Dimension Expansion Method in Non-stationary Spatial Modeling with Examples

STAT8530 Spatial Statistics

Hengrui Luo

Department of Statistics
The Ohio State University

November 13, 2016

How people come up with dimension expansion method? I

- **Motivation**
  - Why stationary covariance is desirable?
    - ★ The space formed by stationary covariance has intrinsic affine structure. Therefore basically we are doing linear inference.
  - What are the means people used for modeling nonstationary covariance?
    - ★ **Kernel based methods.** Find a family of stationary covariance functions supported on sub-domains and represent the nonstationary covariance function as their linear combinations.
    - ★ **Basis function methods.** Use empirical orthogonal functions as a basis for the nonstationary covariance function.
    - ★ **Process convolution methods.** We find a series of locally stationary covariances and proceed like we are doing a partition of unity on the spatial domain. A bit like kernel based method but in a smooth way.
    - ★ **Deformation methods.** Like image wrapping [2], such method requires stringent conditions on the deformation mapping.
    - ★ ***Dimension expansion methods.****(DE) Does it help us to understand the cause of nonstationarity?*

How people come up with dimension expansion method? II

- Dispersion matrix: $K_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \gamma(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|) - \gamma(0)$ for a stationary process where γ is the semi-variogram function characterizing the covariance structure. It can be estimated using existing observations, a typical estimation is $\gamma(h) = (\frac{1}{2N_h})\sum_i [Y(x_i + h) - Y(x_i)]^2$ where N_h is number of pairs of observations that are of distance h from each other.

Theorem

([4], Theorem 1) Let Φ be a function defined on $G \subset \mathbb{R}^n$ by $\Phi(\boldsymbol{x}) = (\boldsymbol{x}, \psi(\boldsymbol{x}))$, where $\psi(\boldsymbol{x})$ is a vector-valued function of the same dimension n such that the transformation

$$h : G \times G \to D - D, (\boldsymbol{x}_1, \boldsymbol{x}_2) \mapsto \Phi(\boldsymbol{x}_1) - \Phi(\boldsymbol{x}_2) = (\boldsymbol{x}_1 - \boldsymbol{x}_2, \psi(\boldsymbol{x}_1) - \psi(\boldsymbol{x}_2))$$

is bijective from $G \times G \backslash \Delta$ onto $D - D \backslash \{\mathbf{0}\}$, where $D = \Phi(G)$ and $\Delta = \{(\boldsymbol{x}, \boldsymbol{x}) : \boldsymbol{x} \in G\}$. Then there exists a reduced, centered and standardized Gaussian stationary random field $\boldsymbol{Y} = \boldsymbol{Y}(u), u \in D \subset \mathbb{R}^n \times \mathbb{R}^n \cong \mathbb{R}^{2n}$ with covariance function $R : D - D \to \mathbb{R}$ such that $\forall \boldsymbol{x}, \boldsymbol{x}' \in G$

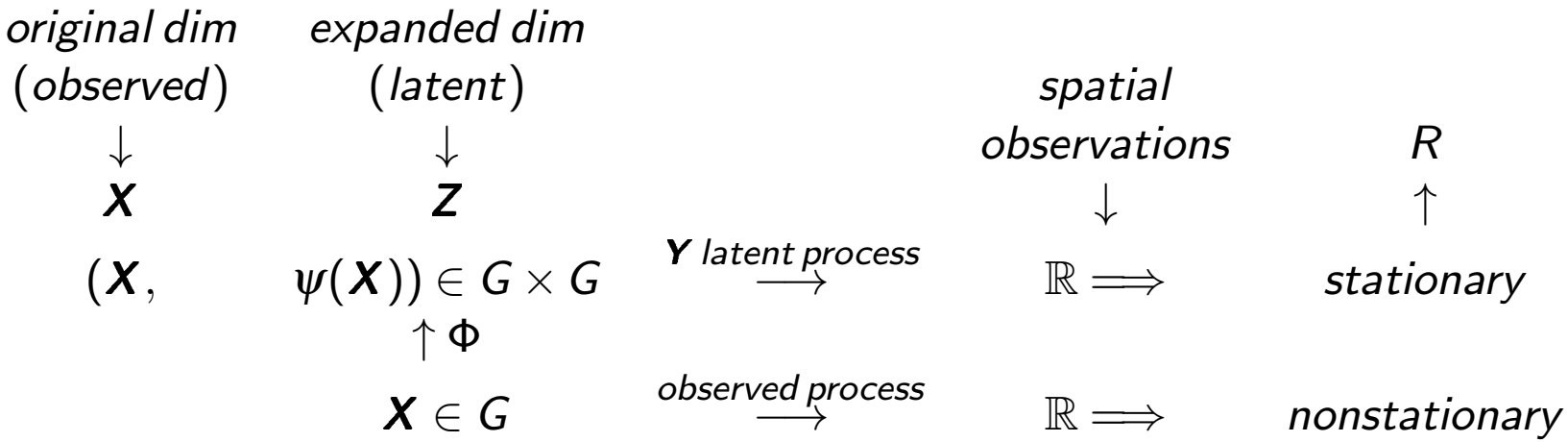
$$R(\boldsymbol{x} - \boldsymbol{x}', \psi - \psi') = R(h(\boldsymbol{x}, \boldsymbol{x}')) = Cov(\boldsymbol{Y}(\boldsymbol{\Phi}(\boldsymbol{x})), \boldsymbol{Y}(\boldsymbol{\Phi}(\boldsymbol{x}')))$$

.

How people come up with dimension expansion method? III

- **Illustration of the idea**
  The goal is to estimate the latent spatial process $\boldsymbol{Y}$, in order to do so we must
  - Estimate Φ, which is the same as estimating ψ. (completed by thin-plate regression)
    In order to estimate ψ, we must estimate the latent dimension value $\boldsymbol{Z}$ and formalize it as a nonparametric regression problem $\boldsymbol{Z} = \psi(\boldsymbol{X})$;
  - Estimate the **stationary** covariance structure of $\boldsymbol{Y}$ (completed by penalized optimization and assuming exponential covariance)
    In order to estimate the covariance structure of a stationary process, we use expanded coordinates (X, Z) as locations in expanded spatial domain and estimate the covariance structure under certain covariance structure.
    Therefore here we have a semiparametric/nonparametric problem of estimating the covariance structure of $\boldsymbol{Y}$.

How people come up with dimension expansion method? IV



- **Statistical analogies**
  - (EM algorithm) EM algorithm used in missing data analysis uses a simpler complete likelihood and integrate out augmented data to infer.
  - (Supported vector machine) SVM uses kernel to map the low dimensional data into a higher dimension in order to find an appropriate separating surface.

Step 1: Figure out the expanded dimension and stationary covariance I

- Procedure
  We used the penalized/regularization method to obtain the latent dimension and yield an estimate of the value of latent dimension.(learned dimension)
  - The optimization procedure used for determination of the expanded coordinate and the stationary covariance function is obtained by matching the empirical covariance matrix

$$\boldsymbol{V}^* = \left(v_{ij}^*\right)_{ij}$$

    where $v_{ij}^* = \frac{1}{|\tau|}\sum_\tau \left|Y(\boldsymbol{X}_i) - Y(\boldsymbol{X}_j)\right|^2$ summing over all available observations.
  - And the stationary covariance matrix in higher dimension defined by the exponential variogram

$$\boldsymbol{V}_{expanded} = \left(\gamma_\phi(\left\|[\boldsymbol{X}_i, \boldsymbol{Z}_i] - [\boldsymbol{X}_j, \boldsymbol{Z}_j]\right\|)\right)_{ij}$$

    where $[\boldsymbol{X}_i, \boldsymbol{Z}_i]$ is the expanded coordinates. The γ_ϕ is a stationary covariance function which in our examples are both exponential covariance function of the form

$$\gamma_\phi(d) := \phi_1\left(1 - exp(-\frac{h}{\phi_2})\right) - \phi_3$$

.

Step 1: Figure out the expanded dimension and stationary covariance II

- ★ A common misunderstanding is that exponential covariance is the only choice due to the specific statement about recovering of stationarity in expanded dimension. The fact is the model works well as long as the chosen covariance structure provides a good approximate to the stationary process recovered in expanded dimension. Interested audience may try Matern covariance kernel for example.

- The objective function in optimization procedure is

$$OBJ\left(\boldsymbol{V}^{*}, \boldsymbol{V}_{expanded}\right) = argmin_{\phi,\boldsymbol{Z}} \sum_{i<j} \left(\boldsymbol{V}^{*}_{\boldsymbol{ij}} - \boldsymbol{V}_{expanded,ij}\right)^{2} - \lambda_1 \left\|\boldsymbol{Z}\right\|_{column}$$

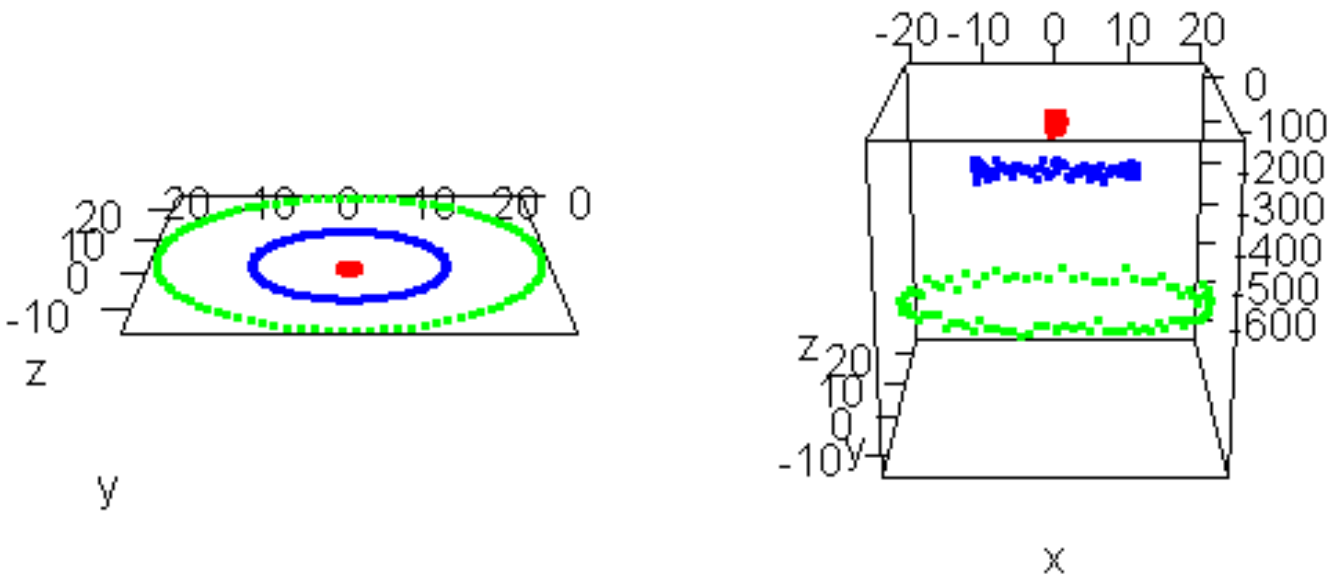
to determine $\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n$ and ϕ in order to determine each term in $\boldsymbol{V}_{expanded}$. (The penalized term is taken as column L^1 norm)

- ★ With estimated $\hat{\phi}$ and $\hat{\boldsymbol{Z}}$ we also have the estimated covariance function (because it is an exponential function which is completely determined by parameter ϕ) if we plug in $\hat{\phi}$into γ_{ϕ},denote such an estimate as $\widehat{\gamma_{\hat{\phi}}}$.

- Difficulties

  - In general we do not know how many augmented dimensions we need, but according to Theorem 1 we need to extend at most $p = s$ latent dimensions in order to obtain a stationary process in $\mathbb{R}^{s+p}$ for a collection of data in $\mathbb{R}^s$.

Figure: Three different elevation groups in $\mathbb{R}^{2+1}$ with stationary covariance becomes three concentric circles with non-stationary variance in $\mathbb{R}^2$.**[8530_example.R]**



Step 2: Thin plate regression using augmented data with penalty of smoothness I

- Procedure
  We used the dimension expansion technique to obtain by matching empirical variogram in observed space $\mathbb{R}^s$and the stationary variogram in augmented space $\mathbb{R}^{s+p}$.
  - We have obtained the expanded coordinates for every observations in original dimension.
  - The prediction of covariance at a certain spatial location can be done in a two-step fashion.
    - ⋆ Given a pair of spatial locations $\boldsymbol{X}_1, \boldsymbol{X}_2$ in original dimension, use $\hat{\psi}$ to obtain latent dimension $\hat{\boldsymbol{Z}} = \hat{\psi}(\boldsymbol{X})$. With this estimated latent coordinate we have $\left[\boldsymbol{X}_1, \hat{\boldsymbol{Z}}_1\right], \left[\boldsymbol{X}_2, \hat{\boldsymbol{Z}}_2\right]$ instead.
    - ⋆ Use the estimated covariance function $\hat{\gamma\phi}(\left[\boldsymbol{X}_1, \hat{\boldsymbol{Z}}_1\right], \left[\boldsymbol{X}_2, \hat{\boldsymbol{Z}}_2\right])$ to estimate the covariance at this pair of locations. The estimated covariance do not have to be stationary in the original dimension because the thin-plate regression does not necessarily give a linear mapping BUT it is stationary on the expanded coordinate space.

Step 2: Thin plate regression using augmented data with penalty of smoothness II

Corollary

The covariance between two locations $\boldsymbol{X}_1, \boldsymbol{X}_2$ of observed process can be estimated by

$$Cov_{observed}(\boldsymbol{X}_1, \boldsymbol{X}_2) = \widehat{\gamma}_{\hat{\phi}}(\left[\boldsymbol{X}_1, \hat{\boldsymbol{Z}}_1\right], \left[\boldsymbol{X}_2, \hat{\boldsymbol{Z}}_2\right]) = \widehat{\gamma}_{\hat{\phi}}([\boldsymbol{X}_1, \hat{\psi}(\boldsymbol{X}_1)], [\boldsymbol{X}_2, \hat{\psi}(\boldsymbol{X}_1)])$$

- Difficulties
  - The scale is very important here. if the latent dimension has a different scale than the original dimension, the covariance might seem to be stationary even in the original dimension, which is not a faithful exposition.
  - However we should maintain the same scale during the step-by-step procedure in order to avoid morbid oscillation of the estimated value as the initial values differ.

Google® Earth Screenshots of Areas under analysis

- Google Earth can adapt to either UTM or latitude-longitude coordinates, and it is a good tool to
  - check whether you have scaled/transformed the data correctly. For example, there might be something wrong if the UTM coordinate you yield from R "fly" you to Pacific Ocean in Google Earth while we are analyzing forest dataset.
  - check whether the estimated latent dimension is somehow interpretable as some geographic features like elevation/altitude.

Example

(Convert Latitude-Longitude into UTM in R) **[Source]**

```
library(rgdal)
xy <- data.frame(ID = 1:length(x), X = x, Y = y)
coordinates(xy) <- c("X", "Y")
proj4string(xy) <- CRS("+proj=longlat +datum=WGS84")
result <- spTransform(xy, CRS(paste("+proj=utm +zone=",zone,"
ellps=WGS84",sep='')))
```

This function illustrates how we can transform the latitude-longitude coordinates into UTM coordinates(m) given the UTM zone.

Figure: The spatial area corresponding to solar radiation dataset.

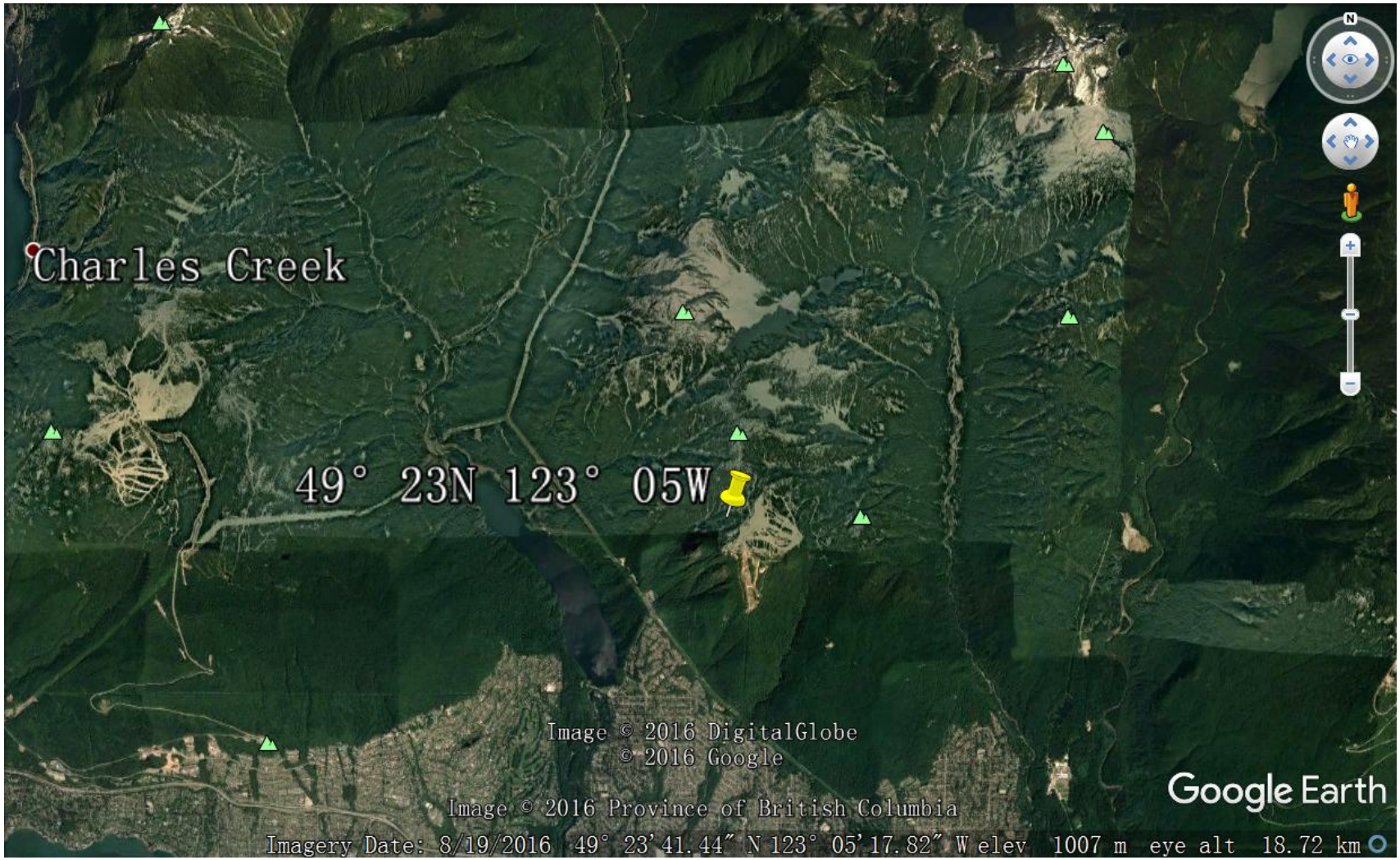


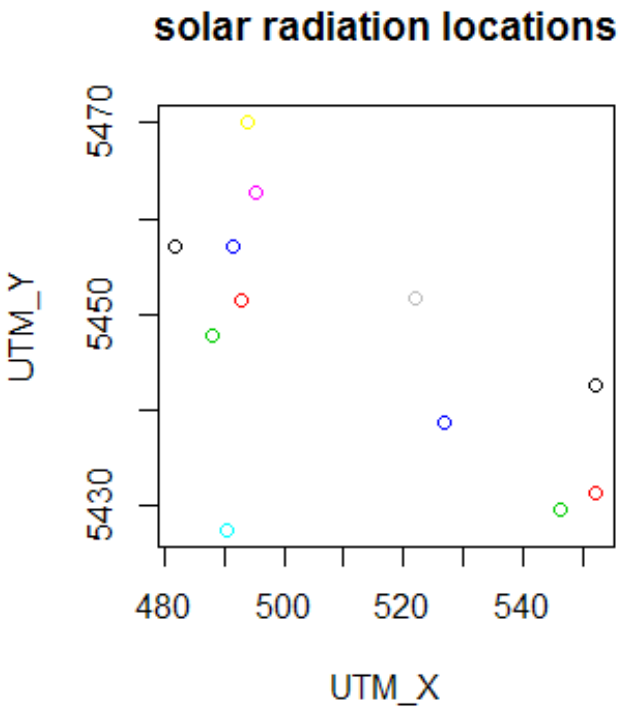
Figure: The spatial area corresponding to biomass dataset.

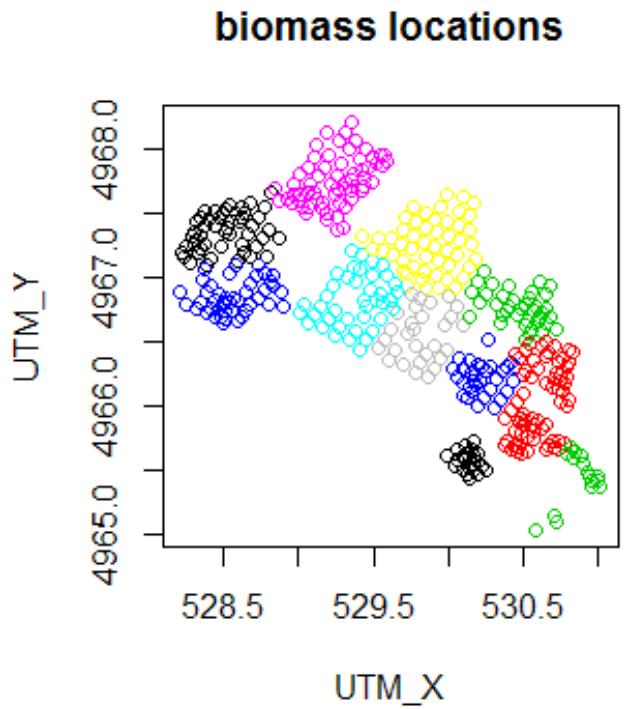
Step-by-step analysis of two spatial datasets I

- **Data**
  - The solar radiation dataset
    The data source scattered in [5](locations, Table1) and [2](observed covariance matrix, Table1). Generally they are of good quality, it contains 12 locations in terms of longitude and latitudes and observed covariance matrix among 12 locations. I transformed the coordinates into UTM coordinates using `rgdal` package.
  - The biomass dataset
    The data source is [6] and I obtained it from our course files. This data set contains 451 locations with observed value of biomasses. To simplify the calculation(To avoid optimizing 451*5 multivariate function), I first use `kmeans` to cluster them into 12 centers and take the mean coordinates as their new coordinates; mean biomasses measurements as their new biomass measurements. And then I use the `variog` function in `geoR` to estimate the observed covariance matrix among these 12 center locations. By doing such a clustering of data

Step-by-step analysis of two spatial datasets II

Figure: Observation locations of two data sets (The different colors means different groups, for solar radiation each cluster contains only one location; the biomass dataset is clustered as described above.)

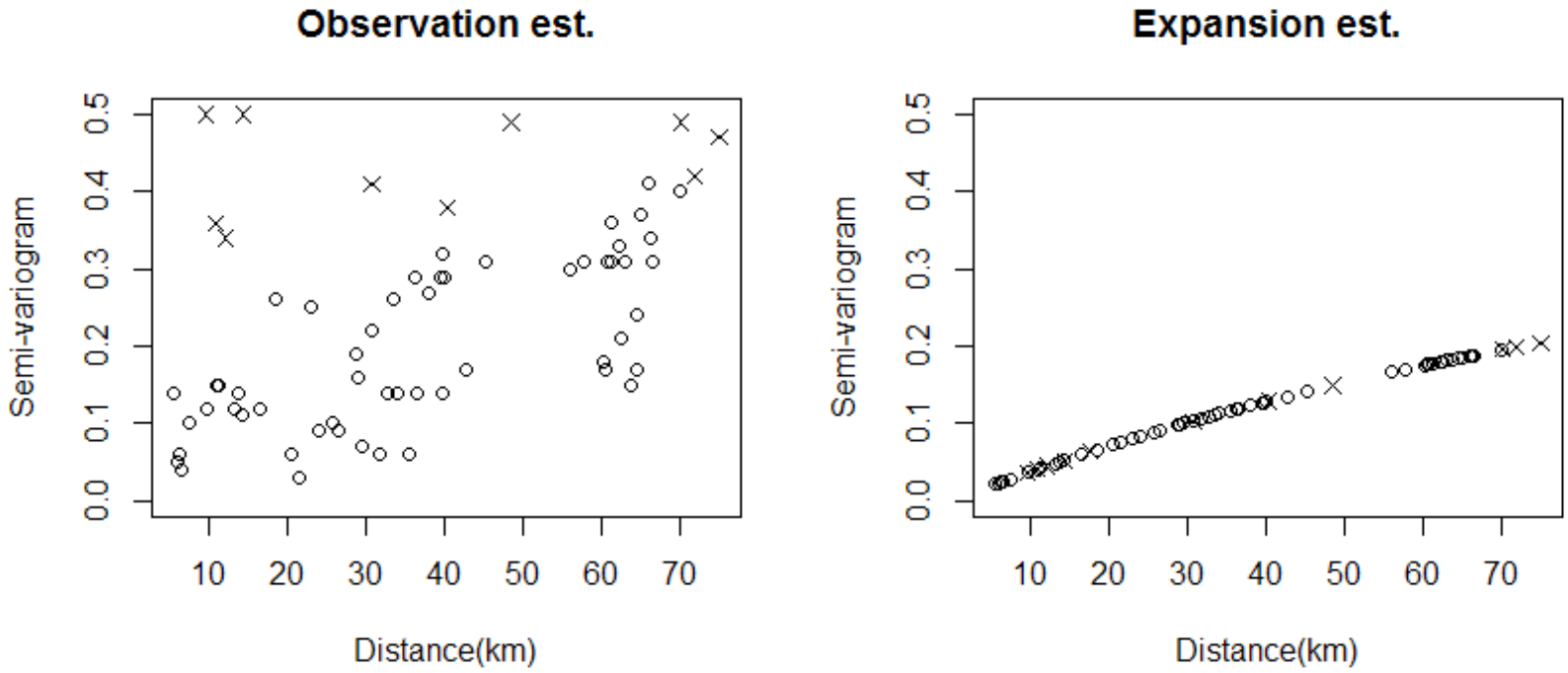




Step-by-step analysis of two spatial datasets III

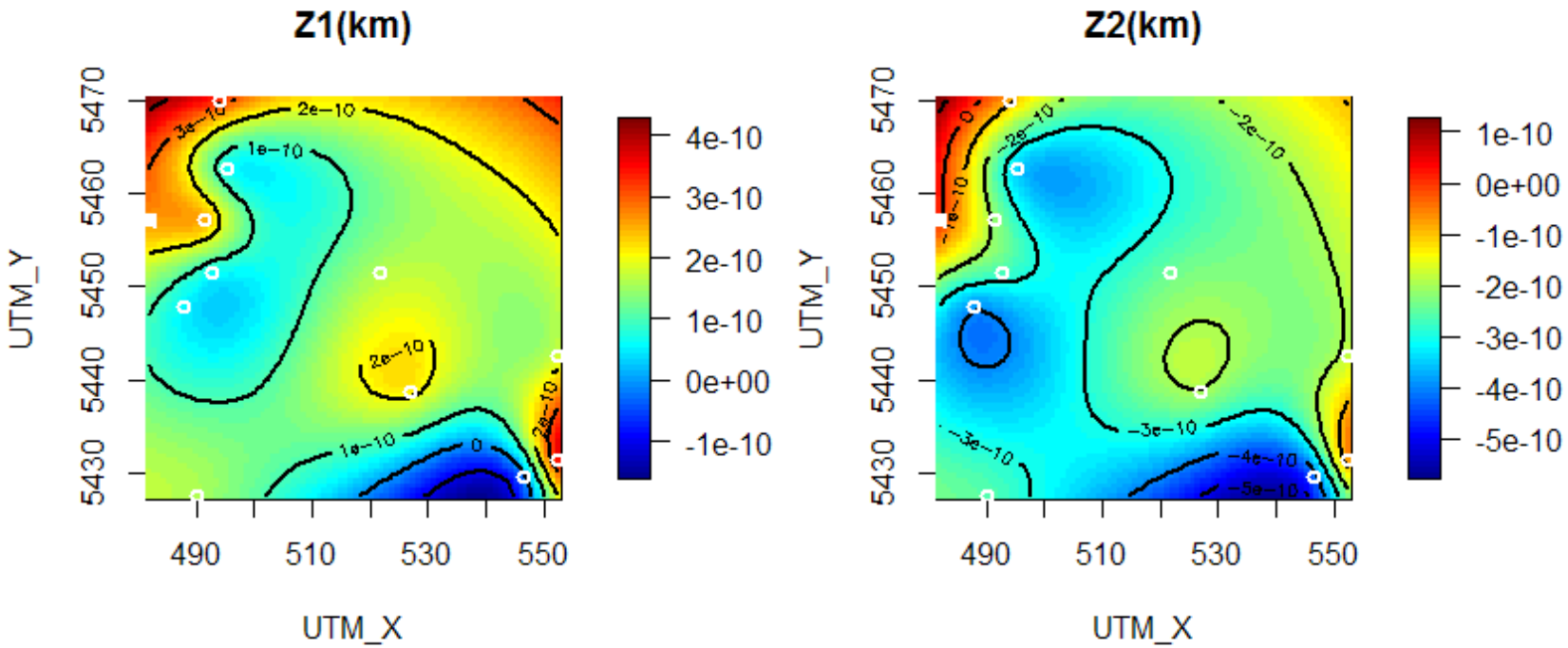
- **Result**
  - **The solar radiation dataset [8530_solarradiation.R]**

Figure: The observed semi-variogram; the estimated semi-variogram via DE. (The result is slightly different from [1] due to the scaling of units, they used m for elevations; I used km.)



Step-by-step analysis of two spatial datasets IV

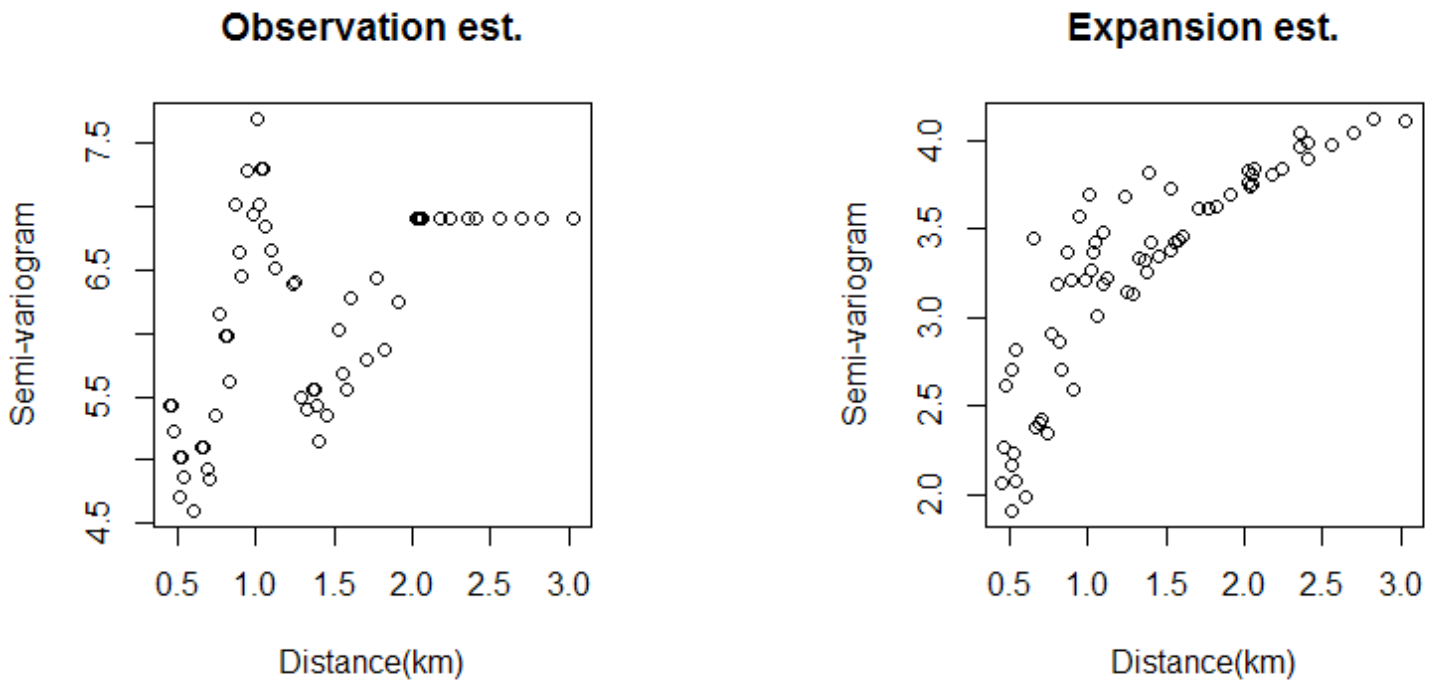
Figure: The estimated latent dimension values for Z_1, Z_2 .



Step-by-step analysis of two spatial datasets V

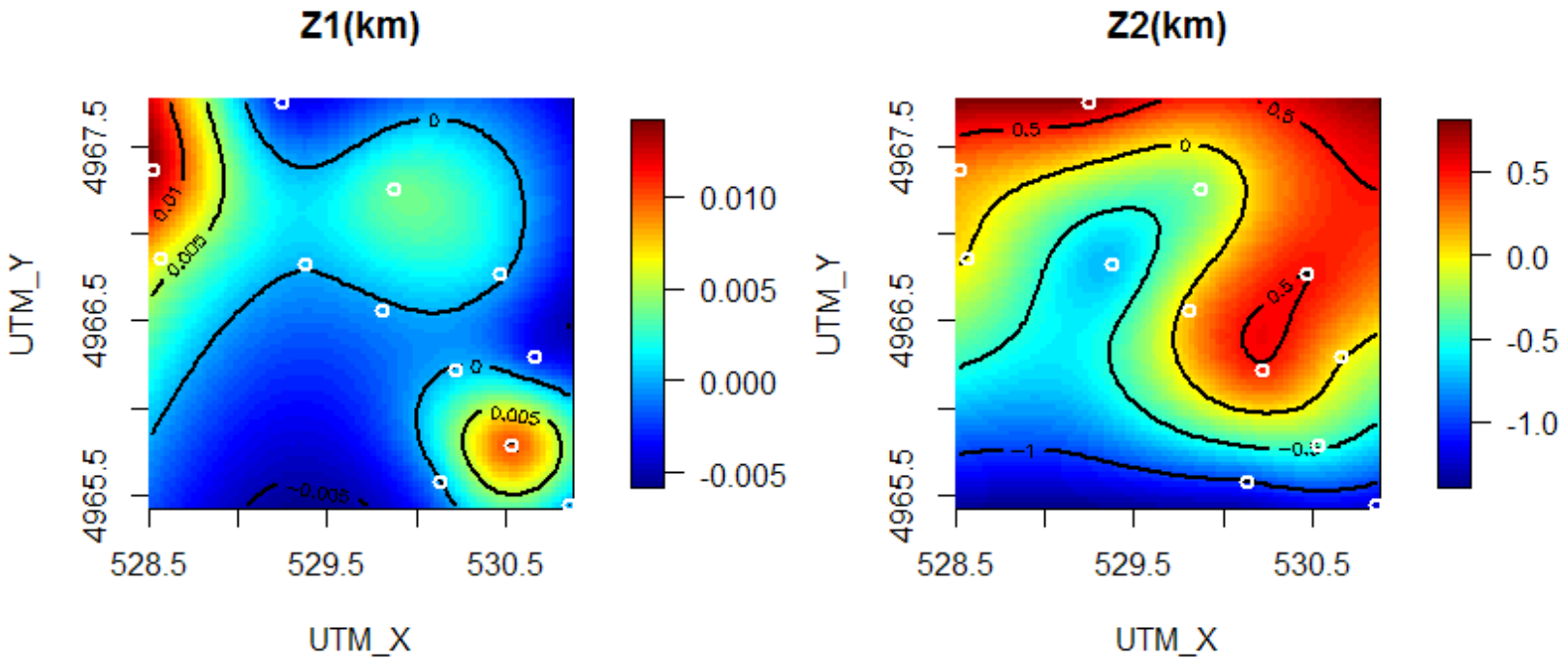
- **The biomass dataset [8530_biomass.R]**

Figure: The observed semi-variogram; the estimated semi-variogram via DE. (The result has a translation of around 2 units)



Step-by-step analysis of two spatial datasets VI

Figure: The estimated latent dimension values for Z_1, Z_2 .



- C

Details in realization of dimension expansion models I

- Optimization methods
  - For moderate sample size (number of locations in data set)
    - ⋆ Nelder-Mead method (Generally not very good, but works when sample size is moderate.)
    - ⋆ BFGS method (Generally good, under suitable initial values)
  - For large sample size
    - ⋆ Gradient projection method
- Estimation of the latent dimension
  - The value of latent dimension is not interpretable.
    For example, when we estimate the biomass data, the latent dimension varies in scale of 10~50(km). Fig 7. of [1] also showed that the latent dimension is not interpretable.
  - The optimization procedure is highly dependent on the scale of the data. However, it turns out that in most cases the $\boldsymbol{V}_{expanded}$ is robust against the choice of initial value. (Also known by the Dr.Bornn via email correspondence.) My suggestion is that we can insert a (informative) prior in order to determine the $\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n$ and ϕ instead of optimizing the penalized objective function. Later I found my idea is partially realized by [7].

Details in realization of dimension expansion models II

- The latent dimension is usually sparse.
  This is to say the estimated dimension $\boldsymbol{Z}_3, \boldsymbol{Z}_4, \boldsymbol{Z}_5$ are zero even if we include these additional dimensions.
  - ⋆ The first is reason is that by Theorem 1 it suffices to include 2 dimensions when the spatial domain is a subset in $\mathbb{R}^2$ when we aimed at recovering stationarity.
  - ⋆ The second reason is that the signs of the first two dimensions $\boldsymbol{Z}_1, \boldsymbol{Z}_2$ can be chosen to be positive/negative by selecting a good initial value. In practice this makes the estimates of additional dimensions minuscule.
    Therefore I believe that it is both computationally efficient and feasible to include 2 dimensions when analyzing 2 dimensional data. By including only 2 dimensions, it also avoids the tendency of constructing an *over-parameterized* model.

    "For most of the problems I have worked on, 1 dimension has been sufficient. So I think it's safe to say 1 or 2 dimensions are usually enough to capture nonstationarity in a reasonable way for most applied problems."(Dr.Bornn, email correspondence)

  - ⋆ Are there recent improvements? Yes, but restricted to the improvement of the optimization procedure, for example [7].

References I

Thanks to Dr.Bornn in helping me with the details of the data source used in [1].

[1] Bornn, Luke, Gavin Shaddick, and James V. Zidek. "Modeling nonstationary processes through dimension expansion." Journal of the American Statistical Association 107.497 (2012): 281-289.

[2] Sampson, Paul D., and Peter Guttorp. "Nonparametric estimation of nonstationary spatial covariance structure." Journal of the American Statistical Association 87.417 (1992): 108-119.

[3] Lindgren, Finn, Håvard Rue, and Johan Lindström. "An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73.4 (2011): 423-498.

[4] Perrin, Olivier, and Wendy Meiring. "Nonstationarity in Rn is second-order stationarity in R2n." Journal of applied probability (2003): 815-820.

[5] Hay, John E. "An assessment of the mesoscale variability of solar radiation at the earth's surface." Solar Energy 32.3 (1984): 425-434.

References II

[6] Banerjee, Sudipto, and Andrew Finley. "Bayesian multi-resolution modeling for spatially replicated data sets with application to forest biomass data." Journal of Statistical Planning and Inference 137.10 (2007): 3193-3205.

[7] Snoek, Jasper, et al. "Input Warping for Bayesian Optimization of Non-Stationary Functions." ICML. 2014.