence assumption applies to any disjoint regions of space





$n(X \cap A)$ = number of successes in a large number of independent trials, each trial having a small probability of success

Famous result in probability theory, $n(X \cap A)$ has a Poisson distribution,

$$\mathrm{P}\{n(X \cap A) = k\} = e^{-\mu} \frac{\mu^k}{k!} \quad where \ \mu = \lambda \ \mathrm{area}(A)$$

Thus, CSR = Homogeneous Poisson Process

Adrian Baddeley, Ege Rubak, Rolf Turner. Spatial Point Patterns. Chapman & Hall/CRC

Inhomogeneous Poisson Process

Most important model for practical purposes

Modification of the CSR in which the average density of the points is spatially varying

Thus, instead of λ we have $\lambda(u)$ for any spatial location *u* and the expected number of points in a region A is given by, $\int_A \lambda(u)du$

Simulation of inhomogeneous Poisson process

Given $\lambda(u)$, find M such that $\lambda(u) <= M$

Generate a homogeneous Poisson process with intensity M

For each random point x_i that was generated, evaluate the probability $p_i = \lambda(x_i)/M$

Randomly delete or retain the x_i with retention probability p_i, independently of the fate of other points

Resulting simulation is an inhomogeneous Poisson process with intensity $\lambda(u)$

Therefore, given a point pattern we can determine the intensity at various *quadrats*, simulate an inhomogeneous Poisson process and compare with original data set (*Monte Carlo tests*) using appropriate test statistic

Common Parametric Models of Intensity

Nonhomogeneous Poisson process a key model
– several popular techniques are equivalent to fitting a Poisson point process model

Homogeneous intensity: $\lambda(u) = \lambda$

Homogeneous in different regions: $\lambda(u) = \beta_j \text{ if } u \in B_j$

Intensity proportional to baseline: $\lambda(u) = cb(u)$
Ex., each member of the population has constant chance '*c*' of contracting a rare disease

Exponential function of *covariate*: $\lambda(u) = \exp(\alpha + \beta Z(u))$
Ex., earthquake modeling, gold deposits modeling

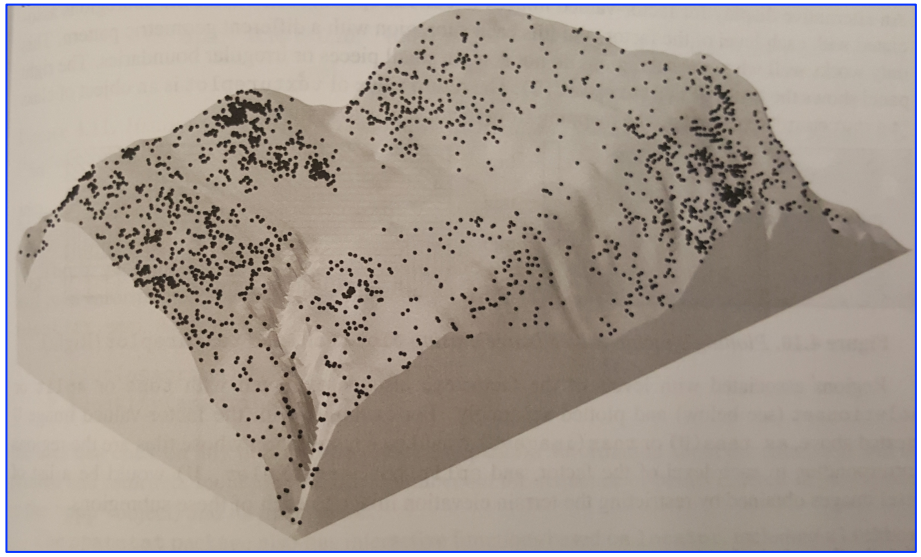
Raised incidence model: $\lambda(u) = b(u)\exp(\alpha + \beta Z(u))$
Ex., spatial epidemiology

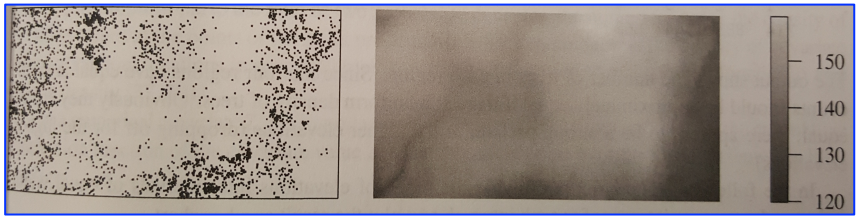
General log-linear model: $\lambda_c(u) = \exp[B(u) + c_1 Z_1(u) + c_2 Z_2(u) + \cdots + c_p Z_p(u)]$

The parameters of these models can be estimated through methods such as maximum likelihood

Covariates

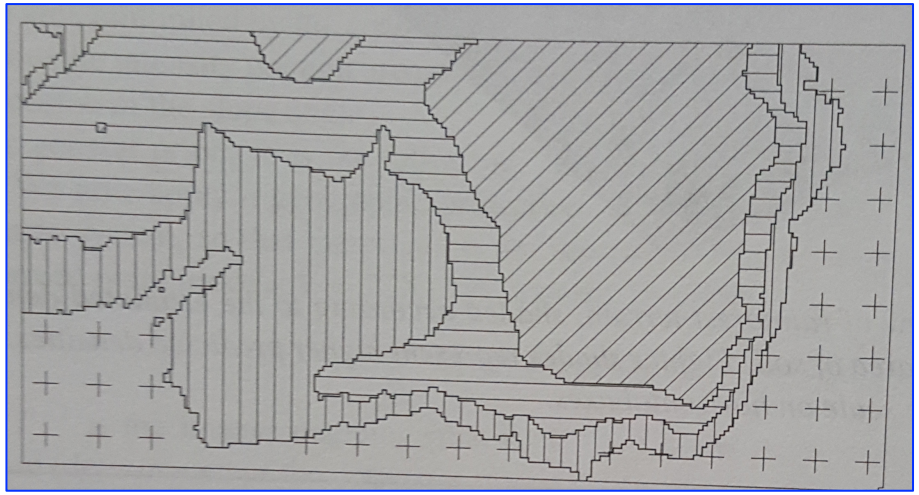
Tropical rain forest terrain with visible Beilschmiedia tree locations (Barro Colorado Island)

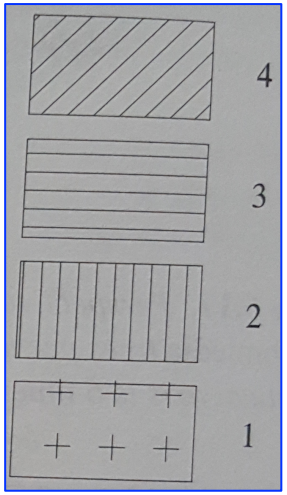




How the intensity of points depend on the values of covariates?

Does Beilschmiedia trees prefer steep or flat terrain?






```
> qb <- quadratcount(bei, tess=V)
> qb
tile
   1    2    3    4
 714  883 1344  663
```

Conclude – Beilschmiedia trees prefer higher elevations

Adrian Baddeley, Ege Rubak, Rolf Turner. Spatial Point Patterns. Chapman & Hall/CRC

Conditional Intensity

- Recall CSR
  - Homogeneity
  - Independence

- Homogeneity can be relaxed
  - Inhomogeneous Poisson process

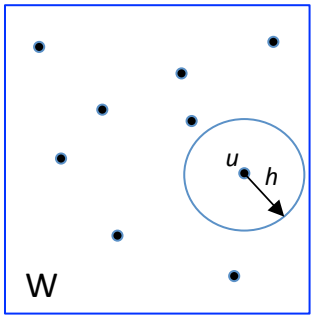
- Notion of independence is too stringent

- Two approaches
  - Make the intensity random (Cox and Cluster processes)
  - Make conditional intensity (Gibbs, Strauss, Geyer processes)

- If the point process exhibits dependence, probability of observing a point *will* depend on the rest of the pattern

$$\lambda(u|X) = \text{intensity at } u \text{ given the rest of the point pattern X}$$

Gibbs Hard Core Process



$$\lambda\left(u\middle|X\right)=\begin{cases}0 & \text{if hard core violated}\\ \beta(u) & \text{if hard core satisfied}\end{cases}$$

- Equilibrium process of spatial birth and death process
- Each small interval of time δt, each cell has a probability *d(t)* δt of undergoing cell death, d(t) is the apoptosis rate
- Cell division takes place in a small region δa, with probability *m(t)* δa δt, where *m(t)* is the mitotic rate
- Provided the dividing cells lies *h* units away, no matter what the initial states are, over long time scales, this process will reach an equilibrium in which any snapshot will be a Gibbs process
- Intensity $\beta=\frac{m(t)}{d(t)}$ with hard core radius *h*

Strauss Process

- Strauss process is a generalization of the Gibbs process - variations of the basic hard core process

- Gibbs hard core process - it is physically impossible for two points to lie closer than *h* distance apart

- Strauss process – *improbable* than impossible
  - introduce an interaction parameter γ and *t(u,h),* number of points less than *h*

$$\lambda\left(u\middle|X\right) = c\ \beta(u)\ \gamma^{t(u,h)}$$

- A collection of cells with many pairs of close cells is much less likely than a collection of cells with only a few close pairs.

- Strauss process is a generalization of the Gibbs hard core process

- Note: $\gamma = 1$ is a Poisson process, $\gamma < 1$ establishes inhibiting phenotype and $\gamma > 1$ can introduce clustering – not allowed

- Probability density cannot be greater than 1

Geyer Saturation Process

- How to introduce clustering?
- Of course, introduce another parameter ☺

$$\lambda\left(u\middle|X\right) = c\ \beta(u)\ \gamma^{\min(s,t(u,h))}$$

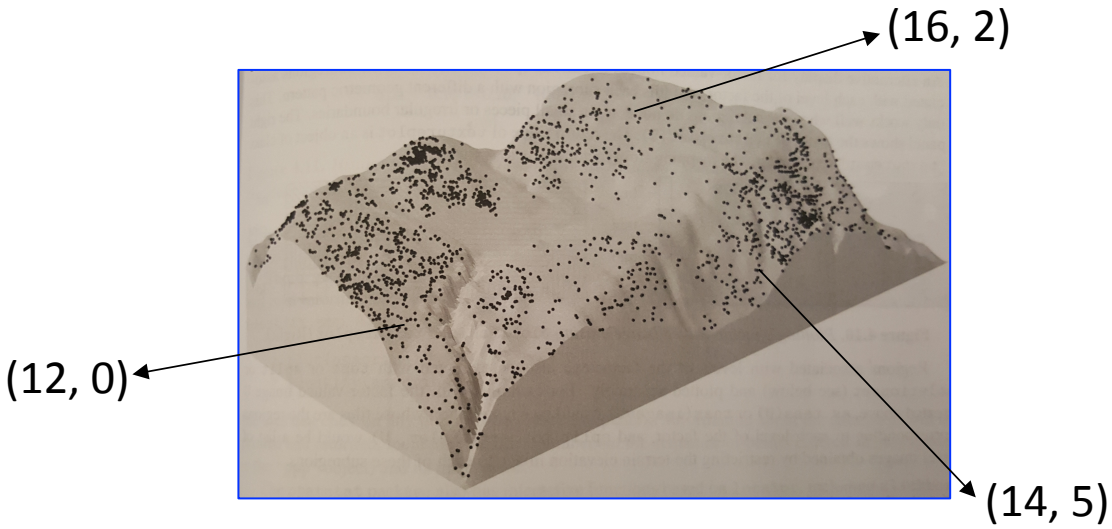
- *s >= 0* − saturation parameter
- Note: *s = 0* is a Poisson process
- $\gamma < 1$ introduces inhibiting phenotype
- $\gamma > 1$ introduces clustering phenotype
- s = ∞, is a Strauss process
- Thus, Geyer process is a generalization of the Gibbs hard core process

Data Integration

Marked Point Processes

- Very powerful framework for biology
  - integrate attributes of cells with the point process representation of cells

- Different from covariates



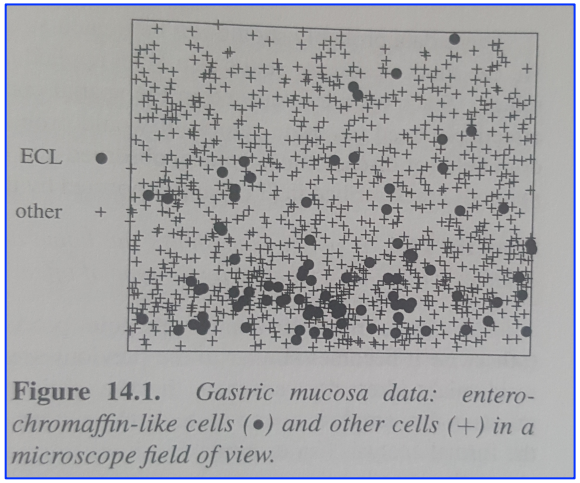
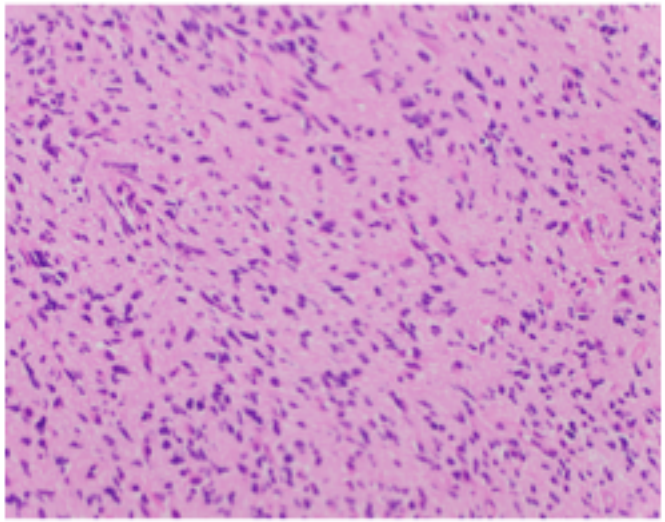


Figure 14.1. *Gastric mucosa data: enterochromaffin-like cells (•) and other cells (+) in a microscope field of view.*

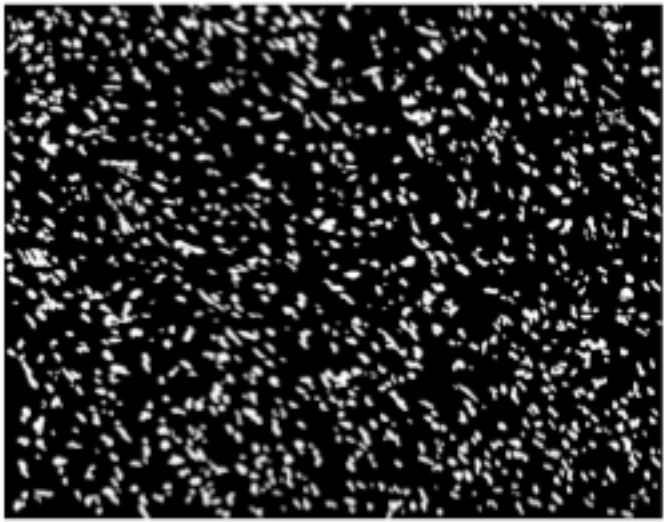
- For example, could be dendritic cells and macrophages

Adrian Baddeley, Ege Rubak, Rolf Turner. Spatial Point Patterns. Chapman & Hall/CRC

Histopathology Image Segmentation



(a) H&E image of astrocytoma



(a) Segmented image of astrocytoma

Segmentation using EBImage[1] and applying the Otsu[2] method for thresholding

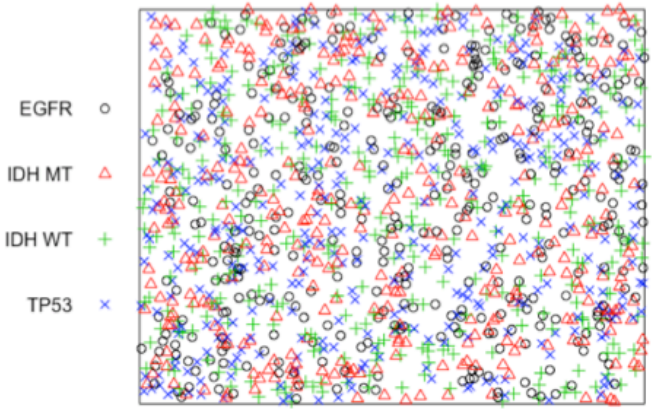
Marked point process representation

1. Pau G, Fuchs F, Sklyar O, Boutros M, Huber W. “EBImage—an R package for image processing with applications to cellular phenotypes.” In: Bioinformatics 26(7) (2010), 979–981.
2. N. Otsu. “A threshold selection method from gray-level histogram.” In: IEEE Transactions on System Man Cybernetics 9(1) (1979), pp. 62–66.

Marked Point Process Representation

- Use centroids from the segmented image to create a point process object corresponding to cell locations
- Use normal distribution to generate marks (lack of experimental data)
- **In reality, one can use RNAi, FISH, IHC, single cell, RNA, Exome, MALDI-MSI, genetics, epigenetics or any omics data sets to mark the cells!!!**

(a) Point process representation of segmented image



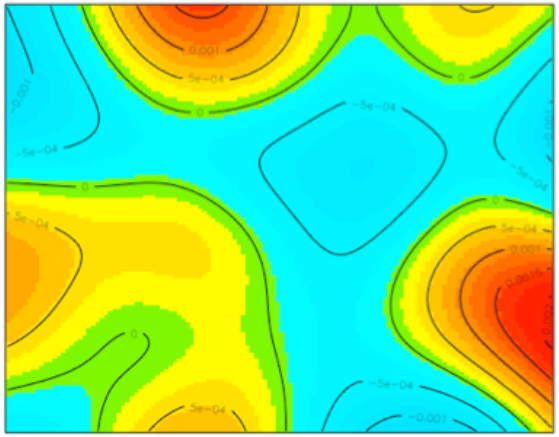
(b) Marked point process representation of segmented image with four marks

Estimating the Density of Marks

- (Density Estimation) What is the tumor density for a specific mark?
- Fit model using Strauss process since the cells are diffuse and infiltrating
- residual = (observed) − (fitted)

```
Nonstationary Strauss process

Log trend:  ~x + y + I(x^2) + I(x * y) + I(y^2)

Fitted trend coefficients:
  (Intercept)            x            y       I(x^2)      I(x * y)       I(y^2)
-9.032894e+00 -3.516460e-04 -6.670171e-04  1.128851e-07  4.580880e-08  3.086753e-07

Interaction distance:   10
Fitted interaction parameter gamma:      1e-06

Relevant coefficients:
Interaction
  -13.86533
```

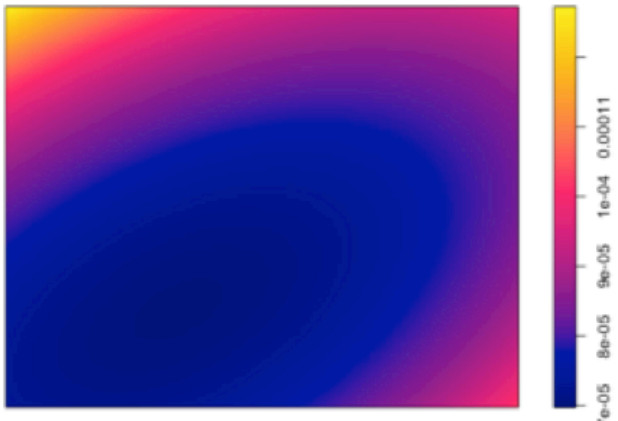
(a) Strauss process fitting for the point pattern



(b) Residual plot for the fit

How Cell Proliferation Varies Across Different Marks?

- Segregation is the estimation of the density of a point pattern with various marks
- Here it denotes the risk of proliferation of various mutation types
- One of central questions in point pattern analysis is to identify regions where a particular mark predominates



(a) Proliferation of IDH mutant cells

(b) Segregation of cells of various marks

- We can compare proliferation patterns across different experimental conditions

How Fast Cell Does Proliferation Happen for a Particular Mark?

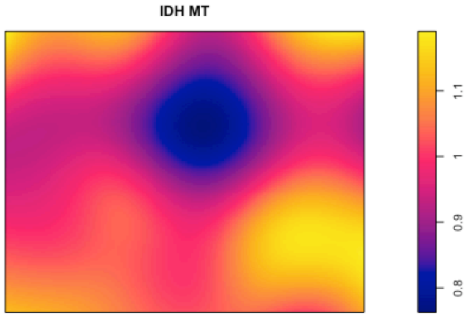
- (Relative Risk Estimation) Do cells of a particular mutation type/mark proliferate faster than any other type?

- Relative measure of proliferation is given by relative probability of different types of marks

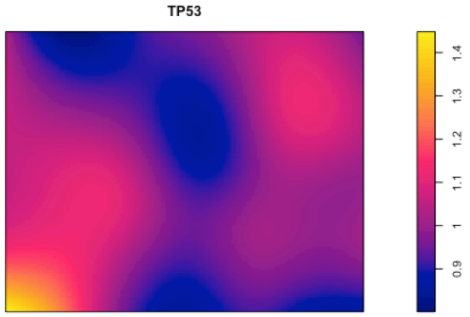
- Epidemiological settings have case-control studies. Relative risk is given by:

$$\rho(u) = \frac{\lambda_D(u)}{\lambda_C(u)}$$

- Similarly, $\rho_{m,k} = \frac{\beta_m(u)}{\beta_k(u)}$



(a) Relative risk of proliferation of IDH mutant cells with respect to IDH wild type cells



(b) Relative risk of proliferation of TP53 mutant cells with respect to IDH wild type cells

What Point Process Does Tumor Heterogeneity Follow?

- Gibbs and Strauss's models are applicable where cell-to-cell 'repulsion' interactions define heterogeneity like in diffuse gliomas

- In tumors of epithelial origin, tumor micro-environment support clustering of cells in a process (*nesting*)

- Cox, Neyman-Scott, Matérn, Thomas, Gauss-Poisson, Cauchy and several other processes can be applied and compared

- Cluster point processes are natural to apply and compare in the setting of clonal evolution
  - model parent offspring associations, mimic the process of cell division

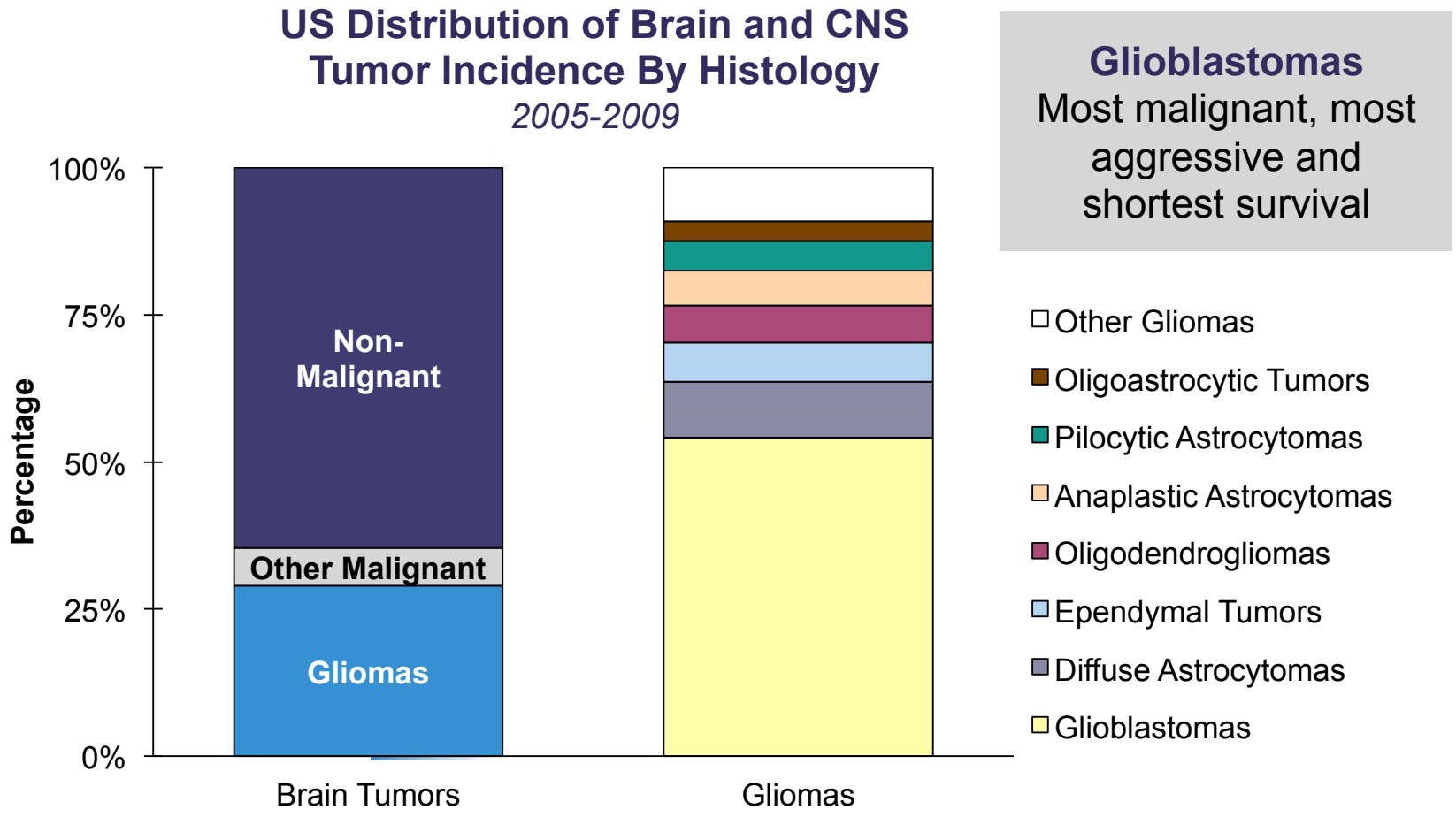
- No such comparative analysis has been performed even purely on histopathology images
  - i.e., even ignoring marked point process representation

- However, comparing control-experiment setup requires just an implementation of any of these methods

An Application of Spatial Point Process

Glioblastoma Multiforme

By annual incidence, glioblastomas are the most common type of malignant brain tumor.

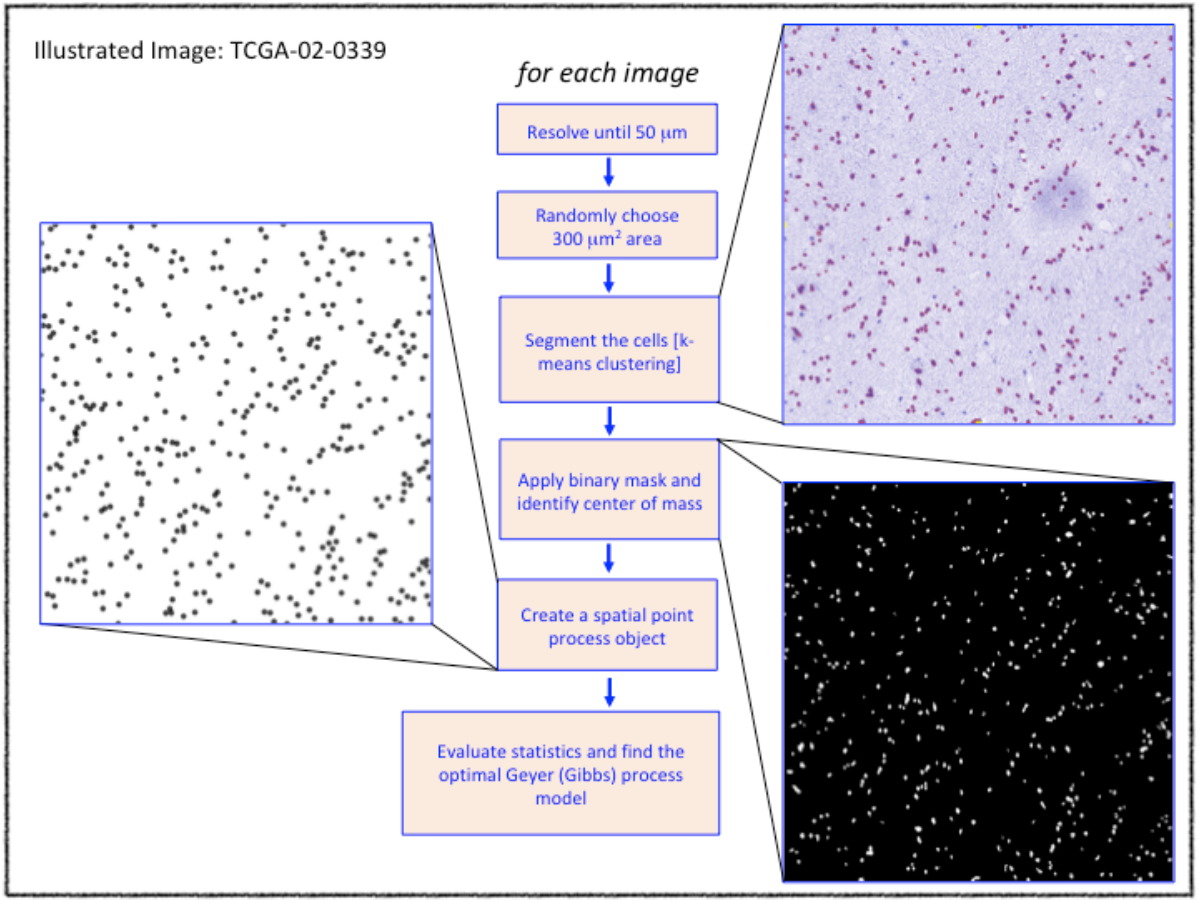


Source: CBTRUS, American Brain Tumor Association website.

Image Segmentation and Point Process Representation

Downloaded and processed 411 Glioblastoma diagnostic images

Used several open source and stand alone packages available in R, Matlab, QuPath



Question

What is the resolution of the image should we work with?

How do we know 50 μm resolution, 300 μm^2 area is appropriate for our analysis?

Ripley's K-function Motivation

Wolf pack example (Central Arctic Region, 1998)

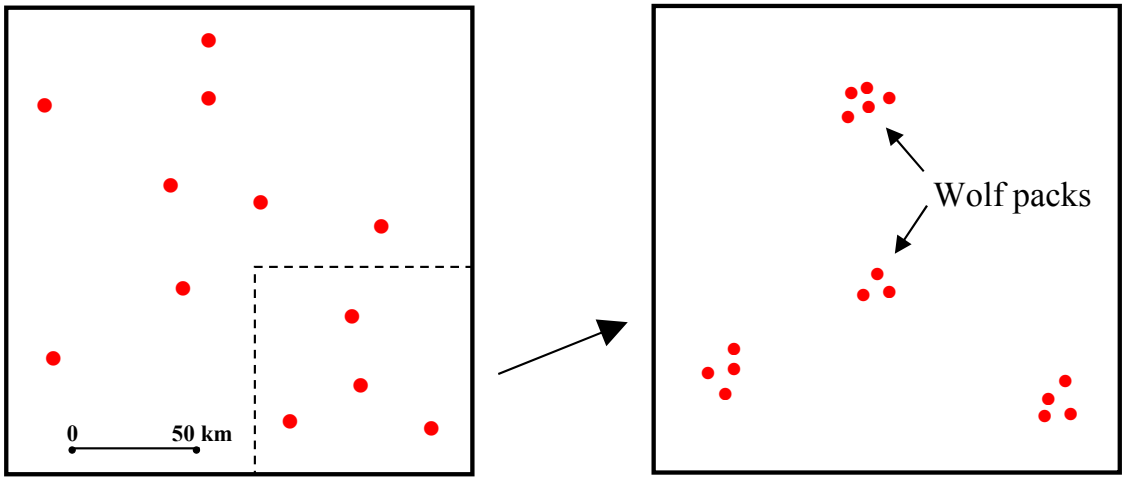


Fig.4.1a. Map of Wolf Packs **Fig.4.1b. Enlarged Portion**

Each pack establishes a hunting territory large enough for its survival (roughly 15 to 20 km in diameter), and actively discourages other packs from invading its territory.

Very clustered at small scales, very dispersed at large scales

Nearest-neighbor techniques can detect only small scale clustering

No way to analyze multiple scales without some form of re-aggregation

K and L-functions

- Key idea of K-functions is incorporating "scale" as a variable in the analysis

- What is the expected number of point events within distance h from any randomly sampled point event?

- This expected number is not very meaningful without specifying the point density, λ
  - It will increase with λ, Therefore, divide by λ

$$K(r) = \frac{1}{\lambda} \mathrm{E}[\text{number of } r\text{-neighbors of any point at a location} \,|\, \text{there is a point in that location}]$$

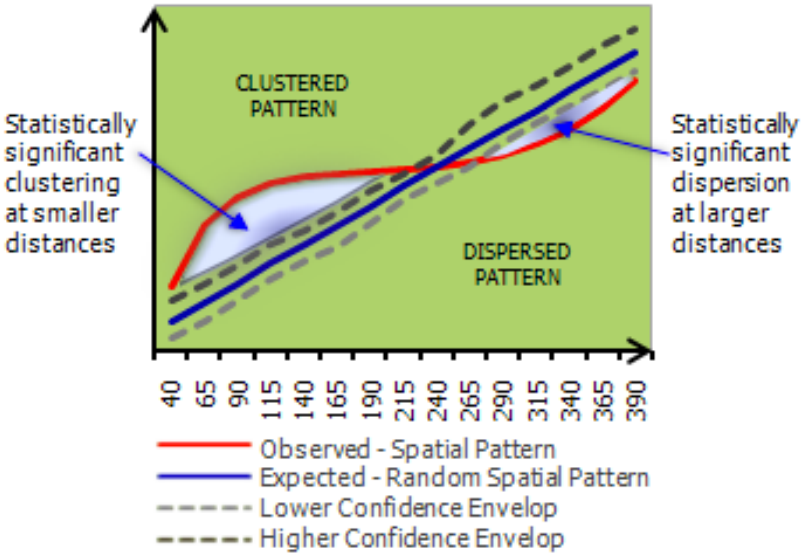
- For homogeneous Poisson process (CSR) since the points are independent, the presence of a random point will have no bearing on other points

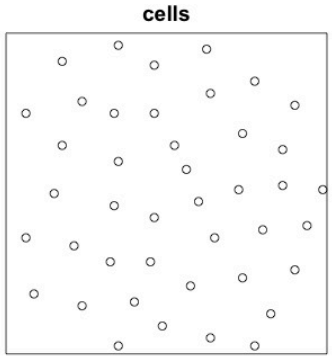
$$K_{poiss}(r) = \frac{1}{\lambda}(\lambda \pi r^2) = \pi r^2$$

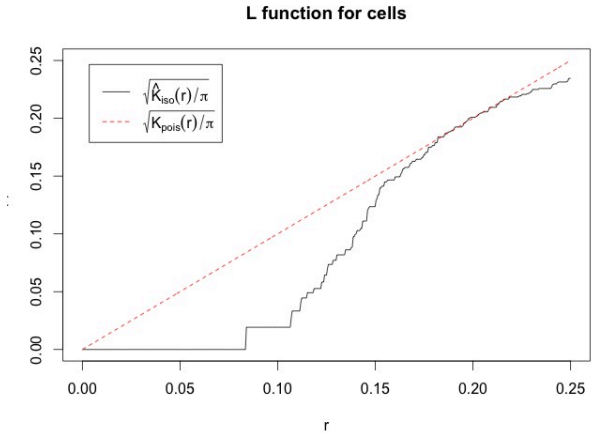
- $K(r) - K_{poiss}(r) = K(r) - \pi r^2$ can be used to determine if the points are random, clustered or dispersed. Provides a scale-free description.

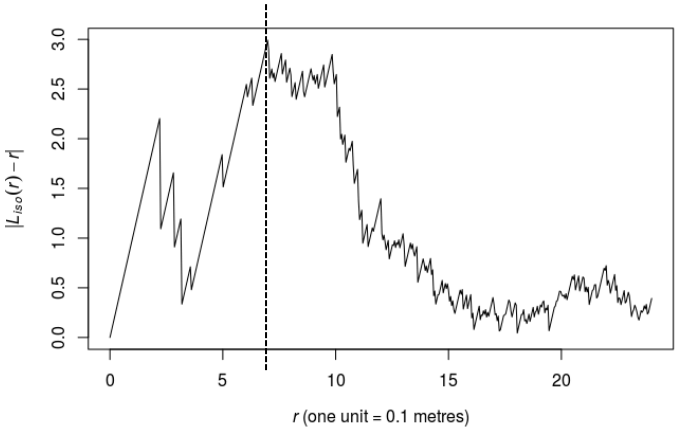
- L- function, a commonly used transformation $L(r) = \sqrt{\frac{K(r)}{\pi}}$

L-function









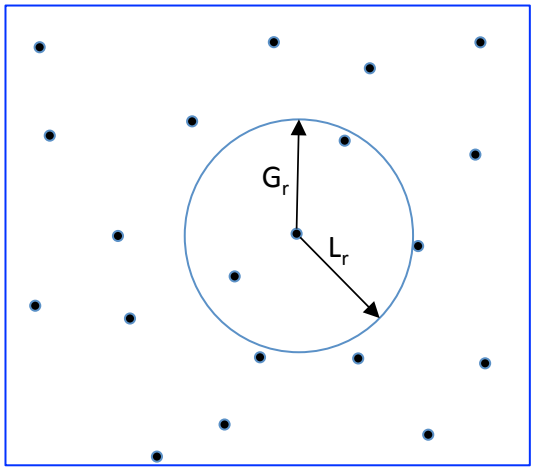
Cells have a dispersed phenotype

http://resources.esri.com/help/9.3/arcgisdesktop/com/gp_toolref/spatial_statistics_tools/how_multi_distance_spatial_cluster_analysis_colon_ripley_s_k_function_spatial_statistics_works.htm
https://stats.stackexchange.com/questions/57384/ripleys-k-function-and-l-function-for-point-patterns/57388#57388

Two Measures

- Optimal L-function and the interaction parameter provided by the Geyer saturation process
  - Call it as M_r and G_r

- If the cells inhibit each other with the interaction radius G_r we can expect the cells be dispersed with the optimal metric $L_r = opt_r|M_r - r|$



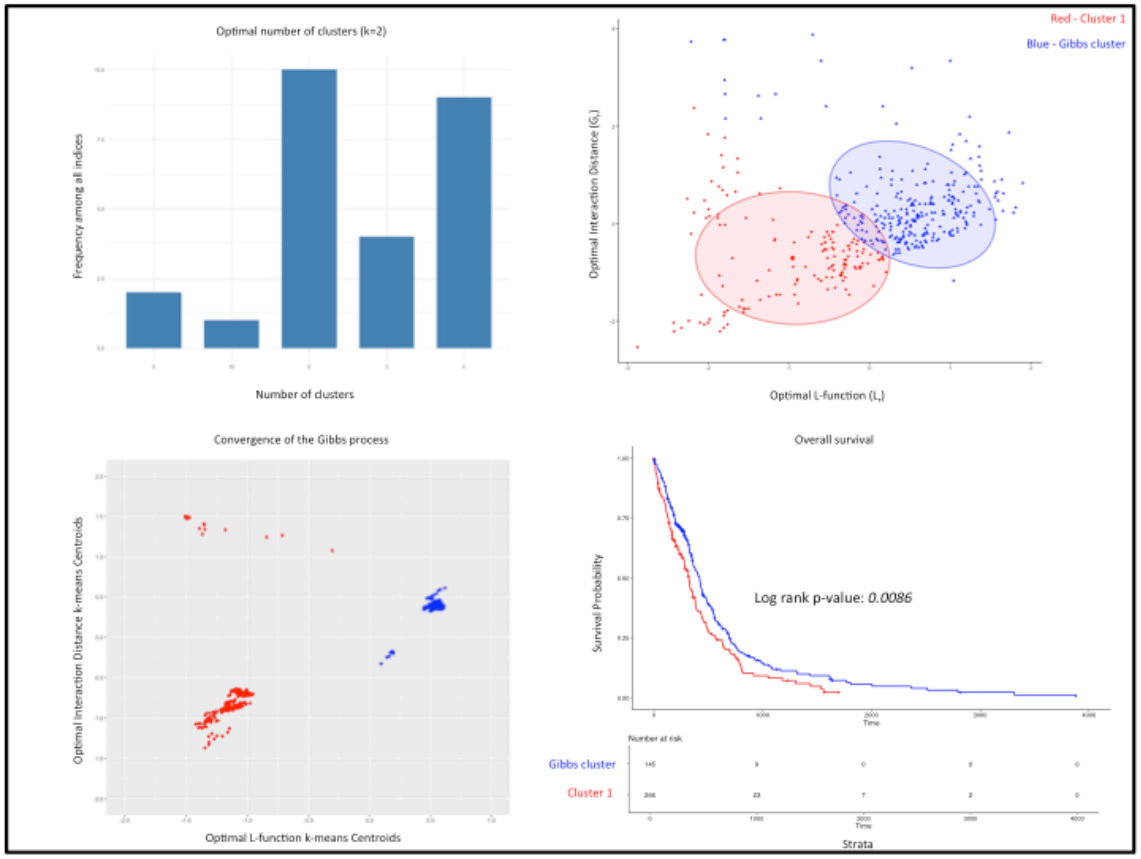
- Therefore, analyze the metric (L_r, G_r)

Subgroups in GBM Cells

Unsupervised analysis - cluster the metric (L_r,G_r)

Optimal number of clusters − 2 (k-means clustering)

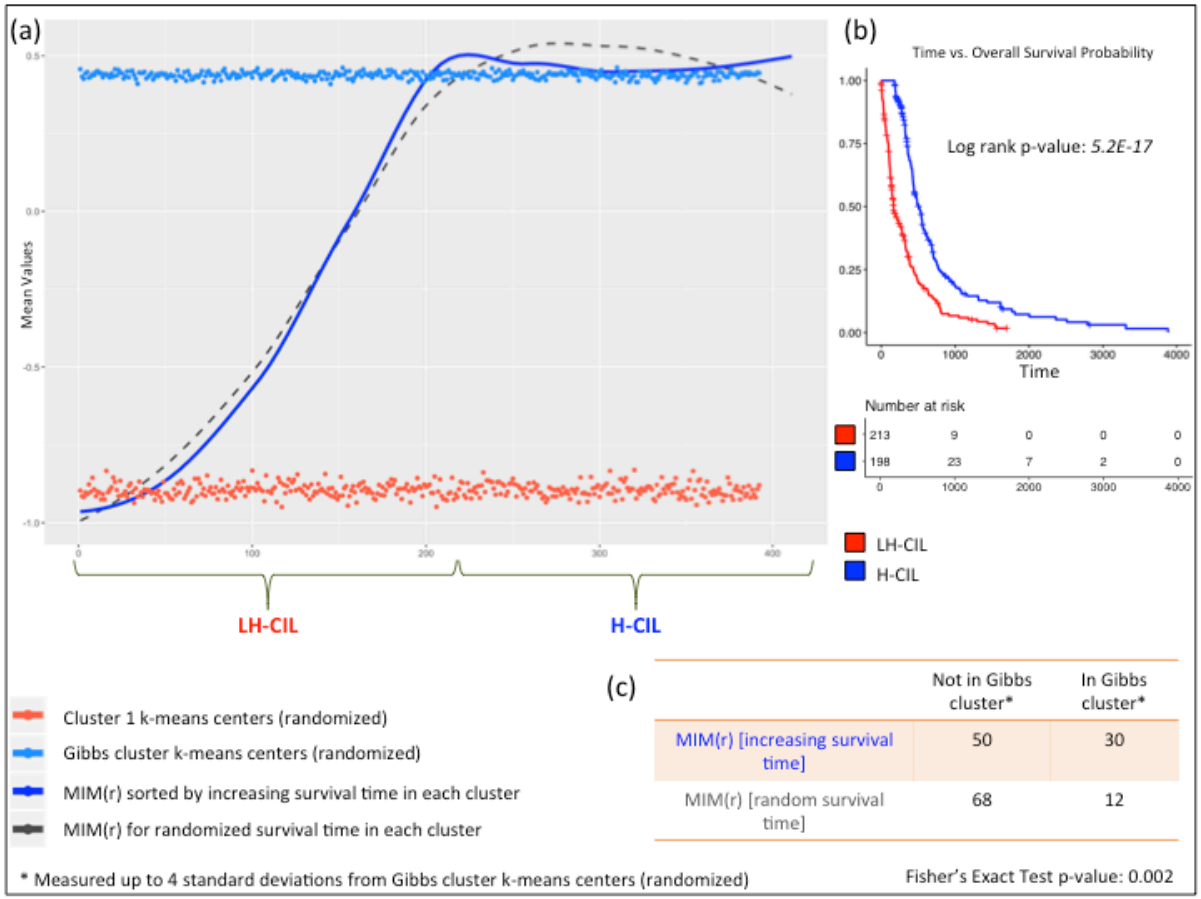
Small sample size with respect to noise - study convergence instead



Gibbs Process Determines Survival

Identify signal from noise

Smooth the data using Loess regression

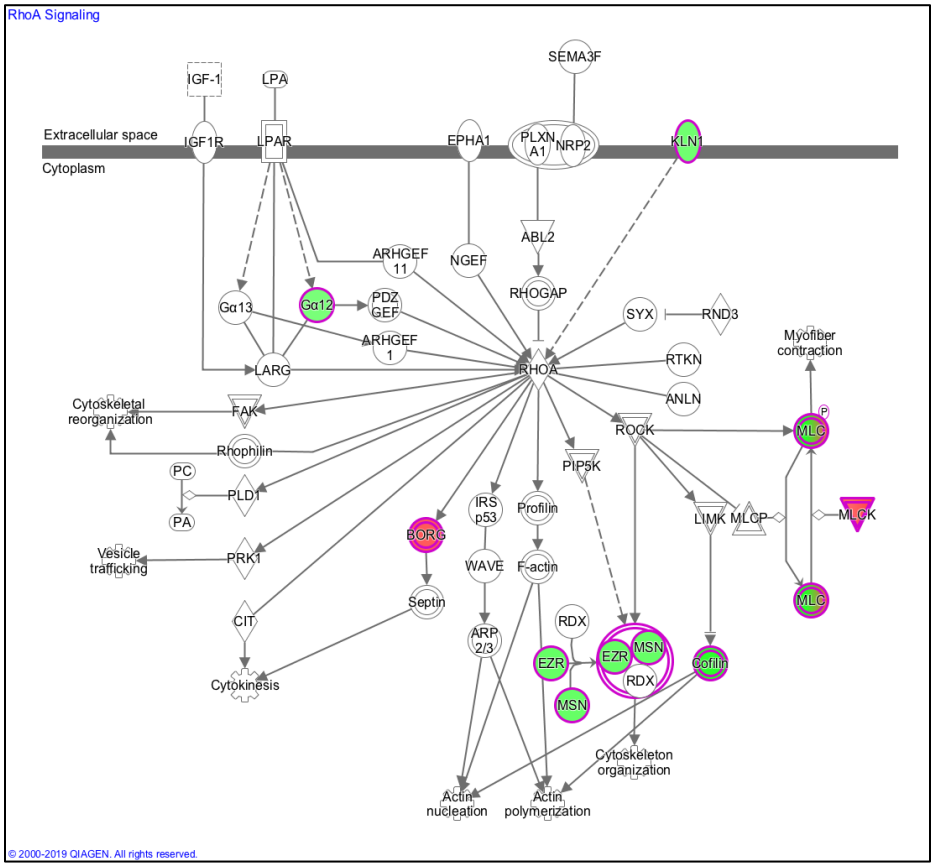


	Not in Gibbs cluster*	In Gibbs cluster*
MIM(r) [increasing survival time]	50	30
MIM(r) [random survival time]	68	12

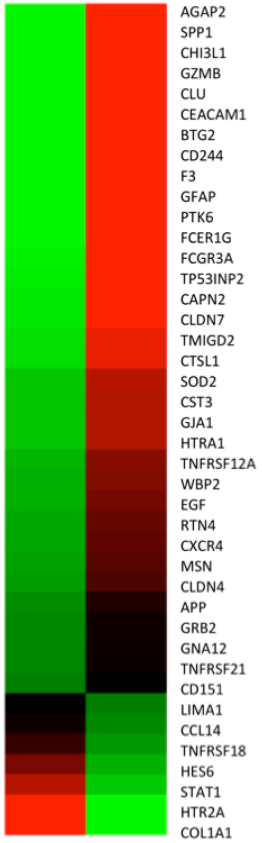
Movement Gene Signatures and Pathways

Analyze RNA-seq. data (69 patients) − 86 gene signature

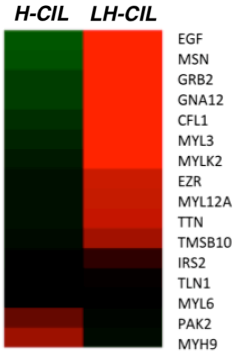
Key genes in cell movement and proliferation pathways



(a)

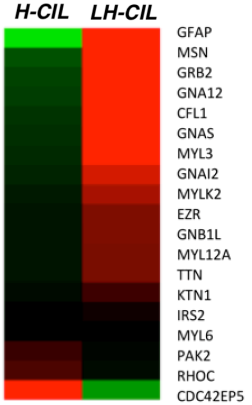


(b)



Actin cytoskeleton pathway signature

(c)



RhoA signaling pathway signature

Contact Inhibition of Locomotion (CIL)

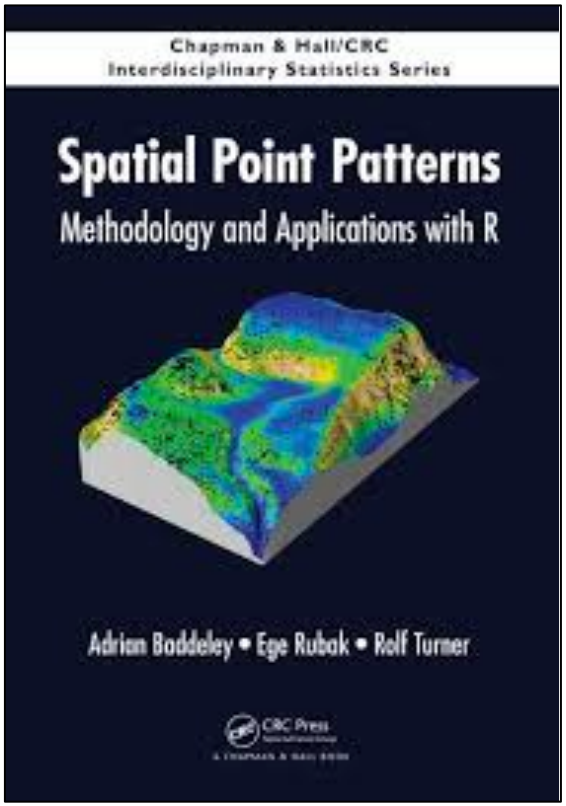
- Both CIL and CIP are important mechanisms in cancer progression
  - CIP is a hallmark of cancer*
- However, CIL has been largely ignored since in vivo experiments in tumors has not been carried out
- CIL first characterized through the work of Abercrombie and Heaysman in the 50's.
- Two modes of CIL: homotypic and hetrotypic CIL
- Loss of heterotypic CIL has been proposed as a mechanism for tumor invasion and metastasis while tumor cells themselves maintain homotypic contact inhibition
- Also makes sense from a mathematical perspective
  - Tumor invasion is a birth and death process of rapidly dividing cells
  - Representing tumor cells as two-dimensional points, we can expect the tumor tissues in histology slides to reflect realizations of spatial birth-and-death process
  - Gibbs process is the equilibrium process of spatial birth and death process

* Hanahan D, Weinberg RA (March 2011). "Hallmarks of cancer: the next generation". Cell. 144 (5): 646–74.

Importance of Spatial Point Processes

- Very powerful mathematical framework for studying the behavior of cells
- Marked point processes with covariates can practically model all aspects of cell behavior through data integration, including omics data
- Extract image frames from videos and use spatial point process
  - Developmental biology (Zebrafish cells diverging to different cell types)
- Logical step is to study immune responses through spatial point process framework
- Estimate relative risk of proliferation (cancer) and immune responses
- Cost effective especially when used with staining techniques
- Translational − bench to bedside, scale-free metrics to determine disease states
- Can be put under the framework of AI and emergent systems − logical next steps for biology

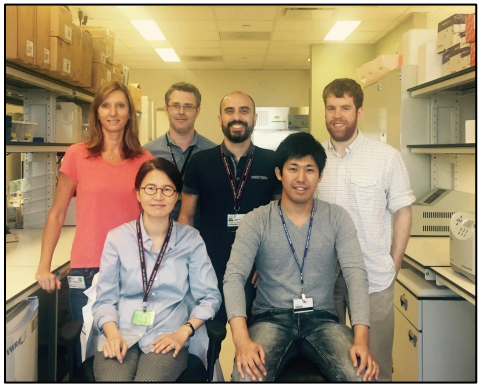
References



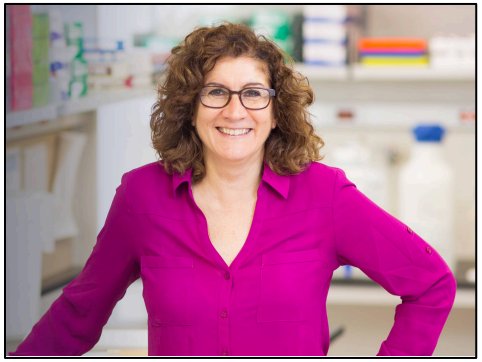
Preprint

Lavanya Kannan, Tarjani Agarwal, Matija Snuderl, David Zagzag, Erik Sulman, Jason Huse, and Kasthuri Kannan. Gibbs Process Determines Survival and Reveals Contact-Inhibition Genes in Glioblastoma Multiforme. bioRxiv 608414; doi: https://doi.org/10.1101/608414

Acknowledgements



Jason Huse, MD, PhD.
Associate Professor of Pathology
Director, Translational Molecular Pathology
MD Anderson Cancer Center



Adriana Heguy, PhD.
Professor of Pathology
Assistant Dean, Division of Advanced Research Technologies
Director, Genome Technology Center
New York University Langone Health

Applied Bioinformatics Laboratories

Questions?