Applied Spatial Data Analysis

Chapter 2 - Second Order Statistics

Dr. Şebnem Er

Table of contents

Welcome

This lecture develops second-order methods for spatial point pattern analysis.

We focus on:

- Ripley's K function; and
- the pair correlation function.

Outline

4. First- and second-order properties
5. Morisita and Fry diagnostics
6. Ripley's K function
7. Pair correlation function

Presenter notes

Emphasise the progression from single-number summaries to scale-dependent functions. The latter retain more information about how the pattern changes with distance.

Recommended resources

- Baddeley, Rubak, and Turner, *Spatial Point Patterns: Methodology and Applications with R*.
- Cressie, *Statistics for Spatial Data*.
- Wiegand and Moloney, *Handbook of Spatial Point-Pattern Analysis in Ecology*.
- The official `spatstat` documentation and vignettes.

Presenter notes

The Baddeley, Rubak, and Turner text is the main reference for the notation and the `spatstat` implementation used here.

First- and second-order properties

First-order structure

The first-order property of a point process is its intensity:

$$\lambda(\mathbf{u}) = \lim_{|d\mathbf{u}| \to 0} \frac{\mathrm{E}[N(d\mathbf{u})]}{|d\mathbf{u}|}.$$

It describes how the expected event density varies over space.

Presenter notes

Before studying interaction, we need a credible model for intensity. Otherwise, broad spatial variation in intensity can be mistaken for attraction or clustering between events.

Second-order structure

Second-order properties describe association between pairs of events.

They address questions such as:

- Are events unusually close together?
- Are short interpoint distances suppressed?
- At which distance scales does dependence occur?

Presenter notes

Second-order summaries are related to pair counts. They describe association, not necessarily causation. Different mechanisms can produce similar second-order patterns.

Why intensity must be handled first

A pattern can look clustered because the intensity is high in one part of the region and low elsewhere, even when events are conditionally independent.

Therefore:

1. model or estimate first-order intensity;
2. examine residual second-order dependence;
3. avoid attributing inhomogeneity directly to point interaction.

Presenter notes

This is one of the most important conceptual points in the lecture. A homogeneous K function applied to a strongly inhomogeneous pattern can show apparent clustering that is entirely caused by the intensity trend.

Quadrat-based second-order diagnostics

Morisita index

Suppose there are n events altogether, and there are $n(n-1)$ number of ordered pairs of distinct events. Assume the events are divided among m quadrats with counts

$$n_1, n_2, \ldots, n_m.$$

The number of ordered pairs in quadrat j is

$$n_j(n_j - 1).$$

The total number of ordered pairs of distinct points which fall inside the same quadrat is thus $\sum_{j=1}^{m} n_j(n_j-1)$. The observed fraction of ordered pairs in the same quadrat is

$$\frac{\sum_{j=1}^{m} n_j(n_j-1)}{n(n-1)}.$$

This is the fraction of all pairs of data points in which both points fall in the same quadrat. In a completely random (homogeneous Poisson) process, where points are independent of each other, two points fall in the same quadrat with probability $1/m$, where m is the number of quadrats, so the fraction above is expected to equal $1/m$.

Presenter notes

Under equal-area quadrats and CSR, two independently selected events fall in the same quadrat with probability $1/m$.

Morisita index formula

The Morisita index is

$$I_M = m \frac{\sum_{j=1}^{m} n_j(n_j - 1)}{n(n-1)}.$$

Interpretation:

- $I_M = 1$: consistent with independence;
- $I_M > 1$: clustering;
- $I_M < 1$: regularity.

Figure 7.1 shows three archetypal point patterns representing 'regularity' (where points tend to avoid each other), 'independence' (complete spatial randomness), and 'clustering' (where points tend to be close together).

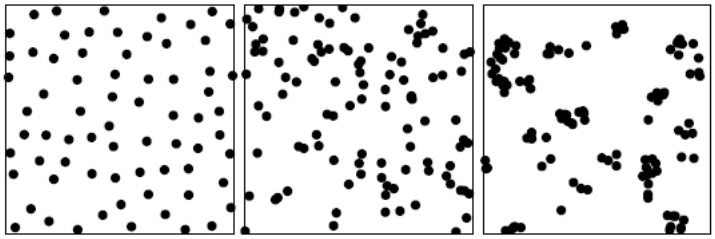


Figure 7.1. *Classical trichotomy between regular (*Left*), independent (*Middle*), and clustered (*Right*) point patterns. All three patterns are in the unit square.*

Ref: - Baddeley, Rubak, and Turner, *Spatial Point Patterns: Methodology and Applications with R*. Page:201

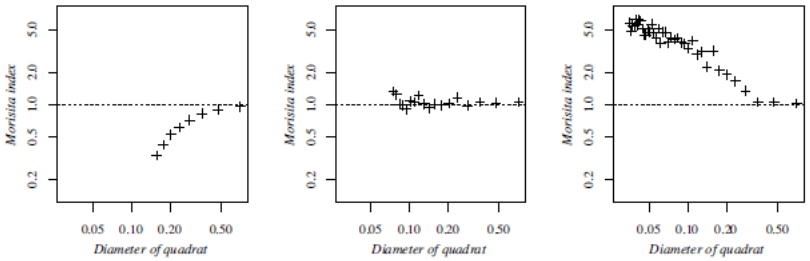


Figure 7.2. *Morisita index plots (on logarithmic scale) for the three patterns in Figure 7.1.*

Ref: - Baddeley, Rubak, and Turner, *Spatial Point Patterns: Methodology and Applications with R*. Page:201

Presenter notes

Like ordinary quadrat counting, the Morisita index depends on quadrat size and placement. Repeating it at several resolutions gives a scale-dependent plot but does not remove the modifiable areal unit problem.

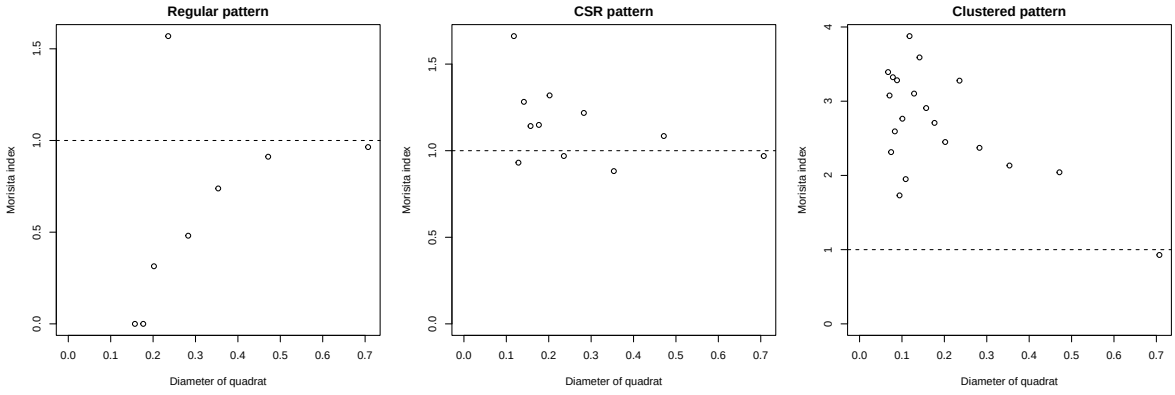
Morisita index in R

```
par(
  mfrow = c(1, 3),
  mar = c(4.5, 4.5, 2, 1)
)

miplot(
  pp_regular,
  main = "Regular pattern",
  xlab = "Diameter of quadrat",
  ylab = "Morisita index"
)

miplot(
  pp_csr,
  main = "CSR pattern",
  xlab = "Diameter of quadrat",
  ylab = "Morisita index"
)

miplot(
  pp_cluster,
  main = "Clustered pattern",
  xlab = "Diameter of quadrat",
  ylab = "Morisita index"
)
```



```r
par(mfrow = c(1, 1))
```

Presenter notes

The horizontal axis is an approximate scale measure. The curve can change substantially with resolution, illustrating the sensitivity of quadrat methods to the chosen spatial partition.

Fry plot

Definition of a Fry plot

Fry plot is a scatterplot of the vector differences $x_i - x_j$ between all pairs of distinct points in the pattern

$$\mathbf{x}_j - \mathbf{x}_i,$$

for every ordered pair of distinct events.

The origin represents a typical event, and the plotted differences show where other events occur relative to it.

Presenter notes

A central hole suggests inhibition at short distances. Directional bands or elongated structure can indicate anisotropy. A dense central concentration is consistent with clustering.

Fry plot in R

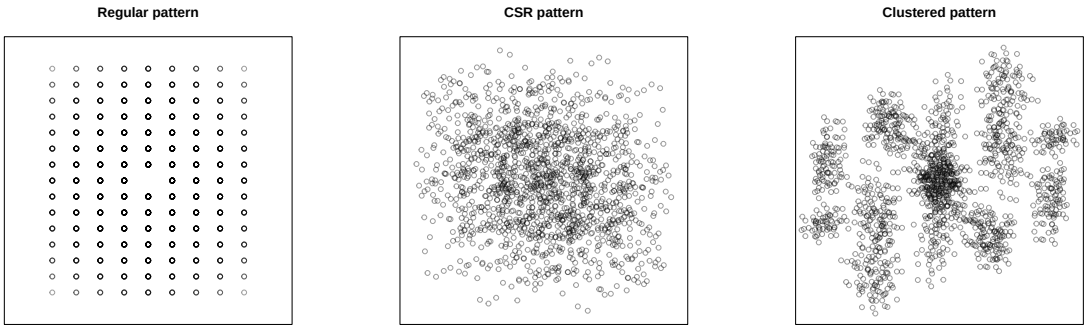
```r
par(
  mfrow = c(1, 3),
  mar = c(4.5, 4.5, 2, 1)
)

fryplot(
  pp_regular,
  main = "Regular pattern"
)

fryplot(
  pp_csr,
  main = "CSR pattern"
```

```
)

fryplot(
  pp_cluster,
  main = "Clustered pattern"
)
```




```
par(mfrow = c(1, 1))
```

Presenter notes

The plot is centrally symmetric because every vector difference is accompanied by its negative. This diagnostic is exploratory and does not replace a formal test.

A Fry plot can be drawn by hand, as follows (see Figure 7.3 in Baddeley book). First print the point pattern on a sheet of paper. Take a transparency or sheet of tracing paper, and mark a cross in the middle. Place the transparency over the printout, so that the cross on the transparency lies on one of the data points. Copy the positions of all the other data points onto the transparency in the same relative position. (Of course some data points may be too far away to be copied onto the transparency, depending on its size.) Now move the cross to another data point, and repeat the copying process. After every data point has been visited, the pattern on the transparency is the Fry plot.

The origin in the Fry plot represents a typical point of the point pattern, and the dots in the Fry plot represent the positions of other nearby points, relative to the typical point.

Ripley's K function

Definition

For a stationary point process with intensity λ, Ripley's K function is defined by

$$K(r) = \frac{1}{\lambda} \mathrm{E}\left[N_0(r)\right].$$

where where $E[.]$ denotes the expectation and $N_0(r)$ represents the number of further events up to a distance r around an arbitrary event.

Presenter notes

The phrase "further events" excludes the event at the centre. The function accumulates neighbours from distance zero up to distance r.

K function under CSR

Under a homogeneous Poisson process, the expected number of further events in a disc of radius r is

$$\lambda \pi r^2.$$

Therefore,

$$K_{\mathrm{CSR}}(r) = \pi r^2.$$

Presenter notes

This benchmark does not depend on intensity because the expected count is divided by λ. That makes the K function useful for comparing patterns with different intensities, provided stationarity is reasonable.

Empirical K function

A generic edge-corrected estimator has the form

$$\widehat{K}(r) = \frac{|W|}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \frac{\mathbf{1}(d_{ij} \leq r)}{e_{ij}},$$

where e_{ij} is an edge-correction weight.

Presenter notes

The indicator counts pairs no farther apart than r. The correction weight compensates for neighbours that could lie outside the observation window and therefore cannot be observed.

Edge effects in K

Events near the boundary have part of their radius-r neighbourhood outside the observation window.

Without correction, their neighbour counts are systematically too small.

Common corrections include:

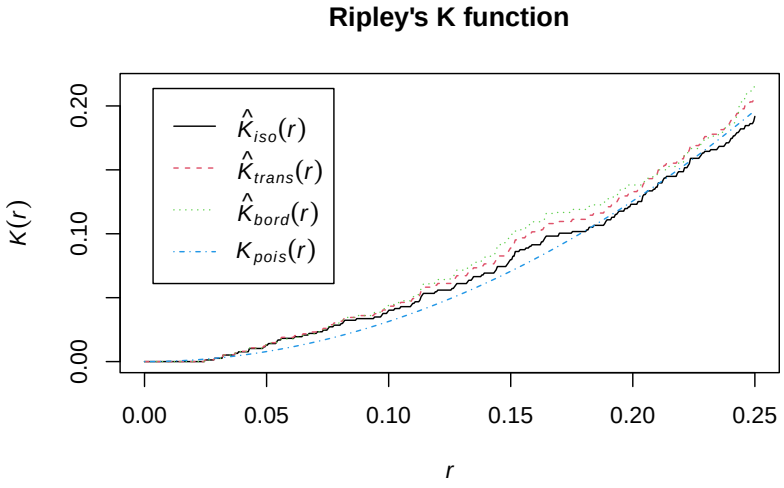
- border correction;
- translation correction;
- isotropic correction.

Presenter notes

Each correction uses a different geometric argument. This lecture does not derive them, but students should understand why an uncorrected estimate is biased downward near the boundary.

K function in R

```
K_csr <- Kest(
  pp_csr,
  correction = c("border", "isotropic", "translate")
)
plot(K_csr, main = "Ripley's K function")
```



Presenter notes

The theoretical CSR curve is labelled `theo`. The corrected empirical estimates are shown separately. Differences among corrections are often largest at larger distances.

Interpreting K

Relative to the CSR benchmark:

$$\widehat{K}(r) > \pi r^2$$

suggests more neighbours than expected and is consistent with clustering.

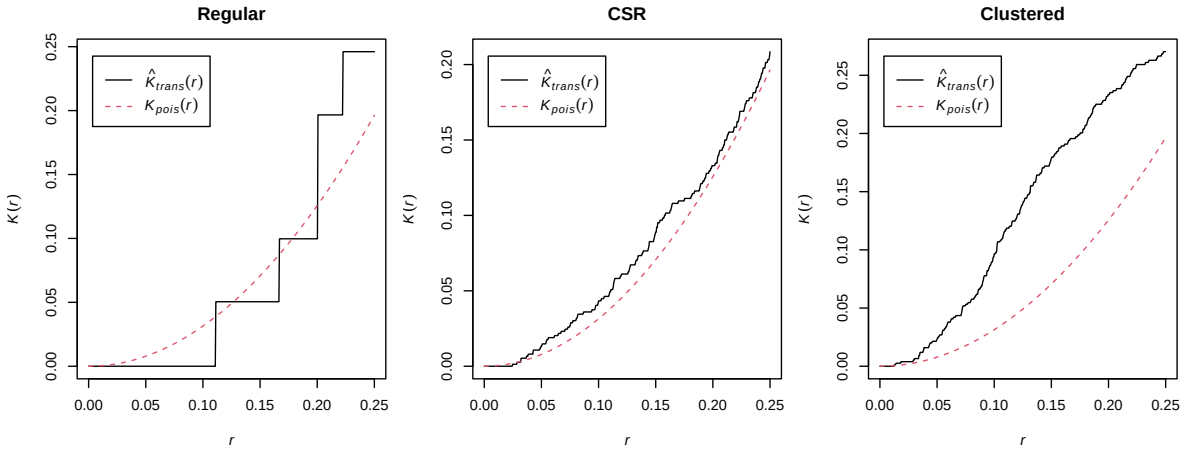
$$\widehat{K}(r) < \pi r^2$$

suggests fewer neighbours than expected and is consistent with inhibition or regularity.

Presenter notes

Interpretation is scale-specific. A pattern may be regular at short distances and clustered at larger distances. Also remember that inhomogeneity can inflate the homogeneous K function.

Comparing K across patterns



Presenter notes

The regular pattern should tend below the theoretical curve at short distances, while the clustered pattern should tend above it. Random simulation means individual plots will not be perfectly idealised.

Important cautions

Stationarity and inhomogeneity

The standard homogeneous versions of G, F, J, and K assume stationarity or homogeneity.

A spatial trend in intensity can produce apparent clustering even when events are independent conditional on the trend.

For inhomogeneous patterns, consider intensity-adjusted methods such as

$$K_{\text{inhom}}(r)$$

and suitable residual diagnostics.

Presenter notes

Do not apply homogeneous second-order methods mechanically. First inspect and model the intensity surface using covariates or non-parametric smoothing, then assess remaining dependence.

Edge correction

Edge correction matters because events or reference locations near the boundary have unobservable neighbourhoods outside the window.

The impact grows as the distance r increases.

Good practice:

- report the correction used;
- avoid over-interpreting very large distances;
- compare corrections when results are sensitive;
- use simulation under the same observation window.

Presenter notes

The observation window is part of the data-generating and observation mechanism. Simulations should use exactly the same window geometry so that edge effects are comparable.

Correlation is not causation

Second-order summaries identify spatial association, not its cause.

Similar clustering can result from:

- environmental heterogeneity;
- social or biological interaction;
- dispersal mechanisms;
- unmeasured covariates;
- observation or sampling processes.

Presenter notes

A statistical summary can reject a simple null model without identifying the true mechanism. Interpretation must be guided by subject-matter knowledge and alternative models.

Summary

Main ideas

- Nearest-neighbour distances describe spacing between events.
- Empty-space distances describe gaps in the pattern.
- The G, F, and J functions retain information across distances.
- Ripley's K function measures cumulative second-order structure.

- The pair correlation function isolates structure near a particular distance.
- Intensity and edge effects must be handled before interpreting dependence.

Presenter notes

Conclude by linking this lecture to formal point-process modelling. The next step is to specify models for intensity and interaction and to assess them using likelihood, residuals, and simulation-based diagnostics.

Reproducibility information

```
sessionInfo()
```

```
R version 4.4.1 (2024-06-14 ucrt)
Platform: x86_64-w64-mingw32/x64
Running under: Windows 10 x64 (build 19043)

Matrix products: default


locale:
[1] LC_COLLATE=English_United Kingdom.utf8
[2] LC_CTYPE=English_United Kingdom.utf8
[3] LC_MONETARY=English_United Kingdom.utf8
[4] LC_NUMERIC=C
[5] LC_TIME=English_United Kingdom.utf8

time zone: Africa/Johannesburg
tzcode source: internal

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] spatstat.model_3.3-6   rpart_4.1.23           spatstat.explore_3.4-3
[4] nlme_3.1-164           spatstat.random_3.4-1  spatstat.geom_3.4-1
[7] spatstat.univar_3.1-3  spatstat.data_3.1-6

loaded via a namespace (and not attached):
 [1] cli_3.6.4             knitr_1.51            rlang_1.1.5
 [4] spatstat.sparse_3.1-0 xfun_0.55             spatstat.utils_3.1-4
```

```
 [7] jsonlite_2.0.0       htmltools_0.5.8.1    tinytex_0.57
[10] rmarkdown_2.29       grid_4.4.1           evaluate_1.0.5
[13] abind_1.4-8          fastmap_1.2.0        yaml_2.3.10
[16] compiler_4.4.1       goftest_1.2-3        polyclip_1.10-7
[19] mgcv_1.9-1           rstudioapi_0.18.0    lattice_0.22-6
[22] digest_0.6.37        tensor_1.5           splines_4.4.1
[25] Matrix_1.7-0         tools_4.4.1          deldir_2.0-4
```

Presenter notes

The session information records package and R versions, helping students reproduce the results and diagnose version-related differences.