A novel bootstrap for Ripley's K

Bruce Tabor and Ross Sparks

CSIRO DIGITAL PRODUCTIVITY FLASGSHIP
www.csiro.au

OUTLINE

1. A weird bootstrap
2. The local K function and marked point bootstrap
3. Loh and Stein's marked point bootstrap
4. Baddeley/Turner marked point bootstrap
5. Loh and Stein's bootstrap in the limit
6. The weird bootstrap: PoissonBootS
7. Variance of Ripley's K & local K functions
8. Gordon Square
9. Conclusions

A weird bootstrap (1)

Bootstrap statistic

An estimate from random sampling of data with replacement.

Classical bootstrap for confidence intervals on independent data

Mean of $iid\ X_i, i = 1, ...N$, $\hat{\theta} = \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$

A bootstrap sample estimate of $\hat{\theta}^*$, $\hat{\theta}^* = \bar{X}^* = \frac{1}{N}\sum_{j=1}^{N} X_j^*$

where X_j^* data points drawn with replacement from $X = (X_1, ...X_N)^T$

For a bootstrap estimate 95 % CIs on $\bar{X}$, draw B such samples and find the 2.5 % and 97.5 % quantiles. Say $N = 100$ and $B = 199$,

Bootstrap sample	b	1	2	3	4	5	4	7	...	199
Draws from X	N	100	100	100	100	100	100	100	...	100

A weird bootstrap (2)

M-of-N bootstrap (prior art)

Let M be the number of draws from X and $M \neq N$. Say $N = 100$, $B = 199$, M is 40 % of N, i.e. $M = 40$:

Bootstrap sample	b	1	2	3	4	5	4	7	...	199
Draws from X	M	40	40	40	40	40	40	40	...	40

A weird bootstrap (3)

Bootstrap with variable draw size (a proposal)

Let M be a variable, m. Each bootstrap sample has $\mathrm{E}[m] = f_s N = M$ draws with a variance $\mathrm{var}[m] = f_v M = f_v f_s N$.

Say $N = 100$, $B = 199$, and $f_s = 0.4$ ($\mathrm{E}[m] = 40$) and $f_v = 0.5$ ($\mathrm{var}[m] = 20$). Using a binomial distribution for m,

Bootstrap sample	b	1	2	3	4	5	4	7	...	199
Draws from X	M_b	35	40	43	51	43	38	38	...	39

Two parameters: f_s and f_v.

If $f_v = 1$, then $\mathrm{var}[m] = \mathrm{E}[m] \Rightarrow$ Poisson distribution

"Poisson-referenced Bootstrap Sampling" (PoissonBootS)

Nomenclature: $PB_{40,50} \Rightarrow$ binomial draw size with $f_s = 0.4$ & $f_v = 0.5$

Why? Instead of drawing X_i, we'll be drawing $X_i(r)$, i.e functions.

The local K function and marked point bootstrap (1)

The empirical Ripley's K function

$$\hat{K}(r) = \frac{A}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} I\{||x_i - x_j|| \leq r\} e(x_i, x_j, r)$$

Is the mean of N "local K" functions:

$$\begin{aligned} \hat{K}(r) &= \frac{1}{N} \sum_{i=1}^{N} \frac{A}{N-1} \sum_{j=1, j \neq i}^{N} I\{||x_i - x_j|| \leq r\} e(x_i, x_j, r) \\ &= \frac{1}{N} \sum_{i=1}^{N} \hat{K}_i(r) \end{aligned}$$

The local K function and marked point bootstrap (2)

We “mark” each point in our observation window with its cumulative distance function $\hat{m}_i(r)$, a simple transformation of its local K function

$$\hat{m}_i(r) = \sum_{j=1, j\neq i}^{N} I\{||x_i - x_j|| \leq r\} e(x_i, x_j, r) = \frac{N-1}{A}\hat{K}_i(r)$$

Resample these marks with replacement: “marked point bootstrap”

Originally suggested by Loh and Stein (2004) for spatial point processes.

Fast $\Rightarrow$ no rearrangement of points or recalculation of $\hat{m}_i(r)$ (or $\hat{K}_i(r)$)

BUT...

How do we resample (or redraw) these marks?

Loh & Stein's marked point bootstrap (1)

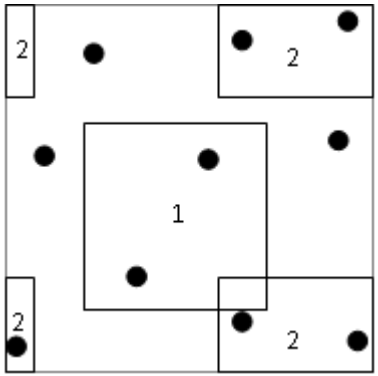
Drawing of marks $\hat{m}_i(r)$ is done in blocks.

Integer number of blocks: Loh & Stein use 4, 16 or 256.

Total block area equals area of window. ($A/4$ for 4 blocks)

Blocks placed randomly; toroidal wrapping.

Marks ($\hat{m}_i(r)$) included in each block are drawn.



Restricted to "regular" windows to avoid bias in draw probabilities.

Loh & Stein's marked point bootstrap (2)

Each bootstrap sample for N points & N_B blocks :

$$\hat{K}^*_{LS}(r) = \frac{A}{(\sum_{k=1}^{N_B} n_k)(\sum_{k=1}^{N_B} n_k - 1)} \sum_{k=1}^{N_B} \sum_{j=1}^{n_k} \hat{m}_{kj}(r)$$

$\hat{m_{kj}}(r)$ $(j = 1, ..., n_k)$ are the marks in block k $(k = 1, ..., N_B)$.

If $n = \sum_{k=1}^{N_B} n_k$ is the total number of marks drawn, we can show:

$$\hat{K}^*_{LS}(r) = \frac{1}{n} \sum_{k=1}^{N_B} \sum_{j=1}^{n_k} \frac{A}{(n-1)} \hat{m}_{kj}(r) = \frac{1}{n} \sum_{k=1}^{N_B} \sum_{j=1}^{n_k} \left(\frac{N-1}{n-1} \right) \hat{K}_{kj}(r)$$

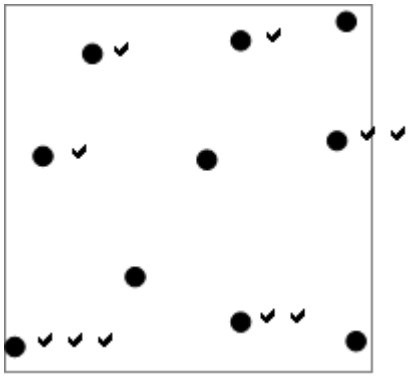
$\hat{K}^*_{LS}(r)$ is approximately the mean of the number of local K functions sampled.

The number of draws in each bootstrap sample (n) is a variable with mean N.

Baddeley/Turner marked point bootstrap (1)

Drawing of local K's is random with replacement.

The number of marks drawn equals the number of points



No restriction on window shape.

Faster than Loh & Stein's bootstrap — avoids calculating inclusion in randomly placed blocks.

Baddeley/Turner marked point bootstrap (2)

Implemented in *spatstat* as the function *lohboot*

Each bootstrap sample has the same form as Ripley's K

Except that the marks are drawn randomly with replacement.

$$\hat{K}^*_{BT}(r) = \frac{1}{N}\sum\nolimits_{j\in I_N} \hat{K}_j(r) = \frac{1}{N}\sum\nolimits_{j\in I_N} \frac{A}{N-1}\hat{m}_j(r)$$

where I_N is a set of numbers of size N chosen randomly with replacement from $\{1, 2, ...N\}$.

The number of draws in each bootstrap sample (N) is constant.

Loh & Stein's bootstrap in the limit (1)

As $N_B \to \infty$ and $A/N_B \to 0$

Probability a point is drawn in a single random block is $p_k = 1/N_B$.

In each block the expected number of points $\mathrm{E}[n_k] = N/N_B$

For N_B blocks the expected bootstrap draw size is $\mathrm{E}[n] = N$.

To a first approximation the sampling is binomial, so $\mathrm{var}[n_k] = Np_k(1-p_k)$

Assume no spatial correlation; variance for N_B blocks is:

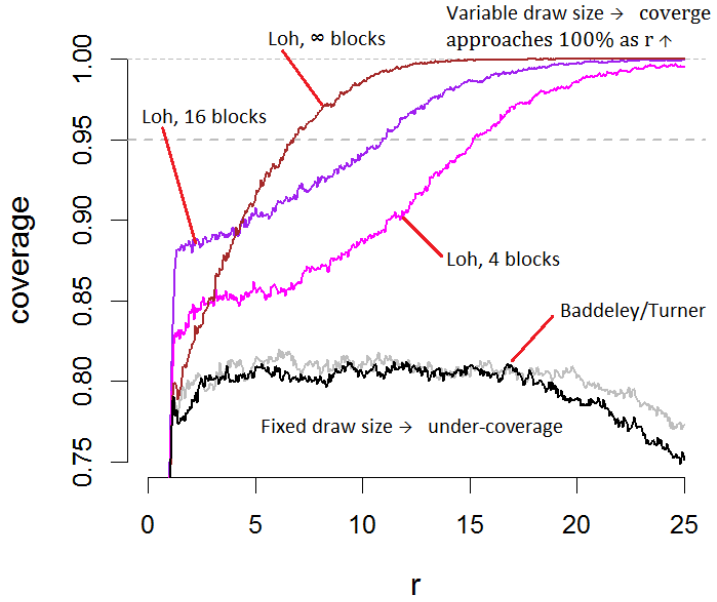
$$\begin{aligned} \mathrm{var}[n] &= N_B N p_k (1-p_k) \\ &= N(1 - \frac{1}{N_B}) \\ \lim_{N_B \to \infty} \mathrm{var}[n] &= N = E[n] \end{aligned}$$

The draw size for Loh's bootstrap converges to a Poisson distribution.

Loh & Stein's bootstrap in the limit (2)

Baddeley/Turner bootstrap vs Loh's bootstrap (4, 16 and ∞ blocks)

95 % CI coverage; HPP mean of 100 points; 100x100 square; 5000 simulations



The weird bootstrap: PoissonBootS (1)

Bootstrap draw size $n_{boot} \neq N$; and vary n_{boot} between bootstraps:

$$\hat{K}^*_{PB_{fs,fv}}(r) = \frac{1}{n_{boot}} \sum_{j \in I_{boot}} \left(\frac{A}{n_{boot} - 1} \right) \hat{m}_j(r) = \frac{1}{n_{boot}} \sum_{j \in I_{boot}} \left(\frac{N-1}{n-1} \right) \hat{K}_j(r)$$

I_{boot} is a set of n_{boot} numbers from $\{1, 2, ...N\}$ chosen randomly with replacement in a manner defined by f_s and f_v

f_s and f_v determine the mean and variance of the distribution from which n_{boot} is drawn : use binomial – Poisson – negative-binomial family

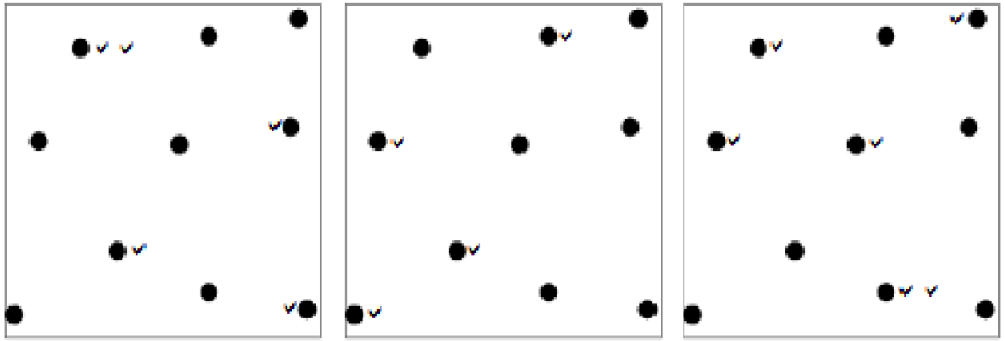
$$\begin{aligned} \mathrm{E}[n_{boot}] &= f_s N \\ var[n_{boot}] &= f_v \, \mathrm{E}[n_{boot}] \\ &= f_v f_s N \end{aligned}$$

The weird bootstrap: PoissonBootS (2)

Drawing of local K's is *still* random with replacement.

Number of marks drawn *is not necessarily equal* to the number of points

Number of marks drawn *may vary* between bootstrap samples



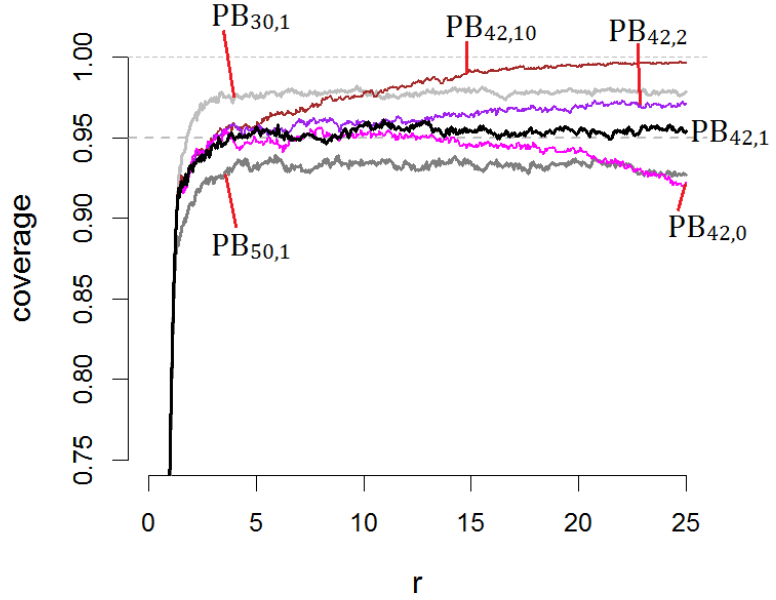
$$\hat{m}_i(r) = \sum_{j=1, j\neq i}^{N} I\{||x_i - x_j|| \leq r\} e(x_i, x_j, r) \quad \text{or} \quad \hat{K}_i(r) = \frac{A}{N-1}\hat{m}_i(r)$$

Still no restriction on window shape. *Still* very fast.

The weird bootstrap: PoissonBootS (3)

After a bit of “tuning”:

95 % CI coverage; HPP mean of 100 points; 100x100 square; 5000 sims.



Variance of Ripley's K and the local K (1)

HPP in a square window $A = d^2$ uncorrected for edge effects

$$\operatorname{var}[K(r)] = \frac{2d^2}{N(N-1)} \operatorname{E} K(r) - \frac{2}{N(N-1)} (\operatorname{E} K(r))^2 + \frac{4d^4(N-2)}{N(N-1)} \operatorname{E} H^2(r)$$

(Lang & Marcon, ESAIM Probability and Statistics, 2013)

$$\operatorname{var}[K_i(r)] = \frac{d^2}{N-1} \operatorname{E} K(r) - \frac{1}{N-1} (\operatorname{E} K(r))^2 + \frac{d^4(N-2)}{N-1} \operatorname{E} H^2(r)$$

(Me , here, now)

Where:

$\operatorname{E} K(r) = r^2 \left(\pi - \frac{8r}{3d} + \frac{r^2}{2d^2}\right)$ and $\operatorname{E} H^2(r) = \frac{r^5}{d^5}\left(\frac{8}{3}\pi - \frac{256}{45}\right) + \frac{r^6}{d^6}\left(\frac{11}{48}\pi - \frac{56}{9}\right) + \frac{8}{3}\frac{r^7}{d^7} - \frac{1}{4}\frac{r^8}{d^8}$

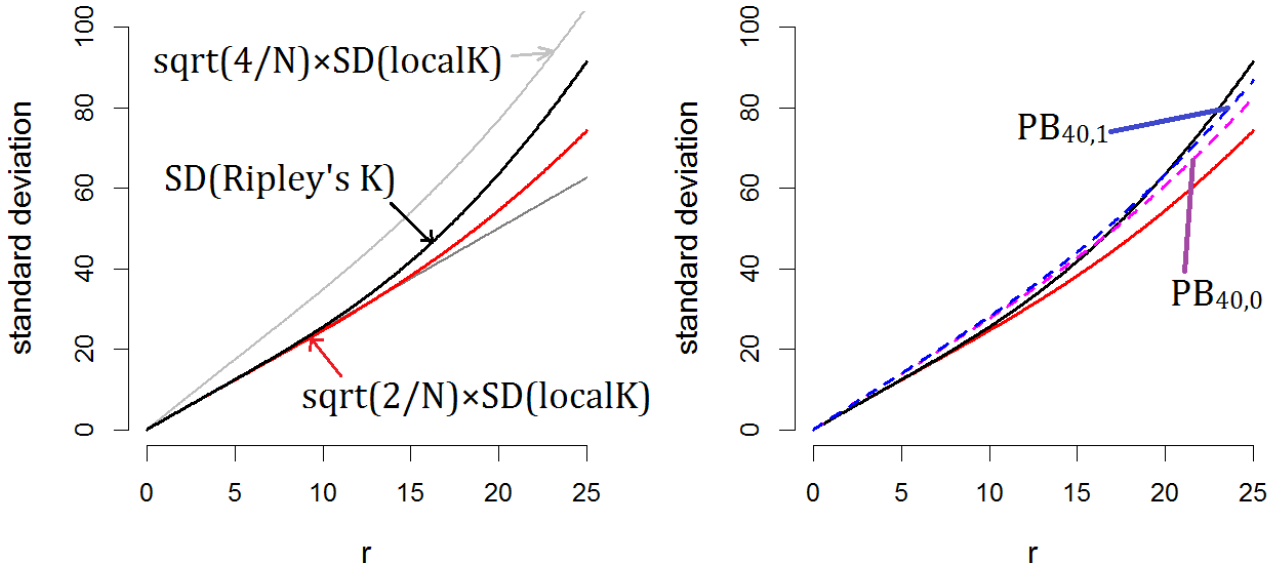
Note: small corrections have been omitted from variances that are negligible when $N > 15$

Variance of Ripley's K and the local K (2)

Ripley's and local K: 100-point (mean) HPP in a 100x100 square

LEFT: SD of Ripley's K (black) vs $\sqrt{2/N}$ SD of the local K (red, $N = 100$).

RIGHT:Two PoissonBootS approximations to SD of Ripley's K.



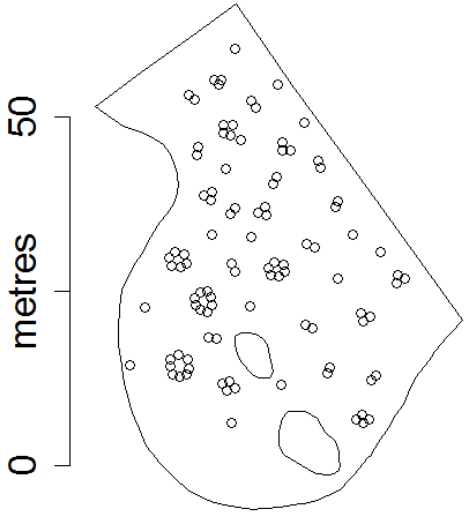
Gordon Square (1)

Can a cluster process explain the (second moment) pattern of people in Gordon Square London?

Gordon Square (2)

People in Gordon Square (*spatstat, data(gordon)*) (Bevan 2012)

"Records the location of people sitting on a grass patch in Gordon Square, London, at 3pm on a sunny afternoon". The 99 people are located singly or in small clusters.



Gordon Square (3)

Hypothesise a Thomas cluster process

Fitting a Thomas model using the *kppm* in *spatstat* produced implausible estimates, viz. mean $\approx$ 5.2 people/cluster & SD $\approx$ 4.2 metres.

By inspection: 42 clusters; mean of 2.4 per cluster; 0.019 clusters per square metre; SD in x and $y \approx$ 0.7 metres.

We take this as our null hypothesis.

Simulate:

$0.019 \;\; clusters/m^2$

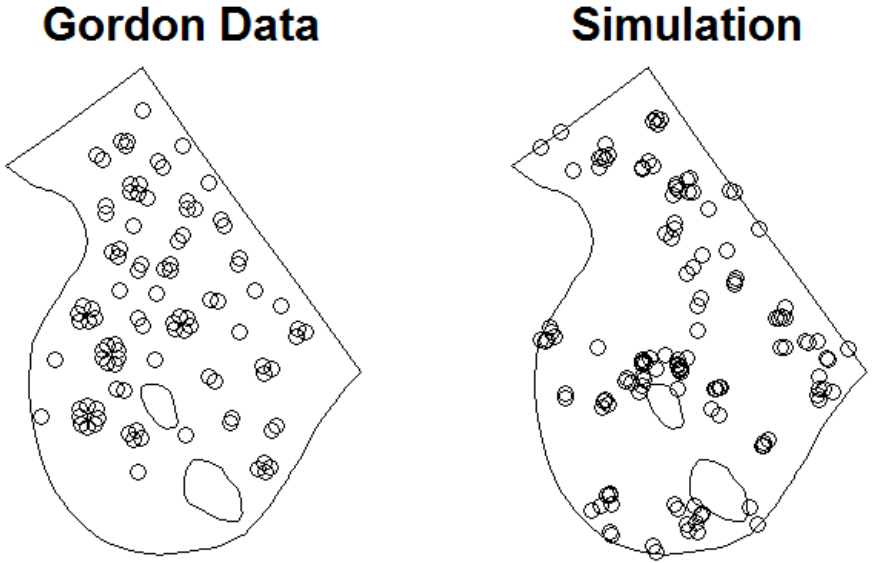
$\sigma(x) = \sigma(y) = 0.7$

$\mu = 2.4 \;\; people/cluster$

rThomas(kappa = 0.019, sigma = 0.7, mu = 2.4, win = gordon$win, nsim=1)

Gordon Square (3)

A simulation...



Gordon Square (4)

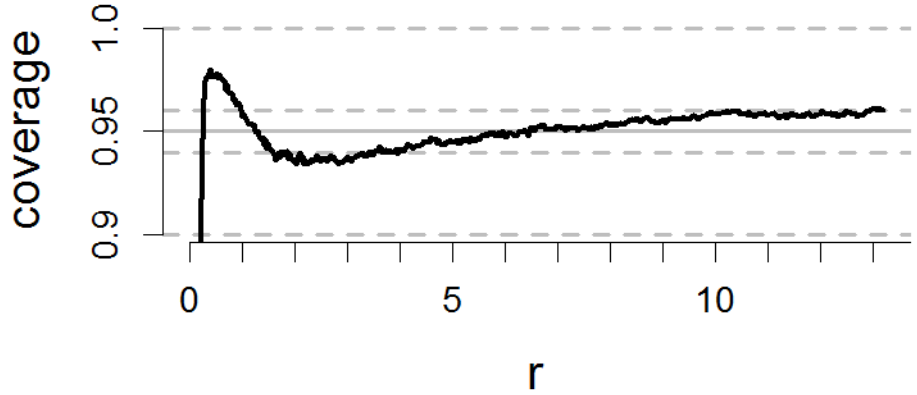
Tuning PoissonBootS: $PB_{20,15}$

1) Start with approximate PoissonBootS tuning parameters, eg. $PB_{40,0}$

2) Repeatedly simulate the Thomas process (the null hypothesis)

3) Calculate coverage

4) Adjust tuning parameters and repeat from (2)

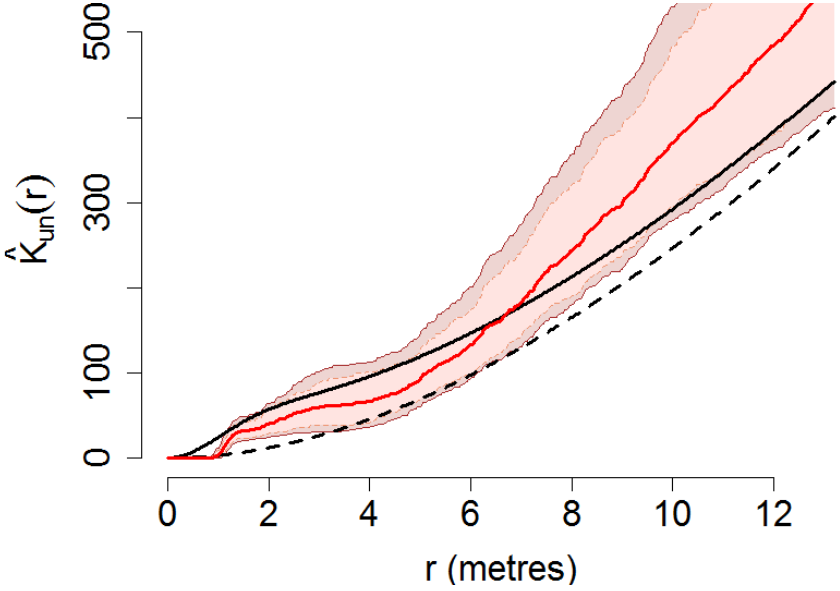


Note: More simulations are required as coverage approaches target (5000 here).

Gordon Square (5)

Uncorrected Ripley's K

Thomas and HPP means (5000 sims, black and dashed); Gordon square K function with local and global 95 % CIs ($PB_{20,15}$)



Conclusions

A novel bootstrap: PoissonBootS

- Mean number of marks drawn *may differ* from N.
- Number of marks drawn *may vary* between bootstrap samples.
- Parameters: mean draw fraction f_s and fraction of Poisson variance f_v.
- Must be tuned to the “null hypothesis” (currently tedious)
- Increases versatility of the bootstrap — e.g. bootstrapping functions.

Applied to estimating confidence intervals on Ripley’s K function.

- Relationship between variance of local K and Ripley’s K changes with r.
- Is 2nd-moment structure of the data consistent with a spatial model (H_0)?
- CI’s for Ripley’s K require an spatial model hypothesis, e.g. CSR, Thomas...

Thank You

CSIRO Digital Productivity Flagship

Bruce Tabor
t +61 2 9325 3192
e Bruce.Tabor@csiro.au
w CSIRO Digital Productivity Flagship web

CSIRO CSIRO Digital Productivity Flagship

Ross Sparks
t +61 2 9325 3262
e Ross.Sparks@csiro.au
w CSIRO Digital Productivity Flagship web