What is the shape of Environmental Engel Curves? Evidence using Panel Data*

Stefan Wöhrmüller

This draft: November 2023

Preliminary - Comments welcome
Please do not quote or redistribute without permission

Abstract

I combine multiple data sources to construct a novel panel dataset in which I observe inter alia US households' income, wealth, and expenditure and greenhouse gas (GHG) consumption. With this dataset, I estimate households' income elasticity with respect to GHG along the income distribution. In other words, I estimate Environmental Engel curves (EECs). I show that making use of the panel dimension of the data, in particular controlling for time-invariant household specific effects, suggests that EECs are flatter and more linear, and reduces the estimated income elasticity for all levels of income. Moreover, I show that these elasticities varies by consumption category. The results imply an attenuated form of the equity-pollution dilemma and suggest differentiated carbon taxes on consumption goods.

Keywords: Environmental Engel curves, greenhouse gas consumption, carbon taxation, environmental economics

JEL codes: C23, D12, D31, Q50

*I am very grateful to Christian Stoltenberg for his comments and advice. I would also like to thank Maurice Bun, Thomas Douenne, Eva Janssens, Franc Klaassen, Davide Pace, Elisabeth Proehl, Timo Schenk, Konstantin Sommer, Evgenii Vladimirov, and Quint Wiersma and other participants in seminars and conferences for many helpful comments and conversations. All errors are my own. *Email:* s.h.p.woehrmueller@uva.nl

1 Introduction

Environmental Engel curves (EECs) describe how household consumption of emissions or pollution (inherent in goods) varies with its income. Their shape has crucial welfare implications: Jacobs and van der Ploeg (2019) show that , under mild conditions, the optimal carbon tax is equal to the Pigouvian level as long as EECs are linear. This holds true even when poorer households spend a larger fraction of their income emissions or pollution and the government values redistribution On the other hand, if EECs are non-linear, the optimal carbon tax needs to be adjusted and might be lower than the Pigouvian level.

Existing approaches to estimate the shape of EECs based on household micro-data rely on repeated cross-sections (Weber and Matthews, 2008; Levinson and O'Brien, 2019; Sager, 2019; Zhang, Shi, Wang, Xue, Song and Sun, 2020). In these studies, researchers typically regress their environmental measure of interest such as greenhouse house gas (GHG) or air pollution consumption on income as well as socioeconomic characteristics. Since EECs are structural relationships between income and consumption, keeping prices fixed, an implicit assumption in these regressions is that there is no omitted variable problem. Otherwise, potential omitted variables, such as household preferences that are correlated with income and consumption, would nullify the structural interpretation of the EECs and introduce bias.

This paper constructs a novel panel dataset which allows the inclusion of household specific effects in the estimation regression and thus to control for time-invariant household characteristics. The main goal is to preserve the structural interpretation of the EEC and to reduce potential endogeneity issues in its estimation, which is not possible with pooled data or a repeated cross-section. For instance, recent studies find evidence that a codependence between income and consumption preferences exist (Alan, Browning and Ejrnæs, 2018; Arellano, Blundell, Bonhomme and Light, 2023). The panel structure of the data, however, to the extent that these preferences do not change over time, can take care of this unobserved omitted variable. Ultimately, the research question of this paper is: What is the shape of Environmental Engel Curves?

I find that, when including household fixed effects to control for time-invariant unobserved omitted variables, EECs are flatter and closer to linearity than standard OLS estimates, which rely on cross-sectional data. Moreover, these estimates also imply that the gap in the income elasticity of GHG consumption between poor and rich households is smaller. Overall, this suggests both that the optimal carbon tax is closer to a Pigouvian tax and that progressive redistribution of income gives rise to a smaller increase aggregate GHG consumption than previously considered.

Construction of the dataset(s) The key idea to construct this novel dataset is to create a mapping between two US household surveys - the Consumer Expenditure Survey (CEX) and the Panel Study of Income Dynamics (PSID). While the idea to impute consumption expenditures itself is not new, but has been used, among other things, to study the pass-through of income to consumption inequality (Blundell, Pistaferri and Preston, 2008), my paper is the first to use it for greenhouse gas consumption to get a better understanding of the shape of EECs. On a broad level, I implement two data-linking steps. First, I create a CEX dataset that contains the GHG emitted when producing the households' consumption basket using the methodology of Levinson and O'Brien (2019), Second, I compute GHG coefficients (GHG consumption per dollar spent) on the household level and impute this measure to the PSID. Given that the PSID has introduced a measure of consumption expenditure in 1999, I can then compute household GHG consumption by multiplying the imputed GHG coefficient with observed consumption.

The imputation methodology I use is multiple hot deck imputation as proposed in Cranmer and Gill (2013). The idea is to compute a distance measure, called "affinity score", between observable household demographic characteristics in the CEX the PSID. For every observation in the PSID, the procedure chooses the observation in the CEX with the highest affinity score and matches the corresponding GHG coefficient. It is possible that several observations in the CEX match. If this is the case, the CEX observation is chosen randomly. Therefore, I repeat this imputation procedure many times and eventually generate multiple (distinct) datasets. I validate the imputation procedure and warrant the comparability between PSID and CEX based on demographic characteristics as well as expenditures. Matching consumption survey data in this way is the main methodological contribution of this paper.

The final panel datasets contain the standard PSID variables as well as the measure for greenhouse gas consumption, which is different for every imputed dataset. In particular, it allows me to study the relationship between total greenhouse gas consumption and income, that is carving out the shape of the EEC, while controlling for both i) time-invariant household specific effects and ii) wealth of households. To the best of my knowledge, this paper is the first to do so using micro-data.

In descriptive terms, I find that three expenditure categories account for almost all of total GHG consumption of US households: housing (incl. utilities), transportation, and food expenditure. In sum, they make up approximately 95% of measured GHG and this number is very robust in all imputed datasets. Vacation expenditure comes fourth with approximately 3%.

Estimation and results I follow the literature and estimate parametric EECs using a model which is linear-quadratic in income. Using the imputed data with this methodology and estimating it with OLS, I recover three established facts (Levinson and O'Brien, 2019; Sager, 2019). First, EECs are upward sloping: richer households consume more GHG in absolute terms. Second, EECs are concave: relative to their income, richer households consume less GHG. Third, EECs shift down over time: household consumption baskets become greener over time. I interpret these results as further validation of my imputation strategy.

The panel data enables me to examine the difference between estimating the linear model with ordinary least squares (OLS) compared to fixed effects (FE) estimation. I estimate the model with both methods for all datasets and find that the FE estimates of the coefficient on (linear) income are about half of the OLS estimates. Moreover, the FE estimates on quadratic income are, on average and in absolute terms, also smaller than the OLS counterparts, however there is some overlap in the distribution of estimates.

These results suggest that the bias of omitting time-invariant variables is pushing up OLS estimates, making EECs steeper and more concave. However, it has been recognized in the literature that FE estimation amplifies measurement error and thus attenuation bias, that is, a bias of coefficients towards zero (Griliches and Hausman, 1986). To exclude this possibility, I provide suggestive evidence based on the comparison of OLS, first difference (FD), and FE estimates as well as higher order differences that this is not the case in my framework. In particular, theory suggests that the bias is stronger for FD estimates than FE estimates. However, I find that the former coefficients are larger than the latter, i.e. less close to zero. Moreover, the results from higher order differences are not significantly different from each other.

Based on the estimated model, I compute income elasticities of GHG consumption. I find elasticities below one over the entire income distribution: GHG consumption is a necessity. Again, I compare OLS and FE estimates over all imputed datasets. According to the (averaged) OLS estimates, households earning income of 51,000$ (expressed in 2012 dollars) have an income elasticity of about 0.27, compared to the (averaged) FE estimates of about 0.12. Income elasticities for rich households (> 300,000$) flatten out at 0.6 (OLS) and at 0.33 (FE), respectively.

While the elasticity estimates in the OLS case seem to be quite robust over the entire range of income, variation over different datasets is larger when the model is estimated using FE. For instance, going back to the household with 51,000$ in earnings, the estimated 10th percentile elasticity over all datasets is 0.09, while the 90th percentile is about 0.15. This difference is increasing with income and reaches 0.21 for households with 500,000$ in earnings. For comparison, the difference in percentiles for the OLS estimates at this level of income is 0.02.

An immediate policy implication of these elasticity estimates refers to the "equity pollution dilemma" (Heerink, Mulatu and Bulte, 2001; Sager, 2019): redistributing income from rich to poor households might increase aggregate GHG emissions. Under the FE estimates, however, this dilemma is attenuated as the difference in elasticities between poor and rich income households is not as large as previously estimated. Moreover, redistributing carbon tax revenue to poorer households to alleviate its regressive impact would have weaker second-round effects on aggregate emissions.

Finally, I study income elasticities for different GHG consumption categories to examine potential heterogeneities in expenditure patterns. Thereby, I focus on the three main categories housing, transportation, and food. Both housing and transportation exhibit a hump-shaped income elasticity, whereas the income elasticity of food is increasing for all income levels. Quantitatively, the elasticity on transportation is lower than housing and food, respectively.

Knowing the income elasticities of different consumption goods is informative for policy makers when they weigh equity and efficiency concerns of carbon taxation, which are present under non-linear Engel curves (Jacobs and van der Ploeg, 2019). In other words, from a non-environmental point of view, carbon tax differentiation based on the type of product might be optimal. General arguments of commodity tax differentiation apply.

Related literature The present paper relates to a burgeoning literature that estimates EECs from household micro-data. My paper builds most notably on Levinson and O'Brien (2019) who connect CEX data to IO tables and thus firm emissions data in a tremendous data linking effort. They use their dataset to estimate EECs with respect to air pollution measures and decompose US households' shift to greener products into an income and substitution effect. Sager (2019) also uses their methodology, but focuses on CO_2 emissions, and studies the link between income inequality and aggregate emissions. The main contribution of this paper is to utilize a panel data framework. Starting from Levinson and O'Brien (2019)'s data building approach, I link the CEX data to the PSID to study the relevance of household fixed effects when estimating EECs.

The methodological contribution lies in the construction of the panel dataset. My strategy is inspired by the estimation of demand systems to connect the CEX and PSID, which has been employed to study the transmission of income shocks to consumption inequality (Blundell et al., 2008). Instead of imputing consumption expenditures directly into the PSID, I can rely on its newly introduced consumption series and impute the GHG intensity of consumption using non-parametric imputation techniques (Cranmer and Gill, 2013). Hence, this paper joins ranks with several other articles that complement existing data sources to account for the missing rich in survey datasets (Nabernegg, Nabernegg and Kopp, 2023), expenditure under-reporting (Hardadi, Buchholz and Pauliuk, 2021), or people living under poverty (Bruckner, Hubacek, Shan, Zhong and Feng, 2022).

Outline The paper is structured as follows. Section 2 discusses estimation of EECs in more detail. Section 3 describes the data, the imputation strategy, and the final sample. Section 4 presents parametric estimates of EECs and compares OLS with FE estimation. Section 5 presents how the income elasticity of GHG consumption varies by income and product category. Section 6 concludes.

2 Estimating Environmental Engel Curves

A common approach to estimate Environmental Engel Curves parametrically is to run the following regression on repeated cross-sectional data (Levinson and O'Brien, 2019; Sager, 2019):

$$G_{it} = \beta_{1,t} Y_{it} + \beta_{2,t} Y_{it}^2 + \mathbf{X}'_{it}\delta + \nu_{it}, \tag{1}$$

where G_{it} is the environmental measure of interest such as greenhouse house gases or air pollution, Y denotes income, and $\mathbf{X}$ is a vector of control variables. The model variables and coefficients have subscript t, because the model is estimated separately for each period. The coefficients of interest are $\beta_{1,t}$ and $\beta_{2,t}$.

The interpretation of the EECs is a structural one: Holding prices constant, an increase in income will give rise to a change in expenditure on G as estimated in Equation (1). Implicit in this interpretation is that G in the EEC as represented above is an endogenous variable, whereas, in particular, Y is considered exogenous (Cameron and Trivdei, 2005, p.21). This implies that the linear model does not suffer from omitted variable bias.

A potential culprit in this respect, as also discussed by Sager (2019), is heterogeneity in household consumption preferences. These (unobserved) preferences, if correlated with income, would constitute an omitted variable that would bias estimates in the linear model eq. (1), rendering the structural interpretation moot. In fact, recent studies find evidence in this respect and stress or model a codependence between income and preferences (Alan et al., 2018; Arellano et al., 2023).

The main contribution of this paper is to make progress in this direction and to include household specific effects in Equation (1). I am able to do so because of the panel structure of a novel dataset that I construct (see below). Related to the discussion in the previous paragraph, the rationale of doing so is to control for potential time-invariant omitted variables which are correlated with consumption and income. For instance, my approach can control for household consumption preferences or any other unobservables, to the extent that these do not change over time.

3 Data

This section consists of four parts. First, I will describe the different data sources, which are environmental data from the Environmental Protection Agency (EPA) in the US, and household survey data from the Consumer Expenditure Survey (CEX) and the Panel Study of Income Dynamics (PSID). I briefly lay out how to connect the CEX to the environmental data to construct household consumption of GHG. Second, I will describe how I map the CEX dataset to the PSID. Third, I will describe the main variables in my analysis and sample selection. Fourth, I show summary statistics from my main sample.

3.1 Data sources and data preparation

Environmental data The collection of the environmental data follows Levinson and O'Brien (2019) with two minor modifications.[1] First, instead of using pollution data to the construct the Environmental Engel curves, I use greenhouse gas data from EPA's Greenhouse Gas Reporting Program (GHGRP). Second, the benchmark year to match environmental and industry level production data is updated to 2012, since this is the first year for which both are available in the same year.

GHGRP The GHGRP collects greenhouse gas emissions data at the facility-level from various sources in the United States. Overall, about 8000 facilities are required to report their emissions every year. This paper uses "direct emissions data" which is reported on-site and includes approximately 50% of total emissions in the United States. In other words, direct emitters describe facilities that directly emit GHGs into the atmosphere, for instance, a power plant that burns natural gas or coal. Greenhouse gases are reported in units of metric tons of CO_2-equivalent.

Most importantly, the GHGRP allows to group facilities by six-digit North American Industrial Classification System (NAICS) which then makes it possible to map it to the Economic Census

[1] I thank Arik Levinson and James O'Brien for sharing their data and programs on concordances between datasets. Some of which were carried together manually with their subjective judgments on UCC-IO mappings. The quality of their documentation in the code is highly appreciated.

and the Census of Agriculture. As a result, I can construct *direct GHG coefficients* in analogy to the pollution coefficients by Levinson and O'Brien (2019). These coefficients measure the GHG intensity of production in each sector - identified by NAICS code - and are computed, for each sector, by dividing the total amount of GHGs emitted by the total value of production. Lastly, I can map each of the coefficients to a particular IO codes which is important for the next step: The mapping from production emissions to total emissions accumulated in final goods. Note that the direct GHG coefficients measure the environmental impact in production irrespective of the place in the supply chain; that is, for both intermediate and final goods. Using input-output tables from the BEA, however, I can compute *total GHG coefficients* which measure for each final consumption good - identified by IO codes - the amount of GHG emissions accruing over the whole supply chain when producing that particular good. Hence, the data contains total GHG emission intensities by IO code.

CEX To map GHG consumption to actual consumption expenditure, I use the Interview component of the CEX. This component is a rotating panel, which follows consumer units (CU), or households, for a period of 15 month and registers information about expenditure in interviews every three months. A strength of this dataset is the amount of detail with which information about expenditure categories is collected. A downside is that even though household are followed for multiple periods, income information is only elicited in the first and fourth interview (Heathcote, Perri and Violante, 2010).

The consumption categories in the CEX are classified by so-called universal classification codes (UCCs). To achieve a mapping between UCCs and IO codes, I rely on the manual concordance by Levinson and O'Brien (2019). I merge the total GHG coefficients from the GHGRP to the consumption data by IO code and can then compute the amount of GHG emissions needed to produce a given consumption basket for every household. In other words, I can compute to what extent a households' consumption basket is relatively clean or dirty. This dataset, which includes consumption expenditure and GHG emissions data on the household level is used for imputation below.

Language note Throughout the entire paper, I will refer to "a households' consumption of GHG" when speaking of the amount of GHG emitted when producing the households' consumption basket. The former is, of course, not technically correct, but facilitates description and avoids unnecessarily complicated sentences.

PSID The PSID is a panel survey of a representative sample of U.S. households. The survey is biennial since 1997 and collects a comprehensive set of consumption expenditure categories since 2005 (Li, Schoeni, Danziger and Kerwin Kofi, 2010). Moreover, it collects information about household members' demographics, income, and wealth. It has low attrition and high response rates (Andreski, Li, Samancioglu and Schoeni, 2014). I will use the 2005-2017 waves of the PSID. The PSID is backward-looking, hence these waves represent the years 2004-2016.

In the following, I will explain how I connect the CEX with its information on GHG consumption to the PSID to create my final panel dataset in which I observe inter alia income, wealth, expenditure, and the amount of greenhouse gases needed to produce a given amount of consumption goods.

3.2 Mapping between CEX and PSID

The mapping between the CEX and the PSID proceeds in four steps. In the first step, I construct GHG coefficients with Levinson and O'Brien (2019)'s strategy and use them to compute GHG consumption on the UCC level. I then aggregate it to the nine categories observed in the PSID. In the second step, I compute household GHG coefficients in the CEX for all nine PSID categories. In the third step, I impute the GHG coefficients from the CEX to the PSID, separately for each consumption category and year. I do this using *multiple hot deck imputation* as proposed in Cranmer and Gill (2013). In the fourth step, I compute GHG consumption in the PSID by multiplying observed consumption expenditures and the imputed GHG coefficient.

Step 1 The CEX data set has both expenditures for goods and services and the amount of GHG necessary to produce them ((Levinson and O'Brien, 2019)). Therefore, to find the level of GHG associated with consumption of each UCC, I multiply the total GHG coefficient, which measures tons of CO_2-equivalents per dollar spent, with the expenditure on UCC consumption goods. Formally, denote the total GHG coefficient for UCC goods by ζ^{UCC} and expenditure on these goods by Z_t^{UCC}. Then for each household i in the CEX, GHG "consumption" by UCC is

$$G_{it}^{CEX,UCC} = \zeta^{CEX,UCC} Z_{it}^{CEX,UCC}. \tag{2}$$

In the CEX, the UCCs are fairly disaggregated. For instance, in 2009 there are over 500 different UCCs, while the PSID has 32 sub- and 9 main consumption categories. Going forward, I will focus on the nine main categories. Hence, to compare consumption in the CEX and the PSID, I rely on a mapping by Andreski et al. (2014) who match multiple UCCs to the PSID categories.[2] It is hence possible to aggregate UCCs from the CEX into consumption categories from the PSID. To be more precise, denote by $\mathcal{C}$ the set of consumption categories in the PSID, with $\mathcal{C} = \{$*Food, Housing, Transportation, Education, Childcare, Healthcare, Clothing & Apparel, Trips & Vacations, and Recreation & Entertainment*$\}$. For each $c \in \mathcal{C}$, sum over all UCCs in the CEX that make up category c:

$$G_{it}^{CEX,c} = \sum_{UCC \in c} G_{it}^{CEX,UCC} \quad \text{and similarly} \quad Z_{it}^{CEX,c} = \sum_{UCC \in c} Z_{it}^{CEX,UCC} \tag{3}$$

For instance, UCC 790410 ("Food prepared by consumer unit on out-of-town trips") and UCC 790240 ("Food and non alcoholic beverages"), among others, both contribute to the category "Food" in the PSID.

[2] I thank Patricia Andreski, Geng Li, Mehmet Zahid Samancioglu, and Robert Schoeni for providing this mapping. See their Table 3 in the Online Appendix.

Step 2 We can now back out the GHG coefficient for each PSID category:

$$\zeta_{it}^{CEX,c} = \frac{G_{it}^{CEX,c}}{Z_{it}^{CEX,c}}. \tag{4}$$

Equation (4) shows that the GHG coefficient $\zeta_{it}^{CEX,c}$ varies on the household level. Why is this the case when $\zeta^{CEX,UCC}$ in Equation (2) did not? The reason is that $\zeta_{it}^{CEX,c}$ now captures differences in households' consumption baskets (in terms of UCC goods).

Step 3 Assume there is a function $f : \mathbb{R}^m \to \mathbb{R}$ which translates demographic characteristics, $\mathbf{X}$, into the CEX-based GHG coefficient:

$$\zeta_{it}^{CEX,c} = f(\mathbf{X}_{it}^{CEX}).$$

To impute the GHG coefficient into the PSID dataset, I approximate this function and apply it on the *same* vector of demographic characteristics:

$$\zeta_{it}^{PSID,c} = \hat{f}(\mathbf{X}_{it}^{PSID}). \tag{5}$$

How do we compute $\hat{f}$? The strategy I use in this paper is called multiple hot deck imputation (Cranmer and Gill, 2013). The idea is to compute a distance measure, called "affinity score", between $\mathbf{X}_{it}^{CEX}$ and $\mathbf{X}_{it}^{PSID}$, which hold the same set of observables. For every observation in the PSID, the procedure chooses the observation in the CEX with the highest affinity score and matches the corresponding GHG coefficient.

It is possible that several observations in the CEX achieve this highest score from which the GHG coefficient should be matched. If this is the case, the CEX observation is chosen randomly. Given the size of the CEX data, having multiple best affinity scores it the rule, rather than the exception. Hence, I repeat the imputation procedure D times, essentially creating a series of functions, $\{\hat{f}_d\}_{d=1}^{D}$. Eventually, using D different $\hat{f}_d$'s on $\mathbf{X}_{it}^{PSID}$ generates D (distinct) PSID datasets that exclusively differ based on the imputed set of GHG coefficients $\{\zeta_{d,it}^{PSID,c}\}_{d=1}^{D}$.

Step 4 For each of these D datasets, I compute household GHG consumption of category c from the imputed GHG coefficient and the *observed* level of expenditure:

$$G_{d,it}^{PSID,c} = \zeta_{d,it}^{PSID,c} Z_{it}^{PSID,c}. \tag{6}$$

3.2.1 Practical implementation

As controls in $\mathbf{X}^{CEX}$ & $\mathbf{X}^{PSID}$ I use the age of the household head, gender, the number of adults in the households, the number of children in the household, marital status, race, region, and education.[3] I recoded the numerical values in both the CEX and PSID to make all variables comparable. Figure B.7 in Appendix B shows histograms for comparison. Overall, the distributions of all variables have similar shape and a high degree of overlap. An exception is the age variable, for the CEX shows an increasing age gradient and the PSID a decreasing one. In total, I set $D = 100$ and run the imputation separately for each wave in the PSID. Since the PSID is biennially I group adjacent years in the CEX; for instance, I pool year 2004 and 2005 in the CEX to compare it to the 2005 wave of the PSID.

3.2.2 Discussion of the imputation procedure

I want to address two points in more detail. First, a comparison of expenditure levels in the CEX and the PSID. Second, the validity of the imputation procedure using my datasets.

Even though I impute the GHG consumption *coefficients*, I also want to verify that the consumption measure is comparable in both dataset. Hence, Figure B.6 compares the expenditure levels of all 9 PSID consumption categories, converted to 2012-$ and adjusted using the OECD equivalence scale. The CEX UCCs have been aggregated accordingly (Andreski et al., 2014). The figure shows that the expenditure levels, excluding outliers, have very similar distributions. This is especially true for food, housing, and transportation, which are the embody the bulk of GHG consumption (see Figure 1 below).

Turning to my second point, Equation (5) implicitly assumes that heterogeneity in $\zeta_{it}^{PSID,c}$ is fully captured by my set observables, $\mathbf{X}$. The idea is that variation present in the UCCs in the CEX, which is lost in the aggregation Equation (3), is (partially) recovered by these demographic characteristics. Since Cranmer and Gill (2013)'s procedure is non-parametric in nature, I do not impose any functional form on $\hat{f}$ to maintain enough flexibility when recovering this variation.

Finally, the hot deck imputation procedure is designed to impute discrete data. However, my GHG coefficient is continuous. While the authors claim that the procedure works also well for continuous data, I validated the performance on CEX data. In particular, I used data from the year 2004, computed total GHG consumption coefficients, and randomly set 20 percent of this variable to missing. I then used $\mathbf{X}^{CEX}$ to impute these missing coefficients, just as in my actual imputation, however now I can compare them to the "true" coefficients. Regressing imputed coefficients on true coefficients, without a constant, yields a coefficient of 0.95 and an R^2 of 0.92 (see also Figure C.8).

3.3 Definitions and sample selection

The following definitions of income, wealth, and consumption are based on the original PSID data. All variables are transformed into 2012-$.

Income I use post-tax household income in my regressions. I define household income as the sum of head and spouse's labor income, transfers, social security income, and food stamps net of taxes (computed with NBER's TAXSIM program).

[3] Hence, note that the demographic characteristics only contain categorical variables, even though this must not necessarily be the case, given our definition of f.

Table 1: Summary statistics

	Mean	Std. dev.		Mean	Std. dev.
Income (10000 2012-$)	6.51	5.72	Wealth (10000 2012-$)	16.98	32.79
Age	41.98	10.28	Family size	2.84	1.46
Gender	0.80	0.40			
Marital status			*Education*		
Married	0.63	0.48	Elementary only	0.10	0.30
Never married	0.19	0.40	High school	0.28	0.45
Widowed	0.01	0.11	Some college	0.27	0.44
Divorced	0.14	0.35	College	0.21	0.41
Separated	0.02	0.14	More than college	0.14	0.35
Race of household head			*Region*		
White	0.88	0.33	Northeast	0.16	0.37
Black	0.10	0.30	Midwest	0.30	0.46
Other	0.02	0.13	South	0.34	0.48
Asian	0.00	0.06	West	0.19	0.40
Observations	20751				

Note. This table shows summary statistics for the benchmark sample from the Panel Study of Income Dynamics 2004-2016 as described in the main text. Income (net of taxes) and wealth is in units of 10000 and expressed in 2012 dollars.

Wealth A household's wealth is defined as the difference between assets and liabilities. Assets include checking and saving accounts, stocks, private annuities/IRAs, present value of one's house/apartment, and vehicles. Liabilities include mortgages, credit card debt, student debt, medical bills, legal debt, and loan from relatives.

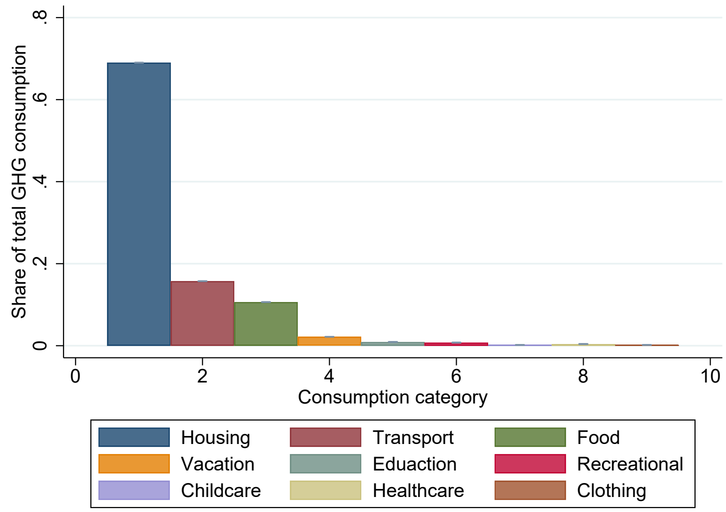
Consumption My consumption variable corresponds to all consumption categories as described above. That is, consumption is the sum of expenditure on food, housing, transportation, education, childcare, healthcare, clothing & apparel, trips & vacations, and recreation & entertainment.

Sample selection I select household heads between 25 and 60 years old. I drop households whose composition has changed throughout my observation period. Furthermore, I exclude observations whose i) total GHG consumption is negative ii) hourly wage is smaller than half the minimum wage iii) labor income is positive, yet have zero working hours iv) wealth is less than minus one million. Lastly, I winsorize the top and bottom 0.5% of my income and wealth variable to account for extreme outliers. Note that all sample selection criteria, except i), apply to the original non-imputed PSID dataset.

3.4 Final dataset(s)

Recall that the final datasets differ only by one variable, namely the imputed GHG consumption coefficient. Hence, Table 1, which shows summary statistics of key household characteristics, represents all D datasets. The average household earns 65,100$ (in 2012 dollars), owns 169,800$ (in 2012 dollars) in wealth, and contains three persons. The average household head is 42 years old, white, married, lives in the south of the US and has a high school degree.

Figure 1: Greenhouse gas consumption by PSID category



Note. This figure shows the share of greenhouse gas consumption across all nine consumption categories relative to the overall greenhouse gas consumption in the PSID dataset. Greenhouse gas consumption is imputed from CEX data as described in the main text.

Figure 1 shows our imputed GHG consumption measure as constructed in Equation (6). The height of the bar indicates mean GHG consumption shares over all datasets. We see that expenditure on housing, transport, and food almost make up all greenhouse gas emissions in consumption baskets of households. Moreover, the small standard errors indicate that there is little variation between datasets.

Finally, it is now possible to estimate the following linear model for every dataset d:

$$G_{d,it} = \beta_{d,1} Y_{it} + \beta_{d,2} Y_{it}^2 + \tilde{\mathbf{X}}'_{it} \delta_d + \alpha_{d,i} + \gamma_{d,t} + \epsilon_{d,it}. \tag{7}$$

The dependent variable is imputed GHG consumption summed over all categories, that is $G_{d,it} = \sum_{c \in \mathcal{C}} G_{d,it}^c$ measured in tons of CO_2-equivalents. As independent variables, I include income, Y, income squared, Y^2, and other covariates, $\tilde{\mathbf{X}}$. These covariates include family size, family size squared, age, age squared, gender of the household head as well as dummy variables for marital status, race, education, and region. Furthermore, I extend existing specifications and include wealth and wealth squared as control variables.

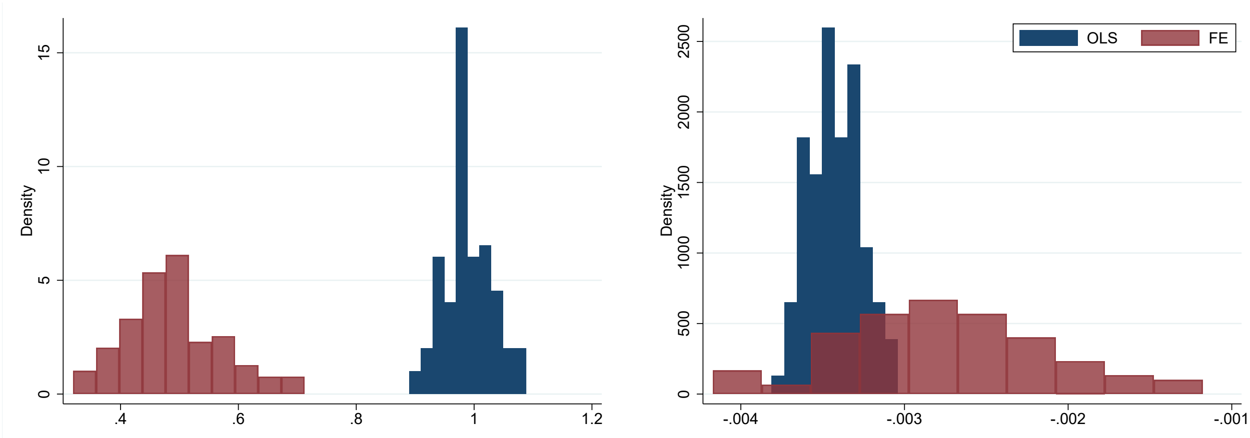
In the following, I will estimate Equation (7) using both pooled OLS and OLS based on demeaned variables; that is, within estimation/fixed effects. For brevity, I will refer to the former as OLS and the latter as FE. Moreover, I will sometimes estimate the model for each dataset and then present the distribution of point estimates from all regressions and not show standard errors. The interpretation is akin to a bootstrap.

4 Results

4.1 Quadratic EECs

Figure 2: Regression estimates for all D datasets

(a) $\hat{\beta}_{d,1}$ (b) $\hat{\beta}_{d,2}$



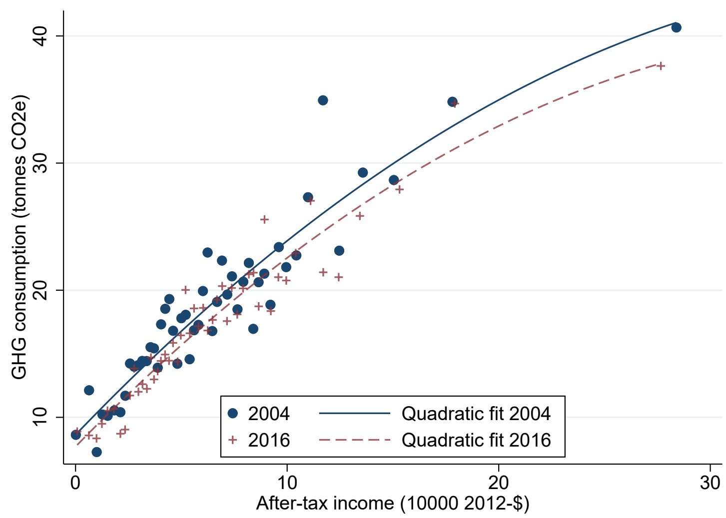
Note. This figure shows OLS (blue) and FE (red) estimates of $\hat{\beta}_{d,1}$ and $\hat{\beta}_{d,2}$ from the linear model eq. (7). The model is estimated for all imputed datasets such that there is an entire distribution of coefficient estimates in both cases.

The blue bars in Figure 2 show the estimates of both $\beta_{d,1}$ and $\beta_{d,2}$ for all datasets D when the linear model is estimated using standard OLS. Furthermore, Figure 3 shows the shape of Environmental Engel Curves for the years 2004 and 2016 (controlling for income (squared) only). From these figures, we can make the following three observations. First, the coefficient on income is positive. Second, the coefficient on income squared is negative. Third, Environmental Engel Curves shift down over time. Taken together, these three results confirm existing results in the literature (Levinson and O'Brien, 2019; Sager, 2019), further validating the imputation approach.

The red bars in Figure 2, on the other hand, show the estimates of both $\beta_{d,1}$ and $\beta_{d,2}$ when the linear model is estimated controlling for household fixed effects. The overall shape of the EEC stays unchanged: it is upward sloping and concave. However, the distribution of point estimates of $\beta_{d,1}$ suggests that the associated increase in GHG consumption due to an increase in income, ceteris paribus, is not as strong as suggested by the OLS regression. Moreover, the EEC seems to be less concave, as the $\beta'_{d,2}s$ are closer to zero, which would indicate a linear relationship. However, the contrast between the OLS and FE estimates is less pronounced for the coefficient(s) on the quadratic term, as the distributions overlap partly.

These results imply that (individual) EECs are i) flatter and ii) more linear as compared to the cross-sectional estimates. This suggests that with respect to the slope there is an upward bias from omitted time-invariant variables. Taking up preference heterogeneity again as a concrete example, it means that household consumption preferences are positively correlated with both income and GHG consumption.

Figure 3: Environmental Engel Curves of GHG consumption



Note. This figure shows Environmental Engel Curves from the first imputed PSID dataset for the years 2004 and 2016, where households have been grouped into 50 income bins, respectively. EECs are represented by the predicted value of GHG consumption from a linear regression of GHG consumption on income and income squared (quadratic fit). Income is net of taxes, in units of 10000, and expressed in 2012 dollars. Greenhouse gas consumption is imputed from CEX data as described in the main text.

4.2 Potential measurement error?

Including household-specific effects in the analysis comes at a potential cost. Griliches and Hausman (1986) show that attenuation bias, due to measurement error in the independent variable, is amplified in fixed effects estimation. Hence, the question arises whether attenuation bias is dominant and pushes my coefficients towards zero (compared to the OLS case).

To study this possibility, I follow a strategy by Griliches and Hausman (1986): "Calculate some differenced estimates (of different lengths) by OLS. If they differ significantly, errors in measurement may well be present" (p.114). Since there are no control variables in their framework, I residualize my variables of interest. In particular, denote by $M_{\tilde{\mathbf{X}},\gamma}$ the residual maker w.r.t the controls and time fixed effects, and denote residuals by lower case variables, that is, $M_{\tilde{\mathbf{X}},\gamma}G_{it} = g_{it}$ (analogous for income) and $M_{\tilde{\mathbf{X}},\gamma}\epsilon_{it} = \varepsilon_{it}$. I then run several regressions on these residuals for different orders of differences $\Delta^x var_{it} \equiv var_{it} - var_{it-x}$ for my quadratic specification in all datasets:

$$\Delta^x g_{d,it} = \beta_{d,1}\Delta^x y_{it} + \beta_{d,2}\Delta^x y_{it}^2 + \Delta^x \varepsilon_{d,it}. \tag{8}$$

Table 2 shows the results of this strategy. The estimated coefficients and standard errors reported here are averages over all datasets. For instance, $\overline{\beta}_1 = \frac{1}{D}\sum_{d=1}^{D}\hat{\beta}_{d,1}$. Standard errors are robust with respect to heteroskedasticity and clustered at the household level.

Table 2: Griliches and Hausman (1986) test regression(s)

Δ^x	$\overline{\beta}_1$	$\overline{\beta}_2$
1	0.532 (0.149)	-0.0156 (0.0123)
2	0.393 (0.146)	-0.0025 (0.0010)
3	0.523 (0.116)	0.0018 (0.0079)
4	0.821 (0.122)	0.0029 (0.0107)
5	0.819 (0.168)	0.0034 (0.0157)
6	0.622 (0.242)	-0.0094 (0.0131)
OLS	0.989 (0.084)	-0.0034 (0.0005)
FE	0.490 (0.130)	-0.0028 (0.0010)

Note. This table shows the average estimates of eq. (8) over all datasets for different orders of the difference operator Δ as well as average OLS and FE estimates. Average standard errors in parentheses are robust and in the case of FE estimation clustered at the household level.

The estimates give rise to two points why I am confident that my results are not driven by (FE-amplified) attenuation bias. First, most of the estimated coefficients for different orders of the difference operator Δ are not significantly different from each other. With respect to $\overline{\beta}_1$, there is some divergence for the fourth and fifth difference, however, the remaining ones agree in magnitude and in sign. With respect to $\overline{\beta}_2$, the coefficients do overlap, as standard errors are larger, but are not significantly different from zero. Hence, based on the first-difference estimates, one could not reject the linearity of EECs.

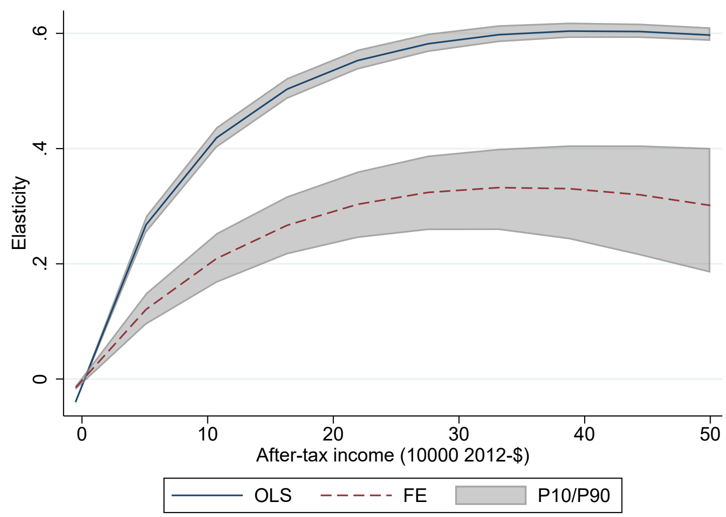
Second, Griliches and Hausman (1986) show that for a positively autocorrelated independent variable with a declining correlogram, such as income in our case, the attenuation bias of the first difference estimate when $T > 2$ should be larger than the FE estimate (see their equation (5)). However, this is not the case for both of the coefficient estimates.

5 Income elasticity of GHG consumption

Another way of studying the shape of EECs and of considerable importance for the public finance literature is to look at income demand elasticities. Hence, Figure 4 plots the income elasticity of GHG consumption based on eq. (7), when all other covariates are fixed at their mean value.[4] Elasticities are evaluated over an equally-spaced ten-point income grid ranging from the minimum income observation to 500,000 2012-$, the interval which covers the major part of the support of the income distribution (Figure D.9). The blue solid line depicts the mean elasticities under OLS for all datasets, whereas the red dashed line depicts the mean elasticities under FE estimation for all datasets. The grey-shaded area covers the 10th and 90th elasticity percentile, respectively, over all estimated elasticities.

[4] Think of a partial equilibrium exercise in which prices are fixed such that income elasticity of demand and income elasticity of consumption can be used interchangeably (Ghoddusi, Rodivilov and Roy, 2021)

Figure 4



Note. This figure shows estimates for the income elasticity of total GHG consumption based on eq. (7) when all covariates are fixed at their mean value. Elasticities are evaluated over an equally-spaced ten-point income grid ranging from the minimum income observation to 500,000 2012-$. The blue solid line shows the mean (over all imputed datasets) elasticities when eq. (7) is estimated using OLS, and the red dashed line shows the mean (over all imputed datasets) elasticities when the model is estimated using FE. The grey-shaded area covers the 10th and 90th elasticity percentile, respectively, over all estimated elasticities.

Three observations stand out. First, both OLS and FE estimates are well below 1 for the entire range of considered income levels. Hence, GHG can be considered a necessary good. Second, however, the elasticity estimate under FE is approximately half that of the estimate under OLS, while the shape of the elasticity curve is largely similar. Third, the FE estimate exhibits a larger band of uncertainty. Obviously, the fact that OLS estimates are larger is an extension from the larger coefficient estimates (Table 2).

These new elasticity estimates give rise to two policy implications. First, the equity-pollution dilemma would be attenuated. This dilemma describes that redistributing income from rich to poor households might increase aggregate GHG emissions, as poorer households have larger marginal propensities to consume (MPC) with respect to GHG consumption (Heerink et al., 2001; Sager, 2019). But note that the income elasticities of consumption can be written as the product of the MPC and the consumption-to-income ratio. Keeping the latter fixed, the FE estimates imply a weaker equity-pollution dilemma, as the difference in elasticities, and hence MPCs, between poor and rich income households is not as large as previously estimated.

Second, and related to this argument, redistributing carbon tax revenue to poorer households to alleviate its regressive impact would have weaker second-round effects on aggregate emissions. This is an immediate implication of smaller elasticities at the lower end of the income distribution (Figure 4).

5.1 Income elasticities of different GHG consumption categories

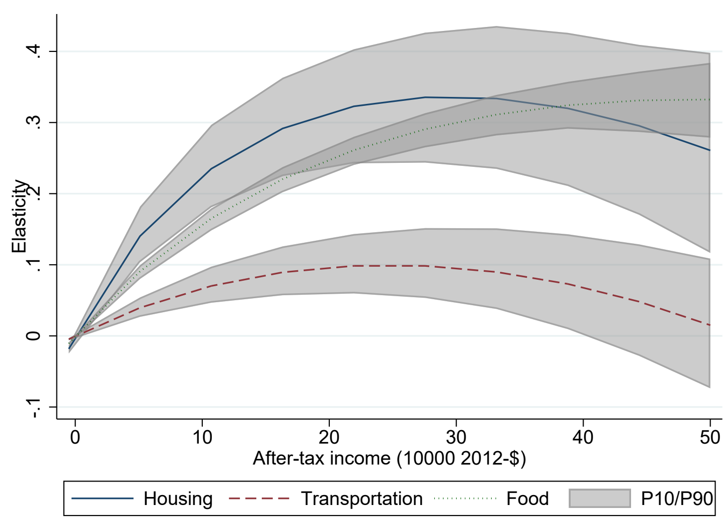
The analysis so far has focused on a total measure of GHG consumption. But we saw in Figure 1 that only a subset of categories drive the total GHG measure. The next decomposition helps us understand to what extent individual categories contribute to the overall income elasticity.

Decomposition 1.

$$\eta_{G,Y} = \sum_{c \in \mathcal{C}} \phi_c \eta_{c,Y},$$

where $\phi_c \equiv \frac{G^c}{\sum_{c \in \mathcal{C}} G^c}$ denotes the GHG consumption share and $\eta_{c,Y} \equiv \frac{\partial G^c}{\partial Y} \frac{Y}{G^c}$ denotes the income elasticity of GHG consumption of category c. Decomposition 1, which is derived in Appendix A, states that the income elasticity of total GHG consumption is equal to the expenditure-weighted elasticities of its components.

Figure 5



Note. This figure shows estimates for the income elasticity of three consumption categories, namely housing, transportation, and food GHG consumption, respectively. The estimates are based on an adjusted linear model eq. (7) where the dependent GHG variable is replaced by the respective category. Elasticities are evaluated over an equally-spaced ten-point income grid ranging from the minimum income observation to 500,000 2012-$. The blue solid line shows the mean (over all imputed datasets) housing elasticities, the red dashed line shows the mean (over all imputed datasets) transportation elasticities, and the green dotted line shows the mean (over all imputed datasets) food elasticities. The grey-shaded area covers the 10th and 90th elasticity percentile, respectively, over all estimated elasticities.

Hence, in the following, I compute different income elasticities as above for the three primary GHG categories, as they are the main contributors to the overall income elasticity. Figure 5 shows the elasticities for housing, transportation, and food; again, over the equally-spaced income grid. Both housing and transportation exhibit a hump-shaped income elasticity, whereas the income elasticity of food is increasing over the entire domain. The income elasticity of housing is subject to considerable uncertainty, whereas the ones for food and transportation have smaller uncertainty bands.

As indicated in the introduction, when Environmental Engel curves are non-linear, governments only have access to distortionary (piece-wise) linear income taxes and household preferences are separable in consumption and labor, carbon taxes do deviate from its Pigouvian level and potentially serve other goals than correcting the climate externality from emissions (Jacobs and van der Ploeg, 2019). In other words, governments should weigh equity and efficiency concerns when setting (potentially different) carbon taxes on consumption categories. Hence, knowing income elasticities of different consumption goods is provides insight for policy makers when carbon taxes also serve redistributive purposes or should help alleviate existing distortions in the economy. For instance, these elasticities are informative about the size of income effects that enter the social marginal utility of income in optimal tax conditions (Diamond, 1975; Jacobs, 2018).

6 Conclusion

In this paper, I have constructed a novel panel dataset by imputing information on household greenhouse gas consumption from the CEX to the PSID using multiple hot deck imputation. The imputation procedure results in multiple distinct datasets that can be used for econometric analysis. Equipped with these datasets, I am able to parametrically estimate Environmental Engel Curves, while controlling for time-invariant household specific effects, such as preference heterogeneity.

Taking into account these household effects (FE) results in flatter and more linear Environmental Engel Curves compared with pooled OLS. The mean (over all imputed datasets) income elasticity of GHG consumption under FE is well below one and about half as large as the one implied by OLS estimation. These elasticities differ by consumption category, as elasticities for GHG emissions in housing and transportation expenditure exhibit a hump-shaped income elasticity, whereas elasticities for GHG emissions in food expenditure is increasing for all levels of income. The estimates have important implications for carbon taxation and redistribution of its revenue.

Lastly, this paper stayed close to existing estimation models to establish the relevance of controlling for household fixed effects. There are other avenues, however, which are well worth exploring in future research. I want to mention two. First, the panel structure allows estimation of entire systems of consumer expenditure functions with better estimation, identification and hypothesis testing characteristics (Aasness, Biørn and Skjerpen, 1993). An application with respect to GHG intensive goods would be particular interesting. Second, survey data often under-represents the extreme rich and thus necessitates to exclude this part of the distribution from the econometric analysis. Incorporating the panel framework while, in addition, taking into account the upper tail of the distribution as in Nabernegg et al. (2023) would provide a more complete representation of Environmental Engel Curves.

References

Aasness, Jørgen, Erik Biørn, and Terje Skjerpen. 1993. "Engel Functions, Panel Data, and Latent

Variables." *Econometrica*, 61(6): 1395–1422.

Alan, Sule, Martin Browning, and Mette Ejrnæs. 2018. "Income and Consumption: A Micro Semistructural Analysis with Pervasive Heterogeneity." *Journal of Political Economy*, 126(5): 1827–1864.

Andreski, Patricia, Geng Li, Mehmet Zahid Samancioglu, and Robert Schoeni. 2014. "Estimates of Annual Consumption Expenditures and Its Major Components in the PSID in Comparison to the CE." *American Economic Review*, 104(5): 132–135.

Arellano, Manuel, Richard Blundell, Stéphane Bonhomme, and Jack Light. 2023. "Heterogeneity of consumption responses to income shocks in the presence of nonlinear persistence." *Journal of Econometrics*, 105449.

Blundell, Richard, Luigi Pistaferri, and Ian Preston. 2008. "Consumption Inequality and Partial Insurance." *American Economic Review*, 98(5): 1887–1921.

Bruckner, Benedikt, Klaus Hubacek, Yuli Shan, Honglin Zhong, and Kuishuang Feng. 2022. "Impacts of poverty alleviation on national and global carbon emissions." *Nature Sustainability*, 5(4): 311–320.

Cameron, Colin A., and Pravin K. Trivdei. 2005. *Microeconometrics: methods and applications.* Cambridge University Press.

Cranmer, Skyler J., and Jeff Gill. 2013. "We Have to Be Discrete About This: A Non-Parametric Imputation Technique for Missing Categorical Data." *British Journal of Political Science*, 43(2): 425–449.

Diamond, P. A. 1975. "A many-person Ramsey tax rule." *Journal of Public Economics*, 4(4): 335–342.

Ghoddusi, Hamed, Alexander Rodivilov, and Mandira Roy. 2021. "Income elasticity of demand versus consumption: Implications for energy policy analysis." *Energy Economics*, 95: 105009.

Griliches, Zvi, and Jerry A. Hausman. 1986. "Errors in Variables in Panel Data." *Journal of Econometrics*, 31: 93–118.

Hardadi, Gilang, Alexander Buchholz, and Stefan Pauliuk. 2021. "Implications of the distribution of German household environmental footprints across income groups for integrating environmental and social policy design." *Journal of Industrial Ecology*, 25(1): 95–113.

Heathcote, Jonathan, Fabrizio Perri, and Giovanni L. Violante. 2010. "Unequal we stand: An empirical analysis of economic inequality in the United States, 1967–2006." *Review of Economic Dynamics*, 13(1): 15–51.

Heerink, Nico, Abay Mulatu, and Erwin Bulte. 2001. "Income inequality and the environment: aggregation bias in environmental Kuznets curves." *Ecological Economics*, 38(3): 359–367.

Jacobs, Bas. 2018. "The marginal cost of public funds is one at the optimal tax system." *International Tax and Public Finance*, 25(4): 883–912.

Jacobs, Bas, and Frederick van der Ploeg. 2019. "Redistribution and pollution taxes with non-linear Engel curves." *Journal of Environmental Economics and Management*, 95: 198–226.

Levinson, Arik, and James O'Brien. 2019. "Environmental Engel Curves: Indirect Emissions of Common Air Pollutants." *The Review of Economics and Statistics*, 101(1): 121–133.

Li, Geng, Robert F. Schoeni, Sheldon Danziger, and Charles Kerwin Kofi. 2010. "New expendilture data in the PSID: comparisons with the CE." *Monthly Lab. Rev.*, 133: 29.

Nabernegg, Markus, Stefan Nabernegg, and Thomas Kopp. 2023. "Environmental Engel Curves with predicted consumption of high-income households, applied to Ecuador." *Manuscript*.

Sager, Lutz. 2019. "Income inequality and carbon consumption: Evidence from Environmental Engel curves." *Energy Economics*, 84: 104507.

Weber, Christopher L., and H. Scott Matthews. 2008. "Quantifying the global and distributional aspects of American household carbon footprint." *Ecological Economics*, 66(2-3): 379–391.

Zhang, Hongwu, Xunpeng Shi, Keying Wang, Jinjun Xue, Ligang Song, and Yongping Sun. 2020. "Intertemporal lifestyle changes and carbon emissions: Evidence from a China household survey." *Energy Economics*, 86: 104655.

Appendix

A Decomposition derivation

Start from the simple identity that total greenhouse gas emissions consumed are made up of different categories, each of which is a function of income:

$$G = \sum_{c \in \mathcal{C}} G^c(Y).$$

Differentiating yields

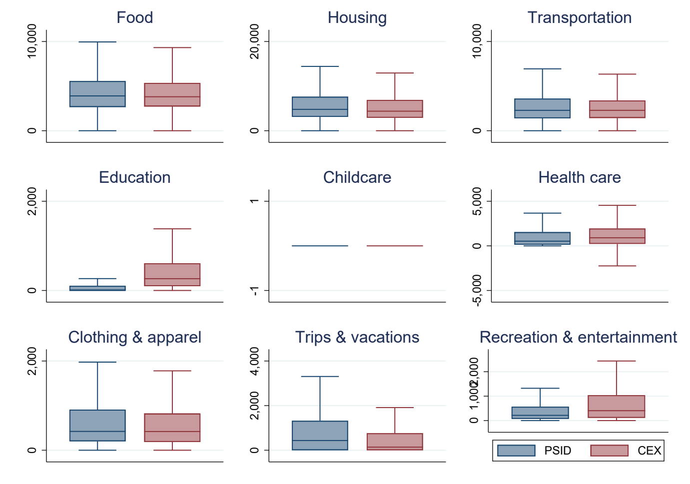
$$dG = \sum_{c \in \mathcal{C}} \frac{\partial G^c(Y)}{\partial Y} dY.$$

Division by dY, expanding terms, and employing the definitions of η and ϕ from the main text yields the decomposition:

$$\begin{aligned} \frac{dG}{dY} &= \sum_{c \in \mathcal{C}} \frac{\partial G^c(Y)}{\partial Y} \\ \underbrace{\frac{dG}{dY}\frac{Y}{G}}_{\eta_{G,Y}} &= \sum_{c \in \mathcal{C}} \underbrace{\frac{\partial G^c(Y)}{\partial Y}\frac{Y}{G^c}}_{\eta_{G^c,Y}} \underbrace{\frac{G^c}{G}}_{\phi_c} \end{aligned}$$

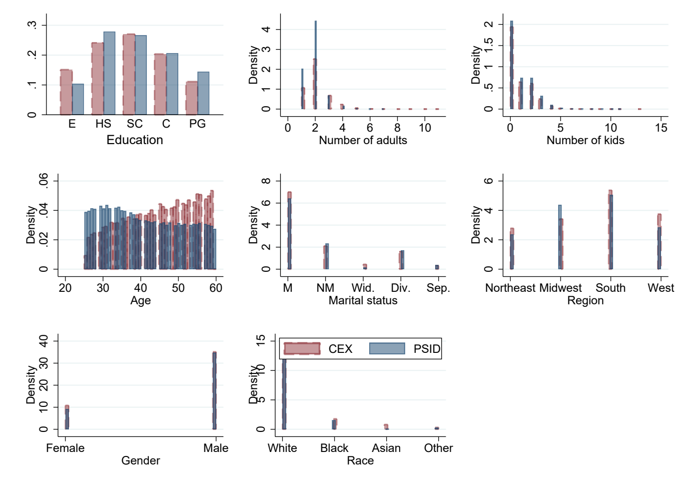
B CEX-PSID Comparison

Figure B.6: Expenditure comparison between CEX and PSID



Note. This figure compares consumption expenditures between the CEX and the PSID. Expenditures in both datasets have been aggregated to nine categories, adjusted using the OECD equivalence scale and are expressed in 2012-$. Outside values are omitted from the boxplots.

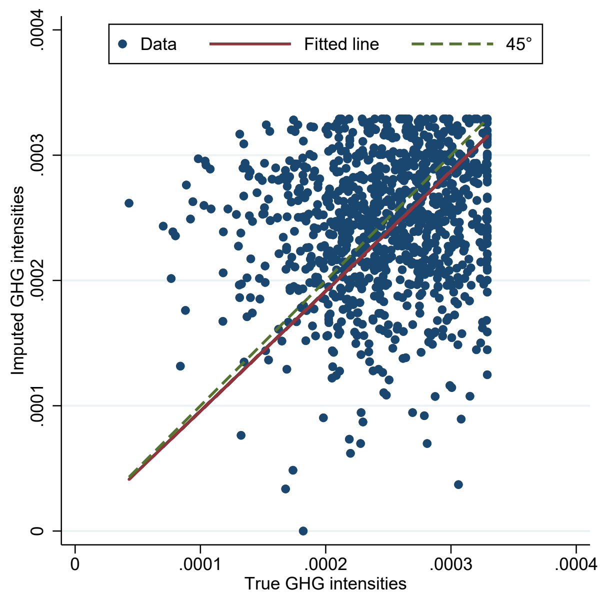
Figure B.7: Covariates comparison between CEX and PSID



Note. This figure compares the distribution of demographic characteristics between the CEX and the PSID. The depicted variables are the ones that enter $\mathbf{X}^{CEX}$ and $\mathbf{X}^{PSID}$, respectively, in the imputation exercise. Education categories are Elementary (E), High School (HS), Some College (SC), College (C), and Postgraduate (PG). Marital status categories are Married (M), Never Married (NM), Widowed (Wid), Divorced (Div.), and Separated (Sep.).

C Hot Deck Matching Validation

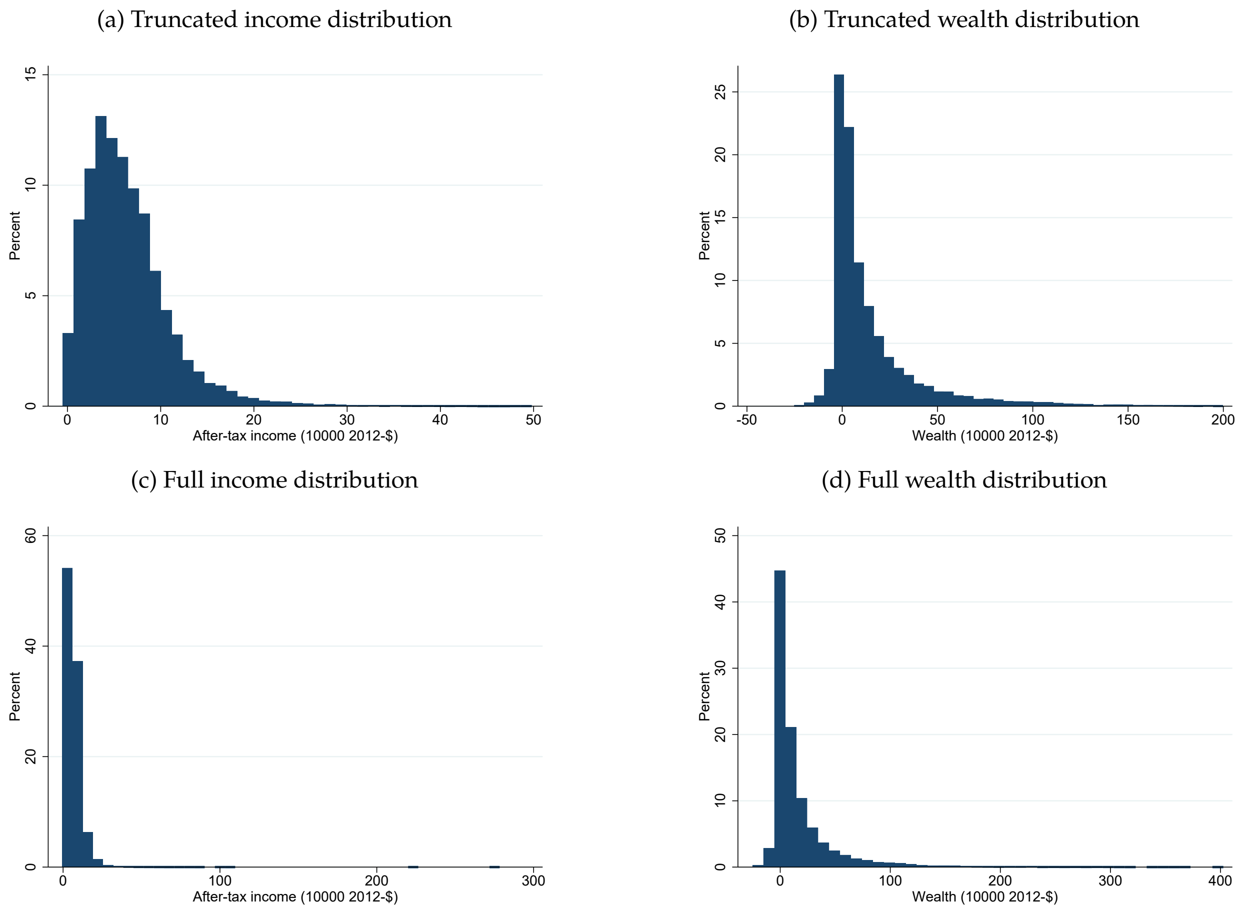
Figure C.8: Hot Deck Matching Validation



Note. This figure shows the relationship between the true GHG coefficients and the imputed GHG coefficients from CEX test data in the year 2004. 20% of the GHG intensities observations in this year have been randomly set to missing and then imputed based on the imputation procedure from the main text. The fitted line is based on a regression of imputed GHG intensities on true GHG intensities, excluding a constant.

D Additional Figures and Tables

Figure D.9: Distribution of income and wealth



Note. This figure shows the distribution of after-tax income and wealth in the PSID, respectively, under the benchmark sample. The top left panel shows the distribution of after-tax income when it is truncated at 500,000 2012-$. The top right panel shows the distribution of wealth when it is truncated at 2,000,000 2012-$. The bottom panels show the full distribution.