Mapping Soil Regolith Depth in Large and Censored Spatial Datasets Using Bayesian Hierarchical Models

Wen-Hsi Yang

2 December 2015

Joint Work With

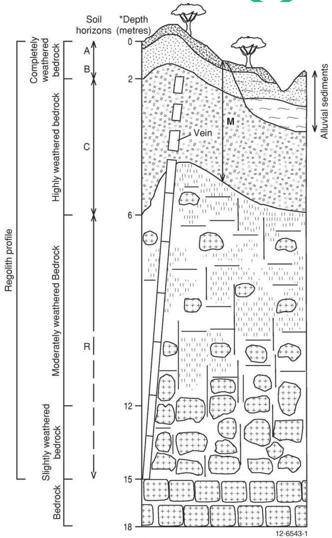
- **Dr. Ross Searle: CSIRO Land and Water**
- **Dr. David Clifford: The Climate Corporation**
- **Dr. John Wilford: Geoscience Australia**

Outline

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

Regolith Depth

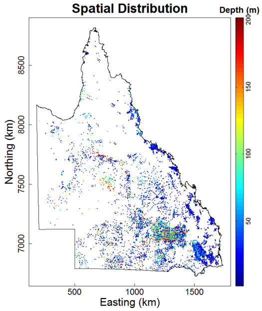
Regolith is a layer from the earth's surface down to unweathered bedrock at depth.

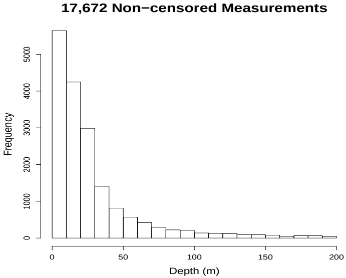


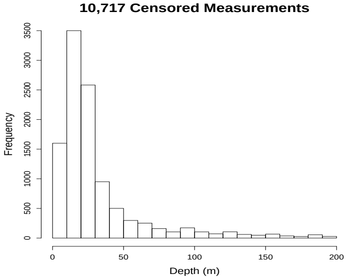
M = *Soil-regolith zone measured in this study*
* = *indicative depth only*

[Wilford and Thomas (2013)]

Depth Measurements in Queensland

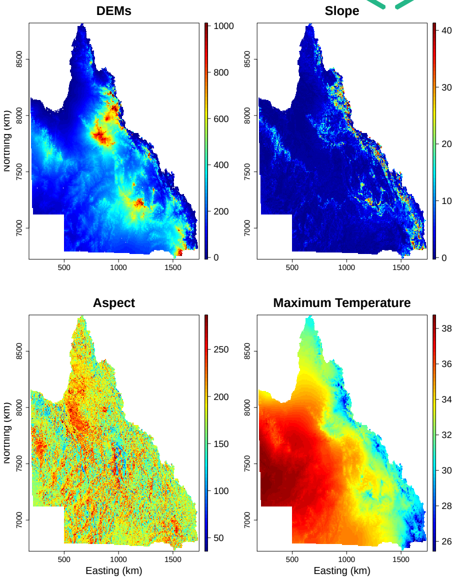






Environmental and Ecological Variables

We consider 17 variables associated with climate, relief, parent material, and time (Jenny, 1941).



Modelling and Mapping Regolith Depth

- 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 $\{z(\mathbf{s}_n)\}_{n=1}^{N}$ be a set of measurements at locations $\mathbf{s}_1, \ldots, \mathbf{s}_N$.
- The sample consist of n_J non-censored and n_K censored measurements. That is, $\{z(\mathbf{s}_j); j \in J\}$ and $\{z(\mathbf{s}_k); k \in K\}$ with $n_J + n_K = N$.
- Data model for non-censored measurements $z(\mathbf{s}_j)$:

$$z(\mathbf{s}_j) \sim N(y(\mathbf{s}_j), \sigma_J^2),$$

where $y(\mathbf{s}_j)$ and σ_J^2 denote the true process at location $\mathbf{s}_j$ and the measurement error variance of the non-censored samples.

- Data model for right censored measurements $z(\mathbf{s}_k)$:

$$z(\mathbf{s}_k) \sim TN(y(\mathbf{s}_k), \sigma_K^2)_{[-\infty, y(\mathbf{s}_k)]},$$

where $y(\mathbf{s}_k)$ and σ_K^2 denote the true process at location $\mathbf{s}_k$ and the measurement error variance of the censored samples.

Hierarchically Spatial Models (Cont.)

- Process model for **Y**:

$$Y(\mathbf{s}_n) = \mathbf{h}(X(\mathbf{s}_n))'\boldsymbol{\beta} + \eta(\mathbf{s}_n),$$

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

Approximate Correlation Matrices

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

$$\mathbf{\Sigma} = [\rho(\mathbf{s}_n, \mathbf{s}_{n'})]_{n,n'=1,\dots,N} \approx \mathbf{\Sigma}_g + \mathbf{\Sigma}_\ell,$$

  where $\mathbf{\Sigma}_g$ and $\mathbf{\Sigma}_\ell$ are a reduced-rank and a sparse approximation matrix, respectively.

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

$$\mathbf{\Sigma}_g = (\Phi\mathbf{\Sigma})^T(\Phi\mathbf{\Sigma}\Phi^T)^{-1}(\Phi\mathbf{\Sigma}),$$

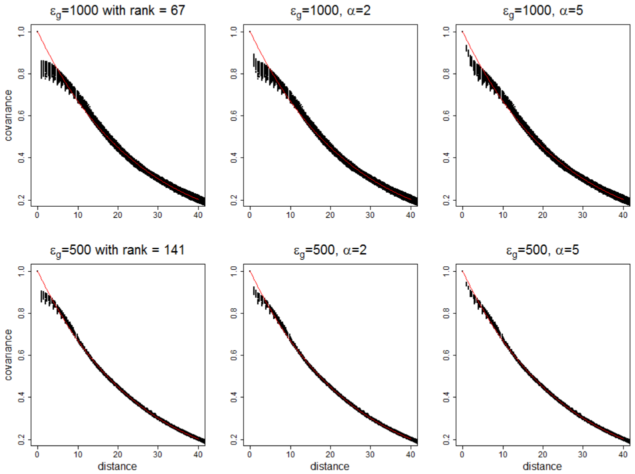
  where Φ is a project matrix.

- Then, we can obtain $\mathbf{\Sigma}_\ell$ as follow

$$\mathbf{\Sigma}_\ell = [\mathbf{\Sigma} - \mathbf{\Sigma}_g] \circ \mathbf{H}_{taper}(\mathbf{s}, \mathbf{s}'; \alpha),$$

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

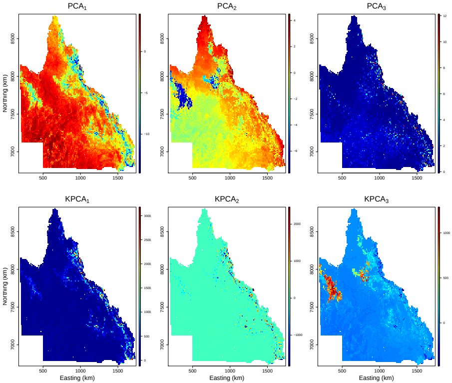
Full Scale Approximations



Given a fixed accuracy level ϵ_g, approximate $\rho(\mathbf{s}_n, \mathbf{s}_{n'}) = \exp(\frac{-||\mathbf{s}_n - \mathbf{s}_{n'}||}{25})$ (red curve) using stochastic matrix approximations and the spherical covariance function $H_{taper}(\mathbf{s}_n, \mathbf{s}_{n'}; \alpha) = (1 - \frac{||\mathbf{s}_n - \mathbf{s}_{n'}||}{\alpha})_+^2 (1 + \frac{||\mathbf{s}_n - \mathbf{s}_{n'}||}{2\alpha})$.

Obtain and Select $\mathbf{h}(X(\mathbf{s}))$

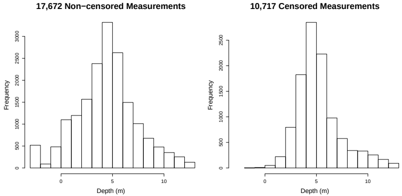
- Use principal component analysis (PCA) and kernel principal component analysis (KPCA) as $\mathbf{h}(X(\mathbf{s}))$.



- 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 θ 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

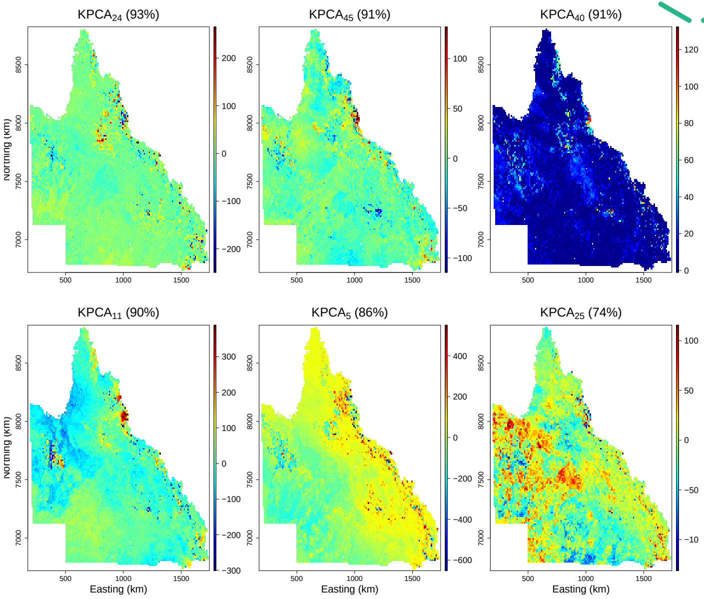
- Use the root-mean-square prediction error (RMSPE) for evaluating model performance for the test set.

$$\text{RMSPE} = \sqrt{\frac{1}{T \times I} \sum_{i=1}^{I} \sum_{t=1}^{T} (z(\mathbf{s}_i) - \widehat{y_t(\mathbf{s}_i)})^2}$$

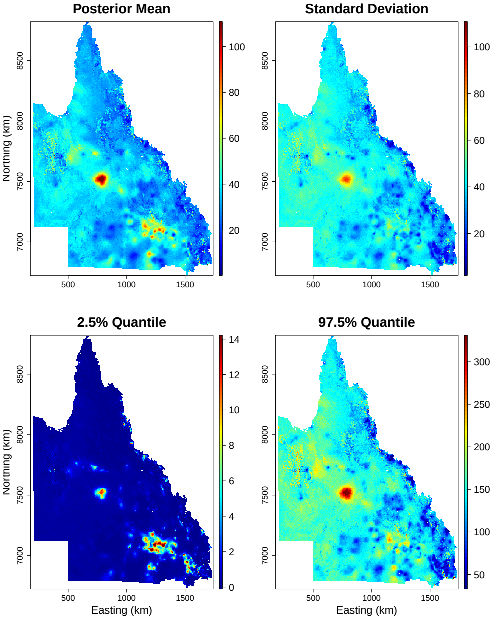
where $\widehat{y_t(\mathbf{s}_i)}$ denotes the t-th prediction from the MCMC iterations at location i.

	15 PCAs		50 KPCAs	
	$\alpha = 0$ km	$\alpha = 0.1$ km	$\alpha = 0$ km	$\alpha = 0.1$ km
$\sigma_\beta^2 = 0.01$	3.6585	3.6786	2.6547	3.6736
$\sigma_\beta^2 = 0.1$	3.5981	3.6141	3.5951	3.6118
$\sigma_\beta^2 = 1$	3.6008	3.6173	3.5981	3.6142
$\sigma_\beta^2 = 10$	3.6012	3.6173	3.5998	3.6155
$\sigma_\beta^2 = 100$	3.6014	3.6178	3.6002	3.6159

Selected KPCAs



Mapping Regolith Depth



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

- Banerjee, A., Dunson, D.B., and Tokdar, S.T. (2013) Efficient Gaussian process regression for large datasets. *Biometrika*. 100: 75–89.
- Cressie, N. & Wikle, C.K. (2001) *Statistics for Spatial-Temporal Data*. John Wiley & Sons.
- De Oliveira, V. (2005) Bayesian inference and prediction of Gaussian random fields based on censored data. *Journal of Computational and Graphical Statistics*. 14: 95–115.
- Jenny, H. (1941) *Factors of Soil Formation: A System of Quantitative Pedology*. McGraw Hill Book Co.
- Sang, H., and Huang, J.Z. (2012) A full scale approximation of covariance functions for large spatial data sets. *Journal of the Royal Statistical Society B*. 74: 111–132.
- George, E.I., and McCulloch, R.E. (1993) Variable selection via Gibbs sampling. *Journal of the American Statistical Association*. 88: 881–889.
- Wilford, J., and Thomas, M. (2013) Predicting regolith thickness in the complex weathering setting of the central Mt Lofty Ranges, South Australia. *Geoderma*. 206: 1–13.