The (un)reliability of contour curves

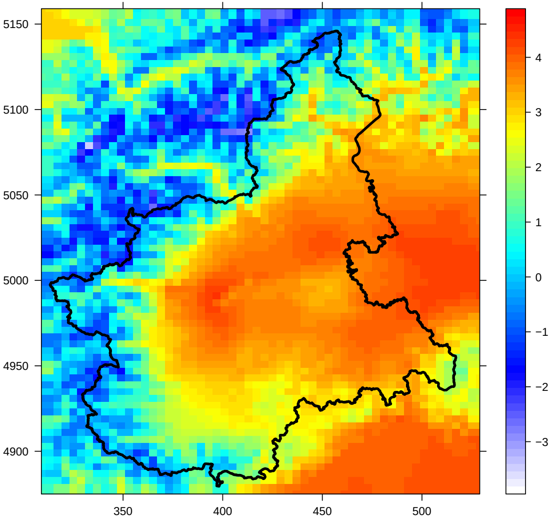
Excursion sets and contour uncertainty regions

David Bolin
Chalmers University of Technology

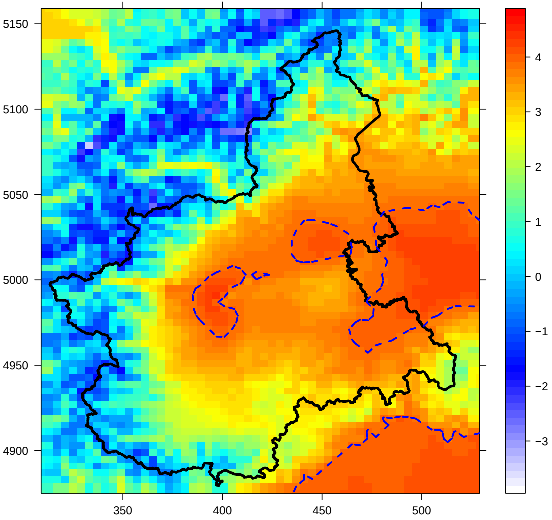
joint work with Finn Lindgren

Búzios, RJ, Brazil,
June 18, 2014

PM_{10} in Piemonte: Where is $PM_{10} > 50$?



PM_{10} in Piemonte: Where is $\mathrm{PM}_{10} > 50$? Uncertainty?



The problem setting

We have observations $\mathbf{y} = (y_1, \dots, y_n)$ at locations $(\mathbf{s}_1, \dots, \mathbf{s}_n)$ of a latent random field $x(s)$. The model is specified through

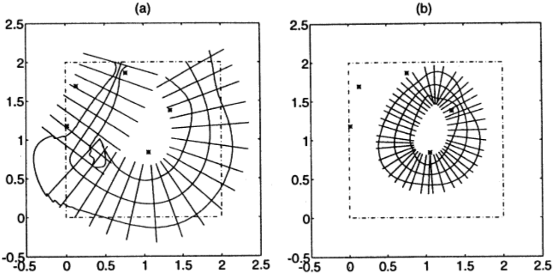
- The (possibly non-gaussian) likelihood $\pi(y_i | x(s_i), \boldsymbol{\theta})$.
- A random field model for $x(s)$, typically including covariates.
- Prior distributions for the parameters.

We estimate the parameters and the posteriors (e.g. using INLA) and use the posterior mean $\mathsf{E}(x(s)|\mathbf{y})$ as a point estimate of the latent field.

We are interested in the uncertainty of contour curves and excursion sets for $x(s)|\mathbf{y}$.

Later, we will assume that $x(s)$ is Gaussian, so that we are in the LGM framework where INLA can be used for estimation.

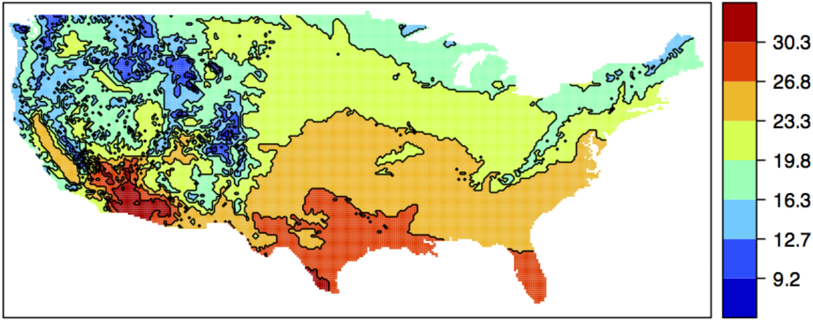
Confidence sets for level contours



Lindgren, Rychlik (1995): *How reliable are contour curves? Confidence sets for level contours*, Bernoulli

- Regions with a single expected crossing
- Method assumes Gaussian likelihood.
- The confidence band is not simultaneous.

Contours maps



Polfeldt (1999), *On the quality of contour maps*, Environmetrics

- How many contour curves should one use in a contour map?
- Based on calculating the marginal probabilities for the field staying between upper and lower contour levels.
- Method assumes Gaussian likelihood.
- Method does not take spatial dependency into account.

Contours and excursions

- A countour curve of a reconstructed field can (almost) be found from the pointwise marginal distributions.
- The *uncertainty* depends on the full joint distribution.
- A credible contour region is a region where the field transitions from being clearly below, to being clearly above.
- An excursion region is a region where the field is clearly above (or below) a given level.
- Finding excursion regions is closely related to multiple testing.
- Solving the problem for excursions solves it for contours.

We now need to

- Give precise definitions for the uncertainty regions.
- Construct a method for finding the regions.

Outline

Definitions for functions

Excursion sets for functions

Given a function $f(s)$, $s \in \Omega$, the positive and negative excursion sets for a level u are

$$A_u^+(f) = \{s \in \Omega; f(s) > u\} \quad \text{and} \quad A_u^-(f) = \{s \in \Omega; f(s) < u\}.$$

Contour sets for functions

Given a function $f(s)$, $s \in \Omega$, the contour set A_u^c for a level u is

$$A_u^c(f) = \left(A_u^+(f)^o \cup A_u^-(f)^o\right)^c$$

where A^o is the interior and A^c the complement of the set A.

Excursion sets for random fields

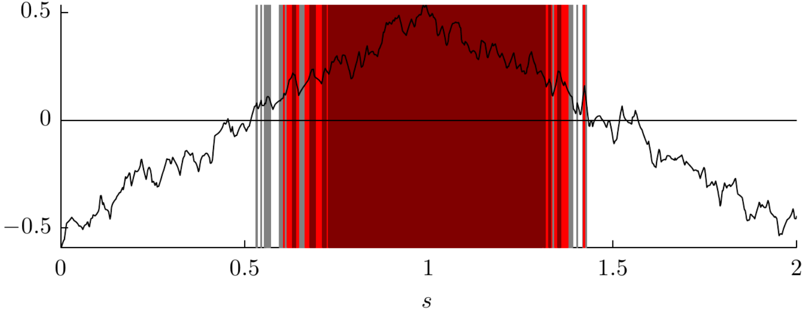
Excursion sets

Let $x(s)$, $s \in \Omega$ be a random process. The positive and negative level u excursion sets with probability $1-\alpha$ are

$$E^+_{u,\alpha}(x) = \arg\max_D \{|D| : \mathsf{P}(D \subseteq A^+_u(x)) \geq 1-\alpha\}.$$
$$E^-_{u,\alpha}(x) = \arg\max_D \{|D| : \mathsf{P}(D \subseteq A^-_u(x)) \geq 1-\alpha\}.$$

- $E^+_{u,\alpha}(x)$ is the largest set so that, with probability $1-\alpha$, the level u is exceeded *at all locations* in the set.
- Another possible definition of an excursion set would be a set that contains *all excursions* with probability $1-\alpha$. This set is given by $E^-_{u,\alpha}(x)^c$.

Example 1: Gaussian process with exponential covariance



- Gaussian process with exponential covariance function.
- $E^{+}_{0,0.05}(x)$ is shown in red.
- The grey area contains $\{s : \mathsf{P}(x(s) > 0) > 0.95\}$.
- The dark red set is the Bonferroni lower bound.
- The black curve is the kriging estimate of $x(s)$.

Contour sets

Level avoiding sets

Let $x(s)$, $s \in \Omega$ be a random process. The pair of level u avoiding sets with probability $1-\alpha$, $(M^+_{u,\alpha}(x), M^-_{u,\alpha}(x))$, is equal to

$$\underset{(D^+,D^-)}{\arg\max}\{|D^- \cup D^+| : \mathsf{P}(D^- \subseteq A^-_u(x),\ D^+ \subseteq A^+_u(x)) \geq 1-\alpha\}.$$

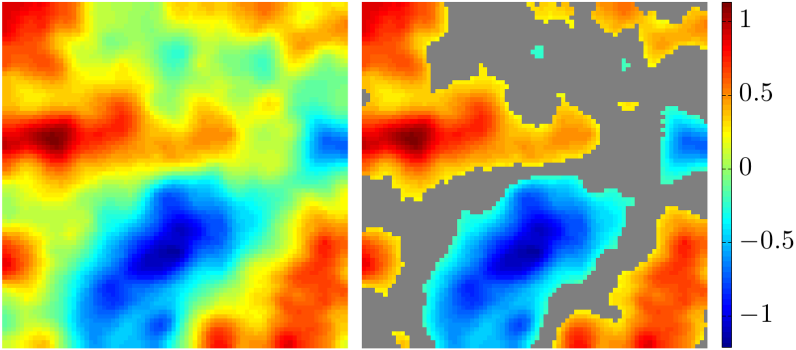
Uncertainty region for contour sets

Let $(M^+_{u,\alpha}(x), M^-_{u,\alpha}(x))$ be the pair of level avoiding sets. The uncertainty region for the contour set of level u is then

$$E^c_{u,\alpha}(x) = \left(M^+_{u,\alpha}(x)^o \cup M^-_{u,\alpha}(x)^o\right)^c.$$

- $E^c_{u,\alpha}$ is the smallest set such that with probability $1-\alpha$ *all* level u crossings of x are in the set.

Example 2: Gaussian Matérn field



- Gaussian Matérn field measured under Gaussian noise.
- Left panel shows the kriging estimate, in the right panel $E^c_{0,0.05}(x)$ is superimposed in grey.
- The complement of $E^c_{u,\alpha}$ is the union of the pair of level avoiding sets.

Excursion functions

- The set $E^+_{u,\alpha}(x)$ does not provide any information about the locations not contained in the set.
- We want a visual tool similar to p-values (i.e. marginal probabilities), but which can be interpreted simultaneously.

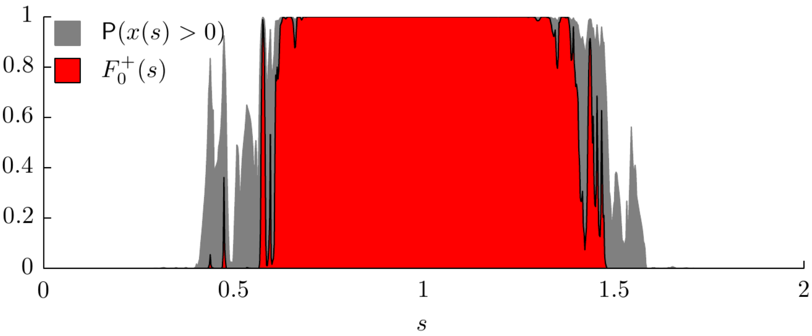
Excursion functions

The positive and negative u excursion functions, contour avoidance functions and the contour function are defined as

$$F_u^+(s) = \sup\{1-\alpha; s \in E^+_{u,\alpha}\}, \quad F_u^-(s) = \sup\{1-\alpha; s \in E^-_{u,\alpha}\},$$
$$F_u(s) = \sup\{1-\alpha; s \in E_{u,\alpha}\}, \quad F_u^c(s) = \sup\{\alpha; s \in E^c_{u,\alpha}\}.$$

Each set $E^\star_{u,\alpha}$ can be retrieved as the $1-\alpha$ excursion set of the function $F_u^\star(s)$

Example 1 (cont): Excursion functions



- $E^+_{u,\alpha}$ is retrieved as the $1-\alpha$ excursion set of $F^+_u(s)$.
- If the function takes a value close to one, the process likely exceeds the level at that location.
- If the value of the function is close to zero, it is more unlikely that the process exceeds the level at that location.

Outline

Calculating excursion sets in practise

- There are, in principle, two main problems that have to be solved in order to find the excursion sets.
    1. Probability calculation: e.g. calculate the probability $\mathsf{P}(D \subseteq A_u^+(x))$ for a given set D.
    2. Shape optimization: find the largest region D satisfying the required probability constraint.
- In practice it may not be computationally feasible to solve the problems separately since the probability calculation requires integration of the joint posterior density.
- We need a method that minimizes the number of probability calculations.
- One way of doing this is to use a parametric family for the possible excursion sets.

Parametric families for excursion sets

- The parametric families are based on the marginal quantiles of $x(s)$, $\mathsf{P}(x(s) \leq q_\rho(s)) = \rho$, which are easy to calculate.

One-parameter family

Let $q_\rho(s)$ be the marginal quantiles for $x(s)$, then a one-parameter family for the positive and negative u excursion sets is given by

$$D_1^+(\rho) = \{s; \mathsf{P}(x(s) > u) \geq 1-\rho\} = A_u^+(q_\rho),$$
$$D_1^-(\rho) = \{s; \mathsf{P}(x(s) < u) \geq 1-\rho\} = A_u^-(q_{1-\rho}).$$

- Using this parametric family reduces the complexity of the shape optimization to finding the correct value of ρ.
- Important: $D_1^\star(\rho_1) \subseteq D_1^\star(\rho_2)$ if $\rho_1 < \rho_2$.
- This simple one-parameter family can be extended in a number of ways, e.g. by smoothing the marginal quantiles.

Gaussian integrals

- For a Gaussian vector $\mathbf{x}$, the probabilities $\mathsf{P}(D \subseteq A_u^+(x))$, $\mathsf{P}(D \subseteq A_u^-(x))$, and $\mathsf{P}(D^+ \subseteq A_u^+(x), D^- \subseteq A_u^-(x))$ can all be written on the form

$$I(\mathbf{a}, \mathbf{b}, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \int_{\mathbf{a} \leq \mathbf{x} \leq \mathbf{b}} \exp(-\frac{1}{2}\mathbf{x}^\top \Sigma^{-1} \mathbf{x}) \, \mathrm{d}\mathbf{x},$$

- $\mathbf{a}$ and $\mathbf{b}$ are vectors depending on the mean value of $\mathbf{x}$, the domain D, and on u.
- There have been considerable research efforts devoted to approximating integrals of this form in recent years[1].
- For GMRFs, we want to use the sparsity of $\mathbf{Q}$.
- We use a method based on sequential importance sampling.

[1] A good introduction given in Genz and Bretz (2009), Computation of Multivariate Normal and t Probabilities, Lecture Notes in Statistics, Springer

A sequential Monte-Carlo algorithm

- a GMRF can be viewed as a non-homogeneous AR-process defined backwards in the indices of $\mathbf{x}$.
- Let L be the Cholesky factor of $\mathbf{Q}$, then

$$x_i|x_{i+1},\ldots,x_n \sim \mathsf{N}\left(\mu_i - \frac{1}{L_{ii}}\sum_{j=i+1}^{n} L_{ji}(x_j-\mu_j), L_{ii}^{-2}\right),$$

- Let I_i be the integral of the last $d-i$ components,

$$I_i = \int_{a_d}^{b_d} \pi(x_d)\int_{a_{d-1}}^{b_{d-1}} \pi(x_{d-1}|x_d)\cdots\int_{a_i}^{b_i} \pi(x_i|x_{i+1:d})\,\mathrm{d}x,$$

- $x_i|x_{i+1:d}$ only depends on the elements in $x_{\mathcal{N}_i\cap\{i+1:d\}}$.
- Estimate the integrals using sequential importance sampling.
- In each step x_j is sampled from the truncated Gaussian distribution $1(a_j < x_j < b_j)\pi(x_j|x_{j+1:d})$.
- The importance weights can be updated recursively.

Putting the pieces together

Calculating excursion sets using a one-parameter family

Assume that $\pi(\mathbf{x})$ is Gaussian and that $D(\rho)$ is a parametric family, such that $D(\rho_1) \subseteq D(\rho_2)$ if $\rho_1 < \rho_2$. The following strategy is then used to calculate $E_{u,\alpha}^{+}$.

- Choose a suitable (sequential) integration method.
- Reorder the nodes to the order they will be added to the excursion set when the parameter ρ is increased.
- sequentially add nodes to the set D and in each step update the probability $\mathsf{P}(D \subseteq A_u^{+}(x))$. Stop as soon as this probability falls below $1-\alpha$.
- $E_{u,\alpha}^{+}$ is given by the last set D for which $\mathsf{P}(D \subseteq A_u^{+}(x)) \geq 1-\alpha$.

Extension to a latent Gaussian setting

- The previous method can only be used in a purely Gaussian setting with known parameters.
- For the more general latent Gaussian setting, the posterior distribution can be written as

$$\pi(\mathbf{x}|\mathbf{y}) = \int \pi(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})\, \mathrm{d}\boldsymbol{\theta},$$

  where $\mathbf{y}$ is data and $\boldsymbol{\theta}$ the parameter vector.
- For Gaussian likelihoods, $\pi(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$ is Gaussian.
- There are a number of, more or less complex, ways we can extend the method to the latent Gaussian setting.
- The simplest is to use an empirical Bayes estimator where $\pi(\mathbf{x}|\mathbf{y})$ is replaced with $\pi_G(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}_0)$, a Gaussian approximation at the mode. Two more accurate methods are:
  - Quantile corrections
  - Numerical integration

Quantile Corrections

The QC method is based on modifying the integration limits in the Gaussian integrals based on the marginal posteriors.

- For each i, replace the lower limits a_i with $\tilde{a}_i = \sigma_i \Phi^{-1}(1 - \mathsf{P}(x_i > a_i|\mathbf{y}))$, where σ_i is the marginal standard deviation for $x_i|\mathbf{y}, \boldsymbol{\theta}_0$ and Φ denotes the standard Gaussian CDF.
- Similarly, the upper limits b_i are replaced with $\tilde{b}_i = \sigma_i \Phi^{-1}(\mathsf{P}(x_i < b_i|\mathbf{y}))$.
- One then has that $\mathsf{P}_G(x_i > \tilde{a}_i|\mathbf{y}, \boldsymbol{\theta}_0) = \mathsf{P}(x_i > a_i|\mathbf{y})$ and $\mathsf{P}_G(x_i < \tilde{b}_i|\mathbf{y}, \boldsymbol{\theta}_0) = \mathsf{P}(x_i < b_i|\mathbf{y})$, where $\mathsf{P}_G(\cdot|\mathbf{y}, \boldsymbol{\theta}_0)$ denotes the probability calculated under a Gaussian approximation of the posterior $\pi(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}_0)$.
- The QC method is exact if the components x_i are independent.

Numerical Integration

In the NI method, one numerically approximates the excursion function as $F_u^\bullet(\mathbf{s}) = \sum_{k=1}^K \lambda_k F_{u,k}^\bullet(\mathbf{s})$.

- Here $F_{u,k}^\bullet(\mathbf{s})$ is the level u excursion function calculated for the conditional posterior $\pi_G(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\theta}_k)$ for a fixed parameter configuration $\boldsymbol{\theta}_k$.
- The configurations $\boldsymbol{\theta}_k$ in the hyper parameter space can, for example, be chosen as in the INLA method and the weights λ_k are chosen proportional to $\pi(\boldsymbol{\theta}_k|\mathbf{y})$.
- Finally, the desired excursion set for a fixed α is retrieved as the excursion set $A_\alpha^+(F_u^\bullet)$ of the excursion function.

The NI method is more accurate than the QC method, but requires K times as many calculations.

Air pollution (PM_{10}) data

- The limit value fixed by the European directive 2008/50/EC for PM_{10} is $50\mu g/m^3$. The daily mean concentration cannot exceed this value more than 35 days in a year.
- A region where this value is periodically exceeded is the Piemonte region in northern Italy.
- Cameletti et al (2012/13)[2] investigated an SPDE/GMRF model for PM_{10} concentration in the region.
- The goal is to analyse exceedance probabilities of the limit value.
- Daily PM_{10} data measured at 24 monitoring stations during 182 days in the period October 2005 - March 2006.

[2] Cameletti, Lindgren, Simpson, and Rue (2012), Spatio-temporal modeling of particulate matter concentration through the SPDE approach, AStA

Model

- The following measurement equation is assumed,

$$y(\mathbf{s}_i, t) = x(\mathbf{s}_i, t) + \mathcal{E}(\mathbf{s}_i, t),$$

where $\mathcal{E}(\mathbf{s}_i, t) \sim \mathsf{N}(0, \sigma_{\mathcal{E}}^2)$ is Gaussian measurement noise, both spatially and temporally uncorrelated.

- $x(\mathbf{s}_i, t)$ is the latent field assumed to be on the form

$$x(\mathbf{s}_i, t) = \sum_{k=1}^{p} z_k(\mathbf{s}_i, t)\beta_k + \xi(\mathbf{s}_i, t),$$

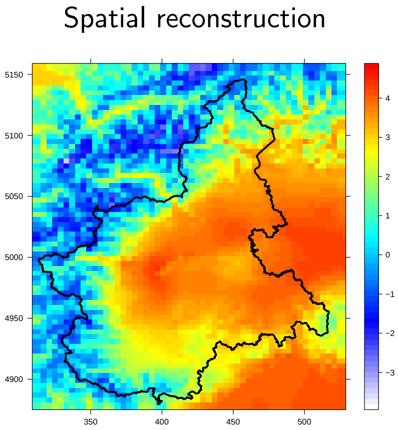
where the $p = 9$ covariates z_k are used.

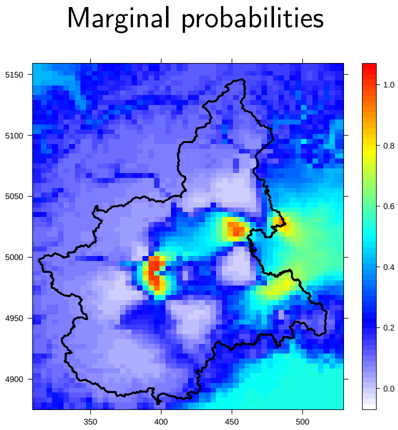
- ξ is assumed to follow first order AR-dynamics in time

$$\xi(\mathbf{s}_i, t) = a\xi(\mathbf{s}_i, t-1) + \omega(\mathbf{s}_i, t),$$

where $|a| < 1$ and $\omega(\mathbf{s}_i, t)$ is a zero-mean temporally independent Gaussian process with spatial Matérn covariances.

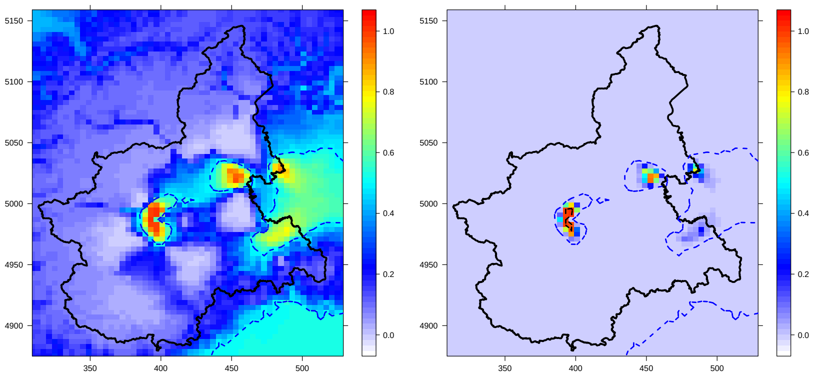
Results for January 30, 2006





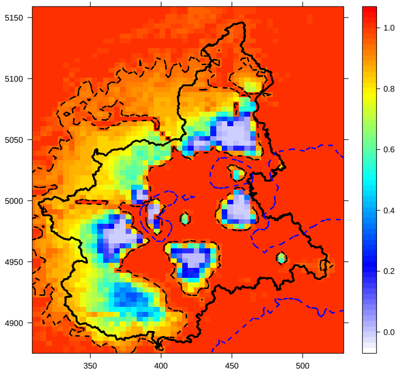
Results for January 30, 2006

Marginal probabilities $F_{50}^{+}(s)$

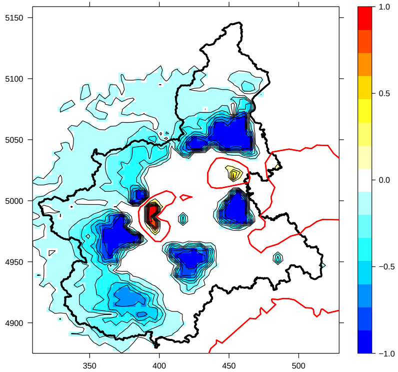


Results for January 30, 2006

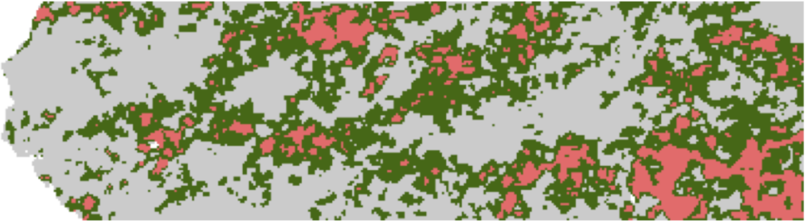
Contour function $F^{c}_{50}(s)$



Signed avoidance $\pm F_{50}(s)$



Further examples: Estimating vegetation increase

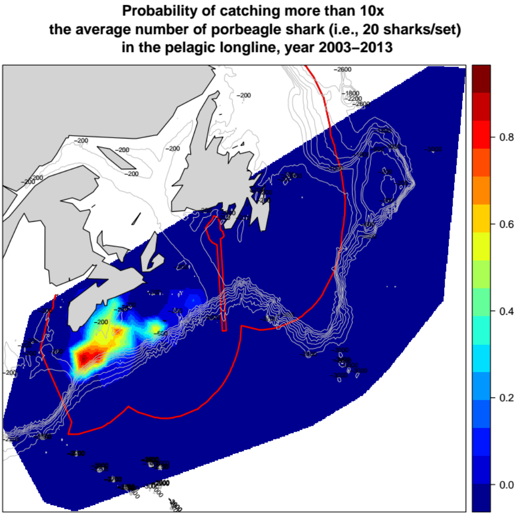


- Estimates of trends in vegetation in the western Sahel for the period 1983 - 1999.
- Marginally significant trends in green.
- Excursion set $E^{+}_{0,0.05}$ in red.
- There has been a vegetation increase in several parts of the region since the drought period in the early 1980s.

Further examples: activation regions in fMRI studies

Joint work with Yue, Lindquist, Lindgren, Simpson, and Rue.

Further examples: estimating bycatch hotspots



Joint work with Godin, Krainski, Worm, Flemming, and Campana.

Remarks

- Excursion sets and contour uncertainty regions are important in many applications.
- For latent Gaussian models, we can find these quantities efficiently.
- R package `excursions`, on CRAN:

```
excursions(alpha=0.05, u=0, type=">",
           mu=field.expectation, Q=precision.matrix)
excursions.inla(result.inla, ind=candidates,
                u=0, type="=", method="NI")
```

- Current and future developments.
  - For excursion sets, compare with other thresholding methods and a sample based method by French and Sain (2013).
  - For contour uncertainty sets, compare with the methods by Lindgren and Rychlik (1995).
  - Combine method with the work by Polfeldt (1999) to make quantitative statements about joint contour map reliability.

References

- Bolin, D. and Lindgren, F.: Excursion and contour uncertainty regions for latent Gaussian models; *JRSS Series B*, 2014, in press. Available on journal webpage: `http://onlinelibrary.wiley.com/doi/10.1111/rssb.12055/abstract`.
  CRAN package: `excursions`
- Cameletti, M., Lindgren, F., Simpson, D., and Rue, H.: Spatio-temporal modeling of particulate matter concentration through the SPDE approach; *AStA*, 2012
- French, J. P. and Sain, S. R.: Spatio-temporal exceedance locations and confidence regions. *Ann. Appl. Statist.*, 2013
- Genz, A. and Bretz, F.: Computation of Multivariate Normal and t Probabilities; *Lecture Notes in Statistics*, 195, Springer 2009
- Lindgren, G. and Rychlik, I.: How reliable are contour curves? Confidence sets for level contours, *Bernoulli*, 1995
- Polfeldt, T.: On the quality of contour maps, *Environmetrics*, 1999