Lattice Models with Spatial Dependencies - An Introduction

Mary C. Christman
Univ. of Florida
Department of Statistics - IFAS

Lattice Models

- Area of interest is subdivided into mutually exclusive and exhaustive plots, strata, or subareas
- Data are aggregated or summary values for each subarea

EXAMPLE: Sudden Infant Death Syndrome statistics for counties in North Carolina in the 1970s

Additional Comments

- Note that unlike geostatistical modeling, in lattice models there is no concept of interpolating between plots or subareas.
- As a result, we are less interested in mapping and more interested is modeling such as regression with correlated data or mixed models with covariance matrices that are not diagonal

Questions

- Classic models are used to test hypotheses about explanatory variables (factors, covariates, etc)
  - Q: Should we worry about spatial autocorrelation?
  - If so, how should the spatially-explicit aspect be incorporated into our modeling effort?
- When planning a study, need to address:
  - Spatial arrangement of treatments if planned experiment
  - Spatial arrangement of plots when observational study

Additional Comments

- Traditionally, the spatial autocorrelation that was presumed to be a potential problem was handled in experimental designs using such techniques as blocking
  - E.g. the Average Distance Balanced Design in which treatments are arranged spatially so that the average distnace between plots of different treatments is approximately constant over all treatments

Classic Model Assumptions

- For General Linear Models
  - Error terms are Normally distributed with constant mean ($\mu = 0$) and variance (σ^2) and
  - Error terms (and hence the responses) are independent
- For Generalized Linear Models
  - Response Variable is distributed appropriately (usually Binomial, Poisson or similar) with a mean that is a function of covariates ($\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$) and variance that depends on the mean.
  - The responses are independent

Additional Comments

- Even with restricted randomization methods to account for spatial arrangement of locations,
  - there may still be spatial autocorrelation and hence the error terms/response variables are not independent
  - and so classical assumptions fail.

Failure of the Independence Assumption

- Due to non-spatial issues such as sampling design
  - E.g. blocking, clustering or temporal effects
- Due to Spatial autocorrelation
  - Correlation between 2 values of the response variable, $Y(s_i)$ and $Y(s_j)$ at locations s_i and s_j, is non-zero and a function of distance
  - How does it arise?

Additional Comments

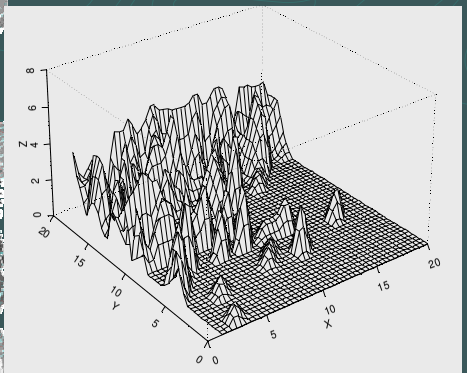
- Non-spatial lack of independence is handled as usual, e.g. random blocks or time series.
- Spatial lack of independence is handled using autocorrelation covariance matrices that require additional information
  - form of the non-independence (as a function of distance),
  - neighborhood structures, etc.
- Note: I assume that distance is Euclidean unless otherwise specified

Sources of Spatial Autocorrelation in Y

- Induced
  - Values close in space could be similar due to an important explanatory variable that varies smoothly in space
    1. E.g. The spatial distribution of bell pepper fungus in a field
       - could be due to spatial distribution of soil moisture
       - could be due to geography (e.g. elevation changes)

Example - Bell Pepper fungus

Leaf Disk Assay



See Larkin et al. 1995. Phytopath 85:191-203

Field plot was subdivided into 400 1x1 m subplots.

In each subplot, 5 leaf disks were assayed for presence of fungus.

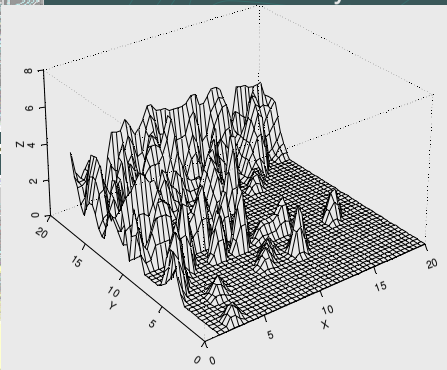
Recorded number that tested positive.

Additional Comments

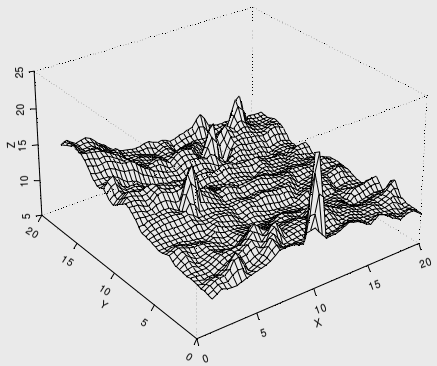
- Graph shows the number of leaf disks assays that tested positive for fungus (out of 5) for each 1x1 m plot within the study area.
- Note the trend (low in SE corner, high in NW corner) as well as grouping of similar values spatially.
- The next slide shows that the pattern may be related to soil moisture, i.e. the spatial patterns show similarity. Is it possible that moisture is a partial predictor for fungus presence?

Example - Bell Pepper fungus

Leaf Disk Assay



Soil Moisture



Sources of Spatial Autocorrelation in Y

- True
  - Intrinsic, underlying covariance that is a function of distance
    - E.g. for the spatial distribution of soil moisture, it could be due to soil characteristics that allow water movement into and through adjacent plots
  - Causal interaction among nearby locations
    - E.g. The spatial distribution of leaf fungus could be due to dispersal mechanism
      - Leaves touching vs. air dispersal

Sources of Spatial Autocorrelation in Y

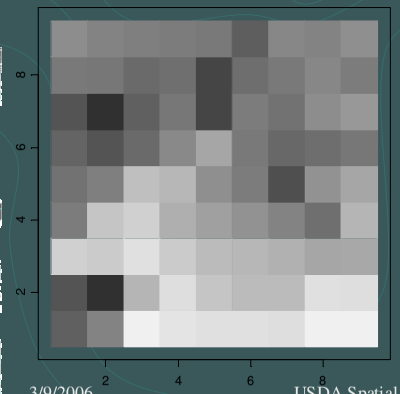
- Spurious
  - Values close in space could be similar due to chance
    - E.g. due to the spatial arrangement of the sampling locations
    - E.g. due to smoothing of the data during preliminary data management
    - E.g. due to the scale at which the data have been aggregated

Additional Comments

- Spurious autocorrelation is unlikely for the bell pepper fungus dataset since plots are small and there is no data manipulation prior to analysis.
- Spurious autocorrelation is the hardest to capture and identify.
  - An example would be in precision agriculture due to the slight delay in recording soil attributes. The recording device often has a delay of 3-4 seconds but the location is recoded not where the data were collected but where the recorder reports the value.
  - See this sometimes in satellite images as well due to interpolation for pixel data

Example

Reflectance Values From An Areal Survey of Pollution Levels Due To Pumping Of Waste Material Into The English Channel



Darker areas represent more polluted spots. This location is closest to the source of the pollution.

Values in any one grid cell are averages over the cell and, due to location error, possibly include values in neighboring cells as well.

Autoregressive Lattice Models

$$Y(s_i) = \mu(s_i) + \sum_{s_j \in N_i} \varpi_{ij}[Y(s_j) - \mu(s_j)] + \varepsilon(s_i)$$

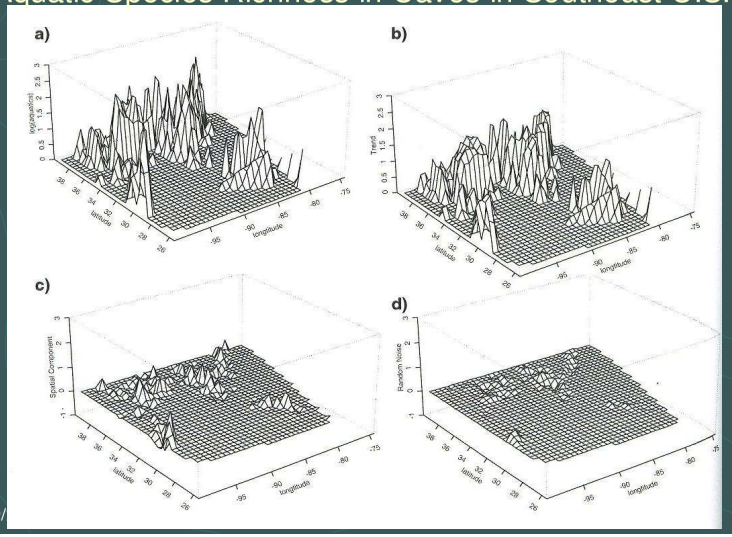
- $Y(s_i)$ is the response variable at location s_i
- $\mu(s_i)$ is the large-scale trend or mean for location s_i
  - may depend on explanatory variables or treatments
- $\sum \varpi_{ij}[Y(s_j) - \mu(s_j)]$ small-scale variation at location s_i
  - Depends on the values in the neighborhood N_i and weights ϖ_{ij}
- $\varepsilon(s_i)$ is the error term, conditionally independent with zero mean and constant variance

Additional Comments

- The large-scale mean is usually dependent on explanatory variables such as covariates or treatment levels or even location (such as a trend surface that is a polynomial in space.
- The small scale variation can be used to calculate the conditional mean, that is the predicted value at a location using the covariates at a location and the values of observations around that location. The conditional mean is the sum of several parts: 1) the mean of the individual subplot, $\mu(s_i)$; 2) the weighted average of the error terms for all of the neighboring subplots.

Example of the Decomposition

Aquatic Species Richness in Caves in Southeast U.S.



Additional Comments

- These perspective plots show the decomposition of species richness values in counties throughout the southeast US.
- $Y(s_i)$ are shown in (a), a plot of the observed values of log (aquatic species richness) in counties in the southeast US.
- $\hat{\mu}(s_i)$ are shown in (b) a plot of the estimated county means of log (aquatic species richness) predicted by the explanatory variable, X=number of caves found in the county.
- $\sum \varpi_{ij}[Y(s_j) - \hat{\mu}(s_j)]$ are shown in (c), a plot of the estimated small-scale variation in each county based on observations of log (aquatic species richness) in contiguous counties. The weights were $w_{ij} = 1$ if counties i and j were contiguous and $w_{ij} = 0$ if they were not.
- $\hat{\varepsilon}(s_i)$ are shown in (d), a plot of the unexplained or residual noise.
- Note that the values in (b), (c) and (d) add up to the observed values shown in (a).

Large-scale Variation $\mu(s_i)$

- Could be a function of factors being manipulated in a planned experiment
  - E.g. a split-plot design with a whole plot factor of crop rotation schedule and a subplot factor of nitrogen source
- Could be explanatory variables being observed
  - E.g. soil moisture in the bell pepper fungus study
  - E.g. the number of caves in a county to predict the species richness of aquatic subterranean animals

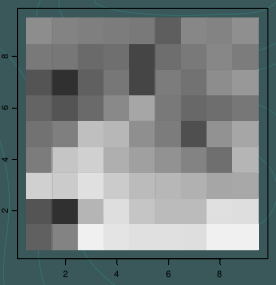
Small scale Variation

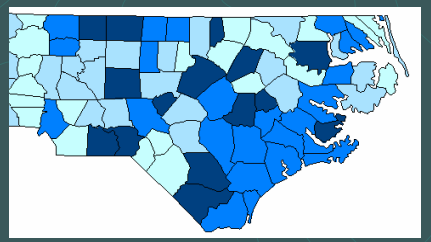
$$\sum_{s_j \in N_i} \varpi_{ij}[Y(s_j) - \mu(s_j)]$$

- Two parts
  - Neighborhood structure N_i
  - Weighting scheme ϖ_{ij}

Constructing Neighborhoods

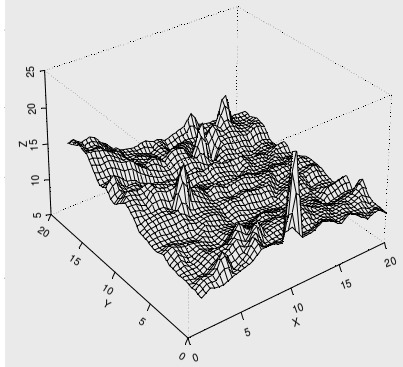
- Depends on whether layout is regular or irregular

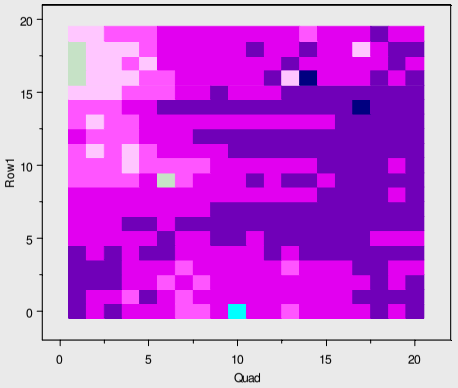




- Every cell (plot, county) must have a defined neighborhood

Example – Bell Pepper Fungus



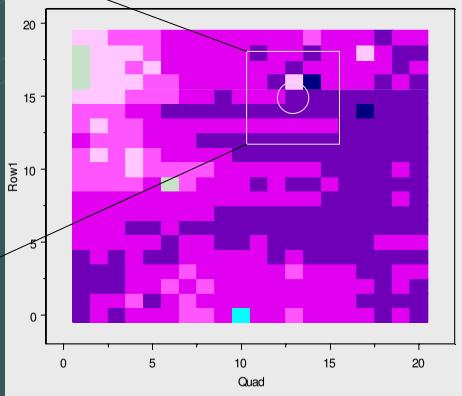


Additional Comments

- The bell pepper fungus data was collected on a regular grid layout with 20 rows (“row1”) and 20 columns (“quad”)–
  - data for each of the 400 cells in the field plot.
  - For example, while soil moisture may in fact vary over a 1x1 m square plot, only a single number is reported for each 1x1 m plot and so represents the value for that plot.
- Showing two graphics here
  - the left one is a perspective plot which shows the variation in soil moisture values
  - The right one shows the same information as color gradations for each cell for which we have data

Examples: Neighborhoods for Square Lattices

10.8	10.9	9.7	10.2	12.0
11.5	9.8	15.0	6.1	10.3
9.7	9.0	9.2	10.4	10.7
8.6	8.8	8.6	8.4	9.5
10.2	11.0	10.3	10.1	10.2



First-order NB

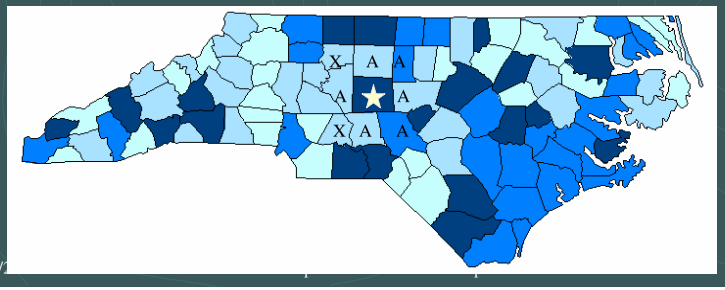
+ Second-order NB

Additional Comments

- These neighborhoods are two of many possible examples – one can further change them or even use different setups.
  - For example, in the case of the water moisture, one might expect that autocorrelation would be higher in the within row direction rather than across rows. This could be due to watering the field by flooding of the pathways between rows or of the beds are raised. In that case, the neighborhood might only be the plots adjacent and within the same row, say the N-S plots only.

Examples: Neighborhoods for Non-Square Lattices

- N_i = *{cells labeled A}* is a neighborhood whose boundaries touch the boundary of the i^{th} cell
- N_i = *{cells labeled A or X}* is a neighborhood whose centroids are within a specified distance from the centroid of the i^{th} cell



Weighting Scheme ϖ_{ij}

- The larger the weight the more that neighboring plot contributes
- Common approaches
  - As a function of Euclidean distance
  - As a function of contiguity
  - Directional weighting (certain directions contribute more than others)
  - As a function of the length of the common boundary
  - Weighting to correct for heterogeneity of variance

Additional Comments

- Weighting can involve some combination of these approaches and is clearly integrally related to the definition of the neighborhood.
- Weights are usually standardized so that they sum to a constant, e.g. $\sum_{j \in N_i} w_{ij} = \eta$
- Negative weights (which imply a negative correlation) are usually avoided but there are times when they are appropriate.

Weighting Scheme ϖ_{ij}

- Crucial to identify appropriate weighting method
- Should have some idea of
  - The range of likely autocorrelation
  - How fast autocorrelation decays as distance increases
  - The direction of likely autocorrelation
    1. The directionality is influenced by both the choice of neighborhood as well as differential weighting by direction.

Weighting Scheme ϖ_{ij}

- Methods for exploring likely form of autocorrelation:
  - Calculate some common autocorrelation statistics such as Moran's I or Geary's C
    - Validity depends on the neighborhood and weighting scheme
    - Try different neighborhoods and weights
  - Do variography using the centroids or nodes of a lattice as the point locations

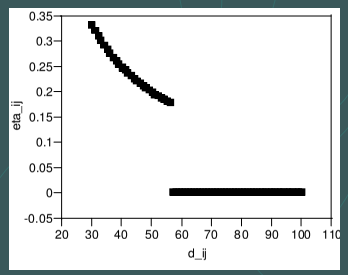
Simple Weighting

- Bell Pepper Fungus
  - Let N_i be the 1x1 m plots having boundaries with the i^{th} plot (first-order NB)
  - Define the weights to be $\varpi_{ij} = \eta \text{ if } j^{th} \text{ plot in } N_i$
  - These weights imply
    - no directionality
    - each neighboring plot is equally autocorrelated with the i^{th} plot
    - The autocorrelation is the same regardless of the location of the i^{th} plot

More Complex Weighting

- Aquatic Cave Species in SE US
  - Defined the neighborhood to be counties with county seats within 56 km of the i^{th} county
  - Uses Euclidean distance to weight closer counties higher than farther counties

$$\eta_{ij} = \begin{cases} 0, & \text{if } d_{ij} > 56\,\text{km} \\ \dfrac{\rho\left\{\dfrac{1}{d_{ij}}\right\}}{\max\limits_{i,j}\left\{\dfrac{1}{d_{ij}} : j \in N_i\right\}}, & \text{otherwise} \end{cases}$$



Additional Comments

- The numerator is a constant times the inverse of the distance between the 2 locations (inverse so that closer neighbors weight higher than further neighbors).
- The denominator is a scaling or standardizing function so that ρ is the correlation between the ith county and its nearest neighbor.
- This approach is a type of "row standardization" and constrains the constant ρ to be less than 1.

Modeling

- So,
  - having identified the explanatory variables for the large-scale variation (trend),
  - the neighborhood structure and weighting scheme for the small-scale variation, and
  - checked for homogeneity of variance,
- the next step is
  - to do the actual model fitting to obtain estimates of the model parameters, means (and SEMs) and, if desired, predictions (and MSPE).

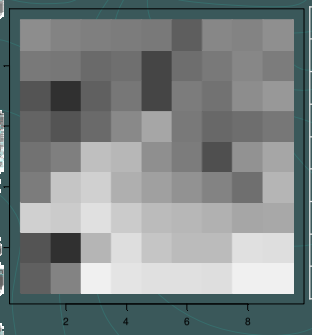
Modeling Approaches

- Two approaches
  - Simultaneous Autoregressive Models (SAR models)
  - Conditional Autoregressive Models (CAR models)
- The difference is in the variance-covariance matrix for the $\{Y(s_1), \ldots, Y(s_n)\}$
- Both can be fitted but fitting the SAR model leads to residuals that are correlated with the neighboring *Y*-values
  - CAR model does not have this problem and is generally preferred

Additional Comments

- Every SAR model can be described in terms of a CAR model but CAR models are not always easily or naturally described as SAR models. This is based on the choices of neighborhoods and variance structure and weights.

Simple Example – Reflectance Values for Pollution in the English Channel



Data Values

32	35	36	37	38	47	34	35	31
38	39	43	41	55	42	38	34	37
50	62	46	39	55	37	40	32	28
45	50	43	33	24	38	44	42	39
40	36	16	18	31	37	52	30	24
37	14	10	21	26	30	35	41	19
10	12	5	12	17	18	20	24	23
50	62	19	6	14	17	17	5	6
46	35	0	4	5	5	6	0	0

from Haining (1990)

Additional Comments

- Like the bell pepper fungus, this dataset is on a regular grid. So, the spatial coordinate system is taken to be
  - the row ID, "r", and
  - the column ID, "c".

Conditional Autoregressive Model

$$Y(s_i) = \mu(s_i) + \sum_{s_j \in N_i} \varpi_{ij}[Y(s_j) - \mu(s_j)] + \varepsilon(s_i)$$

- The error terms are conditionally independent and Normally distributed with mean 0 and constant variance σ^2
- The conditional mean of $Y(s_i)$ is

$$\mu(s_i) + \sum \varpi_{ij}[Y(s_j) - \mu(s_j)]$$

and the unconditional mean is

$$\mu(s_i)$$

Additional Comments

- The conditional mean is the predicted value for an individual observation. The unconditional (marginal) mean is the mean of the trend part only.
- The conditional mean is estimated using the BLUP and the unconditional mean by the BLUE (LSmeans).

Conditional Autoregressive Model

$$Y(s_i) = \mu(s_i) + \sum_{s_j \in N_i} \varpi_{ij}[Y(s_j) - \mu(s_j)] + \varepsilon(s_i)$$

- The conditional variance of $\{Y(s_i) : i = 1,\ldots,n\}$ is

$$\sigma^2 \mathbf{I}$$

and the unconditional variance is

$$(\mathbf{I} - \mathbf{W})^{-1}\sigma^2 \mathbf{I}$$

where $\mathbf{W} = \{w_{ij}\}$ is the matrix version of the weights for the neighborhood

Large-Scale Trend $\mu(s_i)$

$$Y(r,c) = \beta_0 + \beta_1 r + \beta_2 c + \beta_{12} rc + \beta_{11} r^2 + \beta_{22} c^2 + \ldots + \varepsilon(r,c)$$

- Haining (1990) started by ignoring the spatial autocorrelation and fit linear regression models using polynomials in (r, c) where r is the row ID, c is the column ID
- He determined that the linear model had the best fit

$$Y(r,c) = \beta_0 + \beta_1 r + \beta_2 c + \varepsilon(r,c)$$

Small-Scale Variation

- We should now test for autocorrelation in the data, so we'll use the residuals from the large-scale trend fit
- Calculate Moran's I for different neighborhood structures (see next slide) using weight = 1 if grid cell was in the neighborhood and 0 otherwise. From this we can tell
  - If there is autocorrelation
  - Which neighborhood is best (among those reviewed of course)

Small-Scale Variation

Neighborhood	Moran's I	SE	Normal Statistic	Normal p-value	Permutation p-value
Row	0.3215	0.1164	2.869	0.004	0.001
Column	0.5434	0.1164	4.775	0+	0+
Diagonal	0.2043	0.0862	2.514	0.012	0.007
First-order	0.4324	0.0814	5.468	0+	0+
Second-order	0.3251	0.0577	5.848	0+	0+

The highest Moran's I value occurs for the column neighborhood and the second highest for the first-order neighborhood.

Additional Comments

- Under the null hypothesis of no autocorrelation, the expected value of Moran's I is E{I} = -1/(n-1). The stronger the correlation, the closer I is to 1.
- Two approaches for testing autocorrelation using Moran's I are:
- 1) approximate normality holds assuming the number of cells is sufficiently large (also depends on the extent and manner in which the cells are connected by the weights). The usual rule of thumb is at least 20 locations.
- 2) permutation or randomization test in which the Z data are randomly permuted (assigned to different locations) repeatedly and the observed results compared against the expected results.

Fit the CAR model

Model was fit with a linear trend and with weights $w_{ij}=\rho$ if plot j was in the neighborhood and = 0 otherwise.

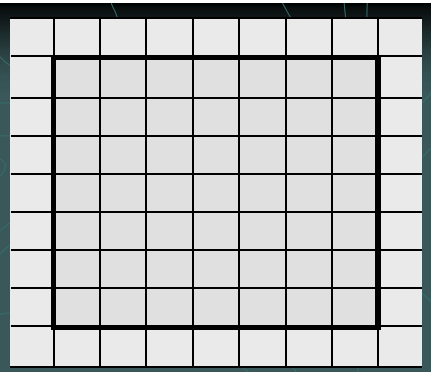
Model	β_0 Estimate	β_1 Estimate	β_2 Estimate	ρ Estimate	Root MSE	Log Likel.
1	50.966**	-2.733**	-2.131*	0.256**	9.02	-362
2	53.755**	-2.966**	-1.757**	0.442**	8.85	-364
1a	60.289**	-3.467**	-2.050*	0.251**	10.00	-211

Models: (1) first-order neighborhood with 9x9 area
(2) column neighborhood with 9x9 area
(1a) first-order neighborhood with 8x8 interior area

Additional Comments

- There is very little difference in the models with the two different neighborhood structures, so for parsimony choose the model using the column neighborhood
- Note that the estimated spatial weight is 0.256 for the first-order neighborhood and 0.44 for the column neighborhood. The difference in values has more to do with the number of neighbors in the neighborhood than with any estimate of autocorrelation.
- The final model is the result of adjusting for boundary effects (next).

Boundary Effects



- The neighborhoods of the cells on the edges are halved
  - Standard errors of predictions at the edges very high
  - Introduces possible estimation bias
- One way to avoid is to analyze only that part of the study region completely within the entire region

Additional Comments

- Choose the subregion within the study area so that every cell in the subregion has a complete neighborhood that can be used in the modeling
- In the bell pepper example, that would be a subregion 19x19 (rather than 20x20) that would be modeled (Y-values on the left side of the model). The remaining cells would appear only on the right side of the model in the small-scale variation.

Fit the CAR model

Model was fit with a linear trend and with weights $w_{ij}=\rho$ if plot j was in the neighborhood and = 0 otherwise.

Model	β_0 Estimate	β_1 Estimate	β_2 Estimate	ρ Estimate	Root MSE	Log Likel.
1	50.966**	-2.733**	-2.131*	0.256**	9.02	-362
2	53.755**	-2.966**	-1.757**	0.442**	8.85	-364
1a	60.289**	-3.467**	-2.050*	0.251**	10.00	-211

Models: (1) first-order neighborhood with 9x9 area
(2) column neighborhood with 9x9 area
(1a) first-order neighborhood with 8x8 interior area

Additional Comments

- Note the difference between model 1 and model 1a in the estimates of the model coefficients. Due to
  - smaller number of observations (64 vs. 81)
  - Better estimation of the spatial autocorrelation since every observations has a full neighborhood

Summary and Conclusions

- When data are collected in aggregate for non-overlapping subregions of the study area and
- The spatial arrangement is such that there are effects due to space (or to spatial covariates that were not measured)
- Then consider models that incorporate an effect due to spatial correlation

Advantages

- Accounts for some additional sources of variation
- Increases understanding of the process of interest
- Overall lattice models are excellent approaches for incorporating spatial correlation and for providing improved predictions

Caveats When Fitting Lattice Models

- If covariates are available that explain the seeming spatial correlation, then these are more appropriately used
- Choice of neighborhood and weighting scheme are critical to good model fitting
- Sample sizes could be too small to adequately estimate the spatial correlation
- Modeling might require a lot of exploratory analyses. Note that this means that the conclusions are only tentative and should be independently tested with a new experiment.