A 'short walk on the spatial side'

- Just a few comments on:

A 'short walk on the spatial side'

- Just a few comments on:
    - the fundamental concepts arising in spatial statistical modelling

A 'short walk on the spatial side'

- Just a few comments on:
  - the fundamental concepts arising in spatial statistical modelling
  - the kind of geostatistical modelling ideas that might be useful in 'post processing' data relating to seasonal forecasting applications

- The term 'geostatistical' data is used to refer to a set of spatially referenced measurements (regular grid or irregular) on some phenomenon which can in principle occur continuously across the study region of interest (precipitation, temperature etc.)

- The term ‘geostatistical’ data is used to refer to a set of spatially referenced measurements (regular grid or irregular) on some phenomenon which can in principle occur continuously across the study region of interest (precipitation, temperature etc.)
- We wish to understand the spatial probability distribution of values over the study region given the values at the *fixed* measurement points (and maybe other covariates observed in the same region).

- The term ‘geostatistical’ data is used to refer to a set of spatially referenced measurements (regular grid or irregular) on some phenomenon which can in principle occur continuously across the study region of interest (precipitation, temperature etc.)
- We wish to understand the spatial probability distribution of values over the study region given the values at the *fixed* measurement points (and maybe other covariates observed in the same region).
- Ultimately, we wish to use such models to obtain good predictions of values at points other than those where measurements are observed.

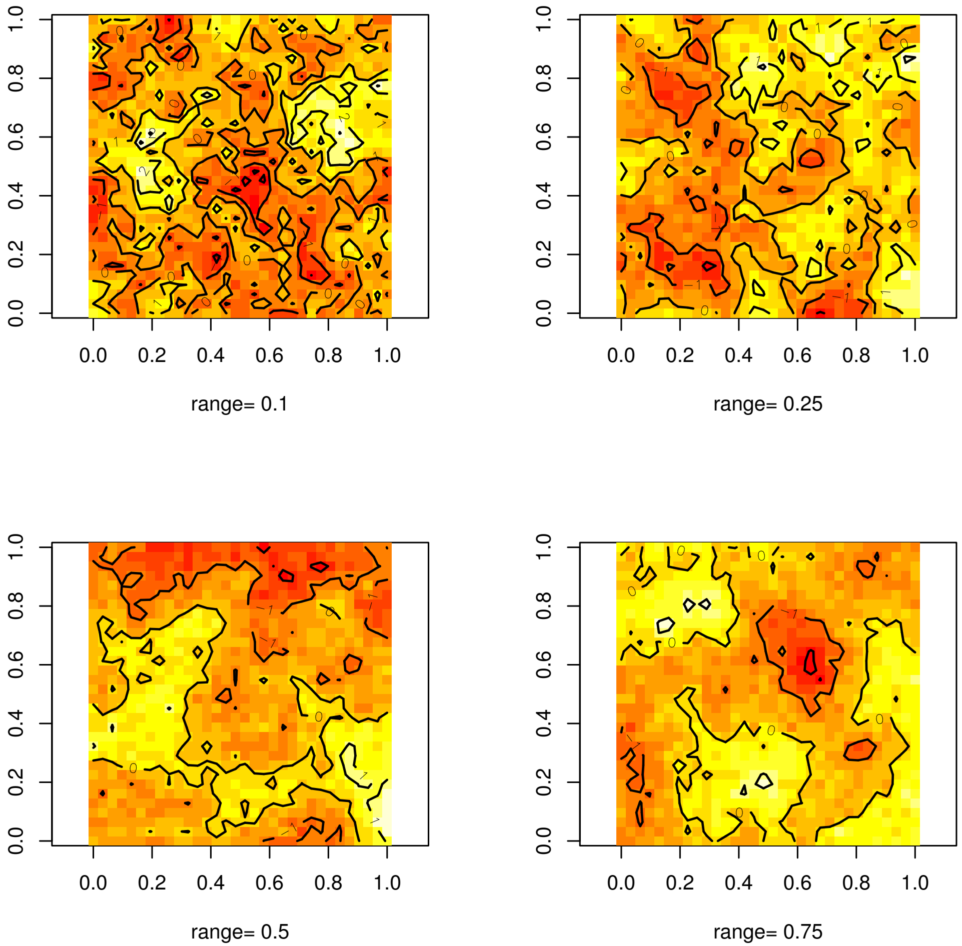
- Represent a *spatial stochastic process* by a set of (usually non-independent) random variables indexed by vector location $\boldsymbol{s} = (s_1, s_2)'$ i.e. $\{Y(\boldsymbol{s}), \boldsymbol{s} \in \mathcal{R}\}$

- Represent a *spatial stochastic process* by a set of (usually non-independent) random variables indexed by vector location $\boldsymbol{s} = (s_1, s_2)'$ i.e. $\{Y(\boldsymbol{s}), \boldsymbol{s} \in \mathcal{R}\}$
- Can talk about *first order* effects, that is behaviour of the process mean - E$(Y(\boldsymbol{s}))$

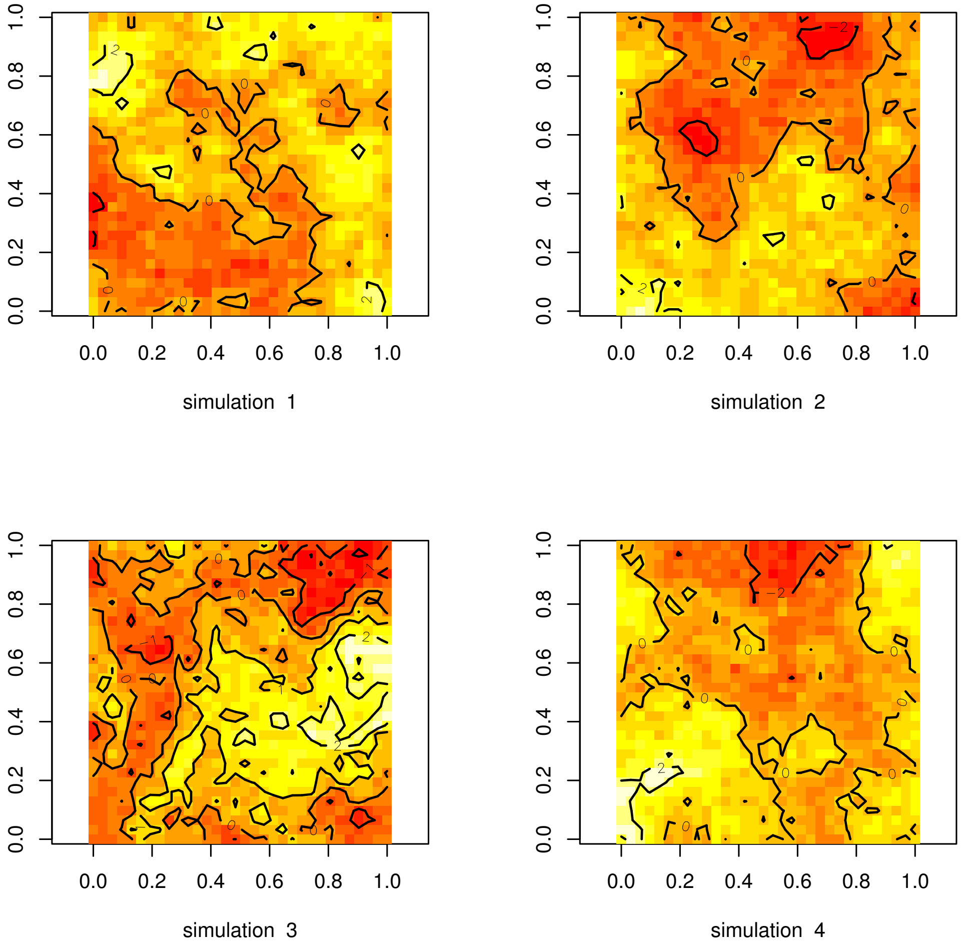
- Represent a *spatial stochastic process* by a set of (usually non-independent) random variables indexed by vector location $\boldsymbol{s} = (s_1, s_2)'$ i.e. $\{Y(\boldsymbol{s}), \boldsymbol{s} \in \mathcal{R}\}$
- Can talk about *first order* effects, that is behaviour of the process mean - E$(Y(\boldsymbol{s}))$
- and *second order* effects, that is behaviour of the spatial correlation structure - Cov $(Y(\boldsymbol{s}_i), Y(\boldsymbol{s}_j))$

- Represent a *spatial stochastic process* by a set of (usually non-independent) random variables indexed by vector location $\boldsymbol{s} = (s_1, s_2)'$ i.e. $\{Y(\boldsymbol{s}), \boldsymbol{s} \in \mathcal{R}\}$
- Can talk about *first order* effects, that is behaviour of the process mean - E$(Y(\boldsymbol{s}))$
- and *second order* effects, that is behaviour of the spatial correlation structure - Cov $(Y(\boldsymbol{s}_i), Y(\boldsymbol{s}_j))$
- Both components are important in understanding a spatial process. In practice the observed behaviour of spatial phenomena involves a confounding of both.

Random Field with varying spatial covariance range



Different realisations of same Random Field



- So the broad objective in spatial modelling is to understand as much as possible about the *stochastic process* giving rise to the data. That is about the joint distribution of $\{Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n)\}$

.

- So the broad objective in spatial modelling is to understand as much as possible about the *stochastic process* giving rise to the data. That is about the joint distribution of $\{Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n)\}$

- This will involve both understanding its mean (first order) behaviour: $\mu(\boldsymbol{s}) = \mathsf{E}(Y(\boldsymbol{s}))$ as a function of space (and possibly other covariates).

- **Also** understanding its spatial correlation (second order) structure: Cov $(Y(\boldsymbol{s}_i), Y(\boldsymbol{s}_j))$ as a function of space

.

- So the broad objective in spatial modelling is to understand as much as possible about the *stochastic process* giving rise to the data. That is about the joint distribution of $\{Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n)\}$
- This will involve both understanding its mean (first order) behaviour: $\mu(\boldsymbol{s}) = \mathrm{E}(Y(\boldsymbol{s}))$ as a function of space (and possibly other covariates).
- **Also** understanding its spatial correlation (second order) structure: Cov $(Y(\boldsymbol{s}_i), Y(\boldsymbol{s}_j))$ as a function of space
- In general this is not going to be easy!

.

- So the broad objective in spatial modelling is to understand as much as possible about the *stochastic process* giving rise to the data. That is about the joint distribution of $\{Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n)\}$
- This will involve both understanding its mean (first order) behaviour: $\mu(\boldsymbol{s}) = \mathrm{E}(Y(\boldsymbol{s}))$ as a function of space (and possibly other covariates).
- **Also** understanding its spatial correlation (second order) structure: Cov $(Y(\boldsymbol{s}_i), Y(\boldsymbol{s}_j))$ as a function of space
- In general this is not going to be easy! The observed behaviour of spatial phenomena usually involves a confounding of both.

.

- So the broad objective in spatial modelling is to understand as much as possible about the *stochastic process* giving rise to the data. That is about the joint distribution of $\{Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n)\}$
- This will involve both understanding its mean (first order) behaviour: $\mu(\boldsymbol{s}) = \mathsf{E}(Y(\boldsymbol{s}))$ as a function of space (and possibly other covariates).
- **Also** understanding its spatial correlation (second order) structure: Cov $(Y(\boldsymbol{s}_i), Y(\boldsymbol{s}_j))$ as a function of space
- In general this is not going to be easy! The observed behaviour of spatial phenomena usually involves a confounding of both. And we usually have very little repetition to work with (often just a single measurement at each spatial location).

- So the broad objective in spatial modelling is to understand as much as possible about the *stochastic process* giving rise to the data. That is about the joint distribution of $\{Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n)\}$
- This will involve both understanding its mean (first order) behaviour: $\mu(\boldsymbol{s}) = \mathsf{E}(Y(\boldsymbol{s}))$ as a function of space (and possibly other covariates).
- **Also** understanding its spatial correlation (second order) structure: Cov $(Y(\boldsymbol{s}_i), Y(\boldsymbol{s}_j))$ as a function of space
- In general this is not going to be easy! The observed behaviour of spatial phenomena usually involves a confounding of both. And we usually have very little repetition to work with (often just a single measurement at each spatial location). And we also have the complexity of real geography to deal with

- $Y(\boldsymbol{s})$ is random variable representing value at vector location $\boldsymbol{s}$. Process is: $\{Y(\boldsymbol{s}),\ \boldsymbol{s} \in \mathcal{R}\}$. We have measurements on this at fixed locations $(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n)$.

- $Y(\boldsymbol{s})$ is random variable representing value at vector location $\boldsymbol{s}$. Process is: $\{Y(\boldsymbol{s}),\ \boldsymbol{s} \in \mathcal{R}\}$. We have measurements on this at fixed locations $(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n)$.
- First order properties—mean value $\mu(\boldsymbol{s}) = \mathsf{E}(Y(\boldsymbol{s}))$

- $Y(\boldsymbol{s})$ is random variable representing value at vector location $\boldsymbol{s}$. Process is: $\{Y(\boldsymbol{s}),\ \boldsymbol{s} \in \mathcal{R}\}$. We have measurements on this at fixed locations $(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n)$.
- First order properties—mean value $\mu(\boldsymbol{s}) = \mathsf{E}(Y(\boldsymbol{s}))$
- Second order properties—covariance function

$$C(\boldsymbol{s}_i, \boldsymbol{s}_j) = \mathsf{Cov}\left(Y(\boldsymbol{s}_i), Y(\boldsymbol{s}_j)\right)$$

- $Y(\boldsymbol{s})$ is random variable representing value at vector location $\boldsymbol{s}$. Process is: $\{Y(\boldsymbol{s}),\ \boldsymbol{s} \in \mathcal{R}\}$. We have measurements on this at fixed locations $(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n)$.
- First order properties—mean value $\mu(\boldsymbol{s}) = \mathsf{E}(Y(\boldsymbol{s}))$
- Second order properties—covariance function

$$C(\boldsymbol{s}_i, \boldsymbol{s}_j) = \mathsf{Cov}\,(Y(\boldsymbol{s}_i), Y(\boldsymbol{s}_j))$$

- Where unambiguous can refer to $Y(\boldsymbol{s}_i)$ as simply Y_i with corresponding observed value y_i. Might have measurements on additional covariates $\boldsymbol{x}_i$ at each $\boldsymbol{s}_i$ to assist in modelling. In which case the mean may be $\mu(\boldsymbol{s}, \boldsymbol{x})$ and the covariance function may be $C(\boldsymbol{s}_i, \boldsymbol{s}_j, \boldsymbol{x}_i, \boldsymbol{x}_j)$

- Process is *stationary* if:

- Process is *stationary* if:
  - $\mu(\boldsymbol{s}) = \mu$ (constant over space)

- Process is *stationary* if:
  - $\mu(\boldsymbol{s}) = \mu$ (constant over space)
  - Var$(Y(\boldsymbol{s})) = \sigma^2$ (constant over space)
  - and covariance depends only on vector difference (distance and direction) between $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ (not on absolute location) viz:

$$\text{Cov}\,(Y(\boldsymbol{s}_i), Y(\boldsymbol{s}_j)) = C(\boldsymbol{s}_i - \boldsymbol{s}_j) = C(\boldsymbol{d})$$

- Process is *stationary* if:
  - $\mu(\boldsymbol{s}) = \mu$ (constant over space)
  - Var$(Y(\boldsymbol{s})) = \sigma^2$ (constant over space)
  - and covariance depends only on vector difference (distance and direction) between $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ (not on absolute location) viz:

$$\text{Cov}\left(Y(\boldsymbol{s}_i), Y(\boldsymbol{s}_j)\right) = C(\boldsymbol{s}_i - \boldsymbol{s}_j) = C(\boldsymbol{d})$$

- Process is *isotropic* if latter dependence is only on length of $\boldsymbol{d}$ (d) and *not direction*.

Essence of statistical modelling of geostatistical data

- Broad thrust is to think of the process being modelled as comprising two components:

Essence of statistical modelling of geostatistical data

- Broad thrust is to think of the process being modelled as comprising two components:
  - A systematic (trend) component (deterministic) representing how the mean value $\mu(\boldsymbol{s})$ varies across space (and with covariates if present).

Essence of statistical modelling of geostatistical data

- Broad thrust is to think of the process being modelled as comprising two components:
  - A systematic (trend) component (deterministic) representing how the mean value $\mu(\boldsymbol{s})$ varies across space (and with covariates if present).
  - A random component (stochastic) comprising a stationary (anisotropic or isotropic) random field with a covariance structure dependent upon spatial separation

Essence of statistical modelling of geostatistical data

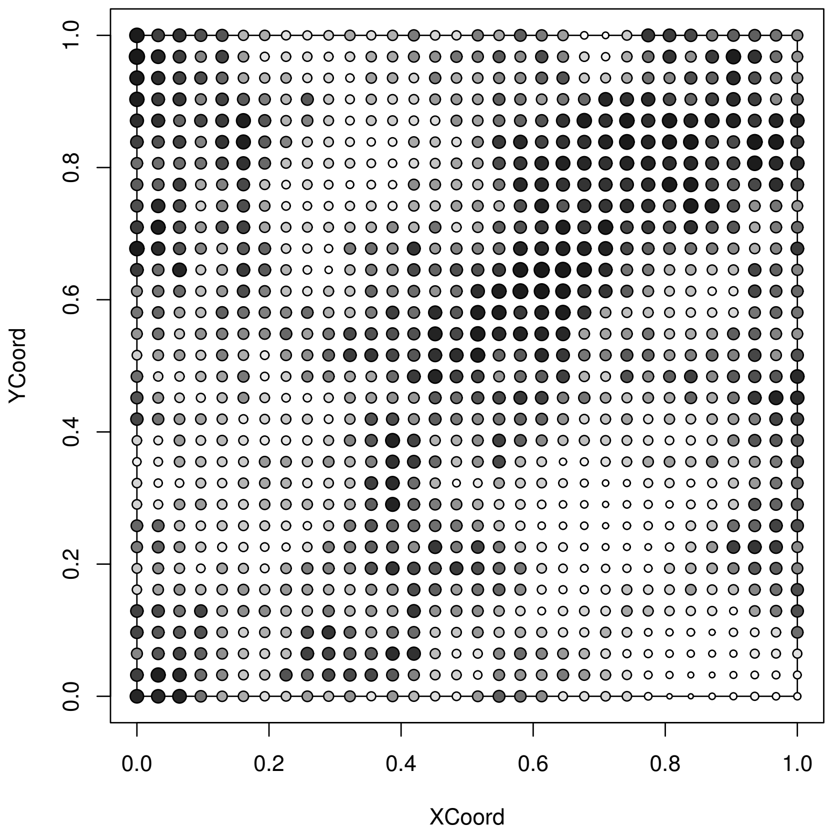
- The typical geostatistical model assumes some simple transformation of the trend to be a linear combination of spatial coordinates and covariates and then the random component to be a stationary Gaussian (normally distributed) **Markov Random Field** (MRF).

Essence of statistical modelling of geostatistical data

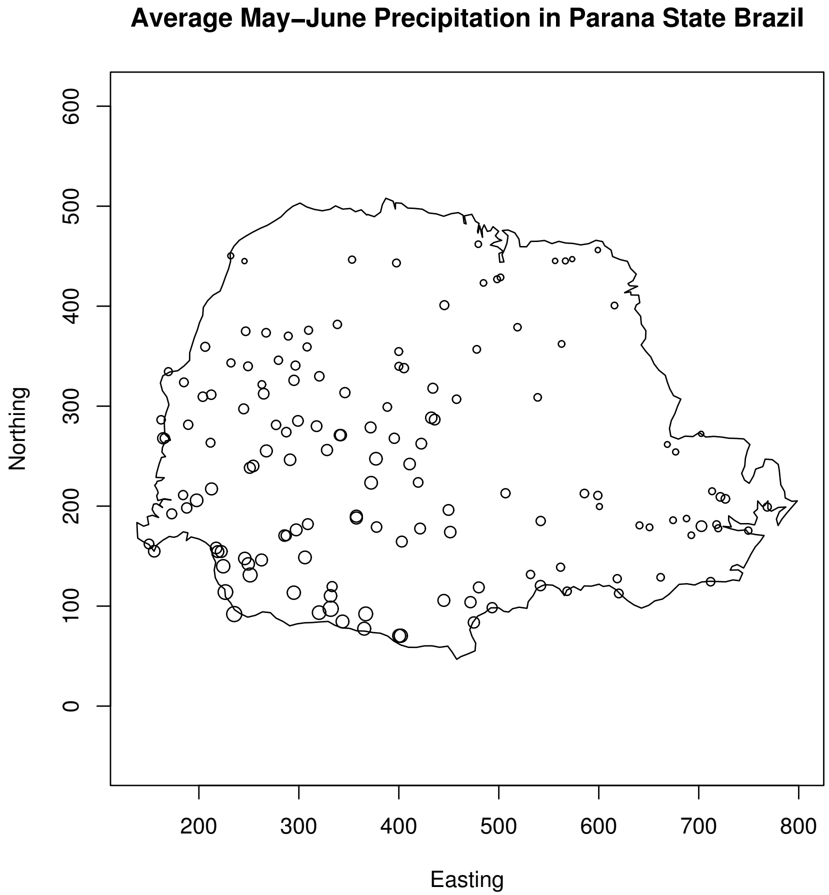
- The typical geostatistical model assumes some simple transformation of the trend to be a linear combination of spatial coordinates and covariates and then the random component to be a stationary Gaussian (normally distributed) **Markov Random Field** (MRF).

- Informally, MRFs are spatial processes where the conditional distribution at any point given the values at all other points is Gaussian with known mean and variance dependent only upon those neighbouring values and the spatial separation from each of them.

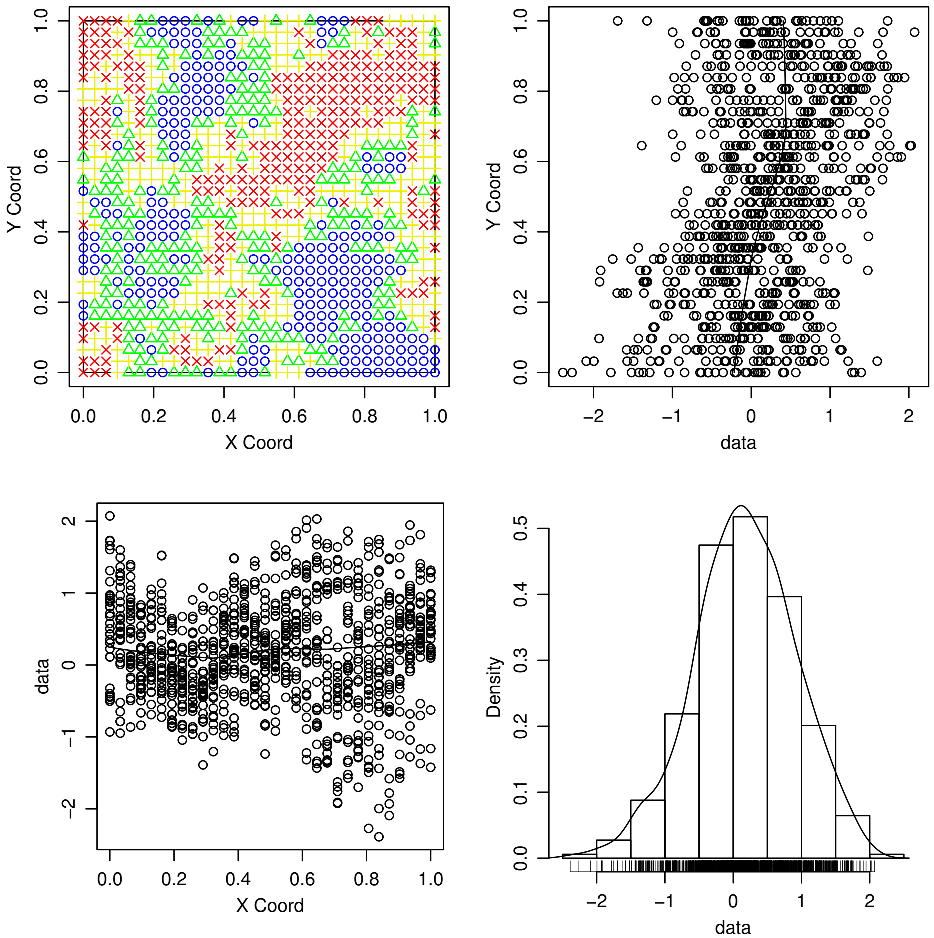
Simulated random field on a regular grid



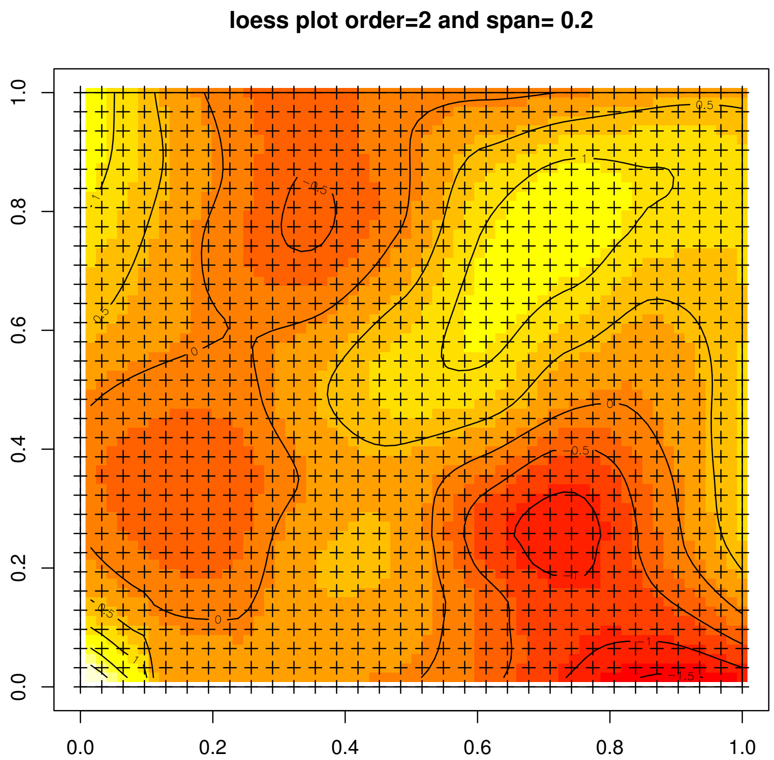
Precipitation in Parana State, Brazil



Exploratory analysis of simulated random field example

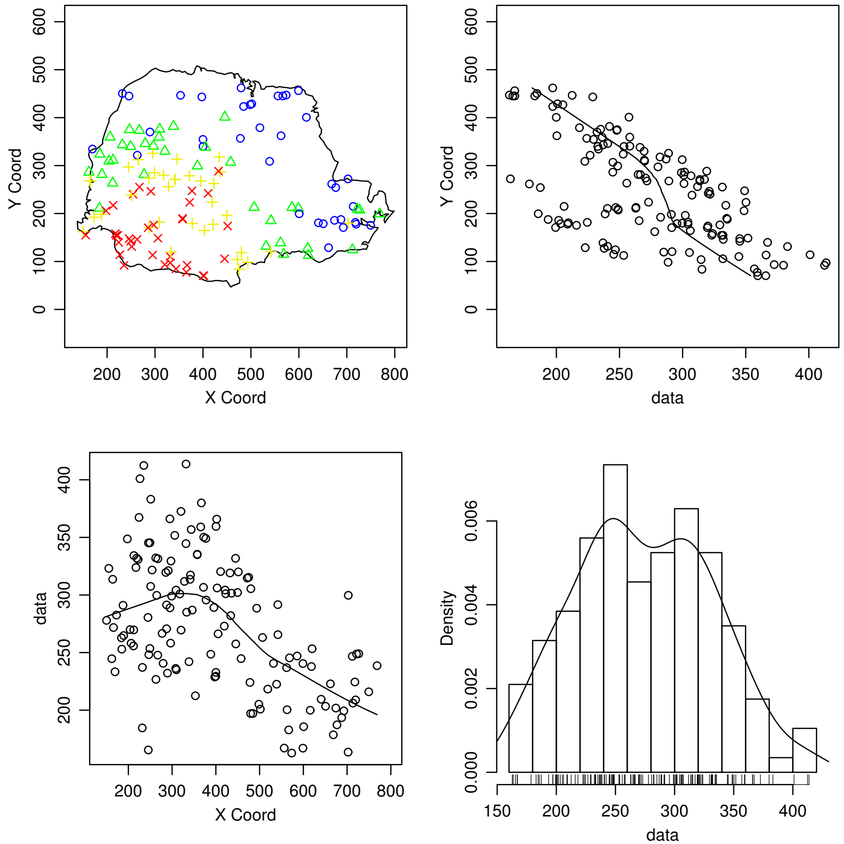


Localised regression applied to simulated random field example

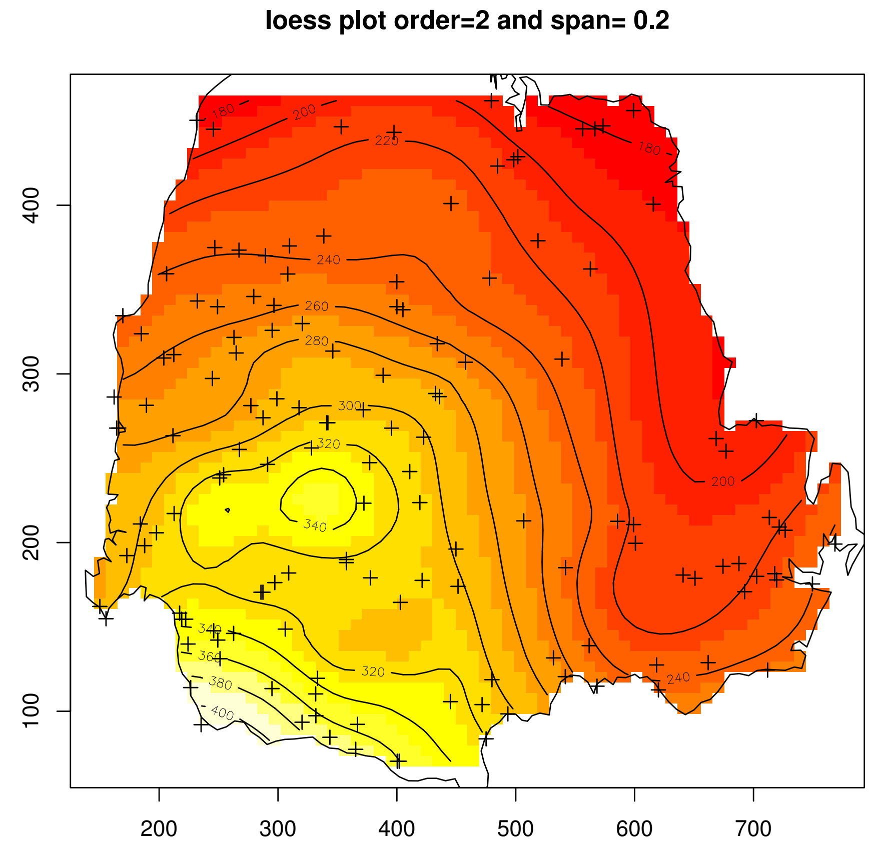


Conclusion might be — no suggestion of simple deterministic spatial structure in $\mu(\boldsymbol{s})$, possibly a constant mean process?

Exploratory analysis of dry season precipitation in Parana State

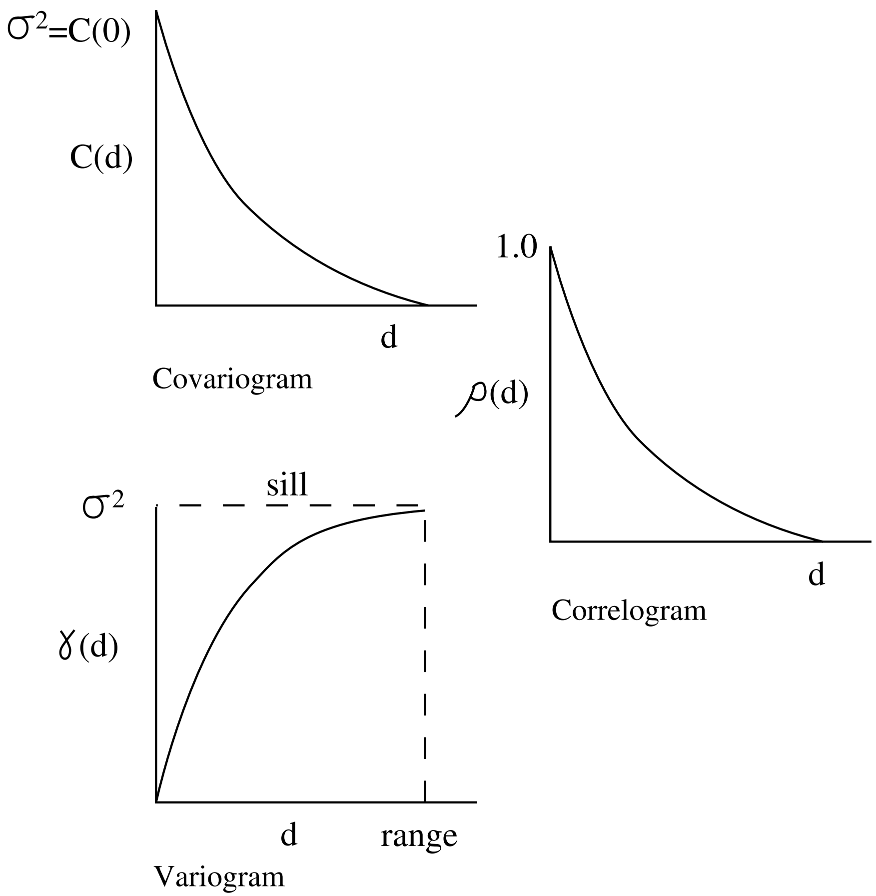


Localised regression of dry season precipitation in Parana State



Conclusion might be — SW trend in $\mu(\boldsymbol{s})$?

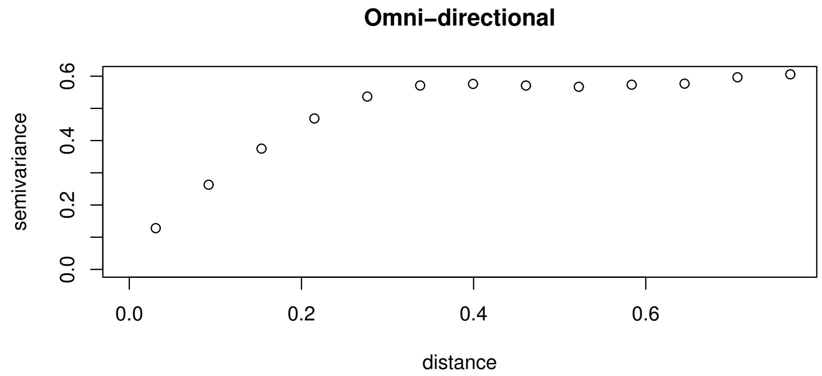
Variogram, Covariogram, Correlogram

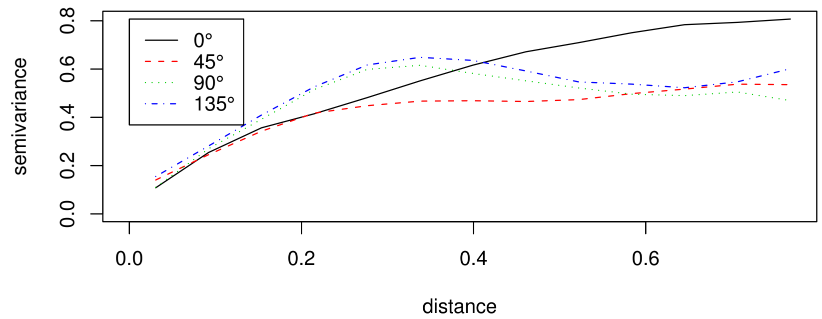


- For stationary process variogram should rise to an upper bound, the *sill* corresponding to underlying variance of the process, σ^2.
- Distance at which this occurs referred to as the *range*

- For stationary process variogram should rise to an upper bound, the *sill* corresponding to underlying variance of the process, σ^2.
- Distance at which this occurs referred to as the *range*
- Note theoretically $\gamma(0) = 0$, but sample values with small separations may be quite dissimilar—*nugget effect*.

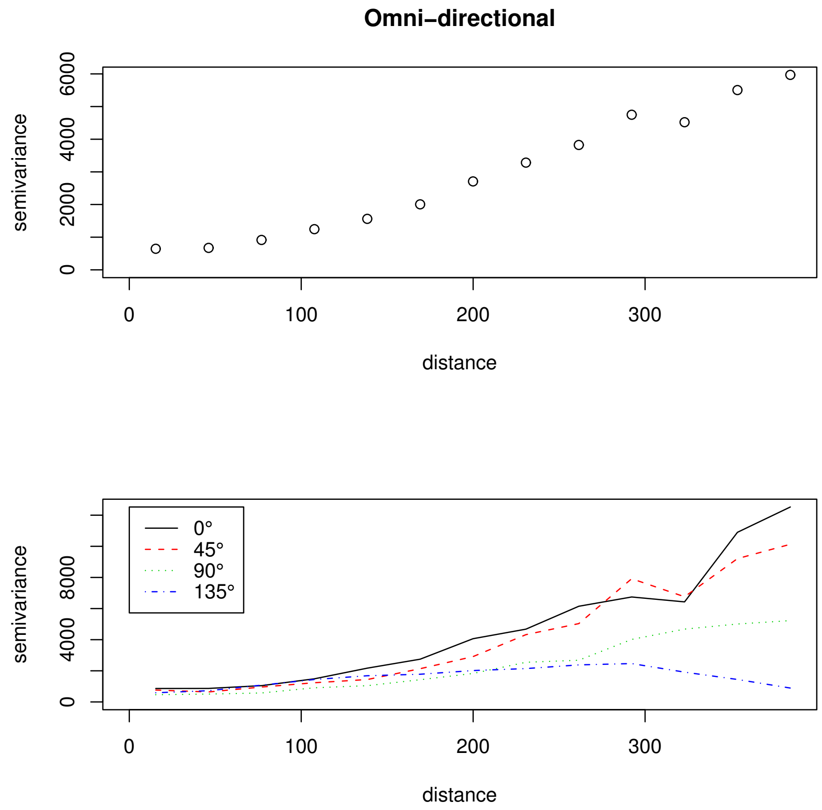
Sample variograms for simulated random field example





Strong spatial (isotropic) correlation structure, limiting behaviour typical of a stationary process,

Sample variograms for Parana State precipitation



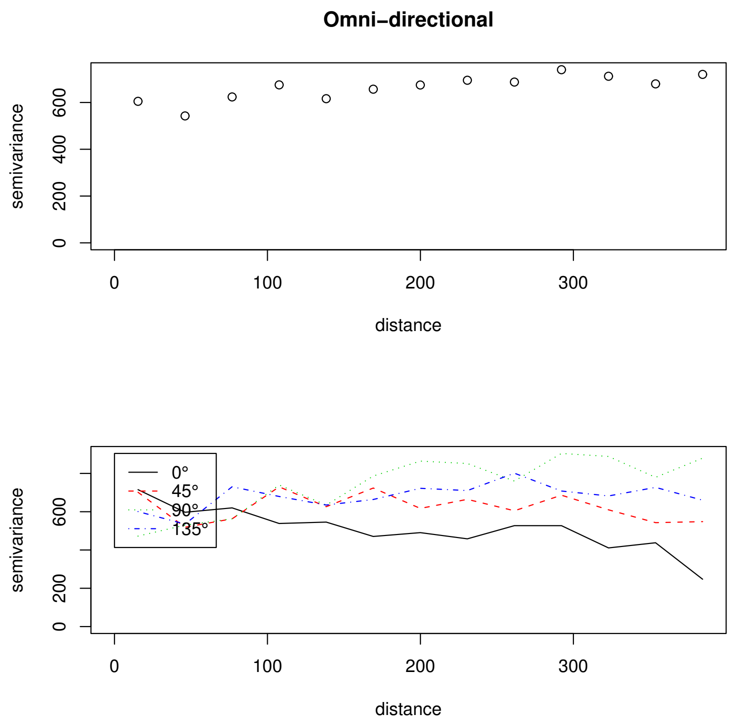
Limiting behaviour typical of a non-stationary process, consistent with earlier conclusions on mean behaviour

- If justified from exploratory analysis, would normally propose a **simple** trend surface to relate $\mu(\boldsymbol{s})$ to a function of $\boldsymbol{s}_i$ (and additional covariates). In the simplest case (where exploratory analysis indicates a stationary mean) the first order model would just comprise a constant mean

- If justified from exploratory analysis, would normally propose a **simple** trend surface to relate $\mu(\boldsymbol{s})$ to a function of $\boldsymbol{s}_i$ (and additional covariates). In the simplest case (where exploratory analysis indicates a stationary mean) the first order model would just comprise a constant mean
- The residuals from this trend can then be considered for modelling second order behaviour.

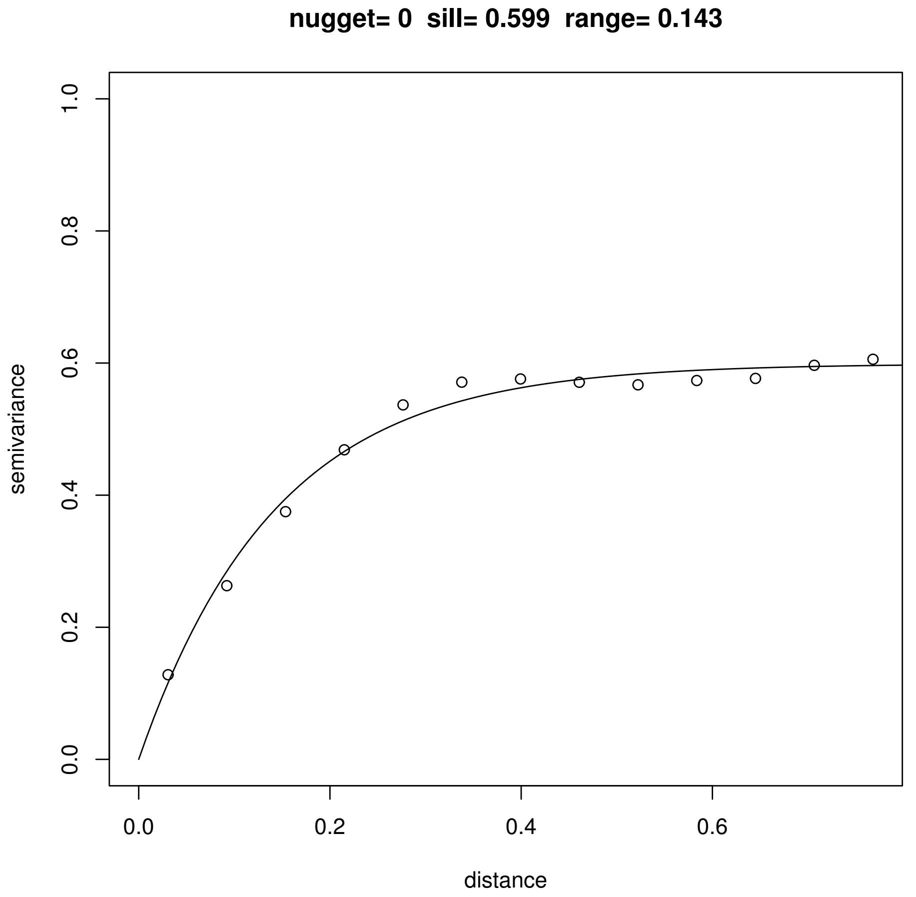
Sample variograms for residuals of Parana State precipitation

If we remove a quadratic trend in the eastings and the northings, then the residuals have a sample variogram:

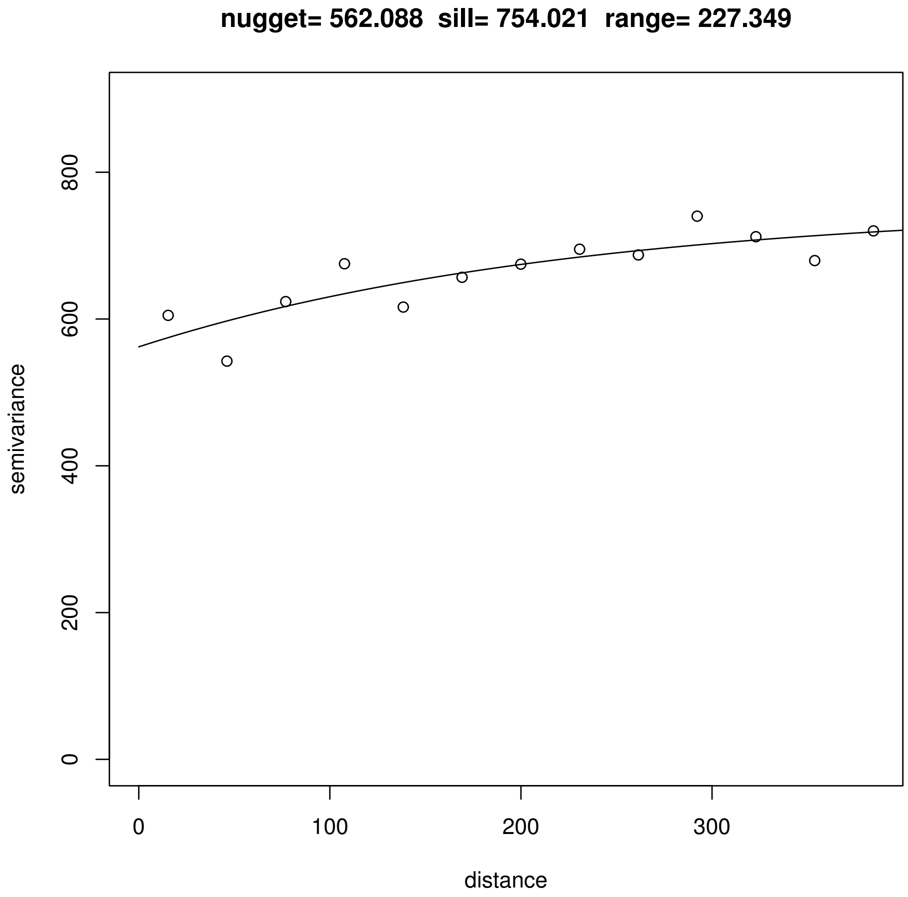


Now suggests weak spatial correlation structure in any direction

Fitted Variogram for simulated random field example (constant mean)



Variogram model of residuals of Parana precipitation (quadratic trend)



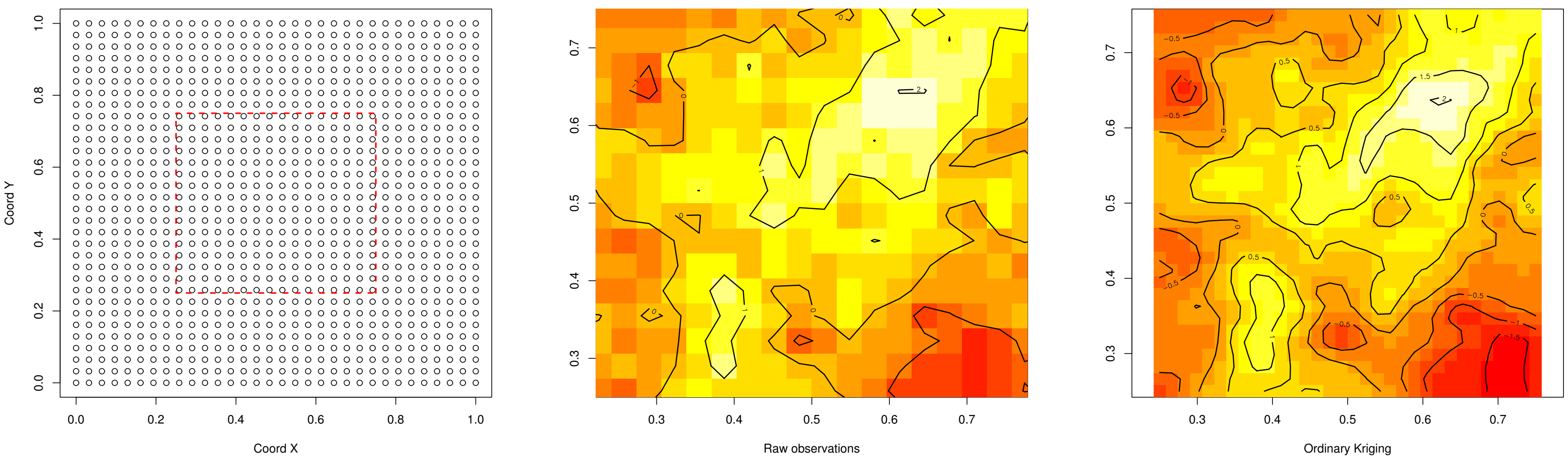
- To form the overall basic geostatistical model, incorporating both first and second order effects, we use a a hierarchical structure:

$$\begin{aligned} y_i &\sim p(\mu_i, \sigma^2) \\ g(\mu_i) &= \beta_0 + \sum_{k=1}^{K} \beta_k x_{ik} + \theta_i \end{aligned}$$

  where $p(\cdot)$ is a suitably chosen probability distribution, $g(\cdot)$ is a suitably chosen 'link function', β_k $(k = 0, \ldots, K)$ are fixed (mean) effects and θ_i are spatially correlated zero mean random effects which are spatially structured i.e. they have a stationary gaussian MRF.

- A hierarchical Bayesian approach, implemented through *Markov Chain Monte Carlo* (MCMC) methods, is used to fit this model. Although in the simple case where $p(\cdot)$ is gaussian and $g(\cdot)$ is the identity function that is not neceassry ('classical kriging').

Geostatistical prediction of simulated random field



Geostatistical prediction of Parana state precipitation