Mapping Soil Regolith Depth in Large and Censored Spatial Datasets Using Bayesian Hierarchical Models
Wen-Hsi Yang
2 December 2015
www.data61.csiro.au
Joint Work With

- Dr. Ross Searle: CSIRO Land and Water
- Dr. David Clifford: The Climate Corporation
- Dr. John Wilford: Geoscience Australia
Outline

- Motivating Examples: Soil Regolith Depth
- Methodology
- Modelling and Mapping Soil Regolith Depth in Queensland
- Summary
Regolith Depth

Regolith is a layer from the earth’s surface down to unweathered bedrock at depth.

[Wilford and Thomas (2013)]
Depth Measurements in Queensland

Spatial Distribution

17,672 Non-censored Measurements

10,717 Censored Measurements

Wen-Hsi Yang
We consider 17 variables associated with climate, relief, parent material, and time (Jenny, 1941).
We propose a Bayesian hierarchically spatial model for large and censored spatial data.

This model can account for the uncertainty of right-censored measurements.

Also, our model can includes environmental and ecological raster data as covariates to explain depth variation.

In addition, we use stochastic search variable selections (SSVS) algorithms to improve model selection and to perform model averaging to enhance prediction.

Lastly, we apply this model to fit and predict regolith depth in Queensland.
Hierarchically Spatial Models

- Let \( \{z(s_n)\}_{n=1}^{N} \) be a set of measurements at locations \( s_1, \ldots, s_N \).
- The sample consist of \( n_J \) non-censored and \( n_K \) censored measurements. That is, \( \{z(s_j); j \in J\} \) and \( \{z(s_k); k \in K\} \) with \( n_J + n_K = N \).
- Data model for non-censored measurements \( z(s_j) \):
  \[
  z(s_j) \sim N(y(s_j), \sigma^2_J),
  \]
  where \( y(s_j) \) and \( \sigma^2_J \) denote the true process at location \( s_j \) and the measurement error variance of the non-censored samples.
- Data model for right censored measurements \( z(s_k) \):
  \[
  z(s_k) \sim TN(y(s_k), \sigma^2_K)[-\infty, y(s_k)],
  \]
  where \( y(s_k) \) and \( \sigma^2_K \) denote the true process at location \( s_k \) and the measurement error variance of the censored samples.
• Process model for $Y$:

$$Y(s_n) = h(X(s_n))'\beta + \eta(s_n),$$

- $h(X(s_n)) = (h_1(X(s_n)), \ldots, h_q(X(s_n)))'$ is a vector of functions of $p$ spatial covariates $X(s_n)$.
- $\beta$ is a $q \times 1$ coefficient vector corresponding to $h(X(s_n))$.
- $\eta(s_n)$ is a mean-zero spatial Gaussian process with a valid covariance function $C_Y(s_n, s_n')$.
- Here, we assume $C_Y(s_n, s_n') = \sigma_Y^2 \rho(s_n, s_n'; \theta)$, where $\sigma_Y^2$ is a constant variance and $\rho(s_n, s_n'; \theta)$ is a correlation function with a set of parameters $\theta$. 
Approximate Correlation Matrices

- **Full scale approximations (Sang and Huang, 2012):**

\[
\Sigma = [\rho(s_n, s_{n'})]_{n, n'=1, \ldots, N} \approx \Sigma_g + \Sigma_\ell,
\]

where $\Sigma_g$ and $\Sigma_\ell$ are a reduced-rank and a sparse approximation matrix, respectively.

- **Stochastic matrix approximations (Banerjee et al., 2013):**

\[
\Sigma_g = (\Phi \Sigma) T (\Phi \Sigma \Phi^T)^{-1} (\Phi \Sigma),
\]

where $\Phi$ is a project matrix.

- Then, we can obtain $\Sigma_\ell$ as follow

\[
\Sigma_\ell = [\Sigma - \Sigma_g] \circ H_{\text{taper}}(s, s'; \alpha),
\]

where $H_{\text{taper}}$ is a correlation matrix defined by a compactly supported correlation function with values equal to zeros when $|s - s'| \geq \alpha$. 

10 | Mapping Soil Regolith Depth in Large and Censored Spatial Datasets Using Bayesian Hierarchical Models | Wen-Hsi Yang
Given a fixed accuracy level $\epsilon_g$, approximate $\rho(s_n, s_{n'}) = \exp\left(\frac{-||s_n - s_{n'}||}{25}\right)$ (red curve) using stochastic matrix approximations and the spherical covariance function $H_{taper}(s_n, s_{n'}; \alpha) = \left(1 - \frac{||s_n - s_{n'}||}{\alpha}\right)^2 \left(1 + \frac{||s_n - s_{n'}||}{2\alpha}\right)$. 
**Obtain and Select \( h(X(s)) \)**

- Use principal component analysis (PCA) and kernel principal component analysis (KPCA) as \( h(X(s)) \).

- 15 PCAs and 50 KPCAs explain 99.62% and 97.79% variations of 17 covariates, respectively.

- Use SSVS algorithms to select components.
Fitting Regolith Depth

- First, we take the Box-Cox transformations on depth.

- Fitting models with the following settings.
  - Use the exponential correlation function.
  - Consider the first 15 leading PCAs with 50 KPCAs.
  - Give vague inverse gamma distributions as priors to all variance parameters.
  - Use a discrete uniform distribution for $\theta$ given a set $\{10, 10.5, \ldots, 35\}$.
  - Use $\epsilon_g = 2000$, $\alpha = \{0, 0.1\}$ km, and $\delta_p = \{0.05, 0.1, 0.5\}$.
  - Use 10-fold cross-validation for model validation.
  - Run 10,000 MCMC iterations with 4,000 discarded as burn-in.
Results

- Use the root-mean-square prediction error (RMSPE) for evaluating model performance for the test set.

\[
\text{RMSPE} = \sqrt{\frac{1}{T \times I} \sum_{i=1}^{I} \sum_{t=1}^{T} (z(s_i) - \hat{y}_t(s_i))^2}
\]

where \( \hat{y}_t(s_i) \) denotes the \( t \)-th prediction from the MCMC iterations at location \( i \).

<table>
<thead>
<tr>
<th></th>
<th>15 PCAs</th>
<th>50 KPCAs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \alpha = 0 ) km</td>
<td>( \alpha = 0 ) km</td>
</tr>
<tr>
<td>( \sigma^2_\beta = 0.01 )</td>
<td>3.6585</td>
<td>3.6786</td>
</tr>
<tr>
<td>( \sigma^2_\beta = 0.1 )</td>
<td>3.5981</td>
<td>3.6141</td>
</tr>
<tr>
<td>( \sigma^2_\beta = 1 )</td>
<td>3.6008</td>
<td>3.6173</td>
</tr>
<tr>
<td>( \sigma^2_\beta = 10 )</td>
<td>3.6012</td>
<td>3.6173</td>
</tr>
<tr>
<td>( \sigma^2_\beta = 100 )</td>
<td>3.6014</td>
<td>3.6178</td>
</tr>
</tbody>
</table>
Selected KPCAs

KPCA_{24} (93%)

KPCA_{45} (91%)

KPCA_{40} (91%)

KPCA_{11} (90%)

KPCA_{5} (86%)

KPCA_{25} (74%)
Mapping Regolith Depth

Posterior Mean

Standard Deviation

2.5% Quantile

97.5% Quantile
Summary

- We consider a case where some spatial measurements are incomplete and the sample size is large.
- We develop a hierarchical model where two data models are constructed for non-censored and censored measurements, and then their true process are combined together in the process model.
- We use stochastic matrix approximations within the framework of full scale approximations to reduce computational burden due to large spatial data and increase the efficiency of the MCMC sampler.
- In data analysis, we uses PCA and KPCA to subtract common features of 17 variables.
- The SSVS helped identify important components relating to the regolith depth in Queensland.
Selected References


