12-17-2017

Bayesian Downscaling Methods for Aggregated Count Data

Clayton P. Michaud

Thomas W. Sproul
University of Rhode Island, sproul@uri.edu

Citation/Publisher Attribution

Michaud, C.P., Sproul, T.W. Bayesian Downscaling Methods for Aggregated Count Data (2018) *Agricultural and Resource Economics Review*, 47(1), pp. 178-194. DOI: 10.1017/age.2017.26

Bayesian Downscaling Methods for Aggregated Count Data

Bayesian Downscaling Methods for Aggregated Count Data

Clayton P. Michaud and Thomas W. Sproul

Policy-critical, micro-level statistical data are often unavailable at the desired level of disaggregation. We present a Bayesian methodology for "downscaling" aggregated count data to the micro level, using an outside statistical sample. Our procedure combines numerical simulation with exact calculation of combinatorial probabilities. We motivate our approach with an application estimating the number of farms in a region, using count totals at higher levels of aggregation. In a simulation analysis over varying population sizes, we demonstrate both robustness to sampling variability and outperformance relative to maximum likelihood. Spatial considerations, implementation of "informative" priors, non-spatial classification problems, and best practices are discussed.

Key Words: aggregated data, agricultural census, Bayesian methods, count data, disaggregation, downscaling, farm counts, posterior distribution

JEL Codes: C43, C11, Q12

Local economic planning often relies on micro-level data that are not always available at the desired level of disaggregation. For example, Federal government-provided economic and employment data for key industry sectors are often reported at the county level, and obtaining city or ZIP-code level data may require time-consuming special requests or considerable expense, or the data may simply be unavailable. In this article, we address the need for micro-level count data by developing a Bayesian methodology to "downscale" aggregated count data to lower levels of aggregation using the information contained in an outside statistical sample.

Suppose a researcher knows the true size of a population (e.g., farmers, voters, customers) and would like to classify members of that population into

Clayton P. Michaud is a graduate student, Department of Environmental and Natural Resource Economics, University of Rhode Island. Thomas W. Sproul is an Associate Professor, Department of Environmental and Natural Resource Economics, University of Rhode Island. *Correspondence: Clayton P. Michaud ▪ Environmental and Natural Resource Economics ▪ University of Rhode Island, Kingston, RI 02881, USA ▪ Phone: 774.219.9198 ▪ Email: claytonmichaud@my.uri.edu*

Dr. Sproul acknowledges financial support from USDA National Institute of Food and Agriculture Hatch Projects: No. RI00H-108, Accession No. 229284, and No. RI0017-NC1177, Accession No. 1011736, and USDA Economic Research Service Cooperative Research Agreement No. 58-6000-5-0091

The views expressed are the authors' and do not necessarily represent the policies or views of any sponsoring agencies.

Agricultural and Resource Economics Review 47/1 (April 2018) 178–194

distinct subgroups (e.g., by farm type, county/region, political party, or demographic attributes) using independent data sampled from the full population. In this setting, we demonstrate a method for estimating the population proportion in each subgroup, in a manner that provides more stable and robust estimates than maximum likelihood estimation (MLE) in the face of sampling variability. The method consists primarily of using simulated random sampling combined with exact calculation of combinatorial probabilities to estimate the posterior distribution over combinations of counts. We leverage two key restrictions: (i) the subgroup counts must add up to the population total, and (ii) the subgroup counts cannot be smaller than their observed counts in the outside sample, nor larger than the population minus the sum of observed samples in the other subgroups. This explicit handling of sampling variability, especially in small- to medium-sized samples, results in smaller normalized errors and, consequently, more reliably accurate estimates.

We are not the first to address the demand for more disaggregated data from aggregated sources. Gocht and Roder (2011), for example, employ a Bayesian procedure to downscale county-level German Agricultural Census estimates of land devoted to agricultural use. Their method incorporates land-use data from GIS to facilitate micro-level environmental impact studies, which would otherwise be hindered by data protection rules (i.e., censoring). Other relevant studies include Chakir (2009), Dendoncker, Bogaert, and Rounsevell (2006), Gärtner, Keller, and Schulin (2013), Howitt and Reynaud (2003), Polasek, Llano, and Sellner (2010), and Purcell and Kish (1980). These papers share a common thread of attempting to estimate land-use patterns using a variety and/or combination of methods, including regression, multinomial logit, maximum entropy, cross-entropy, and various iterative fitting procedures. However, while these procedures perform well in their intended domain, they are ill-suited to solving the downscaling problem for count data. Intuitively, multinomial logit might be mapped to a count model in which sampling probabilities are estimated but many observations and covariates are required. The methods we introduce here are designed to overcome this problem when the outside sample contains only limited categorical information.

Another popular application of downscaling involves disaggregation of global climate data (typically reported at grid levels of 100–200 km^2) to a level of resolution more useful for decision makers and impact assessors. Such procedures are outlined, for example, in Coelho et al. (2006), Fasbender and Ouarda (2010), Hashmi, Shamseldin, and Melville (2009), Murphy (1999), and von Storch, Zorita, and Cubasch (1993). The goals of such estimation procedures, however, are to disaggregate weather/climate data not only spatially, but temporally as well, in order to model various potential weather outcomes for use in forecasting. The procedures outlined by these studies are both unnecessarily complex, given our particular problem of interest, and

potentially ill-suited to the count data problem due to highly detailed data requirements in the outside sample.

In an attempt to balance precision with tractability, we develop a method that is adaptable to the data and computational resources of the applied researcher. Namely, we show that reasonable performance can be obtained using a uniform prior distribution over combinations of counts, but we also demonstrate a method for researchers to incorporate "informative" prior information generated by a simple linear regression or one of the more spatially explicit and computationally demanding methods described above. In our simulation analysis, we demonstrate a means for testing the best performance among MLE, the uniform prior, or an informative spatial prior, over a range of population counts and sample sizes. As might be expected, an informative prior performs best for the smallest sample sizes and smallest population counts. However, we find that the uninformative, uniform prior performs best over an unexpectedly wide range of sample size and population count combinations.

To provide context, we introduce and apply our methods in the setting of estimating spatially disaggregated farm counts by subregion from regional data, using a sample of Rhode Island farms combined with aggregated data from the 2012 United States Department of Agriculture (USDA) Census of Agriculture (herein, "Ag Census"). We explore both county-to-city downscaling and state-to-county downscaling, and show how spatial patterns at higher levels of aggregation might be used to construct an informative prior. We take special advantage of state-to-county downscaling as an example where the true underlying distribution is known and can be used to validate our methods. We also use published estimates of uncertainty in the Ag Census total counts to demonstrate the robustness of our methods to uncertainty in the top-level population count.

Despite the focus of much of the literature, and our own application, on spatial downscaling problems, it is important to note that there is nothing inherently spatial about the mathematics involved. Our method is equally well adapted to arbitrary classification problems in which it is desired to estimate the size of population subgroups according to a number of discrete categories. These applications might include political polling, estimation of workforce participation rates, demographic breakdowns by gender, age, race, or educational attainment, or market segmentation analysis. At the same time, though our method does not require spatial information per se, it is flexible enough to incorporate arbitrarily complex spatial information as an input to the estimation procedure, by way of the informative prior.

The remainder of this article is organized as follows. The next section outlines and derives our estimation methodology, and the following section discusses selection of a prior. The fourth and fifth sections outline our sample data and methods, and the sixth section covers the results. The next section discusses applications of our findings and areas for future research, and the final section concludes.

Bayesian Downscaling of Aggregated Count Data

Consider a source of aggregated count data for which estimated count data are required at the subpopulation (e.g., subregion, demographic classifications) level, for each subpopulation, $s = 1, \ldots, S$. Let N denote counts at the aggregate level, i.e., population size. We denote the counts to be estimated at the subpopulation level as N_s, such that the sum of subpopulations counts is equal to the total population count, $\sum N_s = N$. We supplement this population level data with an outside, independently sampled data set with subpopulation counts, n_s, where $\sum n_s = n < N$. That is, the outside sample of subregion data is a subset of the population to be estimated. The immediate impact of the outside data set is to constrain the range of eligible values, which we will denote as N'_s within each combination. Namely,

$$n_s \leq N'_s \leq N - \sum\nolimits_{s' \neq s} n_{s'}. \tag{1}$$

Thus, we define $\boldsymbol{C}$ to be the set of all valid combinations of integer-valued counts satisfying equation 1. The cardinality of this set is denoted $|\boldsymbol{C}|$ and is given by:

$$|\boldsymbol{C}| = \binom{N - n + S - 1}{S - 1} = \frac{(N - n + S - 1)!}{(S - 1)!(N - n)!}. \tag{2}$$

For example, consider Bristol County, Rhode Island. Bristol County comprises three municipalities and is reported by the Ag Census as containing 42 farms. Our sample counts for these three subregions are (6, 3, 2). A valid combination would therefore be any triple with each value equal to or exceeding the sample count and with the total count equal to 42. Thus, (2, 21, 19) is not a valid combination because there are not enough farms in the first town, and (25, 12, 7) is not valid because there are too many total farms (44), but both (6, 34, 2) and (15, 15, 12) are valid combinations.

We have now developed sufficient notation to outline our estimation procedure. First, recall Bayes' Rule:

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)} \propto \Pr(B|A)\Pr(A). \tag{3}$$

For our purposes, $\Pr(A|B)$ in equation 3 represents the probability of a specific combination of subpopulation counts given the data, or $\Pr(C_i|N, n_1, \ldots, n_s)$, and the other terms translate similarly. That is:

$$\begin{aligned}\Pr(C_i|N, \boldsymbol{n}) &\propto \Pr(N, \boldsymbol{n}|C_i)\Pr(C_i|C^0) \\ &= \Pr(\boldsymbol{n}|C_i)\Pr(C_i|C^0) \\ &\propto \Pr(\boldsymbol{n}|C_i)\end{aligned} \tag{4}$$

where (i) $\boldsymbol{n} = (n_1, \ldots, n_S)$ denotes the vector of subpopulation counts in the outside data, (ii) the equality follows from the constraint in equation 1 because only combinations that sum to N are considered, and (iii) the final proportionality comparison relies on the assumption that the unconditional probability of a combination is uniform across combinations, representing the prior in our Bayesian approach. This is the simplest case of a uniform ("uninformative") prior over combinations, which we will later generalize. Equation 4 therefore tells us that the posterior probability of a given combination is proportional to the probability of our outside data sample conditional on that combination. For the case of a non-uniform prior this proportionality does not hold and the final reduction in equation 4 does not apply.

The analysis is further simplified because the conditional probability of our data given a combination, $\Pr(\boldsymbol{n}|C_i)$, has a closed form according to the formula for sampling without replacement. Namely,

$$\Pr(\boldsymbol{n}|C_i) = \prod_{s=1}^{S}\left(\prod_{k=0}^{k_s-1}\frac{C_{s,i}-k}{N-(s-1)n_s-k}\right). \tag{5}$$

Given equation 5, it is theoretically possible to iterate over all eligible combinations of counts at the subpopulation level and exactly calculate the posterior distribution over those counts given the outside sample data in $\boldsymbol{n}$. Unfortunately, the number of combinations given in equation 2 grows astronomically large rather quickly in real-world applications. Table 1 provides examples for our application.

Table 1. Number of Eligible Combinations per County

County	Farms	Towns	Sample	Number of Eligible Combinations
Bristol	42	3	11	55,278
Kent	126	5	27	4,421,275
Newport	214	6	44	1,291,150,035
Providence	425	16	82	115,508,396,906,738,000,000,000,000
Washington	436	9	65	9,801,540,147,002,170

Because it is not computationally feasible using contemporary hardware to calculate equation 5 for each possible combination, we propose a (pseudo) random sampling procedure in which valid combinations are sample uniformly from $\boldsymbol{C}$. These samples are generated by recognizing that each subpopulation's count falls in a range containing $N-n+1$ consecutive integers, whose lower bound is found in our sample for that subpopulation. Revisiting our previous Bristol County example, wherein $N-n+1=42-11+1=32$, it is only possible for subregional values to fall in the set, $\{n_s (+0),\ n_s+1,\ldots,n_s+31\}$. Since each subregion must have the same size range, the problem reduces to picking uniform integers in this range. If we offset the uniform integers by their minimum values, then all the random choices must add up to the same total (*also* $N-n+1=32$) to be a valid combination as described above.

This is a well-known problem for which the solution is to randomly choose switching points, $s_s \in \{s_2,\ldots,s_S\}$, without replacement from the set of integers, $\{1, 2,\ldots,N-n\}$. The sampled combination is then derived by differencing the switching points after setting $s_1=0$ and $s_{S+1}=N-n+1$. In order to handle the minimum switching interval being size 1, the resulting differences are added to the sample value, minus 1, and the sampled switching points are taken from $N-n+S-1$ candidate values. Downscaling Bristol County into *three* subregions provides $N-n+S-1=42-11+3+1=35$. We would thus randomly sample $S-1=2$ switching points from $\{1,\ldots,35\}$, setting $s_1=0$ and $s_4=36$.

Choosing a Prior

Researchers have two broad choices for estimating the Bayesian prior used in our estimation method: an uninformative (uniform) prior or an informative one. While in its most generalized form, our method has no requirement that subpopulations have additional characteristics from which to estimate a prior, researchers may be able to elicit a more informative prior based on additional characteristics of the sampling units. For example, in the case of classifying farms into subregions, these additional data may include population, land area, demographic data, etc., at the subregion level. We now outline a rigorous procedure for eliciting an informative prior. In cases where such additional characteristic information is unavailable for whatever reason, researchers have little choice but to assume a uniform prior across the subpopulations.

In many cases, the assumption that counted units have an equal probability of occurrence across subpopulations is unrealistic, particularly in our example application of estimating farm counts. If the data available to the researchers consist of aggregate count data for multiple populations, as well as additional covariates at the subregional level (and can thus be summed to the regional level, e.g., population, land area, demographics, spatial information), one can test and identify potential informative priors by regressing these covariates

(summed to the population level) on the population-level count data. By identifying the covariates that are most predictive of (correlated with) counts at the population level, one can use these relationships to estimate subregional farm counts. In this fashion, a more informative prior is elicited than the simple uniform prior. See Figure 1 for an illustration of how the data analysis is structured.

To compare the accuracy of estimates resulting from an informative versus an uninformative prior, we used the above procedure to elicit and compare various informative priors using supplemental subregional data obtained from the 2010 United States Census. Using county-level Census data, we performed a comprehensive regression analysis, regressing various combinations of potentially relevant covariates such as population, area, and population density on our five county-level farm counts. Land area was by far the most predictive covariate, being significantly positively correlated with farm counts (0.928). In calculating our prior, we ignore the constant term in the regression to reduce bias introduced by the fact that the constant term is only meaningful at the aggregated level because we scaled up the data. Removing the constant term also imposes the intuitive constraint that a subregion with zero land area must also contain zero farms. Because of the combinatorial nature of this problem, the regression-based priors are normalized to sum to the known aggregate counts. This makes the elicitation of the informative prior in our example equivalent to distributing N across subregions based on their proportional land area, such that $n_{\mathrm{n}}^{0} = \text{Land Area}_{n}/\text{Land Area}_{N}$.

In what follows, we will explicitly compare predictions of the uniform prior against those of this simple informative prior and compare both against maximum likelihood.

Although geographic downscaling is traditionally a spatial problem, the general form of our method ignores issues of spatial dependency in favor of a more parsimonious method that requires much less data (and less technical expertise in the area of spatial modeling). However, the generalized method can be easily expanded and the use of an informative prior in our procedure makes the incorporation of features such as spatial dependence relatively straightforward. While we opt for a simple, one-parameter, area-based prior as an example here, myriad potential models for eliciting an informative prior exist, including those discussed in the introduction. Researchers with the prior belief that the posterior distribution follows a spatial dependency structure (such as spatial lag or autocorrelation) can easily incorporate such beliefs into this methodology by eliciting their priors using a spatial model such as geographically weighted regression (GWR), among many choices.

Sample Data

Rhode Island has 39 municipalities grouped into five counties: Bristol, Kent, Newport, Providence, and Washington County. Counties range in size from 3 towns in Bristol County to 16 in Providence County. Our aggregated data

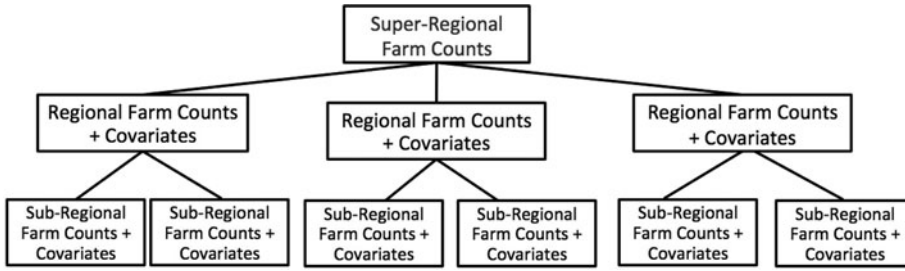

Figure 1. Structure of Sample Data Analysis

source comes from the 2012 Ag Census, which contains farm counts by county and consequently, at the state level. Our outside sample data comes from a survey administered by the University of Rhode Island in 2011–2012, in collaboration with local government agencies and agricultural organizations. It contains a list of addresses for 229 of the 1,243 farms reported in the Ag Census. We further supplement these data with additional subregional data provided by the 2010 Census. of these data, only land area was used in our final estimation procedure (indirectly, to elicit the informative prior). This information is presented in Table 2.

Methods

Clearly the unknowns in our data set are the city-level counts. We focus instead on the county totals, as if unknown, so that we can compare the results of our procedure against the true underlying distribution. By aggregating our regional counts to the state level and aggregating our sample data to the county level, we can compare the accuracy of our estimates using (i) a uniform prior, (ii) a simple spatial prior, and (iii) maximum likelihood. The maximum likelihood estimates (MLE) differ from those of the Bayesian estimates under the uniform prior since because N and n are known, MLE estimates for N_s simplify to $\hat{N}_s^{\mathrm{MLE}} = (n_s/n)N$.

For each methodology, we evaluate the normalized root mean squared error (NRMSE) of our posterior point estimates relative to the values reported in the Ag Census. The NRMSE is simply the familiar root mean squared error (RMSE), scaled by the average region-size, $\bar{N}_s = N/S$. Using NRMSE supports our goals of estimating the effects of both population size, N, and sample size, n/N, on the relative performance of each method.

To obtain estimates for the Bayesian methods, we use the sampling procedure described in "Bayesian Downscaling of Aggregated Count Data" to calculate estimates from 100,000 sample combinations. We report as point estimates the means of the posterior count distributions, which is why decimal values are observed in the estimates despite only considering integer-valued combinations. We repeat the procedure 200 times for each estimate to obtain

Table 2. Farms, Municipalities, and Land Area (mi^2) per County

County	Farms	Land Area (mi^2)
Bristol	42	24.16
Kent	126	168.53
Newport	214	102.39
Providence	425	409.5
Washington	436	329.23

Note: Pearson's Rho = 0.928.

standard errors of our estimates and calculate the NRMSE of our posterior estimates relative to the values reported in the Ag Census.

To test the effect of population size, N, on estimation performance, we simulate five subregions with our county-size proportions and sample proportion (approx. 0.2), at varying population sizes from $N = 50$ up to $N = 5{,}000$. To test the effect of sample size on estimation performance, we again simulate five subregions with our county-size proportions, this time with a fixed population size of $N = 1{,}250$ and with varying sample sizes from $n/N = 0.01$ up to $n/N = 0.50$.

For each comparison, we conduct two simulations, one with observations bootstrapped from our observed Rhode Island sample, and another randomly sampled from a multinomial distribution taken only from the population parameters. The results are nearly identical across the paired simulations, indicating that the sample data we collected do not contain extreme deviations from the projected sampling distribution.

Results

The mean NRMSE and standard error for each method are presented in Table 3, as a function of varying sample size. For a population of 1,250, the Bayesian methods consistently outperform MLE for all sample sizes up to half of the population. Among the Bayesian methods, the simple, area-based, informative prior was best for small samples (due to greater sampling variability), but the uninformative (uniform) prior was best for samples comprising at least 5 percent of the population. These results suggest a somewhat counterintuitive finding: namely, that even in cases where detailed spatial information is available, many applied problems will get more accurate results using the uninformative prior, even when it might not seem realistic to the application at hand. The reasoning is that even relatively small samples will quickly become more representative of the underlying population than a good informative prior, but not so representative as to obviate the need for a Bayesian approach over MLE.

https://doi.org/10.1017/age.2017.26

Table 3. Mean Normalized Root Mean Square Error by Estimate Type: Fixed Population Size (1,250) with Increasing Sample Size

Sample Proportion (n/N)	0.01	0.05	0.10	0.20	0.33	0.50
Uniform prior						
Mean	0.393	**0.207**	**0.143**	**0.1**	**0.077**	**0.055**
Std. dev.	(0.163)	**(0.088)**	**(0.058)**	**(0.042)**	**(0.029)**	**(0.022)**
Overall win %	0.185	**0.535**	**0.61**	**0.64**	**0.495**	**0.465**
Win % vs. Informative	0.19	**0.59**	**0.74**	**0.805**	**0.785**	**0.745**
Win % vs. MLE	0.95	**0.69**	**0.815**	**0.79**	**0.61**	**0.585**
Informative prior						
Mean	**0.245**	0.207	0.173	0.133	0.101	0.069
Std. dev.	**(0.014)**	(0.024)	(0.025)	(0.026)	(0.028)	(0.024)
Overall wins	**0.795**	0.405	0.25	0.19	0.185	0.18
Wins vs. uniform	**0.81**	0.41	0.26	0.195	0.215	0.255
Wins vs. MLE	**0.96**	0.69	0.505	0.315	0.31	0.335
Maximum likelihood						
Mean	0.57	0.265	0.176	0.119	0.086	0.059
Std. dev.	(0.209)	(0.098)	(0.063)	(0.044)	(0.029)	(0.021)
Overall wins	0.02	0.06	0.14	0.17	0.32	0.355
Wins vs. uniform	0.05	0.115	0.185	0.21	0.39	0.415
Wins vs. informative	0.04	0.31	0.495	0.685	0.69	0.665

Note: Boldface indicates estimates with the smallest RMSE compared to the true, known values.

When examining Table 3, it is important to note that the conventional wisdom regarding standard errors and statistical significance does not hold because of correlated testing. That is, we are not testing whether one measure produces a lower NRMSE on average over independent tests, rather, we are testing whether one measure produced a statistically significantly lower NRMSE than the others across the same simulation tests. So, we do not present p-values or a fully developed hypothesis testing framework with our results. Rather, the "winning method" which is bold-faced in each column of Table 3, is determined according to a single transferable vote system as outlined in Tideman (1995) based on the percentage of "wins" (lower observed NRMSE) out of 200 trials. In essence, the winning model is chosen according to winning a simple majority of trials outright, or else winning a "runoff" between the top two candidates.

For the interested reader, pairwise tests of statistically significant better performance between two estimators can be determined as follows. The null hypothesis for pairwise comparison of two identical estimators is that the number of wins for each estimator follows the binomial distribution with $n = 200$ draws and $p = 0.5$. So, for individual, pairwise tests of performance, the threshold for statistical significance at the 95 percent and 99 percent levels are 113 wins (56.5 percent) and 117 wins (58.5 percent), respectively. Clearly, to evaluate every possible pairwise test in Table 3 involves many hypotheses, so p-values would need to be adjusted using either a Bonferroni correction or stepdown methods to control the family-wise error rate (e.g., Romano and Wolf 2005). Explicit testing of multiple hypotheses in this fashion is beyond the scope of this paper.

Table 4 is structured similarly but shows the effect of varying population size, given the sample size held fixed at $n/N = 20\%$ of the population. The table also shows the Bayesian methods consistently outperforming MLE, but show subtly different patterns of performance of the informative prior against the uniform prior. With the sample size held at a fixed percentage, the informative prior outperforms for populations smaller than 500, while the uniform outperforms for larger populations. For populations of exactly 500, the performance of the two priors is not statistically different at conventional levels.

Discussion

The above results are primarily focused on evaluating the performance of our Bayesian methods for a case where the underlying distribution is known. However, our method is only designed to be useful in cases where this information is unavailable. Furthermore, applications of our procedure to new problems will likely involve variation in population size, sample size(s) and availability of an informative prior, distinct from the permutations described here. In this section, we consider the possibility that future researchers have access to data at higher levels of aggregation, similar to

https://doi.org/10.1017/age.2017.26

Table 4. Mean Normalized Root Mean Square Error by Estimate Type: Fixed Sample Proportion (0.20) and Increasing Population Size

Total Farm Count (*N*)	50	100	250	500	750	1,250	2,000	5,000
Uniform prior								
Mean	0.424	0.315	0.194	0.147	**0.124**	**0.1**	**0.079**	**0.053**
Std. dev.	(0.168)	(0.129)	(0.081)	(0.06)	**(0.052)**	**(0.042)**	**(0.033)**	**(0.021)**
Overall wins	0.06	0.105	0.255	0.42	**0.605**	**0.64**	**0.63**	**0.635**
Wins vs. informative	0.06	0.11	0.285	0.455	**0.665**	**0.805**	**0.85**	**0.985**
Wins vs. MLE	0.965	0.905	0.88	0.865	**0.635**	**0.79**	**0.755**	**0.635**
Informative prior								
Mean	**0.258**	**0.203**	**0.159**	**0.138**	0.135	0.133	0.129	0.129
Std. dev.	**(0.097)**	**(0.075)**	**(0.045)**	**(0.038)**	(0.034)	(0.026)	(0.022)	(0.012)
Overall wins	**0.94**	**0.885**	**0.705**	**0.525**	0.335	0.19	0.13	0.005
Wins vs. uniform	**0.94**	**0.89**	**0.715**	**0.545**	0.335	0.195	0.15	0.015
Wins vs. MLE	**1**	**0.975**	**0.885**	**0.7**	0.62	0.315	0.22	0.015
Maximum likelihood								
Mean	0.605	0.42	0.252	0.182	0.15	0.119	0.094	0.061
Std. dev.	(0.225)	(0.142)	(0.103)	(0.071)	(0.055)	(0.044)	(0.036)	(0.022)
Overall wins	0	0.01	0.04	0.055	0.06	0.17	0.24	0.36
Wins vs. uniform	0.035	0.095	0.12	0.135	0.105	0.21	0.245	0.365
Wins vs. informative	0	0.025	0.115	0.3	0.38	0.685	0.78	0.985

Note: Boldface indicates estimates with the smallest RMSE compared to the true, known values.

how we have both state-level and county-level farm counts for Rhode Island from the Ag Census, and county-level land area data from the U.S. Census.

If it can be assumed that spatial (or other group-wise dependence) patterns are likely to hold at higher levels of aggregation, then an informative prior can be calibrated from that data and applied in the downscaling problem. In our application, that would mean calibrating the land-area prior from county-level data and then applying it to the city-level downscaling problem. Depending on the application, however, this assumption may not be palatable. Spatial econometric models can be conceptualized as having a direct effect from the covariates and an indirect effect from the spatial dependence structure. If this indirect effect is relatively smaller at higher levels of aggregation, then calibrating the prior at higher levels will cause it to appear more informative than it actually will be in the downscaled analysis. Identifying when this problem materially affects the analysis is an area for future research. That said, there is no reason why spatial dependence observed in an econometric model would predict non-random sampling, so whenever the spatial prior is suspect, researchers can always default to the uninformative prior for reasonable performance.

Beyond spatial dependence defined econometrically, there is also the possibility that the outside data sample is non-random, in the sense that spatial factors influence response rates. At higher levels of aggregation, this can be tested simply be evaluating the degree to which the sample contains outliers relative to a typical sample from a multinomial distribution. A further verification step is possible using the simulation methods outlined above at higher levels of aggregation. Namely, the bootstrapped analysis can be replicated with counts drawn directly from a multinomial distribution instead of from the sample data. Below, we give an example of simulation results obtained in this fashion in Tables 5 and 6, which replicate our Tables 3 and 4 but do not use our sample data. For our specific application, it can be observed that the results are nearly identical, the desired outcome indicating that systematic sampling bias is unlikely to be a problem in our application.

If we consider the city-level downscaling problem in our application, the above procedures indicate that every county in Rhode Island should be estimated using the informative, area-based prior. Clearly, we do not have the underlying, true distribution of city-level farm counts for verification, so we include this observation only for completeness.

Two issues not previously addressed are (i) the effects of uncertainty in the top-level population estimates, and (ii) the scenarios in which MLE does outperform the Bayesian estimators, according to conventional wisdom based on asymptotic results. While our procedure is designed to mitigate potential estimation error resulting from the increased sampling variability inherent in relatively small samples, it does not account for potential error in the aggregate count data. In our application, for example, the Ag Census farm counts for Rhode Island are reported as 1,243 total farms with a standard

Table 5. Mean Normalized Root Mean Square Error by Estimate Type using Simulated Sample: Fixed Population Size (1,250) with Increasing Sample Size

Sample Proportion (n/N)	0.01	0.05	0.10	0.20	0.33	0.50
Uniform prior						
Mean	0.409	**0.213**	**0.151**	**0.103**	**0.077**	**0.055**
Std. dev.	(0.168)	**(0.076)**	**(0.062)**	**(0.049)**	**(0.032)**	**(0.022)**
Overall wins	0.17	**0.49**	**0.58**	**0.64**	**0.485**	**0.47**
Wins vs. informative	0.175	**0.57**	**0.7**	**0.805**	**0.76**	**0.725**
Wins vs. MLE	0.945	**0.86**	**0.795**	**0.78**	**0.63**	**0.585**
Informative prior						
Mean	**0.245**	0.208	0.174	0.132	0.098	0.069
Std. dev.	**(0.015)**	(0.024)	(0.026)	(0.027)	(0.029)	(0.026)
Overall wins	**0.815**	0.43	0.3	0.18	0.21	0.225
wins vs. uniform	**0.825**	0.43	0.3	0.195	0.24	0.275
Wins vs. MLE	**0.975**	0.66	0.5	0.34	0.37	0.365
Maximum likelihood						
Mean	0.604	0.255	0.181	0.122	0.089	0.061
Std. dev.	(0.228)	(0.086)	(0.066)	(0.044)	(0.034)	(0.023)
Overall wins	0.015	0.08	0.12	0.18	0.305	0.305
Wins vs. uniform	0.055	0.14	0.205	0.22	0.37	0.415
Wins vs. informative	0.025	0.34	0.5	0.66	0.63	0.635

Note: Boldface indicates estimates with the smallest RMSE compared to the true, known values.

error of 236 (USDA 2014). The analysis thus far suggests that incorporating an error term on the total count may have non-linear effects because of simultaneous changes both in the population size, N, and in the sample proportion, n/N.

To address this concern, we repeated the simulation analysis using the uniform prior, with each replication using a different total farm count drawn from a normal distribution with mean and standard deviation according to the reported Ag Census mean and standard error. The mean estimated farm counts arising from this procedure were within 1 percent of the values estimated with $N = 1{,}243$. This suggests that errors in top-level counts are less of a concern, as long as (i) it is recognized that the division of the population into groups will necessarily result in estimates that are proportional to the total used, and (ii) that the estimation error in the total count is not so large as to make the collected sample size unlikely or impossible.

Finally, it is important to give proper context to our finding that these Bayesian methods outperform MLE. Clearly, this is a finite sample result

Table 6. Mean Normalized Root Mean Square Error by Estimate Type using Simulated Sample: Fixed Sample Proportion (0.20) and Increasing Population Size

Total Farm Count (*N*)	50	100	250	500	750	1,250	2,000	5,000
Uniform prior								
Mean	0.394	0.277	0.214	0.15	**0.126**	**0.103**	**0.077**	**0.053**
Std. dev.	(0.173)	(0.111)	(0.09)	(0.063)	**(0.053)**	**(0.049)**	**(0.031)**	**(0.022)**
Overall wins	0.05	0.105	0.245	0.41	**0.54**	**0.64**	**0.72**	**0.675**
Wins vs. informative	0.05	0.115	0.265	0.46	**0.635**	**0.805**	**0.93**	**0.99**
Wins vs. MLE	1	0.98	0.93	0.725	**0.835**	**0.78**	**0.78**	**0.675**
Informative prior								
Mean	**0.242**	**0.193**	**0.153**	**0.142**	0.134	0.132	0.131	0.126
Std. dev.	**(0.095)**	**(0.063)**	**(0.05)**	**(0.036)**	(0.035)	(0.027)	(0.023)	(0.018)
Overall wins	**0.95**	**0.88**	**0.725**	**0.525**	0.35	0.18	0.07	0.01
Wins vs. uniform	**0.95**	**0.885**	**0.735**	**0.54**	0.365	0.195	0.07	0.01
Wins vs. MLE	**1**	**0.98**	**0.93**	**0.725**	0.57	0.34	0.155	0.015
Maximum likelihood								
Mean	0.561	0.397	0.265	0.18	0.148	0.122	0.092	0.059
Std. dev.	(0.198)	(0.165)	(0.096)	(0.067)	(0.055)	(0.044)	(0.037)	(0.021)
Overall wins	0	0.015	0.03	0.065	0.11	0.18	0.21	0.315
Wins vs. uniform	0.01	0.075	0.11	0.14	0.165	0.22	0.22	0.325
Wins vs. informative	0	0.02	0.07	0.275	0.43	0.66	0.845	0.985

Note: Boldface indicates estimates with the smallest RMSE compared to the true, known values.

since, asymptotically, Bayesian updating with a uniform prior converges to MLE, whereas in small samples MLE is equivalent to Bayesian updating with zero sample weight on the prior. Also, it may not be immediately obvious, but our application data set includes considerable variation in the group sizes to be estimated: 42, 126, 214, 425, and 436 (from Table 2). Having extremes in the group sizes, especially on the small end, leads to inherently noisier sampling of the smaller groups. This problem can be conceptualized as arising from the probability that a given sample will be representative of the population conditional on population size and sample size.

To show how variation across group sizes affects the performance of MLE relative to the Bayesian methods described here, we ran some preliminary simulations. The simulated group sizes were all drawn IID from a normal distribution with sigma given by a fraction of the mean value, and samples were then drawn from the resulting multinomial distribution. Our sample data had a standard deviation of 71 percent of the mean count, and MLE did not outperform the Bayesian methods for any population of $N = 5{,}000$ or below. Reducing the standard deviation to 50 percent of the mean count, we found that MLE was statistically significantly best (lowest NRMSE) for populations above 2,000 (as might appear in Table 6). These preliminary results suggest that the efficacy of MLE relative to the Bayesian methods is not only a function of population size and sample size, but also of the degree of heterogeneity in the subpopulation counts to be estimated. We leave exact quantification of these tradeoffs as an area for future research.

Conclusion

Micro-level statistical data are often unavailable at the desired level of disaggregation, despite their critical importance for applied policy research. Herein, we present a Bayesian methodology for "downscaling" aggregated count data to the micro-level, using an outside statistical sample. Our procedure combines numerical simulation with exact calculation of combinatorial probabilities. We motivate our approach with an application estimating the number of farms in a region, using count totals at higher levels of aggregation, and data sourced from the 2012 USDA Ag Census. In a simulation analysis over varying population sizes, we demonstrate both robustness to sampling variability and outperformance relative to maximum likelihood. Our results show that Bayesian methods have better finite sample performance than MLE in many cases relevant to applied research, especially for relatively small populations ($N < 5{,}000$).

We develop a number of methods for applied researchers to calibrate informative prior probabilities, and to estimate whether the combination of sample size and population size in their application will perform best with their informative prior, or with an uninformative alternative. In many cases, the uninformative prior performs reasonably well and can be used as a default in cases where an informative prior is unavailable, or cannot be

reasonably calibrated due to spatial considerations. We also show how the process of calibrating the priors can be simulated to verify that they are not being affected adversely by outside sample data that contains too many outliers. Our methods appear to be robust, both to sampling variability in the outside data sample and also to uncertainty in the top-level population counts. An area for future research is determining the effects of heterogeneity in subpopulation sizes on the relative performance of MLE in smaller populations.

References

Chakir, R. 2009. "Spatial Downscaling of Agricultural Land-Use Data: An Econometric Approach Using Cross Entropy." *Land Economics* 85(2): 238–251.

Coelho, C.A.S., D.B. Stephenson, F.J. Doblas-Reyes, M. Balmaseda, A. Guetter, and G.J. Van Oldenborgh. 2006. "A Bayesian Approach for Multi-Model Downscaling: Seasonal Forecasting of Regional Rainfall and River Flows in South America." *Meteorological Applications* 13(1): 73–82.

Dendoncker, N., P. Bogaert, and M. Rounsevell. 2006. "A Statistical Method to Downscale Aggregated Land Use Data and Scenarios." *Journal of Land Use Science* 1(2–4): 63–82.

Fasbender, D., and T.B.M.J. Ouarda. 2010. "Spatial Bayesian Model for Statistical Downscaling of AOGCM to Minimum and Maximum Daily Temperatures." *Journal of Climate* 23(19): 5222–5242.

Gärtner, D., A. Keller, and R. Schulin. 2013. "A Simple Regional Downscaling Approach for Spatially Distributing Land Use Types for Agricultural Land." *Agricultural Systems* 120: 10–19.

Gocht, A., and N. Roder. 2011. "Salvage the Treasure of Geographic Information in Farm Census Data." Paper presented at the 2011 International Congress of the European Association of Agricultural Economists, Zurich.

Hashmi, M.Z., A.Y. Shamseldin, and B.W. Melville. 2009. "Statistical Downscaling of Precipitation: State-of-the-Art and Application of Bayesian Multi-Model Approach for Uncertainty Assessment." *Hydrology and Earth System Sciences Discussions* 6(5): 6535–6579.

Howitt, R., and A. Reynaud. 2003. "Spatial Disaggregation of Agricultural Production Data Using Maximum Entropy." *European Review of Agricultural Economics* 30(3): 359–387.

Murphy, J. 1999. "An Evaluation of Statistical and Dynamical Techniques for Downscaling Local Climate." *Journal of Climate* 12(8): 2256–2284.

Polasek, W., C. Llano, and R. Sellner. 2010. "Bayesian Methods for Completing Data in Spatial Models." *Review of Economic Analysis* 2(2): 194–214.

Purcell, N.J., and L. Kish. 1980. "Postcensal Estimates for Local Areas (Or Domains)." *International Statistical Review/Revue Internationale De Statistique* 48(1): 3–18.

Romano, J.P., and M. Wolf. 2005. "Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing." *Journal of the American Statistical Association* 100(469): 94–108.

Tideman, N. 1995. "The Single Transferable Vote." *The Journal of Economic Perspectives* 9(1): 27–38.

United States Department of Agriculture. 2014. *2012 Census of Agriculture, Appendix A: Census of Agriculture Methodology*. Washington, DC: U.S. Government Printing Office.

von Storch, H., E. Zorita, and U. Cubasch. 1993. "Downscaling of Global Climate Change Estimates to Regional Scales: An Application to Iberian Rainfall in Wintertime." *Journal of Climate* 6(6): 1161–1171.