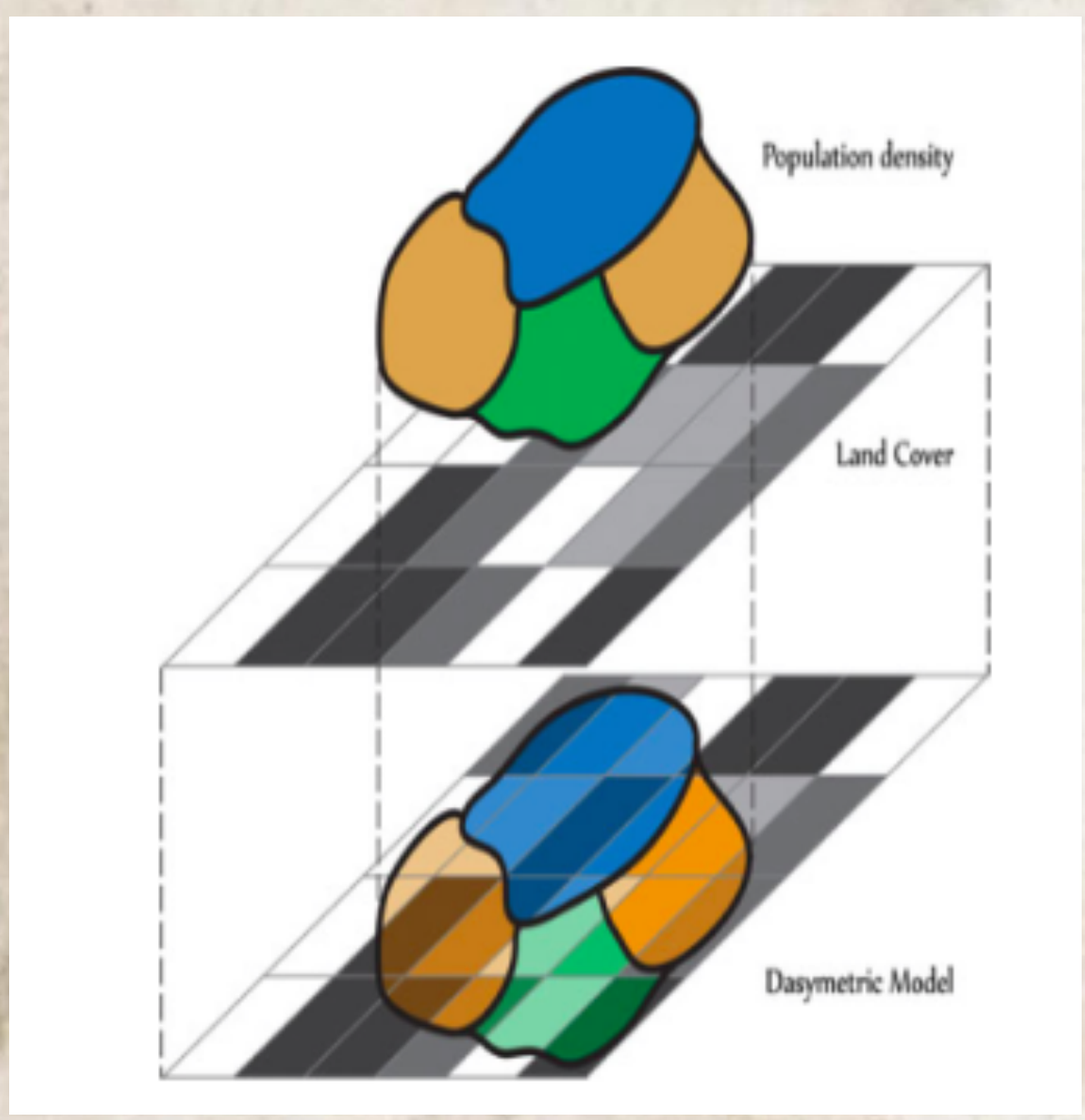
Parallel Method for Dasymetric Mapping

Sam Zhang

Overview



Dasymetric mapping is a commonly-applied algorithm in GIS that can determine the population distribution in a high resolution level. It maps the population data with a coarse resolution with a high-resolution ancillary-level data (most of the time, it's land type data). However, for a large-scale problem, it may take a long time to do this process. So we will propose a parallel method for this problem. And we will use this method to perform dasymetric mapping on Tennessee. If the result is successful, we can expand the source area to the whole U.S.

Size of District Units

- Three levels of district units are included. The largest one is state, the second largest is block group, the smallest is 30m*30m grid. U.S. contains 211267 block groups. Thus a state will contain about 4000 block groups. There are about 10^10 grids in the U.S. Thus there are about 50000 grids in a block group.

Datasets

- NLCD dataset: A 161190x104424 matrix in .img format. (raster file) The value of each entry corresponds to a land type.

- Boundary data of block groups are polygons(vector file) They represent the shape of different block groups in a state.

- ACS Summary File stores the estimate of population in every block group.

Two Stages

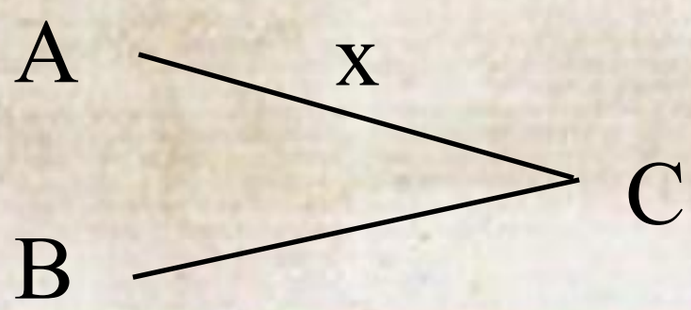
- From tables/raw data to matrices

NLCD Data —extract→ Matrix A QGIS/GDAL

Boundary File for block groups —rasterize→ Matrix B

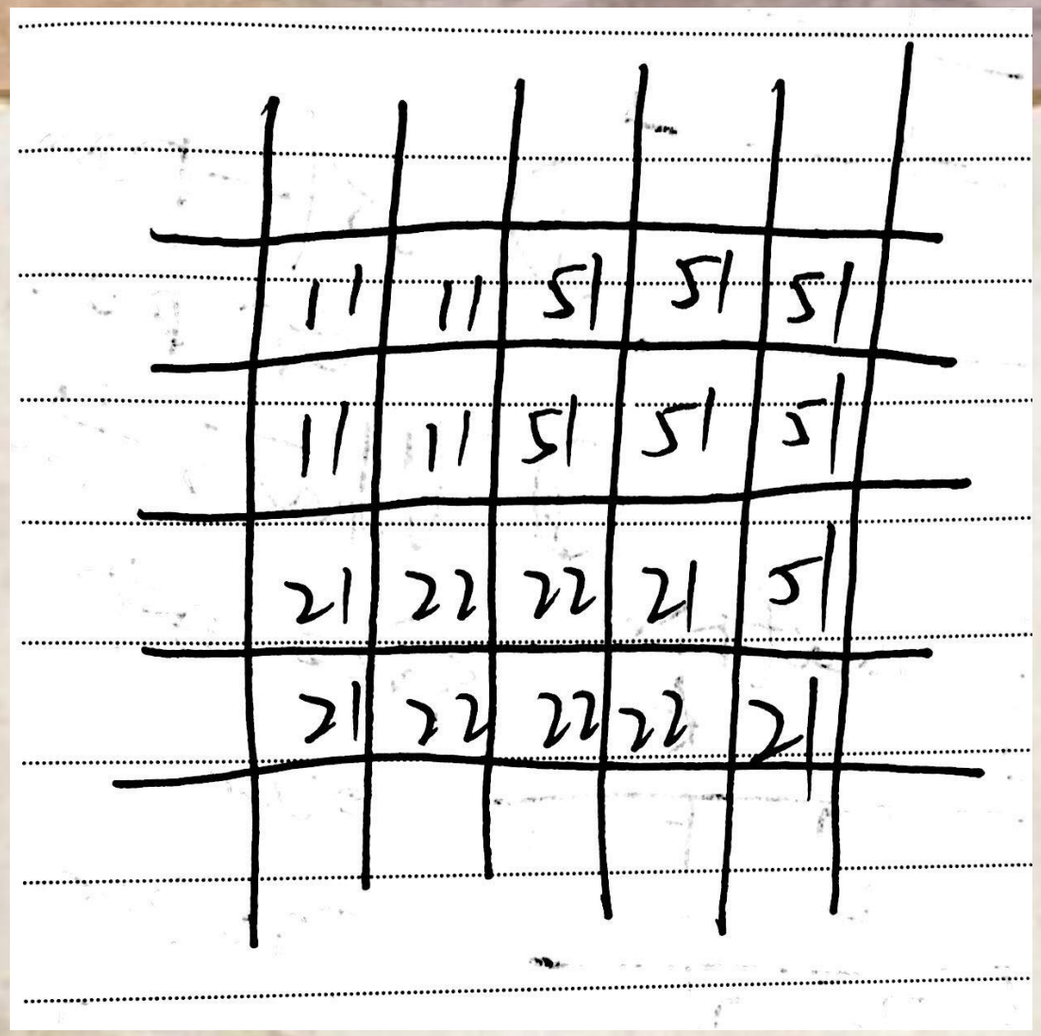
ACS Summary File Table —extract→ Array x

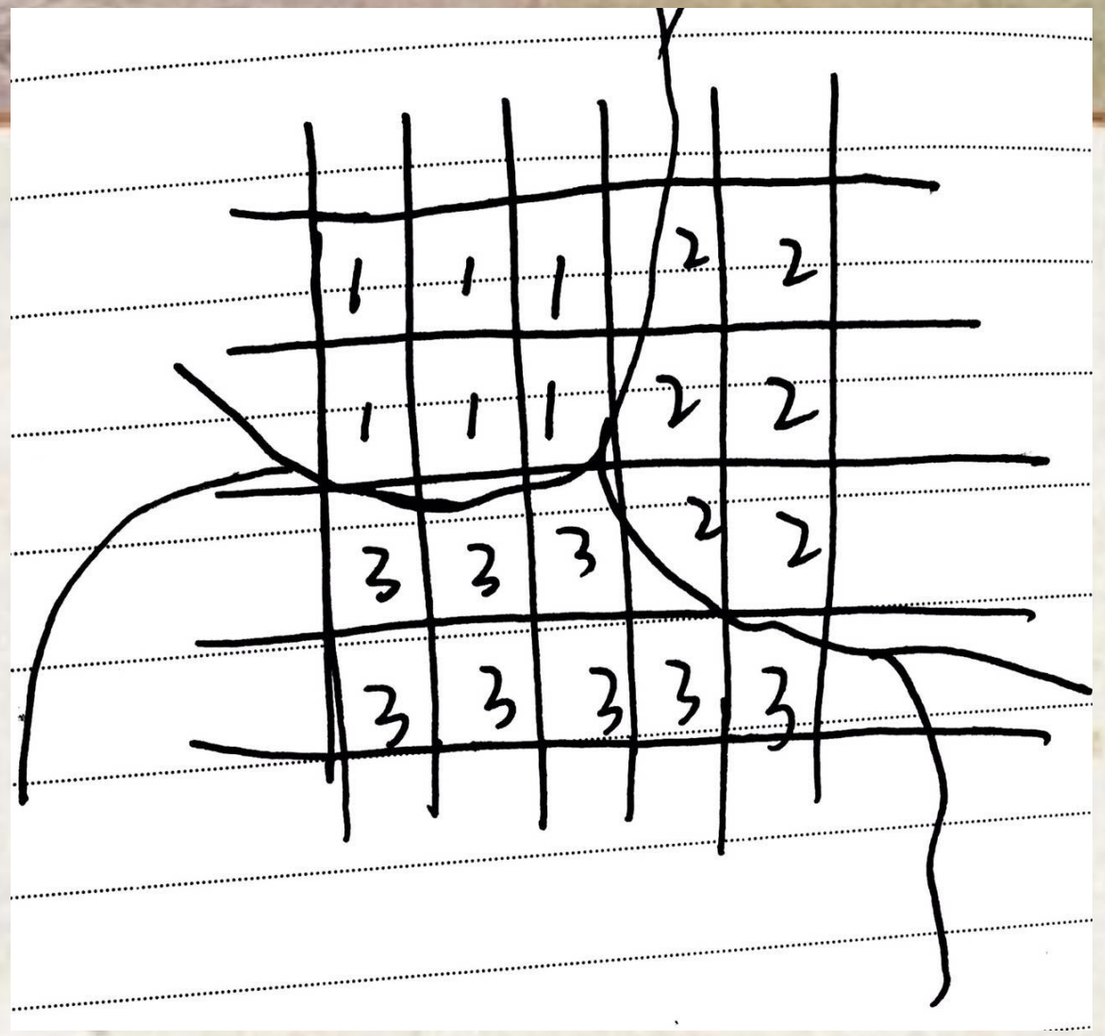
- From matrices to result



What is A,B and x?

- The first step is needed because the three datasets are on different enumeration unit and are in different format. Our output will produce two matrices and an array. They are all on the state level, and the entries of both matrices are on the level of 30*30 grids. So matrix A is the state-level NLCD matrix. Matrix B is the state-level block group matrix(Each entry corresponds to a region on the map, indicating which block group this region belongs to). Array x tells the population in each block group in this state.





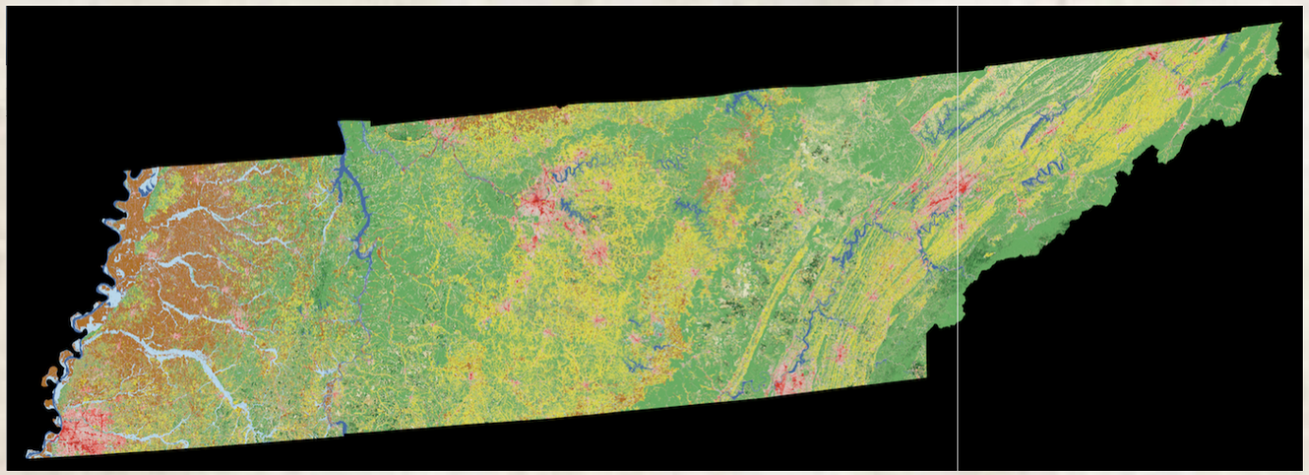
Matrix A

11-Open water
21-Developed, Low Intensity
22-Developed, Medium Intensity
51-Dwarf scrub

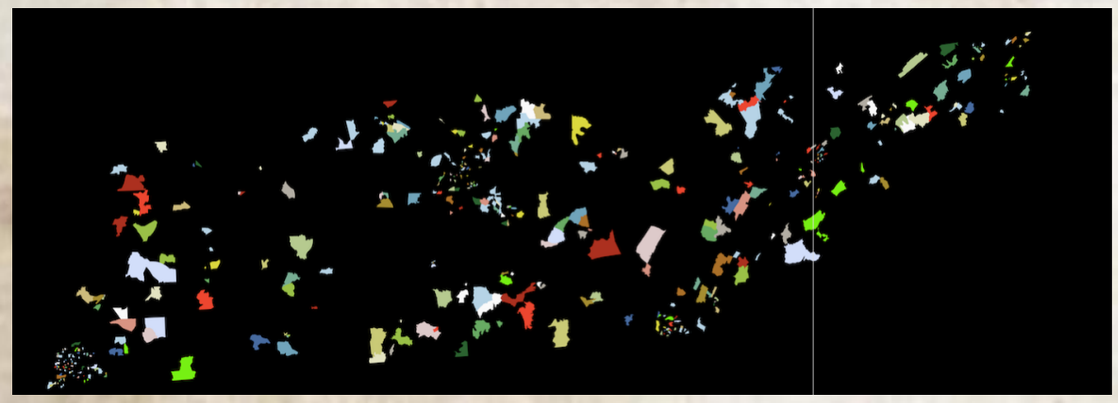
Matrix B

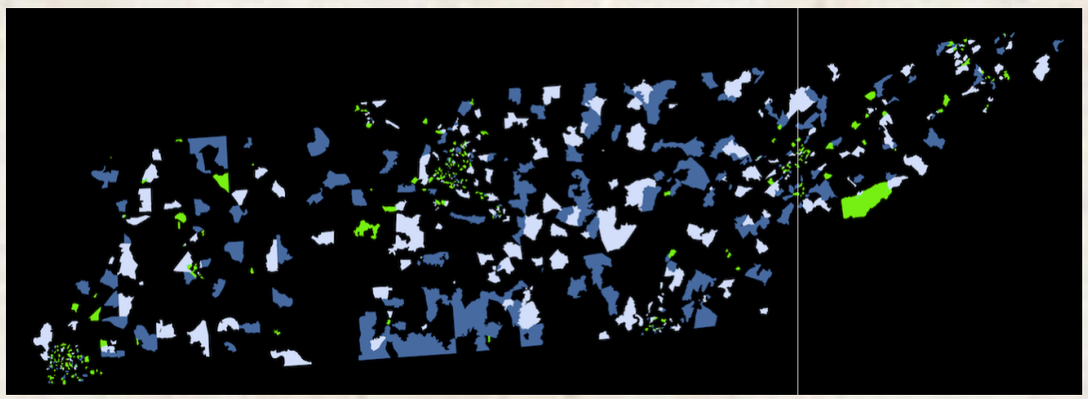
1-Block group 1
2-Block group 2
3-Block group 3

Matrices



Matrix A





Matrix B

Second stage

- The second stage is to get matrix C: A matrix that stores how many people are there in each 30*30 grid.

- First of all, we have to assign weights according to the land type. Each land type will be given a weight according to their ability to reside people. This weight is determined by some regression method. This step changes matrix A to another weight matrix(W) of the same size.

- We can count the total weights in a block group and store this total-weight in another array y(same length as x).

Regression Method

Now we have the land type data for Tennessee, however, it cannot be used directly to determine population distribution. However, it can be used to give a brief estimate of population density in each 30m*30m grid. We can use a linear regression model to model the relation between land type and density:

$$P_S = \alpha + \sum_{c=1}^{C} \beta_c A_{SC} + \epsilon_S$$

where P_S is the population of zone S; α is the intercept term; β_c is the coefficient for land cover c; A_{sc} is the area of land cover c within zone s; C is the number of populated land cover types occurring within study region; and ϵ_S is a random error term.

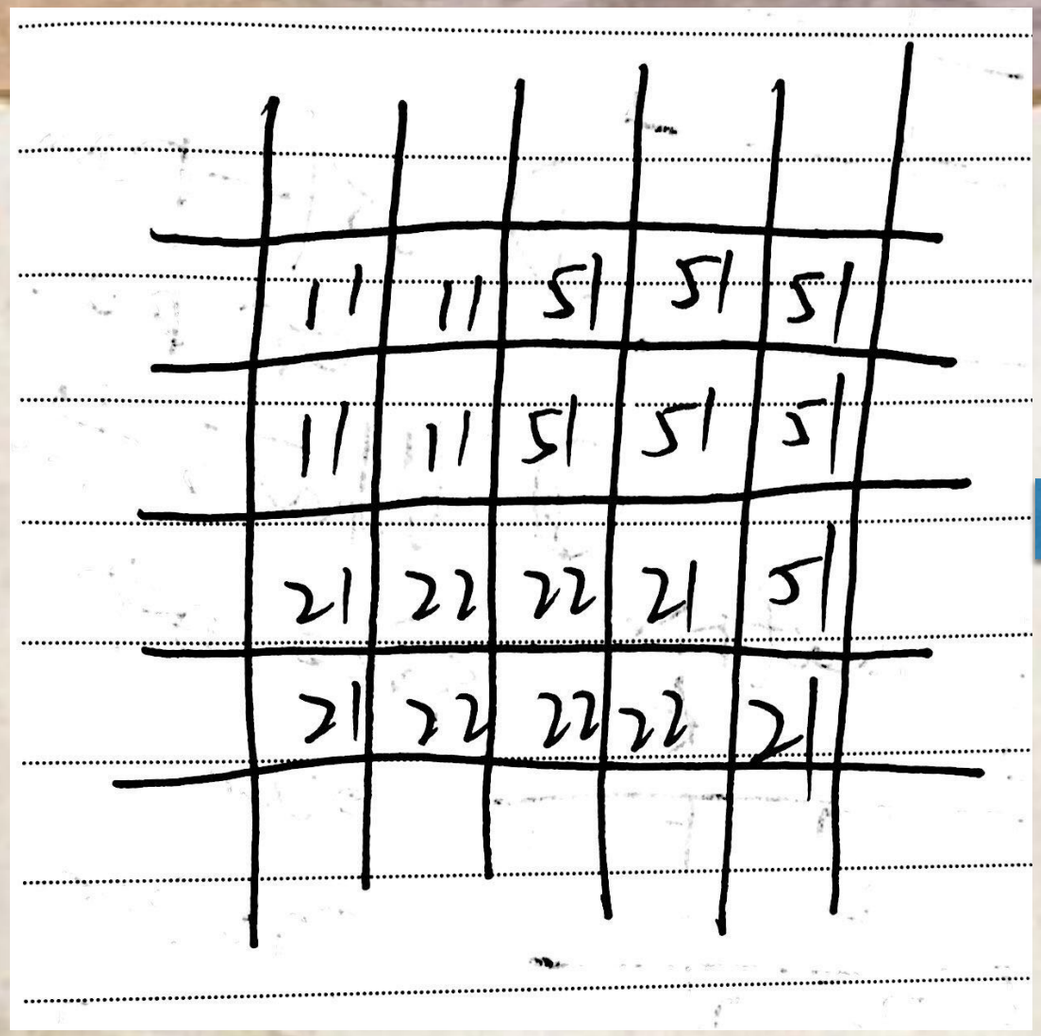
It can also be written as:

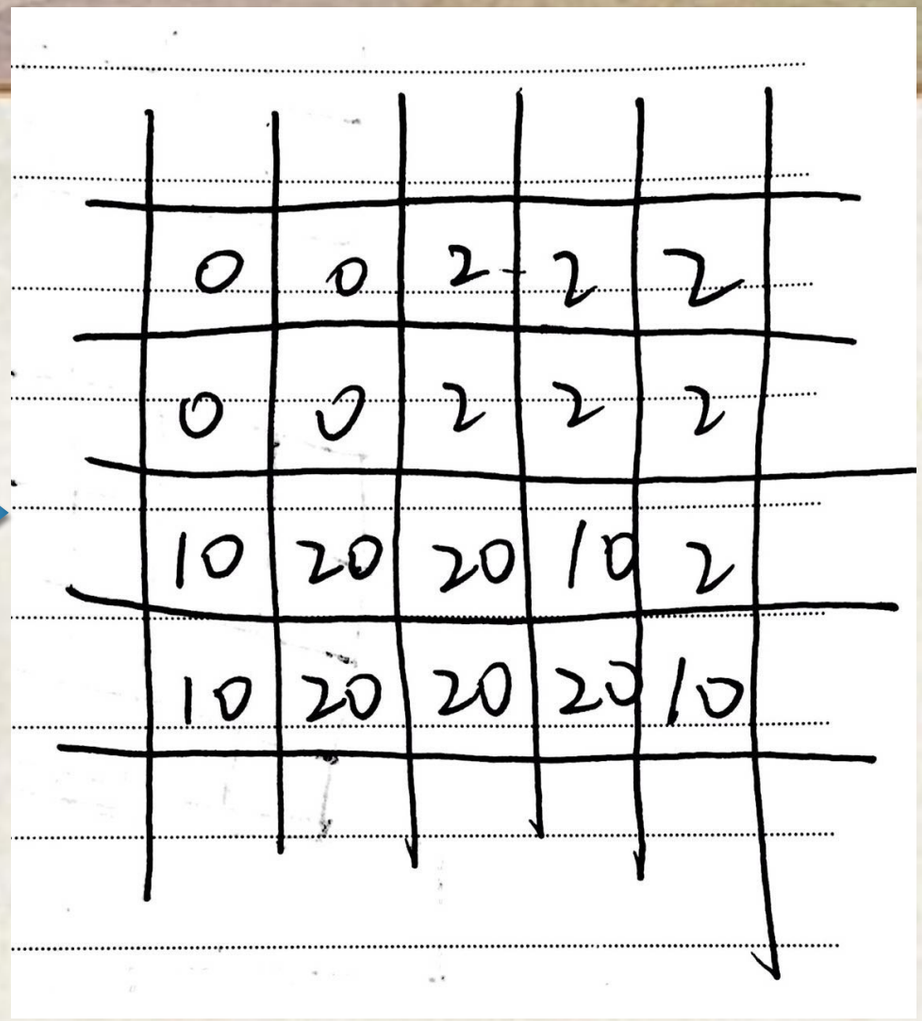
$$E(P_S|A_S) = \alpha + \beta A_S$$

where β and A_S are vectors. Negative population totals can also be avoided by using Poisson regression:

$$E(P_S|A_S) = \exp(\alpha + \beta A_S)$$

Here S(source zone) is a block group. The values in β are the estimated population density for cells with different land types.





Matrix A

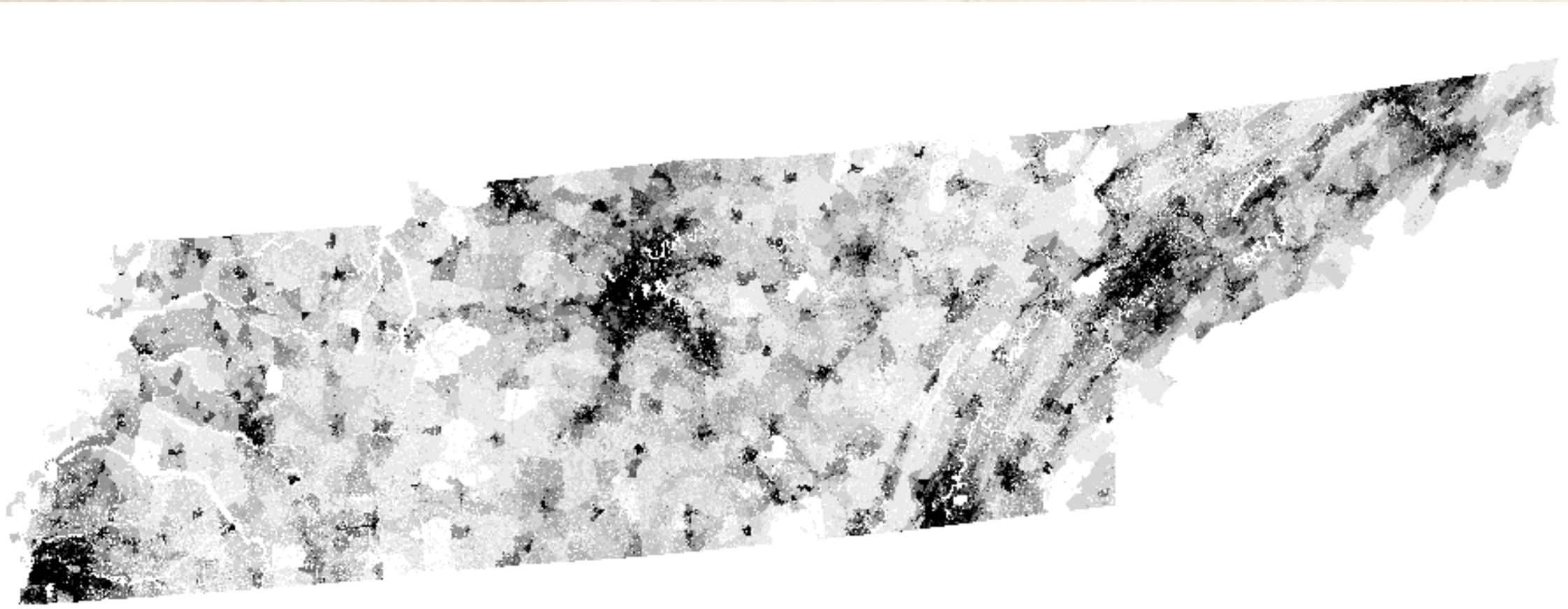
Matrix W

11-Open water
21-Developed, Low Intensity
22-Developed, Medium Intensity
51-Dwarf scrub

Second stage(cont.)

- Now we can easily get our result. For example, given a grid, we know it's from block group a, its weight in W is W1, the total weight of a is W*, total population in block group a is P, then population in this grid will be estimated as W1*P/(W*).All the values in a grid can be computed in this way.

Result



Parallel Method

We would like to use distributed memory to implement the parallel method, so a state will be cut into several boxes. Two parts will be different from serial method. The first part is in determining the brief estimate of population density. Each box will calculate the β for itself. The values in β can be different for different boxes. Also, in the regression process, we need the total area of land cover c within zone s for every block group. However, for those block groups on the intersection between boxes, we cannot know this value for them. Thus we will find these block groups and make sure they will not participate in the regression. The second part is that we need to transfer the weight of cells of block groups on the intersection between boxes to its neighbourhood to get the total weight of that block group.

Due to more than 90% or more block groups will be ‘internal’ area(not on the intersection), and their population density can be calculated using the same method as the serial one inside each box, the data transferring on the boundary will not take a long time in general.

Future Work

For sure, the method above is not the only method. We can try polygon method instead of grid method. Also we plan to implement Prof. Nagle's PMEDM Method for dasymetric mapping and adapt it to become parallel. We will also extend our studying area into the whole U.S. After that, we will make analysis on the improvement of efficiency using parallel computing compared to doing serially.