Fórum Mineiro de Estatística e Probabilidade

Os 30 anos do Curso de Estatística da UFMG

Geostatistical analysis of compositional data: some models and applications

Paulo Justiniano Ribeiro Jr

LEG:Laboratório de Estatística e Geoinformação / UFPR

In collaboration with:
Ernesto Jardim (IPIMAR, Lisboa)
Ana Beatriz Tozzo Martins (UEM/LEG)
Wagner H Bonat (LEG/UFPR)

http://www.leg.ufpr.br
e-mail:paulojus@ufpr.br

Belo Horizonte, MG
20 – 22 de Maio de 2009

54^a RBRAS e 13^o SEAGRO

54^a Reunião da Região Brasileira da Sociedade Internacional de Biometria

& 13^o Simpósio de Estatística Aplicada a Agronomia

1. Região Brasileira da Sociedade Internacional de Biometria
2. 27 a 31 Julho 2009, São Carlos, SP
3. semana seguinte à Escola de Séries Temporais – ESTE (São Carlos)
4. com participação da ABE/Embrapa na programação
5. condições especiais para sócios RBRAS/ABE
6. http://www.rbras.org.br/rbras54

54^a RBRAS e 13^o SEAGRO

54^a Reunião da Região Brasileira da Sociedade Internacional de Biometria

& 13^o Simpósio de Estatística Aplicada a Agronomia

1. Região Brasileira da Sociedade Internacional de Biometria
2. 27 a 31 Julho 2009, São Carlos, SP
3. semana seguinte à Escola de Séries Temporais – ESTE (São Carlos)
4. com participação da ABE/Embrapa na programação
5. condições especiais para sócios RBRAS/ABE
6. http://www.rbras.org.br/rbras54

IBC-2010/Floripa e 55 RBRAS

International Biometrics Conference &

55^a **Reunião da Região Brasileira da Sociedade Internacional de Biometria**

1. Organization: IBS, Rbras, RArg
2. 05 a 10 december de 2009, Florianópolis, SC, Brasil
3. satelite events (opened to proposals)
4. RBRAS meeting on wednesday, 07/12
5. special fees for delegates from *special circunstance countries*
6. `http://www.rbras.org.br/ibc2010` and `http://www.tibs.org`

IBC-2010/Floripa e 55 RBRAS

International Biometrics Conference &

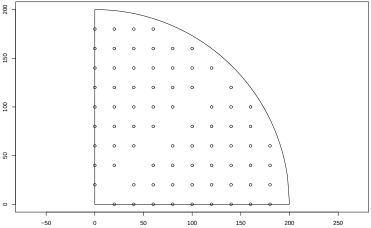
55^a **Reunião da Região Brasileira da Sociedade Internacional de Biometria**

1. Organization: IBS, Rbras, RArg
2. 05 a 10 december de 2009, Florianópolis, SC, Brasil
3. satelite events (opened to proposals)
4. RBRAS meeting on wednesday, 07/12
5. special fees for delegates from *special circunstance countries*
6. `http://www.rbras.org.br/ibc2010` and `http://www.tibs.org`

Outline

- Motivating examples
  - soil fractions
  - fish stocks and age structure
- Joint model for abundance and population structure
  - general setup, strategies and ingredients
  - compositional data analysis
  - geoestatistical analysis
  - results
- Multivariate model for compositional data
- Model specification
- Inference and Prediction
- computational implementation
- Final remarks

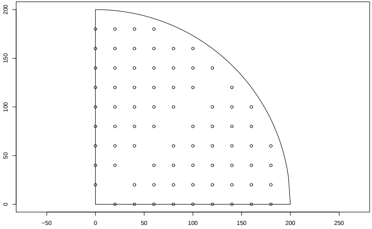
Motivating Examples I: soil fractions



- soil fractions determines management practices (fertilizers, irrigation, . . .)
- spatial description is essential to *precision agriculture*
- definition of soil classes
- typical data structure

X	Y	DenGlob	DenPar	PoroTot	Areia	Silte	Argila
20	0	1.69	2.63	35.61	31	25	44
40	0	1.58	2.87	45.08	24	21	55
60	0	1.44	2.55	43.51	23	27	50
80	0	1.50	2.57	41.48	22	25	53
100	0	1.58	2.56	38.31	22	25	53
120	0	1.45	2.69	46.05	16	25	59

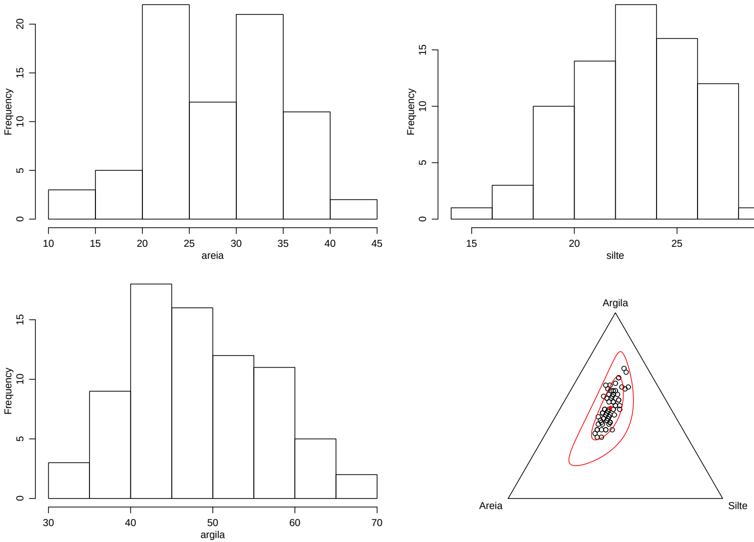
Motivating Examples I: soil fractions



- soil fractions determines management practices (fertilizers, irrigation, . . .)
- spatial description is essential to *precision agriculture*
- definition of soil classes
- typical data structure

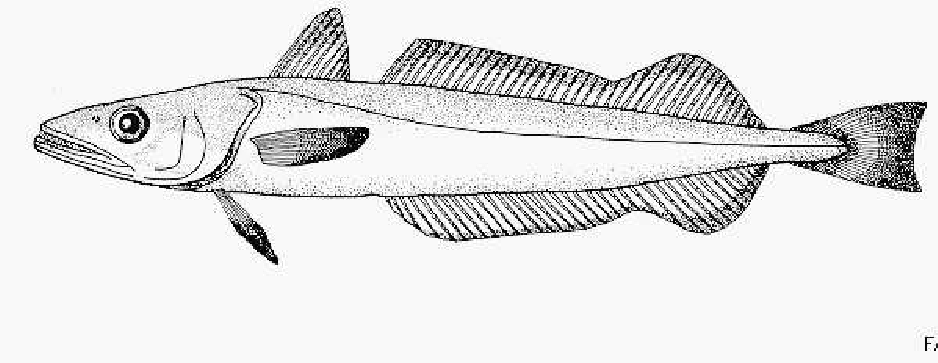
X	Y	DenGlob	DenPar	PoroTot	Areia	Silte	Argila
20	0	1.69	2.63	35.61	31	25	44
40	0	1.58	2.87	45.08	24	21	55
60	0	1.44	2.55	43.51	23	27	50
80	0	1.50	2.57	41.48	22	25	53
100	0	1.58	2.56	38.31	22	25	53
120	0	1.45	2.69	46.05	16	25	59

soil fractions (cont.)

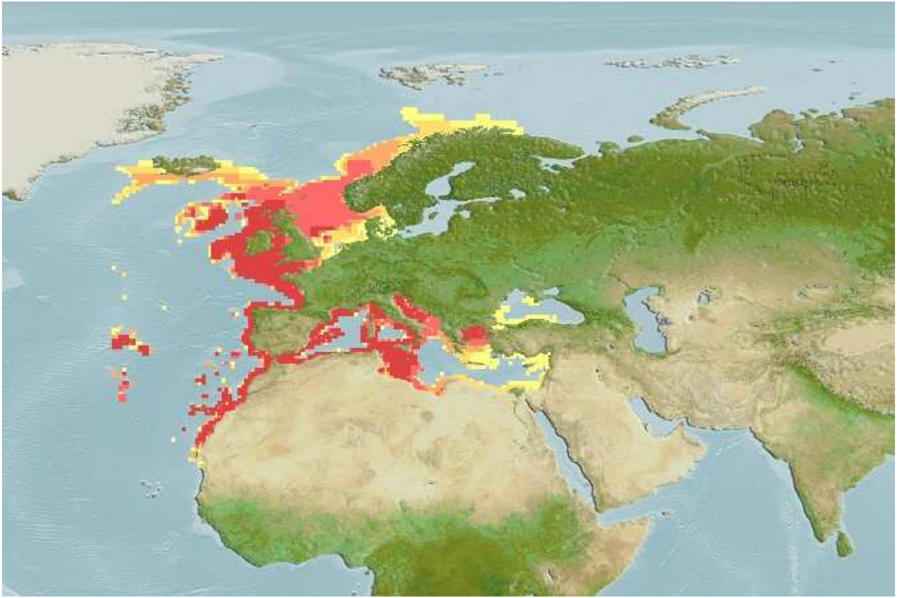


Motivating Examples II: fish stocks and age structure

Hake (*Merluccius merluccius*)



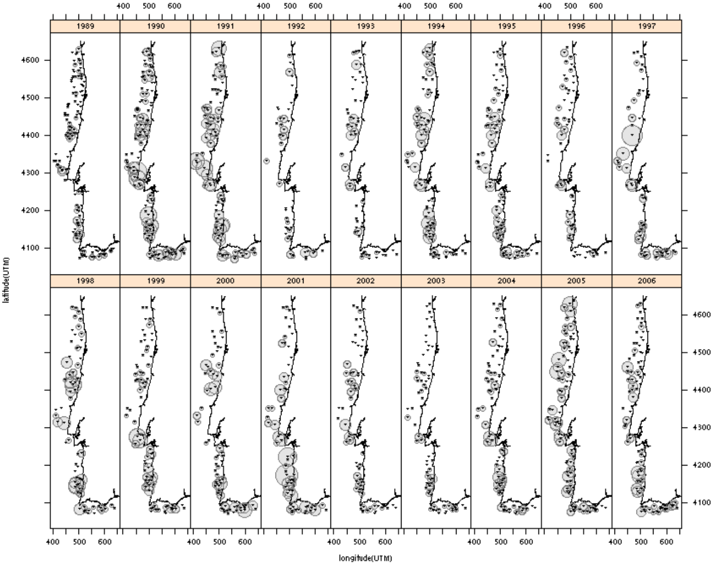
Hake distribution



fish stocks and age structure (cont.)

- assessment of fish stocks and population structure on the Portuguese coast;
- data from BTS (bottom trawl surveys);
- several target species, focus on *Hake*;
- related issues:
  - design (Jardim & Ribeiro Jr, 2007),
  - comparing strategies for stock assessment (Jardim & Ribeiro Jr, 2008),
  - spatial-temporal modelling (Silva & Ribeiro Jr, 2008),
  - wider context (Jardim, 2009): *MSE* (management stratagy evaluation).

Data from BTS



Goals

- General: modelling and prediction of abundance at age;
  - *abundance . . . :* global and local description of stocks
  - *. . . at age:* global and local population structure

- spatial-temporal estimates and predictions

- results as inputs for large simulation frameworks
  MSE (*management strategy evaluation*): scientific advice on fisheries and ecological management.

- confront results from design-based (mostly used) and model-based alternatives

Goals

- General: modelling and prediction of abundance at age;
  - *abundance . . . :* global and local description of stocks
  - *. . . at age:* global and local population structure
- spatial-temporal estimates and predictions
- results as inputs for large simulation frameworks
  MSE (*management strategy evaluation*): scientific advice on fisheries and ecological management.
- confront results from design-based (mostly used) and model-based alternatives

Goals

- General: modelling and prediction of abundance at age;
  - *abundance . . . :* global and local description of stocks
  - *. . . at age:* global and local population structure
- spatial-temporal estimates and predictions
- results as inputs for large simulation frameworks
  MSE (*management strategy evaluation*): scientific advice on fisheries and ecological management.
- confront results from design-based (mostly used) and model-based alternatives

Goals

- General: modelling and prediction of abundance at age;
  - *abundance . . . :* global and local description of stocks
  - *. . . at age:* global and local population structure
- spatial-temporal estimates and predictions
- results as inputs for large simulation frameworks
  MSE (*management strategy evaluation*): scientific advice on fisheries and ecological management.
- confront results from design-based (mostly used) and model-based alternatives

Statistical issues and ”ingredients”

- population structure: proportions of five age classes
  compositional data analysis (Aitchison, 1986)
- spatial structure of total stocks and abundances at age:
  geostatistics (. . . , . . . , . . . , Diggle & Ribeiro Jr, 2007)
- spatial variation of the proportions
  mixing of ”natural” and ”spurious” correlations
  geostatistical analysis of compositions (Pawlowsky-Glahn & Olea, 2004)
- *Bayesian/spatial/compositions* (Tjelmeland & Lund, 2003)
- Joint modelling of abundance **and** compositions
  spatio-temporal
  Jardim & Ribeiro Jr (submitted)

Some (loose) notation

- Primary data: (normalised) catches $c_{ijh}(x)$
  (i-year, j-age, h-haul, x-location)
- aggregated catch $y_{ih}(x) = \sum_j c_{ijh}(x)$
- The variables (dropping i, h and x indexes):
  Abundance at age C_j, proportion at age $P_j = C_j/Y$ and total abundance $Y = \sum_j C_j$
- A natural structure and modelling alternatives

$$[C_1, \ldots, C_m] = [P_1, \ldots, P_m|Y][Y]$$

  - multivariate model for $C's$, or ...
  - full parametric geostatistical model for Y
  - compositional data analysis for multinomial probabilities
- inference by simulation

Inference and Prediction

First Ingredient: compositional

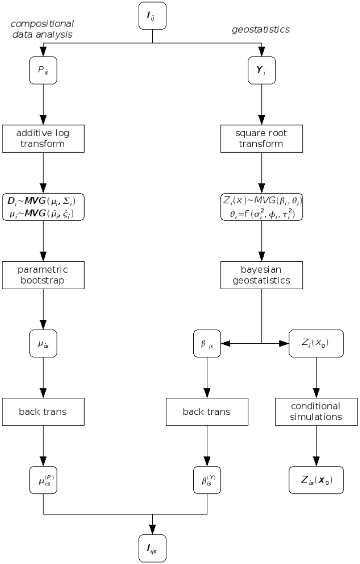
- $G_{ijh} = \log(P_{i,j\neq a,h}/P_{i,j=a,h})$
- $\mathbf{G}_{ij} \sim Gau(\mu_i, \Sigma_i)$, $\mathbf{G} = (G_1, \ldots, G_{D-1})$
- sampling mechanism for each year using estimates
- reference age class a: age 2 (greater abundance)
- 0's: multiplicative replacement strategy (Martín-Fernandez et al, 2003)

Inference and Prediction (cont.)

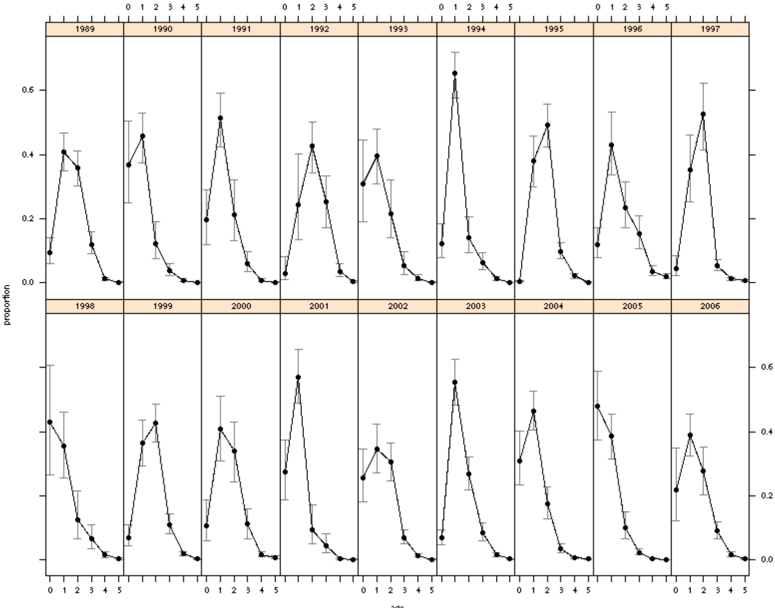
Second Ingredient: (Bayesian) geostatistics

- trans-Gaussian geostatistical model (Christensen et. al, 2001)
- $[g_\lambda(Y(x))|S(x)] \sim Gau(F(x)\beta + S(x), \epsilon)$
  - covariates $F(x)$
  - $S(\cdot)$ Gaussian process with covariance function $\sigma^2\rho(x) = \sigma^2 Corr\,[S(x), S(x')] = \sigma^2 \exp(-\|x - x'\|/\phi)$
  - $\epsilon \sim Gau(0, \tau^2)$
- Priors:
  - $[\beta, \sigma^1|\phi, \tau^2] \propto 1/\sigma^2$
  - $[\phi] \sim \exp(1/20)$
  - $[\tau^2]$ discrete (based on a ZIP) and truncated in $[0, 2]$

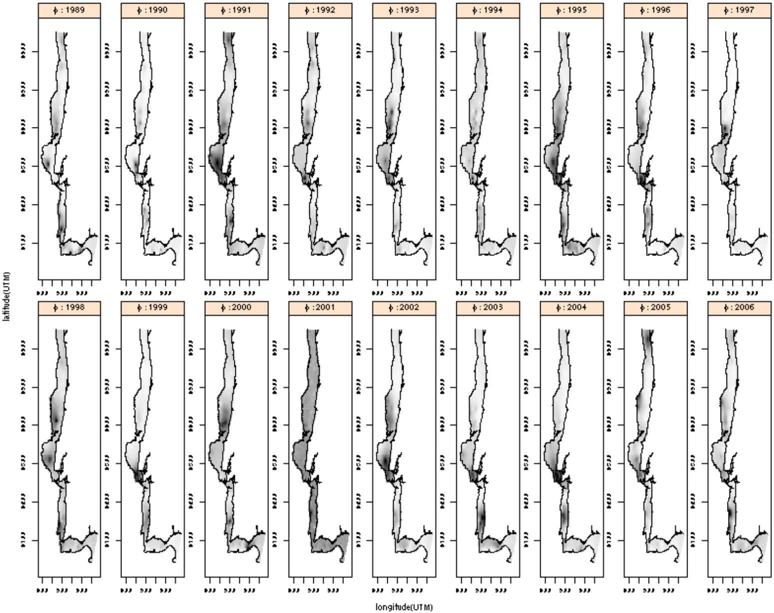
General scheme



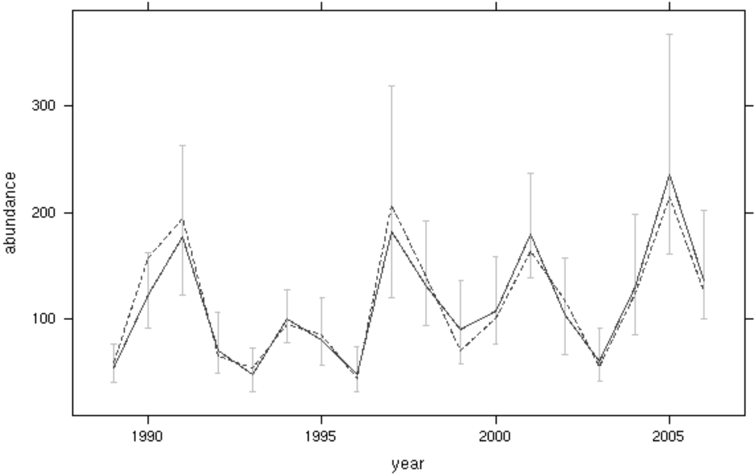
Yearly compositions



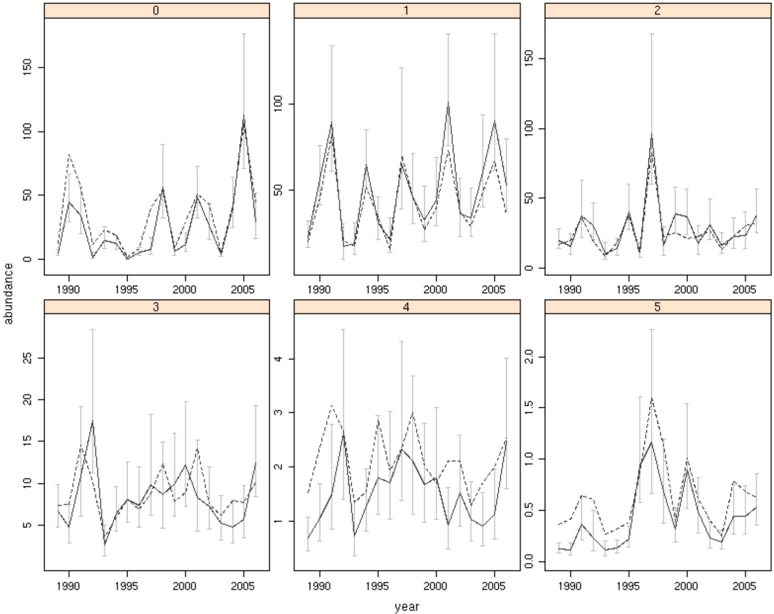
Yearly spatial predictions



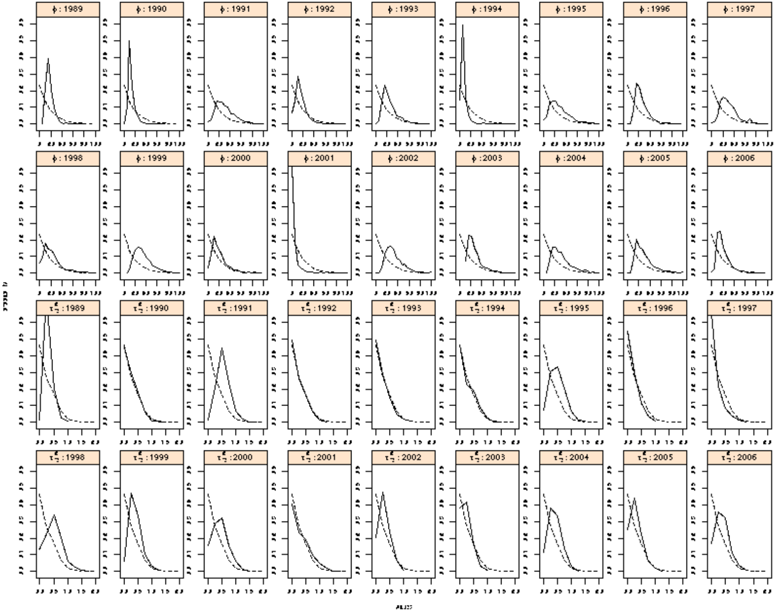
Yearly total abundance



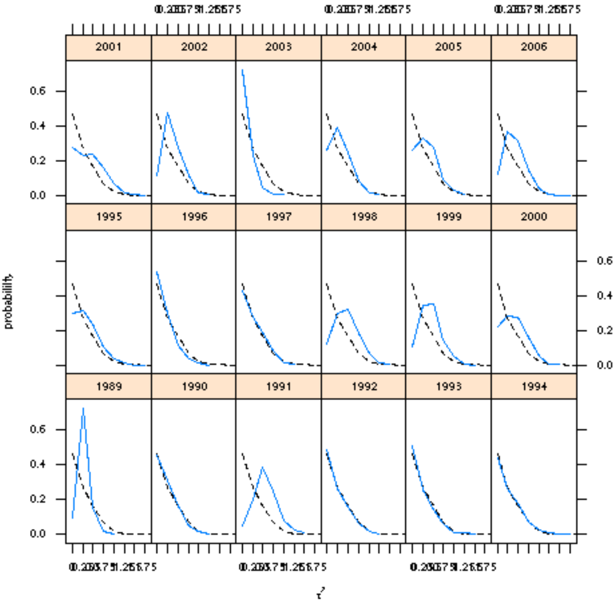
Yearly abundance at age



ϕ: priors and posteriors



τ^2: priors and posteriors



Remarks on the application

- differences between design based and model based estimates
- shifts: differences in abundances at age and/or classification errors
- spatial results highlights persistent areas of high abundance and recruitment spots
- allows identification of competition/cannibalism spots (e.g. recruits vs parental stocks)
- comments on interpretation of correlations from paper
- reasonable and consistent results for ϕ
- little information on data about $\tau_R^2 = \tau^2/\sigma^2$

Remarks on the application

- differences between design based and model based estimates
- shifts: differences in abundances at age and/or classification errors
- spatial results highlights persistent areas of high abundance and recruitment spots
- allows identification of competition/cannibalism spots (e.g. recruits vs parental stocks)
- comments on interpretation of correlations from paper
- reasonable and consistent results for ϕ
- little information on data about $\tau_R^2 = \tau^2/\sigma^2$

Further modelling

- $R(x)$: zero mean, unit variance Gaussian random field
- $Y(x) = \exp\{\mu + \sigma R(x) + \epsilon\}$
- For each age class consider $S_d(x) = \beta_{0d} + \beta_{1d} R(x) + \epsilon_d$
- $q_k(x) = \exp\{S_d(x)\}/(1 + \exp\{S_d(x)\})$
- compositions $p_d(x) = q_d(x)/\sum_d(q_d(x))$
- The abundance at age $C_d(x) = p_d(x) \cdot Y(x)$.
- Parameters:
  - $4 + 3D$ parameters.
  - common τ_d : $5 + 2D$ parameters
  - $\tau_D = \tau$: $4 + 2D$ parameters.

Example 2: soil fractions

Some Algebra for Compositional data

- Original sampling space
  $\mathbb{S}^B = \{\underset{\sim}{Y} \in \mathbb{R}^B;\ Y_i > 0,\ i = 1, ..., B;\ \underset{\sim}{j}'\underset{\sim}{Y} = 1\}$

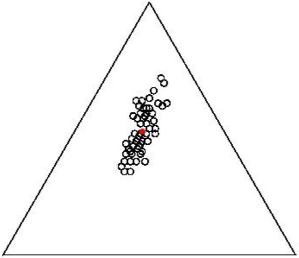
- Base: $\underset{\sim}{W}(\underset{\sim}{x}),\ \underset{\sim}{x} \in \Omega \subset \mathbb{R}^n$

- Transformed sampling space
  $\mathbb{R}^B_+ = \{\underset{\sim}{W}(\underset{\sim}{x}) \in \mathbb{R}^B;\ W_i(\underset{\sim}{x}) > 0,\ i = 1, ..., B\}$

- closure operator: Base $\Rightarrow$ composition

$$\begin{array}{rccl} \mathcal{C}: & \mathbb{R}^B_+ & \longrightarrow & \mathbb{S}^B \\ & \underset{\sim}{W}(\underset{\sim}{x}) & \longrightarrow & \mathcal{C}[\underset{\sim}{W}(\underset{\sim}{x})] = \dfrac{\underset{\sim}{W}(\underset{\sim}{x})}{\underset{\sim}{j}'\underset{\sim}{W}(\underset{\sim}{x})}, \end{array}$$

$\underset{\sim}{j}'$ vetor d



- Amalgamation, partition and sub-compositions

- Aditive log-ratio (ALR):

$$\begin{array}{rccl} \text{ALR}: & \mathbb{S}^B & \longrightarrow & \mathbb{R}^{B-1} \\ & \underset{\sim}{Y}(\underset{\sim}{x}) & \longrightarrow & \text{ALR}[\underset{\sim}{Y}(\underset{\sim}{x})] = \left(\ln \dfrac{Y_1(\underset{\sim}{x})}{Y_B(\underset{\sim}{x})}, \ldots, \ln \dfrac{Y_{B-1}(\underset{\sim}{x})}{Y_B(\underset{\sim}{x})}\right)'. \end{array}$$

- inverse additive generalised logistic
  ($agl = alr^{-1}$)

Analogous and alternative models

- Linear combination of spatial terms

$$S_j(x) = \sum_{j=1}^{d} a_{kj} U_j(x),$$

in matrix form: $S(x) = AU(x)$ with covariance function $\Gamma(x, x') = ARA'$,

- Linear model of corregionalisation

$$S(x) = \sum_{i=1}^{p} A_i U^i(x)$$

- Gelfand et. al. (Test, 2004): *Nonstationary multivariate process modeling through spatially varying coregionalization (with Discussion)*

Bivariate Gaussian Common Component Model (BGCCM)

- Common Components: $S(\cdot) = \{S_1(\cdot), S_2(\cdot)\}$ to be:

$$S_j(x) = S_0^*(x) + S_j^*(x) : j = 1, 2.$$

Por construção, $S(\cdot)$ é um processo válido com covariância:

$$\text{Cov}\{S_j(x), S_{j'}(x-u)\} = \gamma_0(u) + I(j = j')\gamma_j(u)$$

- BGCCM model:

$$Y_1(x_i) = \mu_1 + S_0(x_i) + S_1(x_i) = \mu_1 + \sigma_{01} R_0(x_i; \phi_0) + \sigma_1 R_1(x_i, \phi_1)$$

$$Y_2(x_j) = \mu_2 + S_0(x_j) + S_2(x_j) = \mu_2 + \sigma_{02} R_0(x_j; \phi_0) + \sigma_2 R_2(x_j, \phi_2)$$

Bivariate Gaussian Common Component Model (BGCCM)

- Common Components: $S(\cdot) = \{S_1(\cdot), S_2(\cdot)\}$ to be:

$$S_j(x) = S_0^*(x) + S_j^*(x) : j = 1, 2.$$

Por construção, $S(\cdot)$ é um processo válido com covariância:

$$\mathrm{Cov}\{S_j(x), S_{j'}(x-u)\} = \gamma_0(u) + I(j = j')\gamma_j(u)$$

- BGCCM model:

$$Y_1(x_i) = \mu_1 + S_0(x_i) + S_1(x_i) = \mu_1 + \sigma_{01} R_0(x_i; \phi_0) + \sigma_1 R_1(x_i, \phi_1)$$

$$Y_2(x_j) = \mu_2 + S_0(x_j) + S_2(x_j) = \mu_2 + \sigma_{02} R_0(x_j; \phi_0) + \sigma_2 R_2(x_j, \phi_2)$$

Bivariate Gaussian Common Component Model (BGCCM)

- Notation:
  - C (vector) compositions
  - Y (vector) observations on the additive log-ratio scale
  - Z (vector) standard MV-Gaussian
- Likelihood
  - density:

$$f(C) = (2\pi)^{(-d/2)}|\Sigma_Y|^{-1/2} \exp\left\{-0.5\,(alr(C) - \mu_Y)'\,\Sigma_Y^{-1}\,(alr(C) - \mu_Y)\right\}\left(\prod_{i=1}^{D} C_i\right)^{-1}$$

  - reparametrisation, concentrated likelihood
  - numerical methods, hessian, delta method, ...
- Prediction:
  - usual for geostatistical model on de Y scale based on multivariate-normal
  - particular issues on back-transforming

Bivariate Gaussian Common Component Model (BGCCM)

- Notation:
  - C (vector) compositions
  - Y (vector) observations on the additive log-ratio scale
  - Z (vector) standard MV-Gaussian
- Likelihood
  - density:

$$f(C) = (2\pi)^{(-d/2)}|\Sigma_Y|^{-1/2} \exp\left\{-0.5\left(alr(C)-\mu_Y\right)' \Sigma_Y^{-1}\left(alr(C)-\mu_Y\right)\right\}\left(\prod_{i=1}^{D} C_i\right)^{-1}$$

  - reparametrisation, concentrated likelihood
  - numerical methods, hessian, delta method, ...
- Prediction:
  - usual for geostatistical model on de Y scale based on multivariate-normal
  - particular issues on back-transforming

Back transforming predictions I

- Back-transforming to the original scale (proportions)

$$\mu_C = \mathrm{E}[C] = \int_{S^D} Cf(C)dC$$

$$\Sigma_C = \mathrm{Cov}[C, C] = \int_{S^D} (C - \mu_C)(C - \mu_C)' f(C)dC$$

$$f(C) = (2\pi)^{(-d/2)} |\Sigma_Y|^{-1/2} \exp\left\{-0.5\,(alr(C) - \mu_Y)'\,\Sigma_Y^{-1}\,(alr(C) - \mu_Y)\right\} \left(\prod_{i=1}^{D} C_i\right)^{-1}$$

- Gauss-Hermite integration:

$$\mu_C = \int_{\mathbb{R}^D} g_1(Z)\exp\{-Z'Z\}dZ \quad \Sigma_C = \int_{\mathbb{R}^D} g_2(Z)\exp\{-Z'Z\}dZ$$

$$Z'Z = 0.5\,(alr(C) - \mu_Y)'\,\Sigma_Y^{-1}\,(alr(C) - \mu_Y)$$

$$\Sigma_Y = R'R$$

$$C = agl\left(\sqrt{2}R'Z + \mu_Y\right)$$

$$\mu_C = \int_{\mathbb{R}^D} \pi^{-(D-1)/2} agl\left(\sqrt{2}R'Z + \mu_Y\right)\exp\{-Z'Z\}dZ$$

$$\Sigma_C = \int_{\mathbb{R}^D} \pi^{-d/2)} MM' \exp\{-Z'Z\}dZ \quad ; \quad M = agl\left(\sqrt{2}R'Z + \mu_Y\right) - \mu_C$$

Back transforming predictions I

- Back-transforming to the original scale (proportions)

$$\mu_C = \mathrm{E}[C] = \int_{S^D} Cf(C)dC$$

$$\Sigma_C = \mathrm{Cov}[C, C] = \int_{S^D} (C - \mu_C)(C - \mu_C)' f(C)dC$$

$$f(C) = (2\pi)^{(-d/2)} |\Sigma_Y|^{-1/2} \exp\left\{-0.5\,(alr(C) - \mu_Y)'\,\Sigma_Y^{-1}\,(alr(C) - \mu_Y)\right\} \left(\prod_{i=1}^{D} C_i\right)^{-1}$$

- Gauss-Hermite integration:

$$\mu_C = \int_{\mathbb{R}^D} g_1(Z) \exp\{-Z'Z\} dZ \quad \Sigma_C = \int_{\mathbb{R}^D} g_2(Z) \exp\{-Z'Z\} dZ$$

$$Z'Z = 0.5\,(alr(C) - \mu_Y)'\,\Sigma_Y^{-1}\,(alr(C) - \mu_Y)$$

$$\Sigma_Y = R'R$$

$$C = agl\left(\sqrt{2}R'Z + \mu_Y\right)$$

$$\mu_C = \int_{\mathbb{R}^D} \pi^{-(D-1)/2} agl\left(\sqrt{2}R'Z + \mu_Y\right) \exp\{-Z'Z\} dZ$$

$$\Sigma_C = \int_{\mathbb{R}^D} \pi^{-d/2)} MM' \exp\{-Z'Z\} dZ \quad ; \quad M = agl\left(\sqrt{2}R'Z + \mu_Y\right) - \mu_C$$

Back transforming predictions I

- Back-transforming to the original scale (proportions)

$$\mu_C = \mathrm{E}[C] = \int_{S^D} Cf(C)dC$$

$$\Sigma_C = \mathrm{Cov}[C, C] = \int_{S^D} (C - \mu_C)(C - \mu_C)' f(C)dC$$

$$f(C) = (2\pi)^{(-d/2)} |\Sigma_Y|^{-1/2} \exp\left\{-0.5\,(alr(C) - \mu_Y)'\,\Sigma_Y^{-1}\,(alr(C) - \mu_Y)\right\} \left(\prod_{i=1}^{D} C_i\right)^{-1}$$

- Gauss-Hermite integration:

$$\mu_C = \int_{\mathbb{R}^D} g_1(Z) \exp\{-Z'Z\} dZ \qquad \Sigma_C = \int_{\mathbb{R}^D} g_2(Z) \exp\{-Z'Z\} dZ$$

$$Z'Z = 0.5\,(alr(C) - \mu_Y)'\,\Sigma_Y^{-1}\,(alr(C) - \mu_Y)$$

$$\Sigma_Y = R'R$$

$$C = agl\left(\sqrt{2}R'Z + \mu_Y\right)$$

$$\mu_C = \int_{\mathbb{R}^D} \pi^{-(D-1)/2} agl\left(\sqrt{2}R'Z + \mu_Y\right) \exp\{-Z'Z\} dZ$$

$$\Sigma_C = \int_{\mathbb{R}^D} \pi^{-d/2)} MM' \exp\{-Z'Z\} dZ \quad ; \quad M = agl\left(\sqrt{2}R'Z + \mu_Y\right) - \mu_C$$

Back transforming predictions II

Alternative (simulation)

- sample $Y_s(x)$ from $Y(x)$ predictive distribution
- $C_s(x) = agl(Y_s(x)) : (Y_{s1}(x), \ldots, Y_{s(D-1)}(x), 0) \rightarrow (C_{s1}(x), \ldots, C_{sD}(x))$

Computacional illustration

Back transforming predictions II

Alternative (simulation)

- sample $Y_s(x)$ from $Y(x)$ predictive distribution
- $C_s(x) = agl(Y_s(x)) : (Y_{s1}(x), \ldots, Y_{s(D-1)}(x), 0) \rightarrow (C_{s1}(x), \ldots, C_{sD}(x))$

Computacional illustration

Computational details

- code for fish data analysis at LEG's *paper companions* web page
- code for soil data available at
  geoComp "pre-pre-pre-alpha"R-package (to be added to geoR)
- more general number of compositions
- possible integration with otehr packages sp, INLA, RandomFields, spBayes

Final remarks

- Fish model (hopefully not *fishy*)
  - avoids multivariate models by joint modelling of total and proportions
  - suggests how population structure evolves over time (however with a naïve temporal structure)
  - extended for count (neg. binomial) data
  - needs more general and coherent inferential setup
  - separation of the correlations (spatial and compositional)
  - in summary ... this is just towards a model ...
- Geostatistical models for compositional data
  - More efficient computation
  - Bayesian inference being developed/implemented
  - issues on incorporating covariates
  - more general specification of multivariate models
  - **better understanding of multivariate models**

Final remarks

- Fish model (hopefully not *fishy*)
  - avoids multivariate models by joint modelling of total and proportions
  - suggests how population structure evolves over time (however with a naïve temporal structure)
  - extended for count (neg. binomial) data
  - needs more general and coherent inferential setup
  - separation of the correlations (spatial and compositional)
  - in summary . . . this is just towards a model . . .
- Geostatistical models for compositional data
  - More efficient computation
  - Bayesian inference being developed/implemented
  - issues on incorporating covariates
  - more general specification of multivariate models
  - **better understanding of multivariate models**

Final remarks

Muito Obrigado!

(um privilégio estar aqui!)