Growing Gini Areas - all pairs below threshold gini index combined

Ranjana Rajendran
University of California, Santa Cruz
ranjana@soe.ucsc.edu

Summary of the Algorithm

Each grid cell is paired with each of its four neighboring grid cells and the gini index calculated for each such pair. The neighbor grid cell with which the given grid cell gives the lowest gini index is its best neighbour and for each grid cell its best neighbor and the gini index of the pair is recorded. For each iteration, all pairs of discrete grid cells with gini index below a threshold are combined to form a gini area. As a precondition for the algorithm, each grid cell individually constitute a gini area. Over subsequent iterations, as grid cells combine, the gini areas expand to constitute more than one grid cell and the best neighbor of a gini area is calculated by computing the gini index over the grid cells constituting the given gini area and each of the neighboring gini area of the given gini area. The algorithm stops when there are no more pairs of gini areas which when combined give a gini index below the threshold. The resulting list of gini areas are sorted by the following 3 methods and their precision and recall calculated taking a ground truth obtained by manually marking a rectangular region around Seattle. The 3 methods of sorting the gini areas implemented are (1) Gini index of the individual gini areas (2) The difference in average between the points inside each gini area and the average in a single belt of neighboring cells just outside the gini area (3) The sum of the absolute value of the difference between the mean of the values inside the gini area and the value in each individual cell in a belt of cells outside the gini area.

1. Dataset analysis and preparation

Mean annual air temperature over North America was used for this experiment. Annual and Monthly mean temperature [**?**] consist of station averages of monthly air temperature and precipitation interpolated to a 0.5 degree by 0.5 degree of latitude/longitude grid, where the grid nodes are centered on 0.25 degree. From this a section of data for stations inside North America was used for this experiment.

The entire data set consists of 86609 measurements from all over the world. The attributes of the data are described in [**?**]. The mean annual temperature is the last column attribute. The program climatetoxy.java extracts the required latitude, longitude and the measurement of mean annual temperature at each location. For our plots we take the latitude and longitude without converting it to cartesian coordinates. The scatter plot of cai_temp.txt as shown in Figure **??** is generated by precipscatter.R and the histogram of the measurements as shown in Figure **??** is generated by histogramsclimate.R .

Due to the large size of the original data set we observed that the experiments will be good enough on a small subset of it and we chose a spatial extent defined on North America of 3989 points for our experiments. The scatter plot and histogram of this subset in Figure **??** and Figure **??** show that there are visually identifiable regions in the area of North America which are different from its immediate neighborhood. This subset, tempNA.txt has latitude, longitude and the measured value for 3989 points spread uniformly over North America.

2. Why is this experiment ?

The goal of this experiment is to verify if growing regions based on gini index will give a good precision and recall with respect to a ground truth marked manually in the region of Seattle. This experiment is intended to validate if our algorithm based on growing regions on gini index calculation which will give regions which are homogeneous inside but different from

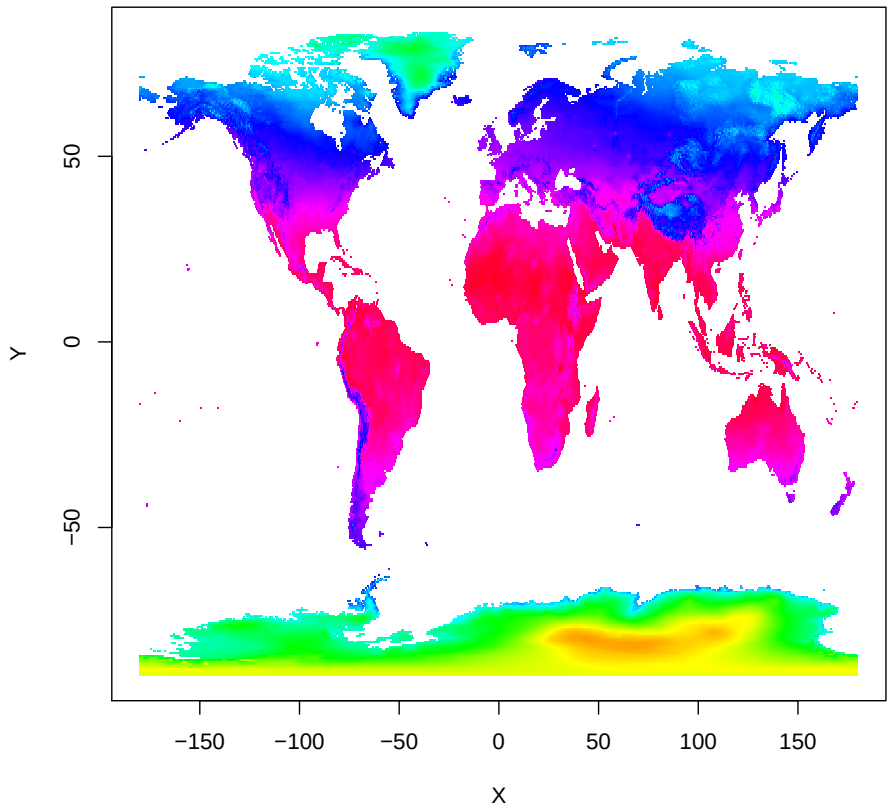

Figure 1. Visualization of dataset cai_temp.txt for mean annual temperature

its immediate neighborhood indeed gives the visually identified anomalous regions. We expect that this experiment will give a better precision and recall than given by SatScan.

3. The program description and source files

Step 1: Load the data set and grow the gini areas The first step is to run giniareathresholdcombined.java. This program requires the dataset tempNA.txt generated as mentioned in section 1. This program basically gives the following files as output:

1. ablineXs.txt and ablineYs.txt : This for the abline() function of the R program to draw the lines of the grid.
2. climateginiarearectanglesthreshold.txt which specifies the coordinates of the lower left and top right corner of the rectangles constituting each gini area separated by spaces. This file is not required for any subsequent step of this program or for the calculation of precision and recall. However, this is required by the R script to plot the rectangles of each gini area. This file also serves the purpose of visually verifying the algorithm and to see which values have been combined to form gini areas. There is an empty line between the entries of two different gini areas.

   The attributes of each column of each line of this file are as follows:

   (a) Longitude of lower left corner of rectangle
   (b) Latitude of lower left corner of rectangle
   (c) Longitude of top right corner of rectangle

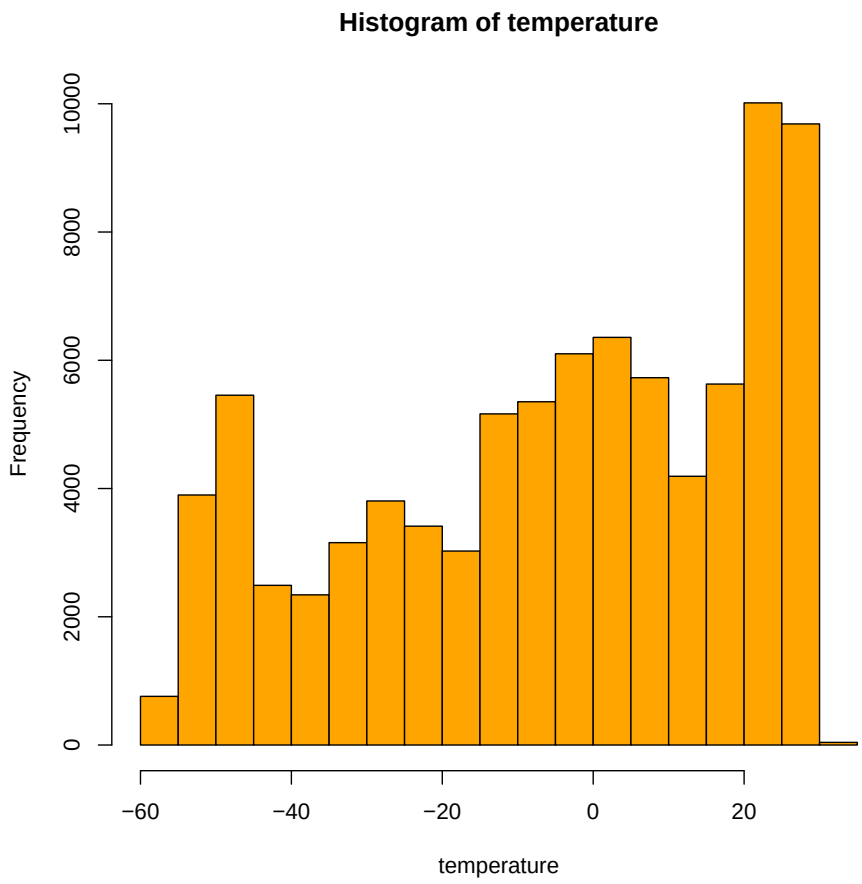

Figure 2. Histogram of dataset cai_temp.txt for mean annual temperature

(d) Latitude of top right corner of rectangle

(e) Gini index of the gini area of which the rectangle is a part of

(f) The measurement of the point in that rectangle

(g) The index of the cell in the array which is (row number) * (number of columns) + (column number)

(h) Row number

(i) Column number

(j) Number of points in that grid cell which is default 1 in this case

(k) Number of rectangles in the gini area of which this rectangle is a part of

(l) The count of the gini area of which this rectangle is a part of

3. giniareastringsginiseattlethreshold.txt which is a text file where each line represents a gini area which is a string of the indices of the cells which are part of the gini area separated by a comma. In this file the gini areas are sorted by the gini indices of the individual gini areas.

4. giniareastringsseattleabswithneighbors.txt : In this file the gini areas are sorted by the sum of the absolute values of the difference between the mean of the values within each gini area and each of the individual values in a belt of cells which form the neighbors of that gini area.

5. giniareastringsseattlewithneighborsavg.txt : In this file the gini areas are sorted by the difference in the average of the values within a gini area and the average of the values in the neighborhood of that gini area.

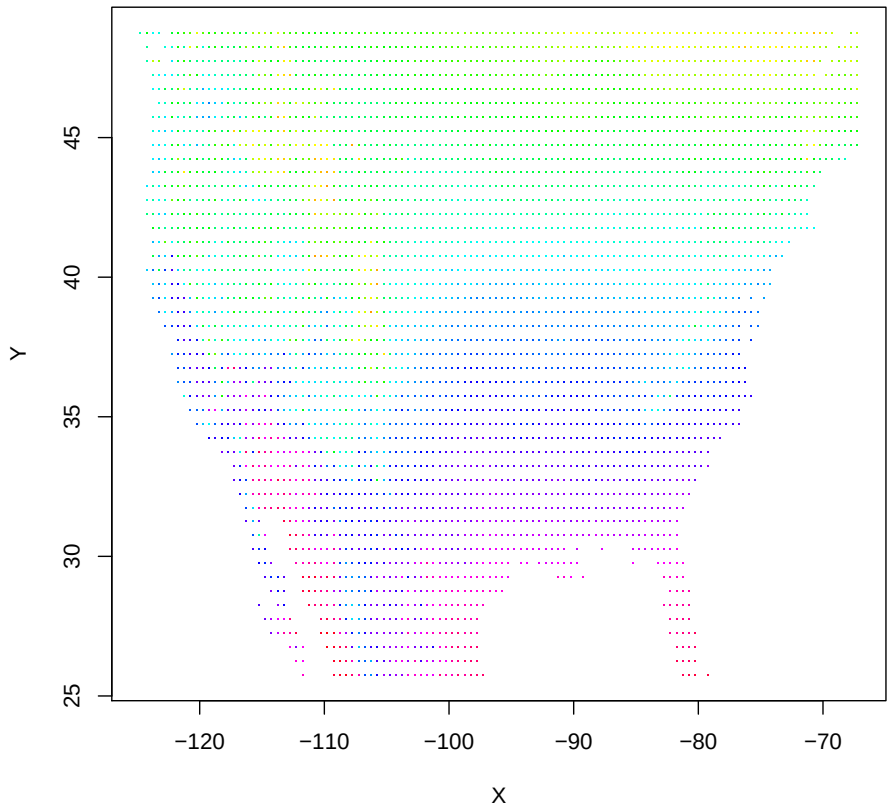

Figure 3. Visualization of dataset tempNA.txt for mean annual temperature in North America

Step 2: Compute the anomalous cells in the Seattle region The outputs generated by giniareathresholdcombined.java has to be further processed by converttoprecinputthreshold.java. The program converttoprecinputthreshold.java require as input a text file in which each line represents a string representation of a gini area. In simple words, feed this program one of the files 3, 4 or 5 listed above as per your requirement. The program converttoprecinputthreshold.java produces only 1 output file ginineighboranomalies.txt which lists the row number and column number of each anomalous cell separated by a comma.

Step 3: Compute precision and recall The precision and recall is calculated by precisionrecallthreshold.java which require two inputs

- ginineighboranomalies.txt which was generated in the previous step by climateginiarearectanglesthreshold.tx
- groundtruthseattle.txt which is manually generated.

These input file have the ground truth and anomalous cells listed following the same format , i.e. the row number and column number separated by a comma. (If new ground truth has to be collected, it has to be in the same way - row number and column number of each cell on a separate line.)

The program precisionrecallthreshold.java gives two output files groundtruthseattlelatlongrectangles.txt and giniarea-cellsseattlerectangles1.txt which give on each line the longitude and latitude of the lower left coordinate and the upper right coordinate of the rectangles constituting the ground truth and gini areas.

Step 4: Plot the graph Now, tempNAgridthreshold.R is the script which plots the graph. This script requires as input the following files:

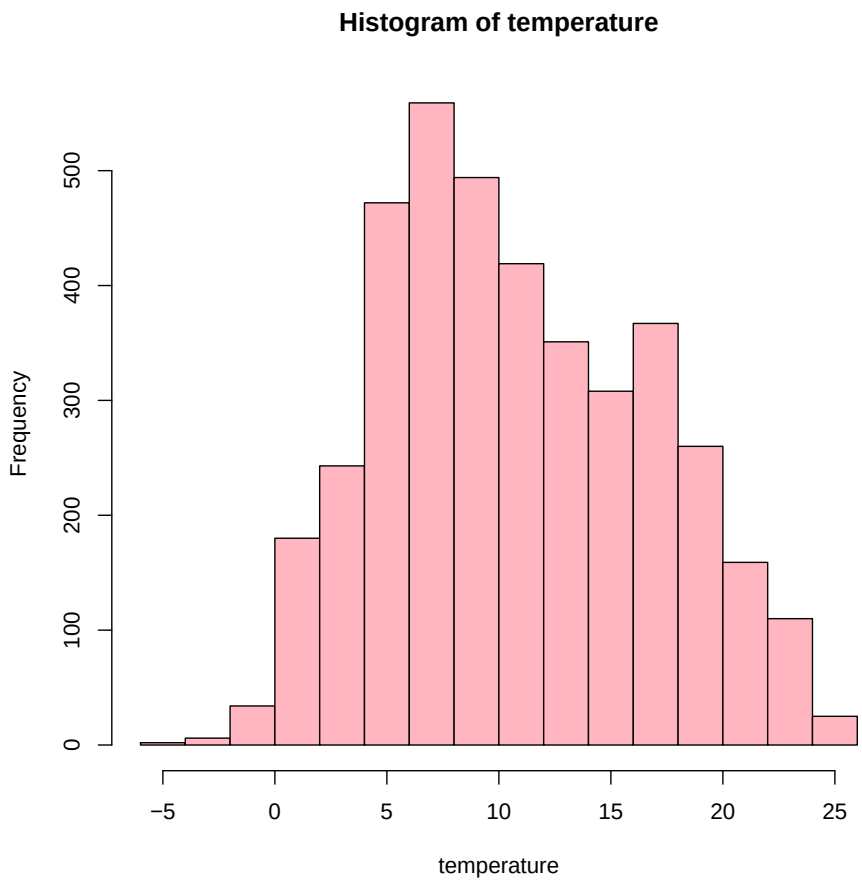

Figure 4. Histogram of dataset tempNA.txt for mean annual temperature in North America

1. NAtemplatlongcasnorm.cas : This is an input file for the SatScan software. This has the latitude , longitude and measurement for each point. Once this data file is loaded, this can be used for all subsequent plots for all experiments using this dataset.

2. ablineXs.txt and ablineYs.txt for drawing the lines of the grid

3. climateginiarearectanglesthreshold.txt for drawing the rectangles of each gini area

4. groundtruthseattlelatlongrectangles.txt for hatching the rectangles corresponding to the ground truth. Since the ground truth is same for all experiments here, once loaded this file does not have to be loaded again.

5. giniareacellsseattlerectangles1.txt for hatching the rectangles for the anomalous cells.

An outline of the program is as follows:

1. Create an array list *ginineighbors* of m x n strings , each entry denotes a cell of the grid. A cell is represented by an integer which is rownum x number of columns + column number of the cell. To start with each gini area is represented by a string for a single cell. As the gini area grows the added cell numbers are concerted to the string separated by a comma.

2. Create an array list for *bestginineighbor* of m X n Integers. Each entry i denotes the index of the best neighbor of gini area of index i in *ginineighbors*. default is -1.

3. Create an array list of *bestginiofneighbor* of m X n Doubles. Each entry i denotes the gini induex of the gini area of index i in *ginineighbors* with its best neighbor. Default is 3.0

4. Create an array list if *gin indices* of m x n Doubles where each entry i is the gini index of the i th gini area in *ginineighbors*.

5. All the above array lists are of equal length after each iteration.

6. Iterate through *ginineighbors* and for each gini area, find its best neighbor and gini index with best neighbor. Record the best of the bestginiofneighbors as minginiindex.

7. If minginiindex is less than 0.05, then go to step 8, else go to step 10.

8. For all those gini areas which forms giniindex less than 0.05 with its best neighbor, the entry for the best neighbor is deleted from the array list *ginineighbors* and concatenated to the entry for that gini area. The change is duplicated in the array lists *bestginiofneighbor, bestginineighbor and ginindices*, so that all the these arrays remain of the same size.

9. Go to step 6.

10. Now we have a an array list of strings *ginineighbors* , each string representing a single gini area with gini index less than 0.05.

4. How to run the programs

Java and R are required for these experiments. Only core java is required for all. I have used Java 1. 6. I would recommend using eclipse.

Install R. I would recommend downloading and installing RStudio. Make sure that the following packages are also installed:

1. plotrix
2. grDevices
3. calibrate
4. graphics

From the Rstudio GUI, take Tools $\rightarrow$ Install packages and type the package name in the space specified. This should do the installation automatically. Once installed, my scripts take care of importing their libraries as required. Sometimes one or more 'devices' remains open and then no more pdfs can be created. In that case, just type dev.off() a couple of times at the command prompt until all 'devices' shut down.

Running the scripts involves the following 4 steps:

1. Open the Rstudio IDE : Has 4 windows in clockwise direction : Editor, Workspace, Plots and Command prompt

2. File $\rightarrow$ Open $\rightarrow$ Navigate to the script $\rightarrow$ Open : This opens the script in the editor

3. In the workspace : Import Dataset $\rightarrow$ Navigate to the dataset $\rightarrow$ Load . The table is now visible in the editor and its internal name appears in the command prompt. Close the table viewer in the editor. Make sure that the script uses the same name as its internal name.

4. Make sure that the script's first line which specifies where the pdf is to be generated is correct.

5. Select "Source on save" in the top left corner of editor.

6. Save the script and this should run the script. (If required, you can stop the script by clicking the stop icon in the window for the command prompt.)

7. Find the pdf where you specified it to be created.

5. Ground truth from different subjects

The ground truth has to be a file named groundtruthseattle.txt and each line of the file should specify a single cell specified as follows:

row number , column number

The subject has to be given the pdf **??** generated by the script markseattle.R which has just the measurement in each cell with the row number and column number so that it will be easy for the subject to mark off cells by just looking at this pdf without being biased by any of the results of the algorithm.

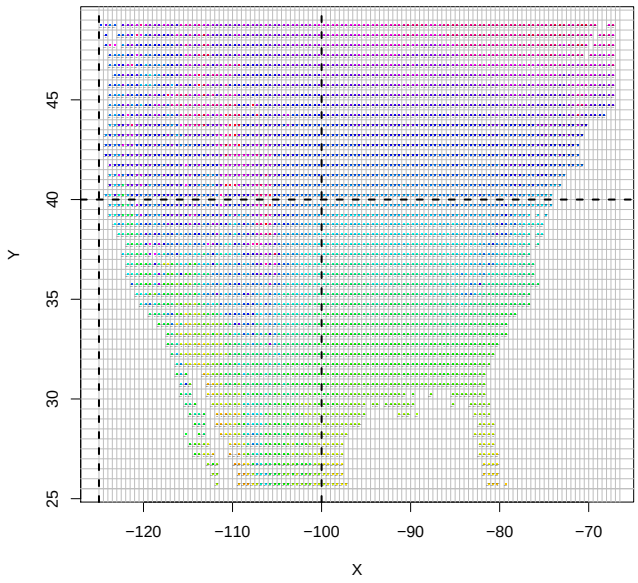

Figure 5. For collecting ground truths from subjects

6. Analysis with Results

The statistics of the algorithm are as follows:

Number of iterations:8
Number of gini areas: 1782
Single cell gini areas :1695
Biggest gini cluster has size: 712

The algorithm is run for the dataset consisting of all points in North America. The extent of the latitude was between 25.75 and 48.75 and that of the longitude was between -127.0 and -67.25. Since this dataset is for gridded data area, this consists of 48 rows and 117 columns. The gini areas are grown over all these points. The java program converttoprecinputthreshold.java crops the top k gini areas intersecting with a rectangular region in Washington for row numbers from 29 to 48 and column numbers from 0 to 49 which is a 19 by 50 grid of 950 cells. The precision and recall are computed over these cells of the cropped top k gini areas.

6.1. Gini areas sorted by gini index

File giniareastringsginiseattlethreshold.txt generated after Step 1 in section 3 contains the strings representing the gini areas sorted by their gini indices. We continue with the rest of the steps with this file as input to step 2. The plot of the anomalous cells taking the top 80 gini areas are as in Figure **??**. The precision and recall are as in the following table for the top k gini areas :

k of Top k	Rank overall	Number of anomalous cells	Precision	Recall
30	1551	30	0.8	0.0916030534351145
40	1690	40	0.825	0.12595419847328243
50	1701	55	0.7454545454545455	0.15648854961832062
60	1725	92	0.6630434782608695	0.23282442748091603
70	1740	152	0.6118421052631579	0.3549618320610687
80	1759	266	0.5225563909774437	0.5305343511450382
90	1775	666	0.3738738738738739	0.950381679389313

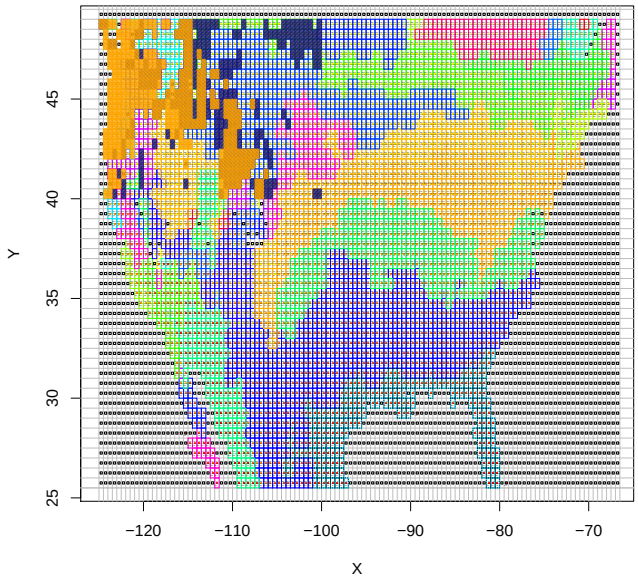

Figure 6. Ground truth(blue) and anomalous cells(orange) taking top 80 gini areas sorted according to gini index

Taking the top k gini areas sorted according to gini index, first all the single cell gini areas of gini index zero get selected and then it selects the gini areas of larger number of cells.

6.2. Gini Areas sorted in decreasing order of the sum of the absolute value of the difference between the value in each cell of the immediate neighborhood and the mean inside the gini area

File giniareastringsseattleabswithneighbors.txt generated after Step 1 in section 3. We continue with the rest of the steps with this file as input to step 2. The plot of the ground truth and anomalous cells taking the top 80 gini areas is shown in Figure **??**. The precision and recall are as in the following table for the top k gini areas :

k of Top k	Rank overall	Number of anomalous cells	Precision	Recall
30	366	127	0.29133858267716534	0.14122137404580154
40	382	156	0.3782051282051282	0.22519083969465647
50	398	238	0.29411764705882354	0.26717557251908397
60	412	336	0.2708333333333333	0.4312977099236641
70	430	472	0.23940677966101695	0.3549618320610687
80	448	515	0.27766990291262134	0.5458015267175572
90	462	625	0.3392	0.8091603053435115
95	1782	867	0.3021914648212226	1.0

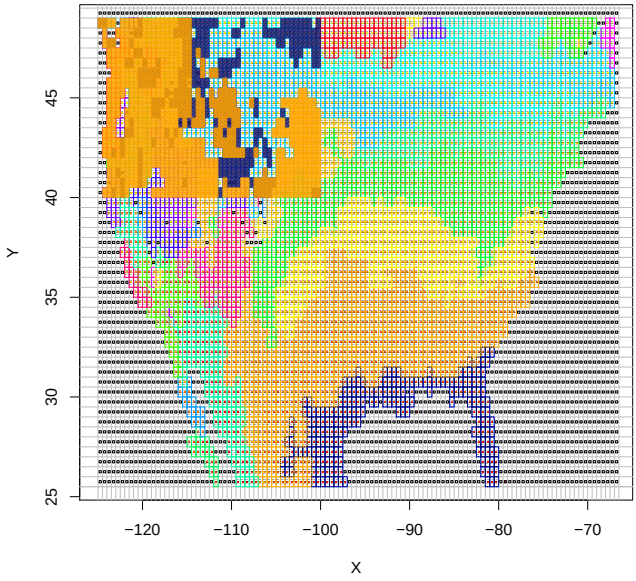

Figure 7. Ground truth(blue) and anomalous cells(orange) taking top 80 gini areas sorted according to sum of absolute value of difference between each value in its neighborhood and the mean of the gini area.

Here, the single cell gini areas lose its rank because here the sum of the absolute difference between the mean inside the gini area and each of its neighbor cells is taken as the sorting statistic. However, gini areas towards the edge with zero valued neighbors get higher up in the rank leading to the selection of gini areas with lesser number of non-zero neighbors.

6.3. Gini Areas sorted by the difference in average of the values in the neighborhood cells and the average within the gini area

File giniareastringsseattlewithneighborsavg.txt generated after Step 1 in section 3. We continue with the rest of the steps with this file as input to step 2. The plot of the ground truth and the anomalous cells are as shown in Figure **??**. The precision and recall are as in the following table for the top k gini areas :

k of Top k	Rank overall	Number of anomalous cells	Precision	Recall
30	364	141	0.3333333333333333	0.17938931297709923
40	379	181	0.35359116022099446	0.24427480916030533
50	398	263	0.30038022813688214	0.3015267175572519
60	413	391	0.23273657289002558	0.3473282442748092
70	427	478	0.23430962343096234	0.42748091603053434
80	448	515	0.27766990291262134	0.5458015267175572
90	462	625	0.3392	0.8091603053435115
95	1782	867	0.3021914648212226	1.0

6.4. Gini Areas sorted by the difference in gini index between each gini area with its best neighbor and without its best neighbor

File giniareastringNAdiffthreshold.txt generated after Step 1 in section 3. We continue with the rest of the steps with this file as input to step 2. The plot of the ground truth and the anomalous cells are as shown in Figure **??**. The precision and recall are as in the following table for the top k gini areas :

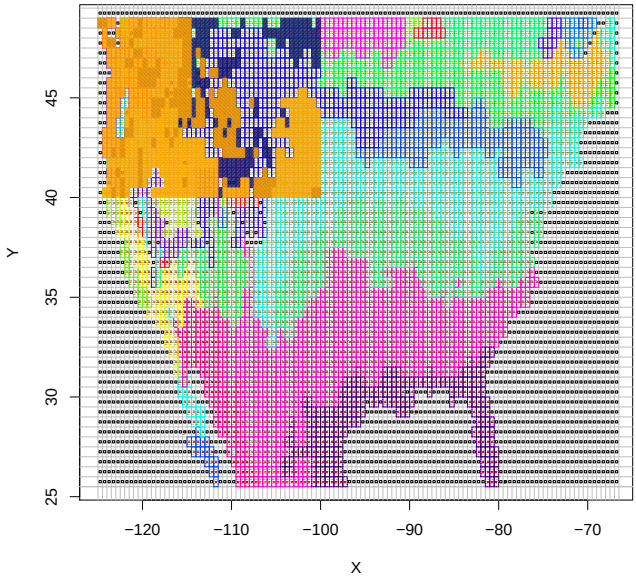

Figure 8. Ground truth(blue) and anomalous cells(orange) taking top 80 gini areas sorted according to difference between the average calculated amongst the neighborhood cells and average calculated from amongst the cells within the gini area.

k of Top k	Rank overall	Number of anomalous cells	Precision	Recall
30	1672	32	0.6875	0.08396946564885496
40	1688	42	0.7380952380952381	0.1183206106870229
50	1702	57	0.6842105263157895	0.14885496183206107
60	1725	86	0.7325581395348837	0.24045801526717558
70	1743	152	0.5789473684210527	0.33587786259541985
80	1760	285	0.4280701754385965	0.46564885496183206
90	1775	465	0.3634408602150538	0.6450381679389313
95	1782	867	0.3021914648212226	1.0

7. A few other experiments conducted earlier

7.1. Growing gini areas combining only the best pair during each iteration

The Java source file for this is climategridginineighbor.java. This program also outputs one file climateginiarearectangles.txt for plotting the rectangles of the gini areas and the respective text files containing the strings of the gini areas sorted using different criteria. The statistics of this algorithm is as follows:

Number of iterations:3871
Number of gini areas: 1745
Single cell gini areas :1679
Biggest gini cluster has size: 593

7.1.1. Gini Areas sorted in increasing order of gini index The file giniareastringsginiNAbestpair.txt generated in Step 1 of section 3 is then used for the rest of the steps. The plot of the ground truth and anomalous cells of the top 60 gini areas is Figure **??**. The precision and recall are as follows:

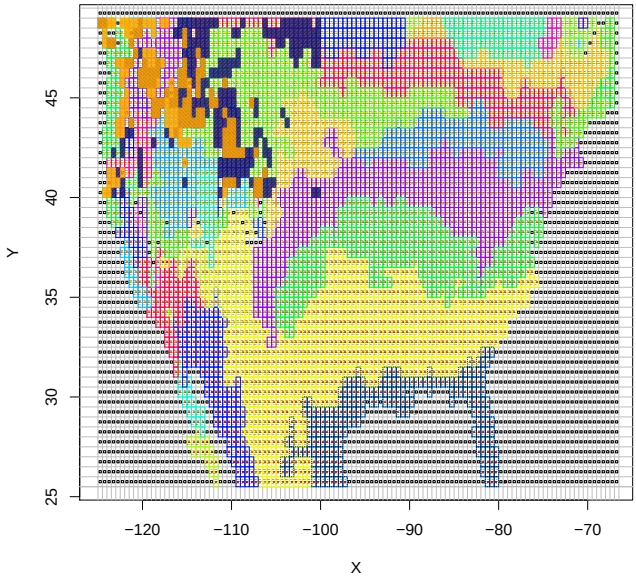

Figure 9. Ground truth(blue) and anomalous cells(orange) taking top 80 gini areas sorted according to difference between the average calculated amongst the neighborhood cells and average calculated from amongst the cells within the gini area.

k of Top k	Rank overall	Number of anomalous cells	Precision	Recall
30	1661	30	0.8	0.0916030534351145
40	1684	45	0.8	0.13740458015267176
50	1705	100	0.63	0.24045801526717558
60	1721	168	0.5476190476190477	0.3511450381679389
70	1738	371	0.3611859838274933	0.5114503816793893
80	1745	867	0.3021914648212226	1.0
90	1745	867	0.3021914648212226	1.0

7.1.2. Gini Areas sorted in decreasing order of the sum of the absolute value of the difference between the value in each cell of the immediate neighborhood and the mean inside the gini area File giniareastringsNAabsbestpair.txt generated after Step 1 in section 3. We continue with the rest of the steps with this file as input to step 2. The plot of the ground truth and anomalous cells taking the top 60 gini areas is shown in Figure **??**. The precision and recall are as in the following table for the top k gini areas :

k of Top k	Rank overall	Number of anomalous cells	Precision	Recall
30	361	80	0.4875	0.14885496183206107
40	376	194	0.30412371134020616	0.22519083969465647
50	393	355	0.2140845070422535	0.2900763358778626
60	408	448	0.2700892857142857	0.4618320610687023
70	422	866	0.29651162790697677	0.9732824427480916
80	1745	867	0.3021914648212226	1.0
90	1745	867	0.3021914648212226	1.0

7.1.3. Gini Areas sorted by the difference in average of the values in the neighborhood cells and the average within the gini area File giniareastringsNAavgbestpair.txt generated after Step 1 in section 3. We continue with the rest of the

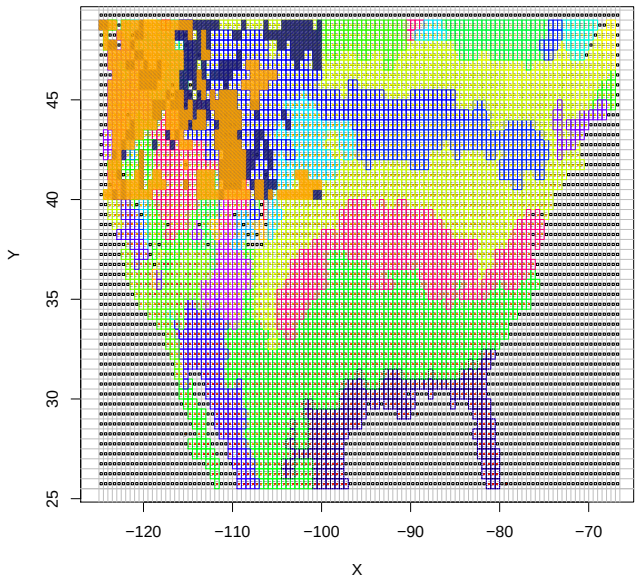

Figure 10. Ground truth(blue) and anomalous cells(orange) taking top 60 gini areas sorted in increasing order of gini index.

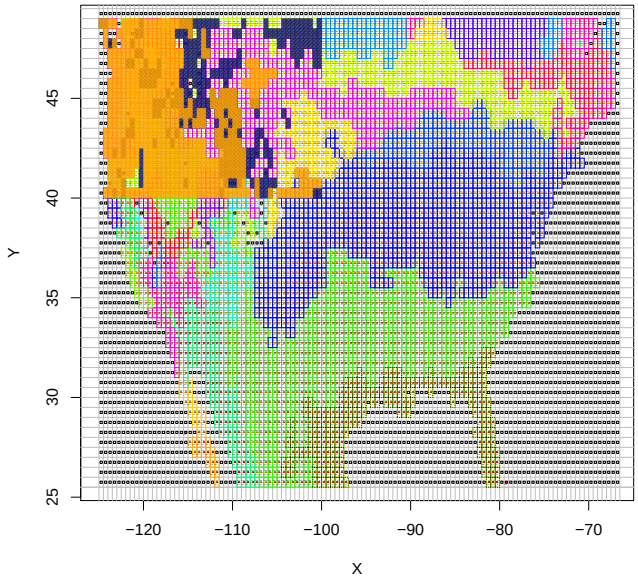

Figure 11. Ground truth(blue) and anomalous cells(orange) taking top 60 gini areas sorted in decreasing order of sum of absolute value of difference between each individual value in the neighborhood and the mean of gini area.

steps with this file as input to step 2. The plot of the ground truth and the anomalous cells are as shown in Figure **??**. The precision and recall are as in the following table for the top k gini areas :

k of Top k	Rank overall	Number of anomalous cells	Precision	Recall
30	360	217	0.2073732718894009	0.1717557251908397
40	373	332	0.20783132530120482	0.2633587786259542
50	392	360	0.2111111111111111	0.2900763358778626
60	408	448	0.2700892857142857	0.4618320610687023
70	422	860	0.2965116279069767 7	0.9732824427480916
80	1745	867	0.3021914648212226	1.0
90	1745	867	0.3021914648212226	1.0

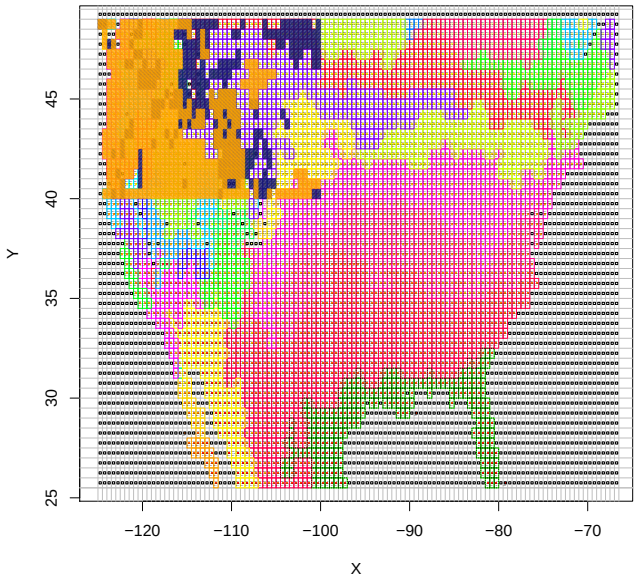

Figure 12. Ground truth(blue) and anomalous cells(orange) taking top 60 gini areas sorted in decreasing of difference in average between neighborhood cells and cells inside the gini area.

7.1.4. Gini Areas sorted in decreasing order of the difference between gini index of each gini area with its best neighbor and without. File giniareastringsNAbestpair.txt generated after Step 1 in section 3. We continue with the rest of the steps with this file as input to step 2. The plot of the ground truth and the anomalous cells are as shown in Figure **??**. The precision and recall are as in the following table for the top k gini areas :

k of Top k	Rank overall	Number of anomalous cells	Precision	Recall
30	1655	143	0.7062937062937062	0.38549618320610685
40	1673	156	0.717948717948718	0.42748091603053434
50	1696	172	0.7383720930232558	0.4847328244274809
60	1721	251	0.6573705179282868	0.6297709923664122
70	1737	433	0.4896073903002309	0.8091603053435115
76	1745	867	0.3021914648212226	1.0

7.2. Growing gini areas taking only the points in the rectangular region in Washington

7.2.1. Gini areas sorted by gini index File giniareastringsginiseattlebpair.txt generated after Step 1 in section 3 contains the strings representing the gini areas sorted by their gini indices. We continue with the rest of the steps with this file as

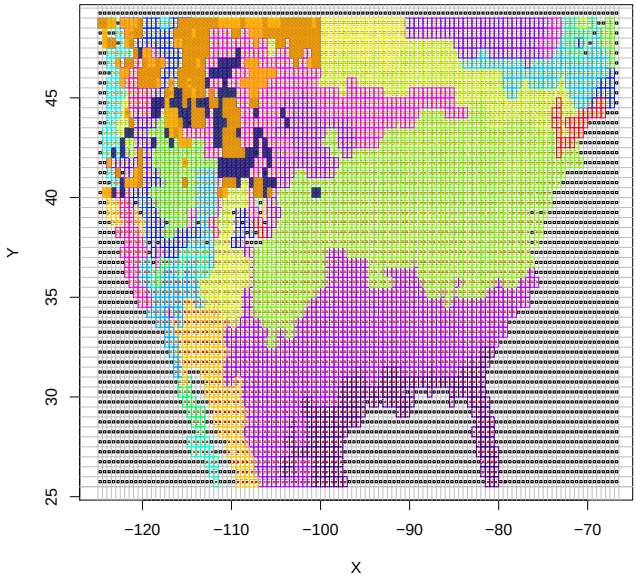

Figure 13. Ground truth(blue) and anomalous cells(orange) taking top 60 gini areas sorted in decreasing order of difference in gini index of the gini area with its best neighbor and without it.

input to step 2. The plot is generated by using same script as in previous experiments. The precision and recall are as in the following table for the top k gini areas :

k of Top k	Rank overall	Number of anomalous cells	Precision	Recall
30	113	31	0.8387096774193549	0.09923664122137404
40	123	41	0.8292682926829268	0.1297709923664122
50	133	51	0.8431372549019608	0.16412213740458015
60	143	89	0.651685393258427	0.22137404580152673
70	153	160	0.55	0.33587786259541985
80	163	221	0.5248868778280543	0.44274809160305345
90	173	537	0.32774674115456237	0.6717557251908397
94	177	867	0.3021914648212226	1.0

7.2.2. Gini Areas sorted in decreasing order of the sum of the absolute value of the difference between the value in each cell of the immediate neighborhood and the mean inside the gini area File giniareastringsseattleabsbestpair.txt generated after Step 1 in section 3. We continue with the rest of the steps with this file as input to step 2. The plot of the ground truth and anomalous cells is generated by the same script as the previous experiments. The precision and recall are as in the following table for the top k gini areas :

k of Top k	Rank overall	Number of anomalous cells	Precision	Recall
30	100	124	0.28225806451612906	0.13358778625954199
40	110	257	0.19455252918287938	0.19083969465648856
50	120	295	0.2847457627118644	0.32061068702290074
60	130	347	0.27089337175792505	0.35877862595419846
70	140	406	0.3275862068965517	0.5076335877862596
80	150	784	0.3022959183673469	0.9045801526717557
90	160	862	0.29814385150812067	0.9809160305343512
94	177	867	0.3021914648212226	1.0

7.2.3. Gini Areas sorted by the difference in average of the values in the neighborhood cells and the average within the gini area File giniareastringsseattleavgbestpair.txt generated after Step 1 in section 3. We continue with the rest of the steps with this file as input to step 2. The plot of the ground truth and the anomalous cells are generated by the same script as the previous experiments. The precision and recall are as in the following table for the top k gini areas :

k of Top k	Rank overall	Number of anomalous cells	Precision	Recall
30	364	141	0.3333333333333333	0.17938931297709923
40	379	181	0.35359116022099446	0.24427480916030533
50	398	302	0.26490066225165565	0.3053435114503817
60	413	391	0.23273657289002558	0.3473282442748092
70	427	478	0.23430962343096234	0.42748091603053434
80	448	515	0.27766990291262134	0.5458015267175572
90	462	625	0.3392	0.8091603053435115
95	1782	867	0.3021914648212226	1.0

7.2.4. Gini Areas sorted in decreasing order of the difference between gini index of each gini area with its best neighbor and without. File giniareastringsseattlediffbestpair.txt generated after Step 1 in section 3. We continue with the rest of the steps with this file as input to step 2. The plot of the ground truth and the anomalous cells are generated using the same script as the previous experiments. The precision and recall are as in the following table for the top k gini areas :

k of Top k	Rank overall	Number of anomalous cells	Precision	Recall
30	113	33	0.7575757575757576	0.09541984732824428
40	123	43	0.7906976744186046	0.1297709923664122
50	133	53	0.7735849056603774	0.15648854961832062
60	143	78	0.7307692307692307	0.21755725190839695
70	153	148	0.5608108108108109	0.31679389312977096
80	163	205	0.5219512195121951	0.4083969465648855
90	173	524	0.44656488549618323	0.8931297709923665
94	177	867	0.3021914648212226	1.0

References

[1] C. J. Willmott and K. Matsuura. *Terrestrial Air Temperature and Precipitation: Monthly and Annual Time Series (1950 - 1996)*, 2000 (accessed July 23, 2012).