Incorporating Spatial Analysis into Agricultural Field Experiments.

Julia Piaskowski

2019-07-26

Contents

Chapter 1

Preface

1.1 Tutorial goals

To help people conducting planned agricultural field trials to understand and incorporate spatial variation routinely into analysis of field trials.

current resources are focused on geospatia appications - LANDSAT data across large area, point information on soils and geology, and other information that typiically requiresa a moderate to deep understanding of obht GIS tools and spatial analytica techniques.

Furthermore, there is not a comprehensive one-stop-shop for spatial analytical techniques for fied experiments that is also freely available.

1.2 Prerequisites

In order to run the scripts in this demonstration, you will to download R, avaiable free through the Comprehensive R Archive Network. While this is sufficient for running R scripts, You may also find it hepful to use RStudio, which provides a nice graphical user interface for R. RStudio can be downloaded here.

If you already have R installed, please make sure you have version 3.5.0 or newer.

This demonstration is not intended to provide instructions on general R usage. However, There are numerous web resources for learning the Basics of R:

-
-
-

-

1.3 Packages we will use

1.4 Acknowledgements

1.5 License

1.5.0.1 OLD STUFF

This is a *sample* book written in **Markdown**. You can use anything that Pandoc's Markdown supports, e.g., a math equation $a^2 + b^2 = c^2$.

The **bookdown** package can be installed from CRAN or Github:

```
install.packages("bookdown")
# or the development version
# devtools::install_github("rstudio/bookdown")
```

Remember each Rmd file contains one and only one chapter, and a chapter is defined by the first-level heading `#`.

To compile this example to PDF, you need XeLaTeX. You are recommended to install TinyTeX (which includes XeLaTeX): https://yihui.name/tinytex/.

Chapter 2

Introduction

2.1 Why care about spatial variation?

Goal in this case is to understand and account for spatial variation. These are lattice data

2.2 Diagnosing spatial autocorrelation

Spatial correlation is similarity of plots that are close to one another. Larger spatial gradients need to modelled indpendently - perhaps through blocking.

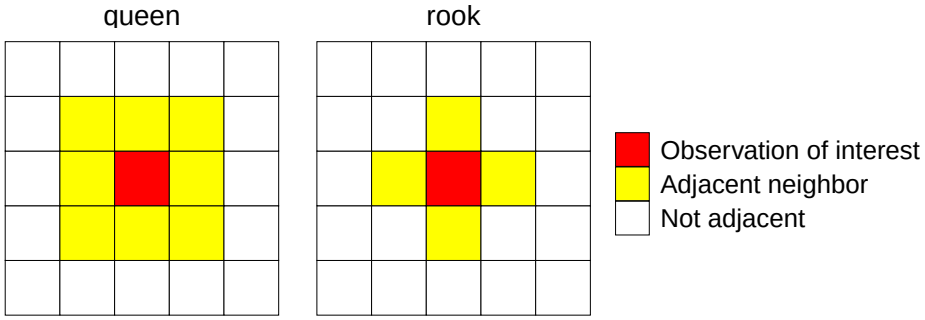
2.2.1 Moran's I

Moran's I, sometimes called "Global Moran's I" is similar like a correlation coefficient. It is a test for correation between units (plots in our case).

$$I = \frac{N}{W} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2} \qquad i \neq j$$

Where N is total number of spatial locations indexed by i and j, x is the variable of interest, w_{ij} are a spatial weights between each i and j, and W is the sum of all weights. The expected values of Moran's I is $-1/(N-1)$. Values greater than that indicate positive spatial correlation (areas close to each other are similar), while values less than the expected Moran's I indicate dissimilarity as spatial distance between points decreases.

There are several options for defining adjacent neighbors and how to weight each neighbors's influence. The two common configuations for defininig neighbors are the rook and queen configurations. These are exactly what their chess analogy suggest: "rook" defines neighbhors in an row/column fashion, while "queen" defines nieghbors in a row/column configuation an also neighbors located diagonally at a 45 degree angle from the row/column neighbors. Determining this can be somewhat complicated when working with irregularly-placed data (e.g. county seats), but is quite unambiguous for gridded data commonly seen in panned field experiments:



Another test for diagnosing spatial correlation is Geary's C:

$$I = \frac{(N-1)}{2W} \frac{\sum_i \sum_j w_{ij}(x_i - x_j)^2}{\sum_i (x_i - \bar{x})^2} \qquad i \neq j$$

These terms have the same meaning in Moran's I. The expected value of Geary's C is 1. Values higher than 1 indicate positive spatial correlation and less than 1 indicate negative spatial correlation.

2.2.2 Empirical variogram & semivariance

An empirical varigram is a visual tool for understanding how error terms are related to each other over spatial distance. It relies on semivariance (γ), a statistic expressing variance as a function of pairwise distances between data points at points i and j.

$$\gamma(h) = \frac{1}{2|N(h)|} \sum_{N(h)} (x_i - x_j)^2$$

Semivariances are binned for distance intervals. The average values for semivariance and distance interval can be fit to correlated error models such a exponential, spherical, gaussian and Matér. How to dothis is explored further in section 2 of this guide.

Three important concepts of an empirical variogram are *nugget*, *sill* and *range*

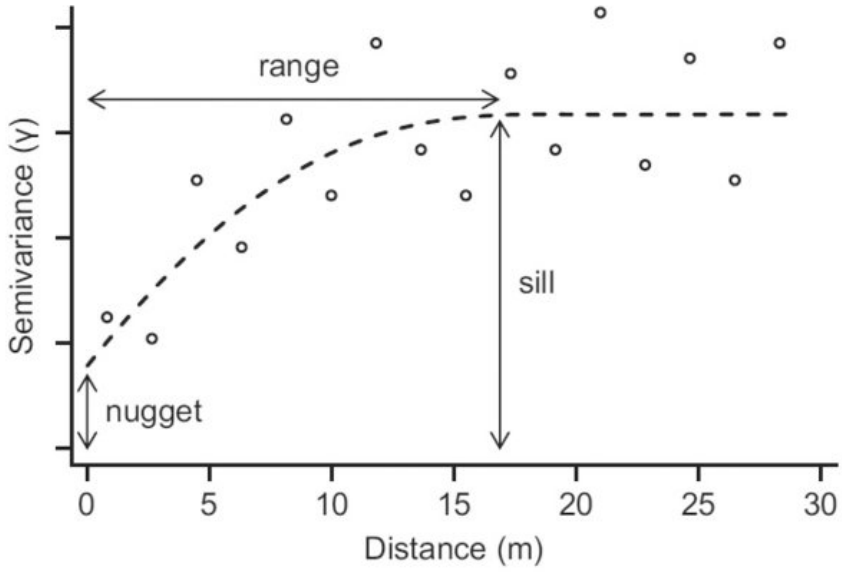


Figure 2.1: Example Empirical Variogram

- range = distance up to which is there is spatial correlation
- sill = uncorrelated variance of the variable of interest
- nugget = measurement error,short-distance spatial variance and other unaccounted for variance

2 other concepts:

- partial sill = sill - nugget
- nugget effect = the nugget/sill ratio, interpreted opposite of r^2

2.2.2.1 OLD STUFF

You can label chapter and section titles using `{#label}` after them, e.g., we can reference Chapter **??**. If you do not manually label them, there will be automatic labels anyway, e.g., Chapter **??**.

Figures and tables with captions will be placed in `figure` and `table` environments, respectively.

```
par(mar = c(4, 4, .1, .1))
plot(pressure, type = 'b', pch = 19)
```

Reference a figure by its code chunk label with the `fig:` prefix, e.g., see Figure **??**. Similarly, you can reference tables generated from `knitr::kable()`, e.g.,

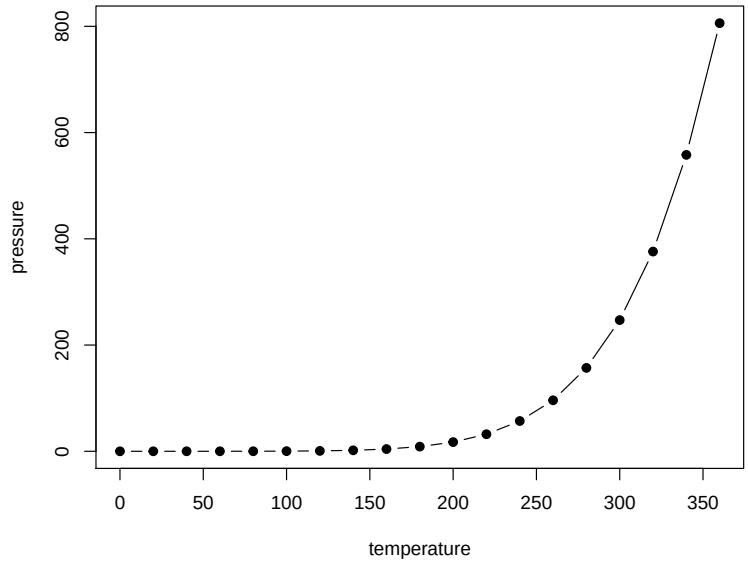


Figure 2.2: Here is a nice figure!

see Table **??**.

```
knitr::kable(
  head(iris, 20), caption = 'Here is a nice table!',
  booktabs = TRUE
)
```

You can write citations, too. For example, we are using the **bookdown** package (**?**) in this sample book, which was built on top of R Markdown and **knitr** (**?**).

Table 2.1: Here is a nice table!

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa
5.4	3.7	1.5	0.2	setosa
4.8	3.4	1.6	0.2	setosa
4.8	3.0	1.4	0.1	setosa
4.3	3.0	1.1	0.1	setosa
5.8	4.0	1.2	0.2	setosa
5.7	4.4	1.5	0.4	setosa
5.4	3.9	1.3	0.4	setosa
5.1	3.5	1.4	0.3	setosa
5.7	3.8	1.7	0.3	setosa
5.1	3.8	1.5	0.3	setosa

Chapter 3

Spatial Models {ch2-bg}

This section contains some of the background behind why spatial models are used and how they work. Understanding this section is not essential, but it is extremely helpful. This section relies on information introduced in part I, s please make sure you have read that section, especially if you are new to empirical variograms and spatial statistics.

General linear statistical models are commonly modeled as thus:

$$Y_i = \beta_0 + X_i\beta_1 + \epsilon_i$$

β_1 is a slope describing the relationship between a continuous variable and the dependent variable, Y_i. If X_i is a categorical variable, such as a crop variety, then there will be $p - 1$ slopes estimated, where p is the number of unique treatements levels in X.

The error terms, ϵ_i are assumed normally distributed with a mean of zero and a variance of σ^2 :

$$e_i \ \sim N(0, \sigma^2)$$

The error terms, or residuals, are assumed to be *identically* and *independently* distributed (abbreviated as "iid"). This implies a constant variance of the error terms and zero covariance between residuals.

If N = 3, the expanded model looks like this:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \beta_0 + \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \beta_1 + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{bmatrix}$$

$$e_i \sim N\left(0, \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix}\right)$$

If spatial variation is present, the off-diagonals of the variance-covariance matrix are not zero - hence the error terms are not independently distributed. As a result, hypotheses test and parameter estimates from uncorrected linear models will provide erroneous results.

3.1 Correlated Error Models

3.1.1 Distance-based Correlation Error Models

There are mathematical tools for modelling how error terms are correlated with each other based on pairwise physical distance between observations. These models can be used to weight observations. Often, the data are assumed to be *isotropic*, where distance but not direction impacts error spatial correlation

There are several methods for estimating the semivariance as a direct function of distance.

3.1.1.1 Exponential Model

$$\gamma(h) \begin{cases} 0 & h = 0 \\ C_0 + C_1 \left[1 - e^{-(\frac{h}{r})}\right] & h > 0 \end{cases}$$

where

$$C_0 = nugget C_1 = partial\ sill r = range$$

$3r = r_p$ is the "practical range", which is 95% of the true value for C_1.

3.1.1.2 Gaussian

(a squared version of the exponential model)

$$\gamma(h) \begin{cases} 0 & h = 0 \\ C_0 + C_1 \left[1 - e^{-(\frac{h}{r})^2}\right] & h > 0 \end{cases}$$

where

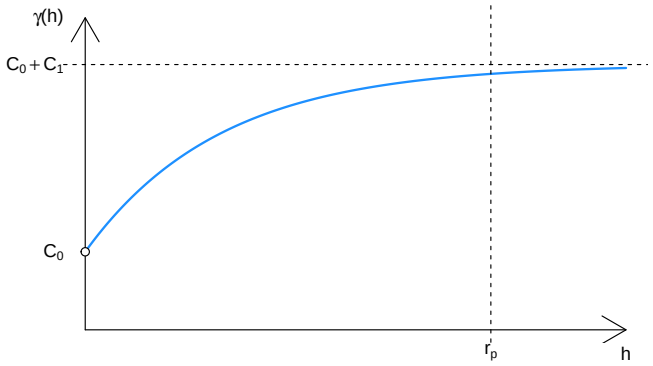


Figure 3.1: Exponential Model

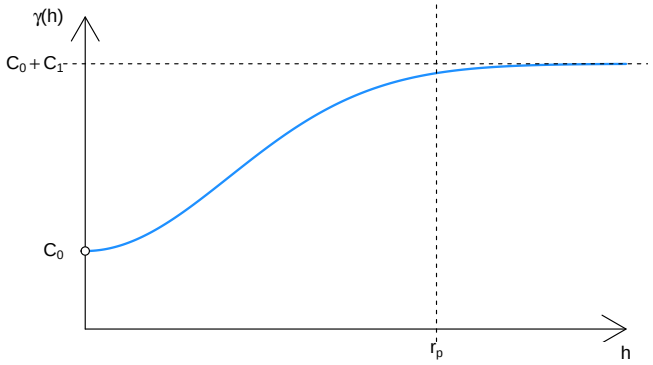


Figure 3.2: Gaussian Model

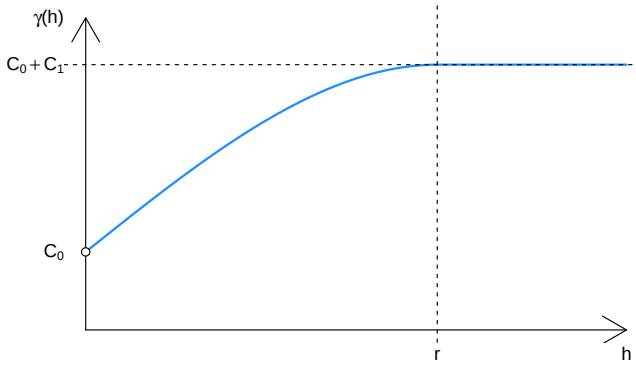


Figure 3.3: Spherical Model

$$C_0 = nuggetC_1 = partial\ sillr = range$$

$\sqrt{3}r = r_p$ is the "practical range", which is 95% of the true value for C_1.

3.1.1.3 Spherical model

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ C_0 + C_1 \left[\frac{3h}{2r} - 0.5 \left(\frac{h}{r} \right)^3 \right] & 0 < h \leq r \\ C_0 + C_1 & h > r \end{cases}$$

where

$$C_0 = nuggetC_1 = partial\ sillr = range$$

3.1.1.4 Other correlated error distance models

There are many more models - Matérn, Cauchy, Logistic - that may describe spatial correlation in a data set.

There are two addition models that have no range or sill, the linear model and power model. If your data fits these, consider doing a trend analysis.

3.1.1.5 Linear model

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ C_0 + C_1 h & h > 0 \end{cases}$$

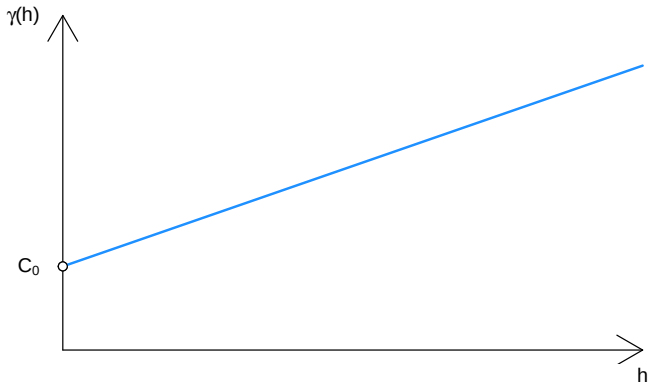


Figure 3.4: Linear Error Model

where

$$C_0 = nugget C_1 = slope$$

There is no sill or range in the linear model, so the variance will continue to increase as a function of distance.

3.1.1.6 Power Model

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ C_0 + C_1 h^{\lambda} & h > 0 \end{cases}$$

where

$$0 \leq \lambda <\leq 2 C_0 = nugget C_1 = scaling\ factor$$

When $\lambda = 1$, that is equivalent to the linear model. Example above is when $\lambda = 0.5$ (i.e. a square-root transformation) and $C_1 = 1$. There is also no sill or range in the power model.

3.1.2 Correlated Error Models for Gridded Data

Planned field experiments often have the advantage of being arranged in regular grid pattern that can be adequately described using Euclidian space. This simplifies aspects of understanding how error terms are related by distance since the data occur in evenly spaced increments. Furthermore, in many agricultural trials, there may be no interest in spatial interpolation between units. Some of these models work with irregular data, but the models presented are simplified forms when experimental units are arranged in regular grid.