Some Analyses of New Brunswick Forest Fire Data

Rolf Turner

Department of Mathematics and Statistics

University of New Brunswick, Fredericton, N. B., Canada

Acknowledgments

Special thanks are due to

- Jeffrey Betts of the N. B. Department of Natural Resources who provided the data sets and helped me to understand their intricacies.
- Jonathan Beaudoin of the U.N.B. Department of Geodesy and Geomatics who gave me a C program to convert latitude and longitude into New Brunswick Double Stereographic Projection coordinates.

Further Acknowledgments

- I owe enormous gratitude to Adrian Baddeley of the University of Western Australia, from whom I learned practically everything I know about spatial point processes.

- I must emphasize however that any flaws or errors in the following presentation are attributable entirely to my own inadequacies . . . .

. . . . and NOT to the fact that I was

Baddeley taught!

Themes — Explicit and Implicit

- The data; much cleaning required; treated year-by-year.
- The importance of good convenient software; the importance of simulation.
- The observation window; the coordinate system.
- EDA; data plots; intensity plots; K functions.
- Spatial models; modelling software; model syntax.
- Model diagnostics; residual plots.
- Conclusions meager; data set is rich; surface barely scratched.

Organizing the Data

- Complete fire records of the N. B. Department of Natural Resources for the years 1987 through 2003.
- Excel spreadsheets, one for each year.
- Each had 97 columns and between 286 and 654 rows (i.e. fires).
- Spreadsheets $\mapsto$ R data frames.
- Considerable cleaning was needed; still more needed.
- Many missing values and anomalous entries.
- Decisions needed as to what variables to retain.

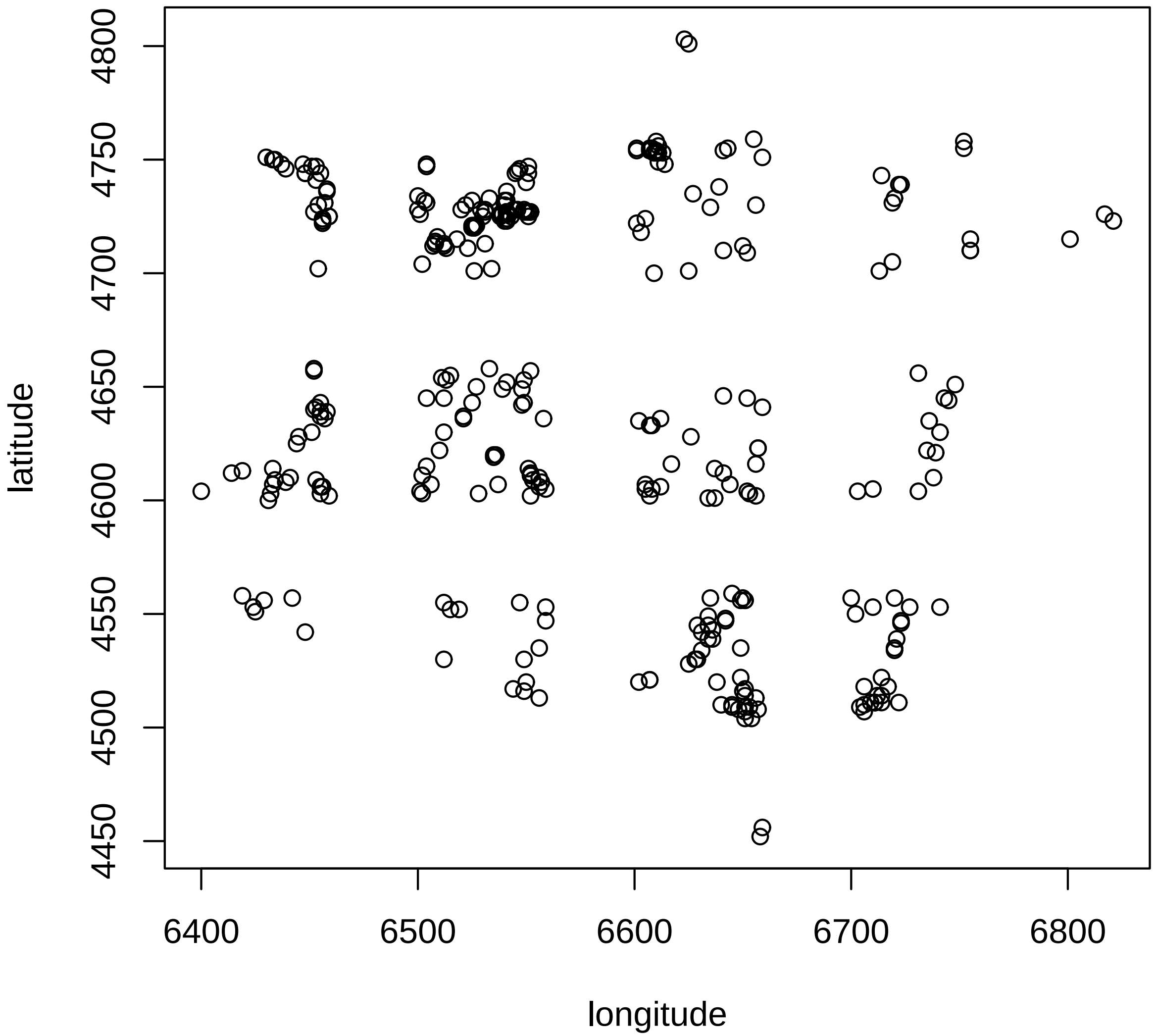
A Slightly Puzzling Plot

- Crucial information: *location* of the fires.
- Given in latitude and longitude, like this:

	latitude	longitude
1	4554	6731
2	4600	6732
3	4618	6544
4	4612	6730
5	4455	6660
6	4519	6707

- Eagerly plotted raw latitude and longitude from the 2000 data.
- Result bizarre; took me a while to see what I'd done.

A *REALLY WRONG* Plot of The Year 2000 N. B. Fires



What’s the explanation?

The `spatstat` *Package*

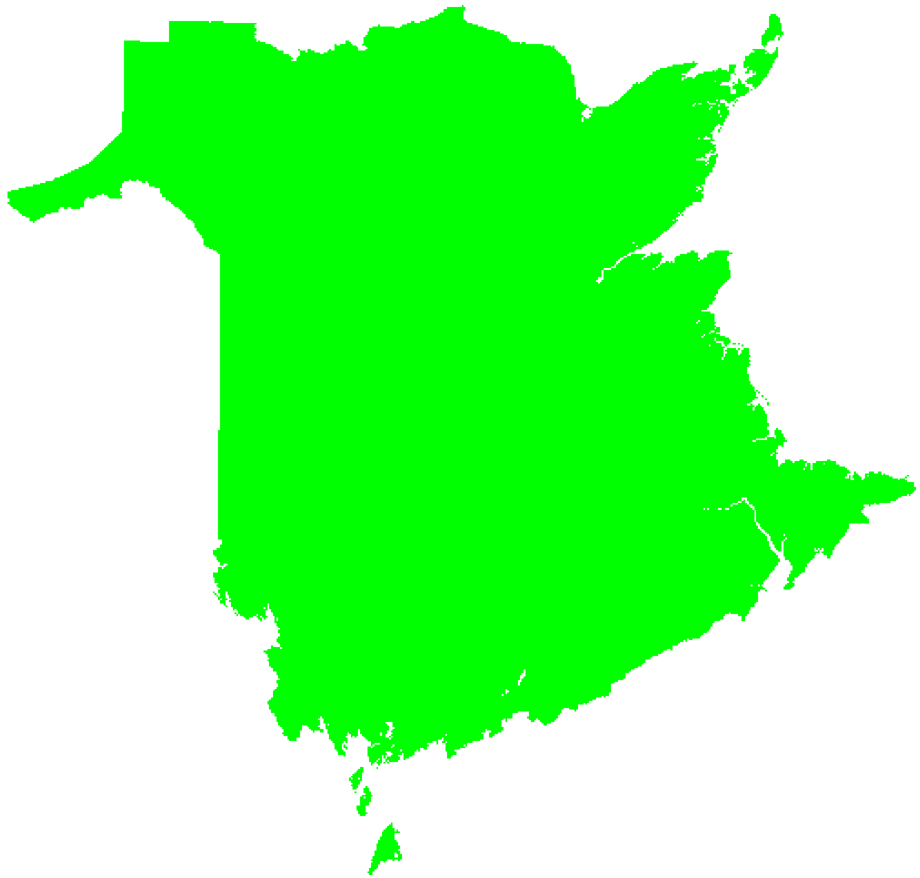
- The `spatstat` package was used for all of my analyses.
- Can get it from `http://www.r-project.org/CRAN` .
- Details — see
  Adrian Baddeley and Rolf Turner (2005). `spatstat`: an R package for analyzing spatial point patterns. *Journal of Statistical Software* **12** no. 6, pp. 1–42,
  URL: `http://www.jstatsoft.org` .

The Observation Window

- We need an observation window in order to properly specify a point pattern.
- Here we need a map of New Brunswick.
- GIS $\mapsto$ "shapefiles" $\mapsto$ collection of polygons $\mapsto$ *mask* type window.
- Mask $=$ pixel array of TRUE/FALSE values.
- Used relatively fine (500×500) pixellation.
- Shapefiles to polygons: used Roger Bivand's `maptools` package from `CRAN`).

The Observation Window as a Mask

New Brunswick Map as a ‘‘Mask’’



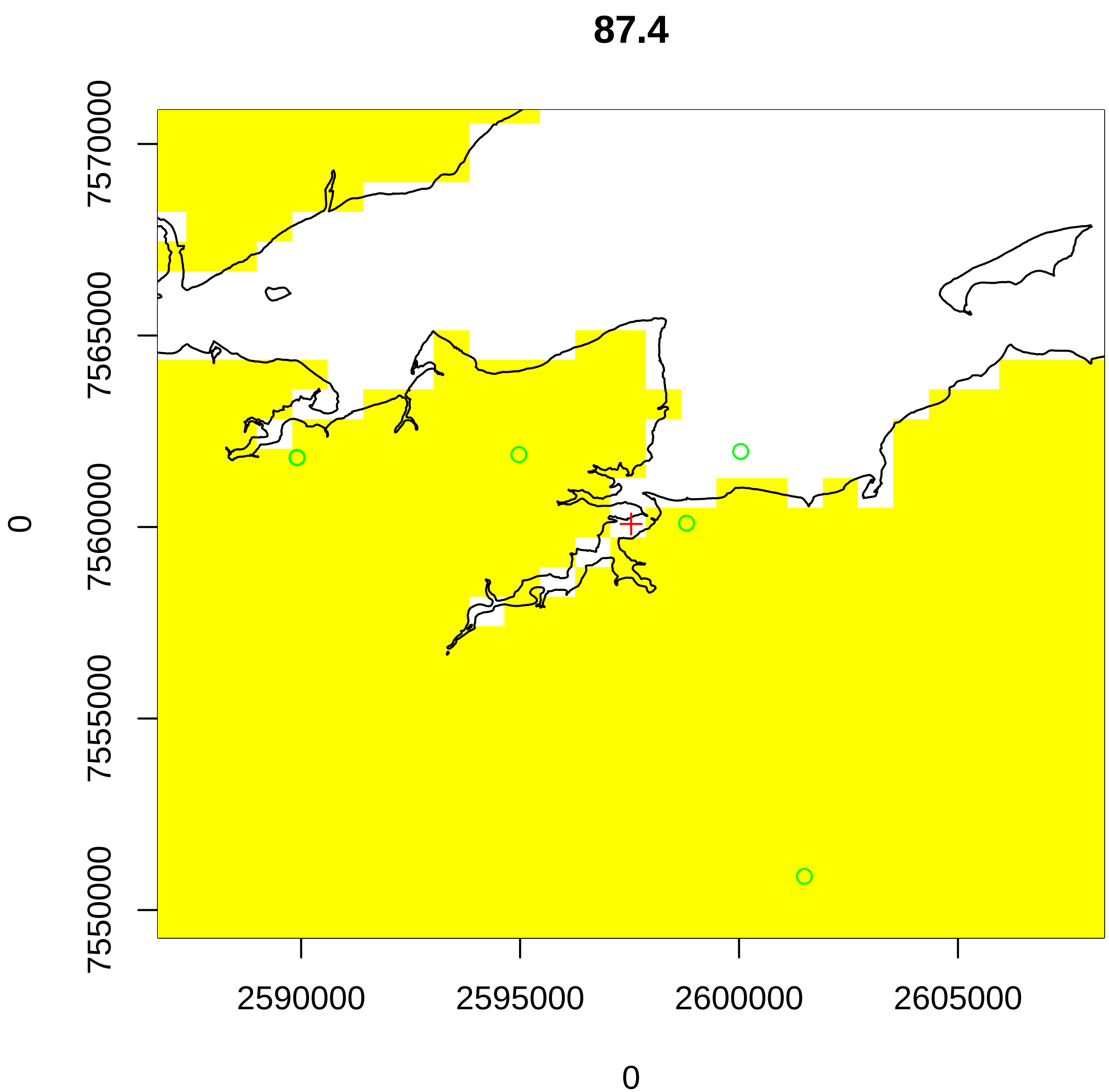
Outsiders

- Constructed window $\Rightarrow$ could construct patterns.
- But many points plotted outside of the observation window.
- Several reasons:
  1. Discretization of the window.
  2. Relative coarseness of fire locations; to nearest minute, ≈ 1 kilometer.
  3. Data entry errors.

Adjusting the Outsiders

- Shifted points which were “mildly” out of place to nearby locations inside the window.
- Deleted points which were “wildly” out of place. (Assumed data entry error.)
- Many points borderline; tossed a coin.
- Some examples follow.

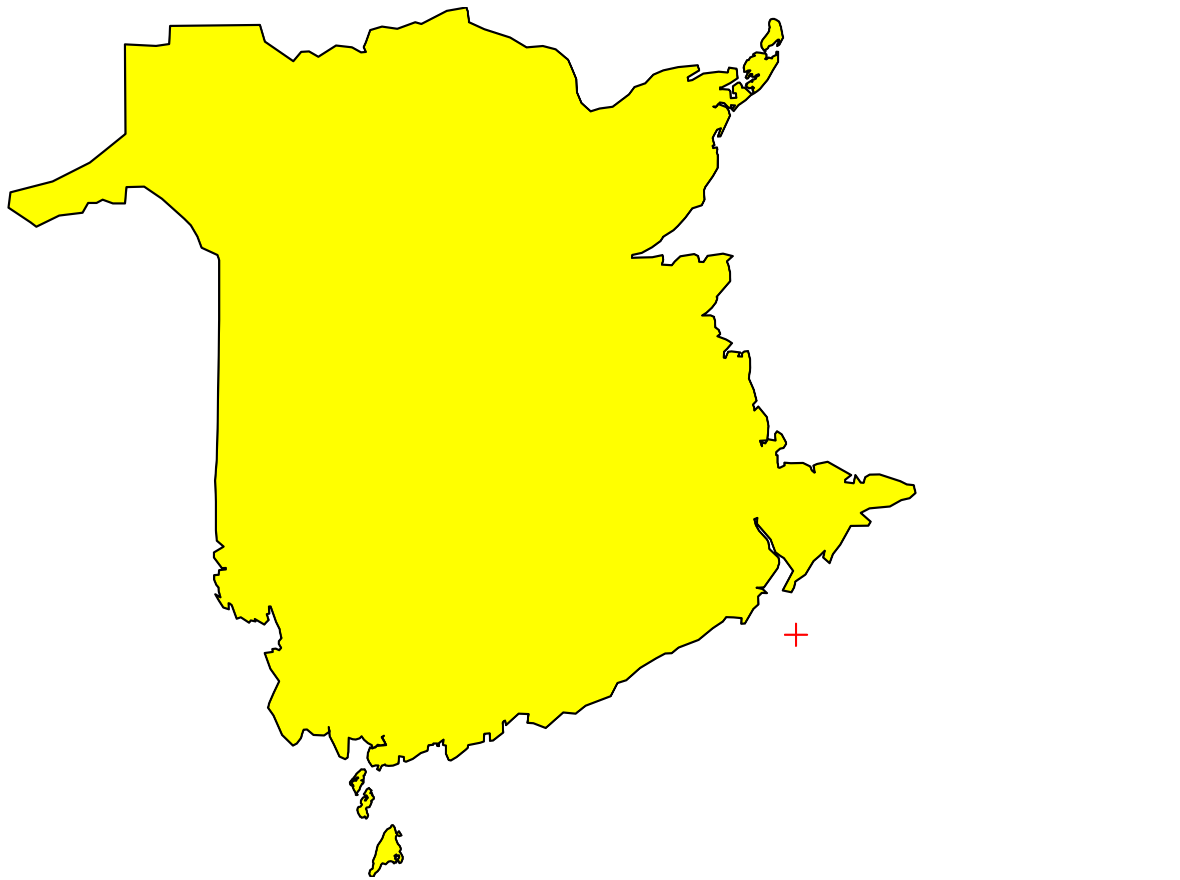
Mildly



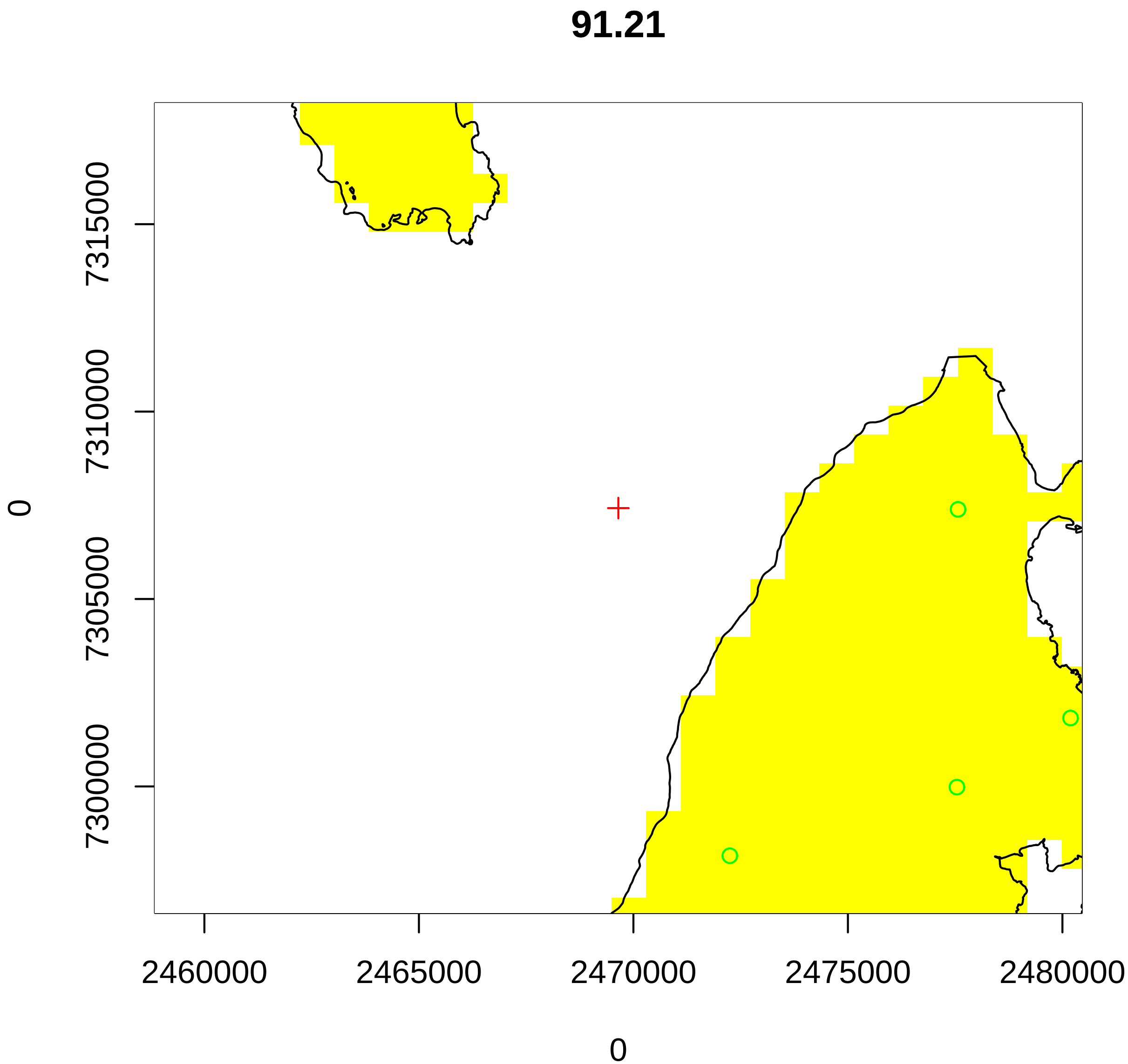
87.4
7550000
7555000
7560000
7565000
7570000
2590000
2595000
2600000
2605000
0
0

Wildly

87.6



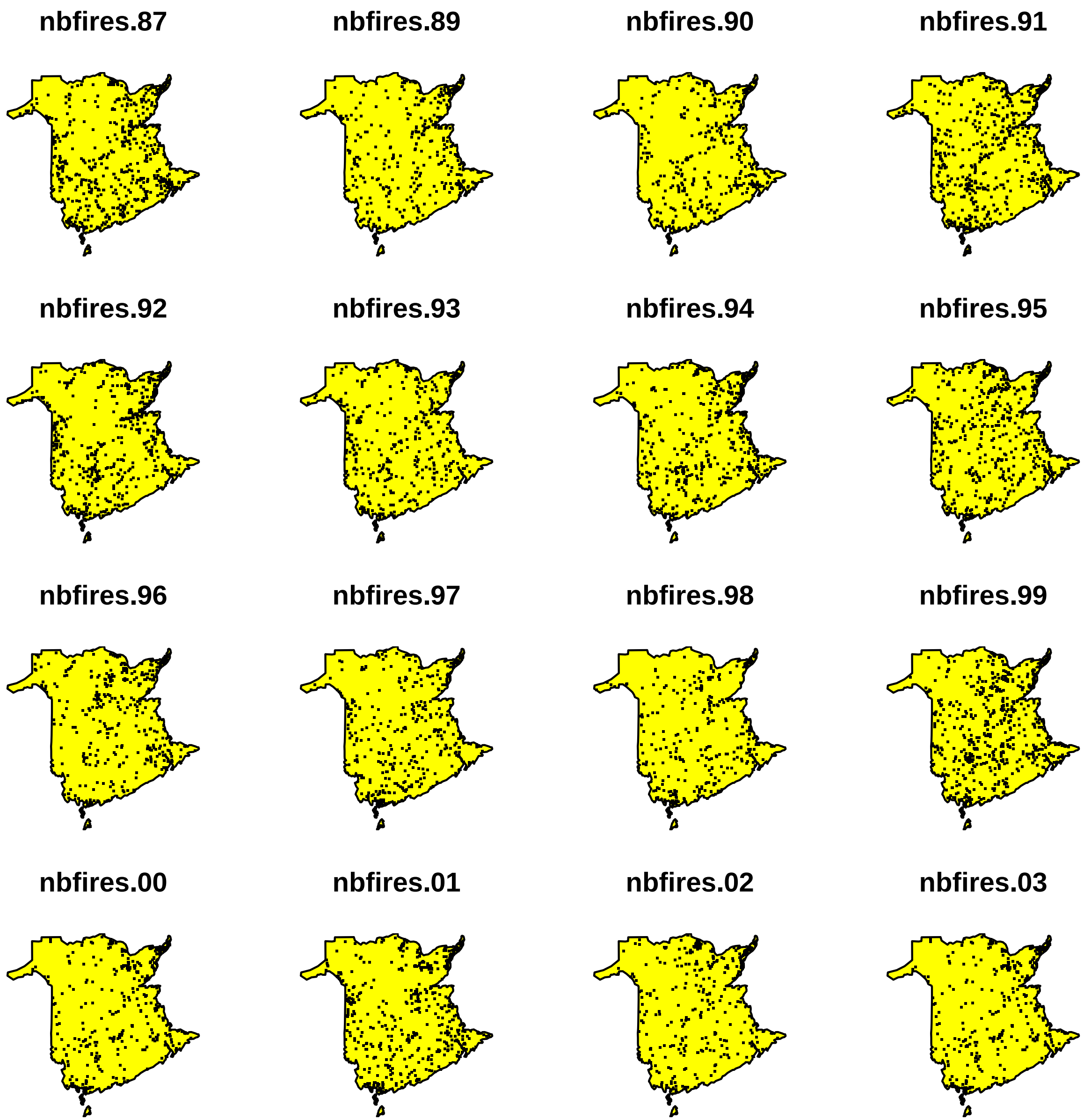
Borderline


91.21
0
7315000
7310000
7305000
7300000
2460000
2465000
2470000
2475000
2480000
0

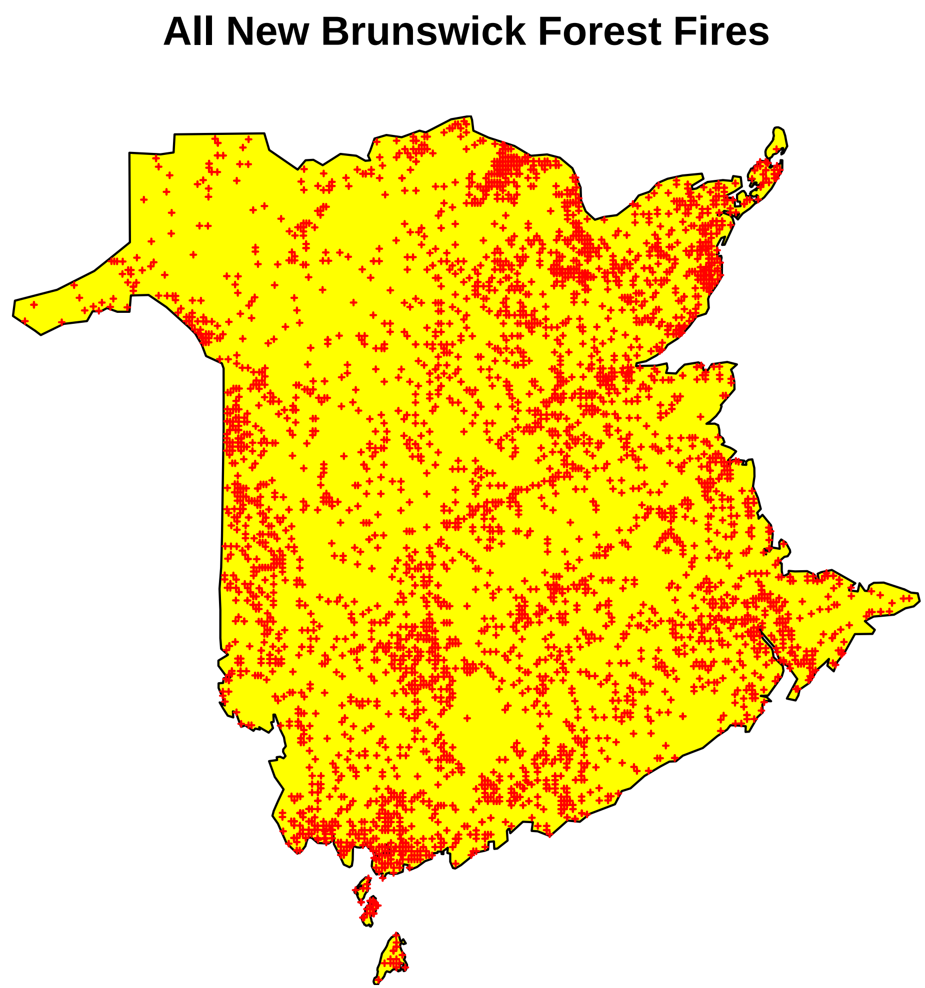
Data Plots

- Data now in some semblance of shape.
- Look at them year-by-year (all fires).
- Then narrow down to *forest* fires only. (Data include “grass”, “dump”, and “other”, as well as “forest” fires.)
- Look at aggregate over all available years; estimate spatial *trend* from the aggregate.
- Estimation done by applying a smoothing kernel.

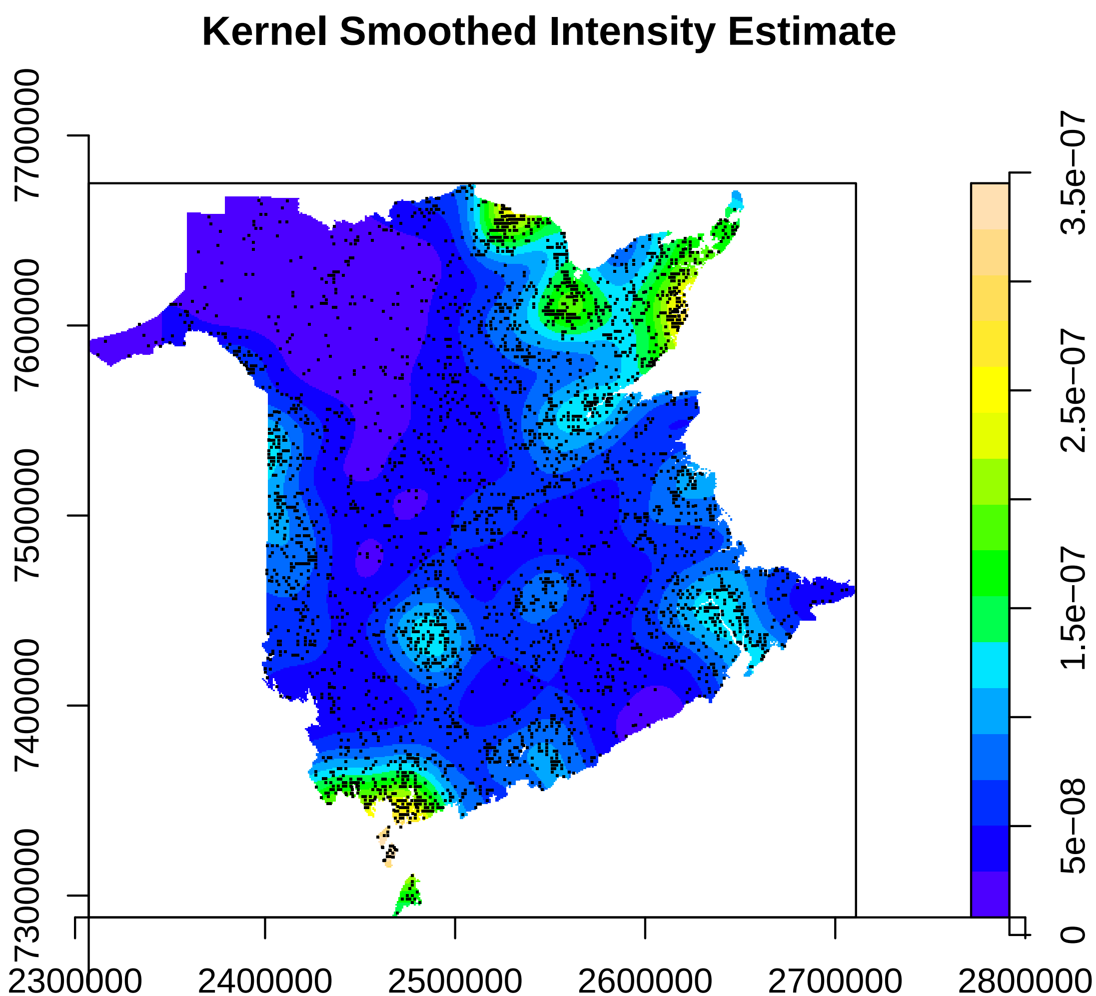
All Fires — Year by Year



Forest Fires — Aggregate



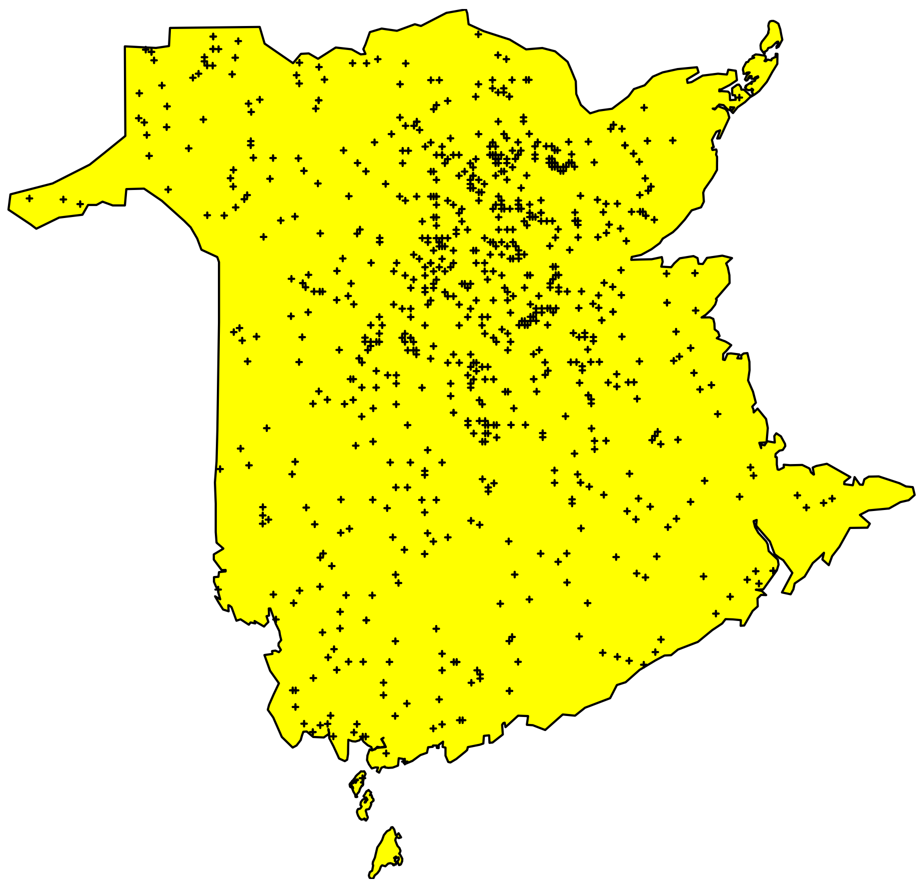
Spatial Trend



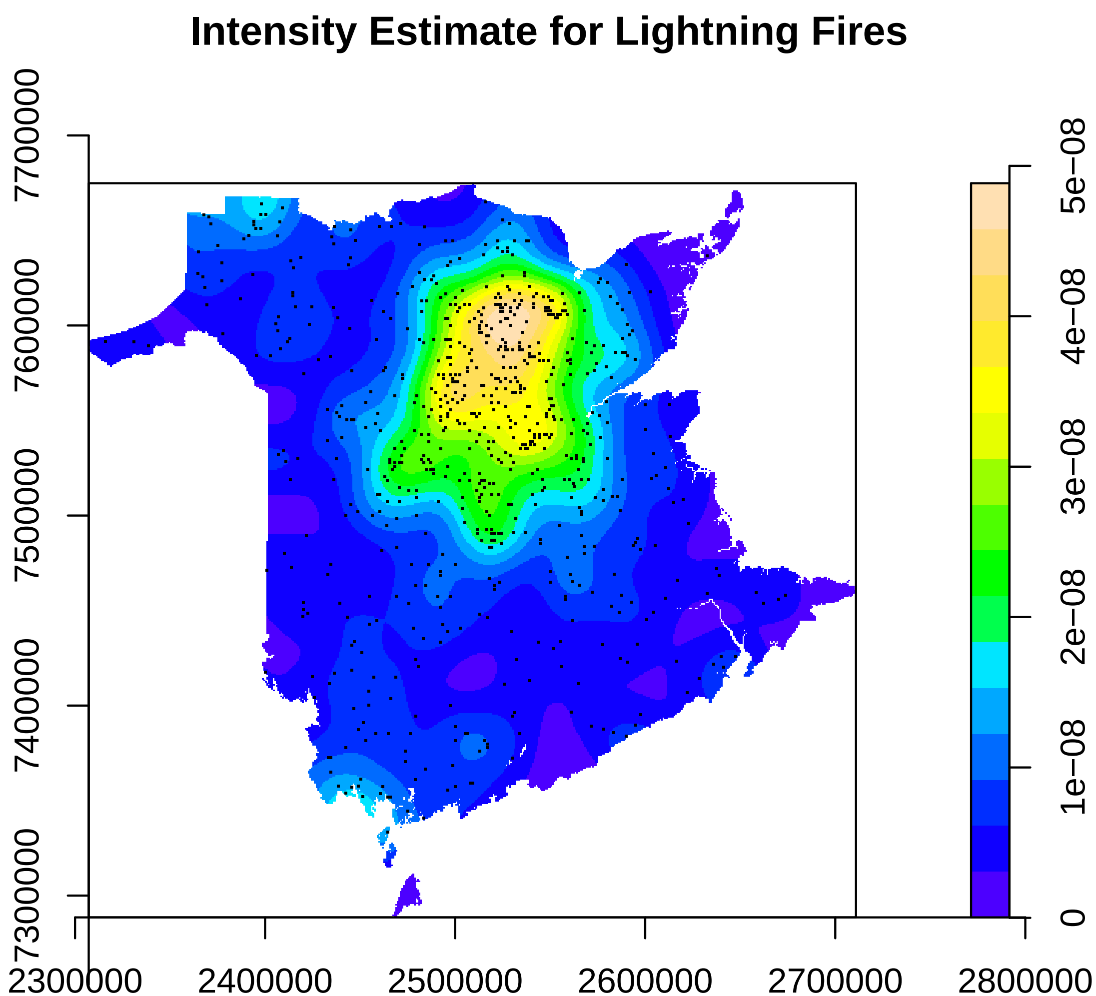
Lightning Fires

- Interesting to examine “naturally” caused fires separately.

Fires Started by Lightning



Lightning Trend



Second Order Effects

- Trend or “inhomogeneity” only a part of the story.
- Process not Poisson $\Rightarrow$ dependence or “interaction”.
- Simplest manifestation: either attraction (aggregation or clustering) or repulsion (“regularity”).
- In detecting such interaction Ripley’s K function is the basic tool.

Interpreting the K Function

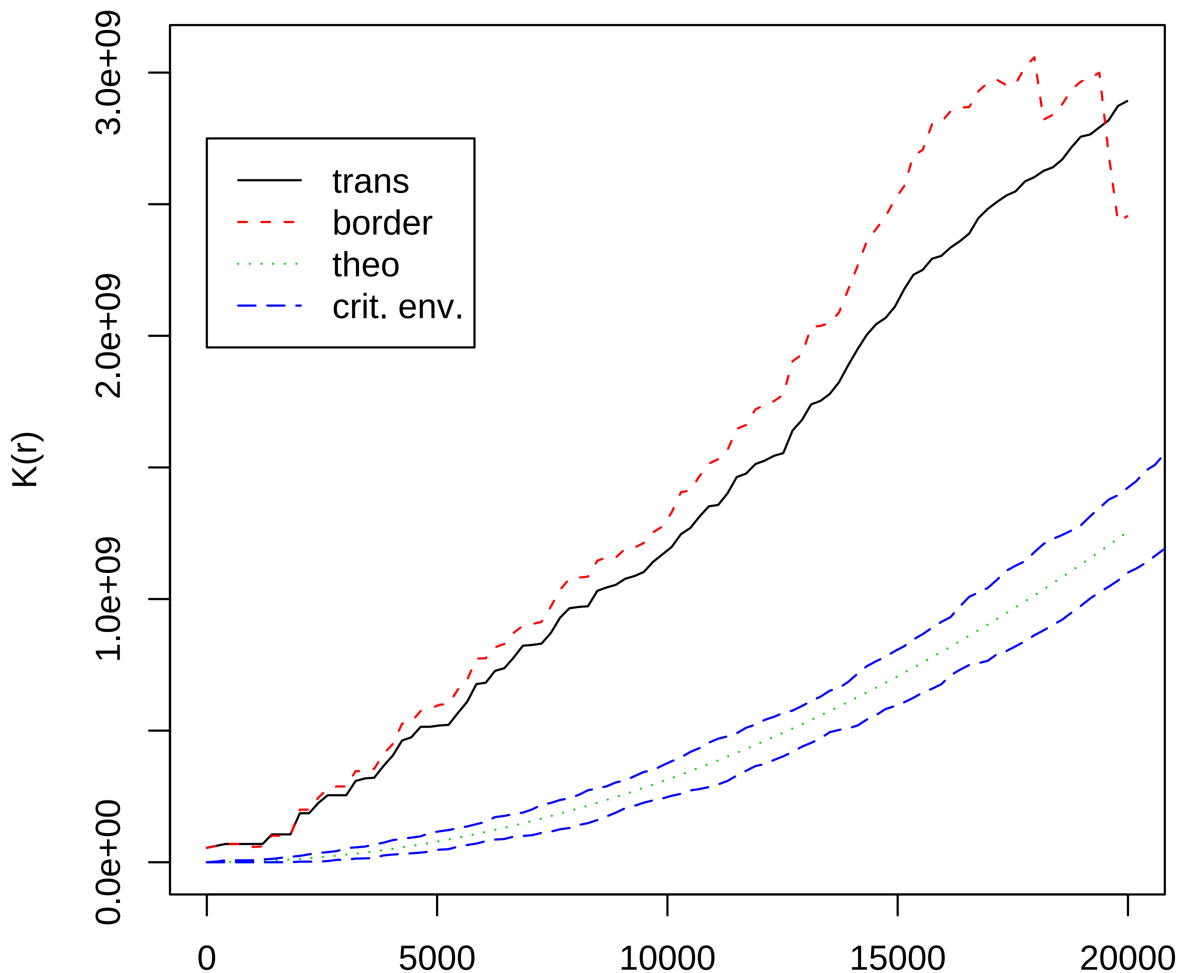
- Basic idea:
- Constant intensity Poisson process, ("complete spatial randomness", "CSR") $\Rightarrow \boldsymbol{K(r) = \pi r^2}$.
- Attraction (with impact at distance $\boldsymbol{r}$) $\Rightarrow \boldsymbol{K(r)}$ larger than under CSR.
- Repulsion $\Rightarrow \boldsymbol{K(r)}$ smaller than under CSR.

Estimating the K Function

- “DO NOT TRY THIS AT HOME.”
- Looks simple; edge effects strongly biasing; allowing for this is subtle.
- Some very clever people (e.g. Brian Ripley, Peter Diggle, Adrian Baddeley) have put a great deal of thought and effort into getting it right.
- Use software written by one of these experts; don’t roll your own!

The K Function for the Year 2000 Data

- For example, an estimate of the K function for the N. B. fires year 2000 data looks like:



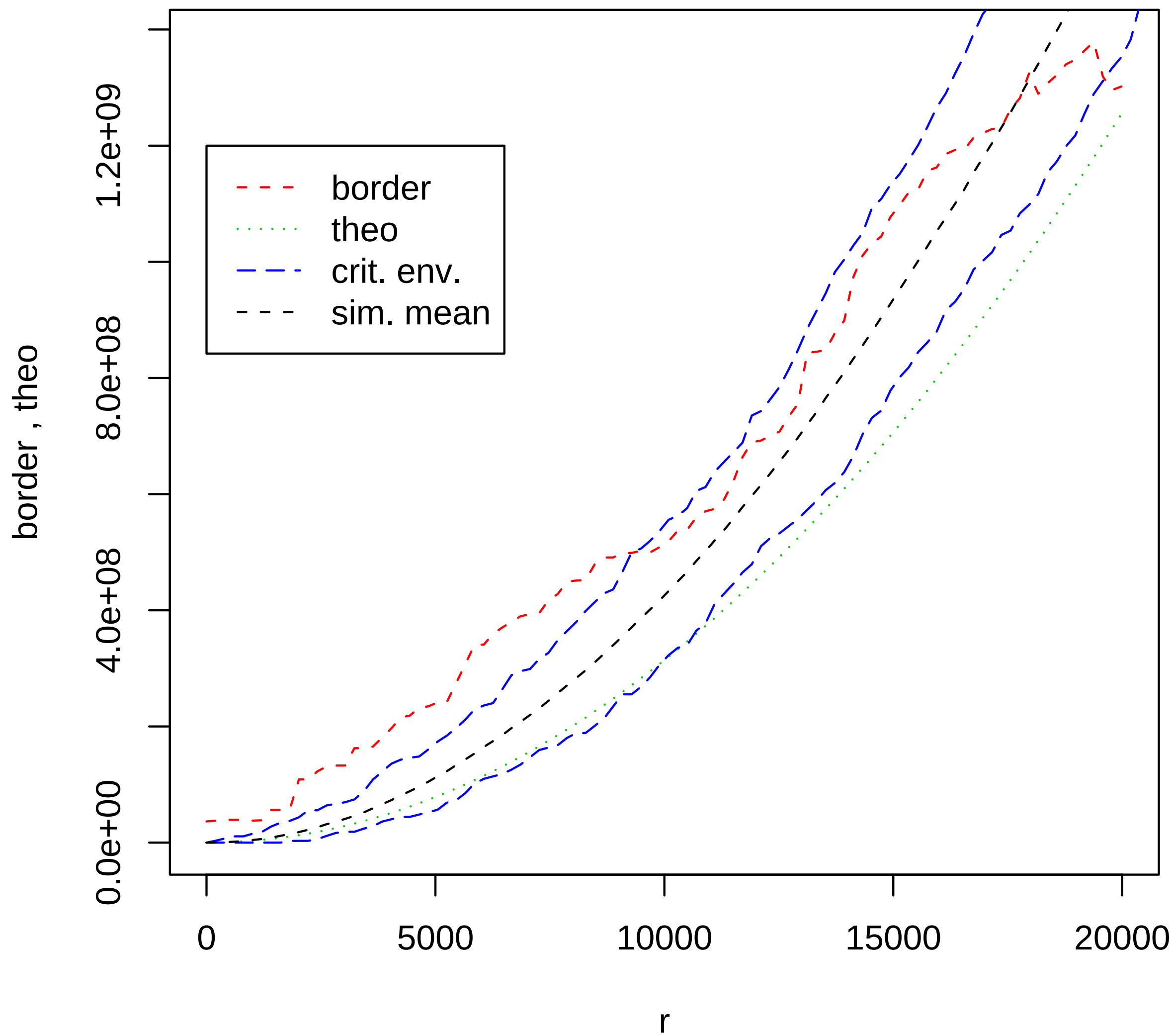
The K Function and Trend

- Both estimates lie entirely outside — far above — the critical envelope.
- Plot seems to shout “attraction” very loudly.
- But this could be due to regions of high concentration in the trend.
- Spatial trend and interaction are in theory *confounded*.

The Inhomogeneous K Function

- Cannot (strictly) be distinguished from looking at a single realization of the process.
- Here we're lucky; we have *multiple* realizations.
- Hence have an estimate of trend (already seen).
- Hence can calculate the *inhomogeneous* K function of Baddeley and Waagepetersen.

Inhomogeneous K Function for the Year 2000 Data



Modelling the Data

- Focus on a purely spatial approach.
- Point of view: each year “Nature” puts down a pattern of fire locations in the observation window $=$ New Brunswick.
- Basic theoretical assumption: these patterns are realizations of a Gibbs point process.
- Haven’t (yet) incorporated time.
- There is at least *some* insight to be gained from the purely spatial approach.

Fitting Models in `spatstat`

- Model fitting function: `ppm()` (“point process model”).
- Method: maximum pseudolikelihood.
- (Huang-Ogata method also available; not yet thoroughly tested.)
- Fits models of exponential family form.
- Models must be expressed in terms of their Papangelou conditional intensity functions.

$\textit{Model} = \textit{Trend} + \textit{Interaction}$

- Assume the Papangelou conditional intensity function has the form

$$\boldsymbol{\lambda}(\boldsymbol{u}, \underline{\boldsymbol{x}}) = \exp\{\boldsymbol{\phi}^{\mathsf{T}}\boldsymbol{b}(\boldsymbol{u}) + \boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{S}(\boldsymbol{u}, \underline{\boldsymbol{x}})\}$$

- $\boldsymbol{\phi}^{\mathsf{T}}\boldsymbol{b}(\boldsymbol{u}) =$ *trend* component.
- $\boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{S}(\boldsymbol{u}, \underline{\boldsymbol{x}}) =$ *interaction* component.
- Syntax of `ppm()` based on this decomposition.
- Syntax analogous with that of `glm()`/GLIM.
- “trend” <---> “linear predictor”, and “interaction” <---> “family”.

Trend Only Model

- Simplest model: trend only.
- Estimate of trend available.
- Exponential family model: assume the intensity for the given year is proportional to the overall trend.

Using an Offset Term

- Explicitly:

$$\boldsymbol{\lambda}(\boldsymbol{u}, \underline{\boldsymbol{x}}) = \boldsymbol{\lambda}(\boldsymbol{u}) = \boldsymbol{\beta\tau}(\boldsymbol{u}) = \exp\{\boldsymbol{\phi} + \log(\boldsymbol{\tau}(\boldsymbol{u}))\}$$

- $\boldsymbol{\phi} = \log(\boldsymbol{\beta})$ only parameter to be estimated.
- $\log(\boldsymbol{\tau}(\boldsymbol{u}))$ called “offset” term. ([Generalized] linear modelling terminology.)

Trend Only Model (Cont'd.)

- `spatstat` syntax:

```
fit1 <- ppm(X.00,~offset(log(intens)),
            covariates=list(intens=intens))
```

- `intens` = non-parametric estimate of the over-all trend.
- `X.00` = point pattern object = forest fire locations for the year 2000.

The Resulting Fit

- “print method” for `ppm` objects in `spatstat` produces:

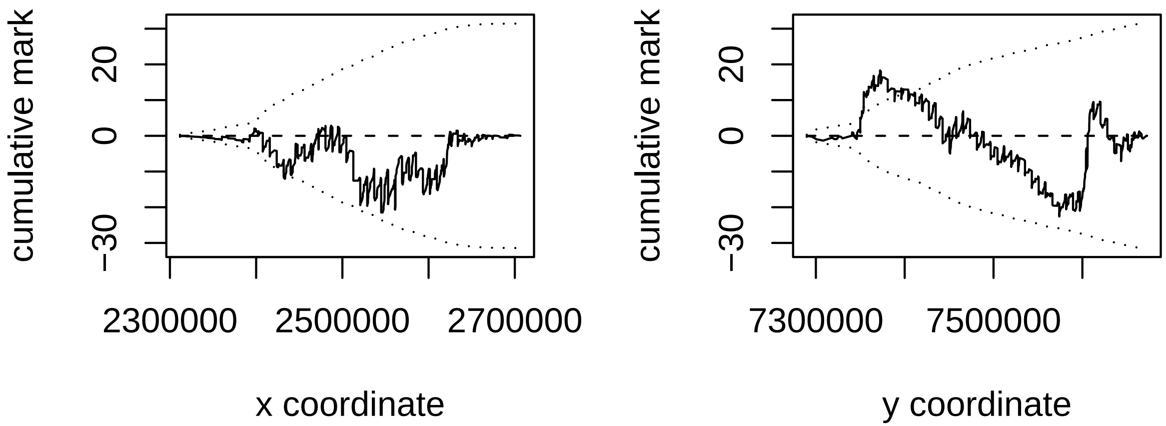
```
        [Stuff omitted.]

Fitted coefficients for trend formula:
(Intercept)
  -2.928027
```

- Note $\exp(-2.928027) = 0.0535$, a bit less than $1/16$.
- Indicates number of fires for the year 2000 is a bit less than average.
- Which is indeed the case.

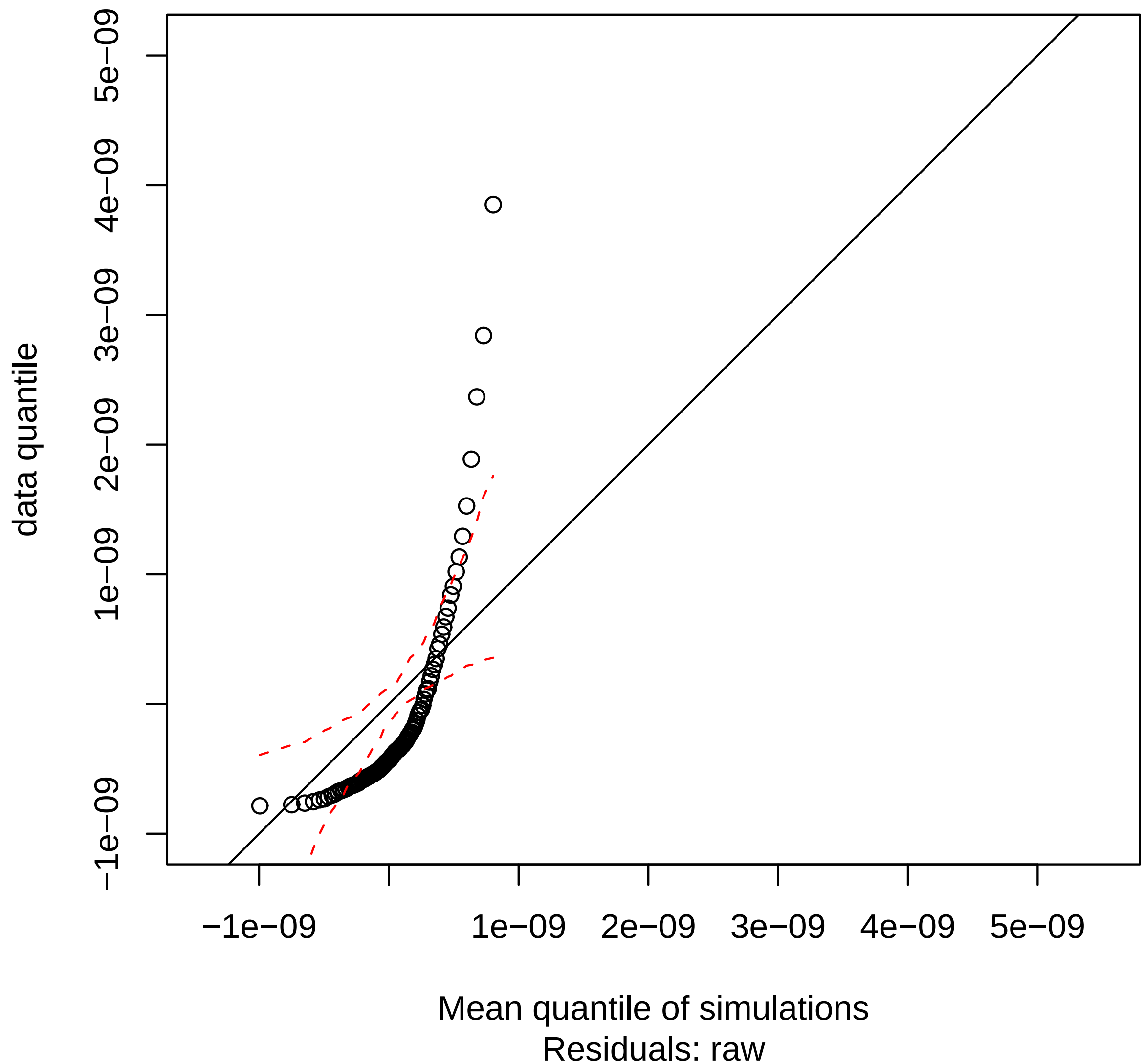
Diagnostics

- Output doesn’t say whether the model is actually sensible.
- Know it isn’t — the K function plots said there is *interaction* as well as trend.
- Goodness of fit is often assessed via residual plots.
- These are now available, for point pattern models, in `spatstat`.
- Lurking variable plot: information about fit of trend component.
- Quantile-quantile plot: information about fit of interaction component.
- Some other plots are available.

Lurking Variable Plot — Trend Only



Quantile-quantile Plot — Trend Only



A Model With Interaction

- “Knew” that the QQ plot would say “No.”
- Try adding a Geyer type interaction to model attraction.
- “Geyer” generalizes “Strauss”.
- Adds “saturation” parameter; makes model well-defined for $\gamma > 1$.
- Hence attraction as well as repulsion can be modelled.
- In `spatstat`:

```
fit2 <- ppm(X.00,~offset(log(intens)),inter=Geyer(10000,5),
            covariates=list(intens=intens))
```

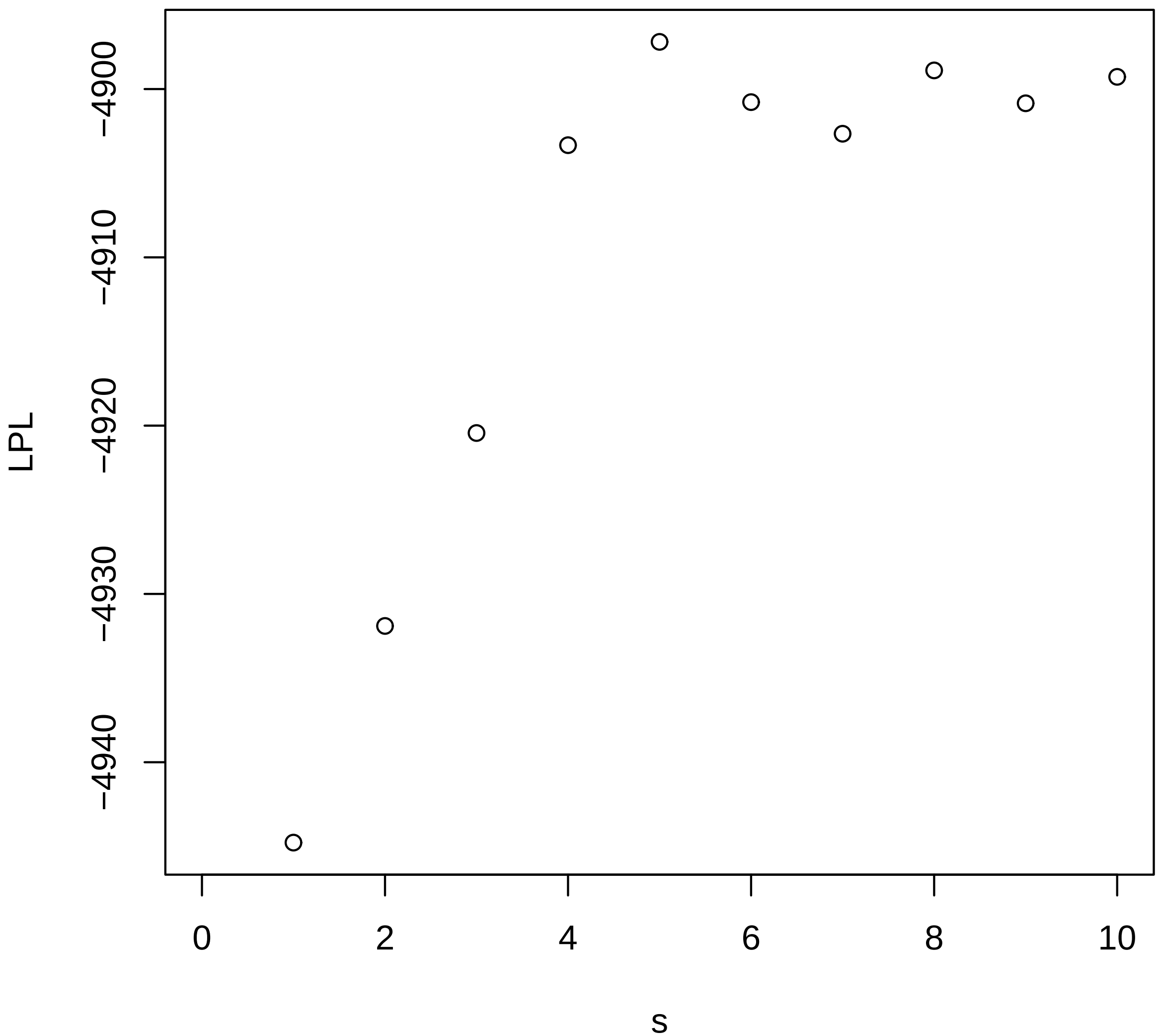
The “Irregular” Parameters

- Interaction radius and saturation parameter are “irregular” parameters.
- These do not conform to the exponential family model.
- Not estimated by `ppm()`.
- Must be estimated/guessed at by other means and pre-specified.
- Guessed at interaction radius from (inhomogeneous) K function plot; K function estimate outside of critical envelope for $r < 10000$.

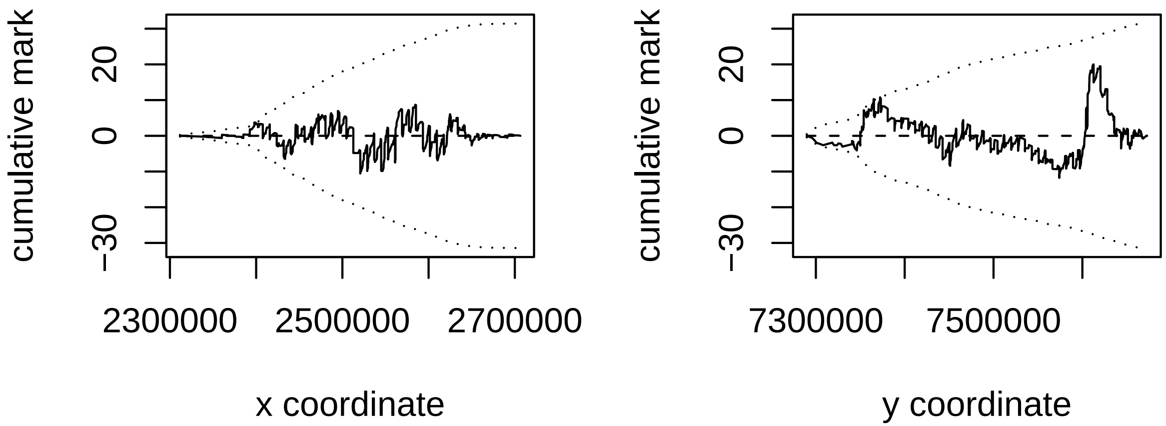
The Saturation Parameter

- Guessed at saturation parameter “s” via rough “profile pseudolikelihood” procedure.
- Fitted models with interaction `Geyer(10000,s)` for s in $\{1, 2, \ldots, 10\}$.
- Plot of “profile” shown on the next slide.

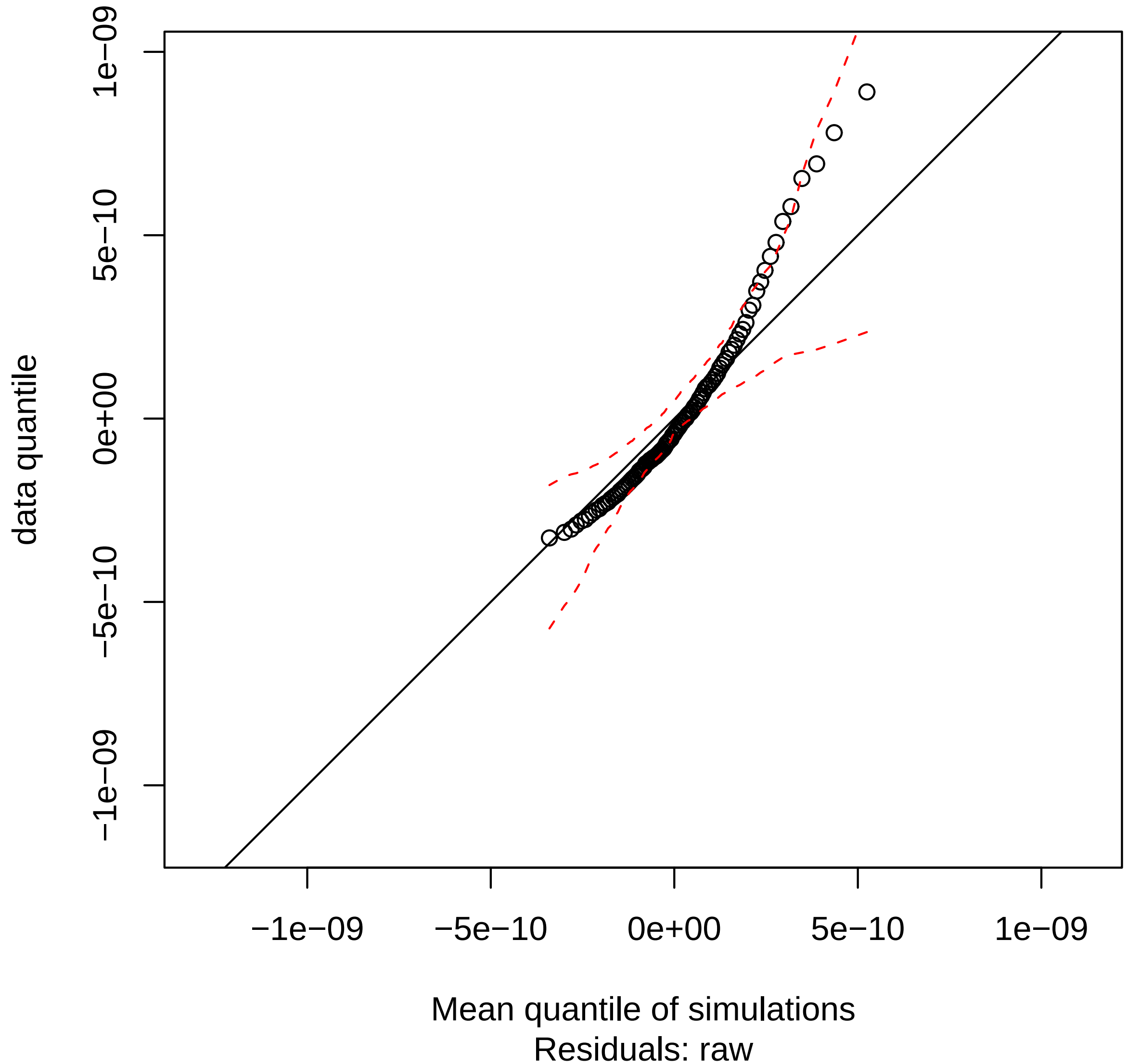
Profile Pseudolikelihood for "s"



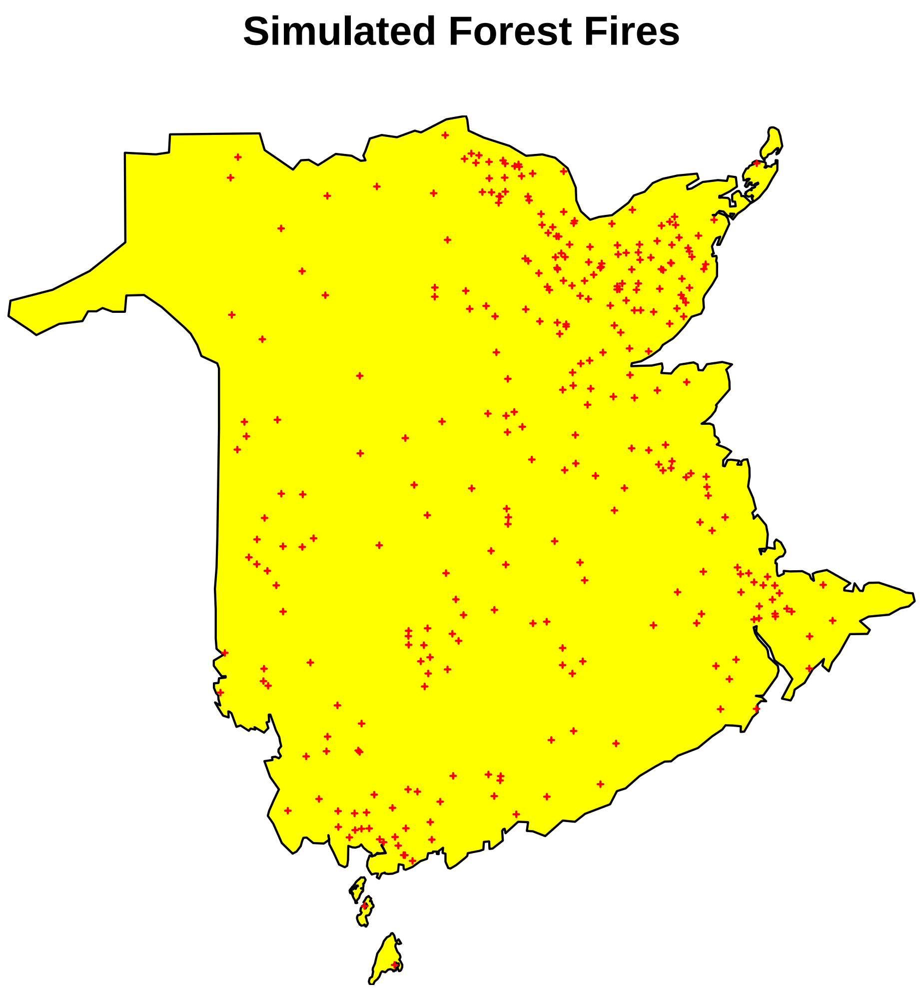
Lurking Variable Plot — Trend + Geyer



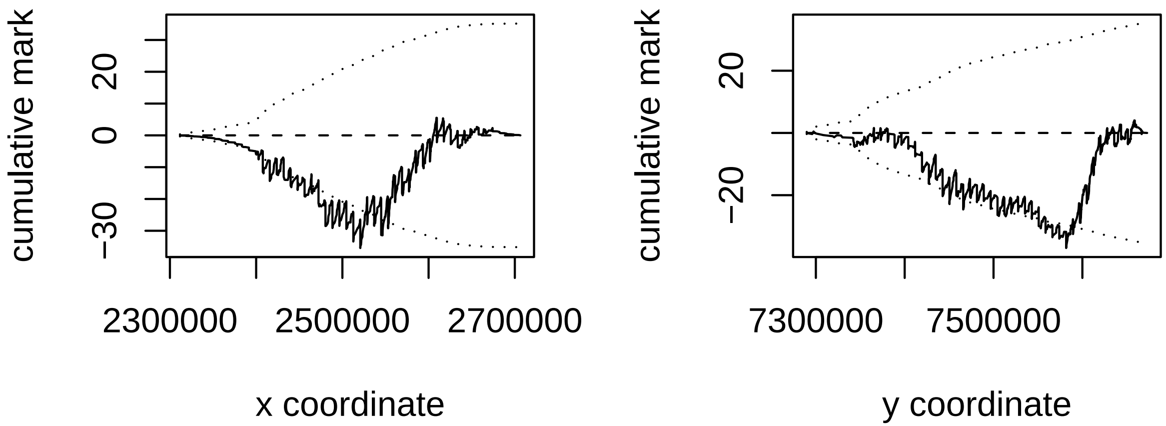
QQ Plot — Trend + Geyer



Simulated Data

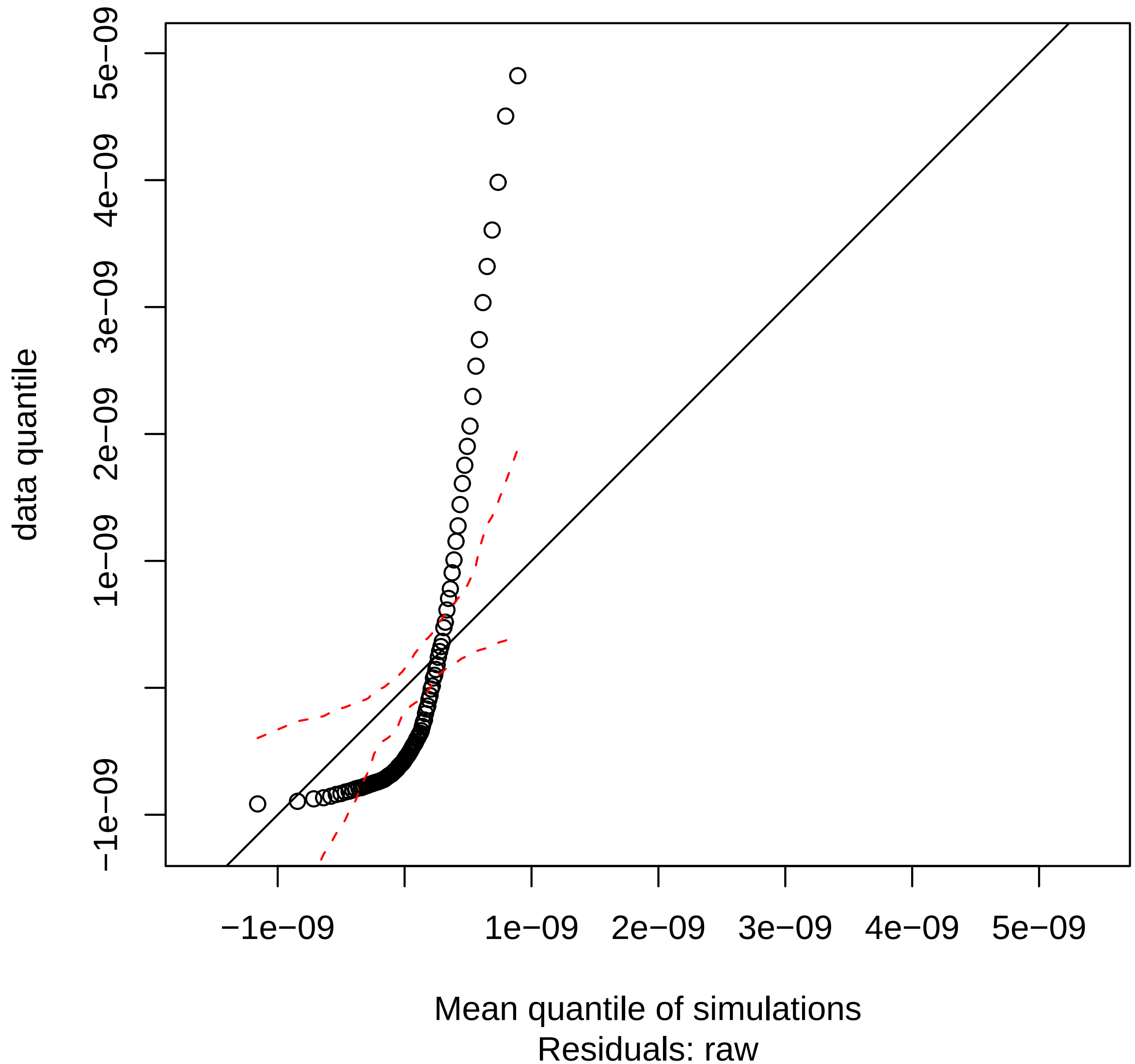


Lurking Variable Plot for Simulated Data



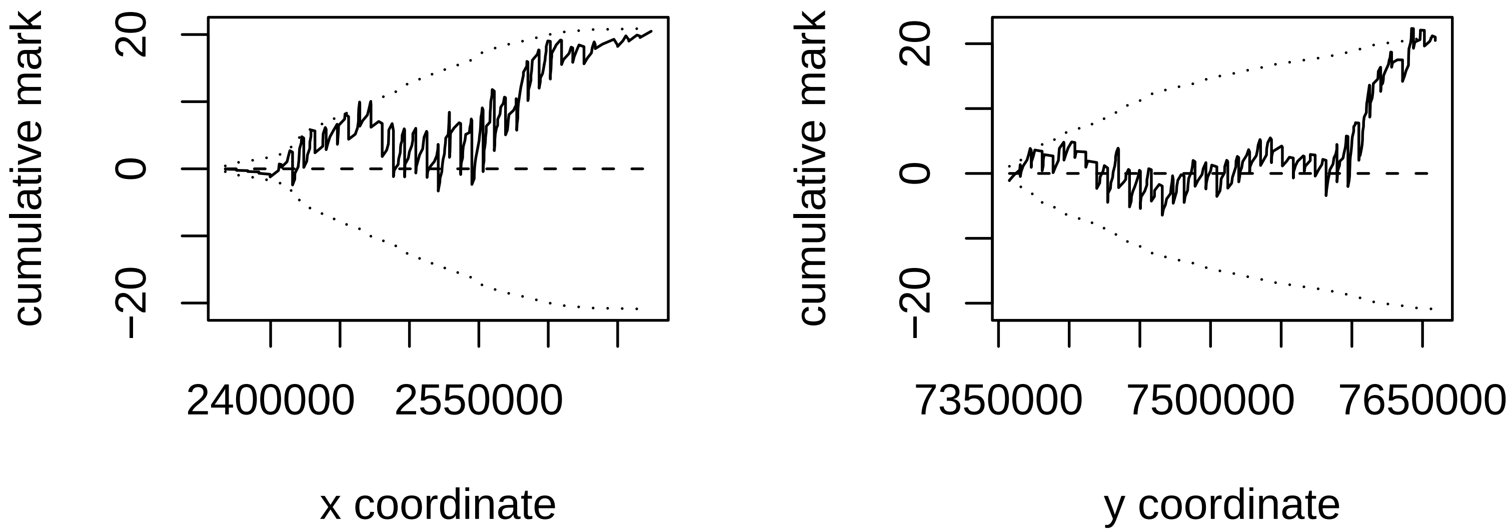
- Simulated from Trend + Geyer; fitted Trend.

QQ Plot for Simulated Data



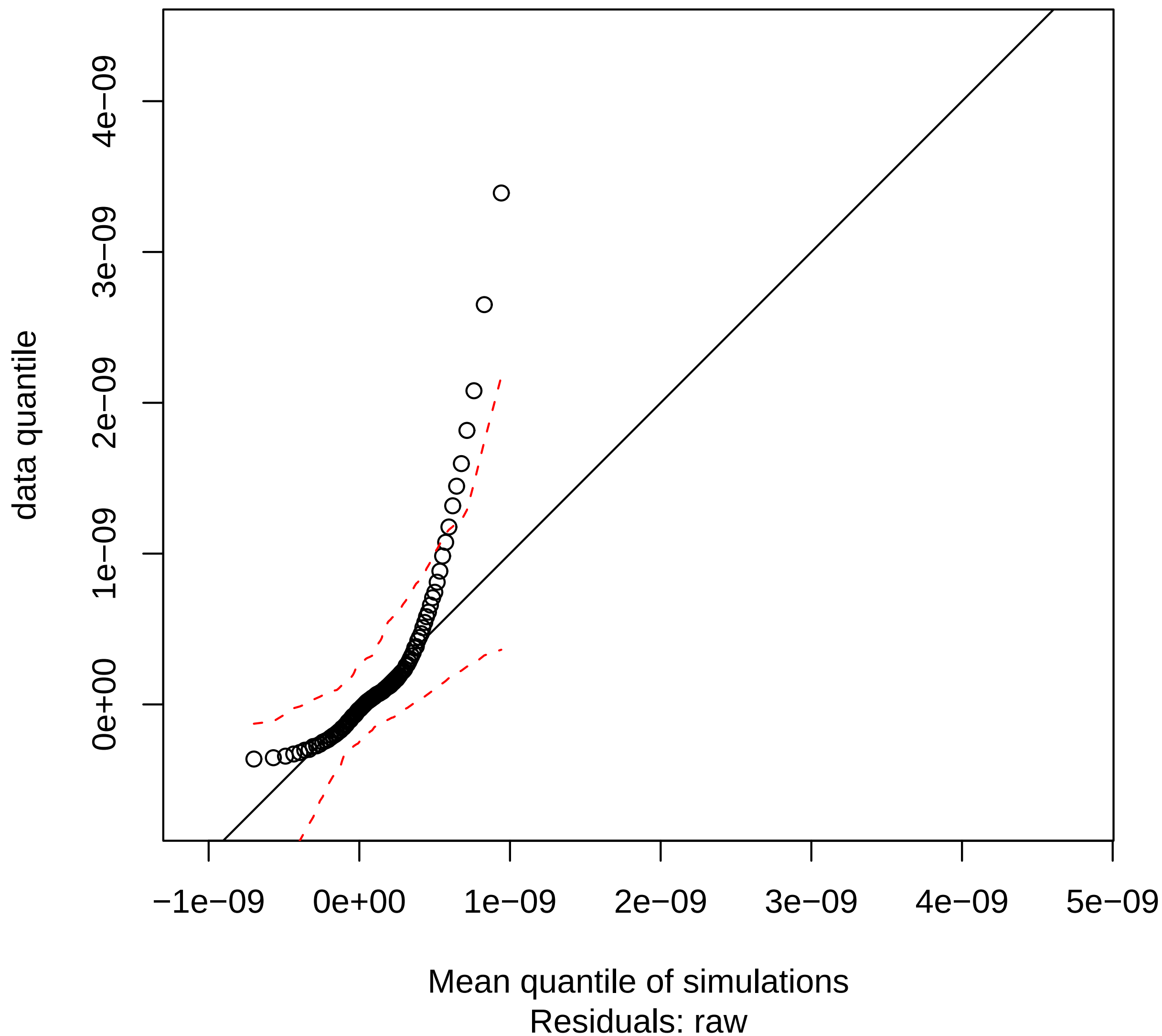
- Simulated from Trend + Geyer; fitted Trend.

Lurking Variable Plot for Simulated Data



- Simulated from Trend $+$ Geyer; fitted Trend $+$ Geyer.

QQ Plot for Simulated Data



- Simulated from Trend + Geyer; fitted Trend + Geyer.

Comments on the Trend + Geyer Model

- Fit not very good.
- OTOH, maybe not too bad compared with fit to simulated data.
- Seem to need to do better with trend as well as with interaction.
- Shouldn't expect wonders of the attempted model.
- It was totally "ad hoc".
- A "good" model would use the temporal/sequential nature of the data.
- As noted, temporal information is available, but not yet brought in to play.

Question on Spatio-Temporal Modelling

- Interesting question: how to relate a spatio-temporal model to a purely spatial model?
- Can we formulate a spatio-temporal model so as to infer a reasonable Gibbs model for the aggregate, end-of-year, process?
- “Reasonable” $=$ having a tractable (computable) Papangelou conditional intensity function.

Further Desiderata for Models

- Make use of “background” information on terrain and vegetation.
- May be possible to get such information from a GIS.
- Make use of weather conditions.
- Some weather information available in the N. B. DNR data.
- However much of this information consists of “missing values” — presumably unrecorded.
- Obtain weather information from Environment Canada?

Cox Process?

- The Cox process seems intuitively plausible as a model for these data.
- It is interesting in theory at least.
- Fitting presents substantial challenges.
- Not clear how much progress can be made.
- Possible interplay between Cox process idea and spatio-temporal modelling?
- E.g. think of the underlying Gaussian random field as varying continuously in time.
- Is there any real sense or any practical mileage in this?

Conclusions

- Substantial evidence of “attraction” between fires for the year 2000 data.
- But a Cox process might provide a better description of the data than an explicit model for attraction.
- [Looking at other years seems to support this.]

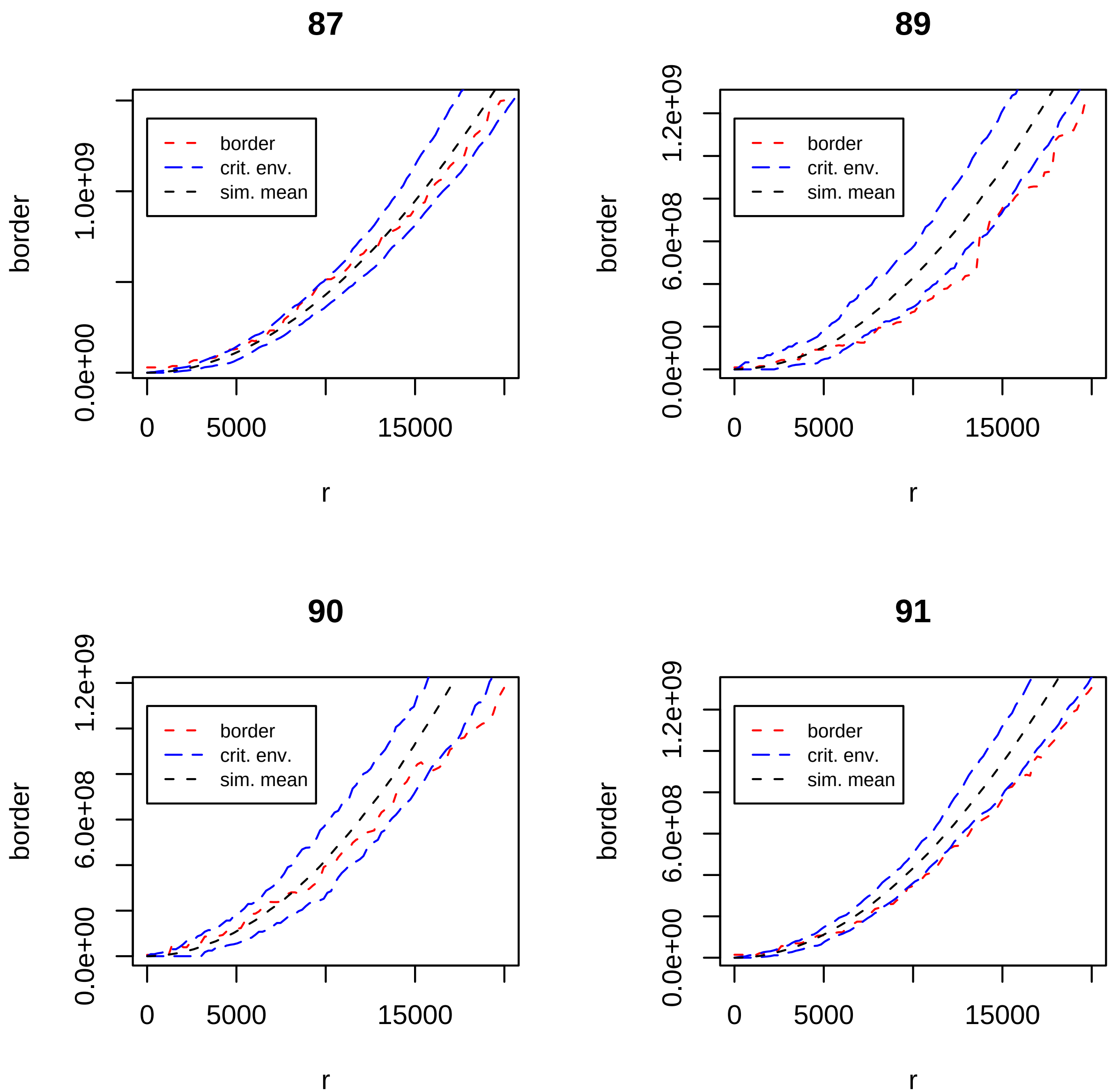
Homework

- Proceed with data cleaning; resolve anomalies, fill in missing data via discussions with N. B. DNR people.
- Obtain, and implement the use of, data on terrain, vegetation, and weather conditions.
- Formulate and fit appropriate spatio-temporal models to the data.
- Investigate relationships between spatio-temporal models and purely spatial models for the yearly aggregate.
- Investigate fitting a Cox process to the data; develop practical methods for this.
- Investigate possible interplay between Cox processes and spatio-temporal modelling.
- Bring peace and harmony to mankind.

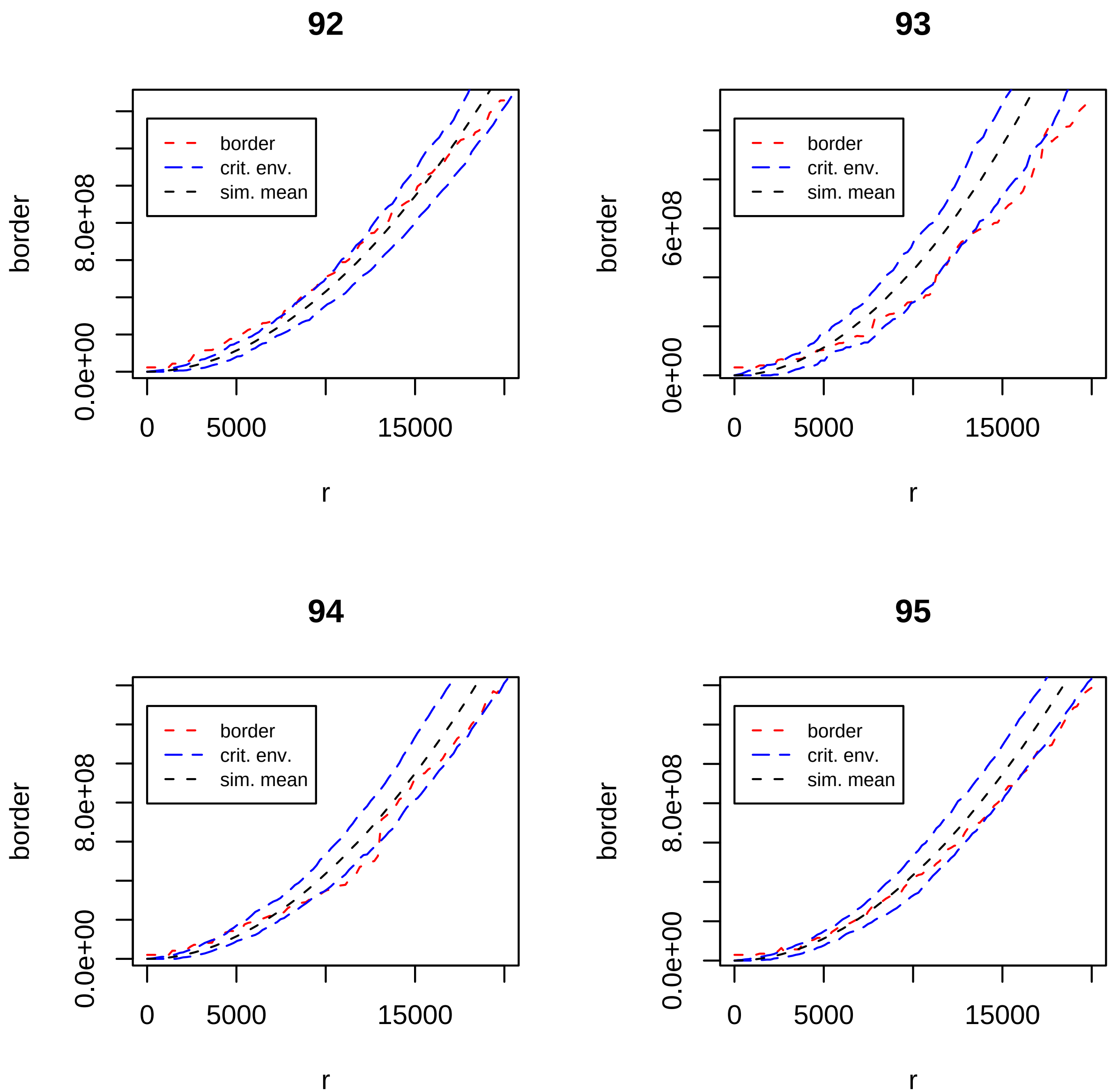
Final Remarks

- Not a lot of information squeezed out of the data so far.
- However, as they say on election night, “It is early days yet.”
- This is a rich collection of data.
- Lots of scope for experimenting with ideas for point process modelling.
- These data will be made generally available as part of the `spatstat` package (obtainable from CRAN) in the near future.

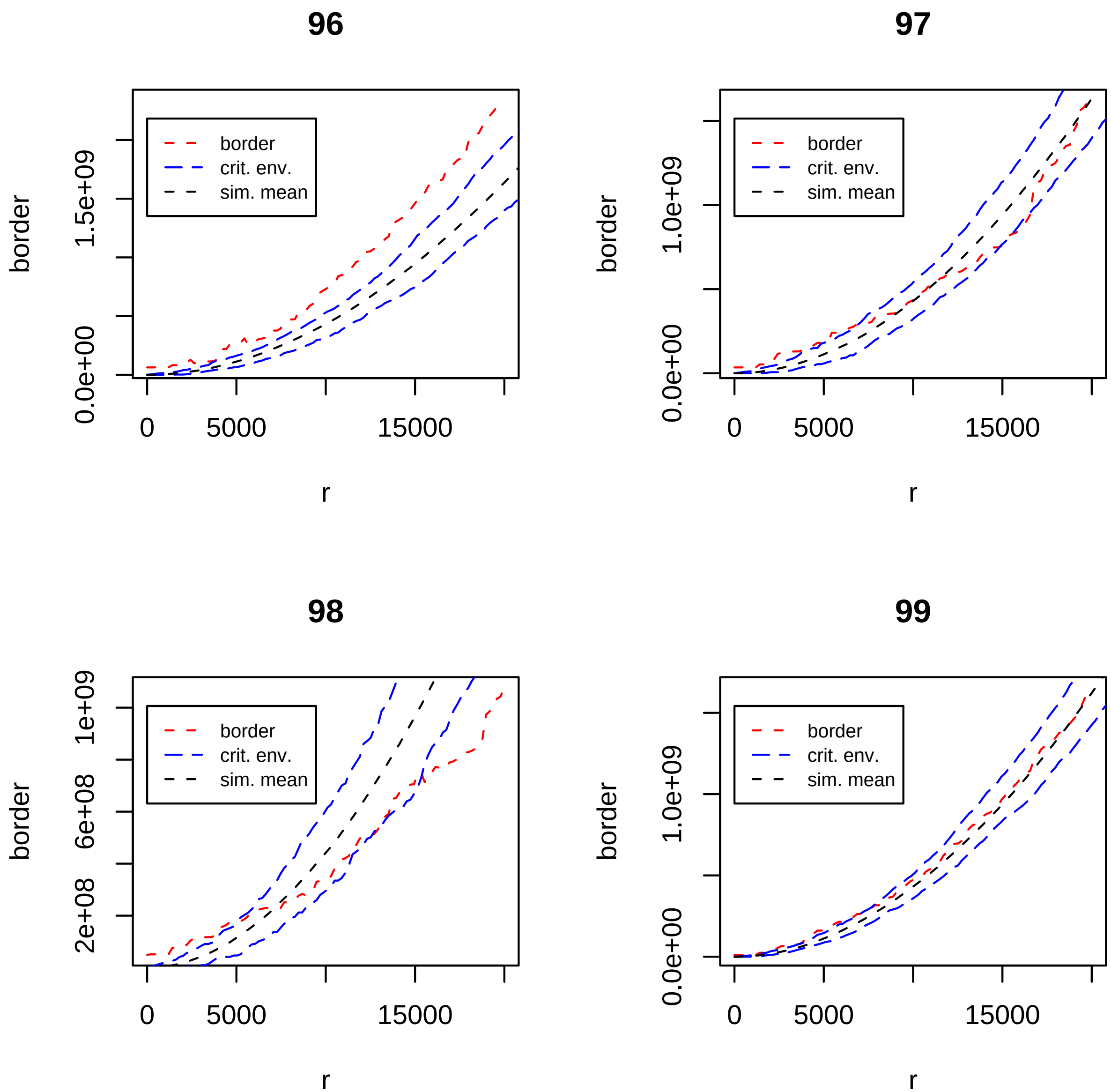
Appendix: Inhomog. K Function Plots (1)



Appendix: Inhomog. K Function Plots (2)



Appendix: Inhomog. K Function Plots (3)



Appendix: Inhomog. K Function Plots (4)