Smooth-CAR mixed models
for spatial count data

María Durbán
Universidad Carlos III de Madrid
Department of Statistics

IBC Dublin 2008
Outline

1. Motivation
2. Penalized splines for spatial data
3. Smooth-CAR model
4. Application: Scottish Lip Cancer Data
5. Conclusions
Motivation

Main interest:

- Study the spatial variation by modelling the correlation structure

Most usual approaches:

- **Geostatistical data:** *Kriging* Methods, Random Fields or Geoadditive Models.

- **Regional/Areal data:** Conditional Autoregressive Models (*CAR*) from a hierarchical Bayes approach.

Our proposal:

- Model spatial regional count data (generally Poisson distributed)
- *P*-splines as mixed models (spatial random effect models).
- "Hybrid model": Smooth-CAR.

To appear "Computational Statistics and Data Analysis".

M. Durban (UC3M)
Motivation

Main interest:

- Study the spatial variation by modelling the correlation structure

Most usual approaches:

- Geostatistical data: Kriging Methods, Random Fields or Geoadditive Models.
- Regional/Areal data: Conditional Autoregressive Models (CAR) from a hierarchical Bayes approach.

Our proposal:

- Model spatial regional count data (generally Poisson distributed)
- $P$-splines as mixed models (spatial random effect models).
- “Hybrid model”: Smooth-CAR.
- To appear “Computational Statistics and Data Analysis”.

M. Durban (UC3M)
1. Motivation

- **Spatial data**

- Response variable, $y_{ij}$
  - measured over geographical locations, $s = (x_i, x_j)$, with $i, j = 1, \ldots, n$
1. Motivation

- **Spatial data**

  - Response variable, $y_{ij}$
    - measured over *geographical locations*, $s = (x_i, x_j)$, with $i, j = 1, ..., n$

- **ISSUE**: huge amount of data available
  - e.g.: Environmental data, epidemiologic studies, disease mapping applications, ...
1. Motivation

- **Spatial data**

- Response variable, $y_{ij}$
  - measured over geographical locations, $s = (x_i, x_j)$, with $i, j = 1, \ldots, n$

- **ISSUE:** huge amount of data available
  - e.g. Environmental data, epidemiologic studies, disease mapping applications, ...

- **Smoothing techniques:**
  - Study spatial trends.
  - Local correlation.
  - “Penalized Splines” (Eilers and Marx, 1996).
2. Penalized splines

▶ “The flexible smoother”

- Methodology:

  ▶ Given the data \((x_i, y_i), i = 1, \ldots, n\).
  ▶ Fit a **sum of local basis functions**: \(f(x_i) = B\theta\)
  ▶ Minimize the **Penalized Sum of Squares**:

\[
\|y_i - f(x_i)\|^2 + \text{Penalty}
\]
2. Penalized splines

- “The flexible smoother”

- Methodology:
  - Given the data \((x_i, y_i), i = 1, \ldots, n\).
  - Fit a **sum of local basis functions**: \(f(x_i) = B\theta\)
  - Minimize the **Penalized Sum of Squares**:
    \[
    \|y_i - f(x_i)\|^2 + \text{Penalty}
    \]
  - The **Penalty** controls the smoothness of the fit.
    - **Smoothing parameter**: \(\lambda\)
    - Apply a **discrete penalty** over coefficients \(\theta\), e.g. in 1d:
      \[
      P = \lambda D' D
      \]
      where \(D\) is a difference matrix acting on \(\theta\).
2. Penalized splines
  ▶ “The flexible smoother”

  • For array data (Currie et al., 2006):
    ▶ Generalized Linear Array Methods (GLAM):
      \[ f(x_1, \ldots, x_d) = B\theta \]
    ▶ where \( B \) is the Kronecker product of \( d \) \( B \)-splines basis:
      \[ B = B_1 \otimes B_2 \otimes \ldots \otimes B_d \]
    ▶ Efficient Algorithms for smoothing on multidimensional grids (e.g. mortality data, images, etc...).
    ▶ Easy representation as a Mixed Model:
      \[ f(x_1, \ldots, x_d) = X\beta + Z\alpha \]
• 2d **Penalty matrix:**

  - Set penalties over the 2d-array \( A \):

    \[
    P = \lambda_1 D_1^T D_1 \otimes I_{c_2} + \lambda_2 I_{c_1} \otimes D_2^T D_2
    \]
    
    row-wise                                      column-wise

  - For **spatial data:**

    \[
    f(\text{longitude, latitude})
    \]
    
    Space

    ✓ **Spatial anisotropy** (\( \lambda_1 \neq \lambda_2 \)), different amount of smoothing for latitude and longitude.

    ✓ However spatial data are **not over a regular grid**.
2. Penalized splines

- **Scattered data smoothing**

  - For scattered data, Eilers et al. (2006), propose:
    - “Row-wise Kronecker” product or **Box-Product** of $B$-spline basis.

  **Def. Box-Product:**

  \[
  B_1 \bigotimes B_2 = (B_1 \bigotimes 1'_{c_2}) \circ (1'_c \bigotimes B_2)
  \]

  where $\circ$ is the element-wise product.

  - We propose the use of $\bigotimes$ for **spatial data**:
    - Although spatial data are not over a grid,
    - the coefficients $\theta$ can be expressed in array form.
    - Choose a moderate number of knots to cover the spatial domain.
2. Penalized splines

- **Mixed Models representation**

  - Reparameterize the basis $\mathbf{B}$ and coefficients $\theta$:

    $$\mathbf{B}\theta = \mathbf{X}\beta + \mathbf{Z}\alpha$$

  - Currie et al. (2006), use the *Singular Value Decomposition (SVD)* over the Penalty $\mathbf{P}$, i.e.:

    $$\mathbf{D}'\mathbf{D} = \begin{bmatrix} \mathbf{U}_n : \mathbf{U}_s \end{bmatrix} \begin{bmatrix} \mathbf{0}_q & \tilde{\Sigma} \\ \tilde{\Sigma} & \mathbf{U}'_n \\ \mathbf{U}'_s \end{bmatrix}$$

  - The **Penalty** becomes *blockdiagonal*, $\mathbf{F} = \lambda \tilde{\Sigma}$

  - Standard mixed model theory (*REML*)
**P-splines for spatial count data**

In many applications:

- Collect count data **observed in regions or areas**.
  - E.g.: # of cases of disease or deaths
- Counts are **Poisson** distributed.

\[
y \sim \mathcal{P}(\mu) \quad \mathbb{E}[y] = \text{Var}[y] = \mu
\]

Real data:

\[
\mathbb{E}[y] \neq \text{Var}[y] = \mu
\]
\textbf{P-splines for spatial count data}

In many applications:

- Collect count data \textbf{observed in regions or areas}.
  - \textbf{E.g.}: \# of cases of disease or deaths
- Counts are \textbf{Poisson} distributed.

\[ y \sim \mathcal{P}(\mu) \quad \mathbb{E}[y] = \text{Var}[y] = \mu \]

\textbf{Real data:}

\[ \mathbb{E}[y] \neq \text{Var}[y] = \mu \]

\textbf{Alternatives to modelling overdispersion:}

1. Add extra parameters to the linear predictor for each observation (\textit{random effects model}).

2. Assume some more general form for the variance:
   - Gamma-Poisson mixture (\textit{Negative Binomial distribution}).
P-splines for spatial count data

Perperoglou and Eilers (2007), Model:

$$\eta = B\theta + \gamma I, \quad \gamma \sim N(0, \kappa^{-1}I)$$

- "Penalized Random Individual Dispersion Effects" (PRIDE).
**P-splines for spatial count data**

Perperoglou and Eilers (2007), Model:

\[ \eta = \mathbf{B}\theta + \gamma \mathbf{l}, \quad \gamma \sim \mathcal{N}(0, \kappa^{-1}\mathbf{I}) \]

- “Penalized Random Individual Dispersion Effects” (PRIDE).

**in Spatial data context:**

2d PRIDE model as a mixed model

\[ \eta = \mathbf{X}\beta + \mathbf{Z}\alpha + \gamma \mathbf{l}, \quad \alpha \sim \mathcal{N}(0, \mathbf{G}), \quad \gamma \sim \mathcal{N}(0, \kappa^{-1}\mathbf{I}) \]

- \( \gamma \) is “spatial random effect” associated to each of the \( n \) areas.
- \( \mathbf{X} \) and \( \mathbf{Z} \) are the mixed model matrices for the spatial case.
**P-splines for spatial count data**

### Spatial count data regression models:

<table>
<thead>
<tr>
<th>Model</th>
<th>log Link ($\eta$)</th>
<th>Inv. Link ($\mu$)</th>
<th>Weight matrix</th>
<th>Overdisp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poisson</td>
<td>$\eta = X\beta + Z\alpha$</td>
<td>$\mu = e^\eta$</td>
<td>$W = \text{diag}(\mu)$</td>
<td>—</td>
</tr>
<tr>
<td>PRIDE</td>
<td>$\eta = X\beta + Z\alpha + \gamma I$</td>
<td>$\mu = e^\eta$</td>
<td>$W^* = \frac{\kappa \text{diag}(\mu)}{\text{diag}(\mu) + \kappa I}$</td>
<td>$\gamma \sim \mathcal{N}(0, \kappa^{-1}I)$</td>
</tr>
</tbody>
</table>

✓ **Note:**

- $X\beta + Z\alpha \implies$ is the spatial $P$-splines for $x_1 =$ “longitude” and $x_2 =$ “latitude”.
Smooth-CAR model

Formulation:

\[ y = X\beta + b, \]

where \( b = (b_1, b_2, \ldots, b_n)' \) is a vector for the spatial effects.
Smooth-CAR model

Formulation:

\[ y = X\beta + b, \]

where \( b = (b_1, b_2, ..., b_n)' \) is a vector for the \textbf{spatial effects}

- Impose a \textbf{spatial dependency structure} by a prior distribution for \( b \):

  \[ b \sim \mathcal{N}(0, G_b) \]

\( G_b \) depends on the \textbf{"neighbourhood structure"} defined by \textbf{Contiguity matrix} (\( Q \))
Smooth-CAR model

Formulation:

\[ y = X\beta + b, \]

where \( b = (b_1, b_2, \ldots, b_n)' \) is a vector for the spatial effects

- Impose a **spatial dependency structure** by a prior distribution for \( b \):
  \[ b \sim \mathcal{N}(0, G_b) \]

\( G_b \) depends on the **“neighbourhood structure”** defined by **Contiguity matrix** (Q)

- \( Q = \{q_{i,j}\} \) is a \( n \times n \) matrix where
  \[ q_{i,j} = \begin{cases} 
  -1 & \text{if } i^{th} \text{ and } j^{th} \text{ regions are neighbours} \\
  0 & \text{otherwise} 
  \end{cases} \]

the diagonal elements \( q_{i,i} \) contains the number of neighbours of \( i^{th} \) region.
Smooth-CAR model

- **CAR Models**

✓ We follow an **Empirical Bayes** approach:

**Intrinsic CAR:**

\[
G_b = \sigma_b^2 Q^- + \kappa^{-1} I
\]  
(Besag, 1991)

- Two independent and separate variance components:
  
  ▶ **Spatially-structured variation:** \( \sigma_b^2 Q^- \)
  
  ▶ **Unstructured non-spatial correlation:** \( \kappa^{-1} I \)
Smooth-CAR model

- CAR Models

✓ We follow an Empirical Bayes approach:

Intrinsic CAR:

\[ G_b = \sigma_b^2 Q^{-} + \kappa^{-1} I \]  

(Besag, 1991)

- Two independent and separate variance components:
  - Spatially-structured variation: \( \sigma_b^2 Q^{-} \)
  - Unstructured non-spatial correlation: \( \kappa^{-1} I \)

Alternative CAR models structures:

\[ G_b = \sigma_b^2 (\phi Q^{-} + (1 - \phi)I)^{-1} \]  

(Leroux et al, 1999)

\[ G_b = \sigma_b^2 (\phi Q^{-} + (1 - \phi)I) \]  

(Dean et al, 2001)

- \( \phi \) measures the relative weight between *structured* and *unstructured* variability
- \( 0 \leq \phi \leq 1 \)
Smooth-CAR model

We propose a “hybrid” model:

- **Spatial $P$-spline** with **CAR** structure: “Smooth-CAR” model
- **Model:**
  \[ \eta = X\beta + Z\alpha + b, \quad b \sim \mathcal{N}(0, G_b) \]
Smooth-CAR model

We propose a “hybrid” model:

- Spatial $P$-spline with CAR structure: “Smooth-CAR” model
- Model:
  \[ \eta = X\beta + Z\alpha + b, \quad b \sim \mathcal{N}(0, G_b) \]

Our approach:

\[ \eta = \underbrace{\text{Spatial Trend}}_{X\beta + Z\alpha} + \underbrace{\text{Local area-level spatial correlation}}_{\text{Spatial Random Effects}} \]

(Large-scale)  
(Small-scale)
Smooth-CAR model

Summary:

<table>
<thead>
<tr>
<th>Model</th>
<th>Linear Predictor</th>
<th>Area-level var.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poisson</td>
<td>$X\beta + Z\alpha$</td>
<td>$-$</td>
</tr>
<tr>
<td>PRIDE</td>
<td>$X\beta + Z\alpha + \gamma I$</td>
<td>$\gamma \sim \mathcal{N}(0, \kappa^{-1}I)$</td>
</tr>
<tr>
<td>CAR</td>
<td>$X\beta + b$</td>
<td>$b \sim \mathcal{N}(0, G_b)$</td>
</tr>
<tr>
<td>Smooth-CAR</td>
<td>$X\beta + Z\alpha + b$</td>
<td>$b \sim \mathcal{N}(0, G_b)$</td>
</tr>
</tbody>
</table>
# Smooth-CAR model

## Summary:

<table>
<thead>
<tr>
<th>Model</th>
<th>Linear Predictor</th>
<th>Area-level var.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poisson</td>
<td>$X\beta + Z\alpha$</td>
<td>$-\gamma \sim \mathcal{N}(0, \kappa^{-1}I)$</td>
</tr>
<tr>
<td>PRIDE</td>
<td>$X\beta + Z\alpha + \gamma I$</td>
<td>$\gamma \sim \mathcal{N}(0, \kappa^{-1}I)$</td>
</tr>
<tr>
<td>CAR</td>
<td>$X\beta + b$</td>
<td>$b \sim \mathcal{N}(0, G_b)$</td>
</tr>
<tr>
<td>Smooth-CAR</td>
<td>$X\beta + Z\alpha + b$</td>
<td>$b \sim \mathcal{N}(0, G_b)$</td>
</tr>
</tbody>
</table>

### The Smooth-CAR:

- Let us model the **spatial trend** ($X\beta + Z\alpha$) along large geographical distances and
- **Local area-level** correlation by a **CAR** component ($b$).
- **Note:** for Leroux and Dean’s **CAR**, when $\phi = 0$, **Smooth-CAR** reduces to **PRIDE**.
Application: Scottish Lip Cancer Data

Example: Scottish Lip Cancer

Clayton and Kaldor (1987)

Breslow and Clayton (1993)

Observed and Expected cases of lip cancer

56 counties in Scotland

Application: Scottish Lip Cancer data

Fitted Models

- Smooth \( P \)-splines models:

\[
\eta = \log(e) + X\beta + Z\alpha \\
\eta = \log(e) + X\beta + Z\alpha + \gamma I
\]

(Poisson)  
(PRIDE)

\( \log(e) \) is the offset term.
Application: Scottish Lip Cancer data

Fitted Models

- **Smooth \( P \)-splines models:**
  \[
  \eta = \log(e) + X\beta + Z\alpha \quad \text{(Poisson)}
  \]
  \[
  \eta = \log(e) + X\beta + Z\alpha + \gamma l \quad \text{(PRIDE)}
  \]
  \( \log(e) \) is the **offset** term.

- **CAR models:**
  \[
  \eta = \log(e) + X\beta + b, \quad b \sim \mathcal{N}(0, G_b),
  \]
  with:
  \[
  G_b = \sigma_b^2 Q^- + \kappa^{-1} l \quad \text{(Besag)}
  \]
  \[
  G_b = \sigma_b^2 (\phi Q^- + (1 - \phi) l)^{-1} \quad \text{(Leroux)}
  \]
  \[
  G_b = \sigma_b^2 (\phi Q^- + (1 - \phi) l) \quad \text{(Dean)}
  \]
Application: Scottish Lip Cancer data

Fitted Models

- **Smooth \( P \)-splines models:**

  \[
  \eta = \log(e) + X\beta + Z\alpha \quad \text{(Poisson)}
  \]

  \[
  \eta = \log(e) + X\beta + Z\alpha + \gamma I \quad \text{(PRIDE)}
  \]

  \( \log(e) \) is the offset term.

- **CAR models:**

  \[
  \eta = \log(e) + X\beta + b, \quad b \sim \mathcal{N}(0, G_b),
  \]

  with:

  \[
  G_b = \sigma_b^2 Q^{-1} + \kappa^{-1} I \quad \text{(Besag)}
  \]

  \[
  G_b = \sigma_b^2 (\phi Q + (1 - \phi) I)^{-1} \quad \text{(Leroux)}
  \]

  \[
  G_b = \sigma_b^2 (\phi Q^{-1} + (1 - \phi) I) \quad \text{(Dean)}
  \]

- **Smooth-CAR model:**

  \[
  \eta = \log(e) + X\beta + Z\alpha + b, \quad b \sim \mathcal{N}(0, G_b)
  \]
## Application: Scottish Lip Cancer data

### Comparison of fitted models

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameters</th>
<th>AIC</th>
<th>BIC</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smooth: Poisson</td>
<td>$\lambda_1$</td>
<td>11.75</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\lambda_2$</td>
<td>3.63</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\sigma^2_s$</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\kappa^{-1}$</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\phi$</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Smooth-CAR:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dean</td>
<td>$\lambda_1$</td>
<td>30.11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\lambda_2$</td>
<td>16.37</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\sigma^2_s$</td>
<td>0.53</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\kappa^{-1}$</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\phi$</td>
<td>0.97</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Observations:

- $\phi \approx 1 \rightarrow$ Overdispersion is due to “structured” spatial correlation ($\sigma^2_b Q^{-1}$).
- **Smooth-CAR** performs better in terms of the selected criteria.
Dean’s CAR model:

(a) Large-scale linear trend: $X\beta$

(b) CAR structured random effects: $b \sim \mathcal{N}(0, G_b)$

(c) $X\beta + b$
Smooth-CAR model:

(a) Smooth large-scale spatial trend: $\mathbf{X}\beta + \mathbf{Z}\alpha$

(b) CAR structured random effects: $\mathbf{b} \sim \mathcal{N}(\mathbf{0}, \mathbf{G}_b)$

(c) $\mathbf{X}\beta + \mathbf{Z}\alpha + \mathbf{b}$
Conclusions and Further Research

- We propose several models from an **unified perspective** for spatial data smoothing
  - Based on non-isotropic P-splines
  - Using a mixed model representation

- **“Smooth-CAR”** model:
  - Identifiability of both smooth $(\lambda_1, \lambda_2)$ and CAR effects $(\sigma^2_b, \varphi)$:
    - Simulation study (work in progress)

- Incorporate relevant covariates (non-linear effects,...)

- Interaction models (spatio-temporal), IWSM08 Utrecht

- Improve estimation procedure using sparse matrix algebra algorithms
Conclusions and Further Research

- We propose several models from an **unified perspective** for spatial data smoothing
  - Based on non-isotropic P-splines
  - Using a mixed model representation

- **“Smooth-CAR”** model:
  - **Identifiability** of both smooth \((\lambda_1, \lambda_2)\) and CAR effects \((\sigma_b^2, \phi)\):
    - simulation study (work in progress)

Incorporate relevant covariates (non-linear effects, ...)

Interaction models (spatio-temporal), IWSM08 Utrecht

Improve estimation procedure using sparse matrix algebra algorithms
Conclusions and Further Research

- We propose several models from an **unified perspective** for spatial data smoothing
  - Based on non-isotropic P-splines
  - Using a mixed model representation
- **“Smooth-CAR”** model:
  - Identifiability of both smooth $(\lambda_1, \lambda_2)$ and CAR effects $(\sigma_b^2, \phi)$: simulation study (**work in progress**)
- Incorporate relevant covariates (non-linear effects,...)
- Interaction models (spatio-temporal), **IWSM08 Utrecht**:
  \[ \eta = f(x_1, x_2, \text{time}) \]
Conclusions and Further Research

- We propose several models from an **unified perspective** for spatial data smoothing
  - Based on non-isotropic P-splines
  - Using a mixed model representation
- **“Smooth-CAR”** model:
  - **Identifiability** of both smooth ($\lambda_1$, $\lambda_2$) and CAR effects ($\sigma_b^2$, $\phi$): simulation study (**work in progress**)
- Incorporate relevant covariates (non-linear effects,...)
- Interaction models (spatio-temporal), **IWSM08 Utrecht**:
  \[ \eta = f(x_1, x_2, \text{time}) \]
- Improve estimation procedure using **sparse matrix algebra algorithms**