Global Spatial Autocorrelation

Luc Anselin

THE CENTER FOR SPATIAL DATA SCIENCE
THE UNIVERSITY OF CHICAGO

http://spatial.uchicago.edu

global spatial autocorrelation

Moran scatter plot

correlogram

variogram

variogram models

Global Spatial Autocorrelation

- Global Spatial Autocorrelation Measures

    combination attribute similarity and locational similarity

    one statistic for the whole pattern

    test for clustering not for clusters (locations)

Moran's I

- Moran's I

  the most commonly used of many spatial autocorrelation statistics

- $I = [\ \Sigma_i \Sigma_j w_{ij} z_i.z_j/S_0\ ]/[\Sigma_i z_i^2 / N]$

  with $z_i = y_i - m_x$: deviations from mean

  cross product statistic $(z_i.z_j)$ similar to a correlation coefficient

  value depends on weights (w_{ij})

- Moran's I examined more closely

  scaling factors in numerator and denominator

  in numerator: $S_0 = \Sigma_i \Sigma_j w_{ij}$

    the number of non-zero elements in the weights matrix, or the number of neighbor pairs (x2)

  in denominator: N

    the total number of observations

- Inference

  how to assess whether computed value of Moran's I is significantly different from a value for a spatially random distribution

  compute analytically (assume normal distribution, etc.)

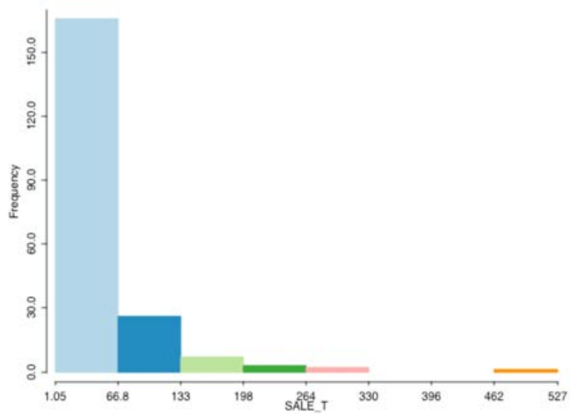
  computationally, compare value to a reference distribution obtained from a series of randomly permuted patterns

- Standardized z-value

    standardize by subtracting mean and dividing by standard deviation, computed from the reference distribution

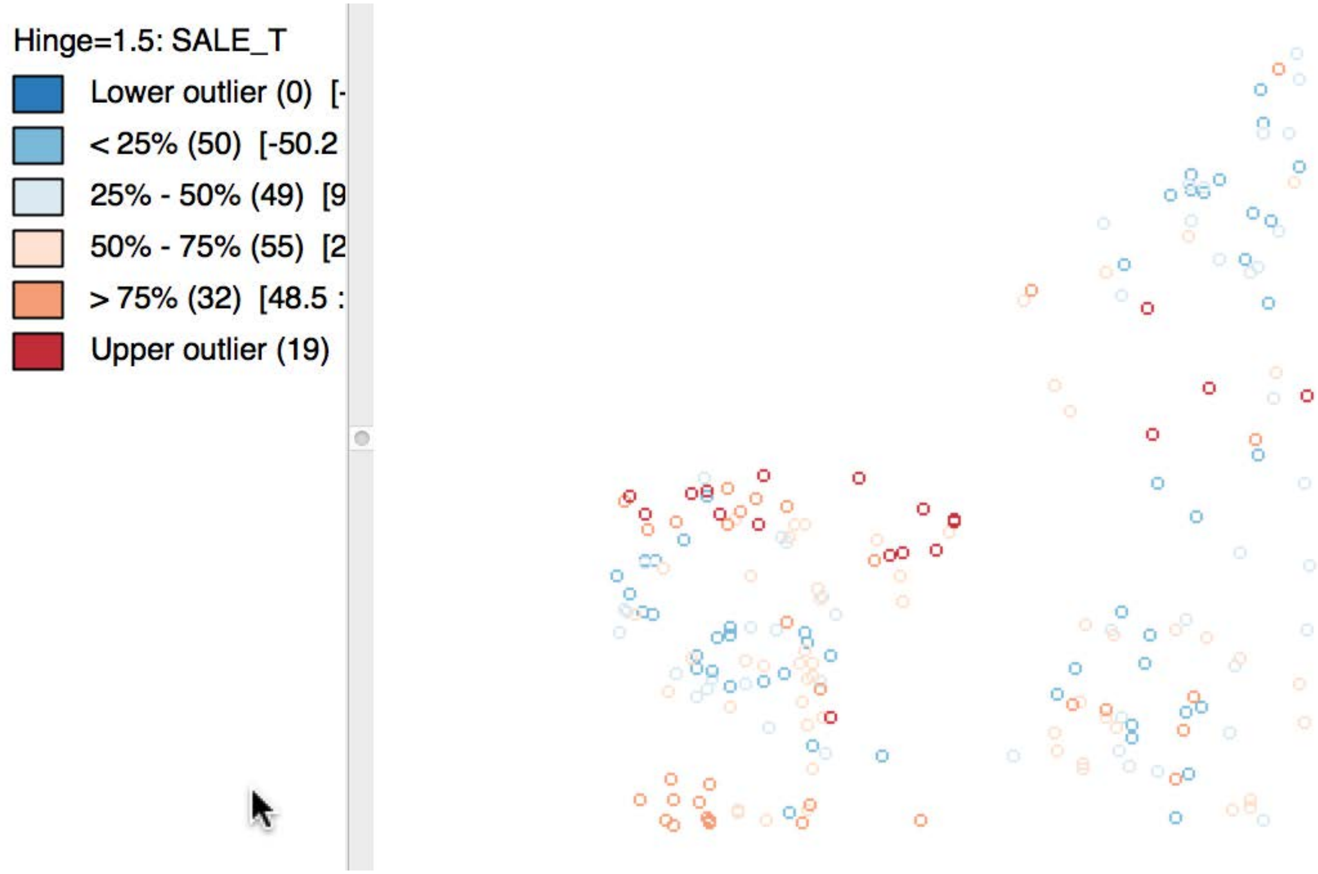
- z = [Observed I - Mean(I)] / Standard Deviation(I)

    z-values are comparable across variables and across spatial weights



from	1.05	66.8	133	198	264	330	396	462
to	66.8	133	198	264	330	396	462	527
#obs	166	26	7	3	2	0	0	1
% of total	81	12.7	3.41	1.46	0.976	0	0	0.488
sd from mean	0	0.411	1.5	2.58	3.67	4.75	5.84	6.92

min: 1.049, max: 527.409, median: 20, mean: 41.8968, s.d.: 60.6543, #obs: 205



Cleveland 2015 q4 house sales prices (in $1,000)

	Normal	Randomization
MI	0.282	0.282
E[MI]	-0.0049	-0.0049
Var[MI]	0.00178	0.00158
z-value	6.81	7.22
p-value	<< 0.000001	<< 0.000001

normal vs randomization inference for Moran's I
queen weights

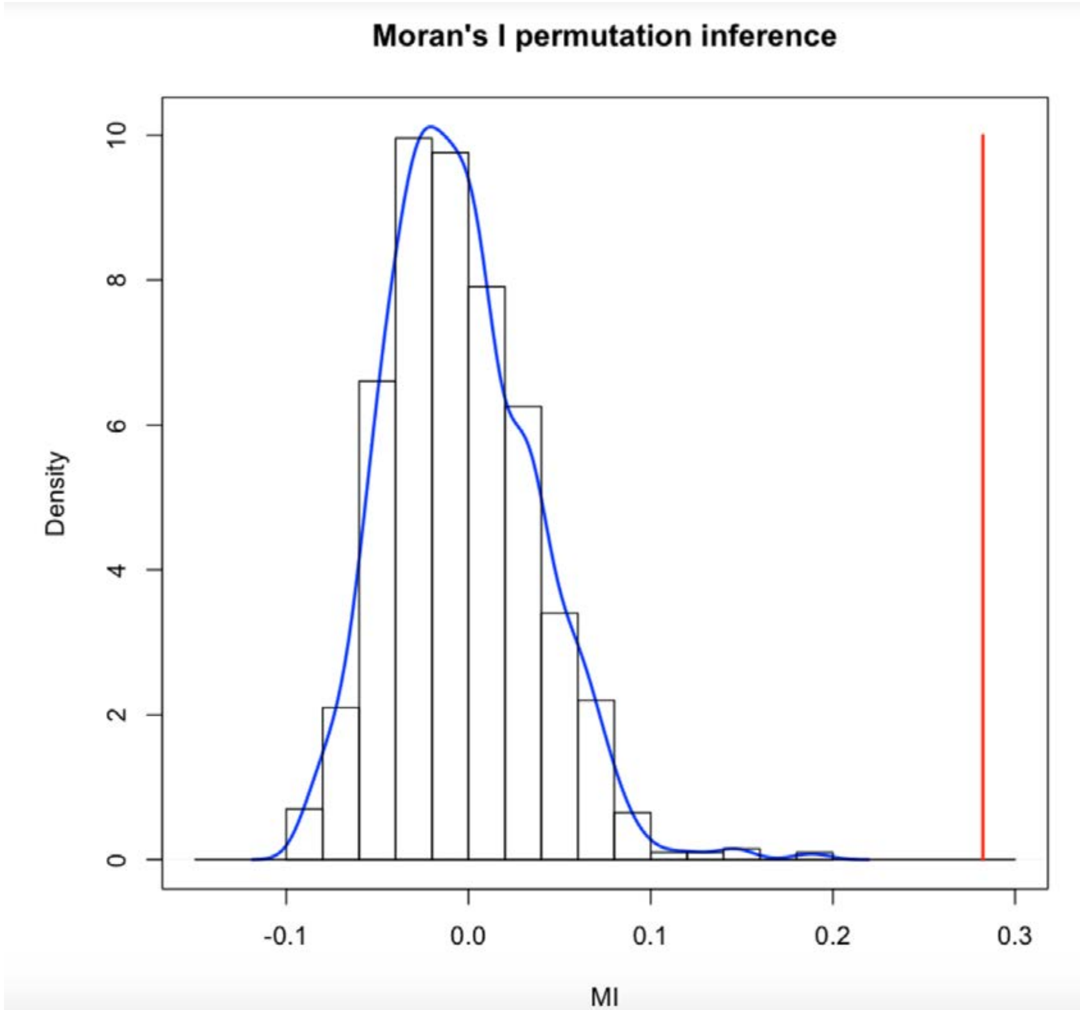
- Permutation Approach

  as such, even a high Moran's I does not indicate significance

  need to construct a reference distribution

  randomly reshuffle observations and recompute Moran's I each time

  compare observed value to reference distribution



999 permutations and reference distribution (queen weights)

MI = 0.282   Mean = -0.0045 s.d. = 0.0401
z-value = 7.15

Interpretation of Moran's I

- Sign of Moran's I
    - theoretical mean is - 1/(N - 1),
        - essentially zero for large N
    - positive and significant = clustering of like value
        - NOT clustering of high or low
        - could be either or a combination
    - negative and significant = alternating values
        - presence of spatial outliers
        - spatial heterogeneity (checkerboard pattern)

- Comparing Moran's I

  Moran's I depends on spatial weights

  relative magnitude for same weights and different variables is meaningful

  but NOT for different spatial weights

  instead, use standardized z-value to compare

	Queen	K6	Distance
MI	0.282	0.343	0.335
E[MI]	-0.0049	-0.0049	-0.0049
Var[MI]	0.00158	0.00124	0.00110
z-value	7.22	9.88	10.3
p-value	<< 0.000001	<< 0.000001	<< 0.000001

Moran's I, different spatial weights
MI is largest for K6, but z is largest for Distance

- Clustering vs Clusters

  Moran's I is a global statistic, i.e., a single value for the whole spatial pattern

  Moran's I does NOT provide the location of clusters

  cluster detection requires a local statistic

- True vs Apparent Contagion

  the indication of clustering does not provide an explanation for why the clustering occurs

  different processes can result in the same pattern

  true contagion: evidence of clustering due to spatial interaction (peer effects, mimicking, etc.)

  apparent contagion: evidence of clustering due to spatial heterogeneity (different spatial structures create local similarity)

Geary's c

- Focus on Dissimilarity

    squared difference as measure of dissimilarity

    similar to notion of variogram (geostatistics)

    values between 0 and 2

- Geary's c Statistic

$$c = (N-1) \Sigma_i \Sigma_j w_{ij} (x_i - x_j)^2 / 2S_0 \Sigma z_i^2$$

alternatively

$$c = [\Sigma_i \Sigma_j w_{ij} (x_i - x_j)^2 / 2S_0] / [\Sigma z_i^2 / (N-1)]$$

with z_i in deviations from the mean

$S_0 = \Sigma_i \Sigma_j w_{ij}$ sum of all the weights

- Interpretation

  positive spatial autocorrelation
  $c < 1$ or $z < 0$

  negative spatial autocorrelation
  $c > 1$ or $z > 0$

  opposite sign of Moran's I

- Inference

  same approach as for Moran's I

    analytical

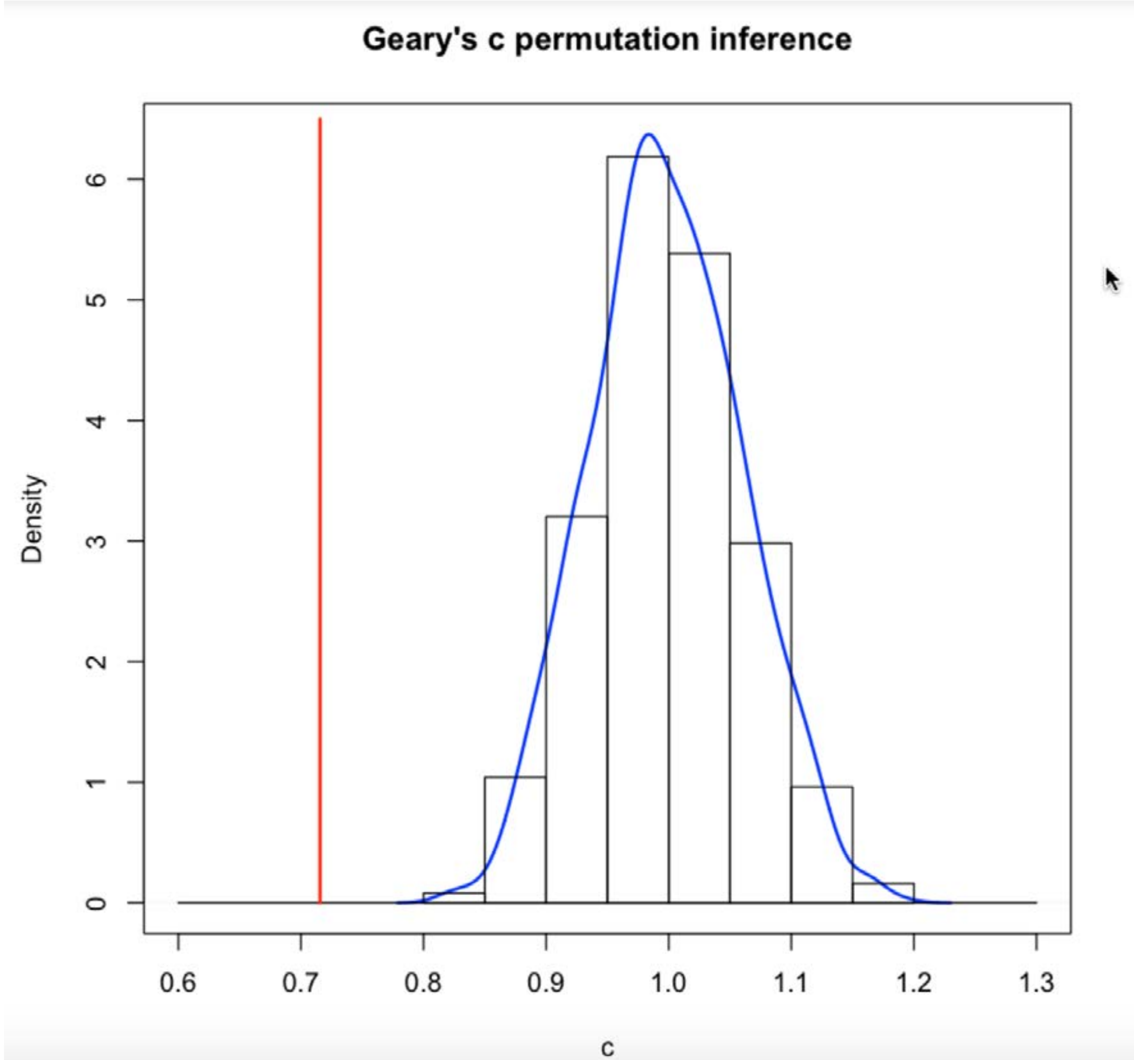
    approximate (normal, randomization)

    permutation

	Queen	K6	Distance
c	0.716	0.496	0.544
E[c]	1.0	1.0	1.0
Var[c]	0.00381	0.00545	0.00285
z-value	-4.61	-6.82	-8.55
p-value	< 0.000004	<< 0.000001	<< 0.000001

geary's c

inference under randomization, different spatial weights



999 permutations and reference distribution (queen weights)

c = 0.716   Mean = 0.998 s.d. = 0.062
z-value = -4.58

Moran Scatter Plot

Recap: Moran's I

- $I = [\, \Sigma_i \Sigma_j w_{ij} z_i.z_j / S_0 \,] / [\Sigma_i z_i^2 / N]$

  with $z_i = y_i - m_x$: deviations from mean

  for row-standardized weights $S_0 = N$

  $I = \Sigma_i \Sigma_j w_{ij} z_i.z_j / \Sigma_i z_i^2 = \Sigma_i z_i \, (\Sigma_j w_{ij}.z_j) / \Sigma_i z_i^2$

  Moran's I is slope in a regression of $\Sigma_j w_{ij}.z_j$ on z_i

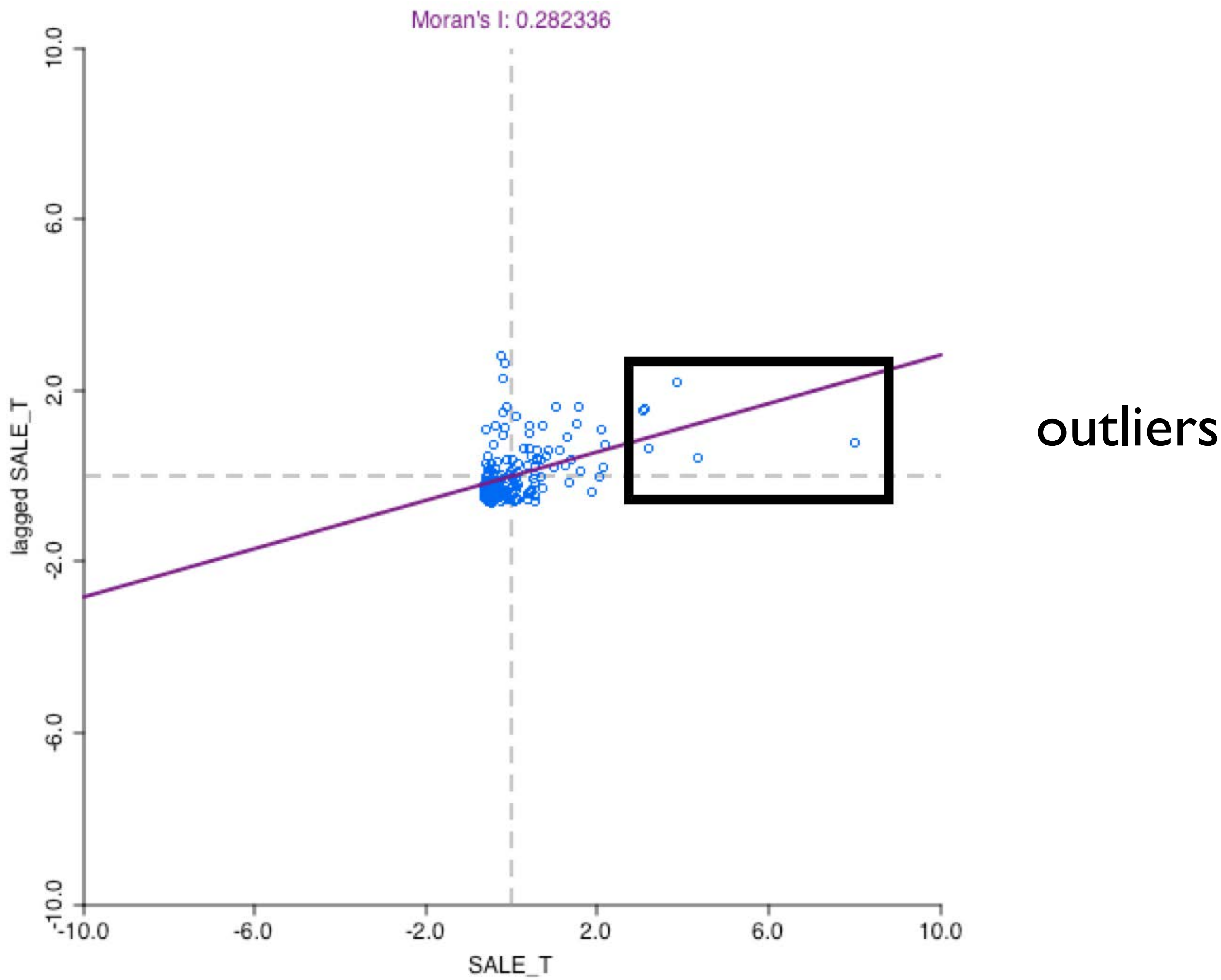
Moran Scatter Plot

- scatter plot of [z_i , Σ_j $w_{ij}.z_j$]

  the value at i on the x-axis, its spatial lag (weighted average of neighboring values) on the y-axis

  slope of linear fit is Moran's I

  use local regression (Lowess) to identify possible structural breaks



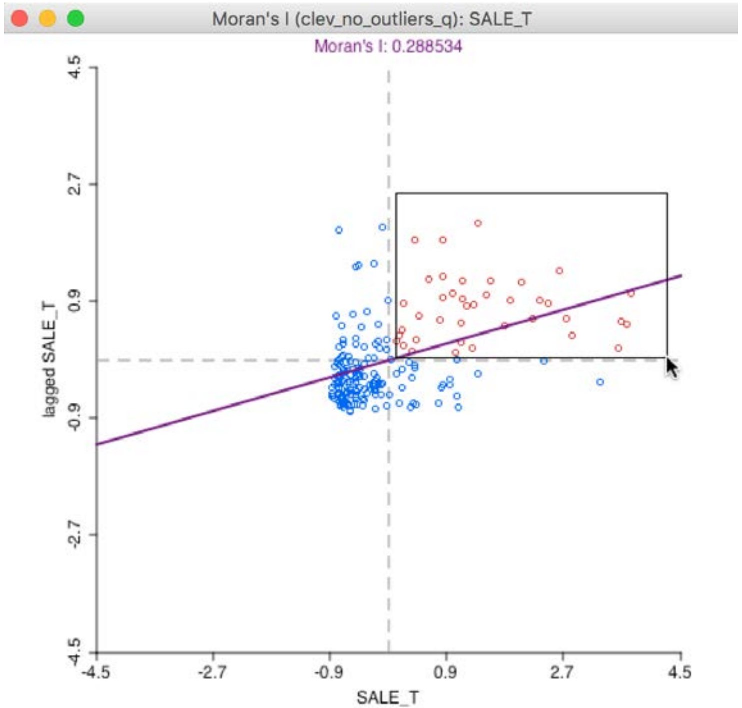
Moran scatter plot - Cleveland house prices - queen weights

- Categories of Local Spatial Autocorrelation
  - four quadrants of the scatter plot
  - upper right and lower left
    - positive spatial autocorrelation
    - clusters of like values
    - locations are similar to their neighbors
  - lower right and upper left
    - negative spatial autocorrelation
    - spatial outliers
    - locations are different from their neighbors

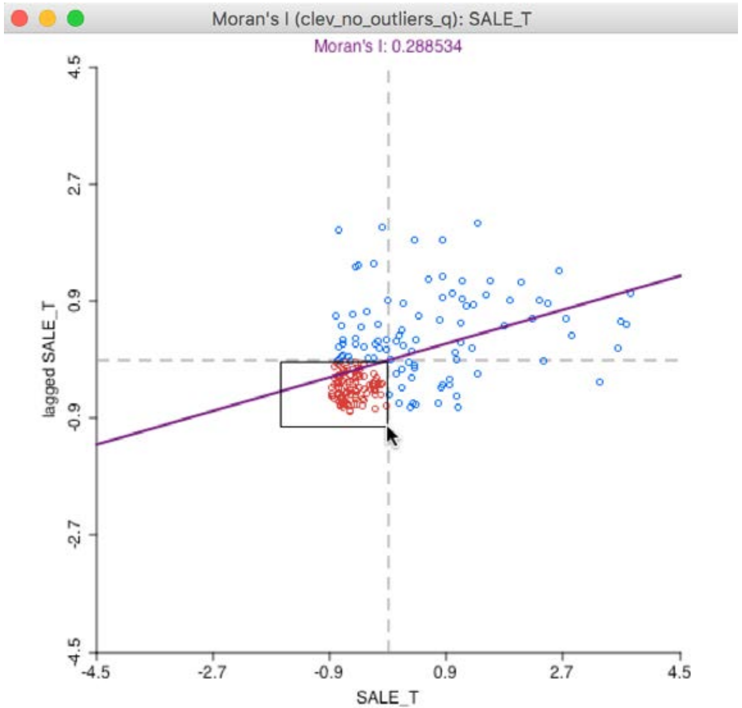
Positive Spatial Autocorrelation

all comparisons relative to the mean

not absolute high or low



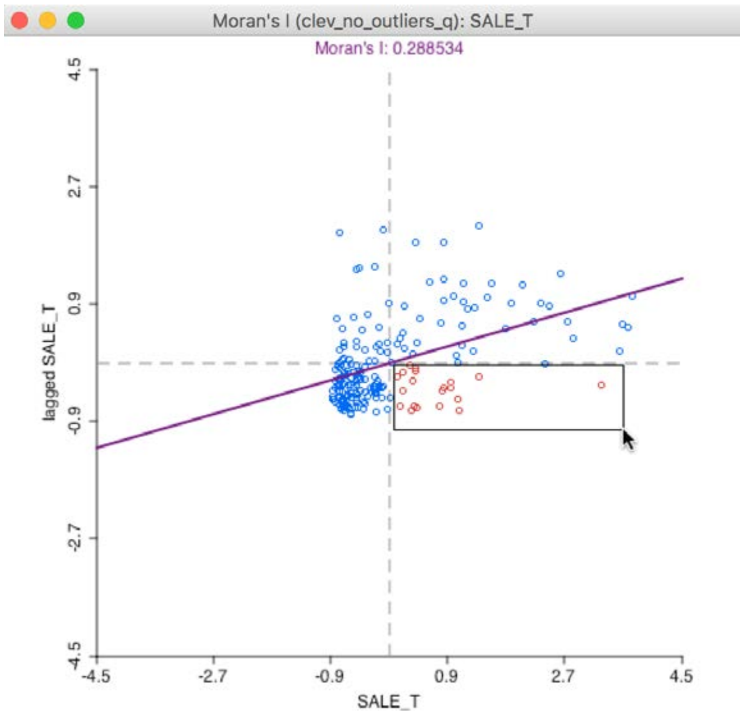
high-high



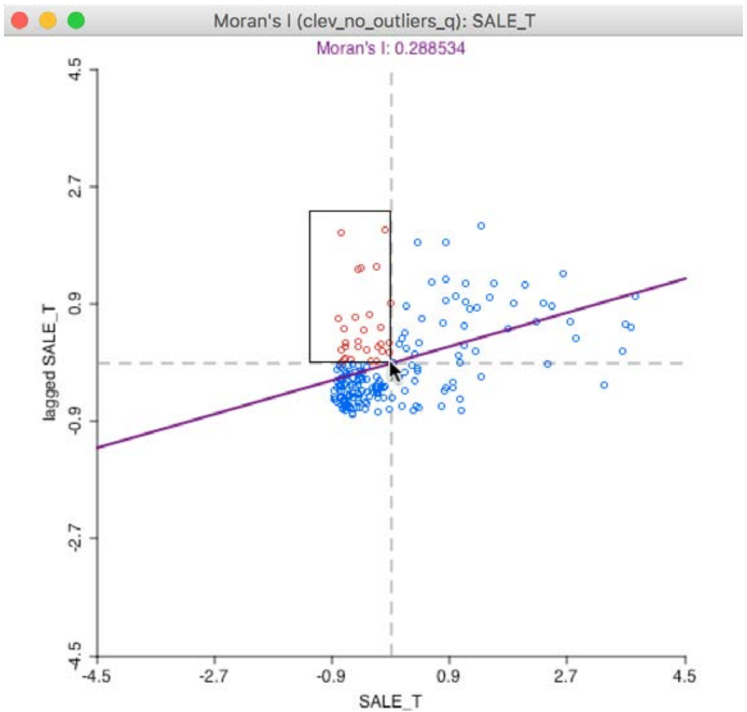
low-low

- Negative Spatial Autocorrelation

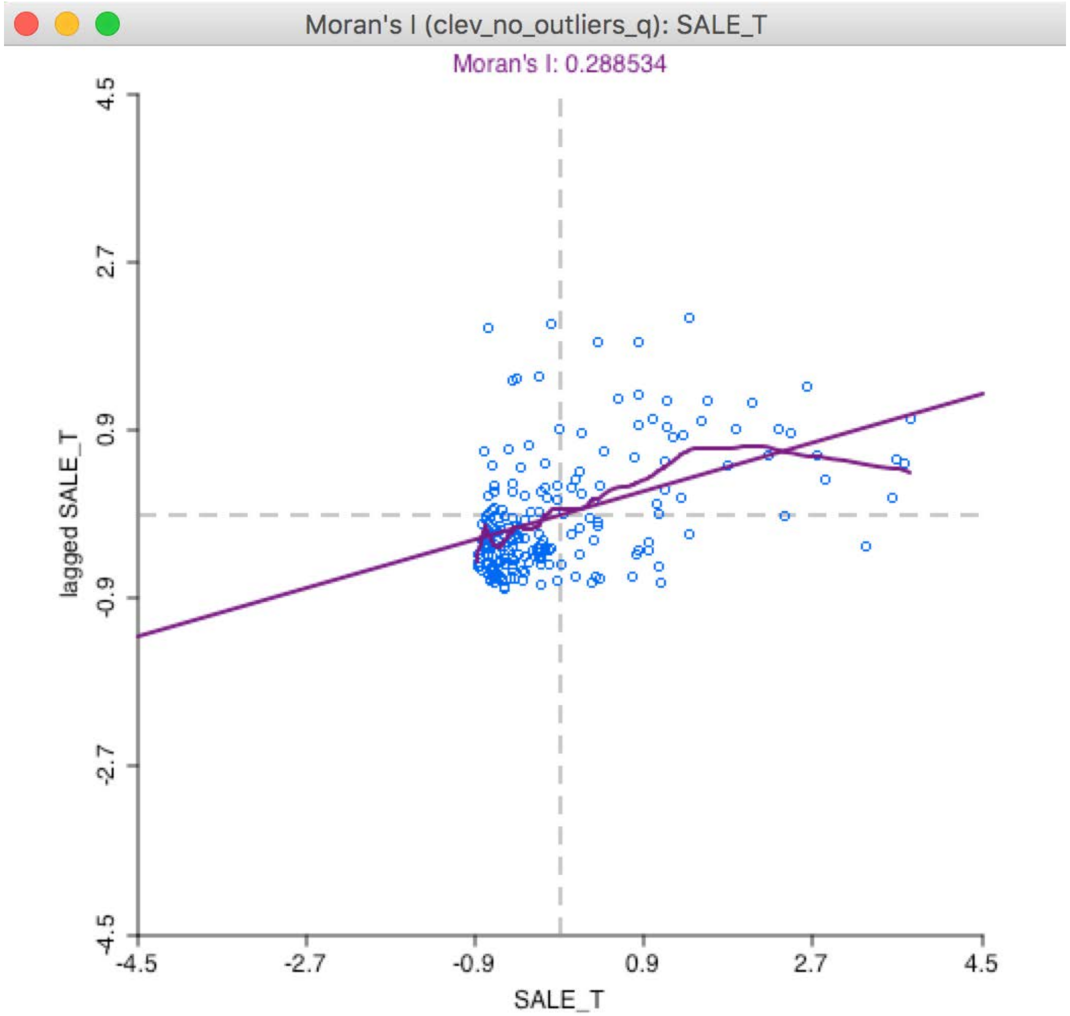
    spatial outliers



high-low



low-high



Moran scatter plot with Lowess smoother

Correlogram

- Alternative Perspective - Nonparametric

  non-parametric approach

    no model for covariance

    let the data suggest the functional form

  based on sample autocorrelation

  covariance function must be positive definite

- Sample Autocorrelation

computed for each pair i, j

$\rho_{ij} = \rho(z_i,z_j) = z_i^* z_j^* / (1/n) \Sigma_h (z_h - z_m)^2$

$z_i^* = z_i - z_m$ deviations from mean

in practice, easier to use standardized z_i

incidental parameter problem

one parameter for each pair i-j

n(n-1)/2 individual values of ρ_{ij}

- Non-Parametric Principle

spatial autocorrelation as an unspecified function of distance

$\rho_{ij} = g(d_{ij})$

how to fit the function? use kernel estimator

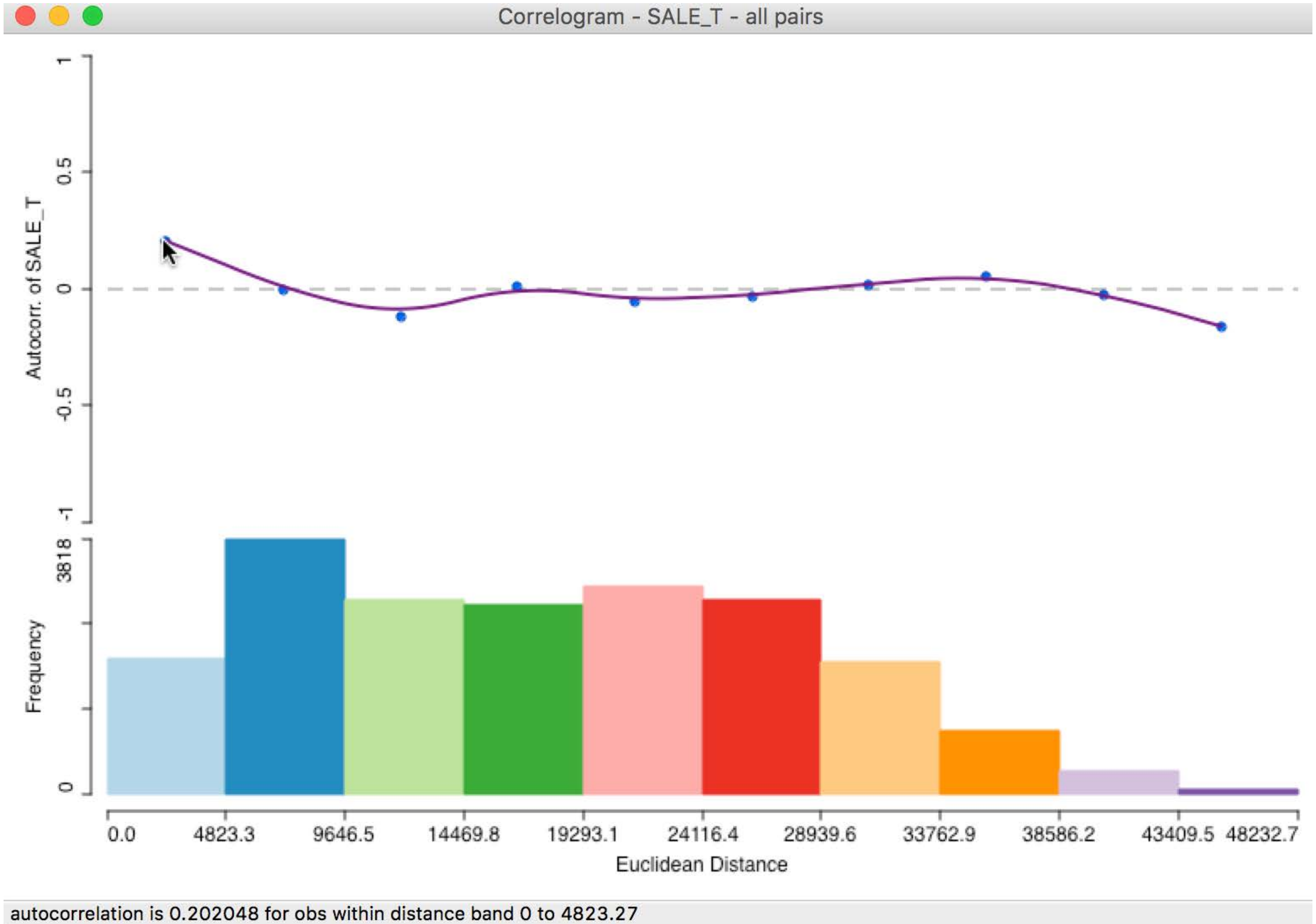
- Kernel Regression

$z_i z_j = g(d_{ij})$

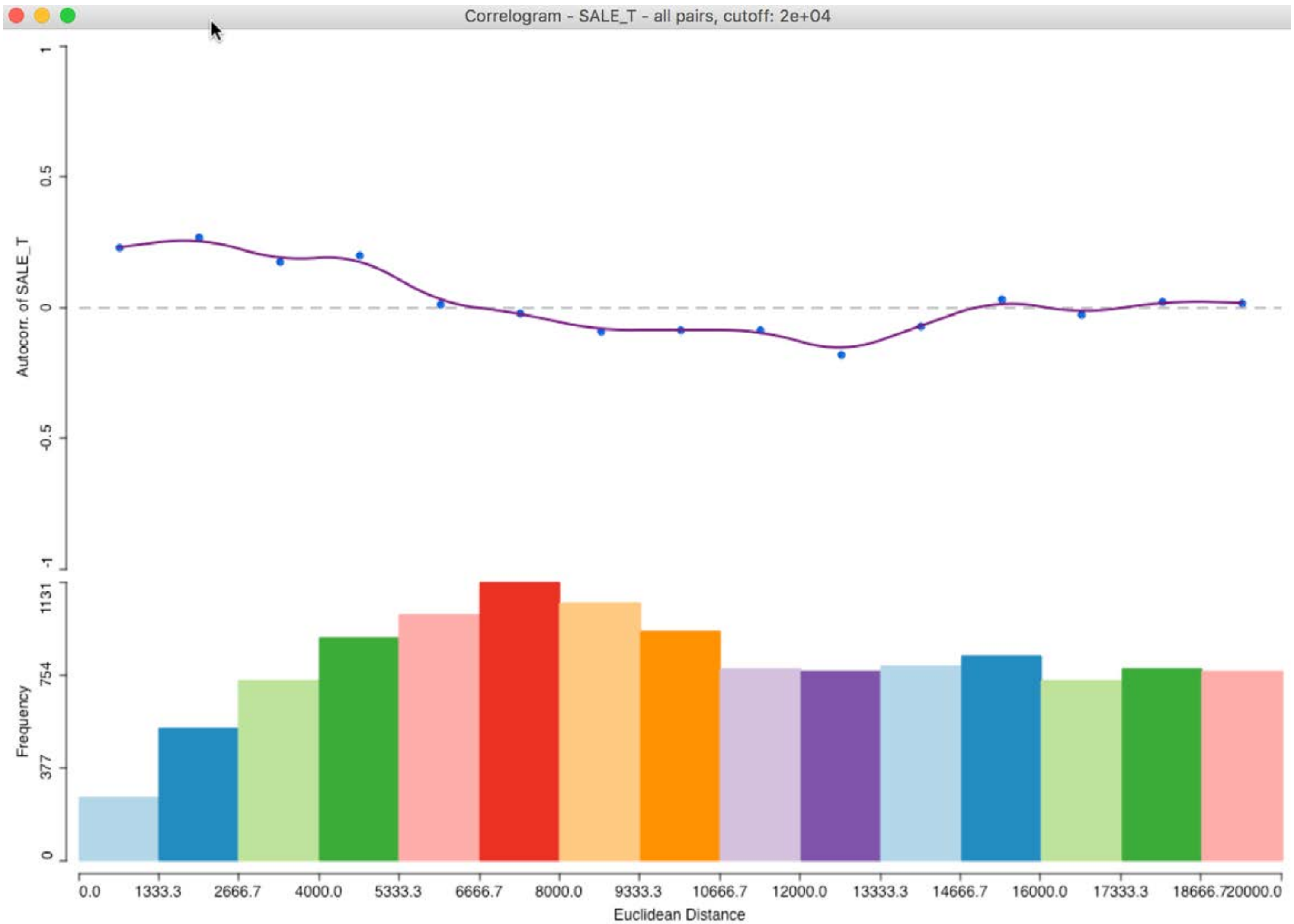
local regression

depends on choice of kernel function and bandwidth

values of the estimated $g(d_{ij})$ do not necessarily result in a valid (positive semi-definite) variance-covariance matrix



correlogram - full distance range



correlogram - 20,000 ft max distance

- Interpretation

    range of spatial autocorrelation

    alternative to specifying spatial weights

    sensitive to kernel fit

    may violate Tobler's law

Variogram

Semi-Variogram

- Variogram Function

magnitude of the variance of the difference as a function of displacement (h)

$$2\gamma(h) = \mathrm{Var}\,[\, Z_{s+h} - Z_s \,]$$

factor 2 is by convention, so half of this function is the semi-variogram

$$\gamma(h) = (1/2)\,\mathrm{Var}\,[\, Z_{s+h} - Z_s \,]$$

- Operational Semi-Variogram

constant mean assumption

$E[Z_{s+h} - Z_s] = E[Z_{s+h}] - E[Z_s] = 0$

such that $Var[Z_{s+h} - Z_s] = E[Z_{s+h} - Z_s]^2 - 0$

semi-variogram

$\gamma(h) = (1/2)\, E[Z_{s+h} - Z_s]^2$

estimate as average of the squared differences

- Variogram Cloud Plot

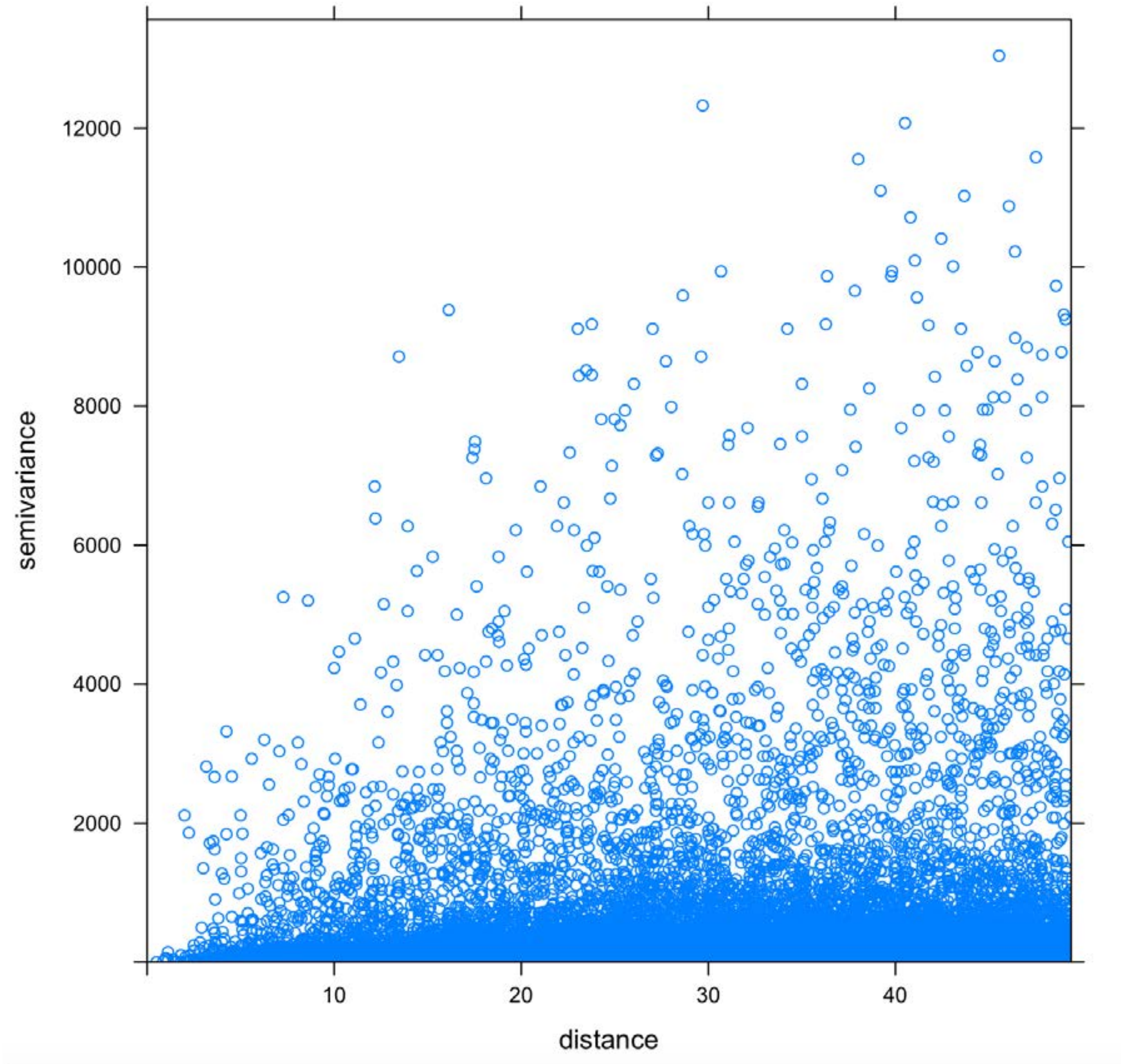
  plot of all squared differences against the distance separating the pair of observations

  exploring the variogram cloud plot

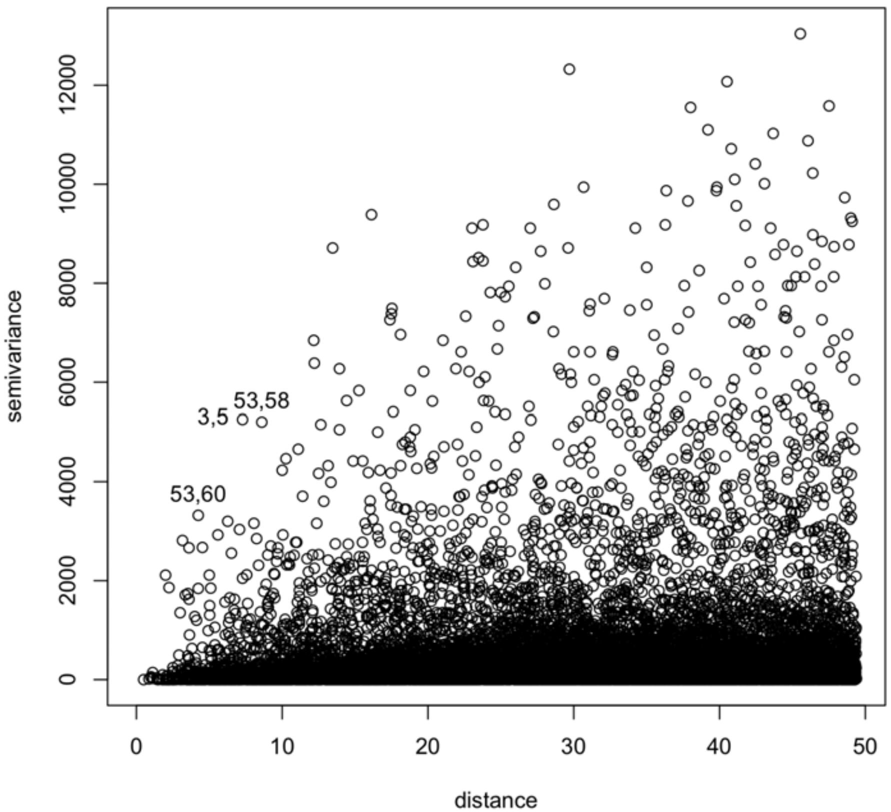
  identify outliers

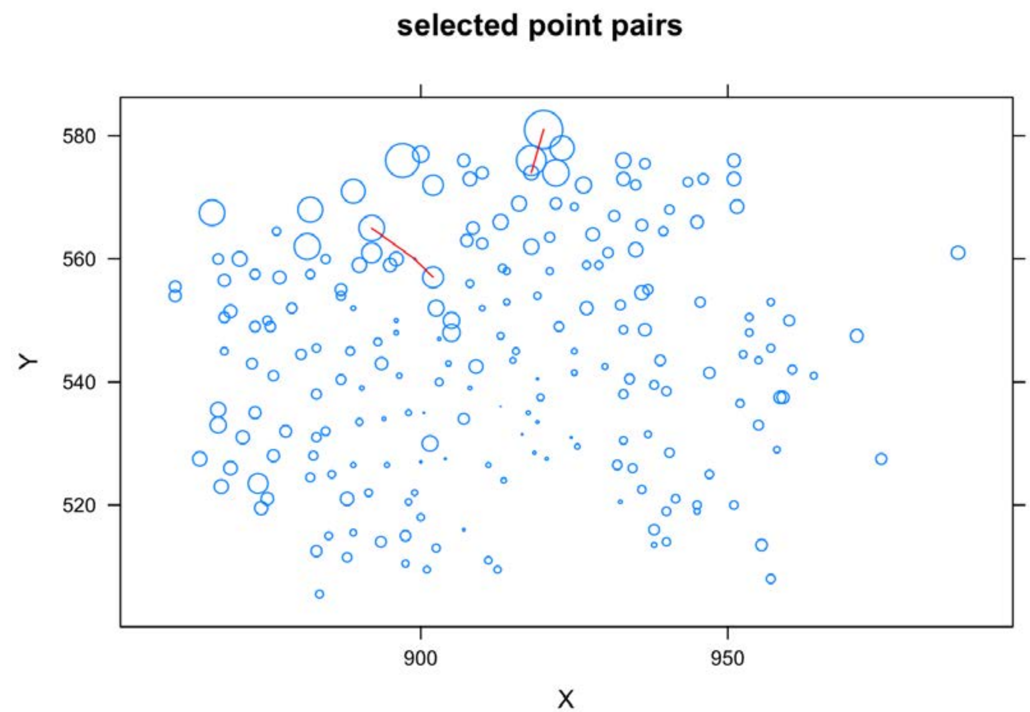
    large difference for small distance

    negative spatial autocorrelation



Variogram Cloud Plot
Baltimore House Prices





Checking for outliers

The Empirical Variogram

- Estimating the Empirical Variogram

  method of moments

  bin the pairs by distance bands

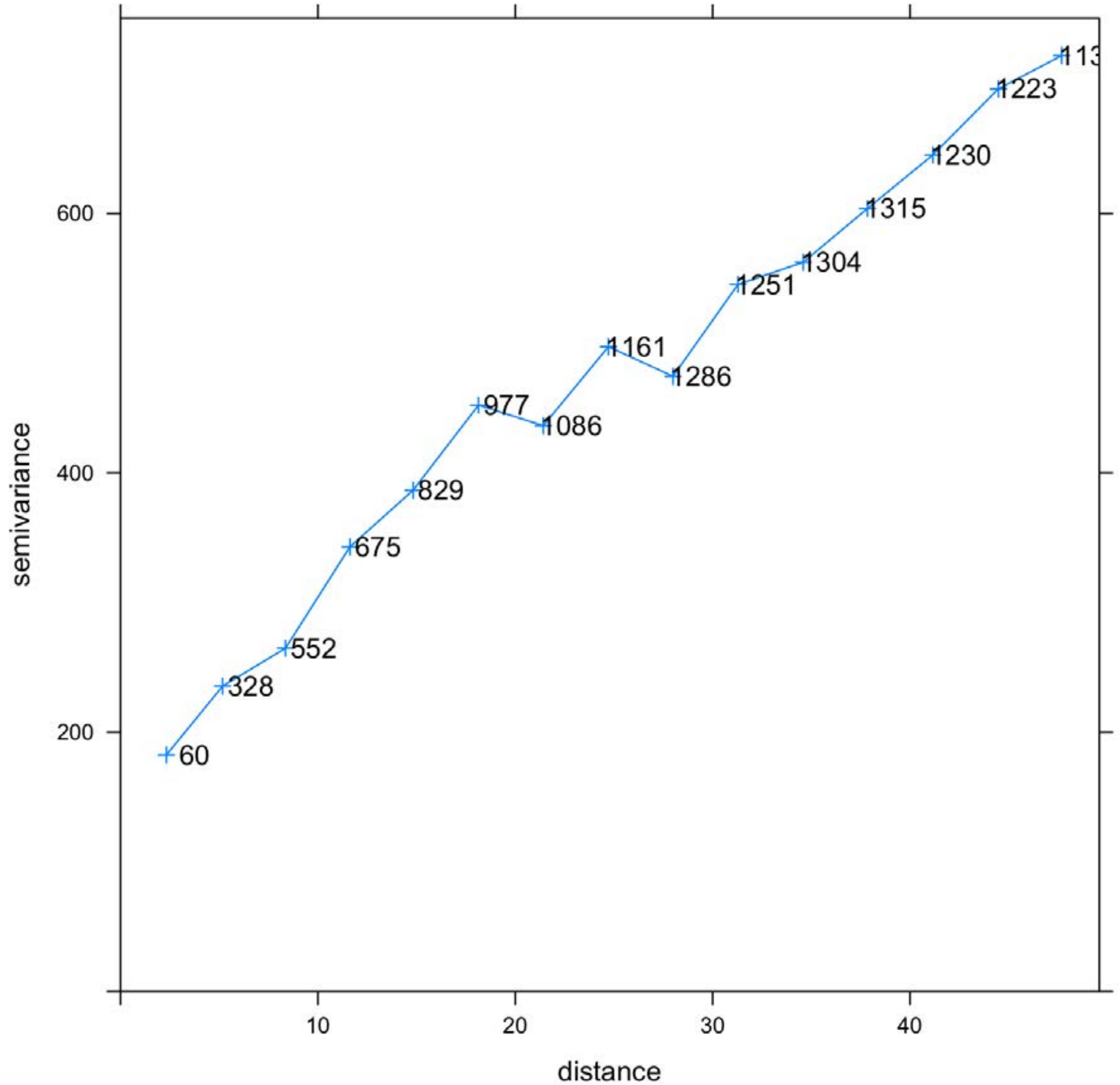
  average squared difference within bin

  - $2\gamma(h) = (1 / |N(h)|)\ \Sigma_h\ [Z_{s+h} - Z_s]^2$

    N(h) number of pairs in distance bin h

- Practical Issues
    - maximum distance = “distance of reliability”
    - different rules of thumb
        - h < D/2 (D is maximum distance)
        - h < d/3 (d is diagonal of bounding box)
    - bin width
        - avoid sparse bins
        - at least 30 pairs
        - precision of estimate will depend on number of pairs in the bin

	np	dist	gamma
1	60	2.325018	182.4068
2	328	5.171430	235.3985
3	552	8.358653	264.7059
4	675	11.626348	342.9250
5	829	14.824826	386.5458
6	977	18.136492	452.1512
7	1086	21.419961	436.3443
8	1161	24.718931	497.1898
9	1286	27.991568	474.3808
10	1251	31.285467	545.2776
11	1304	34.588754	562.3477
12	1315	37.838726	603.7175
13	1230	41.159131	645.1485
14	1223	44.475928	696.0524
15	1138	47.681960	721.7097



Empirical Variogram
Baltimore house prices

- Issues

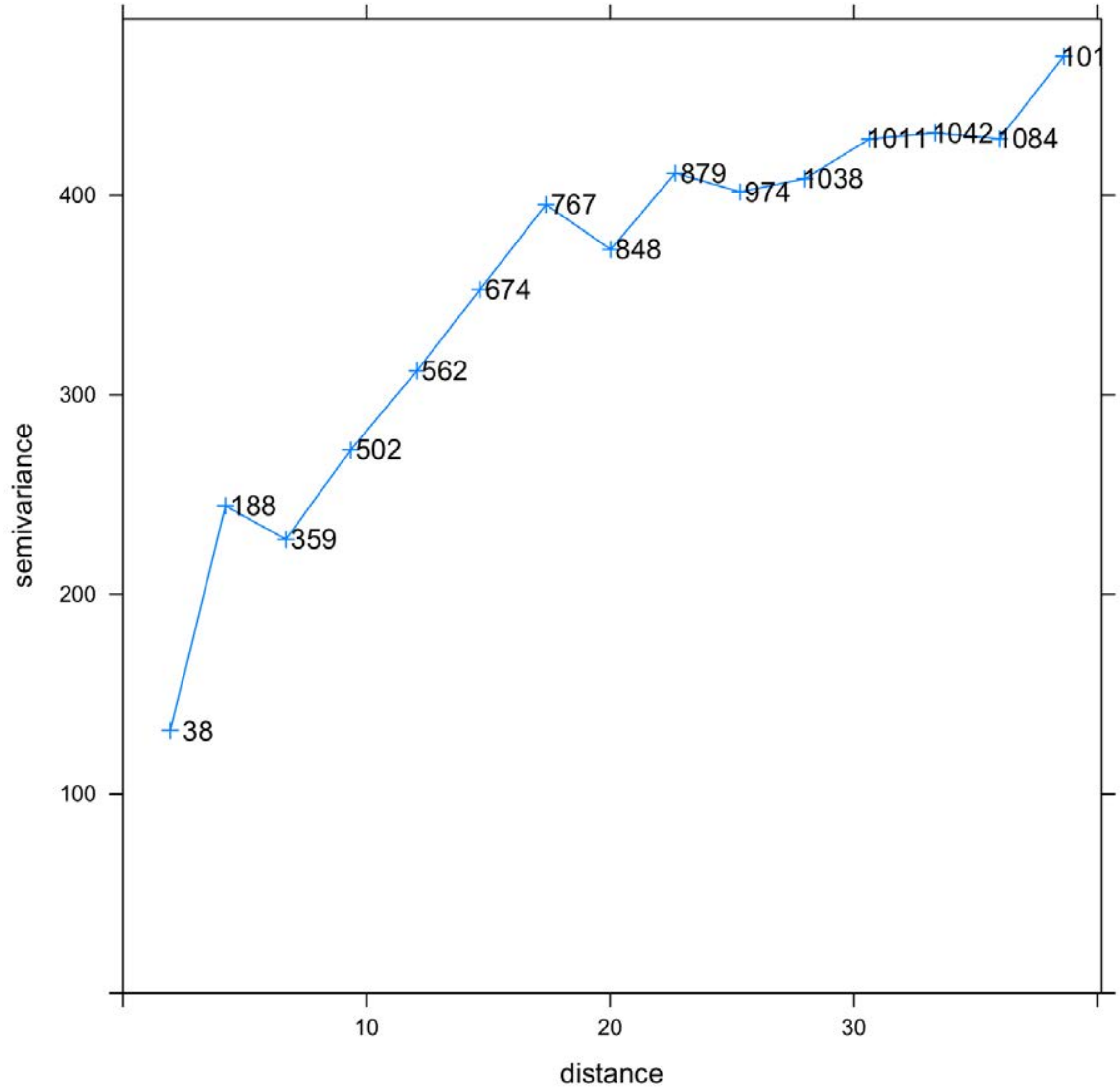
    variogram should increase with distance, but level off at some point

    point beyond it levels off = no more spatial autocorrelation

    when variogram keeps increasing = evidence of a trend in the data

    solution = use regression residuals

	np	dist	gamma
1	38	1.942689	131.6754
2	188	4.212751	244.3271
3	359	6.694554	227.5859
4	502	9.348142	272.4303
5	562	12.074442	312.1512
6	674	14.654980	352.8308
7	767	17.367372	395.3913
8	848	20.025210	372.9660
9	879	22.668438	411.1619
10	974	25.334278	401.7390
11	1038	27.987936	408.3673
12	1011	30.645465	428.3870
13	1042	33.334552	431.3248
14	1084	35.982739	428.5149
15	1014	38.621472	469.8098



Empirical Variogram

Baltimore second order trend surface residuals

- Adjusting the Empirical Variogram

    adjust the cut-off distance

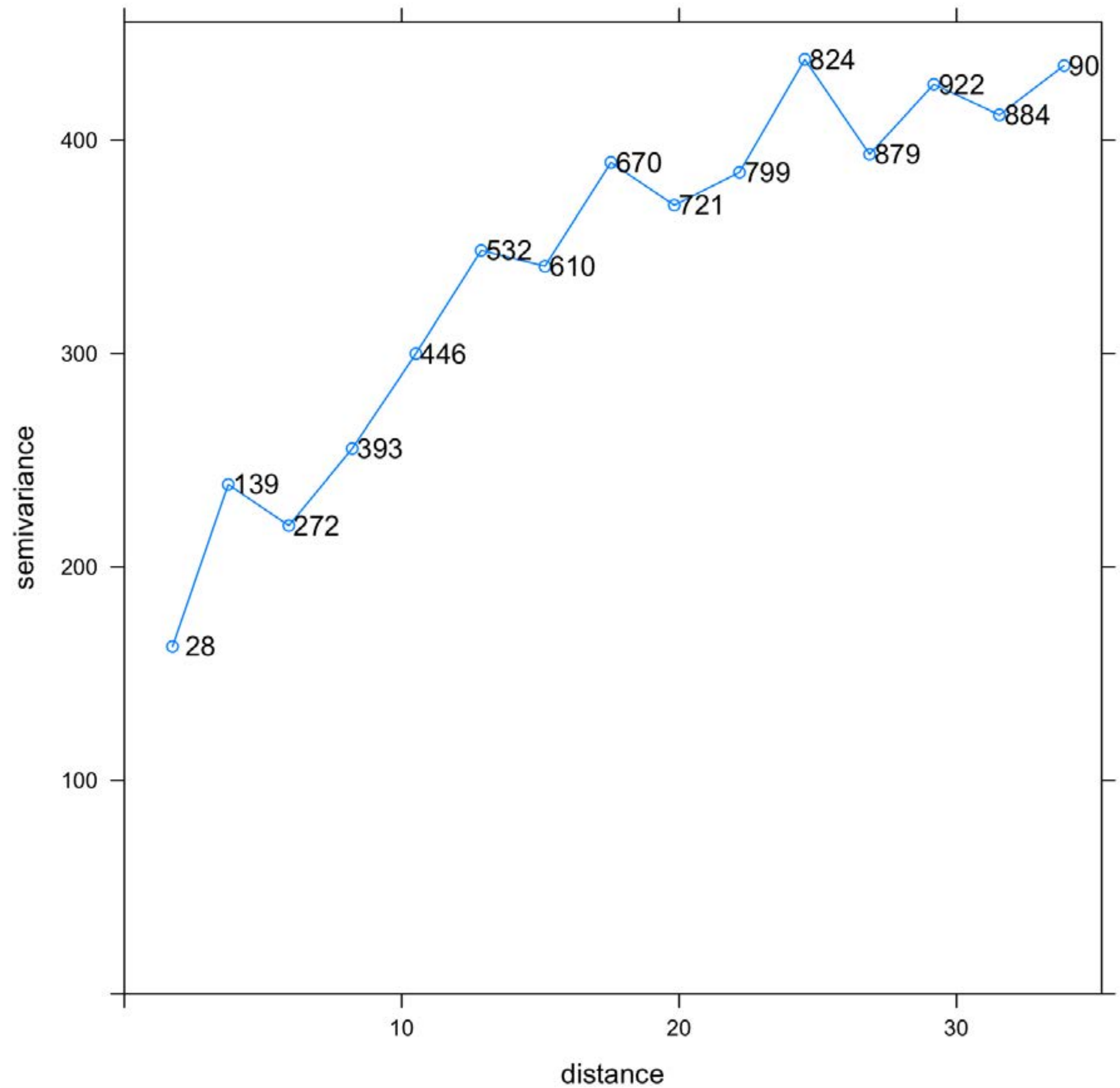
    shorter distances are where the interest is

    larger distances

        tend not to be spatially correlated

        fewer pairs to estimate variogram from

	np	dist	gamma
1	28	1.738344	162.8340
2	139	3.747141	238.7456
3	272	5.932211	219.3918
4	393	8.216204	255.5388
5	446	10.510220	299.9714
6	532	12.864325	348.3107
7	610	15.164216	340.9232
8	670	17.540884	389.5590
9	721	19.826468	369.5056
10	799	22.176412	384.9465
11	824	24.532378	437.8523
12	879	26.866244	393.3482
13	922	29.181277	426.1453
14	884	31.545530	411.8150
15	903	33.877861	434.9531



Empirical Variogram

Baltimore trend surface residuals - cut off d=35

Directional Effects

- Anisotropy - Directional Variogram

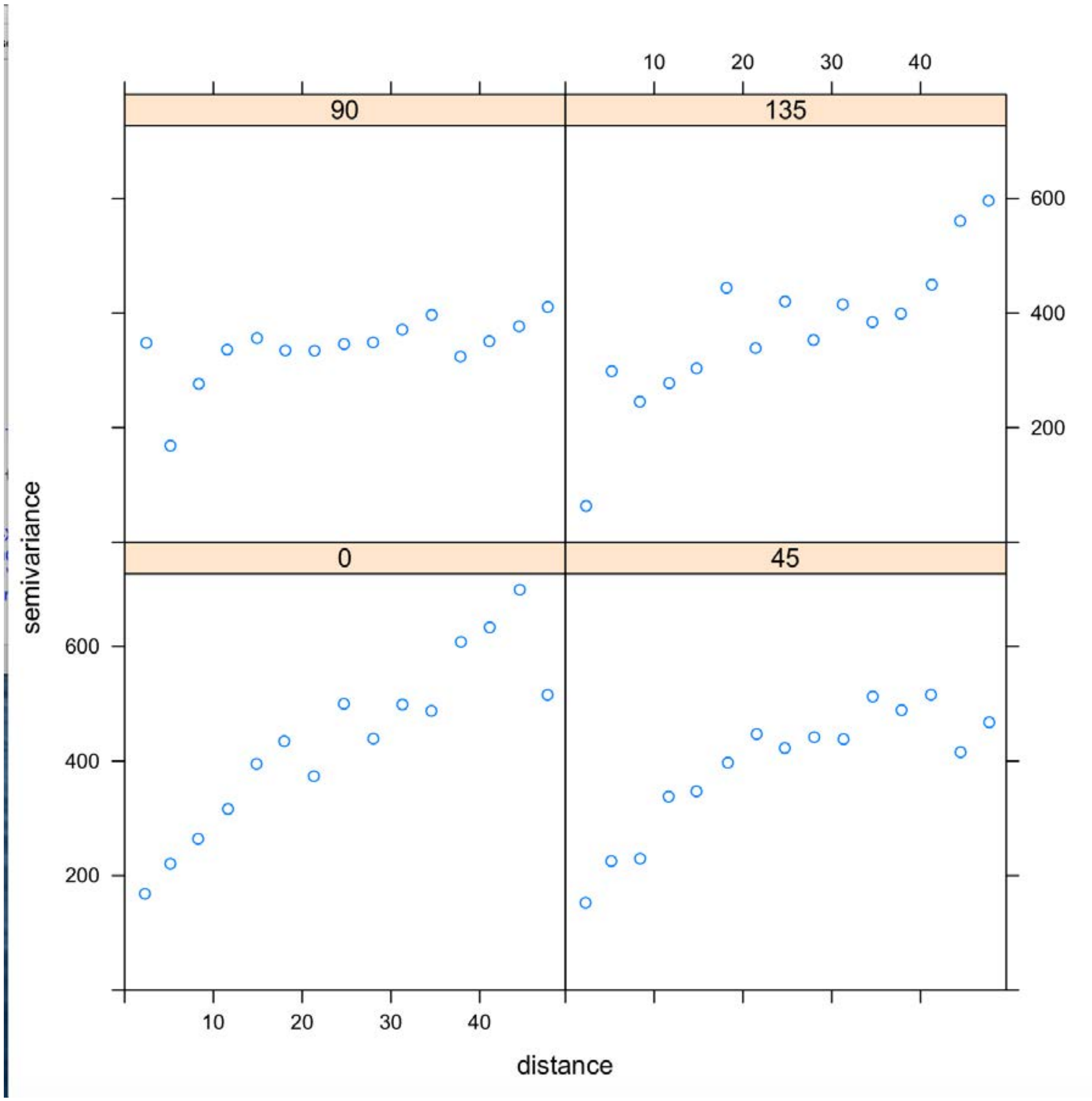
  fundamental assumption = isotropy

    only distance matters, direction doesn't matter

  directional semi-variogram

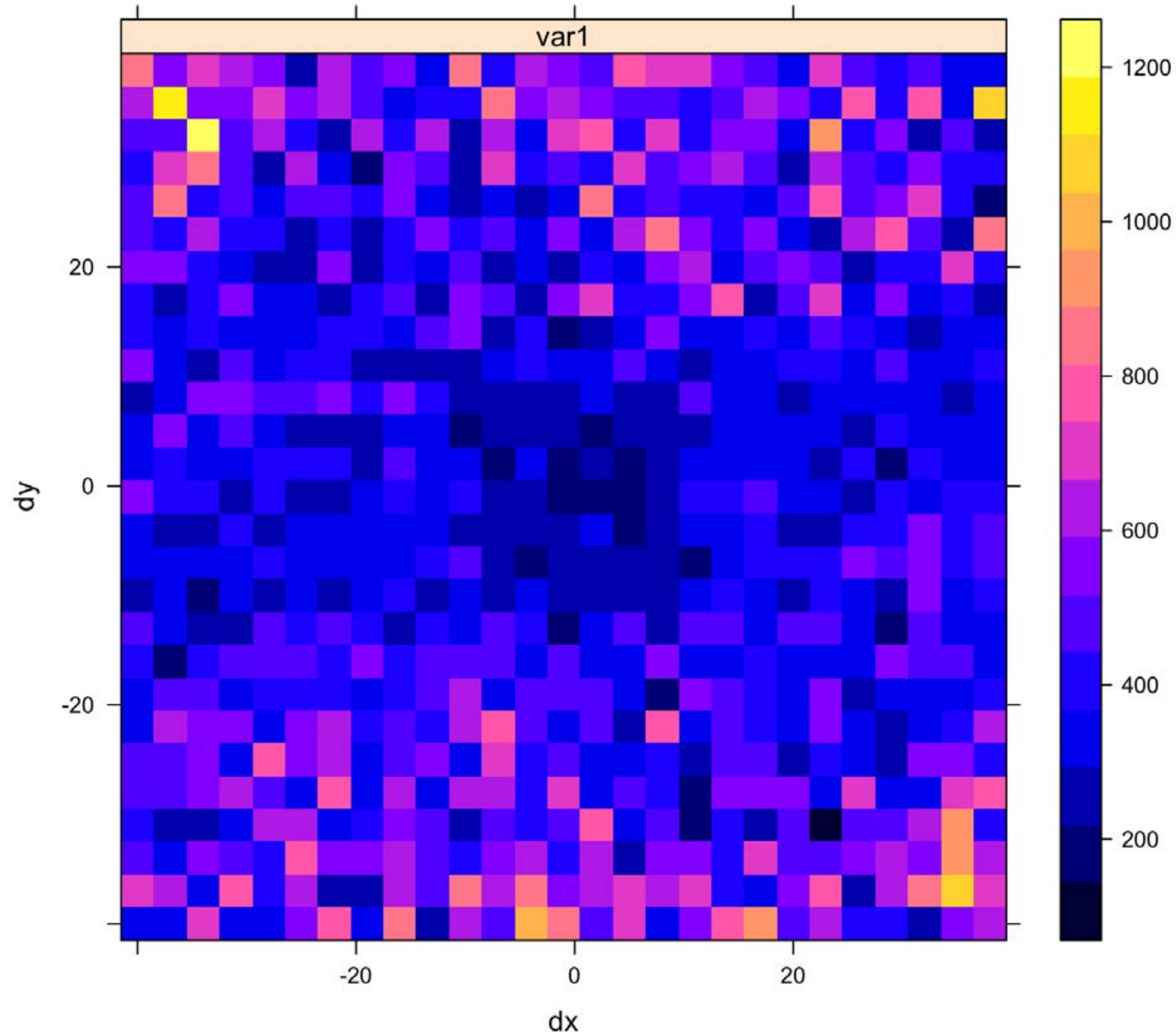
    different variogram for point pairs in angular sections

    typically implemented by 45 degree sections

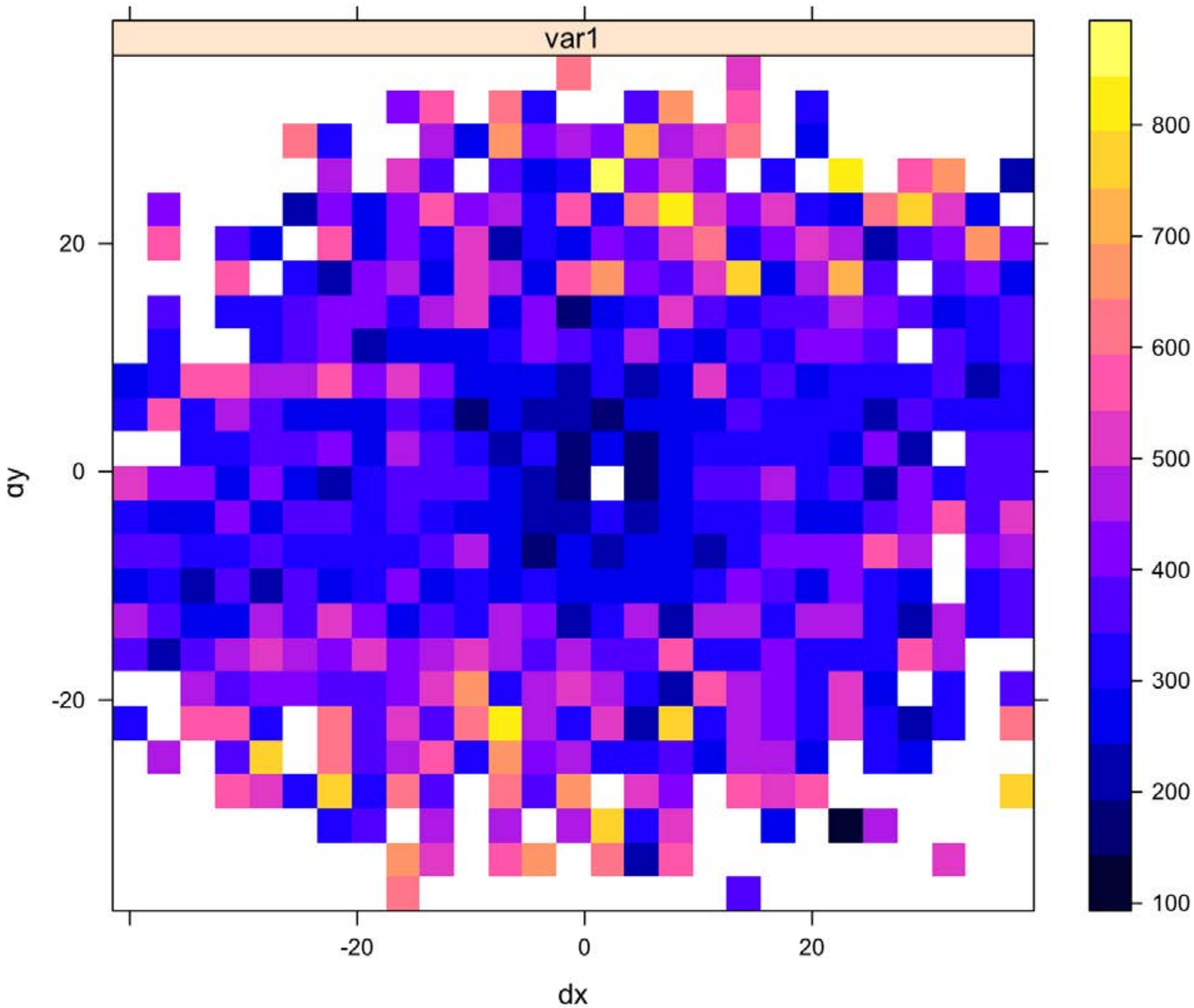


Directional variogram

- Anisotropy - Variogram Map
  - alternative visualization of semi-variogram values
  - centered on 0,0
    - rectangles for distance in E-W direction (dx) and N-S direction (dy)
    - semi-variogram computed for pairs in rectangle
  - variogram map should be concentric
    - deviations point to anisotropy
    - eliminate cells with fewer than “n” pairs



Variogram map - all pairs



Variogram map with more than 30 pairs

Variogram Models

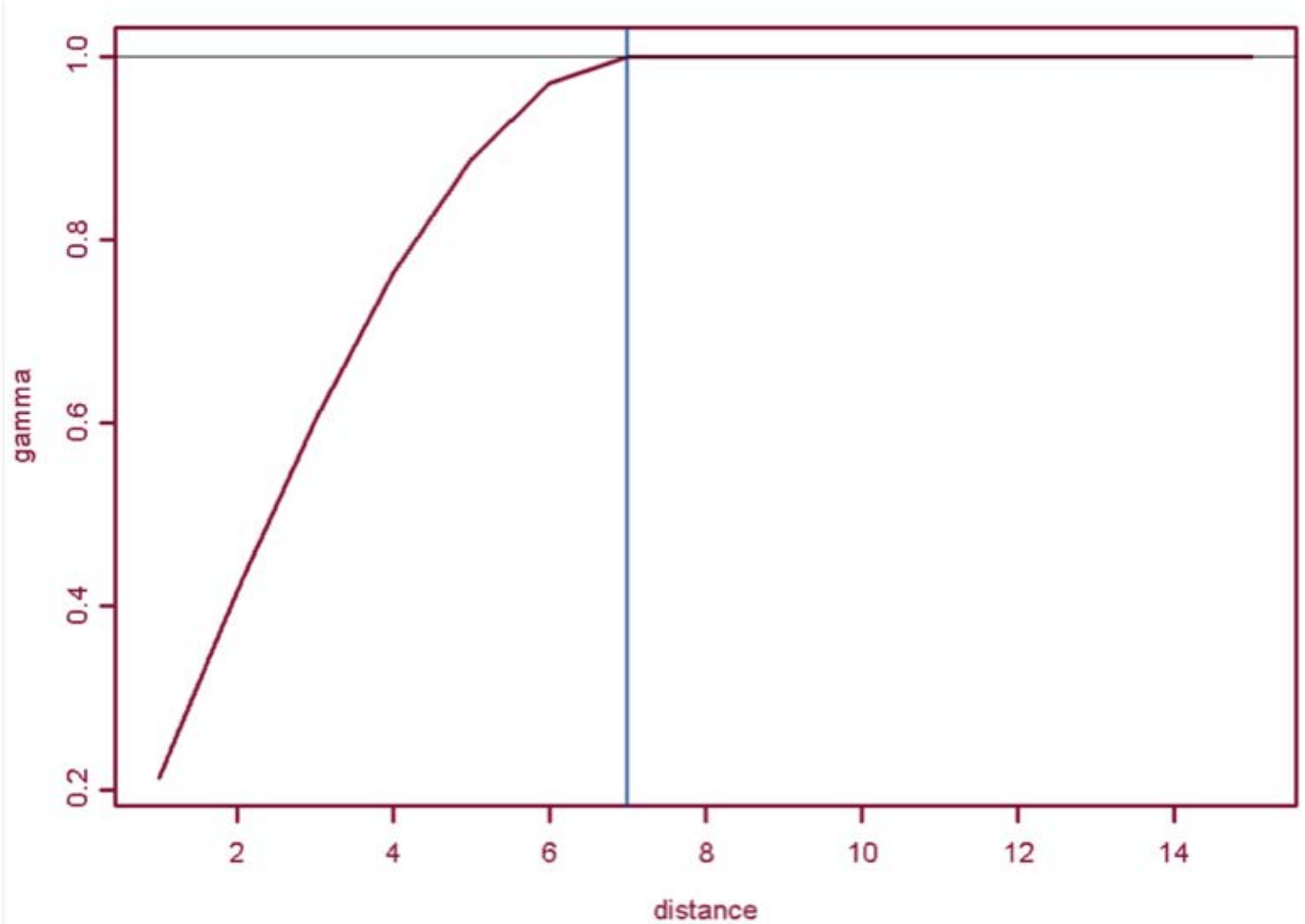
- Terminology

sill = process variance

range = distance beyond which there is no more spatial autocorrelation

nugget = positive variance at distance 0
(variance at 0 should be 0)

sill=1



Range = 7

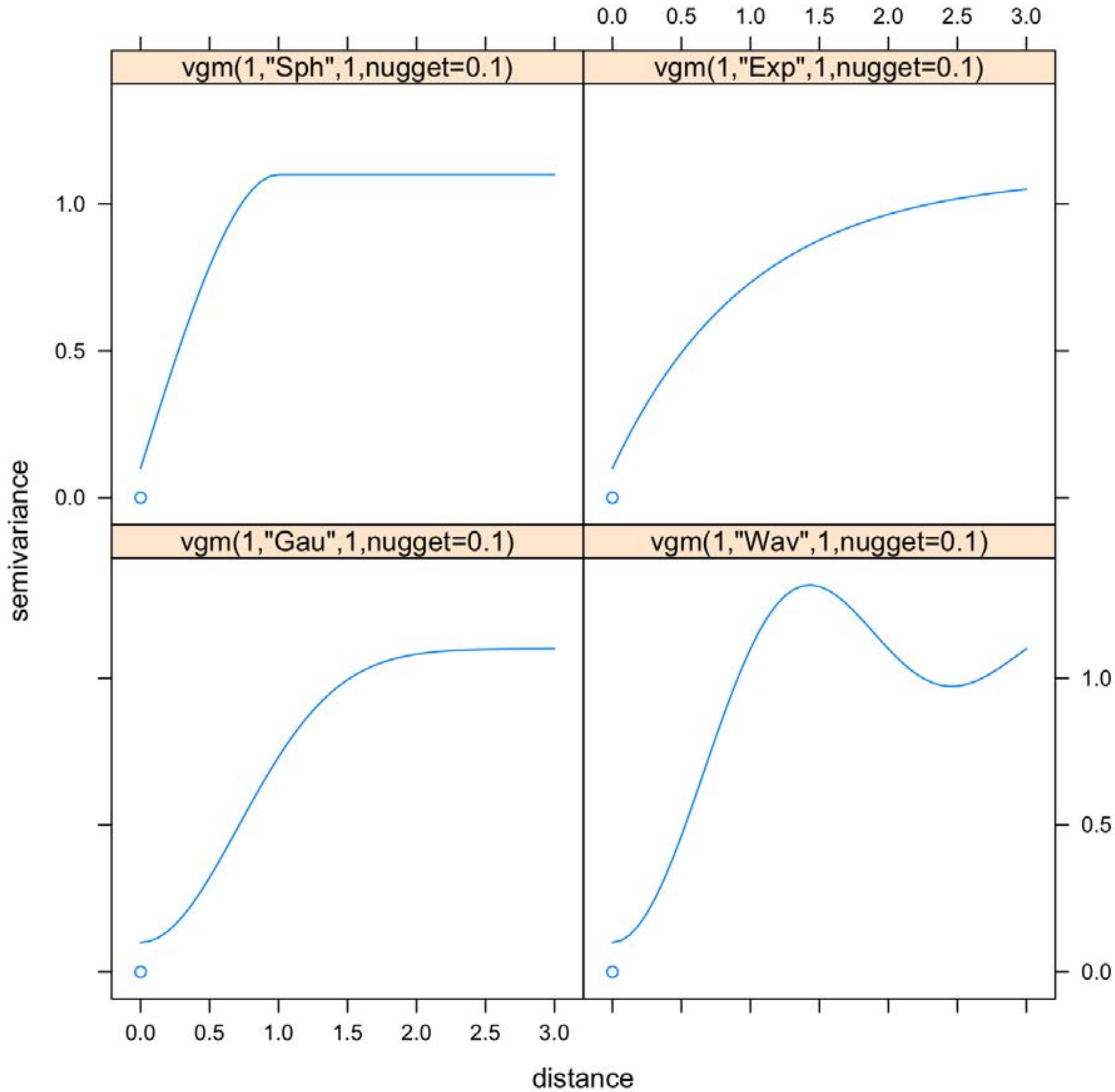
- Valid Variogram Models

  need to ensure that variance-covariance matrix is positive definite

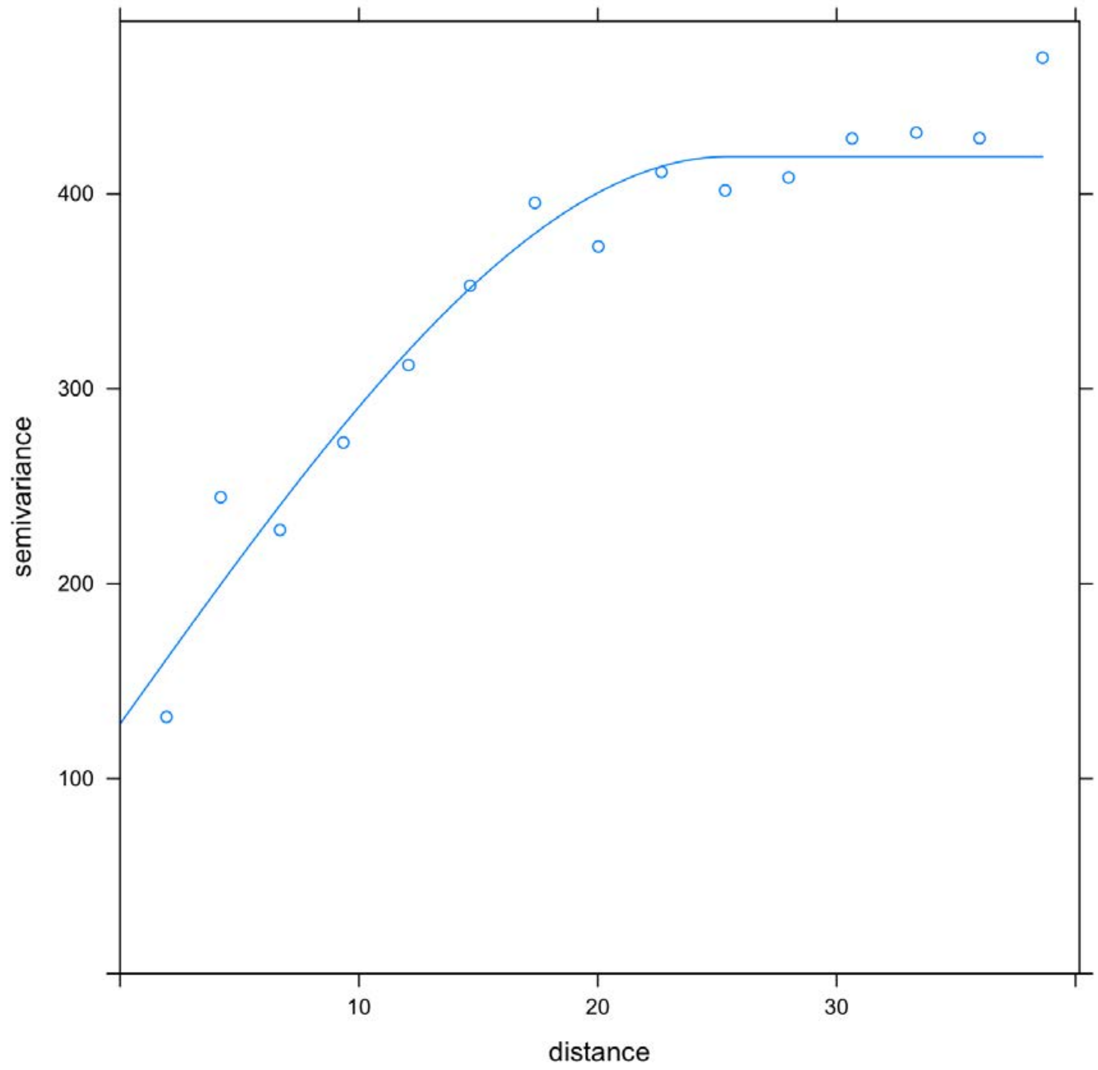
  positive definite $C(h)$ for all h

  negative definite $\gamma(h)$ for all h

  leads to a collection of valid parametric models (with limits on the parameter space)



spherical, exponential, Gaussian and wave variogram models



best fit spherical model
nugget = 128.04, partial sill = 291.01, range = 25.46
SSErr = 36852.6