Why Include Spatial Dependencies?

Mark Otto

U.S. Fish and Wildlife Service

15 March 2006

Statistical Models

Biased Estimates of Standard Errors

Geostatistical Spatial Models

Lattice Models

Experimental Design

Statistical Models
Biased Estimates of Standard Errors
Geostatistical Spatial Models
Lattice Models
Experimental Design

I want to start with the concepts of some basic statistical models. Concepts you are familiar with. Then, I want to build and apply them to models used on correlated data. I'll start with repetition and IID, extend it to regression models and transformation. We can then see how transformations are also used to handle correlations data taken over time and space. Finally, I will talk about the benefits and difficulties of correlated data in mapping and prediction, regression, and experimental design.

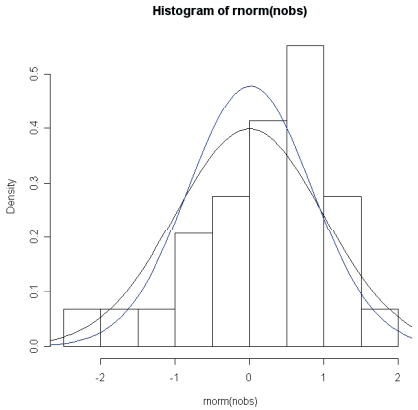
Statistical Models

- Variable data
- Consistent patterns:
  - Parameter estimates
  - Fitted values
  - Predictions
- Repetition
- Model assumptions

Statistical Models
- Variable data
- Consistent patterns:
  - Parameter estimates
  - Fitted values
  - Predictions
- Repetition
- Model assumptions

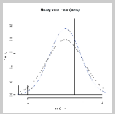
In statistics, we collect variable data that we hope to pull out consistent patterns from. Probability distributions were made to characterize common forms of variability: binomial, and multinomial for categorical data, Poisson data for counts, and with all data no matter what the distribution has estimates the approximate a normal distribution. To extract those patterns, they must repeat in the data many times. In practice we don't know what processes generated the data, so we have to pay attention to the assumptions of the models that we are using and check that the data do not deviate greatly from them

Data from a Normal



6

Data from a Normal



Here are 29 values from a standard distribution, the mean is 0 and the variance is 1. The histogram is of the data, the black line is from standard normal the sample was generated from. The blue line is the normal curve with mean and variance estimated from the sample data. The estimated curve is a reasonable approximation.

Independent, Identically Distributed

- Identically Distributed: each observation is sample from the same distribution
- Independent: One observation does not give information about preceding or succeeding draws

8

Independent, Identically Distributed
- Identically Distributed: each observation is sample from the same distribution
- Indepenent: One observation does not give information about preceding or succeeding draws

IID is one of the most basic assumptions we make for statistical modeling. In term of the covariance independence means no correlation among observations.

Regression

- Mean allowed to vary
- Variance constant

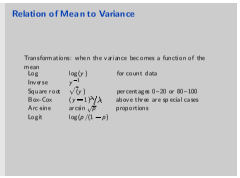
Regression
- Mean allowed to vary
- Variance constant

The first order effects vary, but the relation between the independent variables is constant, $\mathbf{y} = \mathbf{X}\beta$. The variance or second order effects are constant, $\mathbf{I}\sigma^2$. Their is repetition in the data: repetition around a varying mean according to a constant variance. Each observation gives information about β, that relates y to X. The residuals, the data after removing the mean function, are $N(0, \sigma^2)$.

Relation of Mean to Variance

Transformations: when the variance becomes a function of the mean

Log	$\log(y)$	for count data
Inverse	y^{-1}	
Square root	$\sqrt(y)$	percentages 0–20 or 80–100
Box-Cox	$(y-1)^{\lambda}/\lambda$	above three are special cases
Arc-sine	$\arcsin\sqrt{p}$	proportions
Logit	$\log(p/(1-p)$	



For data that does not fit the usual assumptions, we can build more flexible and complex models or we can transform the data back to something that fits the usual assumptions. Here we transform to remove the relation between the mean and variance. Back to IID and a constant variance.

Time Series

Simplified spatial data

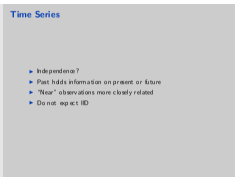
- Both correlated data
- One dimension: time
- Usually sampled regularly: daily, monthly, annually
- Order to observations: past, present, future

14

Time Series

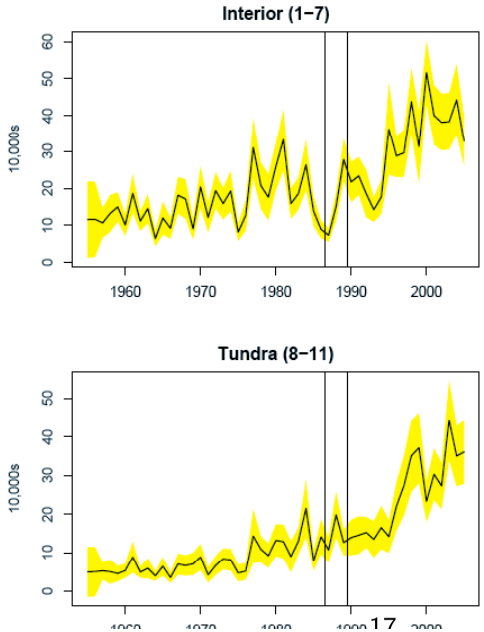
- Independence?
- Past holds information on present or future
- "Near" observations more closely related
- Do not expect IID

15

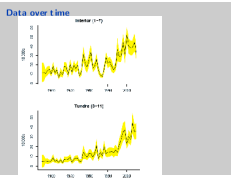


Very reason take measurements periodically is to look for the patterns over time. Look for trends and periodicities. If we did not would not plot the data, just take the information react to it and throw it away.

Data over time



2006-03-07
Why Include Spatial Dependencies?
Statistical Models
Data over time

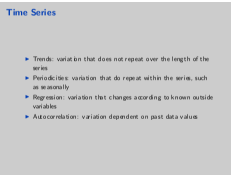


Here we have mallard counts in Alaska. The observations are not independent. Overall the series rises over time. On small scales the deviations are similar. We no longer have IID and do not expect it.

Time Series

- Trends: variation that does not repeat over the length of the series
- Periodicities: variation that do repeat within the series, such as seasonally
- Regression: variation that changes according to known outside variables
- Autocorrelation: variation dependent on past data values

2006-03-07
Why Include Spatial Dependencies?
Statistical Models
Time Series



Time series analysis is the characterization of the correlations as a function of time lag

Where is the Repetition?

- Each observation depends on the observations that came before
- Observation at each time point could be a regression on its past
- Repetitions are of the variation between observations a given number of time lags apart

21

Stationarity

Everything is relative.

- Strong: distribution is the same regardless of where
- Weak: Mean and covariance are the same regardless of position
- Dependence of the data on its past does not change with the mean or with time

22

Stationarity

Everything is relative.
- Strong: distribution is the same regardless of where
- Weak: Mean and covariance are the same regardless of position
- Dependence of the data on its past does not change with the mean or with time

This is a similar situation to the variance being dependent on the mean. Here the autocovariances do not depend on the mean or time. Whether you care about the temporal or spatial effects or not, when working with correlated data, you need to pay attention to the assumptions of the model. The consequences are that your analysis may not make sense (or appear reasonable but be wrong).

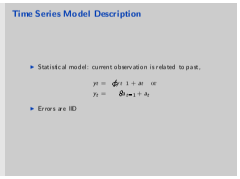
Time Series Model Description

- Statistical model: current observation is related to past,

$$y_t = \phi y_{t-1} + a_t \quad \text{or}$$
$$y_t = \theta a_{t-1} + a_t$$

- Errors are IID

24



Looks like regression but on the series itself, (independent variables are fixed and without error)??? Data are multivariate normal, with a variance as a function of the time series parameters, say ϕ and or θ

$$\mathbf{y} \sim \mathrm{N}(\text{fixed effects} + \text{trend} + \text{periodicities}, \mathbf{V}(\phi, \theta))$$

If we knew the time series parameters this would be generalized least squares (GLM, not GLIM)

Startin' Up

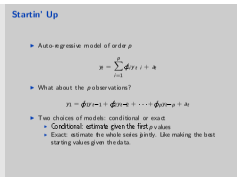
- Auto-regressive model of order p

$$y_t = \sum_{i=1}^{p} \phi_i y_{t-i} + a_t$$

- What about the p observations?

$$y_1 = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + a_t$$

- Two choices of models: conditional or exact
  - Conditional: estimate given the first p values
  - Exact: estimate the whole series jointly. Like making the best starting values given the data.



There is no data for $y_{t-p}, \ldots, y_{t-1}$. Before PCs these were "complex" computations, this was an issue. In fact the first solution was to run the series backwards, forecast the first p values then use them in the original series. The transformation in the next slide will show the exact approach.

Transform Back to IID

Regression residuals, $\mathbf{y} - \mathbf{X}\beta$ have mean $\mathbf{y}$ and general variance $\mathbf{V} = [v_{ij}]$ We can split the variance matrix into two parts like we take the square root

$$\begin{array}{ccccc} \sigma^2 & & = & & \sqrt{(\sigma^2)} & & \sigma \\ \mathbf{V} & & = & & \mathbf{L} & & \mathbf{L}' \\ \begin{bmatrix} v_{11} & \cdots & v_{1n} \\ \vdots & \ddots & \vdots \\ v_{1n} & \cdots & v_{nn} \end{bmatrix} & & = & & \begin{bmatrix} l_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ l_{1n} & \cdots & l_{nn} \end{bmatrix} & & \begin{bmatrix} l_{11} & \cdots & l_{1n} \\ \vdots & \ddots & \vdots \\ 0 & \cdots & l_{nn} \end{bmatrix} \end{array}$$

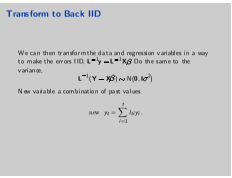
Transform to Back IID

We can then transform the data and regression variables in a way to make the errors IID, $\mathbf{L}^{-1}\mathbf{y} - \mathbf{L}^{-1}\mathbf{X}\beta$ Do the same to the variance,

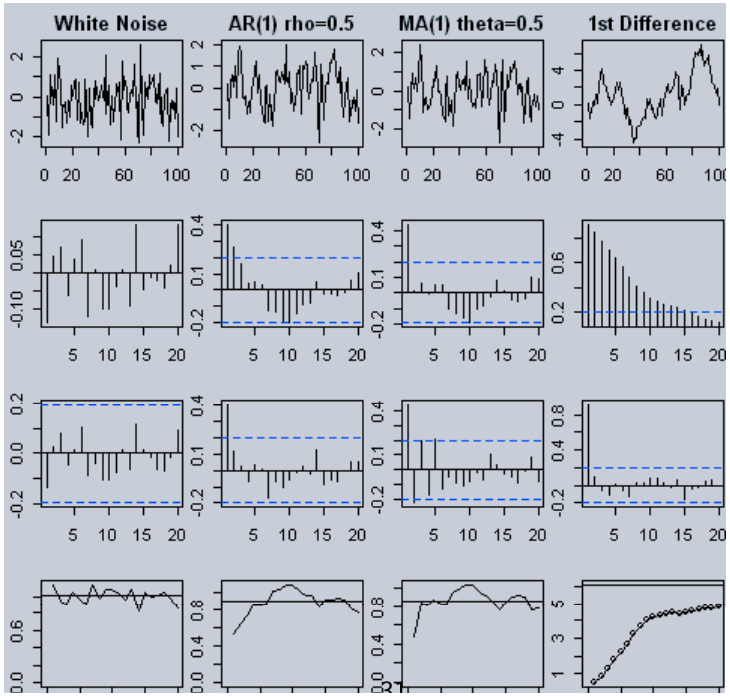
$$\mathbf{L}^{-1}(\mathbf{Y} - \mathbf{X}\beta) \sim \mathrm{N}(\mathbf{0}, \mathbf{I}\sigma^2)$$

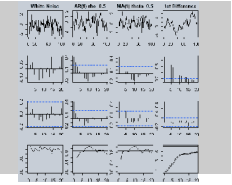
New variable a combination of past values

$$\text{new } y_t = \sum_{i=1}^{t} l_{ti} y_i.$$

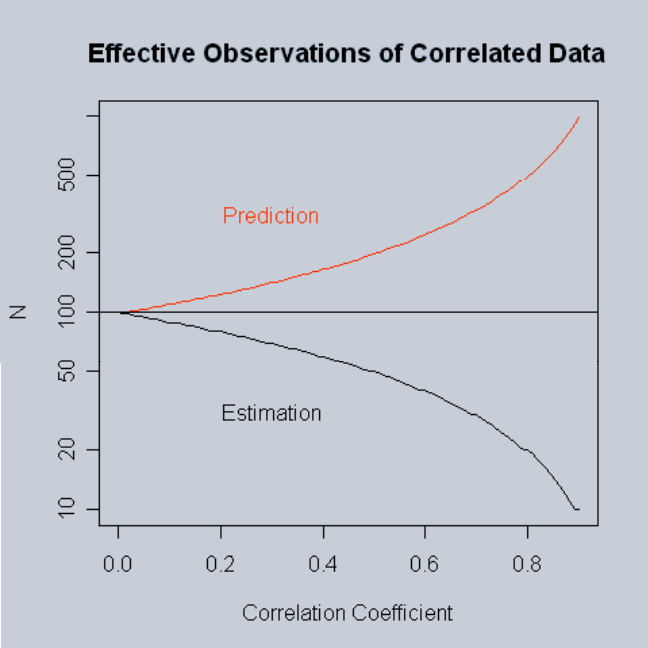


We transformed data and regression back to IID Just make new variables that are linear combinations of data and regression variables, new $y_t = \sum_{i=1}^{t} l_{ti} y_i$. They are combinations of past data. The new variables are approximately transformed back to IID The identification, estimation, and diagnostics are more complicated, but the concepts are the same.

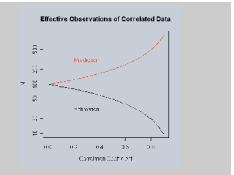




This shows four time series: white noise, autoregressive order 1 where adjacent time points data are correlated and moving average 1 where the adjacent errors are correlated and a first differenced series where the changes between one time point and the last are random.
Rows of plots are: (1)the series; (2) the autocorrelations (ACF), which are correlations between observations a given number of time lags apart; (3) the partial autocorrelations (PACF), like partial correlation coefficients. Auto-regressive-moving-average (ARIMA) models are identified using these functions. AR models by the significant ACFs, MA by the number of PACFs. (4) the variogram is how spatial correlations are identified because they give unbiased estimates and are valid for a larger range of models. The repetitions are here, the relations between observations at different time lags. For the last series, the variogram just increases. It is not stationary. There is no mean. You cannot fit models that assume stationarity to this data.



33



What are the consequences of working with correlated data. This shows the number of IID observations it would take to obtain the same standard error of the mean given the data are all correlated by the same value, ρ. This is the extreme case. Real analyzes will fall within these bounds. Note that as the correlation increases the number of effective observations drops, i.e., each observation provides less information. With prediction the situation is different; the correlation provides information to unrealized observations.

Biased Estimates of Standard Errors

OLS Regression			AR(1) phi=0.36	
Variable	Est	SE	Est	SE
Constant	97.6	12.3	80.8	15.2
72-97	-1.23	0.4	-0.6	0.6
Variance	0.036		0.032	
AIC	1159		1153	

Here is an example of a change in the Scoter breeding population due to a change in harvest regulations. In general the estimates are unbiased but the standard errors are. For positive correlations the standard error are underestimated and the reverse for negative correlations. Even though the correlations are a nuisance they need to be addressed

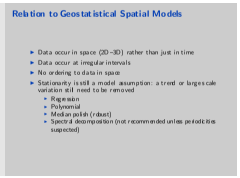
[Plot of Scoter Populations with change in Harvest Regulations]

35

Relation to Geostatistical Spatial Models

- Data occur in space (2D–3D) rather than just in time
- Data occur at irregular intervals
- No ordering to data in space
- Stationarity is still a model assumption: a trend or large scale variation still need to be removed
  - Regression
  - Polynomial
  - Median polish (robust)
  - Spectral decomposition (not recommended unless periodicities suspected)

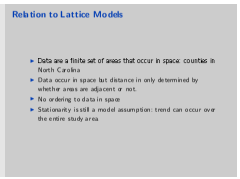
36



Since the data occur more than one dimension, the correlation structure may be different in different directions (anisotropy). Duck populations may be more similar latitudinally then longitudinally.With ARIMA models, we estimate best correlation at given lags. With spatial data, we need to model the correlation as a continuous function of distance between observations. Because there is no ordering of data in space and we are estimating a continuous correlation function, the models are jointly estimated. The correlation function only depends on distance and maybe direction. It does not depend on position in space.

Relation to Lattice Models

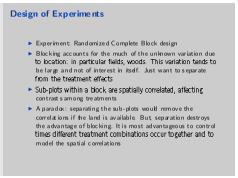
- Data are a finite set of areas that occur in space: counties in North Carolina
- Data occur in space but distance in only determined by whether areas are adjacent or not.
- No ordering to data in space
- Stationarity is still a model assumption: trend can occur over the entire study area.



Using adjacency instead of distance Makes models related to the ARIMA models. The observations or errors are correlated if they are adjacent.Because there is no ordering of data in space, it is possible to have a conditional model where each data point in conditional to all those it is adjacent to.

Design of Experiments

- Experiment: Randomized Complete Block design
- Blocking accounts for the much of the unknown variation due to location: in particular fields, woods. This variation tends to be large and not of interest in itself. Just want to separate from the treatment effects
- Sub-plots within a block are spatially correlated, affecting contrasts among treatments
- A paradox: separating the sub-plots would remove the correlations if the land is available. But, separation destroys the advantage of blocking. It is most advantageous to control times different treatment combinations occur together and to model the spatial correlations



We still need some spatial analysis to know the *range* of the correlations.

Design of Experiments

- Arrangement of treatments controlling the distance between different treatment combinations. Jun Zhu will talk about improving the efficiencies to estimate the correlations due to distance.
- Modeling the spatial correlations in the experiment
- Without modeling the spatial correlations the coverage of the tests (e.g., above time series regression example.)

42

Conclusions

- Repetition is matching pairs of data a given distance apart
- Correlated data can be tranformed back to IID
- Data must be stationary: large scale variation removed
- Correlation affects the amount of information in each observation
- Correlation affects the estimates of standard errors
- Correlation and stationarity are similar in ARIMA and geostatistical models
- Conditional and joint models are similar between ARIMA and lattice models
- Spatial correlations affect arrangement of treatments, and test coverages

43