Spatial point processes in the modern world – an interdisciplinary dialogue

Janine Illian

University of St Andrews, UK and NTNU Trondheim, Norway

Bristol, October 2015

statistical software – past to present

in the past:

- standard statistical software commercial
- complex statistical methods not implemented

⇒ not accessible to non-specialists

⇒ use limited to specialists

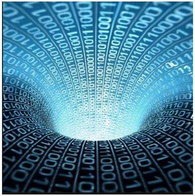
today:

- standard software, such as R, is free
- commonly, “new” methods published as R libraries

⇒ easily accessible to non-specialists

⇒ increasingly used by non-specialists – specialists not involved!

BIG DATA...



at the same time

- increasing amounts of data collected
- open access policies

⇒ increasing need for (complex) analysis

⇒ increasing application of non-statistical methods...

BIG DATA...

so is this a problem? misuse?

- might argue that complex methods should only be used by specialists...

BIG DATA...

so is this a problem? misuse?

- might argue that complex methods should only be used by specialists...
- but: isn't this beside the point?
- it is already happening...
- there are not enough specialists anyway...

⇒ change attitude – make methods usable and communicate

usability

issues:

- methods often developed from the point of view of a statistician not the user
- little thought gone into applicability *in practice*
- producing computationally efficient methods is a first step

usability

issues:

- methods often developed from the point of view of a statistician not the user
- little thought gone into applicability *in practice*
- producing computationally efficient methods is a first step

this talk...

- illustrate that it is possible to improve
  - the *relevance* and
  - the *practicability* of statistical methodology through interdisciplinary dialogue

⇒ spatial point processes in ecology

spatial point processes in ecology

ecology – main interest:

- interactions among individual organisms and their environments
- individuals exist in space and time
- spatially explicit data increasingly available

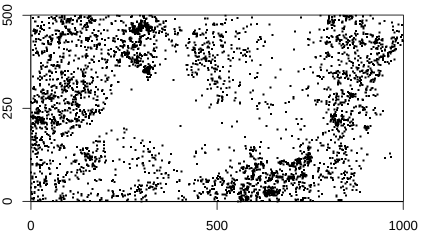
spatial point processes in ecology

ecology – main interest:

- interactions among individual organisms and their environments
- individuals exist in space and time
- spatially explicit data increasingly available

⇒ data: spatial (spatio-temporal) point patterns

⇒ **spatial point process methodology** should be highly relevant!



case study – spatial point processes in ecology

however...

- few ecologists aware of spatial point process methodology
- e.g. models rarely used in practice
- even though very convenient libraries such as `spatstat` available

⇒ not part of the standard statistical toolbox

case study – spatial point processes in ecology

however...

- few ecologists aware of spatial point process methodology
- e.g. models rarely used in practice
- even though very convenient libraries such as `spatstat` available

⇒ not part of the standard statistical toolbox

discuss

- ways of improving relevance of methodology – biodiversity, joint models
- challenges regarding practicality of modelling – prior choice, role of random field

biodiversity – what is it all about?

Convention on Biological Diversity – 2010 target

2010 International Year of Biodiversity

"to achieve . . . a **significant reduction** *of the current* **rate of biodiversity loss** *at a* **global, regional and national level** *. . . ."*

- ecosystem services:
  - species add to the genetic pool...
  - fewer species – problem for the system?

take the rainforests...

- $>$ 600 tree species in a single ha – 16 x as many as in the UK!
- processes maintaining high diversity poorly understood
- is it actually important?

biodiversity – a spatial problem

how is biodiversity is maintained – a debate:

- opposing biodiversity theories
- previously: assessment of theories non-spatial

biodiversity – a spatial problem

how is biodiversity is maintained – a debate:

- opposing biodiversity theories
- previously: assessment of theories non-spatial

- coexistence of species linked to interaction of individuals with local environment

⇒ biodiversity is a spatial problem!

biodiversity – a spatial problem

how is biodiversity is maintained – a debate:

- opposing biodiversity theories
- previously: assessment of theories non-spatial

- coexistence of species linked to interaction of individuals with local environment

⇒ biodiversity is a spatial problem!

simulated data – mimicking mechanisms postulated by theories:

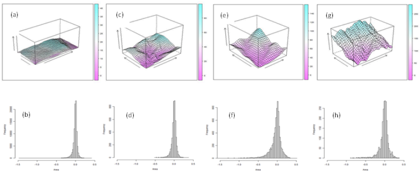
- first order summary characteristics cannot distinguish theories
- marked differences when second order characteristics are used

Brown et al. 2011

biodiversity – a spatial problem

spatial analysis of 14 different large rainforest plots shows:

- pairwise spatial interaction among species more variable in heterogeneous sites
- theories that claim diversity is **not** important **cannot** explain this relationship ⇒ diversity matters

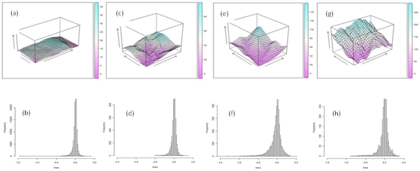


Brown et al. 2013

biodiversity – a spatial problem

spatial analysis of 14 different large rainforest plots shows:

- pairwise spatial interaction among species more variable in heterogeneous sites
- theories that claim diversity is **not** important **cannot** explain this relationship ⇒ diversity matters



Brown et al. 2013

- **interdisciplinary dialogue**: increased awareness of methodology
- spatial structure relevant for understanding coexistence

spatial modelling – Cox processes

Cox processes are spatial point processes with a **random** intensity function.

Log Gaussian Cox processes depend on a (continuous) random field

$$\Lambda(s) = \exp\{Z(s)\},$$

where $\{Z(s) : s \in \mathbb{R}^2\}$ is a Gaussian random field.

- very flexible class of models

spatial modelling – Cox processes

Cox processes are spatial point processes with a **random** intensity function.

Log Gaussian Cox processes depend on a (continuous) random field

$$\Lambda(s) = \exp\{Z(s)\},$$

where $\{Z(s) : s \in \mathbb{R}^2\}$ is a Gaussian random field.

- very flexible class of models
- likelihood is analytically intractable
- LG Cox processes are latent Gaussian models
- can fit them with integrated nested Laplace approximation (INLA) $\Rightarrow$ speeds up and facilitates model fitting

$\Rightarrow$ develop complex models that are relevant in practice

Rue et al. 2009, Illian et al. 2012

INLA in a nutshell

Three main ingredients in INLA

- Gaussian Markov random fields
- Latent Gaussian models
- Laplace approximations

which together (with a few other things) give a very nice tool for Bayesian inference

- quick
- accurate

may be used to fit a large class of models: **latent Gaussian models**

general framework - three-stage hierarchical model

- **observations** $(\boldsymbol{y})$:
  Assumed conditionally independent given $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$

- **latent field** $(\boldsymbol{\eta})$:
  Assumed to be a Gaussian Markov Random Field (details later), sparse precision matrix $\boldsymbol{Q}(\boldsymbol{\theta})$

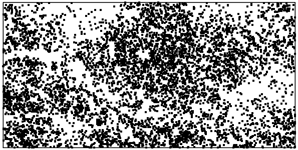
- **hyperparameters** $(\boldsymbol{\theta})$:
  Precision parameters of the Gaussian priors assigned to latent field

general framework - three-stage hierarchical model

- **observations** $(\boldsymbol{y})$:
  Assumed conditionally independent given $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$
  $$\boldsymbol{y} \mid \boldsymbol{\eta}, \boldsymbol{\theta} \sim \prod_i \pi(y_i \mid \eta_i, \boldsymbol{\theta}).$$
- **latent field** $(\boldsymbol{\eta})$:
  Assumed to be a Gaussian Markov Random Field (details later), sparse precision matrix $\boldsymbol{Q}(\boldsymbol{\theta})$
  $$\boldsymbol{\eta} \mid \boldsymbol{\theta} \sim \pi(\boldsymbol{\eta} \mid \boldsymbol{\theta}) = \mathcal{N}(\cdot, \boldsymbol{Q}^{-1}(\boldsymbol{\theta})).$$
- **hyperparameters** $(\boldsymbol{\theta})$:
  Precision parameters of the Gaussian priors assigned to latent field
  $$\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}).$$

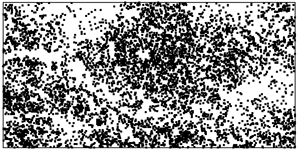
example – spatial pattern relative to covariates

- 50 ha forest dynamics plot at Pasoh Forest Reserve (PFR), Peninsular Malaysia; never logged
- e.g. species *Aporusa microstachya*; 7416 individuals
- tree occurrence relative to environmental covariates



example – spatial pattern relative to covariates

- 50 ha forest dynamics plot at Pasoh Forest Reserve (PFR), Peninsular Malaysia; never logged
- e.g. species *Aporusa microstachya*; 7416 individuals
- tree occurrence relative to environmental covariates

- approximate continuous space by regular grid
- use Gauss Markov Random Field (GMRF) to approximate Gaussian Field

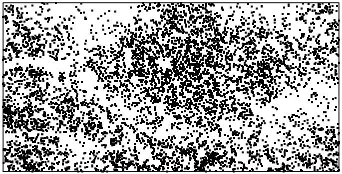
latent field:

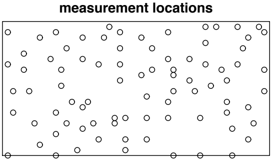
$$\eta_{ij} = \alpha_1 + \sum_{p \in \mathcal{I}} \beta_p z_p(s_{ij}) + \beta_s f_s(s_{ij})$$

relevance – realistically complex

we can make models realistically complex, e.g. **joint models** of:

- two (or more) spatial patterns
- replicated point patterns
- point pattern and multiple, dependent marks
- covariates and the pattern

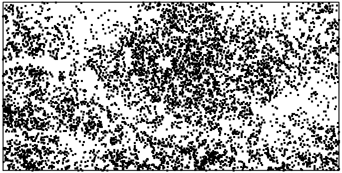


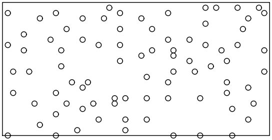


relevance – realistically complex

we can make models realistically complex, e.g. **joint models** of:

- two (or more) spatial patterns
- replicated point patterns
- point pattern and multiple, dependent marks
- covariates and the pattern





⇒ INLA with multiple likelihoods

Illian et al. 2012 a + b, Rajala et al. 2013

flexible models

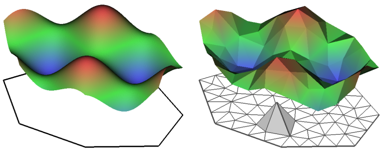
- simple models use a simple gridding approach to approximate the continuous spatial field
  - **computationally inefficient** and
  - not flexible enough (complicated boundaries or domains)

flexible models

- simple models use a simple gridding approach to approximate the continuous spatial field
  - **computationally inefficient** and
  - not flexible enough (complicated boundaries or domains)

⇒ continuously specified, finite dimensional Gaussian random field

⇒ spatial field as solution to a stochastic partial differential equation ("SPDE approach")



- no "binning" of the points – **exact** positions

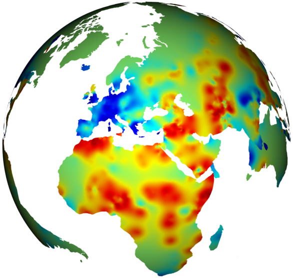
Lindgren et al. 2011, Simpson et al. 2012

flexibility...

model fitting is still fast – and models are very flexible!

we can fit models

- with spatio-temporal random field
- on complicated domains (holes)
- on the sphere
- ...



in practice...

flexible, realistically complex spatial point process models:

we now have all these models; they are flexible, complex and yet can be fitted in a computationally efficient way...
implemented in `R-INLA`

in practice...

flexible, realistically complex spatial point process models:

we now have all these models; they are flexible, complex and yet can be fitted in a computationally efficient way...
implemented in `R-INLA`

So...

- Is this where our role as statisticians ends?

in practice...

flexible, realistically complex spatial point process models:

we now have all these models; they are flexible, complex and yet can be fitted in a computationally efficient way...
implemented in `R-INLA`

So...

- Is this where our role as statisticians ends?
- How can the models be used? – first versus second order processes, prior choice
- How can we communicate them? – interactive visualisation of results

first order versus second order processes

- covariate association models are “first order models” (modelling the intensity)
- random field accounts for spatial structure unaccounted for by covariates

first order versus second order processes

- covariate association models are "first order models" (modelling the intensity)
- random field accounts for spatial structure unaccounted for by covariates

- Is this first order structure (abiotic interactions) or second order structure (biotic interactions)?

first order versus second order processes

- covariate association models are "first order models" (modelling the intensity)
- random field accounts for spatial structure unaccounted for by covariates

- Is this first order structure (abiotic interactions) or second order structure (biotic interactions)?

approach

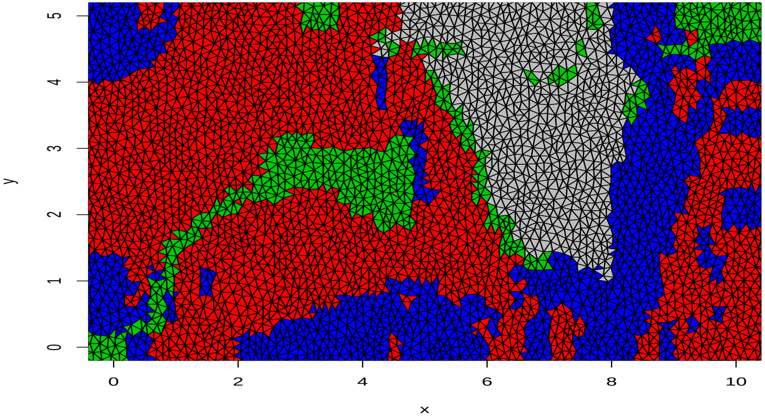
$\Rightarrow$ use SPDE formulation to derive (second order) non-stationary models

- variable range of spatial correlation ("difficulty" = inverse range)

non-stationary models – rainforest species in BCI, Panama

Faramea occidentalis

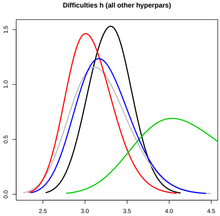
different habitats:



non-stationary models – rainforest species in BCI, Panama

Faramea occidentalis

resulting difficulty (inverse range) h_i:



credibility intervals:

$$h_3/h_1 \in [1.02, 1.82], h_3/h_2 \in [1.06, 1.85], h_3/h_4 \in [1.01, 1.78]$$

Bakka et al. in preparation

in practice – interpretation

- second order structure varies with habitat
- conservative prior on non-stationarity
- first and second order cannot be distinguished without replicates; so focus on second order here

in practice – interpretation

- second order structure varies with habitat
- conservative prior on non-stationarity
- first and second order cannot be distinguished without replicates; so focus on second order here

practical relevance

- communication of difference between first and second structures and (biological) interpretation
- **random fields reflect biological processes**, not just nuisance

prior choice...

significance of covariates depends on **smoothness** of iGMRF

- if too low (= spatial field too wiggly) – overfitting
  ⇒ no covariates significant
- if too high (= spatial field too smooth) – autocorrelation not accounted for
  ⇒ too many covariates significant
- **prior choice:** precision of iGMRF

not an easy question... How do we choose this prior?

⇒ use a principled, practical approach to constructing priors – *pc-priors*

- model component is flexible version of a **base model**
- shrink towards the base model

prior choice in practice...

spatial modelling:
Cox process model and iGMRF (2-dimensional random walk, `rw2d`): deviation from a plane (flat prior)

- scale the magnitude of the model component

⇒ determines how informative a prior is

prior choice in practice...

spatial modelling:

Cox process model and iGMRF (2-dimensional random walk, `rw2d`): deviation from a plane (flat prior)

- scale the magnitude of the model component

⇒ determines how informative a prior is

reparameterise the standard `rw2d` model to include the iid error:

$$\frac{1}{\tau}\left(\sqrt{1-\phi}v_i + \sqrt{\phi}u_i\right)$$

- scaled spatially structured (v_i) and unstructured (u_i) effects
- priors on precision τ and mixing parameter ϕ

in practice – interpretation

- ϕ identifies the fraction of variance explained by the random field
- τ and ϕ have orthogonal interpretation – assign independent (hyper)priors to these hyperparameters

$\Rightarrow$ prior choice conservative and transparent

in practice – interpretation

- ϕ identifies the fraction of variance explained by the random field
- τ and ϕ have orthogonal interpretation – assign independent (hyper)priors to these hyperparameters

$\Rightarrow$ prior choice conservative and transparent

practical relevance

- communication – importance and difficulty of prior choice
- **random fields reflect biological processes**, not just nuisance
- what is is role of the random field?

$\Rightarrow$ what do we want to penalise?

Sørbye, Illian and Simpson in preparation

in practice – communication

communicate

- difference between first and second order structures and relevant random fields
- importance of prior choice and **role** of random field

in practice – communication

communicate

- difference between first and second order structures and relevant random fields
- importance of prior choice and **role** of random field

random field?

in practice – communication

communicate

- difference between first and second order structures and relevant random fields
- importance of prior choice and **role** of random field

random field?

practical relevance

- **random fields reflect (biological) processes**
- how do we explain what a random field is?
- how do we present the model results – i.e. the estimated random field?

crime modelling

spatio-temporal models of crime occurrence

crime modelling

spatio-temporal models of crime occurrence

project involves communication with the police and the Scottish Government – non-specialists...

$\Rightarrow$ communication essential

crime modelling

spatio-temporal models of crime occurrence

project involves communication with the police and the Scottish Government – non-specialists...

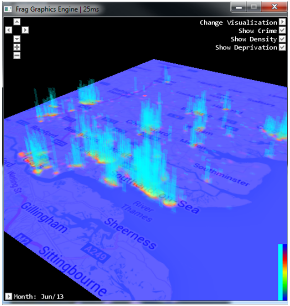
$\Rightarrow$ communication essential

- What can **we** model?
- What are **they** interested in?

EPSRC Impact Acceleration Account (IAA)

crime modelling

computer games technology



- visualisation of results (estimated random field)
- communication – interpret remaining spatial structure

challenge: difference between visualising and modelling

joint work with J. Bown, Abertay University and N. Fyfe, Scottish Institute of Policing Research

interdisciplinarity ...

symbiosis

interaction among two or more "species" with **mutual benefit**

applied statisticians' role:

- not "simply" analysing a data set or using a data set as an illustration of a new method

interdisciplinarity ...

symbiosis

interaction among two or more "species" with **mutual benefit**

applied statisticians' role:

- not "simply" analysing a data set or using a data set as an illustration of a new method

but:

- obtaining inspiration from applied problems
- contributing to the scientific discourse

interdisciplinarity ...

interaction requires **functioning communication**:

interdisciplinarity ...

interaction requires **functioning communication**:

⇒ aiding communication by participating in scientific dialogue **on both sides**

interdisciplinarity ...

interaction requires **functioning communication**:

⇒ aiding communication by participating in scientific dialogue **on both sides**

- developing methods that are
  - statistically interesting and
  - practically relevant and meaningful
- this goes beyond providing software...
- particularly relevant for spatial point processes

Illian and Burslem, in preparation