Appendix A.i

Colonia Selections and the Statistical Analysis in Detail

This appendix sets forth the details of our methodology. It includes the following: (1) a description of the datasets and methods used to estimate the number of housing units in the colonias in the counties studied; (2) an explanation of how sample sizes were determined in each county and the criteria used in order to select colonias; (3) a description of the types of survey materials gathered (face-to-face and mail-back surveys) and the relative rates of return and survey counts; and (4) the methods used to correct for possible sources of bias such as non-responses, unknown eligibility, and the effects of variances on key variables. From the outset of the research project the team was interested **both** in colonias in the six border counties that the TDHCA had specified for inclusion in the study, as well as wishing to research CFDs in colonias and similar informal homestead subdivisions in non-border counties specifically in locations such as Central Texas. Traditionally the policy focus has almost exclusively on border colonias, but based upon more recent research findings the team resolved to broaden the scope beyond the six named counties (Cameron, Hidalgo, Starr, Webb, Maverick, and El Paso) which are those in the border with the highest colonia populations. However, stepping outside the border remit meant that, for certain counties at least we would not start with the same level of access to baseline datasets such as that of the Office of the Attorney general, such that for some counties we would have to construct comparative datasets.

While the TDHCA agreed to our request to broaden the scope of study beyond the six counties, a requirement was that whatever strategy we adopted should allow us in those six counties to arrive at estimates of CFD usage for colonias across the whole county. And while we were confident that we could gather data from the County Clerks' Office about Recorded Contracts for Deed at the county level, and do so over time (since 1989), we also realized that arriving at estimates about unrecorded CFDs could only be achieved through survey analysis and subsequent triangulation of data records that we found in the surveys. Obviously a survey of all households was not feasible, so a sample survey had to be drawn that would allow us to subsequently **extrapolate** our findings to the county level – within acceptable (or clearly stated) confidence limits. Hence the importance of this Methodological Appendix which sets out in much greater detail how the colonias and subdivisions were selected; how we sought to ensure that would be able to extrapolate statistically from those data; and be transparent about the algorithms and weights that we applied in order to conduct that statistical analysis.

Making an extrapolation invariably requires random selection within a specified sample universe (colonias in each county in this instance), as well as the actual selection of households that fall into the survey. As we outline in detail below, while we adopted the random selection principal on both levels to construct our dataset, we also decided to include a number of colonias and subdivisions that were selected **purposively**: principally newer subdivisions many of which had been developed since 1995 often under "model subdivision rules" (i.e. with basic infrastructure) and where we knew developers were most active since restrictions had been placed on their activities post 1991 and especially post 1995. In the dataset, as well as in many parts of the data analysis, we separate (and compare) the findings between the randomly and purposively selected settlements, but the important point to be underscored here is that **extrapolations to the county level may only be made from the colonias that were selected at random**. The crux of this analysis relates to Chapter four where we seek to estimate the frequency of unrecorded contracts for deed for each of the six border counties.

Estimating the Total Number of Housing Units in Each County's Colonias

The Office of the Attorney General (OAG) operates an online viewer containing information about 2000 colonias.[1] However, the population and lot number figures listed on the viewer and which apparently are provided by the Texas Water Board Development Board, are not updated for all colonias. In addition, the figures in the database relate to information gathered from 2000 census data.

Because we wished to base our estimates on 2010 census data, we searched for another way to estimate colonias populations. From the US-Mexico Border Environmental Health Initiative (BEHI), we were able to download a colonias boundaries shapefile[2] containing information about 1808 colonias, the majority of which (1717) are located in the six selected counties. Table 1, below, shows the numbers of colonias listed in each source by county. An advantage of the BEHI dataset is that it has population estimates up to 2005. We compared a recently-updated list obtained from the Secretary of State (SoS) to this BEHI dataset to identify the 109 colonias missing from the list located in the counties selected for study. Of these, however, only 37 (12 in Starr and 25 in Webb) have boundaries and areas accessible via the OAG viewer. Thus we were able to add these colonias to our BEHI-based dataset of possible colonias for analysis -- creating a database of 1754 colonias (Table 1).

Table 1 Number of Colonias in SOS, BEHI, and OAG Databases By County

Counties	# of Colonias (SOS list)	# of Colonias (BEHI shapefile)	Difference	# Colonias missing in shapefile but located in theOAG viewer	# Colonias with no cartographic information
Cameron	178	174	4	0	4
El Paso	321	296	25	0	25
Hidalgo	934	926	8	0	8
Maverick	75	69	6	0	6
Starr	256	221	35	12	23
Webb	62	31	31	25	6
TOTAL	**1826**	**1717**	**109**	**37**	**72**

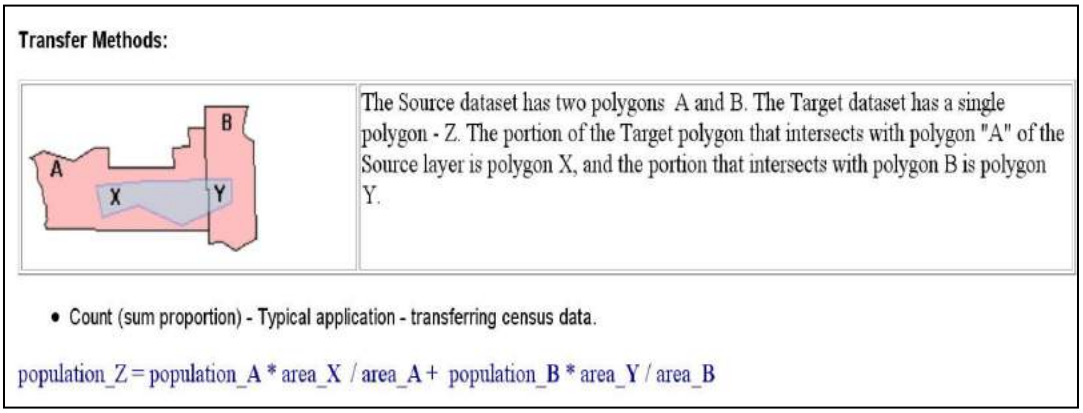
With our BEHI cartographies the OAG cartographic information (from which we derived approximate boundaries using Google Earth™), we then employed a census overlay technique commonly used when boundaries of places of interest do not match exactly. The technique involves an area-based weighting of the target area of analysis, in this case the colonias. Assuming an even distribution of the housing units within each census block, we estimated that the total housing units for colonia j will be given by $\sum n_{ij}$, which is the sum of the estimated number of housing units in all the intersecting areas of the colonia in reference to their corresponding census blocks.[3] This technique is widely available within standard software such

[1] https://maps.oag.state.tx.us/colgeog/colgeog_online.html#

[2] Available online in the following link: http://borderhealth.cr.usgs.gov/datalayers.html. A shapefile is a file format that combines cartographic information with attribute data readable within Arc GIS.

[3] For each of those intersecting areas (n_{ij}), the estimated population will be calculated by:

as Arc GIS extensions like ET Geowizards, through the transfer-data option. An illustration of this technique is provided in Figure Nº 1. Let a polygon z in light blue, composed by the intersecting areas x and y, be the colonia of interest and polygons A and B be the 2010 census blocks. Under the assumption that housing units are evenly distributed in each block, the number of housing units in y will correspond to the proportion of the area that y shares with B multiplied by the total number of housing units in B. The same will be applicable to A. Then, the housing units in the colonia will be the sum of the estimated housing units in x and y in relationship to their corresponding census blocks.



Source: Et Geowizards Help Menu

Figure Nº 1: Illustration of Census Overlay as performed by ET Geowizards

Of course, such calculations may be subject to estimation error: some , high density housing census blocks may lead to an overestimate of the colonia population because the area in the colonia may not, in fact, be as densely populated. Likewise, if a colonia falls within low housing density census blocks, the resulting estimate will contain fewer housing units than are present in the colonia if the colonia lots are fully occupied. Reassuringly, as shown in Table 2, our population and housing estimations differ only slightly from those estimations that exist for Census Defined Places (CDP), which usually contain unincorporated county areas. Since CDP boundaries do not always overlap with colonias boundaries, we opted to rely upon our colonias cartography (columns highlighted in yellow), despite the possibility of estimation error, as they have unique features not captured by the CDP data.[4]

$$n_{ij} = \frac{n_i * A_{ij}}{A_i}$$

Where:n_i = population in the census block i, A_{ij} = Area of the colonia j that falls or intersect the census block i, A_i = Total area of the census block i. The total population living in colonias by county is an aggregate of each of these colonias' population estimates as identified in the Colonias Boundary Cartography.

[4] As described by the Colonias Initiative Program, border colonias meet three criteria: (1) they are located within 50 miles of the border, (2) underserved of basic services, and (3) classified as economically distressed community by the Water Code Section 17.921. In this section, "Economically distressed area" means an area in which: (A) water supply or sewer services are inadequate to meet minimal needs of residential users as defined by board rules;(B) financial resources are inadequate to provide water supply or sewer services that will satisfy those needs; and(C) an established residential subdivision was located on June 1, 2005, as determined by the board.

Table Nº 2 Census Defined Places (CDP) vs. Colonias Population and Housing Units Based in 2010 Census

Counties	CDP			Colonias Study Database		
	# CDPs	Population	Housing Units	# Colonias	Population	Housing Units
Cameron	35	36,788	10,279	174	37,640	11,270
El Paso	13	63,453	17,447	296	46,827	13,716
Hidalgo	35	98,832	28,431	926	88,475	25,525
Maverick	11	25,791	7,577	69	18,077	5,489
Starr	117	22,446	6,603	233	23,414	8,286
Webb	37	5,154	2,010	56	12,543	3,882
TOTAL	**248**	**252,464**	**72,347**	**1754**	**226,976**	**68,168**

Sample Size and Selection of Colonias

To establish a probabilistic sample representative of each county, we used a two-stage sampling technique. First, we calculated for each county a finite population-adjusted sample size that is representative of the housing units in the corresponding county using a confidence interval of 95% and 5% margin of error[5] (see Table 3 for the estimated sample sizes in each county).

Second, we selected a specific number of colonias that was feasible to visit given the logistics of the project by using a Probability Proportional to Size (PPS) technique designed to ensure the selection of **both** smaller and larger colonias within the sampling frame.[6] PPS is a method that involves the ordering of the participating units for selection by size, and randomly selects a specific number of units within a size range that proportionally resembles the distribution of the total population.

Table 3 Colonias Universe and Sample Sizes by County

---

[5] To calculate a finite sample size, which is adjusted to the population from where the sample is drawn, we use the following formula :

$$n_1 = \frac{n_0}{\left(1+\frac{n_0}{TotalHousingunits}\right)};$$

n_0 or the standard sample size given by the selected confidence interval or margin of error is calculated by:

$$n_0 = \frac{Z^2\, p\, (1-p)}{e^2}$$

Where: z is the confidence interval selected, p is the frequency of the expected parameter (in this case the expected number of the current unrecorded transactions is unknown so we use the highest probability of .5) and e is the margin of error.

[6] This was important since we need to ensure that the random selection was not overly weighted to the largest colonias which had often been subject to significant public policy interventions such as title conversions and which were often associated with particular notorious developers. Thus we wanted to ensure that we randomly from within both large and smaller categories.

ounties	Universe		Sample		
	Total Colonias	Housing units in Colonias 2010	Selected Colonias	Randomly Selected Colonias	Sample Size for Housing Units (95% confidence, 5% error)
Cameron	174	11,270	11	5	371
El Paso	296	13,716	12	11	374
Hidalgo	926	25,525	12	7	378
Maverick	69	5,489	7	7	359
Starr	233	8,286	10	10	367
Webb	56	3,882	7	7	350
TOTAL	**1,754**	**68,168**	**59**	**47**	**2,199**

Figure 2 illustrates our use of PPS using Val Verde County as a hypothetical county (it was an additional [7th] border county that was not specified as one of the six in the TDHCA's original request but for which we had data from earlier studies and did gather county data on recorded contracts for deed that are included in chapter three, but which, for resources reasons we decided not to conduct the sample survey). Assuming that the logistics of the project would have allowed us to visit 5 colonias in this county, the colonias in the county would then have been ordered from the smallest to the largest. We would have arrived at a sampling interval (SI) by dividing the total number of housing units in the county by the number of colonias we planned to visit. We next would have asked Excel to generate a random start (RS) to select our first colonia from the cumulative column, here Rio Bravo. The second, and subsequent colonias, would have been determined by adding to the value of the random start (RS+1 SI, RS+2 SI, and so forth). The advantage of this method is that it permits the selection of colonias within different size ranges.

County Name	ID	Community Name	Housing Units 2010	CUMULATIVE
Val Verde	M2330006	Owens Addition #1	3	3
Val Verde	M2330007	Owens Addition #2	3	6
Val Verde	M2330015	Langtry, Texas	4	10
Val Verde	M2330001	Amistad Acres	16	26
Val Verde	M2330009	Rio Bravo	25	51
Val Verde	M2330008	Payment	28	79
Val Verde	M2330005	Los Campos #3 & 4	31	110
Val Verde	M2330002	Box Canyon Estates	41	151
Val Verde	M2330010	Rough Canyon	61	212
Val Verde	M2330016	Los Campos #1,2 & 5	83	295
Val Verde	M2330004	Lake View Addition	91	386
Val Verde	M2330011	Town of Comstock	100	486
Val Verde	M2330012	Val Verde Park	297	783
Val Verde	M2330013	Val Verde Park #2	454	1237

Sampling Interval (SI)	247
Random Start (RS)	42
RS+SI	279
RS+2SI	527
RS+3SI	774
RS+4SI	1022

Figure 2: An Illustration of Probability Proportional to Size Sampling (PPS)

As mentioned above, the estimations of size for each colonia may present some under- or over-estimations. To address this issue, we updated the number of housing units for the PPS-selected colonias using Google Earth™. We then altered our estimations of total housing units in each of the selected colonias accordingly. Having updated these calculations with more accurate information, we established a sampling size for each colonia at a fixed rate calculated by dividing the total target sample for the county by the total number of housing units for the selected colonias in the county.[7]

In Chapter Two and earlier (above) we describe how given that our sampling frame did not include colonias platted after 2000, the research team decided to incorporate into the study newer colonia type subdivisions[8] in three counties: El Paso (1), Hidalgo (5) and Cameron (6). Because we lack data for full universe of such newer communities, which do not appear in our three sources of data (BEHI, OAG, and SOS), we did could not include them in our sampling frame. Instead, we selected them purposively. Therefore, we cannot attest that they are statistically representative of similar new subdivision communities across the border. For these reasons, we do not include the data gathered in these communities in our extrapolative sample.

[7] An equally fixed sample size for each selected colonia was not possible because, in many cases, the size of smaller colonias fell under the established fixed size.

[8] Given that these subdivisions are developed with basic infrastructure many observers and policy makers would not regard them as colonias, and they would not be analyzed as such, nor are they included in colonias datasets. However, despite having services they comprise very poor housing that is often identical or worse in dwelling conditions as other traditional colonas in their early phase of development. Moreover, this is where developers are concentrating their current practices. Thus we felt it important to include this new frontier of "colonia" type development into our study.

The same applies to all of the informal homestead subdivisions that we selected for interview in Guadalupe and Hays county in Central Texas.

Description of the Survey Returns and Margins of Error

Table 4. Characteristics of Survey Returns for Randomly Selected Colonias by County

County	Total Survey Returns	Completed Surveys (c)		Margin of error after survey implementation	Eligible Non Respondents (nr)	Ineligible Respondents (ir)			Unknown eligibility (ue)
		In-person	Mailbacks		Refusals	Vacant or abandoned dwelling	Rental Units	Commercial Establishments	Not Home
Cameron	128	126	2	8.61	50	110	15	13	171
El Paso	183	172	11	7.20	15	72	6	16	203
Hidalgo	181	176	5	7.26	25	99	32	11	351
Maverick	182	182	-	7.14	84	327	29	33	430
Starr	131	131	-	8.47	69	263	43	19	575
Webb	194	191	3	6.82	28	135	13	16	136
TOTAL	**999**	**978**	**21**		**271**	**1,006**	**138**	**108**	**1,866**

Given the limitations encountered once the survey was fielded, -- such as refusals, absences of probable eligible respondents at the moment of the survey, and either unoccupied or ineligible dwellings -- we were not able to achieve the target sample size. The overall response rate for the survey is 29%.[9] More survey returns or completed surveys were obtained for Webb and fewer for Cameron. As shown in Table 4, a total of 999 surveys were completed in the randomly-selected colonias, mainly through in-person interviews.[10] Most of the non-responses were due to vacant or abandoned dwellings, non-eligible properties such rental units or commercial establishments, and cases where the suitable respondent (the head of the household or the spouse) was not home. The overall refusal rate, based on the cases where the responded declined to participate in the survey, is 0.08%.[11]

[9] The response rate is calculated by: $\frac{C}{C+NR+IR+e*U}$, where C=completed surveys, NR=eligible non respondents (refusals), IR=ineligible respondents, U= unknown eligibility (not homes) and e=estimated proportion of eligibility (calculated from $\frac{C+NR}{C+NR+IR}$)

[10] As we describe elsewhere we designed the surveys such that where a selected lot respondent was not home after three visits, we left a copy of the survey with a letter of explanation and a pre-paid envelope in the hope that the household would return the completed question to us. This was a practice that we had used with some success in an earlier study (LBJ Rancho Vista and Redwood study www.lahn.utexas.org). However, perhaps because of the relative complexity of the instrument, and the focus of study – title papers past and present – the proportion of mail backs returned to us was disappointingly was low (around 2%). Thus we felt vindicated in having decided to use the face-to-face method of survey application.

[11]Non-response rates are calculated using the similar formula of response rates, but switching C by NR in the dividend.

Because fewer surveys were obtained than expected, margins of errors increased from ± 5 to between ± 6.82 (Webb) and ± 8.61 (Cameron). However, one of the advantages of using a probabilistic sampling approach is that post-sample weights may be used to make adjustments for non-response rates, unknown eligibility, and the impact of variances.

Sampling Weights and Adjustments for Extrapolation

The design implemented (PPS at a fixed sample rate) did not entail a self-weighting sample. This is partially due to the differences between our census overlay estimations and the actual size of the colonias. For the randomly selected colonias we calculated initial sample weights, adjusted by those differences in probabilities of selection, in order to be able extrapolate the results obtained from the survey to the county level. The pre-survey weights for each colonia (w_{ij}) were calculated as the inverse of two probabilities: the probability of selection of a housing unit given the number of colonias selected for the county relative to their sizes (P_i),[12] and the probability of selection for a housing unit in a selected colonia given the fixed rate established for its selection ($P_{j\,(i)}$). These initial sample weights were adjusted for unknown eligibility (given the large number of potential respondents that were not home at the time of the survey); non-response (given differences between the target sample size for a selected colonia and the number of completed surveys); and for variance effects (a design effect commonly present in PPS). For the statistical purist, as well as for reasons of full disclosure, these variance effects are described below. But for many readers the discussion may be overly detailed in which case the Appendix can end here.

a) Adjustment for unknown eligibility

During the survey, sampled housing units were visited at least twice. When the interviewer did not find anyone but the housing unit appeared to be occupied (e.g. house seem to be visibly taken care of), it was classified as "not at home." As the interviewer did not get to talk to the potential interviewee, and could not verify directly whether it was occupied or not, these assumptions may become a source of bias. Where the eligibilities of some sampled units are unknown, weights must be adjusted to account for this fact.

We used the proportion of sampled housing units known to be either eligible or ineligible in each colonia to make assumption about how many housing units of unknown eligibility could be considered eligible. Then, we calculated an adjustment factor as follows:

$$F_{ue} = \frac{\sum_c W_{ij,b} + \sum_{nr} W_{ij,b} + \varepsilon * \sum_{ue} W_{ij,b}}{\sum_c W_{ij,b} + \sum_{nr} W_{ij,b}}$$

where ε refers to the proportion of the unknown eligibility cases that are estimated to be eligible, *c* = the sum of completed surveys, *nr* refers to the sum of eligible non respondents, and *ue* refers to the sum of unknown eligibility. The adjusted base weights of housing units with complete interviews and eligible non-respondents are then obtained by multiplying their initial base weights ***wij,b*** by the factor *Fue*.

[12] $P_i = n * {}^{m_i}/_{m_{total}}$, where n=number of selected colonias in the county, m_i=the number of housing units in the selected colonia)

b) Adjustment for non response

To adjust for non-response, we calculated the non-response adjusted weight for the each sampled colonia as:$W_i = W1_i * F_{nr}$
where ***w1i*** is the initial weight (in this case the weight already adjusted by unknown eligibility) and ***Fnr***is which is the non-response adjustment factor. The adjustment non-response rate factor can be defined as the ratio of: the weighted number of surveys completed with eligible sampled cases, to the weighted total number of eligible known sampled cases. After the adjustment, we re-scaled the weights to match for the estimated county population.

c) Adjustment for variances

Table 5. Variance Inflation Factor by County

County	Variance Inflation Factor (L) Pre-Adjustment
Cameron	1.85
El Paso	2.84
Hidalgo	3.44
Maverick	1.64
Starr	1.59
Webb	1.70

Even though the use of weights in the analysis of survey data tends to reduce the bias in the estimates for extrapolation, it can also inflate the variances of such estimates. Indeed, given the PPS approach adopted this is likely the case, since each colonia may have different weights, which, in turn, may lead to greater variances in the population mean across colonias in the county. We identified the variance inflation factor (L) due to the use of the adjusted weights for each county.[13] As shown in Table 5, for Cameron, Maverick, Starr and Webb, there are increases in variances of 85, 64, 59 and 70 percent respectively, due to the use of weights. For El Paso and Hidalgo, the increase of variance due to weights is greater, 184 and 244 percent respectively.

To correct the inflation of variances, we used a technique to trim the weights based on the observation of mean square error (MSE) for two key variables from with the database: the year of move to the neighborhood and the total market value of the housing unit.[14] We calculated the MSE for each of the selected variables at different levels of truncation of the largest weight in the county. This procedure plots the values of the estimated MSE versus the percentiles of

[13] Note that this is not normally reported in most surveys unless the methodological report is specifically required to explain the adjustment for variances as a technique. To identify the factor of variance inflation (L), we used the following formula: $L = n * \frac{\sum_h n_h w_h^2}{(\sum_h n_h w_h)^2}$

[14] A detailed description of this procedure is provided by Frank Potter (1988) "Survey of procedures to control extreme sampling weights." Research Triangle Institute. CET CITE

truncation levels to visually determine a cut-off value that achieves adequate reductions of MSE. The lowest point implies the minimum cut-off point to reduce MSE. After selecting the adequate cut-off points (this being the one that reduces MSE the most), extreme weights were replaced by the value of the cut-off point and rescaled to permit the extrapolation to the estimated county housing units. Occasionally in the erosion or exclusion of cases included in a particular analysis of selected variables raises the margin of error above the minimum cut off point to ±10, in which case the increased is noted in the text or Table.

* * * *