Analyzing small geographic area datasets containing values having high levels of uncertainty

Daniel A. Griffith
School of EPPS
University of Texas at Dallas
Richardson, TX, USA
dagriffith@utdallas.edu

Robert P. Haining
Department of Geography
University of Cambridge
Cambridge, UK
rph26@cam.ac.uk

Abstract—**Data collected and then post-stratified by small geographic areas frequently result in small, or even zero, sample sizes for some areas. Government agencies faced with this outcome commonly suppress many of the sample-based attribute measures for confidentiality reasons. Meanwhile, modeling concerns of spatial scientists faced with this situation include the accompanying large standard errors for parameter estimates obtained with such data, as well as how to deal directly with any missing values. This paper addresses these two issues in the context of spatial statistical modeling that accounts for high levels of uncertainty for some data values in two specific datasets. Its purpose is to demonstrate ways of handling high levels of uncertainty in georeferenced data. In doing so, empirically-based findings summarized in this paper illustrate selected approaches that can be employed to account for high levels of uncertainty for some data values in a dataset. Its implications should be of interest to users in government, the private sector, and the academic community who engage in the modeling of georeferenced data.**

Keywords: Bayesian; error; imputation; missing data; Poisson regression; sample size; uncertainty

I. INTRODUCTION

All data contain error as a consequence of imperfections in the processes of taking and recording measurements. The nature and sources of error are many and varied. Here we address "extreme" or a "high level" of data uncertainty.

One extreme type of data uncertainty arises when some data values for one or more variables in a data set are missing. This case might be referred to as "total" or "near total" uncertainty with regard to the values. Another type of extreme uncertainty arises when data are obtained through sampling. Data values may have large sampling standard errors if sample sizes are very small. This situation can arise in the case of spatial data when sampling is not stratified, or stratification is only defined in terms of large spatial units so that the number of samples falling inside smaller areas can be very small or even zero, or sampling standard errors vary from one geographical area to another because sample sizes are not uniform.

In these preceding situations, analysts may want to estimate missing values, whilst in the case of data with large sampling standard errors they might want to improve the precision of data values—especially those with very large standard errors. Whether a dataset contains missing data values or some data values with low precision (or both), an analyst may also wish to fit a model to some part of the data for the purpose of hypothesis testing. Typically, solutions to these types of problems use the information contained in the rest of a dataset to either estimate the missing values or improve precision. If a dataset is geographical, so that each observation has a geo-reference, methods typically attach greater weight to data values that are geographically nearby—borrowing more information from these data values than from those further away. Nearby values tend to be more highly correlated than those further apart so that data values at nearby locations may carry considerable information about a missing value or a value that has low precision. Different approaches fall broadly into two categories: non-parametric and parametric.

The spatial moving average (MA) method is an example of a non-parametric approach. For example, data values in areal units that abut a polygon with no data value are averaged and the average used as an estimate of the missing value. An adjacent polygon that is "closer" to the polygon with the missing value might be given more weight in computing such a local average. Averaging may be extended to more distant neighbors, in which case differential weighting is essential to reflect the assumption that values for more distant polygons carry less information about an unknown data value than values for those polygons that are nearby. If the data value in a polygon is known but has low precision, then its value may be included in the spatial averaging process. The term "borrowing strength" is typically used to refer to these types of procedures.

Weaknesses of non-parametric approaches are addressed by parametric approaches for which an analyst makes assumptions about the distribution of a variable. In this case, distinguishing between continuous valued variables (for which the normal assumption might be appropriate) and discrete valued variables such as count data (for which the Poisson, binomial or negative binomial distribution might be appropriate) is important. In these approaches, data values, whether missing or with low precision, are treated as parameters of a distribution. Typical central tendency parameters are the mean, median or mode. By specifying a multivariate distribution derived from a permissible model for spatial variation of an attribute, a parametric approach exploits the spatial correlation structure in data to return an estimate that recognizes that nearby values carry more information about a

missing or low precision data value than values further away. An advantage of the parametric approach is that it captures the uncertainty associated with estimates, providing not only confidence intervals but also the associated distribution. In a modeling context, this uncertainty can be recognized in the estimation of other parameters (and their standard errors), which may be important in hypothesis testing. The disadvantage with parametric methods is that typically they are much more difficult to implement than non-parametric approaches, and require an analyst to make and then check distributional assumptions.

II. Example: Estimating Suppressed/Missing Corn Harvest Data by County in Ohio, 2005-2008

The National Agricultural Statistics Service (NASS) of the United States Department of Agriculture (USDA) has a standing interest in estimating agricultural commodities, especially crop area planted, area harvested, and yields for a variety of selected crops by small geographic area. Data collected and then post-stratified by small geographic areas frequently result in small, or even zero, sample sizes for some areas. Government agencies faced with this outcome commonly suppress many of the sample-based attribute measures for confidentiality reasons. Analysts may be interested in estimating the missing values. The variable analyzed here for illustrative purposes is acres of corn harvested in Ohio for the years 2005-08.

A. Non-Parametric Imputations

Figure 1 shows the 88 Ohio counties aggregated into 9 districts, as well as the counties containing a total of 31 missing values across 4 annual production periods. The nature of these data suggests that they contain both temporal and spatial autocorrelation.

Figure 1. The state of Ohio partitioned into districts (heavy black lines) and counties (thin black lines).

Missing values occur in three regional clusters. Thus, the MA estimate of one in a cluster requires the estimate of the others in that cluster. In order to capture both direct and indirect effects, this situation results in k equations in k unknowns—where k is the number of counties in a cluster—which can be solved algebraically with software such as Mathematica or Maple.

The general form of the spatial MA estimation equation for each of a pair of adjacent counties that have the only missing values in a district is as follows:

$$A = \frac{\text{County\#1\_acres}[\sum(\text{reported neighbor counties' harvested acres}) + \text{County\#2\_missing\_value}]}{\sum(\text{neighbor counties' total acres})},$$

$$B = \frac{\text{County\#2\_acres}[\sum(\text{reported neighbor counties' harvested acres}) + \text{County\#1\_missing\_value}]}{\sum(\text{neighbor counties' total acres})},$$

and

$$\text{County\#1\_imputation} = [\text{residual district total}]\, A/(A+B),$$

where "residual district total" is the known total harvested acreage of corn in the district minus the sum of the harvested acreages of corn for those counties in the district with reported values.

B. Parametric Imputations

In the *E*xpectation-*M*aximization (EM) algorithm (Dempster et al., 1977), an iterative procedure for computing maximum likelihood estimates when datasets are incomplete, the E-step replaces missing data by their conditional expectations (Flury and Zoppè, 2000; Little and Rubin, 2002; Rao, 2003). Yates (1933) shows for analysis of variance that if each missing observation is replaced by a parameter to be estimated (i.e., the conditional expectation for a missing value), the resulting modified analysis becomes straightforward by treating the estimated missing value as a parameter to be estimated (i.e., an imputation). This specification relates a linear model formulation of the missing value problem to the regression problem of predicting new values. This paper extends Yate's formulation (see Griffith, 2010) to the case of a spatially autocorrelated binomial random variable drawing on generalized linear and generalized linear mixed model (GLM and GLMM; see Hsiao, 2003; Baltagi, 2005) theory. Acres of harvested corn (Y) is a binomial random variable because it constitutes a percentage of each county's total acreage (i.e., an upper limit exists on the number of possible acres).

For any county, the model may be specified as follows (Griffith, 2004):

$$E(Y) = N \frac{e^{\alpha + F\beta + \sum_{k=1}^{K} E_k \gamma_k + \varphi}}{1 + e^{\alpha + F\beta + \sum_{k=1}^{K} E_k \gamma_k + \varphi}}, \quad (1)$$

where α denotes the intercept, β denotes the binomial regression coefficient for the covariate F (here, the number of farms in a county), E_k denotes eigenvector k (see Griffith, 2003; Tiefelsdorf and Griffith, 2007) and γ_k denotes its binomial regression coefficient (the sum of these products is the spatial filter)—this term also represents the spatially

structured random effects—and φ denotes the spatially un-structured random effects. Equation (1) is the full model specification; setting $\varphi = 0$ reduces the specification to a fixed effects GLM; setting $\varphi = 0$ and $\gamma_k = 0$, for all k, reduces the specification to a simple bivariate GLM. Table 1 reports the 95% confidence intervals (CIs) for the imputations based upon the GLMM.

TABLE I. APPROXIMATE 95% CIS FOR THE GLMM BASED IMPUTATIONS

County	2005		2006		2007		2008	
Cuyahoga	380	420	315	333	380	420	759	833
Lake	180	220	149	167	180	220	370	434
Summit	*	*	*	*	*	*	695	769
Belmont	*	*	926	961	1037	1111	926	996
Jefferson	*	*	704	739	789	863	704	774
Gallia	*	*	1029	1068	1079	1159	*	*
Jackson	*	*	*	*	*	*	2849	3019
Lawrence	*	*	993	1032	1041	1121	3031	3205
Pike	*	*	*	*	*	*	3012	3186
Scioto	*	*	*	*	*	*	4185	4373
Meigs	*	*	*	*	*	*	970	1068
Monroe	*	*	402	426	506	560	1020	1118
Noble	*	*	350	374	440	494	891	987
Vinton	*	*	*	*	*	*	935	1031

C. Comparisons

Two sets of comparisons are of interest. The first concerns the correspondences between the non-parametric MA and the GLM imputations. Diagnostic statistics appear in Table 2. None of these correspondences are very good; the space-time MA and the GLMM imputations are the closest.

TABLE II. CORRESPONDENCES BETWEEN THE NON-PARAMETRIC MA AND THE GLM IMPUTATIONS.

GLM	statistic	Non-parametric MAs		
		space	time	space -time
GLM with # of farms covariate	MAXAD	3477	3471	3451
	MAD	557	556	554
	RMSE	960	714	683
GLM with # of farms covariate & spatial filter	MAXAD	2544	2544	2493
	MAD	428	427	423
	RMSE	959	714	682
GLMM with # of farms covariate & spatially structured and un-structured random effects	MAXAD	2438	2438	2387
	MAD	416	415	411
	RMSE	953	704	673
MAXAD denotes maximum absolute difference MAD denotes mean absolute difference RMSE denotes root mean squared error				

The second set of comparisons is with data patterns. The first and last years of data for Gallia are known. Given these known end values, GLM/GLMM imputations for the two intervening years of missing data appear to be too small. Meigs, Pike, Scioto, Summit, and Vinton all have their first three years of data known; in each case, the GLM/GLMM imputed values for the fourth year appear to be too small. In contrast, Jackson, which also has its first three years of data known, has a GLM/GLMM imputed value for its fourth year that appears to be too large. In contrast, the space-time MA imputations appear quite reasonable for Meigs, Scioto, Summit, and Vinton, and too large for Jackson and Pike. These findings reflect the tendency for the statistical imputations to shrink toward a sample mean, which in this case are the residual district means for the missing values (because they are known and serve as estimation constraints), whereas the non-parametric space-time MA includes a dominant time term (weighted 13-to-1 with the spatial term).

III. EXAMPLE: MODELING EXPOSURE TO AIR POLLUTION AND STROKE MORTALITY

In a recent ecological study, the risk of stroke arising from exposure to nitrous oxide (NO_x) was estimated by small census areas in Sheffield, England (Maheswaran et al. 2006). Estimating this risk requires controlling not only for the age and sex composition of the small area populations, but also for deprivation and smoking prevalence, both of which are potential confounders. Analysis proceeded by using Poisson regression with spatial random effects to address the problem of overdispersion, which is possibly the result of missing covariates. The smoking prevalence data were obtained from a city-wide survey, which resulted in some areal units having small (and even zero) sample sizes, and hence low precision. This type of uncertainty in these data may be captured by Bayesian hierarchical modeling in which local non-parametric smoothing is applied to the smoking prevalence data, with resulting smoothed values being treated as data. Another approach is to attach a probability distribution to the smoking prevalence data associated with each census area. A comparison of these modeling results allows an assessment of the sensitivity of findings to different ways of handling this feature of the data.

Because the stratification was by ward, many enumeration districts (EDs) had small sample sizes, and some had no data at all. The non-parametric solution to this problem is to calculate a spatially averaged ratio. This means summing the observed counts in an ED and its adjacent neighbors and dividing by the sum of the expected count in an ED and its adjacent neighbors. Then this ratio is assigned to the ED and the process replicated for all EDs. The resulting model is

$$\log[p_i] = b_0 + S_i + U_i + b_1 X_{1,i} + b_2 X_{2,i}^{AVE} \qquad (2)$$

where p_i denotes the underlying true area specific relative risk of stroke mortality in area i, $X_{1,i}$ is the NO_x level in area i and $X_{2,i}^{AVE}$ is the spatially averaged smoking prevalence ratio. The term S_i is a spatially structured (zero mean intrinsic conditional autoregressive) random effect. The term U_i is a spatially unstructured (independent zero mean normal) random effect. The regression coefficients are b_0, b_1 and b_2.

In order to recognize the sampling error associated with the smoking covariate, the following errors in variable model was used for $X_{2,i}$:

$$\log[p_i] = b_0 + S_i + U_i + b_1 X_{1,i} + b_2 X_{2,i}^{EST} \qquad (3)$$

where $X_{2,i}^{EST}$ is the true log-smoking-prevalence ratio. This variable is not observed data; rather, it is a parameter to be estimated in the model using the observed count of smokers in area i (smoke.O_i) and the expected count of smokers in area i (smoke.E_i). Spatially structured and unstructured random effects terms were added to allow for spatial autocorrelation and random measurement error in the observed smoking count. Accordingly,

$$\text{smoke.}O_i \sim \text{Poisson}(\text{smoke.}\mu_i)$$

$$\text{smoke.}\mu_i = \text{smoke.}p_i \times \text{smoke.}E_i$$

$$X_{2,i}^{EST} = \log(\text{smoke.}p_i) = \text{smoke.}b_0 + \text{smoke.}S_i + \text{smoke.}U_i$$

where smoke.μ_i is the mean of the observed number of smokers and smoke.p_i is the smoothed smoking prevalence ratio. The variable smoke.S_i is a zero-mean intrinsic conditional autoregressive spatially structured random effect, and the variable smoke.U_i is an independent zero-mean normally distributed spatially unstructured random effect. Full details of these models and their fitting in WinBUGS, including discussion of computational issues, appear in Maheswaran et al. (2006).

Table 3 reports the results that we wish to contrast here. Two points are worth stressing. First, parameter estimates for the relative risk (RR) associated with different levels of exposure to NO_x and the effects of smoking are nearly the same in both models. Second, although parameter standard errors are, as expected, wider in the case of model (3) compared to model (2), reflecting the additional variability associated with the way the smoking prevalence data is used in model fitting, the differences are small. Larger standard errors are associated with the parameter estimate for the smoking variable, which again is to be expected.

TABLE III. TABLE 3. 95% CIS FROM FITTING MODELS (2) AND (3).

variable	Model (2) RR	Model (3) RR
NO_x category 5	1.27 (1.05-1.53)	1.27 (1.03-1.54)
NO_x category 4	1.16 (0.96-1.39)	1.16 (0.95-1.40)
NO_x category 3	1.04 (0.86-1.24)	1.04 (0.85-1.25)
NO_x category 2	1.08 (0.89-1.29)	1.07 (0.89-1.29)
NO_x category 1	1	1
smoking	1.03 (0.88-1.21)	1.05 (0.79-1.40)
Spatial fraction	0.005	0.006
Smoking spatial fraction		0.99
DIC	3929.11	3927.77
RR denotes relative risk. CI denotes credible interval. DIC denotes deviance information criterion.		

IV. CONCLUDING REMARKS

In the Ohio corn production example, results obtained with the non-parametric estimator of harvested acres are substantially different from those obtained with the parametric estimators. In the air pollution and stroke example, results obtained with the non-parametric estimator of the smoking prevalence ratio are nearly identical to those obtained with the parametric estimator. These two examples illustrate that the non-parametric and parametric estimators do not always render similar results. Furthermore, even when non-parametric and parametric estimates are very similar, circumstances may exist where even small differences in regression parameter standard errors can have a significant effect on hypothesis testing and the decisions following from such testing. Consequently, under what circumstances might it be necessary to use only the parametric methods also described here?

Perhaps one of the most appealing advantages of the parametric estimators is that they are accompanied by standard errors. For imputations, the resulting confidence intervals can be made quite narrow by knowing the totals of the missing values. Is the shrinkage due to this ancillary information similar to that achieved with a Bayesian solution to the problem (see Bellow, 2007)?

It is our intention to pursue these types of questions in future work, and also to undertake a more systematic analysis of missing data estimation through the use of conditional cross-validation. This involves deleting additional, known, data values from the Ohio county data, and then evaluating how well these known values are estimated by the parametric and non-parametric methods described here.

REFERENCES

Baltagi, B. (2005). *Econometric analysis of panel data.* Hoboken, NJ: Wiley.

Bellow, M. (2007). Comparison of methods for estimating crop yield at the county level, *Research Report RDD-07-05.* Washington, DC: National Agricultural Statistics Service, United States Department of Agriculture.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society.* Series B 39, 1-38.

Flury, B., and A. Zoppè. (2000). Exercises in EM. *The American Statistician.* 54, 207-209.

Griffith, D. (2003). *Spatial autocorrelation and spatial filtering: gaining understanding through theory and scientific visualization.* Berlin: Springer-Verlag.

Griffith, D. (2004). A spatial filtering specification for the auto-logistic model. *Environment & Planning A.* 36, 1791-1811.

Griffith, D. (2010). Some simplifications for the expectation-maximization (EM) algorithm: the linear regression model case, *InterStat,* March article 2 (interstat.statjournals.net/YEAR/2010/articles/1003002.pdf), 23 pp.

Hsiao, C. (2003). *Analysis of panel data.* New York: Cambridge University Press.

Little, R., and D. Rubin. (2002). *Statistical analysis with missing data,* 2nd ed. New York: Wiley.

Maheswaran, R., R. Haining, P. Brindley, J. Law, T. Pearson, and N. Best. (2006). Outdoor NO_x and stroke mortality—adjusting for small area level smoking prevalence using a Bayesian approach. *Statistical Methods in Medical Research.* 15, 499-516.

Rao, J. (2003). *Small area estimation.* New York: Wiley.

Tiefelsdorf, M., and D. Griffith. (2007). Semi-parametric filtering of spatial autocorrelation: the eigenvector approach. *Environment & Planning A.* 39, 1193-1221.

Yates, F. (1933). The analysis of replicated experiments when the field results are incomplete. *Empirical Journal of Experimental Agriculture.* 1, 129-142.