Spatial Data Analysis in R

Dealing With Spatial Dependence 1
Overview

Eco 697DR – University of Massachusetts, Amherst – Spring 2022
Michael France Nelson

Plan for next 3 lectures

Today

- Review of dual model paradigm and regression models

Tuesday

- Survey of spatially-aware regession methods

Next Thursday

- Spatially-aware regression examples

Course Wiki

Reminder:

- If you haven’t yet contributed to the course wiki on Moodle, you should do so soon!

Preview of chapter 6: dealing with spatial dependence

- What is the primary statistical problem posed by spatial dependence?
  - Non-independent observations
- How could we deal with it? Some possibilities:
  - Ignore
  - Interpret dependence as topic of interest
  - Spatially-aware regression

Avoiding Spatial Dependence: Sampling Design

Simple options?

- Ignore spatial dependence
  - observed dependence/correlation is low
- Avoid spatial dependence
  - Design sampling scheme

Sampling design for avoiding spatial dependence:

- You can use a correlograms or variograms to guess a critical distance, above which spatial dependence does not occur.
- Simply space your sampling locations greater than this critical distance.
- Any challenges or problems with this approach?

Methods for dealing with spatial dependence

We often want to do inference and/or prediction:

- spatially-aware regression

Some considerations for spatial dependence in regression-like models include:

- Consider exogenous and endogenous factors
- Consider dependence in responses and predictors
- Consider dependence in residuals
- Consider model structure

Endogenous and exogenous factors

What are some potential exogenous contributors to spatial dependence?
What are some potential endogenous spatial contributors to spatial dependence?

- Consider two general classes of techniques:
  - coordinate-based
  - distance-based

Coordinate and distance paradigms

- What are the fundamental conceptual differences?
- How does each paradigm consider space?

Coordinate-based models

We can use the explicit spatial (x, y) coordinates in models via:

- Polynomial model terms of the spatial coordinates
- Fourier or wavelet methods
- Eigenvector mapping techniques

These may be effective for large spatial scale exogenous factors.

- examples?

When might projections matter?

Additive models

General Additive Models - GAMs

Local regression

- Distance weighting
- Splines and knots

Distance-based models

We can define distance in many ways, including

- Euclidean
- Neighborhood

Neighborhoods

- first order neighbors
- larger neighborhoods
- distance-decay function

Implementation via distance and weight matrices

Spatial dependence in factors

Spatial dependence can occur in the

- Predictors
- Responses
- Model residuals

How do we consider each?

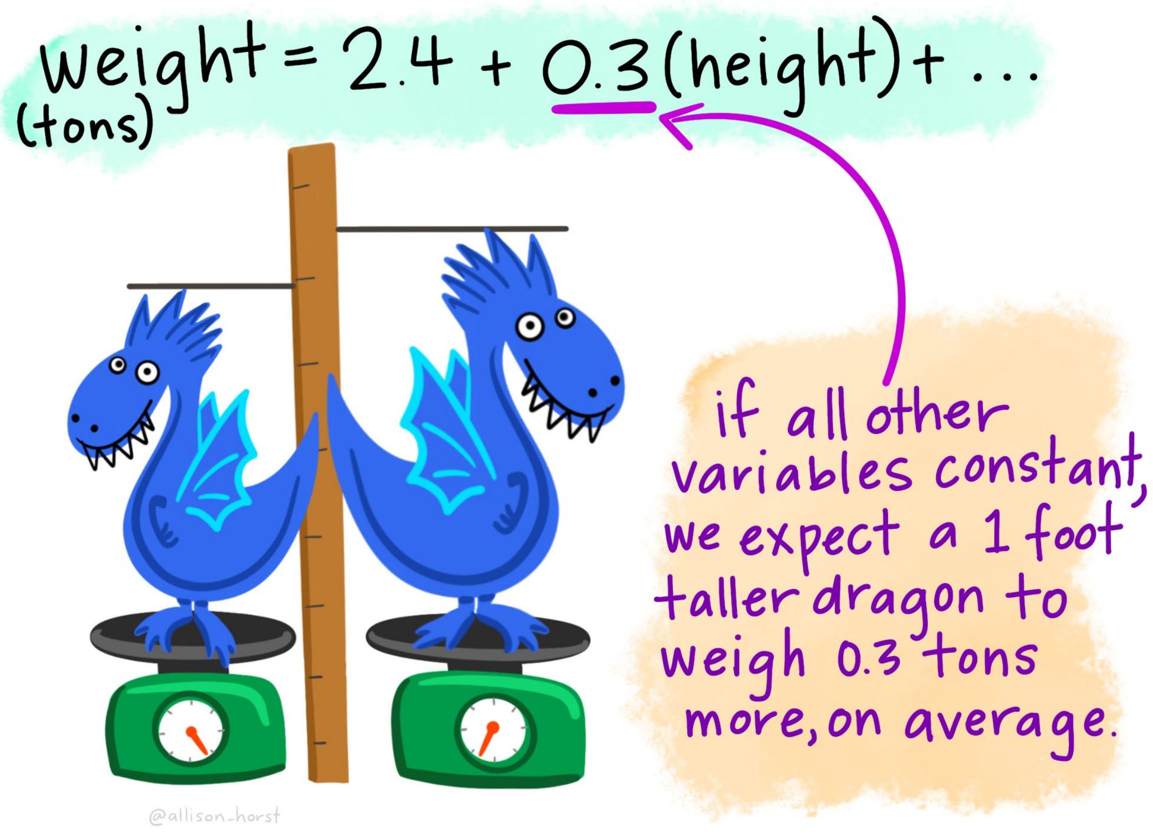
Regression and the Dual Model Paradigm

What is a Regression?

Regressions embody the dual-model concept

Regression is a modeling paradigm in which we specify a mathematical relationship between independent and dependent variables.

- A regression includes a *deterministic model* to specify the average behavior.
- It specifies a *stochastic model* to describe the variability around the average behavior.



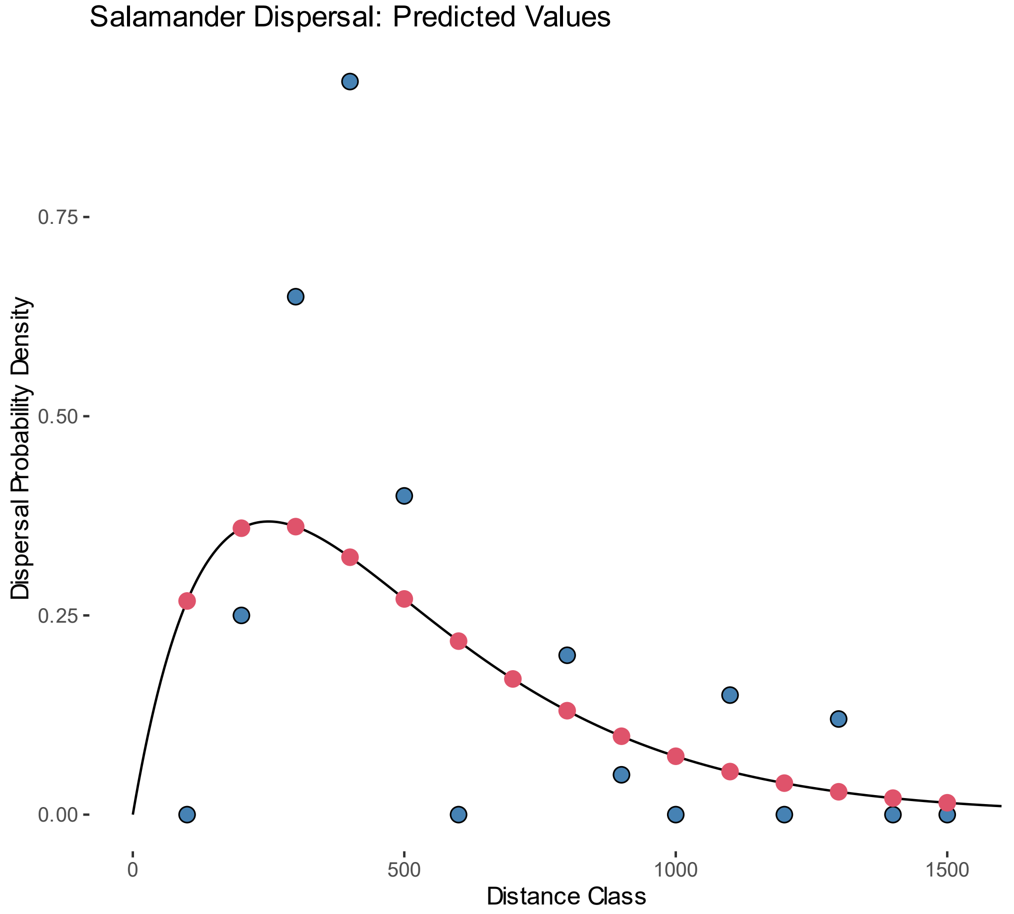
Artwork by @allison_horst

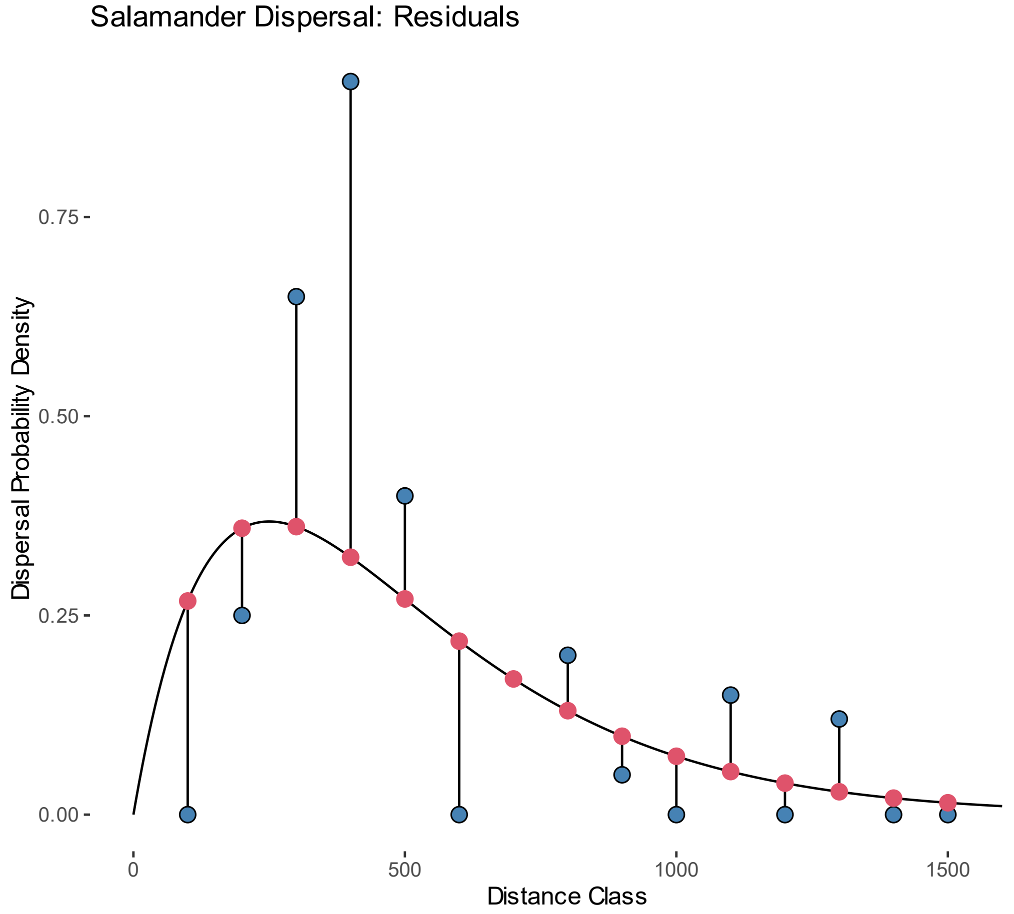
What is the Dual Model Paradigm?

Deterministic Model	Stochastic Model
• The outcome of a deterministic process is always the same, there's no uncertainty. • We can use mathematical functions to model a deterministic process. • For example: a linear equation	• A stochastic process features uncertainty in its outcome. • Every *realization* of a stochastic process has a different outcome. • We can use a stochastic model to understand uncertainty. • Stochastic models are often described by probability distributions.

Model Residuals: Salamander Dispersal

Deterministic Model: The Predicted Values	Stochastic Model: The Residuals!





Dual Models

Deterministic Model: The Regression Line	**Stochastic Model: The Residuals**
The regression line describes the model's **predicted values**. We don't expect that the observed values fall exactly on the regression line.	The residuals are the variation in the response that the model can't explain. In a linear regression, we assume that the residuals are Normally distributed.

What is regression?

Let’s step back and consider a simple, yet deep questions:

- What is regression?
- Conceptually simple goal: fit a *mathematical functions* to data to gain insight. Two main goals:
  - Inference: Using the form of the functions for means and variations to gain insight about the larger population.
  - Prediction: Using the form of the functions for prediction.
- Concept is simple: details can be tricky

Key regression concepts

- Form of the mean model
- Form of the error model
- Predictor and response variables
- Sums of squares, least squares, likelihood
- Measures of **center** and **spread**
- Assumptions
- Model diagnostics

Two Models

The Means Model: Response Curves	The Error Model: Sources of Error

The Means Model: Response Curves

- The shape of the curve of the predictor/response relationships
- Doesn’t have to be linear
- Options include:
    - Linear
    - Power functions: integer, rational, real exponents
    - Exponential: fixed base, power is variable
    - Logarithmic: diminishing returns
    - Hybrid: e.g. Holling functional response curves

The Error Model: Sources of Error

- **Error** is an unfortunate term
- Error is the unexplained variability in the model
- Some possible sources:
    - Lurking variables: model mis-specification
    - Measurement error: imperfect observations
        - response
        - predictors
    - Process/system error: inherent variability

Dual models

Fitting a regression model means fitting two separate models:

- Function to describe means (deterministic model)
- Function to describe noise/variability (stochastic model)
- Fitting models
  - Means model is often the easier one to fit
  - Error model is often trickier

Observations: N and DF

- Inference is associated with degrees of freedom (DF)
- Number of observations: how many sampling units did you observe
  - Defining SUs might not be as straightforward as you think
- Effective degrees of freedom
  - Are the SU observations independent?
- Experimental design
  - Analysis of Variance - categorical variables: balanced design?
  - Sampling units
- Hierarchical structure, autocorrelation
  - May reduce information content
  - Effective DF

Regression Types: A Non-Exhaustive Summary

Different classes of regression models have been devised to accommodate various data types and relationship structures. Some of the most familiar include:

- Linear Models
- Generalized Linear Models
- Random Effect and Mixed Models
- Generalized Least Squares

Key regression concepts

- Predictor and response variables: data types
- Sums of squares, least squares, likelihood
- Measures of **center** and **spread**
- Assumptions and when they can be relaxed
- Independent observations
- Model diagnostics
- Function to describe means (deterministic model)
  - Form of the mean model
- Function to describe noise/variability (stochastic model)
  - Form of the error model
- Fitting models
  - Means model is often the easier one to fit
  - Error model is often trickier

Linear Models

Simplest class

Normal distribution: theoretical basis

- two-parameter distribution
- Variance does not depend on the mean
- Variance is constant for all levels/ranges of predictors

Form of the regression equation is **linear in the predictors**.

- This concept can be confusing
- Inferred model parameters can only multiply the predictors
- Response doesn’t have to be linear: certain forms of nonlinearity are allowed

Linear Models

The regression equation:

$$response = \alpha + \beta_1 \times predictor1 + \beta_2 \times predictor2 + \ldots. + error$$

- Linearity means that the beta terms can only multiply the predictors, and their values don't depend on the specific values of the predictors.
- The predictors may be modified by functions, but the predictor function parameters must be fixed, i.e. **inference** is not performed on the predictor expressions.
  - Polynomial, exponential, logarithmic terms are allowed.
  - Must have constant exponent, bases, powers.

Linear Models: Key Concepts and Assumptions

Assumptions arise from Normal-distribution probability theory

- Independence of observations
- Assumptions for errors:
    - Errors are independent of one another
    - Errors are **identically** distributed: there is constant variance

Transformations

- Response and/or predictors can be transformed to make the relationships **linear** in the predictors
- Transformations can help stabilize variance

Response is numeric and [ideally] continuous

Linear Models: Normality

A frequent misconception:

"Your data have to be Normal to use linear regressions."

- This is misleading.
- The responses are assumed to be Normal **at each value of the predictors**.
- In other words, the residuals need to be Normally distributed

Elaborating the Linear Model

The Constellation

Regression Models

The constellation of statistical models includes many paradigms.

Some commonly used regression types, each devised to accommodate different sets of assumptions, data structures, and other factors include:

- Linear Models
- Generalized Linear Models
- Mixed Models, Generalized versions
- Generalized Least Squares
- Additive Models

What are some of the key properties of each of these classes?

Generalized Linear Models

Generalized Linear Models are an elaboration of Linear Models. They are well suited to certain situations that pose issues for linear models:

- GLMs relax the requirement that residuals/errors must be Normally distributed
  - Residuals must still be independent and identically distributed
- Error distribution must belong to a member of the **exponential** family of distributions.
  - This family contains **many** common discrete and continuous distributions, including the Normal.
  - Fitting of non-normal distributions is via a **link function.**
  - Response is modeled in terms of a **mean function.**
  - Mean function is the inverse of the

GLMS: Uses

- GLMs can better describe discrete outcomes like counts or presence/absence.
- Some pros and cons:
  - A very flexible and useful class of models
  - Coefficients are less intuitive to interpret than LM
  - Models are easy to fit in R
  - Model diagnostics may be more difficult than LM
  - Great for discrete data
  - Retains the independence of observations assumption

Random Effects

Question: What is a fixed effect?

Another question: What is a random effect?

Yet another: How do we tell the difference?

NOTE: the distinction is not always straightforward!

Random Effects accommodate hierarchical structure in our data.

What are some examples?

Random and Fixed Effects

This is a very simplified description of the differences:

Fixed Effects:	**Random Effects:**
• Are what you want to do inference on. • Usually the focus of your research question. • You are interested in coefficients for the **specific** coefficients. • You are interested in the variability	• Represent hierarchical or grouping structures in the data. • Factors in data/system that you want to 'control for', but don't care about the specific observed levels. • Can model groups in experimental design: • Blocks, Latin Squares, etc.

Mixed Effects Models

"Welcome to our world, the world of mixed effects modelling. The bad news is that it is a complicated world. Nonetheless, it is one that few ecologists can avoid, even though it is one of the most difficult fields in statistics." (Zuur et al., 2009)

- They are really powerful, but also complex.
- Mixed effects can help account for some types of spatial autocorrelation.
- Mixed effects models are more complicated to implement and interpret.
- Simulation, or other non-analytically methods may be needed to estimate model parameters.

Generalized Least Squares

What if my data/errors aren’t independent?

What if I can’t get rid of heteroskedasticity?

(What does that mean?)

All of the previous models had restrictive assumptions about the errors.

Sometimes we can’t massage our data to fit the other frameworks.

GLS: Errors and Variance/Covariance

Other models may obfuscate the **variance/covariance matrix** concept.

- Independent, identically-distributed errors: we can simplify to a single number.

In reality, we have a:

- Variance/covariance matrix:
  - Variance on the diagonal: may be all equal
  - Covariance on off-diagonals

GLS works by estimating the variance/covariance matrix

Spatially-Aware Regression

- Finally!
- The problem:
  - autocorrelated data
- Two possible solutions:
  - Include spatial covariates
    - Additional fixed effects
    - Random effects
    - Explicitly spatial covariates: trend surface, etc.
  - Model a spatially-aware variance/covariance structure (GLS)

Spatial Autocorrelation: Regression Perspective

Autocorrelation can be caused by:	**We can accommodate autocorrelation**
Model mis-specification • Unavailable covariates are not included: covariates we're not able to measure or are unaware of • Available, yet missing covariates: covariates that we know about, but don't have data for • Incorrect understanding of system Real spatial dependence	• Model of the means • Include additional covariates • Incorporate a trend surface • Include linear or polynomial functions of the coordinates • Include autocovariates • Model of the errors • include variance/covariance structures

Errors and Variance/Covariance

- Other models obfuscate the **variance/covariance matrix** concept.
- Independent, identically-distributed errors: we can simplify to a single number.
- Variance/covariance matrix:
  - Variance on the diagonal: may be all equal
  - Covariance on off-diagonals
- Weighted regression
  - Heteroskedasticity
  - diagonal elements are not all equal

Spatially-Aware Regression Workflow

A general workflow for spatially-aware regression

1. Data import, prep, exploration, cleaning
2. Examine autocorrelation in predictors and response.
3. Fit nonspatial model, examine autocorrelation in residuals
4. Propose spatially-aware options
    - non-spatial covariates such as:
        - random effects, polynomial terms, additional predictors
    - spatial covariates such as:
        - terrain/landscape covariates, spatial lag terms
    - variance/covariance structures
5. Fit nonspatial model, examine autocorrelation in residuals
6. Iterate 4 and 5 until you are satisfied

Spatial Autocorrelation: a Regression Perspective

Two important sources of autocorrelation

Model mis-specification	Real spatial dependence
• Unavailable covariates are not included • Available, yet missing covariates • Incorrect or incomplete understanding of system	• endogenous dependence • exogenous dependence

Spatial Autocorrelation: a Dual Model Perspective

What is the dual-model paradigm?

- Two important model components to address spatial autocorrelation
  - Deterministic model: covariates
  - Stochastic model: variance/covariance structures