Warwick Statistics

WARWICK
THE UNIVERSITY OF WARWICK

Julia Brettschneider

Statistical methodology and data science in science and engineering I

1. Applications involving point processes

1.1 Hawkes process modelling of the pandemic

1.2 Dead pixel formations on digital X-ray detectors https://warwick.ac.uk/fac/sci/statistics/staff/academic-research/brettschneider/

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**

This was my first slide at my Warwick job talk in 2007!

My path

Research in applied statistics:

methods for statistical analysis of high-dim. molecular measurements (pre-processing, QA)

now

Research in probability/ theoretical statistics:

Measure valued diffusion proc. & quasilinear PDE,

Master's

Large deviations for random fields

PhD

Postdoc

Learning genomics and statistics:

molecular biology basics, genetics, genomics, high-throughput measurement technology

"Wege entstehen indem man sie geht." Franz Kafka
(Paths are created by walking them.)

Preamble: Applied mathematics as bridge

*"The **instrument that mediates between theory and practice**, between thought and observation, is mathematics; it builds the **connecting bridge** and makes it stronger and stronger. Thus it happens that our entire present-day culture, insofar as it rests on intellectual insight into and harnessing of nature, is founded on mathematics."*

David Hilbert

In Königsberg on 8 September 1930, David Hilbert addressed the yearly meeting of the Society of German Natural Scientists and Physicians (Gesellschaft der Deutschen Naturforscher und Ärzte).
Full text of the speech in English and German at url below, including audio file:

http://math.sfsu.edu/smith/Documents/HilbertRadio/HilbertRadio.pdf

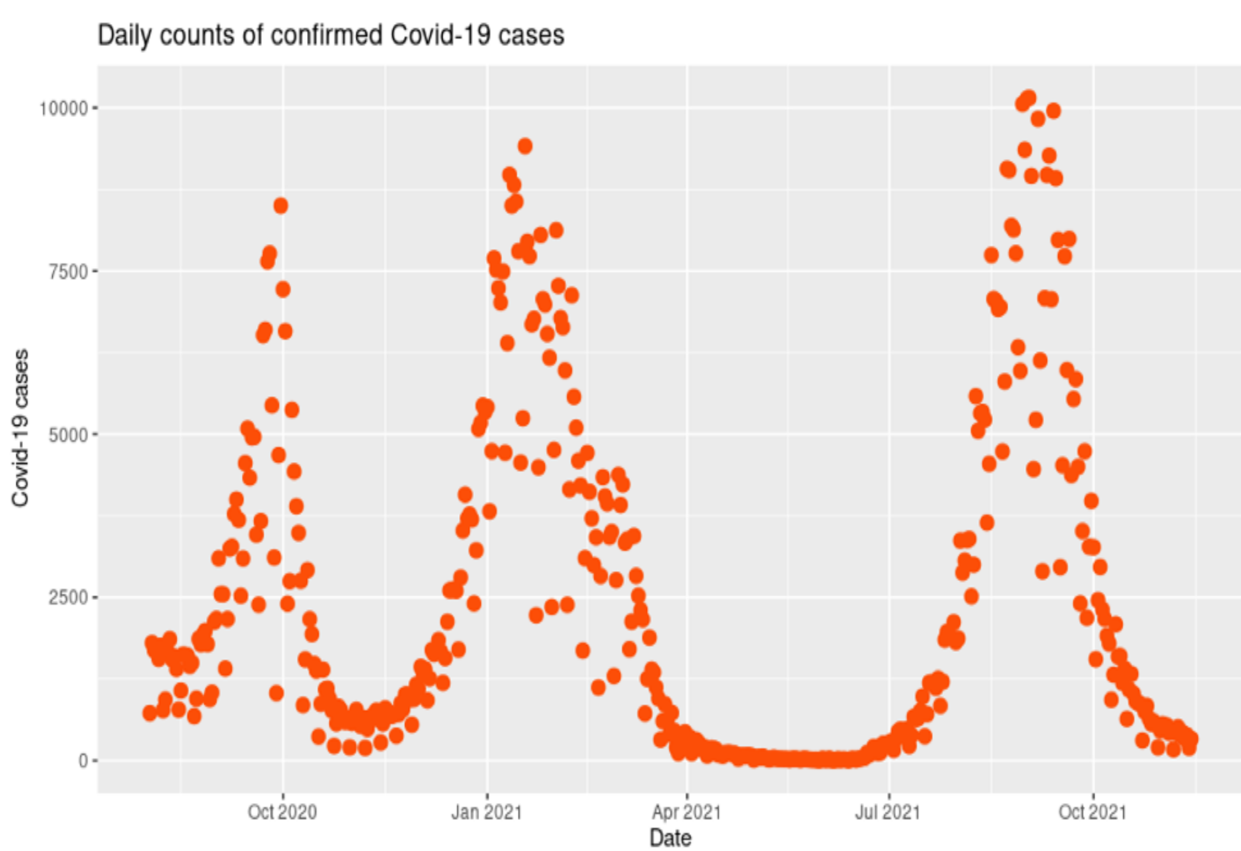
Hawkes process modelling of the pandemic

Ongoing joint projects with integrated Master's students on COVID-19 data

1. Data from Israel

Vaccine for data deal with Pfizer/BioNTech

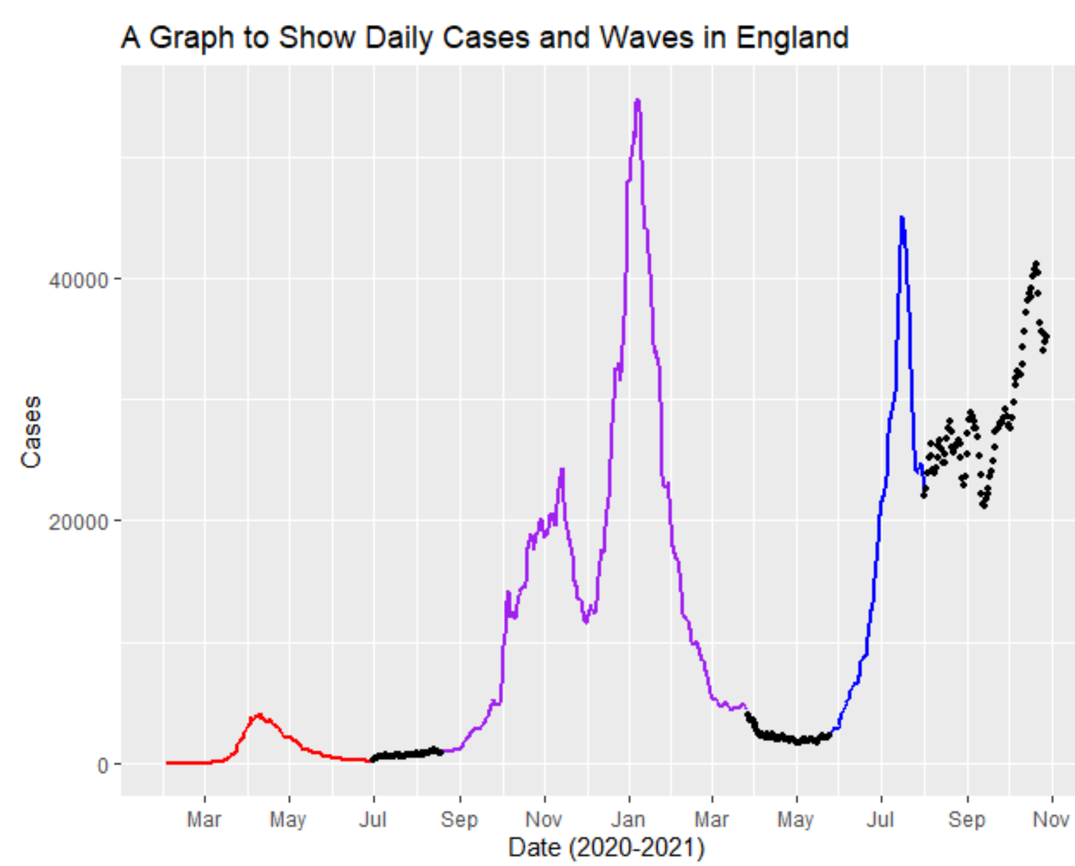
(Marianna Mavroleftherou's project)



2. Data from England

ONS/Oxford survey

(Adam Davison's project)



Hawkes Process I

Definition: Point Process

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

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

Definition: Temporal Point Process

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

i.e. a sequence of events $t = \{t_1, t_2, \ldots, t_d\}$ s.t. $0 \leq t_1 < t_2 < \ldots < t_d$

Hawkes Process II

<u>Definition: Hawkes Process</u>

A Hawkes Process *N* is defined to be a self-exciting temporal point process, where *N(t)* represents the number of events **<u>up to time *t.*</u>**

And is fully described by the conditional intensity function λ:

$$\lambda(t) = \underbrace{\mu}_{\substack{\text{Background intensity}\\ \text{Or}\\ \text{Baseline mean}}} + \underbrace{\alpha \sum_{i:t_i<t} y_{t_i} \overbrace{g(t-t_i)}^{\substack{\text{Triggering kernel}\\ \text{Or}\\ \text{decay function}}}}_{\text{Excitation intensity}}$$

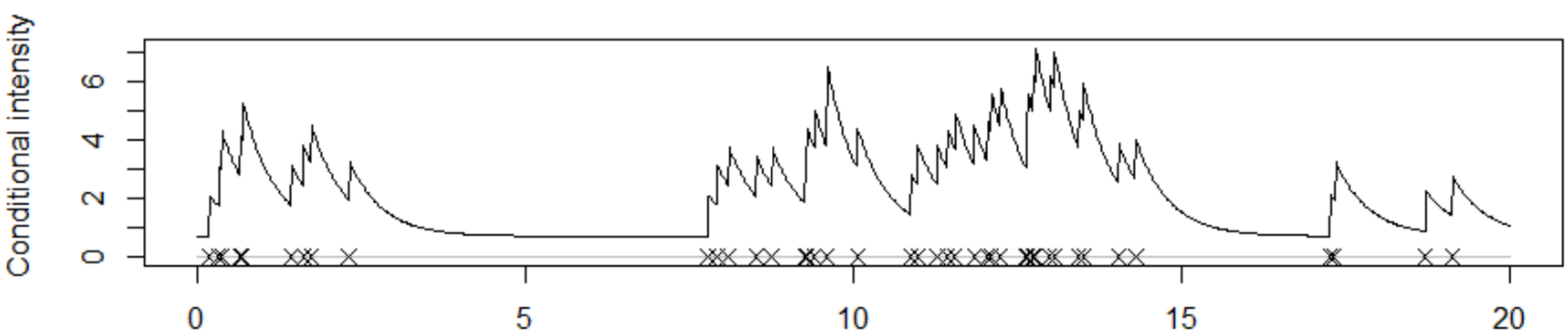
ETAS Hawkes Process model

ETAS = Epidemic-type aftershock sequence

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

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

Example:



Model Setup I

- We want to model the number of cases per day,
  i.e. $Y(t)$ where each time period t is a day.

- Self – excitation property: Hawkes Process suitable for modelling Covid-19 cases.

- We focus on ***Discrete Time Hawkes Process (DTHP)***
  - Enables the use of data collected in daily counts.
  - Avoids artificially imputed data

Model Setup II

- It is a natural response to assume that ***Y(t)* follows *Poisson(λ(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:

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

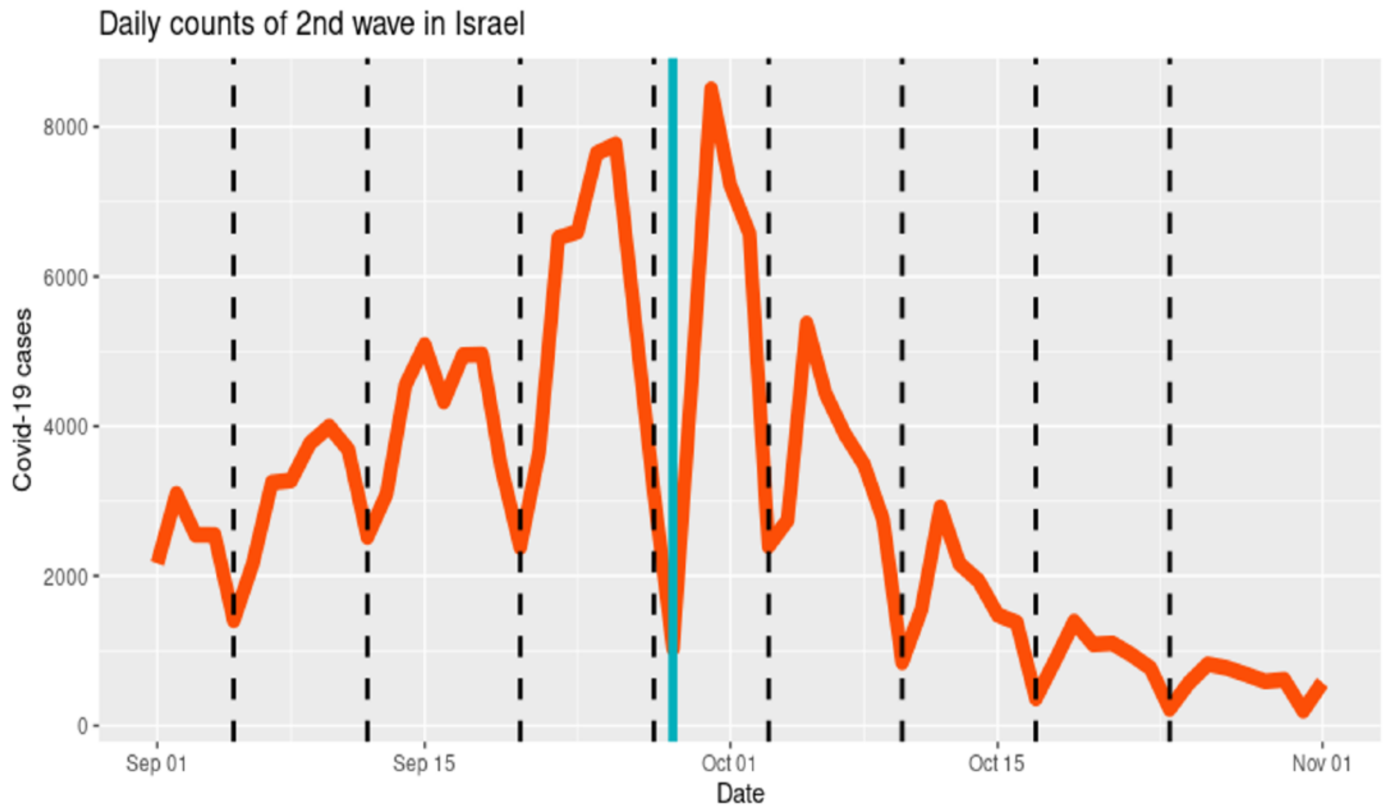
Since it can be shown to be the generalization of the exponential distribution in discrete time.

Hawkes process modelling of the pandemic

Model fitting:

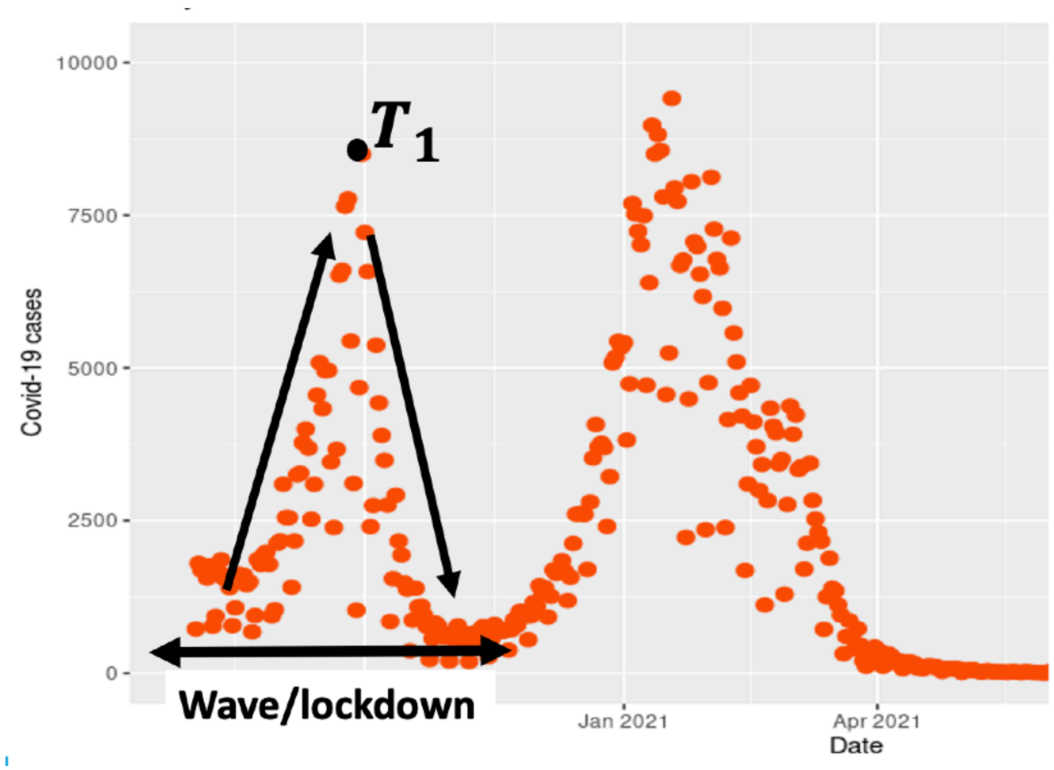
1. Data preprocessing

Small counts, smoothing



2. Waves

Piecewise fitting, piques



3. Parameter estimation

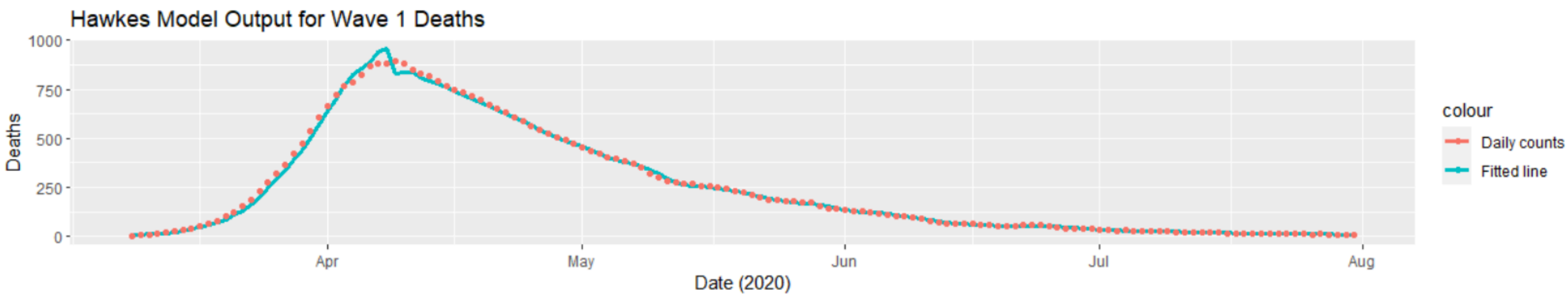
Likelihood function

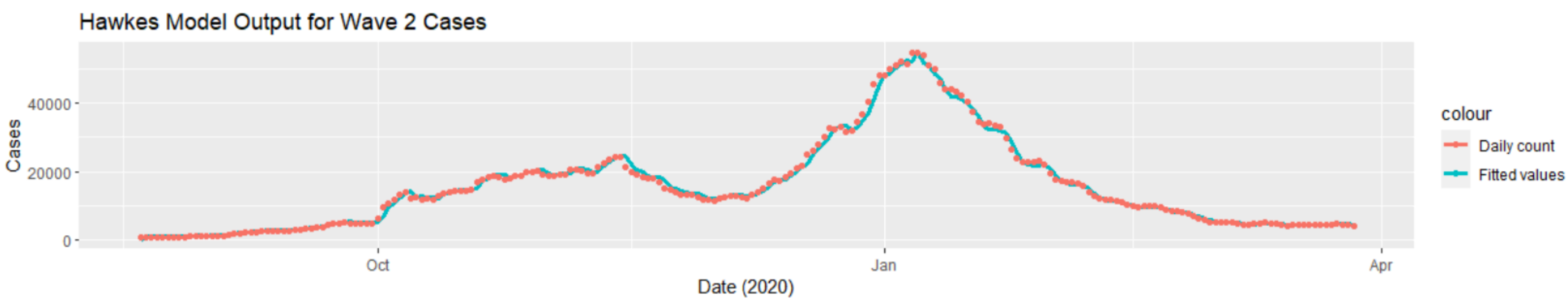
4. Maximise likelihood

Optimisation algorithm
(e.g. Nelder)

Hawkes process modelling of the pandemic

Preliminary results (English data)





Hawkes process modelling of the pandemic

Future work (e.g. miniproject):

- Alternative models including age
- Including vaccination rates
- Behavioural indicators (e.g. google searches or mobility)
- Lockdown effects
- Events (e.g. football, holidays) and interaction with other factors
- Mixture population
- Region (England data has regional resolution and postcode, Israel has regions and cities)
- Additional countries (e.g. Germany, also has regional)
- Different algorithms to determine parameters (e.g. MCMC)

About right and wrong models

What are models for? Prediction and explanation. Answering questions…

> *"Far* ***better an approximate answer to the right question****, which is often vague, than the exact answer to the wrong question, which can always be made precise."*
>
> John Tukey (1915-2000)
> American statistician (FFT, various statistical tests, EDA)

How good is a model?

> *"All models are wrong, some are* ***useful****."*
>
> George Box, FRS (1919-2013)
> English statistician (quality control, time series, design of experiments, response surfaces, Bayesian inference etc)

Rephrase "how good":

How good is it *at the task what you want it to do (prediction and/or explanation)*?

Dead pixel formations on digital X-ray detectors

Inside-out

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

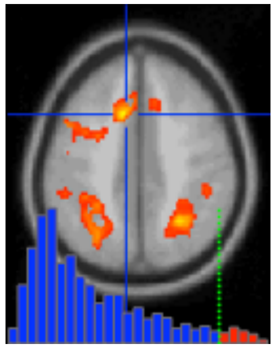
PI: Prof W Kendall
Other investigators: Prof M A Williams, Dr G J Gibbons,
Dr J Brettschneider, Prof T Nichols

EPSRC grant
3 years

10/2013 -
9/2016

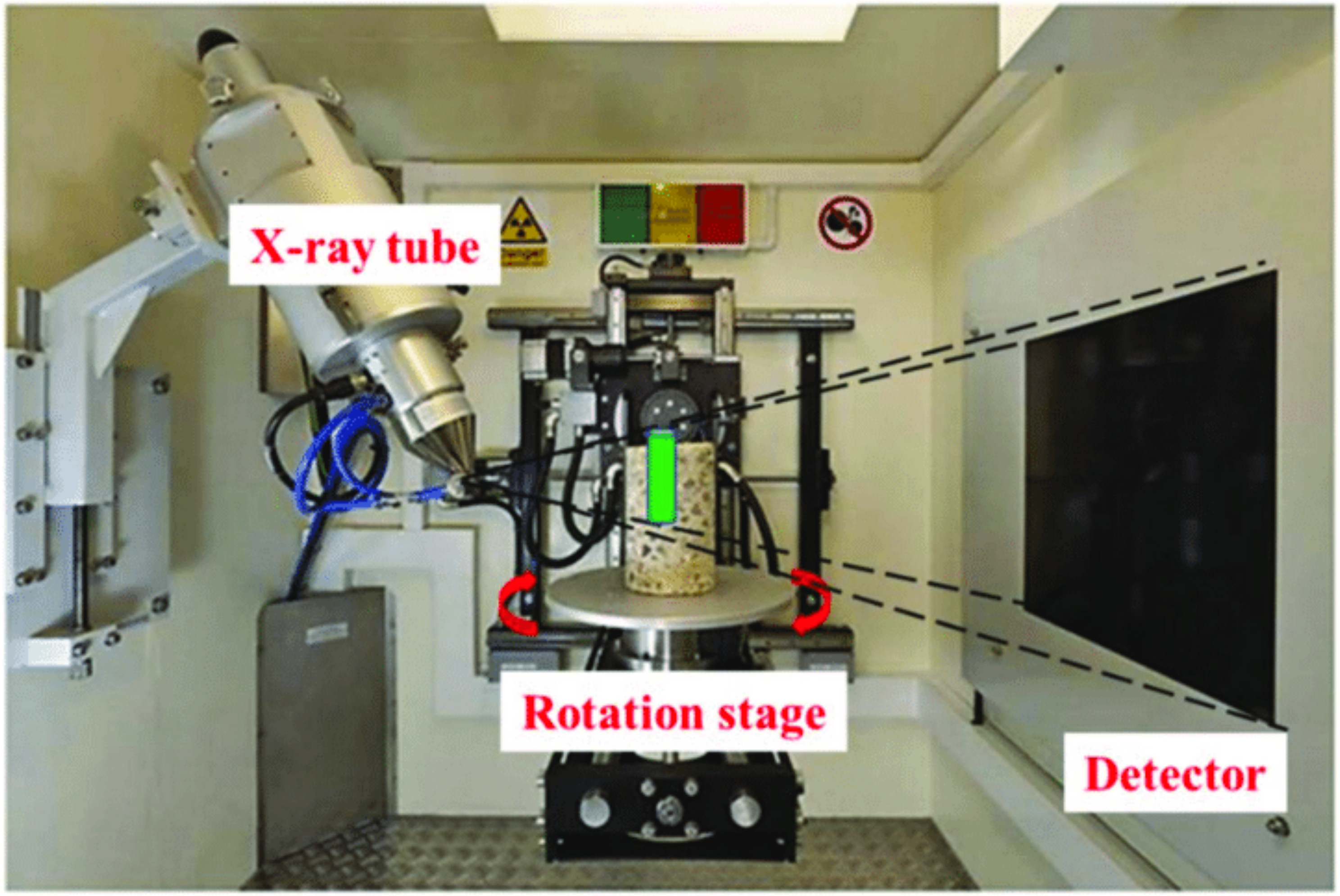
EP/K031066/1

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

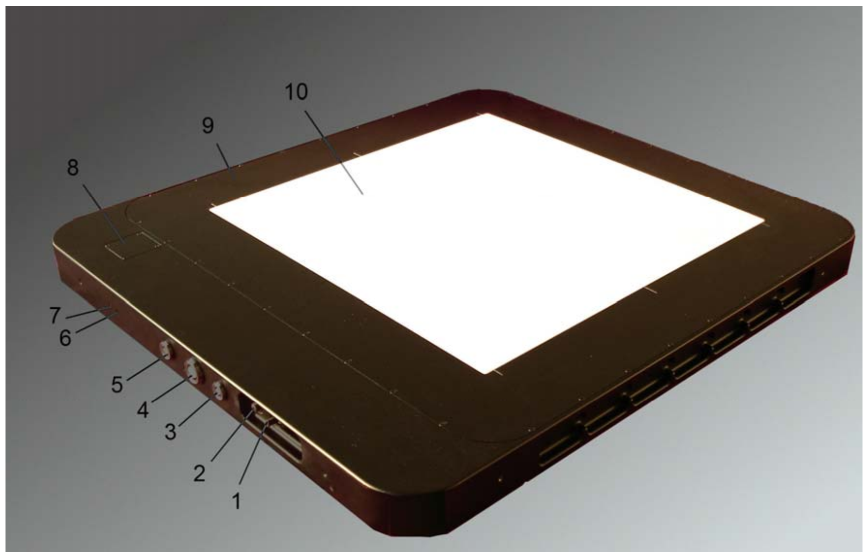



X-ray chamber



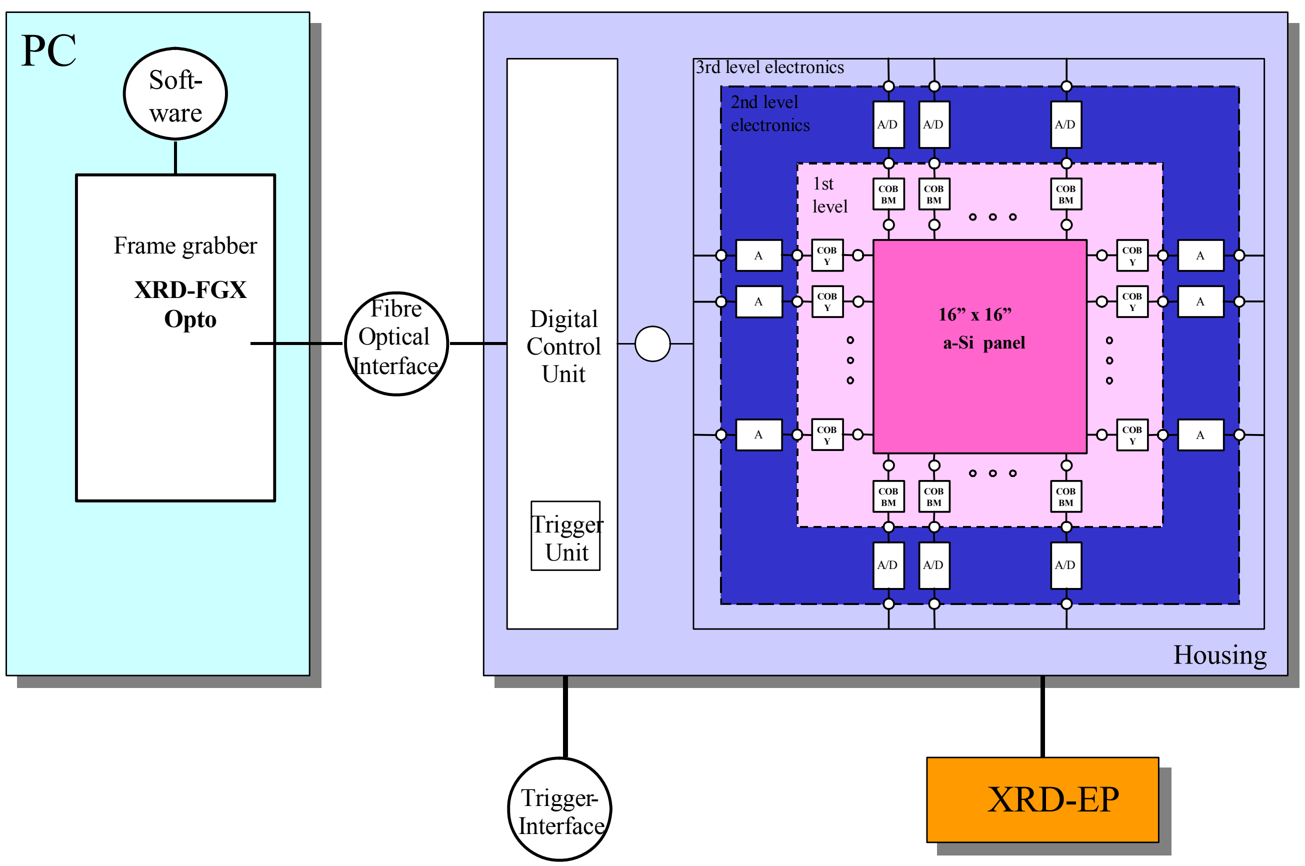
https://www.researchgate.net/figure/Example-of-an-industrial-computed-tomography-CT-system_fig1_324511614

X-ray detector



Perkin Elmer
XRD 1621

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32



Readout groups (ROG):
Upper groups transferred first, starting read out from the upper row.
Lower groups starting from the last row.

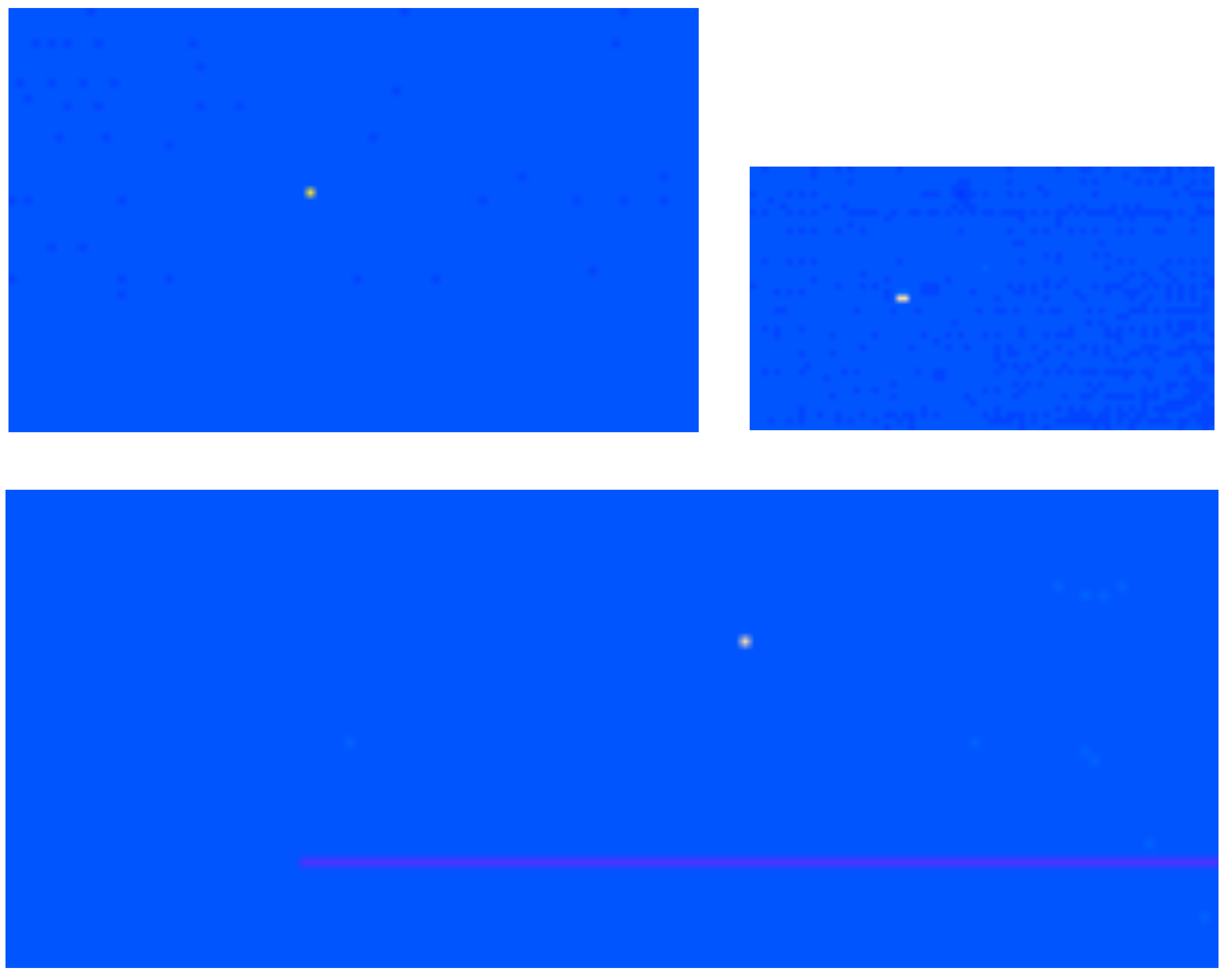
Bad pixel maps

- Criteria for "underperforming" (Perkin Elmer):
  - Signal sensitivity (at different energies)
  - Noise observed in sequence of 100 bright/dark images
  - Uniformity (global, local)

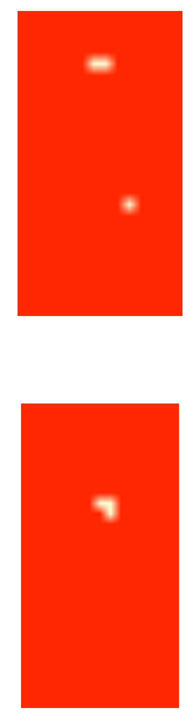
- Each bad pixel map consist of a total of 10 files:
  - White images: mean, min, max, sd (.tif)
  - Grey images: mean (.tif)
  - Black images: mean, min, max, sd (.tif)
  - Bad pixel list of locations (.xml)

Local defects: Isolated dead pixels

Singles, doubles, small clusters



A_0: Grey image [R]

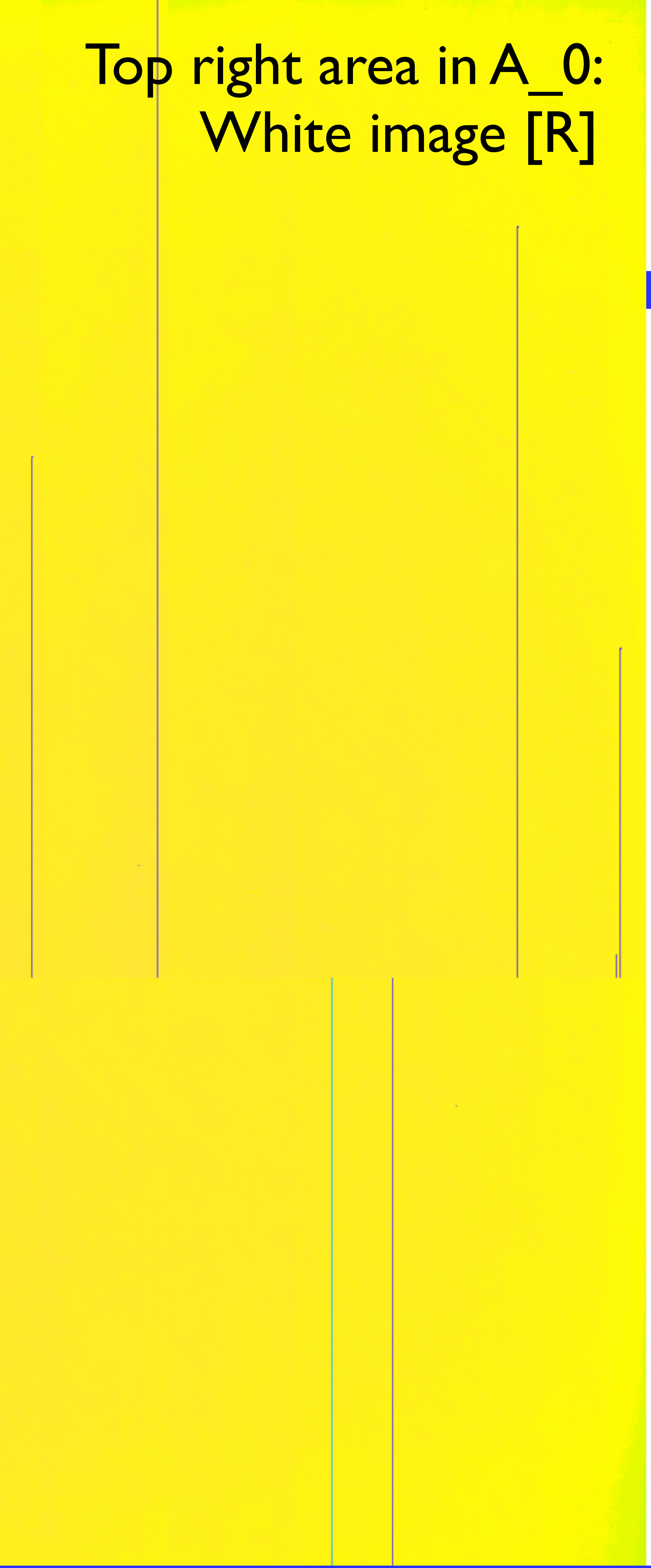


A_0:
bp binary
image [R]

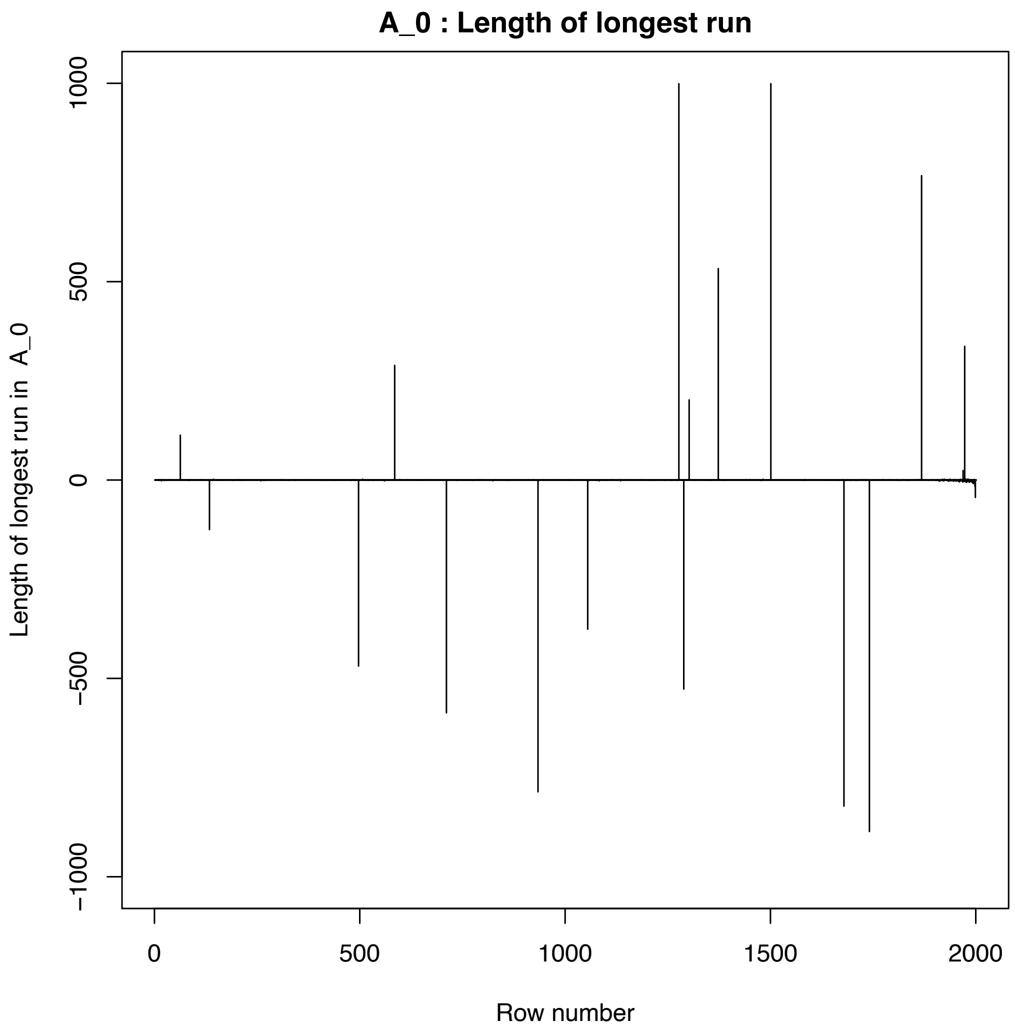
A_0: Black
image [R]

Local defects: Dead lines

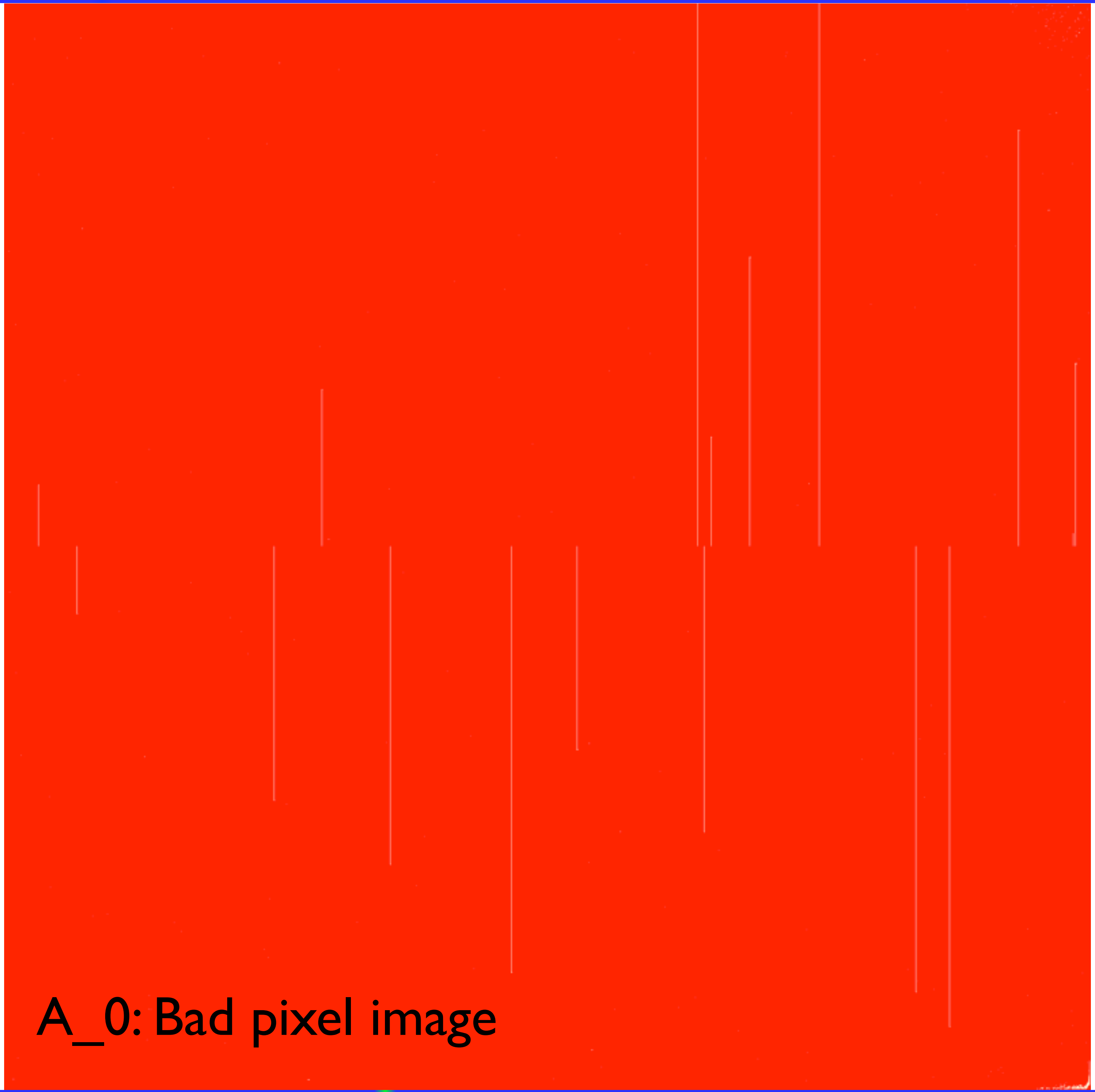
- Lines on bad pixel images
- From centre horizontal line outwards
- Visible on tif images of channel(s), too



Local defects: Locations of dead lines



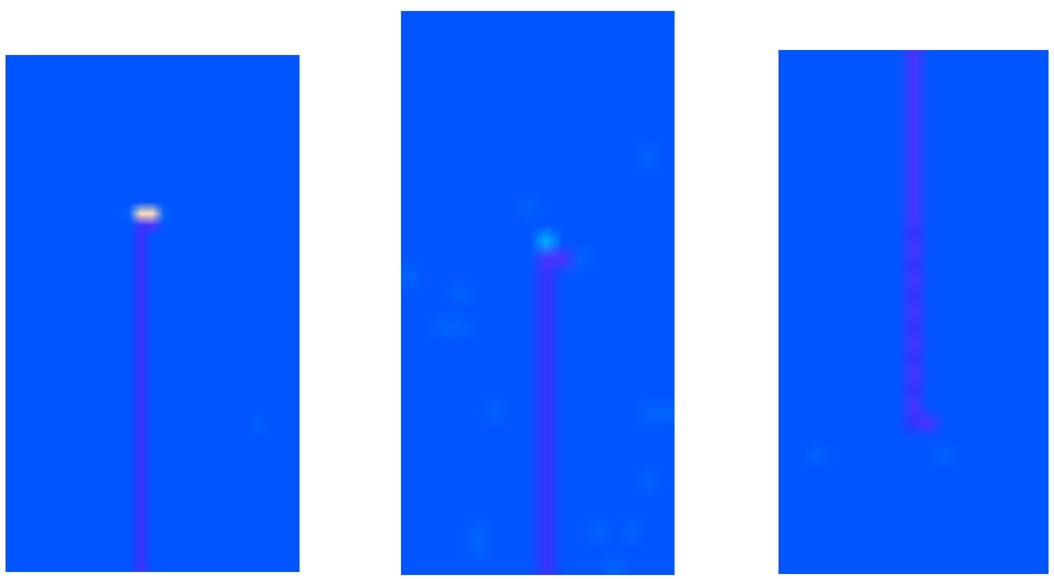
A_0: Graph of bad pixel images



A_0: Bad pixel image

Local defects: Ends of dead lines

bmp binary image

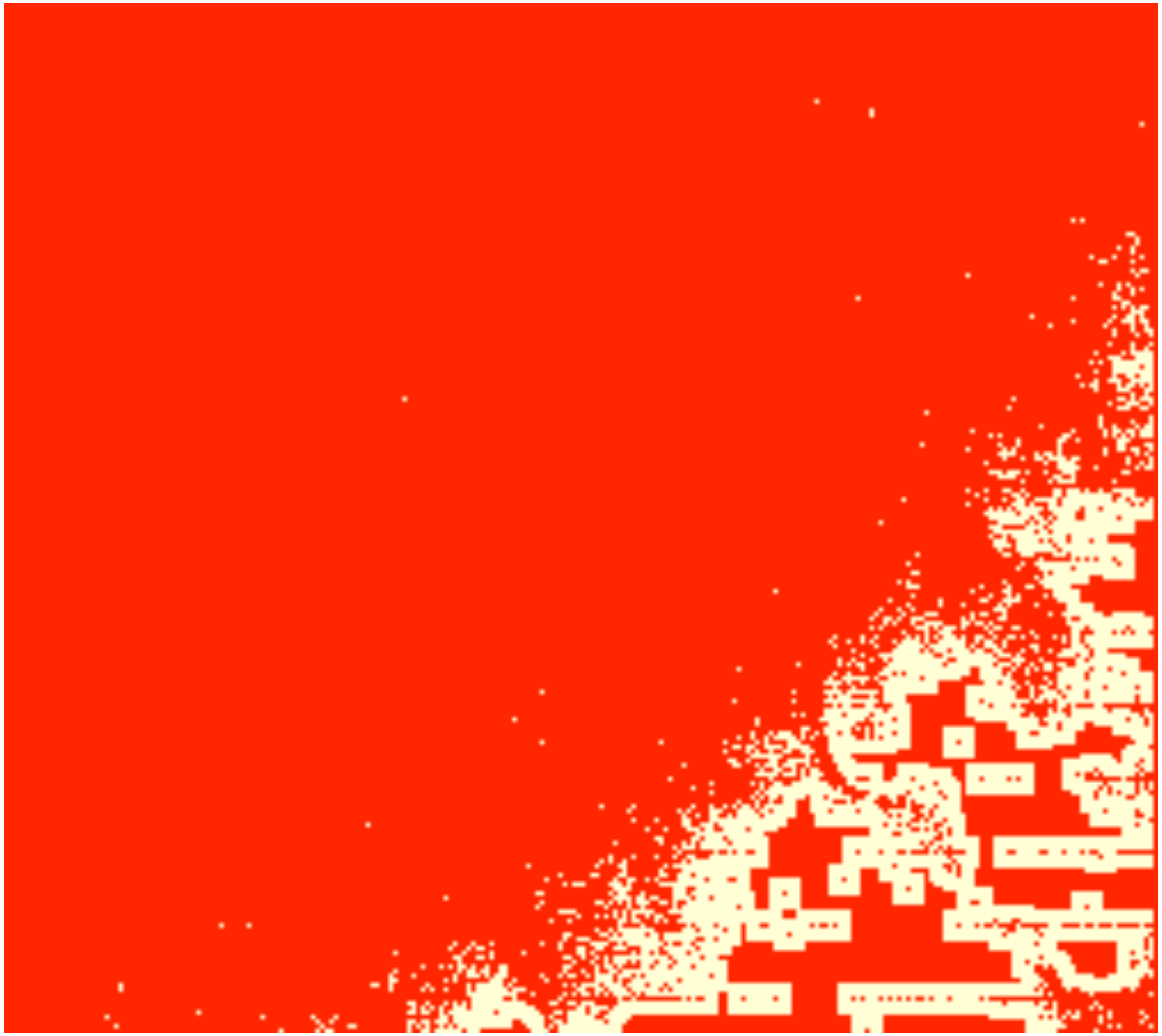


Black image

- Most lines end in small cluster pointing to the right
- Lines are composed of dark pixels
- Lines have constant intensity, except end may differ

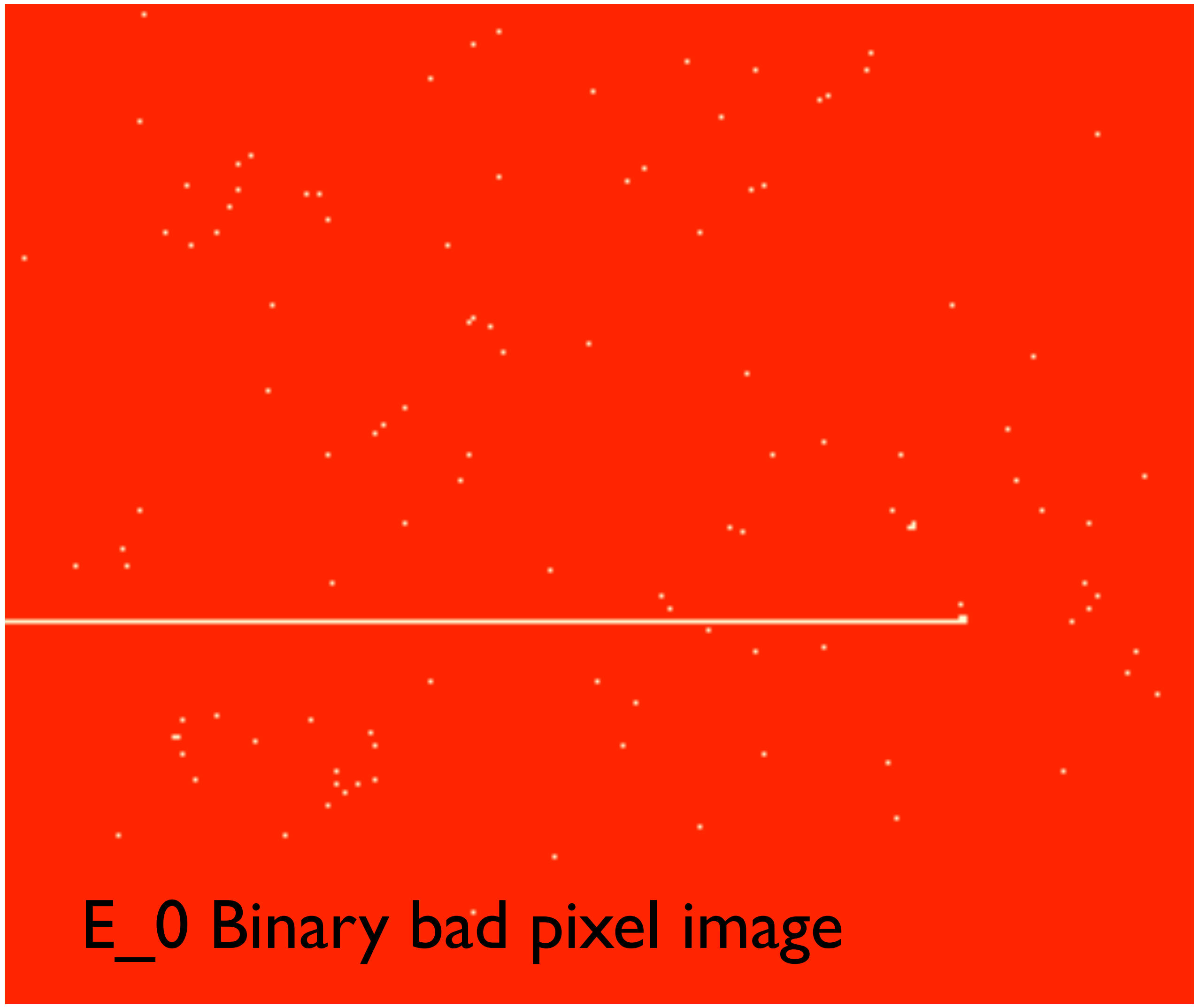
Local defects: Corners

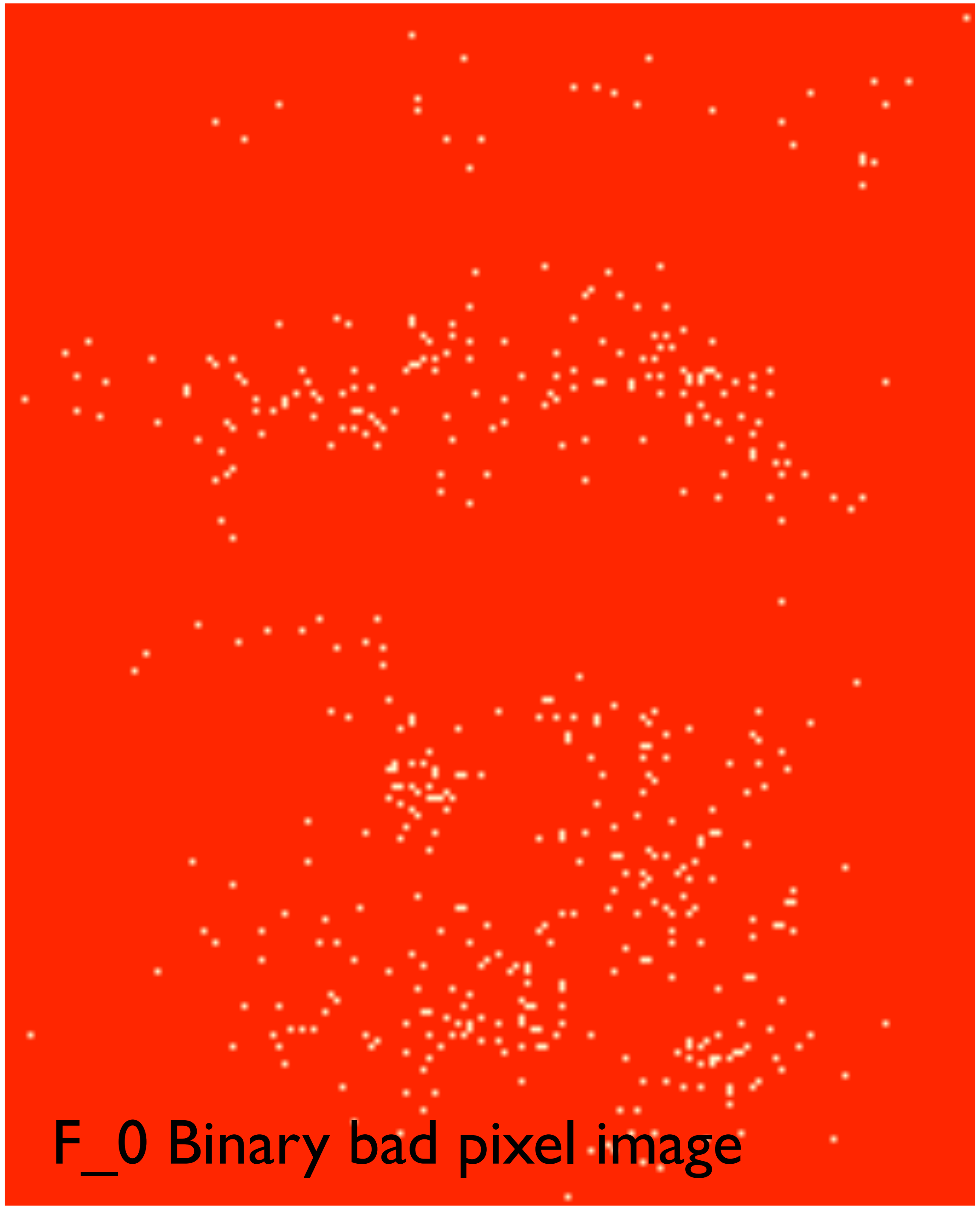
B_0: Binary bad pixel image [R]



Local defects: Patches

- Areas with high density area of bad pixels





Which model for this type of spatial data ?

Three common types described by Cressie (1993)

Geostatistical data:
Fixed study region with a random variable (observed or unobserved) in every location.
e.g. *UK with rainfall*

Lattice data:
Collection of fixed (nonrandom) set of points in study region with a random variable defined in each of them.
e.g. *Ising model on a lattice, crime in snap points*

Spatial point patterns:
Spatial locations of the observations are random, with observations itself deterministic (=1) or itself random variables.
e.g. *locations of bird nests, same with number of eggs in each nest*

Spatial model for dead pixels

Lattice or point pattern?

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)

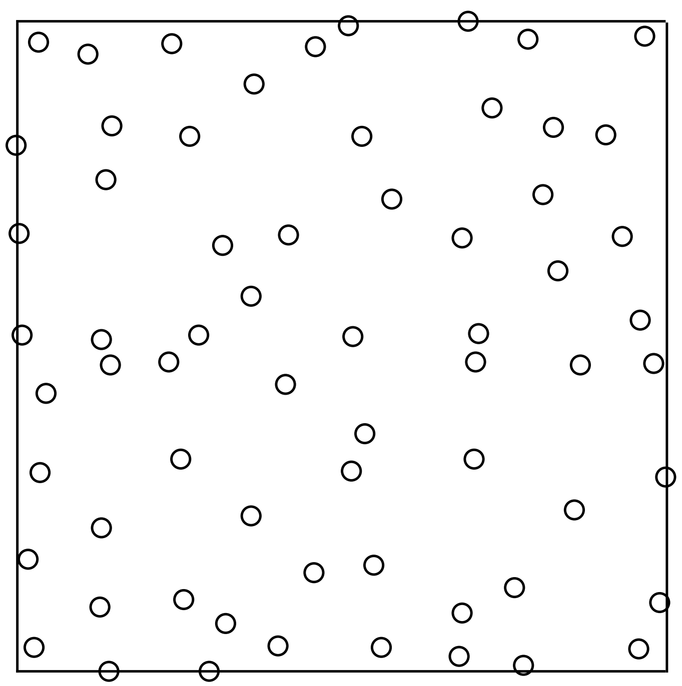
Objectives:

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

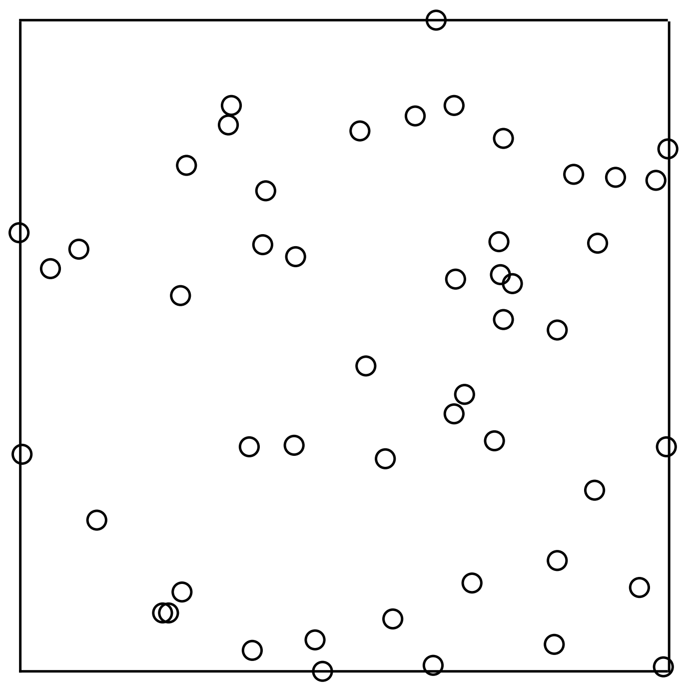
For example, look at CSR…

Complete spatial randomness (CSR)

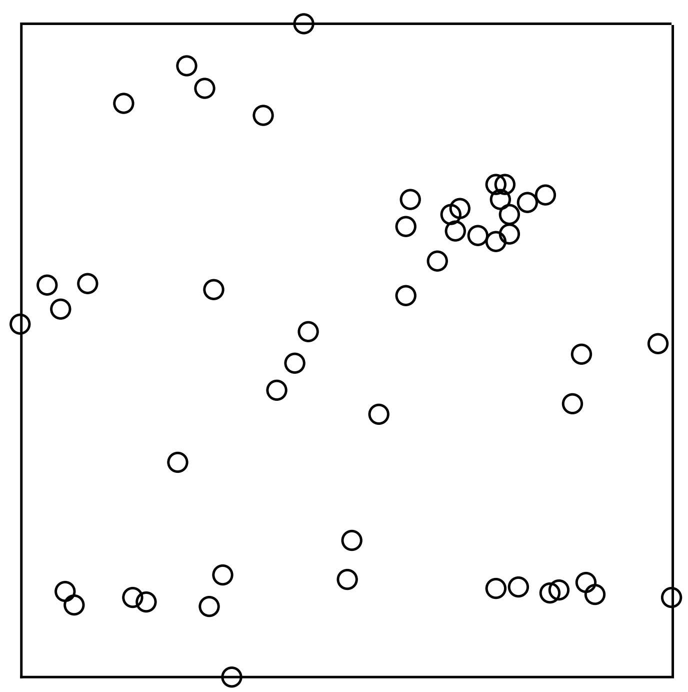
CSR: Points are distributed independently and homogeneously, as in a homogenous Poisson process.



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

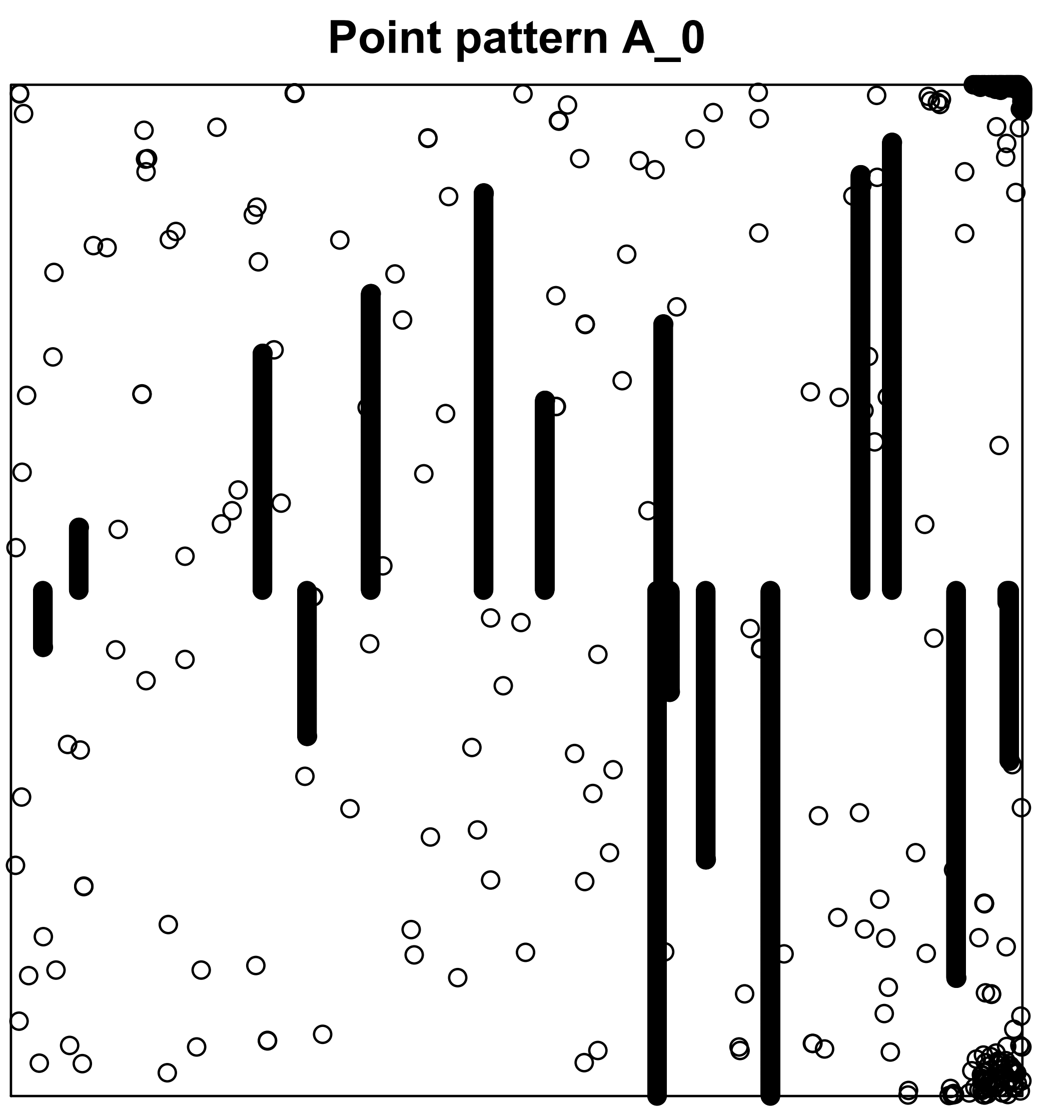
$$K(r) = \lambda^{-1} E[N_0(r)]$$

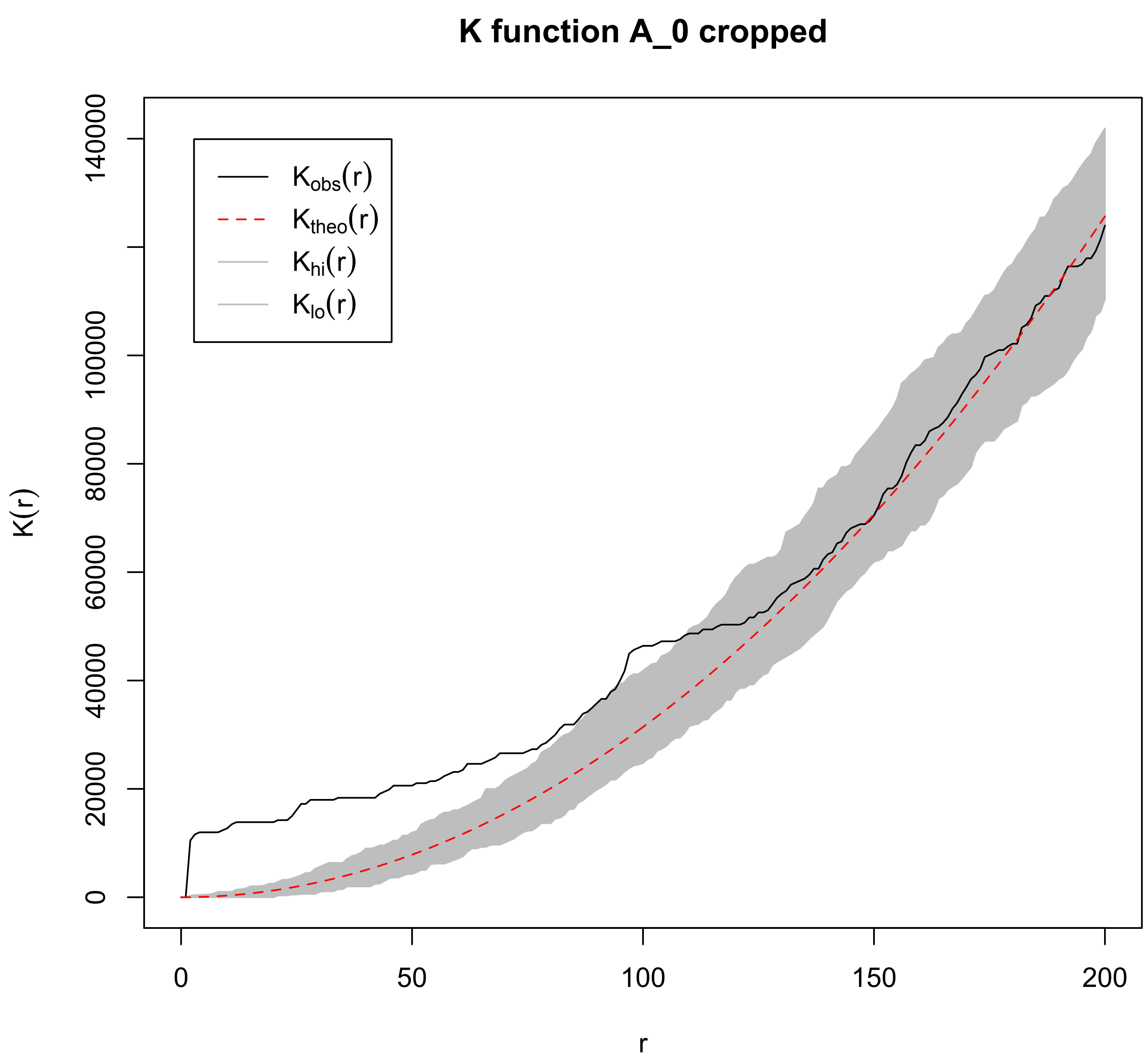
$N_0(r)$ number of points within distance r from arbitrary point

λ globally estimated density

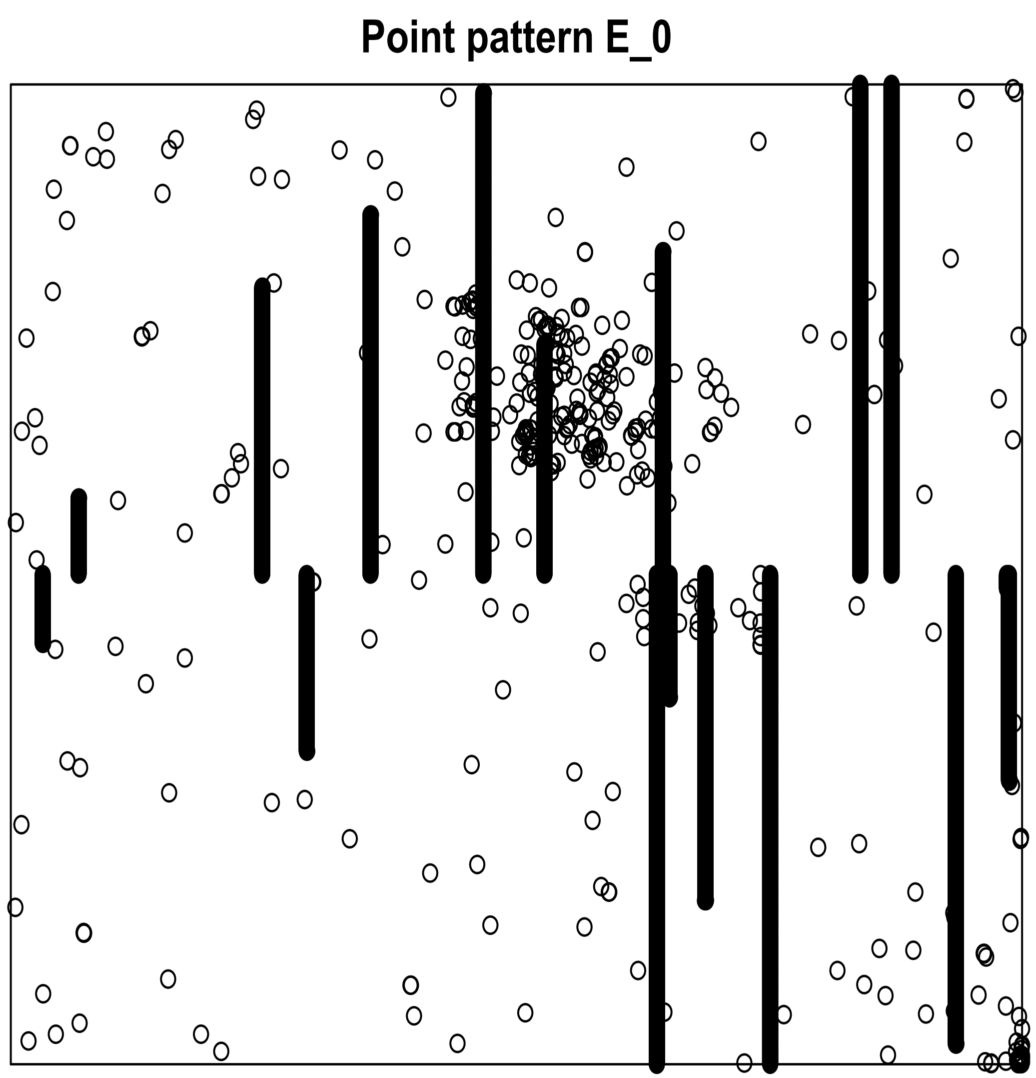
Under CSR: $K(r) = \pi r^2$

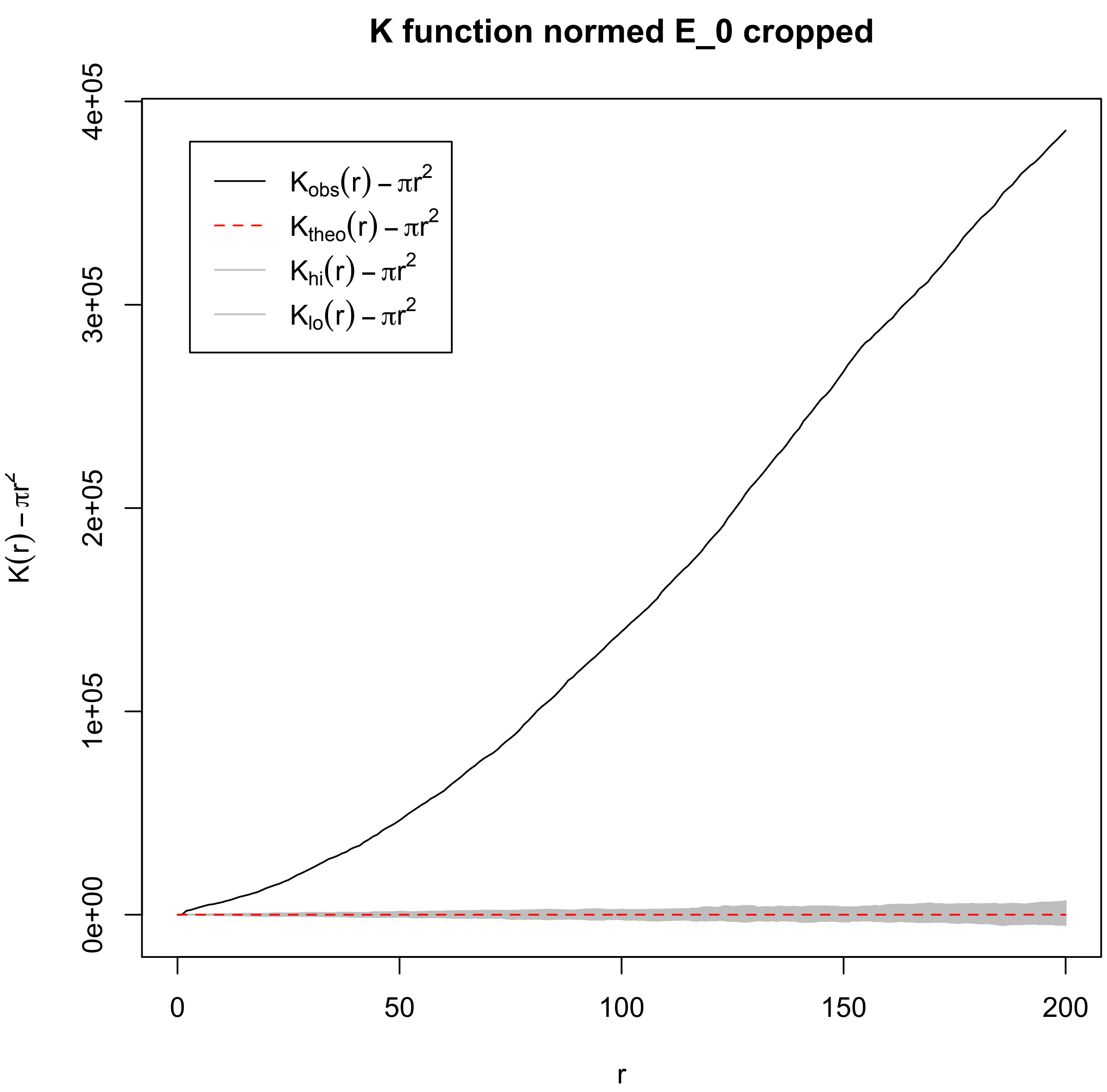
Point pattern and K-function





Point pattern and K-function





Are we asking the right question?

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

Step 1:

Convert point process into *event process* by

- Reducing lines to their endpoint
- Reducing clusters to their centre point

Are we asking the right question?

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

Step 1:

Convert point process into *event process* by

- Reducing lines to their endpoint
- Reducing clusters to their centre point

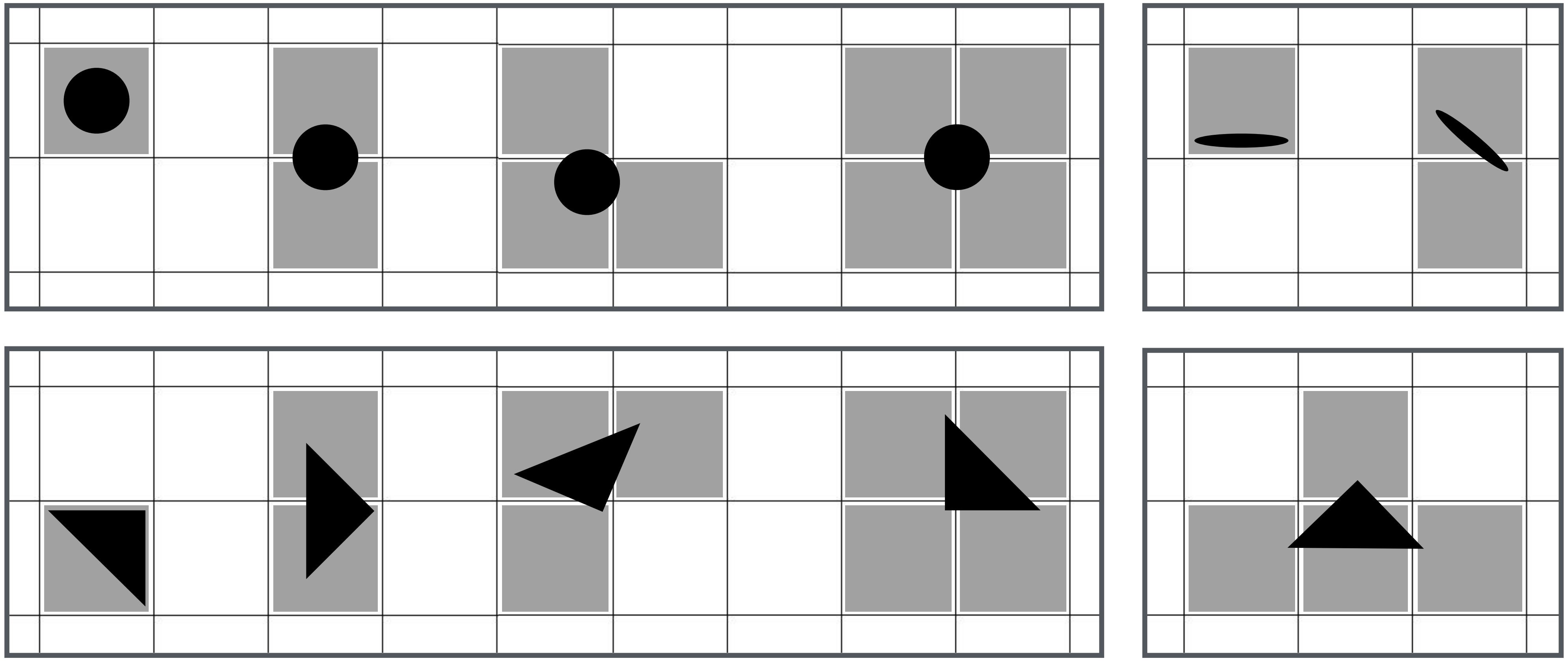
Step 2:

- Fit inhomogeneous density
- Cut out areas above threshold

Model for cause versus model for effect

Detector is based on a lattice, but damage occurs independently of the lattice structure.

The ***same cause for damage*** shape can hit 1, 2, 3 or 4 pixels, depending on position and orientation.



Higher level model: dead events

Solution: Model the damage by summarising neighbouring dead pixels into one dead event.

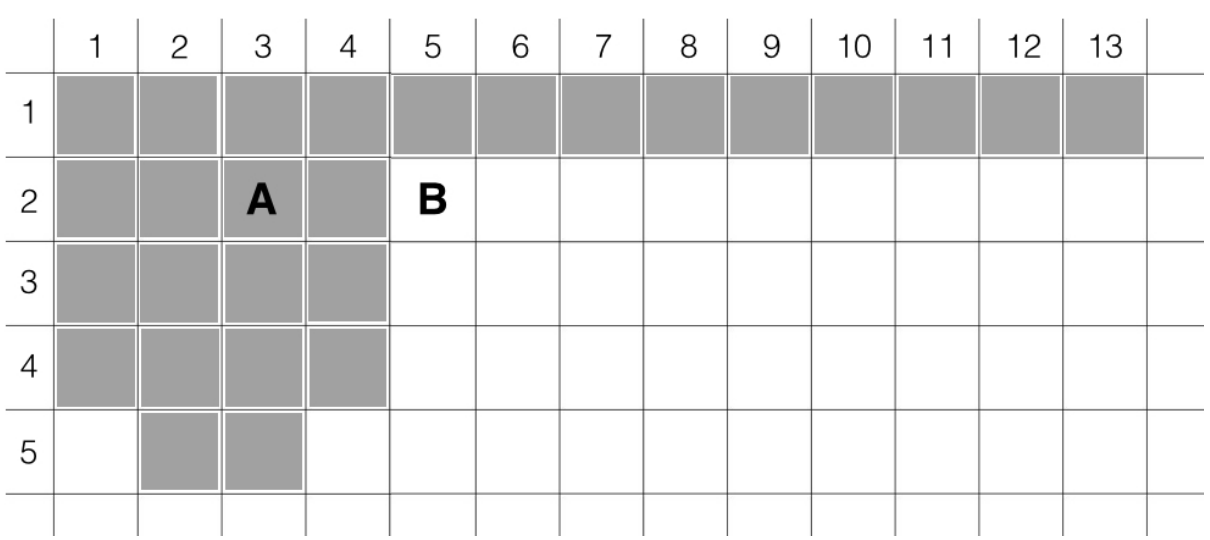
Convert pixel point pattern X into event point pattern Y by replacing each cluster of pixels $\mathcal{C}$ in X by

$$i^{\text{median}}(\mathcal{C}) = \big(\rho(\text{median}_1(\mathcal{C})), \rho(\text{median}_2(\mathcal{C}))\big)$$

Using median because of robustness:

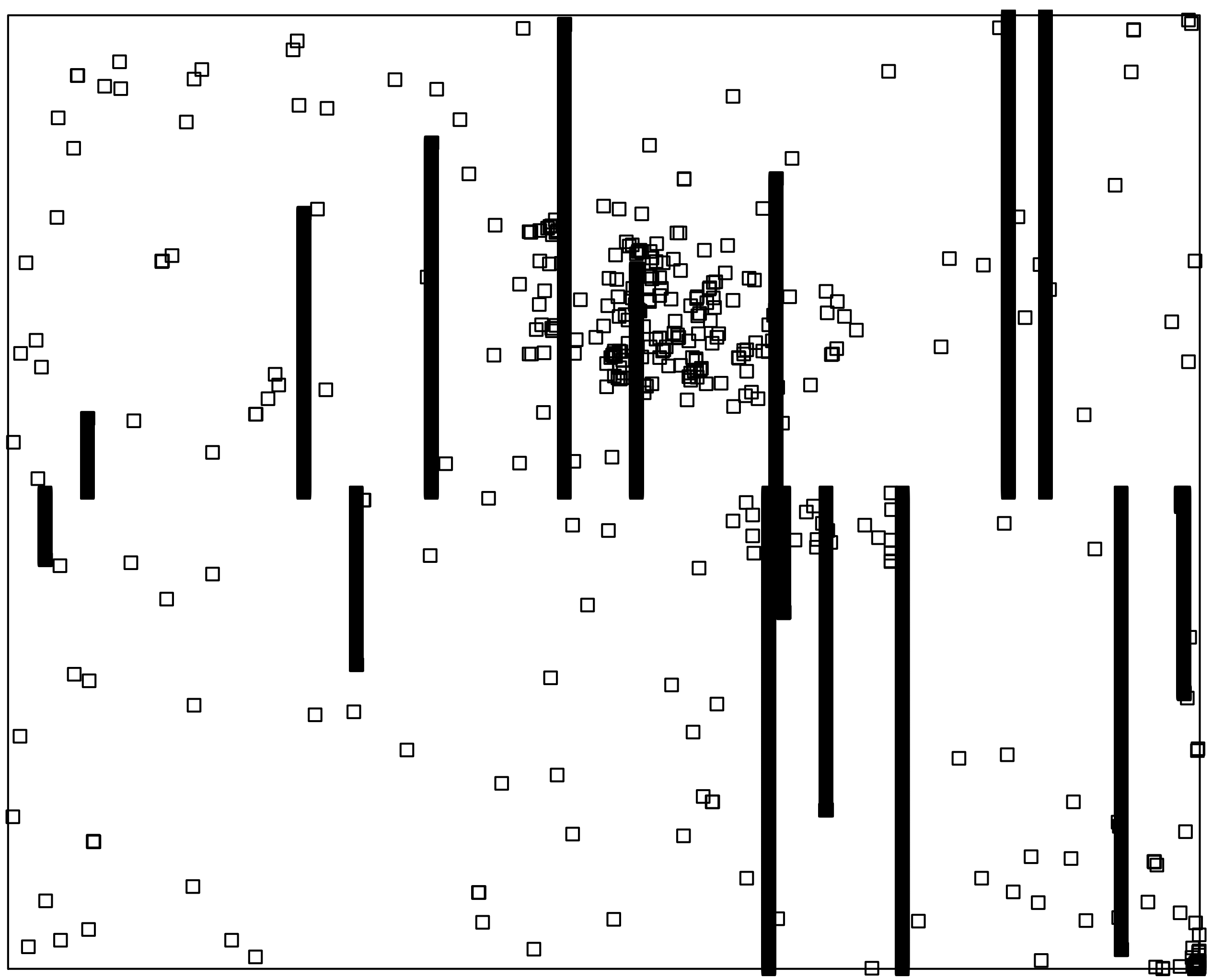
e.g. $i^{\text{mean}}(\mathcal{C}) = (5, 2)$

$i^{\text{median}}(\mathcal{C}) = (3, 2)$



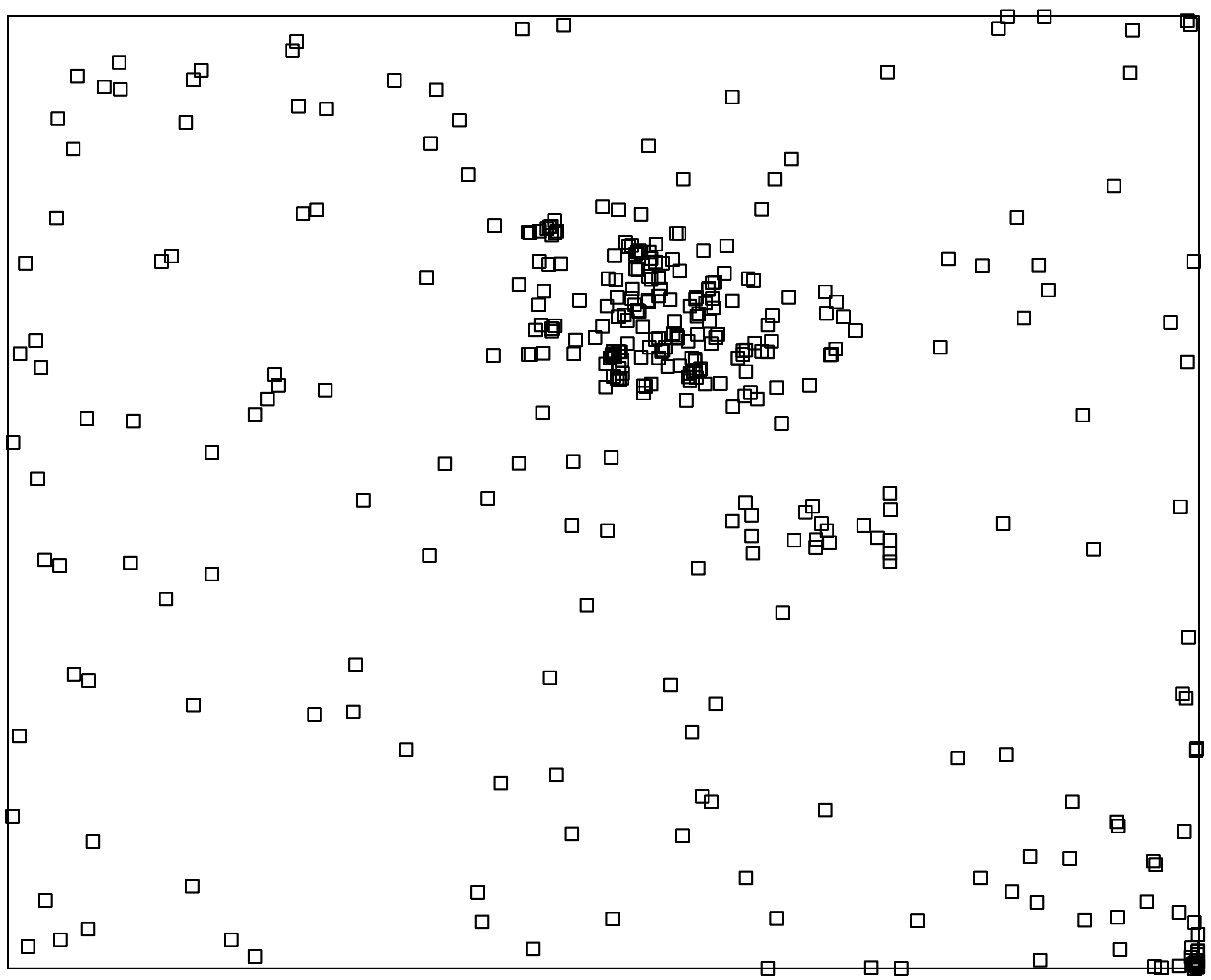
Dead pixels versus dead events

X (dead pixels)



(a) Pixel process

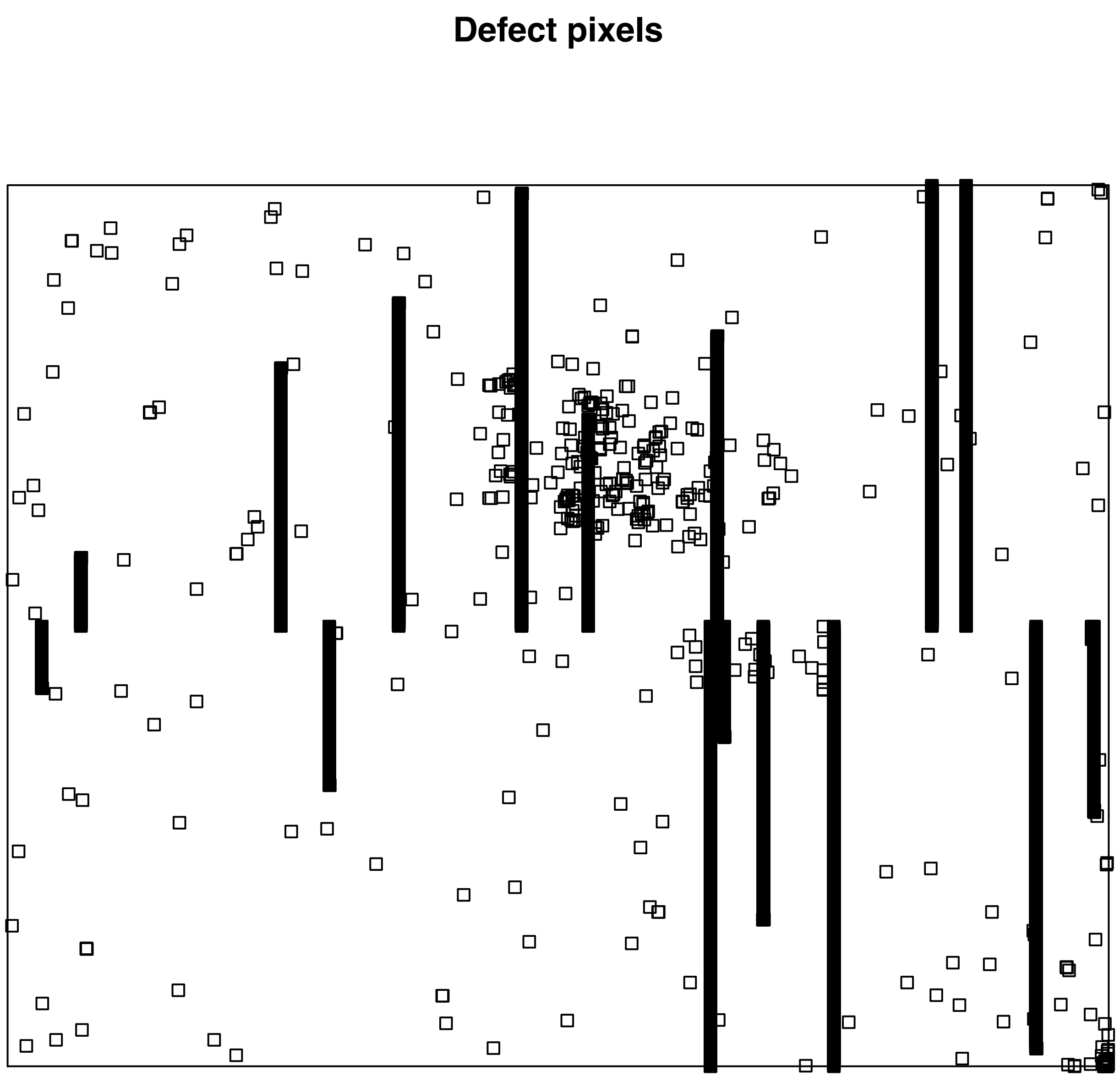
Y (dead events)

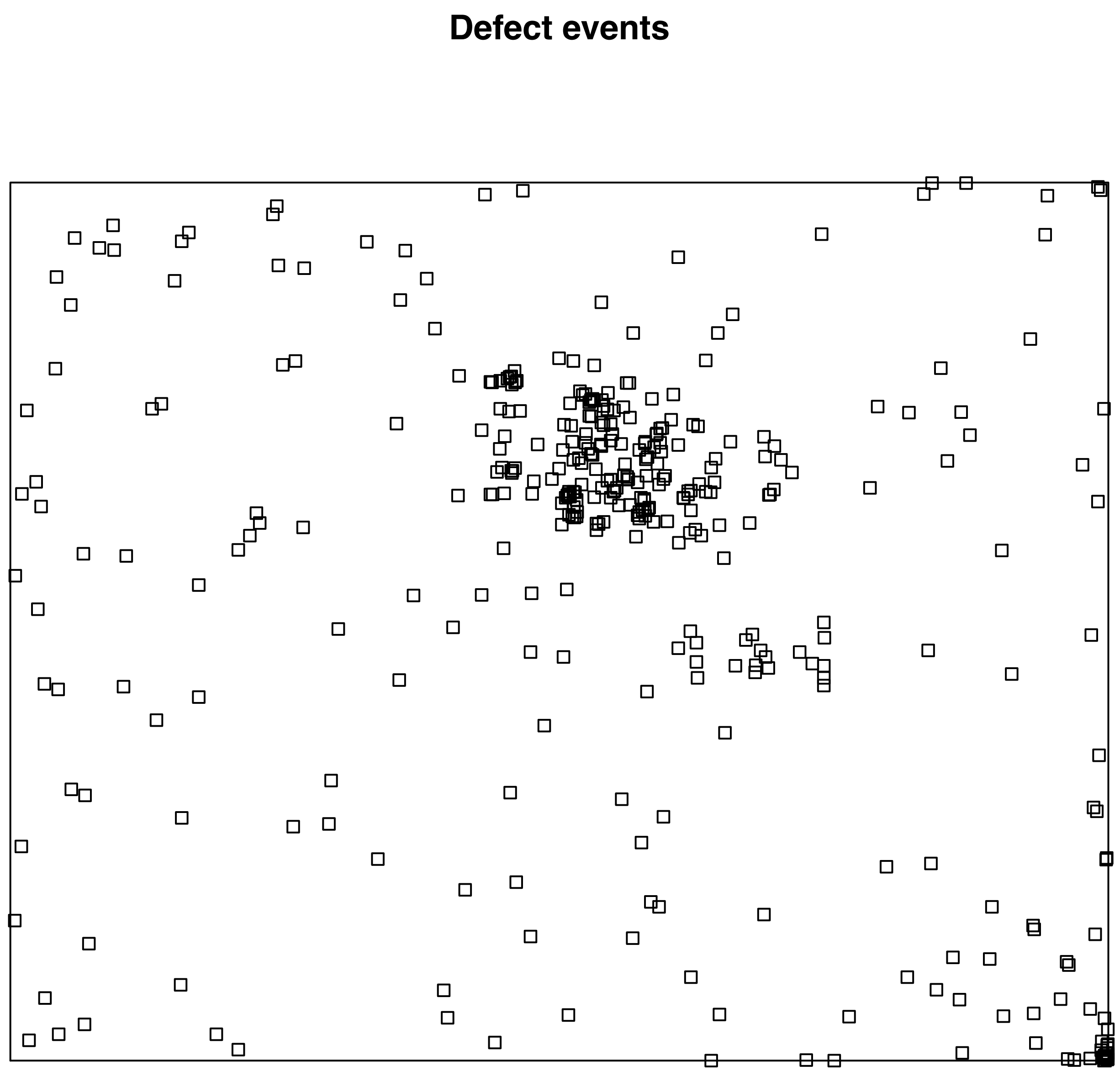


(b) Event process (marks not visualised)

Higher level defect model (Step 1)

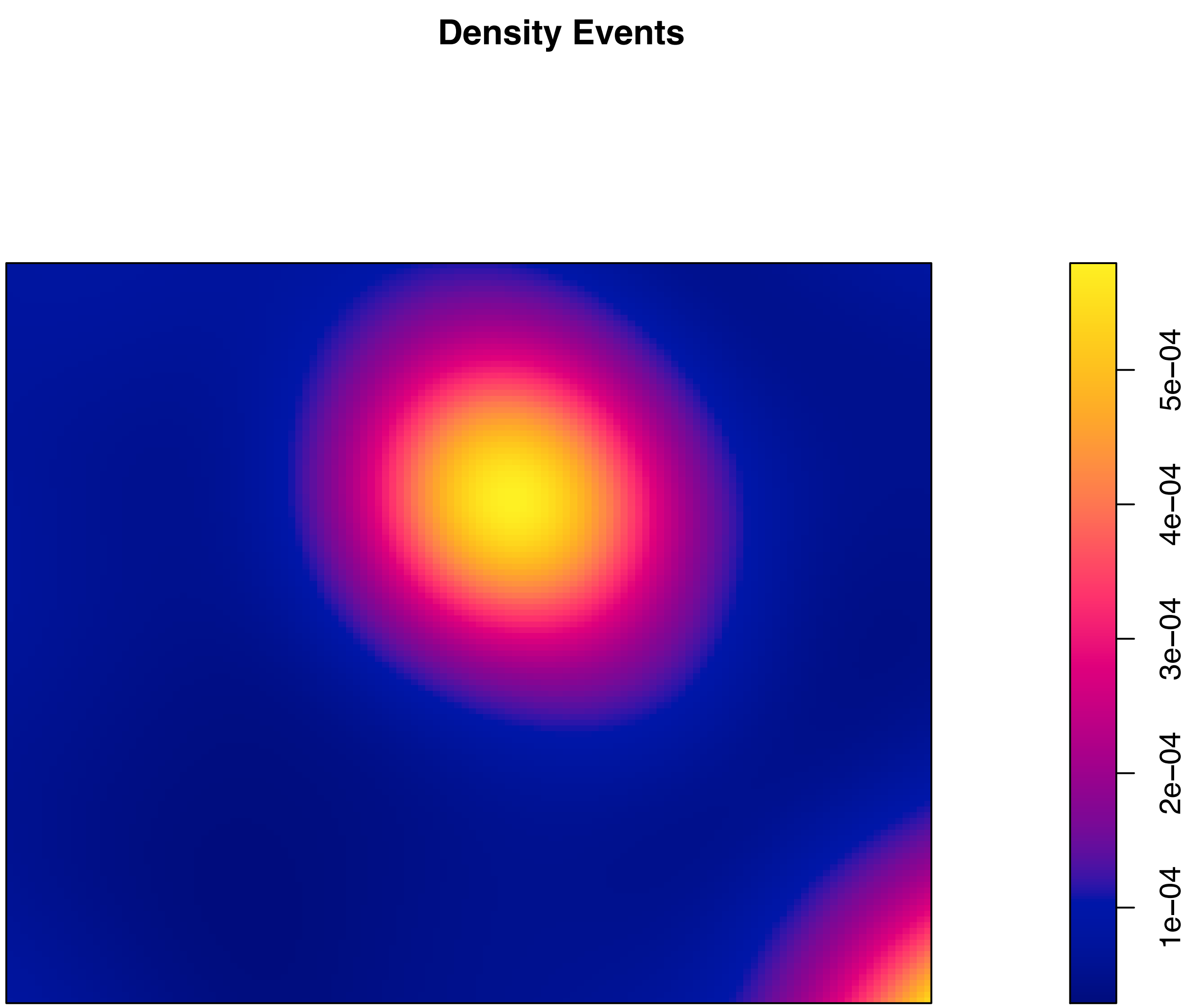
Conversion of point process to *event process*

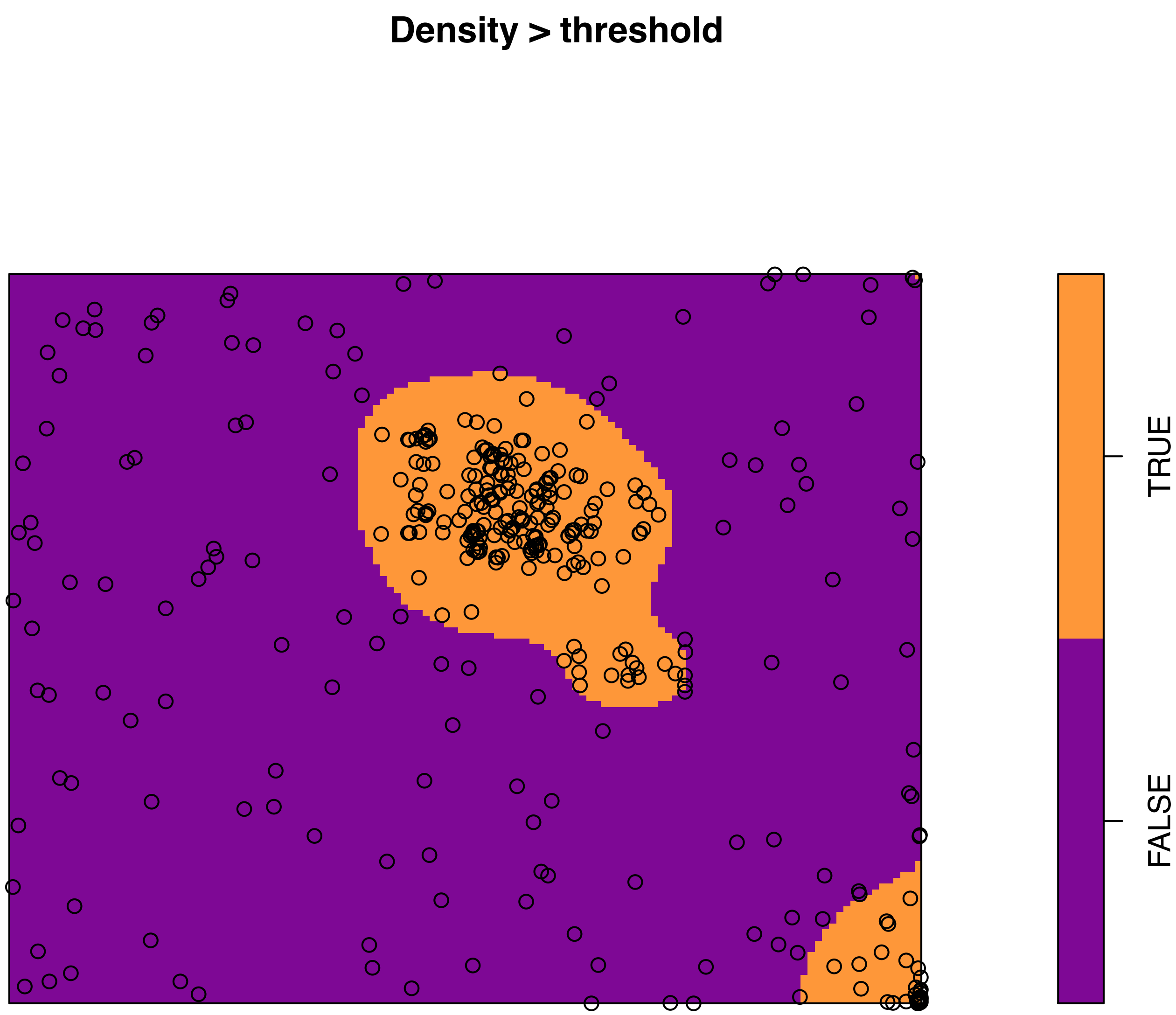




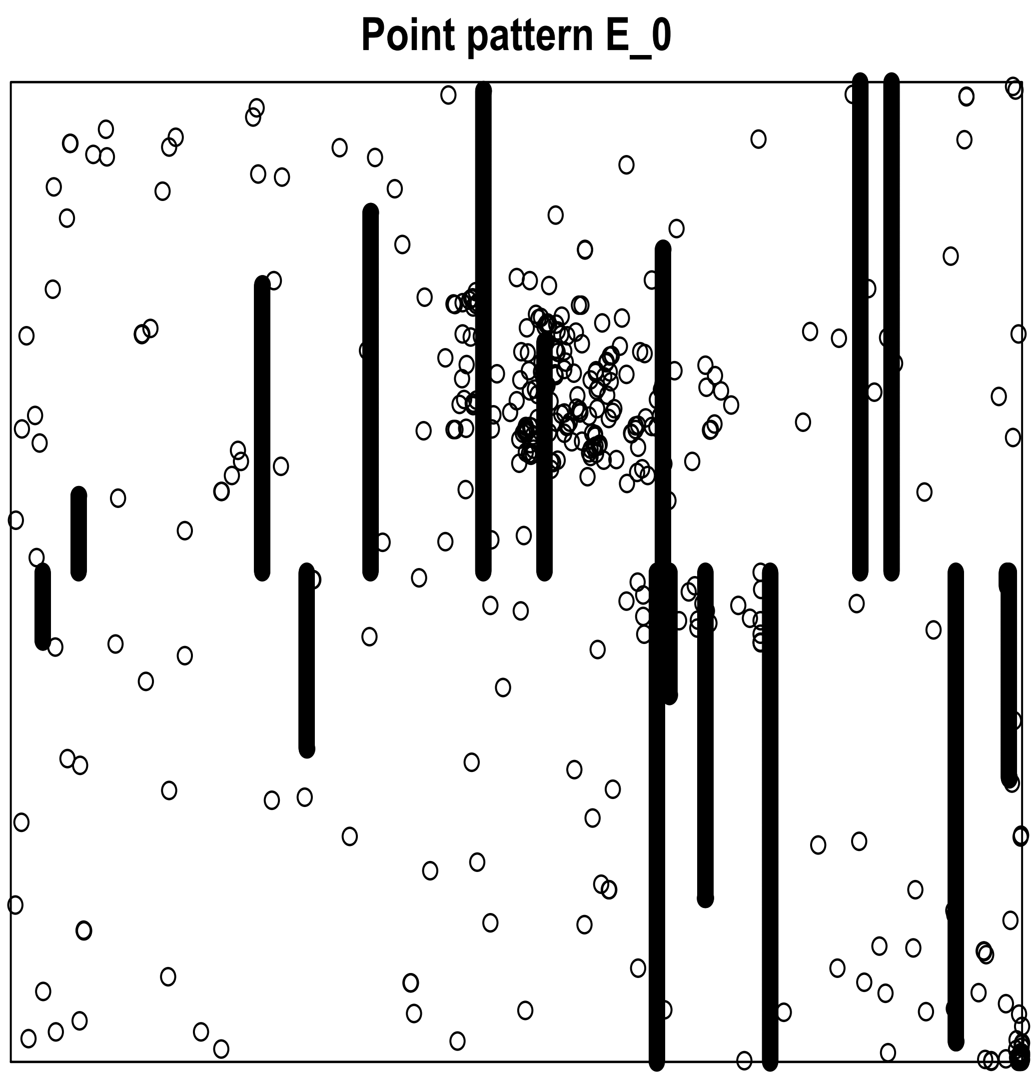
Density based thresholding (Step 2)

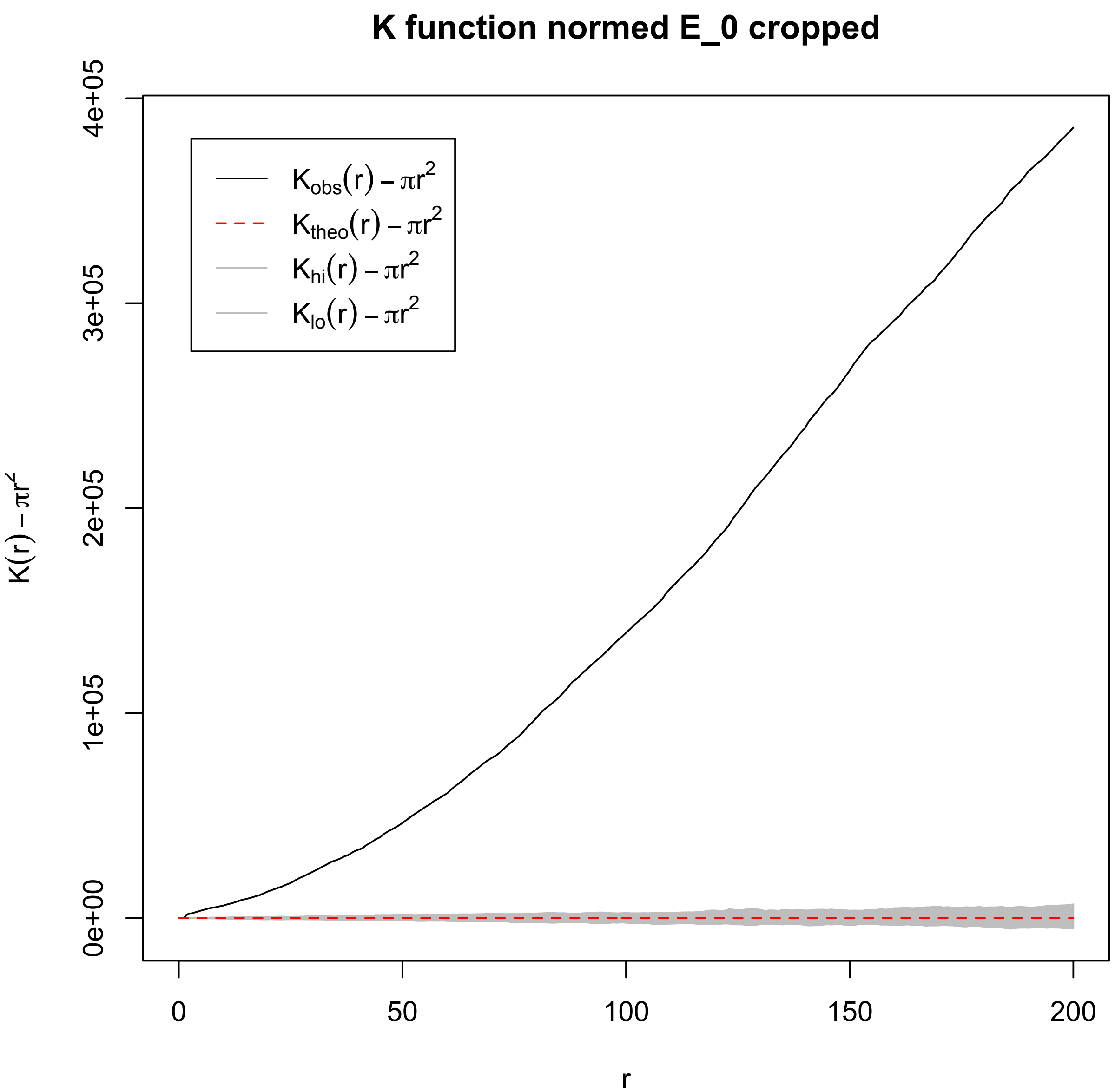
Remove areas with local density above threshold (median +1.5 IQR)



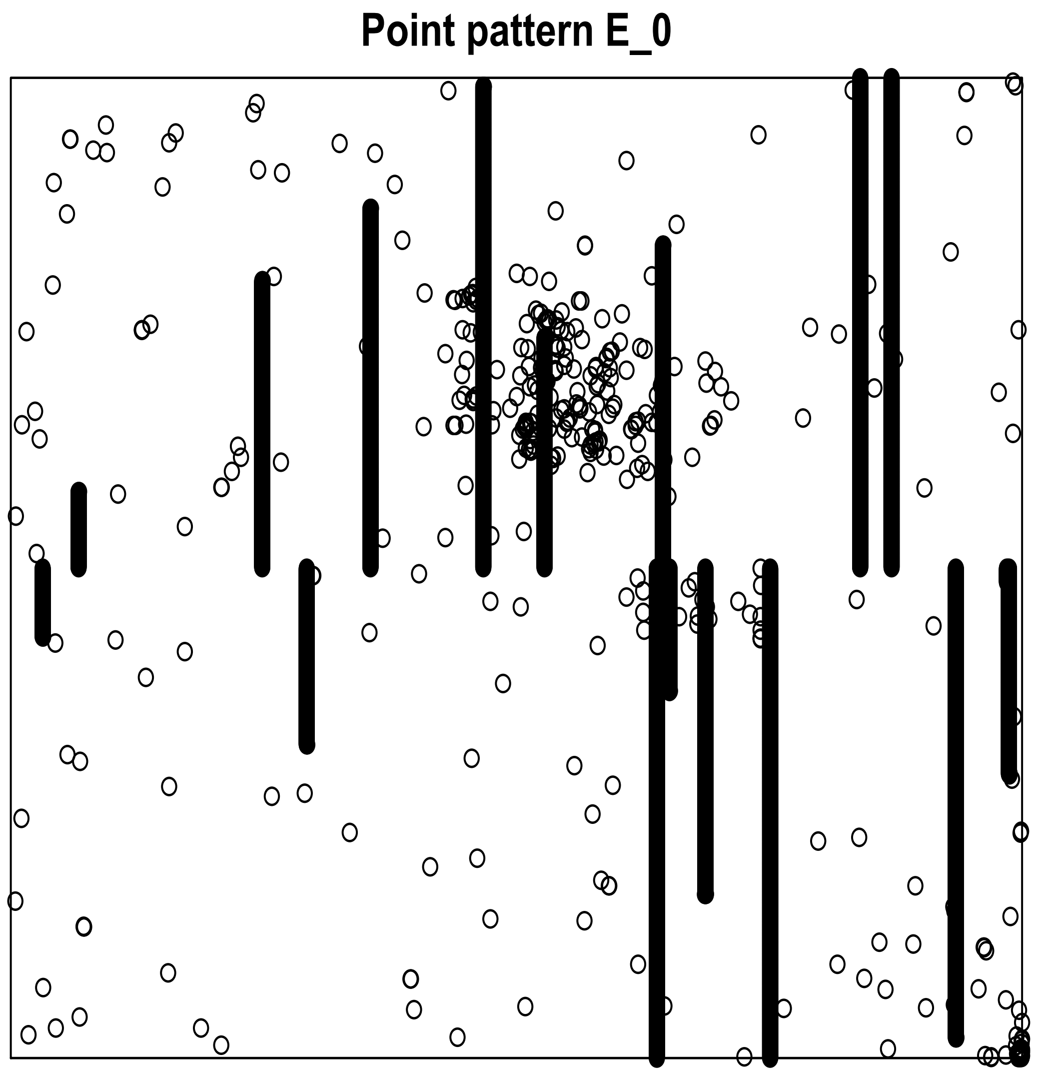


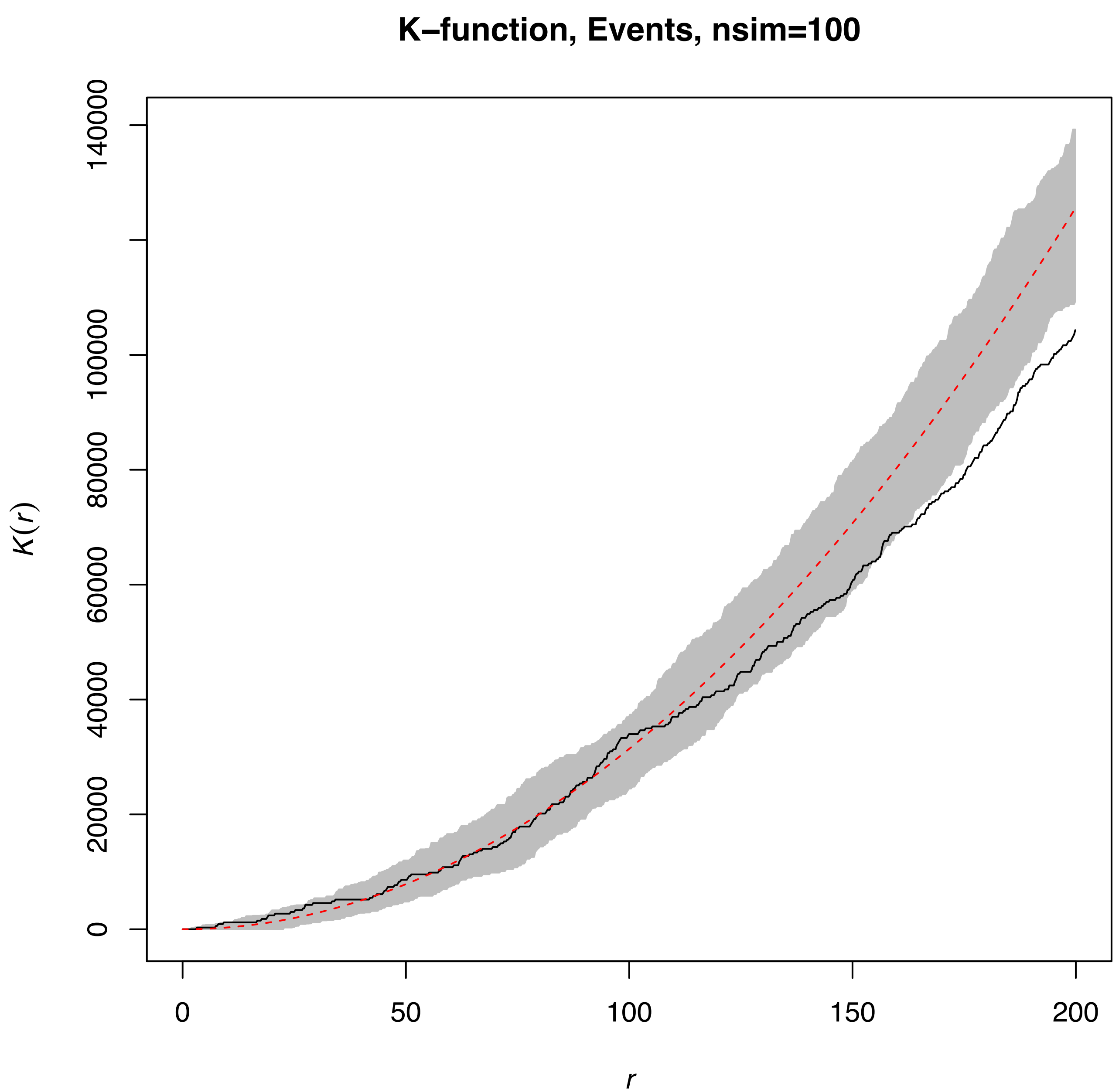
Before modification: K-function





After modification: K-function





Measurement quality assessment/improvement

- 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

Software project with the Alan Turing Institute

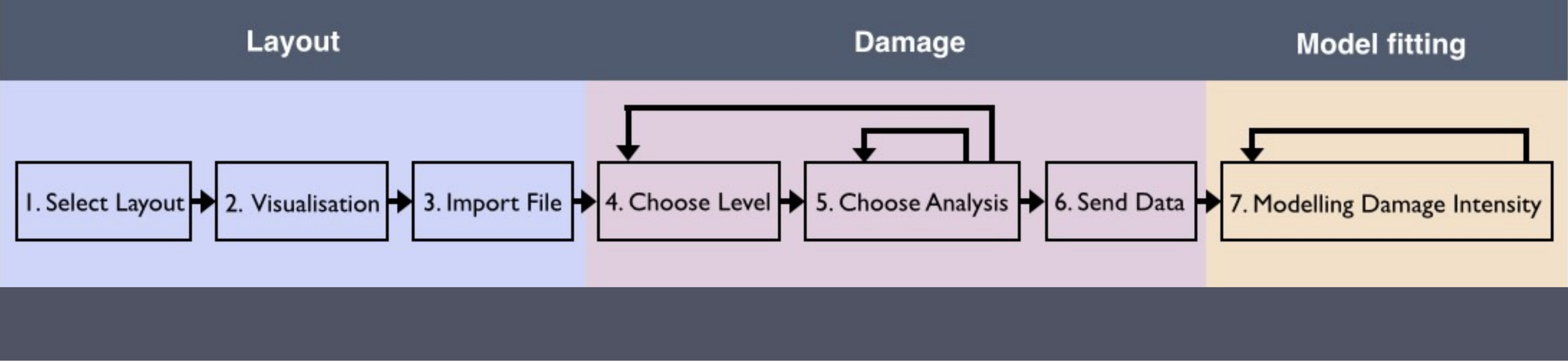
Objectives:

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

Seed funded project:

- 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

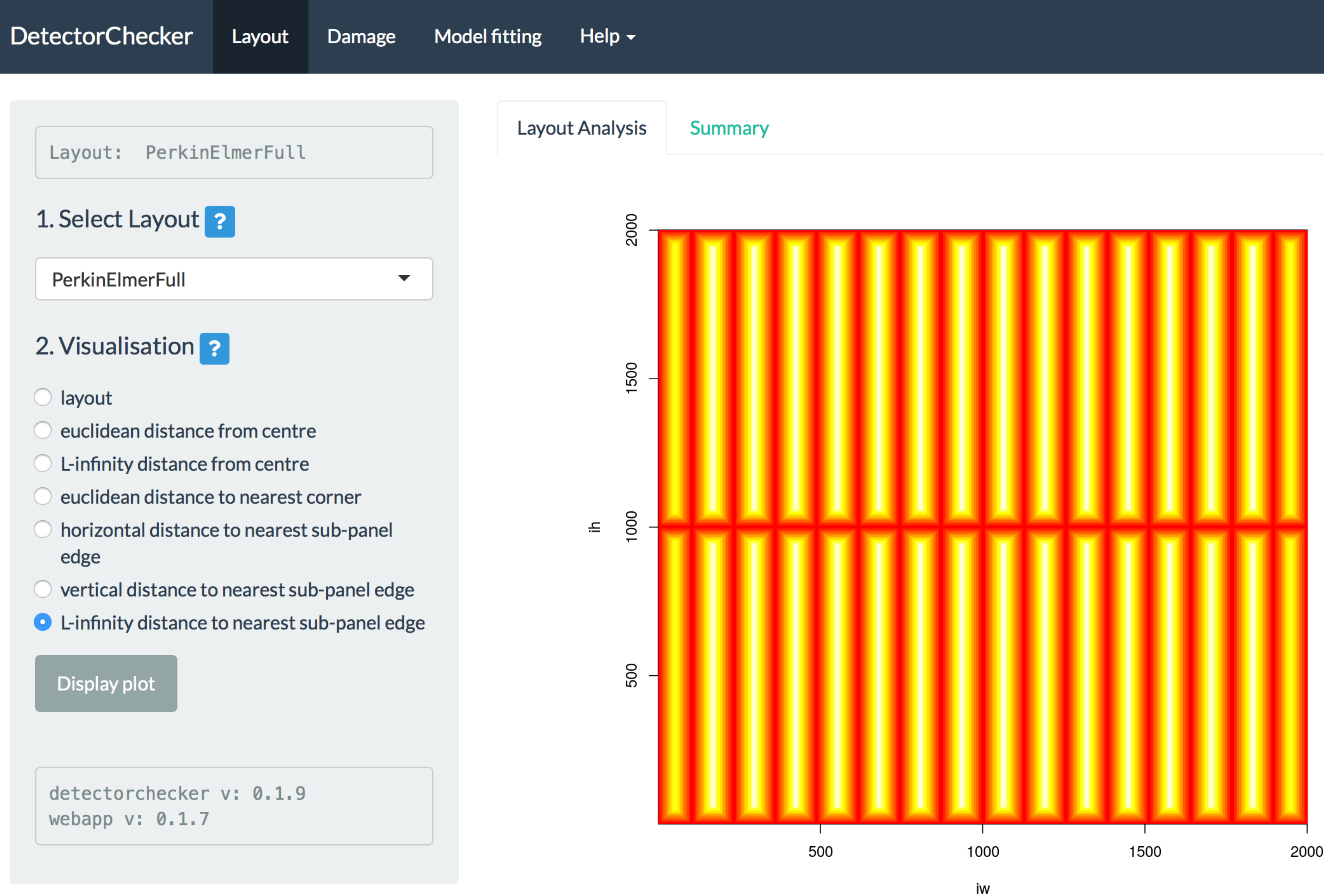


https://detectorchecker.azurewebsites.net

Team
Dr Julia Brettschneider (University of Warwick)
Dr Oscar Giles (The Alan Turing Institute)
Dr Tomas Lazauskas (The Alan Turing Institute)
Prof Wilfrid Kendall (University of Warwick)

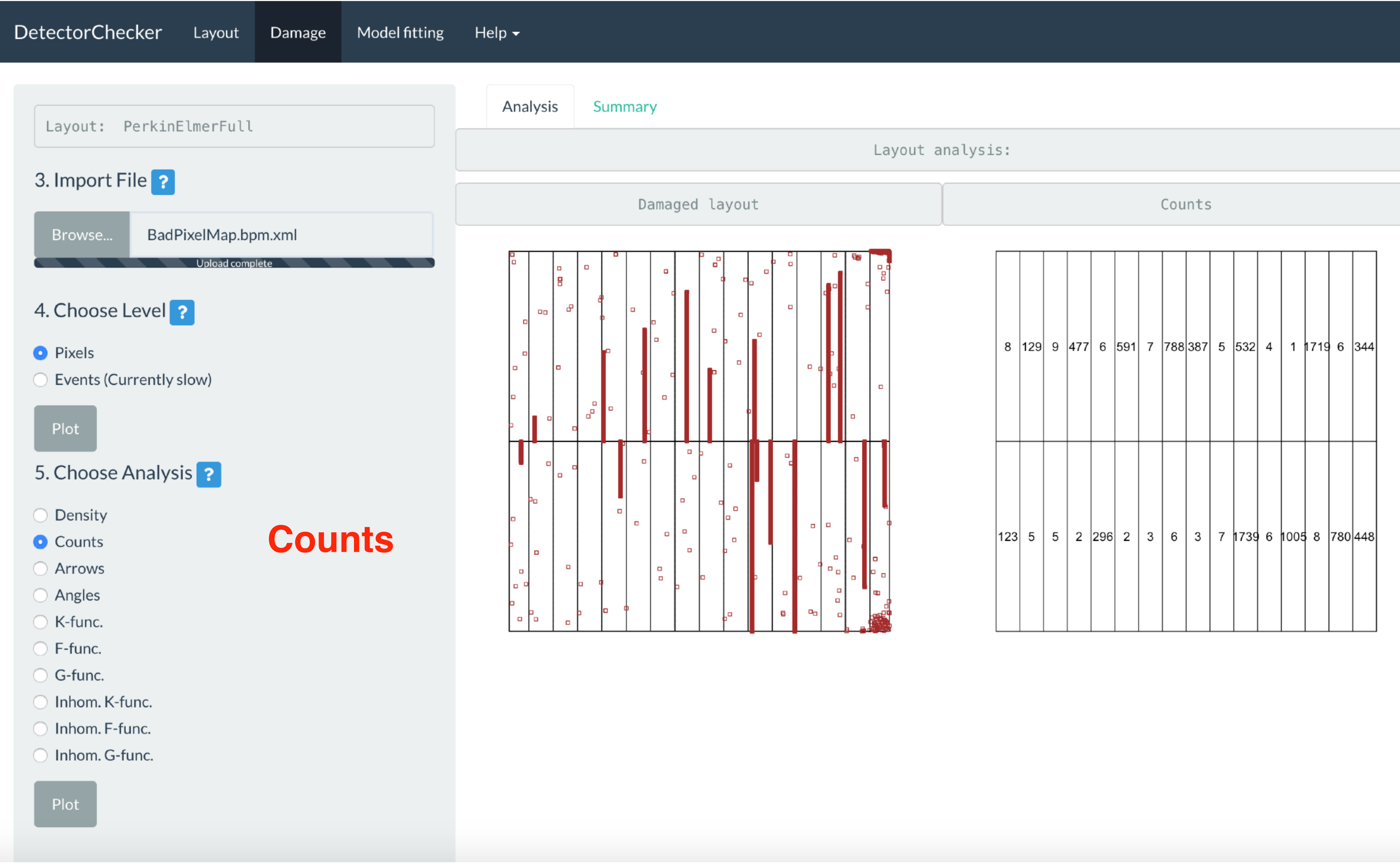
Contacts
julia.brettschneider@warwick.ac.uk
W.S.Kendall@warwick.ac.uk

Warwick Statistics



https://detectorchecker.azurewebsites.net

Warwick Statistics



https://detectorchecker.azurewebsites.net

Warwick Statistics

DetectorChecker Layout Damage Model fitting Help

Layout: PerkinElmerFull

3. Import File ?

Browse... BadPixelMap.bpm.xml

Upload complete

4. Choose Level ?

- Pixels
- Events (Currently slow)

Plot

5. Choose Analysis ?

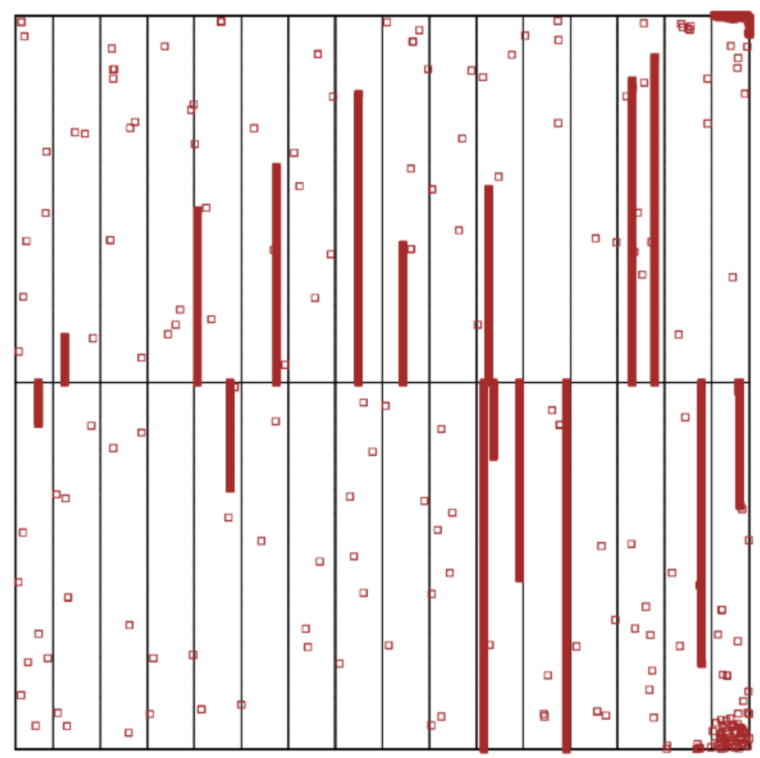
- Density
- Counts
- Arrows
- Angles
- K-func.
- F-func.
- G-func.
- Inhom. K-func.
- Inhom. F-func.
- Inhom. G-func.

Plot

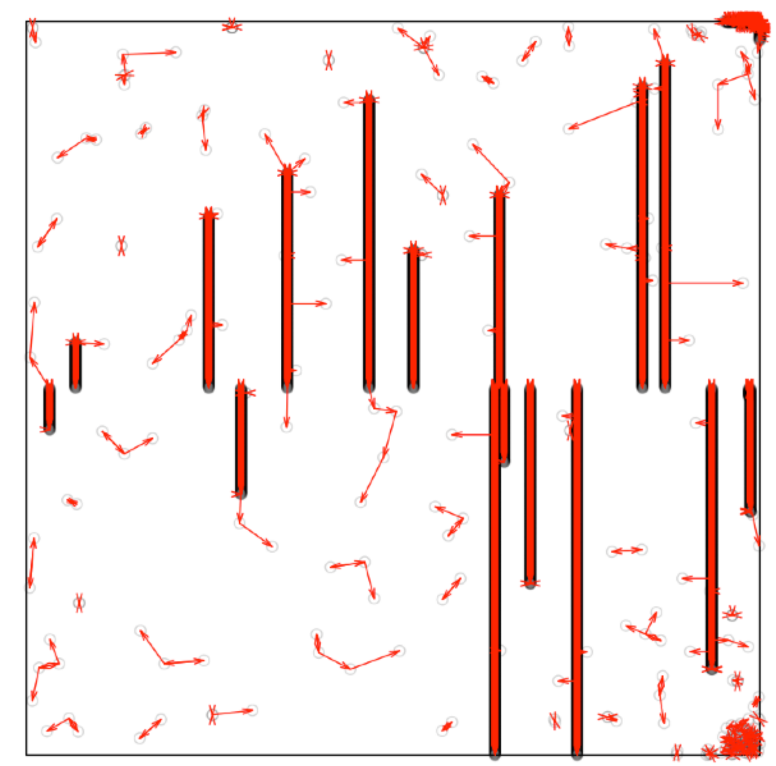
Analysis Summary

Layout analysis:

Damaged layout



Arrows



Arrows
Angles?

Arrows pointing at nearest neighbour
Angles distribution dominated by lines

https://detectorchecker.azurewebsites.net

Warwick Statistics

DetectorChecker | Layout | Damage | Model fitting | Help

Layout: PerkinElmerFull

3. Import File ?

Browse... BadPixelMap.bpm.xml

Upload complete

4. Choose Level ?

- (•) Pixels
- ( ) Events (Currently slow)

Plot

5. Choose Analysis ?

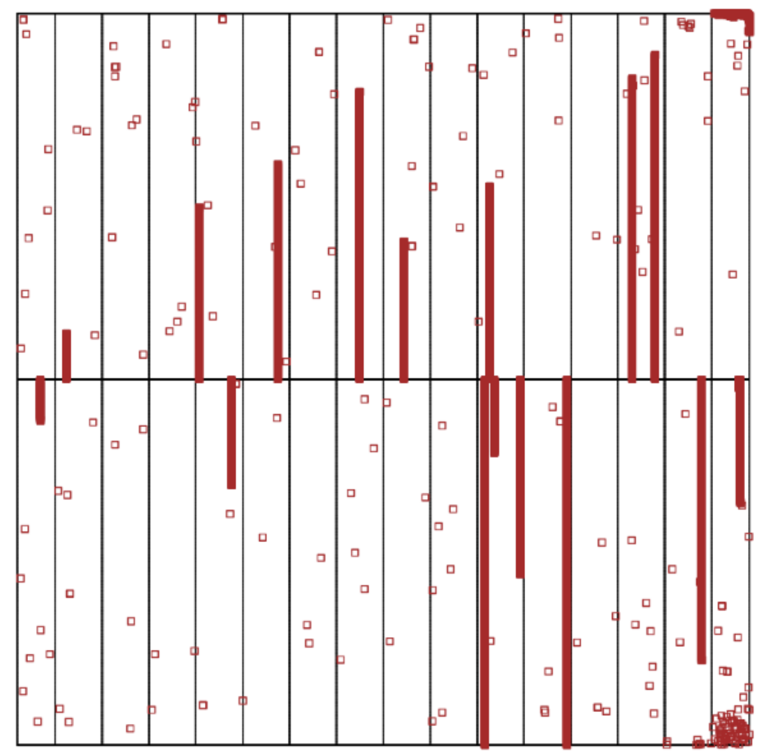
- ( ) Density
- ( ) Counts
- ( ) Arrows
- ( ) Angles
- (•) K-func.
- ( ) F-func.
- ( ) G-func.
- ( ) Inhom. K-func.
- ( ) Inhom. F-func.
- ( ) Inhom. G-func.

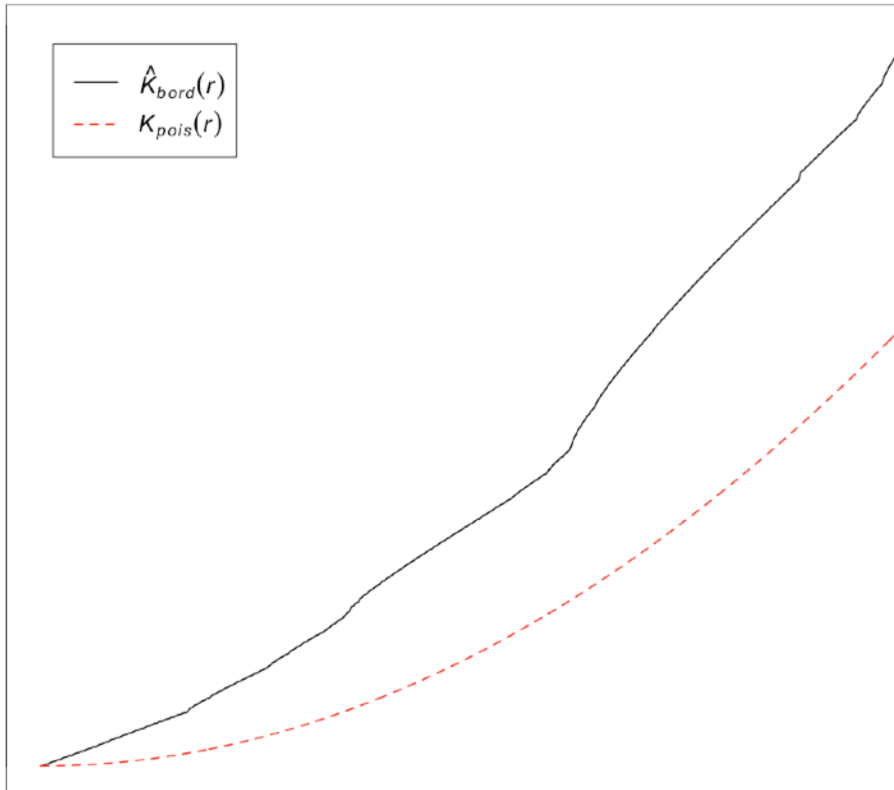
Plot

Analysis | Summary

Layout analysis:

Damaged layout | K-func.

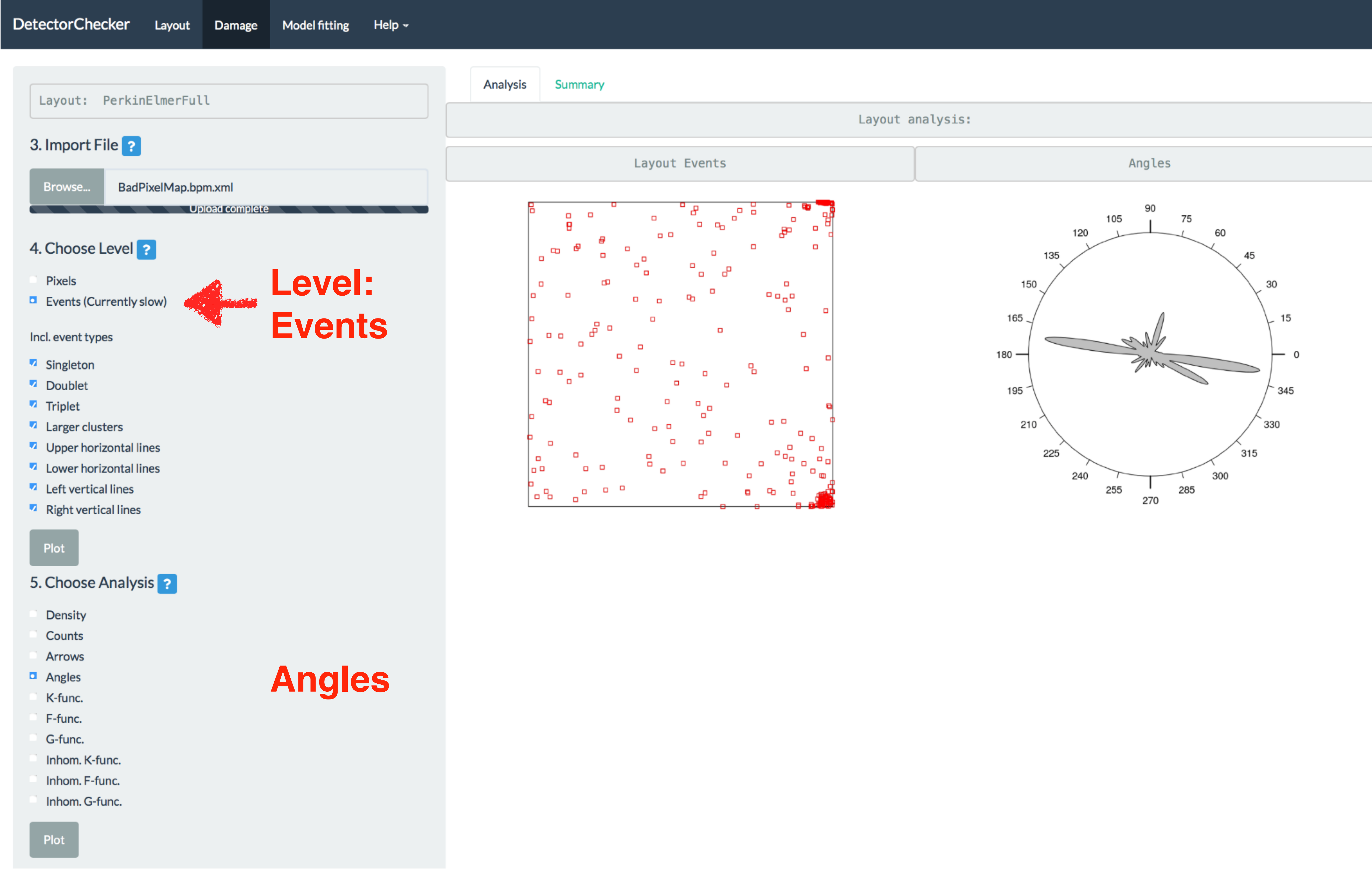




K-function

Not CSR (completely spatially at random)

https://detectorchecker.azurewebsites.net



https://detectorchecker.azurewebsites.net

Warwick Statistics

DetectorChecker Layout Damage Model fitting Help

Layout: PerkinElmerFull

7. Modelling Damage Intensity ?

- euclidean distance from centre
- L-infinity distance from centre
- horizontal distance to nearest sub-panel edge
- vertical distance to nearest sub-panel edge

Fit model

Output

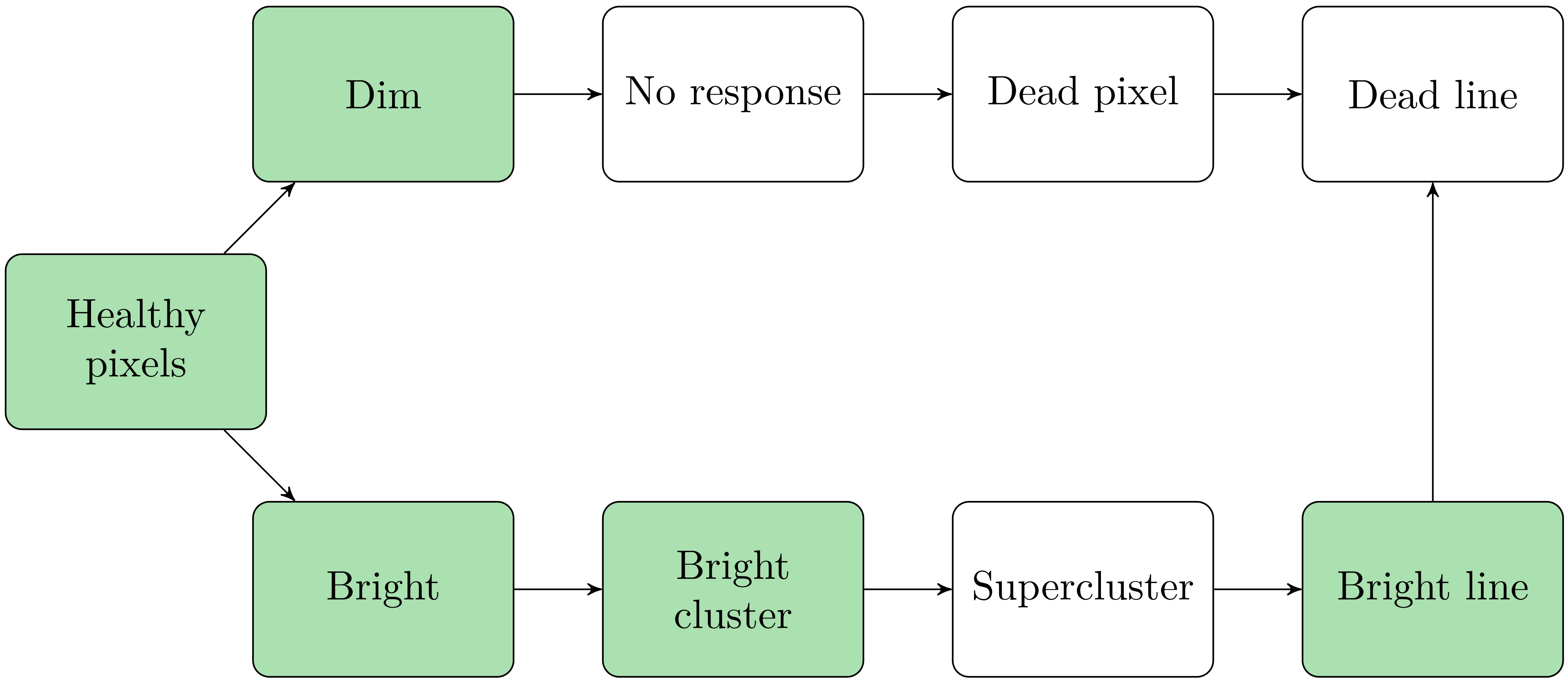
```
Call:
glm(formula = as.vector(pix_matrix) ~ as.vector(dist), family = binomial(link = logit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.0988  -0.0745  -0.0662  -0.0605   3.6700

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -5.320e+00  2.605e-02 -204.22   <2e-16 ***
as.vector(dist) -1.002e-03  3.511e-05  -28.54   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 133198  on 3999999  degrees of freedom
Residual deviance: 132394  on 3999998  degrees of freedom
AIC: 132398

Number of Fisher Scoring iterations: 9
```

https://detectorchecker.azurewebsites.net

Warwick Statistics

Refined states (more than just dead)

Using grey, white and black images define a variety of dysfunctional states and look at transitions.



Model for temporal development

Markov model with transition probabilities estimated from data:

		New state										
		Normal	No response	Dead	Hot	V. bright	Bright	Bright line	Screen spot	Edge	V. dim	Dim
Initial state	normal	99.91	-	-	0	0	0	0	0.1	0	-	0
	no response	-	98.83	-	-	-	-	-	-	-	1.17	-
	dead	-	-	100	-	-	-	-	-	-	-	-
	hot	-	-	-	96.72	3.28	-	-	-	-	-	-
	v.bright	0.89	-	-	2.74	88.62	7.69	-	-	0.06	-	-
	bright	18.07	-	-	-	8.44	73.44	0.05	-	-	-	-
	line.b	-	-	-	-	-	-	100	-	-	-	-
	screen spot	84.7	-	-	-	-	-	-	16.59	0.56	-	-
	edge	0.06	-	-	-	0	-	0	0.06	99.89	-	-
	v.dim	-	10	-	-	-	-	-	-	-	90	-
	dim	15	-	-	-	-	-	-	-	-	-	94.44

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
Journal of Open Source Software, 2020, 5(56), 2474

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

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