High-Resolution Spatial Disaggregation using a Conditional Autoregressive Model

A Project Report Submitted in Partial Fulfilment of the Requirements of the course MTH598A for the Degree of

MASTER OF SCIENCE

in

STATISTICS

by

Anis Pakrashi[1]

(Roll No. 211264)

under the supervision of

Prof. Arnab Hazra[2]

to

DEPARTMENT OF MATHEMATICS AND STATISTICS

INDIAN INSTITUTE OF TECHNOLOGY

KANPUR

November, 2022

[1]Student, M.Sc. Statistics, IIT Kanpur

[2]Assistant Professor, Department of Mathematics and Statistics, IIT Kanpur

DECLARATION

I hereby declare that the work presented in the project report entitled **High-Resolution Spatial Disaggregation using a Conditional Autoregressive Model** contains my own ideas in my own words. At places, where ideas and words are borrowed from other sources, proper references, as applicable, have been cited. To the best of my knowledge, this work does not emanate from or resemble other work created by person(s) other than mentioned herein.

Name: **Anis Pakrashi**

Date: November 01, 2022

ACKNOWLEDGEMENT

I want to extend a sincere and heartfelt gratitude towards all the personages, without whom the succesful completion of the project would have been a distant dream. I express my profound thankfulness and deep regards to Professor Dr. Arnab Hazra, Assistant Professor, IIT Kanpur for his guidance, valuable feedback and constant encouragement throughout the project.

I am immensely grateful to Prof. Dr. Debasis Sen who has been the convener of this project and has allowed me to take up the project. I thank the review committee who has spent their valuable time behind evaluating my project. I also thank all the professors of the department for being the source of inspiration and helping me to acquire knowledge on diverse fields of the statistical realm.

I am also grateful to the people of Indian Institute of Human Settlements, who have provided me with the necessary data to work on.

Lastly, I thank my family and friends for being the continuous moral support, essential for smooth completion of the project.

Anis Pakrashi

ABSTRACT

With technological advancement, there has been a growing demand for spatially detailed data. Gradually, people have started going beyond low resolution data and have begun to take interest in the study of high resolution data. An analysis of only aggregated data may distort the true picture underlying the scenario. This project mainly aims at a methodology for high-resolution mapping using areal data and a conditional auto-regressive modeling. We consider different spatial fields, say 20 times 20 grids and 200 times 200 grids for the simulation. The independent and identically distributed (IID) model and the conditionally autoregressive (CAR) models are fitted to the aggregated data and compared. The fitted model has been used for estimating the pixel-level disaggregated values and also to demonstrate the predictive ability on another simulated cluster set. Finally, we use a ward-level population dataset on Bangalore city, where the ward level population values are known and we need to estimate the population values at the pixel level. In the dataset, we have a total of 786702 pixel values and a total of 198 wards. After the study we have seen that the estimated population values in different wards follow nearly the same pattern as in case of the data. However, there has been a significant overestimation in the values of population.

Keywords: spatial disaggregation, high resolution maps, spatial pixels, Ccnditionally autoregressive model, areal data

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 General Background

Spatial data is something which deals with data across various aspects but connected by a common link that is purely geographic in nature. *Areal data*[1] arise when a fixed region is partitioned into many sub-regions with aggregated outcomes. The primary requirements of spatial data study is that the data must be spatially correlated and observations close to a certain area are more similar as compared to observations which are distant. This is termed *spatial pattern*. Spatial data analysis poses serious problems in nature. Information collection about detailed processes using aggregation alone can be an uphill task in research involving geo-spatial data, which encompass different fields like forestry, agronomy, meteorology, public health, epidemiology, soil science and others. While the geographical space and several processes are continuous, numerous data provide a summarizing function of the underlying phenomena. The definition of boundaries depends on the problem under consideration, with census data being likely the most common setting. The procedure of converting data from a higher (or finer) to a lower (or coarser) resolution is called *aggregation*, and the collection of aggregated data can be motivated by technical, administrative and other physical constraints. The benefits of collecting aggregated data comes at a cost. Often, data are collected over large regions where heterogeneity is inherent. This hinders the process of making fine-scale inferences. It becomes almost completely impossible to conclude about the variations at a higher resolution level. This drawback is called *misalignment*[2]. The potential drawbacks of models for aggregated data have motivated the search for options to recover the original (pixel-level) information from the coarser resolution observations. Such a reverse procedure is called *spatial downscaling* or *disaggregation*. In all its applications, the disaggregated models generate a set of information at a higher spatial resolution from the data at a lower spatial resolution. (Refer to Figure 1.1 for illustration). In this project, we address the problem of disaggregating spatial data and focus on a particular model to do so, namely the conditional autoregressive (CAR) model and also include a few applications of Bayesian approaches

[1] also known as lattice data
[2] also called Modifiable Areal Unit Problem (MAUP)

to reach our goal. Then, we have used a large dataset provided by Indian Institute of Human Settlements (IIHS), which is mainly based on population data in Bangalore City, and applied spatial diaggregation on the data.

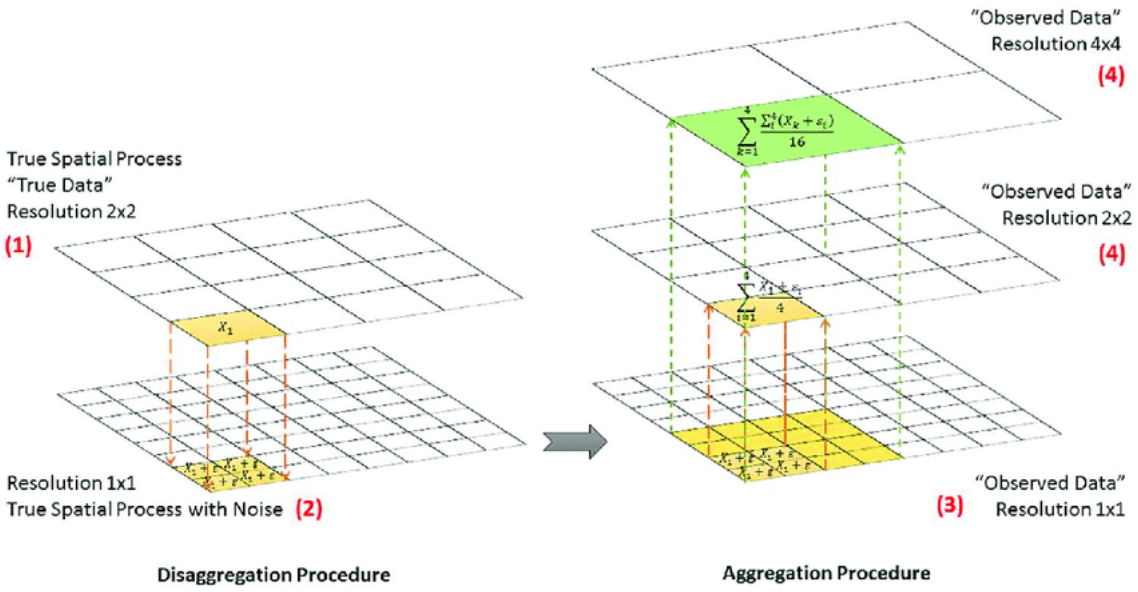


Figure 1.1: *Aggregation and Disaggregation Procedure. Example of a spatial process with true level = 4x4 resolution. (doi:10.1371/journal.pone.0167945.g003)*

1.2 Objectives

The main objectives of this project :

- High resolution spatial simulation assuming that true rates are observed and estimation of the model parameters
- simulation from a spatial structure, assuming that the rates are known only after some aggregation and then fitting a CAR model to the aggregates. Try to estimate the values at a higher resolution and predict the aggregates for a different set of clusters
- For a ward-level population dataset of Bangalore city, fitting a disaggregation model (under certain assumptions) to find out the population values at a finer resolution

Chapter 2

Mathematical and Statistical Background

2.1 Disaggregation Modelling

Suppose we have response data, y_i, for N polygons, which corresponds to data for the quantity of interest within that polygon. The process that is being measured occurs in continuous space that we model as a high-resolution, square lattice. The data, y_i, are assumed to be created by the aggregation of the response value over the finer units of the polygon, i.e. the data value of the polygon is given by the sum of the data values for all the pixels within that polygon. The *rate* is defined such that the number of cases in this pixel can be calculated by multiplying the rate by the aggregation raster. For the disaggregation model, we model the rate at pixel level, with the likelihood for the observed data given by aggregating these pixel level rates. The rate in pixel j of polygon i at location s_{ij} is given by:

$$link(rate_{ij}) = \beta_0 + \beta_1 X_{ij} + GP(s_{ij}) + u_i \tag{2.1}$$

where, β are the regression coefficients, X_{ij} are the covariate values, GP is a Gaussian random field and u_i is a polygon-specific iid effect. The link function is considered to be normal link in our case. The Gaussian random field has a Matern covariance function parameterised by ρ, the range and σ. The predictions at the pixel level are then aggregated to the polygon level, by weighted sum:

$$cases_i = \sum_{j=0}^{N_i} a_{ij} rate_{ij} \tag{2.2}$$

where, a_{ij} is the aggregation raster[1].

$$rate_i = \frac{cases_i}{\sum_{j=0}^{N_i} a_{ij}} \tag{2.3}$$

[1] helps to create a new raster layer with lower (coarser) resolution, by aggregating values over pixels.

Here, we use the Gaussian likelihood:

$$y_i \sim Normal(cases_i, \sigma_i) \tag{2.4}$$

where, $\sigma_i = \sigma\sqrt{\sum_j a_{ij}}$, σ being the pixel level variance.

2.2 Conditional Auto-Regressive (CAR) Model

Consider $\boldsymbol{Y} = (y_1, y_2, ..., y_n)$ and consider the set $\{p(y_i|y_j, j \neq i\}$. We define a **proximity matrix** or an **adjacency matrix** W with entries $w_{ij}(w_{ii} = 0)$ such that the choices of w_{ij} can be -

- $w_{ij} = 1$, if i, j share a common boundary, possibly a common vertex
- w_{ij} is an inverse distance between units
- $w_{ij} = 1$, if distance between units is less than K, for some pre-fixed K
- $w_{ij} = 1$, for m nearest neighbours

Consider the Gaussian case: $\{p(y_i|y_j, j \neq i\} = \text{Normal}\left(\sum_j b_{ij} y_i, \tau^2\right)$

Using Brook's Lemma, we can obtain

$$p(y_1, y_2, ..., y_n) \propto exp\left\{-\frac{1}{2} y^{'} D^{-1}(I - B)y\right\} \tag{2.5}$$

where, $B = \{b_{ij}\}$ and $D_{ii} = \tau_i^2$

The essential idea here is that the probability of values estimated at any given location are conditional on the level of neighboring values. The standard CAR model for the expectation of a specific observation, y_i, is of the form:

$$\{E(y_i|y_j, j \neq i\} = \mu_i + \rho \sum_{j \neq i} w_{ij}(y_j - \mu_j) \tag{2.6}$$

where μ_i is the expected value at i, and ρ is a spatial autocorrelation parameter that determines the size and sign of the spatial neighborhood effect. The summation term is the weighted sum of the mean adjusted values at all other locations j.

Chapter 3

Simulation Study

3.1 Theory behind the study

Suppose there are P pixels in a rectangular spatial grid, such that the region has the pixels arranged in N rows and N columns. For N=20, refer to Figure 3.1 below.

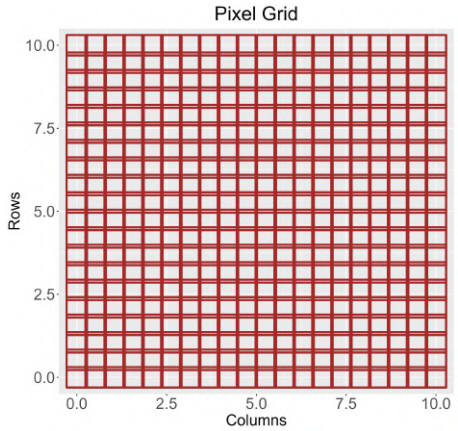


Figure 3.1: Illustration of spatial pixel structure

Let A denote the adjacency matrix at the pixel level. Define

$$M = \begin{bmatrix} \sum_j A_{1j} & 0 & 0 & \dots & 0 \\ 0 & \sum_j A_{2j} & 0 & \dots & 0 \\ 0 & 0 & \sum_j A_{3j} & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & \dots & \sum_j A_{Pj} \end{bmatrix} \tag{3.1}$$

Consider a correlation coefficient of ρ and a dispersion of σ^2. Now, we have a P-component error vector, which follows a P-variate normal distribution with mean vector $\mathbf{0}$ and dispersion matrix $\sigma^2(M - \rho A)^{-1}$, i.e.

$$\boldsymbol{\varepsilon} \sim N_P\left(\mathbf{0}, \sigma^2(\mathbf{M} - \rho\mathbf{A})^{-1}\right) \tag{3.2}$$

Let $\mathbf{y} = (y_{1,1}, y_{1,2}, \ldots, y_{1,n_1}, y_{2,1}, y_{2,2}, \ldots, y_{2,n_2}, \ldots, y_{N,1}, y_{N,2}, \ldots, y_{N,n_N})'$ denote the vector of unobserved rates. Let $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)'$ denote the vector of aggregate outcomes of the N blocks or clusters. The problem is to estimate $y_{i,j}$, with respect to the following constraints:

- All $y_{i,j}$ are constrained in some interval, $a < y_{i,j} < b$
- $\mathbf{y}$ is a function of one or more explanatory variables
- $\mathbf{y}$ is potentially spatially-correlated, i.e., closer pixels are more likely to have similar values than farther pixels

Consider X to be the matrix of covariate values, such that each pixel has one or more covariate values attached to it. Each column of the matrix indicates a covariate.

We use *Cholesky Decomposition*[1] to obtain the true rates, using the backsolve technique. For this purpose, we use the **chol()** function in R. Note that, **we do not observe the true rates**.

Now, in order to generate a region with aggregated pixels, we use *k-means clustering*[2] and obtain a set of K clusters. These clusters will act as the only units for which the values of response will be known. Now, we aggregate the rates as per the clusters and obtain the case values of the clusters. Finally, we have our response values as aggregates of the true rates, as per the cluster structure. The responses are induced with a randomness with varability τ^2.
Note that the values of Y are the observed values that we have at hand.

Using numerical methods, we obtain the maximum likelihood estimators of ρ and σ^2 and hence obtain the estimates of β_0 and β_1. Finally we obtain the estimates of the rates at pixel level from the above information and estimates. We consider the estimates as the mean of the posterior distribution of the rates.

The prior distribution of the rates is as follows:

$$\mathbf{R} \sim N_P\left(\mathbf{X}\beta, \sigma^2(\mathbf{M} - \rho\mathbf{A})^{-1}\right) \tag{3.3}$$

The likelihood function is as follows:

$$\mathbf{Y}|\mathbf{R} \sim N_K\left(B\mathbf{R}, \tau^2 I\right) \tag{3.4}$$

The posterior distribution is also obtained as a multivariate normal distribution with mean vector $\mathbf{m}$ and covariance matrix Σ, where

[1]The **Cholesky decomposition** of a Hermitian positive-definite matrix A is a decomposition of the form $A = L.L^T$, where L is a lower triangular matrix with real and positive diagonal entries, and L^T denotes the conjugate transpose of L. Every Hermitian positive-definite matrix (and thus also every real-valued symmetric positive-definite matrix) has a unique Cholesky decomposition.
[2]K-means clustering is an unsupervised learning algorithm that groups an unlabelled dataset into groups

$$\Sigma = \left(\frac{\mathbf{B}'\mathbf{B}}{\tau^2} + \frac{\mathbf{M} - \rho\mathbf{A}}{\sigma^2} \right)^{-1} \tag{3.5}$$

$$\mathbf{m} = \left(\frac{\mathbf{B}'\mathbf{B}}{\tau^2} + \frac{\mathbf{M} - \rho\mathbf{A}}{\sigma^2} \right)^{-1} \left(\frac{\mathbf{B}'\mathbf{Y}}{\tau^2} + \frac{(\mathbf{M} - \rho\mathbf{A})\mathbf{X}\beta}{\sigma^2} \right) \tag{3.6}$$

Thus,

$$\mathbf{R}|\mathbf{Y} \sim N_P(\mathbf{m}, \Sigma) \tag{3.7}$$

The estimated rates at pixel level are thus obtained as the posterior mean of the distribution in Equation 3.9 above.

Let there be another set of clusters for which we wish to find the aggregated values of the variable under study. Let the number of such clusters be L. Note that, for this set of clusters, we may compute the aggregate values of our response, using the pixel-level estimates. Also, we find out the true aggregates using the original rates (which were assumed to be unknown in the process), for the sake of comparison.

3.2 Parameter Estimation for small and simple Spatial Structure

We consider a 20 x 20 spatial grid (with 400 pixels). Each pixel has two covariates associated with it. The true values for the following quantities are to be assumed as follows:

The correlation parameter, ρ is 0.5

The dispersion parameter, σ^2 is 1

The three coefficients $\beta_0, \beta_1, \beta_2$ of the model are 5, 1, 2 respectively.

Using the conditional autoregressive model, the true rates are generated. Here, we assume that we observe the true rates. Consider the Figure 3.2 below:

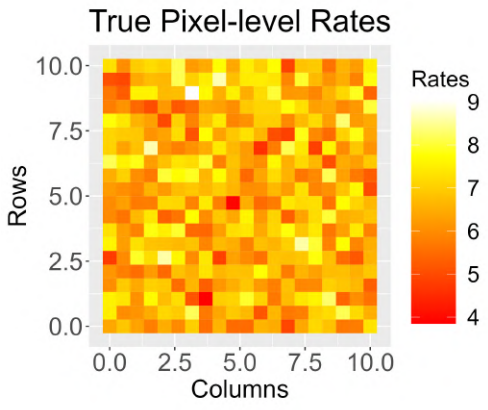


Figure 3.2: Illustration of True outcomes at the high resolution level (400 pixels)

Assume that,

$$\mathbf{Y} \sim N\left(\mathbf{X}\beta, \sigma^2(\mathbf{M} - \rho\mathbf{A})^{-1}\right) \tag{3.8}$$

Using numerical optimization, with the help of **optim()**, function in R, we obtain the estimates of ρ, σ^2, β_0, β_1, β_2. Note that the **optim()** function minimizes the objective the function in its code-block. Suppose, a function is dependent on ρ only. Within the code-block, we obtain the parameter coefficients of the model and the model variance and finally the negative log-likelihood.

We repeat the entire process several times (say, 250 times) to obtain *statistical regularity*. The following figures illustrate the summary of the estimates over the iterations:

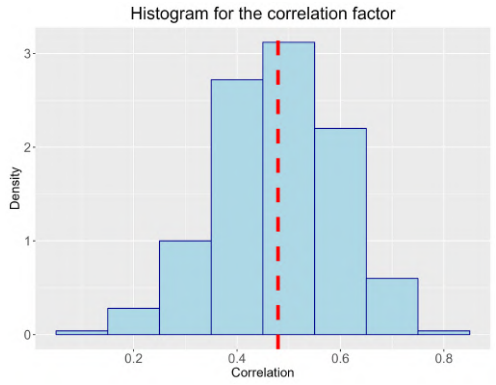


Figure 3.3: Distribution of Values of ρ (20 x 20 grid)

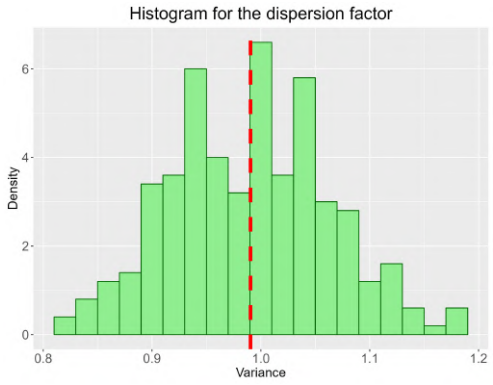
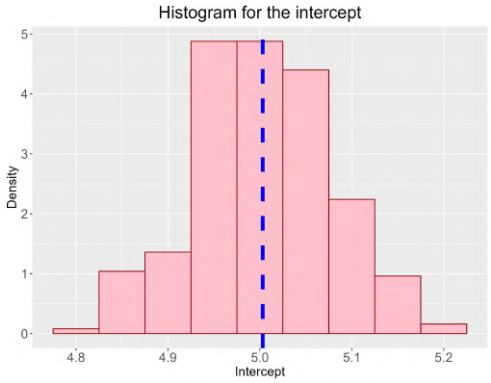
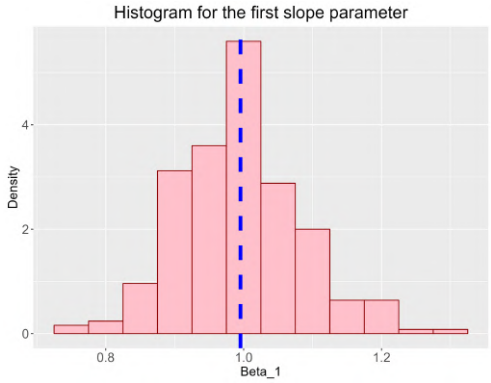


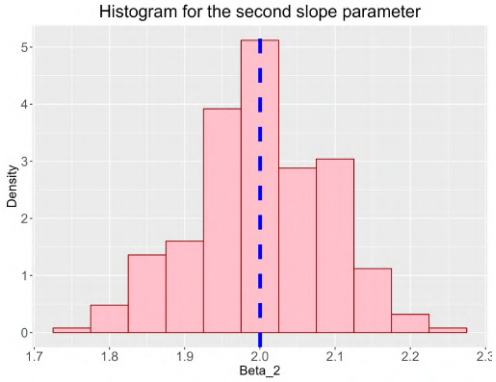
Figure 3.4: Distribution of Values of σ^2 (20 x 20 grid)



(a) Values of β_0



(b) Values of β_1



(c) Values of β_2

Figure 3.5: Distribution of values of the model parameters (20 x 20 grid)

The table below (Table 3.1) compares the true values along with the estimates.

Parameter	True Value	Estimate
β_0	5	5.003126
β_1	1	0.995414
β_2	2	2.000348
σ^2	1	0.9904422
ρ	0.5	0.4792831

Table 3.1: Comparison between True and Estimated values of the parameters for 20 x 20 grid

From the figures and the table, it is apparent that the estimates lie close to the true value of the parameters. The CAR model performs quite well in case the grid size is small and the model is simple and this is quite apparent from the figures and table.

3.3 Parameter Estimation for large Spatial Structure

We consider a 200 x 200 spatial grid (with 40000 pixels). This is primarily an extension of the small grid case in the previous section. Each pixel has two covariates associated with it. The true values for the following quantities are to be assumed as follows:

The correlation parameter, ρ is 0.5

The dispersion parameter, σ^2 is 1

The three coefficients $\beta_0, \beta_1, \beta_2$ of the model are 5, 1, 2 respectively.

Using the conditional autoregressive model, the true rates are generated. Here too, we assume that we observe the true rates. Consider the Figure 3.6 below:

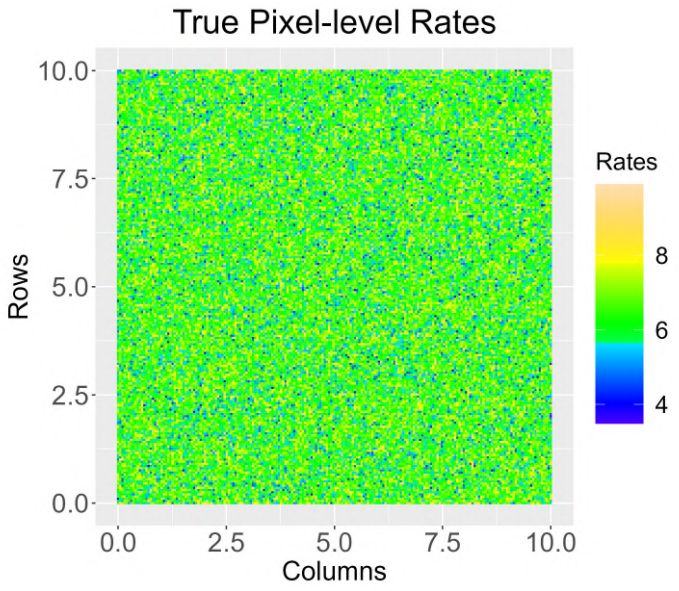


Figure 3.6: Illustration of True outcomes at the high resolution level (40000 pixels)

Assume that,

$$\mathbf{Y} \sim N\left(\mathbf{X}\beta, \sigma^2(\mathbf{M} - \rho\mathbf{A})^{-1}\right) \tag{3.9}$$

The large grid case cannot be handled in R as easily as the small grid case. The usual functions for storing matrices and other high dimensional data structures will not work as the memory requirement is too high. Here, we introduce the Sparse Matrix concept, which allocates memory only to non-zero values in the matrix, resulting in lesser memory consumption. This is implemented using the **Matrix** and **spam** packages.

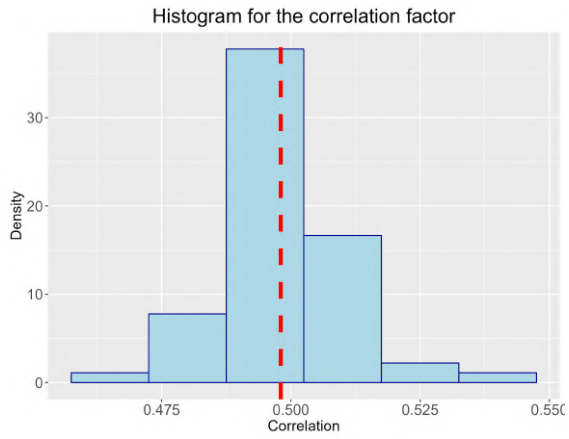
Using numerical optimization, with the help of **optim()**, function in R, we obtain the estimates of ρ, σ^2, β_0, β_1, β_2. Note that the **optim()** function minimizes the objective the function in its code-block. Suppose, a function is dependent on ρ only. Within the code-block, we obtain the parameter coefficients of the model and the model variance and finally the negative log-likelihood.

We repeat the entire process several times (say, 60 times) to obtain *statistical regularity*. The table below (Table 3.2) compares the true values along with the estimates.

Parameter	**True Value**	**Estimate**
β_0	5	4.9991144
β_1	1	0..9999662
β_2	2	2.0015976
σ^2	1	1.000835
ρ	0.5	0.4980341

Table 3.2: Comparison between True and Estimated values of the parameters for 200 x 200 grid

Consider the following figures, which illustrate the summary of the estimates over the iterations, when the spatial area considered is large:



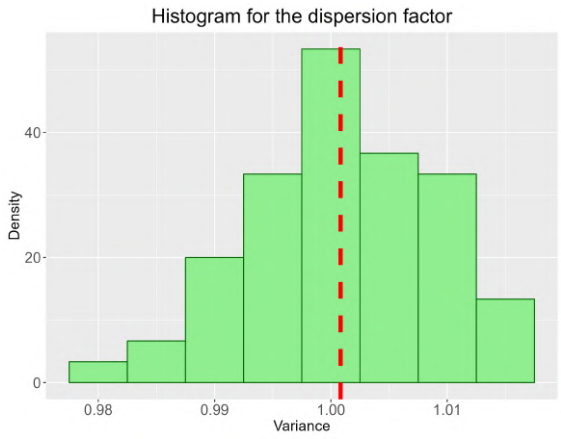
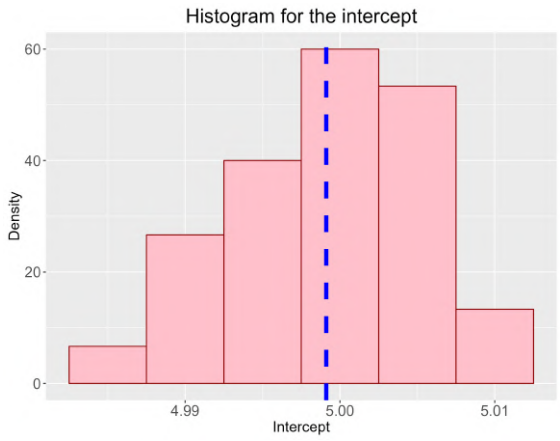
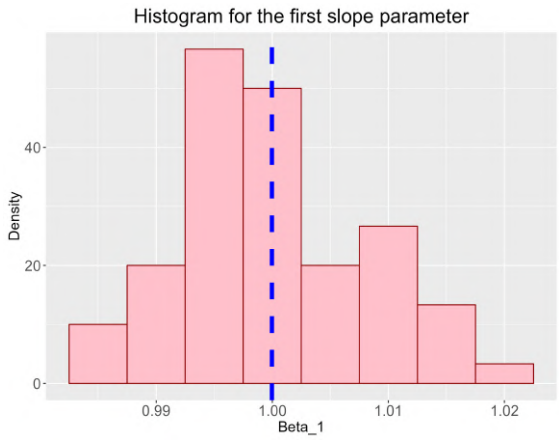


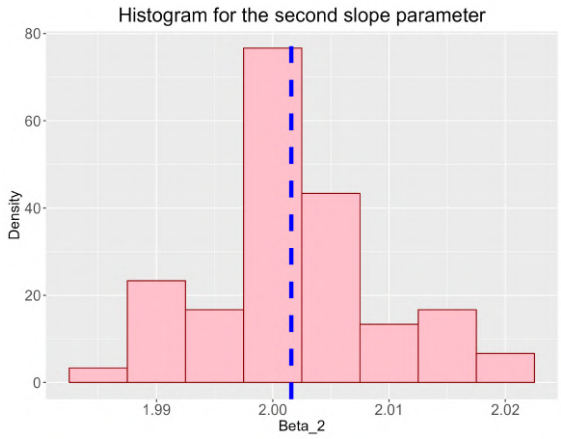
Figure 3.7: Distribution of Values of ρ (200 x 200 grid)Figure 3.8: Distribution of Values of σ^2 (200 x 200 grid)



(a) Values of β_0



(b) Values of β_1



(c) Values of β_2

Figure 3.9: Distribution of values of the model parameters (200 x 200 grid)

From the figures and the table, it is apparent that the estimates lie close to the true value of the parameters. The CAR model performs well even in case of large grid size.

3.4 Disaggregation modelling in small Spatial Structure

The setup is similar to Section 3.2. We have 20 x 20 grid structure with 400 pixels. Each pixel has two covariates associated with it. The true values for the following quantities are to be assumed as follows:

The correlation parameter, ρ is 0.5

The dispersion parameter, σ^2 is 1

The three coefficients $\beta_0, \beta_1, \beta_2$ of the model are 5, 1, 2 respectively.

We have another parameter, τ^2 which gives an idea of the variabilty in the response. The true value is 0.1, for this parameter.

Using the conditional autoregressive model, the true rates are generated. Now, we *donot* observe the true rates in this case. Hence it is more like a realistic situation where the finer resolution outcomes remain unobserved. A k-means clustering algorithm is implemented, with $K = 100$. Refer to Figure 3.10 below for an illustration :

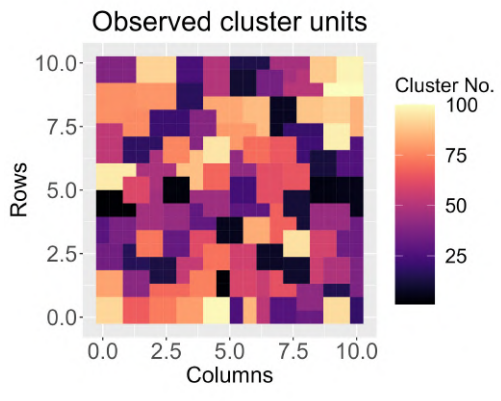


Figure 3.10: Illustration of Blocking or clustering (20 x 20 grid)

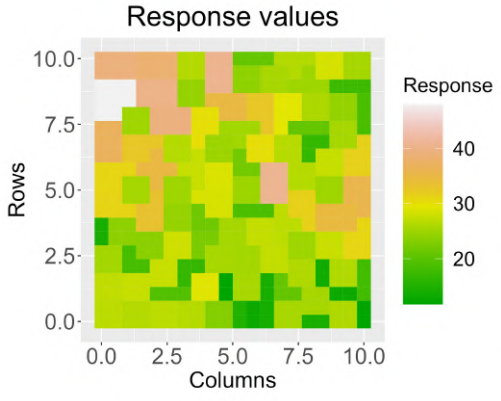


Figure 3.11: Distribution of response values across clusters (20 x 20 grid)

Now, the **observed** responses are the aggregates of the true rates as per the cluster structure. A randomness is assigned to the response by adding $rnorm(K, 0, \tau^2)$. Consider Figure 3.11 above.

At this juncture, we consider two modelling frameworks:

1. the Independently Distributed(ID) model :

$$\mathbf{Y} \sim N\left(\mathbf{X}\beta, \sigma^2\mathbf{I}\right) \tag{3.10}$$

2. the CAR model :

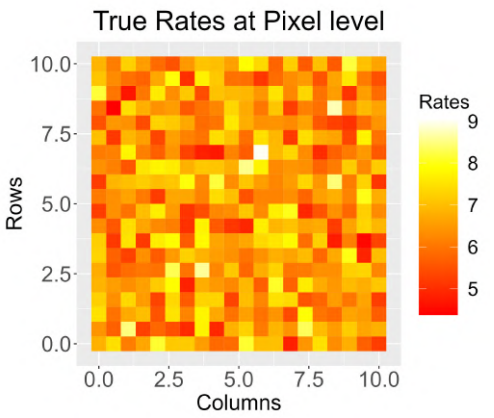
$$\mathbf{Y} \sim N\left(\mathbf{X}\beta, \sigma^2(\mathbf{M} - \rho\mathbf{A})^{-1}\right) \tag{3.11}$$

We model using both the above situations in parallel. Using 200 iterations, we summarize our estimates as shown below in Table 3.3 in the next page. We observe that the two models perform relatively close-by for small spatial samples.

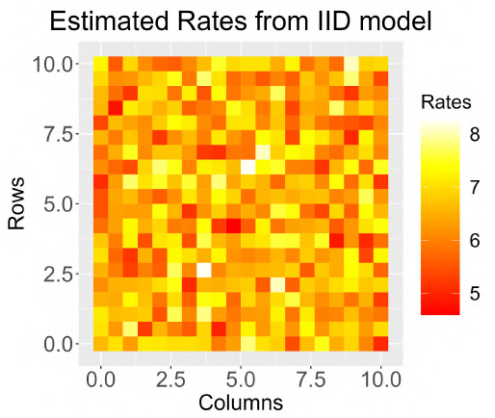
Next, we obtain the estimated rates (**disaggregated values**) as the mean of the posterior distribution of the Rates (As explained in section 3.1 before). The rates for both the models are stored for future comparisons and re-aggregations.

Parameter	**True Value**	**IID Model Estimate**	**CAR Model Estimate**
β_0	5	4.971995	4.972366
β_1	1	1.013098	1.012527
β_2	2	2.044339	2.044381
σ^2	1	0.3227822	0.3490421
ρ	0.5	-	0.4671472
τ^2	0.1	0.2676889	0.1707561

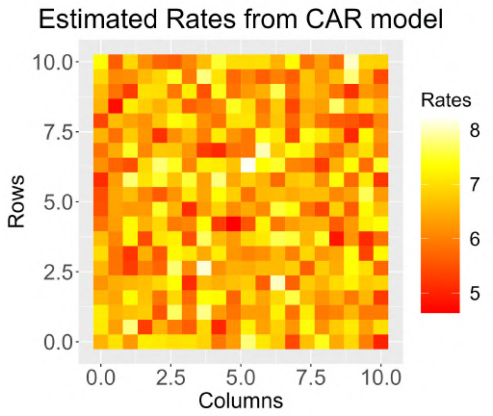
Table 3.3: Comparison among True value and Estimates of the parameters for 20 x 20 grid for disaggregation modelling



(a) True Rates



(b) IID Model Estimates



(c) CAR Model Estimates

Figure 3.12: True Values and Estimates of pixel rates (20 x 20 grid)

We go on to use the estimated rates for predicting the aggregate values for an entirely new pattern of blocking. We now obtain this new set of clusters by k-means clustering algorithm with $K = 75$, i.e., a reduced number of blocks. Consider the illustration in the next page (Figure 3.13):

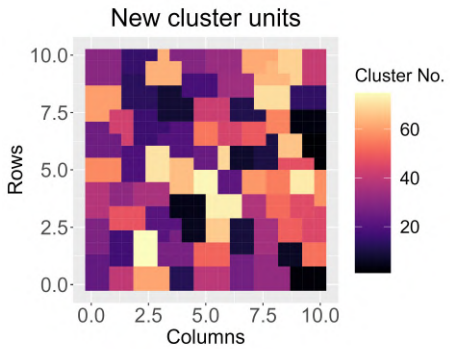
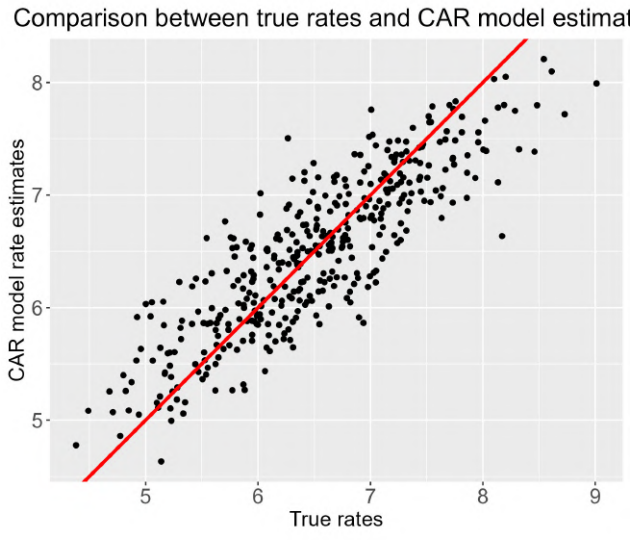


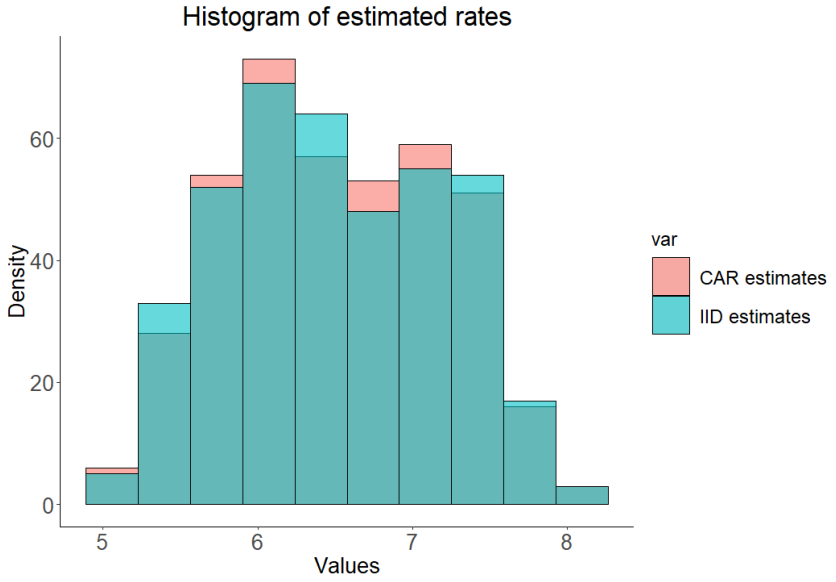
Figure 3.13: A new pattern of Blocking for checking re-aggregation (20 x 20 grid)

The following figures in the next page show the relations between the true model rates and estimated model rates. Refer to Figure 3.14 :

(a) True vs IID estimates



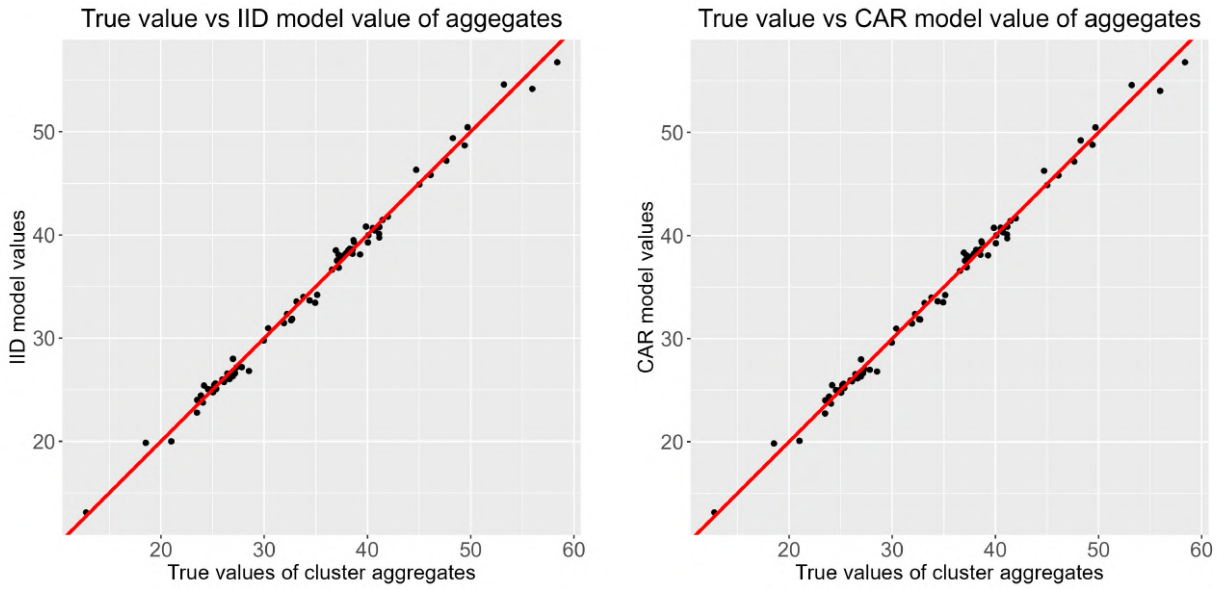
(b) True vs CAR estimates



(c) Histogram for the estimates

Figure 3.14: Relation between true and estimated rates (20 x 20 grid)

Under the present setup of new clusters, we compute the block values for the true case, IID estimate case and the CAR estimate case. The relation of the estimated aggregates with the true aggregates are given below:



(a) True vs IID aggregates (b) True vs CAR aggregates

Figure 3.15: Relation between true and estimated aggregates (20 x 20 grid)

3.5 Disaggregation modelling in large Spatial Structure

The setup is similar to Section 3.3. We have 200 x 200 grid structure with 40000 pixels. Each pixel has two covariates associated with it. The true values for the following quantities are to be assumed as follows:

The correlation parameter, ρ is 0.5

The dispersion parameter, σ^2 is 1

The three coefficients $\beta_0, \beta_1, \beta_2$ of the model are 5, 1, 2 respectively.

We have another parameter, τ^2 which gives an idea of the variabilty in the response. The true value is 0.1, for this parameter.

Using the conditional autoregressive model, the true rates are generated. As in the small sample case, a k-means clustering algorithm is implemented, with $K = 200$. Refer to Figure 3.16 below for an illustration :

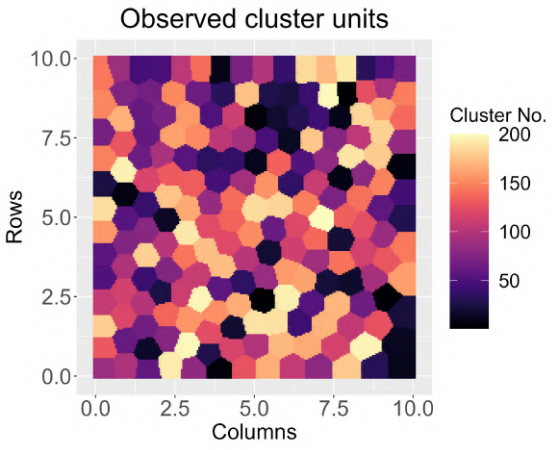


Figure 3.16: Illustration of Blocking or clustering (200 x 200 grid)

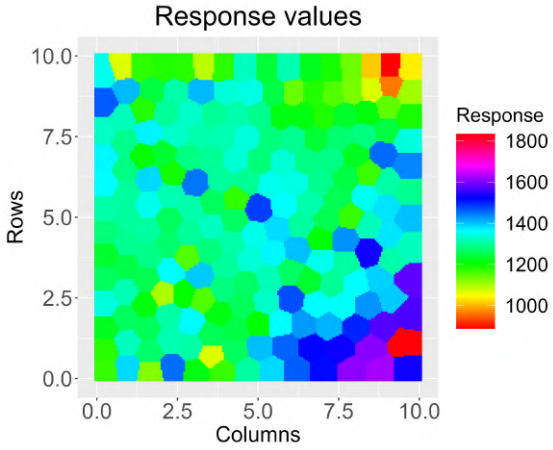


Figure 3.17: Distribution of response values across clusters (200 x 200 grid)

Now, the **observed** responses are the aggregates of the true rates as per the cluster structure. A randomness is assigned to the response by adding $rnorm(K, 0, \tau^2)$. Consider Figure 3.17 above.

In this scenario, we consider the ID modelling framework (The CAR model is difficult to implement for large spatial data and the software breaks down in the process of handling the heavy data. We have left this case to be probed further in future):

$$Y \sim N\left(X\beta, \sigma^2 I\right) \tag{3.12}$$

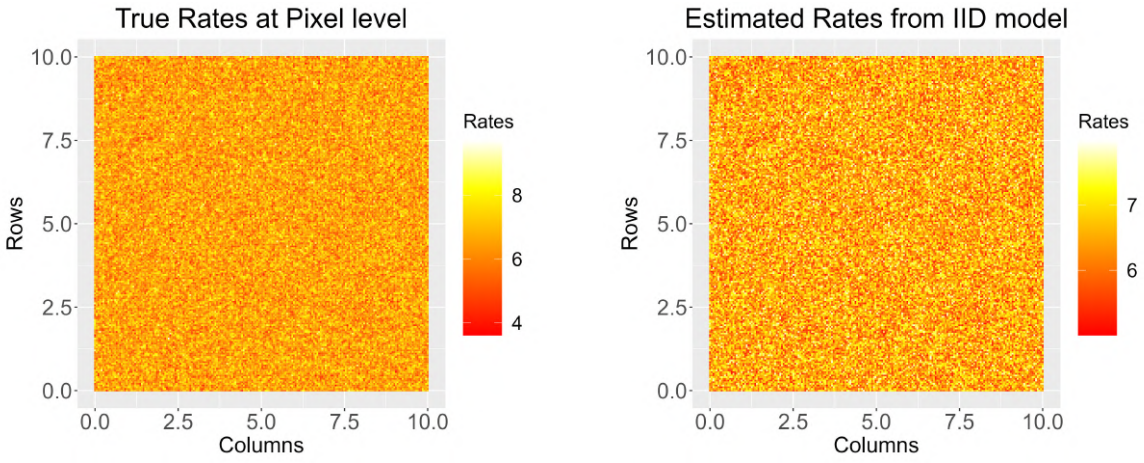
We model using the above situations. Using 100 iterations, we summarize our estimates as shown below in Table 3.3 in the next page. We observe that the two models perform relatively close-by for small spatial samples.

Next, we obtain the estimated rates (**disaggregated values**) as the mean of the posterior distribution of the rates:

$$\mathbf{m} = \left(\frac{B'B}{\tau^2} + \frac{I_P}{\sigma^2}\right)^{-1}\left(\frac{B'Y}{\tau^2} + \frac{X\beta}{\sigma^2}\right) \tag{3.13}$$

Parameter	**True Value**	**IID Model Estimate**
β_0	5	5.0464074
β_1	1	0.9719178
β_2	2	1.9357025
σ^2	1	0.3433005
τ^2	0.1	24.7076

Table 3.4: Comparison among True value and Estimates of the parameters for 200 x 200 grid for diaggregation modelling



(a) True Rates (b) IID Model Estimates

Figure 3.18: True Values and Estimates of pixel rates (200 x 200 grid)

On similar lines as the previous section, we use the estimated rates for predicting the aggregate values for an entirely new pattern of blocking. We now obtain this new set of clusters by k-means clustering algorithm

with $K = 150$, i.e., a reduced number of blocks. Consider the illustration in Figure 3.19:

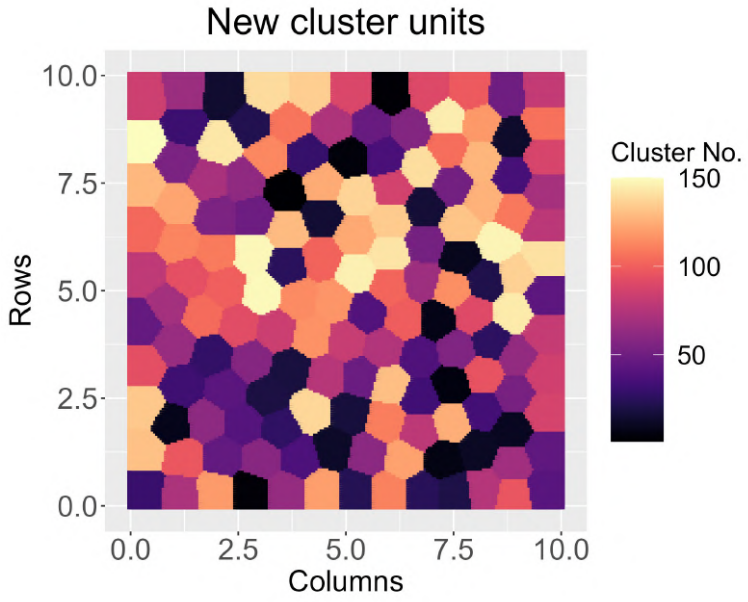


Figure 3.19: A new pattern of Blocking for checking re-aggregation (200 x 200 grid)

The following figure shows the relation between the true model rates and estimated model rates. A regression line is passed through the scatterplot. Refer to Figure 3.20 :

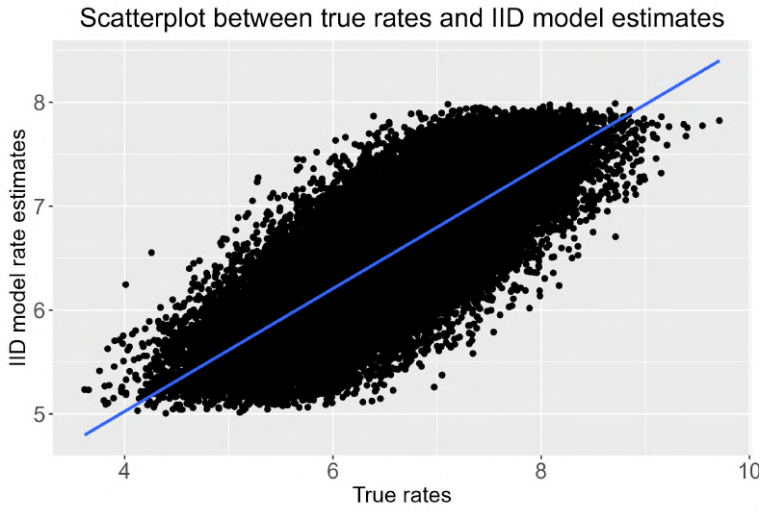


Figure 3.20: Scatterplot with Regression Fit for true vs estimated rates

Under the present setup of new clusters, we compute the block values for the true case and the IID estimate case. The relationship between the estimated aggregates and the true aggregates are given below:

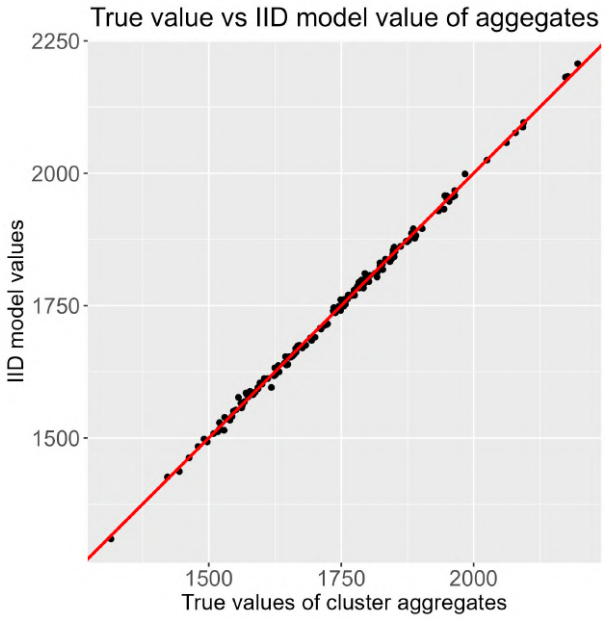


Figure 3.21: Relation between true and estimated aggregates (200 x 200 grid)

Chapter 4

Data Study

4.1 Data Source

Our objective is to use covariates at the resolution of 30m X 30m cells to disaggregate the population of Bangalore city. Specifically, we do this for Bruhat Bengaluru Mahanagara Palike (BBMP), the administrative body for the Bangalore metropolitan area. We use the 2011 Census data for BBMP, which was divided into 198 wards. The data has been taken from the database of **Indian Institute of Human Settlements**[1].

The data comprises 3 response variables in total, but we focus on the population counts only. We consider all the covariates.

4.2 Data Description

4.2.1 Covariates

1. **2011landcover@1** : Land cover category [1: Built-up; 2: Vegetation; 3: Water; 4: Vacant]
2. **residential@1** : Binary indicator of land use [1: Residential; 0: Non-residential]
3. **StreetDensity@1** : Continuous measure of steet density in the pixel
4. **BuildingHeight@1** : Building height (in metres), estimated from stereo imagery
5. **BuiltCount@1** : Indicator of number of sub-pixels (5m x 5m) that are built-up
6. **VegetationCount@1** : Indicator of number of sub-pixels (5m x 5m) which are covered with vegetation

[1] IIHS is a national education, research, practice capacity development institution, committed to the transformation of Indian cities settlements

7. **VacantCount@1** : Indicates how many sub-pixels (5m x 5m) are vacant

8. **flowAcc@1** : Continuous measure indicating density of drainage network in the pixel

Refer to the maps below for a comprehensive plot of the covariate layers:

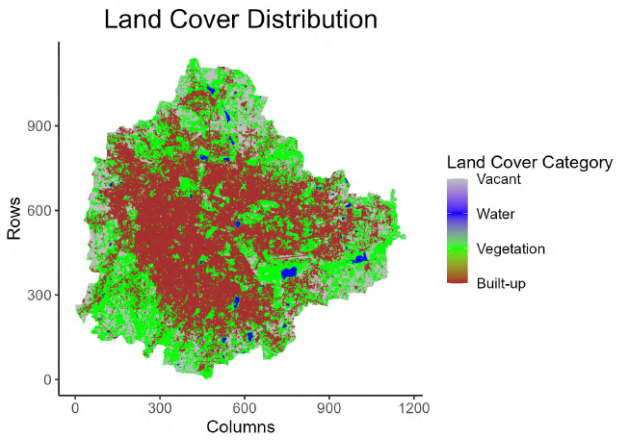


Figure 4.1: Plot of Land Cover in Bangalore

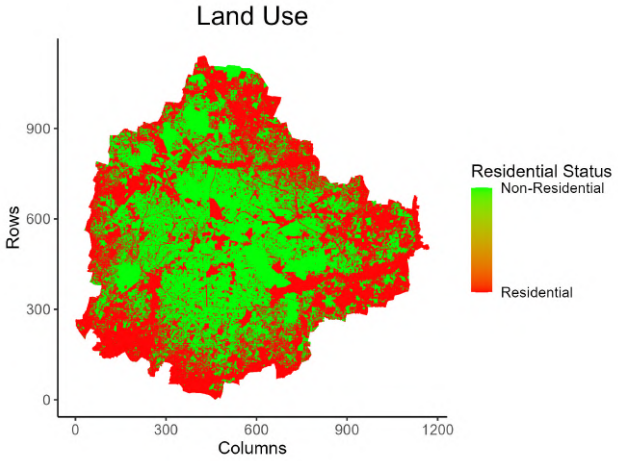


Figure 4.2: Plot of Land Use in Bangalore

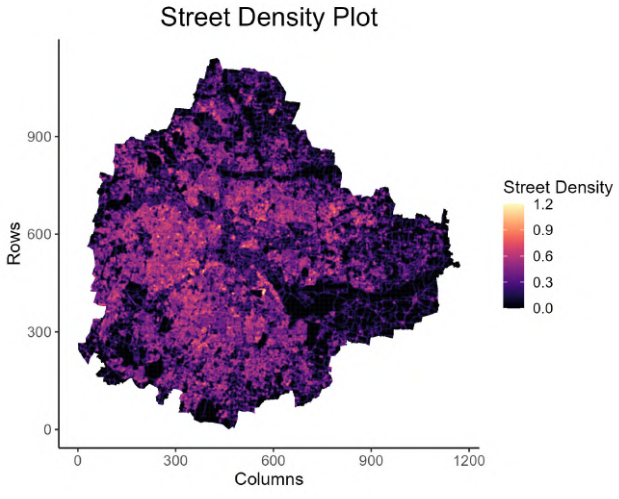


Figure 4.3: Plot of Street Density in Bangalore

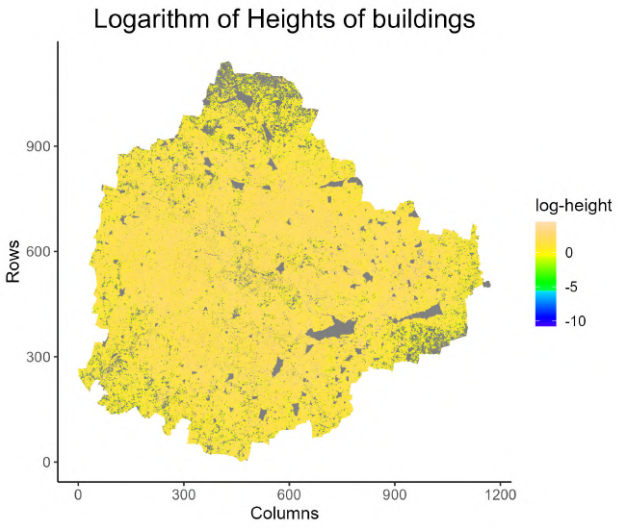


Figure 4.4: Plot of Building Heights in Bangalore

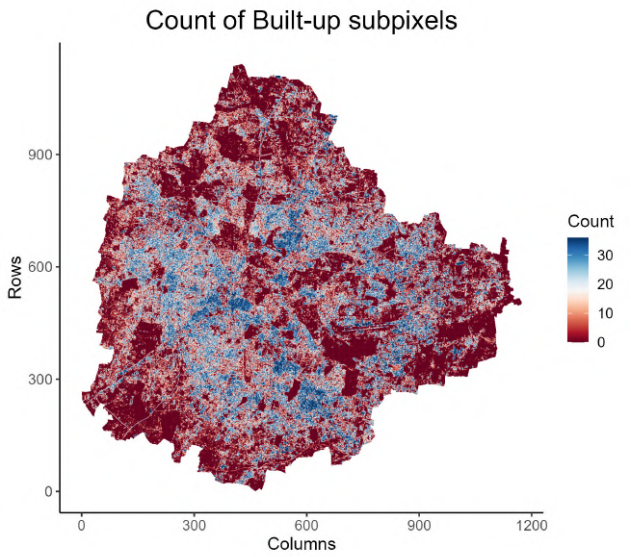


Figure 4.5: Indication of built-up sub-pixels

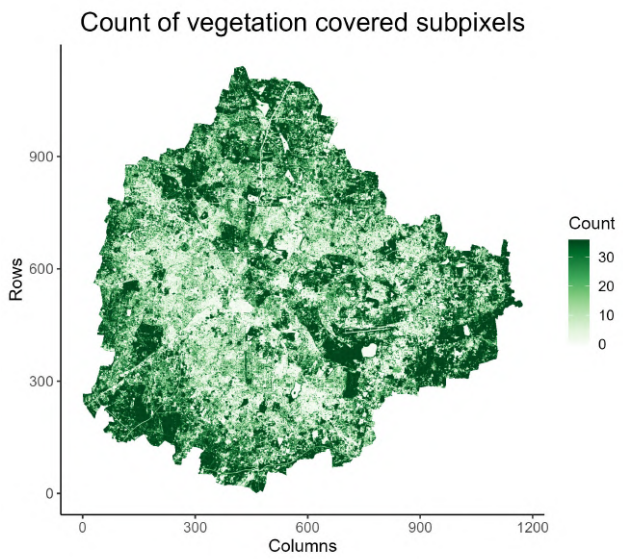


Figure 4.6: Indication of sub-pixels with vegetation cover

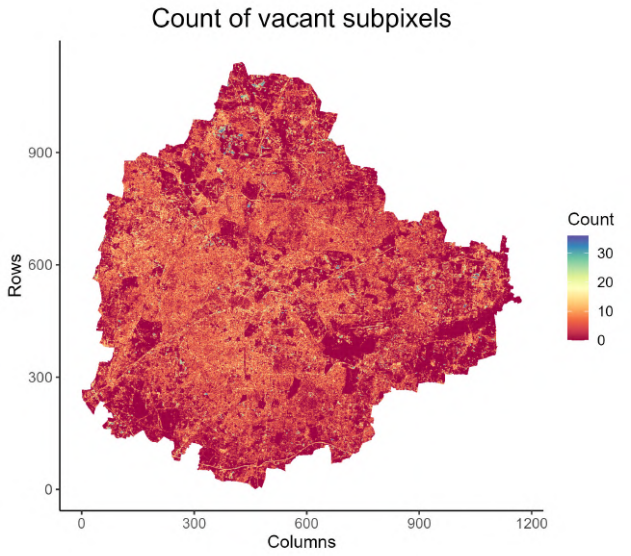


Figure 4.7: Indication of vacant sub-pixels

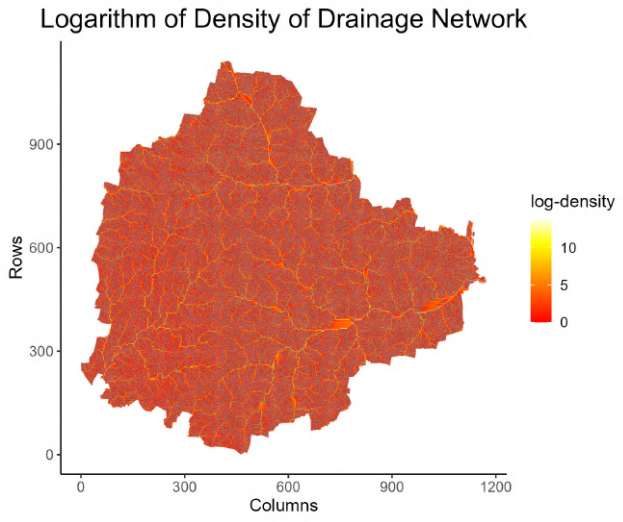


Figure 4.8: Measure of drainage density

4.2.2 Response

The response variable under study is the population count in the wards, as per the National Census, 2011. In fact, it is the only observable value of the response that we are provided with. The variable is denoted by **BBMPPopulation@1**.

We are also given a variable named **BBMPWard@1**, which is a useful indicator of the ward to which the pixel belongs.

In our dataset, we have a total of **786702** pixel values and a total of **198** wards, into which the pixels are aggregated.

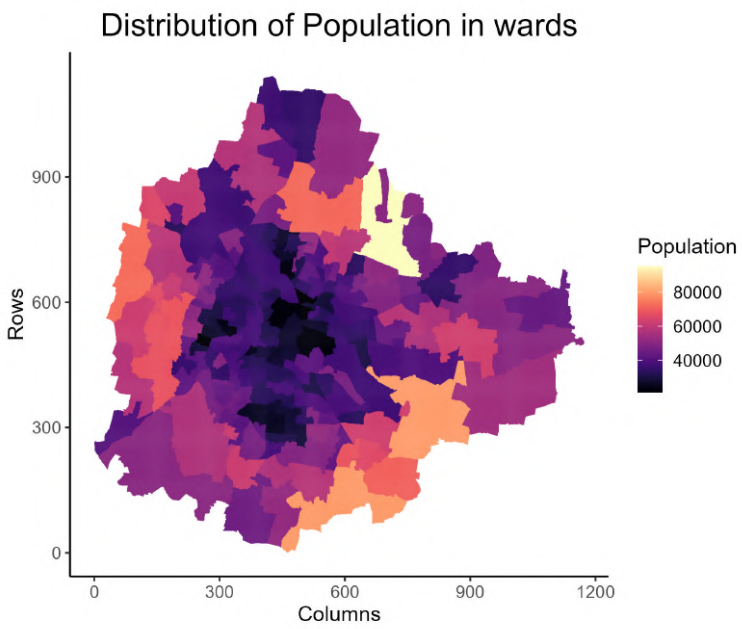


Figure 4.9: Ward level population counts in Bangalore

4.2.3 Relation between response and covariates

We introduce scatterplots between the response and each of the covariates at the ward level to show the true relationship between the pair of variables:

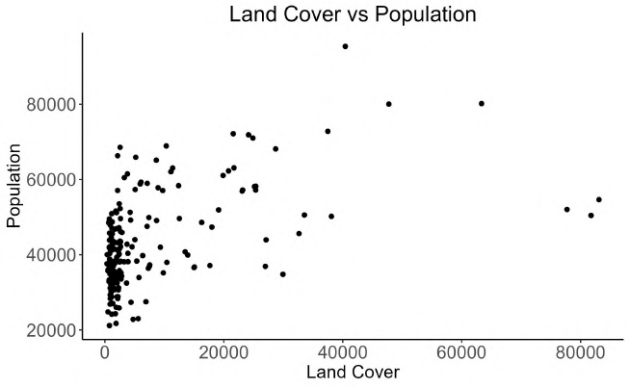


Figure 4.10: Land Cover vs Population

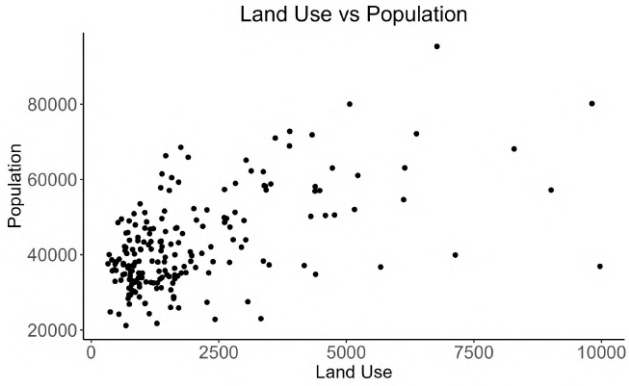


Figure 4.11: Land Use vs Population

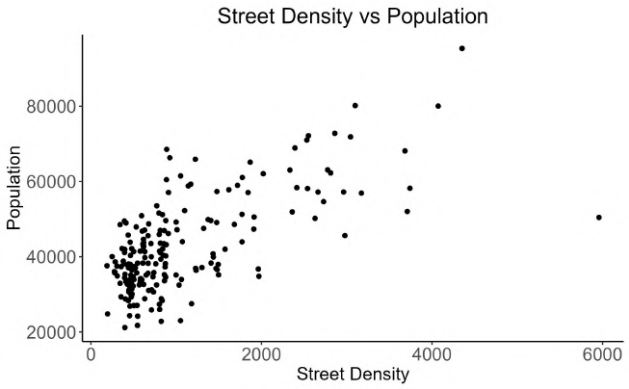


Figure 4.12: Street Density vs Population

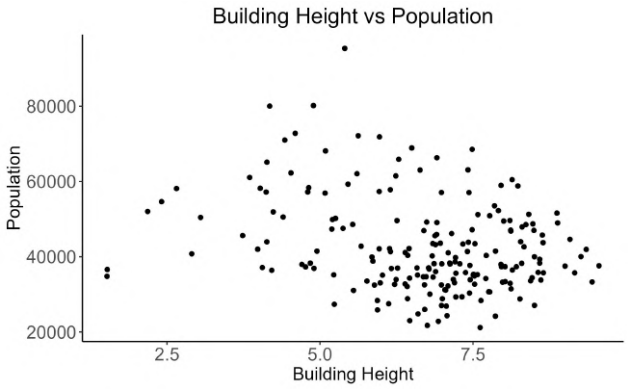


Figure 4.13: Building Height vs Population

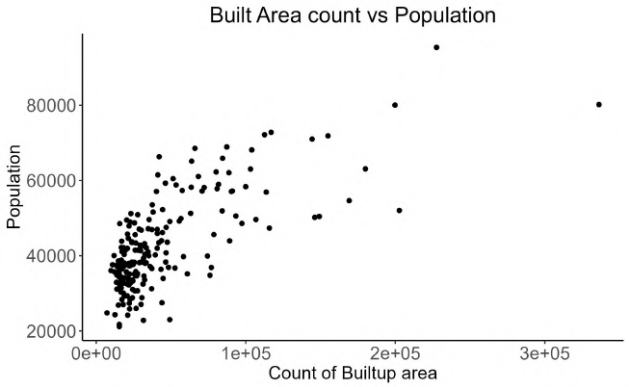


Figure 4.14: Built Area count vs Population

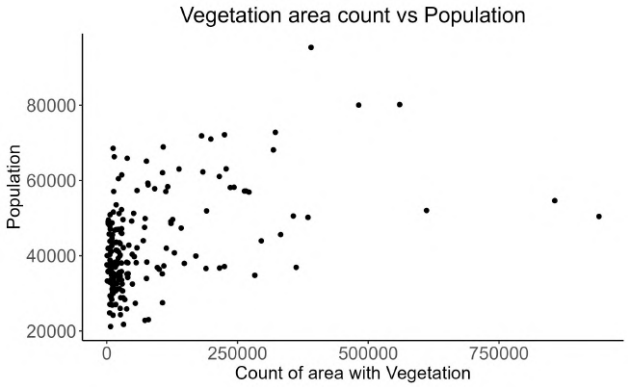


Figure 4.15: Vegetation Area count vs Population

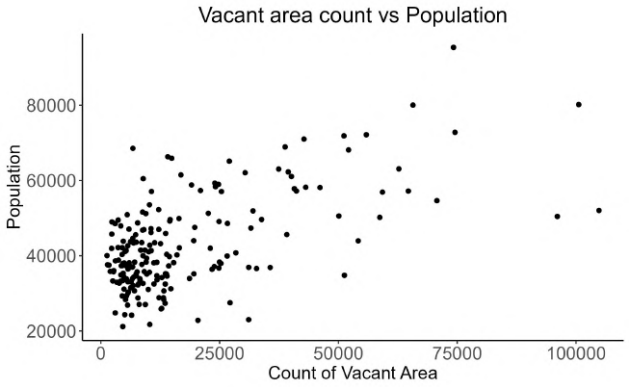


Figure 4.16: Vacant area count vs Population

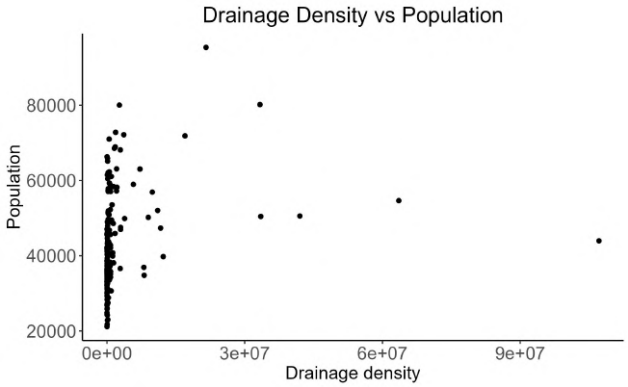


Figure 4.17: Drainage Density vs Population

4.3 Application of Diaggregation modelling

The diaggregation procedure, as described in the theory and simulations, is followed here to obtain estimates of the rates at pixel level. The model used is the ID model, which assumes no spatial correlation.

The CAR model is difficult to implement for large spatial data and the software breaks down in the process of handling the heavy data. We have left this case to be probed further in future

We have made use of the **optim()** function in R and also colour packages like **RColorBrewer**, **viridis**, **RPMG** and **wesanderson** to modify the spatial maps, in the process of extensive use of **ggplot2** package.

We required the use of **Matrix** package to account the use of sparse matrices[1].

n case of ID model, we consider the estimates as the mean of the posterior distribution of the rates :

$$\mathbf{R}|\mathbf{Y} \sim N_P(\mathbf{m}, \Sigma)$$

$$\left[\Sigma = \left(\frac{\mathbf{B}'\mathbf{B}}{\tau^2} + \frac{I_P}{\sigma^2}\right)^{-1}, \mathbf{m} = \left(\frac{\mathbf{B}'\mathbf{B}}{\tau^2} + \frac{I_P}{\sigma^2}\right)^{-1}\left(\frac{\mathbf{B}'\mathbf{Y}}{\tau^2} + \frac{\mathbf{X}\beta}{\sigma^2}\right)\right]$$

To calculate the value of the complicated inverse function in the process of obtaining the estimates above, we have also made use of the *Woodbury matrix identity*, given by

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U\left(C^{-1} + VA^{-1}U\right)^{-1}VA^{-1} \tag{4.1}$$

Suppose the left hand side of the above equation is the inverse of large (P x P) matrix. Now, using the above identity, we can avoid computing the massive inverse. The inverse component on the right hand side is the inverse of a small matrix (N x N) (where, say, P is about 4000 times larger than N!). This inverse is easier to calculate and requires lesser computational time. Also, in our case, the inverse on the right side turns out to be a diagonal matrix and hence the entire difficulty is removed.

Using the identity, we obtain

$$\mathbf{\Sigma} = \sigma^2\mathbf{I}_P - \sigma^4\mathbf{B}'\left[\tau^2\mathbf{I}_N + \sigma^2\mathbf{B}\mathbf{B}'\right]^{-1}\mathbf{B}$$

We have obtained the estimates of the rates as the *posterior mean* of the distribution of rates.

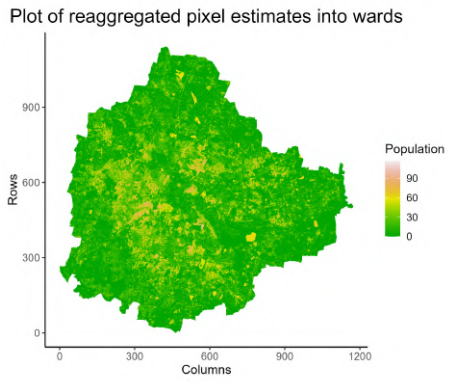
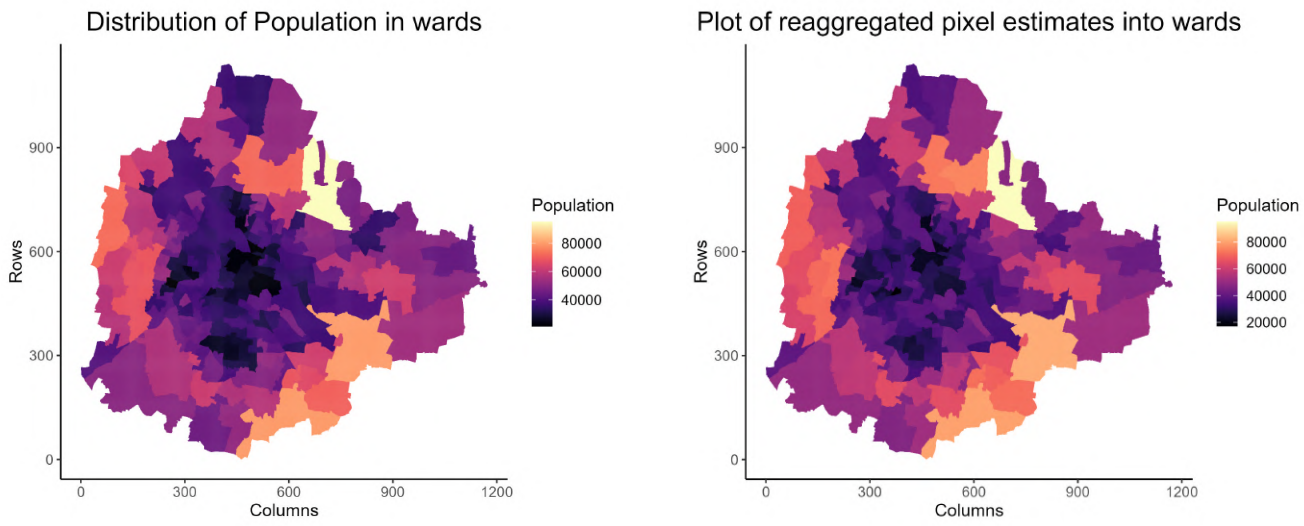


Figure 4.18: Population estimates across pixels

We note that the plot seems quite reasonable with high population counts towards the centre of Bangalore and lower counts near the fringes. Thus, the model is apparently successful in predicting the pixel-level rates.

[1]A sparse matrix is one that is comprised of mostly zero values. They are distinct from matrices with mostly non-zero values, which are called as dense matrices.



(a) True Values

(b) Estimated Values

Figure 4.19: Comaprison between true and estimated population values at ward level

Now, we compare the true ward population values with the estimated values, obtained by aggregating the estimates of the rates as per the ward structure of Bangalore City. Refer to Figure 4.19 for the maps. We note that the plot of aggregates of estimates seems quite in line with the plot of true values (data at hand). The population values are also quite similar and the overall pattern of population distribution seems to be close enough.

To compare the true and the fitted population values at the ward level, we consider the scatterplot below:

Figure 4.20: Scatterplot of true and fitted population values at ward level

We observe that the ID model seems to fit the data quite well and gives estimates which are close to true value.

Chapter 5

Conclusion

5.1 Final Comments

The Census of India data sets for cities and towns are usually available in the form of aggregate counts or fractions at the ward level. However, wards are generally too large and too heterogeneous for many practical situations. Our aim throughout the project had been to take the ward-level aggregate variables and estimate the values of those variables at the level of each of the pixels that constitute the respective ward. In the project, We have tried to consider different types of spatial structures. The simulation study revealed that estimates are more robust in case of a larger grid structure where we have many data points. While fitting a disaggregation model, the model performances were also better for the larger spatial structure case. Besides, in the small scale spatial map, both the ID model and the CAR model worked more or less well, with the CAR model performing slightly better. We expect that the difference in performance will increase and the CAR model will perform better in higher resolution and large spaces, due to its natural assumptions of spatial correlation. Finally, the data study indicated that the ID model, when fit to the data, works well with the data, giving good estimates.

5.2 Future Scope

The project has been completed under many assumptions, the most significant of which is the use of ID models whenever the spatial dimension increases. This has been done keeping in mind the technical shortcomings. However, there is ample scope of improvisation and we believe that the CAR model may well be used with ease in case of data spatial fields, if handled strategically. This would definitely help to obtain better and more robust and natural estimates, which will boost further research.

Bibliography

[1] Krishnachandran Balakrishnan. **A method for urban population density prediction at 30m resolution**. *Cartography and Geographic Information Science, 47:3, 193-213.*

[2] Arthur Nicolaus Fendrich, Elias Salomão Helou Neto, Lucas Esperancini Moreira e Moreira, and Durval Dourado Neto. **A scalable method for the estimation of spatial disaggregation models**. *Computers Geosciences 166 (2022) 105161.*

[3] Anita K Nandi, Tim CD Lucas, Rohan Arambepola, Peter Gething, and Daniel J Weiss. **disaggregation: An R Package for Bayesian Spatial Disaggregation Modelling**. *arXiv:2001.04847v1 [stat.CO] 9 Jan 2020.*

[4] Franz Schug, David Frantz, Sebastian van der Linden, and Patrick Hostert. **Gridded population mapping for Germany based on building density, height and type from Earth Observation data using census disaggregation and bottom-up estimates**. *PLoS ONE 16(3): e0249044.*

[5] Forrest R. Stevens, Andrea E. Gaughan, Catherine Linard, and Andrew J. Tatem. **Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data**. *PLoS ONE 10(2): e0107042.*

[6] CE Utazi, J Thorley, VA Alegana, MJ Ferrari, K Nilsen, S Takahashi, CJE Metcalf, J Lessler, and AJ Tatem. **A spatial regression model for the disaggregation of real unit based data to high-resolution grids with application to vaccination coverage mapping**. *Statistical Methods in Medical Research 2019, Vol. 28(10–11) 3226–3241.*