Estimating seal pup abundance with SPDE-INLA

What to do when you don't get the amazing results you were hoping for?

Martin Jullum

Oslo, 01.02.17

Background

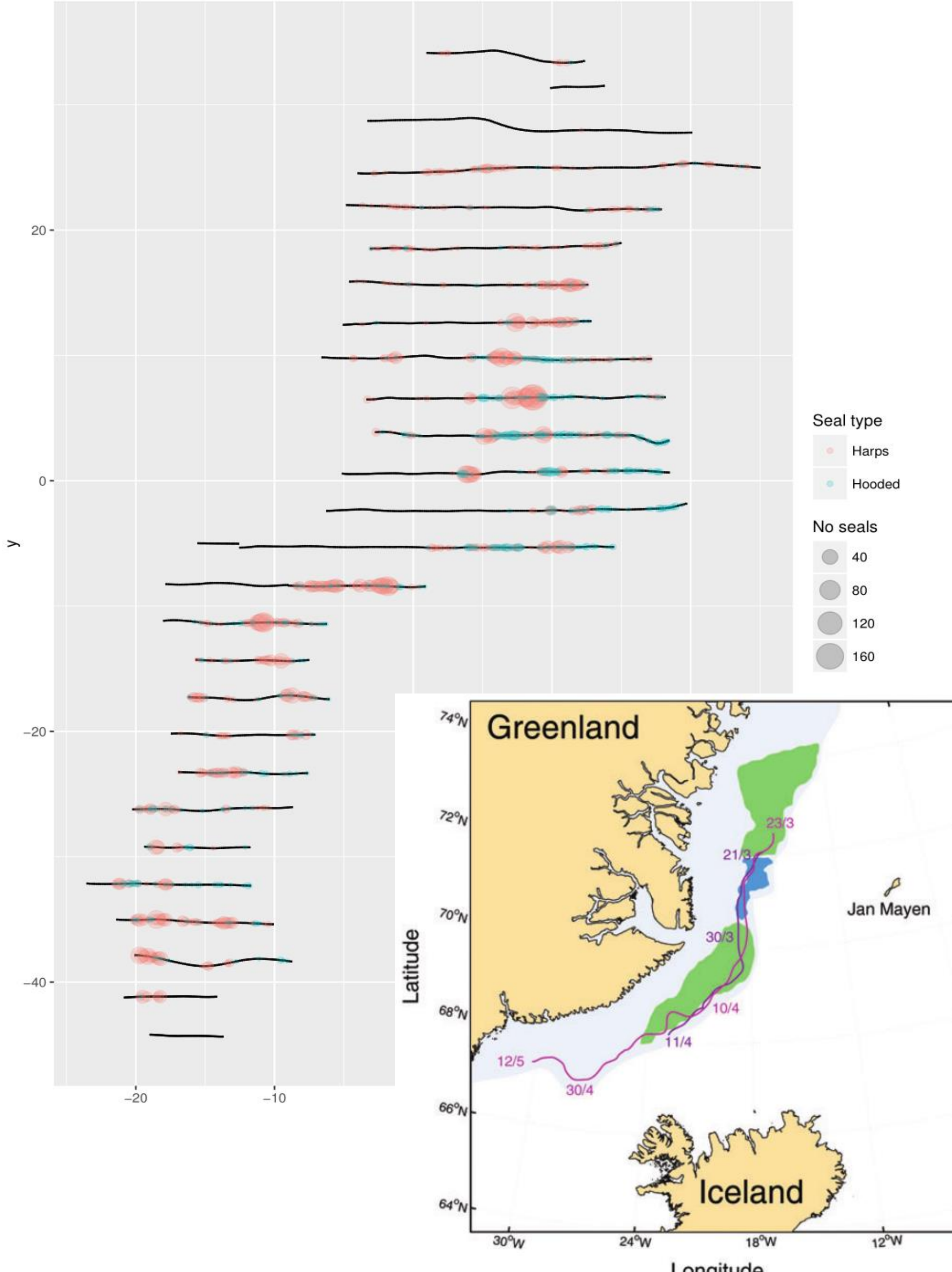
- Thordis & Martin’s original project: Write *some* paper on point processes using INLA (Dec 2015)
- Tor Arne contributed with the idea to use point processes in seal abundance modeling (Jan? 2016)
- Fabian Bachl (Uni. Edinburgh) joined the team (July? 2016)
- People
  - Thordis: Point process expert
  - Fabian: INLA expert
  - Tor Arne: Seal data expert
  - Martin: No experience with point processes, INLA, nor seals

Problem

- Ultimale goal: Monitor seal abundance in the North Atlantic
- Well established dynamic abundance model for seals
  - Key component is **estimate + uncertainty of the number of seal <u>pups</u>**
    - Typically based on oversimplified method
    - Do not trust the uncertainty
- **Our task: Propose a method to estimate the total number of <u>pups</u> with uncertainty**

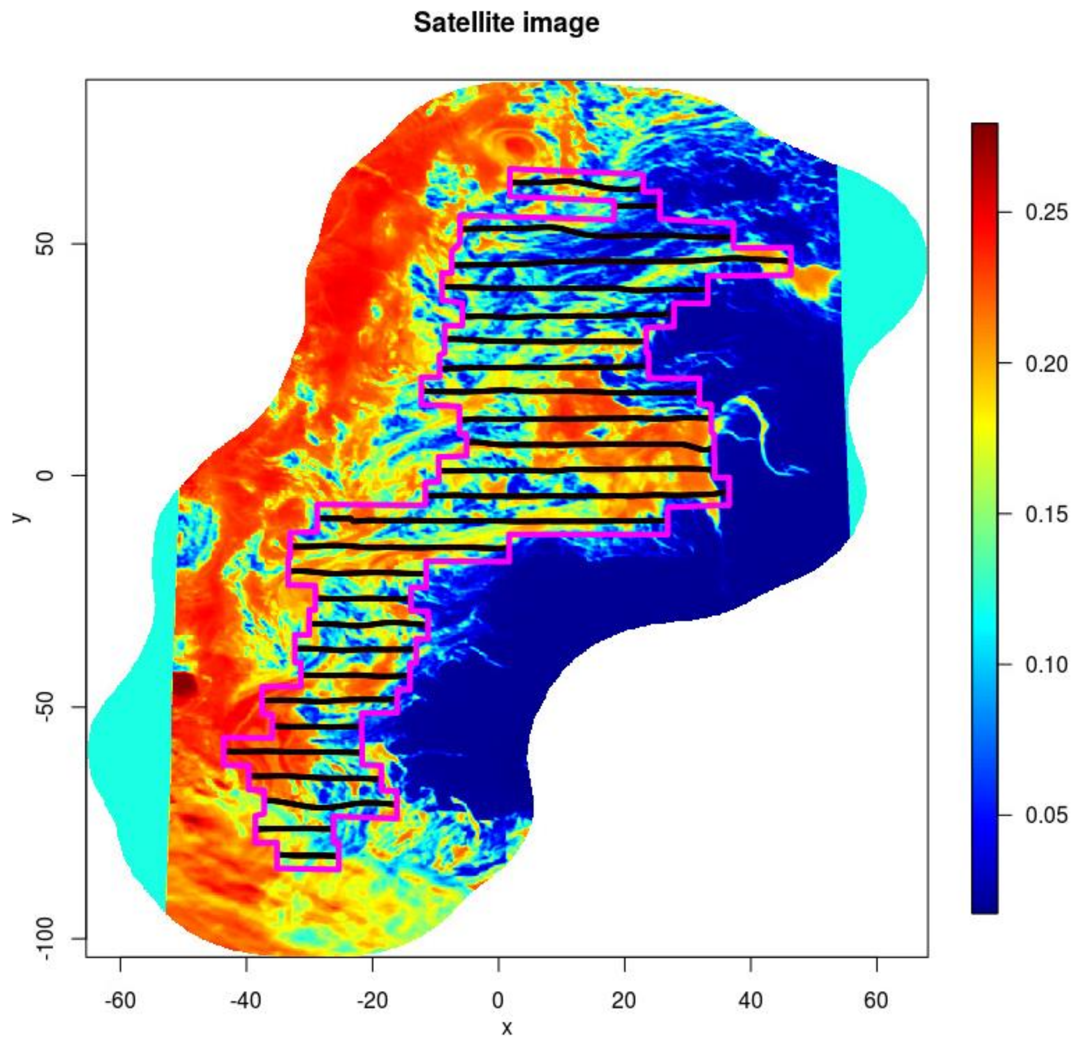
Data

- From an aerial photo survey conducted east of Greenland in 2012
- Number of pups in 2792 photos in 27 transects sparsely covering the seal domain
- 2 seal types: Harps and hooded



Data

- From an aerial photo survey conducted east of Greenland in 2012
- Number of pups in 2792 photos in 27 transects sparsely covering the seal domain
- 2 seal types: Harps and hooded
- Additional info: Quantified satellite image to indicate ice thickness
- Seal domain Ω shown in pink



Solution idea

- Handle the seal types separately
- Model the spatial distribution of the pups with a Log-Gaussian Cox Process (LGCP)
  - Gaussian latent field Z
  - Point pattern $Y|Z \sim \mathrm{PoissonProcess}(\lambda(s) = \exp(Z(s)))$
  - Log-likelihood: $|\mathrm{A}| - \int_A \exp\big(Z(s)\big)ds + \sum_i Z(s_i)$, s_i the location of photo i
  - LGCP property: Given Z, counts $N(B)$ in disjoint Borel sets B indep. and distributed as $\mathrm{Poisson}(\lambda = \int_B \exp\big(Z(s)\big)\, ds)$
- The Bayesian solution to our problem is the «**posterior predictive distribution**» of pup counts in the seal domain $p(N(\Omega)|Y)$
  - Easy to compute with samples from $p(Z|Y)$
- We use
  - $\mathrm{Z}(s) = \alpha + \beta^t \boldsymbol{x}_s + f(s)$, $\boldsymbol{x}_s$ satellite information, $f(s)$ zero-mean Gaussian field
  - Vague priors on all parameters
- Main challenges:
  - Observe only a small part of the seal area
  - Data are aggregated counts per photo
  - How can such a model be fitted in INLA?

3 approximation approaches

- 1: Poisson regression on a regular grid (Rue et al. 2009)
  - Grid the modeling area
  - Let $f(s)$ be a GMRF on that grid
  - Per LGCP property: Given Z, the counts in each cell are indep. Poisson with $\lambda_i = \int_{A_i} \exp(Z(s))ds$
    - $\lambda_i \approx \lambda_i^* = |A_i|\exp(Z(s_i^*))$ for s_i^* the mid-point in cell A_i
  - Model the counts in each cell as Poisson distributed with offset $|A_i|$

3 approximation approaches

- 1: Poisson regression on a regular grid (Rue et al. 2009)
  - Grid the observation area
  - Let $f(s)$ be a GMRF on that grid
  - Per LGCP property: Given Z, the counts in each cell are indep. Poisson with $\lambda_i = \int_{A_i} \exp(Z(s))ds$
    - $\lambda_i \approx \lambda_i^* = |A_i|\exp(Z(s_i^*))$ for s_i^* the mid-point in cell A_i
  - Model the counts in each cell as Poisson distributed with offset $|A_i|$
- 2: Direct likelihood approximation with SPDE field (Simpson et al. 2016)
  - Let let $f(s)$ be a SPDE-based GMRF
  - Apply deterministic integration approximation + utilize FEM representation of Z
  - Gives approx likelihood equivalent to a Poisson likelihood for certain pseudo-observations
  - Can be fitted directly by contructing the pseudo-observations

3 approximation approaches

- 3: Poisson regression with SPDE field
  - Per LGCP property: Given Z, the counts per **photo** are indep. Poisson with $\lambda_i = \int_{A_i} \exp(Z(s))ds$
    - $\lambda_i \approx \lambda_i^* = |A_i|\exp(Z(s_i^*))$ for s_i^* the mid-point in photo A_i
  - Construct a triangulated mesh in a special way
    - One mesh point in the center of each photo
    - Other mesh points distributed such that the extent of each photo $|A_i|$ corresponds to the Voronoi tesselation of the center mesh point
  - Fit the counts per photo as a Poisson regression with $|A_i|$ as offset, using INLAs SPDE approach

We chose the 3rd approach

- Since
  - Allows us to use the aggregated photo counts directly
  - Model defined also without the LGCP argument (data are not actually a point pattern)
  - Regular grid is restrictive

Alternative model

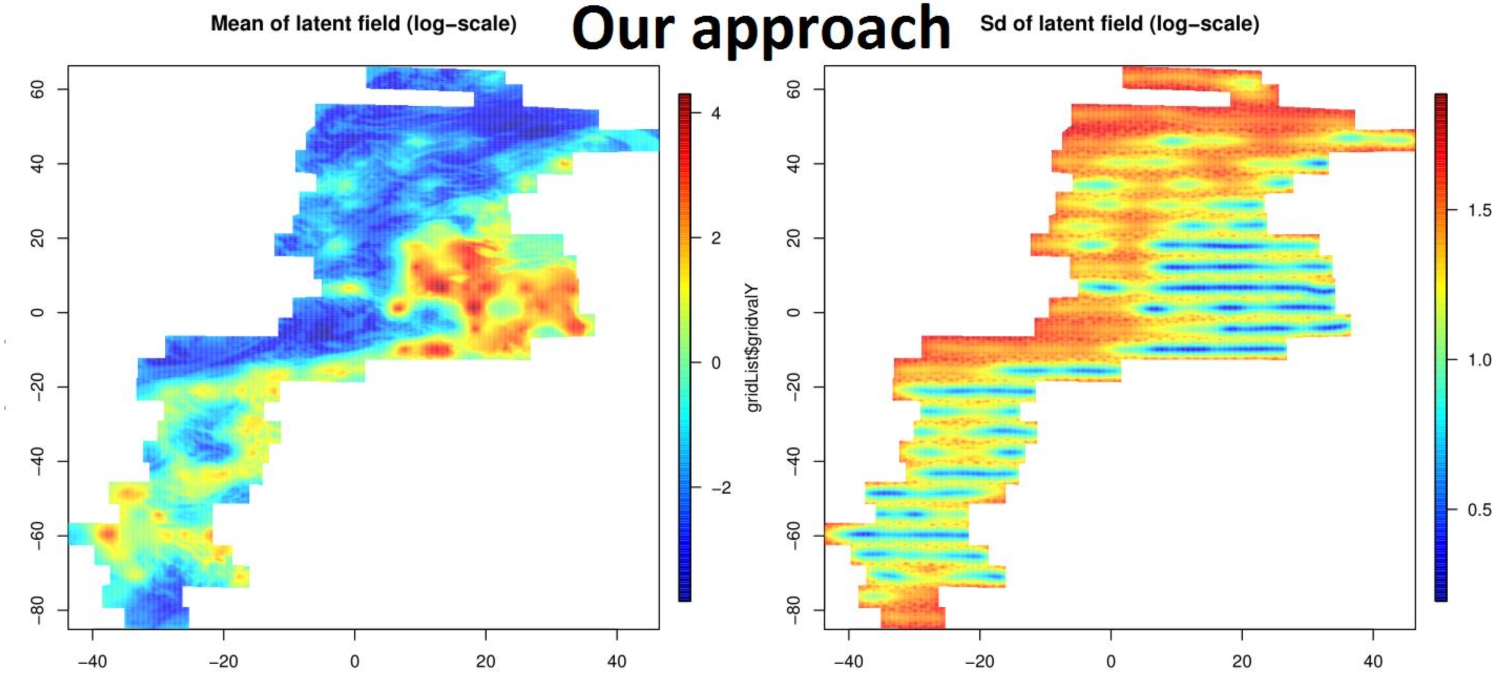
- Arnt-Børre + Tor Arne + others (2009)
  - Let $Z_0(s) = f_{GAM}(s)$, with $f_{GAM}(s)$ a (spatial) smooth spline.
  - $\mu_i = |A_i| \exp(Z_0(s_i^*))$ for s_i^* the mid-point in cell A_i
  - Fit the counts per photo as a negative binomial regression with constant shape κ and $|A_i|$ as offset
  - Frequentist approach
  - Smoothness of $f_{GAM}(s)$ chosen through generalized CV

  - We test this formulation, also with satellite data and Poisson distributions
  - Use sampling to produce predictive distribution of total pup counts for comparison

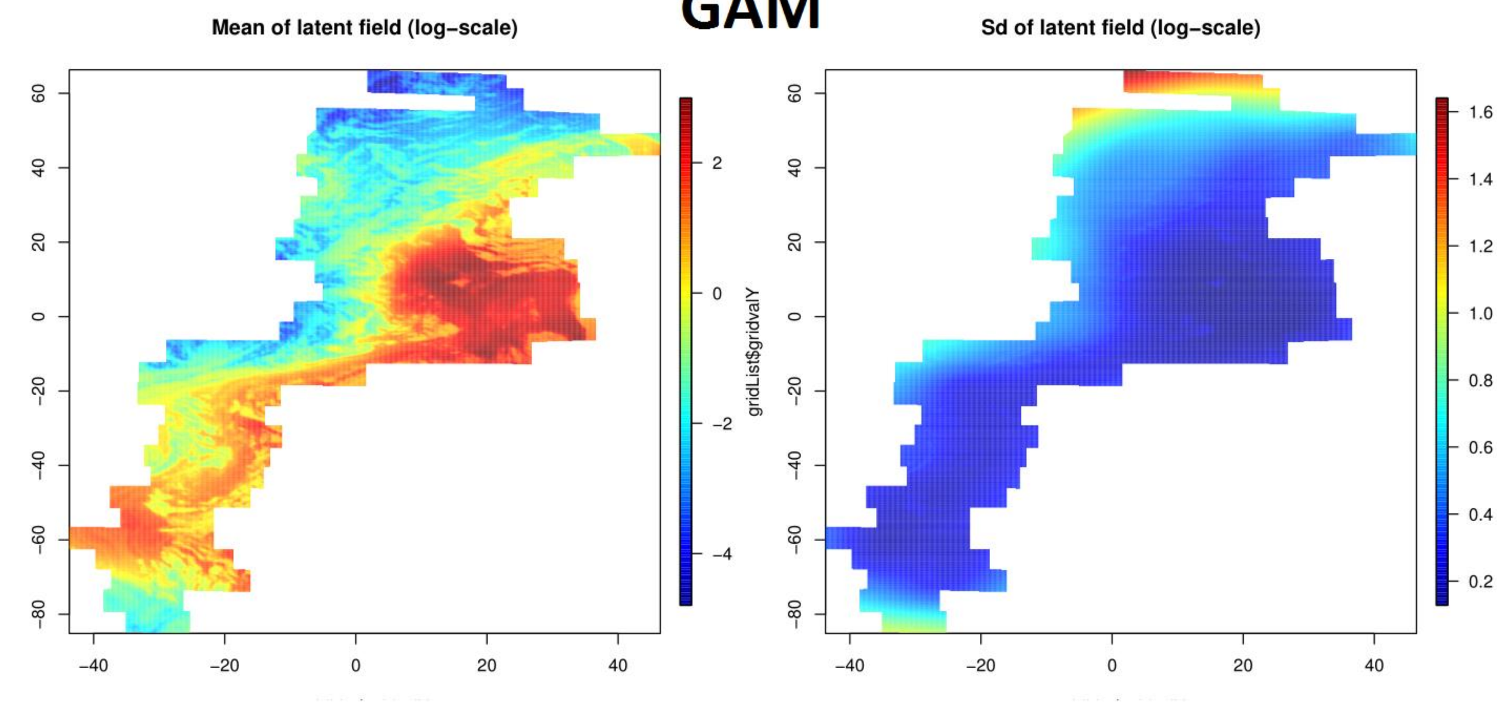
Model comparison

- GAM slightly smaller point predictions and much smaller uncertainty
- 2 cross validation schemes
  - Leave out one full transect at a time
    - Predict photo and transect counts
  - Regular 10-fold CV
    - Predict left out individual photos and sum of photo counts
- Performance measures
  - CRPS = $\int \left(\hat{F}(n) - \mathbf{1}_{\{n-n_{true}\}}\right)^2 dn$
  - LogScore = $\log(\hat{f}(n_{true}))$
- Performance results
  - Our approach better at photo level
  - GAM better/as good on aggregated/transect level

Our approach



GAM



Conclusion

- What have we learned
  - It can be hard to know exactly what you are doing in INLA
  - Doing non-standard things in INLA requires a lot of coding
  - It takes time to
    - learn something completely new
    - try out a lot of different approaches
- Further work
  - Joint model harps and hooded seals?
- Subprojects
  - Exact analytical formula for LGCP likelihood for SPDE fields

Discussion

- What to do when you don’t get the results you want?
  - Can we get a decent paper out of this?
    - The only fundamentally new thing is using satellite images as covariates for this application
      - Bayes + SPDE + proper validation procedure are also new, but not fundamentally
    - What we care about is total pup abundance, our approach is only superior on a finer scale
    - Do we need to try out a joint model for harps and hooded seals?
      - Are we dependent on great results for such a model?
- What could we have done differently?