An Analytical Evaluation of the Co-variations among Local Moran's I_is in Autocorrelated Map Patterns

by

Michael Tiefelsdorf
tiefelsdorf@utd.edu
The University of Texas at Dallas

Sweta Jinnagara Puttaswamy
The University of Texas at Dallas

Paper presented at the North-American Regional Science Association
Denver, November 10-13, 2010

INTRODUCTION: LISA

- Local indicators of spatial association (LISA) are frequently used to identify
  - Local misspecifications of the first order trend component.
  - Heterogeneities within the stochastic second order component of a spatial processes.
- Communal assessment of a set of LISA's and multiple comparisons requires
  - knowledge of their joint distribution, which has to be
  - conditional to an underlying global spatial process.
- Knowledge of the pairwise correlations among LISAs is a step towards a solution of the multiple comparison problem.

<u>Focus of this study</u>

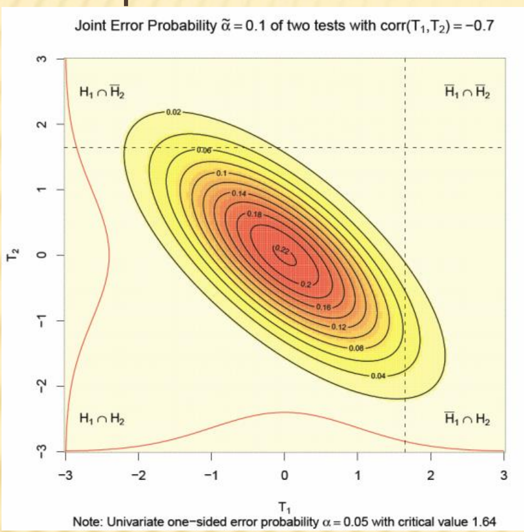
- Results are shown for the correlation among local Moran's I_is subject to a global spatial process in hexagonal tessellations.
- Several cross-checks were performed, analytical results against simulation results.

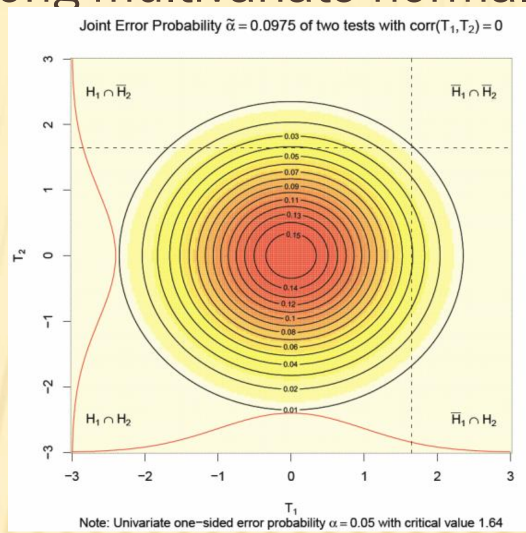
Developments Around Moran's *I*

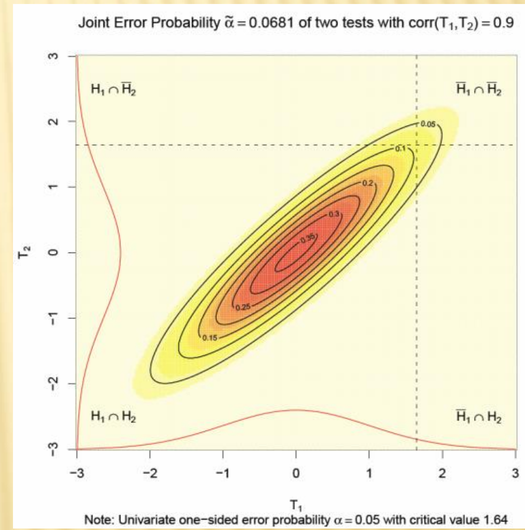
- **Past:**
  - *Moran*, 1950. Proposal of the spatial autocorrelation test statistic
  - *Imhof*, 1961. Univariate distribution through numerical inversion
  - *Sawa*, 1978. Higher order conditional moments
  - *Shepard*, 1991. Joint distribution through multivariate numerical inversion.
  - *Anselin*, 1995, Introduction of the local Moran's I_i concept.
    $\Longrightarrow$ Reformulated by *Tiefelsdorf* , 1996, as a ratio of quadratic forms.
  - *Tiefelsdorf*, 1998. Exact conditional distribution
  - *Tiefelsdorf*, 1998. First exploration of conditional correlation.
- **Present:**
  - Feasibility study and systematic exploration of correlation.
  - Implementation in an open source environment.
- **Future:**
  - Relaxation of the underlying Gaussian assumptions
  - Efficient approximations of the joint distribution based on pairwise correlation $\Longrightarrow$ addressing the multiple comparison problem
  - Increased numerical performance.

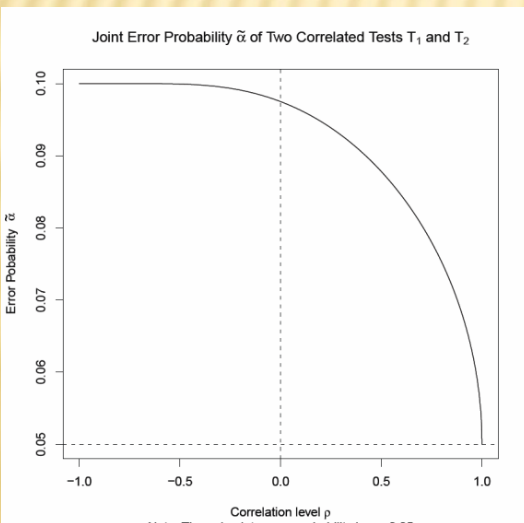
Review: Correlation and the Multiple Comparison Problem

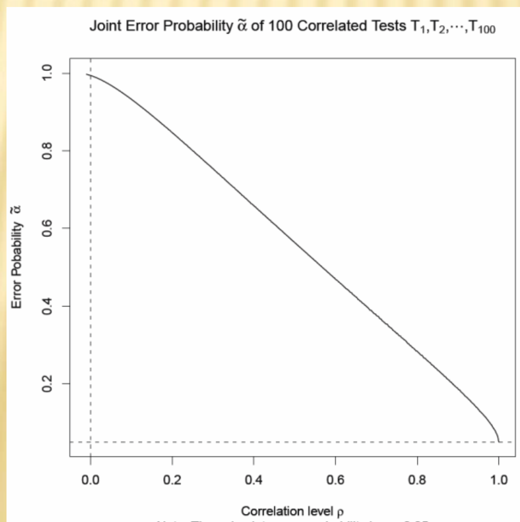
- The correlation among test statistics influences their joint probabilities.
- Example: Correlation among multivariate normal distributed test statistics

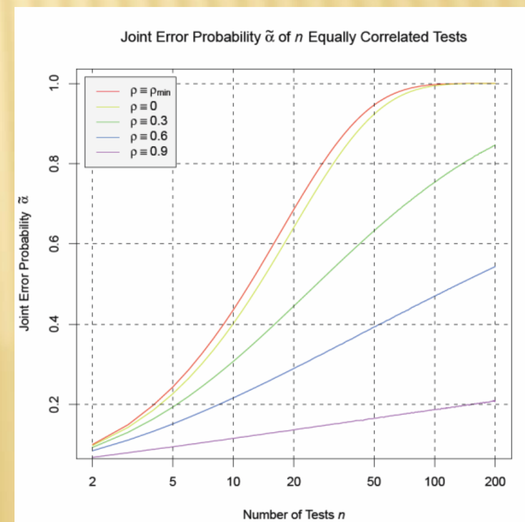












Review: Global and Local Moran's I_i

- Moran's *I* measures the degree of autocorrelation in regression residuals $\hat{\boldsymbol{\varepsilon}}$.
- Its support is bounded — but usually $I \notin [-1,1]$.
- Its observed value and distribution depend on the specification of [a] the first order spatial trend, [b] the coded spatial structure and [c] the underlying spatial process

$$I^{obs} \leftarrow f\big(\mathbf{y}(\mathbf{X},\mathbf{V},\boldsymbol{\Omega})\big) \text{ and } F(I) \leftarrow g(\mathbf{X},\mathbf{V},\boldsymbol{\Omega})$$

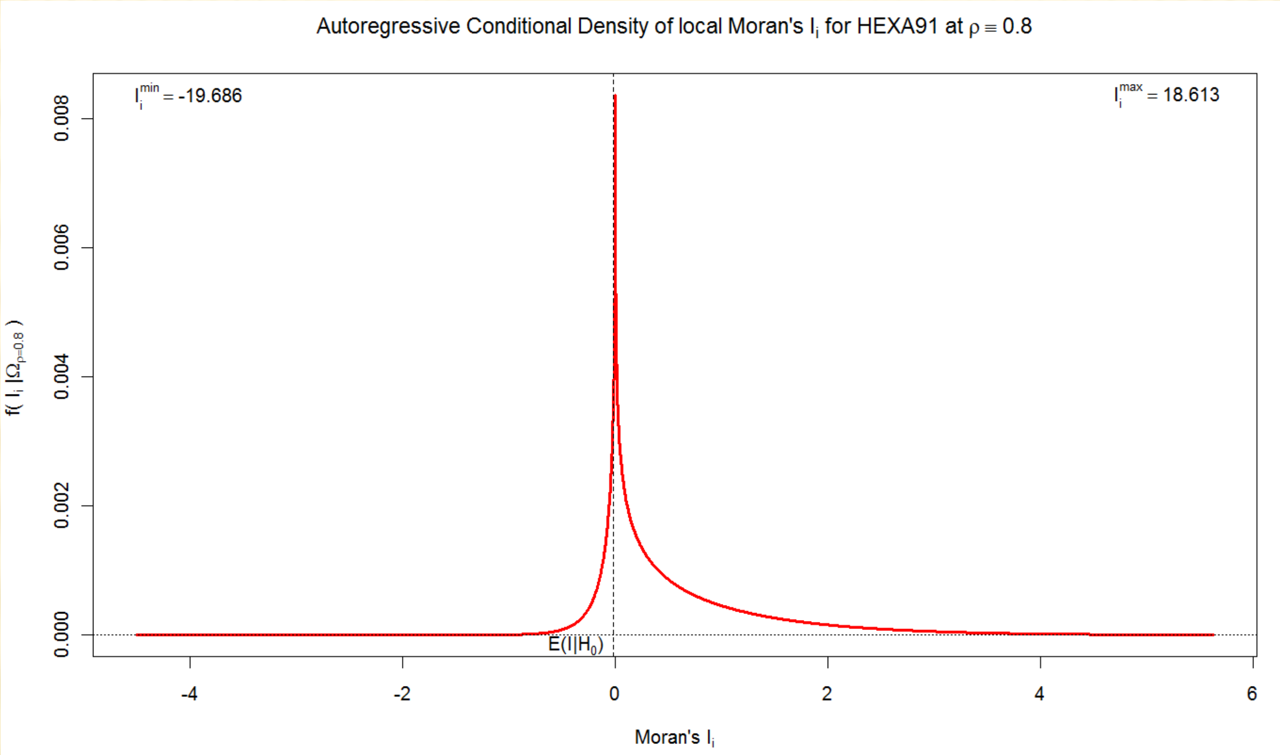
- The set of local Moran's I_is in a map pattern relate to global Moran's *I*

$$E(I|\boldsymbol{\Omega}) = \sum_{i=1}^{n} E(I_i|\boldsymbol{\Omega}) \text{ and}$$

$$Var(I|\boldsymbol{\Omega}) = \sum_{i=1}^{n}\sum_{j=1}^{n} Cov\big(I_i, I_j\big|\boldsymbol{\Omega}\big)$$

Review: Conditional Univariate Distribution of Local Moran's I_i

- Local Moran's I_i is not approximately normally distributed.
- Its mode is invariant of the global autocorrelation level $\rho \in [\rho_{min}, \rho_{max}]$.



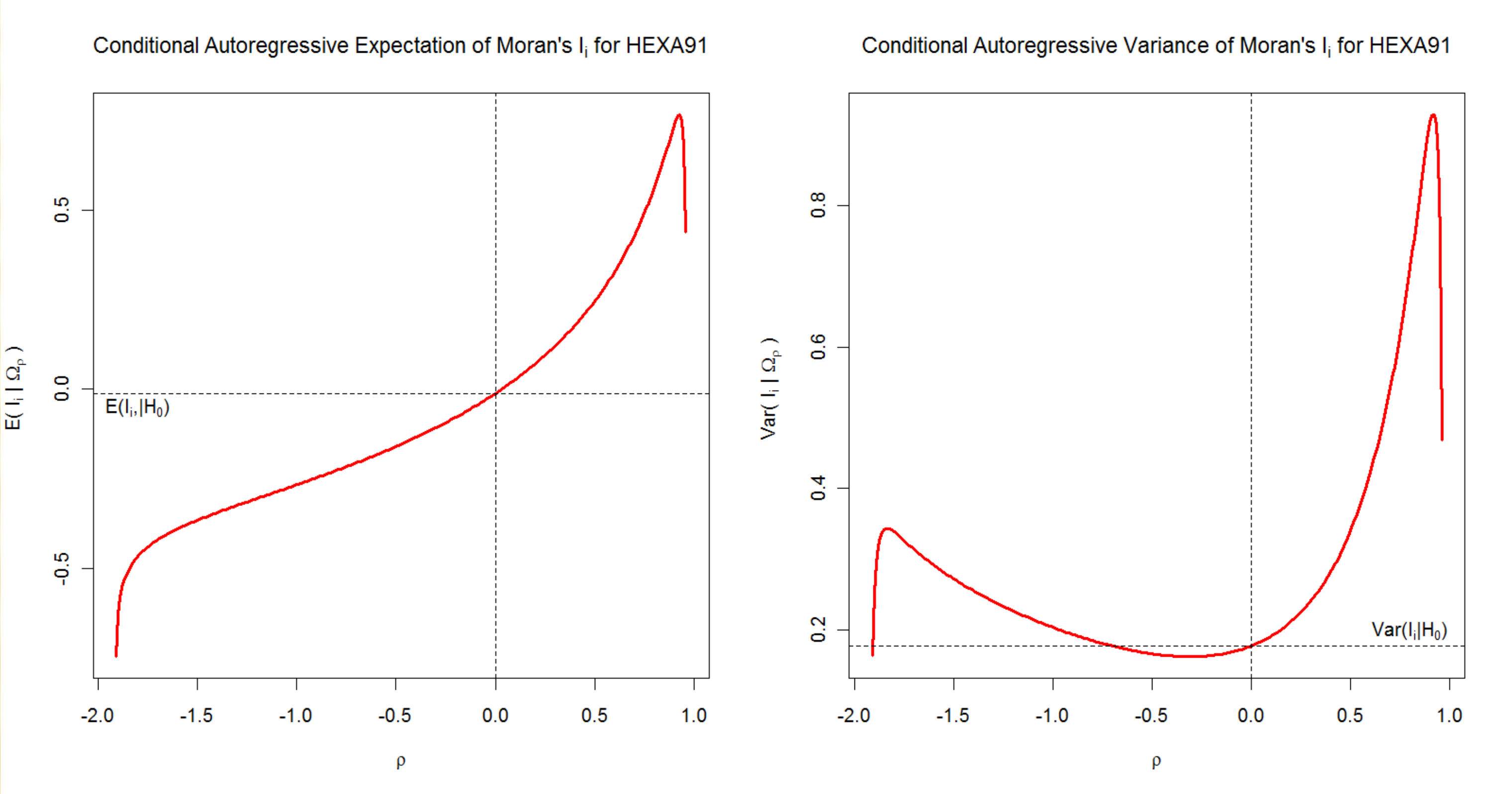
Example:

- Interior cell with 6 neighbors
- Tessellation size $n = 91$.
- SAR process with $\rho = 0.8$.

- The probability mass shifts into the direction corresponding to the sign of the global autocorrelation level ρ.
- The feasible range and kurtosis increase with increasing tessellation size n.

Review: Conditional Univariate Moments of Local Moran's I_i

- Local Moran's I_i conditional expectation and variance depend on the autocorrelation level ρ of the underlying global process.



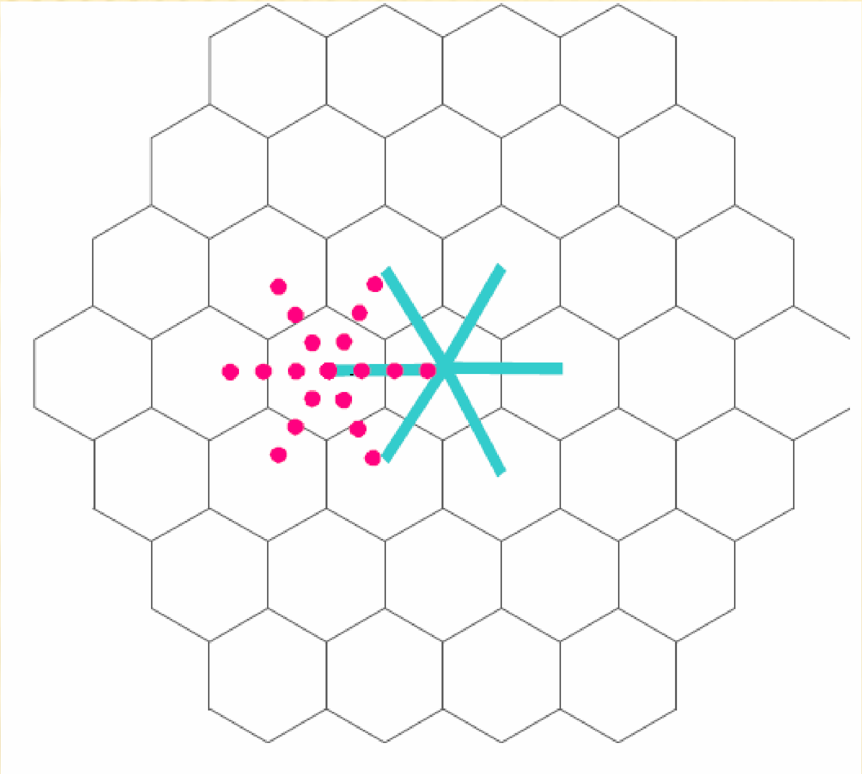
- Notice the boundary effects at the extremes of the stationary domain of the autocorrelation levels ρ.

Review: Local Moran's I_i p –Value Distribution

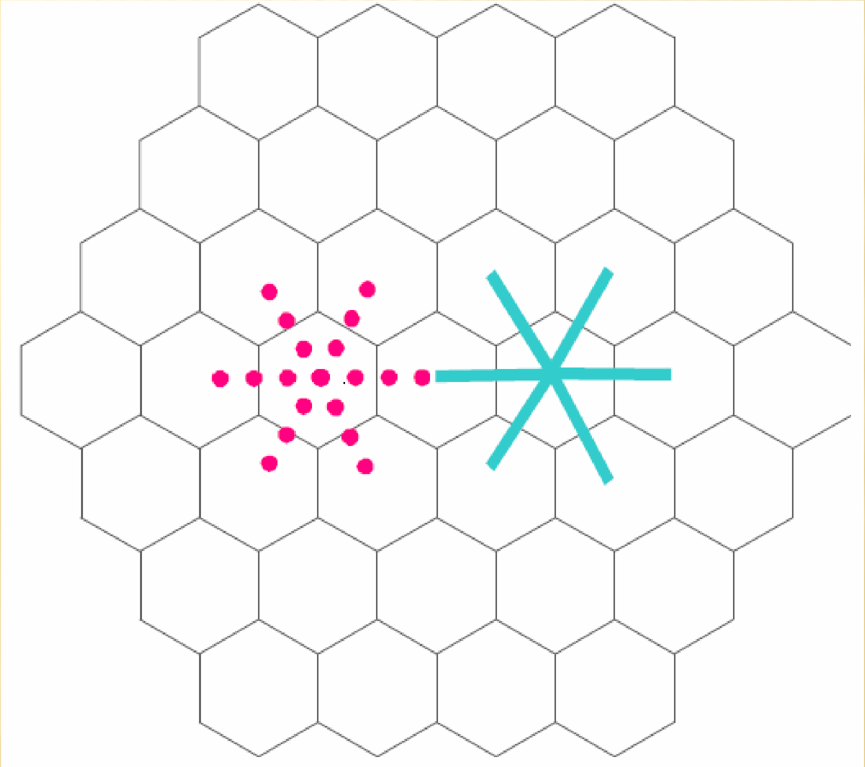
- If a model is ***properly specified*** then the p –values follow $p \sim U[0,1]$ with
$$p \equiv \Pr\left(I_i \leq I_i^{obs} | \mathbf{X}, \mathbf{V}, \boldsymbol{\Omega}\right)$$
(see Murdoch et al. *The American Statistician*, 2008, pp 242-245; Rice, 2007, p 63; or Tiefelsdorf, 2000, p 136)
- Therefore, it is likely to see extreme values of local Moran's I_is even though H_0 holds true.
- The p-values of a ***mis-specified*** model are triangular distributed.
- This concept can be applied to an ensemble of local Moran's I_is p –values:
  - proper model specification implies uniform distribution
  - the pooled p –values are correlated.

Induced Correlation: Topological Relationships

- The common information at cells, which are shared by the local structures of I_i and I_j, induces correlation among local Moran's I_is irrespectively of an underlying spatial process.



Shared cells at lag 1



Shared cell at lag 2

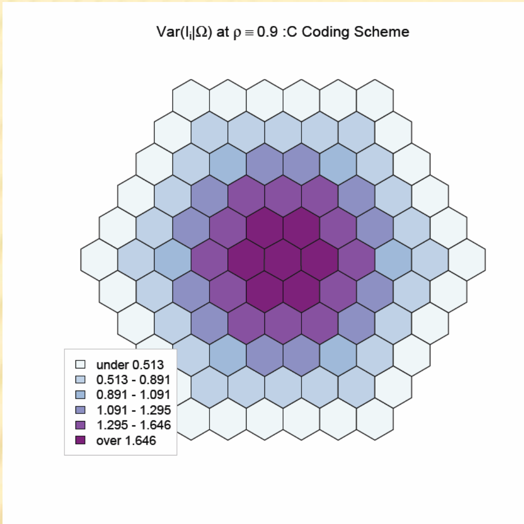
Induced Correlation: Global Processes

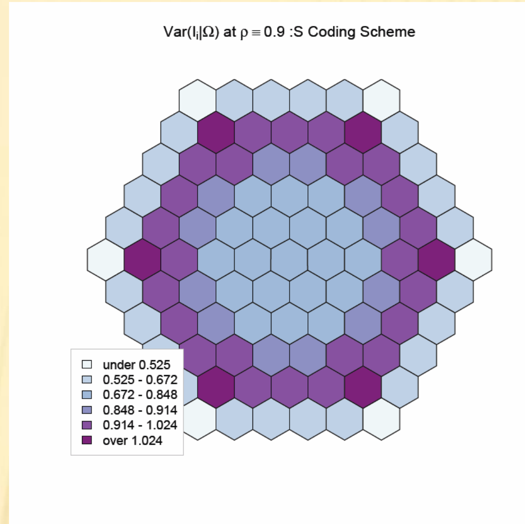
- Premise: A spatial process governs the diffusion or spill-over mechanisms of a stochastic inputs within a tessellation.
  $\Longrightarrow$ it will have an impact on the correlation among local I_is.
- Process strength: larger absolute autocorrelation levels lead to stronger signals.
  $\Longrightarrow$ Stronger signals will tie more distant I_is together.
- Process specification: long range SAR process *versus* short range MA process.
- Coding Schemes: in dependence of a cell's connectivity a signal at that cell is either enhanced or dampened.
- Edge effects (closed system perspective):
  [a] on average edge cells have lower connectivity
  $\Longrightarrow$ an edge signal is either enhanced or dampened depending on the coding scheme
  [b] spill-over signals from outside the tessellation cannot diffuse into the closed system.

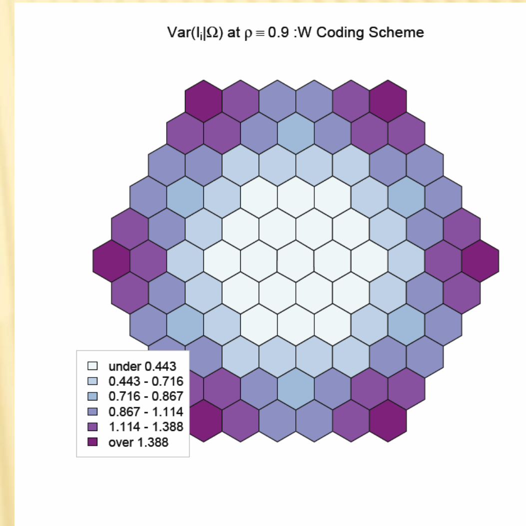
Review: Local Conditional Variances by Coding Scheme

***SAR*:**

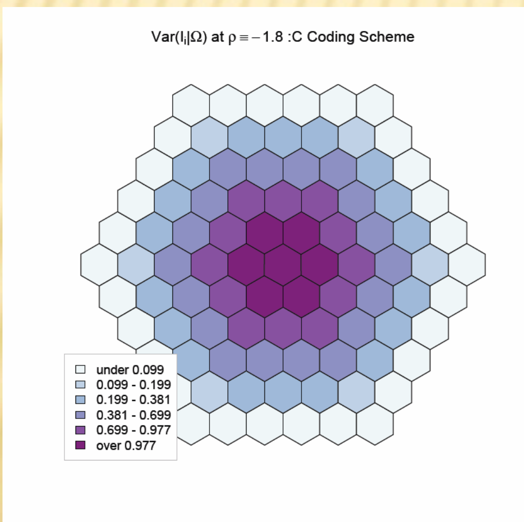
$\rho = 0.9$

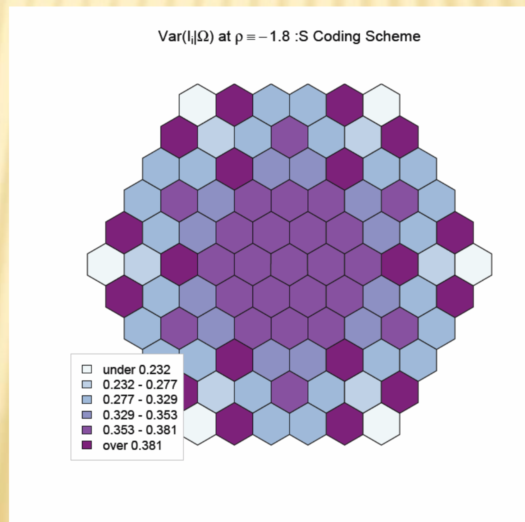


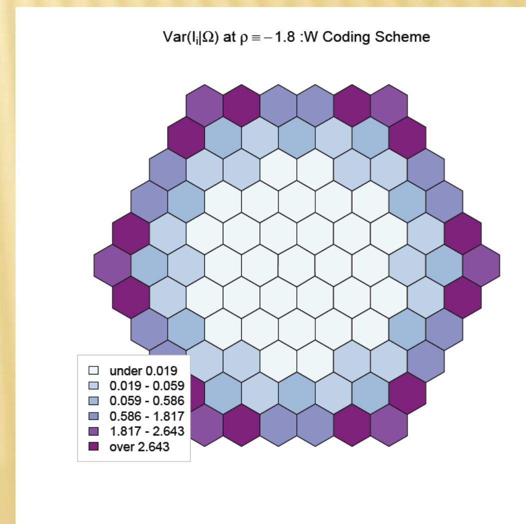




$\rho = -1.8$







$Var(I_i|\mathbf{\Omega})$ is most stable in the S-coding scheme

Key Equations to Evaluate $Cov(I_i, I_j|\boldsymbol{\Omega})$

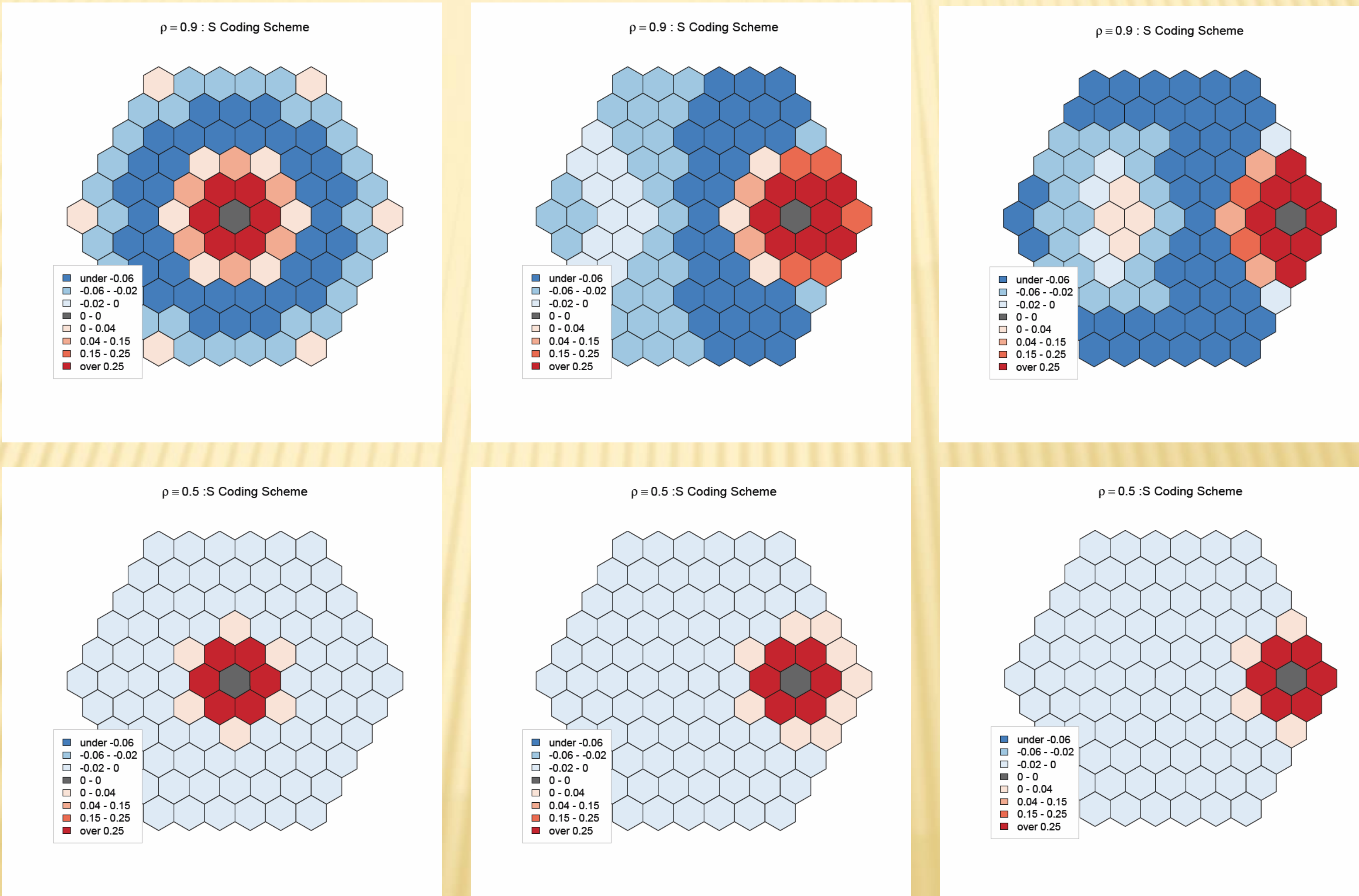
- Numerical evaluation of two well-behaved infinite integrands using Sawa's approach to get the moments:
  - The expectations $E(I_i|\boldsymbol{\Omega})$ and $E(I_i - I_j|\boldsymbol{\Omega})$.
  - The second moments $E\left({I_i}^2|\boldsymbol{\Omega}\right)$ and $E\left(\left[I_i - I_j\right]^2|\boldsymbol{\Omega}\right)$
  - Both integrals depend on elaborated matrix expressions, which are not easily vectorized.
- Calculation of variances:
  - $Var(I_i|\boldsymbol{\Omega}) = E\left({I_i}^2|\boldsymbol{\Omega}\right) - E(I_i|\boldsymbol{\Omega})^2$
  - $Var\left(\left[I_i - I_j\right]^2|\boldsymbol{\Omega}\right) = E\left(\left[I_i - I_j\right]^2|\boldsymbol{\Omega}\right) - E(I_i - I_j|\boldsymbol{\Omega})^2$
- Calculation of the co-variance:
  - $Cov(I_i, I_j|\boldsymbol{\Omega}) = \frac{1}{2} \cdot \left[Var(I_i|\boldsymbol{\Omega}) + Var(I_j|\boldsymbol{\Omega}) - Var\left(\left[I_i - I_j\right]^2|\boldsymbol{\Omega}\right)\right]$
- In total $n \cdot (n+1)$ integrals need to be evaluated for all pairwise cell combinations.

Results I: Correlation Pattern by Autocorrelation Level ρ

SAR:

$\rho = 0.9$

$\rho = 0.5$



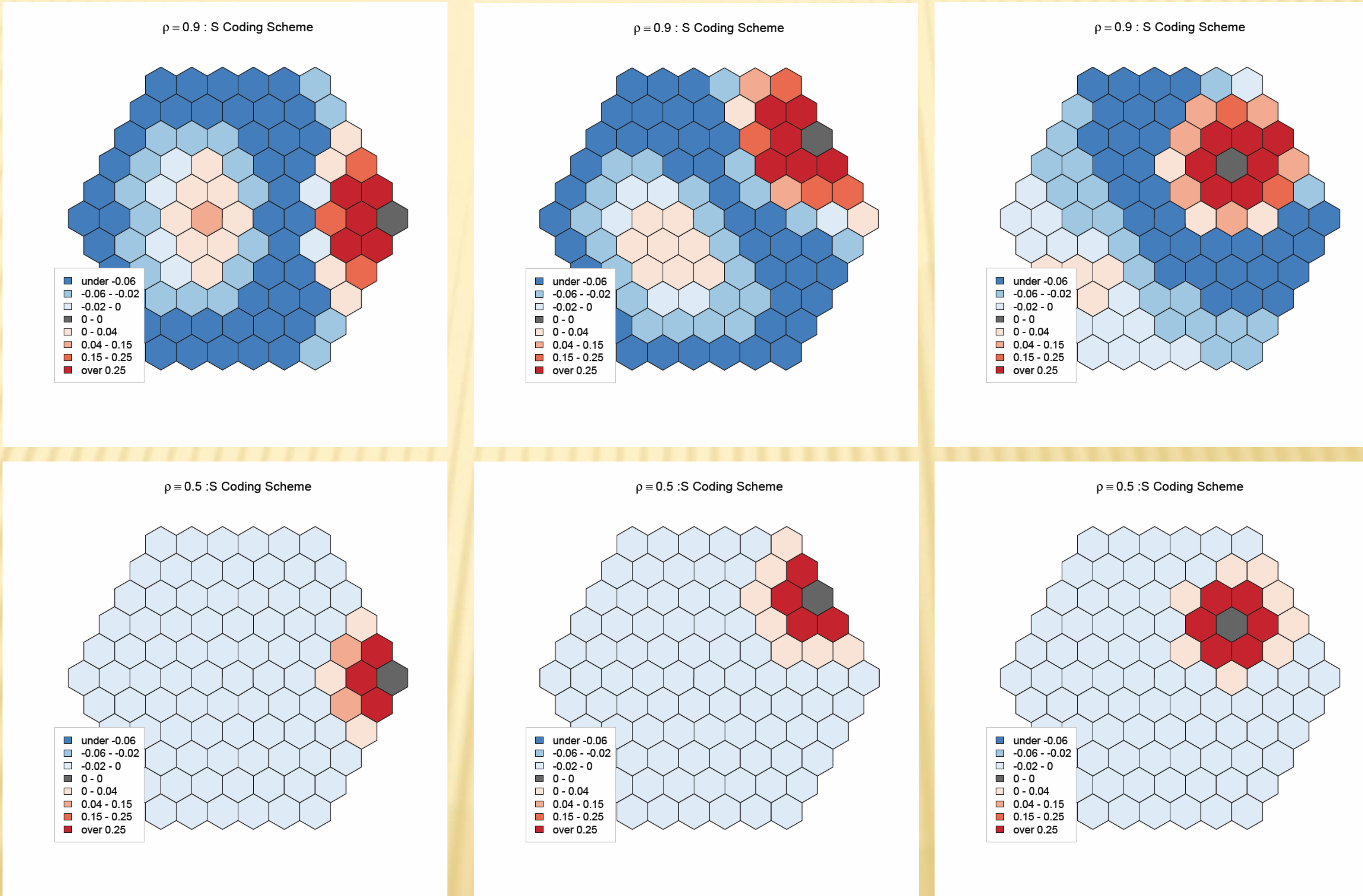
Correlation for selected reference cells in the *S*-coding scheme

Results II: Correlation Pattern by Autocorrelation Level ρ

SAR:

$\rho = 0.9$

$\rho = 0.5$



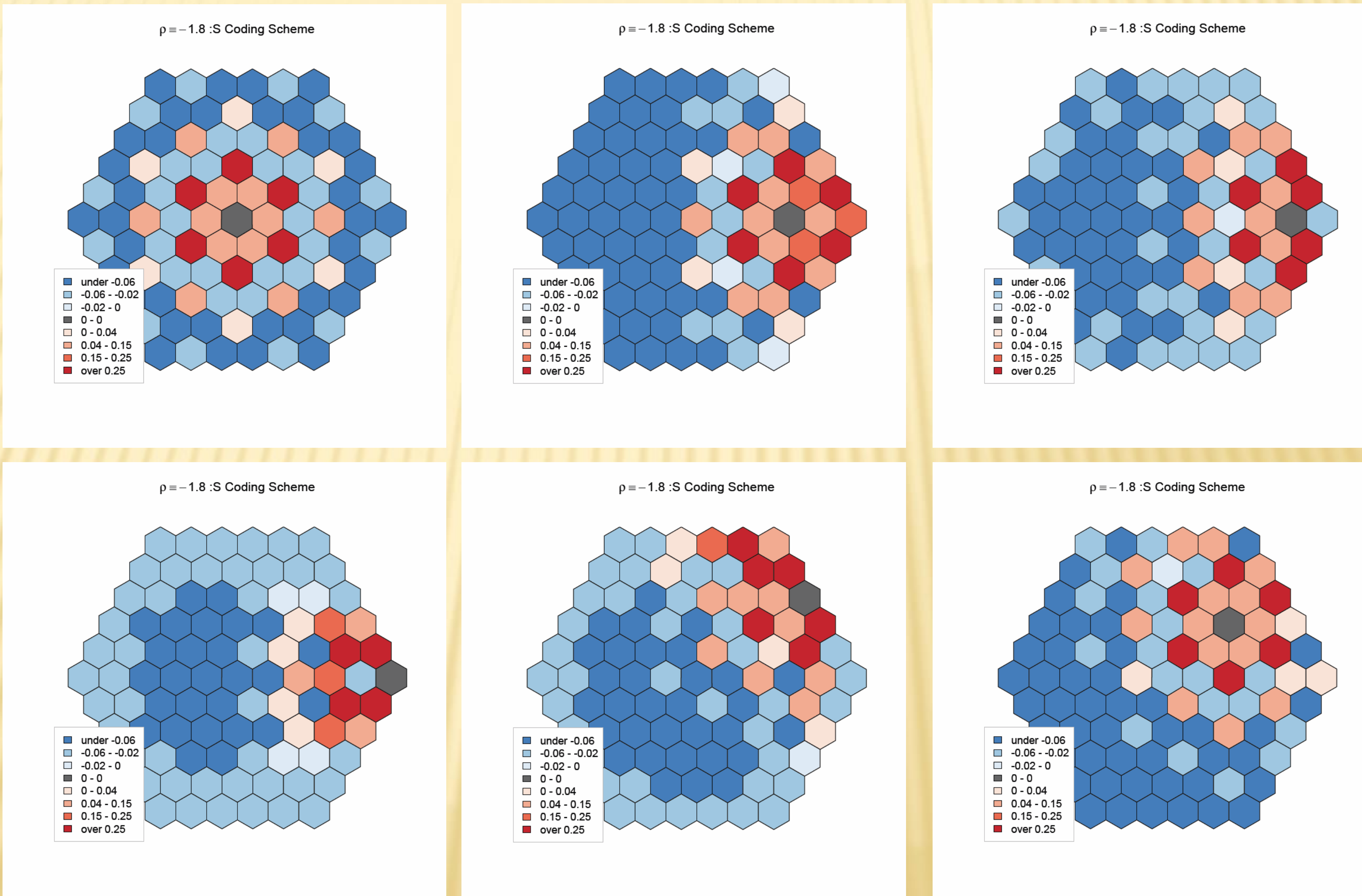
Correlation for selected reference cells in the S-coding scheme

Results III: Correlation Pattern by Autocorrelation Level ρ

SAR:

$\rho = -1.8$

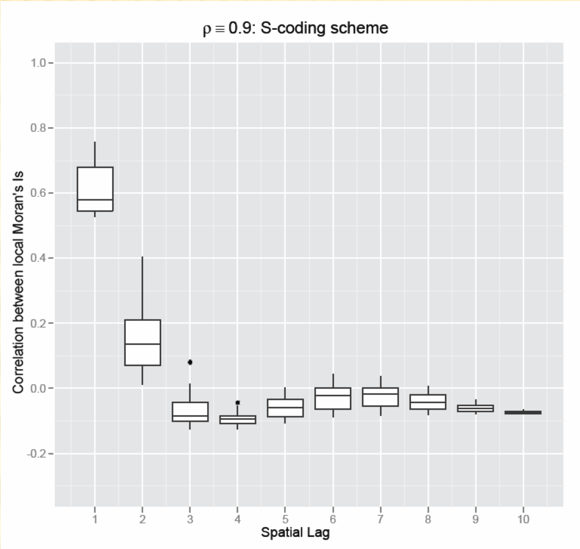
$\rho = -1.8$



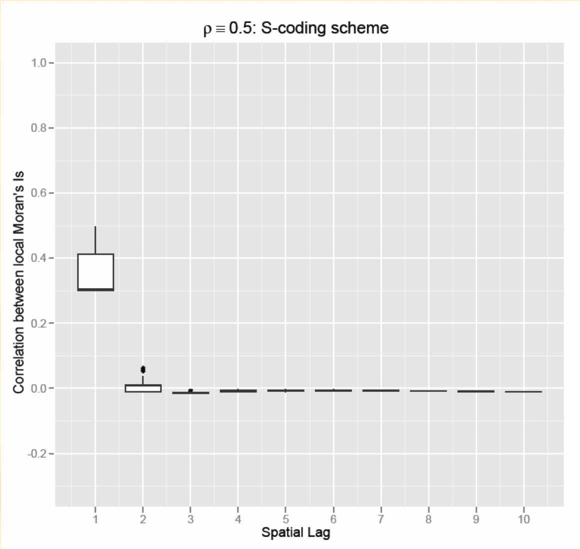
Correlation for selected reference cells in the S-coding scheme

Results IV: Overall Strength of Correlation Pattern

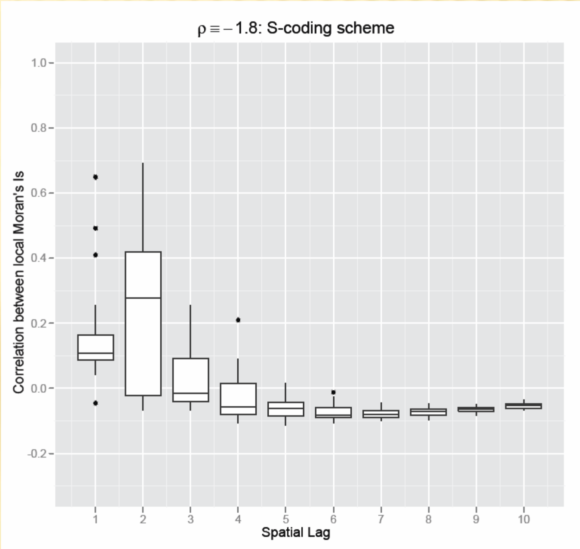
SAR:



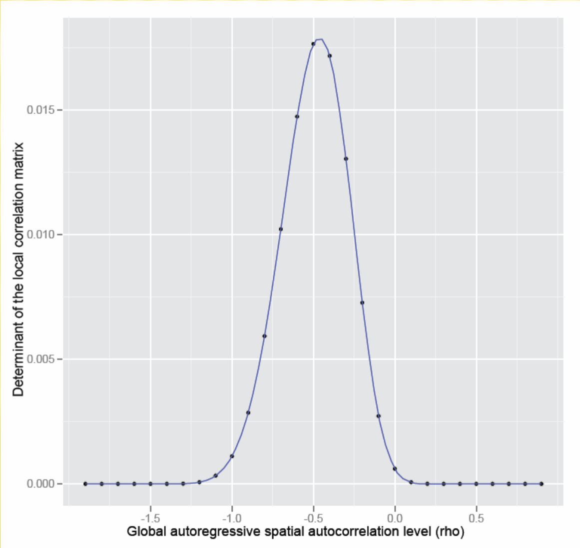
$\rho = 0.9$



$\rho = 0.5$



$\rho = -1.8$

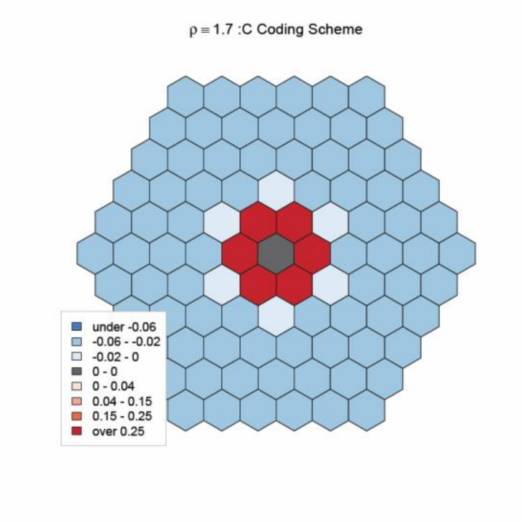


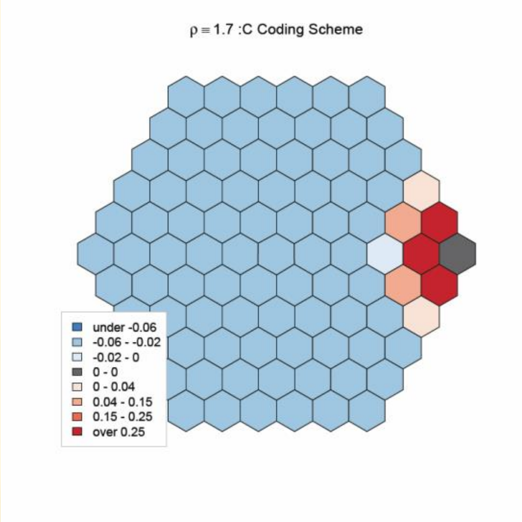
The highest degree of independence among local Moran's I_is is observed around $\rho = -0.5$ rather than at an spatial independence level $\rho = 0.0$.

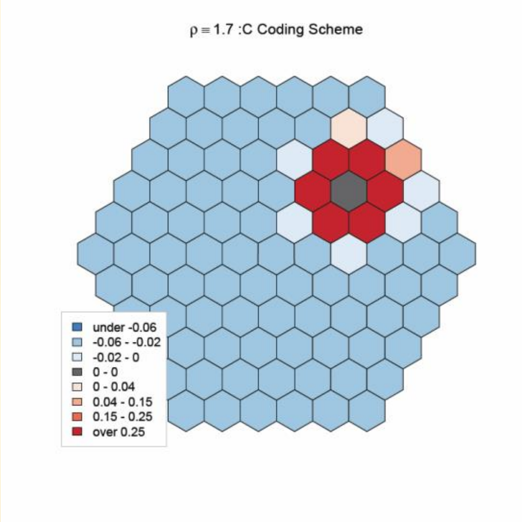
This holds for all coding schemes and sizes of spatial tessellations.

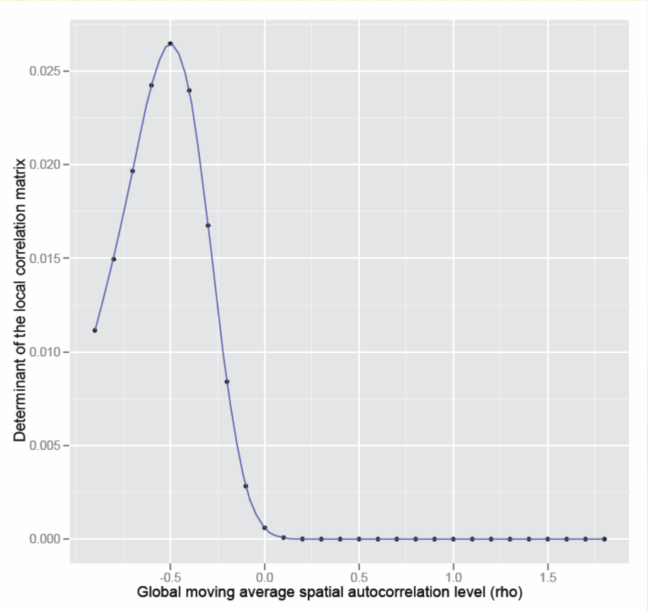
Results V: Correlation Patterns for Moving Average Processes

MA:







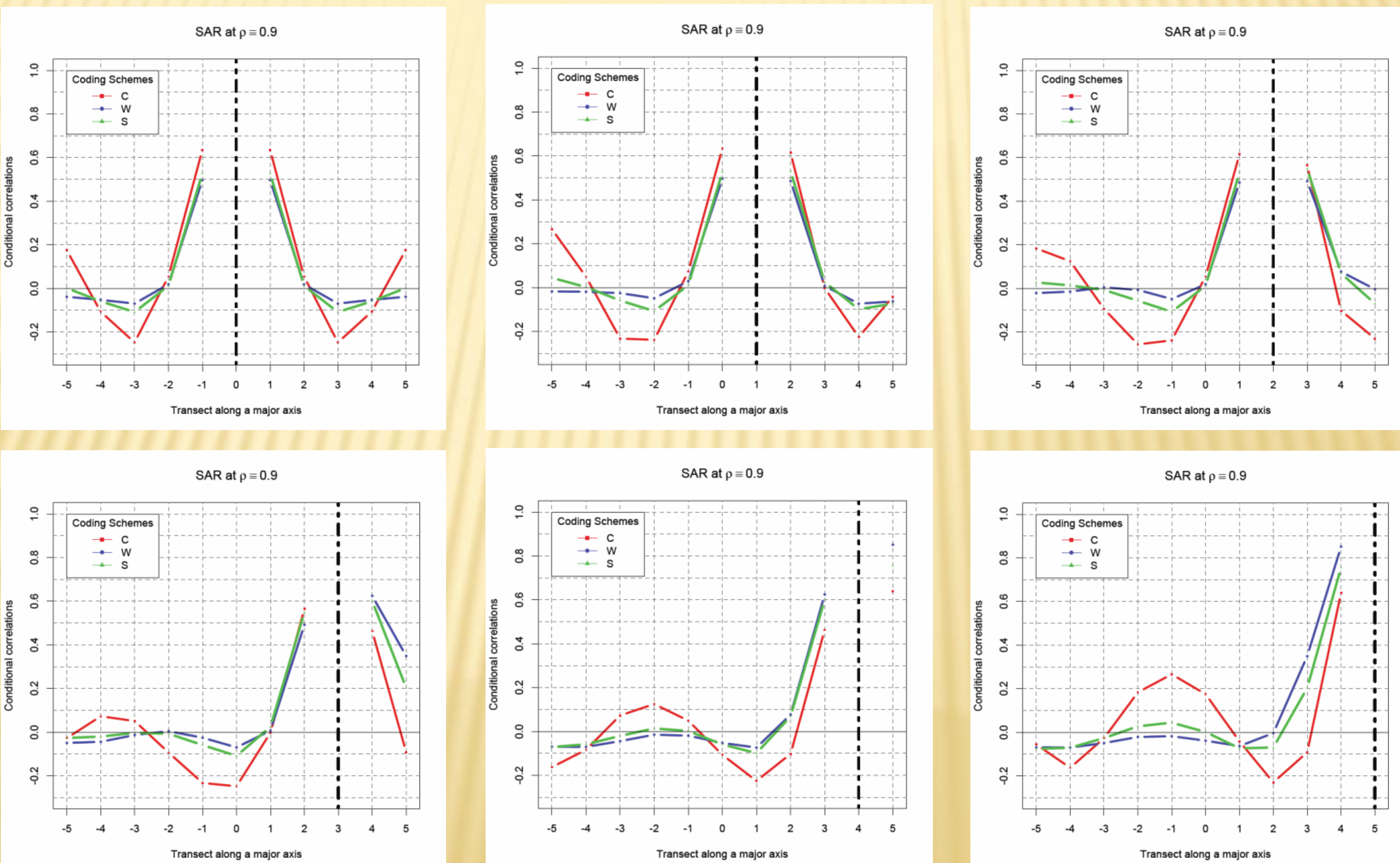


- As ***theoretically expected*** moving average spatial processes have a maximal topological range up to spatial lag 2.
- The degree of correlation among local Moran's I_is depends again on the spatial autocorrelation level.

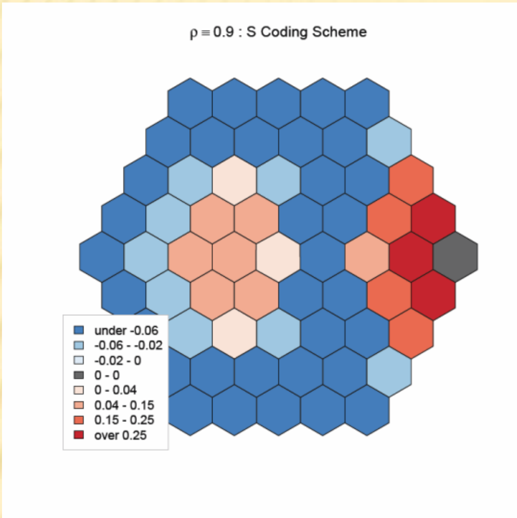
RESULTS VI: IMPACT OF THE CODING SCHEMES ON THE CORRELATION

Correlation levels of cell pairs along a major axis transect (SAR with $\rho = 0.9$).

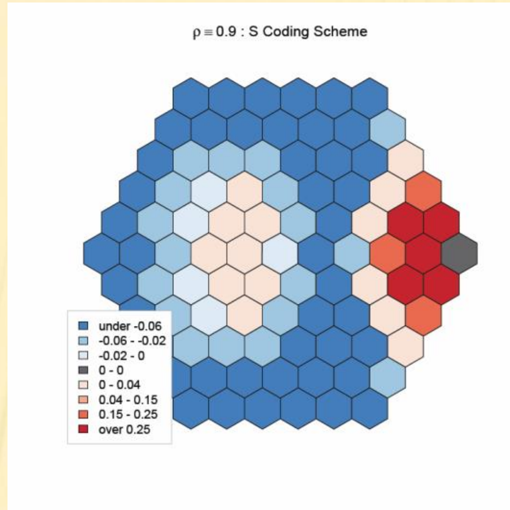
$\Longrightarrow$ Correlation patterns of *S*-coding scheme is in-between the other coding schemes.



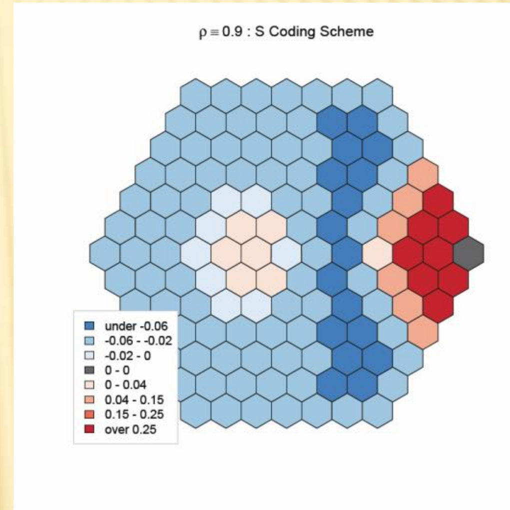
RESULTS VII: CORRELATION PATTERNS FOR EXPANDING SPATIAL DOMAINS



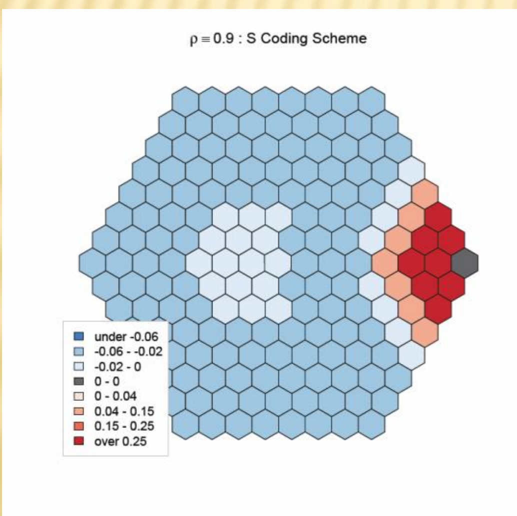
$n = 61$



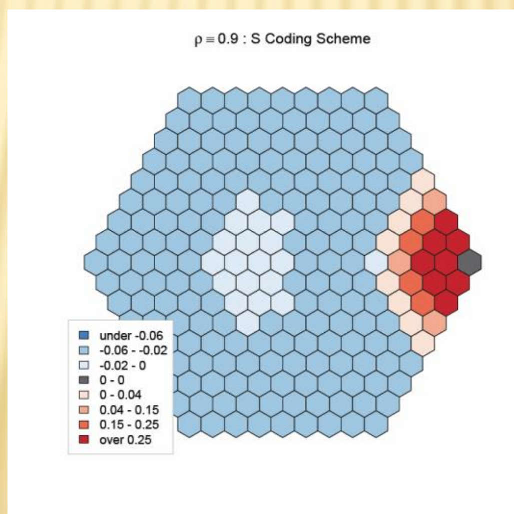
$n = 91$



$n = 127$



$n = 169$



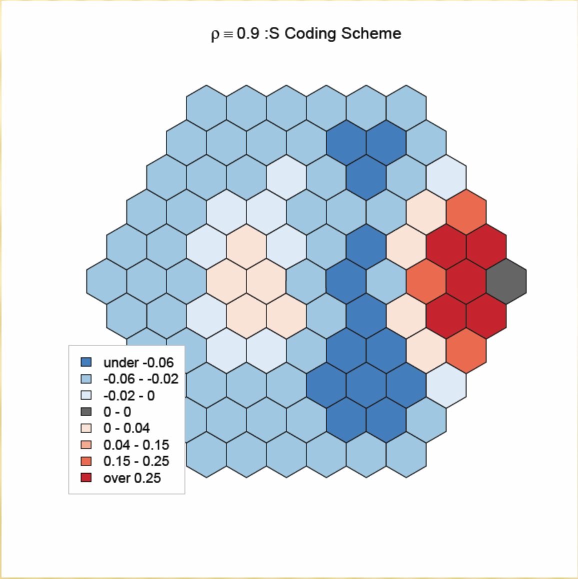
$n = 217$

- The general “echo” pattern at half the diameter due to edge reflections persists in an expanding spatial domain.
- The long-range intensity diminishes because the process exhausts its “effective” reach.

Results VIII: Accuracy of Analytical Versus Simulation Results

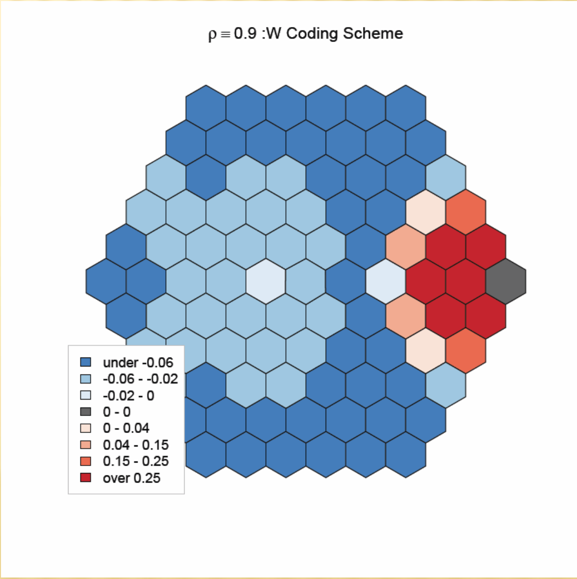
Design of the Simulation Experiment:

- Generate 10,000 random white noise map patterns $\boldsymbol{\varepsilon} \sim N(0, \sigma)$ with $\boldsymbol{\varepsilon} \perp \mathbf{1}$ that satisfy the underlying assumptions of spatial independence $\mathrm{z}[I(\boldsymbol{\varepsilon})] \sim 0$
- Transform the set of white noise inputs $\boldsymbol{\varepsilon}$ into a set of autocorrelated map patterns that follow a particular spatial process specification.
- Calculated for each map pattern its associated set of local Moran's I_is
- Evaluate at each cell the univariate distribution and for all locations the joint distribution.



Simulated Correlation Pattern

versus



Analytical Correlation Pattern

Results IX: Computational Efficiency

- Computational efficiency of the time consuming rejection sampling simulation approach (10,000 out of ~80,000 proposals inputs) is compared against the analytical calculations.
- A fine granular implementation with a ***parallel Basic Linear Algebra System*** (BLAS) is compared against an explicit parallel evaluation in ***independent threads*** of the correlations among local Moran's I_is.
- Results for a hexagonal tessellation with 91 cells on a 4 cores Intel I7 processor:
  - The rejection simulation approach uses the 4 cores parallel BLAS
  - The sequential analytical approach uses the 4 cores parallel BLAS
  - The parallel analytical approach uses four thread and a one core BLAS
- Computational efficiency:

Rejection Simulation	Analytical Sequential	Analytical Parallel
393 sec.	222 sec.	87 sec.

Conclusions

- With a few exceptions the conditional distribution of the Moran statistic is overlooked in the literature.
- Several open issues in spatial statistics can be addressed with the knowledge of pairwise correlations among local Moran's I_is.
- The conditional correlation among local Moran's I_is behaves as theoretically expected.
  - Counter-intuitively however, the lowest degree of inter-correlations is not found at spatial independence.
- A properly specified simulation experiment is computationally more demanding and less accurate than its analytical counterpart.
- The parallel implementation of the algorithm in different R environments requires further improvements.
  - E.g., it takes over 25 hours to calculate all 129,286 pairwise covariances of local Moran's I_is for a model based on the 508 US State Economic Areas .
    However, many higher order spatial lag correlations are virtually zero and therefore can be ignored.