Alternative Approaches to Spatial Rescaling[1]

William D. Nordhaus
Yale University
February 28, 2002
Version 2.2.2

I. Background

It is often necessary to rescale variables between different boundaries, a technique we call "spatial rescaling." An important example, which is investigated here, comes when we are given data with political boundaries and need to rescale them to geophysical boundaries, or sometimes in the other direction. For example, we may be given data on climate by latitude-longitude grids and need to convert this to country or state levels. Another important issue is determining the location of population or economic activity, where we would like to estimate the population or output for grid cells. There are many approaches to doing spatial rescaling, and these are examined in the present study.

We can envisage the issue graphically using Figure 1.[2] The figure shows five irregular regions (call them "states"). We have data for each of the states (population, income, output, and the like) and need to convert it to data for the grid cells. If we consider the central state, four grid cells lie entirely in the state, while parts of 12 others lie partially within the state.

In the field of quantitative geography, the techniques involved in spatial rescaling data are known as "cross-area aggregation"[3] or as "areal

---

[1] The author is grateful for comments from Robert Mendelsohn, Nadja Makarova, and Alexandra Miltner. This research was supported by the National Science Foundation and EPRI. [version is scaling_paper_022802.wpd]

[2] Important note to readers of the cellulose versions of this paper: Many of the figures in this study are best read in the electronic color version. This is available on line at http://www.econ.yale.edu/~nordhaus/homepage/recent_stuff.html .

[3] See A. Stewart Fotheringham, Chris Brunsdon, and Marton Charlton, *Quantitative Geography,* Sage, London, 2000, pp. 59-60.

interpolation."[4] Areal interpolation or cross-area aggregation arises in a number of different contexts, such as when data from census tracts are aggregated into legislative districts.

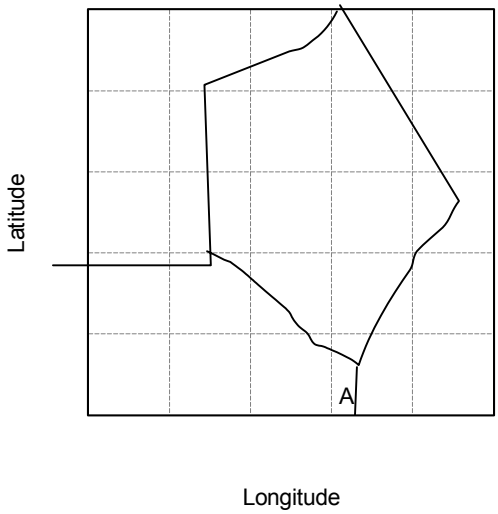


Figure 1. Example of Rescaling Problem

This shows a typical example in which data need to be converted from states (here five irregularly shaped regions) to grid cells. Often the mapping is from points to regions as well.

For purposes of this analysis, we call this problem "spatial rescaling" to indicate that it is generally not aggregation but inferring the distribution of the data in one set of spatial aggregates on the basis of the distribution in another set of spatial aggregates, where neither is a subset of the other. In environmental economics, spatial rescaling is common because environmental data often are scaled or aggregated on a geophysical basis (such as latitude and longitude) while socioeconomic data are generally available aggregated within political boundaries (countries, states, cities). Sometimes, the primary data are collected as individual observations and are aggregated into one or the other scales either to preserve confidentiality or because the original data are samples (this would be the case for virtually all economic data, such as

[4] See Robin Flowerdew and Mick Green, "Statistical Methods for Inference Between Incompatible Zonal Systems," in Michael Goodchild and Sucharita Gopal, eds., *The Accuracy of Spatial Databases,* Taylor and Francis, London, 1989, 239-247. Also see P. F. Fisher and M. Langford, "Modelling the errors in areal interpolation between zonal systems by Monte Carlo simulation," *Environment and Planning A*, vol. 27, 1995, pp. 211-224.

state unemployment rates, output, and price indexes). Often, the geographic scale is diffuse or heterogeneous and makes no sense unless it is aggregated, as in the case of country size, length of coastline, network size, corporate profits, or population density.

The practical application for the rescaling arises in the "G-Econ" project at Yale. The purpose of this project is to develop a global, geophysically scaled economic data set gridded at a 1-degree longitude by 1-degree latitude resolution. This is approximately 100 km by 100 km, which is somewhat smaller than the size of the major subnational political entities for most large countries (e.g., states in the United States, Länder in Germany, or oblasts in Russia). The scaling problem arises in this context because all economic data are collected and presented on the political scale. In general, we are using data at the subnational level (corresponding to states and counties for the United States) and converting these to gridded data.

The purpose of this paper is the practical one of determining the best technique for spatial rescaling along with the advantage of using disaggregated political-boundary data as the primary element for rescaling. We begin with discussing some statistical issues, next turn to a discussion of a simulation strategy, and then present the results of the simulations.

II. Statistical Issues in Spatial Rescaling

A. The setup

In general, country data are averaged over a two-dimensional space. To facilitate discussion and analysis, we treat our "countries" as one-dimensional strips rather than two-dimensional areas. Suppose that we have a continuous data set or a densely spaced data set in one dimension, {x(i)}. The data set is aggregated into N aggregates, which take the value y_1 , ... , y_N. These observed values are for the *source regions* or *source data*. The N aggregates might be countries, states, or counties, in which case the length of the aggregation (or the number of observations if the data are discrete) might be uneven. We wish to rescale the data into M aggregates, z_1 , ... , z_M , with different ranges of aggregation. These unobserved values are for the *target regions* or *target data*. If the z variables are geophysical, then the size of the aggregates might be equal, but this is inessential.

There are few examples of rescaling in the economics literature. One

example is the project to develop gridded data on population.[5] In earlier papers by Robert Mendelsohn, DaiGee Shaw, and the present author, we used regression analysis to rescale data on climate from individual observations into county data.[6]

B. Spatial rescaling under simplified distributions

Before undertaking the simulations, it is worth considering whether the correct form of spatial rescaling can be determined on the basis of assumptions about the underlying process generating the individual observations. For this purpose, assume that the observations are generated by a process such as the following moving-average representation:

(1) $x(i) = a + f(L)\, e(i)$

where a is a constant, $e(i)$ are independent errors with mean zero, and $f(L)$ is a polynomial lag function of the form $f(L) = b_{-k} L^{-k} + \cdots + b_0 + \cdots + b_g L^g$ in which L^i is the ith lag of $e(i)$, where minus signs represent lags, plus represents lags, and b_i is the coefficient on the ith lag.

In certain limiting cases, the optimal form of aggregation is straightforward. The first case is where $f(L) = 0$, or where the observations are independent. In this case, the observation on the state is the best estimator for relevant grid cell, so the optimal spatial rescaling simply takes the area-weighted average for each cell. This technique is known more generally as the "polygon overlay" method, in which variables are interpolated by area-weighted averages. We also call this the "proportional representation" technique below.

Unfortunately, once we move beyond the independent case, extracting the distribution becomes more difficult. Consider the "simple" case where the data-generating process is a simple moving average process, for example where $f(L) = L^{-1} + 1 + L^1$. We can solve this as

---

[5] See http://www.ciesin.org/datasets/gpw/globldem.doc.html .

[6] See particularly Robert Mendelsohn, William Nordhaus, and Dai Gee Shaw, "The Impact of Global Warming on Agriculture: A Ricardian Approach," *American Economic Review*, September 1994, vol. 84, No. 4, pp. 753-771.

(2) $e(i) = L^{-1}(1 - L)(1 - L^3)]\, u(i)$

This is difficult to solve analytically because it involves a large number of leads and lags. In practice, an analytical derivation of (2) is not a useful route to pursue for two reasons. First, the size of political boundaries is uneven, so it would be necessary to develop estimates that accounted for this fact. Second, an examination of the correlation pattern of the underlying income process for U.S. counties reveals that the distribution is of a very high order, that it is poorly determined, and that it varies among different variables (such as wage rates, per capita income, and population density). For example, if counties are sorted by proximity, the log of median income has a significant moving-average representation for up to 12 moving-average terms. The logarithm of county wages has 10 significant moving-average coefficients, and the logarithm of population density has 13 significant terms. Moreover, each of the distributions is quite distinct. Because of the complexity of the spatial correlation of economic relationships, we concluded that a simulation approach would be the most fruitful way of analyzing the relative merit of different spatial rescaling techniques.

III. Data for the simulation

A. Background

Given the difficulties of solving the rescaling problem analytically, we next turn to a simulation study of different techniques. We note at the outset that the simulations are designed for scaling economic data and will not necessarily pertain to all spatial data. However, by using a variety of distributions, some general principles do emerge.

The underlying data examined here are county data for the United States. These data are the most disaggregated data for which a comprehensive set of national accounts can be constructed. The basic approach is to treat the county data as the "true" underlying data, to aggregate the county data up in different ways, and then to test different rescaling techniques by comparing the calculated aggregates with the "true" aggregates. This approach differs from standard Monte Carlo approaches only in that we use actual data rather than generated data as our underlying data.

B. Construction of the Data Set

The data for this experiment are data for 3105 counties of the United

States. This data set was gathered by the author for a study of climate amenities.[7] It contains a number of economic, demographic, and geophysical measures for each of the counties of the United States.

The major difficulty is to construct a simulated county distribution where the counties are "close" to each other. In principle, we would like to line the counties up so that the path through the counties minimizes the cumulative distance of the path. To get the minimal path would require solving what is today an infeasible traveling salesman problem, so we settled for "pretty-good" paths. The preferred data set was one where the counties were chosen so that the distance between adjacent counties was relatively small. As we show below, experiments with different paths found relatively little change in the results.

We assume that the underlying county data are accurate but will generally be unobserved by the analyst. We further assume that the counties are of equal size (so that they correspond to the individual elements in the x(i) series discussed above). We then aggregate the county data into both gridded data and state data by simple averages. We take the ("observed") state data and rescale them to the gridded data using different techniques. We then compare the constructed ("unobserved") gridded data with the "true" gridded data. On the basis of these comparisons, we can then rank different methods. The advantage of this approach is that economic data tend to be skewed and have significant but variable spatial correlation, so we can best evaluate the accuracy of different approaches using actual data rather than by assuming tractable but unrealistic statistical distributions.

C. The variables

For the experiments, we begin with five different sets of variables that are available at the county level and can be aggregated into the state level:

1. Average hourly earnings index
2. Median family income
3. Population density
4. Median family income (alternative locational sort)
5. Population density (alternative locational sort)

---

[7] William Nordhaus, "Climate Amenities and Global Warming," in *Climate Change: Integrating Science, Economics, and Policy*, N. Nakicenovic, W. Nordhaus, R. Richels, and F. Toth, eds., IIASA, CP-96-1, 1996., pp. 3-45.

In addition, we compare the results with other geographic or randomly generated variables:

6. Temperature
7. Latitude
8. Precipitation
9. A normal random variable *(rand)*
10. A 5-period two-sided moving average process generated from a normal random variable *(randma)*
11. A highly skewed random variable *(pareto)*
12. A 5-period two-sided moving average process generated from the highly skewed random variable *(paretoma)*

Most of the experiments were conducted with the logarithm of median family income because this was most closely related to the measure of county output. We tested the major results for the other variables as well. For the simulations, we chose different combinations of grid sizes and state sizes. For example, we might examine grid size of 47 and state size of 41. With this choice, the state boundaries occur at (41, 82, 123, …) and the grid boundaries at (47, 94, …). In general, the sizes were taken to be prime numbers so that no spurious correlations arose from the internal harmonics of the choice of sizes. In other words, the choice was made to reflect the fact (illustrated in Figure 1) that internal political boundaries seldom coincide with the boundaries of grid cells; by choosing prime numbers for the multiplies, we ensure that the boundaries generally only coincide at country boundaries.

D. Rescaling algorithms

The final issue is to devise rescaling algorithms. All techniques share the function that they map the "observed" data aggregated by states into estimates of the data aggregated by grid cells and then compare the "constructed" cell data with the "true" cell data. The literature on spatial rescaling contains two general kinds of techniques. The first, which is the most widely used, is the polygon overlay method mentioned above. We call this the "weighted average" below. The other general type of technique is curve fitting, such as regression. We separate regressions into "local" techniques, which rely only on local attributes, and "global" techniques, which rely upon the entire sample. (By "global" we mean those techniques that use the entire sample; by "local" techniques we mean those which only examine data in the neighborhood of the source and target regions.)

The techniques examined here are the following:

1. *Weighted average or proportional representation (polygon overlap).* This technique apportions the target grid cell among the different states and then sets the value for the grid cell as the weighted average of the values of the different states, where the weights are proportional to the area of each state in the grid cell. This approach is data-intensive and computationally burdensome because it requires estimating the fraction of each grid cell belonging to each state. For example, look back to the state in the center of Figure 1. To calculate the weighted-average approach, it would be necessary to apportion the central state among the 16 grid cells into which it falls.

2. *Median or "plurality rule."* This technique takes as the value of the grid or target region the value for that state or source region which has the largest area in the grid. That is, the grid cell takes on the value of the state, which has the "plurality" of area in the grid. (For example, in Figure 1, the value for the grid cell marked by *A* is determined by the value for the state in which *A* lies since that is the largest subregion of that grid cell.) This approach has the same data requirements as weighted regression and is therefore relatively simple. In practice, it can be accomplished without any statistical analysis as an analyst can simply examine a map to determine the "plurality" state for each grid cell. This approach was apparently taken in the original version of the gridded world population database. In preliminary work done at Yale to prepare the G-Econ economic data for grid cells, we relied upon the median approach, but after investigating different approaches to spatial rescaling, we switched to the weighted-average approach.

3. *Local kernel regression (six alternatives).* This technique uses a local kernel regression for each state. The technique fits a kernel regression, which covers the state plus n counties (or source regions) beyond the boundary of the state. The purpose of extending the regression beyond the boundary is to allow for the possibility of moving-average effects. The tests used 2, 4, 6, 8, 10, and 12 counties beyond the state boundary, and experiments showed that extending the boundary further generally produced a deterioration in the fit. This technique is in principle relatively simple to employ; the major disadvantage of this approach in practice is that it requires using kernel regression techniques, which are somewhat temperamental and not always compatible across different statistical packages. The procedure can easily be automated within a statistical package that has programming capabilities and nonparametric regression features. Since these features are rapidly being included in statistical software packages, this approach is likely to become

increasingly feasible in the future.[8]

6. *Global kernel regression (three alternatives).* This technique uses a global kernel regression for each state. This uses the same approach as 5 but uses the entire sample rather than a local sample. The major issue here is the bandwidth. Intuition suggests a very bandwidth, and we experimented with bandwidths of 2, 10, and 100 observations. Not surprisingly, the global kernel regression performed worse than local techniques for virtually all tests.

7. *Weighted regression.* In an earlier stage of this analysis, we analyzed global weighted regressions. This technique, which is closely related to the kernel regression, uses regressions in which the state observations are weighted inversely to the distance of the center of the state from the center of the grid cell. These regressions are constructed by using weights which are $wt(i) = d(i)^{-\gamma}$, where wt(i) is the weight on a given observation, d(i) is the distance of the observation from the midpoint of the observation to be predicted, and $\gamma \geq 0$ is an exponent on the distance. Note that where $\gamma = 0$, the model reduces to the crudest estimate, which is simply the country average. A fair amount of experimentation in the early stages went into looking at alternative weights for the weighted regressions. For most variables, the root mean squared error was minimized with a distance parameter of $\gamma = 1$. Weighted regression has the advantage of being relatively simple; it requires primarily geophysical data for both grids (which is trivial) and states (which is not always readily and accurately available). This approach was used in the Mendelsohn, Nordhaus, and Shaw studies cited above. We decided not to analyze this technique for this round of experiments to reduce the number of experiments, because it had proven unfruitful in earlier rounds, and because it is logically a subset of kernel regressions. We will show some of the earlier results in a later section.

8. *Country average.* This approach uses the mean of the entire sample and applies it to each grid cell. In other words, the mean of the variable for the entire "country" is applied to each grid cell. This approach has the advantage of being extremely simple, as it requires no subnational data.

---

[8] The kernel regressions were performed in EViews 4.0. Analysts should be warned that it is necessary to tune the parameters very carefully. For these results, we used local linear regression, with "exact" fitting, with the number of grid points equal to the total sample size, and with a bandwidth equal to 2 to 12 observations, using an Epanechnikov kernel. This is a much finer (and slower) setting than is usually set as a default. Using the default settings will produce very smooth results and poor fits and defeat the purpose of the kernel regressions.

9. *Pycnophylactic smoothing.* This approach, devised by Waldo Tobler, is widely used to create maps from areal data. It imposes Laplacian smoothness on the surface by minimizing the curvature; that is, it requires that the sum of the squared second differences of the variable in both directions be minimized.[9] We have up to now been unable to implement the full pycnophylactic technique in our statistical study. Instead, we undertook a pilot test using a pycnophylactic variation of the most accurate technique (weighted average). This experiment was performed by taking the simulation for the weighted-average technique and applying pycnophylactic smoothing. The smoothing was achieved by minimizing the squared second difference between states with a Newton method with quadratic central search using the "Solver" program.

IV. Results

Box 1 provides a summary of the major results, and we now proceed with a discussion of each of the points. It should be emphasized that all these results are conditional on the experiments and data actually examined.

[9] W.A. Tobler, "Smooth Pycnophylactic Interpolation for Geographical Regions," *Journal of the American Statistical Association*, 1979, pp. 519-529.

1. The weighted average technique (polygon overlap) is the most robust technique and generally gives the most accurate estimate of the true state values.
2. Local techniques are generally more accurate and more robust than global techniques.
3. There is generally little difference among local techniques (except for the "median" estimates and pycnophylactic interpolation), and the choice should be made on the basis of ease of implementation.
4. Global techniques can be extremely inaccurate because they do not respect local variation in economic typography.
5. Very large improvements in accuracy can be obtained by disaggregating the source data; that is, by reducing the size of the source regions, the accuracy of the spatial rescaling for the target regions is increased significantly.
6. Large increases in accuracy are also gained by considering larger target regions.
7. Pycnophylactic smoothing was a disappointment. It tended to smooth the results far too much.
8. The rankings of the different algorithms are similar across a wide range of economic and statistical series.

Box 1. **Summary of Results of Simulations**

For reference purposes, the 11 techniques that were scored are shown in Box 2.

avg = weighted average
med = median of counties
ker2 = Local kernel estimate overlapping 2 counties into adjacent states
ker4 = Local kernel estimate overlapping 2 counties into adjacent states
ker6= Local kernel estimate overlapping 2 counties into adjacent states
ker8= Local kernel estimate overlapping 2 counties into adjacent states
ker10= Local kernel estimate overlapping 2 counties into adjacent states
ker12= Local kernel estimate overlapping 2 counties into adjacent states
kertot2 = Global regression with a window size of 2 counties
kertot10 = Global regression with a window size of 10 counties
kertot100 = Global regression with a window size of 100 counties

Box 2. **Algorithms Examined in This Study**

1. <u>The weighted average technique (polygon overlap) is the most robust technique and generally gives the most accurate estimate of the true state values.</u>

One surprise is that "simple is best." Of all the approaches from trivial through fancy to esoteric, the simplest and most intuitive is to take a weighted average of the variable where the weights are the areas of the source region (state). This produced not only the lowest error but also the most reliable algorithm across a variety of distributions.

Figure 2 shows a summary of the results of different techniques. The bars show the average root mean squared error (RMSE) of different rescaling algorithms. The results apply to two different economic variables (the log of median income and the log of wages) for seven different combinations of grid and state sizes. The size of the bars measures the average RMSE for that algorithm relative to the minimal RMSE across all algorithms. The results are clear: the average technique has the lowest average error. Moreover, it is robust in that it has at most a 2 percent penalty relative to the most efficient algorithm.

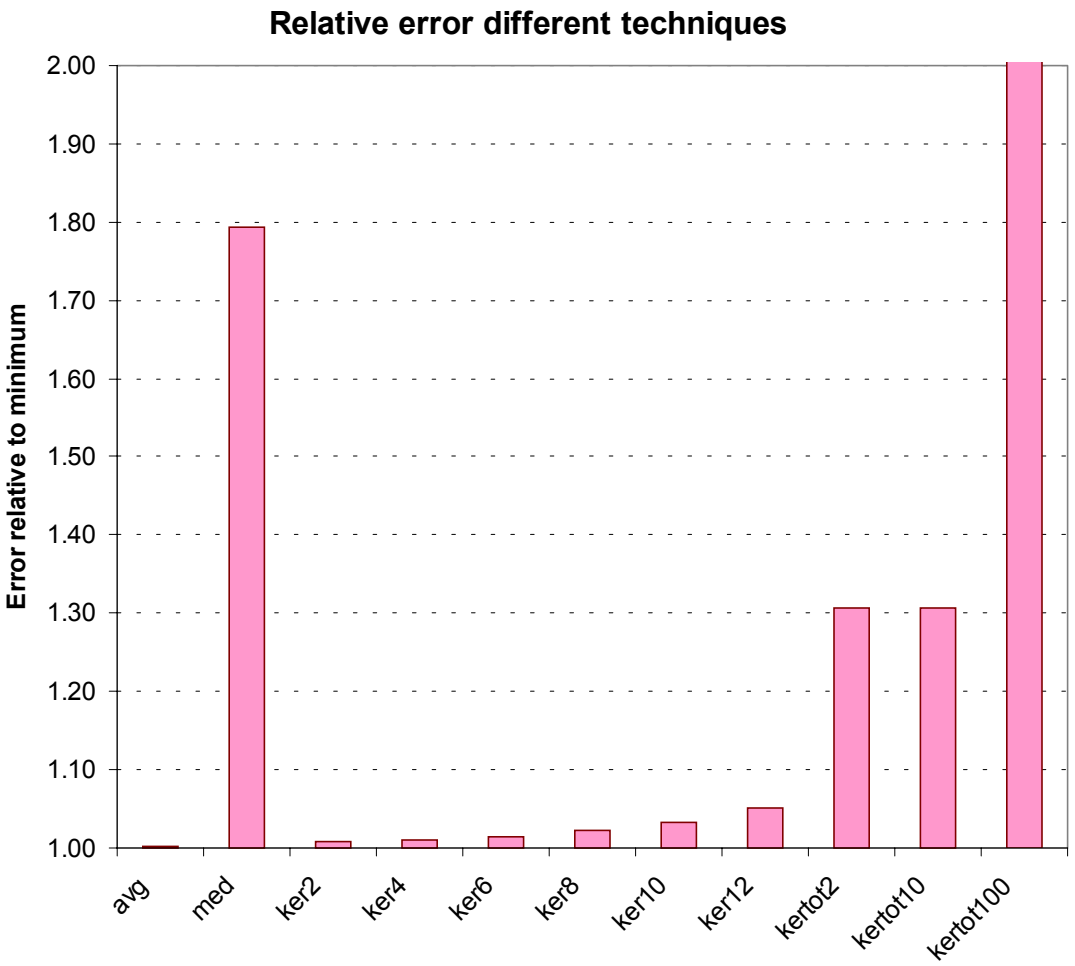


Figure 2. **Relative error for different rescaling algorithms.** This graph shows the root mean squared error for 11 different algorithms defined in Box 2 and the text. These show the average RMSE for each algorithm relative to the minimum RMSE of all algorithms, averaged across seven different grid combinations and two different variables.

Figure 3 shows the range of errors (i.e., the ratio of maximum to minimum RMSE) across the different algorithms for the 14 different experiments, where these are ranked from left to right by the range for "local" algorithms. The first bar shows the range for the local algorithms, the second bar shows the range for all algorithms, and the (virtually invisible) third bar shows the ratio of the RMSE for the average technique to the RMSE of the most efficient technique. The chart makes it clear that there can be a substantial penalty from choice of an inefficient algorithm. In the worst case, an inefficient algorithm can increase the RMSE by a factor of almost six.

A further point can be seen by examining the labels at the bottom of Figure 3. These show (size of grid, size of state). For example, the first set of bars represents a simulation where each grid contains 13 counties while each state contains 79 counties. The range is generally greatest when grid size is large relative to state size.

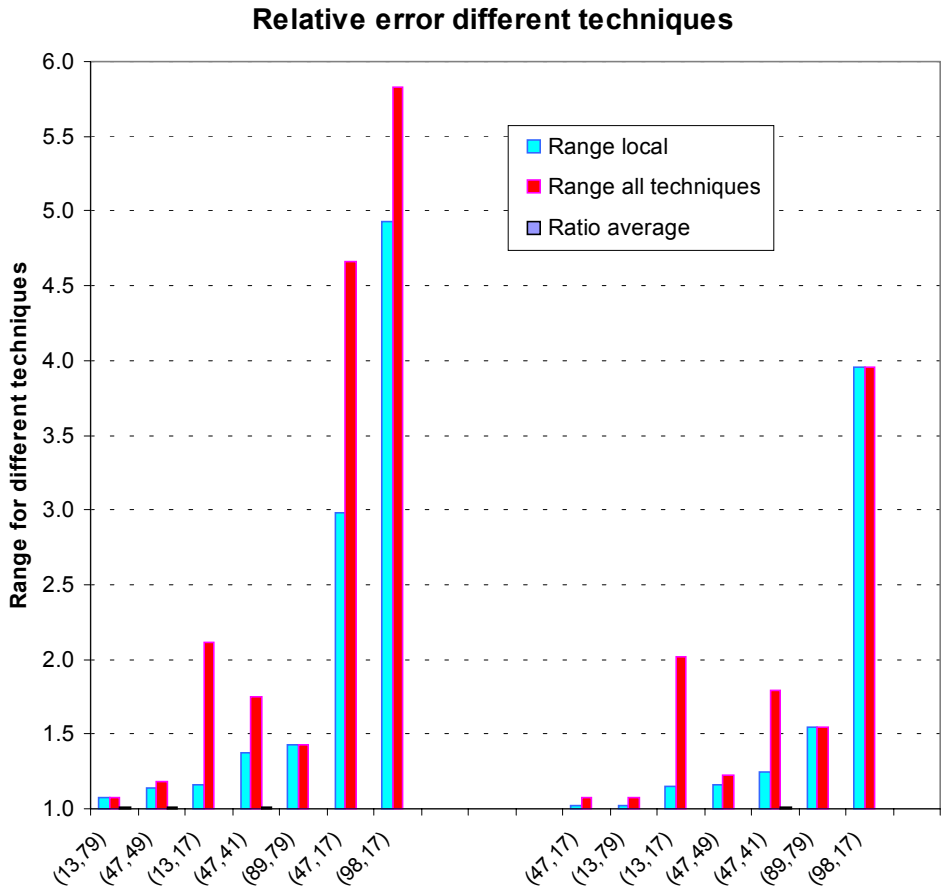


Figure 3. Range of errors for different algorithms. The figure shows the range of RMSE for the different algorithms for the grid and state size shown at the bottom. The numbers at the bottom of the figure are (size of grid, size of state), where the "size" is measured as the number of counties comprising the grid or state.

---

2. Local techniques are generally more accurate and more robust than global techniques

One of the key results of the simulations is that "local" techniques

generally dominate "global" techniques. We designate algorithms as local when they examine only the local information and discard information that is far from the local topography. This result is somewhat paradoxical because it is usually presumed that more information is better than less.

The results for different algorithms were shown in Figure 2. The first 8 bars are local algorithms. With the exception of the median, they all perform better than the global algorithms shown in the last 3 bars. Other global algorithms, such as ones that smooth over the entire country, generally do even worse, as we discuss below.

3. There is generally little difference among local techniques (except for the "median" estimates and pycnophylactic interpolation), and the choice should be made on the basis of ease of implementation

Looking more closely at the local algorithms, we see that there is little difference in their performance across different simulations, the exceptions being the median and the pycnophylactic algorithms (the latter is discussed below). The median and pycnophylactic algorithms aside, the local algorithms have a range of RMSE between best and worst of less than 10 percent. While there is not much to choose among them in terms of performance, as we discuss below the ease of implementation may be quite different using existing software packages. Therefore, we conclude that (aside from the median and pycnophylactic) the choice among local estimators should be made largely on the ease of the implementation.

4. Global techniques can be extremely inaccurate because they do not respect local variation in economic topography

Global techniques often perform very poorly. For these experiments, the global techniques were limited to different kernel regression models. In an earlier set of experiments, we tested a variety of weighted regression approaches and found that these were generally very inaccurate relative to local techniques. Figure 4 shows a comparison of three global weighted regressions with three local techniques. The global regressions have much higher average error.

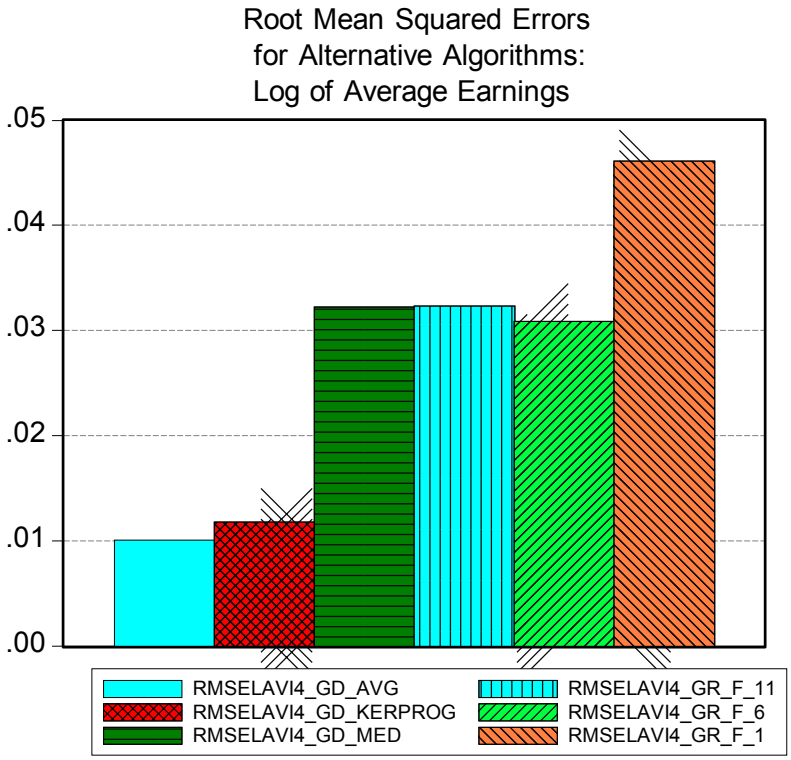


Figure 4. Error of Different Global Algorithms.

This simulation compares the RMSE for two local regressions (the first two bars) with three global-regression algorithms for the log of average hourly earnings. The global regressions weight the observations by the inverse distance to the power shown divided by 10. Hence, RMSELAVI4_GR_F_1 represents a weighted regression of the log of median income where observations are weighted by the distance from the prediction to the power 0.1 (see the discussion above).

---

5. Large improvements in accuracy can be obtained by disaggregating the source data

Two important decisions that are under the control of researchers involve the size of the source region (here the "states") and the size of the target region (here the "cells"). How far down should the source economic data be disaggregated? Into how fine a grid should the gridded or target region be subdivided?

Often the answers to these questions will be dictated by the availability

of the data. For example, in our G-econ project at Yale, we began with the difficulty that most countries prepare output data only at the national level. For large countries like the United States, Russia, or Brazil, such large geographical aggregates for the source data set might lead to very large errors in grid cell data. Similarly, in choosing a size of target grid cell, we were torn between the desire for higher resolution and the recognition that the errors would be larger at smaller resolutions.

Our simulations cast some light into the tradeoffs here. Starting with the first question, Figure 5 shows the gains from disaggregating the source regions with different target grid cells. The elasticity of the RMSE with respect to state size is 0.52 ($\pm$.10), so a decrease in source size by 10 percent would lower the error rate by 5 percent. There are therefore clear gains from disaggregating the source data where the errors from disaggregation are relatively small.

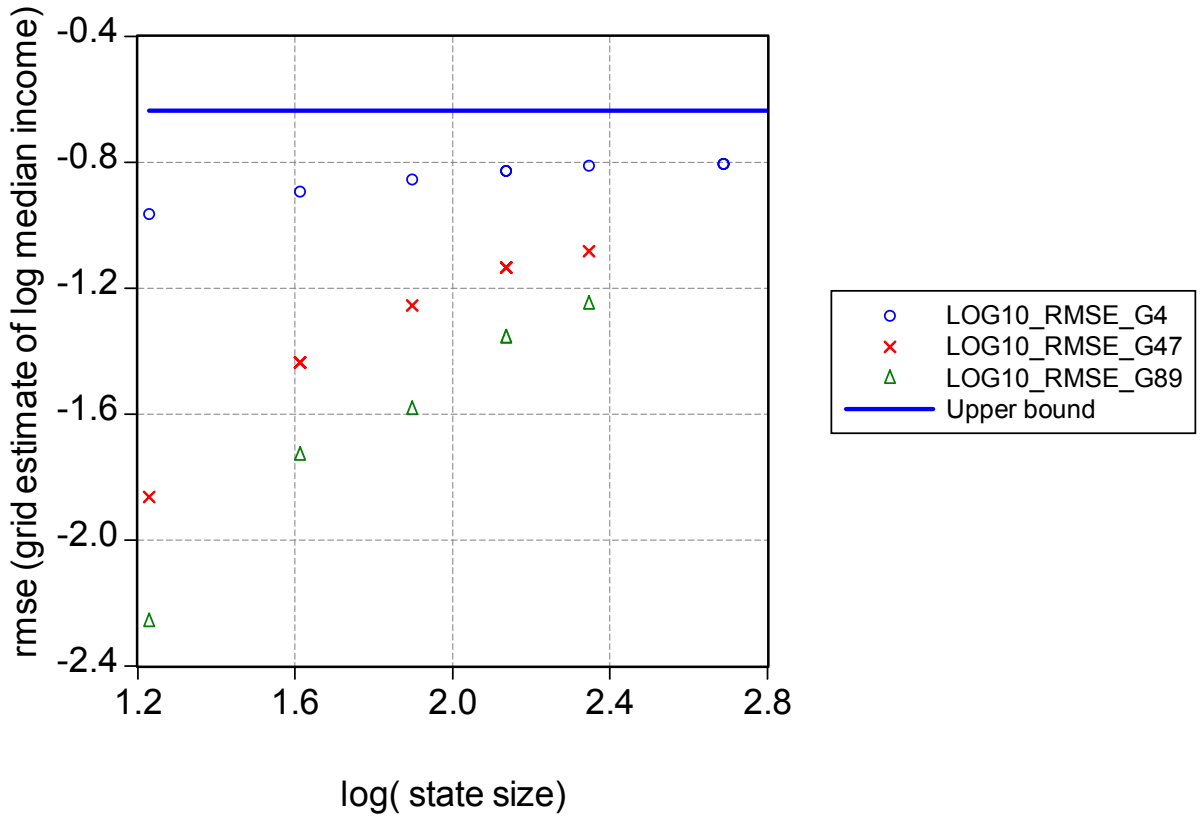


Figure 5. Impact of state size on accuracy. This figure shows the relation between $\log_{10}$ of state (source-region) size and $\log_{10}$ of RMSE for three different grid cell (target region) sizes. For example, in changing the source region from U.S. states to U.S. counties, the size of the source declines by a factor of sixty (or a $\log_{10}$ change of 1.8). This produces a substantial gain in accuracy.

---

6. Large increases in accuracy are gained by considering larger target regions.

The second tradeoff is shown in Figure 6, which shows the impact of grid cell size on RMSE. The estimated elasticity of error with respect to grid cell size is -0.72 (±.059). These results indicate that disaggregating can raise the error rate markedly. For example, many current estimates use a 1-minute by 1-minute resolution of population. Using these results suggests that the error from moving to such a fine resolution would be approximately 130 percent higher than using a 1-degree by 1-degree grid aggregation.

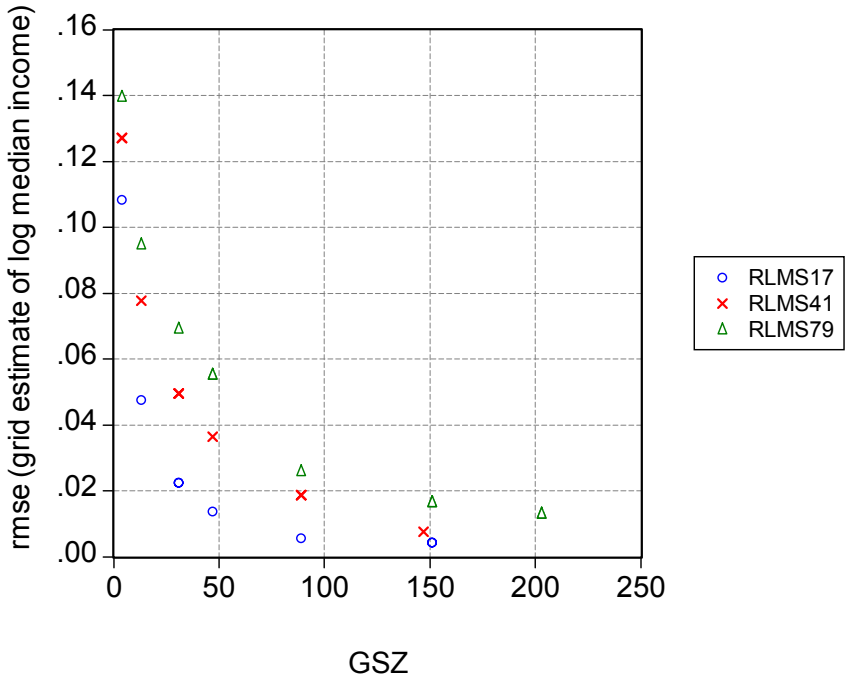


Figure 6. **Impact of grid cell size on error.**
This figure shows the same relationship as Figure 5, where the grid (target-region) size is changed and the state sizes are held constant for the different series.

7. Pycnophylactic smoothing produces very poor results because it smoothes the results far too much

Pycnophylactic smoothing is widely used and often praised. The criterion for selecting the target data (here, the grid cells) is that it minimizes the sum squared second-difference in the variable from one cell to the next. Figure 7 shows the results of pycnophylactic smoothing for the case of log median income for the first 470 observations. The heavy block line is the observed state data, the light block lines are true state and average estimates, and the smooth curve is the pycnophylactic estimator. It smoothes nicely but tends to miss fine gradations in the underlying data. A good example is the grid cell between lying approximately between 190 and 235. The average technique gets the exactly correct answer, while the pycnophylactic technique is shooting up to smooth the data. The average error was 76 percent larger for the pycnophylactic smoother as compared with the average technique. Given the dismal performance, we choose not to pursue this approach further. Pycnophylactic interpolation produces pretty pictures but the jagged estimates produced by the local algorithms yield more accurate estimates of the underlying distributions.

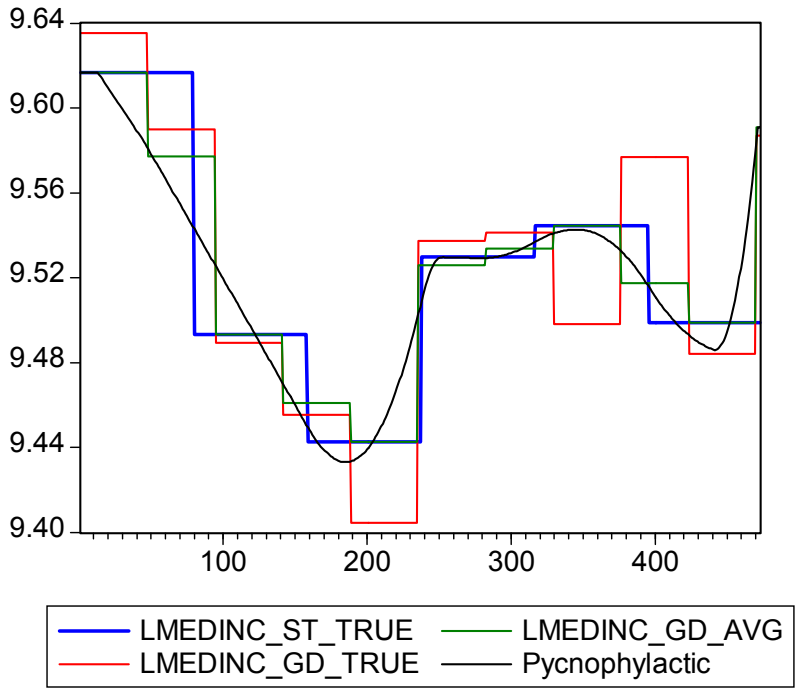


Figure 7. Comparison of Average and Pycnophylactic-Smoothing Algorithms

8. The rankings of the different algorithms are similar across a wide range of economic and statistical series

The results for distributions of several other variables were tested. For this purpose, we display only one set of simulations, that with the number of counties per grid equal to 47 and the number of counties per state equal to 49. This choice was taken to represent a case where the sizes of grid cells and counties are approximately the same, which holds for the U.S., and to minimize the sampling error.

The results are shown in the Appendix, with the key to the different series given at the end of the Appendix. Each bar shows the ratio of the RMSE for the algorithm listed in the key. The results are clear: local techniques are always better than global algorithms; the local kernel-regression algorithms are virtually identical to the AVG algorithm. We have not shown the MED algorithm, but it was uniformly worse than the AVG, and often worse than the global algorithms.

V. Implementation Issues

The ranking of the different techniques is clear. The next question concerns the ease of implementing each of the techniques. We can group the difficulty into three categories: very easy, moderately difficult, and most difficult.

Very easy

Country-average (unweighted) regression techniques are very easy to implement. This technique requires only country data and no data on individual states (source region). This is the approach that has been taken in rescaling in most existing studies using economic data. The RMSE for country averages is, however, very large. For example, if there are 50 states and 3000 counties (which is the case for the U.S.), then, for median income, using national data has a RMSE of 0.24, using state data has an RMSE of 0.13, and using county data has an RMSE of 0.08. Using country data is generally not recommended unless subnational data are simply unavailable or there is a low premium on accuracy.

Moderately difficult

Weighted regression, kernel regression, and "plurality rule" are in the "moderately difficult" category. These techniques require economic (source) data for each of the states, but do not require mapping the targets or sources (grids or states) in detail. We see, however, that these techniques have significantly higher error that the best techniques. Looking at Figures 2 and 3, we see that the error of the median techniques and efficient weighted regressions have errors between 80 and 300 percent larger than the most accurate algorithms.

Most difficult

The most difficult techniques are ones that require both regional data and mapping out each state and grid cell in detail. Take Figure 1 above as an example. These techniques require obtaining detailed data for the variable of interest for each state and also determining exactly how much of each state lies in each grid cell. The first of these is often very time intensive; for example, in the G-Econ project, it requires developing regional output accounts for many countries. For the second, detailed mapping is in principle straightforward with computerized mapping programs, but it appears that no programs currently exist that can perform this task for all countries. There are, however, major gains from moving to the most difficult algorithms.

Figure 8 shows a comparison of very easy (country average) algorithms, more difficult (global regression) algorithms, and most difficult (efficient local) algorithms. For this example, country approach has a RMSE about 120 percent higher than the efficient technique; global techniques have RMSE from 10 to 50 percent higher than the most efficient technique; and the four local algorithms are essentially equivalent. Clearly, there are substantial gains in accuracy from a careful scaling at the subnational level.

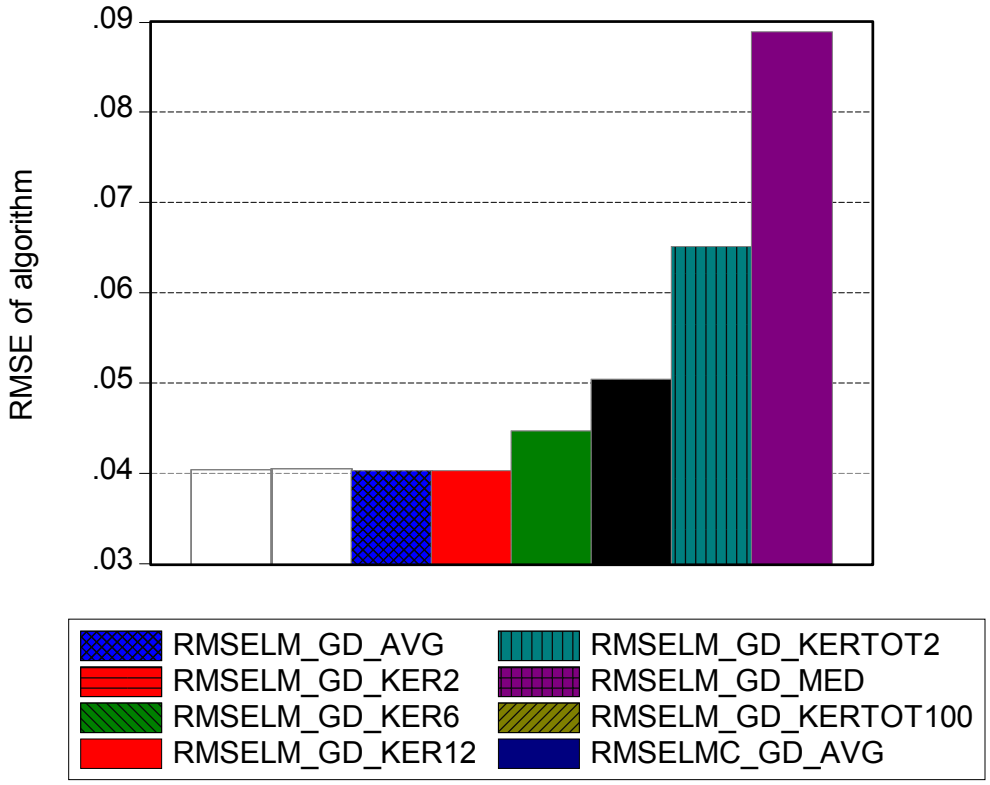


Figure 8. RMSE for different algorithms and for country average

This figure compares the relative error of efficient local algorithms on the left with the error from using a country average, shown on the last bar on the right. This example, which is for the United States, probably understates the differences in many countries with larger regional inequality.

It is useful to put these results in international perspective. The data used here are for the United States. Regional differences in incomes in the United States are small relative to many other large countries. Therefore the gains from using subnational data are likely to be even larger in other countries.

V. Summary

Spatial rescaling is an area of growing importance in empirical studies of countries and regions with the advent of spatially based data, such as that produced by satellites. At the same time, researchers who cross the disciplinary boundaries between physical and natural sciences increasingly need to rescale variables produced with one scaling system into another scaling system. The present study investigates the properties of alternative rescaling algorithms using both artificial data and county data for the United States. It finds that the error from not using appropriate techniques is

substantial. Moreover, it finds that the errors in different rescaling techniques will differ depending upon both the techniques used and the extent of regional structure in the variable under investigation.

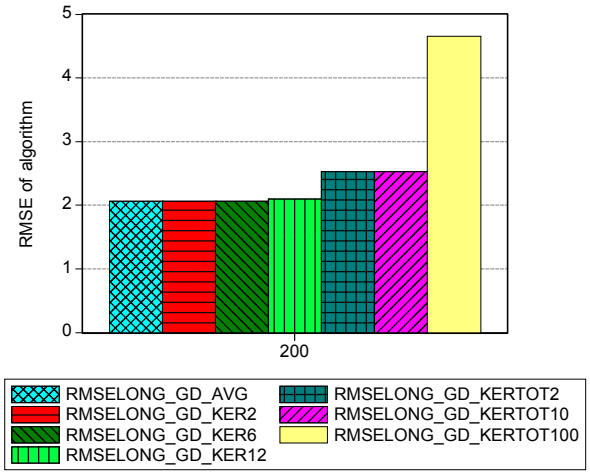
Techniques using country averages (which are "very easy" to implement) are likely to perform poorly, especially where there is systematic variation. Techniques that are "moderately difficult," needing only average data for the individual states, improve the results by a large factor, especially where there is systematic variation in the data that can be exploited. Techniques that are "most difficult" to apply require detailed geophysical data for each state; they improve the performance markedly for the different variables that we investigated.

What are the final recommendations that emerge from this study? If researchers have either lots of research time or the ability to write specialized programs that can capture the detailed contours of the political boundaries and gridded regions, then the "most difficult techniques" are definitely recommended. In particular, the "weighted-average" technique (a.k.a. polynomial interpolation in the spatial statistics literature) stands out as the most robust and most accurate. Local kernel regressions with very small bandwidths perform almost as well. The choice between these two techniques will depend upon which is easier to implement.

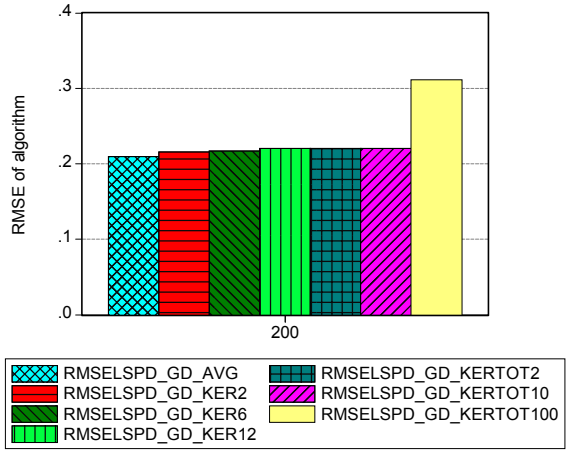
The other finding is that using national averages of data is likely to provide substantial errors for most data series when they are rescaled to relatively fine grid cells. Put differently, there are major gains to disaggregating the source data below the national levels and to the lowest level at which reasonable accuracy in the regional data can be obtained.

Appendix. **Results for Different Variables**

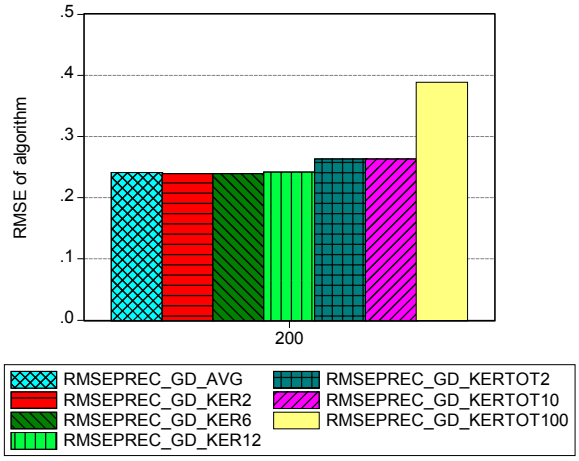
Longitude



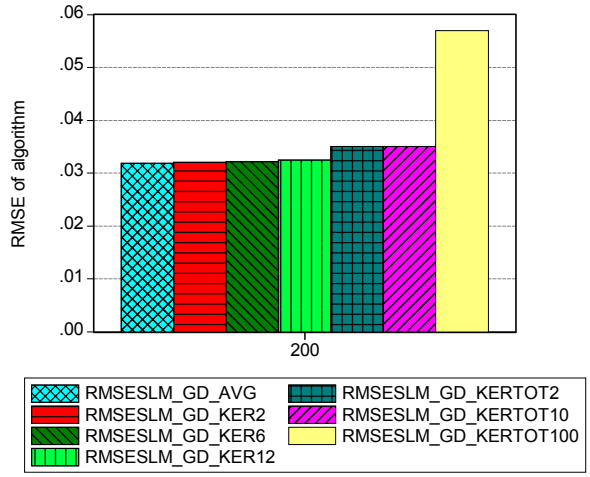
Alternative distribution of population density



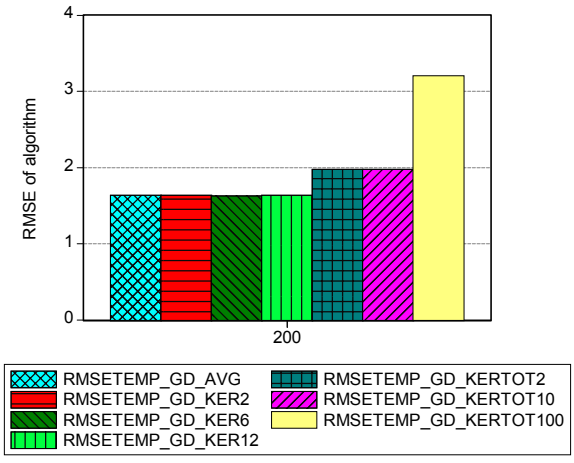
Precipitation



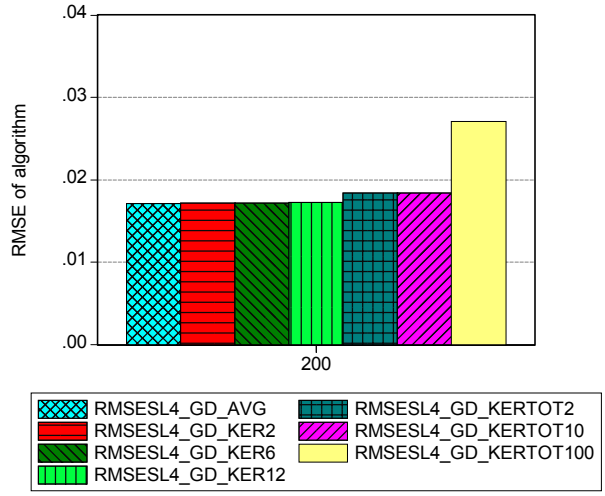
Alternative distribution of median income



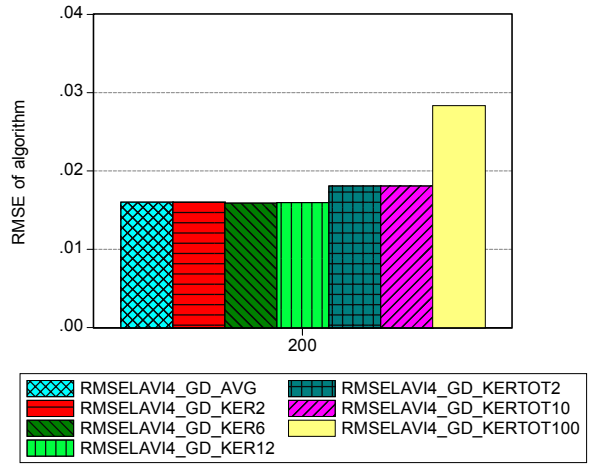
Temperature



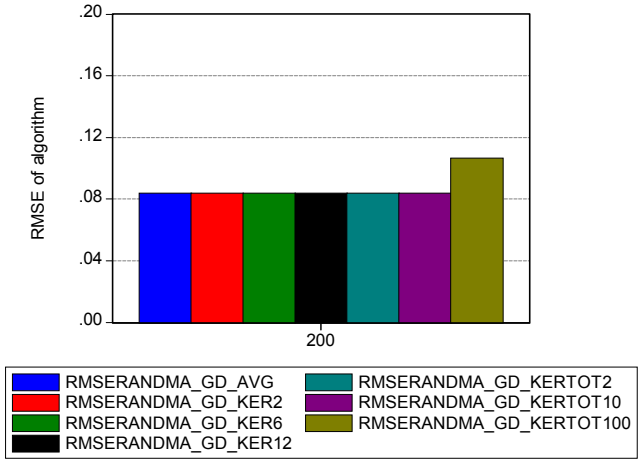
Alternative distribution of wage rate



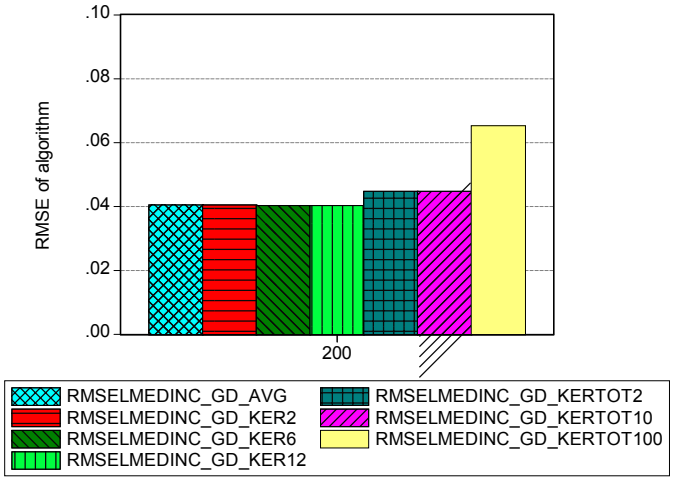
Wage rate



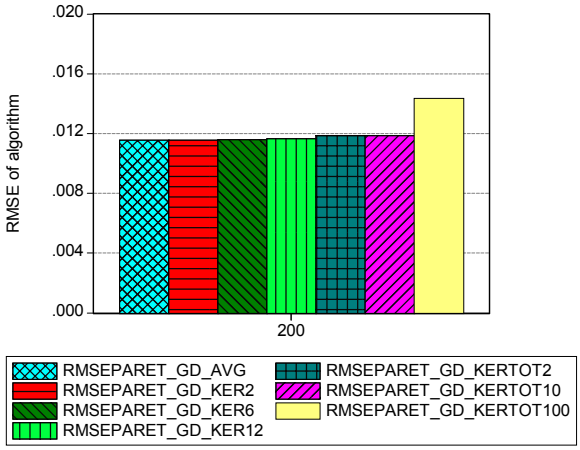
Random normal variable, moving average



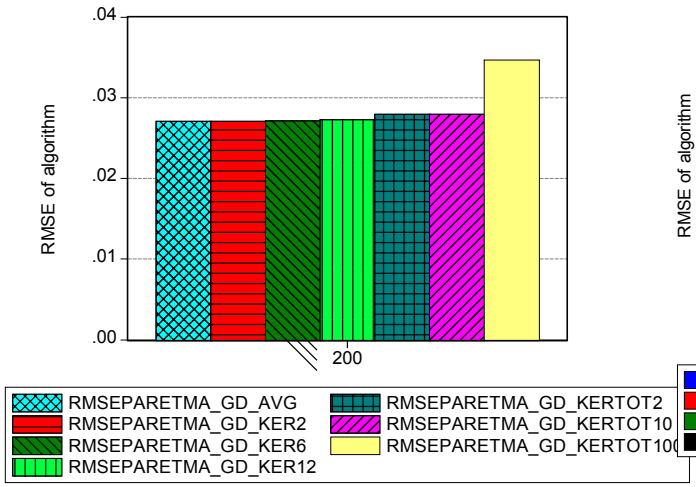
Median income



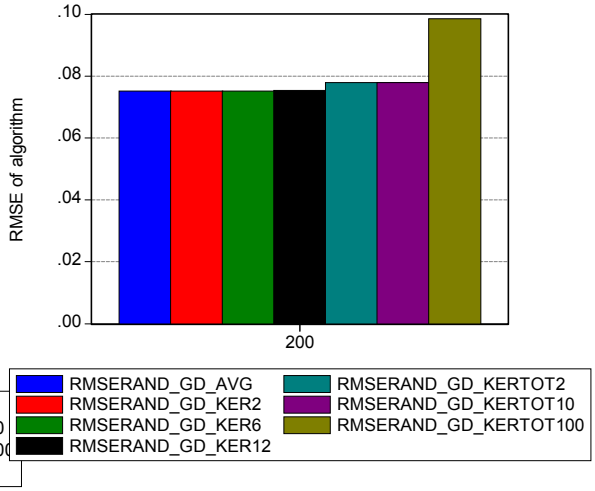
Pareto variable



Pareto moving average



Random normal variable



Key to variables in appendix:

Variables have the following structure:

"RMSE[variable name]_GD_[algorithm name]"

The algorithm name is given in Box 2. "GD" means that it is scaled to the grid cell. The key to variable names is below:

1. Average hourly earnings index *(lavi4)*
2. Median family income *(lmedinc)*
3. Population density *(lpd)*
4. Median family income (alternative locational sort) *(slm)*
5. Population density (alternative locational sort) *(sl4)*
6. Temperature *(temp)*
7. Longitude *(long)*
8. Precipitation *(prec)*
9. A normal random variable *(rand)*
10. A 5-period two-sided moving average process generated from a normal random variable *(randma)*
11. A highly skewed random variable *(pareto)*
12. A 5-period two-sided moving average process generated from the highly skewed random variable *(paretoma)*