Estimating seal pup abundance with LGCP

Work in progress (!)

Martin Jullum

Joint with Thordis Thorarinsdottir and Fabian Bachl

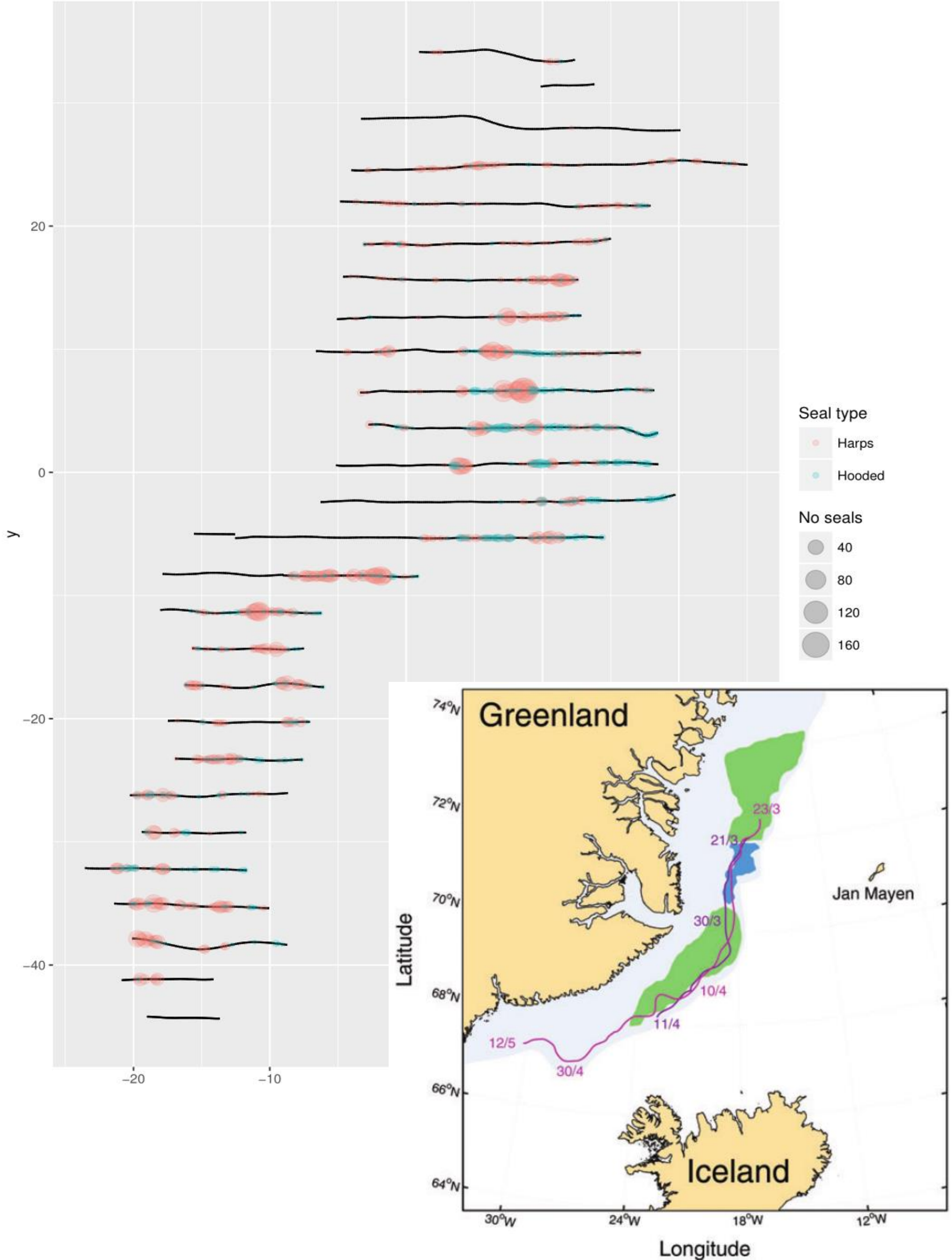
Trondheim, 10.11.16

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 pups**
    - Typically based on oversimplified method
    - Do not trust the uncertainty
- **Our task: Propose a method to quantify the total number of pups with uncertainty**

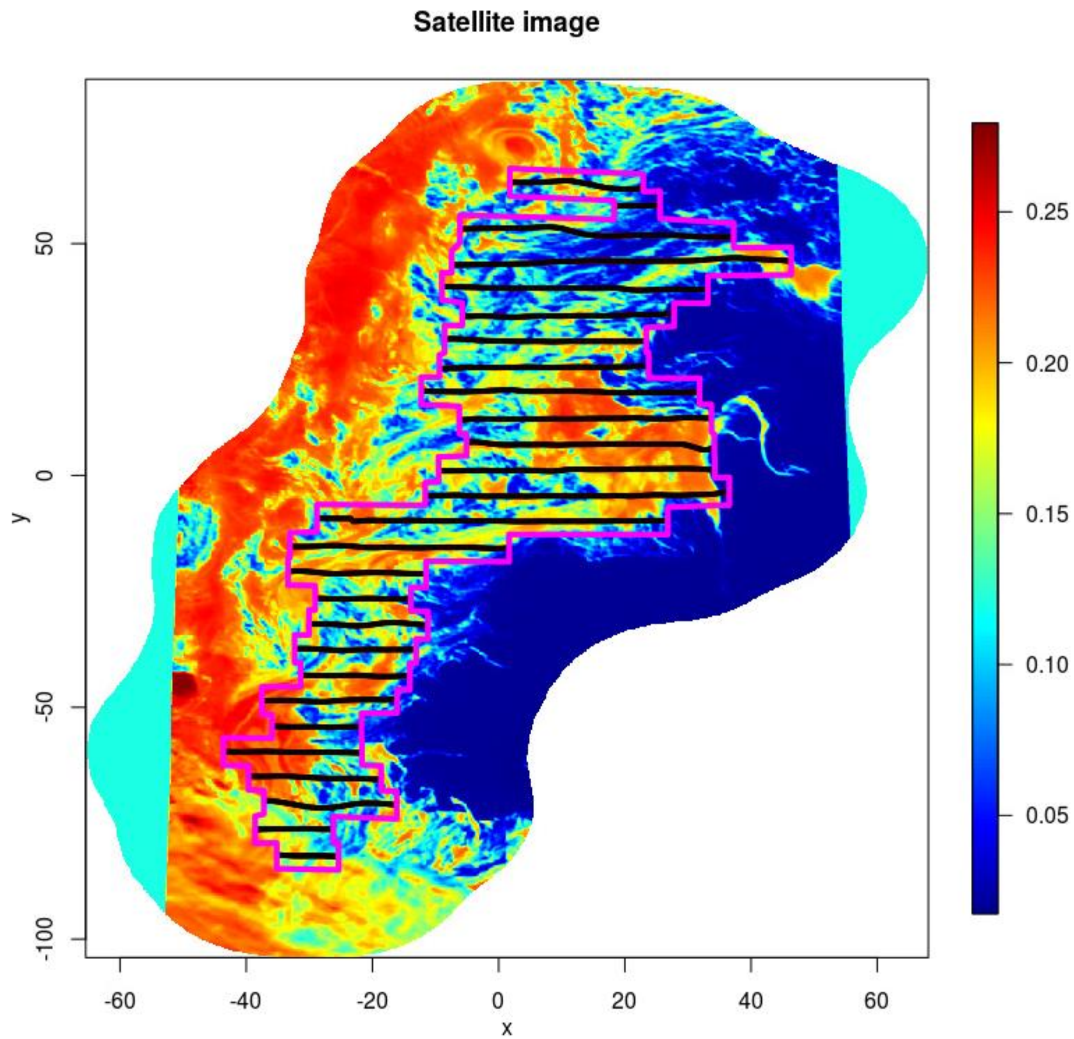
Data

- From an aerial photo survey conducted east of Greenland in 2012
- Number of pups in 2792 photos (A) 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 (A) 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



Main approach

- Model the spatial distribution of the pups with an Log-Gaussian Cox Process (LGCP)
  - Gaussian latent field Z
  - Point pattern $Y|Z \sim \text{PoissonProcess}(\lambda(s) \ = \exp(Z(s)))$
  - Given Z, counts $N(B)$ in disjoint Borel sets B indep. and distributed as $\text{Poisson}(\lambda \ = \int_B \ \exp\big(Z(s)\big)\, ds)$
- 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 $p(Z|Y)$
- Wish to utilize INLA framework and the SPDE approach
  - $\mathrm{Z}(s) = \alpha + \beta^t \boldsymbol{x}_s + f(s)$, $\boldsymbol{x}_s$ satellite information, $f(s)$ SPDE-based Matern GMRF
- Main challenges:
  - Observe only a small part of the seal area
  - Data are aggregated counts per photo
  - Several ways to fit approximated versions using INLA
    - Are the approximations sufficiently accurate?

LGCP approximation approaches

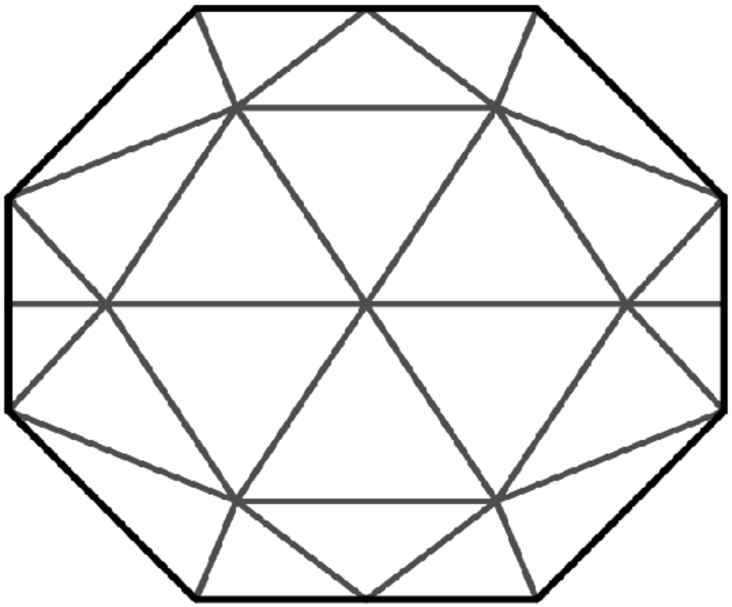
- Poisson regression formulation
- Direct likelihood approximation (Simpson et al. 2016, Biometrika)
  - LGCP Log-likelihood

$$|A| - \int_A \exp(Z(s))\mathrm{d}s + \sum_{i=1}^{n} Z(s_i),$$

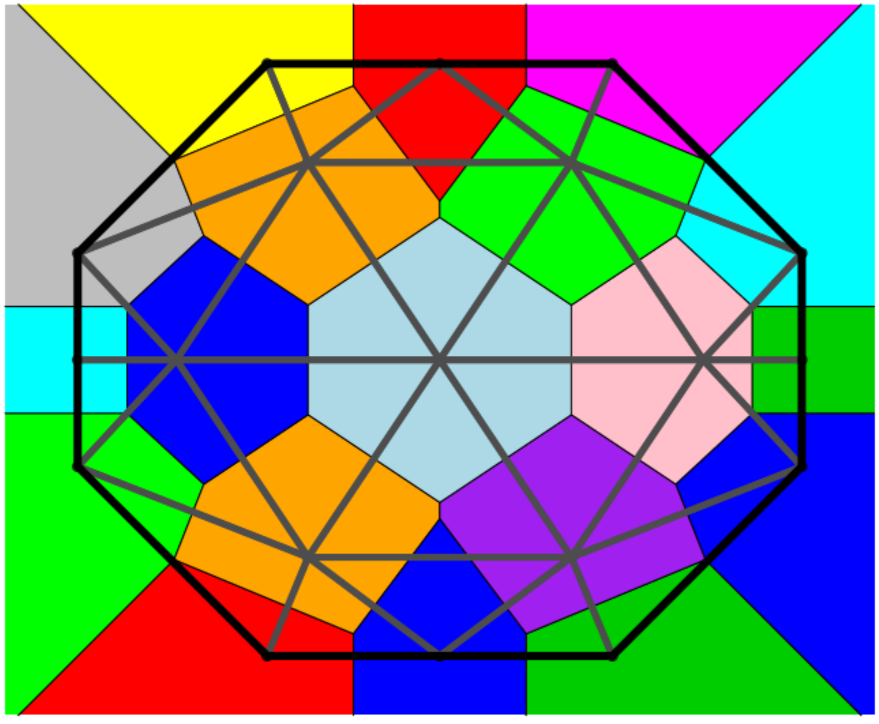
  - We consider different variations of this approach

Voronoi tesselation

Delaunay triangulated mesh



Voronoi tesselation



Poisson regression formulation

- Per LGCP formulation: Given Z, the counts per photo N_i 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 mesh such that the extent of each photo A_i corresponds to the Voronoi tesselation of a mesh point
- Fit the counts per photo $(N_1, \ldots, N_{2792})$ as a Poisson regression with $|A_i|$ as offset, using INLAs SPDE approach

Direct likelihood approx.

- Recall $$\log(p(y|Z)) = C - \int_A \exp(Z(s))\mathrm{d}s + \sum_{i=1}^{n} Z(s_i),$$

- Approx. integration: $\int_A \exp\big(Z(s)\big)ds \approx \sum_{k=1}^{K} \alpha_k \exp(Z(\tilde{s}_k))$ where the $\tilde{s}_k$ are deterministic integration points with weights α_k, and $\sum_k^K \alpha_k = |A|$

- When Z is SPDE-based with q mesh points: $Z(s) = \sum_{j=1}^{q} z_j\, \phi_j(s)$, for z_j indep. Gaussian and $\phi_j(s)$ piecewise linear basis functions

- $\log\big(p(y|Z)\big) \approx C - \sum_k^K \alpha_k \exp\left(\sum_{j=1}^{q} z_j\, \phi_j(\tilde{s}_k)\right) + \sum_{i=1}^{n} \sum_{j=1}^{q} z_j\, \phi_j(s_i)$

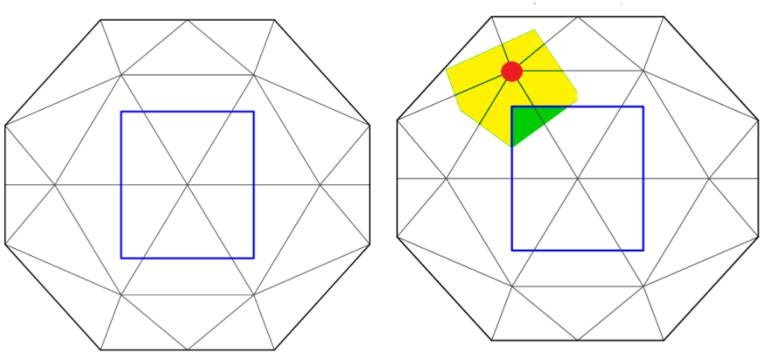
- Reformulation with pseudo-observation y_i^* and linear predictor η_i shows this can be written on Poisson form
$$p(y|Z) \approx \prod_{i}^{n+K} \eta_i^{y_i^*} \exp(-\alpha_i \eta_i)$$

Direct likelihood approx: Data usage and integration scheme

- Data usage. Either
  - All counts in photo located in photo center
  - Randomly distribute counts within each photo

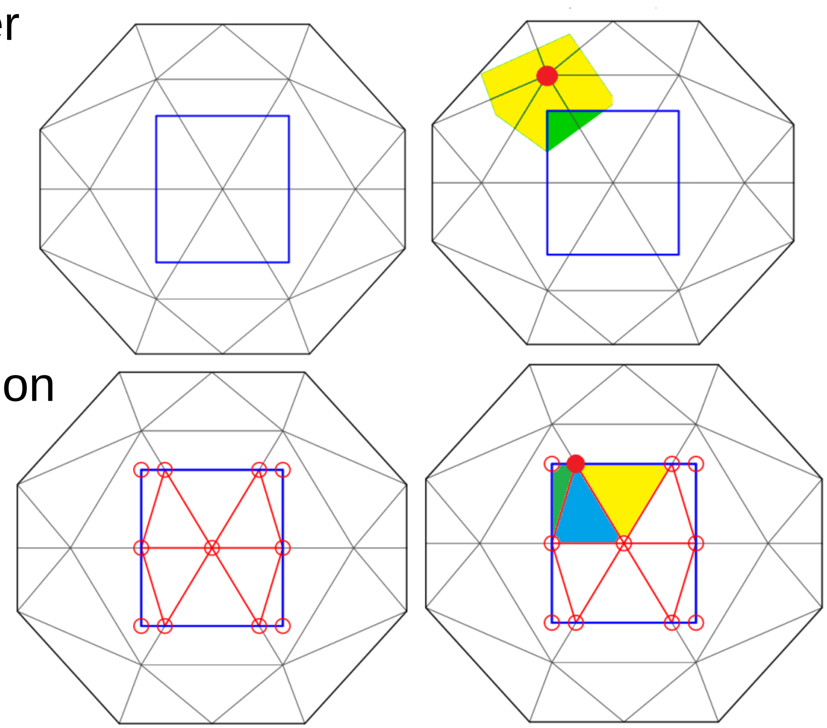
Direct likelihood approx: Data usage and integration scheme

- Data usage. Either
  - All counts in photo located in photo center
  - Randomly distribute counts within each photo
- Integration approach 1
  - $\tilde{s}_k$ = Mesh points
  - α_k = Area of assoicated Voronoi tesselation



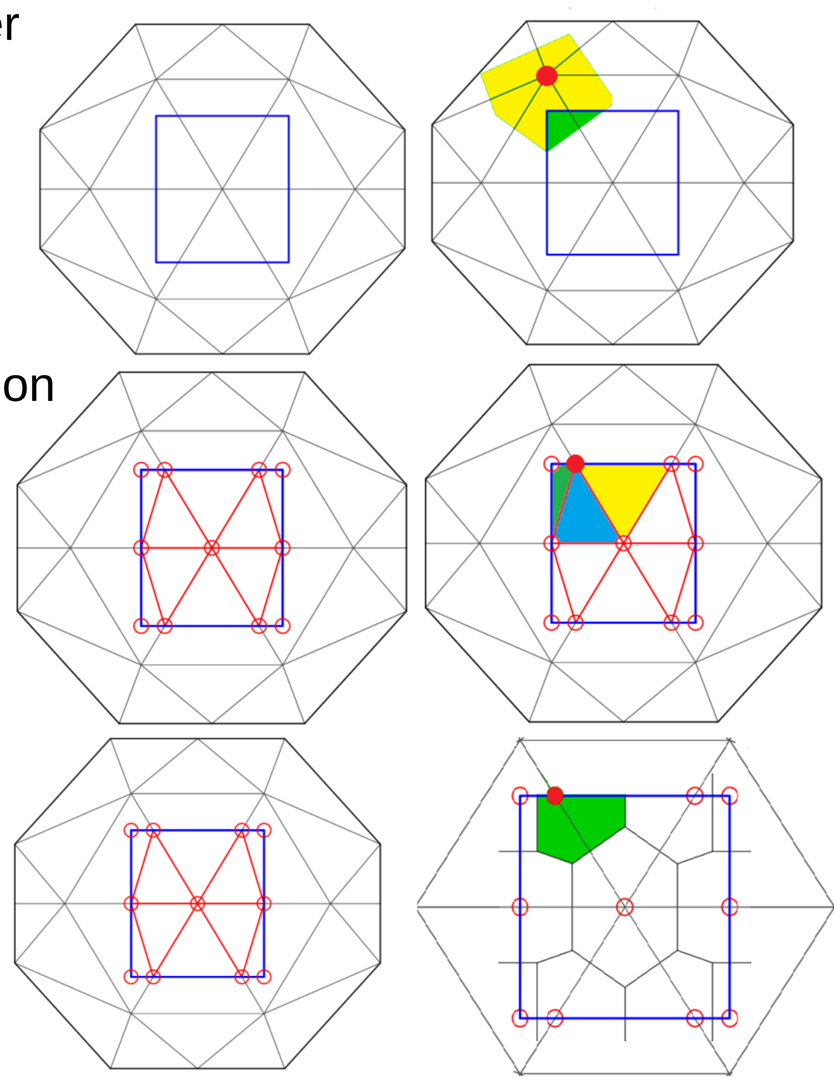
Direct likelihood approx: Data usage and integration scheme

- Data usage. Either
  - All counts in photo located in photo center
  - Randomly distribute counts within each photo
- Integration approach 1
  - $\tilde{s}_k$ = Mesh points
  - α_k = Area of assoicated Voronoi tesselation
- Integration approach 2
  - Create a refined mesh within each photo
  - $\tilde{s}_k$ = Vertices in refined mesh
  - α_k = 1/3 of area of connected triangles



Direct likelihood approx: Data usage and integration scheme

- Data usage. Either
  - All counts in photo located in photo center
  - Randomly distribute counts within each photo
- Integration approach 1
  - $\tilde{s}_k$ = Mesh points
  - α_k = Area of assoicated Voronoi tesselation
- Integration approach 2
  - Create a refined mesh within each photo
  - $\tilde{s}_k$ = Vertices in refined mesh
  - α_k = 1/3 of area of connected triangles
- Integration approach 3
  - Create a refined mesh within each photo
  - $\tilde{s}_k$ = Vertices in refined mesh
  - α_k = Area of assoicated Voronoi tesselation (within photo)



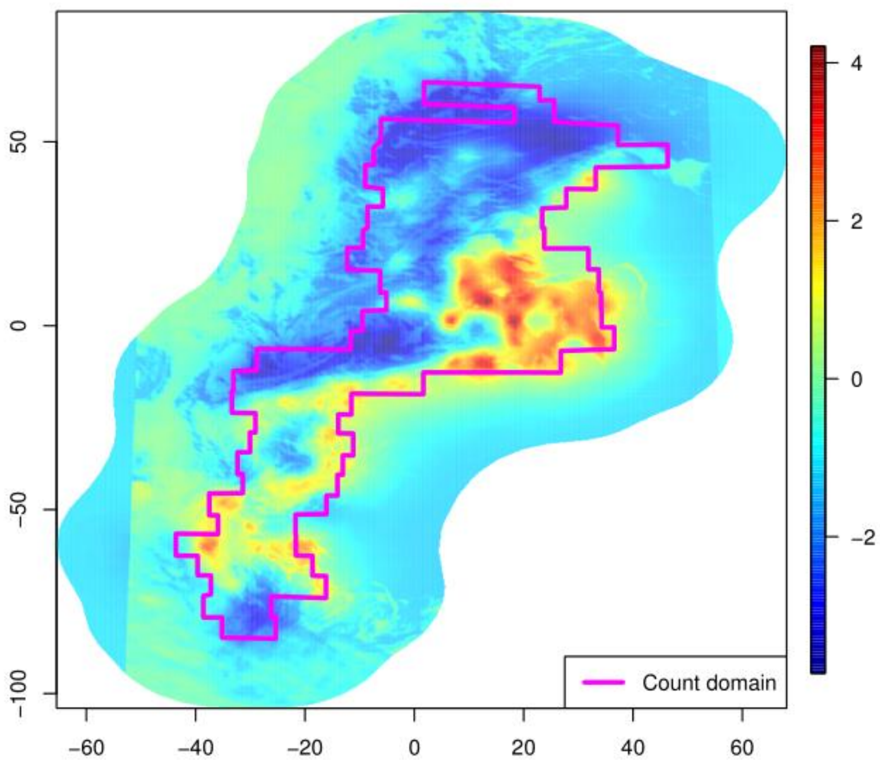
Results

- The current simple model
  - Pup abundance estimate:
    10928, 95% CI = (8120,13844)

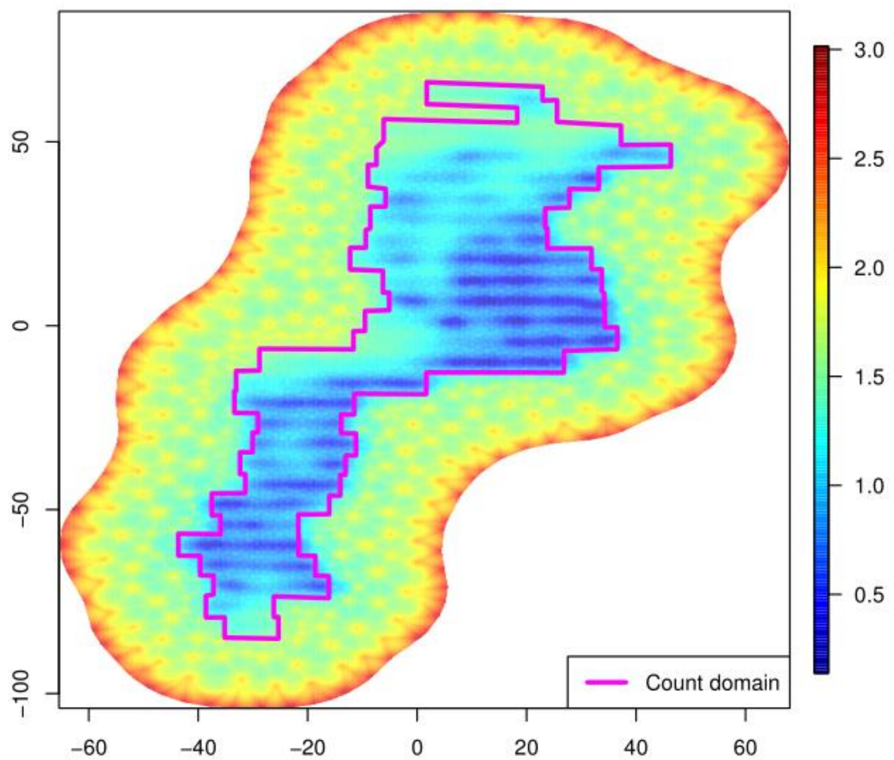
Results

- The current simple model
  - Pup abundance estimate:
    10928, 95% CI = (8120,13844)
- Latent field for our methods (only minor differences between approaches)

Mean of latent field (log−scale)

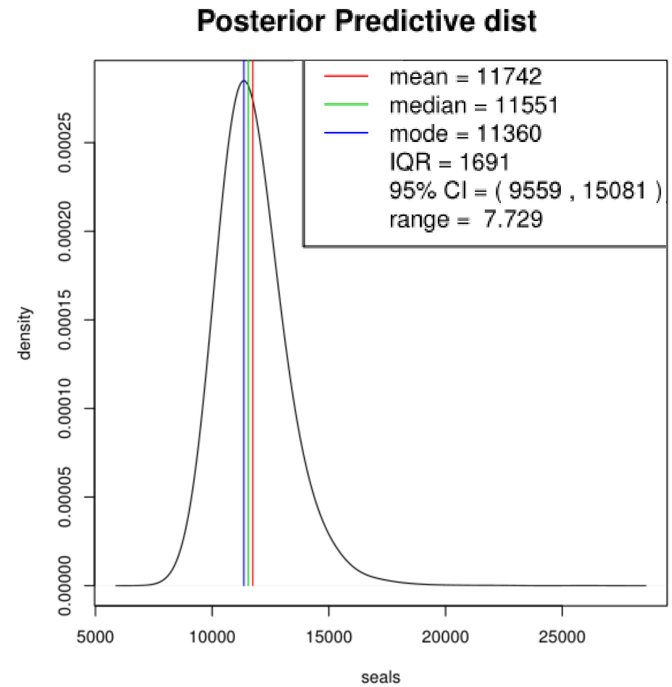


Sd of latent field (log−scale)

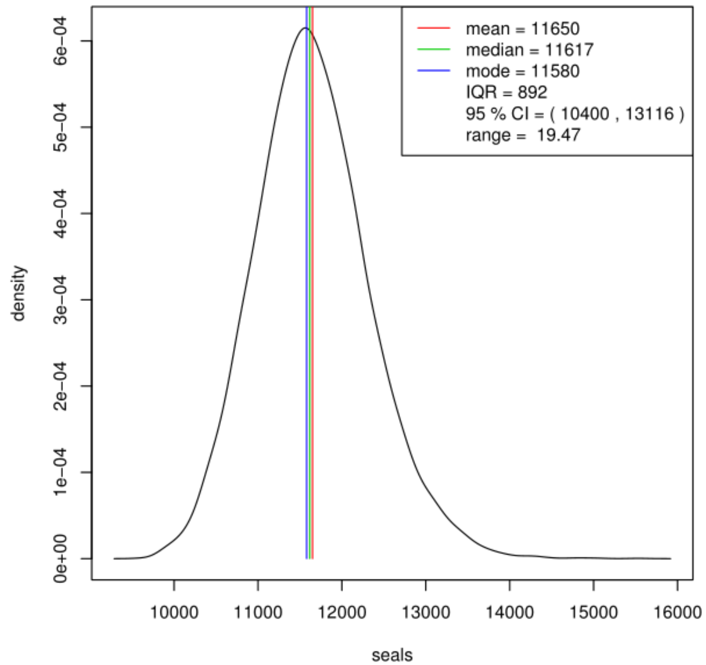


Results

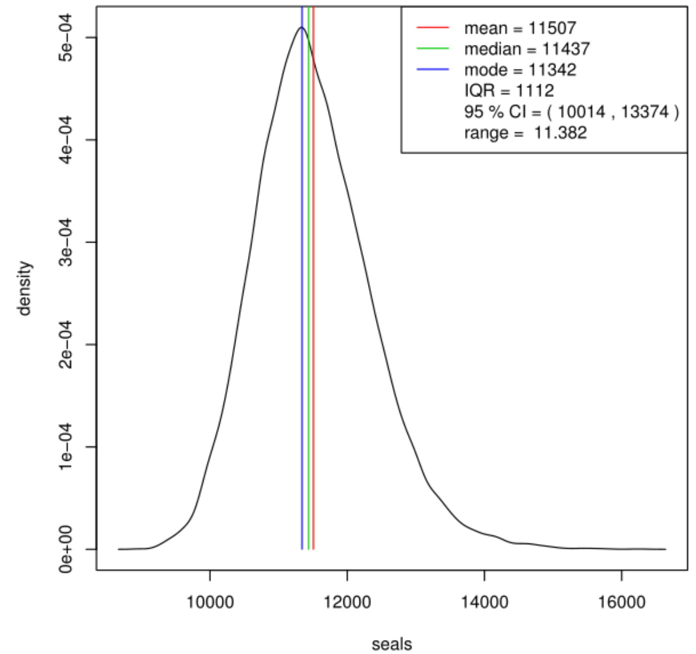
- The current simple model
  - Pup abundance estimate: 10928, 95% CI = (8120,13844)
- Poisson regression formulation
- Direct likelihood approx.



$\tilde{s}_k$: Original mesh, α_k: Voronoi



$\tilde{s}_k$: Refined mesh, α_k: 1/3 splitting

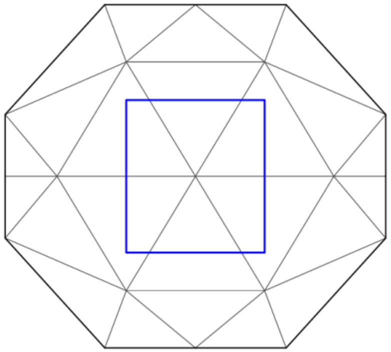


Integration constant

- When $Z(s) = \sum_{j=1}^{q} z_j \, \phi_j(s)$, we can actually solve $\int_A \exp\big(Z(s)\big)ds$ analytically for every realization of Z!

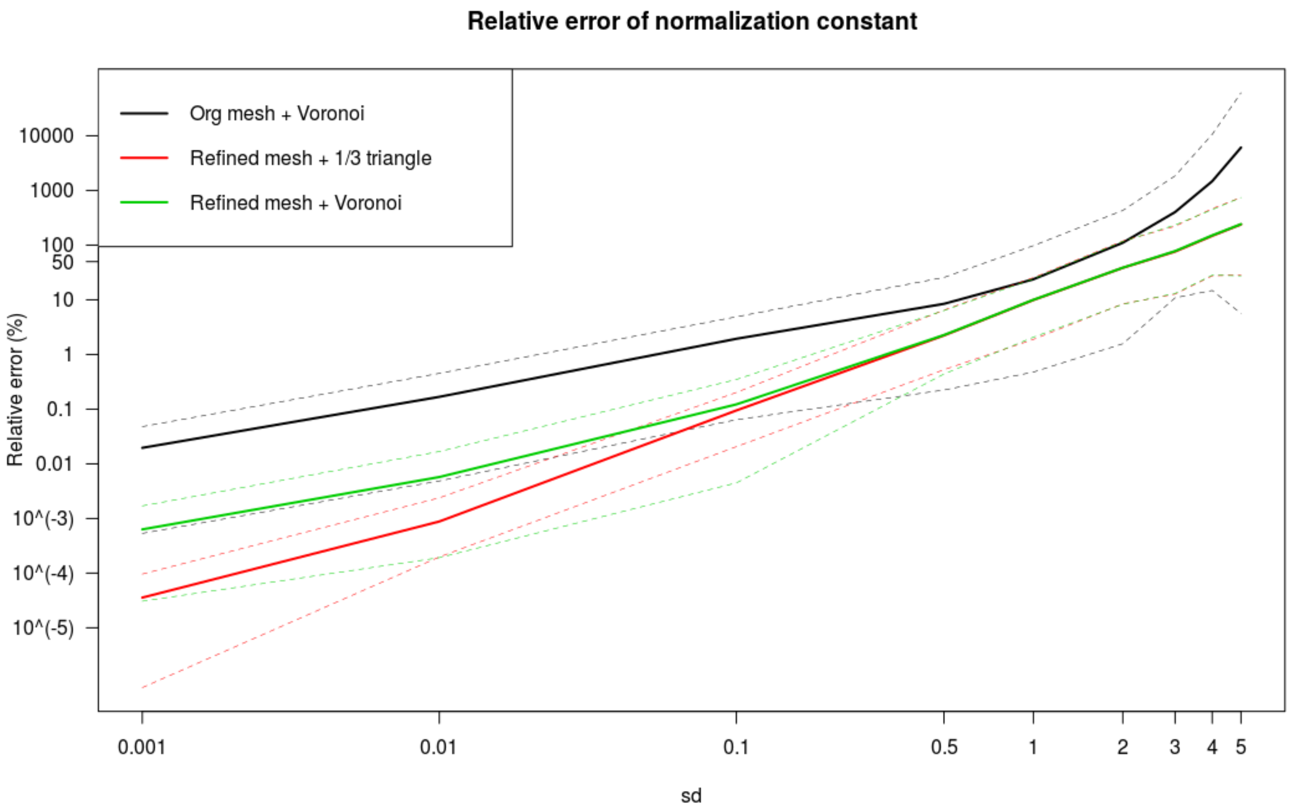
- $\int_A \exp\big(Z(s)\big)ds = \sum_l^L g(\boldsymbol{z}; \eta_{l1}, \eta_{l2}, \eta_{l3})$, for g a non-linear function, $\boldsymbol{z} = (z_1, \ldots, z_q)$ and η_{lj} linear predictors dependent on mesh points and the observation region, $\mathrm{L} = O(q)$

- Basic simulation study
  - Fixed mesh and integration region
  - Sample different types of Z and check performance of different integration variants



Integration constant

- All methods overestimates $\int_A \exp\big(Z(s)\big)\, ds$
- Performance varies with sd of Z, not mean



- Implication: Likelihood contribution will be too small where the variability of Z is large

Further work

- Cross validation performance test on seal data

- Open methodological questions
  - Benefits of joint modeling of the seal types?
  - What are the implications of the error in the integration constant?
    - MCMC-run with exact integration constant ?
  - Can a modified version of INLA run with the exact integration constant?
    - Requires 3 linear predictors, rather than 1