Parameter Estimation for the Spatial Autoregression Model: A Rigorous Approach [*]

Mete Celik [†] Baris M. Kazar [‡] Shashi Shekhar [†] Daniel Boley [†]

Abstract

The spatial autoregression (SAR) model is a knowledge discovery technique used for mining massive geo-spatial data in many application domains. Estimation of the parameters of the exact SAR model using Maximum Likelihood (ML) theory is computationally very expensive because of the need to compute the logarithm of the determinant (log-det) of a large matrix in the log-likelihood function. In this paper, we developed a faster, scalable and ***NO****vel p****R****ediction and estimation* ***T****ec****H****nique for the exact* ***S****pa****T****ial* ***A****uto* ***R****egression model solution (NORTHSTAR). In this heuristic, the SAR model parameters are first estimated using a computationally more efficient sum-of-squared errors (SSE) term of the log-likelihood function. Next, starting from an initial estimate very close to the optimal estimate, the computationally more expensive log-det term is embedded into the estimation process to save log-det computations. Experimental results show that the NORTHSTAR algorithm outperformed the previous exact SAR model solutions.*

1 Introduction

Explosive growth in the size of spatial databases has highlighted the need for spatial data analysis and spatial data mining techniques to mine the interesting but implicit spatial patterns within these large databases. Many classical data mining algorithms, such as linear regression, assume that the learning samples are *independently and identically distributed (i.i.d.)*. This assumption is violated in the case of spatial data due to spatial autocorrelation [1, 21]. In such cases classical linear regression yields a weak model with not only low prediction accuracy [22] but also residual error exhibiting spatial dependence. Modeling spatial dependencies improves overall classification and prediction accuracies.

Nasa Relevance: The spatial autoregression model (SAR) [4, 6, 21] is a generalization of the linear regression model to account for spatial autocorrelation. The model yields better classifica-

[*] This work was partially supported by the Army High Performance Computing Research Center (AHPCRC) under the auspices of the Department of the Army, Army Research Laboratory (ARL) under contract number DAAD19-01-2-0014, the NSF grant IIS-0208621, and the NSF grant IIS-0534286. This work received additional support from the University of Minnesota Digital Technology Center and the Minnesota Supercomputing Institute.

[†] Computer Science and Engineering Department, University of Minnesota {mcelik,shekhar,boley}@cs.umn.edu

[‡] Oracle Corporation {baris.kazar@oracle.com}

tion and prediction accuracy [3, 22] for many spatial datasets exhibiting strong spatial autocorrelation. NASA's Earth Observing System (EOS) generates one terabyte of data every day with satellites observing the earth's surface. Spatial data mining techniques, such as SAR, can provide a tool to analyze these datasets to extract information on land use, land cover, and parameters, such as temperature, pressure, precipitation, etc.

Challenge: However, it is computationally expensive to estimate the parameters of SAR. For example, it can take an hour of computation for a spatial dataset with 10,000 observation points on a single IBM Regatta processor using a 1.3GHz pSeries 690 Power4 architecture with 3.2 GB memory [7, 8]. This has limited the use of SAR to small problems, despite its promise to improve classification and prediction accuracy for larger spatial datasets.

Related Work: There are two families of SAR model solutions, one based on ML Theory [12, 14, 20, 13, 18, 15, 23, 19, 16, 17] and the second based on Bayesian Statistics [11, 2, 10, 22]. ML Theory-based SAR model solutions can be classified into exact and approximate solutions, based on how they compute the log-det and least-squares (SSE) term of the SAR solution procedure. This study covers only ML based exact SAR model solutions. Previous approaches first compute the log-det term of the SAR model to estimate the SAR parameters which is computationally complex and then compute SSE term of the SAR solution. So, the previous approaches are not scalable to the large problem sizes. In this paper we developed a faster, scalable and **NO**vel p**R**ediction and estimation **T**ec**H**nique for the exact **S**pa**T**ial **A**uto **R**egression model solution (NORTHSTAR). In the NORTHSTAR heuristic, SAR model parameters first estimated using much less computationally complex SSE term of the log-likelihood function. A second computationally more complex step is required only if the parameters obtained in the first step are not accurate enough; in this case, the log-det term is embedded into the estimation process.

Contributions: A faster, scalable and **NO**vel p**R**ediction and estimation **T**ec**H**nique for the exact **S**pa**T**ial **A**uto **R**egression model solution (NORTHSTAR) was developed. Second, we experimentally showed that the proposed algorithm outperforms the eigen-value approaches (EV) and straight log-det approach (SLD).

2 Problem Statement and Background on SAR Model

2.1 Problem Statement

Given a spatial framework S for the underlying spatial graph G, a collection of attribute functions $f_{\mathbf{x}_k}$ over S, a dependent function $f_{\mathbf{y}}$, a family $\mathbf{F}$ of learning model functions, and the neighborhood relationship R, build the SAR model and find its parameters by minimizing the concentrated log-likelihood (objective) function. Constraints are, geographic space S is a multi-dimensional Euclidean Space, the values of the explanatory variables $\mathbf{x}$ and the dependent function (observed variable) $\mathbf{y}$ may not be independent with respect to those of nearby spatial sites, i.e., spatial autocorrelation exists, the domain of explanatory and dependent variables are real numbers, SAR parameter ρ varies in the range $[0, 1)$, and the neighborhood matrix $\mathbf{W}$ exhibits sparsity.

2.2 Background on SAR Model

The SAR model [4, 1], also known in the literature as spatial lag model or mixed regressive model, is an extension of the linear regression model and is given in equation (1).

$$\mathbf{y} = \rho \mathbf{W} \mathbf{y} + \mathbf{x}\beta + \epsilon \tag{1}$$

Here ρ is the spatial autocorrelation parameter, $\mathbf{y}$ is an n-by-1 vector of observations on the dependent variable, $\mathbf{x}$ is an n-by-k matrix of observations on the explanatory variable, $\mathbf{W}$ is the n-by-n neighborhood matrix that accounts for the spatial relationships (dependencies) among the spatial data, β is a k-by-1 vector of regression coefficients, and ϵ is an n-by-1 vector of unobservable error. The *spatial autocorrelation* term $\rho\mathbf{W}\mathbf{y}$ is added to the linear regression model in order to model the strength of the spatial dependencies among the elements of the dependent variable, $\mathbf{y}$. One can use Moran's I index [5] in order to see whether there is significant spatial dependency in the given dataset.

The neighborhood matrices used by the SAR model are the neighborhood relationships on one-dimensional regular and irregular grid spaces with two neighbors and two-dimensional regular or irregular grid space with ”s” neighbors, where ”s” is four, eight, sixteen, twenty-four and so on neighbors [5, 9]. The rows of the neighborhood matrix $\mathbf{W}$ sum to 1, which means that $\mathbf{W}$ is row-standardized. A non-zero entry in the j^{th} column of the i^{th} row indicates that the j^{th} observation will be used to adjust the prediction of the i^{th} row where i is not equal to j.

3 Experimental Evaluation

Due to the limited space, the details of the NORTHSTAR algorithms are introduced in the appendix section. We compared the proposed algorithm, NORTHSTAR, with the exact EV and SLD approaches using a real dataset. The dataset is a spring Landsat 7 scene, taken May 31, 2000, satellite remote sensing data which is belong to Carlton County, Minnesota. The region is predominantly forested, composed mostly of upland hardwoods, and low-land conifers. This scene was clipped to the Carlton county boundaries, which resulted in an image of size 1343 lines by 2043 pixels and 6-bands. Out of this we took a subset image of 1200 by 1800 to eliminate boundary zero-valued pixels. This translates to a $\mathbf{W}$ matrix of size 2.1 million x 2.1 million (2.1M x 2.1M) points. The observed variable $\mathbf{x}$ is a matrix of size 2.1M by 6. We chose nine thematic classes for the classification.

Table 1. The execution time and the memory usage

Problem Size(n)	Time		
	Exact EV	Exact SLD	NORTHSTAR
400x400 (160,000)	Intractable	32 minutes	24 minutes
1000x1000 (1,000,000)	Intractable	72 hours	45 hours

Problem Size(n)	Memory (MB)		
	Exact EV	Exact SLD	NORTHSTAR
50x50 (2,500)	50	1	1
100x100 (10,000)	2400	4.5	4.5
400x400 (160,000)	$\sim 6.14 * 10^5$	70	70
1000x1000 (1,000,000)	$\sim 8 * 10^6$	450	450

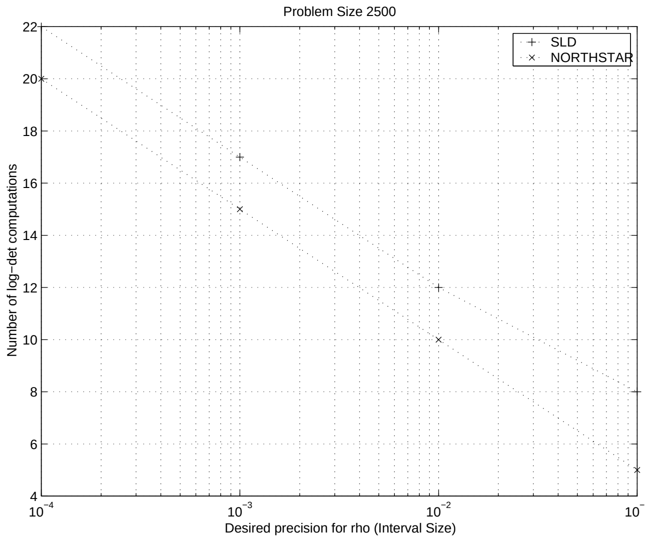

Figure 1. The savings from log-det computation rho (ρ) $4.7293 * 10^{-1}$.

We, first, tested the scalability (computation time) and memory usage of the algorithms on an IBM Regatta 1.3GHz Power4 processor. The computation (CPU) time of the NORTHSTAR outperforms the other algorithms (Table 1). The memory usage is very low due to the sparse representation of the neighborhood matrix $\mathbf{W}$ as a sparse matrix. However, this is not possible for the EV approach since it has to use the dense representation of the matrix. As can be seen, NORTHSTAR is the most scalable algorithm among the exact SAR model solutions.

We compared NORTHSTAR and SLD algorithms to see the savings from the log-det computations for various precision of ρ parameter (e.g. ρ=$4.7293 * 10^{-1}$.) Since step (ii) of NORTHSTAR starts with an initial ρ search space of $(0, \rho_{init-est})$ where $\rho_{init-est}$ is the estimate of ρ from the first step of NORTHSTAR, the algorithm saves from the log-det computations (Figure 1). Experiments showed that when the precision decreased savings from log-det computation increases.

4 Accomplishments and Future Work

We developed a ML based exact SAR model solution. In this algorithm, the SAR model parameter is first estimated using a computationally more efficient *SSE* term of the log-likelihood function. Next, starting from an initial estimate very close to the optimal estimate, the computationally more expensive log-det term is embedded into the estimation process to save log-det computations. Experiments show that the NORTHSTAR algorithm outperformed the previous exact SAR model solutions.

In the future, we plan to investigate to put bounds to the ρ parameter, to use optimization algorithms other than GSS to reach the large problem sizes, to use different neighborhood structures to evaluate the proposed algorithm, and to test the behavior of the NORTHSTAR for different problem sizes and varying ρ parameters.

References

[1] L. Anselin. *Spatial Econometrics: Methods and Models*. Kluwer Academic Publishers, Dorddrecht, 1988.

[2] R. Barry and R. Pace. Monte carlo estimates of the log-determinant of large sparse matrices. *Linear Algebra and its Applications*, 289:41–54, 1999.

[3] S. Chawla, S. Shekhar, W. Wu, and U. Ozesmi. Modeling spatial dependencies for mining geospatial data. *1st SIAM International Conference on Data Mining*, 2001.

[4] N. Cressie. *Statistics for Spatial Data*. Wiley, New York, 1993.

[5] D. Griffith. *Advanced Spatial Statistics*. Kluwer Academic Publishers, 1998.

[6] B. Kazar, S. Shekhar, and D. Lilja. Parallel formulation of spatial auto-regression. *AHPCRC Technical Report No: 2003-125*, 2003.

[7] B. Kazar, S. Shekhar, D. Lilja, and D. Boley. A parallel formulation of the spatial auto-regression model for mining large geo-spatial datasets. *SIAM International Conf. on Data Mining Workshop on High Performance and Distributed Mining (HPDM2004)*, April 2004.

[8] B. Kazar, S. Shekhar, D. Lilja, D. Boley, D. Shires, J. Rogers, and M. Celik. A parallel forumulation of the spatial autoregression model. *II International Conference and Exhibition on Geographic Information*, 2005.

[9] B. Kazar, S. Shekhar, D. Lilja, R. Vatsavai, and R. Pace. Comparing exact and approximate spatial auto-regression model solutions for spatial data analysis. *Third International Conference on Geographic Information Science (GIScience2004)*, October 2004.

[10] J. LeSage. Solving large-scale spatial autoregressive models. *Second Workshop on Mining Scientific Datasets*, 2000.

[11] J. LeSage and R. Pace. Using matrix exponentials to explore spatial structure in regression relationships (bayesian mess). *http://www.spatial-statistics.com*, 2000.

[12] B. Li. Implementing spatial statistics on parallel computers. *Practical Handbook of Spatial Statistics, CRC Press*, pages 107–148, 1996.

[13] R. Martin. Approximations to the determinant term in gaussian maximum likelihood estimation of some spatial models. *Statistical Theory Models*, 22(1):189–205, 1993.

[14] R. Pace and J. LeSage. Closed-form maximum likelihood estimates for spatial problems (mess). *http://www.spatial-statistics.com*, 2000.

[15] R. Pace and J. LeSage. Semiparametric maximum likelihood estimates of spatial dependence. *Geographical Analysis*, 34(1):76–90, 2002.

[16] R. Pace and J. LeSage. Simple bounds for difficult spatial likelihood problems. *http://www.spatial-statistics.com*, 2003.

[17] R. Pace and J. LeSage. Spatial auto-regressive local estimation (sale). *Spatial Statistics and Spatial Econometrics, ed. by Art Getis*, 2003.

[18] R. Pace and J. LeSage. Chebyshev approximation of log-determinant of spatial weight matrices. *Computational Statistics and Data Analysis*, 2004.

[19] R. Pace and J. LeSage. Closed-form maximum likelihood estimates of spatial auto-regressive models: the double bounded likelihood estimator (dble). *Geographical Analysis*, Forthcoming.

[20] R. Pace and D. Zou. Closed-form maximum likelihood estimates of nearest neighbor spatial dependence. *Geographical Analysis*, 32(2), 2000.

[21] S. Shekhar and S. Chawla. *Spatial Databases: A Tour*. Prentice Hall, 2003.

[22] S. Shekhar, P. Schrater, R. Raju, and W. Wu. Spatial contextual classification and prediction models for mining geospatial data. *IEEE Transactions on Multimedia*, 4(2):174–188, 2002.

[23] O. Smirnov and L. Anselin. Fast maximum likelihood estimation of very large spatial autoregressive models: A characteristic polynomial approach. *Computational Statistics and Data Analysis*, 35(3):301–319, 2001.

5 Appendix

5.1 NORTHSTAR Algorithm

The log-likelihood function of ML based SAR model solution basically contains two terms, such as, log-det term and SSE term (Equation 2).

$$\min_{|\rho|<1} \frac{-2}{n} \underbrace{\ln|\mathbf{I}-\rho\mathbf{W}|}_{log-det} + \underbrace{\ln((\mathbf{I}-\rho\mathbf{W})\mathbf{y})^T(\mathbf{I}-\mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\,\mathbf{x}^T)^T(\mathbf{I}-\mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\,\mathbf{x}^T)((\mathbf{I}-\rho\mathbf{W})\mathbf{y})}_{SSE} \tag{2}$$

In contrast to the previous studies, in the NORTHSTAR algorithm, the SAR model parameters are first estimated using a computationally more efficient *SSE* term of the corresponding log-likelihood function of SAR model. Next, starting from an initial estimate very close to the optimal estimate, the computationally more expensive log-det term is embedded into the estimation process to save log-det computations. NORTHSTAR algorithm limits range of the SAR parameter ρ with the initial estimate of the ρ parameter using SSE term. Since the range of the ρ limited, the number of the log-det computations also decreases.

The pseudocode of NORTHSTAR algorithm is given in Figure 2, where GSS stands for golden section search. Instead of the GSS, which is not sensitive to the derivative of the optimized function, a derivative-sensitive search algorithm can be used for faster convergence to the optimal SAR parameter ρ but we need to compute inverse of the large matrix $(\mathbf{I} - \rho\mathbf{W})$ which is as costly as log-det computation in Step (ii). Since the likelihood function is uni-modular, the GSS always

```
Input:   ($\rho_{start}$,$tol$, $\mathbf{W}$)
Output:  ($\rho_{opt}, \beta_{\mathbf{opt}}, \sigma^2_{opt}$)
Steps:
Step (i) {
    $\rho_{init-est}$=GSS($range = [0, 1]$
                 $start = \{\rho_{start}\}$
                 $f_{loglike} = SSE$) }
Step (ii) {
    $\rho_{opt}$=GSS($range = [0, \rho_{init-est}]$
                 $start = \{\rho_{init-est}\}$
                 $f_{loglike} = \frac{-2}{n} \cdot \ln |\mathbf{I} - \hat{\rho}\mathbf{W}| + SSE$)
    Compute ($\beta_{opt}, \sigma^2_{opt}$) }
return ($\rho_{opt}, \beta_{opt}, \sigma^2_{opt}$)
```

Figure 2. The NORTHSTAR algorithm.

finds the global minimum of the log-likelihood function. Thus, we have an optimal parameter estimation for the ML-based SAR model solutions. We plotted the log-likelihood function in order to see its extrema for the problem size of 2500 in Figure 3. We can also check the magnitudes of the components of the log-likelihood function, namely the log-det term and the *SSE* term, which will lead to our NORTHSTAR heuristic.

Figure 2 explicitly reveals, first, that the log-likelihood function is uni-modular and, second, that the log-det term in equation 2 is *very small* with respect to the *SSE* term. Table 2 shows the magnitudes of the log-det and *SSE* ter ms at the optimal ρ value where log-likelihood function is minimum for different neighborhood structures. The cost of the NORTHSTAR algorithm is dominated by the sparse LU factorization operation, which is $(j - m)(2nb_u b_l) + 9n^2 + 2j - 3$. The parameters b_u and b_l correspond to the upper bandwidth and lower bandwidth of the neighborhood matrix $\mathbf{W}$ respectively. The parameter $(j - m)$ is the number of log-det computations for NORTHSTAR algorithm.

Table 2. The magnitudes of the log-det and SSE terms at the optimal ρ value where the log-likelihood function is minimum

Problem Size(n)	Neighborhood	ρ_{opt}	abs(Log-Likelihood)	abs(log-det)	abs(*SSE*)
2500	4-N	0.467	15.185	15.125	0.061
2500	8-N	0.430	15.267	15.238	0.028

The total computational complexity (the operation counts) of our NORTHSTAR heuristic is listed in Figure 4 and it should be noted that $j \ll n$. The parameter j in Figure 4 is around 8 due to the fact that the estimated ρ from step (i) of NORTHSTAR is in the ± 0.1 range of the optimal ρ value. The variable m is the number of savings from log-det computations (i.e., 3 on average). EV computation based SAR model solution cannot go beyond problem size of $10K$ due to memory

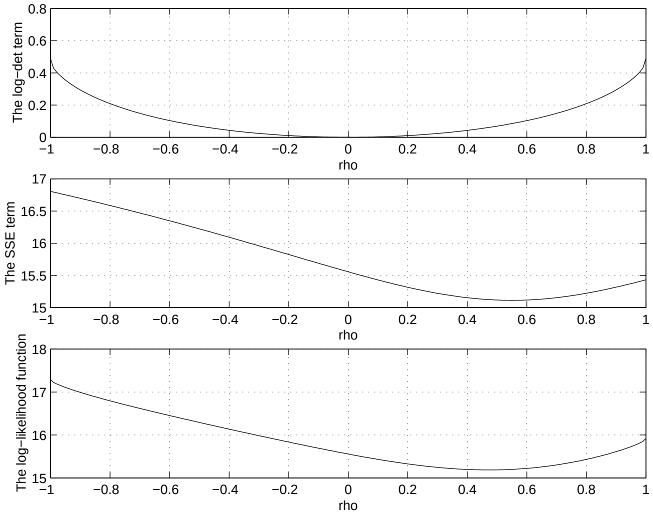

Figure 3. The components of the log-likelihood function

Problem Size	NORTHSTAR	EV	SLD
n	$(j-m)(2nb_ub_l)+9n^2+2j-3$	$\frac{2}{3}n^3+529n^2+j$	$j(2nb_ub_l)+9n^2+j$

Figure 4. The total computational complexity of NORTHSTAR, EV, and SLD

constraints. For large problem sizes, our approach is much more computationally efficient than the SLD approach and the EV approach.