Amandine Teagho
Haoming Sun
Siqi Lyu
Lukas Dargel
Abderraman

Spatial reallocation of areal data

Amandine Teagho
Haoming Sun
Siqi Lyu
Lukas Dargel
Abderraman

11, March 2019

Amandine Teagho
Haoming Sun
Siqi Lyu
Lukas Dargel
Abderraman

Introduction

- Spatial interpolation((or Reallocation)) is the process of using points with known values to estimate values at other unknown points;
- **Aim:** Spatial interpolation is used to combine information from data set ;
- The basic role of spatial interpolation is to fill in the missing data for those areas where real world observations are not available.

Amandine Teagho
Haoming Sun
Siqi Lyu
Lukas Dargel
Abderraman

Motivation

When do we need to spatially reallocate data?

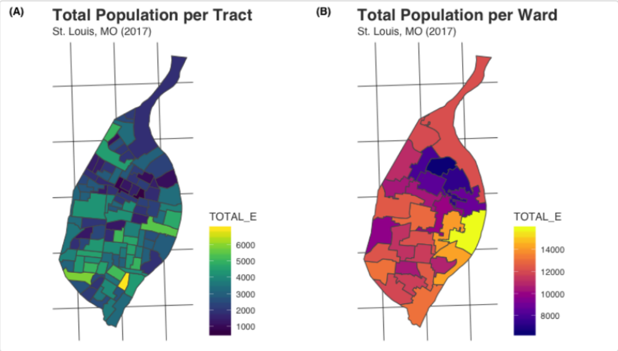


Figure: Illustration of spatial reallocation (example from https://cran.r-project.org/web/packages/areal/vignettes/areal.html)

Amandine Teagho
Haoming Sun
Siqi Lyu
Lukas Dargel
Abderraman

Amandine Teagho
Haoming Sun
Siqi Lyu
Lukas Dargel
Abderraman

Objectives

Throughout this work, we are going to:

- Describe 3 class of methods of spatial reallocation:
  - An elementary method: Areal weigthing,
  - Dasymetric weigthing techniques,
  - Regression techniques;
- Apply these methods on a data set and compare them.

Amandine Teagho
Haoming Sun
Siqi Lyu
Lukas Dargel
Abderraman

Notations

Throughout this work, we consider the following notations:

- Y_A: target variable on the region A;
- $|A|$: the area of the region A;
- S: a set of source zones S_s, $s = 1, \cdots, S$;
- T: a set of target zones T_t, with $t = 1, \cdots, T$;
- $A_{s,t} = S_s \cap T_t$: the intersection zones;
- For simplicity, we set:
  - $Y_{S_s} := Y_{s,}$,
  - $Y_{T_t} := Y_t$,
  - $Y_{A_{s,t}} := Y_{s,t}$.

Definitions

- Extensive: The variable Y is extensive if its value for a target zone t is equal to the sum of its value for each zone st which intersects.

$$Y_t = \sum_s Y_{s,t}$$

- Intensive: if its value for a target zone is a weighted average of its value for the intersections zones, often the weights will be the areas involved:

$$Y_t = \frac{\sum Y_{s,t} A_{s,t}}{\sum A_{s,t}}$$

(e.g. percentage, proportion, ratio)

Amandine Teagho
Haoming Sun
Siqi Lyu
Lukas Dargel
Abderraman

Properties

There are 2 mains properties to

Homogeneity

- An extensive variable is homogeneous in a given zone A if it is evenly distributed within A.
- An intensive variable is homogeneous in a given zone A when the variable is constant in each sub-zone of A.

Properties: -Pycnophylactic Property

- It ensures the preservation of the initial data;
- The predicted value $\hat{Y}_s$ on source S_s obtained by aggregating the predicted values on intersections with S_s should coincide with the observed value Y_s on S_s.

Extensive variable

$$Y_{s,t} \sim \mathscr{P}(\mu_{s,t})$$

Intensive variable

$$Y_{s,t} \sim \mathscr{N}(\mu_{s,t}, \frac{\sigma^2}{n_{s,t}})$$

Amandine Teagho
Haoming Sun
Siqi Lyu
Lukas Dargel
Abderraman

Areal weighting interpolation: How does it work?

- **Input:** A set of source zones;
- **Idea:** Allocate a set of source zones S into a set of target zones T;
- We distinguish 2 cases for computing the intersection zone prediction;

Extensive variable

It is based on the homogeneity assumption stated as follows;

- **Assumption 1:** For each sub-region s(with $s \cap t \neq \emptyset$), the estimated value of Y in $A_{s,t}$ is:

$$\hat{Y}_{s,t} = \frac{|A_{s,t}|}{|S_s|} Y_s$$

- The target zone prediction is:

$$\hat{Y}_t = \sum_{s:s\cap t\neq\emptyset} \frac{|A_{s,t}|}{|S_s|} Y_s$$

Areal weighting interpolation: second case

Intensive variable

- **Assumption 2:** Y is uniform on the source zones:

$$\hat{Y}_{s,t} = Y_s$$

- The target zone prediction is:

$$\hat{Y}_t = \sum_{s:s \cap t \neq \emptyset} \frac{|A_{s,t}|}{|T_t|} Y_s$$

where $\hat{Y_{s,t}}$ is the estimated value for the intersection zones and $\hat{Y}_t$ is the the estimated value for the target zone.

Amandine Teagho
Haoming Sun
Siqi Lyu
Lukas Dargel
Abderraman

Areal weighting interpolation: Pros and Cons

Pros

- It is the simplest technique used for areal interpolation;
- It can be used for extensive or intensive variable;
- It satisfies the pycnophylactic property;

Con

- But, it does not require additional auxiliary information;
- Hence, we cannot make use of external knowledge about the data with this method.

Dasymetric weighting: Overview

- **Objective:** Extension and improvement of the areal weighting methods.
- **Idea:** It uses other relevant and available information X to distribute Y accordingly.

Notes:

- This method is adapted to the case of both intensive and extensive variable Y.
- This method also satisfies the pycnophylactic property.

Spatial reallocation of areal data

Amandine Teagho
Haoming Sun
Siqi Lyu
Lukas Dargel
Abderraman

Ordinary dasymetric weighting

- **Assumption:** auxiliary information is known at intersection level and it has quantitative nature.

Case of an extensive target variable with an extensive auxiliary variable X:

- $$\hat{Y}_t = \sum_{s:s\cap t\neq\emptyset} \frac{X_{s,t}}{X_s} Y_s$$
- This formulae extend the formulae of the areal weighting interpolation by substituting X for the area

Amandine Teagho
Haoming Sun
Siqi Lyu
Lukas Dargel
Abderraman

Case of an intensive target variable with weights given by $\omega_A = \frac{Z_A}{Z_\Omega}$ for a given variable Z and an extensive auxiliary variable X:

- We define two variables $\tilde{Y}_A = Z_A Y_A$ and $\tilde{\tilde{X}}_A = \frac{X_A}{Z_A}$
- The formula is obtained using the correspondence intensive-to-extension

$$\hat{Y}_t = \sum_{s:s\cap t\neq\emptyset} \frac{X_{s,t}}{X_s}\frac{Z_s}{Z_t} Y_s$$

Amandine Teagho
Haoming Sun
Siqi Lyu
Lukas Dargel
Abderraman

Dasymetric weighting with control zones

Amandine Teagho
Haoming Sun
Siqi Lyu
Lukas Dargel
Abderraman

- Input: a set of source zones and a categorical ancillary data set——control zone
- Idea: In order to provide a more accurate depiction of how the demographic data is distributed, we redistribute the data to a set of target zones formed from the intersection of the source and ancillary zones. e.g. redistribute the population based on levels of urbanization.
- Assumptions:
  1. Auxiliary information is known at intersection level and it has qualitative nature.
  2. The count density is uniform throughout control zones.
  3. Intersection units are nested within control zones (not restrictive)

Dasymetric weighting with control zones

Amandine Teagho
Haoming Sun
Siqi Lyu
Lukas Dargel
Abderraman

- Method:
  1. Estimates these densities D_c for control zone c by sampling a subset of the total source zones using:

$$\hat{D}_c = \frac{\sum_{s \in c} Y_s}{\sum_{s \in c} |S_s|}$$

  2. The intersection zone prediction is given by:

$$\hat{Y}_{s,t} = \frac{|A_{s,t}|\hat{D}_{(s,t)}}{\sum_{t': s \cap t' \notin \varnothing} |A_{t',s}|\hat{D}_{c(t',s)}} Y_s$$

- Advantage: The method is guided by use of ancillary land-use data, in contrast to the use of simple areal weighting.
- Limitation: The accuracy of dasymetric maps is highly dependent on the data being used as inputs.

Amandine Teagho
Haoming Sun
Siqi Lyu
Lukas Dargel
Abderraman

Regression with auxiliary information at target level

- Translate the target variable from source to intersection level ($Y_s \to Y_{s,t}$)
- Requires auxiliary information at target level X_T
- Distinguish two cases:

Extensive variable (count data)

- **Assumption 1:**

$$Y_{s,t} \sim \mathscr{P}(\mu_{s,t})$$

Intensive variable (shares or ratios)

- **Assumption 1:**

$$Y_{s,t} \sim \mathcal{N}(\mu_{s,t}, \frac{\sigma^2}{n_{s,t}})$$

- we assume:

$$Cov(Y_{s,t}, Y_s) = \sigma^2 / n_s$$

Regression: How to estimate the model? 1 (case of an intensive target variable)

- We have the relation between X and Y at target level:

$$E[Y_T|X] = \mu_T = X_T\beta$$

- Given the aggregation equation:

$$\mu_s = \sum_t \frac{n_{s,t}}{n_s}\mu_{s,t}$$

- Given the uniformity at target assumption:

$$\mu_t = \mu_{st}$$

- And the weight matrix W whose elements are given by:

$$w_{s,t} = \frac{n_{s,t}}{n_s}$$

.

Regression: How to estimate the model? 2 (case of an intensive target variable)

- We get the following regression equation:

$$\mu_S = WX_T\beta \qquad \text{with} \qquad \mu_S = Y_S$$

- From this equation it is easy to estimate beta using weighted least squares
  - Weights are given by n_s (usually population)
- Use the model to obtain $\hat{\mu}_t$ as a predictor for Y_t:

$$\hat{\mu}_t = X_T\hat{\beta}$$

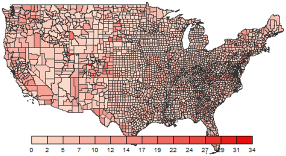
- (For an extensive target one has to use Poisson regression and adapt the aggregation equation and the weight matrix.)

.

Amandine Teagho
Haoming Sun
Siqi Lyu
Lukas Dargel
Abderraman

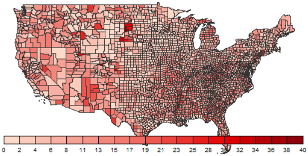
Absolute Error

Absolute errors of AIW methode



(Income per capita)

Absolute errors of regression methode



(per capita income)

Amandine Teagho
Haoming Sun
Siqi Lyu
Lukas Dargel
Abderraman

Amandine Teagho
Haoming Sun
Siqi Lyu
Lukas Dargel
Abderraman

EM algorithm

The interpolation problem is cast as a missing data problem

- E-step: the algorithm compute an expected values of the intersection values of the target variable given the model and the source values.
  - $\mu_{s,t}$ at first iteration it can be estimated by areal weighting
- M-step: the previous values obtained are considered as i.i.d observation from $\mathscr{P}(\hat{\mu}_{s,t})$ in case of extensive variable or $\mathscr{N}(\hat{\mu}_{s,t})$ for intensive

Amandine Teagho
Haoming Sun
Siqi Lyu
Lukas Dargel
Abderraman

Regression with control zone

- Assumption: means on controls zone are uniform $\mu_{s,c} = \mu_t$
  The homogenity are on controls ($C \leq T$)

- We have also the relation between Y and X at controls level:

$$\mu_C = X_C \beta$$

- And the aggregation equation:

$$\mu_s = \sum_c \frac{n_{s,c}}{n_s} \mu_{s,c}$$

- We get :

$$\mu_C = X_C \beta$$

$$\mu_S = W X_C \beta$$

with $w_{s,c} = \frac{n_{s,c}}{n_s}$ element of W

Regression with control zones 2

- We obtain $\hat{\mu_C}$ as above :

$$\hat{\mu}_C = X_C \hat{\beta}$$

and a predictor of Y_t by this relation:

$$\hat{Y}_t = \sum_c \frac{n_{c,t}}{n_c} \hat{\mu}_c$$

Amandine Teagho
Haoming Sun
Siqi Lyu
Lukas Dargel
Abderraman

Amandine Teagho
Haoming Sun
Siqi Lyu
Lukas Dargel
Abderraman

Application

- Now, we move on to the application part to apply each of the method we just described.
- For this purpose, let us describe the data used:
  - We use socio-demographic variables given by USA' census (2016):
  - incomepercap(GDP) as extensive variable, car numbers as intensive variable.
  - population on county
  - medianage, popmale, popfemale , popwhite , popblack, popasia, medianincome,unemployed,laborforce, incomepercap, ginicoef, cars, bartenders , as auxiliary's information on state.

Amandine Teagho
Haoming Sun
Siqi Lyu
Lukas Dargel
Abderraman

Spatial reallocation of areal data

Amandine Teagho
Haoming Sun
Siqi Lyu
Lukas Dargel
Abderraman

Introduction

Objectives

Objectives

Notations

Definitions and Properties

Areal weighting interpolation

Dasymetric weighting

Ordinary dasymetric weighting

Dasymetric weighting with control zones

Regression techniques

Regression with auxiliary information at target level

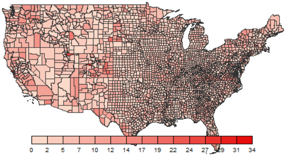
Comparison of the methods

EM algorithm

Regression with control zones

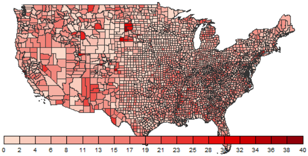
Absolute Error

Absolute errors of AIW methode



(Income per capita)

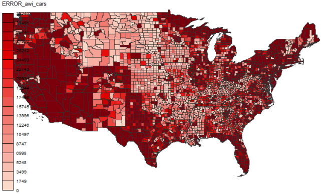
Absolute errors of regression methode



(per capita income)

Absolute Error

Absolute errors of DAX method



(number of the cars)

Amandine Teagho
Haoming Sun
Siqi Lyu
Lukas Dargel
Abderraman

Thanks for your attention!