Comparing exact and approximate spatial auto-regression model solutions for spatial data analysis

Baris M. Kazar, Shashi Shekhar, David J. Lilja, Ranga R. Vatsavai and R. Kelley Pace

Magnus Udd
Spatial data mining
Spring 2007

Introduction

- The spatial auto-regression (SAR) model is a popular spatial data analysis technique
- Computationally quite expensive
- In this paper 2 solutions for estimating SAR model parameters for large spatial data analysis is presented
- Using *Taylor series expansion* and *Chebyshev polynomials*
- Compared with an exact solution for the same model
- Tested on satellite image data
- I will not cover equations and lemmas/proofs..

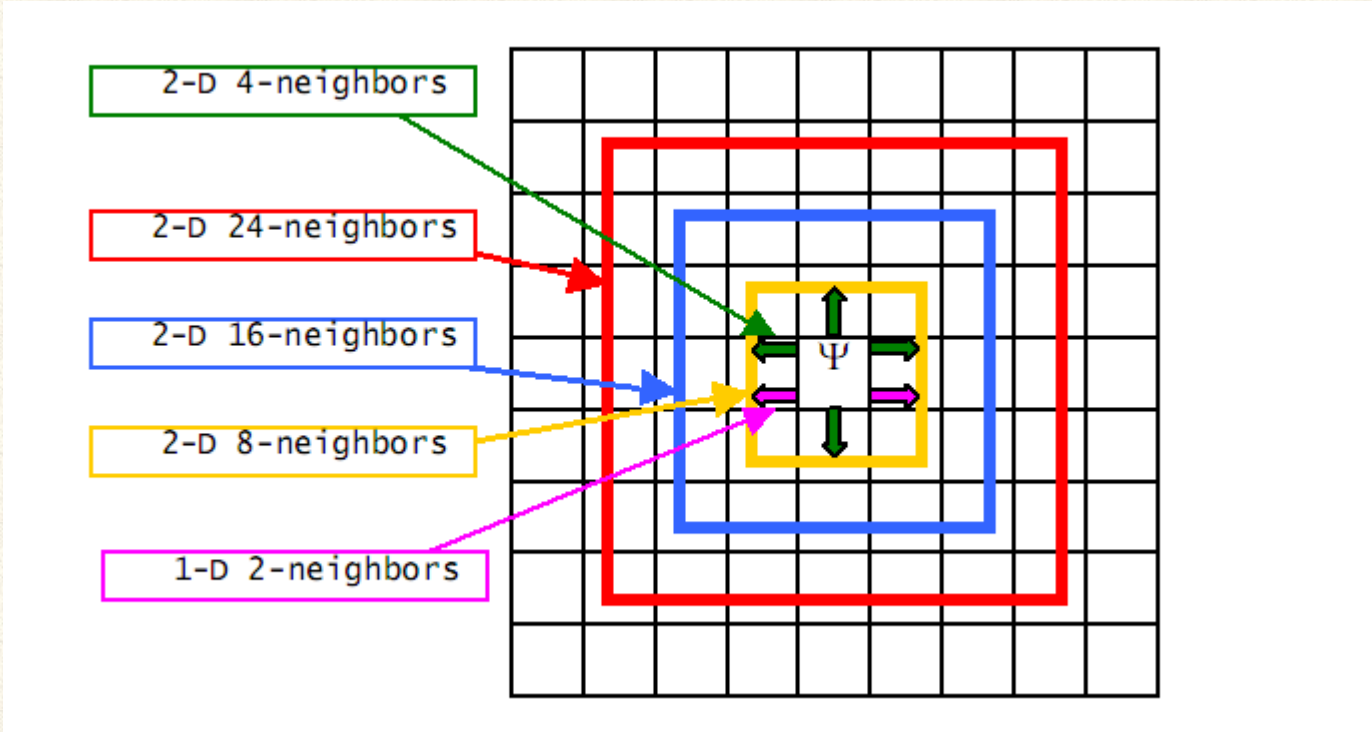
Problem statement

- There exists a solution for one dimensional geospatial datasets (Dense matrix approach)
- We **need** a multidimensional solution for spatial data
- An extension of linear regression model
- The equation consists of lin.reg. model+spatial autocorrelation term (pWy)
- Spatial autoregression parameter ρ is 0...1

Problem statement cont.

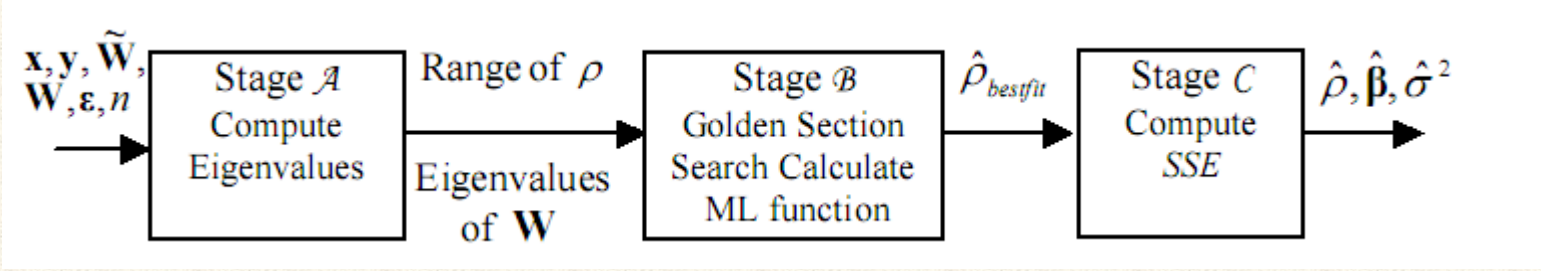
$$y = \rho Wy + x\beta + \varepsilon \, .$$

Problem statement cont.



Exact SAR model solution

- Parameters ρ and β in the aforementioned equation can be found using bayesian statistics or *maximum likelyhood*, the later is used in this study
- Three stages: computing eigenvalues most time-consuming (>99 %)



Exact SAR model solution cont.

Table 3. Measured serial response times of stages of the exact SAR model solution for problem sizes of 2500, 6400 and 10K. Problem size denotes the number of observation points

Problem size (n)	Machine	Serial Execution Time (sec) Spent on		
		Stage $\mathcal{A}$	Stage $\mathcal{B}$	Stage C
		Computing Eigenvalues	ML Function	Least Squares
2500	SGI Origin	78.10	0.41	0.06
	IBM SP	69.20	1.30	0.07
	IBM Regatta	46.90	0.58	0.06
6400	SGI Origin	1735.41	5.06	0.51
	IBM SP	1194.80	17.65	0.44
	IBM Regatta	798.70	6.19	0.42
10000	SGI Origin	6450.90	11.20	1.22
	IBM SP	6546.00	66.88	1.63
	IBM Regatta	3439.30	24.15	0.93

Two approximate SAR solutions

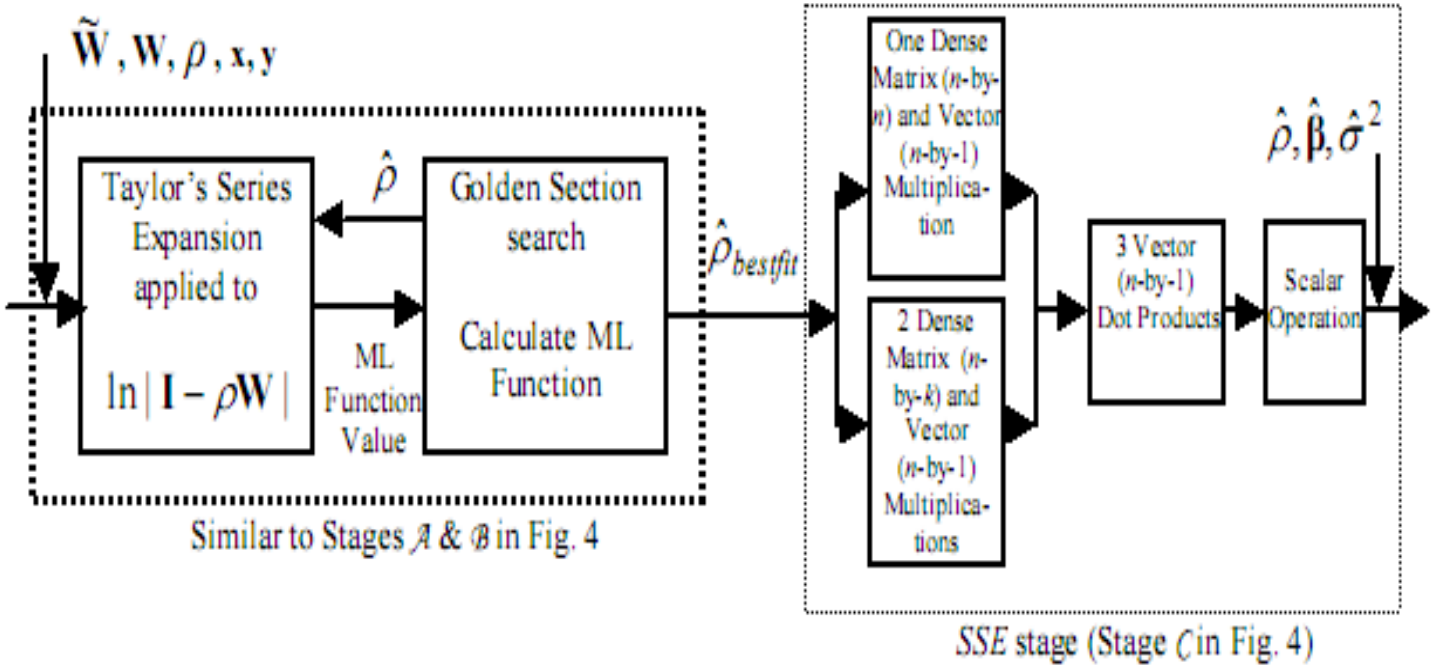


Fig. 6. The system diagram for the Taylor's Series expansion approximation for the SAR model solution. The inner structure of Taylor series expansion is similar to that of Chebyshev Polynomial except that there is one more vector sum operation, which is very cheap to compute

Two approximate SAR solutions

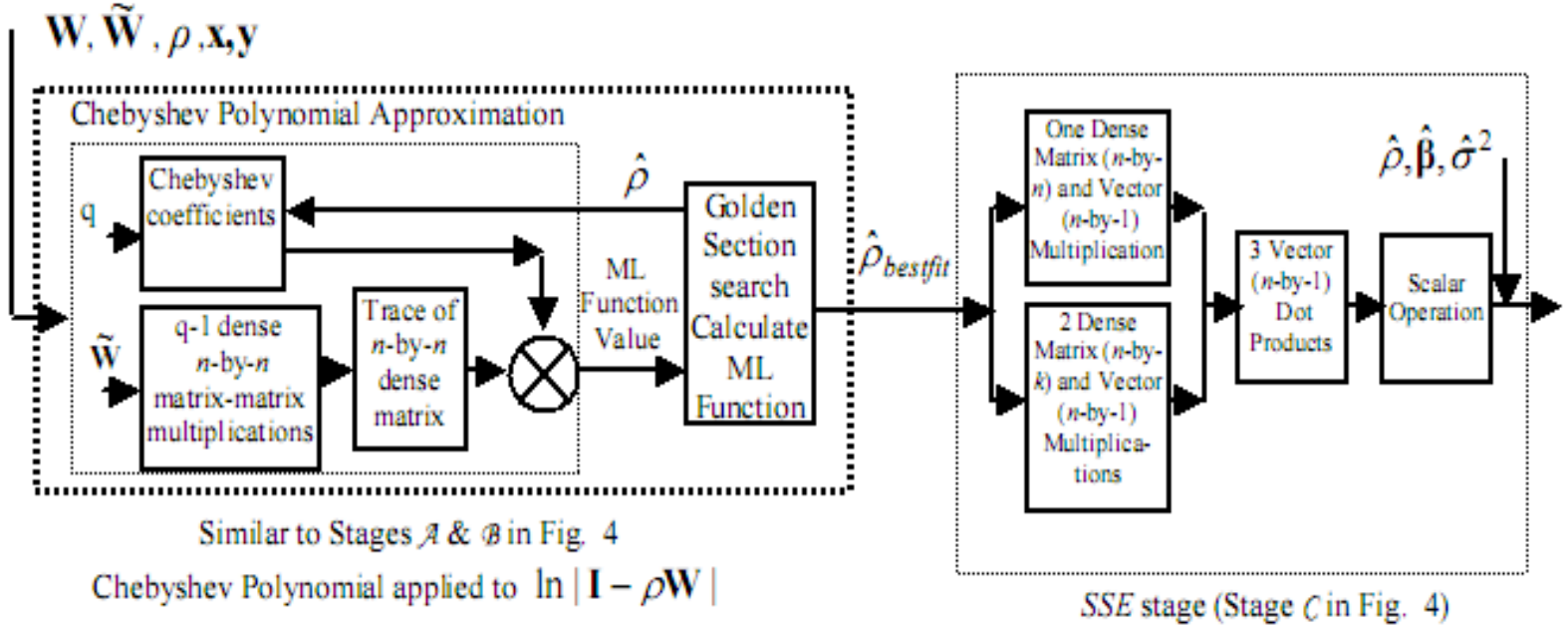


Fig. 7. System diagram of the approximate SAR model solution, where $\ln(\mathbf{I} - \rho\mathbf{W})$ is expressed as a Chebyshev polynomial. The term "q" is the degree of the Chebyshev Polynomial

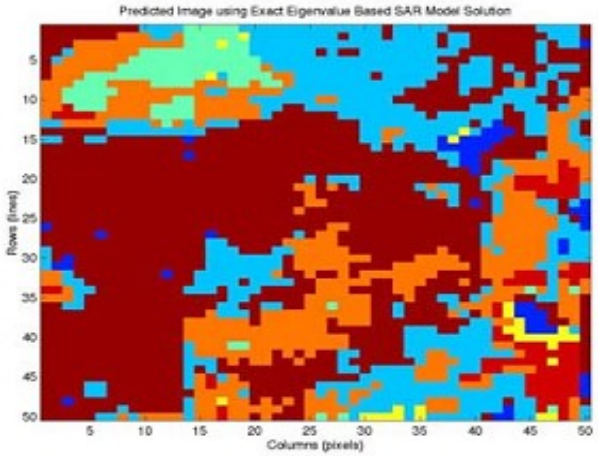
Experimental design

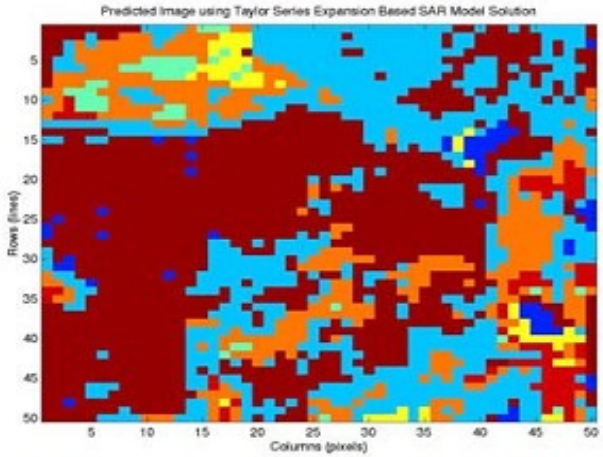
- Performed on Landsat satellite images from forest area in Minnesota, USA
- ScaLAPACK software/libraries was used
- Scalability (computational time), accuracy and memory usage of the 3 models was measured
- Both approximate solutions much faster&less memory intensive, and yet accurate predictors
- Some differences in the two:
- One performs better on low autocorrelation parameters, the other on high values

Experimental design cont.

- Results from image prediction using one exact and two approximate methods
- Some differences, why?
- when predicting thematic class labels the models performed quite similar
- 48.32, 48.4 and 50.4 % accuracy (exact, Chebyshev, Taylor series)

Experimental design cont.





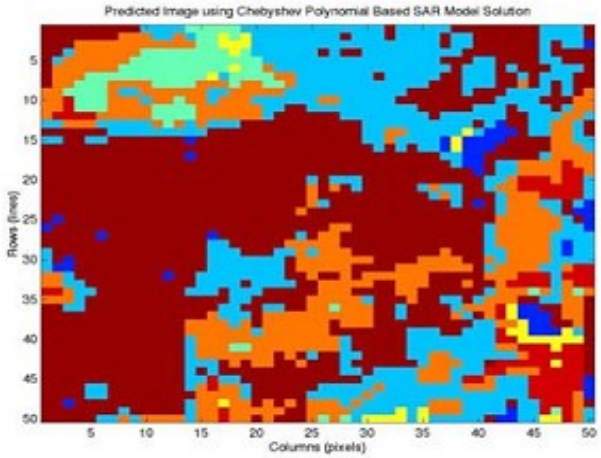


Fig. 9. The images (50x50) using exact and approximate solutions

Experimental design cont.

Table 7. The execution time in seconds and the memory usage in mega-bytes (MB)

Problem Size (n)	*Time (Seconds)*			*Memory (MB)*		
	Exact	*Taylor*	*Chebyshev*	*Exact*	*Taylor*	*Chebyshev*
50x50 (2500)	38	0.014	0.013	50	1.0	1.0
100x100 (10K)	5100	0.117	0.116	2400	4.5	4.5
1200x1800 (2.1M)	Intractable	17.432	17.431	$\sim 32 * 10^6$	415	415

Conlusion

- This study focused on scalability of the SAR model on large geospatial data sets
- Compared exact and approximate solutions
- Future challenges: comparing SAR model vs. other models, eg. Markov random fields