Who Pollutes? A Household-Level Database of America's Greenhouse Gas Footprint

Kevin Ummel

Abstract

This paper describes the creation of a database providing estimated greenhouse gas (GHG) footprints for 6 million US households over the period 2008-2012. The database allows analysis of footprints for 52 types of consumption (e.g. electricity, gasoline, apparel, beef, air travel, etc.) within and across geographic regions as small as individual census tracts. Potential research applications with respect to carbon pricing and tax policy are discussed. Preliminary analysis reveals:

• The top 10% of US polluters are responsible for 25% of the country's GHG footprint. The least-polluting 40% of the population accounts for only 20% of the total. The average GHG footprint of individuals in the top 2% of the income distribution is more than four times that of those in the bottom quintile.

• The highest GHG footprints are found in America's suburbs, where relatively inefficient housing and transport converge with higher incomes. Rural areas exhibit moderate GHG footprints. High-density urban areas generally exhibit the lowest GHG footprints, but location-specific results are highly dependent on income.

• Residents of Republican-held congressional districts have slightly higher average GHG footprints than those in Democratic districts – but the difference is small (21.8 tCO_2e/person/year in Republican districts; 20.6 in Democratic). There is little relationship between the strength of a district's party affiliation and average GHG footprint.

JEL Codes: Q53, Q57

Keywords: air pollution, greenhouse gases, climate change, environment

Center for Global Development
www.cgdev.org

Working Paper 381
October 2014

Who Pollutes? A Household-Level Database of America's Greenhouse Gas Footprint

Kevin Ummel
Center for Global Development

The author would like to thank Glen Peters, Scott Curtin, Art Diem, Dallas Burtraw, and two anonymous reviewers for their assistance and suggestions. All remaining errors are the author's alone.

CGD is grateful for contributions from its board and funders in support of this work.

Kevin Ummel. 2014. "Who Pollutes? A Household-Level Database of America's Greenhouse Gas Footprint." CGD Working Paper 381. Washington, DC: Center for Global Development.
http://www.cgdev.org/publication/who-pollutes-household-level-database-americas-greenhouse-gas-footprint-working-paper

Center for Global Development
2055 L Street, NW
Fifth Floor
Washington, DC 20036

202.416.4000
(f) 202.416.4050

www.cgdev.org

Contents

Introduction

Climate-changing greenhouse gas (GHG) emissions ultimately result from the demand for goods and services. In theory, almost all anthropogenic GHG emissions can be traced back to consumption on the part of households.[1] This includes "direct" emissions due to the consumption of energy (e.g. electricity to light or cool a home, gasoline for personal automobiles, natural gas for cooking and residential heating, etc.) and "indirect" emissions released during the production or manufacture of food, apparel, air travel, services, etc.

So-called "consumption-based" emissions differ from production-related or territorial emissions. In the case of consumption-based emissions, the location at which GHG's are released into the atmosphere – whether a power plant in Ohio or a factory in China – is irrelevant. The goal is to trace emissions back to the household consumption choices that ultimately led to their production. See Wiedmann (2009) for a review of methodological issues in consumption-based emissions accounting.

Consumption-based GHG "footprints" provide a more relevant metric for assigning responsibility for climate change across societies and individuals. They are also a necessary input to the analysis of household-level effects (i.e. "incidence") of potential carbon pricing policy. This paper does *not* attempt to estimate the incidence of a carbon tax on U.S. households; see Williams III et al. (2014) for an example of such work. However, the data developed here are central to such efforts. The implications and applications of this database with respect to carbon pricing policy are discussed later.

The objective of this study is to create a high-resolution, household-level database of U.S. GHG footprints to enable future analysis across any relevant sectoral, demographic, or spatial dimension. While previous work has modeled average household footprints for specific geographic entities (Jones and Kamen 2014), this study develops GHG footprints for all individual households in a large, nationally-representative sample. Further, while the methodology used here shares basic features with previous research in this vein (e.g. Burtraw et al. 2009, Hassett et al. 2009, Grainger and Kolstad 2010), new data and techniques introduced here offer significant improvements.

[1] Some GHG emissions are attributable to consumption by governments (e.g. public administration, military, etc.) and account for about 8% of the total U.S. GHG footprint (Andrew and Peters 2013). There is no obvious way to allocate these emissions to individual households. This analysis focuses on the remaining 92% of emissions that can be attributed to specific consumption choices on the part of households.

Data and Methodology

Household level, consumption-based GHG footprints are most readily estimated from expenditure data. If expenditures for various goods and services are known, they can be converted to GHG emissions using emission factors for specific types of consumption (e.g. GHG per dollar of air travel expenditure). This study proceeds in two steps: 1) creation of a large, household-level expenditure database and 2) conversion of those expenditures to GHG footprints. Each step is explained below.

Simulation of household expenditures

The Bureau of Labor Statistic's Consumer Expenditure Survey (CEX) uses interviews and diaries to collect expenditure, income, and demographic data for a representative sample of American households.[2] This study utilizes all available CEX data over the period 2008-2012 to compute inflation-adjusted (real) annual expenditures for 52 spending categories across 23,552 unique households. The 52 expenditure variables are listed in Table 1.

Table 1: Household expenditure variables constructed from CEX survey data

Expenditure variable
Air travel
Alcoholic beverages
Apparel
Beef
Cash contributions
Cereals and baked goods
Dairy
Drugs
Education
Electricity
Fees and admissions
Food away from home
Fruits and vegetables
Furniture
Gasoline
Health insurance
Heating oil
Home insurance
Home maintenance and repairs
Household textiles
Laundry and cleaning supplies

[2] http://www.bls.gov/cex/

LPG
Major appliances
Medical services
Medical supplies
Miscellaneous household equipment
Mortgage interest
Natural gas
New car and truck net outlay
Nonalcoholic beverages
Other entertainment supplies, equipment, and services
Other food at home
Other fuels
Other household expenses
Other shelter
Other vehicle net outlay
Personal care products and services
Personal insurance and pensions
Personal services
Pets, toys, and playground equipment
Pork
Poultry and fish
Public transportation
Rent
Small appliances, miscellaneous house wares
Telephone services
Television, radios, sound equipment
Tobacco products and smoking supplies
Used car and truck net outlay
Vehicle maintenance and repairs
Vehicle rental, leases, licenses, other charges
Water and other public services

Although the most exhaustive survey of its kind, the CEX is not large enough to allow valid analysis of household expenditures at high spatial resolution. In order to analyze expenditure patterns at higher spatial resolution, it is necessary to use the CEX data to impute or (more accurately, in this particular case) *simulate* expenditures for a larger sample of households.

The Census Bureau's American Community Survey (ACS) is the largest ongoing household survey in the United States.[3] The ACS provides information on household demographics, sources of income, and some expenditures (e.g. cost of housing). The Public Use Microdata Sample (PUMS) is a representative subsample of the complete ACS made public for research purposes. To ensure confidentiality, the PUMS identifies the location of individual

[3] In a given 5-year period, the ACS obtains completed surveys from about 7.5% of all U.S. households. Surveys are sent out monthly; responses are provided by mail, telephone, and in-person interview (http://www.census.gov/acs/www/).

households by Public Use Microdata Area (PUMA), of which there are more than 2,000 nationwide. This study uses the 2008-2012 5-year PUMS release containing data for 6 million households or ~5% of the total population.[4]

It is possible to identify or construct 50 household variables common to both the CEX and PUMS. It is also possible to assign 8 third-party variables to households in each survey based on geographic location: population density, heating and cooling degree-days[5], and local and state fuel prices.[6] Table 2 provides a complete list.[7]

Table 2: Household variables that can be derived for both the CEX and ACS surveys

Household characteristics
Number of people
Age of primary earner or householder
Race of primary earner or householder
Sex of primary earner or householder
Employment status of primary earner or householder
Educational attainment of primary earner or householder
Occupation of primary earner or householder
Number of adults 18 or older
Number of adults 18 to 44 years old
Number of adults 45 to 64 years old
Number of adults more than 64 years old
Number of children under 18
Number of children less than 6 years old
Number of children 6 to 12 years old
Number of children 13 to 17 years old
Number of college students
Number of people in labor force
Number of unemployed workers
Number of active military
Average age of adults
Average age of children
Average age of workers
Average years of education completed by adults
Average hours worked per week per worker (last 12 months)

[4] The analysis here is limited to the non-group quarter population and excludes individuals housed in correctional facilities, juvenile facilities, nursing homes, and health care facilities.

[5] Degree-days data from NOAA (http://www.cpc.ncep.noaa.gov/products/analysis_monitoring/cdus/degree_days/)

[6] State-level fuel prices from EIA SEDS database (http://www.eia.gov/state/seds/)

[7] The aforementioned third-party variables are naturally assigned to PUMS households using the designated PUMA. Since the CEX does not reveal household location with precision, it is necessary to assign households to a PUMA on the basis of both location and household characteristics. A series of state-specific boosted regression tree classification models are fit to the PUMS data, predicting PUMA assignment using household income, race, education, housing status/cost, property value, etc. along with the geographic information that is provided in the CEX. In cases where the state is unidentified in the CEX, models specific to the Census region are used. Having assigned the most likely PUMA to each CEX household, it is possible to attach the third-party variables.

Family type (couple both in labor force, single person not in labor force, etc.)
Food stamp recipient flag
Number of vehicles
Household income
Total household income
Wages and salary
Self-employment net income
Social Security, Railroad Retirement, and Supplemental Security Income
Retirement income (pensions, IRA, etc.)
Interest, dividend, and rental income
Welfare or public assistance income
Other income (unemployment, child support, etc.)
Housing situation
Housing status (rent, own, etc.)
Housing type (single family home, apartment building, etc.)
Number of rooms
Housing age
Heating fuel
Housing tenure (years; owned housing only)
Rent expenditure
Mortgage expenditure
Homeowners insurance
Property tax
Property value (self-reported; owned housing only)
Flags (4) indicating if electricity, natural gas, heating fuel, or water included in rent
Third-party variables
Population density
Heating degree-days (monthly average; last 12 months)
Cooling degree-days (monthly average; last 12 months)
Local electricity price (average; last 12 months)
State gasoline price (average; last 12 months)
State natural gas price (average; last 12 months)
State heating oil price (average; last 12 months)
State LPG price (average; last 12 months)

The goal is to use the common variables (Table 2) to simulate a valid value for each CEX expenditure variable (Table 1) for each household in the PUMS. One can approach this problem in a number of ways. Simple statistical imputation (e.g. mean response linear regression) will generate the *expected* expenditure for a given variable and household, but it will not preserve the population-wide *distribution* of expenditures. Alternatively, a sample-to-sample "matching" algorithm attempts to assign observed expenditure values for a single CEX household to a statistically similar household in the PUMS. However, the CEX's small size and limited geographic information (relative to the PUMS) make matching algorithms a poor choice.

To avoid these problems, the 52 CEX expenditure variables are simulated for each of the 6 million households in the PUMS using boosted quantile regression trees in conjunction with a 6 million x 52 matrix of correlated random uniform variates.

A boosted regression tree (BRT) model is fit to the CEX sample for each expenditure variable (the response/dependent variable). The variables in Table 1 are provided as potential regressors (independent variables). The R *gbm* package (Ridgeway 2013) is used to implement the gradient boosting machine of Friedman (2001) with a quantile regression loss function. Quantile regression allows prediction of conditional quantiles rather than the conditional mean of the response variable.

A BRT model is fit for each of 52 expenditure variables and 14 percentile values ranging from 0.05 to 0.999. Three-quarters of the CEX observations are randomly chosen to fit the model (i.e. training data), and the remaining observations are used to evaluate the loss function at each boosting iteration. Trees are added until the loss function shows no significant improvement. The use of BRT models allows for non-linear relationships between the response and independent variables; multiple degrees of interaction (up to 5, in this case); and automated variable selection with little risk of over-fitting to noise. The reader is directed to Brieman et al. (1984), Friedman (2002), and Koenker and Bassett (1978) for more technical descriptions of regression trees, gradient boosting, and quantile regression, respectively.

The resulting 728 BRT models are used to predict quantile values for each expenditure variable and PUMS household. The quantile values describe the cumulative distribution function for a given expenditure variable, conditional upon each household's unique characteristics.

A naïve simulation approach is to draw expenditure values randomly from the conditional distributions. This would ensure a plausible value for each variable-household combination and preserve the population-wide distribution for each variable, but it would *not* preserve the observed *correlation* among expenditure variables. For example, households with air travel expenditures at the 90th percentile of the conditional distribution are also likely to exhibit relatively high hotel expenditures (e.g. 80th percentile of the conditional distribution). That is, some expenditures remain correlated across households *even after controlling for observable household characteristics*.

To capture this phenomenon, I draw expenditure values from the conditional distributions using uniformly-distributed random variates that are appropriately correlated across expenditure categories. For each expenditure variable and household in the CEX data, the BRT models are used to identify the location (percentile) of the observed expenditure within the conditional distribution. A 52 x 52 weighted correlation matrix is computed from the percentiles. The algorithm of Schumann (2009) is then used to generate a 6 million x 52 matrix of random uniform variates that retain the correlations observed in the CEX.

This matrix is used to draw values from the conditional distributions computed for each ACS household and expenditure variable. The end result is a 6 million x 55 matrix of simulated expenditures. Initial household expenditure estimates for electricity, natural gas, and heating oil are then adjusted state-by-state to ensure the totals match reported residential revenues in the EIA State Energy Data System (SEDS).

Conversion of expenditures to GHG footprints

Conversion of expenditures to GHG emission footprints requires an assumed emission intensity for each expenditure variable (i.e. GHG per dollar). Table 3 provides assumed emission factors (kgCO2e per USD 2012) for each of the 52 expenditure variables. Some emission factors exhibit spatial variation due to geographic variability in the GHG-intensity of consumption (e.g. electricity) and/or observable spatial variation in prices (e.g. gasoline). In these cases, Table 3 reports the national minimum and maximum emission factors. These factors reflect estimates of the fully-accounted "life-cycle" emissions.

In some cases, the GHG-intensity of consumption can be calculated directly from available environmental and economic data. In other cases, previous input-output and life-cycle studies are drawn upon to generate plausible values. National, consumption-based GHG emissions data from Andrew and Peters (2013) are used to provide "top down" checks on the magnitude of emission factors; that study constructs a multi-region input-output table from the Global Trade Analysis Project (GTAP) database to estimate consumption-based emissions for 129 countries and regions across 58 trade categories.

The following sections briefly describe the creation of the emission factor values in Table 3 for major types of consumption. More details are provided in the Annex.

– Electricity

The GHG-intensity of electricity supply varies across space. The EPA eGRID program provides GHG emission factors for 26 power grid subregions in year 2010, reflecting emissions released at power plants during fuel combustion (zero in the case of renewable generators).[8] Additional "upstream" emissions from associated construction, mining, processing, and transport are introduced via technology-specific emission factors from NREL's Life Cycle Assessment Harmonization Project.[9]

The subregion emission factors are further adjusted for grid line losses between generators and consumers. The eGRID-derived emission factors do *not* account for inter-regional electricity flows that could impact the true GHG-intensity of electricity consumed. Up to 30% of electricity consumed in some grid subregions originates elsewhere (Diem and Quiroz 2012), but there is currently no simple way to account for these flows.

A dataset provided by eGRID linking grid subregions to zip codes is used to estimate mean GHG-intensity of electricity supply for individual PUMA's in 2010. This is merged with PUMA-level electricity price data to calculate the implied GHG footprint per dollar of expenditure.

– Gasoline, natural gas, heating oil, and LPG

For these direct energy expenditures, state-specific emission factors are calculated using EIA fuel price data and life-cycle GHG emission estimates from the literature (see Annex for relevant references).

– Air travel

Analysis of data from the MIT Airline Data Project[10] and Andrew and Peters (2013) implies an average emission factor of 1.35 kgCO2e per dollar of air travel expenditure (2012 USD). This figure is based on reported passenger revenue and jet fuel consumption for U.S. airlines over 2008-2012 and an estimate of upstream pollution attributable to capital formation in the airline industry (e.g. construction of airplanes).

[8] http://www.epa.gov/cleanenergy/energy-resources/egrid/index.html
[9] http://www.nrel.gov/analysis/sustain_lcah.html
[10] http://web.mit.edu/airlinedata/www/default.html

– *Food and drink*

The expenditure variables include 10 food- and drink-related categories. Life-cycle GHG emissions per kg for associated raw foods are extracted from Venkat (2011) and combined with average U.S. price data for similar foods from the BLS Consumer Price Index database to estimate the GHG footprint per dollar of expenditure. For variables where a direct calculation is not possible (e.g. alcoholic beverages), Weber and Matthews (2008) is used to estimate the emission factor relative to that of beef. The magnitude of the initial emission factors is then adjusted to ensure that the national average food- and drink-related household GHG footprint matches that calculated by Weber and Matthews (2008).

– *Other consumption*

For all other expenditure variables, initial emission factors are taken from Shammin and Bullard (2009), which uses the U.S. Department of Commerce Economic Input-Output Life Cycle Analysis database to derive emission factors (CO_2 per dollar) for the same CEX expenditure categories used here. The original emission factors are then scaled to ensure that the total GHG footprint from the underlying consumption categories matches that implied by Andrew and Peters (2013). This also converts the original emission factors to CO_2-equivalence, under the assumption that the allocation of non-CO_2 GHG's across categories is proportional to that of CO_2.

Table 3: Assumed GHG emission factor for each expenditure variable

Expenditure variable	Emission factor kgCO_2e per USD (2012)
Electricity	0.9 to 11.51
Natural gas	1.43 to 10.86
Heating oil	2.65 to 5.67
LPG (propane)	1.86 to 4.84
Gasoline	2.73 to 4.61
Beef	3.02
Home maintenance and repairs	2.37
Dairy	1.91
Other vehicle net outlay	1.82
Water and other public services	1.64
Used car and truck net outlay	1.37
Air travel	1.35
Household textiles	1.32
Pork	1.32
Major appliances	1.28

New car and truck net outlay	1.28
Apparel	1.19
Pets, toys, and playground equipment	1.19
Furniture	1.14
Laundry and cleaning supplies	1.09
Miscellaneous household equipment	1.05
Other entertainment supplies, equipment, and services	1.00
Small appliances, miscellaneous house wares	1.00
Poultry and fish	0.95
Other food at home	0.91
Other shelter	0.91
Food away from home	0.84
Television, radios, sound equipment	0.82
Vehicle maintenance and repairs	0.78
Alcoholic beverages	0.76
Nonalcoholic beverages	0.76
Public transportation	0.75
Cash contributions	0.73
Other household expenses	0.73
Personal care products and services	0.73
Rent	0.68
Drugs	0.64
Medical supplies	0.64
Cereals and baked goods	0.63
Education	0.55
Fruits and vegetables	0.54
Medical services	0.50
Personal services	0.50
Vehicle rental, leases, licenses, other charges	0.50
Telephone services	0.46
Mortgage interest	0.41
Tobacco products and smoking supplies	0.36
Home insurance	0.32
Personal insurance and pensions	0.32
Health insurance	0.18
Fees and admissions	0.05
Other fuels	0.00

Results

The final database contains estimated GHG footprints for 52 types of consumption across 6 million households over the period 2008 through 2012, in addition to the full range of demographic variables inherent to the ACS. The high resolution allows analysis of a wide range of phenomena. In this section, I illustrate relationships between mean per-person

GHG footprints and per-person income, population density, and congressional district political affiliation.[11]

Figure 1 shows average per-person GHG footprint by income group across seven aggregated emission categories. Utilities (electricity, natural gas, heating oil, and LPG) and gasoline constitute the "direct" component of the footprint; other categories reflect "indirect" emissions. Indirect emissions account for about 64% of the average GHG footprint, but this varies considerably with income. Indirect emissions account for 59% of the average footprint among individuals in the lowest income quintile, but this rises to 75% for those in the top 2% of the income distribution.

These estimates of the "indirect" component of household GHG footprints are somewhat higher than reported elsewhere (for example, Grainger and Kolstad 2010). However, most previous studies have not used a top-down, consumption-based national emissions data source to constraint estimates derived from household surveys alone. Moreover, the indirect component reported here is supported by analysis of the EPA's U.S. GHG Inventory[12] and the 2009 National Household Travel Survey (NHTS).[13] Together, these independent data sources imply indirect household emissions equal to ~64% of total GHG emissions – in accordance with the results presented here.[14]

The disproportionate growth in indirect emissions at higher levels of income leads to significant inequality in GHG footprints across individuals. On average, persons in the top 2% of the income distribution exhibit footprints more than four times larger than those in the bottom quintile. It is likely that higher-income individuals greatly underestimate their personal GHG footprints, given that indirect emissions are less apparent than those associated with direct consumption of electricity, natural gas, and gasoline.

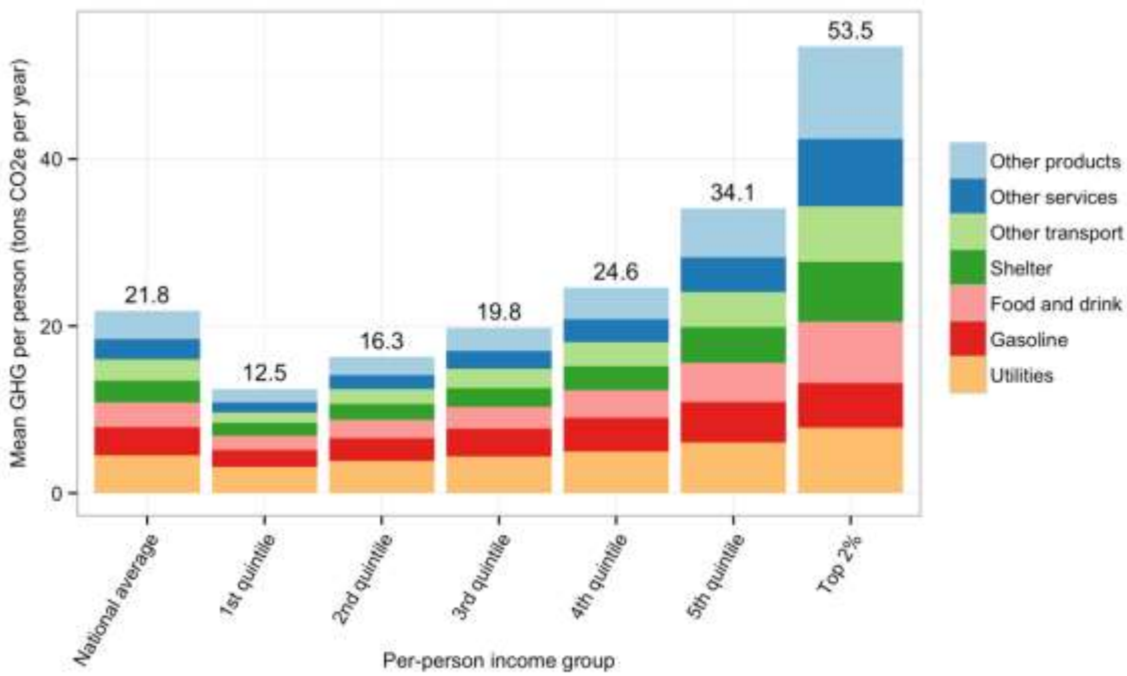
[11] Per-person measures are used instead of household averages in order to account for differences in household size. Per-person averages are computed by dividing the relevant household value (e.g. total GHG footprint) by the number of people and weighting the result by the product of the household sample weight and number of people.

[12] http://www.epa.gov/climatechange/ghgemissions/usinventoryreport.html

[13] http://nhts.ornl.gov/

[14] For 2009, the EPA reports total production-related U.S. emissions of 6.66 $GtCO_2e$, of which 1.18 $GtCO_2e$ are due to residential electricity- and combustion-related activities. The 2009 NHTS implies total household gasoline consumption was responsible for 1.21 $GtCO_2e$. Combined this suggests household "direct" emissions account for 35.8% of total emissions.

Figure 1: Average GHG footprint and income per person[15]



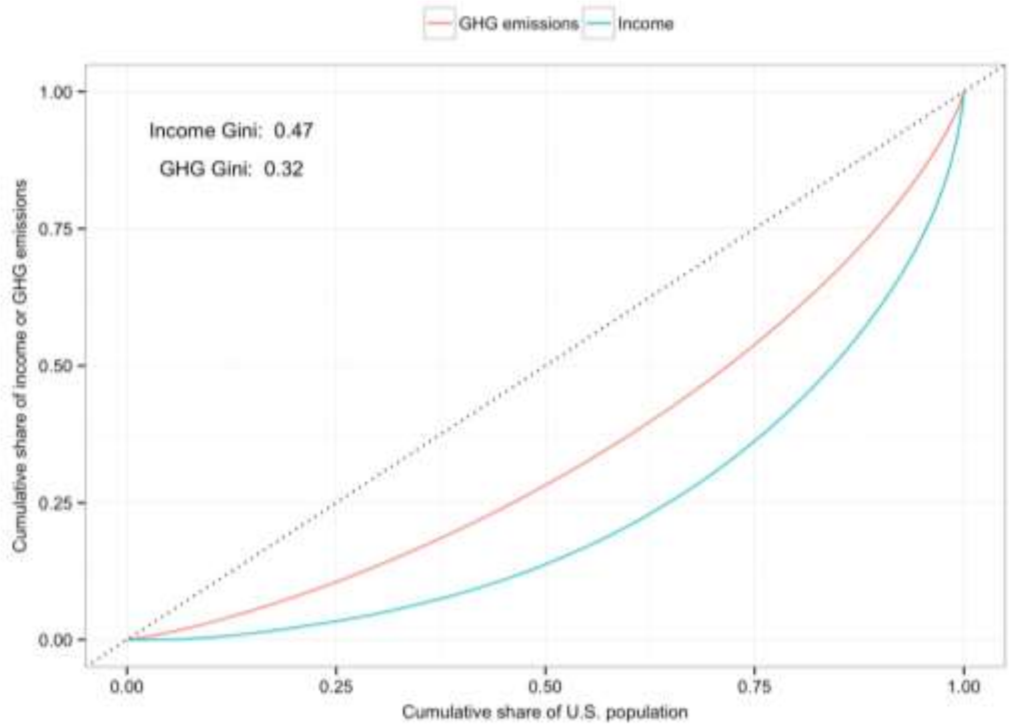
Since GHG footprints are estimated at the level of individual households, it is possible to explicitly chart the degree of GHG inequality nationally. Figure 2 provides the well-known Lorenz curve for both GHG emissions and income. The curves describe the proportion of each quantity assumed by a given proportion of the population, ranked from lowest to highest.

GHG emissions are less unequally distributed (Gini coefficient = 0.32) across the U.S. population than income (Gini coefficient = 0.47). However, considerable pollution inequality remains. The top 10% of polluters are responsible for nearly 25% of the national GHG footprint, and the top 20% of polluters account for 40% of all GHG pollution. Conversely, the lowest-emitting 40% of the population (largely individuals in lower income groups) are responsible for just 20% of the total burden.

---

[15] The 5th income quintile reported here does not include the top 2%. It includes only percentiles 80 through 98. Student households are excluded, as are those where expenditures greatly exceed reported income; these are likely to be cases of under- or mis-reported income.

Figure 2: Lorenz curves and Gini coefficients for income and GHG emissions



It is also possible to use the household-level results to visualize average GHG footprints across space. In order to calculate statistics for alternative geographic regions (e.g. zip codes or congressional districts), it is necessary to compute new sample weights that reflect the likelihood of a given household being located in a given region. A sample weight "raking" algorithm is employed to assign and re-weight households for any given geographic region, using region-specific marginal household counts from ACS and 2010 Census summary files.

This technique ensures that the subsample of households assigned to a given zip code or congressional district, for example, reflects the *actual* distribution of households across income, age, race, housing tenure, and household size. It also allows for households to be re-weighted for analysis down to the level of individual census tracts, though zip code results are presented here. More details are provided in the Annex.

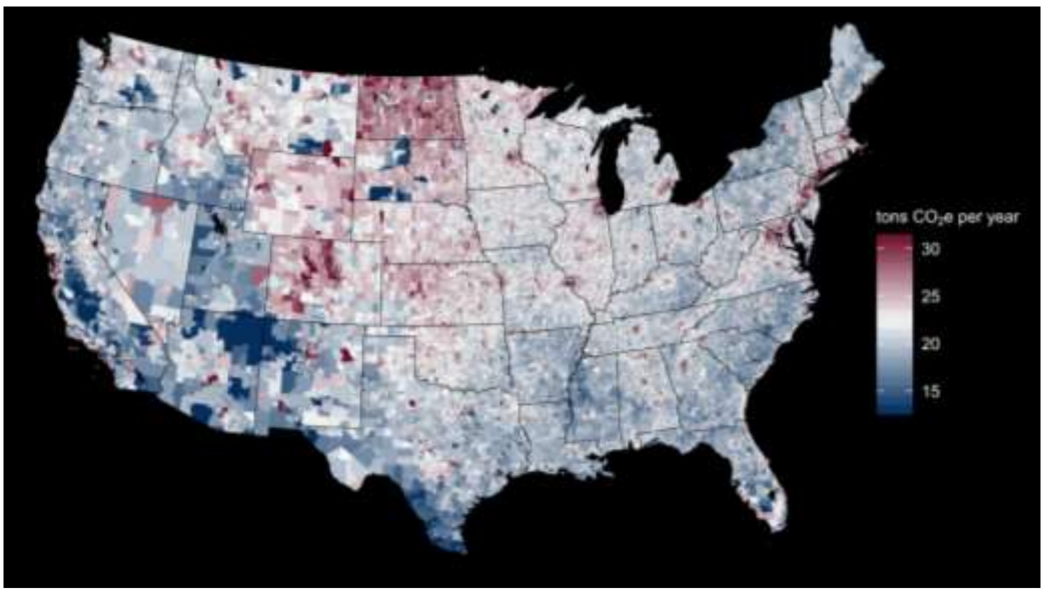
Figure 3 displays average per-person GHG footprints for more than 30,000 individual zip codes.[16] Broadly speaking, higher GHG footprints are observed across the north-central areas of the country – especially North Dakota, South Dakota, Nebraska, Kansas, Colorado,

[16] If the number of households assigned to a zip code was less than 60, the original estimate of average, per-person GHG footprint was arbitrarily dropped to reduce the chance of erroneous values. Values for those zip codes are spatially interpolated in Figure 3.

and Wyoming. Relatively low GHG footprints are (again, *generally*) found in the western, southern, and northeastern states.

However, intra-state variation is generally more significant than inter-regional differences. This phenomenon is driven by local spatial variability in population density and income. Throughout the country, a general pattern is noticeable in and around major urban areas. Footprints are often quite low within urban core areas but increase as one moves outward geographically. The highest footprints are found in suburban communities characterized by higher incomes and less efficient transport and housing. Footprints then decline as one moves beyond the suburbs into relatively poorer rural areas.

Figure 3: Average GHG footprint per person, by zip code (2008-2012)[17]



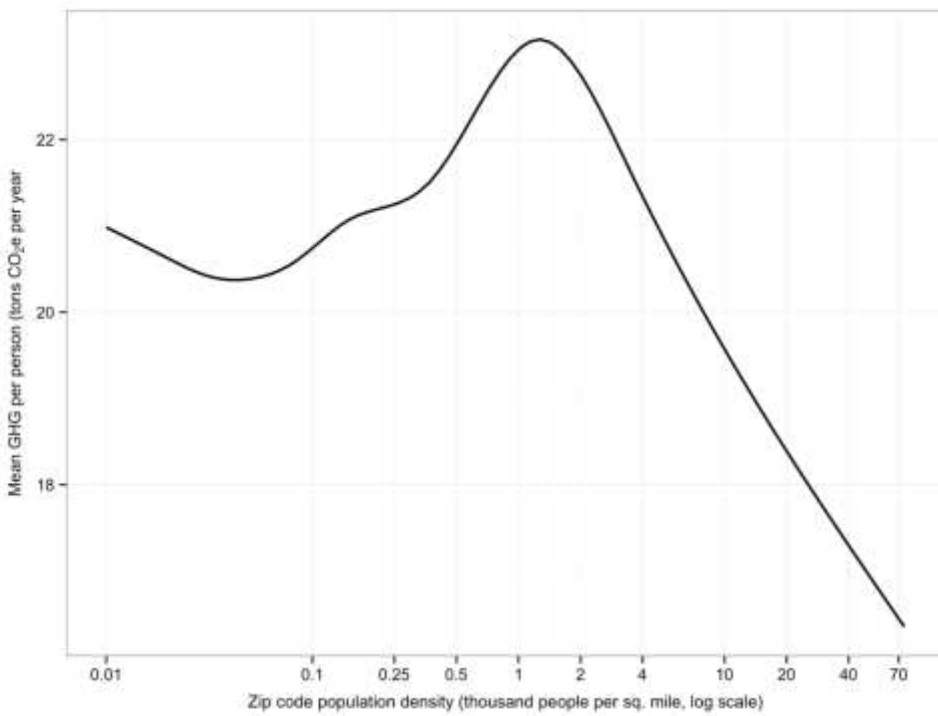
This relationship can be seen clearly in Figure 4, which shows the generalized relationship between average per-person GHG footprint and population density. The curve is the result of fitting a generalized additive model to the 30,000 zip code data points plotted in Figure 3. Note that a log scale is used for the x-axis.

The results suggest that per-person GHG footprints actually *increase*, on average, with greater population density up to about 2,000 persons per square mile. At densities beyond that threshold, average GHG footprints decline. Overall, footprints are typically *highest* at

[17] A high-resolution version of this map is available at: https://www.dropbox.com/s/5zzarkilwfxu0ty/Map%20of%20mean%20footprint%20by%20zip%20code%20%28high-res%29.png?dl=0

densities between about 250 and 4,000 persons per square mile. It is only beyond about 6,000 people per square mile that greater population density is significantly associated with lower GHG footprints.

Figure 4: Generalized relationship between GHG footprint and population density



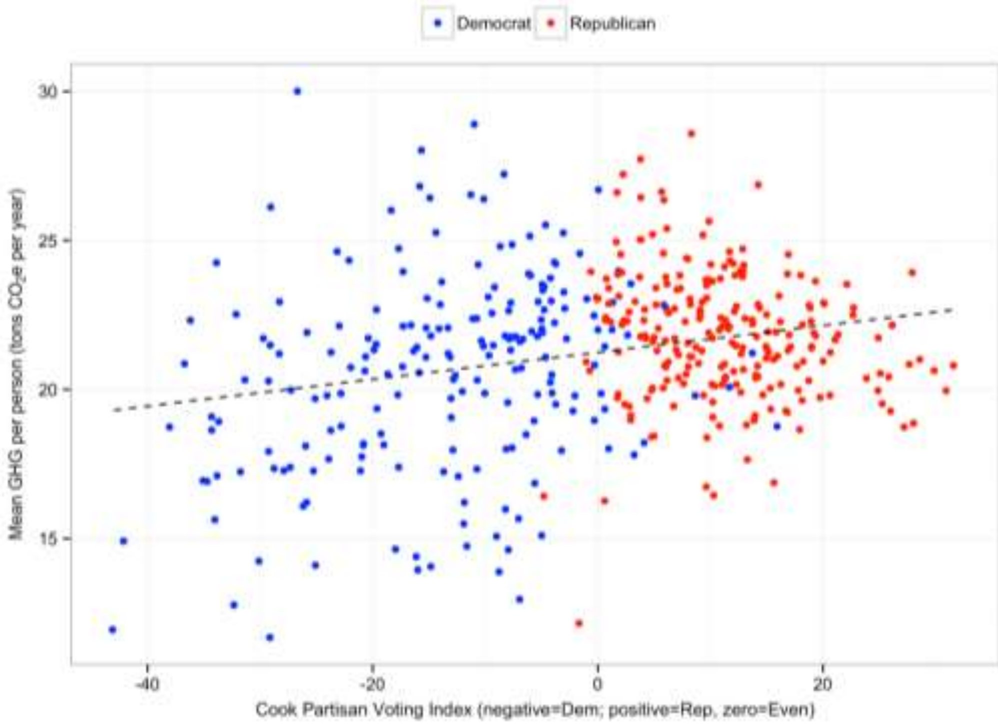
Finally, Figure 5 displays the observed relationship between average GHG footprints and the political allegiance of individual congressional districts. Each point represents one of 435 districts, color-coded by party affiliation of the current representative (i.e. 113th Congress). Each point identifies the average per-person GHG footprint in the district (y-axis) and the degree of partisanship as measured by the Cook Partisan Voter Index (PVI) using results from the 2012 election (x-axis).[18] Negative or positive PVI values indicate that a district is Democrat- or Republican-leaning, respectively. The dashed line is a linear best fit between PVI and average GHG footprint.

Most striking is the lack of obvious patterns. Individuals in Republican-held districts do, on average, exhibit slightly higher GHG footprints than those in Democratic districts – but the difference is less than 6% (mean of 21.8 tCO_2e per year in Republican-held districts; 20.6 in

[18] http://cookpolitical.com/house/pvi

Democrat-held). And the relationship between PVI and average footprint is quite weak (R^2 = 0.06).

Figure 5: Average GHG footprint by congressional district and degree of partisanship



Discussion

This paper describes the creation of a database containing estimating GHG footprints across 52 consumption categories for a sample of 6 million individual U.S. households. The data are relevant to a wide range of research questions and simulation exercises. A technique for raking household weights allows representative household subsamples to be created for geographic entities as small as individual census tracts.

While lower than income inequality, GHG pollution inequality is substantial. The top 10% of U.S. polluters are responsible for nearly 25% of the national GHG footprint, while the lowest-emitting 40% of the population are responsible for just 20% of the total burden. On average, persons in the top 2% of the income distribution exhibit GHG footprints more than four times larger than those in the bottom quintile.

Surveys like the CEX and PUMS often have difficulty adequately sampling households at the "tails" of the income distribution, especially among wealthier households. It is likely that expenditures are under-reported for households at the upper end of the income distribution. If this is true, then *actual* GHG footprints among the richest households could be significantly higher than reported here.

On the other hand, the current methodology makes no adjustment for spatial variation in prices (except for electricity and, to a lesser extent, other fuels) or sales tax. This likely overstates indirect emissions for households in places with higher than average prices and understates emissions in low-cost areas. Since higher prices are typically associated with high-income areas, households in those areas may be unduly penalized with the current approach. It is also possible that economy-wide emission factors are inappropriate at the upper end of the income distribution.[19] Unfortunately, there is no simple way to tease out the relative size of these competing effects.

With respect to spatial variation in footprints, I show that zip codes with population density between about 250 and 4,000 people per square mile exhibit the highest average per-person GHG footprints. These are typically suburban and quasi-rural areas where relatively inefficient transport and housing converge with higher incomes. Half of the entire U.S. population lives in such areas.

GHG footprints typically decline at population densities beyond 6,000 people per square mile. But these are average results, and location-specific footprints are highly income-dependent. For example, estimated mean per-person GHG footprints for zip codes 10014 (Manhattan, New York City) and 10457 (Bronx, New York City) are 36 and 11.7 tCO_2e per year, respectively, despite being located in the same city and having similar population density.

The unambiguously lowest GHG footprints are found in places with *both* high density and lower incomes and consumption. That is, where the energy efficiency benefits of density are not offset by higher incomes and a resulting increase in indirect emissions.

Residents of Republican-held congressional districts have slightly higher average GHG footprints than those in Democratic districts – but the difference is less than 6% (21.8 tCO_2e/person/year in Republican districts; 20.6 in Democratic). There is little relationship

[19] For example, physically-similar cotton shirts purchased at Walmart and Abercrombie & Fitch will differ significantly in price, but the production process and associated life-cycle GHG emissions may be quite similar.

between the strength of a district's party affiliation (i.e. Partisan Voter Index) and average GHG footprint.

The absence of marked differences between Republican and Democratic districts suggests that the parties' constituents are about equally exposed to the household financial effects of potential carbon pricing. That is, carbon pricing (e.g. a national carbon tax) would likely impose similar average financial costs on Republican and Democratic households through higher prices for carbon-intensive goods and services, ignoring effects on local employment and returns to capital. However, this is not what one would surmise given the parties' divergent positions on climate policy.

It should be noted that moving from the database developed here to a fully-accounted estimate of the incidence of a carbon tax requires adjustments for a number of important considerations, including: 1) the effect of household substitution away from carbon-intensive goods and services; 2) the degree to which the tax is passed through to consumers or borne by producers, and 3) the effect of a tax on returns to capital.

The database introduced here provides a basis for identifying – with considerable spatial and demographic detail – how carbon tax revenues might be returned to taxpayers in a way that is amenable to members of both parties. A carbon tax alone imposes a cost *roughly* proportional to a household's GHG footprint, subject to the considerations mentioned above. If a policy could use the new revenue to direct tax cuts, expanded Social Security benefits, and/or per-person "dividends" such that key constituencies experience a net financial *benefit* under the policy, it might be possible to forge a political coalition. The high level of spatial and demographic detail provided by this new database makes such an analysis possible.

References

Andrew, R.M. and Peters, G.P. 2013. A multi-region input-output table based on the Global Trade Analysis Project database (GTAP-MRIO). *Economic Systems Research* 25(1): 99-121.

Brieman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. 1984. *Classification and regression trees*. Chapman and Hall. Monterey, CA.

Burnham, A., Jeongwoo, H., Clark, C.E., Wang, M., Dunn, J.B., and Palou-Rivera, I. 2012. Life-cycle greenhouse gas emissions of shale gas, coal, and petroleum. *Environ. Sci. Technol.* 46(2): 619-627.

Burtraw, D., Sweeney, R., and Walls, M. 2009. The incidence of U.S. climate policy: Alternative uses of revenues from a cap-and-trade auction. Discussion Paper 09-17, Resources for the Future, Washington D.C.

Diem, A. and Quiroz, C. 2012. How to use eGRID for carbon footprinting electricity purchases in greenhouse gas emission inventories. Environmental Protection Agency, July. <http://www.epa.gov/ttn/chief/conference/ei20/session3/adiem.pdf>

Friedman, J.H. 2001. Greedy function approximation: a gradient boosting machine. *Annals of Statistics* 29(5): 1189-1232.

Friedman, J.H. 2002. Stochastic gradient boosting. *Computational Statistics and Data Analysis* 38(4): 367-378.

Grainger, C.A. and Kolstad, C.D. 2010. Who pays a price on carbon? *Environmental Resource Economics* 46: 359-376.

Hassett, K.A., Mathur, A., and Metcalf, G.E. 2009. The incidence of a U.S. carbon tax: a lifetime and regional analysis. *Energy Journal* 30(2): 155-177.

Jones, C. and Kammen, D.M. 2014. Spatial distribution of U.S. household carbon footprints reveals suburbanization undermines greenhouse gas benefits of urban population density. *Environ. Sci. Technol.* 48: 895-902.

Koenker, R.W. and Bassett, G.W. 1978. Regression quantiles. *Econometrica* 46: 33-50.

Lattanzio, R.K. 2014. Canadian oil sands: life-cycle assessments of greenhouse gas emissions. Congressional Research Service, March.

Ridgeway, G. 2013. gbm: generalized boosted regression models. R package version 2.1-0.3.

Schumann, E. 2009. Generating correlated uniform variates. COMISEF wiki: http://comisef.wikidot.com/tutorial:correlateduniformvariates/

Shammin, M.R. and Bullard, C.W. 2009. Impact of cap-and-trade policies for reducing greenhouse gas emissions of U.S. households. *Ecological Economics* 68: 2432-2438.

Stratton, R.W., Hsin, M.W. and Hileman, J.I. 2010. Life cycle greenhouse gas emissions from alternative jet fuels. PARTNER Project 28 Report, Partnership for AiR Transportation Noise and Emissions Reduction, June.

Venkat, K. 2011. The climate change and economic impacts of food waste in the United States. *Int. J. Food System Dynamics* 2(4): 431-446.

Weber, C.L. and Matthews, H.S. 2008. Food-miles and the relative climate impacts of food choices in the United States. *Environ. Sci. Technol.* 42: 3508-3513.

Weidmann, T. 2009. A review of recent multi-region input-output models used for consumption-based emission and resource accounting. *Ecological Economics* 69(2): 211-222.

Williams III, R. C., Gordon, H., Burtraw, D., Carbone, J. C., and Morgenstern, R. D. 2014. The initial incidence of a carbon tax across income groups. Discussion Paper 14-24, Resources for the Future, Washington D.C.

Annex

GHG emission factor details

Electricity

The EIA provides monthly average residential electricity prices at the state level.[20] However, intra-state electricity prices can exhibit considerable spatial variation.[21] To better capture these patterns, a dataset from NREL linking utility companies to zip codes along with additional EIA data on residential electricity revenues and deliveries for the same utilities is used to estimate average residential electricity prices for individual PUMA's in the year 2010.[22,23] The spatial variation observed in 2010 is used to downscale state-level, monthly prices to individual PUMA's over the study period.

Additional "upstream" emissions from associated construction, mining, processing, and transport are introduced via technology-specific emission factors weighted by each technology's share of subregion generation. The upstream emission factors are based on data from NREL's Life Cycle Assessment Harmonization Project.[24] The assumed upstream emission factors are (gCO2e/kWh):

Coal: 49

Natural gas and oil: 150

Nuclear: 12

Hydropower: 7

Solar: 42

Wind: 11

Biopower: 40

Geothermal: 40

---

[20] http://www.eia.gov/beta/api/qb.cfm?category=1012
[21] http://en.openei.org/wiki/File:2012_12_14_Electricity_Price-01.jpg
[22] http://en.openei.org/datasets/node/899
[23] http://www.eia.gov/electricity/data/eia861/
[24] http://www.nrel.gov/analysis/sustain_lcah.html

Gasoline, natural gas, heating oil, and LPG

The EIA provides monthly average residential fuel and retail gasoline prices at the state level (prices for heating oil and LPG during the heating season only). The assumed life-cycle GHG emission factors are (gCO2e per MJ LHV):

Gasoline: 92 (Burnham 2012; Lattanzio 2014)

Natural gas: 78 (Burnham 2012)

Heating oil: 83 (ICF International analysis of New York City heating oil supply[25])

LPG: 82 (Burnham personal communication)

Air travel

Analysis of data provided by the MIT Airline Data Project[26] on total operating revenue, passenger revenue, and jet fuel consumption over the period 2008-2012 results in an average *direct* emission factor of 1.153 kgCO2e per dollar of expenditure. This figure assumes fuel consumption accounts for 98% of the airline industry's operational GHG footprint[27], and jet fuel exhibits a life-cycle GHG emission factor of 87.5 gCO2e/MJ (Stratton 2010).

A comprehensive emission factor should also account for pollution attributable to capital formation in the airline industry (e.g. construction of airplanes). Analysis of data from Andrew and Peters (2013) suggests inclusion of capital formation increases the direct emission factor by ~17% to a final emission factor of 1.35 gCO2e per dollar of air travel expenditure (2012 USD).

Other consumption

Shammin and Bullard (2009) calculate an emission factor for "Public transportation" inclusive of air travel. In this study, air travel is treated separately. Consequently, the original emission factor is reduced by ~80% prior to re-scaling to reflect the removal of GHG-intensive air travel from the category. The reduction is based on observed U.S. airline jet fuel consumption and passenger revenue in 2003 (Shammin and Bullard's base year) and the fact that total air travel expenditures are about twice that of other public transportation spending.

---

[25] http://www.nyc.gov/html/planyc2030/downloads/pdf/nyc_combined_natural_gas_report.pdf
[26] http://web.mit.edu/airlinedata/www/default.html
[27] https://www.united.com/web/en-US/content/company/globalcitizenship/environment_faq.aspx

Raking household weights for specific geographic regions

The PUMS only identifies individual households by PUMA. In order to analyze patterns for other geographic entities, it is necessary to assign households new sample weights that reflect the likelihood of the household being located in the desired entity.

Linkage between PUMA's and other geographic entities is provided by the Missouri Census Data Center's MABLE/Geocorr12 system.[28] This provides population-based weights for allocating a PUMA's population to a particular entity. An initial estimate of a household's revised sample weight is simply the product of the original weight and the allocation weight provided by MABLE.

Since the characteristics of households may vary considerably within a PUMA, the initial revised weights are then adjusted (or "raked") to ensure that the entity-specific sample resembles the true population in key respects (i.e. to create a balanced sample). Census Bureau "summary files" provide a description of the true population for a geographic entity. Summary files are derived from full samples of Census data (either 2008-2012 ACS or 2010 Census, in this case) and so provide the most complete information about a specific entity (e.g. a zip code). Two summary files are used: the ACS B19037 and 2010 Census H16. The former provides household marginal counts by income group, householder age, and race; the latter by housing status (owner or renter), household size, and race.

For a given entity, the initial revised weights are then iteratively "raked" until the marginal household counts of the PUMS-derived subsample (closely) match those of the actual population (i.e. summary files). This provides a defensible way of assigning surveyed households to individual, small-scale geographic entities, even though the native PUMS provides only a moderate level of geographic detail.

[28] http://mcdc.missouri.edu/websas/geocorr12.html