Bayesian Hierarchical Model Estimates of Local Crime Perceptions

Brent D. Mast

U.S. Department of Housing and Urban Development

The contents of this article are the views of the author and do not necessarily reflect the views or policies of the U.S. Department of Housing and Urban Development or the U.S. government.

I Introduction

This study uses survey data to estimate a Bayesian hierarchical model of local crime perceptions. Data are employed from two sources.

Top level prior hyperparameters are based on crime perception responses from the American Housing Survey (AHS)[1]. The AHS is actually two surveys, metro and national, taking place in different years. I employ data from the national AHS for 2001.

I also employ data aggregated to the county level from HUD's survey of Section 8 Housing Choice Voucher (HCV) households[2]. Dubbed the Customer Satisfaction Survey (CSS), it was a three year national survey conducted between 2000 and 2002.

Nearly one-half million households returned questionnaires, answering a wide variety of questions regarding the condition of their housing and neighborhoods. The large sample was stratified by public housing agency and year. This paper focuses on responses to a question regarding neighborhood crime and drug problems.

Results indicate that the Bayesian approach yields more robust local estimates. Compared to estimates solely based on CSS data, the Bayesian estimates have lower variance and correlate more highly with published county crime rates.

The data are described in more detail in the following sections. The Bayesian Hierarchical model is then described. Estimates are presented next. Correlation of survey estimates with county crime rates is then explored. The final section summarizes my analysis.

II National Data

This section summarizes national responses to crime questions employed from the 2001 (national) AHS and CSS. It is possible to compute estimates for select counties that participate in the metro AHS in other years. My analysis only employs national AHS data. CSS responses aggregated to the county level are described in the next section.

Exhibit 1 summarizes responses to variable "crimea" from the 2001 AHS. This variable indicates responses to a question asking households if their "neighborhood has a neighborhood crime problem". Estimates are reported for HCV households and all occupied rental units. 31.4 percent of voucher households report a crime problem, as do 22.1 percent of all renters.

Exhibit 1: 2001 AHS Crime Question Responses

	Voucher households			All Occupied Rental Units		
Survey Response	Responses (N)	Weighted Frequency	Weighted % of households	Responses (N)	Weighted Frequency	Weighted % of households
Yes	101.000	267473.706	31.410	2799.000	7069411.615	22.092
No	234.000	562890.296	66.101	9831.000	24228208.443	75.714

Don't know	8.000	19202.527	2.255	248.000	613130.331	1.916
No, Don't know	242.000	582092.823	68.356	10079.000	24841338.774	77.630
No Response	1.000	1992.405	0.234	36.000	89029.032	0.278

Exhibit 2 summarizes CSS responses to a questionnaire item asking households to indicate if crime or drugs "is a big problem in (their) neighborhood." An estimated 45.1 % do not perceive a problem. An estimated 19.8 percent of households do not know if their neighborhood has a crime or drug problem. An estimated 22.6 percent of households report somewhat of a problem with crime or drugs. 10.6 percent are estimated to perceive a major problem with crime or drugs in their neighborhood.

Exhibit 2: CSS Crime Question Responses

Survey response	Responses (N)	Weighted Frequency	Weighted % of households
No response	8812	92434.8	1.927%
No Problem	231993	2162695.1	45.083%
Don't know	89800	951830.1	19.842%
Subtotal: No problem, don't know	321793	3114525.2	64.924%
Some problem	92435	1082558.0	22.567%
Big problem	36258	507632.4	10.582%
Subtotal: some problem, big problem	128693	1590190.5	33.149%
Total	459298	4797150.5	100.000%

For statistical analysis, I recode the crime responses as binary indicators. For the AHS, "Yes" responses are set to one, while "No" and "Don't know" responses are treated as zeros. For the CSS, "some problem" and "big problem" responses are set to one, and "no problem" and "don't know" responses are set to zero. Non-responses for both surveys are set to missing.

Exhibit 3 reports summary statistics for the binary indicators of perceived crime problems. Standard errors are reported as measures of standard deviation for survey data with unknown population means. All standard errors are adjusted for finite population, and CSS standard errors are adjusted for the stratified survey design. They were computed by SAS software using the Taylor expansion (linearization) method. The software uses the same formula for standard errors of the sample mean and sample proportion.[3] For this reason, and because the posterior likelihood formulas I employ are based on means, I will use the term "mean" instead of "proportion" for the remainder of this study.

Exhibit 3: Binary Crime Indicator Summary Statistics

Source	Mean	Standard Error	Lower 95 % confidence limit	Upper 95 % confidence limit
CSS (HCV Households)	0.338	0.002	0.335	0.341
AHS - HCV Households	0.315	0.028	0.261	0.369
AHS - All Rental Units	0.221	0.004	0.213	0.228

AHS estimates for all occupied rental units have a weighted mean of .221 and standard error of .004. AHS estimates for HCV occupied rental units have a weighted mean of .315 and standard error of .028. 95 % confidence limits for the weighted mean are .261 to .369 for HCV households, and .213 to .228 for all occupied rental units.

I exclude owner-occupied households from my analysis because all HCV households in the CSS rent, and AHS estimates are much lower for owners (mean of .120) versus renters (mean of .221).

The CSS indicator has a weighted mean of .338, with a standard error of .002. The 95 % confidence limits for the weighted mean are .335 to .341.

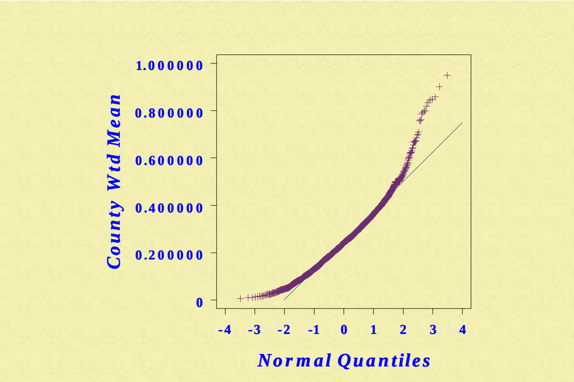
III. County Data

The CSS sample is large enough to produce accurate estimates for local areas. For this study, I focus on a subset of counties with means that can reasonably be treated as normal. The AHS prior distributions employed are also easily large enough to be treated as normal. This results in posterior distributions that can also be assumed normal. My analysis could be extended to more counties and smaller areas (such as census tracts or block groups) with a binomial/beta conjugate model.

I assume the CSS weighted county mean crime measures are normally distributed for 1158 counties meeting the following three criteria: 1) the weighted count of responses is at least 30; 2) the weighted count of households reporting a crime problem is at least 10; and 3) the weighted count of households not reporting a crime problem is also at least 10. The weights used for the criteria sum to total responses (not sampling frame).

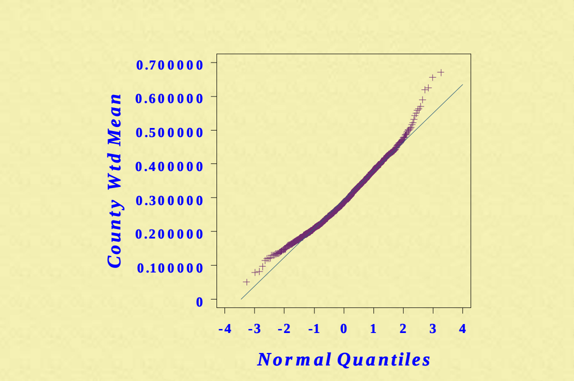
Exhibit 4 depicts a QQ plot of 2490 counties with 0<mean<1. Exhibit 5 depicts a QQ plot of the 1158 counties meeting the normality criteria. The first plot shows significant deviations from normality. The plot for counties meeting the normality criteria deviates from linearity somewhat in the tails of the distribution. But it is much closer to theoretical normal than the all county plot.

Exhibit 4: QQ Plot of County Means, All Counties



Note: N=2490, μ=.252, σ=.125.

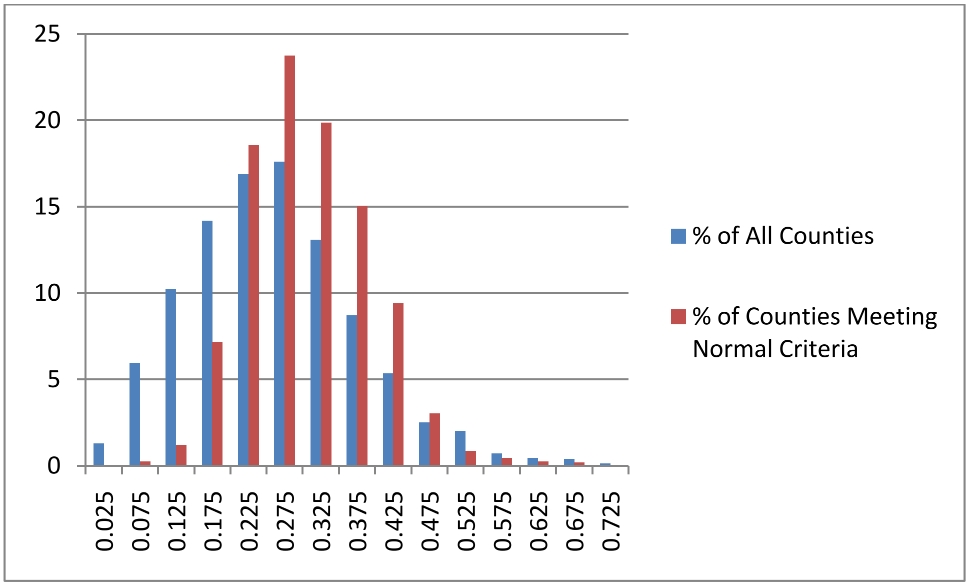
Exhibit 5: QQ Plot of County Means, Counties Meeting Normality Criteria



Note: N=1158, μ=.294, σ=.085.

Exhibit 6 depicts a histogram of county weighted means for all counties with 0<mean<1, and those meeting the normality criteria. The height of the histogram equals the percentage of counties in each category. The categories have width .05, labeled by their midpoints. Very few counties have means above .8; these categories are not depicted. The distribution for all counties is skewed left, with a thick lower tail. The distribution for counties meeting the normality criteria is more normal.

Exhibit 6: Histogram of County Means, All Counties and Those Meeting Normality Criteria



N=2490 for all counties (with 0<mean<1), and 1158 for subsample meeting normality criteria.

Exhibit 7 reports summary statistics for the 1158 counties meeting the normality criteria. For the remainder of the study, my analysis is restricted to this subsample. Responses range from 4 (Hodgeman County, KS for example) to 6282 in Los Angeles County. Mean responses equal 339.332 with a standard deviation of 401.017. Weighted responses range from 30.394 to 19,546.916, with a mean of 363.221 and standard deviation of 921.702.

Exhibit 7: County Summary Statistics, Counties Meeting Normality Criteria

Variable	Minimum	10th Percentile	Median	Mean	Standard Deviation	90th Percentile	Maximum
Responses	4.000	60.000	247.000	339.332	401.017	667.000	6282.000
Weighted Responses	30.394	51.584	133.663	363.221	921.702	815.050	19546.916
Weighted Mean Crime Problem	0.051	0.192	0.285	0.294	0.085	0.407	0.671
Standard Error of the Mean	0.012	0.020	0.032	0.042	0.030	0.075	0.268

N=1158.

Weighted mean crime measures range from .051 to .671, with a mean of .294 and STD of .085. The means reported are a simple average for the "normal" subsample of counties. Thus the mean of the county weighted means (.294) does not equal the grand weighted mean of .338 reported in Exhibit 3. Standard errors range from .012 to .268, with a mean of .042 and standard deviation of .030.

IV Bayesian Hierarchical Model

I employ a Bayesian hierarchical model adopted from Gelman et al. (2004: 131-135). Top level hyperparameters are used as priors for generating county posterior estimates of mean household crime indicators.

Top Level Priors

I use two prior hyperparameters M for the mean. The first prior of .315 is the AHS mean for voucher households. The second is the AHS mean for all rental units equal to .221. The HCV prior is slightly larger than the mean of the county means (.294). The alternative prior based on all rental units is considerably smaller.

My goal is to generate posterior estimates of county crime risk. The CSS is a survey of HCV households. Thus estimates using the HCV prior yield estimates solely based on HCV households. I consider an alternative broader prior in order to generate posterior estimates more representative of the general county population. In fact, one could argue that the mean for all AHS households (including owner-occupied households) is the most appropriate prior. Yet the overall AHS mean of .151 is so far from most of the CSS county means that I deem it unsuitable. As such, my posterior estimates are based solely on crime perceptions of renters.

For less restrictive prior variance than the actual AHS estimates, I use a common standard deviation T=.065 for both prior means. This common standard deviation is 2.355 times greater than the AHS HCV standard error, and 17.105 times greater than the AHS standard error for all renters.

While less informative than the actual AHS variance estimates, T is fairly close to the average of the CSS county crime measure standard errors (.042). The CSS county average could be used as an informative prior measure of variance. Yet this empirical Bayes approach uses the CSS data twice, ignoring prior information from the AHS. A non-informative uniform prior could also be used as a much more flexible prior. T is a fairly informative prior measure of variance loosely based on the AHS. It balances the desire for independent prior information with the conflicting goal of model flexibility.

Conditional Posterior Distributions

Conditional on M and T, the county mean crime measures are assumed to be independent draws from a Normal(M, T) distribution. This is admittedly an oversimplification. A more realistic model might use regression to adjust for socioeconomic factors.

Given M and T, the posterior distribution for county c is Normal(μ_c*, ρ_c*) with ρ_c*= $[1/T^2 + n_c/s^2]^{-.5}$ and μ_c*=$[M/T^2 + m_c n_c/s^2]\rho^{*2}$, where m_c equals the CSS county weighted mean, and n_c equals county weighted responses. S represents the overall HCV population standard deviation from which the CSS responses were drawn.

I assume a value of .473 for s. This equals the weighted standard deviation of the CSS binary crime indicator computed using the standard formula, ignoring the survey nature of the data. Using the standard error of the CSS mean equal to .002 would give the CSS enormous weight relative to the AHS for most counties. Thus there would be little reason to compute Bayesian posterior estimates.

The posterior county mean μ_c* is a weighted average of prior mean M and sample county mean $m_{c.}$ The weights are the respective precisions (inverse variances) $1/T^2 = 1/.065^2$ =236.686 and n_c/s^2. For the typical county with 363.221 weighted responses, precision n_c/s^2 would equal $363.221/.473^2$ =1623.487. The county mean would receive (1623.487/236.686) or 6.895 times the weight of the prior mean when computing μ*.

For counties with the fewest responses, the AHS dominates the CSS. Holmes County, MS has 31.725 weighted responses and a CSS mean of .366. The CSS mean receives .599 the weight of the AHS prior when computing μ*. The posterior mean based on the HCV prior of .315 is .334, and μ*based on the all renter prior of .221 is .275.

Los Angeles County has 19,547 weighted responses. The CSS receives 369 times as much weight as the AHS. Rounding to three decimal places, both posterior means equal the CSS county mean of .353.

IV Posterior Estimates

Exhibit 8 reports summary statistics for posterior estimates of mean μ* under both priors, and common standard deviation ρ*. For comparison, data for the CSS county means and standard errors are reprinted from Exhibit 7. The distributions of posterior means are more compact than the CSS county means. The CSS sample means range from .051 to .671. Posterior mean estimates under the HCV prior of .315 range from .086 to .617. Posterior means under the broader prior of .221 range from .073 to .615.

Exhibit 8: Posterior Summary Statistics

Variable	Minimum	10th Percentile	Median	Mean	Standard Deviation	90th Percentile	Maximum
CSS Mean Crime Problem	0.051	0.192	0.285	0.294	0.085	0.407	0.671
Standard Error of the CSS Mean	0.012	0.020	0.032	0.042	0.030	0.075	0.268
Posterior Mean with HCV Prior=.315	0.086	0.232	0.296	0.304	0.064	0.385	0.617
Posterior Mean with All Rental Prior=.221	0.073	0.202	0.264	0.277	0.069	0.369	0.615
Posterior Standard Deviation	0.003	0.016	0.035	0.033	0.011	0.046	0.052

N=1158.

Posterior estimates of standard deviation ρ* are smaller on average than the CSS standard errors. Posterior standard deviations average .035, compared to .042 for the CSS standard errors.

Exhibit 9 depicts kernel density plots for the CSS county means (labeled “Mean”), posterior means with the HCV prior (labeled “mu_v”), and posterior means with the all rental prior (labeled “mu_r”). Compared to estimates based only on CSS data, Bayesian estimates have more counties near the center of their distributions.

Exhibit 9: Kernel Density Plot of County Mean Estimates

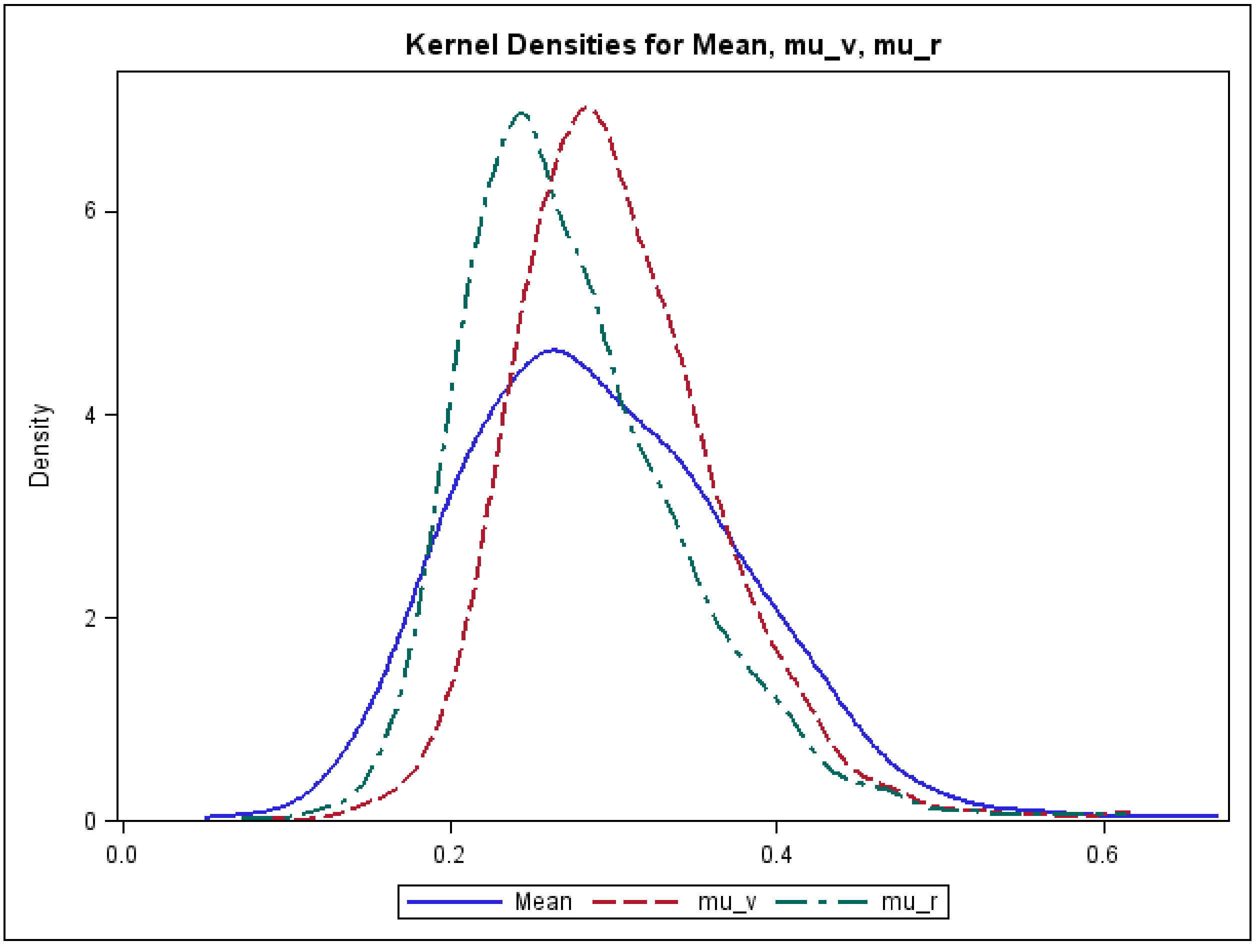


Exhibit 10 depicts the prior, sample, and posterior distributions for Clackamas County, OR under the HCV prior mean of .315. Clackamas was chosen because weighted responses=365.2 are closest to average. The CSS county weighted mean is .246 with standard error of .033. The 95 % confidence interval is .181 to .310, with width .129. The posterior distribution has mean μ*=.254 and standard deviation ρ*=.023. The posterior 95 % credible interval is .209 to .300, with width .091.

Exhibit 10: Clackamas County, OR Estimates with HCV Prior Mean=.315

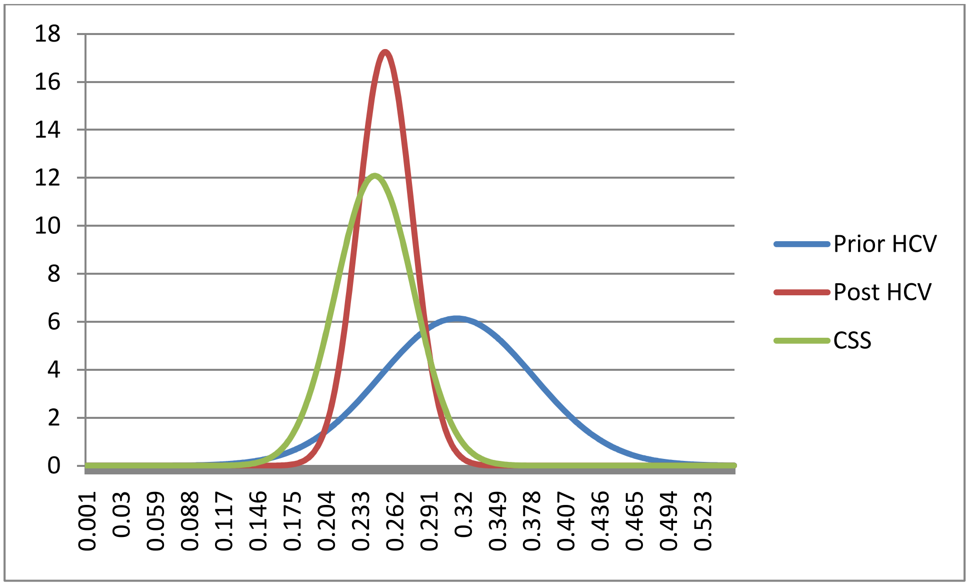
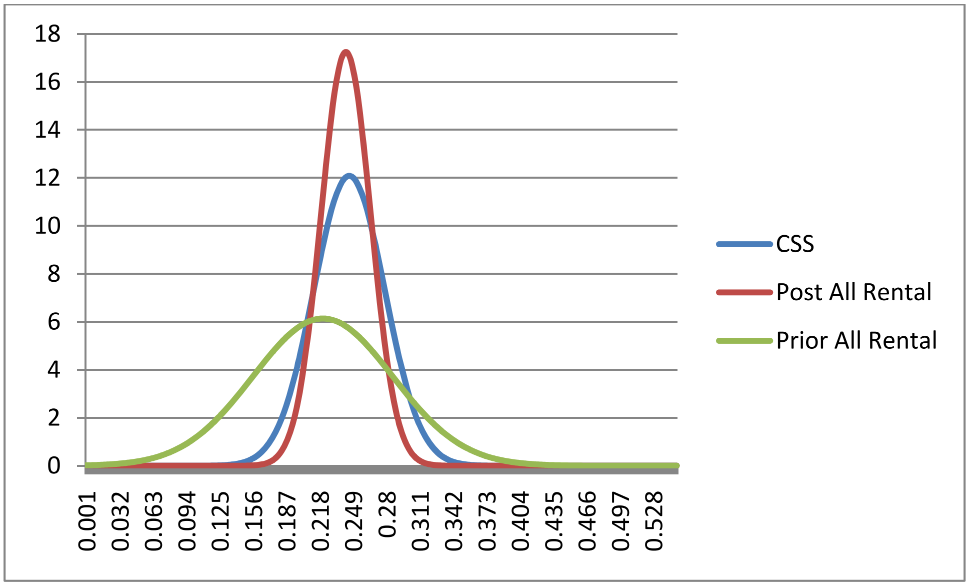


Exhibit 11 depicts the analogous distributions under the prior mean of .221 for all rental units. The posterior distribution has mean μ*=.242 and standard deviation ρ*=.023. The posterior 95 % credible interval is .197 to .288, with width .091.

Exhibit 11: Clackamas County, OR Estimates with All Rental Prior Mean=.221.



V Crime Perception Estimates and Crime Rates

In this section, I compare the three county mean survey estimators with published county property and violent crime rates per 10,000 population averaged over 2000-2002. Crime rate data are from FBI

Uniform Crime Reports[4]. Summary statistics are reported in Exhibit 12 for the 1076 counties with available data. Violent crimes per 10,000 population average 39.613, with a standard deviation of 28.067. The property crime rates average 350.462, with a standard deviation of 150.756.

Exhibit 12: County Crime Rate Summary Statistics

Crime Rate per 10,000 population	Min	10th Percentile	Median	Mean	Std	90th Percentile	Max
Violent	1.449	11.697	31.881	39.613	28.067	76.095	225.869
Property	42.784	180.875	327.507	350.462	150.756	559.883	1242.172
N=1076							

Hipp (2007) studies the relationship between AHS household crime perceptions and county crime rates. He finds household perceptions of crime are more strongly related to violent crime than property crime. A simple correlation analysis of CSS household data with county crime rates confirms this result. The Pearson correlation coefficient of the CSS crime indicator with county violent crime is .129, versus .089 with property crime. Both are highly significant.

This relationship can vary locally. For instance, Mast (forthcoming), using CSS data, estimates that West Virginia crime perceptions relate more strongly with property crime.

Of course, these results could be an artifact of crime rate measurement error. Violent crimes (other than rape) may be reported more consistently to police. In high violent crime areas, police and prosecutors may not take property crime as seriously. If victims believe property offenders are less likely to be apprehended and punished, they may be less likely to report. Accordingly, reporting rates for property crime may vary more than reporting rates for violent crime.

Regardless, I validate I my county mean estimates by comparison with violent and property crime rates. Exhibit 13 reports Pearson correlation coefficients of my three county mean survey estimators with property and violent crime rates. Consistent with Hipp (2007), estimated correlation is higher with violent crime than with property crime. All coefficients are significant at the .0001 level.

Exhibit 13: Pearson Correlation Coefficients

	CSS Mean	Posterior Mean - HCV Prior	Posterior Mean - All Rental Prior
Property Crime Rate	0.3655	0.3812	0.4243
Violent Crime Rate	0.3770	0.4136	0.4452

N=1076

More germane to this study, for both property and violent crime, correlation coefficients are lowest for the estimator based solely on CSS data. Correlation is highest for the Bayesian estimates with a prior based on all AHS rental units. Along with evidence of less county variance (shrinkage), this may imply that the Bayesian hierarchical model produces more reliable local estimates.

VI Conclusions

This study uses crime perception survey data from two sources: the American Housing Survey (AHS) and HUD's Customer Satisfaction Survey (CSS) of Housing Choice Voucher households. National AHS data and county CSS data are used to estimate a Bayesian hierarchical model of local crime perceptions.

Results indicate the Bayesian approach yields more robust local estimates. Compared to estimates solely based on CSS data, the Bayesian estimates have lower variance and correlate more highly with published county crime rates.

Estimates are subject to an assumption of known prior mean and variance. A more thorough analysis would account for uncertainty in these hyperparameters. This is an interesting avenue for further research.

References

Gelman, Andrew et al. *Bayesian Data Analysis: 2nd Edition*, 2004 Chapman and Hall/CRC, London.

Hipp, John. Resident Perceptions of Crime: How Similar are They to Official Crime Rates? Center for Economic Studies, U.S. Census Bureau, Working Papers. December, 2007 http://www.ces.census.gov/index.php/ces/cespapers?detail_key=101783 .

Mast, Brent D., "Measuring Housing Quality in the Housing Choice Voucher Program with Customer Satisfaction Survey Data" *Cityscape: A Journal of Policy Development and Research Vol. 11(2)* July, 2009 http://www.huduser.org/periodicals/cityscpe/vol11num2/index.html .

[1] AHS data and information are available at http://www.huduser.org/datasets/ahs.html .

[2] For more information on the CSS, see Mast (2009).

[3] SAS version 9.2 Proc SurveyMeans was used to compute standard errors. It uses the same formula for standard errors of the sample mean and sample proportion (http://support.sas.com/documentation/cdl/en/statug/59654/HTML/default/statug_surveymeans_a0000000218.htm.).

[4] http://www.fbi.gov/ucr/ucr.htm .