



Julia Brettschneider

Statistical methodology and data science in science and engineering: Point processes

I Hawkes process modelling of the pandemic

2 Dead pixel formations on digital X-ray detectors

3 Microscopic image based modelling of biological processes

Applied statistician/data scientists

Collaborators: engineers, life scientists, clinicians...

Domains: genomics, microscopy, detectors, cancer, screening, finance, OR...

Methodological topics: data quality, spatial statistics, decision theory,

concepts of probability and risk

Short biography

- Reader (since 2021), Associate Professor (2010-2021), Assistant Professor (2007-2010),
 Dept of Statistics, University of Warwick, UK
- Turing fellow since 2017
- Assistant Professor, Dept of Math/Stats & Dept of Community Health/Epidemiology & Cancer Research Institute, Queen's University, CN
- Visiting Assistant Professor and Research Statistician, Dept of Statistics at University of California at Berkeley, USA
- Postdoctoral fellow in Computational Biology at Eurandom, NL
- PhD (2001) in Mathematics, thesis supervisor Prof H Föllmer, Humboldt Uni Berlin, D
- Masters in Mathematics (with Computer Sciences and Psychology), thesis supervisor Prof H Föllmer, University Bonn, D

Point processes

Definition: Point Process

Point processes are a class of random process whose <u>realisations</u> are a set of points on some given space.

<u>i.e.</u> A sequence of random variables $t = \{t_1, t_2, ..., t_d\}$ taking values in a subset of \mathbb{R}^d .

Definition: Temporal Point Process (also called counting process)

Temporal Point process is a point process over time: It describes the occurrence of random events over time.

<u>i.e.</u> a sequence of events $t = \{t_1, t_2, ..., t_d\}$ s.t. $0 \le t_1 < t_2 < ... < t_d$

Example: Poisson point process

1d: Stochastic process N(t) $(t \ge 0)$ with

- N(0) = 0
- independent increments
- number of events at interval of length t is a Poisson random variable with mean λt

Higher d: Stochastic process $\ N(A)\ (A\in\mathcal{B}(\mathbf{R}^d))$ with

- $N(\emptyset) = 0$
- ${\cal N}(A)$ and ${\cal N}(B)$ are independent for disjoints sets A,B
- N(A) is a Poisson random variable with mean $\mu(A)$

Interpretation:

Number of points in a set is distributed proportionally to the size of the set, i.e. the area (plane d=2), volume (space d=3), and what happens in a region does not effect what happens in a different regions (if there is no overlap).

Example: Hawkes process

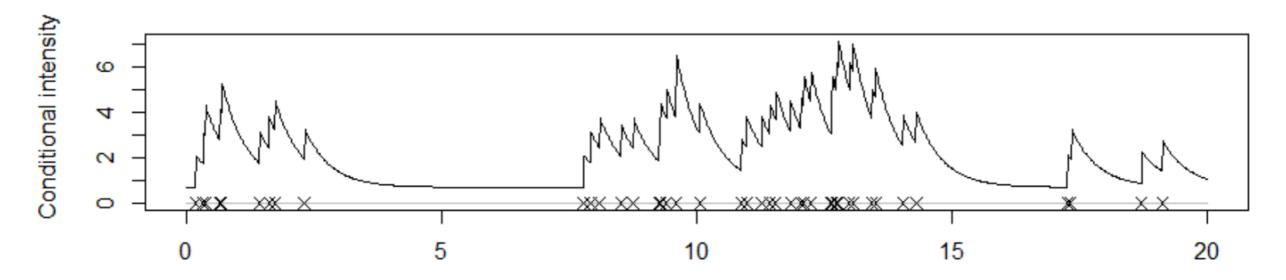
ETAS Hawkes Process model

ETAS = Epidemic-type aftershock sequence

- Most widely used example of Hawkes Process.
- Model:

$$\lambda(t) = \mu + \sum_{t_i < t} \alpha e^{-\beta(t - t_i)}$$
Exponential triggering function

Example:



Joint projects with integrated Master's students:

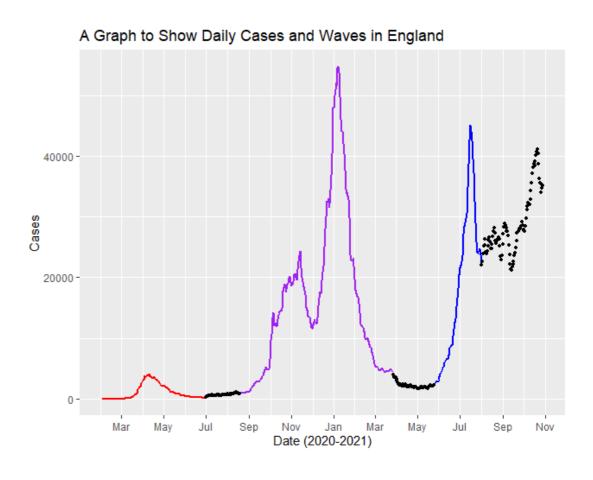
Marianna Mavroleftherou and Adam Davison

1. Data from Israel

Vaccine for data deal with Pfizer/BioNTech (Marianna Mavroleftherou's project)

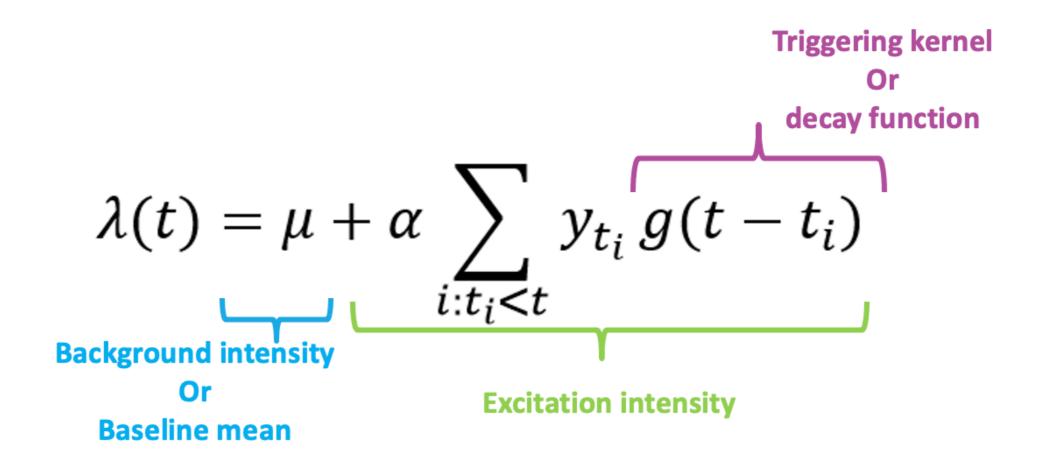
2. Data from England

ONS/Oxford survey (Adam Davison's project)



Hawkes Process

Self-exciting temporal point process described by the conditional intensity function:



Model Setup

Number of daily cases (or deaths) [e.g. Rousseau et al 2020]

 \succ It is a natural response to assume that Y(t) follows Poisson($\lambda(t)$)

Since Poisson distribution is governed by one parameter λ , which is the expected number of times an event occurs in an interval of time or space.

Hence, we set:

$$P(Y(t) = y \mid \lambda(t)) = \frac{\lambda(t)^{y} e^{\lambda(t)}}{y!}$$

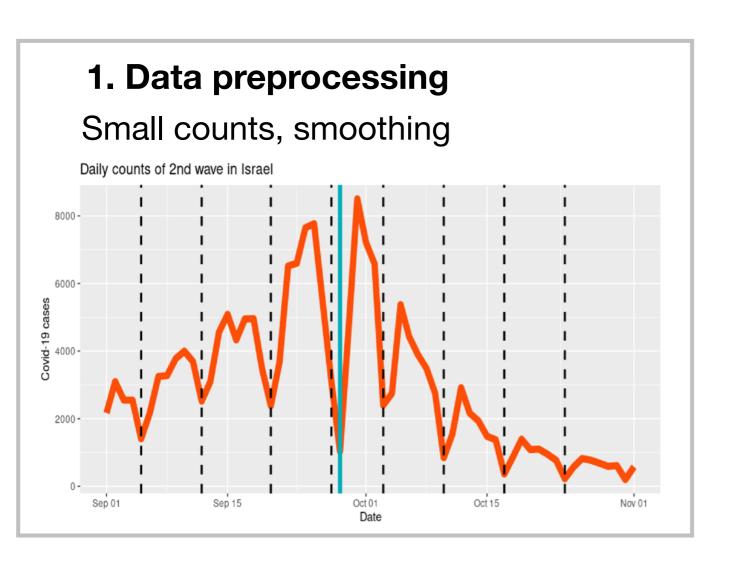
> Further to this:

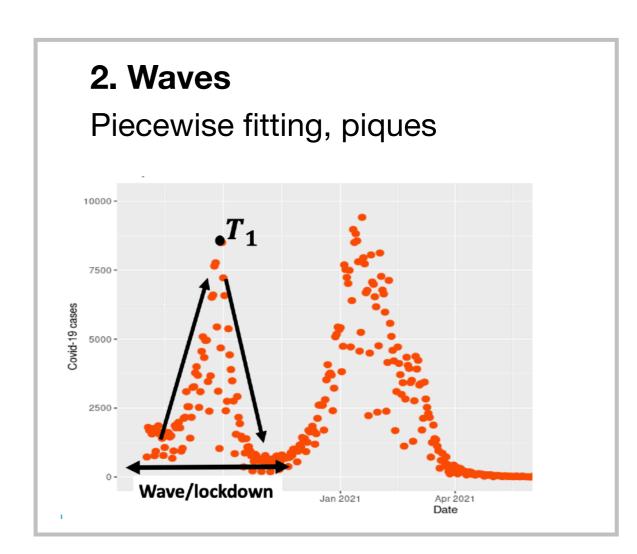
we choose the triggering kernel g() to be the geometric excitation kernel:

$$g(t - t_i | \beta) = \beta (1 - \beta)^{t - t_i - 1}$$

Since it can be shown to be the generalization of the exponential distribution in discrete time.

Model fitting steps in practice:



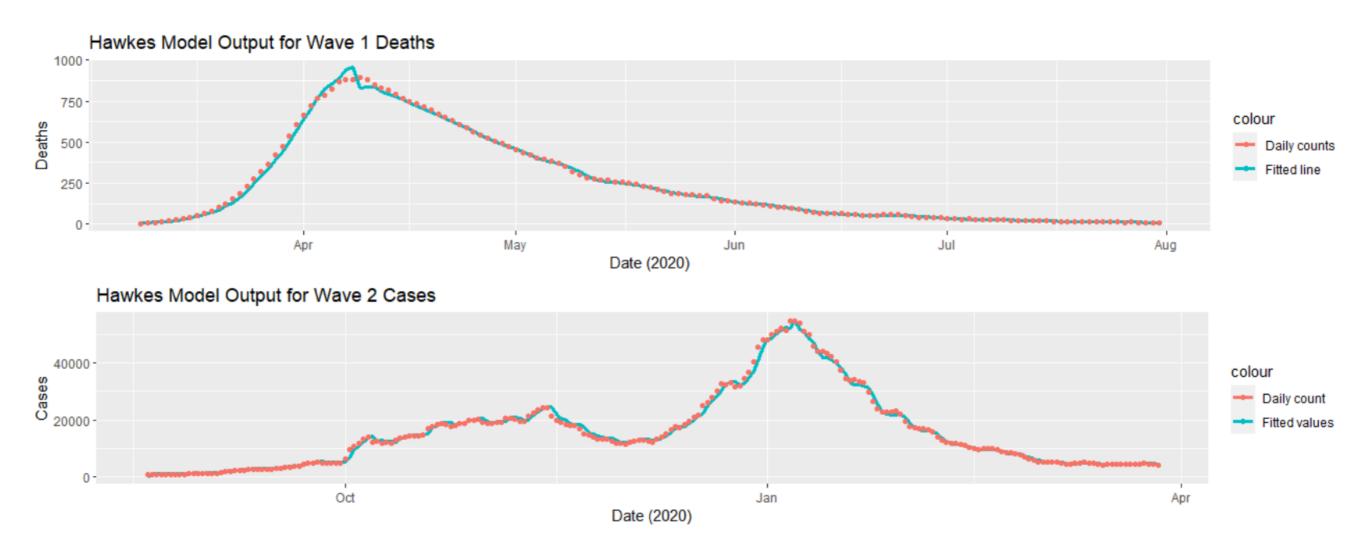


3. Likelihood function

4. Inference

- ML with Nelder-Mead
- Bayesian posterior MCMC

English data:



Further and future work:

- Models including age
- Including vaccination rates
- Testing coverage/reliability
- Behavioural indicators (e.g. google searches or mobility)
- Lockdown effects
- Events (e.g. football, holidays) and interaction with other factors
- Mixture population
- Regional models

Dead pixel formations on digital X-ray detectors

Inside-out

Statistical Methods for Computed Tomography Validation of Complex Structures in Additive Layer Manufacturing

PI: Prof W Kendall

Other investigators: Prof M A Williams, Dr G J Gibbons,

Dr J Brettschneider, Prof T Nichols

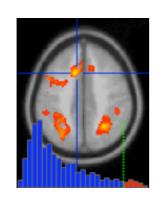
EPSRC grant 3 years

10/2013 -9/2016

EP/K031066/1

Further team members: Clair Barnes, Jay Warnett, Audrey Kueh Industrial partners: Nikon Metrology, Remishaw, EOS systems















Spatial analysis of dead pixels

Objectives:

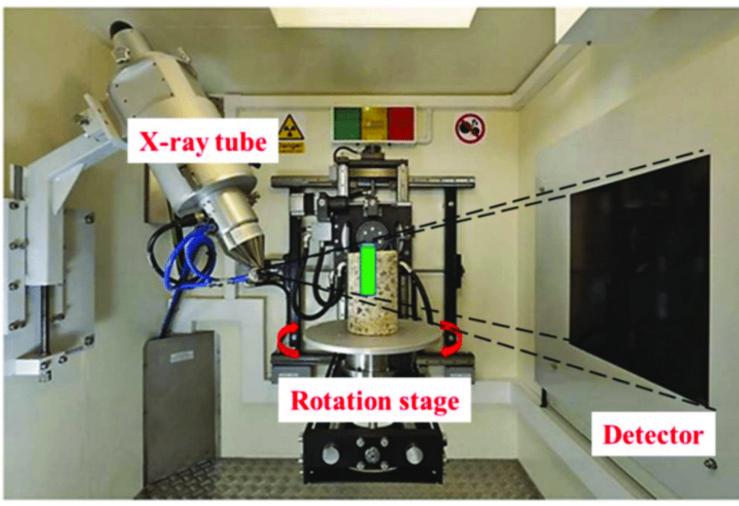
- Feedback about state of detector through spatial pixel damage analysis
- Detector data repository

Applications:

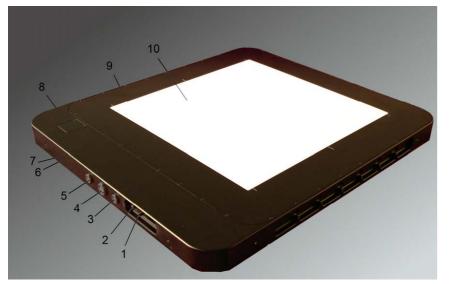
- Identify poor quality regions (patches with high dead pixels density)
 through density thresholding
- Remaining area CSR means no special causes of poor quality
- Identify causes of poor quality
- Monitor over time
- Conclusions for usage modes
- Conclusions about weakness of detector construction

X-ray chamber

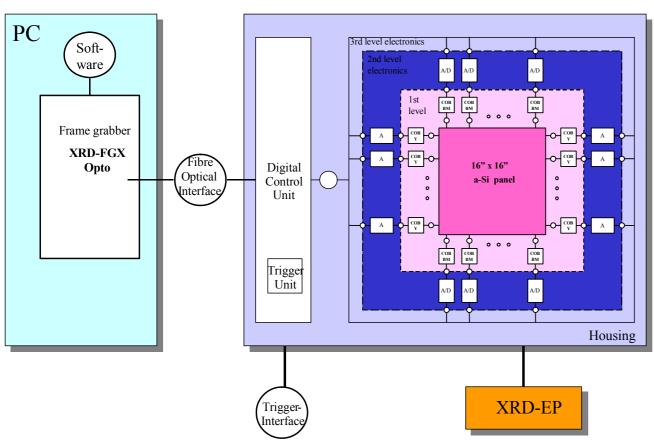


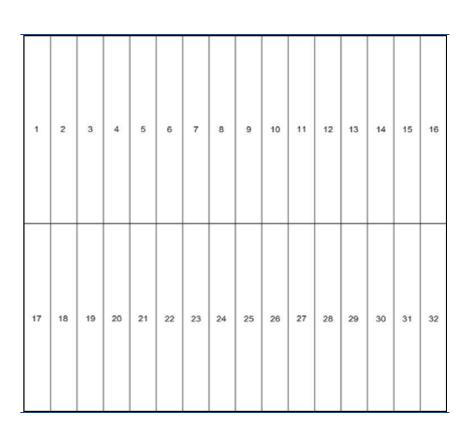


X-ray detector



Perkin Elmer XRD 1621



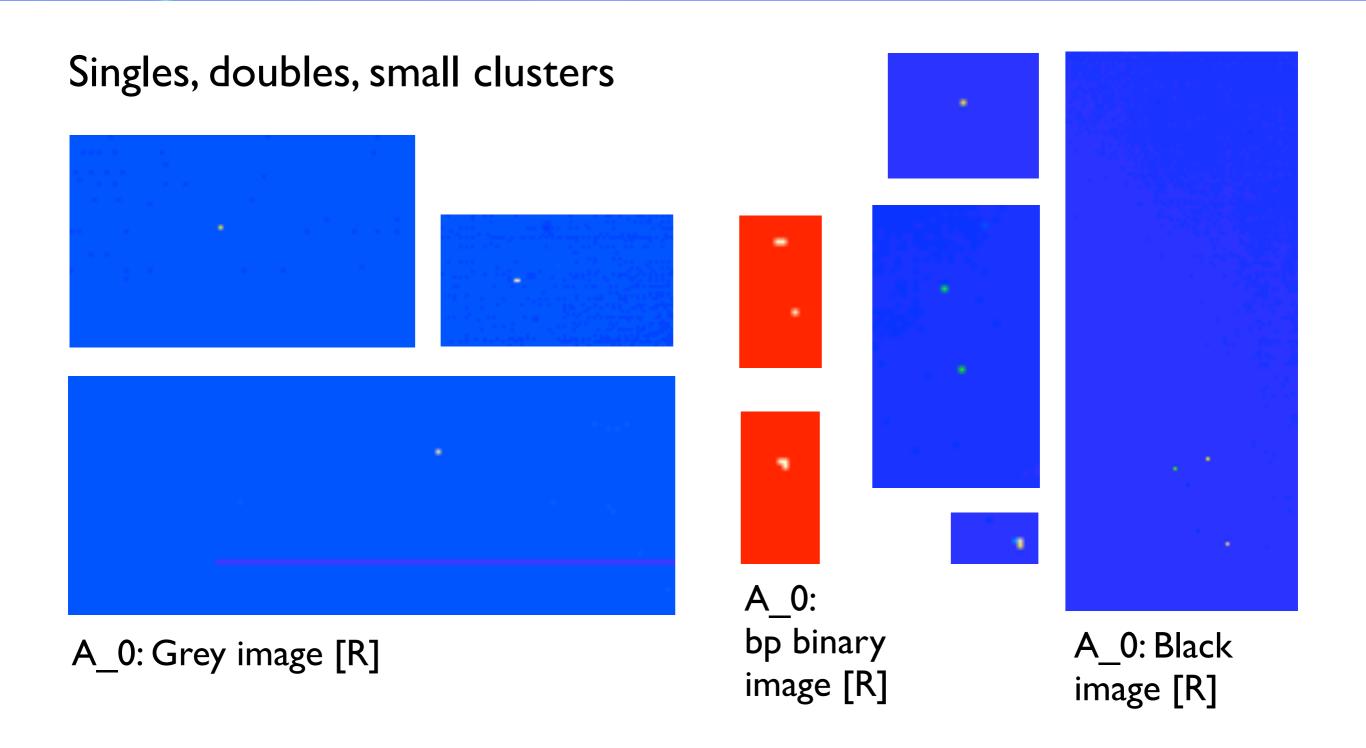


Readout groups (ROG):

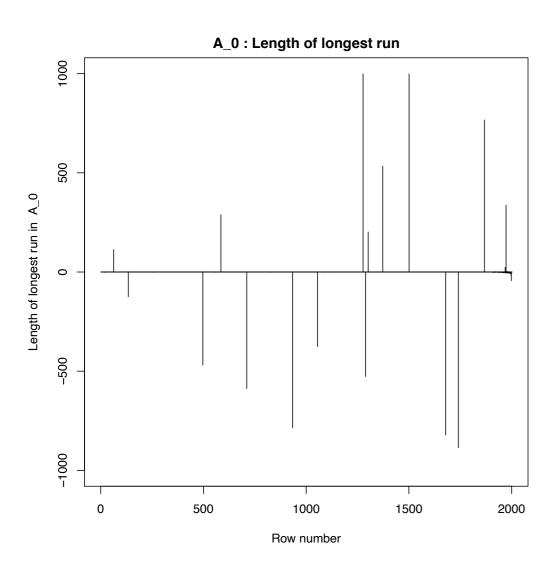
Upper groups transferred first, starting read out from the upper row.

Lower groups starting from the last row.

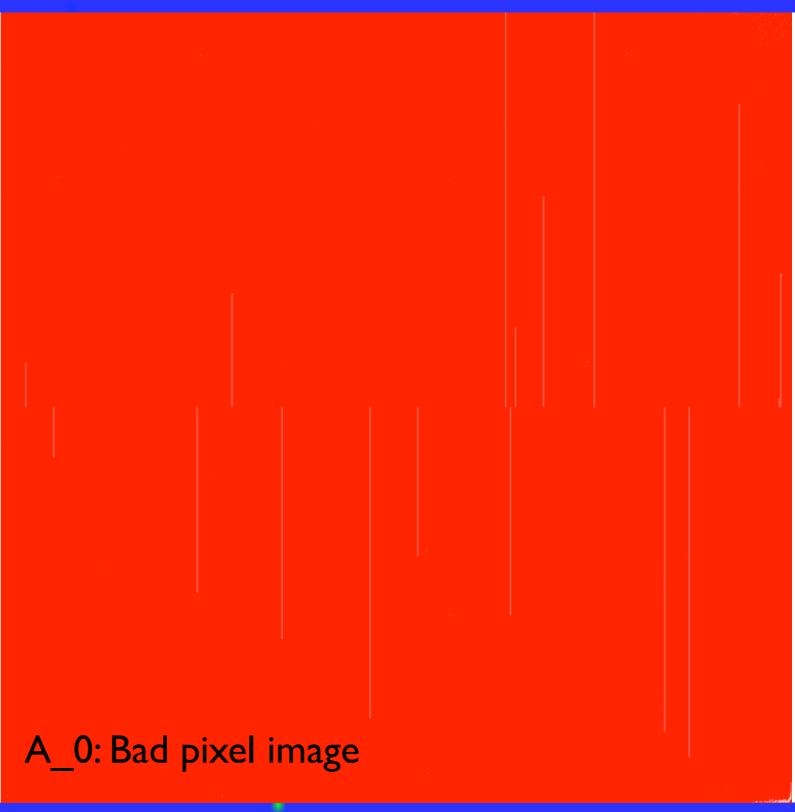
Local defects: Isolated dead pixels



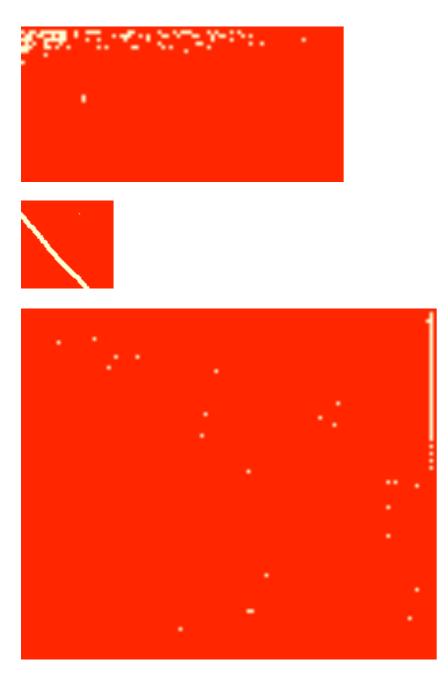
Local defects: Dead lines



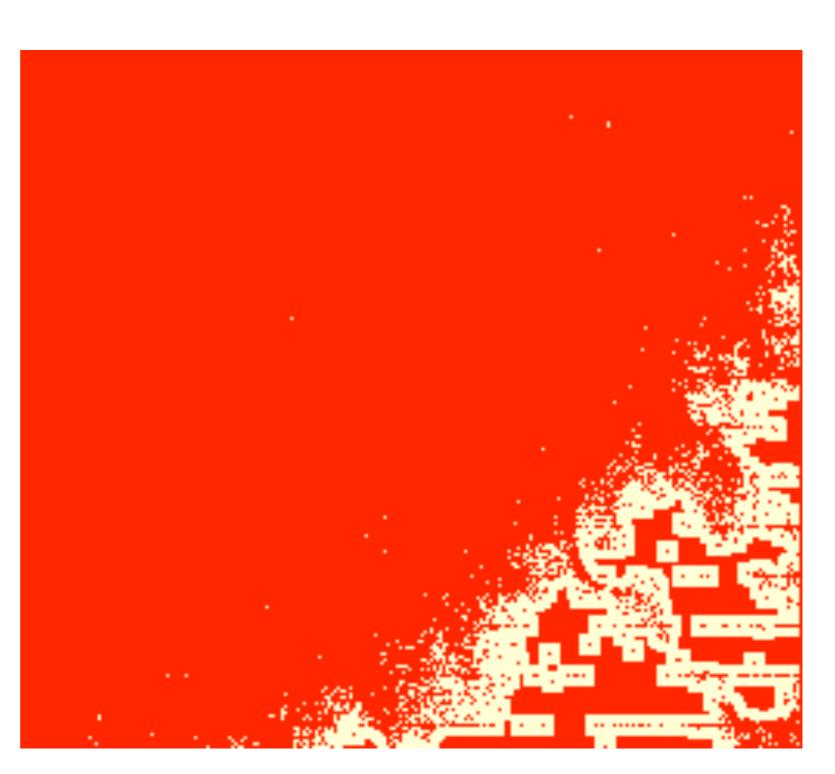
A_0: Graph of bad pixel images



Local defects: Corners

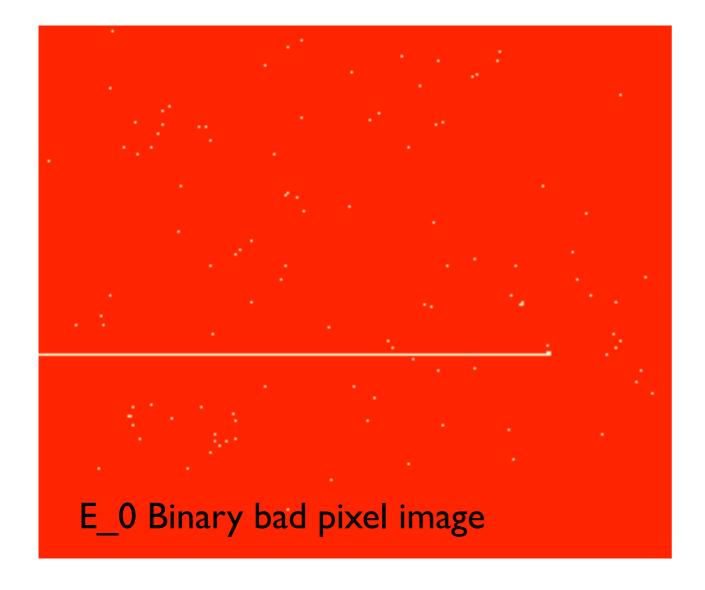






Local defects: Patches

Areas with high density area of bad pixels





Spatial model for dead pixels

Dead pixel set as point process

Detector is based on a lattice, but **our interest** is in **locations** of dead pixels and these are relatively few. Hence, use a spatial point pattern model, but with reduced resolution (given by the detector lattice).

Point pattern X: random locations of dead pixels (2 dimensional)

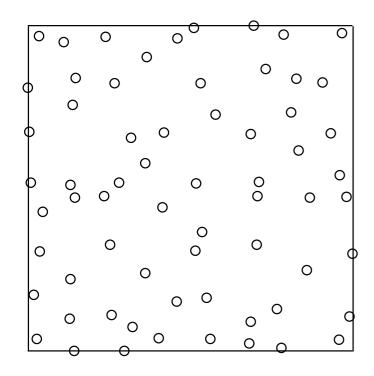
Objectives:

- describe spatial distribution of dead pixels
- hypothesise causes for dead pixels

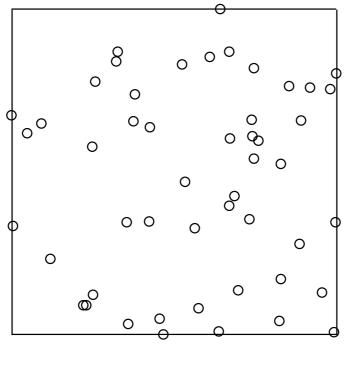
For example, look at CSR...

Complete spatial randomness (CSR)

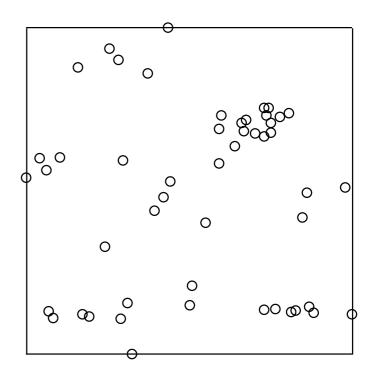
CSR: Points are distributed independently and homogeneously, as in a homogenous Poisson process.



Regular (nearly)



CSR



Clustering

Exploring CSR using F- and G-functions

Nearest neighbour function G:

Cumulative distribution function of the distance from an arbitrary point to its nearest point

Under CSR:
$$G(r) = 1 - \exp(-\lambda \pi r^2)$$

Empty space function F:

Cumulative distribution function of the distance from an arbitrary location to its nearest point

Under CSR:
$$F(r) = 1 - \exp(-\lambda \pi r^2)$$

Exploring CSR using Ripley's K-function

K-function:

expected number of extra points in circle of radius r rescaled by density

$$K(r) = \lambda^{-1} E[N_0(r)]$$

 $N_0(r)$ number of points within distance r from arbitrary point λ globally estimated density

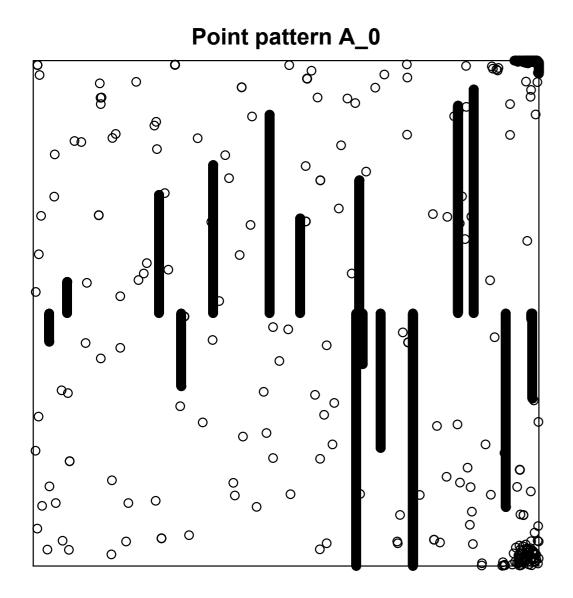
Under CSR:
$$K(r) = \pi r^2$$

, nsim=100

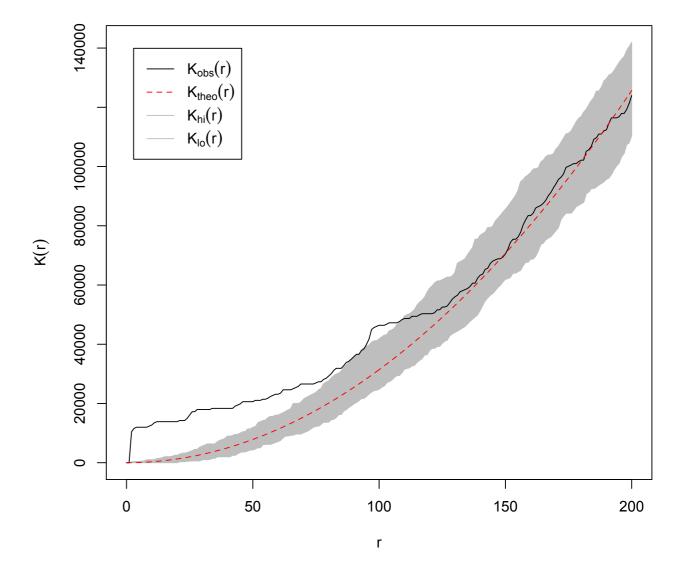
G-function, Events, nsim=100



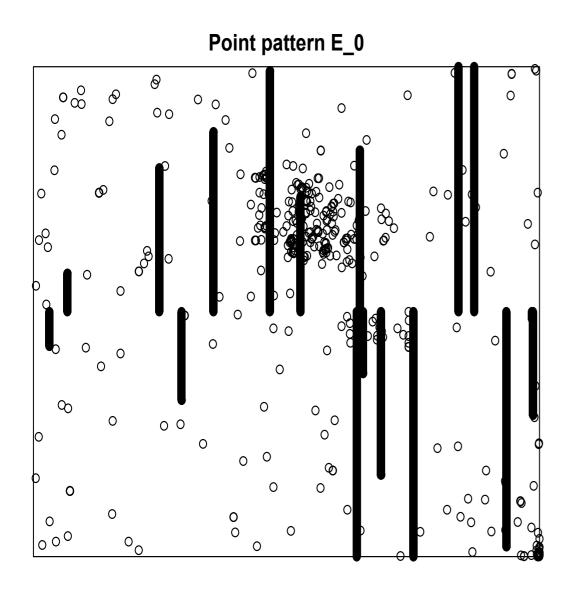
Point pattern and K-function

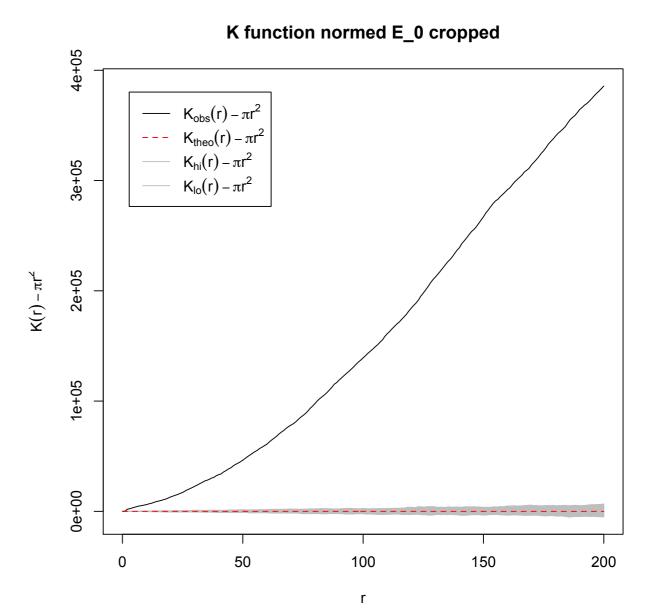


K function A_0 cropped



Point pattern and K-function





Are we asking the right question?

Modified question: Is it CSR after we remove all specific (known) problems?

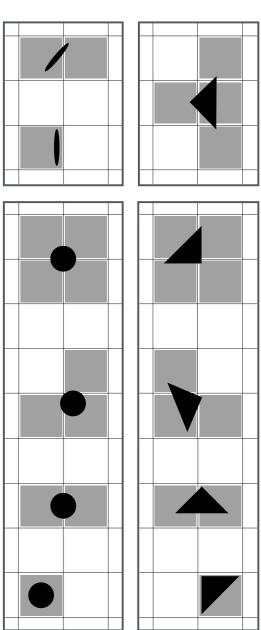
Step I:

Convert point process into event process by

- Reducing a line to one endpoint
- Reducing a clusters to its centre point

Step 2:

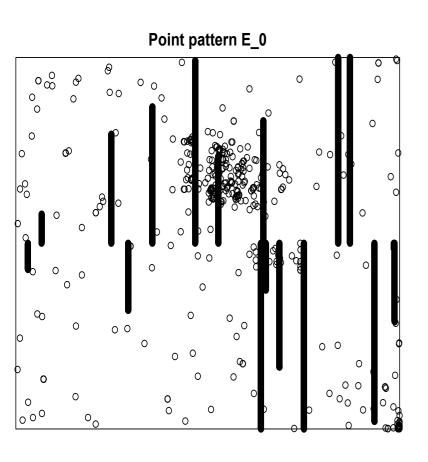
- Fit inhomogeneous density
- Cut out areas above threshold

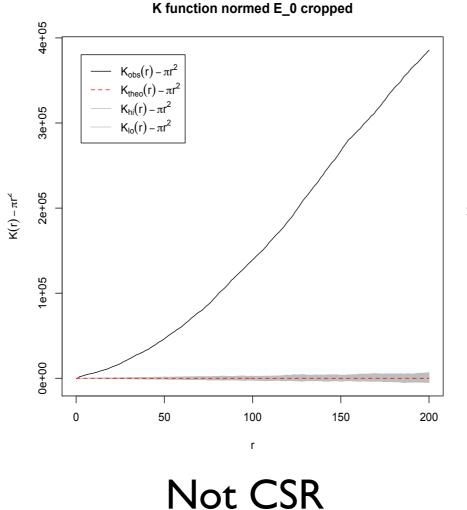


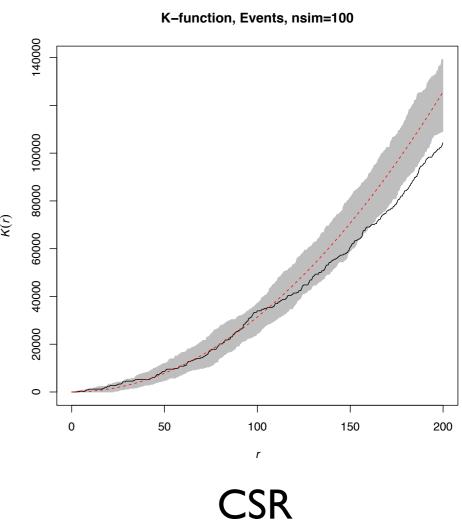
Modified process: K-function



Event level



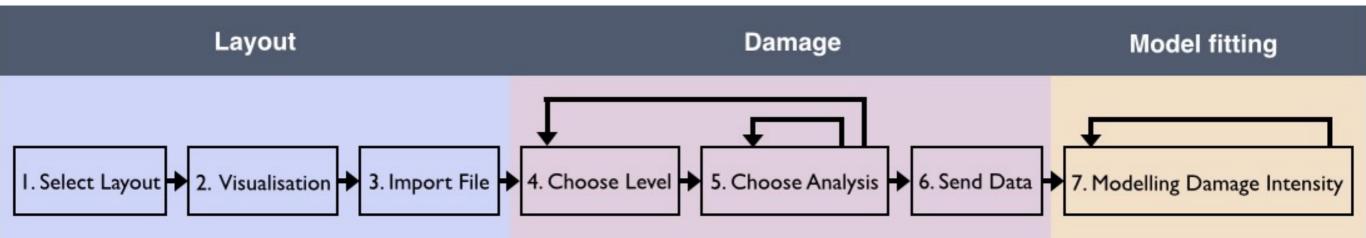




Seed funding for a software project with Turing

- Working with Turing Research Software Engineer Group
- DetectorChecker R package for statistical analysis of pixel damage in CT scanners available at https://github.com/alan-turing-institute/DetectorChecker
- DetectorCheckerWebApp for useful initial graphical/analysis, available at https://detectorchecker.azurewebsites.net
- Facility to upload data in different formats (crowd sourcing)

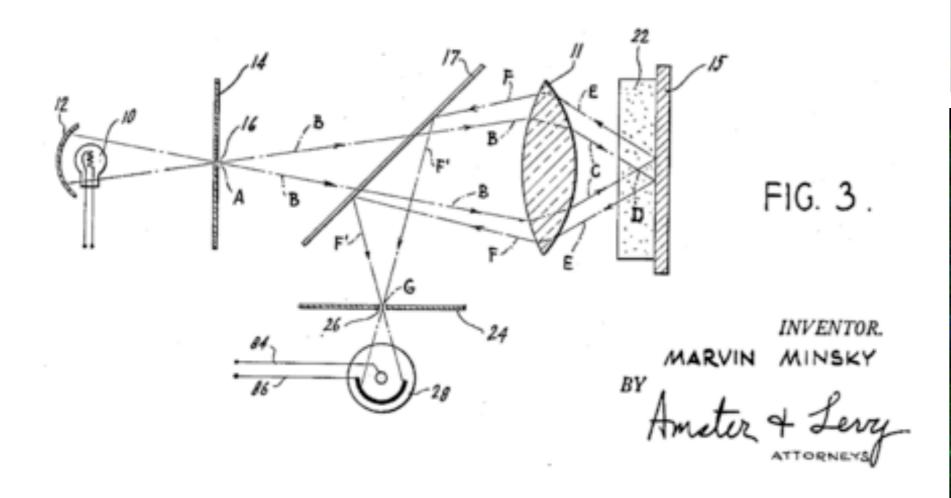
Brettschneider, Giles, Kendall, Lausaskas, (2020). DetectorChecker: analyzing patterns of defects in detector screens. Journal of Open Source Software, 5(56), 2474

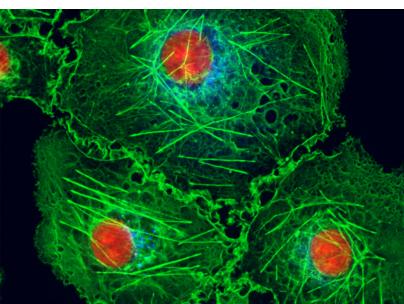


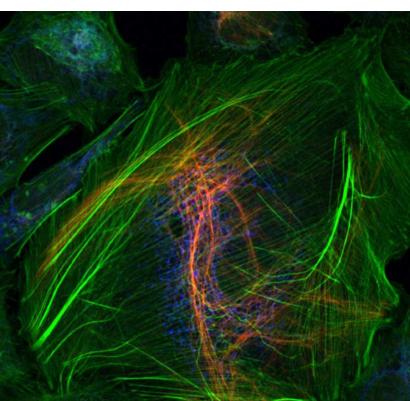
Microscopic image based modelling of biological processes

Quantifying spatial abundance of proteins in living cells

Confocal fluorescent laster microscopy (live cells)
Electron microscopy (dead cells, higher resolution)







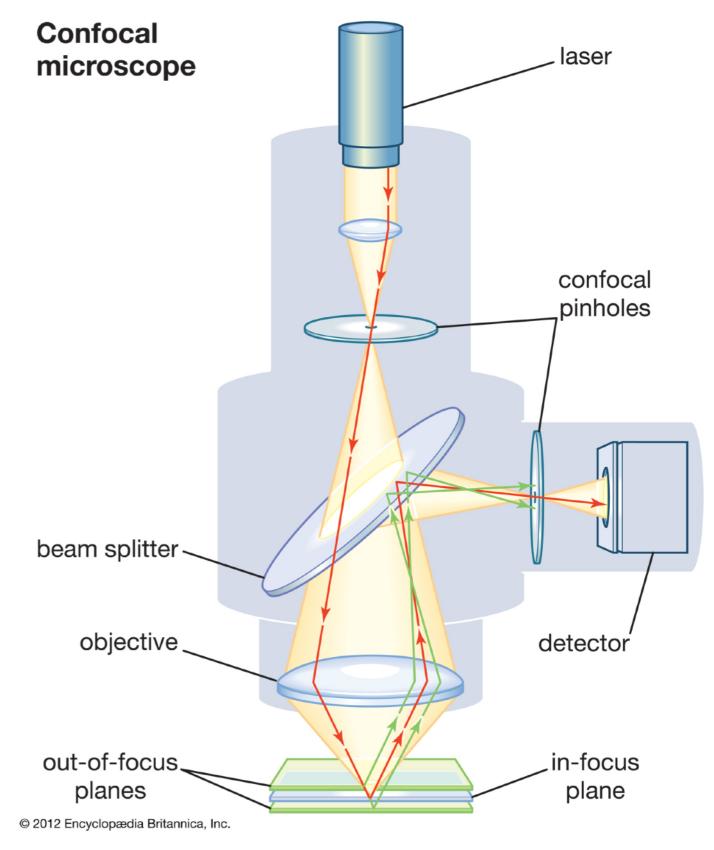
Confocal microscope

Wide-field microscopy:

- All of specimen excited at the same time
- Large unfocused background

Confocal microscope:

- Field of view limited by geometric optics
- Pinhole in front of the detector to eliminate out-of-focus signal
- Long exposure required
- Scanning arrangement to build up image of larger region
- Better resolution

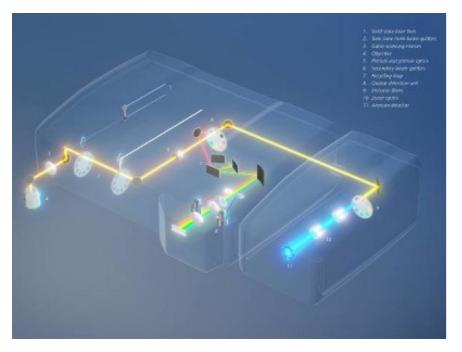


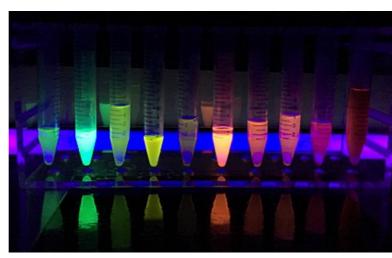
https://www.britannica.com/technology/microscope/Confocal-microscopes

Confocal fluorescent laser microscope

Fluorescent confocal microscope:

- Combination of two ideas in microscopy technology
- High resolution images
- Life cells
- 2D or 3D through scanning schemes
- Multi-channel through use of range of fluorescent proteins



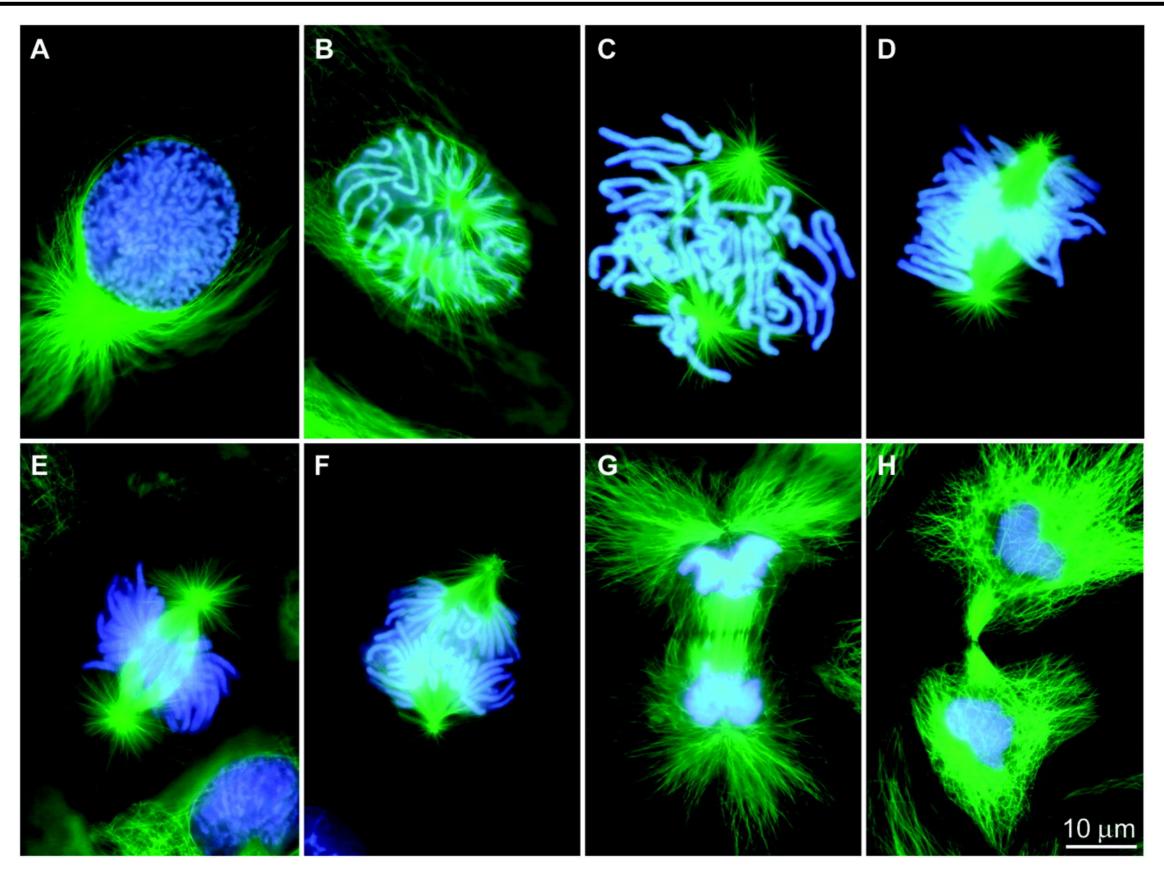




https://www.biocompare.com/25608-Microscopes-and-Cell-Imaging-Systems/14617250-ZEISS-LSM-980-Confocal-Laser-Scanning-Microscope/?pda=25608|14617250_0_1|2254289,2254327|1|&dfp=true

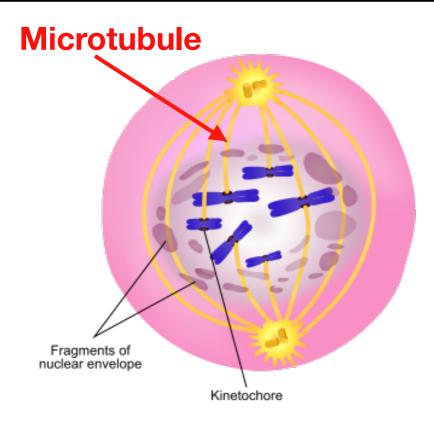
https://en.wikipedia.org/wiki/Green_fluorescent_protein#/media/File:Fluorescence_from_Fluorescent_Proteins.jpg

Microtubules formation during mitosis



https://science.sciencemag.org/content/300/5616/91/tab-figures-data

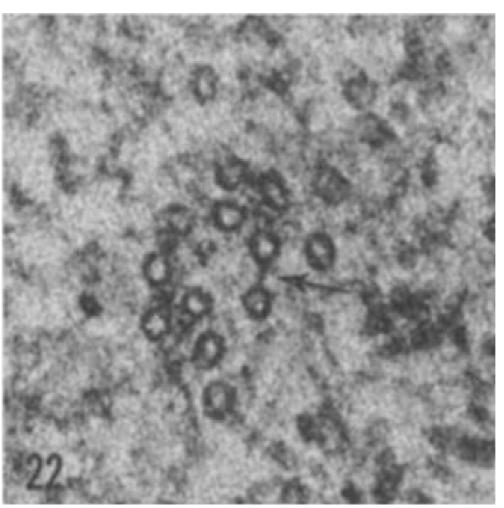
Microtubules during mitosis (cell division)



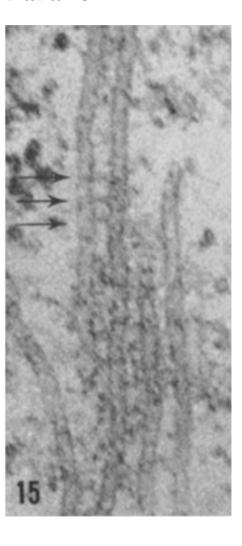
- Centrosomes = centrioles + microtubules
- Centrioles help the spindle into proper formation
- Spindle microtubules are arranged in K-fibers
- Intertubule bridges formed by mesh

E

Perpendicular to the microtubule axis



Parallel



https://www.scienceabc.com/nature/mitosis-vs-meiosis-how-does-cell-division-work.html

Microtubules locations as point patterns

Stephen Royle's Lab (Centre for Mechanochemical Cell Biology) asks: What is the role of TACC3 protein for the structure of microtubules within K-fibres and mesh?

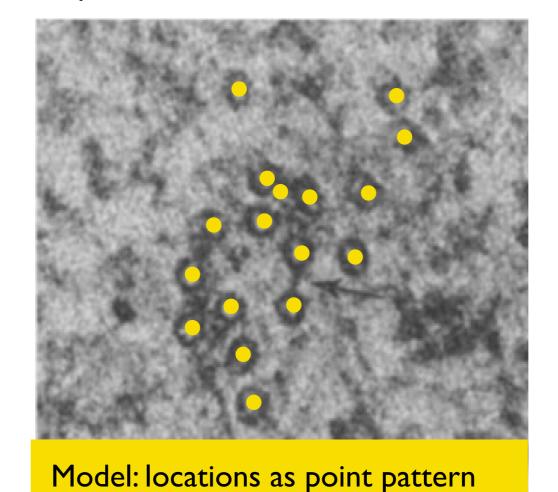
Experiment: Overexpression of TACC3 through treatment versus control.

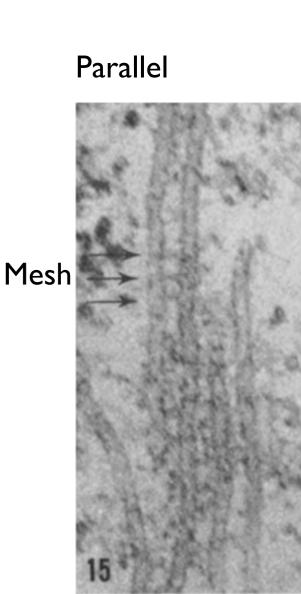
Perpendicular view

Data: Microscopic images collected in planes perpendicular to the fibre axes.



Team:
Steve Royle
Tom Honnor (now at UCL)
Adam Johnson
Julia Brettschneider



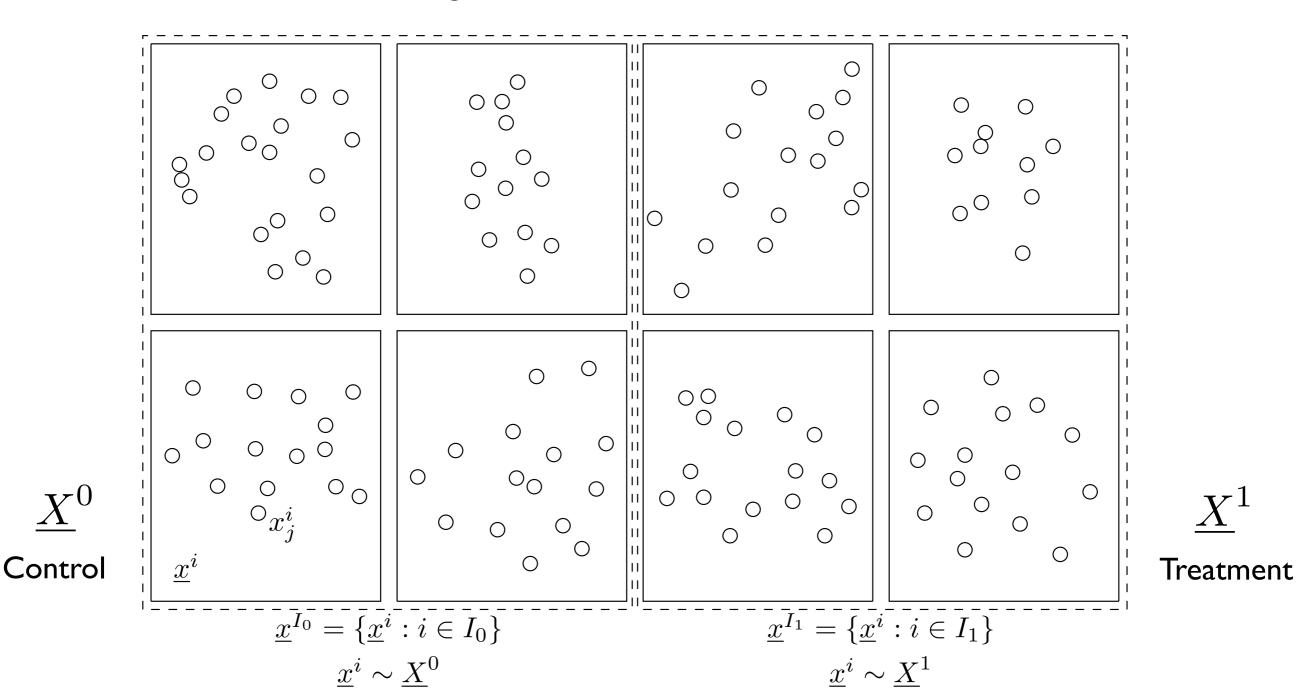


Study I: Microtubules - mathematical model

Data:

Microscopic images of treatment (n=37) versus control (n=26)

Observation window surrogate for cross sectional area of K-fibres



Point patterns models

Set of point patterns:

$$\chi_2 := \{ (\underline{x} = x_1, x_2, \dots, x_{n(\underline{x})}) : n(\underline{x}) \in \mathbb{N}, x_i \in \mathbb{R}^2 \text{ for } i = 1, 2, \dots, n \}$$

Model pattern as realisations of a point process:

Random subset \underline{X} on \mathbb{R}^2 .

For B in Borel σ -algebra $\mathcal{B}(\mathbb{R}^2)$ on $\mathbb{R}^2 : \underline{X}_B = \underline{X} \cap B$

Counts (random variable): $N(B) = n(\underline{X}_B) = \text{number of points of } \underline{X} \text{ in } B$

Intensity measure μ

$$\mu(B) = \mathbb{E}[N(B)], \quad \forall B \in \mathcal{B}(\mathbb{R}^d).$$

If for some function $\rho: \mathbb{R}^2 \to [0, \infty)$

$$\mu(B) = \int_{x \in B} \rho(x) dx, \quad \forall B \in \mathcal{B}(\mathbb{R}^d),$$

then ρ is referred to as the intensity function of X.

Summary statistics: basics

Let \underline{x} be a realisation of \underline{X} on the observation window W. Estimator for the **intensity** of \underline{X} :

$$\hat{\rho} = \frac{n(\underline{x})}{|W|}$$

Let $\operatorname{nn}(x_j)$ be the (set of) nearest neighbours of point x_j .

$$\operatorname{nn}(x_j) = \{x_k : k = \operatorname{argmin}_l ||x_l - x_j||\},\,$$

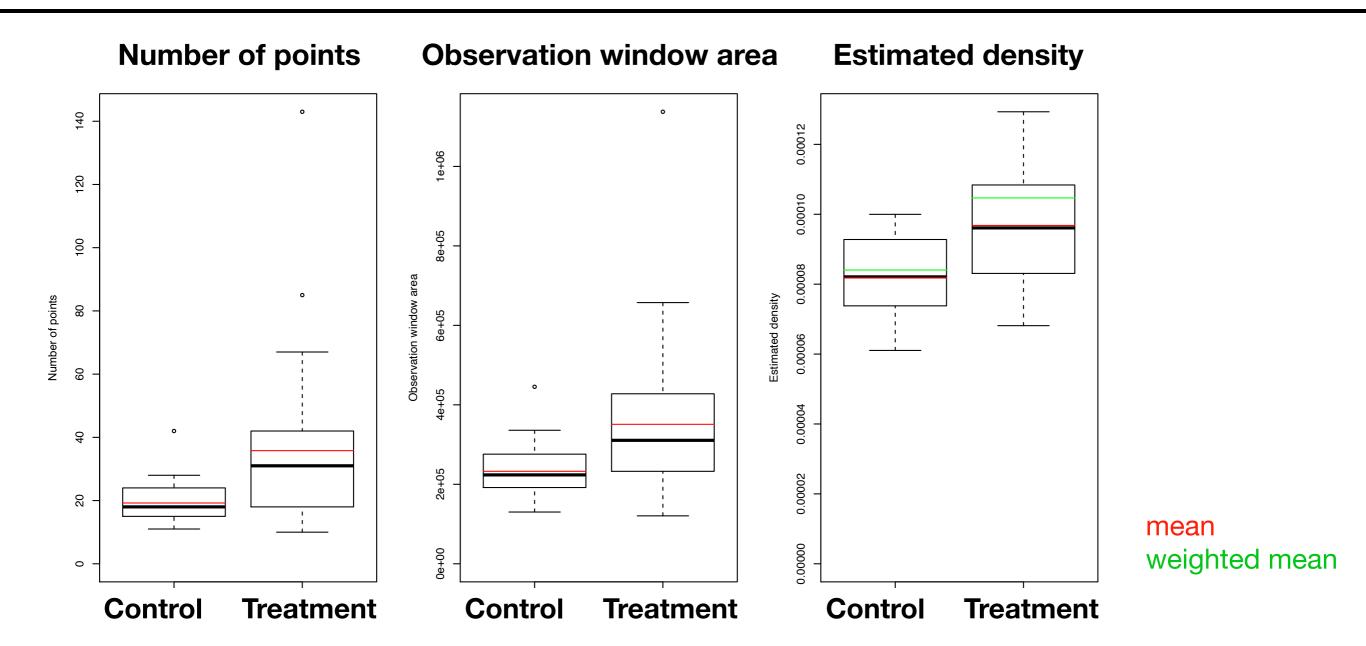
and $\operatorname{nnd}(x_j)$ its nearest neighbour distance

$$nnd(x_j) = \inf_{x \in nn(x_j)} \{ ||x_j - x|| \}.$$

Estimator for the **mean nearest neighbour distance** for X:

$$\overline{\operatorname{nnd}}(\underline{x}) = \frac{1}{n(\underline{x})} \sum_{i=1}^{n(\underline{x})} \operatorname{nnd}(x_j)$$

EDA: First order statistics



- All means/medians are greater for treatment
- Treated K-fibers are made up of a greater number of microtubules which are more closely separated within thicker K-fibers
- Weighted mean densities greater than unweighted means densities
 (i.e. K-fibers with greater numbers of microtubules are more tightly packed

Summary statistics: K-function

K-function (Ripley 1977) (scaled neighbourhood count function):

$$K(r) = \frac{1}{\rho} \mathbb{E} \left[\frac{1}{N(S)} \sum_{x_j \neq x_k \in \underline{X}} 1_{\{||x_j - x_k|| < r\}} \right]$$

Estimate:

$$\hat{K}(\underline{x},r) = \frac{|W|}{n(\underline{x})^2} \sum_{j \neq k} e_{j,k} 1_{\{||x_j - x_k|| \le r\}}$$

where $e_{j,k}$ is the proportion of the circumference of the circle with centre x_j and radius $||x_j - x_k||$ in W (edge correction).

 $K(r) = \pi r^2 : \text{CSR (complete spatial randomness)}$

 $K(r) > \pi r^2$: aggregation at distances less than r

 $K(r) < \pi r^2$: repulsion at distances less than r

Summary statistics: G-function

Nearest neighbour function (Diggle 2003):

$$G(r) = \frac{1}{\rho |B|} \mathbb{E} \left[\sum_{x \in \underline{X}_B} 1_{\{\underline{X} \setminus x) \cap b(x,r) \neq 0\}} \right]$$

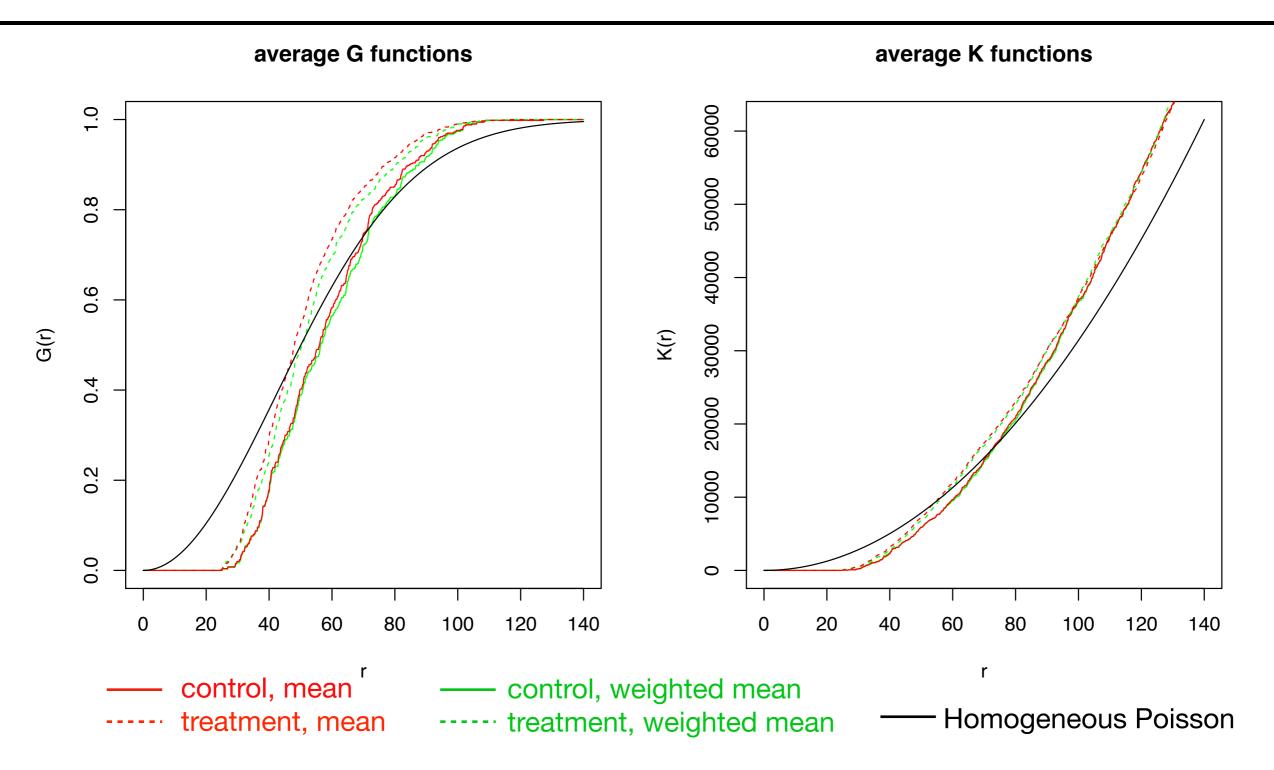
for finite B in \mathbb{R}^2 , and b(x,r) the disc centred at x with radius r. (For stationary X it is independent of B.)

Distribution of distance of randomly selected point to its nearest neighbour. Estimate:

$$\hat{G}(\underline{x}, r) = \frac{1}{n(\underline{x})} \sum_{j=1}^{n(\underline{x})} 1_{\{\operatorname{nnd}(x_j) \le r\}}$$

If <u>X</u> is completely spatially at random then $G(r) = 1 - \exp(-\rho \pi r^2)$

EDA: Spatial functions



- Some evidence of clustering at larger length scales
- Effect of limitation of nnd in [25,105]
- Difference between weighted mean and unweighted mean negligible

Test statistics based on basic observations

Pattern size test statistic:

$$\delta_N(I) = \frac{1}{|I_0|} \sum_{i \in I_0} n(\underline{x}^i) - \frac{1}{|I_1|} \sum_{i \in I_1} n(\underline{x}^i)$$

Observation window statistic:

$$\delta_W(I) = \frac{1}{|I_0|} \sum_{i \in I_0} |W^i| - \frac{1}{|I_1|} \sum_{i \in I_1} |W^i|$$

Intensity test statistic:

$$\sum_{i \in I_0} \omega_0(\underline{x}^i) \hat{\rho}(\underline{x}^i) - \sum_{i \in I_1} \omega_1(\underline{x}^i) \hat{\rho}(\underline{x}^i)$$

where $\delta_{\rho}(I)$ denotes unweighted case using $\omega_{k}(\underline{x}^{i}) = 1/|I_{k}|$ (k = 0, 1) $\delta_{\rho,\omega}(I)$ denotes weighted case using $\omega_{k}(\underline{x}^{i}) = n(\underline{x}^{i}) / \sum_{j \in I_{k}} n(\underline{x}^{j})$ (k = 0, 1)

Test statistics based on G-functions

Estimated nearest neighbour functions averaged over the collection of point patterns \underline{x}^J with weights ω_J as above:

$$\hat{G}(\underline{x}^{J}, r) = \sum_{i \in J} \omega_{J}(\underline{x}^{i}) \hat{G}(\underline{x}^{i}, r)$$

Nearest neighbour distribution test statistic statistics:

$$\delta_{G,1}(I) = ||\hat{G}(\underline{x}^{I_0}, r) - \hat{G}(\underline{x}^{I_1}, r)||_1 = \int_0^\infty |\hat{G}(\underline{x}^{I_0}, r) - \hat{G}(\underline{x}^{I_1}, r)| dr$$

$$\delta_{G,\infty}(I) = ||\hat{G}(\underline{x}^{I_0}, r) - \hat{G}(\underline{x}^{(I_1}, r))||_\infty = \sup_r |\hat{G}(\underline{x}^{I_0}, r) - \hat{G}(\underline{x}^{I_1}, r)|$$

For comparison of $\hat{G}(\underline{x}^{I_0}, r)$ and $\hat{G}(\underline{x}^{I_1}, r)$ across the range of distances r > 0.

Also, scaled neighbourhood count test statistic (Diggle 2000).

Significance quantification

- Based on permutation tests (nonparametric)
- Need exchangeability under the Null under suitable set of operations
- Statistics under permutations are identically distributed
- p-values are uniformly distributed (test e.g. with KS)
- Exact or approximate (subset of operations)
- Evaluated in simulations studies

Test statistics

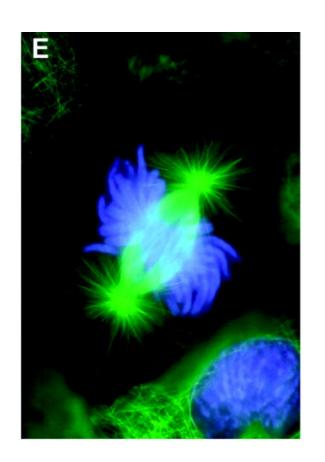
Observations of exploratory analysis can be confirmed by formal testing. All proposed test statistics show significant results:

δ_N	0.0005	$\delta_{ m nnd}$	0.0057	δ_K	0.1092	δ_{EFT}	0.0011
δ_W	0.0018	$\delta_{\mathrm{nnd},\omega}$	0.0005	$\delta_{G,1}$	0.0061	$\delta_{EFT,\omega}$	0.0005
$\delta_{ ho}$	0.0001	$\delta_{ m msd}$	0.0019	$\delta_{G,1,\omega}$	0.0005		
$\delta_{ ho,\omega}$	0.0002	$\delta_{\mathrm{msd},\omega}$	0.0005	$\delta_{G,\infty}$	0.0087		
				$\delta_{G,\infty,\omega}$	0.0013		

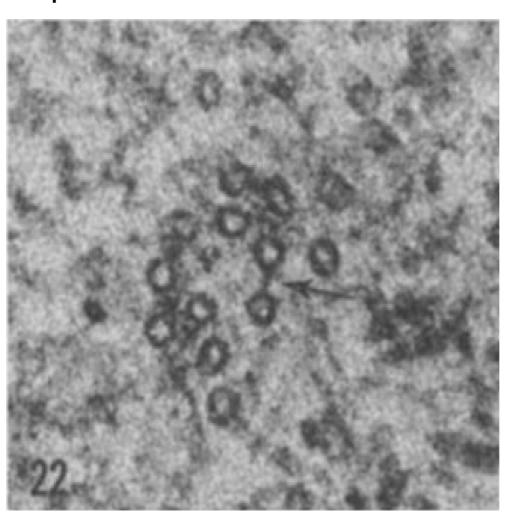
Results remain significant after multiple testing adjustment of critical p-value (using Bonferroni).

What did we find?

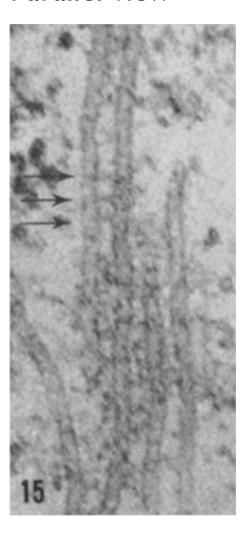
- Microtubules are **bound together** (in K-fibers, by mesh-like structure)
- TACC3 overexpression is associated with an impact on the mesh
- Detection of treatment effects not visible by eye



Perpendicular to the microtubule axis



Parallel view



References

The (ongoing) work on modelling Covid with Hawkes process is joint with Adam Davison and Marianna Mavroleftherou (integrated Masters students at Warwick Statistics)

Dead pixels and other work related to digital X-ray imaging:

JA Brettschneider, OT Giles, WS Kendall, T Lazauskas *DetectorChecker: analyzing patterns of defects in detector screens*<u>Journal of Open Source Software, 2020, 5(56), 2474</u>

JA Brettschneider, JW Warnett, TE Nichols, WS Kendall

Higher level spatial analysis of dead pixels on detectors based on local grid geometry

CRiSM Working Paper Series No. 17-02, 2017

Kueh A, Warnett JM, Gibbons GJ, Brettschneider J, Nichols TE, Williams MA, & Kendall WS *Modelling the Penumbra in Computed Tomography*Journal of X-ray science and technology, 24 (4), 2016, 583-97 (Gold Access)

Brettschneider J, Thornby J, Nichols TE and Kendall WS Spatial analysis of dead pixels CRiSM Working Paper Series No. 14-24, 2014

Software

<u>DetectorCheckerWebApp</u> is an interactive WebApp for analysing pixel damage in CT scanners using the associated R-package <u>DetectorChecker</u>. Both have been developed jointly with The Alan Turing Institute and Prof Wilfrid Kendall. The project emerged as a spin off from our EPSRC <u>inside-out blog</u>. Initial versions of the software were issued on 29.3.2019 and new versions were released on 30.6.2020.

References

Modelling for microscopic images is joint work with Tom Honnor (Warwick Statistics, now UCL), Adam Johnson (Warwick Statistics), Steve Royle (Warwick Medical School)

Thomas R. Honnor, Julia A. Brettschneider, Adam M. Johansen (2017), **Differences in spatial point patterns with application to subcellular biological structures**

https://warwick.ac.uk/fac/sci/statistics/crism/research/17-01/

Honnor TR, Johansen AM and Brettschneider JA. (2017)

A nonparametric test for dependency between estimated local bulk movement patterns

https://warwick.ac.uk/fac/sci/statistics/crism/research/17-03/

Nixon*, F.M., Honnor*, T.R., Starling, G.P., Beckett, A.J., Johansen, A.M., Brettschneider, J.A., Prior, I.A. & Royle, S.J. J. Cell Science, April 2017

Microtubule organization within mitotic spindles revealed by serial block face scanning EM and image analysis