Modeling Spatial Dependencies for Mining Geospatial Data

Sanjay Chawla, Shashi Shekhar and Weili Wu

Department of Computer Science

University of Minnesota

{chawla, shekhar,wuw }@cs.umn.edu

http://www.cs.umn.edu/Research/shashi-group

Overview

- Motivation
- Problem Formulation
- Basic concept and Related works
- Our contributions and approach
- Experiment Design
- Conclusion and Future works

Motivation

- Widespread use of spatial databases
  - *Triggers increasing interest in mining spatial patterns.*
  - *The 1855 Asiatic Cholera in London[Griffith]*
    - Asiatic Cholera was sweeping through London
    - Epidemiologist marked on a map all the spatial locations
    - Formed a cluster and the centroid turned out to be a water-pump
    - Turned-off the water-pump and the cholera epidemic subsided
- Fair Lending [NY, R. Nader]
  - *Correlation of bank locations with loan activity in poor neighborhoods.*
- Retail Outlets[NY, Walmart, McDonald, etc]
  - *Determining locations of stores by relating neighborhood maps with customer databases.*
- Crime Hot Spot Analysis[NY, NIJ CML]
  - *Explaining clusters of sexual assaults by locating addresses of registered sex-offenders.*
- Ecology[Uygar]
  - *Explaining location of bird nests based on structural and environmental variables.*

Problem Formulation

- Given
  - *A spatial framework* S *consisting of sites* $\{s_1, \ldots, s_n\}$ *for an underlying geographic space* G.
  - *A collection of explanatory functions* $f_{X_k} : S \to R^k, k = 1, \ldots K$. R^k *is the range of possible values for the explanatory functions.*
  - *A dependent function* $f_Y : S \to R^Y$
  - *A family* $\mathcal{F}$ *of learning model functions mapping* $R^1 \times \ldots R^K \to R^Y$.
- Find
  - *A function* $\hat{f}^Y \in \mathcal{F}$.
- Objective
  - *maximize similarity*$\Big(map_{s_i \in S}(\hat{f}^Y(f_{X_1}, \ldots, f_{X_K})), map(f_Y(s_i))\Big)$
    $= (1-\alpha)$ *classification_accuracy*$(\hat{f}^Y, f_Y)+(\alpha)$*spatial_accuracy*$((\hat{f}^Y, f_Y)$
- Constraints
  1. **Geographic Space S is a multi-dimensional Euclidean Space [1].**
  2. **The values of the explanatory functions, the f_{X_k}'s and the response function f_Y may not be independent with respect to those of nearby spatial sites, i.e. spatial autocorrelation exists.**
  3. **The domain R^k of the explanatory functions is the one-dimensional domain of real numbers.**
  4. **The domain of the dependent variable, $R^Y = \{0, 1\}$.**

[1]The entire surface of the Earth cannot be modeled as a Euclidean space but locally the approximation holds true.

Example of bird nests prediction

- Example

  - *Given:* $attributes$,***Dominant Vegetation Type(V)*** , ***Distance to Open Water(DOW)*** **and** ***Water Depth(WD)***

  - *Find:*$Prob(C = nestsite|v, dow, wd)$

  - *Objective: Maximize class(nest, no-nest) accuracy.*

  - *Constraint: Location of a nest influences location of neighboring nests*

Bird habitat maps(dependent variable)

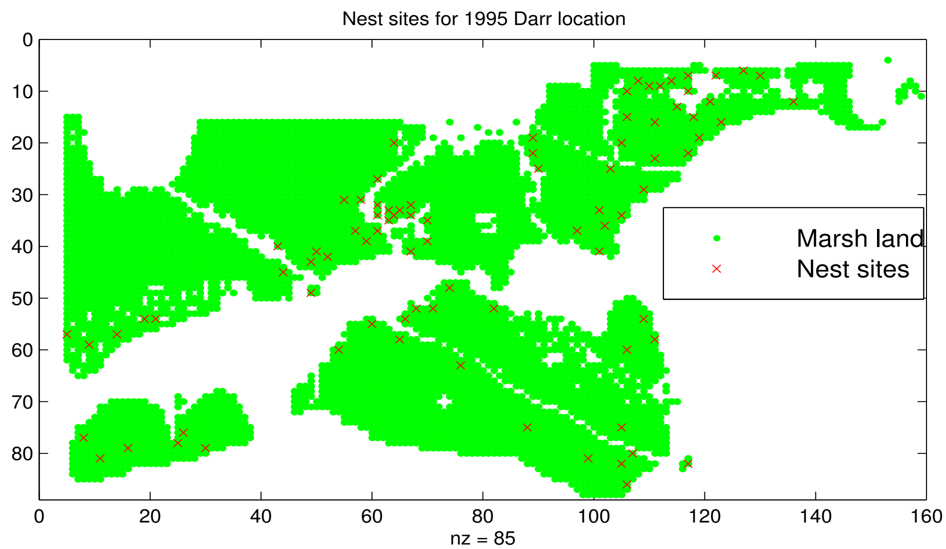


Figure 1: Nests site locations in Darr Marshland in 1995(Learning Data)

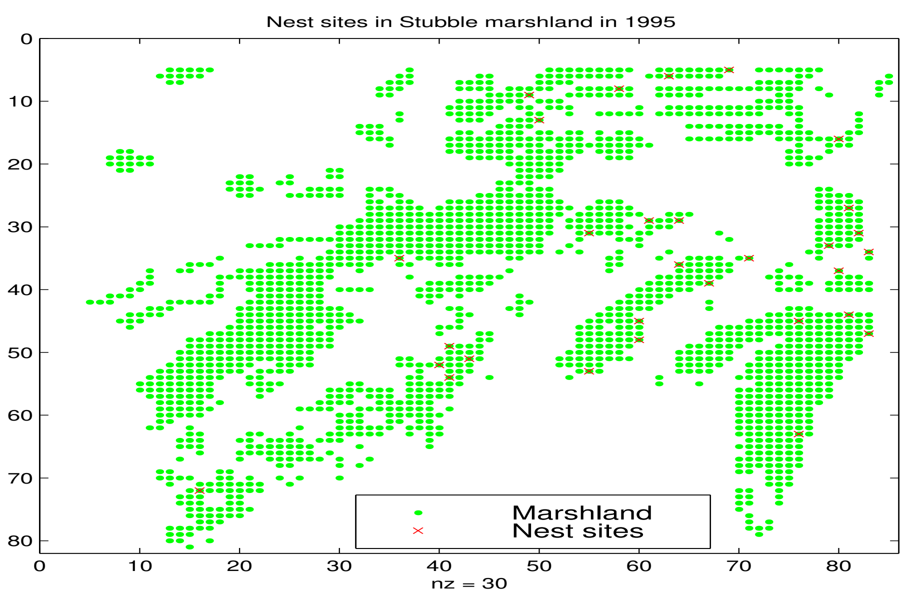


Figure 2: Nests site locations in Stubble Marshland in 1995(Test Data)

Explanatory variables maps (Learning data)

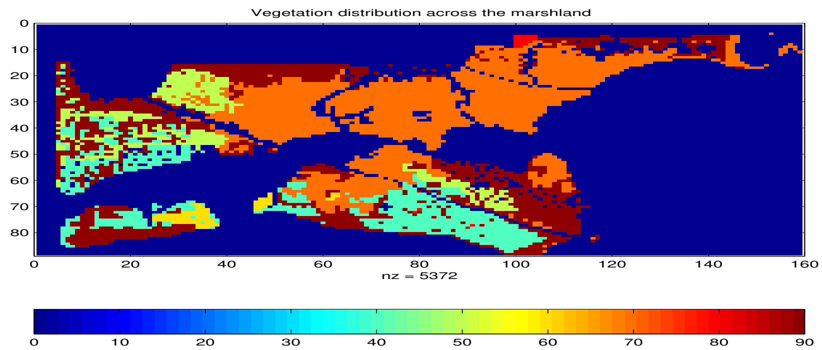


Figure 3: Vegetation layer for 1995 Darr marshland

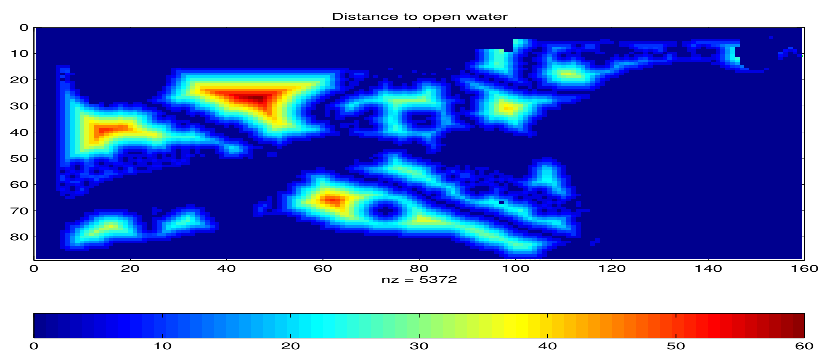


Figure 4: Distance to open water layer for 1995 Darr marshland

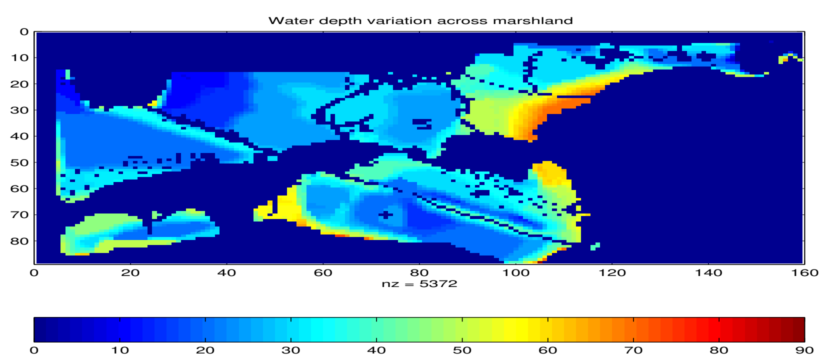


Figure 5: Water depth layer for 1995 Darr marshland

Explanatory variables maps (Testing data)

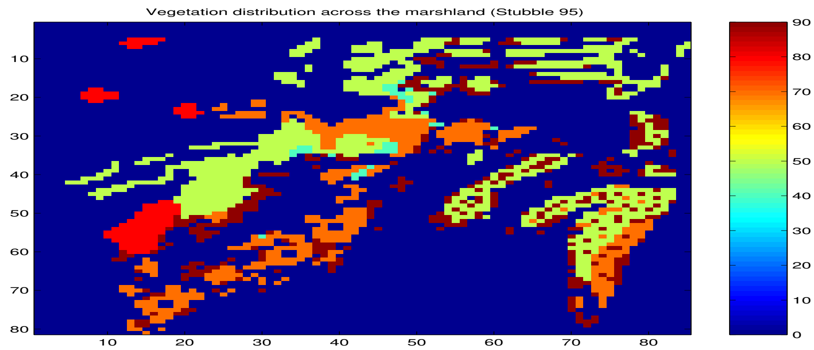


Figure 6: Vegetation layer for 1995 Stubble marshland

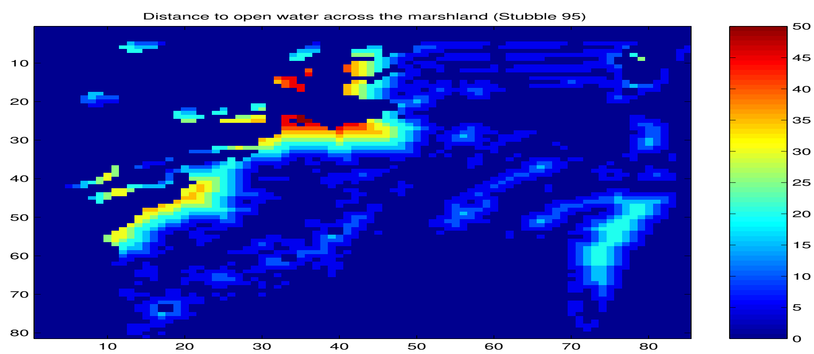


Figure 7: Distance to open water layer for 1995 Stubble marshland

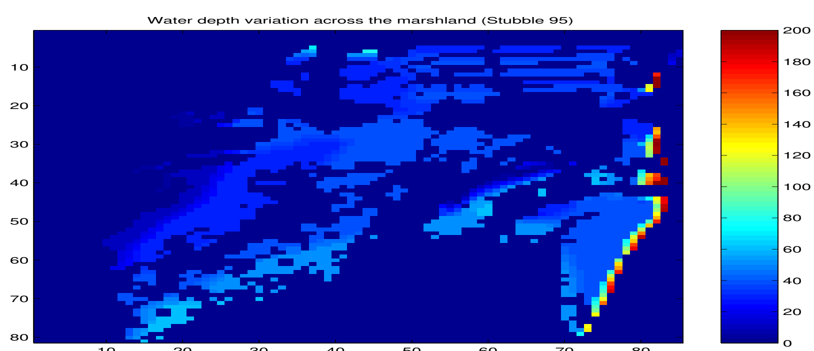


Figure 8: Water depth layer for 1995 Stubble marshland

Basic Concept and Related works

- Basic Concept
  - *Classical linear regression model and its assumptions*
  - *Classical assumptions in context of spatial data*
  - *Spatial Autocorrelation*

- Related works
  - *Spatial Autoregressive Model (SAR)*
  - *Logit regression model*
  - *Probit regression model*
  - *Artificial neural network model*
  - *Benefits of modeling spatial autocorrelation*

Classical linear regression and its assumptions

- Given a vector $\vec{y}$ of n observations and a matrix X of explanatory variables
- Encapsulates their interrelationship using the standard linear equation

$$y_i = \beta_0 + \sum_j \beta_j X_{ij} + \epsilon_j \quad (i = 1, \ldots, n \; j = 1, \ldots, m)$$

- Vector form:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon.$$

- Example from bird habitat
  - *Index i varies over pixels of space*
  - $y_i = 1$ **(nest site),** $y_i = 0$ **(no nest)**
  - X_0: *constant*
  - X_1: *Vegetation*
  - X_2: *DistWater*
  - X_3: *Depth*
- Definition of TSE

$$TSE = \frac{\sum |y_{actual} - y_{predict}|^2}{n}$$

- Assumptions
  - *Error terms* $\epsilon_i \approx N(0, \sigma^2)$
  - *Error terms are independent*
  - *Identically distributed normal random variables*

Related works

- Binary response: Generalized Linear Model [Flurry 95]
  - *Logit*
    - Logit regression model

$$\theta = Prob(y=1) = \frac{e^{X\beta}}{1+e^{X\beta}}$$

    - Called Generalized Linear

$$\lambda_i = \log(\frac{\theta_i}{1-\theta_i}) = X\beta.$$

    - Logistic distribution
  - *Probit regression Model*

$$Prob(y=1) = \frac{1}{2\pi}\int_{-\infty}^{X\beta} e^{-\frac{t^2}{2}} dt$$

    - Cumulative Normal Distribution
- ROC Curve measure
  - *Receiver Operating Characteristic(ROC) curve*
  - *True Positive Ratio(TPR) vs. False Positive Ratio(FPR)*
  -

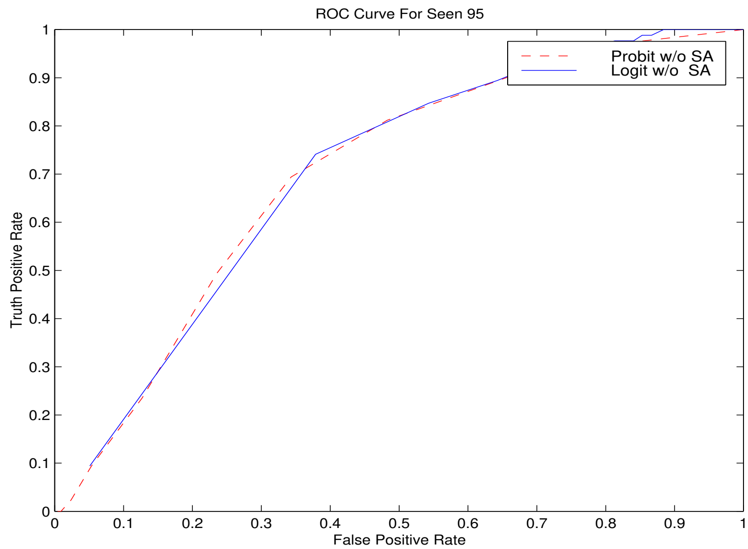
$$TPR = \frac{\sum_i I_{y_p|p>b} I_{[y_i=1]}}{\sum_i y_i}$$

$$FPR = \frac{\sum_i I_{y_p|p>b} I_{[y_i=0]}}{\sum_i (1-y_i)}$$

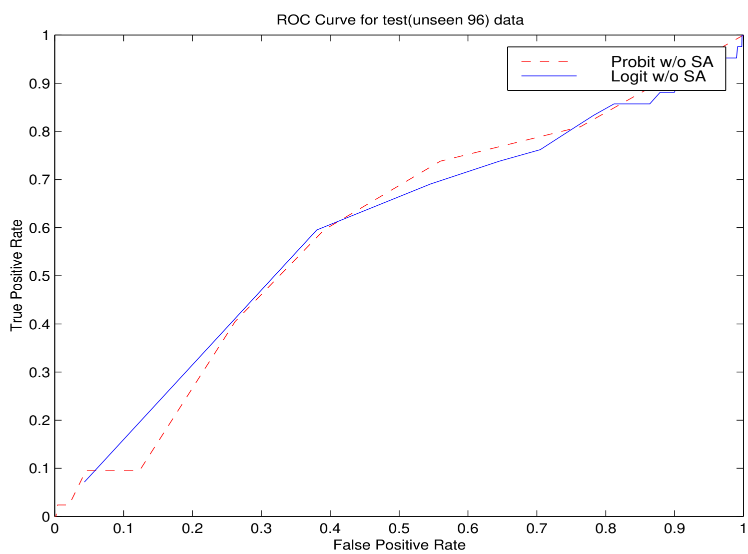
  - $0 \le b \le 1$ *and* I *is an indicator function*
  - *Area above* $TPR = FPR$ *is the percentage better than random*

Results for Logit and Probit

- Model via learning 1995 data



- Testing model via 1996 data



The assumptions may not be valid for spatial data and observations

- The residual error may have systematic variation over space
- The occurrence of one feature is influenced by the distribution of similar features in the adjacent area
- This property is called spatial autocorrelation
- May yield biased and inconsistent estimate of the relationship between $\vec{y}$ and $\tilde{X}$
- In addition,may lead to poor fit of the model
- Moran's I results

Explanatory Variable	**Four Neighbor**	**Eight Neighbor**
Distance to Edge	0.7606	0.9032
Distance to Open Water	0.7342	0.8022
Depth	0.6476	0.7408
Vegetation Height	0.7742	0.8149
Stem Density	0.6267	0.7653
Vegetation Durability	0.3322	0.4851

Table 1: Moran I coefficient of explanatory variables to predict nest locations for the Red Winged Blackbird

Spatial Data Mining

- Reduce assumptions behind classical data mining
  - *Independent identical distributed error term*
- Quantify the spatial dependence
- Factor it into techniques for estimation
- Spatial data is more complex
  - *including extended objects such as points, lines, and polygons [egen-93]*
  - *Spatial predicates (e.g. overlap) and attributes (e.g. distance, spatial auto-correlation) are often implicit [kope-95].*

Spatial Autocorrelation (SA)

- Autocorrelation:
  - *From the field of time series analysis*
  - *Study similarity/dissimilarity of the same variables between intervals*
  - *Intervals: temporal or spatial nature*

- Spatial Autocorrelation (SA) :
  - *Measure the spatial influence of values across space*
    - e.g. How the soil fertility varies across the spatial landscape?
  - *First Law of Geography: Everything is related to everything else but nearby things are more related than distant things*
  - *Measures*
    - Moran's I [A. Cliff 73]
    - Nearest neighbor index (NNI) [Chou 97]

Moran's I measure

- Definition:

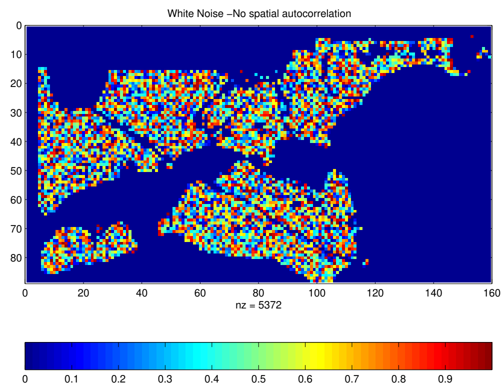
$$MI = \frac{n}{\sum_{i=1}^{i=n}\sum_{j=1}^{j=n} W_{ij}} \cdot \frac{\sum_{i=1}^{i=n}\sum_{j=1}^{j=n} W_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{i=n}(x_i - \bar{x})^2}$$

  - *n: the number of data points*
  - *$x_i's$: data values*
  - *$\bar{x}$: mean of x*
  - *W: the design or contiguity matrix*

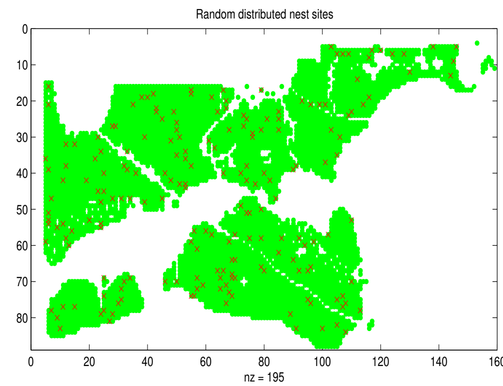
- Ranges between -1 and +1
  - *higher positive value $\Rightarrow$ high SA, Cluster, Attract*
  - *lower negative value $\Rightarrow$ interspersed, de-clustered, repel*
  - *zero $\Rightarrow$ random*

Examples of data w/o SA and data with SA

- Random distributed data (data w/o SA)



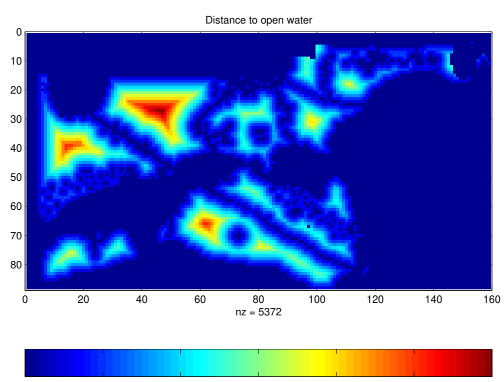
(a) pixel property with independent identical distribution



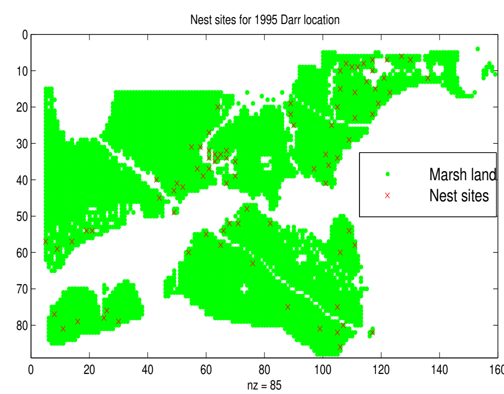
(b) Random nest locations

Figure 9: Spatial distribution satisfying distribution assumptions of classical regression

- Cluster distributed data (data with SA)



(a) pixel property with spatial autocorrelation



(b) Cluster nest locations

Figure 10: Spatial distribution NOT satisfying distribution assumptions of classical regression

Related works

- Spatial Autocorrelation
  - *Spatial autoregressive model (SAR)[Anselin 88]*

$$y = \rho W y + X\beta + \epsilon.$$

  - *Assume: depend values* y_i' *are related to each other*

$$y_i = f(y_j)\ i \neq j,\ j \in nbd(i)$$

  - *When* $\rho = 0$, **SAR** $\approx$ *Std regression model*

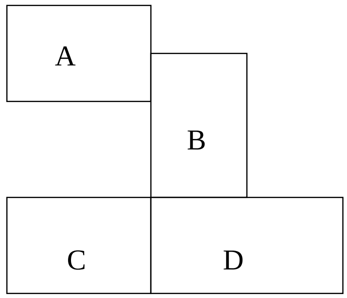
- Combining SA with Logit and Probit
  - *Logit with SA*

$$Prob(y = 1) = \frac{e^{\rho W y + X\beta}}{1 + e^{\rho W y + X\beta}}$$

  - *Probit with SA [Lesage 97, McMillen 95]*

$$Prob(y = 1) = \frac{1}{2\pi} \int_{-\infty}^{\rho W y + X\beta} e^{-\frac{t^2}{2}} dt$$

Contiguity Matrix W

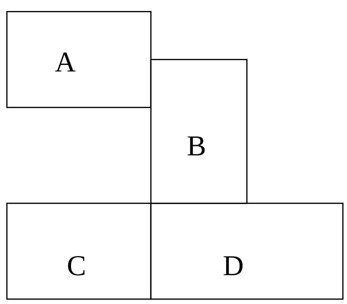


(a)

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

(b)

Figure 11: An example contiguity matrix



(a)

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1/3 & 0 & 1/3 & 1/3 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 1/2 & 1/2 & 0 \end{bmatrix}$$

(b)

Figure 12: A normalized example contiguity matrix

Example of including spatial autocorrelation in regression model

- Data Set:
  - *Crime in 49 neighborhoods in Columbus Ohio[Anselin 88]*
  - *Dependent variable:* **number of crime incidents**, y
  - *Independent variable:* **mean income** *and* **mean house value**.

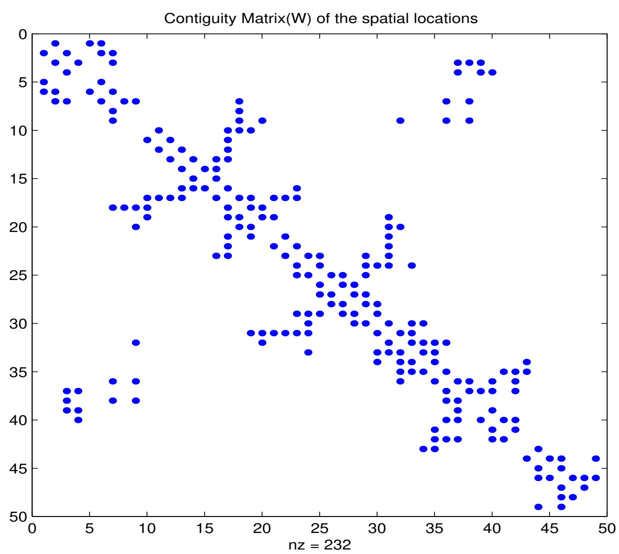


Figure 13: The contiguity matrix of locations

	R-square	Moran I(residual)
Ordinary Regression	**0.5521**	**0.23**
Spatial Auto Regression	**0.6518**	**0.04**

Benefits of modeling spatial autocorrelation

- Have a better fit
  - *Lower total square error of fit*
- More accurate coefficient (β) for independent variables
- Reduce spatial autocorrelation (SA) in residual error
  - *With proper contiguity matrix w, no SA in residual error*

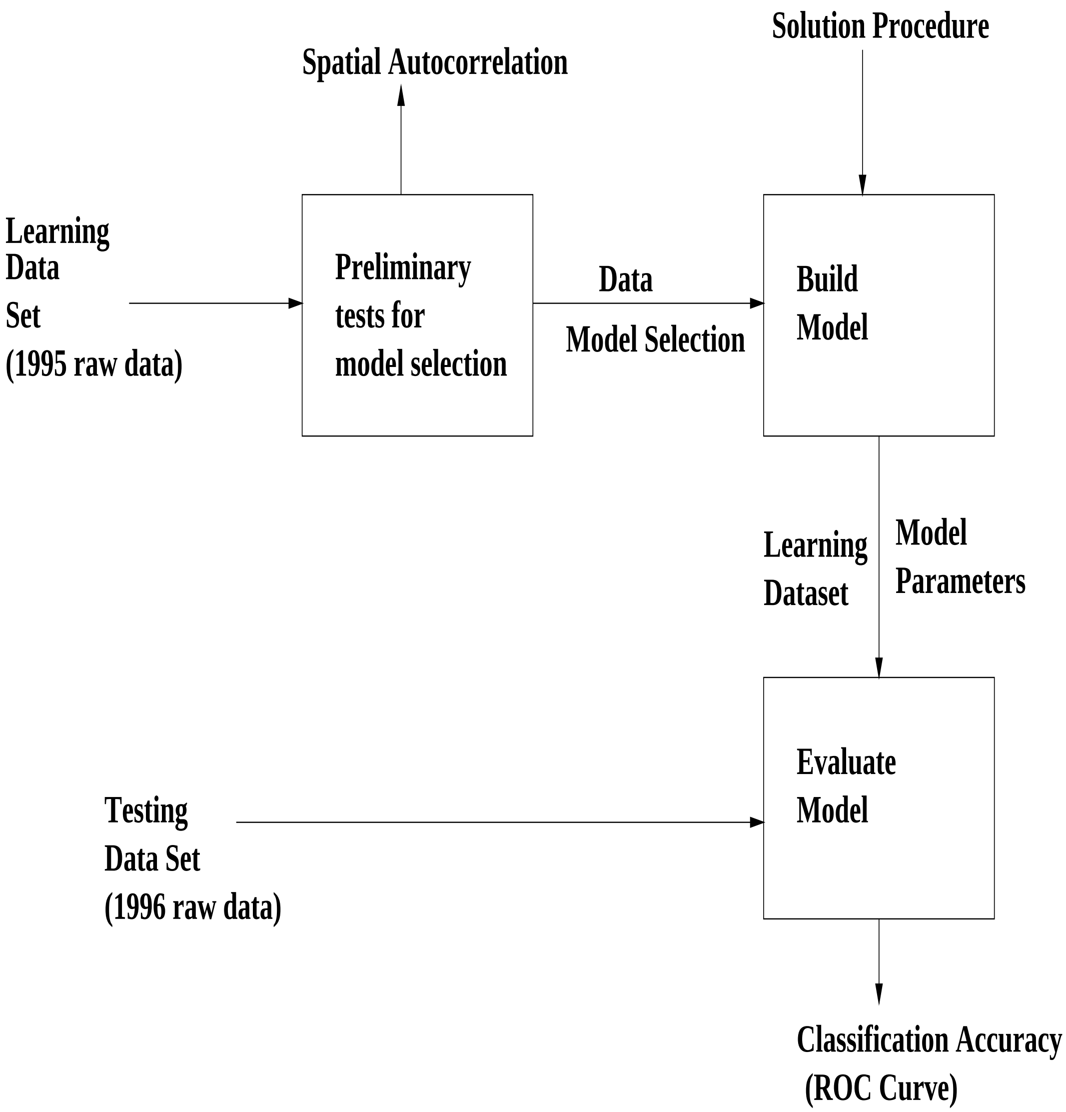
Survey of related work

- Criteria of comparison
  - *Model*
  - *Measure optimized*
  - *Solution procedure (search algorithm)*
- List of different models

Method Name	Model Type	Spatial AC	Dependent Var. Type	Accuracy Measure	Solution Procedure
Linear Regression	Linear	No	Numeric	Total Square Error(TSE)	Closed Form
Neural Networks	Nonlinear	No	Numeric/ Categorical	TSE	Gradient Descent Back-Propagation
Probit	Gen. Linear	No	Binary	TPR/FPR	Gradient Descent
Logit	Gen. Linear	No	Binary	TPR/FPR	Gradient Descent
SAR + Probit	en. Linear	Yes	Binary	TPR/FPR	ML/EM/Gibbs [2]

- Limitation
  - *Choice of w non-trivial, but crucial to success*
  - *Effect of incorrect w [Griffith]*

Experiment Design



Result for Probit(w/o SA) and SAR(probit with SA)

- Temporal Generalization
  - *Learning model via 1995 data and Testing model via 96 unseen data*

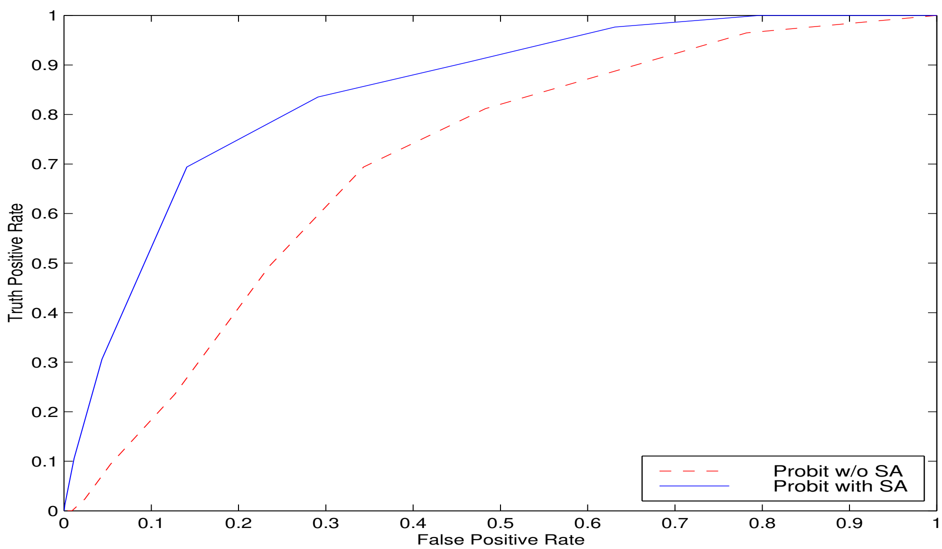


Figure 14: ROC curve of learning model via Darr95 data

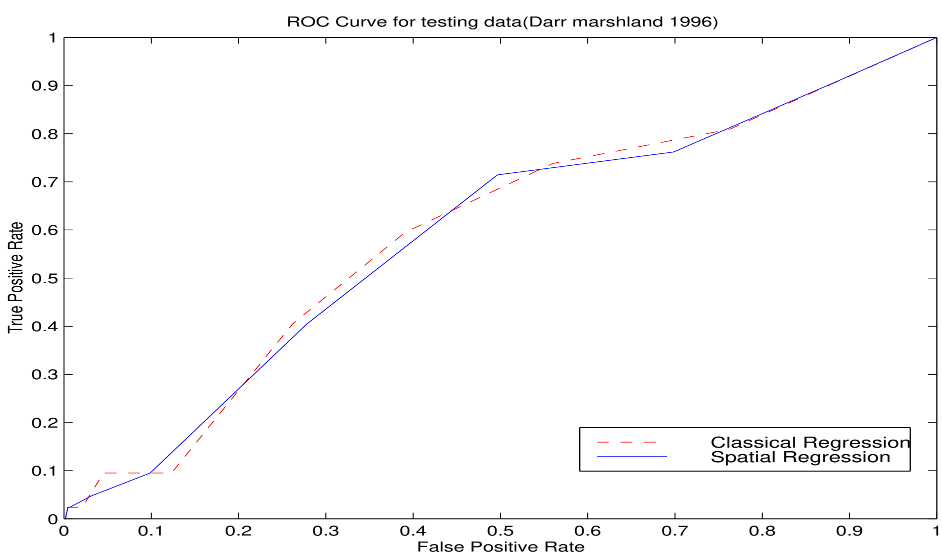


Figure 15: ROC curve of testing model via Darr96 data

Result for Probit(w/o SA) and SAR(probit with SA)

- Spatial Generalization

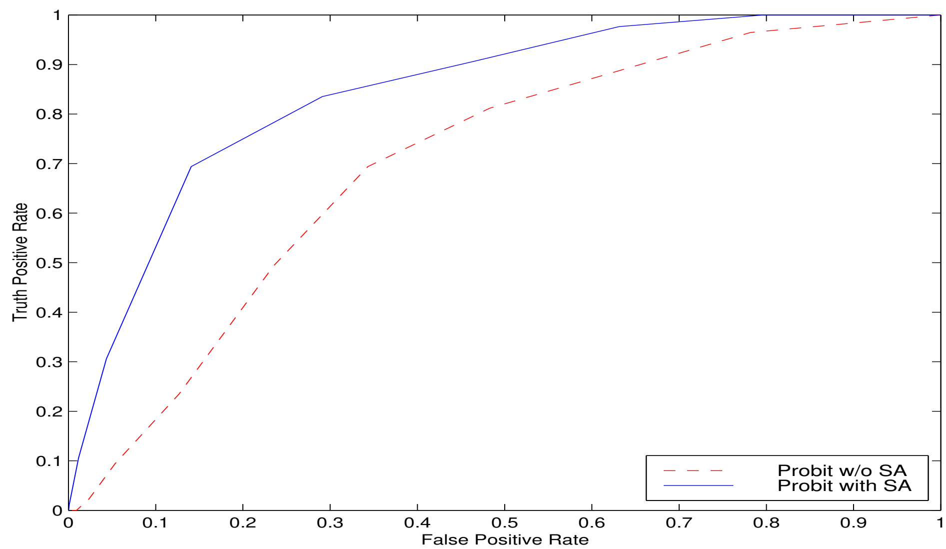


Figure 16: ROC curve of learning model via Darr95 data

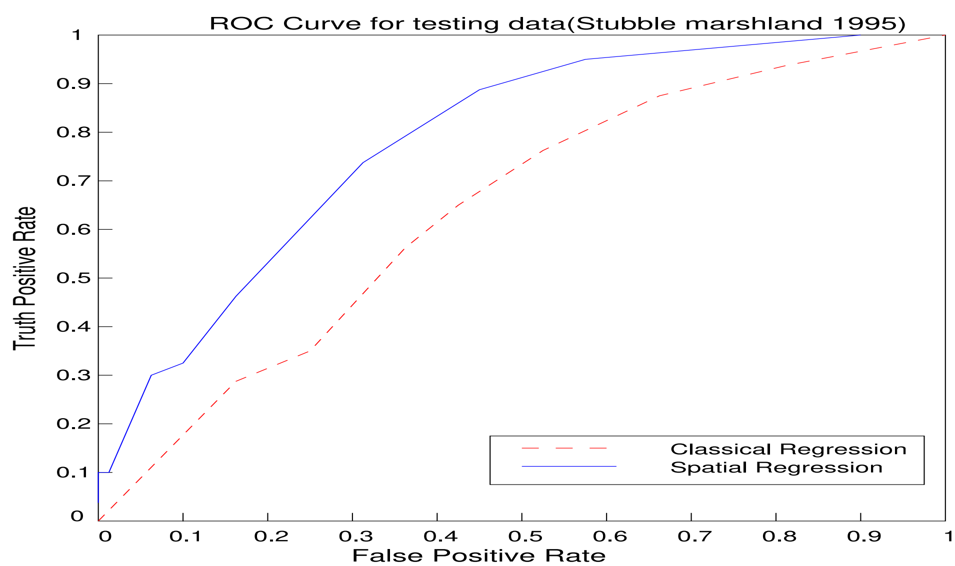
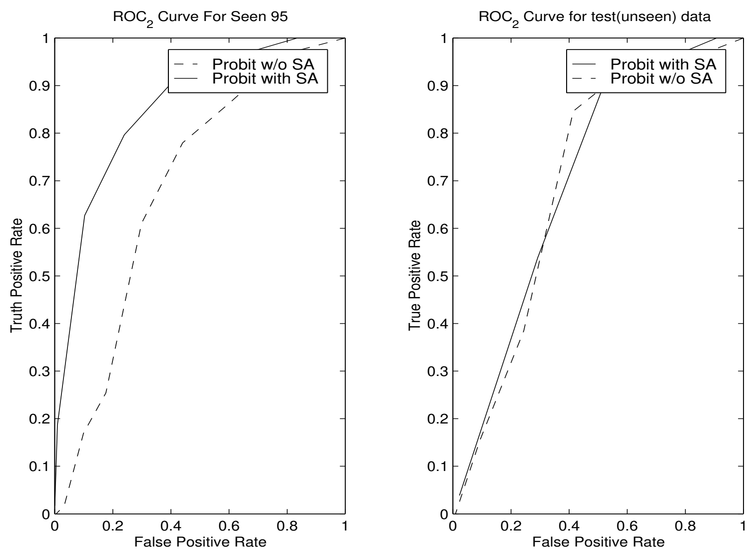


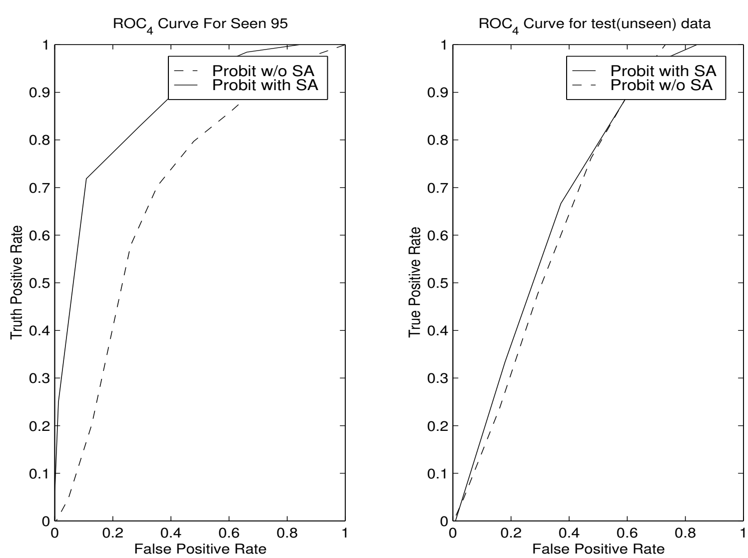
Figure 17: ROC curve of testing model via Stubble95 data

Result for Probit(w/o SA) and SAR(probit with SA) with different W

- Learning and testing Models with Four-neighborhood contiguity matrix



- Learning and Testing models with Eight-neighborhood contiguity matrix



Result of Coefficient β for diff models

	Logit(w/o SA)	Probit(w/o SA)	Probit with SA
Constant	-4.2814	-2.2442	-3.2245
Vegetation durability	0.0097	0.0041	0.2618
Distance to open-water	-0.0891	-0.0341	-0.2742
Water depth	0.0146	0.0062	0.1872

Table 2: Coefficient for different models

- Computation of Moran I is expensive
- Interpretation of Moran I for residuals not natural(binary dependent variable)

Critique of SAR model

- Very slow, not scalable
- Memory Hog
- 10^6 pixel map $\rightarrow$ 10^{12} element W
- Choice of W is non-trivial
- Classification accuracy(ROC) is a weak measure

Basic Idea:Motivation

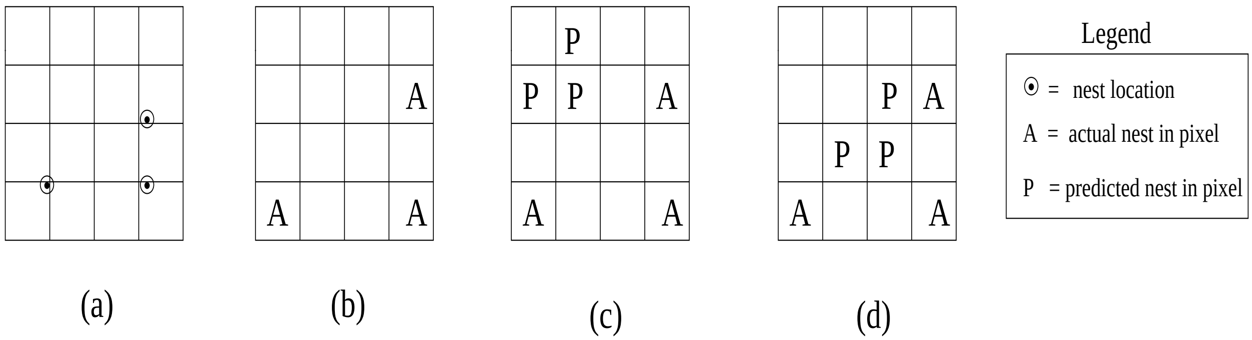


Figure 18: (a)The actual locations of nest, (b)Pixels with actual nests, (c)Location predicted by a model, (d)Location predicted by another mode. Prediction(d) is spatially more accurate than (c). Classical measures of classification accuracy will not capture this distincition.

- Figure (a): Actual location of nests
- Figure (b): Pixels with actual nests
- Figure (c): Locations predicted by a model
- Figure (d): Locations predicted by another model
- Analysis:
  - *Current measure fails to distinguish between the two models*
  - *Model shown in Figure (d) is spatially more accurate than (c)*

Basic Idea:New measure

0.71	0.6	0.5	0.6	0.71
0.6	0.6	0.7	0.6	0.6
0.6	0.8	0.8	0.8	0.7
0.5	0.5	0.5	0.5	0.6
0.73	0.5	0.6	0.4	0.72

EXAMPLE PROBABILITY SURFACE

0.71	0.6	0.5	0.6	0.71
0.6	0.6	0.7	0.6	0.6
0.6	0.8	0.8	0.8	0.7
0.5	0.5	0.5	0.5	0.6
0.73	0.5	0.6	0.4	0.72

FOUR SITES SELECTED ON BASIS OF HIGHEST WEIGHT

0.71	0.6	0.5	0.6	0.71 ?
0.6	0.6	0.7	0.6	0.6
0.6	0.8	0.8 ?	0.8	0.7
0.5	0.5	0.5	0.5	0.6
0.73	0.5	0.6	0.4	0.72

FOUR SITES SELECTED ON BASIS OF WEIGHTED ZONE OF INFLUENCE

Our contributions

- New measure !
- New algorithm
- Simple implementation
- Preliminary experiment evaluation

Our approach

- Basic Idea
- New measure
- Model choices
- Solution procedure choices
- Current approach
- Implementation details

New measure

- Average distance to nearest predicted neighbor (ADN)

$$ADN = \frac{\sum_{actual_i} minDist(actual_i, anyPredicated)}{NumOfActual}$$

- Connection with NNI

$$NNI = \frac{Ad}{Ed}$$

  - Ad:avg nearest neighbor dist of the point pattern
  - Ed: expected avg nearest dist

$$Ed = \frac{1}{2}\sqrt{\frac{A}{n}}$$

  - A: Map area
  - n: num of point
  - *Assume: Num of actual nests = Num of predicted*

Model choices

Method Name	Model Type	Spatial AC	Dependent Var. Type	Accuracy Measure	Solution Procedure
Linear Regression	Linear	No	Numerical	Total Square Error(TSE)	Closed Form
Neural Networks	Nonlinear	No	Numeric/ Categorical	TSE	Iteration Back-Propagation
Probit	Gen. Linear	No	Binary	TPR/FPR	Iterative
Logit	Gen. Linear	No	Binary	TPR/FPR	Iterative
Spatial Auto Reg. (SAR)	Linear	Yes	Numeric	TSE	Maximum Likelihood(ML)
SAR + Probit	Gen-Linear	Yes	Binary	TPR/FPR	ML/EM/Gibbs [3]
Our Method (ADNP)	Gen. Linear	No	Binary	Distance/ Proximity	Greedy Search/ Iterative

- Our Approach
  - *Generalized linear model*
    - Probit w/o spatial autocorrelation
    - Logit w/o spatial autocorrelation
  - *New Measure*

Solution procedure choices

- Closed form
  - *Classical linear regression model*

$$\beta = (X'X)^{-1}X'y$$

- Gradient descent search, Newton-Raphson method
  - *Artificial neural network(e.g. back-propagation,SAR, Probit/Logit)*
  - *Iterative technique*
  - *minimizing the error*
- Greedy search
  - *Iterative technique*
  - *A**
  - *Dijkstra shortest path*

PLUMS Approach

- Framwork for the location prediction process

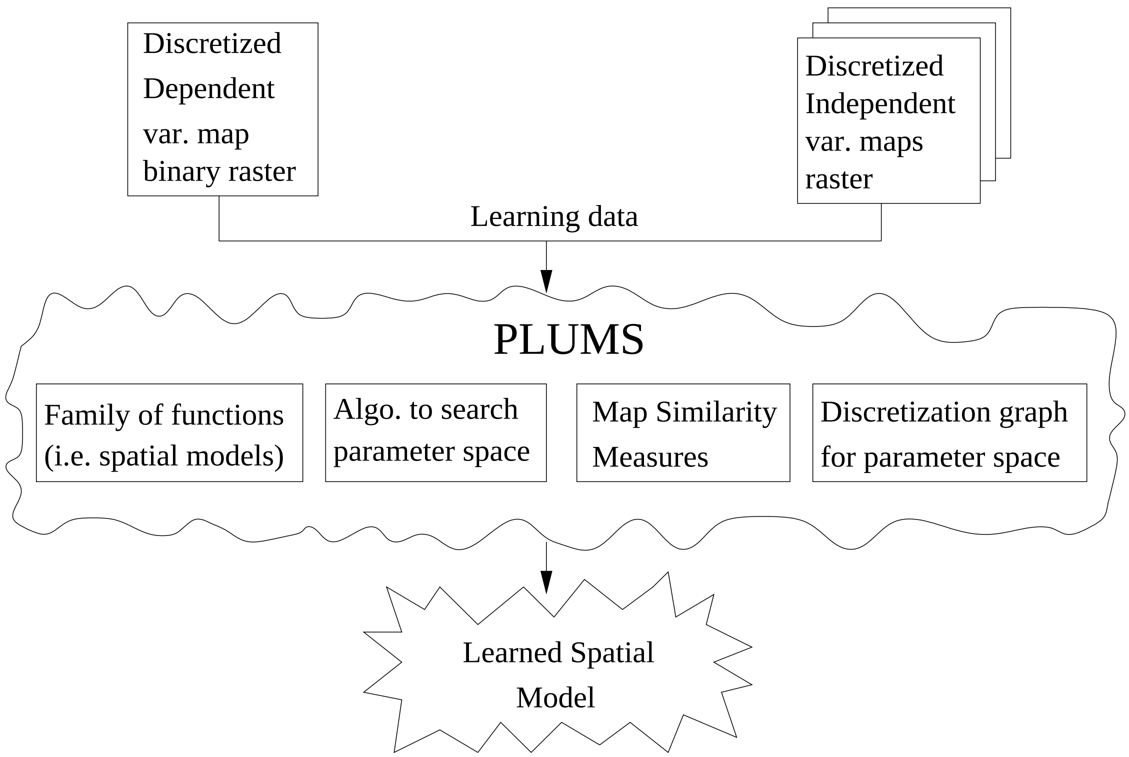


Figure 19: The framework for the location prediction process

Configuration Space

- Space of design choices for PLUMS components

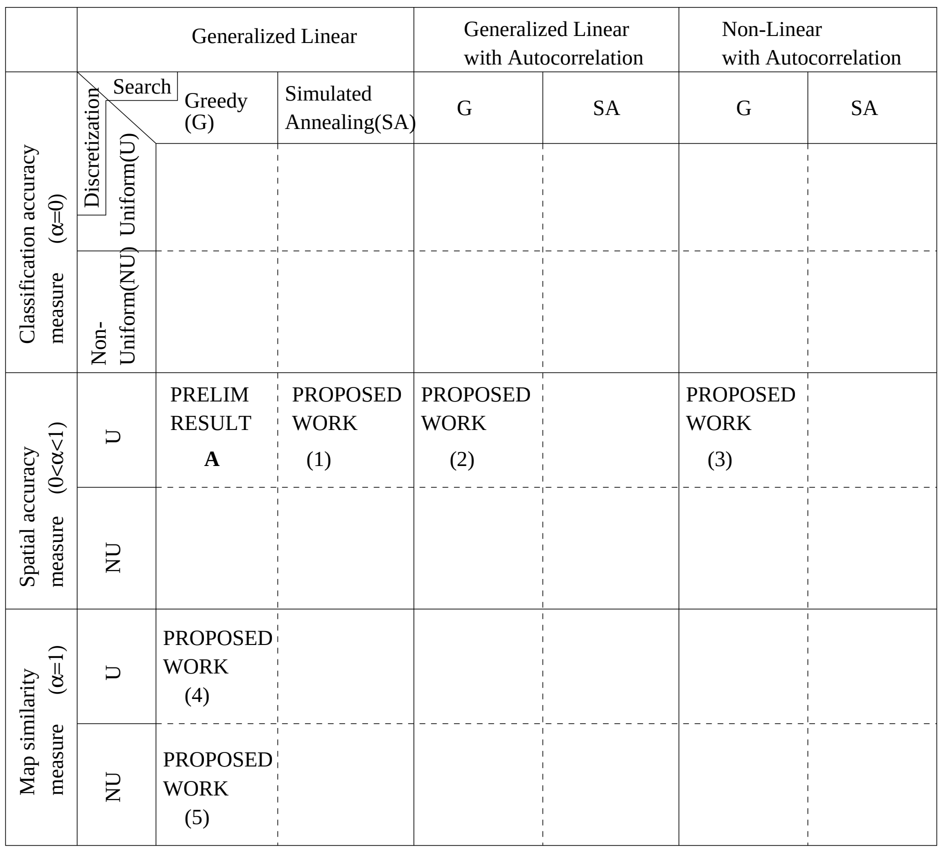


Figure 20: Space of design choices for PLUMS components:*function family, map-similarity measure, search algorithms* and *discretization.***G** refers to Greedy search and **SA** r efers to Simulated Annealing. **U** and **NU** refer to uniform and non-uniform grid based discretization of parameter space respectively.

PLUMS Component Choices	
Component	Choices
Map similarity	avg. distance to nearest prediction from actual, nearest neighbor index, ...
Search algorithm	greedy, gradient descent, simulated annealing, ...
Function family	generalized linear(GL) (logit, probit), non-linear, GL with autocorrelation
Discretization of parameter space	Uniform, non-uniform, multi-resolution, ...

Table 3: PLUMS Component Choices

Greedy search algorithm psuedo-code

```
parameter-value-set find-A-local-maxima(parameter-value-set PVS, discretization-of-parameter-space SF, map-similarity-measure-function MSM, learning-map-set LMS) {
    parameter-value-set best-neighbor, a-neighbor;
    real best-improvement=1, an-improvement;
    while(best-improvement > 0) do {
        best-neighbor = PVS.get-a-neighbor(SF);
        best-improvement = MSM(best-neighbor,LMS) - MSM(PVS,LMS);
        foreach a-neighbor in PVS.get-all-neighbors(SF) do {
            an-improvement = MSM(a-neighbor,LMS) - MSM(PVS,LMS);
            if(an-improvement > best-improvement) {
                best-neighbor = a-neighbor; best-improvement = an-improvement;
            }
        }
        if (best-improvement > 0) then PVS=best-neighbor;
    } /* found a local maxima in parameter space */
    return PVS;
}
```

Current Approach

- **Given:**
  - *y a spatially indexed binary dependent variable and*
  - $X = (X_1, \ldots, X_m)$, *explanatory variable .*
  - *Let* $A = \{s_k | y_{s_k} = 1\}$ *be the spatial locations*
  - *with positive response* $(y = 1)$.
  - $|A| = K$.

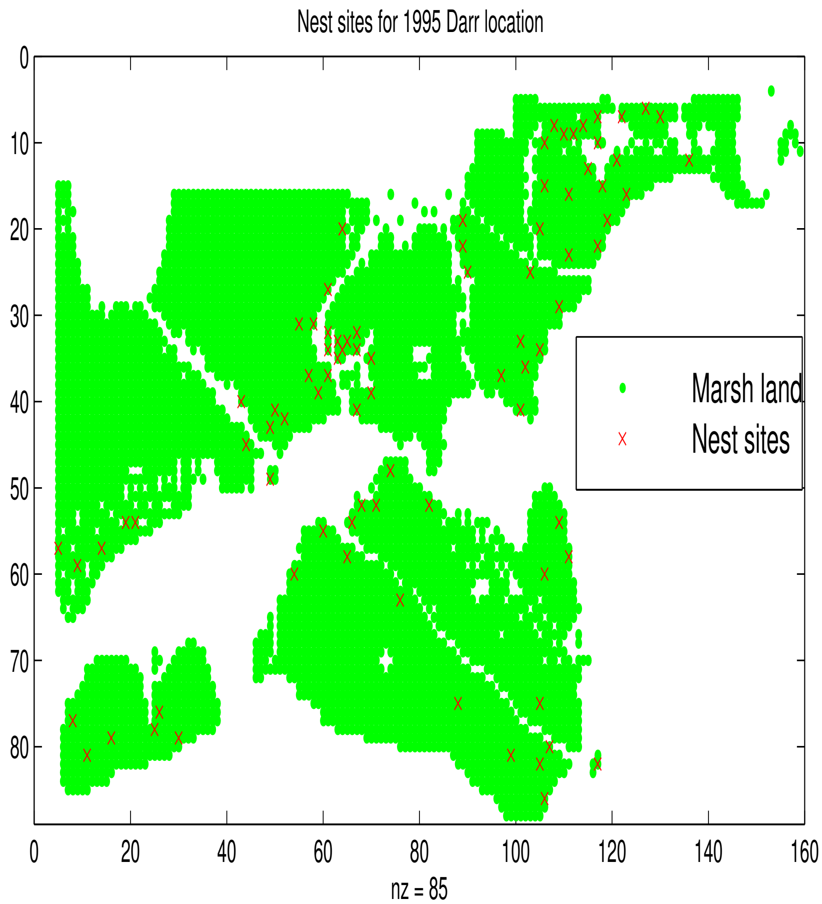


Figure 21: Nest location map for 1995 marshland

Explanatory variables maps

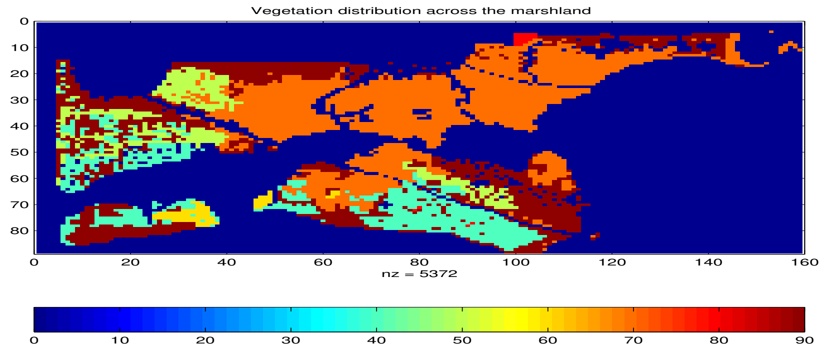


Figure 22: Vegetation layer for 1995 marshland

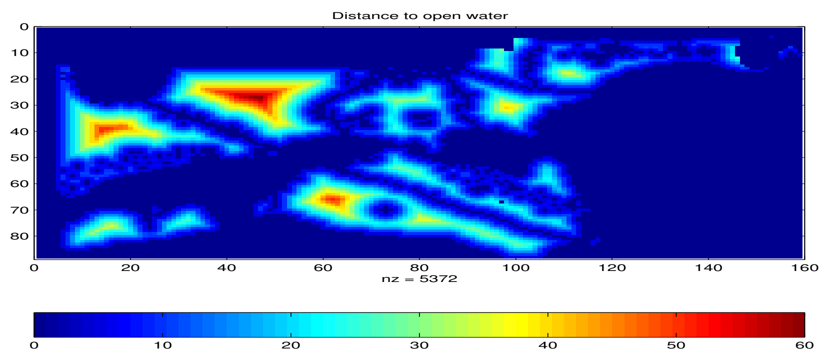


Figure 23: Distance to open water layer for 1995 marshland

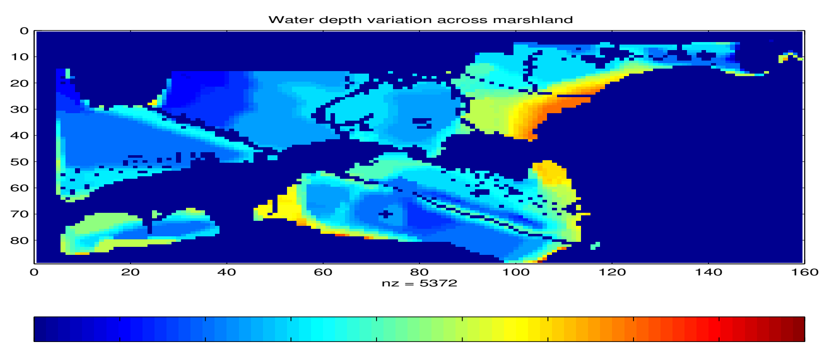


Figure 24: Water depth layer for 1995 marshland

Current approach (Cont.)

- **Find:** $\vec{\beta}^T = (\beta_1, \ldots, \beta_m)$ **such that**

$$Prob(y=1) = \frac{1}{2\pi}\int_{-\infty}^{X\beta^T} e^{\frac{-t^2}{2}} dt$$

- **Objective: Minimize**

$$\frac{1}{K}\sum_{k=1}^{K} d(s_{y_k}, s_{y_k}.nearest(P))$$

where

1. $P = \{p_1, \ldots p_k\}$ **are spatial locations on the predicted probability surface with the K-highest probability values.**
2. $s_{y_k}.nearest(P)$ **is the nearest predicted location to** s_{y_k}**.**
3. $d(.,.)$ **is a distance function(Euclidean, Manhattan, etc.)**

- **Constraints:**

1. $L_i \le \beta_i \le U_i, \; i = 1, \ldots, m$
2. $|\beta_i^p - \beta_i^{p+1}| = \delta_i$

Implementation details

- Find actual nest sites (A) from nest layer pixels
  - *Find nonzero entries from "Nest" attribute values and it's index num*
  - *Num of actual nest site is 85 for year 1995*
  - *Get it's corresponding coordinate via index num*
- Construct predicate nest sites (P) with initial β^0
  - *pred_Location = Solve_Probit($X\beta^0$) [leSage 99]*
  - *Assume: num of actual nest sites = num of predicted nest sites*
  - *Sort the pred_Location values in descending order*
  - *Take highest 85 probability surface and it's corresponding index p_i*
  - *Get predicted nest locations from coordinate file via p_i*

Implementation details (Cont.)

1. **Compute** rad_0

   - $rad_0 = RAD(A, P(\beta^0))$
     $RAD(A, P) = \frac{ADN(A,P)}{\sqrt{ADN(A,A) \cdot ADN(P,P)}}$

2. **Find local minima via ADNP algorithm**

   - *Define neighborhood of* $< \beta_1, \beta_2 >$
   - *Includes* $< \beta_1 + \delta, \beta_2 >, < \beta_1, \beta_2 + \delta >, < \beta_1 - \delta, \beta_2 >$ *and* $< \beta_1, \beta_2 - \delta >$
   - *For each neighbor* β_i **of** β^0
   - *Compute* RAD_i

3. **Repeat step 1, 2 to find local minima**

4. **Repeat step 1,2,3 with 10 initial weights**

Experiment Results

- Experiment design
- Candidates
- Data set
- Data graphs, data tables, data maps

Experiment design

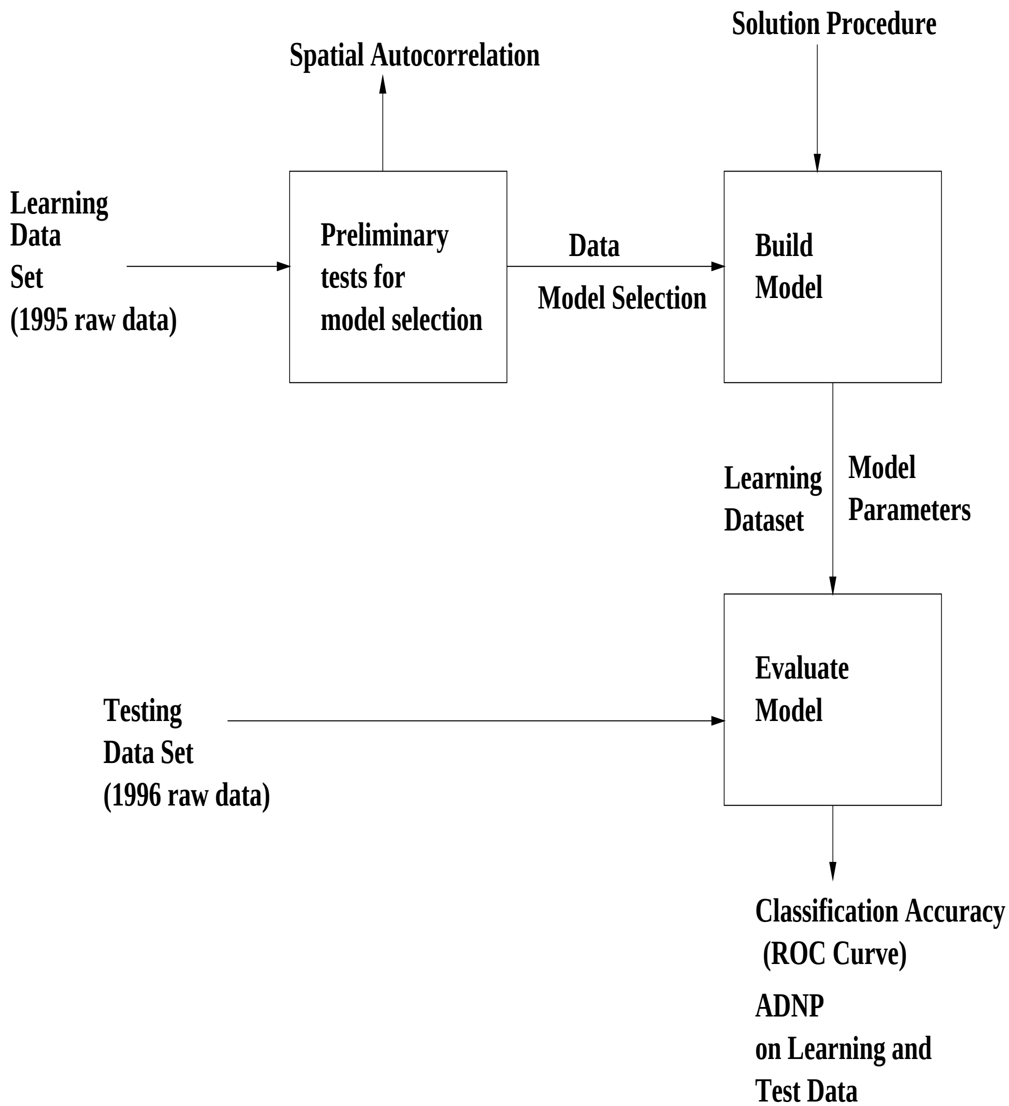


Figure 25: The data block diagram

Candidates

- Logistic regression w/o SA
  - *Software : matlab 5.2*
- Probit regression w/o SA
  - *Econometrics toolbox [leSage 99]*
- Probit with SA
  - *Econometrics toolbox [LeSage 99]*
- ADNP algorithm
  - *Implemented via Matlab 5.2*

Workload

- Download and adapt Econometrics toolbox
- Understand and modify existing functions
- Run different models with all variables in different combinations
- Wrote new code to implement new algorithm (ADNP)
- Wrote new code for our new measure (ADN)

Data set

- Data source
  - *Conservation Ecology*
  - *Marsh-breeding red-winged blackbird habitat [Uygar Ozesmi 1996]*
- Nest
  - *Binary dependent variable*
  - 0*: no nest site*
  - 1*: nest exists*
- Vegetation durability and stem density
  - *1x1 m quadrats*
- Vegetation height
  - *Measured to the nearest 5 cm*
- Water depth
  - *Recorded weekly with installed gauges*
  - *Metric is meter*
- Location index metrics
  - *Pixel unit*
  - *Each unit is 5 m*

Data Tables

- Data set
  - *1995 Darr location index file (matrix: 5372 x 2)*
  - *1995 attributes' values (matrix: 5372 x 8)*
  - *1995 Stubble location insex file (matrix: 1817 x 2)*
  - *1995 Stubble attributes' value (matrix file: 1817 x 8)*
  - *1996 darr location index file(matrix: 4921 x 2)*
  - *1996 attributes' values (matrix:4921 x 8)*

- *Attribute and Class Table*

Case_ID	Veg	StemD	StemH	DoW	DEdge	Nest(1/0)
1	10	100	5	10	10	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
200	10	40	5	12	10	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
5372	10	100	5	10	10	0

- *Location Index*

Case_ID	X-coord	Y-coord
1	2	3
⋮	⋮	⋮
5372	88	100

Results

- Spatial accurancy results

Data set		PLUMS(A)	Probit	Logit	SAR(with Probit)
Learning	spatial accuracy	16.90	47.16	47.19	13.96
Testing	spatial accuracy	19.19	41.43	41.5	19.30
Learning	Run-time(Seconds)	80	10	10	19420 [1]

Table 4: Learning time and spatial accuracies for learning and test data set

- Observations
  - *Given same model(w/o SA), PLUMS algorithm has better performance*
  - *Even w/o SA, our results are comparable to Probit with SA*
  - *SAR(probit with SA) takes two orders of magnitude more computation time relative to PLUMS(A)*

[1] 10,000 draws for Gibbs sampling, 1000 burn-outs

Results for learning models

- *The probability distribution of nest sites via ADNP*

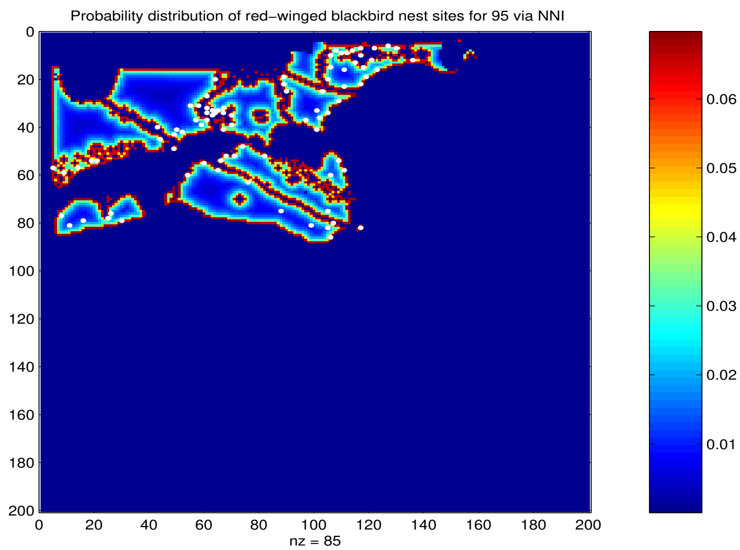


Figure 26: The probability distribution of nest sites via ADNP(95)

- *The probability distribution of nest sites via Probit*

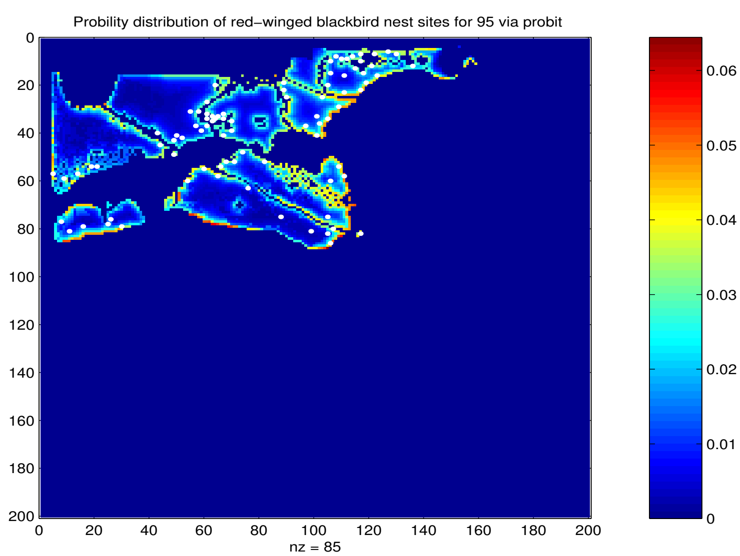


Figure 27: The probobility distribution of nest sites Probit

Results for testing models

- *The probability distribution of nest sites via ADNP*

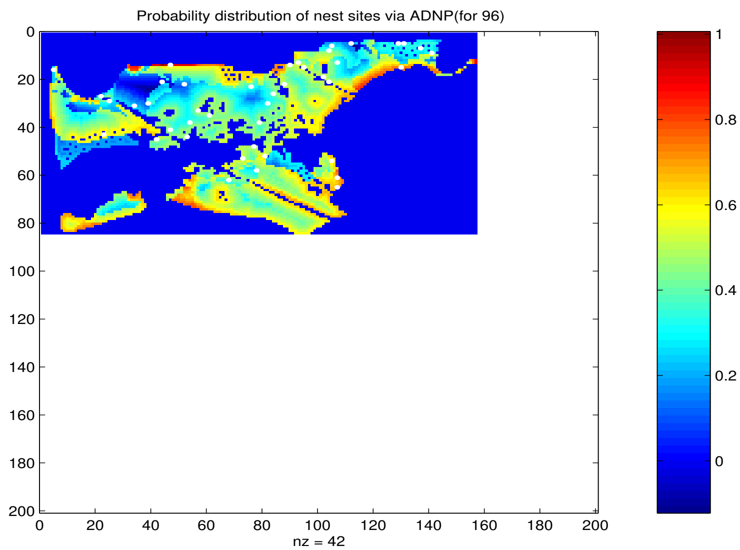


Figure 28: The probability distribution of nest sites via ADNP(95)

- *The probability distribution of nest sites via Probit*

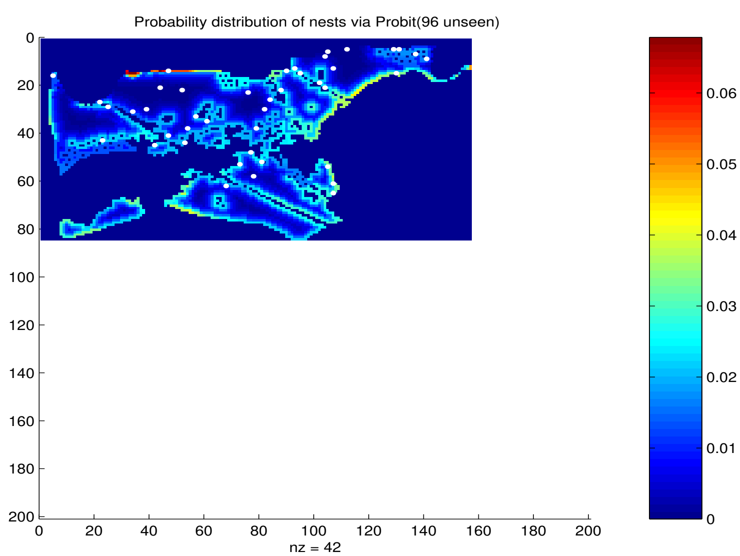


Figure 29: The probability distribution of nest sites via Probit(96)

Results for testing models via Stubble95

- *The probability distribution of nest sites via ADNP*

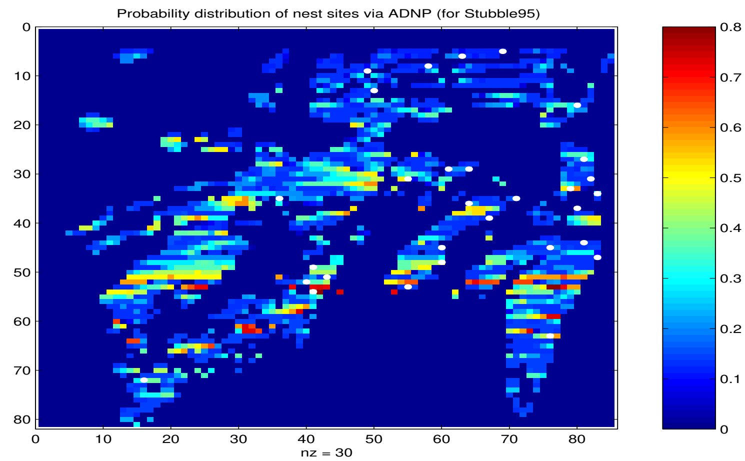


Figure 30: The probability distribution of nest sites via ADNP(Stubble95)

- *The probability distribution of nest sites via Probit*

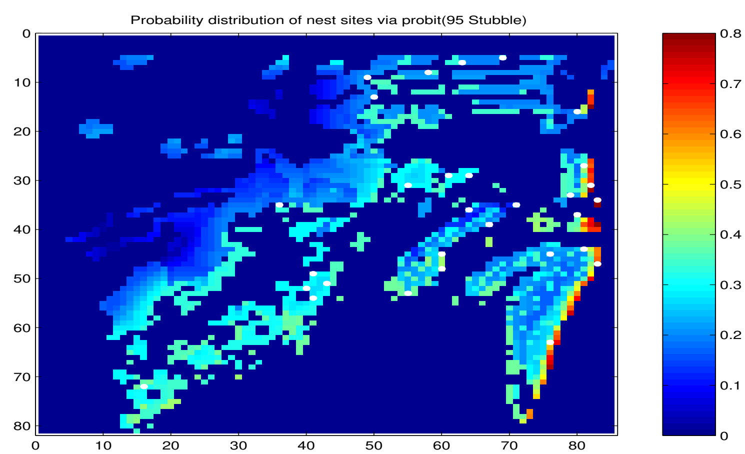


Figure 31: The probability distribution of nest sites Probit(Stubble95)

Results for testing models via Stubble95

- *The probability distribution of nest sites via Probit with SA*

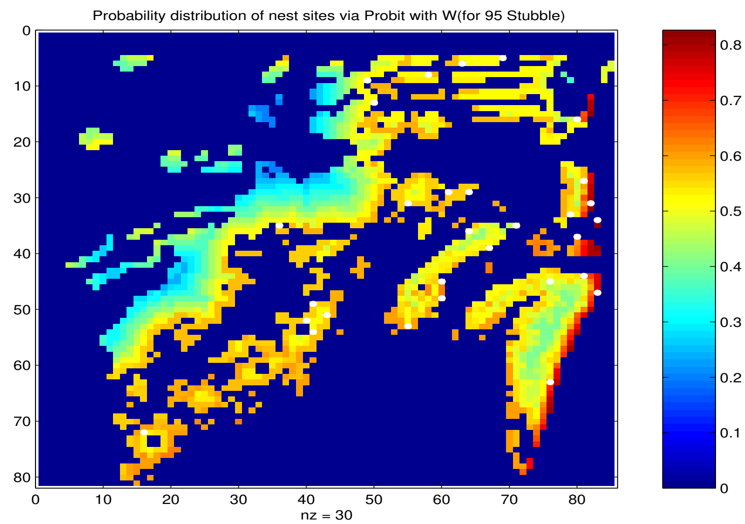


Figure 32: The probability distribution of nest sites via Probit with SA(Stubble95)

Conclusion and Future Works

- Conclusion
- Future work 1
  - *Better model*
  - *Better search algorithm*
- Future work 2
  - *New benchmark*
- Future work 3
  - *Formalize*
- Timetable

Conclusion

- Presented an overview of stat methods
- Proposed new algorithm (ADNP)
- Our approach significantly improves spatial classification accuracy compared to classical methods without using SA
  - *Better performance by two order of magnitude*
- Carried out experiments using data from "conservation ecology"
- Demonstrate the usefulness of spatial effects

Future Work 1

- Select better model
  - *Our current approach uses simple linear model w/o spatial autocorrelation*
  - *Select more complex nonlinear model*
  - *Factor spatial autocorrelation term into selected model*
  - *Quantify spatial autocorrelation*

- Better search algorithm
  - *Current approach uses simple greedy search*
  - *For huge geo-spatial data set, design efficient search algorithm*
  - *Design spatial join-index to efficient process our approach*
    - Search through relevant spatial join index
    - Avoid access to the spatial objects

- Sparse matrix computation

Future Work 2

- Design new benchmark
  - *Limitation of current benchmark*
    - Not take into account factor of zone influence
    - Scale sensitivity: Distance based measure are known to be sensitive to scale
  - *Ranking_measure*
    - ranking_measure = height * size_of_zone_influence
    - height = Pred_Location(β^0, X)
    - size_of_zone_influence = Num of pixels in zone of influence
- Main idea of the new measure
  - *For each pixel compute zone_of_influence(pixel);*
  - *size_influence_zone=count pixels in zone_of_influence;*
  - *rank_of_peak=size_influence_zone * L(pixel);*
  - *Sort pixels in descent order of rank_of_peak;*
  - *Take highest N pixels as pred centers*

Future Work 3

- Formalize

  - *Spatial data is a linear combination of Gaussian*

$$F = \sum_{i=1}^{c} w_i G(x_i, y_i, u_i, \sigma_i)$$

  - *Goal of SDM is to find the* w_i*(weights),* (x_i, y_i)*(centers),* μ_i*(mean),* σ_i*(variance)*
  - *Our method finds the centers*
  - *Goal of formalization is to make our assumptions explicit*

Future Works

- New Measures for map similarity
  - *Map similarity measure is a combination of classification accuracy and spatial accuracy*
  - *raster-vector conversion*
- Efficient Search Algorithms
  - *Preliminary implementation of PLUMS is based on greedy search of the parameter space*
  - *explore other search paradigms to speed up and to find solutions better than local maximas of map similarity*
  - *explore search algorithms which allow better exploration of parameter space*
- I/O Performance Tuning
  - *Many applications of spatial data mining are characterized by large datasets*
  - *redesign PLUMS to scale upto larger geo-spatial datasets*
  - *explore map compression techniques to further speed up computation of classification and spatial accuracy*
- Formalization
  - *Proposed a new objective of map similarity*
  - *evaluate the proposed objective from a theoretical view-point based on statistical learning theory*
- Experimental Evaluation