

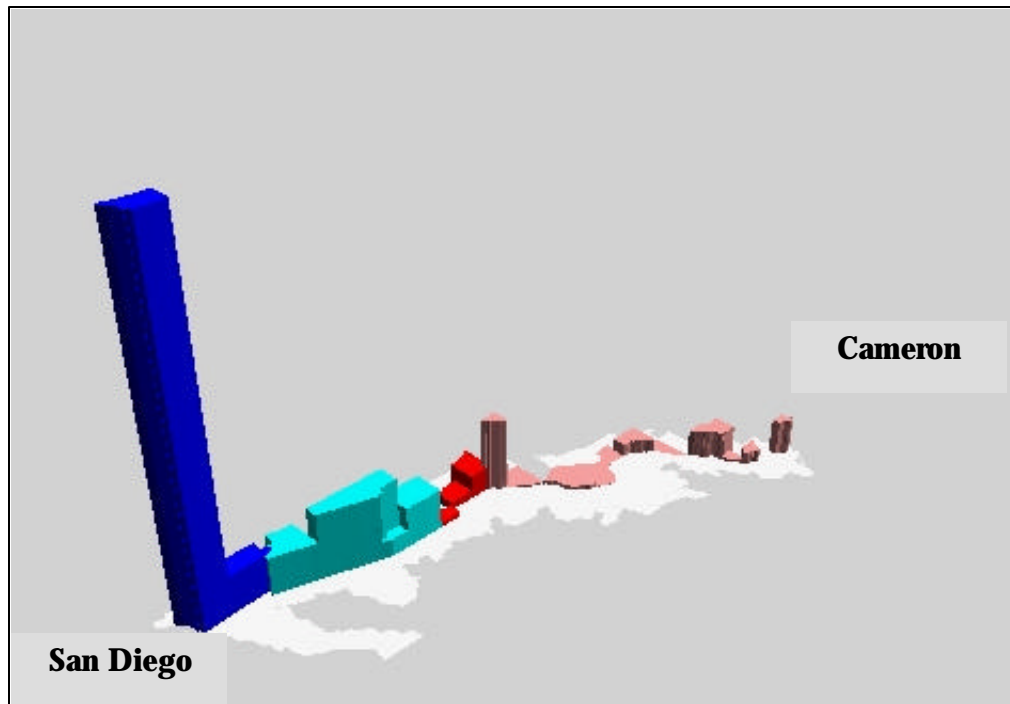
A PREDICTIVE MODEL

A predictive model has been generated as part of the effort to produce a protocol for estimating the impact of criminal illegal aliens on the law-criminal justice system. This model is based on a least-squares regression approach in which calculations of the impact on each county are treated as the dependent variable, with a variety of predictor variables assessed as to what combination of those variables would best predict the impact as determined by this study. If a combination of variables seems to fit the data well, then a first step has been taken toward a model that might be used on an annual basis to: (a) estimate the impact for each county without having to replicate the more extensive and expensive personal interview and analysis that was implemented in this study; and (b) estimate the impact on counties for which data are not otherwise available.

Data and Methods

The dependent variable in the analysis is thus the estimate of the fiscal impact (in Fiscal Year 1999) of the processing of criminal illegal aliens, reflecting the costs associated with county functions, including the Sheriff's department, detention, District Attorney, Public Defender, Probation, Coroner, and the Courts. These costs include a calculation for general government (overhead) costs associated with the operation of each of those county functions. Emergency medical costs were not included. The criminal justice-law impact varies from a high of \$39.4 million in San Diego County, California to a low of zero impact in Terrell County, Texas. No data were available for estimation in Maverick County, Texas, so it is excluded from the analysis. Figure P1 illustrates the spatial variability in the total dollar impact of processing criminal illegal aliens in each county along the length of the border.

Figure P1. Total Dollar Fiscal Impact of Processing Criminal Illegal Aliens, by County



More populous counties, such as San Diego, Pima, and El Paso, can be expected to have a larger absolute impact due to the larger economies and thus the larger draw of illegal immigrants. In general the counties toward the west are more populous than those in the east, and there has been a general trend for illegal migration to be directed to the western counties.

The predictor variables included in the initial model include general crime data, border patrol data, demographic information, geographic information, and local fiscal data, as shown below in Table P1. Only those variables that are statistically significantly related to the estimate of fiscal impact are included in the final model.

The general model is that the fiscal impact of processing criminal illegal aliens is a function of the amount of criminal activity in a county, the volume of apprehensions of illegal aliens by the border patrol, the number of ports of entry, the number of legal crossings at those ports of entry, the population on the U.S. side of the border, the percent of the population in the U.S. county that is Hispanic, the per person income in the U.S. border county, the population on the Mexican side of the border at each U.S. county, the percent of the population within 10 kilometers of the border on the U.S. side, the approximate miles of border that each county shares with Mexico, and the size of the county's general fund. Also, because the budgeting for law enforcement and criminal justice expenses varies from state to state, a set of dummy variables will be introduced into the model to reflect the state in which each county is located, with Texas omitted as the reference variable.

Table P1. Variables in the Analysis

Predictor Type	Variable	Abbreviation	Source
General crime	Total Part I FBI crimes in 1998-99	ARREST	FBI Uniform Crime Reports
Border patrol	Apprehensions of undocumented aliens 1999	APREHND	USDOJ, INS
	Number of ports of entry	PORTS	USDOJ, INS
	Number of crossings at ports of entry	CROSS	USDOJ, INS
Demographic	Population as of 1 January 1999	POP_1_99	State population units ¹
	Percent of the population that was Hispanic as of 1999	PCT_HISP	US Census Bureau
	Per person income	INCOME	US Census Bureau
	Mexican municipio population across the border as of March 2000	MEXPOP	INEGI, Mexico Census 2000
Geographic	Percent of county population within 10km of the border	W_IN10KM	Estimates by SDSU International Population Center based on road density
	Approximate miles of border with Mexico ²	BRDR_MI	Calculations by SDSU International Population Center
Fiscal	Total County General Fund	GEN_FUND	Individual counties
Dependent (1)	Fiscal Impact of Processing Criminal Illegal Aliens	TOTIMPCT	Research Team
Dependent (2)	Per Person Impact of Processing Criminal Illegal Aliens	PERIMPCT	Research Team

¹In California: *The Demographic Research Unit of the State of California Department of Finance*; in Arizona: *The Population Unit of the State of Arizona Department of Economic Security*; in New Mexico: *The Bureau of Business and Economic Research at the University of New Mexico*; in Texas: *The Texas State Data Center, Department of Rural Sociology, Texas A & M University.*

² Note that this calculation may not be exactly the same as mentioned in each county's analytic chapter.

The population in the contiguous Mexican *municipios* is not always exactly known because there is overlap between some *municipios* and some counties. Thus, a Mexican municipio may share a border with two or more U.S. counties. The population that was adjacent to the U.S. county was calculated on the assumption that the distribution of population on the Mexican side of the border was similar to the distribution on the U.S. side of the border. For example, the Mexican *municipio* of Juarez, in the state of Chihuahua, is contiguous to both El Paso County, Texas, and Doña Ana County, New Mexico. So, the population of Juarez that was estimated to be contiguous to El Paso County was calculated as the ratio of El Paso's population to the population of El Paso and Doña Ana combined, multiplied by the 2000 census enumeration of population in Juarez. A comparable calculation was performed to estimate the population of Juarez that was contiguous to Doña Ana County.

The percent of each county's population residing within 10 kilometers of the border was calculated indirectly in the ArcView geographic information system (GIS) by assuming a one-to-one relationship between street length and population density. Thus, the total street length in a county's non-federally owned territory was calculated, and then the proportion of that total that lay within 10 kilometers (6 miles) of the border was calculated, and estimated to be equal to the proportion of the population within that distance of the border. The length in miles of border shared by each county with Mexico was also calculated within the ArcView GIS environment. The value of each variable for each county is presented in the Appendix Table AP1.

The basic regression model to be tested is shown below, where the model predicts the total dollar impact (TOTIMPCT):

$$\text{TOTIMPCT} = a + b_1 \cdot \text{ARREST} + b_2 \cdot \text{APREHND} + b_3 \cdot \text{PORTS} + b_4 \cdot \text{CROSS} + b_5 \cdot \text{POP-1_99} + b_6 \cdot \text{PCT_HISP} + b_7 \cdot \text{INCOME} + b_8 \cdot \text{MEXPOP} + b_9 \cdot \text{W_IN10KM} + b_{10} \cdot \text{BRDR_MI} + b_{11} \cdot \text{GEN_FUND} + b_{12} \cdot \text{CALIF (dummy)} + b_{13} \cdot \text{ARIZ (dummy)} + b_{14} \cdot \text{NM (dummy)} + e$$

Where a is the constant y -intercept and e is the residual error term.

Results for Total Dollar Impact

Table P2 shows the zero-order (bivariate) correlation coefficients between the total dollar impact and each of the predictor variables, as well as the correlations among the predictor variables themselves. The data show that there are several interrelated variables that are highly correlated with the fiscal impact on a county of processing criminal illegal aliens. The size of the general fund, the annual number of felony arrests in the county, per person income, being in California, and the size of the population in the bordering Mexican *municipio* are the most important—all with correlation coefficients of 0.89 or higher. The percent Hispanic and the number of Border Patrol apprehensions in the county are also related to the impact with correlation coefficients higher than 0.70. The other statistically significant correlations are with the number of border crossers, the population within 10 kilometers of the border, the number of miles of border shared with Mexico, and being in Arizona.

Table P2. Correlations Among Variables

	TOT IMPCT	ARREST	APREHND	PORTS	CROSS	PCT_HISP	INCOME	MEXPOP	W_IN10KM	BRDR_MI	GEN_FUND	CALIF	ARIZ	NM
TOTIMPCT	1.000	.988	.755	.126	.683	-.767	.920	.894	-.534	-.476	.997	.946	-.398	-.197
ARREST		1.000	.697	.127	.633	-.803	.946	.850	-.574	-.392	.987	.915	-.325	-.228
APREHND			1.000	.006	.484	-.640	.692	.675	-.462	-.308	.754	.829	-.194	-.267
PORTS				1.000	.694	.370	-.106	.398	.599	-.569	.098	.119	-.614	-.440
CROSS					1.000	-.154	.451	.894	.144	-.753	.649	.660	-.750	-.317
PCT_HISP						1.000	-.932	-.505	.842	-.107	-.769	-.685	-.230	.076
INCOME							1.000	.744	-.708	-.151	.915	.845	-.069	-.172
MEXPOP								1.000	-.174	-.684	.866	.869	-.624	-.208
W_IN10KM									1.000	-.302	-.568	-.522	-.325	-.148
BRDR_MI										1.000	-.463	-.494	.916	-.120
GEN_FUND											1.000	.952	-.392	-.184
CALIF												1.000	-.457	-.176
ARIZ													1.000	-.086
NM														1.000

Coefficients in **bold** are statistically significant at or beyond the .05 level

It can also be seen in Table P2 that there are a number of significant intercorrelations among the predictor variables that will have to be dealt with. However, the initial model is run with all variables entered into the equation, and the results are displayed in Table P3.

Table P3. Results of Regression Analysis - Initial Model*

	Unstandardized Coefficients		Standardized Coefficients	t-score	Significance
	B	Standard Error	Beta		
(Constant)	2975750.095	2693430.470		1.105	.293
ARREST	-71.307	529.223	-.062	-.135	.895
APREHND	7.264	4.665	.069	1.557	.148
PORTS	-362899.874	410476.974	-.060	-.884	.396
CROSS	.01276	.029	.031	.438	.670
POP_1_99	5.785	8.141	.435	.711	.492
PCT_HISP	-27802.167	26521.106	-.072	-1.048	.317
INCOME	-90184.295	105178.525	-.041	-.857	.409
MEXPOP	3.042	1.578	.128	1.927	.080
W_IN10KM	15520.125	18118.010	.039	.857	.410
BRDR_MI	3373.928	4933.907	.019	.684	.508
GEN_FUND	10049.830	3830.342	.507	2.624	.024

*Ordinary Least Squares (OLS), all variables entered into the equation
 Dependent Variable: TOTIMPCT impact of law-justice costs; $R^2 = .99$

The regression results in Table P3 indicate that the combination of these eleven predictor variables explains 99 percent of the overall variation among counties in the total dollar impact of the law-justice costs of processing illegal aliens. However, only one of the eleven variables—the size of the county’s general fund—is statistically significant at the .05 level. There are several problems with this model: (1) the disproportionate size of San Diego County means that its variance dominates the model; (2) the large number of variables (11) relative to the number of counties (23) obfuscates the results; and (3) the high level of intercorrelation among the predictor variables complicates the results.

The disproportionate size of San Diego County is dealt with initially by implementing a weighted least-squares model (WLS) in which the population size of each county is withdrawn as a predictor variable and is used instead as a weighting factor. Each data point is thus weighted by the reciprocal of its variance. In this way, observations with large variances (such as San Diego) have less impact on the analysis than do observations associated with small variances (such as most counties in New Mexico and Texas). The results of this regression are shown in Table P4.

**Table P4. Results of Regression Analysis – Revised Initial Model
Weighted by County Population Size***

	Unstandardized Coefficients		Standardized Coefficients	t-score	Significance
	B	Standard Error	Beta		
(Constant)	10256220.734	3101752.802		3.307	.006
ARREST	401.882	192.375	.311	2.089	.059
APREHND	5.161	4.275	.020	1.207	.251
PORTS	-1423637.523	513277.894	-.065	-2.774	.017
CROSS	.06098	.017	.080	3.660	.003
PCT_HISP	-64366.779	19314.965	-.098	-3.332	.006
INCOME	-338949.201	193713.084	-.097	-1.750	.106
MEXPOP	2.630	1.049	.079	2.506	.028
W_IN10KM	40461.765	11728.066	.050	3.450	.005
BRDR_MI	-1038.094	6368.992	-.002	-.163	.873
GEN_FUND	11077.577	2301.494	.603	4.813	.000

*Weighted Least Squares Regression - Weighted by POP_1_99; all variables entered into the model
Dependent Variable: TOTIMPCT impact of law-justice costs; $R^2 = .99$

This revised model also explains virtually all of the variability from county-to-county in the total dollar impact, but the weighting scheme produces a set of additional statistically significant predictors: general fund, the number of border crossers, the percent Hispanic, the population within 10 kilometers of the border, the number of ports of entry, and the size of the population on the Mexican side of the border. Nonetheless, there remains a high level of multicollinearity.

There is no completely satisfactory solution to the multicollinearity problem, but the usual solution is to drop the offending variables. In order to decide which variables to drop, it was necessary to quantify the exact nature of the multicollinearity. This was accomplished through the use of principal components factor analysis. In this procedure, those variables that vary together are clustered together into distinct components. From each of the distinct components, one of the variables with a high factor loading was chosen to be included in the revised model. The results of the factor analysis are shown in Table P5. Four components emerged from the factor analysis, and cumulatively these four account for 80 percent of the overall variation among all variables. The first component is the most important and it can be seen that five variables load very high on this component. The Mexican population, the number of Part I arrests, the county general fund, being in California, and the number of border crossers all have component coefficients higher than 0.80. Following a strategy of choosing variables that are easiest to obtain and update, the general fund variable is the obvious choice among these variables for both methodological and conceptual reasons and it is shown in bold to reflect its choice for the revised regression model. The rationale is that the greater is the size of the county's general fund, the more money there is available to spend

on processing criminal illegal aliens and, at the same time, the larger is the budget, the larger is the economy, indicating greater opportunity for criminal activity.

Table P5. Results of Factor Analysis

	Component			
	1	2	3	4
MEXPOP	.926	.263		
ARREST	.899			
GEN_FUND	.875	-.304		
CALIF	.817			
CROSS	.804	.469		
W_IN10KM		.870		
PCT_HISP	-.291	.837	-.325	
PORTS	.607	.658	.262	
INCOME	.566	-.621		
ARIZ			.852	-.231
BRDR_MI			.741	
APREHND	.542		.564	-.202
NM				.936

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization. Rotation converged in 8 iterations.

The second component has only two variables with coefficients higher than 0.80—the population within 10 kilometers of the border and the percent of the population that is Hispanic. The percent of the population within 10 kilometers of the border is chosen as the representative variable in this group, on the rationale that the more people there are close to the border, the greater will be the opportunities for criminal activity on the part of people illegally crossing the border.

The third component has only one coefficient that is greater than 0.80—being in Arizona. Thus, that variable is included in the revised model. The fourth component also has only one variable with a high coefficient—being in New Mexico. Therefore, that variable is included in the revised model. Table P6 shows the results of this revised model.

Table P6. Results of Regression Analysis – Revised Model Using Results of Factor Analysis

	Unstandardized Coefficients		Standardized Coefficients	t-score	Significance
	B	Standard Error	Beta		
(Constant)	105905.249	527256.609		.201	.843
GEN_FUND	19761.024	701.676	.997	28.163	.000
W_IN10KM	37926.447	14088.880	.095	2.692	.014
ARIZ	1742010.292	734300.559	.083	2.372	.028

Dependent Variable: TOTIMPCT impact of law-justice costs; R² = .98

Step-wise regression was used in this model to eliminate variables that were not statistically significant. In this model, being in New Mexico was not statistically significantly related to impact and so it dropped out of the model. The results show that the size of the general fund emerges as the clearly most important predictor variable. The three variables of general fund, population within 10 km of the border and being in Arizona combine to explain 98 percent of the variation in county

impact, but general fund alone explains 96 percent of that variability. The weighted regression results were essentially the same and are not shown. Of some interest is the fact that even after the various adjustments that have been made, the revised model shown in Table P6 produces the same conclusion as did the original model shown in Table P3—the total dollar impact is related to the size of a county’s general fund.

There remains a concern that the size of San Diego County is dominating the results and the residuals reflect the fact that the absolute level of predicted impact is not nearly as good for the other counties as it is for San Diego. For this reason, San Diego County was deleted from the model and the factor analysis and the regression results were recalculated without San Diego. The results of the factor analysis are shown in Table P7.

Table P7. Factor Analysis Excluding San Diego County

	Component				
	1	2	3	4	5
PORTS	.904				
CROSS	.869			.239	
MEXPOP	.843			.390	
ARREST	.706	.596		-.201	
W_IN10KM	.680	-.504			.211
INCOME		.833			
PCT_HISP	.429	-.779	-.255		
GEN_FUND	.540	.693	.266		
BRDR_MI			.811		
ARIZ		.366	.750		.315
CALIF	.213			.875	
APREHND	.204		.477	.740	
NM					-.943

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization. Rotation converged in 16 iterations.

The results are considerably different when San Diego is removed. The size of the general fund no longer emerges as an overwhelmingly important variable. In component 1, the number of ports is the single most important variable, whereas in component two, the per person income is most important. In component 3 the number of miles of border shared with Mexico is most important, while in component 4 being in California emerges as most important. However, since there is only one county in California after removing San Diego, that variable was not used. Rather, the number of Border Patrol apprehensions was taken as the representative variable from component 4. In component 5, not being in New Mexico was the significant variable. Note that two variables—general fund and arrests—load fairly high (coefficients of at least 0.50) on more than one factor. This overlap was consistent across several different factoring methods, indicating that it is something that is inherent in the data. Because of that, and because of its importance in the previous model, the size of the general fund was included in the revised regression model which excludes San Diego County. Arrests were also included since they couldn’t otherwise be excluded as being potentially important on their own. Table P8 shows the results of the regression analysis, excluding San Diego County.

Table P8. Regression Analysis Excluding San Diego County

	Unstandardized Coefficients		Standardized Coefficients	t-score	Significance
	B	Std. Error	Beta		
(Constant)	435480.836	260705.230		1.670	.111
ARREST	575.949	65.077	.770	8.850	.000
APREHND	12.555	2.749	.397	4.566	.000

Dependent Variable: TOTIMPCT impact of law-justice costs; R² = .85

The data in Table P8 show that two variables emerge as influencing the total dollar impact on counties, when San Diego is excluded from the data set. These two variables are the number of Part I arrests in the county (an index of crime overall) and the number of Border Patrol apprehensions (an index of the volume of illegal immigrants). Together they explain 85 percent of the county-to-county variation in total dollar impact of processing criminal illegal aliens. Of these two variables, the number of arrests is more important than the number of apprehensions, but both variables make important contributions to the impact. The model predicts the following impact:

$$\text{TOTIMPCT} = \$435,480 + (\$576 * \text{ARREST}) + (\$13 * \text{APREHND})$$

The constant indicates a “bottom-line” of \$435,480 for any border county to process criminal illegal aliens, and then it goes up from there depending upon the number of arrests and the number of apprehensions. Each arrest averages an additional \$576 in impact, whereas each Border Patrol apprehension averages an additional \$13 in impact.

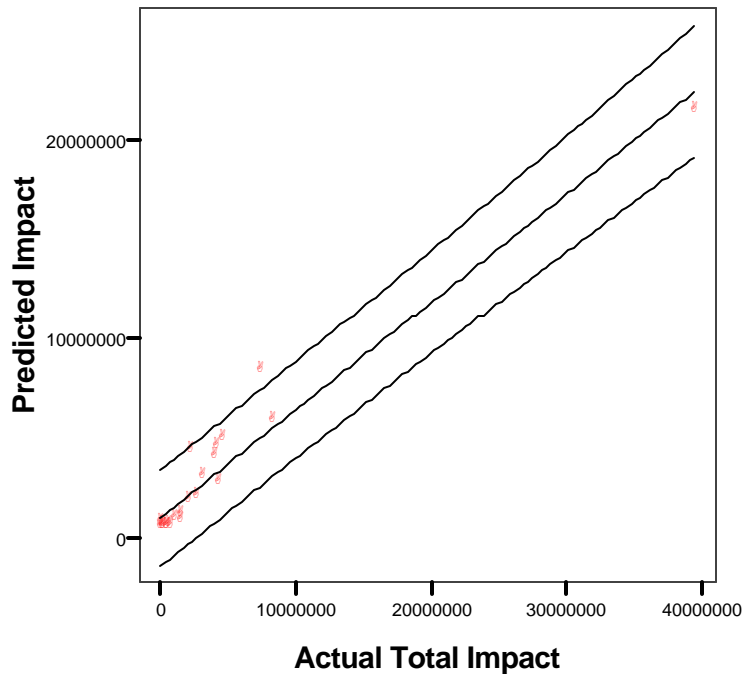
Table P9 shows the actual impact calculated for each county by the researchers compared to the values predicted by the above regression model, and it shows the absolute residual (the difference between actual and predicted) for each county, as well as the standard residual (the size of the residual for this county compared to the average residual for all counties). Also calculated is the absolute ratio of the actual impact (TOTIMPCT) to the predicted impact. Although San Diego County was not in the data that produced this regression model, the predicted value for San Diego was calculated and is included in the table.

If the regression model were a perfect fit, then there would be no difference between the actual and the predicted values. As can be seen in Table P9, nine of the twenty-three have predicted values that are quite close to the actual values (defined as between 0.75 and 1.25 of the actual value and shown in bold in the table). The smaller counties tend to have an impact that is less than the constant in the equation (\$435,480) and so the predicted values are too high. On the other hand there are two counties—Starr and San Diego—which have considerably higher relative impacts than would be predicted by the model. Figure P2 shows the scatterplot of the actual impact costs compared to the predicted impact, and the data show a good fit to the model.

Table P9. Actual and Predicted Total Impact, by County

COUNTY	Std. Residual	TOTIMPCT Impact of law -justice costs	Predicted Value	Residual	Ratio of Actual to Predicted
Cameron	-0.024	3,928,294	3,950,628	-22,334	0.99
Hidalgo, TX	-2.299	2,162,102	4,292,049	-2,129,947	0.50
Starr	0.747	1,440,443	748272	692,170	1.93
Zapata	-0.122	345,574	458,338	-112,764	0.75
Webb	0.169	3,121,601	2,965,245	156,355	1.05
Maverick	.	.	1,577,387	.	
Kinney	-0.716	13,673	676,988	-663,315	0.02
Val Verde	0.459	1,445,587	1,020,626	424,960	1.42
Terrell	-0.490	0	454,356	-454,356	0.00
Brewster	-0.467	39,425	472,088	-432,663	0.08
Presidio	-0.015	453,293	466,911	-13,618	0.97
Jeff Davis	-0.431	40,801	440,088	-399,287	0.09
Culberson	0.136	610,104	483,882	126,221	1.26
Hudspeth	-0.428	120,525	517,049	-396,524	0.23
El Paso	2.652	8,226,973	5,770,299	2,456,673	1.43
Dona Ana	0.772	2,663,760	1,948,143	715,616	1.37
Luna	0.059	924,866	870,260	54,605	1.06
Hidalgo, NM	-0.062	480,041	537,300	-57,259	0.89
Cochise	-0.444	4,475,828	4,887,438	-411,610	0.92
Santa Cruz	0.197	1,962,711	1,779,834	182,876	1.10
Pima	-1.004	7,302,866	8,233,220	-930,354	0.89
Yuma	1.606	4,190,003	2,702,352	1,487,650	1.55
Imperial	-0.295	4,152,052	4,425,146	-273,094	0.94
San Diego	N/A	39,459,572	21,406,239	18,053,333	1.84

Figure P2. Scatter Plot of the Actual Impact Compared to the Predicted Impact

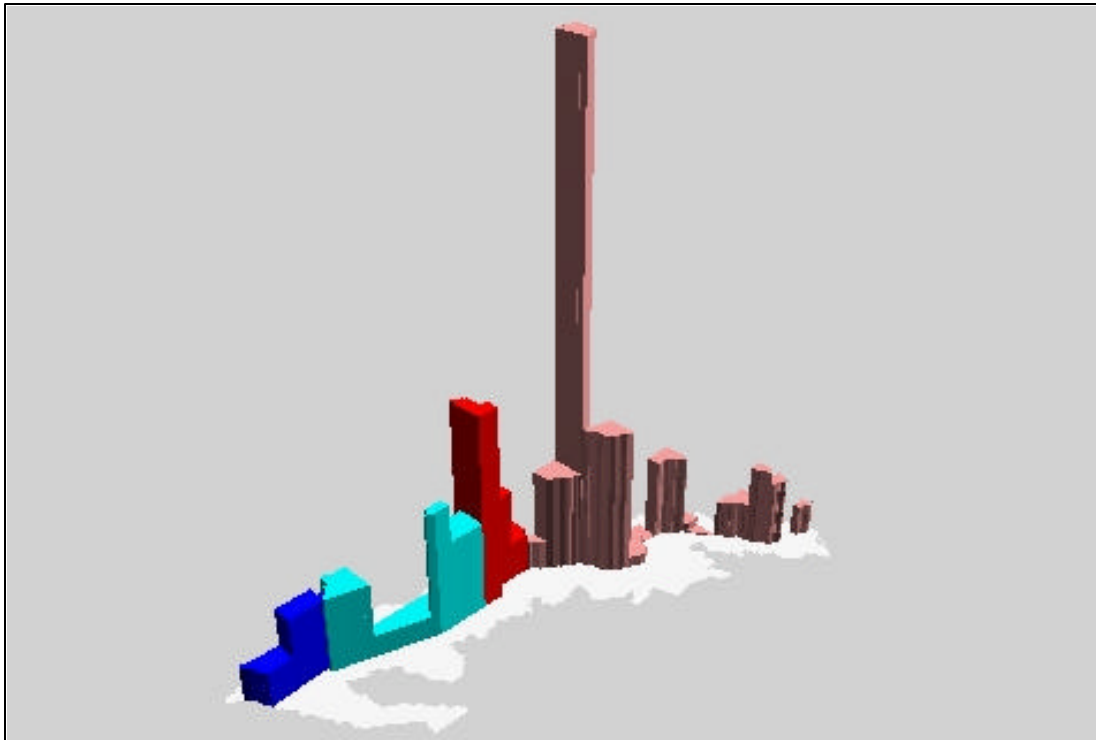


Results Based on Per Person Impact

Another way of viewing the impact is to normalize the impact by the size of the county's population and thus examine the per person impact for each county. These figures show a very different pattern than the total impact by county, as can be seen in Figure P3. Although San Diego has by far the highest total dollar impact, its per person impact of \$13.82 is well below the border county average of \$32.98. The lowest non-zero impact is found in Kinney County, Texas (\$3.90), although Hidalgo County, Texas (\$4.10) and Brewster County, Texas (\$4.31) are also very low. By far the highest impact was found in Culberson County, Texas, where the impact averaged \$194.67 per person. The next closest was Hidalgo County, New Mexico at \$78.69.

The regression model using per person impact as the dependent variable failed to find a statistically significant predictor variable. In other words, none of the variables that were available could produce a statistically significant relationship to per person impact. Because of the unusually high value for Culberson County, that county was removed from the data set and the regression recalculated. However, that did not change the result. The predictor variables used in the analysis apparently do not capture the reasons why per person impact varies from one county to another.

Figure P2. Per Person Impact of Processing Criminal Illegal Aliens, by County



Summary and Conclusion

The results suggest that it is possible to quite accurately model the total dollar law-justice impact of processing criminal illegal aliens in U.S.-Mexico border counties. The disproportionate size of San Diego County means that the results are best for all other counties when San Diego is removed from the data set, but in all counties the combination of Part I arrests and Border Patrol apprehensions explains a very large fraction of the county-to-county variability in the impact. It would be almost unheard of for a statistical model of this type to explain 100 percent of the variation, and the high percentage explained is indicative of a clear pattern of impact. There will always be variability among counties in how resources are allocated and it would be very unusual to be able to predict all such variability.

The data do not support an interpretation of the pattern of variability in per person impact by county, but this is perhaps less important to policy-planners than is the total dollar impact. Nonetheless, the considerable variability in per person impact is indicative of the variation by county in how criminal illegal aliens are dealt with, including variability in the resources available to cope with criminal illegal aliens. This is an issue worthy of additional investigation.

The results of the predictive model represent a baseline study which can provide surrogate measures of the likely impact on each county on an annual basis. However, a study of this type should be replicated on a regular (such as 5-year) cycle to evaluate trends and changes.