Small-area poverty estimates

Dr Hector Najera

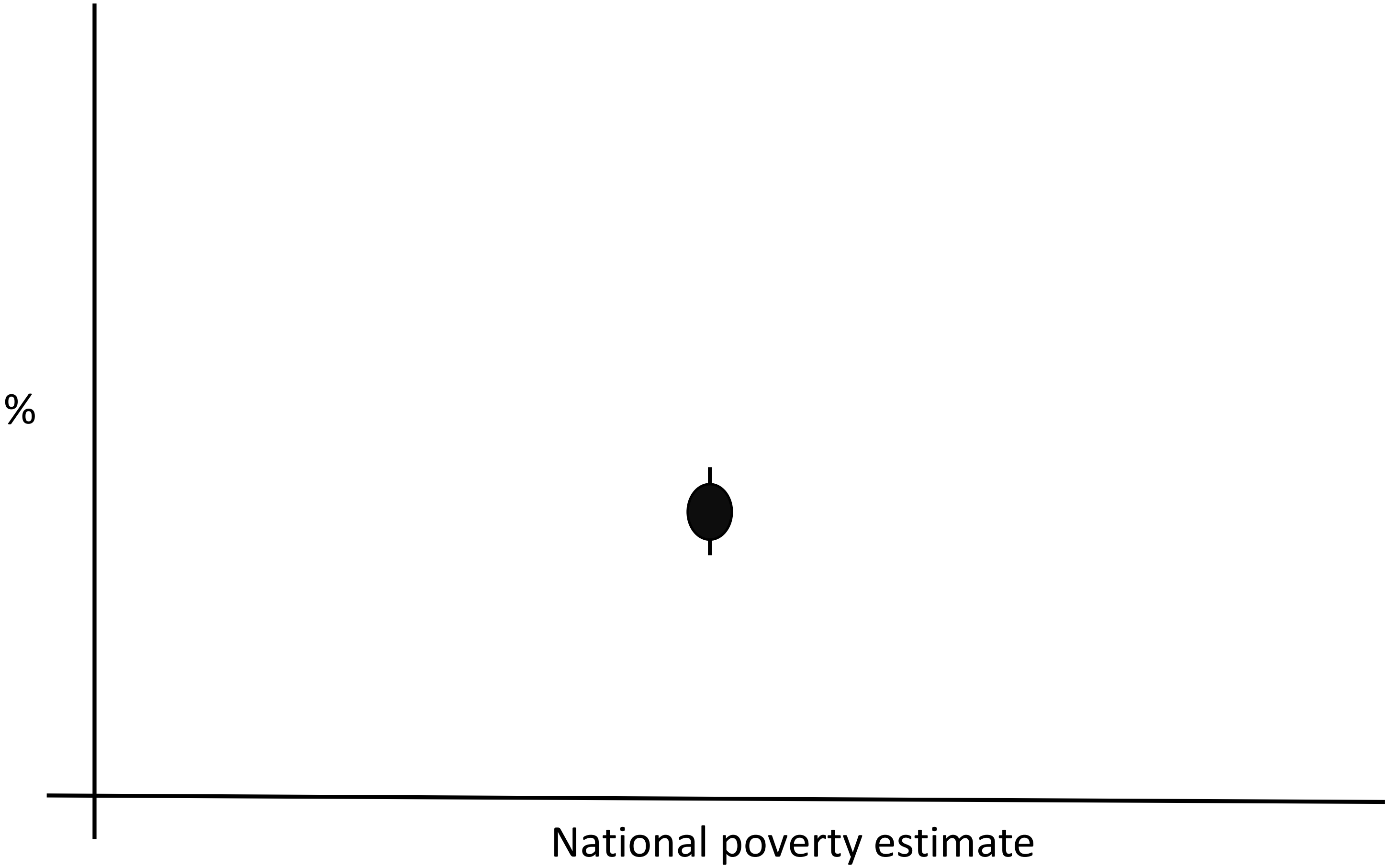
19 July 2019

Introduction: Small-Area Estimation

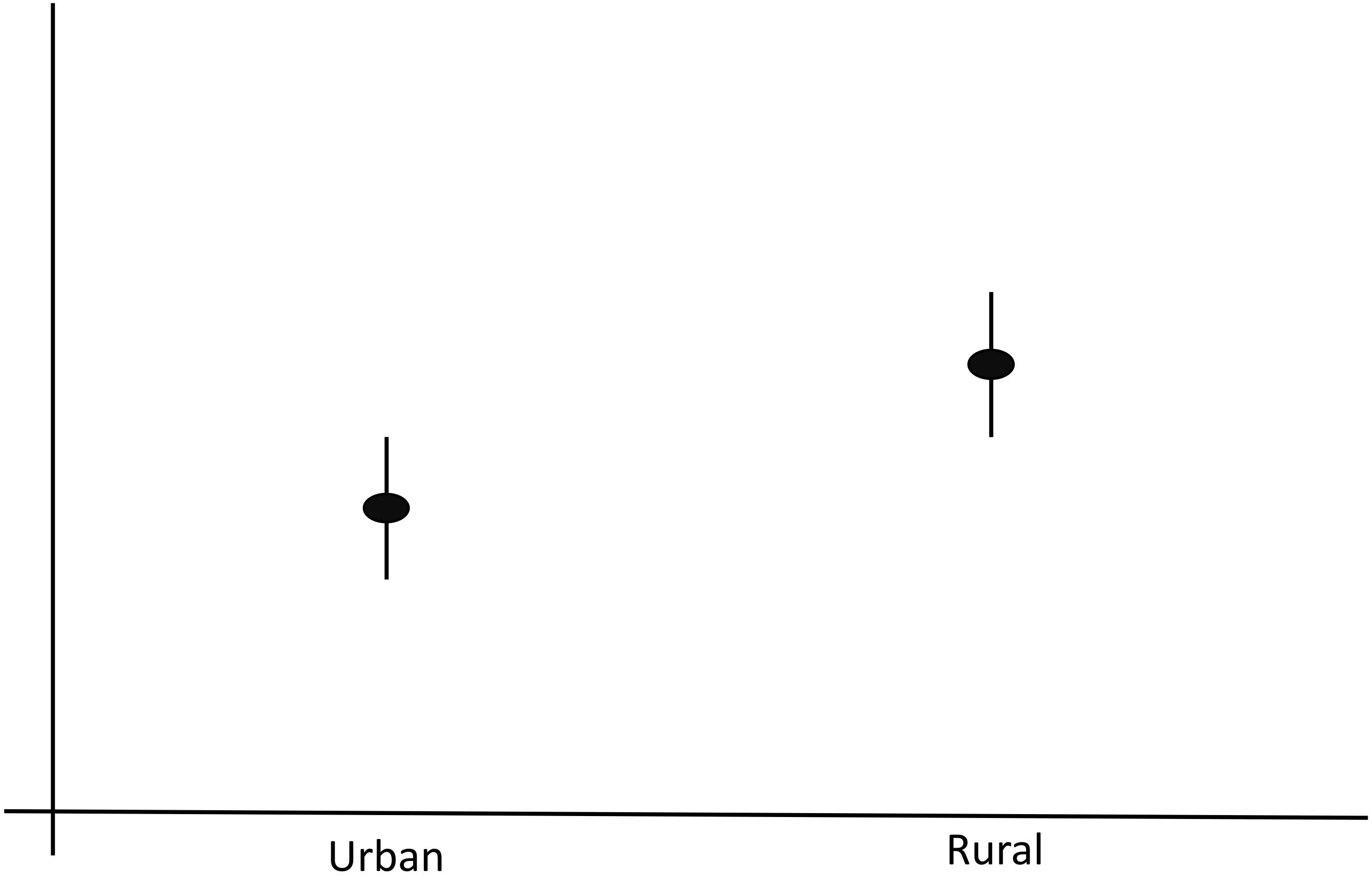
1. Survey data is a cost-effective way to measure complex phenomena such as **poverty** at national and some sub-national levels (urban rural)
2. Survey data, however, will produce unreliable estimates (i.e. biased and with high variance) for smaller areas
3. Furthermore, survey data cannot produce estimates for out-of-sample areas
4. Policy makers need indicators and maps of poverty to formulate and implement policies, (re)distribute resources, and measure the effect of local policy actions

Survey-data estimaton

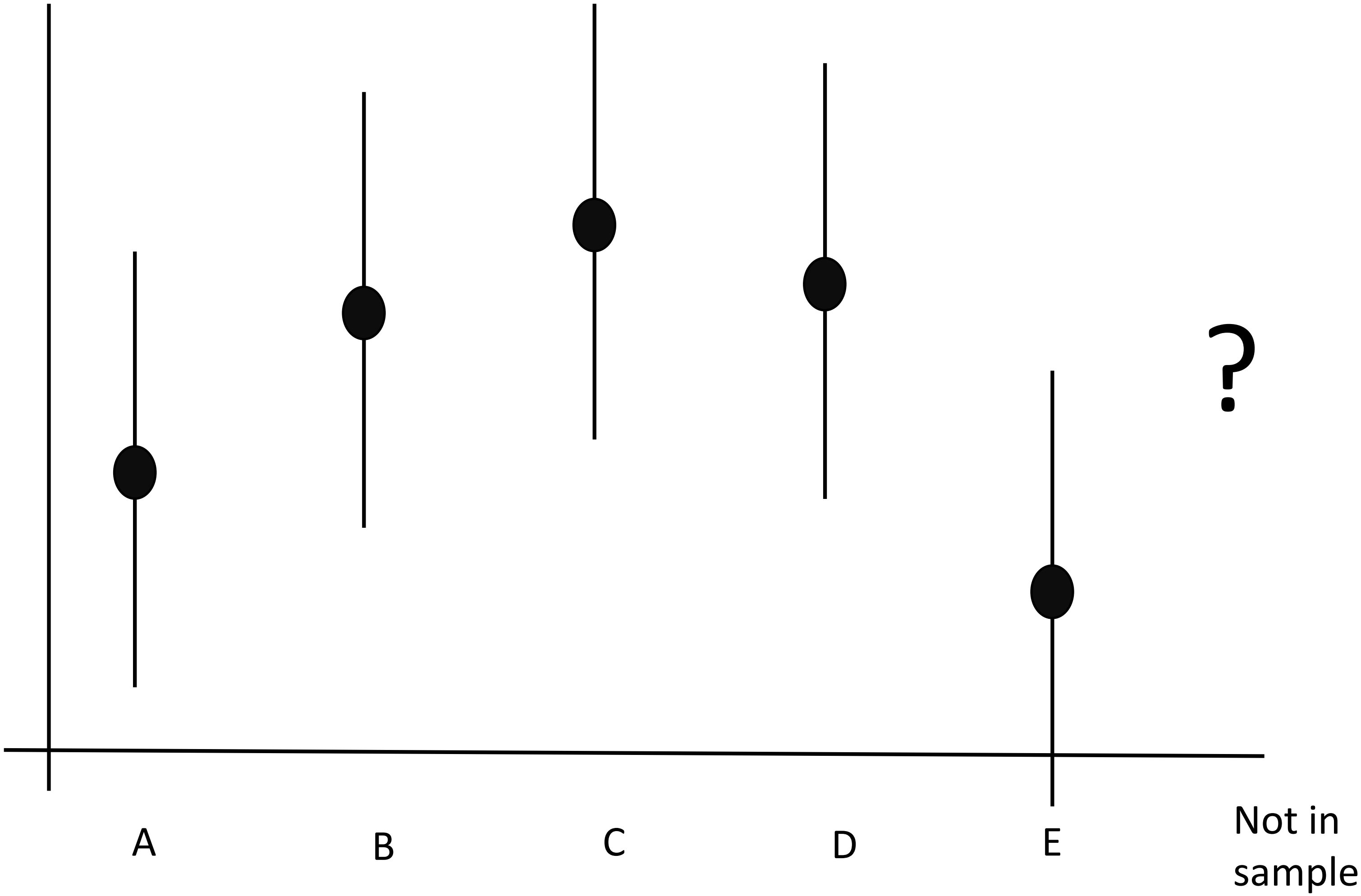
National prevalence rate: Point estimate with some uncertainty around it (Confidence Interval)



If you have a representative sample for urban and rural areas



For smaller areas you will have a lot of uncertainty and bias



Small-area estimation (SAE)

- Small-area estimation is the academic response to the problem of small samples (or not data at all).
- The difficulty is that SAE offers not one but several competing answers to the same problem
- SAE provides several contested answers because there are different technical proposals (i.e. **estimators**) to address the problem of producing data for areas with tiny n's or $n = 0$
- All these proposals aim to do something very simple: Use the available data in the best possible way.

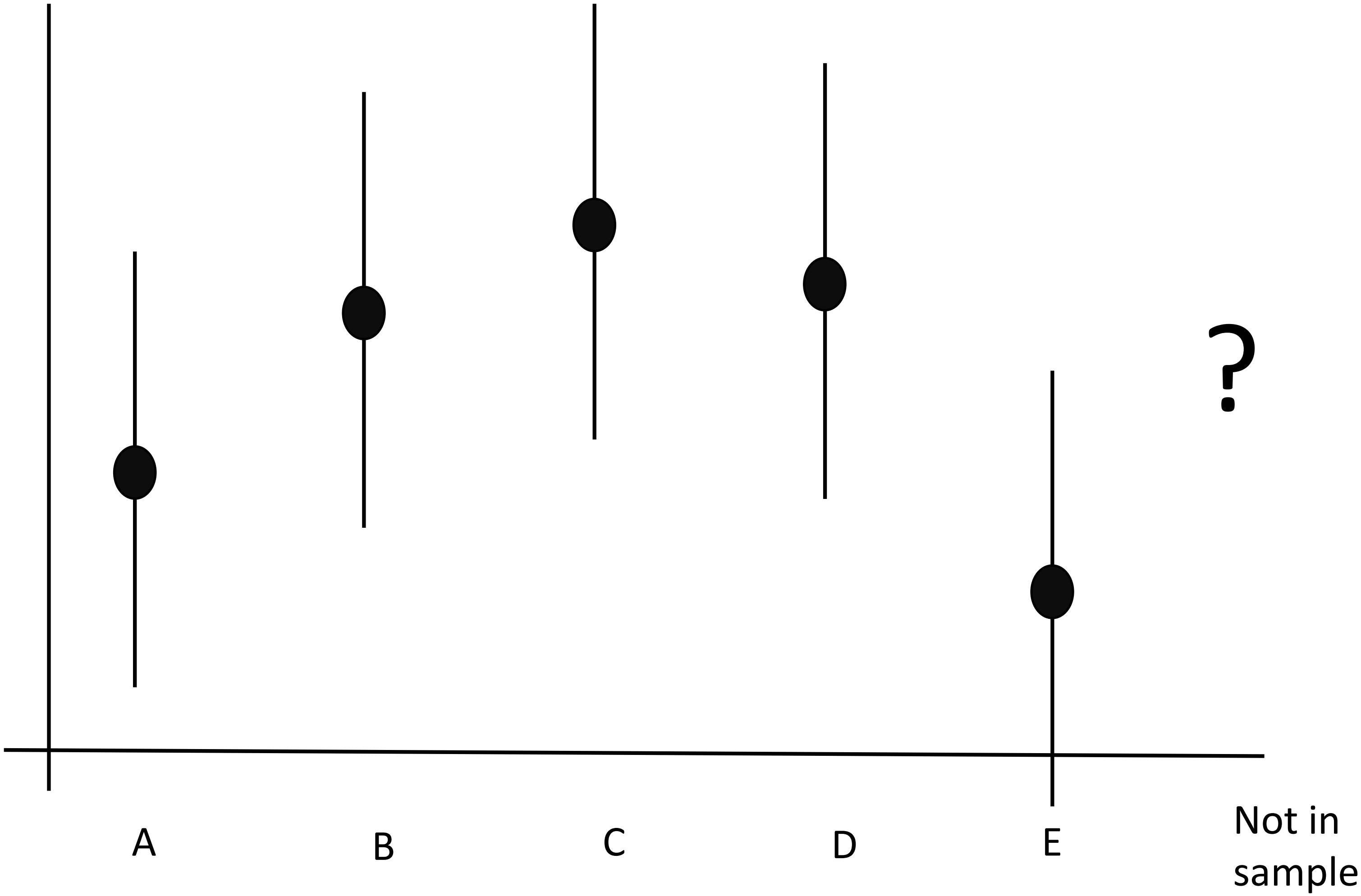
Approaches in SAE

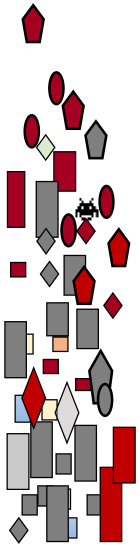
- The literature crudely classifies the SAE **estimators** in two types:
  - Direct
  - Indirect
- For many years, mainly due to lack of computing power, direct estimators were predominantly used
- These days indirect or model-based estimators are more commonly used

Direct SAE estimators

- A direct estimate is a survey/census-based *figure*, i.e. a prevalence rate, a ratio, etc.
- When we use survey data, it is a prevalence rate taken from a sample.
- If this sample is representative, then the **direct estimator** is unbiased with some error (sampling error). So, it is a point estimate with some uncertainty around it (CI, CrI, SE). This is also called a **design estimate**
- When our sample is not representative for some areas we can calculate a **direct** estimator but this will be biased and it will have a lot of uncertainty around it.
- **Direct** estimators are thus useless when $(n = 0)$

Direct estimators under unrepresentative sampling

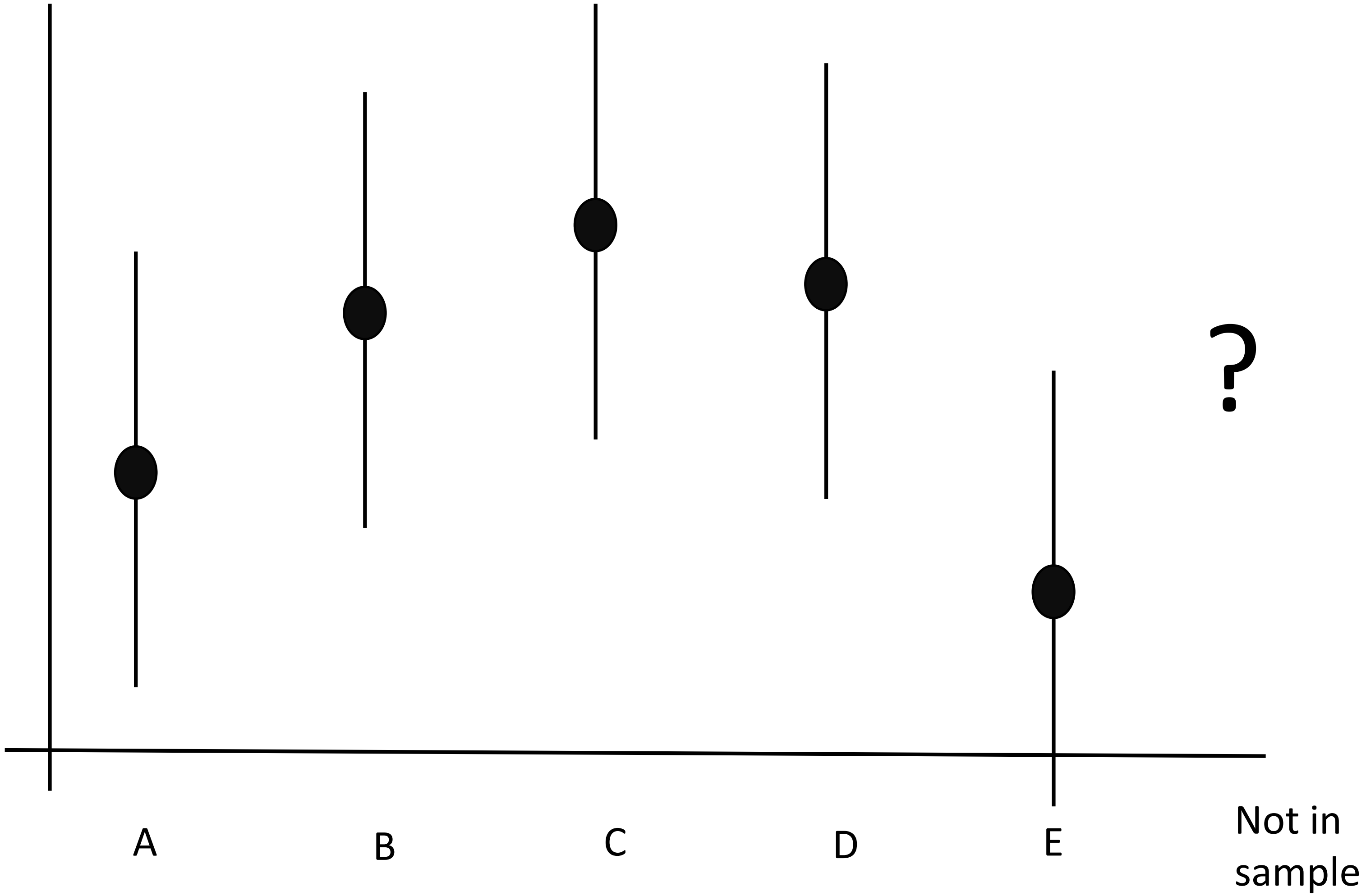




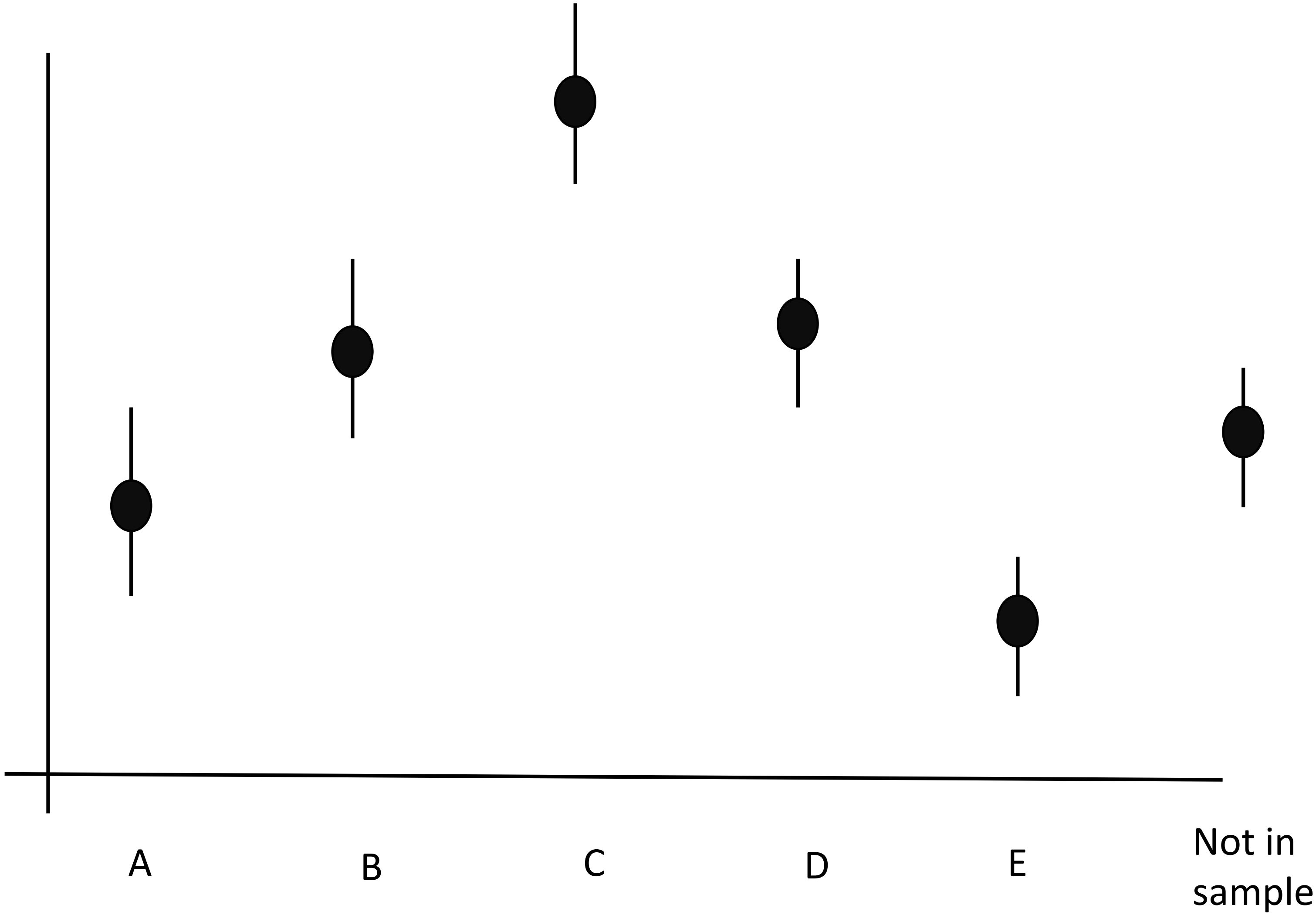
Indirect SAE estimators

- The use of a model is the main feature of **indirect** estimators.
- Both statistical theory and computer power have increased and this has resulted in the development of several types of **indirect estimators**.
- Indirect estimators are thus based on a model, that tries to use the available data in the most efficient way.
- That means fitting a predictive model (try to approximate the data generative process behind a phenomenon), i.e. find the best predictors of poverty and estimate the probability for a person in a sample.
- Then that predictive model is applied to the Census data, i.e. each person in the census has a probability of being poor.

Indirect estimators aim to correct direct estimation



Indirect estimators aim to correct direct estimation



Intuition of indirect small-area estimation

- A representative sample is taken from a population X_{ij}. Where i are individuals/households and j are areas (i.e. counties).
- P_{ij} is the poverty status for each person/household in the sample {1=Poor, 0=Not poor}.
- Remember that you don't have a sample of all $j's$ in the survey
- But you have data for all $j's$ in the Census
- SAE aims to link the Survey data with the Census using a predictive model as follows

Intuition of indirect small-area estimation

- Think about small-area estimation as a missing data problem
- You have a couple of options: Multiple imputation or a predictive model
- In SAE multiple imputation makes little sense because we are interested in the **true** values
- What is a predictive model?

Intuition of indirect small-area estimation: Predictive models

- You produce a variable P_i, i.e. the poverty status for each person in the sample {1=Poor, 0=Not poor}
- What are the predictors of poverty?

$$P_i = \alpha + \beta_1 Educationyrs + \beta_2 Employmentstatus + \beta_3 rurality + \beta_4 age + \eta_i$$

- If you have a very good model, you will be able to predict quite accurately poverty

$$P_i \approx \hat{P}_i$$

- This is fine but remember that you don't have all j in your sample, you need to apply the same model to a new data set (which all areas)
- So you need the same set of common variables in the new data to produce:

$$\hat{P}_i$$

- Small-area estimation has put forward several ways to have better predictive models

Intuition of indirect small-area estimation

1. Fitting a predictive model using survey data

$$\begin{array}{c} P_{ij} = \alpha + \underset{\Big\downarrow}{\beta_1 X_{1i}} + \underset{\Big\downarrow}{\beta_1 X_{2j}} + U_j \\[1em] P'_{ij} = \alpha + \beta_1 X_{1i} + \beta_1 X_{2j} + U_j \end{array}$$

2. Applying the coefficients of the model to a different data set where you have all $j's$

In this case P_j is not present in our data but the $X's$ are available

So what you do is to apply the estimated coefs to different data

- Survey

$$P_{ij} = \alpha + \beta_1 Educationyrs_i + \beta_1 Unemploymentrate_j + U_j$$

- Census

$$P_{ij} = \alpha + \beta_1 Educationyrs_{1i} + \beta_1 Unemploymentrate_j + U_j$$

So you can estimate P_{ij} for each person in the Census (IF YOU HAVE MICRODATA!)

Indirect SAE estimators

The SAE literature has proposed several estimators to fit such a predictive model.

In order of complexity, theoretical and experimental accuracy:

1.Fay-Herriot estimator

2.EBLUP: Empirical Best Linear Unbiased Predictor

2.1 ELL - World Bank method

3. EB: Empirical Best (Non-linear) Unbiased Predictor

4. Hierarchical Bayes (HB) !This is what we are proposing

5. M-quantile + spatial component

6. EBLUP/HB + spatial component

Yes, but: What’s a Good model?

- It is a model that reproduces the true estimates
- It should have good fit (R^2, Pseudo-R^2, AIC, WAIC, LOO)
- If it has Good fit, it is going to have low error
- It is a model that is parsimonious (predict well with few variables)
- It is a model that is computationally feasible

So, finding a Good model is difficult because you have to fit several models to find a good one

For example: Fay-herriot estimate

Fay-Herriot (FH) model (Fay and Herriot, 1979) is still widely used because it has some advantages:

1. its very simple (it is just a regression model)
2. its does not require microdata
3. its ability to produce design-consistent estimators.
4. it takes into account the sampling design(level 1 model)
5. it only requires area auxiliary variables that, in general, are more easily avail-able in practice than unit (i.e. individual or household-level) auxiliary variables.

Fay-herriot estimate

$$P_j = \alpha + \beta_1 X_{1j} + \beta_1 X_{2j} + U_j$$

- In this case j are areas not individuals. So you fit a model using areas as unit of analysis.

X_1j = % of people with secondary education
X_2j = % of people unemployed

- U_j is a the specific intercept for each area.

EBLUP model

- Remember that in regression you attempt to produce BLUP estimates of β, i.e. unbiased and accurate coefficients
- The **E** means **Empirical** because you use a clever way to estimate area-level specific variability. Given a model:

$$P_{ij} = \alpha + \beta_1 Educationyrs_i + \beta_1 Unemploymentrate_j + U_j$$

- The U_j are random intercepts, i.e. the intercept for each area j. Estimate these intercept is not easy!
- U_j can't be directly estimated in one step. First you fit a model that estimates the variability across areas σ_j^2.
- That is, how much of the variability of poverty is due to individual-level observed variables and how much is explained by differences across areas

EBLUP model

- In a second step using Empirical Bayes estimation (Pseudo-Bayesian estimation), you can use the variance across areas to estimate the specific intercept for each area

- The EBLUP model is, thus, a better predictive model than a common or synthetic model because it uses both individual and contextual data and estimates a specific parameter for each area.

EBLUP model

Guadarrama et al. (2014) conclude the following:

- It is based on unit level data, which are richer than the area level data and uses much larger sample size to fit the model.
- Best estimators are model-unbiased.
- Once the model is fitted, estimates can be obtained at whatever subarea level.
- However, it could be unfeasible for complex models and for very large datasets

The WB method

- It is also known as the ELL method (Elbers et al., 2003)
- It is a special case of the EBLUP model but worse!
- The WB re-invented multilevel/hierarchical models
- It does not work with *j* but with clusters, primary sampling units.
- It does not rely on a sensible approach like Empirical Bayes.

The WB method

- The WB method proposes to fit a model and then use bootstrapping
  - Produces many replications of the same predictive model to approximate the **true** parameter
  - It produces several synthetic surveys from the Census, estimates poverty and repeats this several times
- It is based on the frequentist tradition, i.e. if you have many and big samples you are likely to hit the target

The WB method

- The WB has been heavily criticised because it will fail in countries with high heterogeneity
- The main reason is that for the out-of-sample areas bootstrapping will produce a biased estimate
- Its upper bound is a two-level model (EBLUP).
- With the EBLUP you can have proper areas *j* and then the *k* primary sampling units in a three-level model.

The WB method

Guadarrama et al. (2014) conclude the following:

ELL estimates perform poorly and can even perform worse than direct estimators when unexplained between-area variation is significant, see Molina and Rao (2010). In fact, for the estimation of domain means, ELL estimates are basically equal to regression-synthetic estimators, which assume the regression model without further between-area variation.

They are not design-unbiased and can be seriously biased under informative sampling.

They can be seriously affected by unit level outliers.

If cluster effects are included in the model instead of area effects, but area effects are significant, ELL estimates of the model MSE can seriously underestimate the true MSE. Even if area effects are included in the model, ELL estimates of MSE do not track correctly the true MSE for each area.

The HB estimator

- SAE is a problem of high-dimensions, i.e. large data sets and many parameters
- The HB is just an general case of the EBLUP estimator
- It is also known as fully Bayesian estimator
- The random intercepts or slopes are not estimated ex-post but under the same model. This has the advantage that estimation error is lower
- It does not uses Maximum Likelihood estimation but Bayesian computation, i.e. Monte Carlo Marcov Chains
- MCMC is more efficient for high-dimensional problems and recent breakthroughs in Bayesian computation have boosted the MCMC estimator via the Hamiltonian Monte Carlo (HMC)

The HB estimator

You could have:

$$P_{ij} = \alpha + \beta_1 Educationyrs_i + \beta_1 Unemploymentrate_j + U_j$$

But also add some prior information like

$$\beta_1 \sim (0,1)$$

You can say that you expect -from prior evidence- that the employment rate at area level is going to have a small effect.

Summary of advantages and disadvantages

Estimator	Type of variable	Auxiliary data	Computational efficiency	Complexity
SE	Continous/Categorical	Individual data	Very fast	Limited
EBLUP	Continous	Individual + Contextual	Fast	Limited
ELL	Continous	Individual + Contextual	Very slow	Limited
EBUP	Categorical	Individual + Contextual	Slow	Limited
HB	Continous/Categorical	Individual + Contextual	Slow	Very flexible
M-quantile	Continous	Individual + Contextual	Slow	Flexible

Which estimator is better?

- Most of what we know about the behaviour of these estimators is via *Monte Carlo* simulation
- That is, one simulates data and checks if the model finds the correct answer.
- It is well now that the **FH** has the worst performance on average.
- It has been well established that the World Bank method is outperformed by the EBLUP, HB and M-quantile
- The reason is that it does not work too well with heterogenous data. For example, in countries with very high spatial inequalities, we will see the ELL fail.
- The HB tends to have lower error than the EBLUP and it is more flexible. The literature is moving toward Bayesian methods.
- For outliers the M-quantile method is far superior. It could be implemented via Bayesian modelling too.

Summary of the accuracy of SAE models

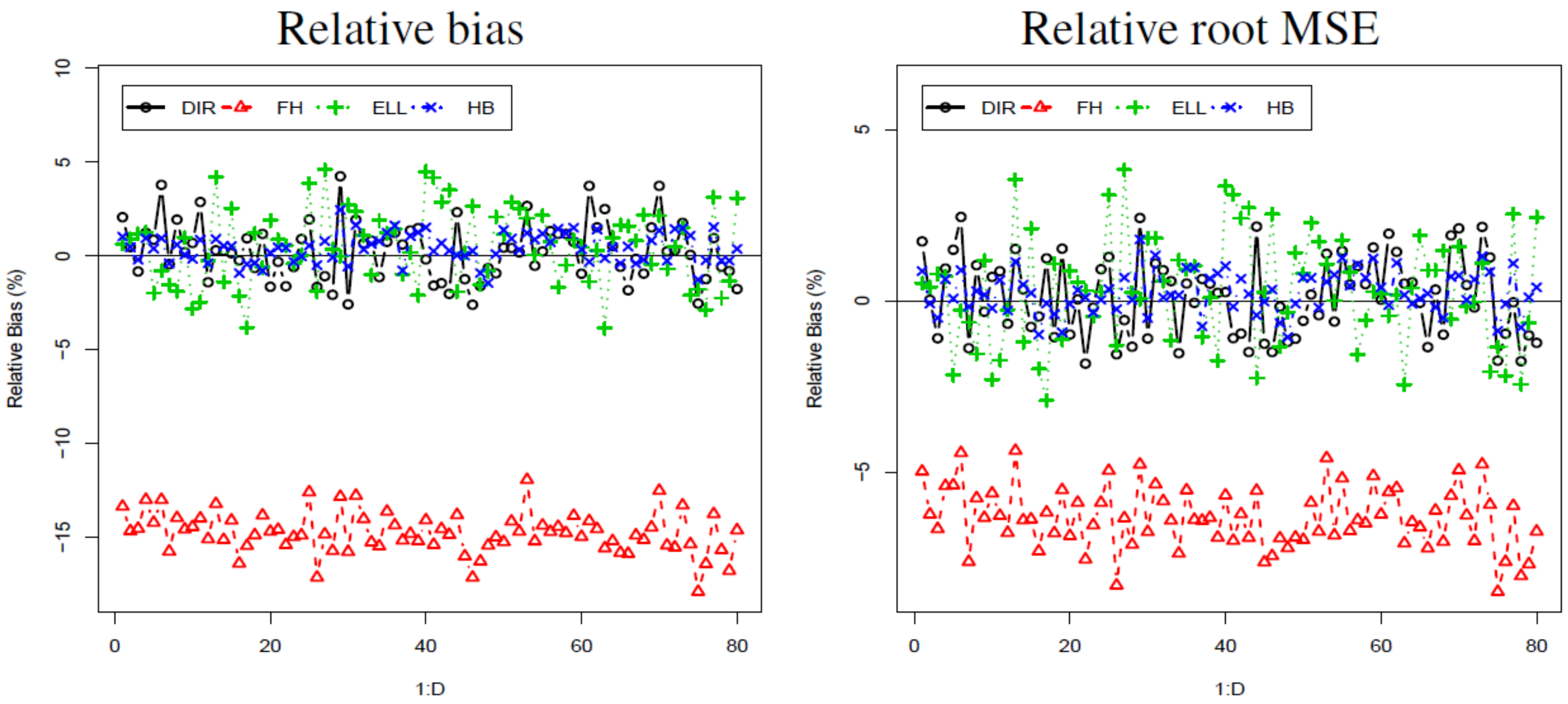


Figure 1. Percent relative bias (left) and relative root MSE (right) of direct, FH, HB and ELL estimators of poverty gap F_{1d} for each area d under the nested error model with simple random sampling.

Summary of SAE estimators

Based on Guadarrama et al. (2014)

1.EB and HB methods perform practically the same, and are the best among the considered estimators when the nested error model with normality holds and sampling is noninformative.

2.They are not very much affected by mildly informative sampling and small proportion of mild outliers, but might be severely affected by highly informative sampling or severe outliers in large proportions.

3.If your sample is small and with a very complex design you need to include the survey design for both the EB and HB

4.Census-EB estimators of poverty indicators are practically the same as EB estimators and avoid linking the survey and census data files.

5.ELL method under a nested error model with random area effects performs the worst in all scenarios

What are we doing in Bristol?

We have successfully implemented the HB for Mexico (poverty and malnutrition/Stunting) and for Tonga.

Nájera, H, Fifita, V and Faingaanuku, Winston (2019) Small-Area Multidimensional Poverty Estimatesfor Tonga 2016: Drawn from a HierarchicalBayesian Estimator. Applied Spatial Analysis and Policy https://link.springer.com/article/10.1007/s12061-019-09304-8

Nájera, H. (2019). Small-area estimates of stunting. Mexico 2010: Based on a hierarchical Bayesian estimator. Spatial and Spatio-Temporal Epidemiology, 29, 1–11. https://doi.org/10.1016/j.sste.2019.01.001.

- To give you an idea Mexico used the ELL and the EBLUP. It took them two years to produce a good model. We have a model with 12 variables that is just as good!
- The estimates for Tonga have been validated in situ

CONEVAL SAE project

- Combined the HIES data and a sample of the CENSUS. Both have common variables. A good model for the HIES should work on the Census
- CONEVAL applied an *ad hoc* version of the following estimators: Empirical Bayes, Elbers and colleagues (World Bank Method) and Hierarchical Bayes (CONEVAL, 2017)
- The models are very complex~ around 50 predictors some unclear “ad hoc” decisions regarding the estimation procedure
- The EB and the ELL produced the most sensible results. Unexpectedly, the HB did not work (Rao and Molina, 2015)
  - In principle, the HB should as good as the EB and the ELL
- It took CONEVAL more than 2 years to produce its estimates and a research team devoted to this task

The question is:

- Is there a more effective way to produce SAE estimates for Mexico?
  - Use a simpler model (Not 50 but around 10 fixed effects and some random slopes)
  - Faster computation
  - Robust results (at least as good as the EB and ELL).
  - Easy to replicate
  - External validation based on simulations

The Hybrid Bayesian alternative

- One of the main challenge in SAE is dealing with high-dimensional problems
- The HB relies on standard MCMC techniques that are not suited to deal efficiently with complex problems (ad infinitum… yes) (Betancourt, 2017)
- The Hamiltonian Monte Carlo has been put forward as a potential solution but it has not implemented for a real-data SAE problem (Hoffman et al., 2014)
- Even better the amazing work of the STAN group makes the HMC fairly easy to implement
- The HMC in theory is faster and more accurate than the standard MCMC
- Therefore, it should be better than CONEVAL's HB implementation
- If successful it should give the possibility of calibrating many models at a reasonable cost and it should also produce good results with a somewhat simple model (replicability)

Implementation

Data

HIES 2010 and 2014 (n=.25 million). Sample of the Census 2010 (n=12 million) and Intercensal survey 2015 (n=22 million)

Method

Hamiltonian Hierarchical Bayes (HHB)- Rao and Molina (2015) HB estimator based on the HMC

Three-level model fitted to the HIES data (People, Municipalities, States)

10 level-1, 2 level-2 and 1 level-3 variables plus random effects (intercepts only)

Informative and slightly weaker priors were utilized for the fixed effects

Then, the coefficients used for prediction based on the Census

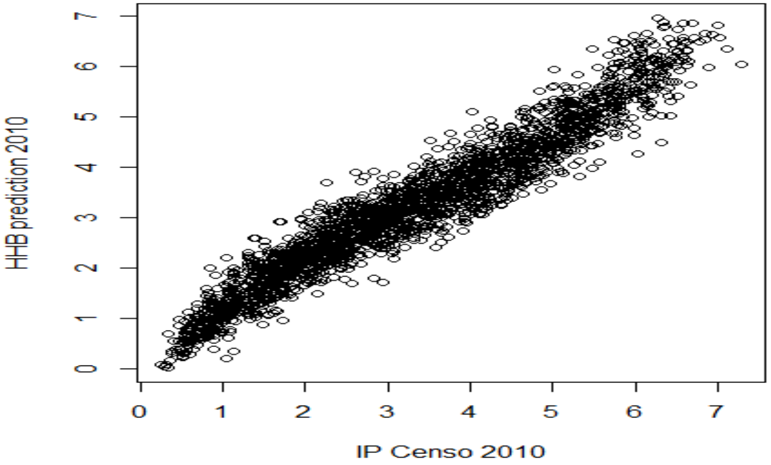
Out-of-sample areas (Only State random effect was used plus the fixed effects)

Mean prediction adjusted by post-stratification following recent approaches that draw on Gelman (1997)

Cross-validation of the HHB with a population parameter generated from the CENSUS 2010 and other two models

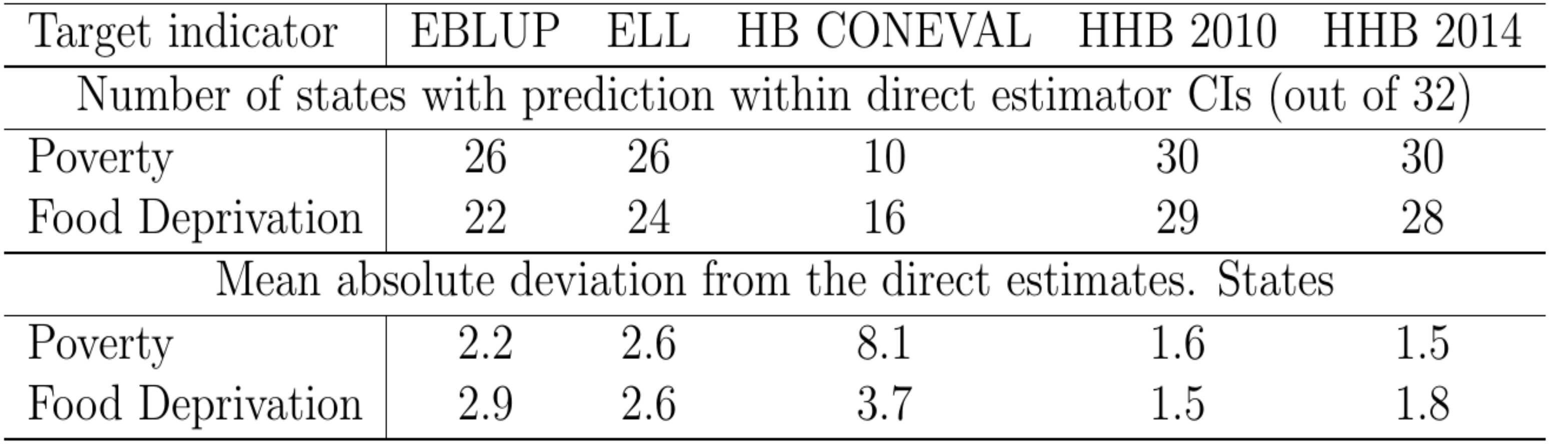
Results

- Validation model. A population parameter (deprivation score) derived from the Census
- The HHB did a very good job in reproducing the population parameter: correlation = .93, population mean = 3.3, HHB prediction = 3.4, Mean Squa



Cross-validation (CONEVAL criteria)

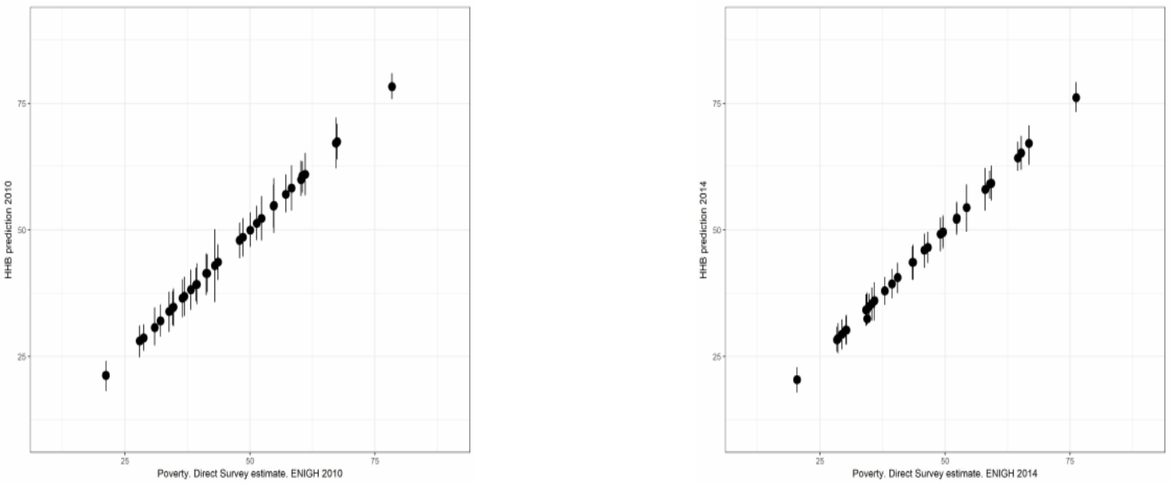
Table 3: Cross-validation of the HHB with CONEVAL's SAE methods using direct estimates as reference



Target indicator	EBLUP	ELL	HB CONEVAL	HHB 2010	HHB 2014
Number of states with prediction within direct estimator CIs (out of 32)					
Poverty	26	26	10	30	30
Food Deprivation	22	24	16	29	28
Mean absolute deviation from the direct estimates. States					
Poverty	2.2	2.6	8.1	1.6	1.5
Food Deprivation	2.9	2.6	3.7	1.5	1.8

Results

Figure 1: State-level poverty prevalence direct and HHB estimates 2010 - 2014. Based on ENIGHs

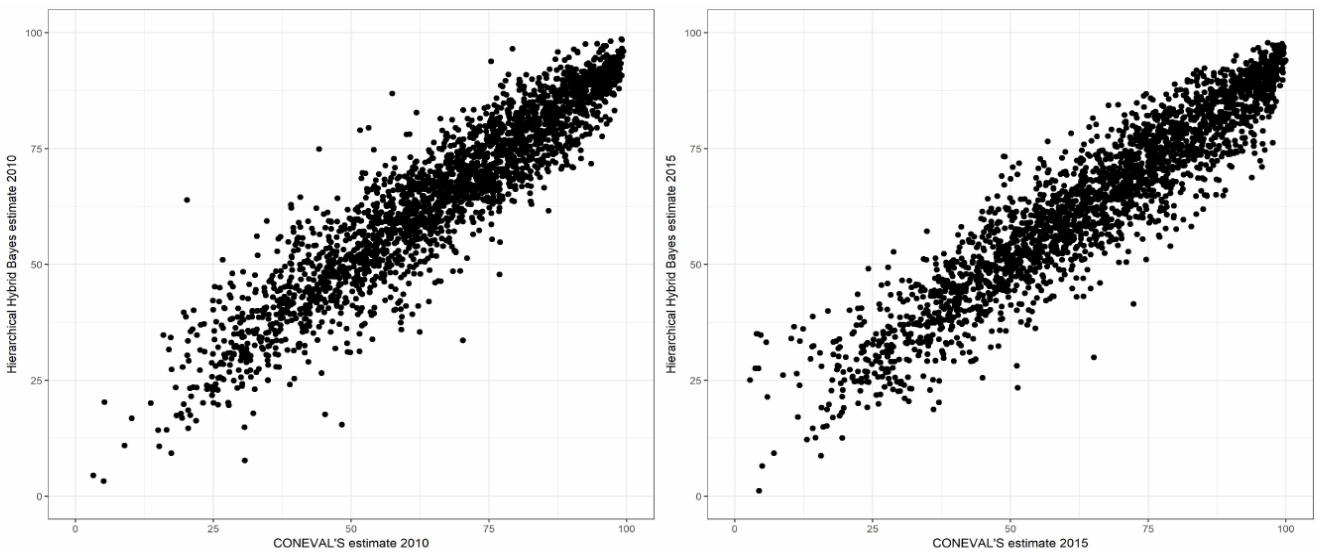


(a) HHB estimator 2010 (y-axis). 32-States. CI's Direct Estimate x-axis

(b) HHB estimator 2014 (y-axis). 32-States. CI's Direct Estimate x-axis

Municipal-level estimation

Figure 3: Municipal-level poverty prevalence CONEVAL's and HHB estimates 2010 - 2015

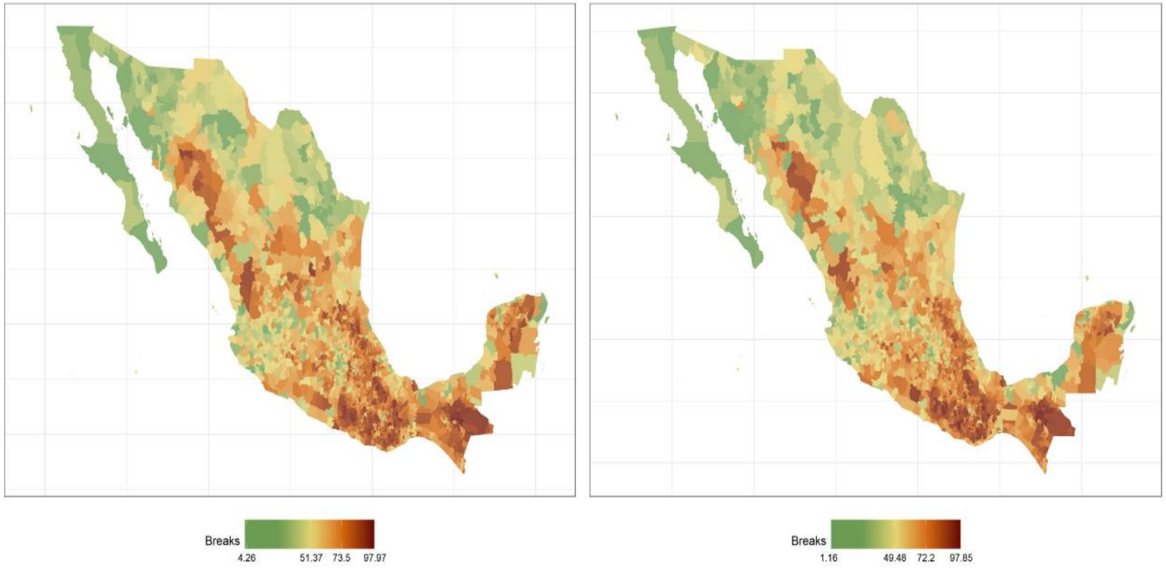


(a) HHB estimator 2010 (y-axis). CONEVAL's x-axis Municipalities

(b) HHB estimator 2015 (y-axis). CONEVAL's x-axis Municipalities

Municipal-level estimation

Figure 4: Small Area Estimates. Based on the HHB method. 2010 - 2015. Municipalities. Mexico



(a) HHB estimator 2010. Municipalities

(b) HHB estimator 2015. Municipalities

According to the HHB estimates poverty increased in 28% and decreased in 38% of the municipalities between 2010 and 2015.

Formally, with Moran's I poverty is clustered. Likewise the increase and decrease followed a spatial pattern

Does it mean it will almost always outperform others with similar data?

- This was the question asked by the reviewers of the paper
    - I absolutely agree with them but I hadn’t done it… So…

What we should expect for similar situations in SAE?

- Often the literature offers examples for small data sets with very few predictors
- The examples focus on continuous variables to illustrate the properties of their estimators (if not, the example is far too simple)
- Rarely this reflects a real-data problem (at least in large developing countries)
- This can be assess using simulations
- The next is working progress…

Simulations

- A Census of 15 million people
- 30 states
- 100 municipalities within state
- Response variable from a Bernoulli process (poverty is often treated as a binary outcome)
- A representative sample was taken using circa 70,000 cases (stratified sampling)
- We should expect poverty to be modelled by a non-trivial model. The data was generated with variables coming from the three levels:
  - 6 individual-level variables (with within state variation- random slopes)
  - 6 level-2 variables
  - 3 level-4 variables

The models for the simulated data

- I decided to compare the EB (which is widely used) with the HHB. I still have to run the standard HB.
- 8 increasingly complex models were fitted to the data to compare the EB and the HHB
  - The first 4 models used random intercepts (level-2 and 3) and only included fixed effects
  - Models 4-8 included random slopes (level-1 across states)

The results were the following:

Results: Time / Feasibility

Computation time in Minutes

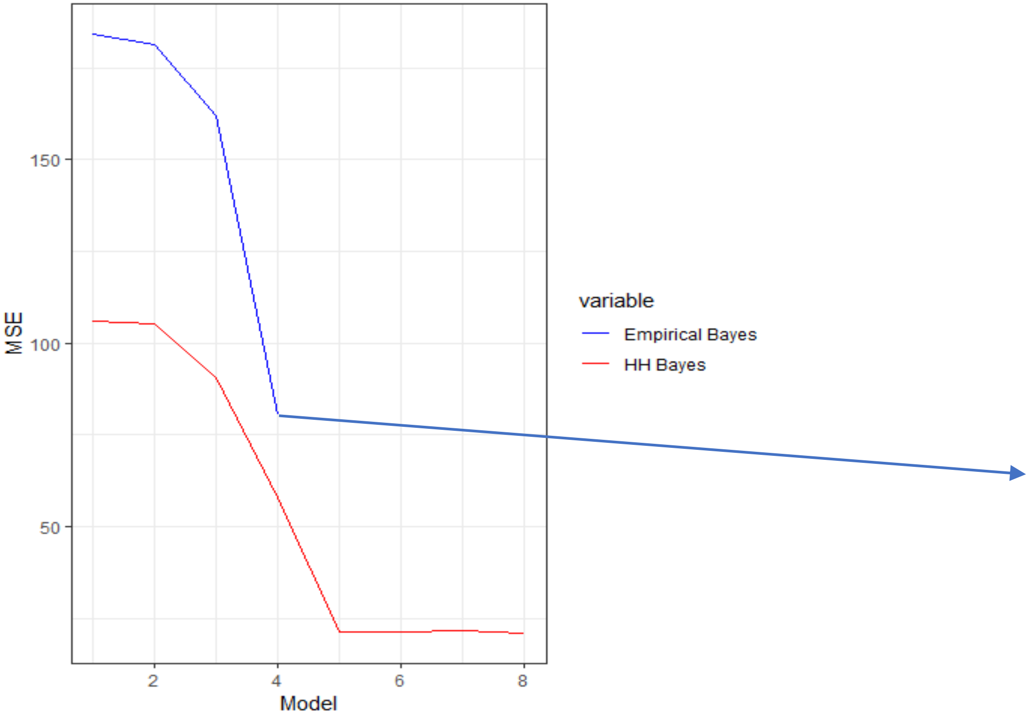
Models	Parameters	EB	HHB
1	4i's, 2j's, 1k	12	50
2	M1 + 2i's	12	55
3	M2 + 1i, 2j's	13	78
4	M3 + 2j's, 1k	14	59
5	M4 + 1j, 1k	180	84
6	M5 + 1 rs_i	na	96
7	M6 + 2 rs_i	na	156
8	M7 + 4 rs_i	na	168

i=individual-level
J=level-2
K=level-3

Random slopes added. The EB underperforms relative to the HHB. It is unlikely that a model without random slopes will work for these kind of data

Adding just one random slope resulted in serios problems for the ML estimator. This is the value for the third iteration!!!!

What were the effective gains?



Unfeasible for the Empirical Bayes

Second session: Poverty mapping and spatial analysis

- So we have produced small-area estimates and then what?
- What are the possibilities with these new data?
- Often people produce a map but that’s pretty much it.
- This is a shame because there is a lot of information in there that you could use to inform policies

Poverty mapping

- Everyone can draw and understand a map… because they have a rough idea of where things are located
- Maps often are easier to explain than a table or a plot
- The first poverty study utilized mapping…
- Mapping for many years was very time very time consuming
- Now we have computer software that makes things easier for us
- For many years the available software was very expensive and few people knew or could use the tools to map out things

Why mapping?

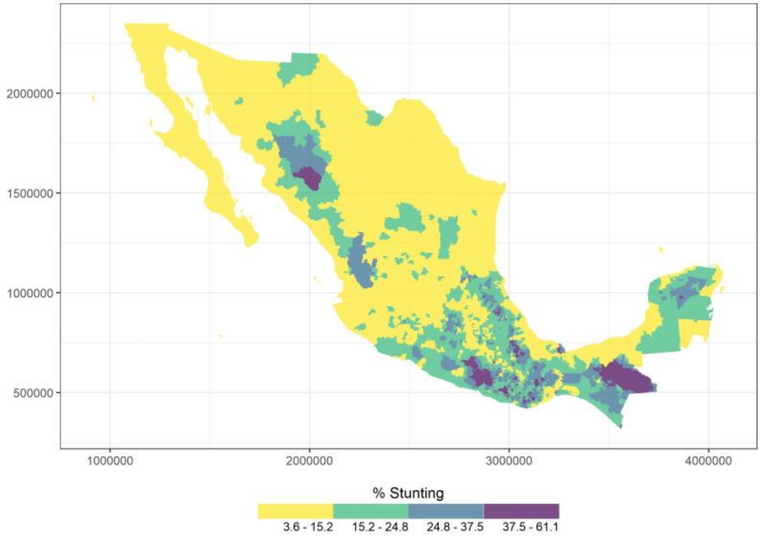
Toble’s law of geaography:

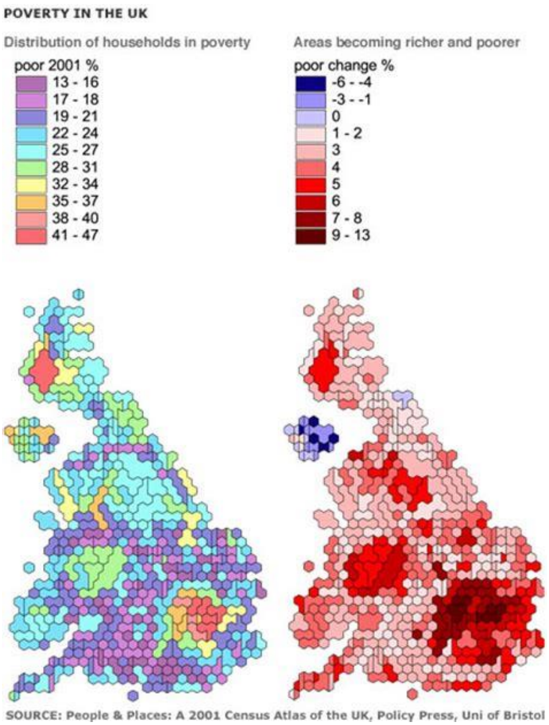
“Everything is related to everything else, but near things are more related than distant things”

Location is one of the best predictors of almost anything you can think about

Certainly, it is one of the best predictors of poverty

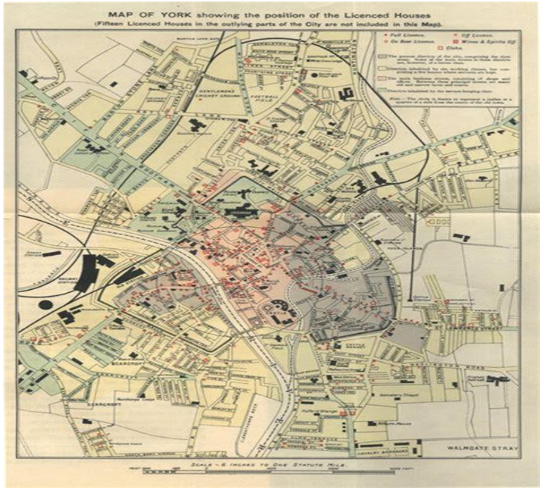
Why mapping?

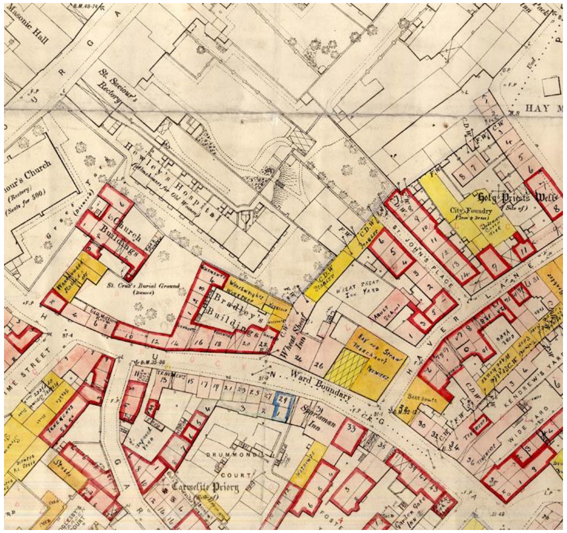



POVERTY IN THE UK

SOURCE: People & Places: A 2001 Census Atlas of the UK, Policy Press, Uni of Bristol

Now.. The first poverty maps look like this:


MAP OF YORK showing the position of the Licenced Houses
(Fifteen Licenced Houses in the outlying parts of the City are not included in this Map).
WALMGATE STRAY



Of course you can produce a map by hand. But how long is going to take you? How accurate is going to be?

Changes in computation have shifted things quite dramatically in recent years.

Now is very easy!

Data preparation for mapping

- We need two things:
  - A database with our TARGET (i.e. the indicator we want to map).
  - Cartography data. A series of files that basically draw the boundaries of a map

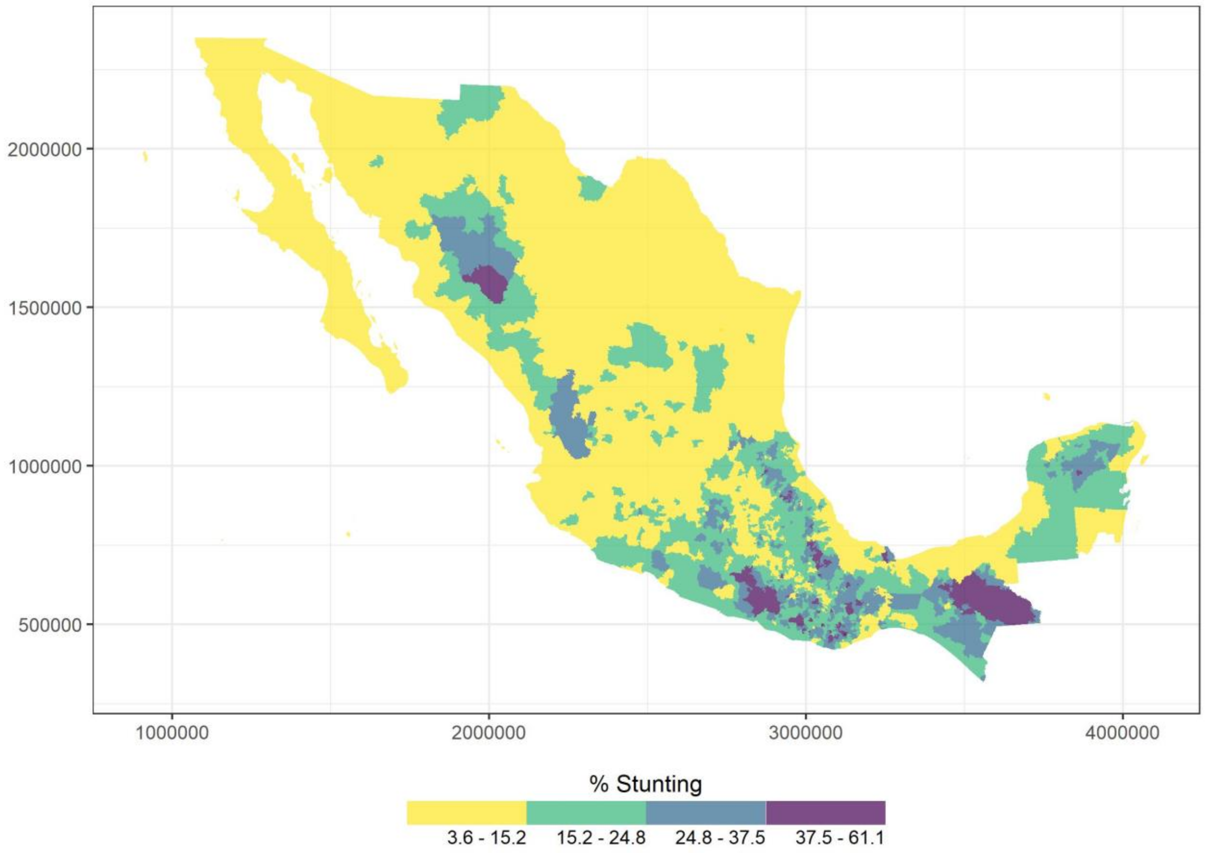
- Therefore, both kinds of data need to be matched in some way. Poor data preparation for mapping is what gets people frustrated

Database with our target variable

- A key when mapping geographical data is to know the level at which the variable is measured:
  - Area-level data (Villages, Regions, States, Countries, etc). These can of data are also called “polygon” data

Let’s see some examples!!!

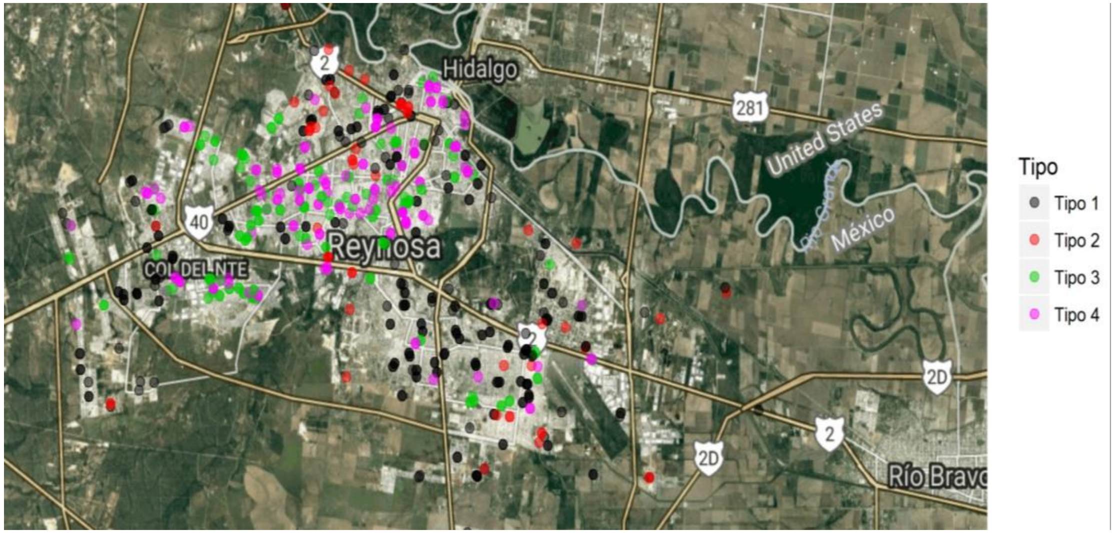
Example: Area-level data. Stunting across the Mexican municipalities



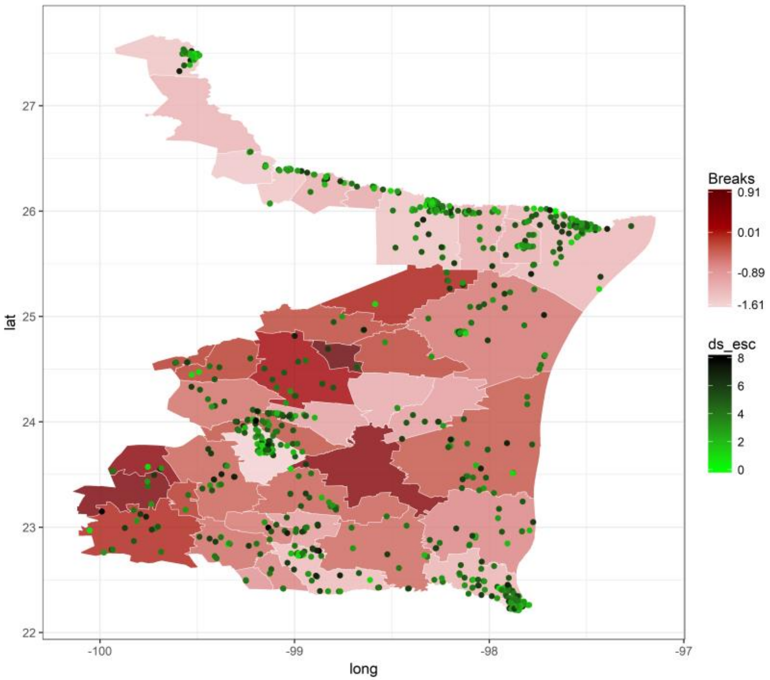
The second main type of spatial data:

Point data (Households, industries, shops, etc). These can of data are rarely available in social statistics due to confidentiality

Sometimes we have access to school-level data, for example



Of course you could combine both



Schools in Tamaulipas by poverty rate. Municipalities. Mexico 2018

Area-level data

- Area-level data, therefore, often are proportions, rates, aggregate numbers, counts, etc
- This is due to the fact that for an area we are targeting the set of observations that are inside the area:
  - Households in a block, village, etc….
  - Number of people in a country

  - This sounds obvious but the implication is that the Census data and Survey data are aggregated in some way to represent certain area (Islands for example)

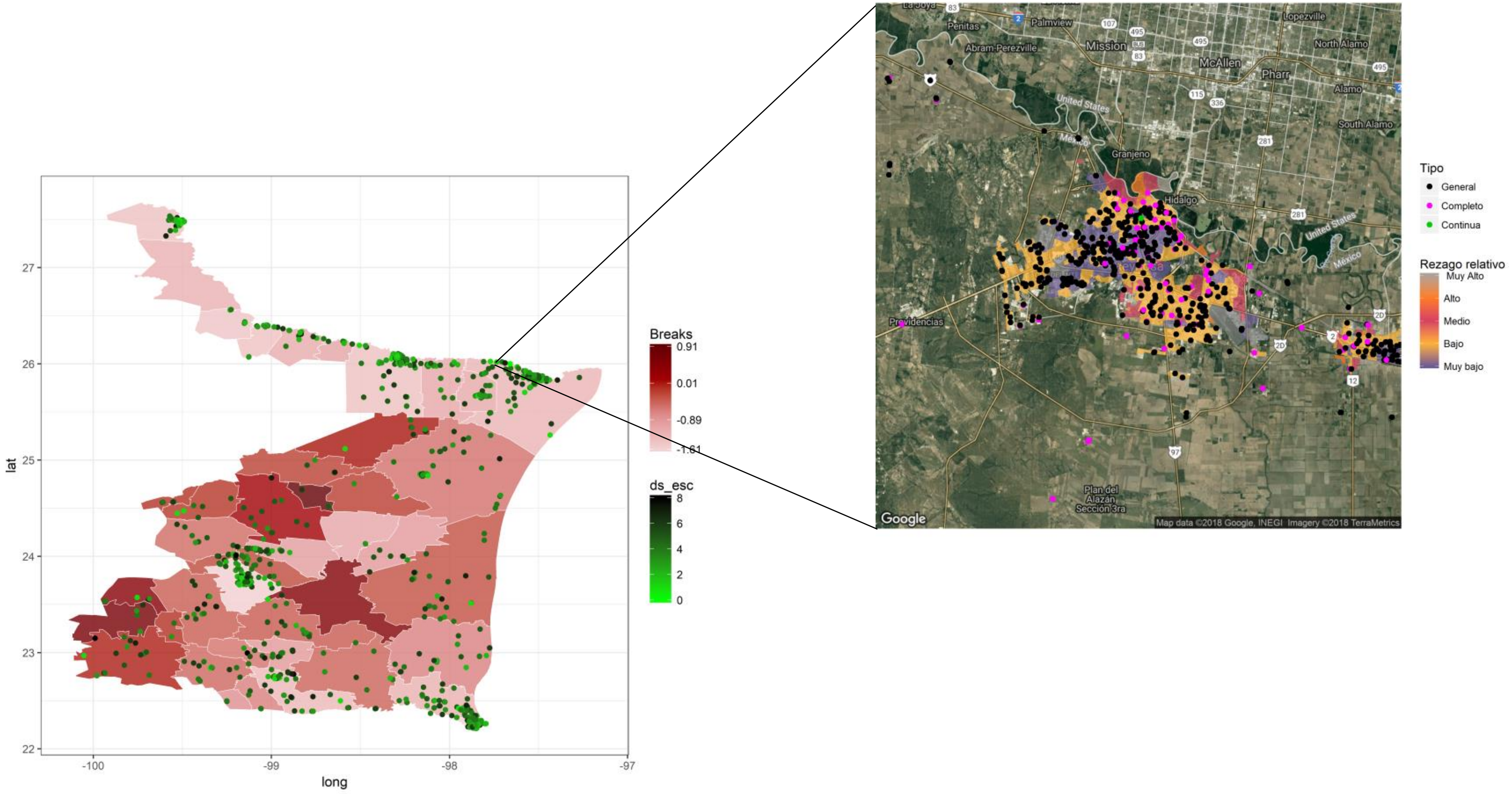
Area-level data

- The fact that we aggregate our data by certain administrative division has, nonetheless, some disadvantages:
  - Modifiable areal unit problem (MAUP): Your conclusions will change depending on the unit of analysis
  - Violation of the first law of geography because of the use of arbitrary boundaries

MAUP problem

- Let’s say you have two blocks in the same constituency
- One has high poverty rate and one low poverty rate
- If you map both using a constituency level map… Their differences will disappear and your conclusions are going to change

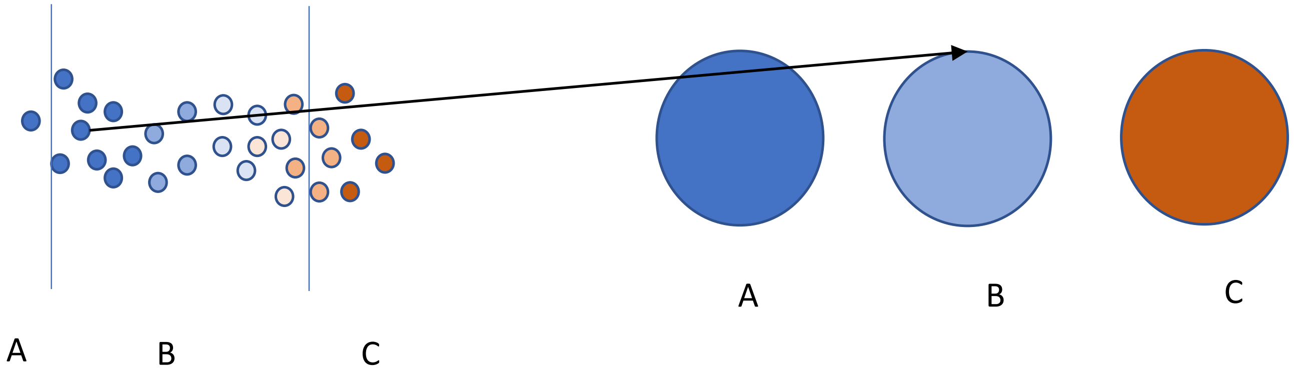
MAUP problem



Tobler’s law of geography

- Now imagine that you zoom in to the dividing line between both blocks… What you will see is that the households close to the line are going to be similar. Therefore, the division line is also influencing our conclusions

Actually you have lots of darks in that group



Poverty mapping

Producing a map these days is very simple. You need the following things

1. Geocoded tabular data, i.e. the prevalence rate of poverty at county level with the codes for each county
2. The shapefiles (cartography) that contain the polygon data (i.e. the boundaries of each county)
3. You join these two sources and you are good to go!

- There are several software alternatives out there. You can classify them into two groups

Point-and-click mapping

1. ArcGIS: Very good but you have to pay for it
2. QGIS: Very good and open source
3. GeoDa: Very good, sligthly limited but open source

- The disadvantage with point-and-click is that reproducing your findings is not easy!

Syntax-based

1.QGIS + Phyton: Good and free

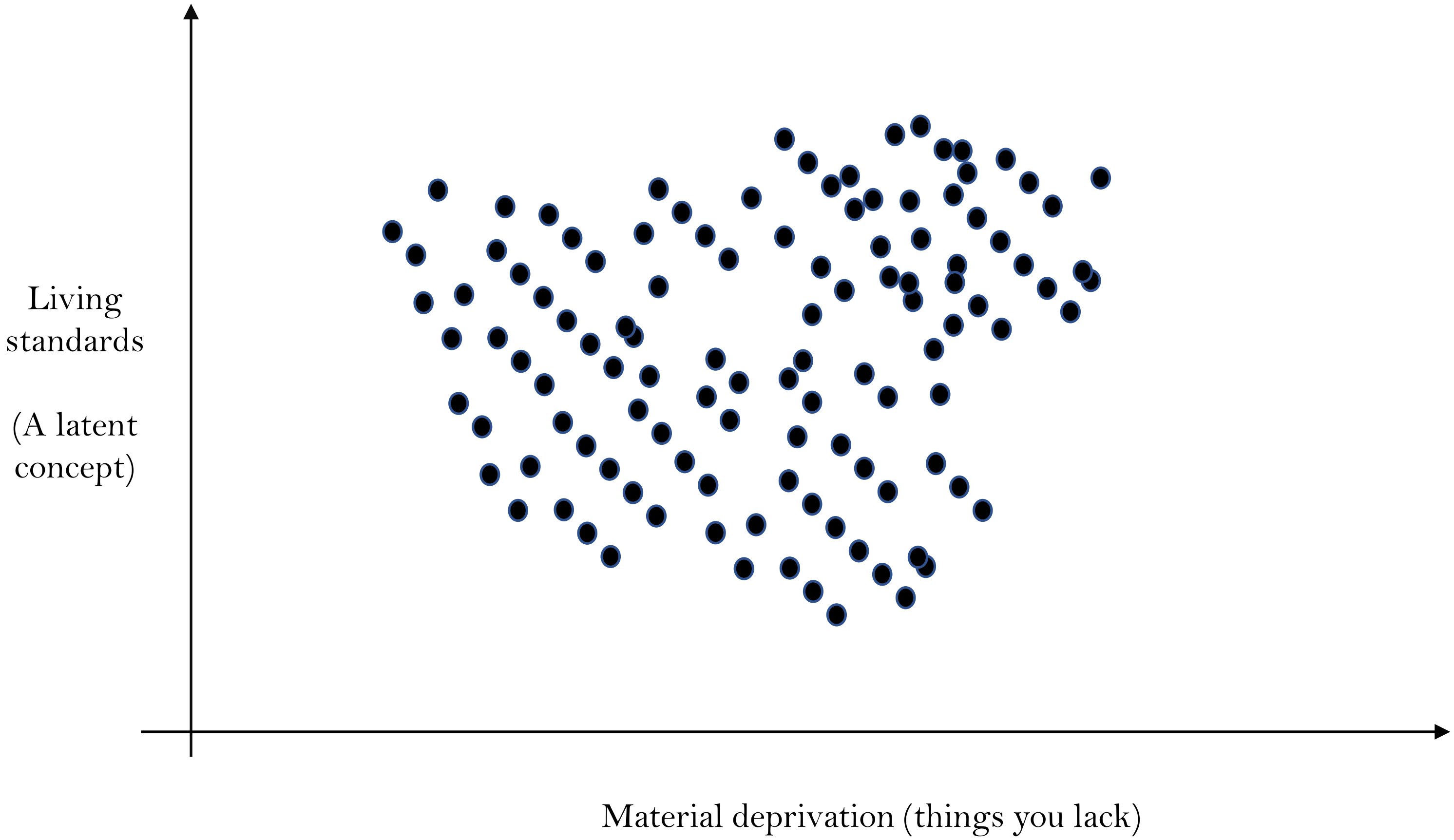
2.R (ggplot2): Several nice packages to produce maps

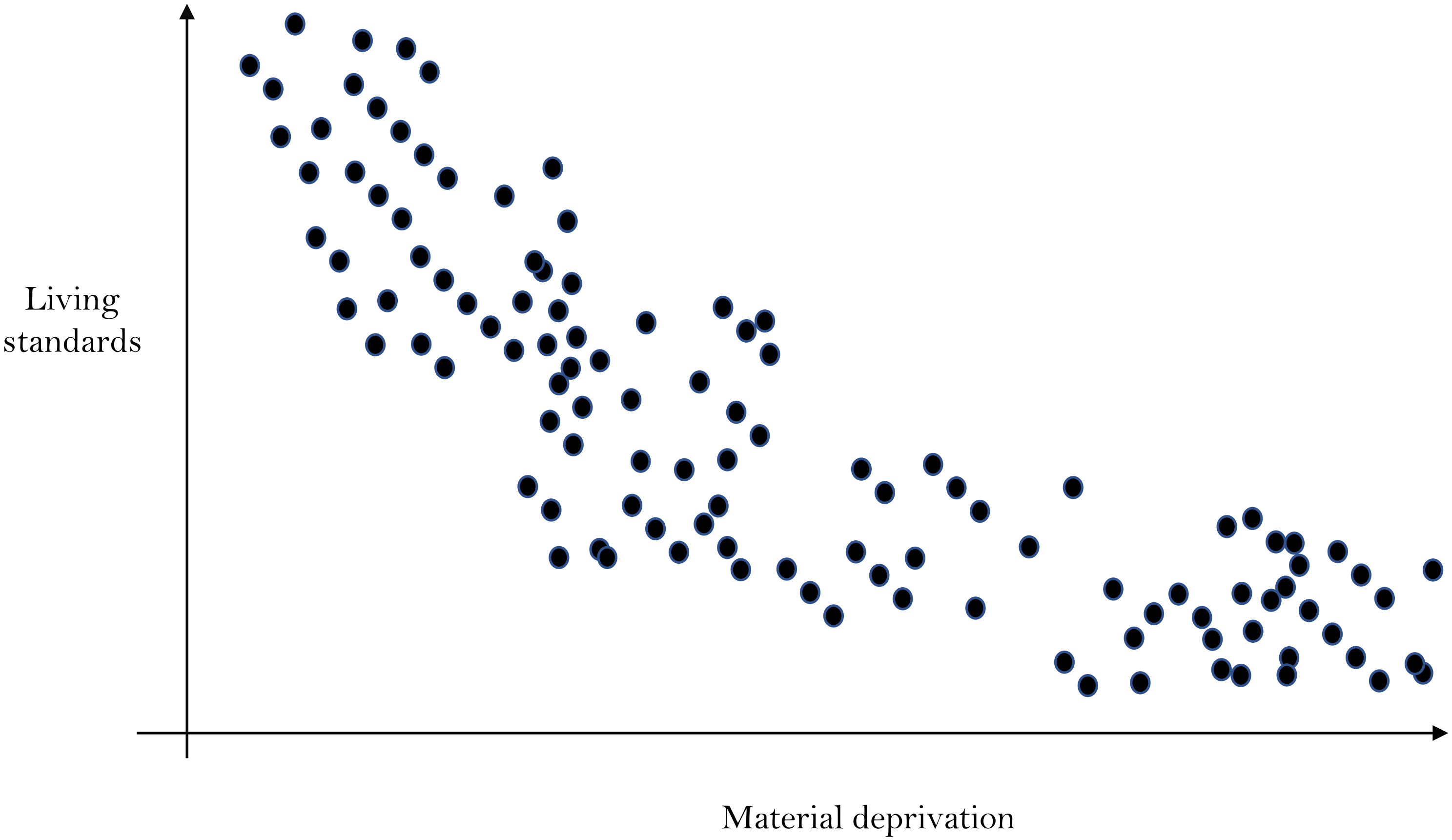
3.SPSS and Stata: No idea, but very few people use them…

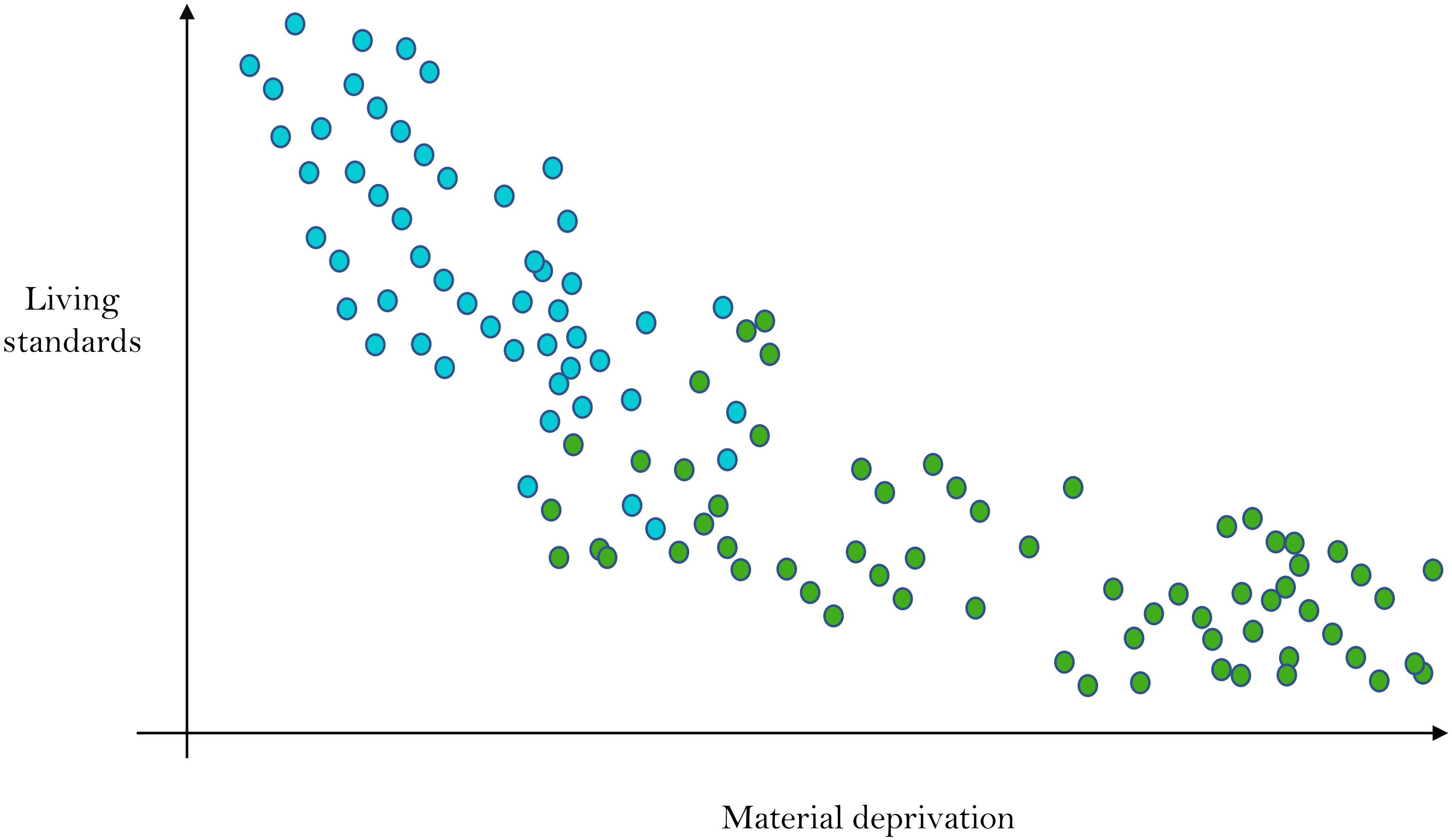
- The advantage is that you just need to rerun your syntax to produce a map!
- I would recomend to start with QGIS and then move into one of the syntax-based options

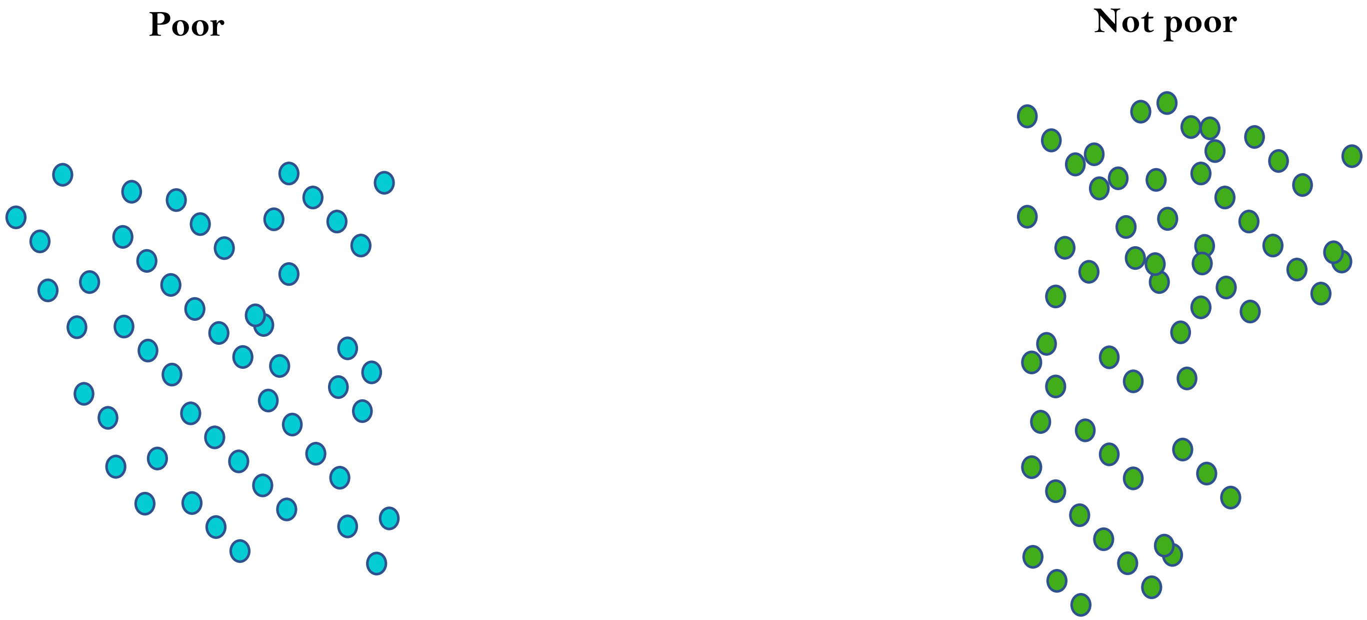
Spatial analysis of poverty

- Looking at a geographical pattern is fine
- However, this constitutes just a very basic step in spatial analysis
- It is the equivalent of a tabulation
- What kind of questions you can answer when you work with spatial data?
  - Is poverty spatially autocorrelated?
  - Where is poverty more spatially autocorrelated?
  - What are the spatial predictors of poverty?


Living standards
(A latent concept)
Material deprivation (things you lack)


Living standards
Material deprivation

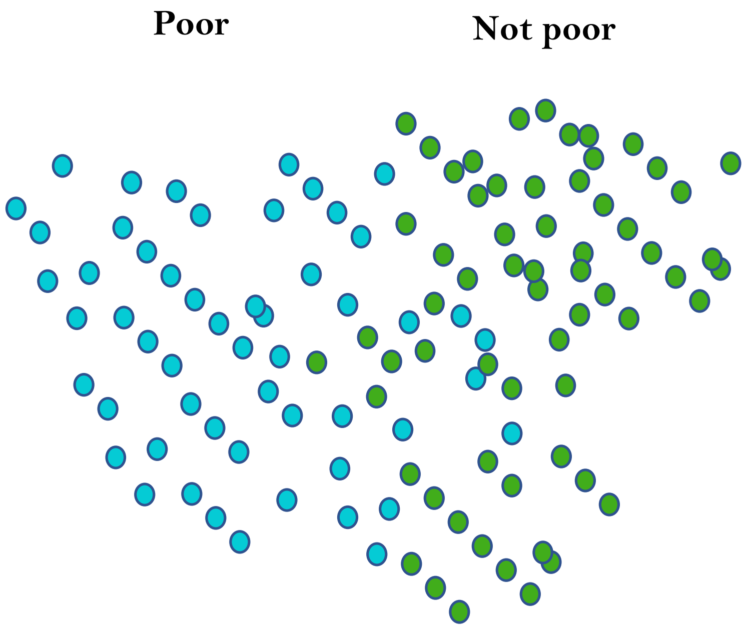

Living standards
Material deprivation



Poor

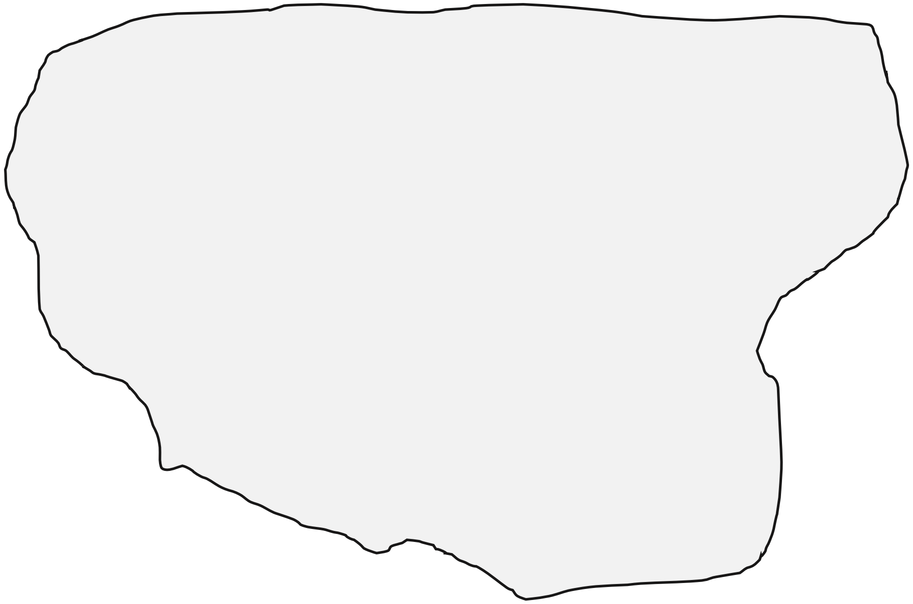
Not poor

Ideal classification

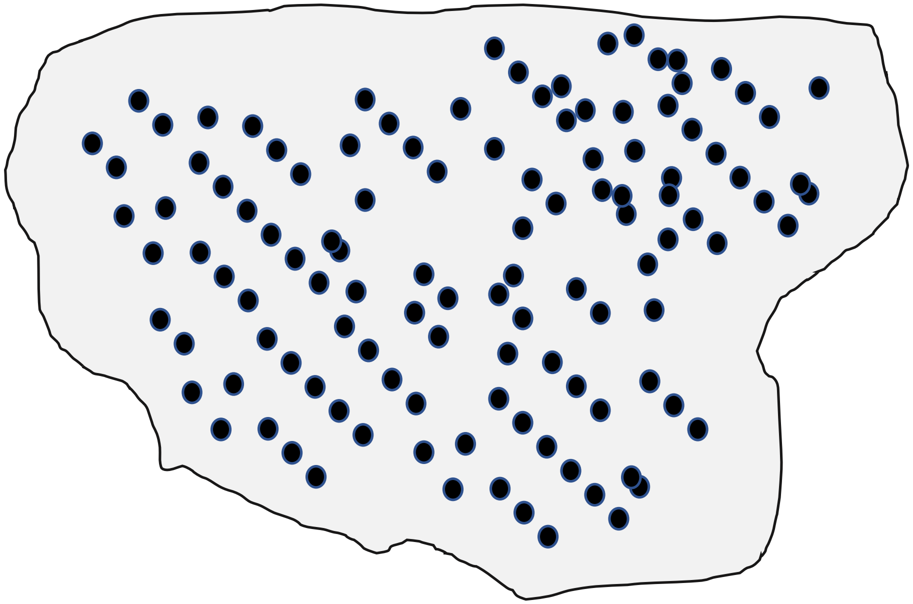


In reality we will have something like this

Imagine a grid (i.e. a piece of land)

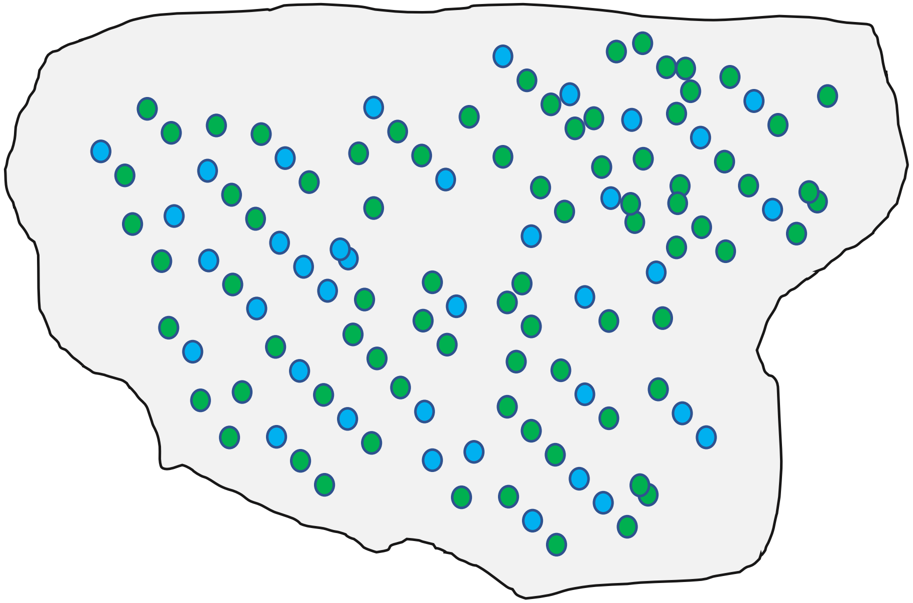


No pattern (random allocation)



No pattern (random allocation)

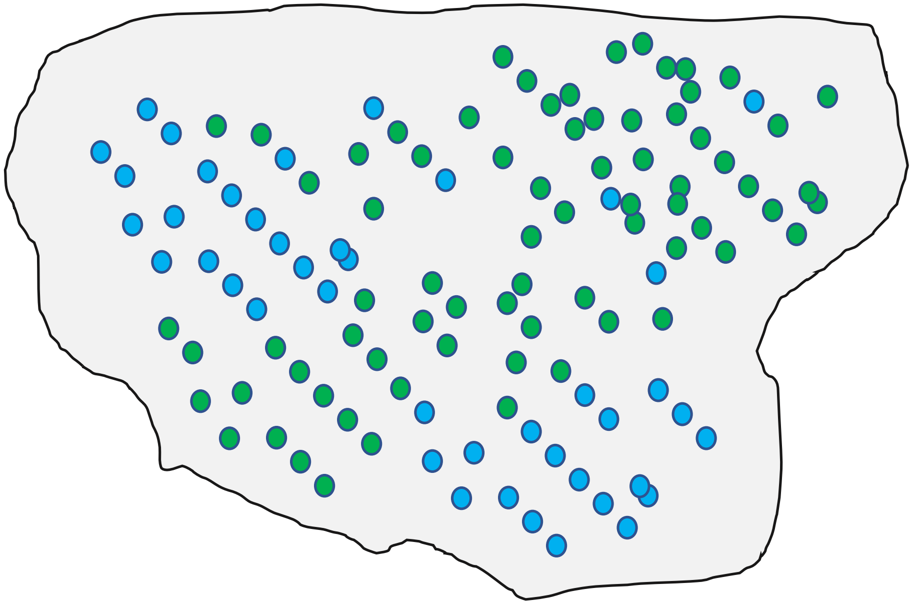
Blues and greens are equally likely to appear on a given location on the grid



Tobler’s prediction

Blues and greens are NOT equally likely to appear on a given location on the grid

Blues are more likely to appear “North west and south east”



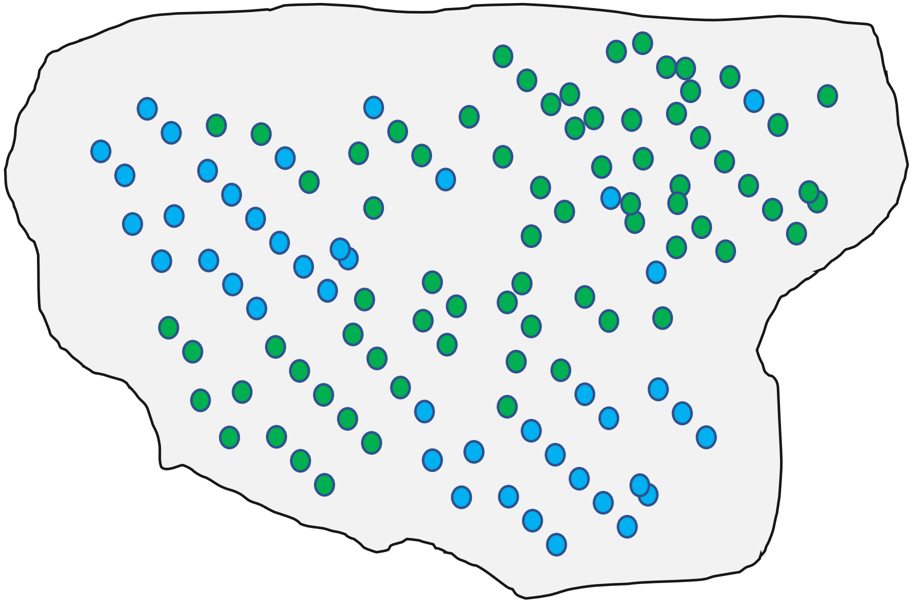
Two types of questions: Descriptive and explanatory

Visually the pattern seems to exist BUT:

How can we formally know that a pattern exists?

Statistically speaking

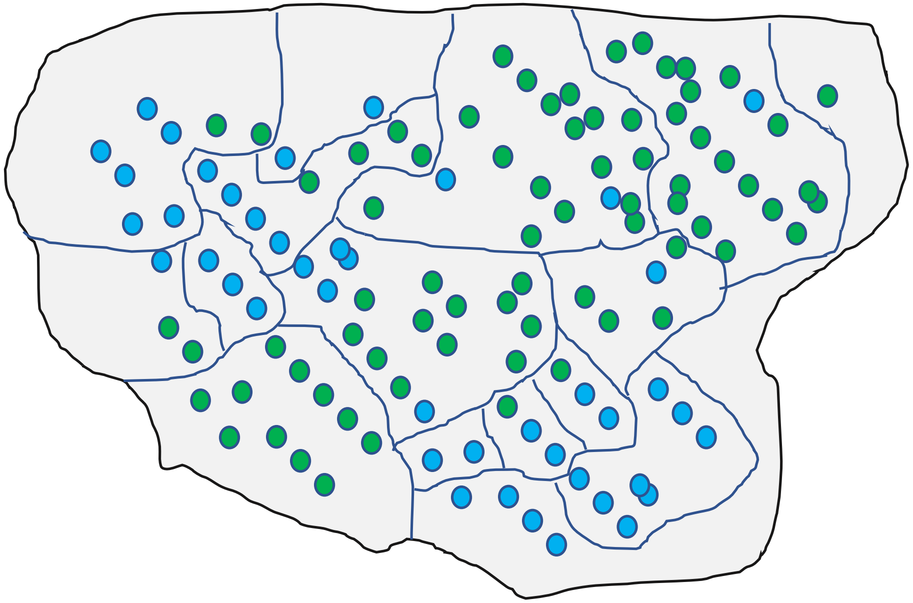
Is there any spatial autocorrelation?



Two types of questions: Descriptive and explanatory

How our conclusions are affected by the unit of analysis?

How the (arbitrary) administrative locations influence our conclusions?

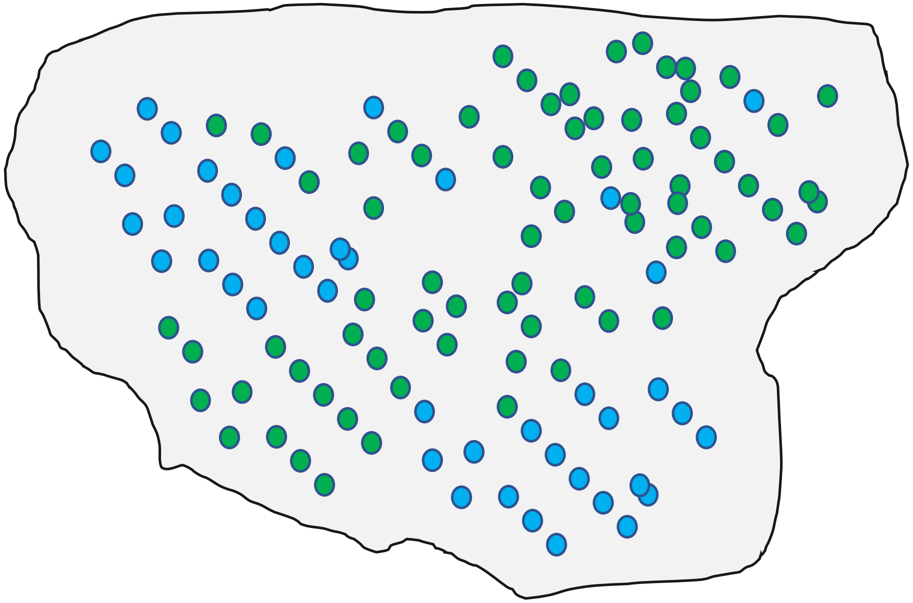


Explanatory

Why blues and greens
Are distributed this way?

Why blues in the north west?

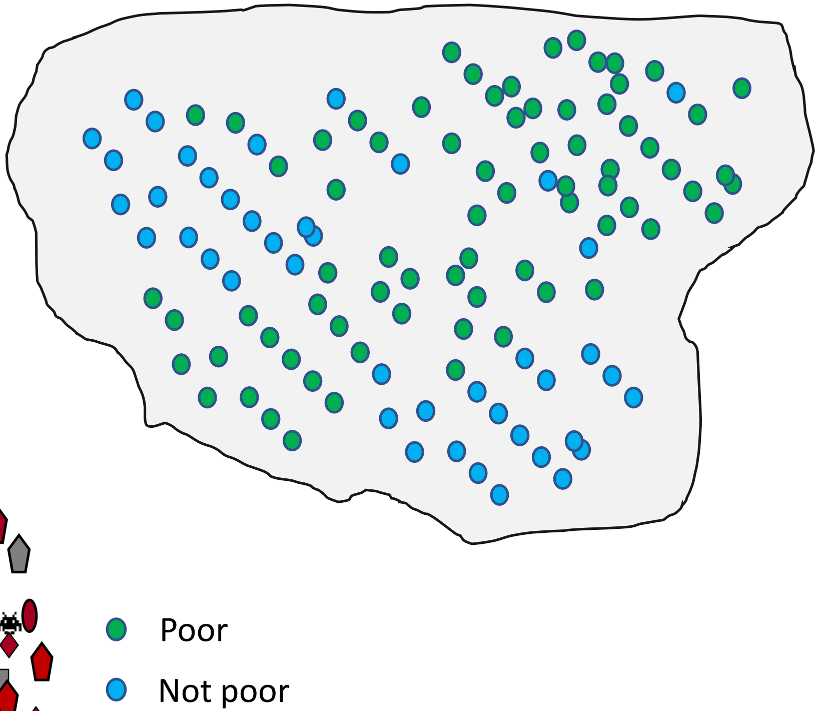
Is it geographically or
individually driven?



These are questions about the explanations of spatial inequalities… why does it exist?

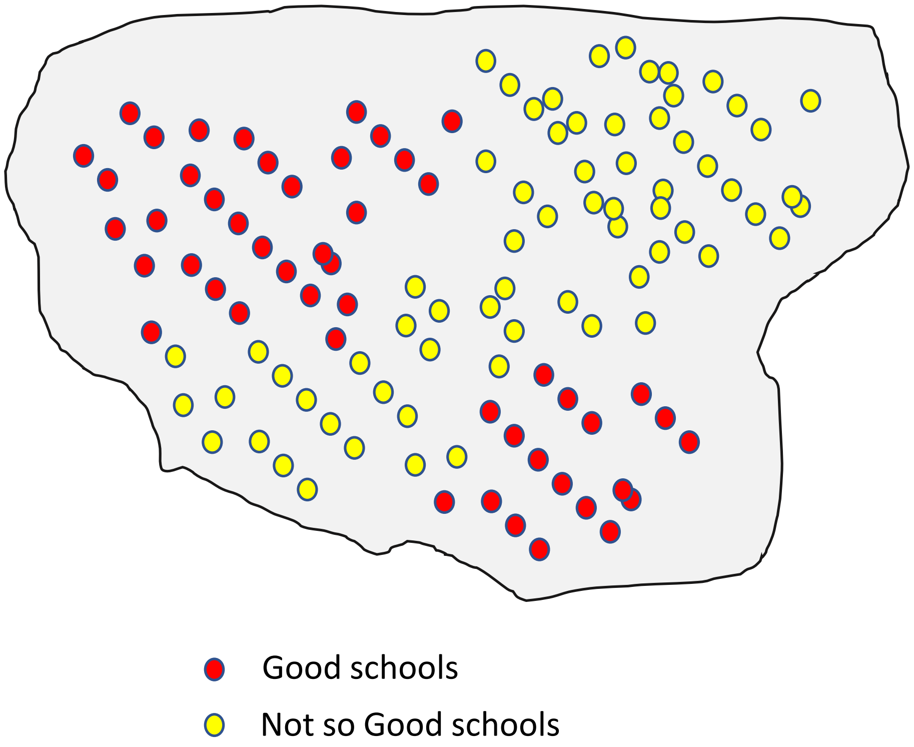
Explanatory

There is some (bivariate) “ecological” spatial autocorrelation



🟢 Poor

🔵 Not poor



🔴 Good schools

🟡 Not so Good schools

Measures of spatial inequality

- There are several measures of inequality (GINI, Theil, Entropy)
- These measures help to have a general (non-spatial) idea of inequality
- However, when studying spatial inequalities we would like to consider location

Formal analysis of clustering and spatial autocorrelation

- Just as in a simple correlation, one looks at the extent to which the values of a variable change due to distance
- What distance? There are several measures (neighbourhood, linear kms, radius, etc).

- The most popular measure is Moran’s I. Which is just a simple spatial correlation coefficient

Spatial Correlation

Think about this room

Think about your location

Think about the distribution of gender

Think about whether there is some clustering

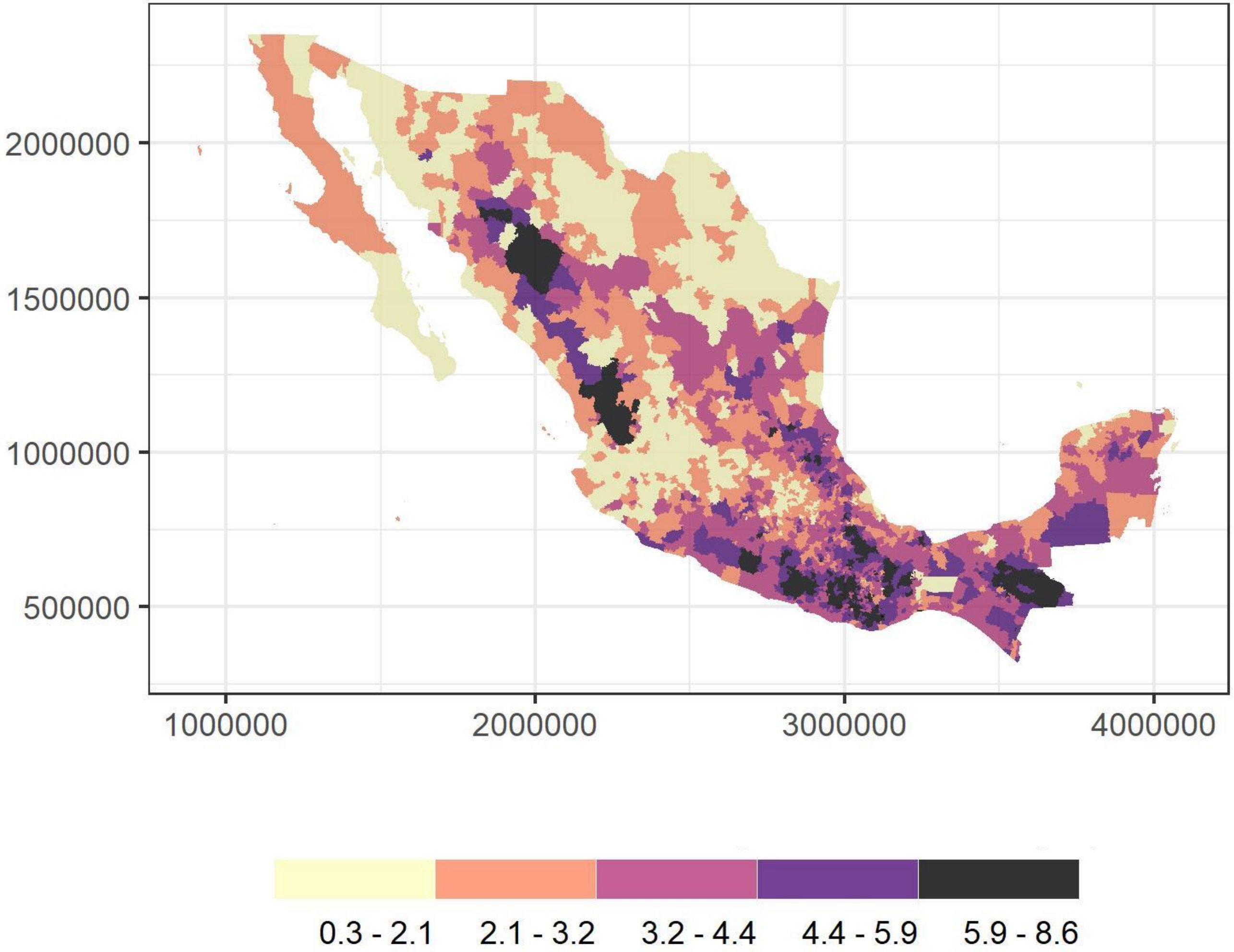
Moran's I considers your gender and the gender of the person next to you and it does so for the rest of the group

If everyone has someone that is next her/him of the same gender… Moran's I will be high (close to 1)

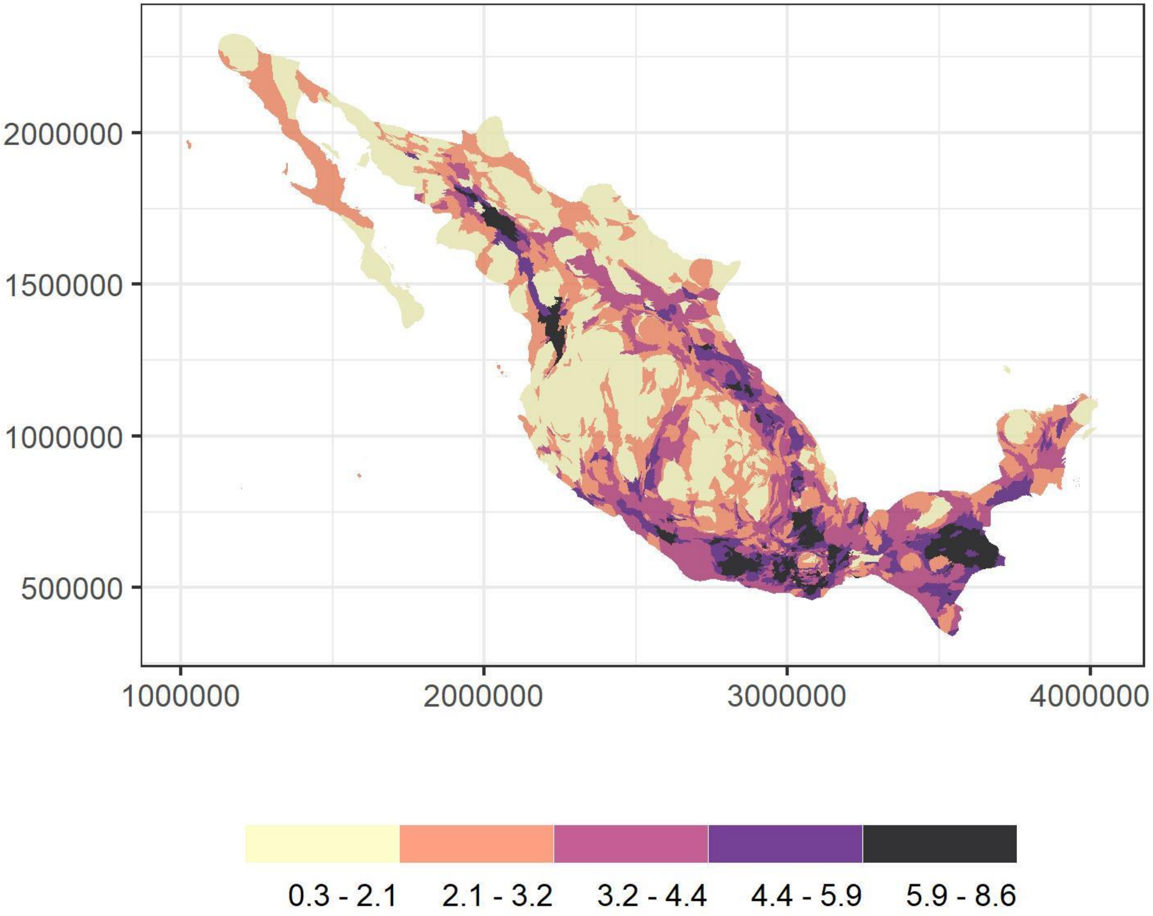
Correlation is good but far from enough

- We would like to know whether there is some clustering but also the “location of the clusters”
- Local Indicators of Spatial Association (LISA) use local Moran’s I to assess where the correlation is stronger.
- Let’s see an example

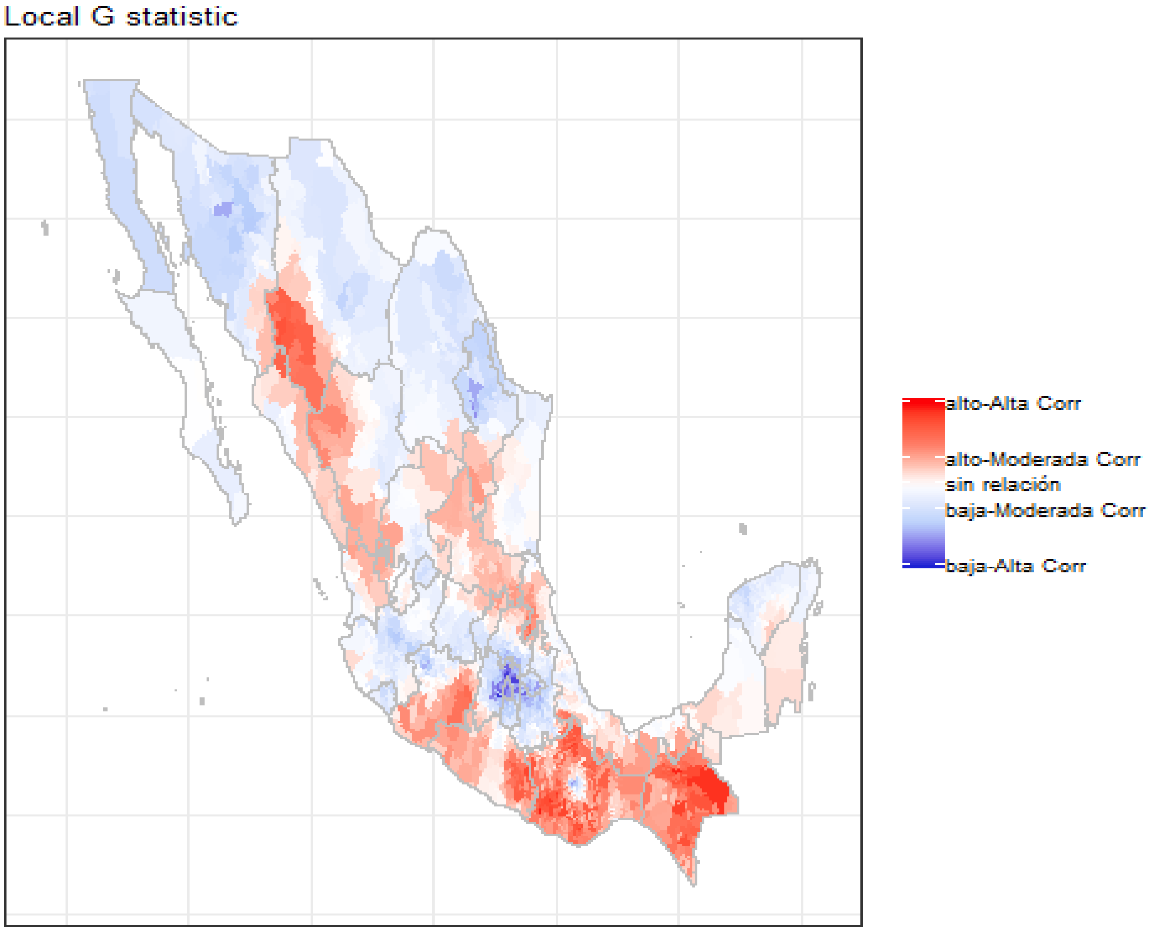
Average child material deprivation in Mexico. 2015



Cartogram. Child material deprivation. Mexico 2015



Local spatial autocorrelation. Child material deprivation



Thanks!

- Dr. Héctor Nájera

hecatalan@Hotmail.com