Hotspot Geoinformatics Software for Detection and Prioritization

Est. 15th Aug 2006

JalaSRI

Watershed Surveillance and Research Institute

Sponsored by

K.C.E. Society's

Moolji Jaitha College, Jalgaon

Software Development Team

1. Dr. Sharadchandra Joshi
2. Kishor Mahajan
3. Sanjay Pawde
4. Hemlta Patil
5. Rajesh Koli
6. Leena Nemade
7. Ekta Sapkale
8. Tushar Chaudhari
9. Umesh Pawar

Outline

- Introduction
- Need of Hotspot Detection
- Algorithm of ULS Detection
- Implementation details of ULS Scan Detection Software
- Case Study : Hotspot Detection for Credit Cooperative Societies, which are in vulnerable condition.
- Need of Prioritization
- Features of newly implemented Prioritization Software
- Prioritization algorithm
- Case Study : Prioritization and Ranking of Hotspots detected for Credit Cooperative Societies, which are in loss.
- Conclusions
- The Path Forward …!

Introduction

- Geoinformatic surveillance for spatial and temporal hotspot detection and prioritization is a critical need for the 21th century

Introduction (Contd…)

- What is a Hotspot ?

  A hotspot can mean an unusual

  - Phenomenon
  - Anomaly
  - Aberration
  - Outbreak
  - Elevated cluster
  - Critical area

Introduction (Contd…)

- The hotspot geoinformatic software has two methodological components-
  - Detection
  - Prioritization
- The goal of detection software is identifying, delineating and assessing the significance of hotspots
- The goal of the prioritization software is to canonically transform a partial order into a linear order of the objects.

Introduction (Contd…)

- This white paper presents and demonstrates the ***implementation of algorithms*** for detection – ULS Scan and for prioritization - Starter, Markov Chain Monte Carlo (MCMC), POSET 1R, POSET2R, Final Rank and it's data processing results

Need of Hotspot Detection

- Why Hotspot Detection ?

  The declared need may be for

  - Early warning
  - Monitoring
  - Etiology
  - Sustainable Management

ULS Scan Detection Software

- Martin Kullback and associates use circular scan statistics in their well known software for hotspot detection.

- The software being presented is based on the concept of Upper Level Set Scan Tree as proposed by ***Dr. G.P.Patil and Taillie*** for handling arbitrarily shaped cells.

- The ULS scan approach examines spatially connected components of a tessellation as a threshold and is moved from the highest level in the data to the lowest level.

- This version of the computer program that assumes the ***Binomial and Poisson Distribution Models*** and are implemented in **VC++** environment.

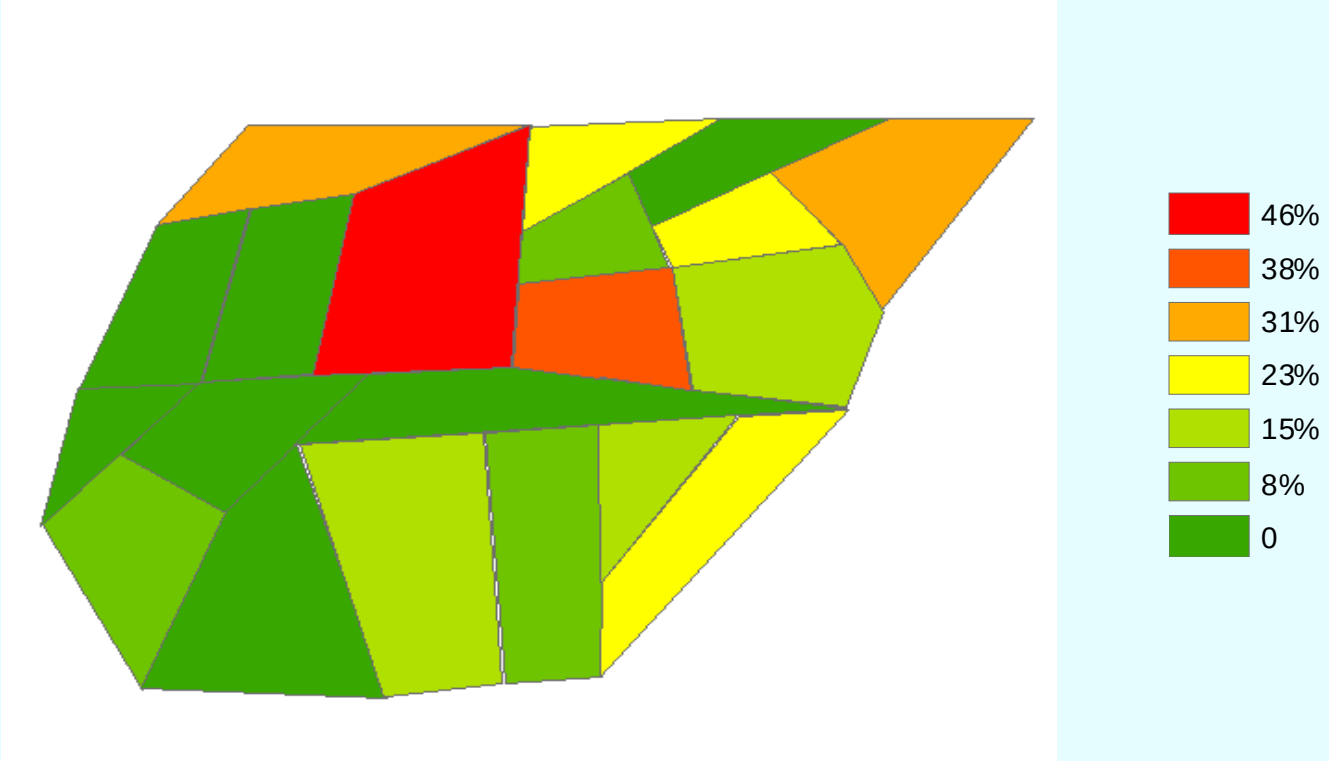
ULS Scan Detection Software (Contd…)

Assumptions –

- The region is partitioned into units called cells. We assume that 'size' of each cell is known
- For a given cell, ‘A’ represents its ‘size’ and ‘Y’ represents the ‘response’ for the cell.
- If one or more contiguous groups of cells or a connected zone is found to have significantly higher response, then it is called a hotspot.
- Hotspots are defined in terms of higher values of ‘Y’.

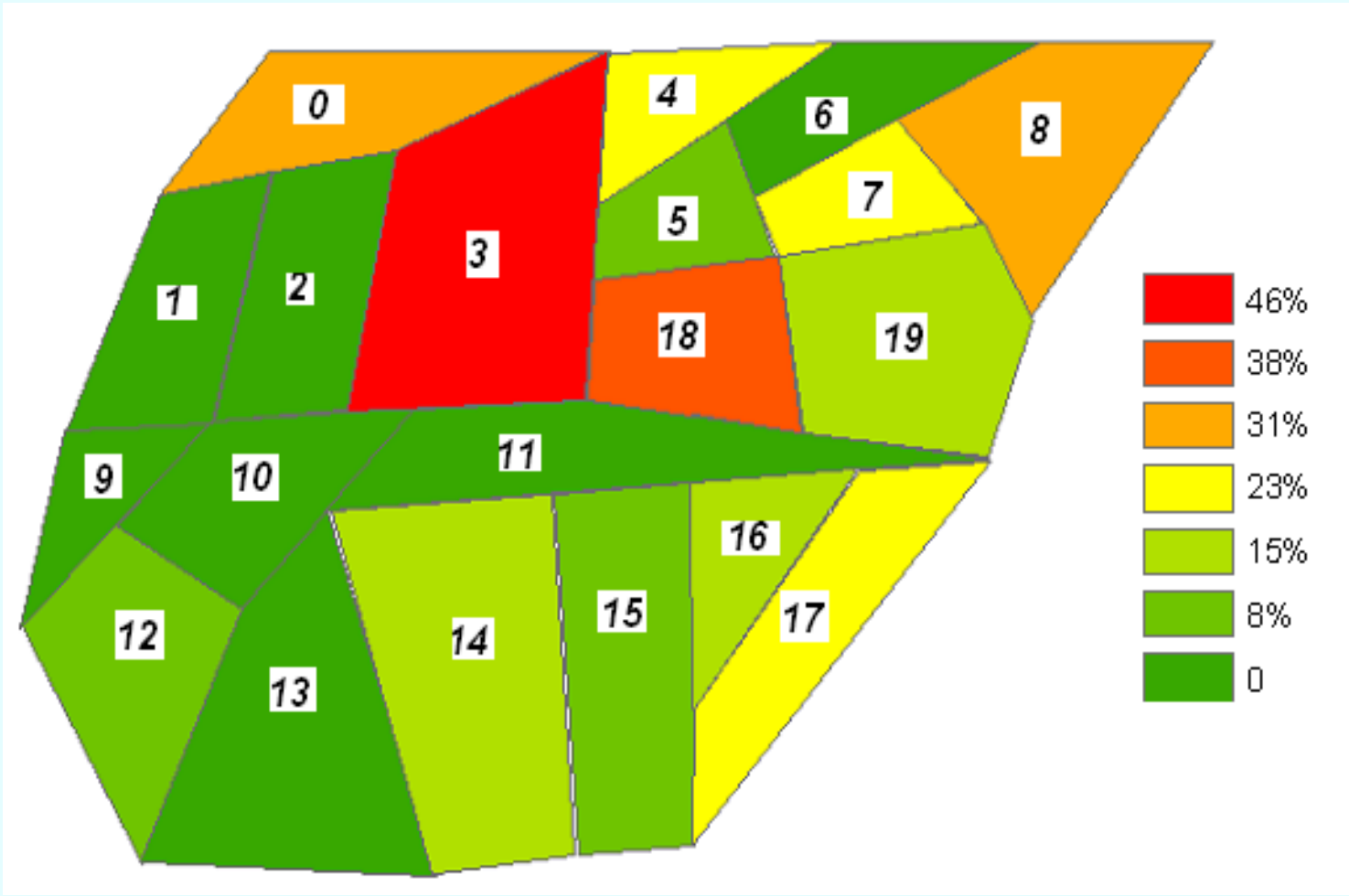
ULS Tree Example

Consider the region below. Response values shown in the legend are rounded off to the nearest integer. Thus response values for cells in the same category are not necessarily equal.



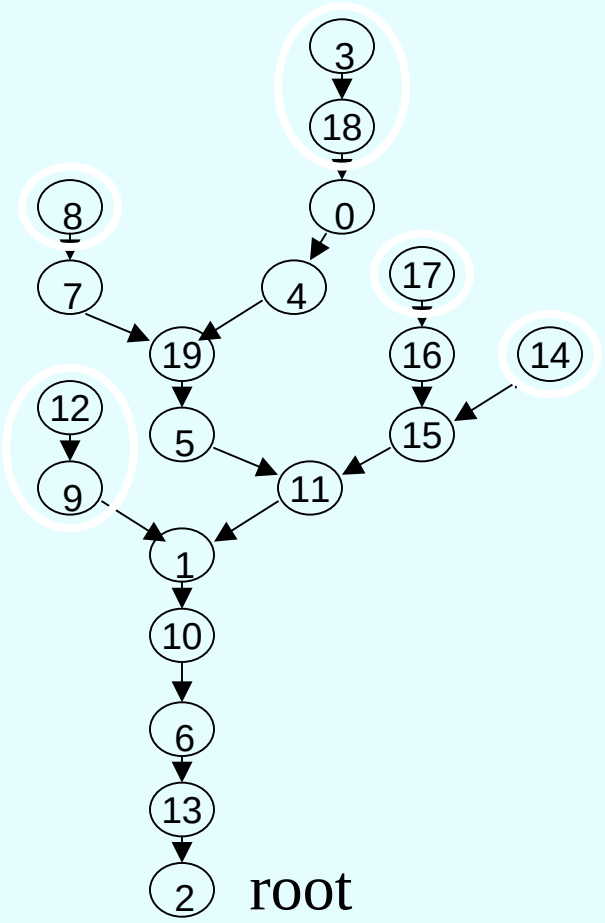
Example Continued

First we number cells for identification



Example Continued

Tree showing topological sort of cells



Algorithm for ULS Scan Detection Software

- **Main Steps**

1. Construction of ULS tree.
2. Finding the candidate zones.
3. Computation of zonal log likelihood.
4. Determination of hotspots.

- **Algorithm**

1. **START**
2. **READ** the input from the text files contains following information
   a) Cell ID b) Cell size c) Cell response d) Cell neighborhoods
3. **Construct the ULS tree with following steps (For observed data)**
   a) **CALCULATE ,** level g = Y/A for each cell , where A is the cell size and Y is the cell response
   b) **SELECT** a cell that has maximum g.
   c) **PUT** this cell in tree as a node.

Algorithm for ULS Scan Detection Software (Cont…)

4. **SELECT** the next Cell ID that has maximum g from remaining one, **IF** this selected cell is connected to cells from the tree already built, **THEN** combine newly selected cell and previously selected cells together to form one zone.
5. **PUT** this zone in the tree with connection to connected cell node of the tree.
6. **IF** there is no connected cell to newly selected cell, that is present in already generated tree, **THEN** put this cell in tree as a separate node.
7. **IF** there are cells remaining in the list, repeat steps 4 to 6.
8. **FIND** Candidate zones of ULS tree
   a) **Start** from bottom of ULS tree
   b) **Select** the Last node of tree as zone, **if** zone size is greater than predetermined default size of 0.5, **then** expose this one.
   c) **Select** next upper level zone and **do** steps b
   d) **Display** the candidate zones

Algorithm for ULS Scan Detection Software (Cont…)

9. COMPUTE the zonal log likelihood

Each candidate zone is rated by the logarithm of its likelihood. The zone with the maximum likelihood is the maximum likelihood estimate of the region.

a) To determine if a candidate zone is a hotspot or not, we generate 999 simulated maximized likelihood values, for the zone , and one from the observed data.

b) These 999 maximized likelihood values are compared with the likelihood of each of the candidate zones computed from the observed data.

10. COMPUTE the Hotspot

a) **If** candidate zone has n out of 1000 maximized likelihoods which are equal or greater than maximized likelihood values of the candidate zone, **then** compute p-value = (n/1000)

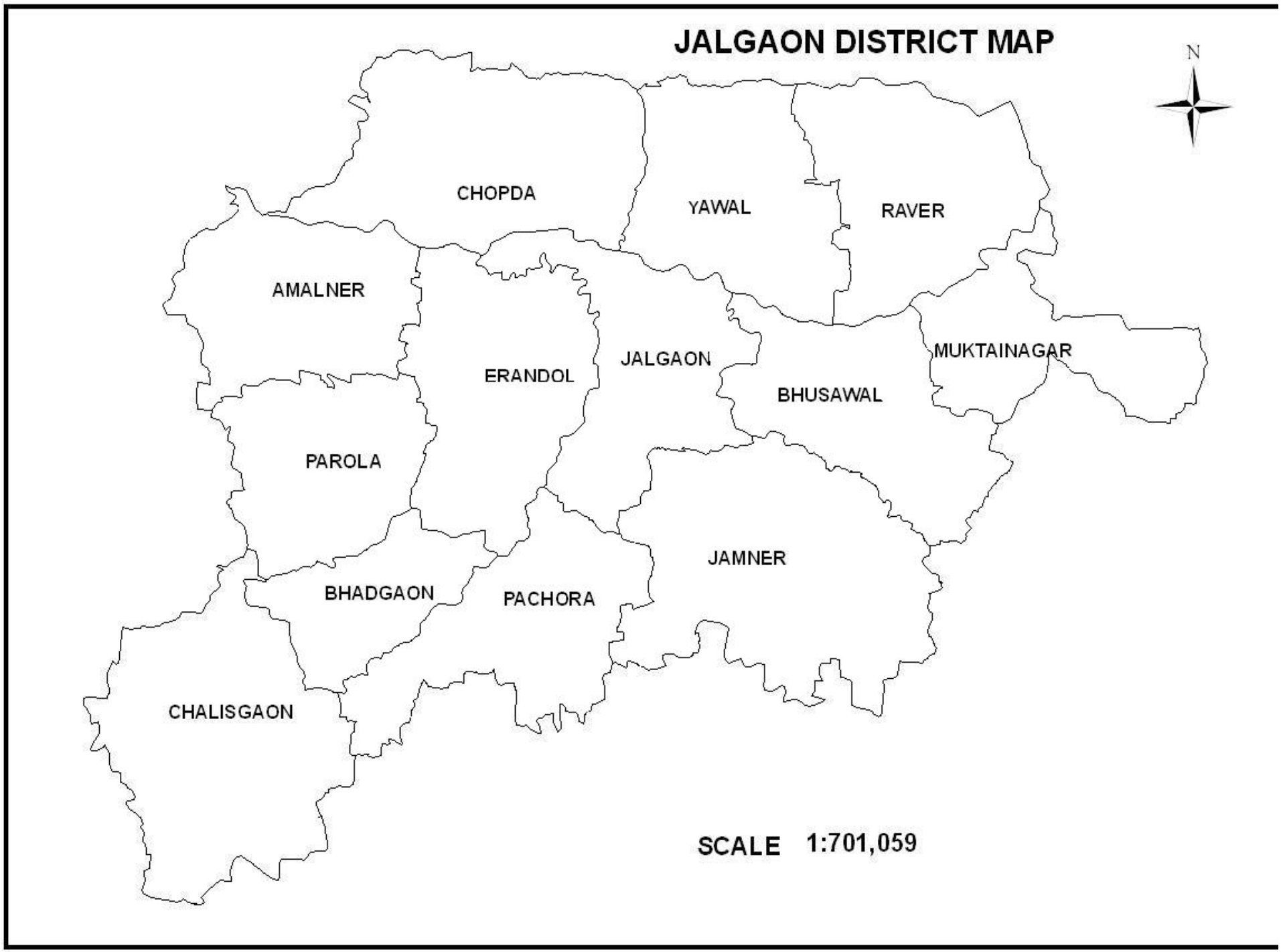
b) **If** p-value is less than or equal to level of significance i.e. 0.05, **then** the candidate zone is a Hotspot

11. **END**

Case study – To detect the hotspots in the Jalgaon district of Maharashtra state at taluka level, for the cooperative societies, which are in vulnerable condition for the year 2005-2006

Information about Jalgaon district -

- Jalgaon district is located in the northwest region of the state of Maharashtra. It is bounded by Satpuda mountain ranges in the north, Ajanta hill ranges in the south, Dhule district in the west and Buldhana district in the east
- Jalgaon is rich in volcanic soil, which is well suited for cotton, banana and sugar cane production. It is a major business center for tea, gold and pulses
- With an area of about 11,700 sq km, Jalgaon District has a population of more than 3.5 Million


JALGAON DISTRICT MAP
N
CHOPDA
YAWAL
RAVER
AMALNER
ERANDOL
JALGAON
MUKTAINAGAR
BHUSAWAL
PAROLA
JAMNER
BHADGAON
PACHORA
CHALISGAON
SCALE 1:701,059

Data matrix for the Credit Cooperative Societies in Jalgaon district

(Amount in Rs Millions)

Sr. No.	Taluka	No. of Co-op. Societies	No. Of Co-op. Cr. Societies in Loss	Deposit Amount	Loan Amount	Overdue Amount	Loss Making Societies (%)	Credit To Deposit Ratio (%)	Ratio of Overdue Amount to Loan Amount (%)
0	Amalner	53	22	4348.6	252.8	145.8	41.50943	58.14053	57.68292152
1	Bhadgaon	28	15	188.0	177.5	496.3	53.57143	94.40718	27.96799279
2	Bhusawal Bodwad	126	28	4053.7	2834.3	716.5	22.22222	69.91854	25.27919444
3	Chalisgaon	45	8	410.1	203.4	67.5	17.77778	49.60136	33.15850529
4	Chopda	59	7	1675.1	1410.1	91.5	11.86441	84.17836	6.489814212
5	Erandol Dharangaon	54	20	387.3	186.3	126.1	37.03704	48.09879	67.69048232
6	Jalgaon	207	20	3019.0	1425.4	409.9	9.661836	47.21439	28.75519499
7	Jamner	47	7	1235.5	555.4	161.5	14.89362	44.95655	29.07799886
8	Muktainagar	20	10	179.6	142.6	25.8	50	79.30459	18.11995791
9	Pachora	57	30	633.7	337.4	28.4	52.63158	53.28989	8.414588606
10	Parola	57	22	645.2	567.9	248.2	38.59649	88.00698	43.70987113
11	Raver	67	17	1534.7	1234.4	282.6	25.37313	80.42887	22.89392543
12	Yawal	64	23	1300.8	903.5	60.3	35.9375	69.45912	6.674133775

Attributes selected for Hotspot Detection

- Sr. No (Cell ID Number)
- Total Number of Credit Cooperative Societies
- Credit Cooperative Societies in Loss
- Neighborhoods

Attribute Table & Values

Cell ID	Taluka	Total No Of Co-op Cr. Societies	Total No Of Co-op Cr. Societies In Loss	Neighborhoods
0	Amalner	53	22	4,5,10
1	Bhadgaon	28	15	3,5,9,10
2	Bhu-Bod	126	28	6,7,8,11,12
3	Chalisgaon	45	8	1,9,10
4	Chopda	59	7	0,5,7,12
5	Dha-Era	54	20	0,1,4,7,9,10
6	Edlabad	20	10	2,11
7	Jalgaon	207	20	2,4,5,8,9,12
8	Jamner	47	7	2,7,9
9	Pachora	57	30	1,3,5,7,8
10	Parola	57	22	0,1,3,5
11	Raver	67	17	2,6,12
12	Yawal	64	23	2,4,7,11

Output of Hotspot Detection Software

- Total number of cases: 229
- Total population: 884
- p under the null hypothesis: 0.259050
- Log likelihood under the null hypothesis: -1011.403250

CELL ID	HOTSPOT NO.	LIKELIHOOD
10091 5	0 0 0 0 0	54.471602

Geoinformatics Hotspot Detection Software Output

Upper Level Set Tree Scan Program Report

Hotspot Type: High Values

Input File Name: C:\CREDITSOCIETY.TXT

Maximum Zone Size as percentage of region size:50.000

Number of cells in the region: 13

Total size (area or population) of the region: 884.000000

Total response of the region: 229.000000

Model Used: Binomial

Maximum Likelihood Estimate, pNull, of p: 0.259050

Global log-Likelihood under null hypothesis:-1011.403250

ULS Tree Levels

Level	g-value	Exposed Size Total	Exposed Response Total	Member Zones
1	0.5	28.0	15.0	0
2	0.5	85.0	45.0	1
3	0.5	105.0	55.0	2
4	0.4	158.0	77.0	3
5	0.4	215.0	99.0	4
6	0.4	269.0	119.0	5
7	0.4	333.0	142.0	6
8	0.3	400.0	159.0	7
9	0.2	526.0	187.0	8
10	0.2	571.0	195.0	9
11	0.1	618.0	202.0	10
12	0.1	677.0	209.0	11
13	0.1	884.0	229.0	12

Candidate zones

Zone#	g value	Member Count	Size Total	Response Total	log-likeli-hood ratio	p value
5	0.4	5	249.0	109.0	54.4716	0.001
4	0.4	4	195.0	89.0	46.9493	0.001
9	0.2	6	294.0	117.0	42.7111	0.001
1	0.5	2	85.0	45.0	31.5435	0.001
0	0.5	1	28.0	15.0	10.0122	0.082
3	0.4	1	53.0	22.0	6.5159	0.292
2	0.5	1	20.0	10.0	5.42595	0.412
7	0.3	3	151.0	50.0	4.72838	0.505
6	0.4	1	64.0	23.0	3.39149	0.751
8	0.2	4	277.0	78.0	1.05745	0.995
10	0.1	11	618.0	202.0	-1e+010	1.000
11	0.1	12	677.0	209.0	-1e+010	1.000
12	0.1	13	884.0	229.0	-1e+010	1.000

Output of Hotspot Detection Software (cont…)

ULS likelihood values with actual

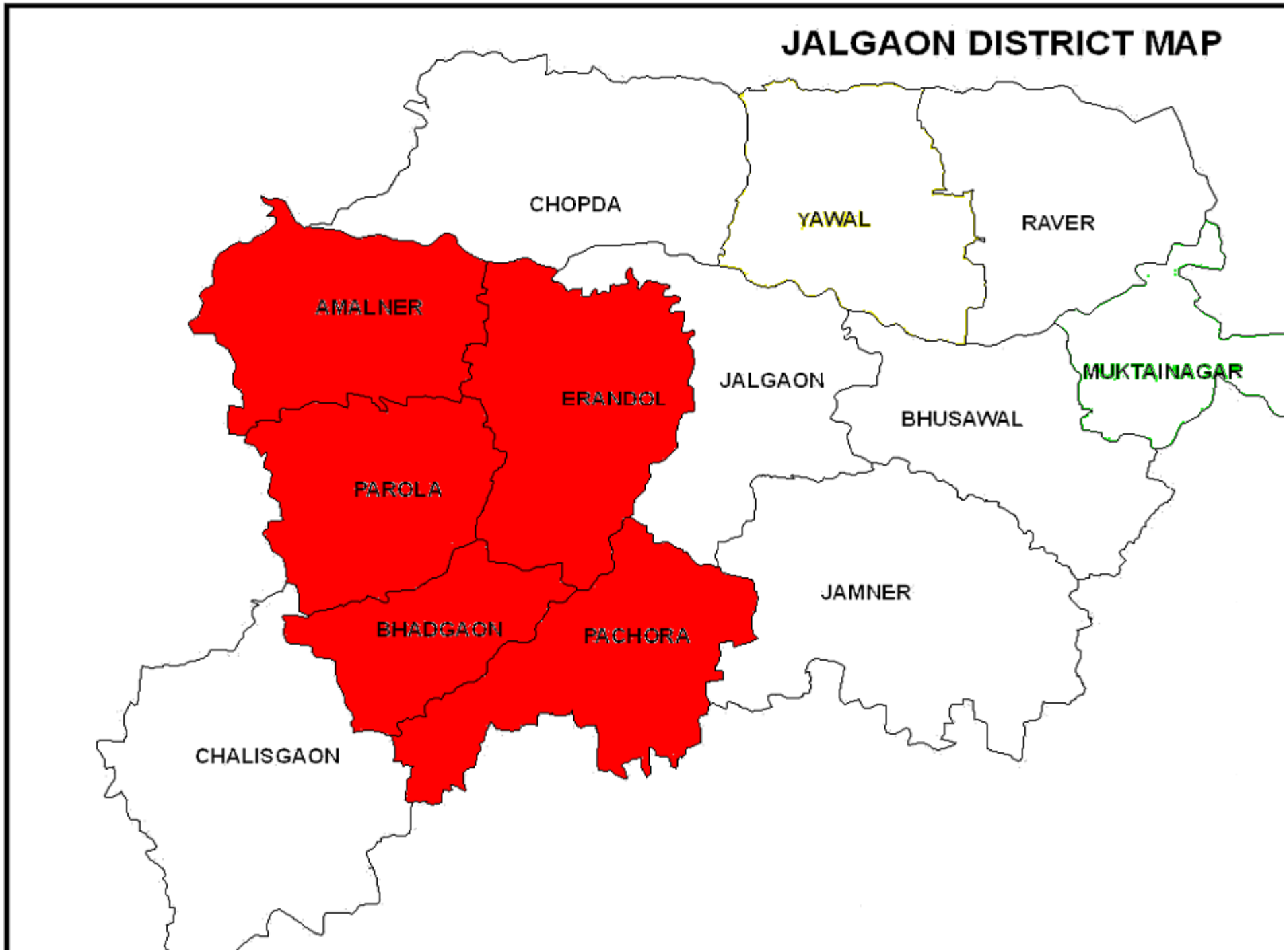
0	10.012220
1	31.543474
2	5.425945
3	6.515903
4	46.949280
5	**54.471602**
6	3.391494
7	4.728377
8	1.057450
9	42.711067
10	-1.000000
11	-1.000000
12	-1.000000

Performing 1000 simulations

MLE for real data is 54.471602

MLE for simulations

pvalue for **hotspot 0** is 0.001


JALGAON DISTRICT MAP
CHOPDA
YAWAL
RAVER
AMALNER
ERANDOL
JALGAON
MUKTAINAGAR
BHUSAWAL
PAROLA
JAMNER
BHADGAON
PACHORA
CHALISGAON

Need of Prioritization

- Once we identify hotspots, we want to know, which hotspots need monitoring, which ones can be used for analysis, which ones are ready for some assessment and management, so we want to prioritize and rank these hotspots.

Prioritization Software

- A conventional solution is to assign a composite numerical score to each hotspot by combining the indicator information in some fashion.
- Rather than trying to combine indicators, we take the view that the relative positions in indicator space determine only a partial ordering and that a given pair of hotspots may not be inherently comparable.
- Using partial order sets, rank frequency distributions, and cumulative rank frequency operators, we accomplish the desired prioritization of the hotspots.

Prioritization Software (contd..)

Prioritization software contains following computer programs –

- **STARTER :**It computes ZETA Matrix and generates first linear extension.

- **MCMC :** It samples the linear extensions for data matrix with more than a few objects. It computes the Cumulative Rank Frequency (CRF) .

- **POSET 1R :** It checks whether the CRF are stacked or not.

Prioritization Software (Contd….)

- **POSET 2R :** It is logically same as MCMC program, only the difference is that the input to POSET 2R program is the output of POSET 1R,. if more than 10 million iterations are required then this program samples the linear extensions again.

- **Final Rank :** It computes the final ranks of an objects. It is implemented in R-Script. The R script is an interpreted computer language, which allows branching, looping as well as modular programming using functions.

Features of Prioritization Software

- All the programs are uniformly implemented in Visual C++
- The nature of programming follows an object oriented approach.
- There is uniform and single window GUI for all the programs.
- One can execute any program randomly by using Tab Press Event. There is no need to execute each option(program) in sequence.
- Due to VC++ environment, the numbers of source lines in each individual program are less and the code is easy to manage and change.
- The output of all the programs are visible simultaneously, for the purpose of comparison and study.

Algorithm of Prioritization Program

1. **START**
2. **INPUT** data matrix.
3. Generate the ***Zeta matrix,*** from data matrix.

   IF x is one object and y is another, **THEN**, evaluate comparison between them such that, **IF** all corresponding attributes of x are greater than or equal to corresponding attributes of y, **THEN** put 1 at the location (x, y) in zeta matrix, **ELSE** put 0.
4. Using ***H-method*** create a ***first linear extension*** that work as starter to ***Markov Chain Monte Carlo program (MCMC)***.
5. Create linear extensions using ***Markov Chain Monte Carlo***.

Algorithm of Prioritization Program (Cont…)

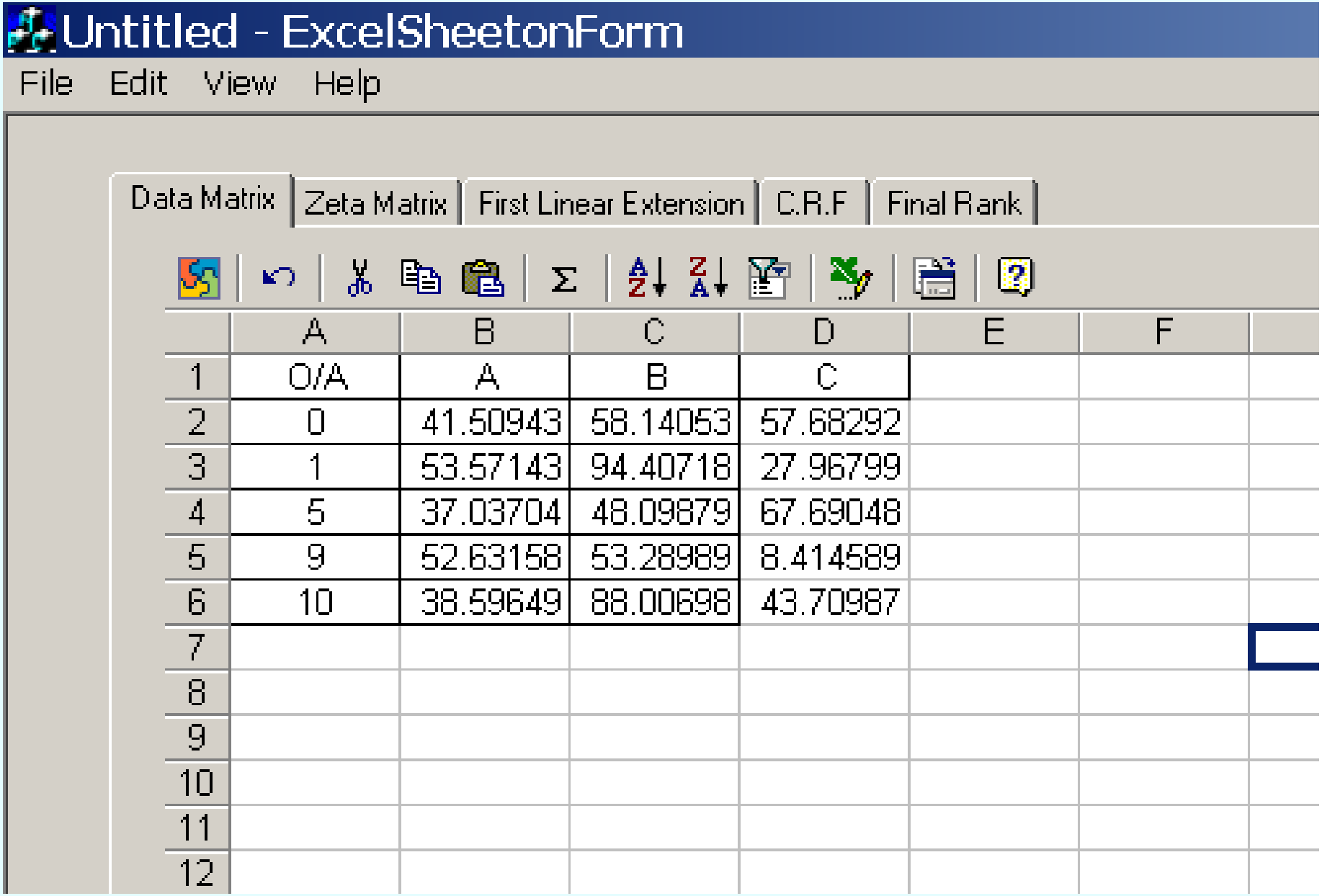
6. **CHECK** whether CRF created are ***STACKED*** or not, using module ***POSET1R***..
7. **IF** it is not STACKD successfully, **THEN** execute ***POSET2R*** and **REPEAT** step 7 again.
8. **IF STACKD** successfully, **THEN** compute ***FINAL RANK***
9. **END**

Case Study - To prioritize and rank the hotspots detected in the Jalgaon district of Maharashtra state at taluka level for the cooperative societies, which are in vulnerable condition for the year 2005-2006

- **Attributes selected for Hotspot Prioritization**

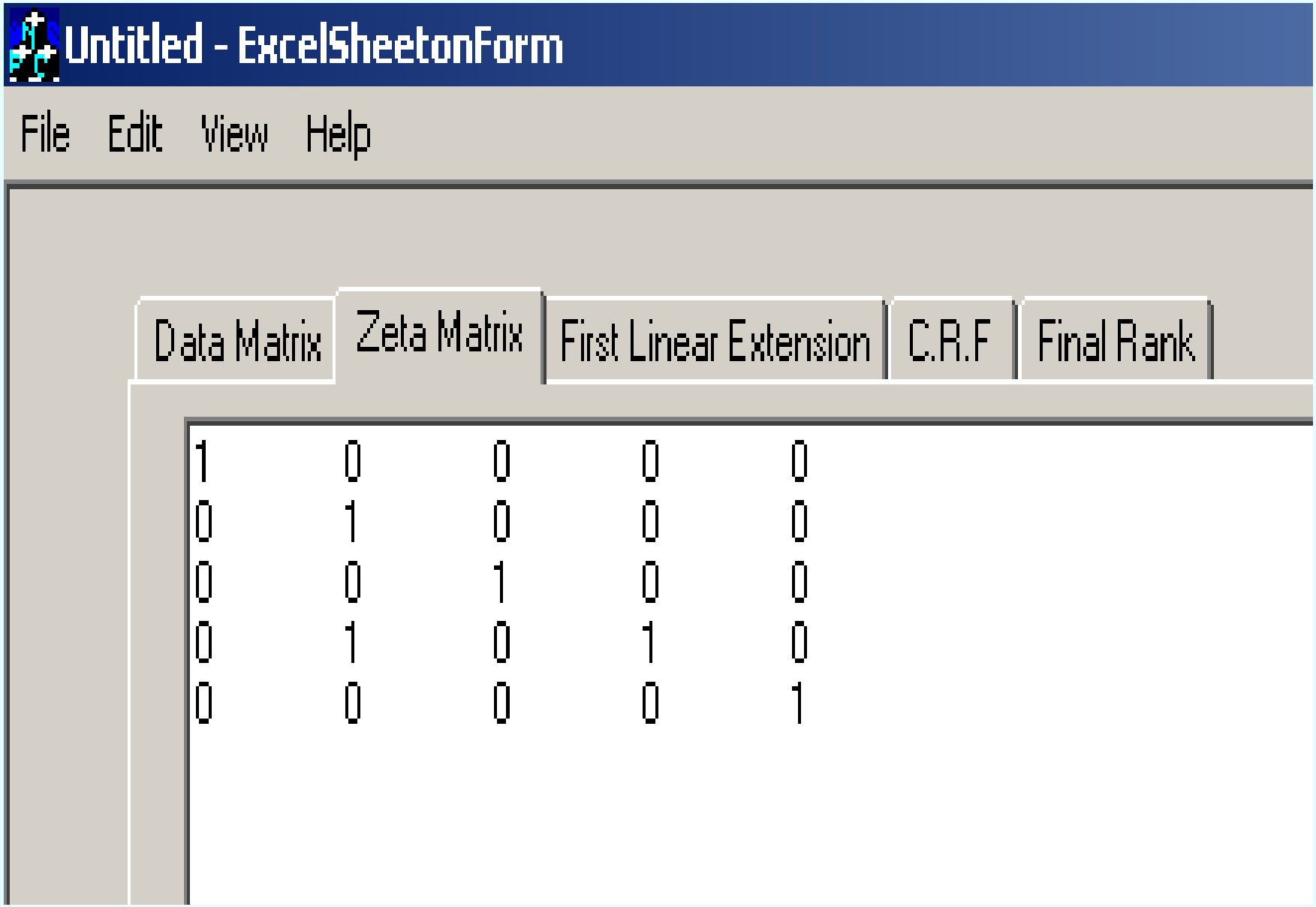
Cell ID	Loss Making Societies (%)	Credit To Deposit Ratio (%)	Ratio of Overdue Amount to Loan Amount
0	41.5094339622642	58.1405296018	57.682921
1	53.5714285714286	94.4071754986	27.967992
5	37.037037037037	48.0987881307	67.690482
9	52.6315789473684	53.2898903448	8.414588
10	38.5964912280702	88.0069803267	43.709871

Snapshot - Input Data Matrix

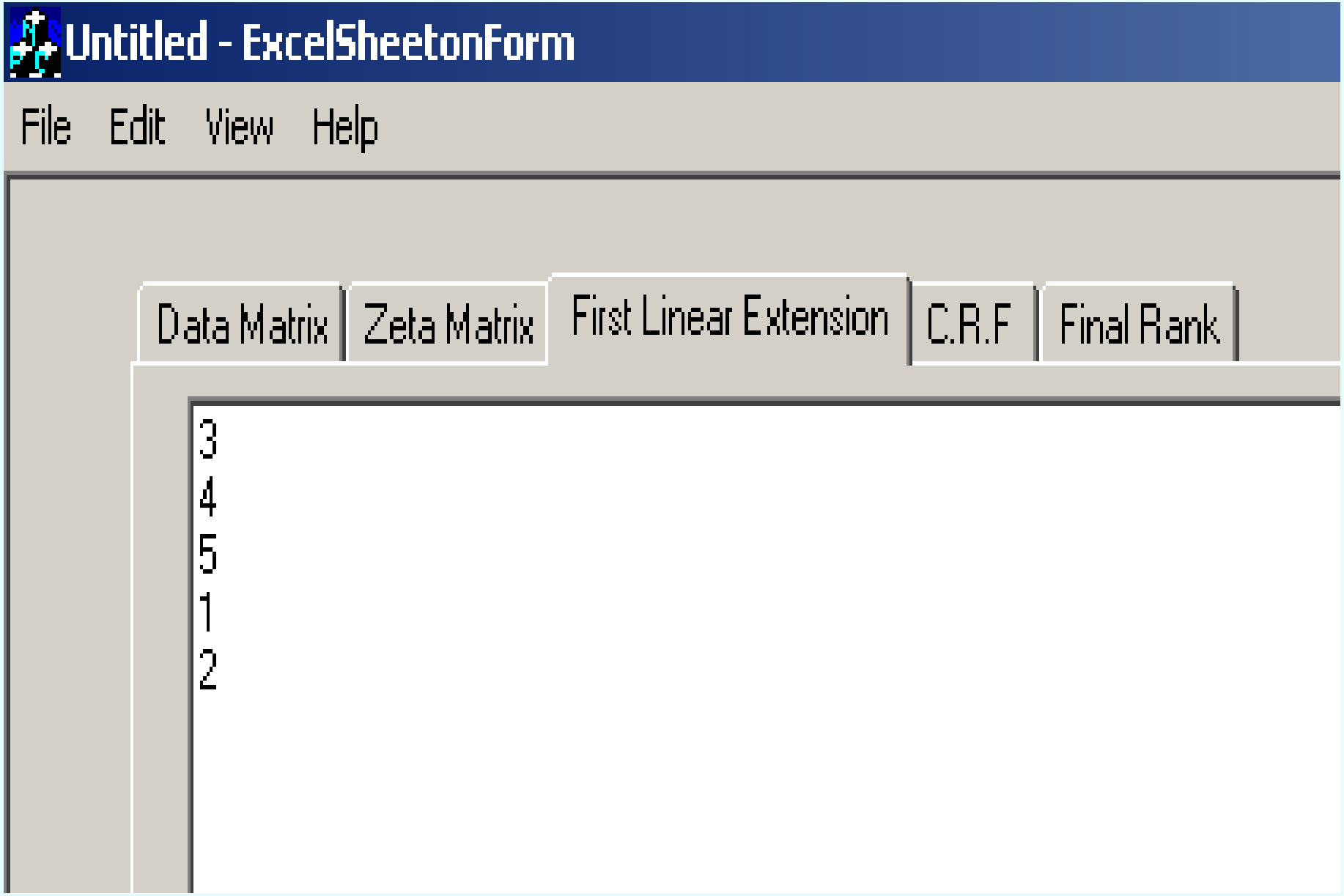


	A	B	C	D	E	F
1	O/A	A	B	C		
2	0	41.50943	58.14053	57.68292		
3	1	53.57143	94.40718	27.96799		
4	5	37.03704	48.09879	67.69048		
5	9	52.63158	53.28989	8.414589		
6	10	38.59649	88.00698	43.70987		
7						
8						
9						
10						
11						
12						

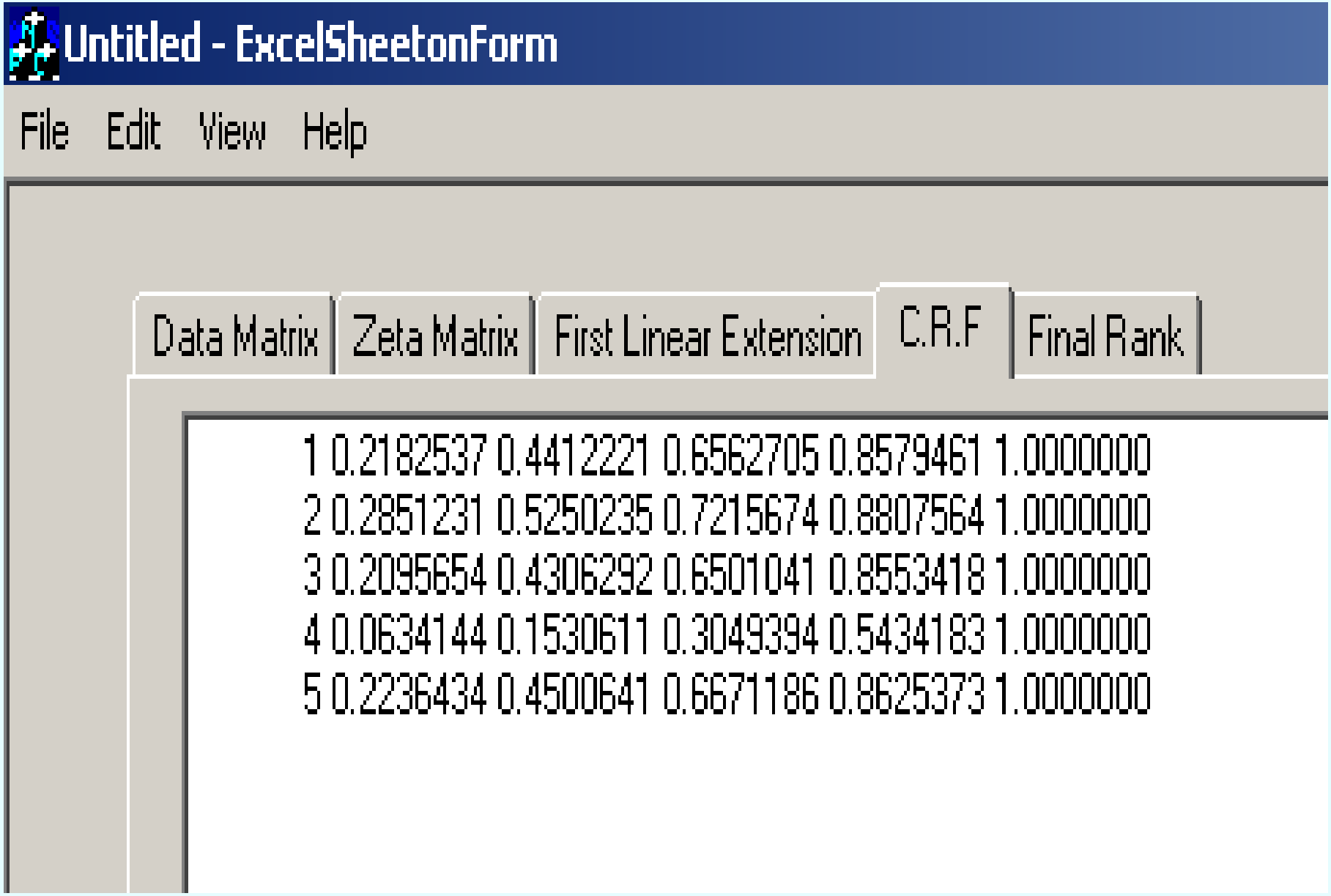
Snapshot - Resulting Zeta Matrix



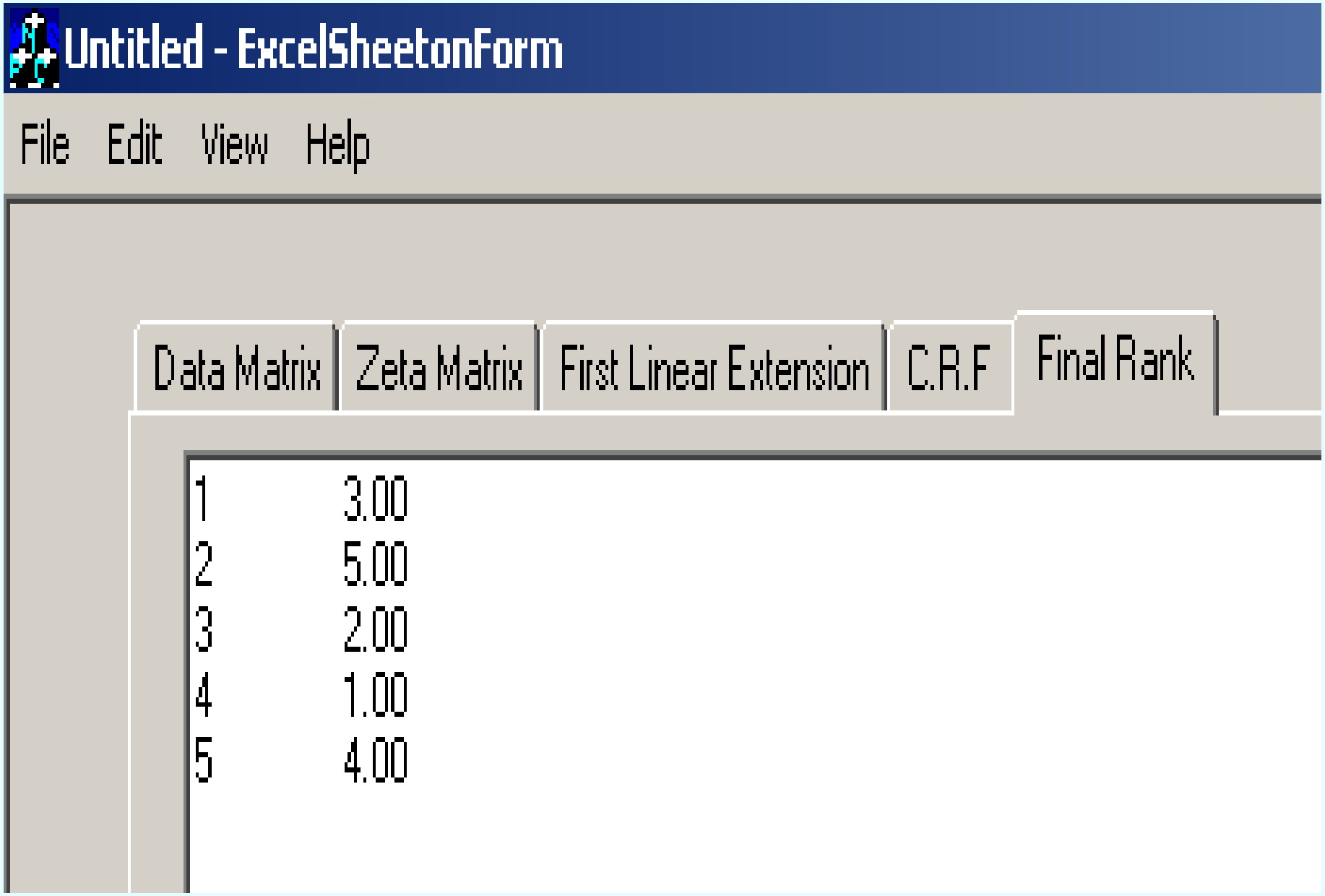
Snapshot - Resulting First Linear Extension



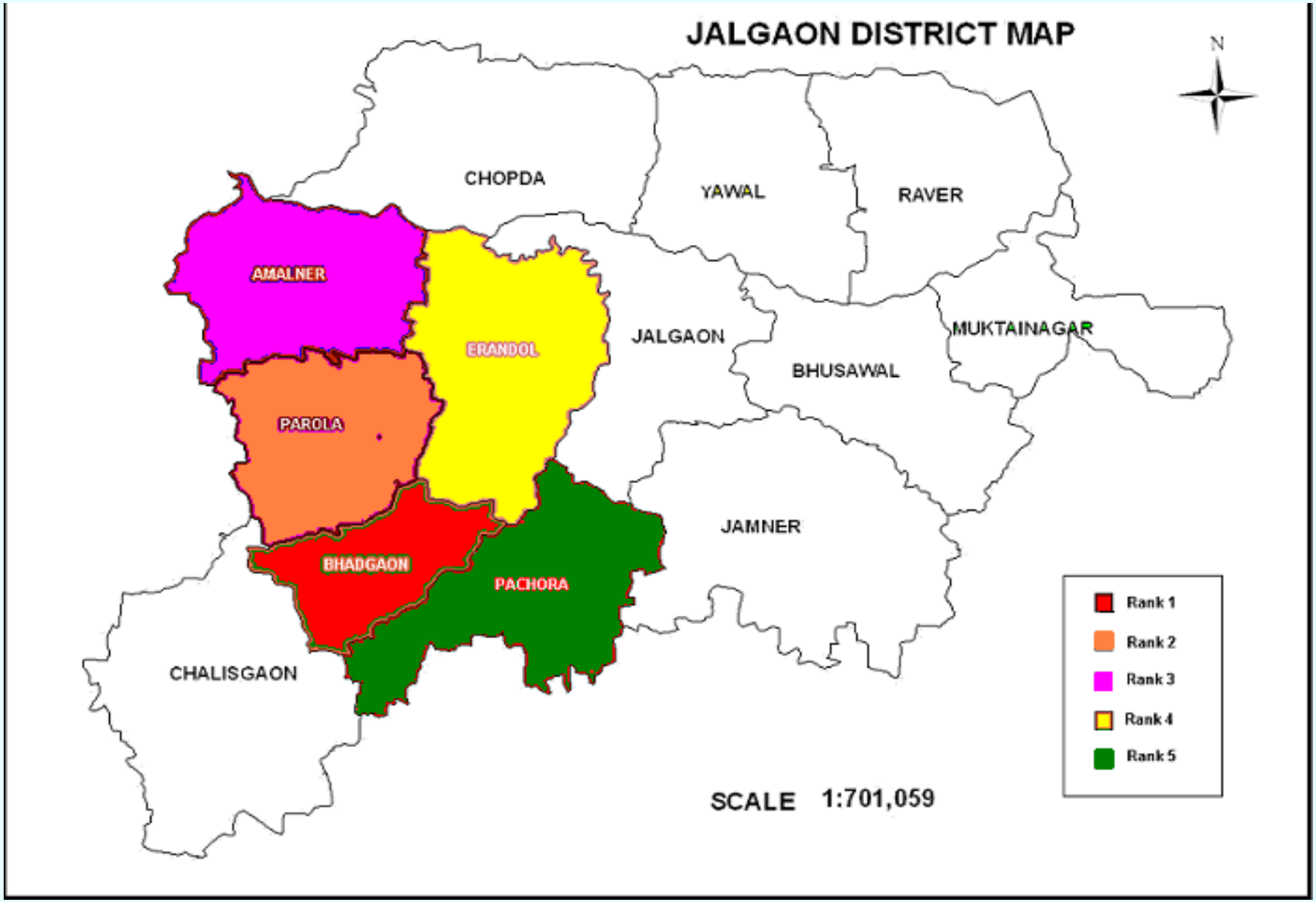
Snapshot - Resulting C.R.F.



Snapshot - Resulting Final Ranks



Prioritized Hotspots



Conclusion

- We have demonstrated through a computer algorithm and computer software, how the tree structure of ULS is formed and how the candidate zones are detected.

- We have illustrated a **case study for application of ULS Detection algorithm** for detection of hotspots in the Jalgaon district of Maharashtra State at 'taluka level' for the cooperative societies, which are in vulnerable condition. (2005-2006)

- We have presented computer algorithms and its implementation for prioritization scheme based on multiple indicator and stakeholder criteria without having to integrate indicators into an index, using Hasse diagrams and partially ordered sets. It is accordingly called ***Poset Prioritization and Ranking System***

Conclusion (Cont..)

- We have illustrated a **case study for application of Prioritization algorithm** for prioritization and ranking of hotspots in the Jalgaon District of Maharashtra State at 'taluka level' for the cooperative societies, which are in vulnerable condition. (2005-2006)

The Path Forward…!

- We propose a cross-disciplinary collaboration to further extend, design and build the complete prototype system for geoinformatic hotspot **detection** and **prioritization software**.

- Further extensions to prioritization software will ***be the implementation of computer algorithms,*** *which* will facilitate the application of **POSAC, MPOSAC and METEOR**.

- ***Single Window Interface*** to the integrated and customizable software modules comprising of Detection , Prioritization and Ranking programs implemented on relevant algorithms discussed.

The Path Forward…! (Cont….)

- With increasing importance of Internet one should also think how POSETIC information can quickly transferred via Internet to other user. A standardized format needs to be defined for an easy exchange.
- Furthermore, with increasing importance and availability of MATLAB and software package R, a interface is needed to be clarified, as to how there can be established a good communication platform.
- The ***complete methodological toolbox and the software toolkit*** thus developed will support and leverage core missions of several agencies as well as their interactive counterparts in the society .

Acknowledgements

Junior authors acknowledge the guidance of Dr. G. P. Patil, Pennsylvania State University; University Park; PA 16802-2112, Dr.S.W.Joshi, Slippery Rock University, PA , Dr.Rainer Bruggumann, Professor, Leibniz Institute of Freshwater Ecology, Berlin, Germany, Khandesh College Education society, Jalgaon and Researchers Team of JalaSRI , M. J.College, Jalgaon, Maharashtra, India.

References

[1] PATIL, G.P., AND TAILLIE, C. Upper level set scan statistic for detecting arbitrarily Shaped hotspots. *Environmental and Ecological Statistics, 11* (2004), 183-197.

[2] PATIL, G.P., AND TAILLIE, C. *Multiple indicators, partially ordered sets, and linear extensions: Multi-criterion ranking and prioritization. Environmental and Ecological Statistics 11* (2004), 199-228.

[3] PATIL, G.P, P. PATANKAR, R.ACHARYA, S.L. RATHBUN, C.TAILLIE and REZAMODARRES *Geoinformatic Surveillance of Hotspot Detection, Prioritization and Early Warning*

[4] PATIL,G.P.,"IT research and collaboration tools: Geoinformatic surveillance: Hotspot Detection and Prioritization across geographic regions and networks for digital government in the 21st century."

References (Cont…)

[5] PATIL, G. P., and TAILLIE, C. 199-228 Multiple Indicators, Partially Ordered Sets, and Linear Extensions: Multi-criterion Ranking and Prioritization.

[6] MYERS WL, KURIHARA K, PATIL GP, AND VRANEY R (2002). Finding upper level sets in cellular surface data using echelons and SaTScan

[7] PATIL,G.P., AND TAILLIE C (2002). Multiple indicators, partially ordered sets, and linear extensions: multi-criterion ranking methods. Invited Paper, International Environ metrics Society, Geneva.

[8] PATIL, G.P, RAJ ACHARAYA, AMY GLASMEIER, WAYNE MYERS, SHASHI PHOHA, AND STEPHEN RATHBUN. Hotspot Detection And Prioritization Geoinformatics For Digital Governance.

References (Cont…)

[9] PATIL, G.P, S.W. JOSHI, AND S.L. RATHBUN – Hotspot Geoinformatics, Environmental Risk, And Digital Governance

[10] Web Resources
http://www.stat.psu.edu/~gpp/Powerpoint/PosePrioritization

Thank You !