Exploring Accurate Spatial Downscaling using Optimization

Michael Poss[*3] **and Didier Josselin**[1,2]

[1]UMR ESPACE 7300, CNRS, Université d'Avignon, France
[2]LIA, Université d'Avignon, France
[3]UMR CNRS 5506 LIRMM, Université de Montpellier, France

*Corresponding author: michael.poss@lirmm.fr

Abstract

This papers proposes to combine several downscaling processes based on expert knowledge and objective functions from Operations Research to fill a table of social data about illiteracy in the USA during the Thirties.

Keywords

Downscaling, social data, Operations Research, optimization, data about illiteracy, USA, 1930

I THE PROBLEM OF SPATIAL DOWNSCALING ON CENSUS DATA

This paper deals with the well-known downscaling problem (Bierkens et al. (2000)) related to the ecological inference fallacy (King (1997)), change of support problem (King et al. (2004)) or aggregation problem (Josselin et al. (2008)) in general. Many applications and research fields (environment, social science, geography) are concerned and this problem notably impairs census data analysis (Gehlke and Biehl (1934); Robinson (1950); Tranmer and Steel (1998)).

1.1 Example

Let us consider that we have information (attribute values) for a complete statistical population of a territory. We want to disaggregate this information into a set of smaller sub-areas that compose the whole territory, at a finer scale. To do so, we need to make assumptions on the way we distribute the values into the spatial partition elements. Fortunately, we also have a few relevant information about the complete population and about individuals aggregates, for all the categories of the variable, and also in each sub-area. But we do not know the exact distribution of the category-specific values in each area. We have to make assumptions to estimate it. This is illustrated in the Table 1 with an example about housing.

Region	**House owner**	**Tenant**	**Free accommodation**	***Inhabitants (millions)***
PACA	?	?	?	5
Rhône-Alpes	?	?	?	6.5
Centre	?	?	?	2.5
Quantity	8	5.6	0.4	**14**

Table 1: 14 millions of people from 3 French regions to distribute in 9 cells of the table according to different types of housing.

Depending on the number of sub-areas (e.g. regions) and classes of the attribute (e.g. housing types), there exist many ways to fill the partition of the Table 1. At this point, those are all equivalent solutions to the problem since they respect the constraints of the (known) summed values in the row and column margins. For instance, the most simple table (that would correspond to a freedom degree of only one in a contingency table) with only 20 individuals and 4 cells, i.e. 2 categories for both variables, has already 5 possible solutions.

To find the solution that fits the reality the best way –in the unreachable case of knowing the complete table content, i.e. how many people really live in a given region with a given type of housing–, we sometimes need to use statistical tools or to fix complementary assumptions related to complementary knowledge.

However, we are not completely deprived of means. What is interesting here is that we can firstly set simple constraints based on aggregated information. Indeed, we know that:

- the sum of the values in every line or column should be equal to the corresponding margin values;
- the sum of the margin values should be equal to the total number of individuals.

To find what is the "good" solution, then we have to fix an objective to maximize or minimize, according to assumptions from experts. This is the main idea of this paper. On the one hand, getting reliable information using census is very costly and cannot be made very frequently. On the other hand, there are experts in social science who may know a lot about socio-spatial characteristics of people. It can strongly help in making assumptions to drive disaggregation process, although we know that we definitely cannot find the real and exact distribution of the data at a given time. However, it seems anyway better than reading in a crystal ball.

In this paper, we propose an approach mixing statistics and optimization to find an optimal data distribution in a table, according to aggregated constraints and expert knowledge. Our objectives are the following:

- getting a better accuracy in statistical data due to a downscaling process driven by a statistical criterion optimized using a mathematical solver;
- finding the solution the most related to a given statistical criterion or to expert knowledge translated to a new matrix of expected values;
- assessing the quality of the solution when taking into account uncertainty in downscaling estimations, providing min and max values in each cell of the contingency table.

This method is tested on data studied by Robinson (1950), for which we know the complete matrix. Thence it enables to compare the results of our method to the real observed data.

1.2 Our data set: USA census data about illiteracy in 1930 used by Robinson (1950)

We use the data about the illiteracy in the population of USA in 1930, which are reference data in the field of sociology [1]. They include the number and the percentage of illiterate people in each State and in regions, by population color and nativity. We study four groups of population:

- *native white with native parentage* people;
- *native white with foreign or mixed parentage* people;
- *foreign born white* people;
- *black* people.

In this paper, optimization methods are used for the whole matrix M and results are provided and only discussed about black people in USA in 1930. Indeed, we test a complementary hypothesis on the role of history (slavery in confederate states before civil war) in black population illiteracy (for instance, see map provided in the Figure 1).

[1] http://www.ru.nl/sociology/mt/rob/downloads/

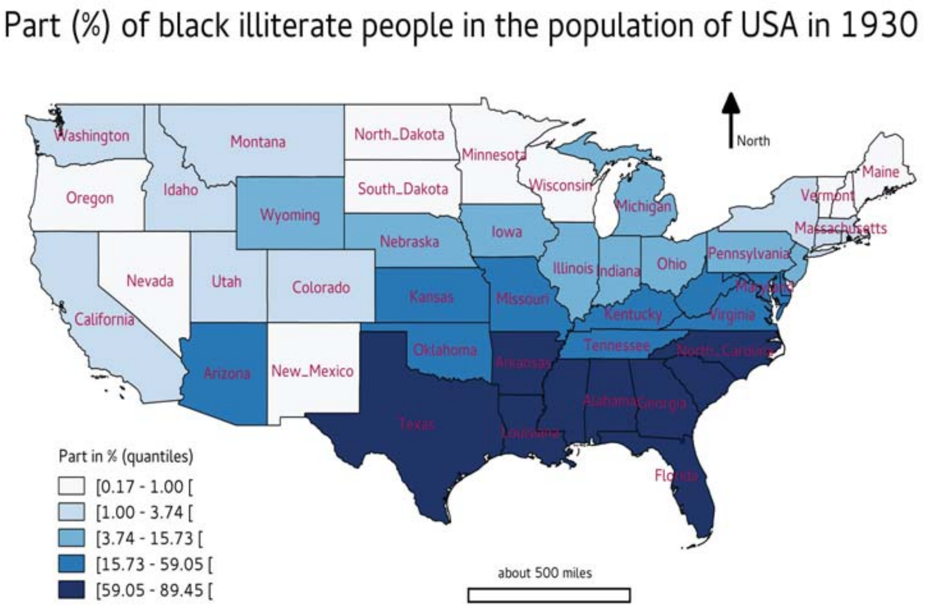

Figure 1: Illiterate black people in USA in 1930: a location related to history (unified *vs* confederate states in East-Southern).

Using these data, Robinson (1950) noticed that the correlation between illiteracy rate of black people and the different regions was close to 1. But if we consider the statistics at another more accurate scale (the States), it goes down to 0.2. This lead Robinson to define the *ecological fallacy* in social science and more generally on census data. On a practical point of view, it seems indeed very difficult to apply a knowledge learnt using the same statistics provided at a certain aggregated level, on a more disaggregated and finer level. That is the purpose of our proposition: trying to improve local estimation accuracy and downscaling process by modifying the data according to hypothesis and new constraints from aggregated statistics and from expert knowledge. Because we know the complete information about illiteracy in each type of population and in all the States in USA in 1930, we can compare and discuss the results of our method to the real data of the census. In this first stage, we fix the objective function criteria and the expert assumptions by ourselves to test the method.

II A METHOD TO DOWNSCALE CENSUS DATA

2.1 Downscaling in areal units

In geography, it is very common to disaggregate spatial data into more accurate layers. Usually, a hypothesis of proportionality is applied. The assumption is the following: space is somehow isotropic and densities are the same all over geographical space or inside a given type of characterized area. This corresponds to the $H0$ hypothesis in Chi^2 statistics. For instance, as in a contingency table, the expected number of individuals $\hat{x_{ij}}$ in a cell of the table is estimated by the product of the margin values of the corresponding column $x_{i.}$ and row $x_{.j}$, divided by the total number of individuals $x_{..}$. Here we consider that illiteracy and regions where individuals are located are independent statistical variables. If there are significant differences in the number of individuals between regions or between classes of the variable, this can lead to very various values in the cells of the table.

Another way to compute downscaling may be to find the solutions minimizing a certain statistical criterion. For instance, variance of the values in the matrix M of the table can be minimized. In this case, experts consider a kind of homogeneity in downscaling, i.e. that cell values should not differ too much from each others whatever the margin value weights. This approach is slightly different from the previous one, because a global statistical criterion is now considered. Moreover, it is possible to minimize many different functions $f(\theta)$ on such a table, according to several Lp-norms as we shall see in the next sections.

In an analog way, experts can use information about sub-areas peculiarities to provide probabilities or ranges (probability min and max) in each cell. Then the problem leads to find the solution that is the closest to the probability matrix M given by the expert. Practically, experts can provide bounds in which (s)he expects the values to be, in each cell of M. Whatever the function to minimize and levels of uncertainty (intervals), the optimal solution can be reached using an optimizer, as we propose in the following section.

2.2 Variability criteria to optimize

We denote by x_{ij} the value of row i and column j in the resulting matrix M and $\bar{x}$ the mean of all x_{ij}.

$$\theta_{ij} = |x_{ij} - \bar{x}| \tag{1}$$

We aim at minimizing 4 different objective criteria $f(\theta)$, each of them representing one approach to weight the variability among the different components of x. Then, given each specific objective criteria, we analyze the resulting optimal distributions of $\hat{x}_{ij}$ in M.

A first criterion is based on the L_∞-norm, which illustrates an equity process: outliers have an important weight in this solution.

$$f(\theta) = \max_{i,j}(\theta_{ij}) \tag{2}$$

A second is similar but takes into account squared residuals θ_{ij}. Outliers are even more considered.

$$f(\theta) = \max_{i,j}(\theta_{ij}^2) \tag{3}$$

The following case corresponds to the least absolute deviation case (L_1-norm) which depicts an efficiency purpose.

$$f(\theta) = \sum_{i,j} \theta_{ij} \tag{4}$$

This last criterion is the variance and corresponds to the L_2-norm (equality).

$$f(\theta) = \sum_{i,j} \theta_{ij}^2 \tag{5}$$

The drawback of the objective functions (2) and (4) is that many optimal solutions may exist, and in particular, optimal solutions that contain many components of x set to 0 which does not provide interesting insights. To avoid that situation, we smoothed the objective functions by replacing θ with θ' defined as

$$\theta'_{ij} = \theta_{ij} + 0.001.\theta_{ij}^2 \tag{6}$$

Smoothing ensures unicity of the optimal solutions and, more importantly, increase the number of positive coefficients.

These functions are tested with different data relating to different assumptions and in two cases:

- No information is provided about the range of the value in each cell (no information known or expressed about uncertainty of $\hat{x}_{ij}$);
- In each cell, the value is bounded and the solution must take into account this constraint to be included in the interval (uncertainty modeling for $\hat{x}_{ij}$).

2.3 Mathematical model: optimization under constraints

Let C_j be the given value for the sum of all elements in column $j \in \{1, \ldots, m\}$ and R_i be the sum of all elements in row $i \in \{1, \ldots, n\}$. We denote by x_{ij} the value of row i and column j in the resulting matrix M. The problem of finding the optimal values for x can be stated as the following optimization problem

$$(P) \equiv \begin{cases} \min & f(\theta) \\ \text{s.t.} & \displaystyle\sum_{i=1}^{n} x_i = C_j \qquad j \in \{1, \ldots, m\} \\ & \displaystyle\sum_{j=1}^{n} x_j = R_i \qquad i \in \{1, \ldots, n\} \\ & x \geq 0 \\ & \theta_{ij} \geq x_{ij} - \bar{x} \quad i \in \{1, \ldots, n\}, j \in \{1, \ldots, m\} \\ & \theta_{ij} \geq -x_{ij} + \bar{x} \quad i \in \{1, \ldots, n\}, j \in \{1, \ldots, m\} \end{cases}$$

where the function f models the preference used in the construction of the matrix M as explained in the previous section, and the last two groups of inequalities are a linearization for the definition of θ from (1). Notice that the objective functions involving $\max$ can be linearized by introducing the artificial optimization variable Θ linked to θ_{ij} through

$$\Theta \geq \theta_{ij} \quad i \in \{1, \ldots, n\}, j \in \{1, \ldots, m\},$$

and optimizing the objective function $f(\Theta) = \Theta$. Finally, the model can be completed by adding bounds on variables x_{ij} as explained in the previous section.

Since the four cases of functions f presented in the previous section are convex quadratic functions, (P) turns to a convex quadratic and linearly constrained optimization problem, which can be solved very quickly with state-of-the-art optimization solvers (e.g. CPLEX, Gurobi).

2.4 Several cases and hypothesis to test

For all the studied cases, the process is similar: we set an hypothesis that will change the estimation of each value x_{ij} in M. Then we recompute the sums of individuals in rows and columns and we adjust the constraints accordingly (e.g. bounds). Several types of solutions are searched, based on margin values and according to different hypothesis:

- Models without optimization:
  - *[Variable independence $H0$].* As explained in the introduction, this model is based on the $H0$ table of contingency; so $\hat{x}_{ij} = x_{i.}x_{.j}/x_{ii}$;
  - *[Global Illiteracy Rate GIR].* In the second model, we apply the global rate (%) of illiterate people observed in the whole USA on each type of population in every State;

  - *[Population Illiteracy Rate PIR].* In this model, we take into account the part of illiteracy in each type of population (%);
  - *[Effect of Old Confederacy EOF].* Here we add a historical hypothesis based on whether or not a given State belong (or was close) to the Confederation area during the civil war in USA (1861-1865); indeed, this could partly explain the capacity of a State to potentially integrate black or foreign population including language learning facilities. The probability of illiteracy is multiplied by 2 for the Confederate states and by 1.5 for a few bordering States which used to allow slavery and to belong to the Union however.
- Models with optimization but without any local information about value range (no information about uncertainty, in all of them, we only use recalculated margin values in rows and columns):
  - *[Raw Data RW].* We try to find out an optimal solution;
  - *[Global Illiteracy Rate GIR].* We look for an optimal solution using GIR;
- Models with local bounded values (the solver finds a solution given new local constraints of uncertainty in each matrix cell); we use recalculated margin values in rows and columns and we fix min and max bounds in each cell of M using two different formula 7 and 8; are concerned by this procedure:
  - *[Global Illiteracy Rate GIR]* data (for the whole USA);
  - *[Population Illiteracy Rate PIR]* data (by type of population);
  - *[Population Illiteracy Rate PIR]* data (by type of population) updated by *[Effect of Old Confederacy EOF]* data.

For the models with local bounded values, we apply a function to find the minimum $inf(\hat{x}_{ij})$ and maximum $sup(\hat{x}_{ij})$ values of the interval centered on the estimated $\hat{x}_{ij}$:

$$sup(\hat{x}_{ij}) = round(\hat{x}_{ij}.k^{1/length(\hat{x}_{ij})}) \tag{7}$$

and

$$inf(\hat{x}_{ij}) = round(2\hat{x}_{ij} - sup(\hat{x}_{ij})) \tag{8}$$

These functions (equations 7 and 8) allow to have symmetric intervals, larger for low values of $\hat{x}_{ij}$ (indeed $length$ counts the digits of $\hat{x}_{ij}$) depending on k.

For models dealing with uncertainty, we consider two cases:
- $k = 2$; for instance $[109 - 185]$ and $[6023 - 8835]$
- $k = 4$; resp. $[0 - 294]$ and $[2363 - 12495]$

III RESULTS

We focus on the relation between illiteracy and black people in USA in 1930. To do so, we compute a LS regression model between observed data and estimations we obtained. We study the coefficient of determination r^2 and the slope of the regression. Closer to 1, better the quality solution, for both indicators.

In the Figure 2, we can notice that using optimization increases the estimation quality when expert assumption is weak (case with $H0$ and Global Illiteracy Rate GIR). For other hypothesis (Population Illiteracy Rate PIR and Effect of Old Confederacy EOF), optimization has no effect compared to estimations based on accurate hypothesis from experts.

Globally, Figures 3 and 4 show that:

 [ISBN: 978-2-9105-4510-5]

- The results are better for bounded optimization methods;
- The difference between objective functions is not marked; indeed they all somehow compute a sort of variability minimization;
- Generally, solutions are better with $k = 2$ (more narrow intervals);
- Most of the time, functions including the max operator seem to get better estimates;
- The best optimization reaches a coefficient of determination r^2 of 0.86 and a $slop$e of 0.76, both quite close to 1.

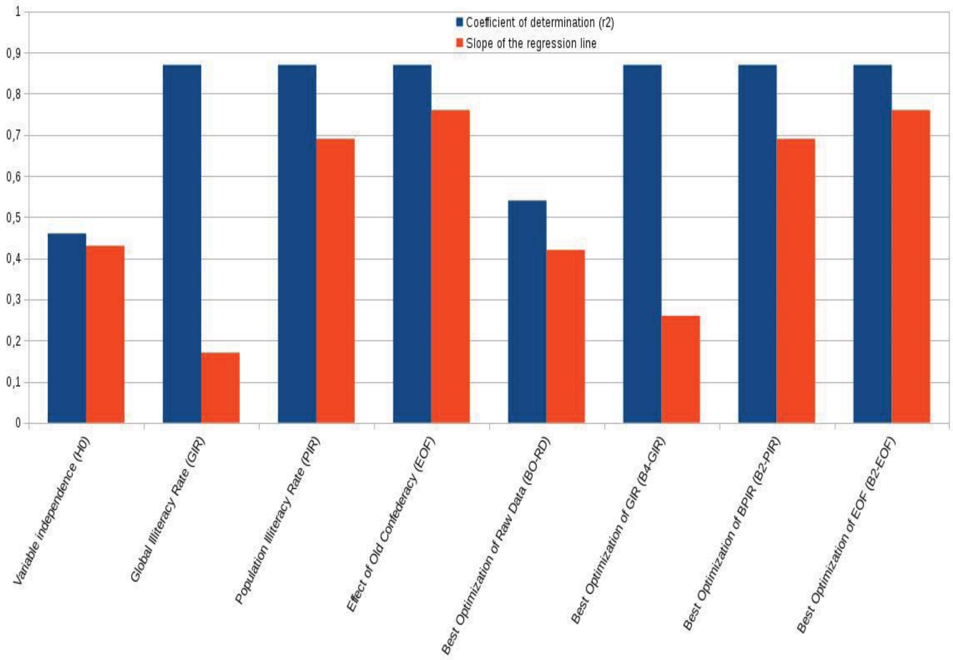

Figure 2: Comparison of different estimation methods according to the values of the coefficient of determination r^2 and of the slope of regression line (on left, 4 solutions without optimization; on right, 4 best solutions using optimization). Only the black population in USA in 1930 is considered.

IV CONCLUSION

In this paper, we explore how we can mix optimization methods and expert knowledge translated in statistical method to improve data estimation in contingency table downscaling. Although the results do not strongly demonstrate that the use of optimization is determinant on the result quality compared to other current methods from experts of the domain, it shows that, in certain conditions (weak assumptions or bounds knowledge from experts for instance), estimations can be improved, especially prediction quality (slope of the regression line between estimations and observed data).

Moreover, using optimization allows to find the most optimal solution among a large set of possible solutions, according to a given objective. These solutions can be then compared and participate in the data exploration.

Other experiments will be made, on the whole set of data whatever the type of population nativity, also on other types of data and with other compared methods (multiple regressions for example) to assess in which conditions, such a mixed approach may be useful for predicting data in contingency tables.

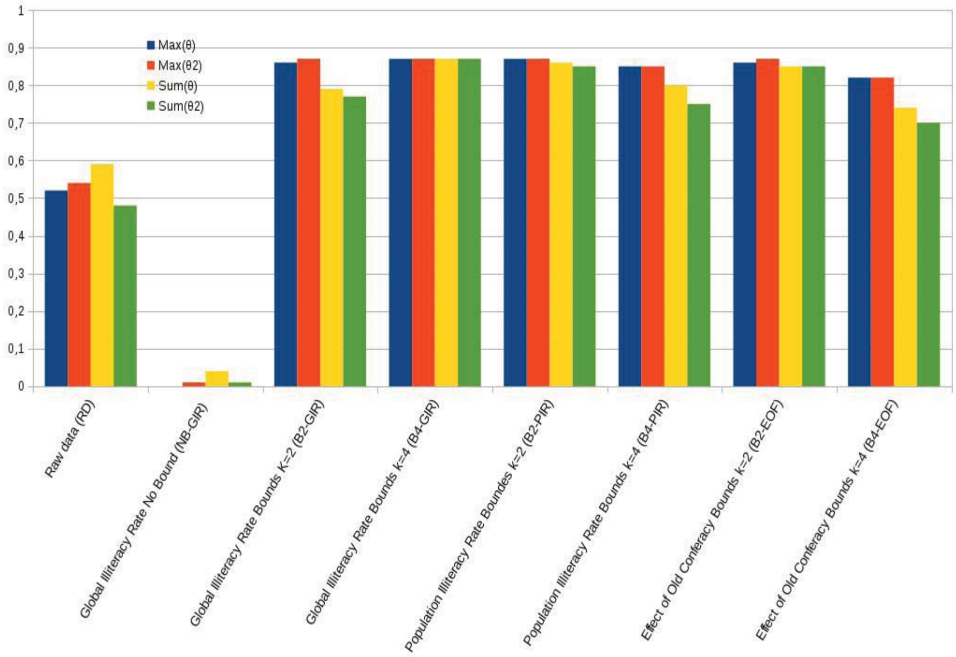

Figure 3: Comparison of different estimation methods according to the r^2 of the matrix M and to 4 different objective functions of θ. Only the black population in USA in 1930 is considered.

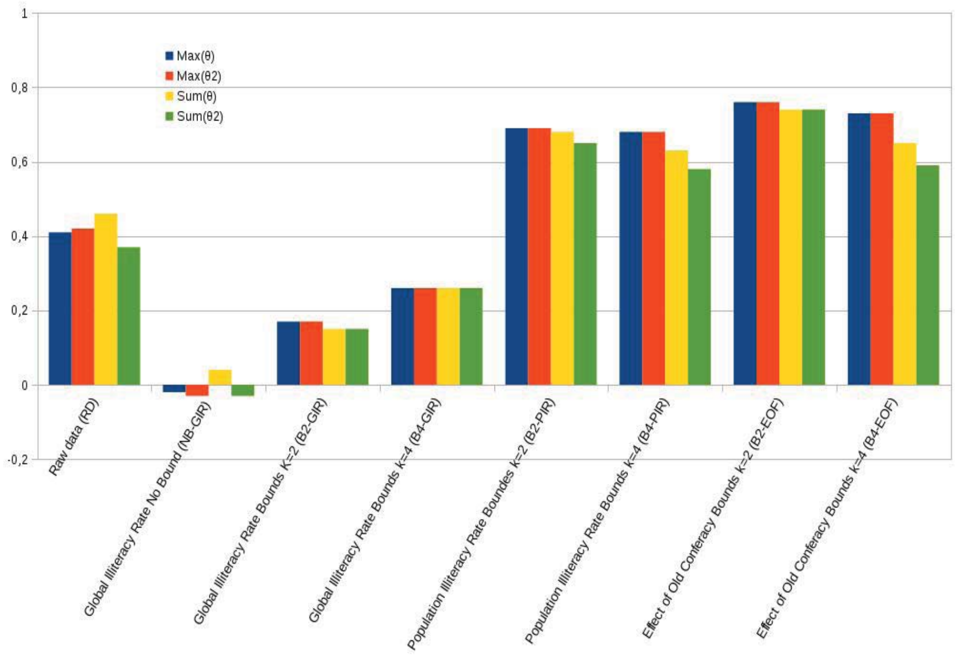

Figure 4: Comparison of different methods to estimate the slope of the regression line according to 4 different objective functions of θ. Only the black population in USA in 1930 is considered.

References

Bierkens M. F., A. F. P., de Willigen P. (2000). *Upscaling and Downscaling Methods for Environmental Research*. Springer, Series Developments in Plant and Soil Sciences.

Gehlke C., Biehl H. (1934). Certain effects of grouping upon the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association supplement 29*, 169–170.

Josselin D., Mahfoud I., Fady B. (2008). Impact of a change of support on the assessment of biodiversity with shannon entropy. In *Spatial Data Handling, SDH'2008"*, Montpellier, June, 23-25, pp. 109–131.

King G. (1997). *A solution to the ecological inference problem. Reconstructing individual behaviour from aggregate data*. Princeton University Press.

King G., Rosen O., Tanner A. M. (Eds.) (2004). *Ecological Inference. New Methodological Strategies*. Cambridge University Press.

Robinson W. (1950). Ecological correlations and the behaviour of individuals. *American Sociological Review 15*, 351–357.

Tranmer M., Steel D. (1998). Using census data to investigate the causes of the ecological fallacy. *Environment and Planning A 30*, 817–831.