SUPPORTING INFORMATION FOR

"CLOSE-UPS AND THE SCALE OF ECOLOGY: LAND USES AND THE GEOGRAPHY OF SOCIAL CONTEXT AND CRIME"*

ADAM BOESSEN

JOHN R. HIPP

*Published in *Criminology*, volume 53, issue 3, 2015.

doi: 10.1111/j.1745-9125.2011.00245.x

SYNTHETIC ESTIMATION TO IMPUTE VALUES FOR SMALL UNITS OF ANALYSIS

When data are available at larger geographic units but not at smaller geographic units, a technique for imputing values is synthetic estimation for ecological inference (Cohen and Zhang, 1988; Steinberg, 1979). The synthetic estimation approach relies on the assumption that the relationship between variables at one level of analysis is similar at a different level of analysis, which is certainly not ideal. Nonetheless, whereas researchers often simply impute values from the larger units to the smaller units assuming homogeneity within the larger units, the synthetic estimation approach is more principled in attempting to build a model to predict such values. A brief treatment of the topic can be found in Steinberg (1979), whereas a longer discussion of the issues involved can be found in Cohen and Zhang (1988). A more recent treatment of the ecological inference problem can be found in King (1997).

For ecological inference from larger to smaller units, the three main issues to confront are as follows: 1) the necessity to build a prediction model at the next highest level of aggregation that contains valid values of the variable, and then use this model to predict the values of the variable at the smaller unit; 2) the values of the variable of interest in the smaller units must be constrained to sum to the observed total in the larger unit; and 3) the need to account for the uncertainty in this prediction. We adopt such an approach here by building a regression model at the higher level of aggregation using the coefficient estimates of this model to obtain predicted values in the smaller units, adjusting the imputed values for the smaller units such that they sum to the value in the larger unit they are contained within, and then adding uncertainty to the predicted values based on the uncertainty in the imputation model at the higher unit of analysis.

To demonstrate this approach, we used U.S. Census data and crime data for the city of Los Angeles. We used data aggregated to tracts (N = 1,053) to estimate predicted values at the block-group level and compared those with the true values. We used the following four measures in models separately, as well as combined as a measure of *concentrated disadvantage*: 1) percentage single-parent households, 2) percentage below the poverty level, 3) average household income, and 4) percentage with at least a bachelor's degree. The measure is created using regression scoring of the factor loadings from a confirmatory factor analysis: This measure has a mean of 0 and a standard deviation equal to that of the percentage of poverty measure (because this is used to scale the factor). We estimated negative binomial regression models with aggravated assault, and then robbery, as the outcome variables, controlling for standard measures used in the neighborhood context of crime literature.

RESULTS

The top half of table S.1 presents the results for several models with aggravated assault as the outcome measure; the bottom half of the table displays similar model results with robbery as the outcome measure. Each row represents the models using a particular variable as the key independent variable of interest, and each column presents the models using a particular imputation strategy (each model also contains all control variables). For example, column 1 displays the various model results when using the actual block group aggregated data. Thus, these are essentially the “gold standard” results as we actually have these various measures aggregated to block groups. Column 2 displays the results when adopting the common strategy of simply imputing the value of a measure for a tract to each of the block groups within that same tract. Column 3 uses our synthetic estimation approach with a single imputation, and column 4 uses our synthetic estimation approach with multiple (five) imputations.

We observe in row 1 that when using the percentage single-parent households as the single measure to capture concentrated disadvantage, it has a positive, but nonsignificant, effect on aggravated assaults in the true model. In column 2, the approach that simply imputes the mean value of the larger tracts to the block groups results in a much stronger, and significant, effect. Notably, across virtually all models, the coefficient estimates are always much *larger* than the true values. Thus, this approach of simply imputing the value from the larger unit into the smaller units within it always results in overestimates of the true relationship in our example data set. We see in this table that the coefficient for the measure in column 2 is always larger than the corresponding coefficient in column 1 using the “true” measure. These coefficients are typically 50% to 100% larger than the true coefficients, and sometimes much larger than this. The exception is that this approach actually yields a coefficient of the *opposite* sign in the model with single-parent households as a covariate in the robbery model. For example, whereas the true measure of percentage single-parent households did not have a significant effect on aggravated assault in the model using the true measure in column 1, it seems to have a significant positive effect with a coefficient nearly eight times larger in column 2 when using this mean imputation approach. Clearly, this pattern of results is unsatisfactory.

In column 3, we display the results for our synthetic estimates but only using a single imputation, and column 4 displays the same results when using multiple imputations. Whether using a single imputation or multiple imputations, the coefficient estimates are similar and often are reasonably close to the true values (and typically closer than the approach that simply imputes the mean value to the smaller units). The standard errors are larger for the multiple imputation approach, as expected, given that this approach accounts for the uncertainty of not actually having the measures at the smaller unit of analysis.

CONCLUSION

Whereas the synthetic estimation approach is certainly not ideal, we argue that it is preferred to the most common imputation approach of simply imputing the value from a larger unit into the subunits within that unit. We have shown here that this common strategy is undesirable. Not only does it produce standard errors that are too small, but also in these examples, it consistently produced coefficient estimates that were severely upwardly biased. It is also worth emphasizing that another common strategy, simply omitting a variable because it is missing in the smaller units of analysis, is not desirable: This will result in the well-known omitted variable problem for regression analysis, which yields biased estimates. It is therefore essential that researchers address this missing data problem directly. Although the synthetic estimation approach certainly has limitations, we argue that it is more principled than many of the existing strategies employed by applied researchers.

REFERENCES

Cohen, Michael Lee and Xiao Di Zhang. 1988. *The Difficulty of Improving Statistical Synthetic Estimation*. Washington, DC: Bureau of the Census.

King, Gary. 1997. *A Solution to the Ecological Inference Problem*. Princeton, NJ: Princeton University Press.

Steinberg, Joseph. 1979. Synthetic estimates for small areas: Statistical workshop papers and discussion. In *National Institute on Drug Abuse Research Monograph Series*, vol. 24. Washington, DC: National Institute on Drug Abuse.

Table.S.1. Using synthetic estimation to create block group level variables based on tract level measures: Negative binomial regression coefficients for aggravated assault and robbery models

		(1)		(2)		(3)		(4)	
	Aggravated assault models	True value		Value of larger unit		Single imputation		Multiple imputation	
(1)	Single parent households	0.0031		0.0238	**	0.0029		0.0030	
		(0.0020)		(0.0034)		(0.0023)		(0.0024)	
(2)	Poverty	0.0209	**	0.0325	**	0.0157	**	0.0156	**
		(0.0017)		(0.0018)		(0.0018)		(0.0022)	
(3)	Average household income	-0.0043	**	-0.0077	**	-0.0024	**	-0.0024	**
		(0.0006)		(0.0007)		(0.0006)		(0.0007)	
(4)	Education level	-0.0216	**	-0.0293	**	-0.0096	**	-0.0100	**
		(0.0019)		(0.0019)		(0.0018)		(0.0021)	
(5)	Concentrated disadvantage index	0.0445	**	0.0506	**	0.0222	**	0.0223	**
		(0.0030)		(0.0027)		(0.0028)		(0.0030)	
	Robbery models								
(1)	Single parent households	-0.0093	**	0.0163	**	-0.0066	*	-0.0058	*
		(0.0024)		(0.0042)		(0.0029)		(0.0029)	
(2)	Poverty	0.0201	**	0.0374	**	0.0196	**	0.0196	**
		(0.0021)		(0.0022)		(0.0022)		(0.0024)	

(3)	Average household income	-0.0021	**	-0.0090	**	-0.0029	**	-0.0026	**
		(0.0007)		(0.0009)		(0.0008)		(0.0009)	
(4)	Education level	-0.0153	**	-0.0222	**	-0.0063	**	-0.0065	*
		(0.0023)		(0.0024)		(0.0022)		(0.0027)	
(5)	Concentrated disadvantage index	0.0304	**	0.0515	**	0.0184	**	0.0191	**
		(0.0038)		(0.0035)		(0.0035)		(0.0036)	

*Note: ** p < .01; * p < .05; † p < .1. T-values in parentheses. All models control for: percent vacant units, percent owners, percent African American percent Latino, racial/ethnic heterogeneity, population density, and the percent aged 16 to 29. N=2,759 block groups*