Combining spatial data derived from conventional research protocols and social media platforms

A story of two dolphin species

Sara Martino¹

Department of Mathematical Science (NTNU)

¹Giovanna Jona Lasinio, Daniela Silvia Pace, et al
Introduction

Statistical Tools

Modeling the intensity

Results

Summary and conclusions
Introduction
Introduction I

▶ **Goal:** Understand the spatial distribution of wild species

How: Traditional data sources

→ go out and search for dolphins!!

The observation process introduces a bias...

We know the searching protocol...

..we can correct for such bias

There are more data available....could we use them?
Introduction I

- **Goal**: Understand the spatial distribution of wild species
- **How**: Traditional data sources → go out and search for dolphins!!
Introduction I

- **Goal:** Understand the spatial distribution of wild species
- **How:** Traditional data sources → go out and search for dolphins!!

- The observation process introduces a bias...
Introduction I

- **Goal:** Understand the spatial distribution of wild species
- **How:** Traditional data sources $\rightarrow$ go out and search for dolphins!!

- The observation process introduces a bias...
  - We know the searching protocol...
Introduction I

- **Goal:** Understand the spatial distribution of wild species
- **How:** Traditional data sources → go out and search for dolphins!!
- The observation process introduces a bias...
  - We know the searching protocol...
  - ..we can correct for such bias
Introduction I

- **Goal:** Understand the spatial distribution of wild species
- **How:** Traditional data sources → go out and search for dolphins!!
- The observation process introduces a bias...
  - We know the searching protocol...
  - ..we can correct for such bias
- There are more data available....could we use them?
Social Data

Many people are out in the sea with leisure boats
Many people are out in the sea with leisure boats.
People like to take pictures of dolphins if they spot one...such pictures are often put on social media...
Social Data

- Many people are out in the sea with leisure boats
- People like to take pictures of dolphins if they spot one...such pictures are often put on social media...
- This can be a valuable source of data
Social Data

- Many people are out in the sea with leisure boats
- People like to take pictures of dolphins if they spot one...such pictures are often put on social media...
- This can be a valuable source of data

...but there is no “searching protocol”
Social Data

- Many people are out in the sea with leisure boats
- People like to take pictures of dolphins if they spot one...such pictures are often put on social media...
- This can be a valuable source of data

...but there is no “searching protocol”

How can we correct for the bias?
- All our data are presence-only
- We want to merge all data sources...
- ...accounting for each specific bias!
## The Data

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Years</th>
<th>N.Campaigns</th>
<th>N.Sightings Stenella</th>
<th>N.Sightings Tursiope</th>
</tr>
</thead>
<tbody>
<tr>
<td>FERRY</td>
<td>2007-2018</td>
<td>311</td>
<td>133</td>
<td>16</td>
</tr>
<tr>
<td>UNIRM</td>
<td>2017-2019</td>
<td>73</td>
<td>14</td>
<td>98</td>
</tr>
<tr>
<td>Social</td>
<td>2008-2019</td>
<td>??</td>
<td>136</td>
<td>465</td>
</tr>
</tbody>
</table>

Notes:

- We have many “Social media” data
- We have both a “Spatial” and a “Temporal” bias!!
Observations
Statistical Tools
Statistical tools

- Log Gaussian Cox Processes (presence only data)
  - SPDE representation of Gaussian fields
  - Inference using INLA

- Thinned point process (observation bias)
  - Detection function
  - Needs more than just INLA \(\rightarrow\) inlabru

- Joint modeling (merging of all data sources)
  - Easy with INLA+inlabru
Log Gaussian Cox Processes

- We observe $N$ points in the domain $\Omega$.
- Given the intensity $\lambda(s)$ the likelihood is given by
  \[
  \pi(Y|\lambda) = \exp\left\{ |\Omega| - \int_{\Omega} \lambda(s)ds \right\} \prod_{i=1}^{N} \lambda(s_i)
  \]
- The log-intensity is a Gaussian process
  \[
  \log(\lambda(s)) = Z(s)
  \]
- Not analytically tractable
Discretize the domain into a grid

\[ N_{ij} = \# \text{ of observation in cell } (i, j) \]

\[ N_{ij} \sim \text{Pois}(\Lambda_{ij}) \text{ where} \]

\[ \Lambda_{ij} = \int_{s_{ij}} \lambda(s) ds \approx |s_{ij}| \exp(z_{ij}) \]

Possible models for \( Z(s) \):

- Continuous GF \( \rightarrow \) dense covariance matrix
- GMRF \( \rightarrow \) sparse covariance matrix

The grid serves both to approximate the latent field and to approximate the likelihood
Use the SPDE approach over a mesh to represent the GF

\[ Z(s) = \sum_{i=1}^{n} z_i \phi_i(s) \]

Approximate the GF
Do not need to approximate the observation location
Efficient computationally
Use INLA

\(^2\text{in Going off grid: Computationally efficient inference for log-Gaussian Cox Processes, Simpson et al 2011}\)
Thinned point process

\[ \lambda(s) \]

\[ g(s) \]

\[ \lambda(s)g(s) \]
Thinned point process

- “True” intensity: $\lambda(s)$
- Thinned intensity $\lambda^*(s) = \lambda(s)g(s)$
  - $g(s)$ is the thinning (detection) function
  - Unless $g(s)$ is log-linear in all parameters the INLA framework does not work!
  - `inlabru` is an extension of INLA that allows for non-linear terms
Modeling the intensity
Modeling the intensity

▶ The “true” (unthinned) intensity:

\[ \lambda(s, t) = \beta_0 + \beta^T X(s, t) + \sum_k f_k(x_k(s, t)) + u(s) \]

▶ \( u(s) \) is a GRF with Matern correlation function
▶ could be spatio-temporal but would need more data!

▶ The observed intensity:

\[ \lambda_j(s, t) = t_j \lambda(s, t) g_j(s); \quad j = 1, \ldots, 4 \]

where

▶ \( \lambda(s, t) \) is the true density
▶ \( g_j(s) \) is the thinning function for observation process \( j \)
▶ \( t_j \) is the time-scaling factor (this is known for all observations processes except for the social data!)
Accounting for spatial bias: detection functions

▶ For FERRY data

\[ g_{ferry}(s) = \exp \left( - \frac{1}{\sigma_{ferry}^2} d(s)^2 \right) \]

where \( d(s) \) is the perpendicular distance to the transect

▶ How about the SOCIAL MEDIA data?
Accounting for spatial bias: detection functions

- For FERRY data

\[ g_{\text{ferry}}(s) = \exp \left( -\frac{1}{\sigma^2_{\text{ferry}}} d(s)^2 \right) \]

where \( d(s) \) is the perpendicular distance to the transect.

- For UNIRM data

\[ g_{\text{unirm}}(s) = \begin{cases} 
1 & \text{for } d(s) < K \\
0 & \text{for } d(s) > K
\end{cases} \]
Accounting for spatial bias: detection functions

- For FERRY data

\[ g_{\text{ferry}}(s) = \exp \left( -\frac{1}{\sigma_{\text{ferry}}^2} d(s)^2 \right) \]

where \( d(s) \) is the perpendicular distance to the transect.

- For UNIRM data

\[
 g_{\text{unirm}}(s) = \begin{cases} 
 1 & \text{for } d(s) < K \\
 0 & \text{for } d(s) > K 
\end{cases}
\]

- How about the SOCIAL MEDIA data?
Modeling spatial bias for social data

- We assume that the sightings are biased towards area where there are more leisure boats.

- Distance from the coastline
- Boat density data from EmodNET platform
- Use animal intensity as proxy for small boat intensity
Modeling spatial bias for social data

- We assume that the sightings are biased towards areas where there are more leisure boats.
- but we do not have data about that...

- Distance from the coastline
- Boat density data from EmodNET platform
- Use animal intensity as proxy for small boat intensity
Modeling spatial bias for social data

- We assume that the sightings are biased towards areas where there are more leisure boats. . .
- but we do not have data about that. . .
- Three different ideas:
  - Distance from the coastline
  - Boat density data from EmodNET platform
  - Use animal intensity as proxy for small boat intensity
Distance from the coastline

Assume that the closer to the coast there are more small boats, hence a higher detection probability close to the coast.

- This in is not necessarily true, people like islands.
- This is also a covariate often used to model species density.
EMODnet data for boat density

- EmodNET (European Marine Observation and Data Network) records boats using AIS (Automatic Identification System, mandatory above 15m length)

- Detection probability is higher where boat intensity is higher
  - Does not consider small boats which are often those reporting sightings
Social data sightings for all species

- Use all sightings as a proxy for boat density
- Data include species with very different behavior
- Detection probability is higher where boat intensity is higher
Putting things together

- The “true” intensity:

\[ \lambda(s) = \beta_0 + \beta X(s) + u(s); \]
\[ u(s) \sim GRF(\rho, \sigma_u^2) \]

- The observed intensity:

\[ \lambda_{FERRY}(s) = t_{FERRY} \lambda(s) g_{FERRY}(s); \]
\[ \lambda_{UNIRM}(s) = t_{UNIRM} \lambda(s) g_{UNIRM}(s); \]
\[ \lambda_{SOCIAL}(s) = t_{SOCIAL} \lambda(s) g_{SOCIAL}(s); \]

Four choices for \( g_{SOCIAL}(s) \)

- No (constant) detection \( g_{SOCIAL}(s) = 1 \) (benchmark)
- Detection based on distance from the coastline
- Detection based on boat intensity
- Detection based on sightings intensity
Is the model identifiable?

-Low sigthings intensity can result from:
  ▶ There are no animals in the area
  ▶ There are no observer in the area
  ▶ How to solve this?
    ▶ Gather information about the observation process
    ▶ Use informative prior to “guide” inference
Prior for the parameters in the detection functions
Results
Reconstructed intensity surface (Stenella)
Reconstructed intensity surface (Tursiope)
Summary and conclusions
Summary and conclusions

- Complex but very topical problem
Summary and conclusions

- Complex but very topical problem
- What can we do:
  - Model several data sources jointly
  - Correct for the bias induced by the observation process
  - Recover known covariate effects
  - Estimate intensity surface with associated uncertainty

INLA + inlabru give a huge model flexibility...with great power

comes great responsibility!!!
Summary and conclusions

- Complex but very topical problem
- What can we do:
  - Model several data sources jointly
Summary and conclusions

- Complex but very topical problem
- What can we do:
  - Model several data sources jointly
  - Correct for the bias induced by the observation process

INLA + inlabru give a huge model flexibility... with great power comes great responsibility!!!
Summary and conclusions

- Complex but very topical problem
- What can we do:
  - Model several data sources jointly
  - Correct for the bias induced by the observation process
  - Recover known covariate effects
Summary and conclusions

- Complex but very topical problem

- What can we do:
  - Model several data sources jointly
  - Correct for the bias induced by the observation process
  - Recover known covariate effects
  - Estimate intensity surface with associated uncertainty
Summary and conclusions

- Complex but very topical problem
- What can we do:
  - Model several data sources jointly
  - Correct for the bias induced by the observation process
  - Recover known covariate effects
  - Estimate intensity surface with associated uncertainty
- INLA + inlabru give a huge model flexibility....with great power comes great responsibility!!!