The London School of Economics and Political Science

Essays in Public and Environmental Economics

Nicolas Chanut

London, June 2022

Thesis submitted to the Department of Economics of the London School of Economics for the degree of Doctor of Philosophy.

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I declare that my thesis consists of approximately 48,604 words, including footnotes and appendices but excluding bibliography.

Acknowledgements

This work would not have been possible without the support I received from many people. I would like to take this opportunity to thank them.

My deepest appreciation goes out to my supervisors Xavier Jaravel and Philippe Aghion for their invaluable support, guidance, and for their insightful comments and suggestions throughout the PhD. They have been instrumental in the development of this thesis.

I would like to offer my sincere gratitude to the management of the French private retailer for allowing access to their data, which underlies the first two chapters of this thesis, and for their support in understanding the data. The OpenFoodFact team also provided great help and support in understanding their open source data. I am grateful to the ESRC, STICERD, Collège de France and RES for their financial support, and to the peaceful and friendly work environment provided by the Collège de France in Paris.

I am grateful to many other faculty members at LSE including Daniel Sturm, Camille Landais, Johannes Spinnewijn, Daniel Reck, Alan Manning and many others. Their supportive comments and feedback during various meetings and seminars played an important role in improving this thesis. I would also like to thank the professional services staff at the LSE, especially Mark Wilbor, Emma Taverner, Mike Rose and Hitesh Patel, for their outstanding administrative and technical support.

My friends and PhD colleagues Alexandre Desbuquois, Nicola Fontana, Maxi Guennewig, Thomas Minten, Derek Pillay, Bhargavi Sakthivel, Hugo Villaresh and many others have brightened up my PhD days. Outside LSE, I will fondly remember my intellectually and physically stimulating evening runs in Hyde Park with Hendrik Klaus, as well as my regenerating brunches and dinners with Theo Bourgery and Willy Bonneuil.

Finally, I would like to thank my friends back home and my family for their support on all possible fronts throughout my PhD years, especially my parents, grand-parents and my sister Mathilde. Above all, this journey would not have been possible without my wonderful wife and *co-équipière* Marie. I owe her everything for her unconditional trust and endless patience.

Abstract

Economists and policy makers are increasingly concerned with the heterogeneous impact of economic policies, on top of their overall effectiveness. However, understanding precisely how they impact different population groups often requires specific approaches and data sets. This thesis contributes to this issue by leveraging novel data sets and methods to improve the economic analysis of carbon emissions, the measurement of inflation and the measurement wage inequality.

The first chapter analyzes to what extent countries can reduce their CO2 emissions by changing the composition of consumption rather than the underlying technology, with a focus on food products. I document that carbon intensity is heterogeneous even within detailed product categories. I then show that well-targeted taco deliver large efficiency gains, and that the impact of carbon taxes across households varies strongly with their exposure to high-carbon products, but not with their expenditure level. In addition, the welfare cost of reducing carbon emissions varies strongly across product categories, providing a new justification for granular taxes or, equivalently, carbon markets.

The second chapter concerns the inflation dynamics during the Covid-19 lockdown in France. Using scanner data on fast-moving consumer goods from a large retailer, I find that the lockdown lead to an important, generalized but temporary increase in price levels across product categories. Further, I find that this inflation shock was asymmetric across products, households and cities, and that this asymmetry did not vanish on the medium-term.

The third chapter focuses on the contribution of firm heterogeneity to wage inequality and its measurement in two-way fixed effect models. I provide evidence that firm-side drivers of wage inequality are overestimated by at least 25% because of model overfitting. I then provide a simple procedure to recover the correct measures of interest, show that the correction matters quantitatively and derive more precise estimates of firm effects using shrinkage methods.

Contents

List of Tables

List of Figures

Chapter 1

Demand-side substitution and the welfare cost of reducing carbon emissions

Abstract

Can countries reduce their carbon emissions by changing the composition of consumption instead of, or in addition to, changing the underlying production technology? If so, what type of policy would be most effective? I contribute to answering these questions by introducing a novel dataset linking barcode level food consumption choices to carbon emissions, describing new stylized facts and analyzing their policy implications. I first document that about a third of the variation of carbon intensity across products and households comes from differences within, rather than between, detailed product categories. Building on this fact and on an estimation of the demand system, I derive three sets of results. First, well-targeted taxes deliver large efficiency gains: a product-level tax achieves the same carbon emissions reduction as a tax across product departments, at a 64% lower utility cost. Second, the impact of carbon taxes across households varies strongly with their consumption's average carbon intensity, but not with their expenditure level. Third, the welfare cost of reducing carbon emissions varies strongly across product categories: an exogenous 5% price increase on two detailed categories would reduce food carbon emissions by about 1% at a utility cost of 1.5€ per household, or 0.1% of the average annual food expenditure in this data. Overall, this paper highlights the need for environmental policies to leverage the substitution patterns within detailed product categories, and thus provides a new justification for granular taxes or, equivalently, carbon markets.

1.1 Introduction

Can countries reduce their carbon emissions[1] by changing the composition of consumption instead of, or in addition to, changing the underlying production technology? If so, what type of policy would be most effective? In this paper, I contribute to answering these questions by introducing a novel dataset linking barcode level food consumption choices to carbon emissions, describing new stylized facts and analyzing their policy implications.

This newly assembled data set results from the combination of scanner data from a large French retailer with publicly available data from the French agency for environmental transition (ADEME) on direct and indirect carbon emissions. Detailed information on price, quantity and carbon emissions are available for $27,000$ products and $160,000$ households over 2017 to 2019, which allows me to document the relationship between consumption and carbon emissions at an unprecedented level of granularity. This data addresses a meaningful share of the overall problem of reducing carbon emissions, as about 26% of all anthropogenic carbon emissions comes from the food supply chain (Poore and Nemecek, 2018).

This data is first used to document new stylized facts focusing primarily on carbon intensity across products and households, defined as the amount of CO2-equivalent (CO2eq) carbon emission per real euro spent. I emphasize carbon intensity for three reasons. First, because from an economic policy perspective, changes in carbon intensity summarizes the degree of decoupling in an economy, that is the degree at which output and carbon emissions grow at different rates and ultimately in different directions. Put otherwise, as long as policy makers have as dual objective to decrease carbon emissions and promote economic growth, carbon intensity is a major parameter of interest. Second, because carbon intensity is a well-defined measure both at the macro and at the micro level, and therefore enables me to draw meaningful macro conclusions from micro data. In particular, carbon intensity scales well. Economy-wide carbon intensity is simply the market-share weighted average of carbon intensity of all the products : $\frac{\bar{E}}{Y} \equiv \bar{e} = \sum_i s_i \cdot e_i$, where $\bar{E}$ denotes the overall carbon emissions, Y denotes the aggregate real output, $\bar{e}$ is by definition the average carbon intensity at the aggregate level, i sums over products of the economy (or sectors, or households), s_i is the market share of sector i and $e_i = \frac{E_i}{Y_i}$ is the carbon intensity of product i. Last, focussing on carbon intensity is also justified from an optimal taxation perspective. I show below that optimal carbon taxes on products, expressed as a percentage of their price, are in general a weighted average of carbon intensity.

I document two empirical patterns on the distribution of carbon intensity across products. In this paper, barcode-level products ($N \approx 27,000$) are nested into products modules ($N \approx 1,000$), which are themselves nested into product sub-groups ($N = 115$), themselves

[1] Throughout this paper, I use interchangeably the terms carbon, carbon dioxide (CO2) and green-house gas (GHG). All are expressed in units of CO2 equivalent (CO2eq) emissions.

nested in product groups ($N = 62$), which are finally nested in product departments ($N = 11$). Taken together, this set of results shows that carbon intensity is very heterogenous across products, and that this heterogeneity comes from differences within narrow product categories, rather than differences between product categories. Specifically, food consumption is skewed towards high-carbon intensity products, even within detailed product categories. In addition, within-group variation in carbon intensity explains a significant share of the overall variation, even within these detailed product categories. More precisely, 57% of the variance of carbon intensity across products comes from within product sub-groups. In the appendix, I complement these finding by documenting the fact that carbon intensity of supplied product is very heterogenous, and this patterns holds even within narrow product categories. Furthermore, targeting a small number of detailed product categories has the potential to achieve significant reduction in carbon emissions. Leveraging data from an environmental label, back-of-the-envelope calculations suggest that a policy focusing on only two detailed product categories, which would incentivize consumers to shift low environmental quality products to high environmental quality products within the same category would reduce carbon emissions from food by close to 6%.

Looking at the distribution of carbon intensity across households suggests that consumption choices *within* product categories, rather than *between* product categories explains a significant share of the difference in carbon intensity. Variation in carbon intensity within detailed product category accounts for between 21% and 46% of the difference between households in the top and bottom decile, depending on whether they are ranked by expenditure or carbon intensity. In the appendix, I complement this finding by relating it to the previous literature. There is a modest but statistically significant negative relationship between carbon intensity and expenditure across households: a 10% increase in monthly expenditure is associated with a decrease of 0.003 points of carbon intensity, or 0.5% of the average carbon intensity. While these estimates of CO2-expenditure elasticity are in line with previous literature, this hides important heterogeneity across products: almost half of the product categories have elasticities significantly greater than 1, which is the upper bound of elasticity estimates in Chancel et al. (2015). Overall, the heterogeneity in carbon intensity documented by these stylized facts suggests that there is room for policies incentivizing households to shift their consumption patterns towards less carbon intensive alternatives. However, these stylized facts alone do not consider the utility cost to the households of making those changes. This is why this paper then turn to a model of food demand to address this questions.

I first build a simple framework in which a social planner chooses a set of optimal commodity taxes subject to a given carbon budget and highlight under which conditions she should care about the preference structure and the distribution of carbon intensity across products. Motivated by the stylized facts, I focus on a specific type of constrained optimal taxes, in which the social planner can set an optimal tax rate at a given product category level only, rather than at the product level. Optimal taxes depend not only on a properly weighted average carbon intensity of a product category, but also on a weighted average of the uninternalized externality from substitute products. An optimal tax will put emphasis on indirect carbon emission targeting over direct emission targeting when the appropriately weighted cross-elasticity is high, or when the product category's own price elasticity is low. The main features of this framework, namely that only final goods are taxed, and that not all of them can be taxed, can be justified by the necessity to think about second best policies. In a first best setting, there would be a uniform tax on carbon emissions and all sources of carbon would be taxed. However, this is not the case in practice. For instance, current supply-side carbon pricing, such as the European Emission Trading Scheme (EU-ETS) is restricted to the energy and the manufacturing industry, so that several sectors, such as transportation and agriculture, are not included. In addition, there is currently no policy mechanism dealing with emissions from imported goods, even though a border carbon-adjustments mechanism is being currently discussed in the European Union. As a matter of fact, this mechanism would likely take the form of a product- or industry-specific tax, not on the source of carbon. Further, one of the rationale of a uniform carbon pricing is to provide a price-signal to incentivize agents to shift their behaviors. While taxing sources of carbon can be effective in incentivizing producers to shift their production processes towards less polluting inputs, market imperfections make it unlikely that downstream consumers would face the same price-signal, and thus would be less incentivized to shift their consumption. Therefore, in a real world setting, taxes on goods are likely to be a useful policy tools to complement taxes on sources of carbon. Last, another important way to interpret this optimal taxation framework is to consider it as a way to specify the impact on household's utility and consumption of uniform carbon pricing on the supply side, if all sources of carbon were taxed, if incidence were fully borne by consumers and if there were no other supply response.

I then add additional structure to the model to discuss the impact of tax changes on private utility and carbon emissions, and use this setting to derive formulas quantifying the relative efficiency of policy targeting. Specifically, I borrow insights from the inflation measurement literature (Broda and Weinstein, 2010; Sato, 1976; Vartia, 1976) to show that under nested constant elasticity of substitution (CES) utility, I can express the welfare impact of a change in tax schedule as a function of substitution elasticities and market share only. This setting can then be used to derive formulas quantifying the relative efficiency of finer policy targeting. These formulas can be expressed as a form of market-share weighted covariance between the within product category elasticity of substitution and a measure of quality of policy targeting within this product category.

Elasticities of substitution are crucial ingredients of the model. They are estimated by exploiting the time and geographic variation of product market shares and prices. Specifically, I implement an instrumental variable approach close to the one of DellaVigna and Gentzkow (2019), in which local changes in log prices is instrumented with changes in national level prices. The very detailed set of fixed effects, as well as controlling for promotions reduces the risk of instrument endogeneity. The median estimated elasticity of substitution at the product module level is 3.04, which is consistent with the previous literature.

Three main results are derived from the empirical implementation.

First, well-targeted taxes are crucial in delivering a significant reduction in carbon emissions at a small utility cost. At a social cost of carbon of 50€ per ton, product-level optimal tax would reduce carbon emissions by 10.2%. To put this numbers into perspective, France's official objective, as set out in its National Low Carbon strategy[2], is to reduce carbon emissions by $\frac{1-0.4}{1-0.185} \approx 26\%$ between 2018 and 2030, in order to meet its 40% reduction target in 2030 relative to 1990. Indeed, emissions were already 18.5% lower in 2018 relative to 1990. In this context, depending on the product category at which the policy is set, optimal taxes would close between 20% and 40% of gap between current carbon emissions and the 2030 objective. In addition, setting taxes at a granular level is much more efficient than the same policy implemented at a coarser level, because the former leverages the high elasticities of substitution between more substitutable products. Specifically, given a specific carbon reduction target, a product-level optimal tax achieves the same objective as a uniform tax a utility loss at a 67% lower utility cost. This also holds for modest changes in the flexibility of the policy instrument: a product-group level carbon tax achieves the same carbon reduction target as a department-level tax at 25% lower utility cost.

Second, heterogeneity across households matters, but only across the carbon intensity distribution, not across the expenditure distribution. When households are ranked according to their average monthly expenditure level, the impact of optimal and constrained optimal taxes on carbon emission reduction and private utility is essentially the same across quartiles of the distribution. However, this impact is quite heterogenous across quartiles of the carbon intensity distribution: when optimal taxes are set at the product level, the bottom quartile reduces carbon emissions by 6.7%, against 13.5% for the top quartile. 60% of this difference is explained by different preferences, as parametrized by the estimated CES, while the rest comes from differences in spending patterns impacting the exposure of households to optimal taxes. Combined with the fact that much of the variation in carbon intensity comes from within bins of the expenditure distribution of households, this result suggests that the distributional impact of carbon taxes is important but takes place mostly across households of similar expenditure level. Third, the welfare cost of reducing carbon emissions is very heterogenous across product categories. To see this, I define an efficiency measure of total carbon emissions reduction following a price increase in a particular product category as

[2] Available at: https://www.ecologie.gouv.fr/strategie-nationale-bas-carbone-snbc; accessed 1st October 2021

the change in utility (or social welfare) expressed in euros per kilogram of carbon emissions avoided. This measure can be interpreted as a general equilibrium efficiency measure, as it accounts for all substitution patterns with other products, within and between product categories (under the maintained assumption of no supply-side response). A 5% price increase from two product categories only, namely raw beef and meat-based raviolis, would reduce carbon emissions by 7.5 kgCO2eq per household, or slightly less that 1% of their overall annual carbon emissions. Further, this would be achieved at a utility cost of less than 1.5€, or 0.1% of their annual expenditure on food in my data. Accounting for the value of carbon emissions avoided as well as for the value of revenue transfers, I find that taxing meat-based ravioli, pre-made chili con carne, pre-made meat-based tomato sauce and tripes would be welfare improving at a social cost of carbon of 50€ per ton.

These findings have a number of implications. First, the fact that carbon emissions can be reduced from shifting consumption patterns within detailed product categories also implies that we can do so without fundamentally changing our dietary mix, at least for a small decrease in carbon emissions. Second, these results also shed a new light on the long-established optimality of carbon markets. Under the assumptions discussed above, carbon markets are equivalent to product-level optimal taxes, and might also be easier to implement across the whole product space in practice. This paper thus also quantifies the gap between first- and (a number of) second-best policies. It suggests that carefully designed environmental policies leveraging the documented heterogeneity in carbon intensity across and within product categories, and accounting for substitution patterns across products could lead to meaningful carbon emission reduction while improving overall welfare.

This paper contributes to three strands of the literature. First, it can be linked to a growing literature studying policies aimed at correcting environmental externalities and mitigating climate change. While the literature on correcting environmental externalities dates back to Pigou (1920) and is too large to be reviewed here, many recent studies on climate change focus on supply-side responses, including supply-side carbon taxation (for a literature review, see for instance Bovenberg and Goulder (2002)). Other studies consider demand-side policies aimed at mitigating environmental externalities, but most of them focus on energy (Allcott, 2011; Allcott and Greenstone, 2012; Allcott et al., 2014; Ganapati et al., 2020), or focus on only one good (Allcott and Taubinsky, 2015; Holland et al., 2016). This paper instead focuses on a setting in which all goods have a negative externality. More generally, it takes the supply side of the economy as given and instead focuses on demand-side substitution between product categories as a source of aggregate reduction in carbon emissions. In particular, this paper takes technology as given and assumes perfect competition. Three reasons drive this choice of neglecting the supply side. First, it is a logical first step when analyzing the role of consumption choices in reducing carbon emissions. Second, price- and tax-driven shifts in consumption are likely to deliver more immediate gains in terms of carbon emissions reduction than supply-side policy instruments. Indeed, the implicit time frame when considering market-size driven innovations (Acemoglu et al., 2012, 2019; Acemoglu, 2002) or preference-driven innovation (Aghion et al., 2021; Besley and Persson, 2020) is set in years, if not decades. Third, even though panel data on sector-wide emissions exist (Stadler et al., 2018), such data is currently not available at product level.

This paper also contributes to a growing literature aimed a measuring and documenting the distribution of carbon emissions and carbon intensity. (Shapiro, 2021; Chancel et al., 2015; Stadler et al., 2018). Most of the literature to date focuses on aggregate statistics and discusses the distribution of carbon emissions at best across broadly defined sectors. Instead, this paper leverages much more granular data and unpacks heterogeneity in carbon emissions at the product module level, at the cost of restricting itself to mass retail food products. To do so, I use Life Cycle Analysis (LCA)-based measures of carbon emissions, rather than environmentally extended input-output (EE-IO) frameworks (Shapiro, 2021; Stadler et al., 2018). One significant advantage of LCA-based measures is that they allow for a much finer characterization of carbon intensity by product. This comes at the cost of two disadvantages: first, unlike EE-IO models, these measures are not designed to be consistent at the macro-economic level, such that there exists a risk of double counting carbon emissions (Chancel et al., 2015). However, appendix figure 1.C.1 shows at the average carbon intensity of food products in Exiobase, one of the leading EE-IO data base is 0.68 kilogram of carbon per real euro, which is exactly the average carbon intensity in my data. While not a definitive proof, this is suggestive evidence that the risk of double counting is mitigated in my data. The second disadvantage is the reduced scope of the analysis. This paper considers only food consumption in France, whereas EE-IO frameworks include most of the consumption basket. However, recent empirical work shows that it is possible to obtain barcode level data on a large fraction of the consumer basket in many countries (Beck and Jaravel, 2021), so that advances on this front could be made in the coming years. Thanks to this level of granularity, I am the first to provide evidence that while income does explain most of the difference in overall carbon emissions, it is not the most important dimension to explain carbon intensity and hence to predict exposure to optimal carbon taxes. This also highlights the importance of using micro data to understand the potential reactions to carbon taxes or more generally to climate change policies.

Finally, this paper relates to the theory of the welfare impact of second-best environmental taxation (Ganapati et al., 2020; Jacobsen et al., 2020; Holland et al., 2016). This paper explores coarse taxation as a specific form of second best-taxation, and consider the impact of many tax changes, instead of a marginal change in one tax rate, at the cost of imposing more structure to the problem. While the model is simple, it makes two contributions. First, it underlines that an optimal Pigouvian tax depends on substitution elasticities with untaxed products. Importantly, it shows how in a second-best world where not all products can be taxed, the optimal tax rates depend heavily on substitution patterns and the distribution of market shares. It suggests that appropriate, second-best climate tax design must account for industry structure and consumer preferences. While this is not a new result per se, it is often left implicit in the literature, mostly because other frameworks assume only one polluting good. Second, the micro-founded definition of the social cost of carbon stresses that this cost crucially depends on the type of policy instruments available to the social planner. It formalizes the idea that the cost to society of carbon emissions actually depends on how difficult it is to adapt. This is a fundamental insight often overlooked in the literature and implies that if individual preferences change, or equivalently if the demand structure of the economy change, so does the social cost of carbon. While I focus on taxes as policy instruments, this framework and its insights could be easily extended to other policy instruments.

The reminder of the paper is organized as follows. Section 2.2 describes the data set and the data construction process. Section 1.3 documents the salient stylized facts from this data set. Section 1.4 sets up a theoretical framework, whose key components, the elasticities of substitution, are estimated in section 1.5. Last, section 1.6 discusses the results and its policy implications.

1.2 Data

This section presents the different data sources used for this paper and describes the data sets resulting from the matching process.

1.2.1 Raw data

Agribalyse. I use Agribalyse[3], an open source and public database created and maintained by ADEME, the French agency for environmental transition and INRAE, the National Research Institute for Agriculture, Food and Environment. Agribalyse documents the environmental impact of about $2,500$ food categories, using 14 quantitative indicators on climate change (in particular CO2 equivalent emissions), fine particles, water use, fossil resource use, land use, mineral and metal use, ozone depletion, acidification, ionizing radiation, photochemical ozone formation, terrestrial eutrophication, marine eutrophication and fresh water eutrophication (ADEME, 2020a). For each food category and indicator, data is available for all stages of the product lifecycle: farming, transportation, packaging, transport, distribution and consumption. These quantitative indicators have been estimated by experts from ADEME and INRAE using a Life Cycle Assessment (LCA) method framed by the ISO 14044 standard ADEME (2020a). The LCA-based nature of the data imply that these indicators are consumption-based and include both direct and indirect pollution. In

[3] AGRIBALYSE data v3.0 - 2020, ADEME

particular, they account for both domestic and imported pollution.

One significant advantage of LCA-based measures of pollution is that they allow for a much finer characterization of carbon intensity by product, down to the barcode in my case. This comes at the cost of two disadvantages: firstly, unlike EE-IO models, these measures are not designed to be consistent at the macro-economic level, such that there exists a risk of double counting (Chancel et al., 2015). However, I document in appendix figure 1.C.1 that this risk seems limited. This figure plots the average carbon intensity of different good categories coming using the Exiobase 3, a state of the art EE-IO data base. In Exiobase, the average carbon intensity for food products in France, accounting for direct and indirect emissions, is 0.68, which exactly the average carbon intensity in my data. Secondly, the scope of the analysis, which is in this paper food products in France, is smaller than with EE-IO models, even though recent empirical work show that it is possible to obtain barcode level data on a large fraction of the consumer basket in many countries (Beck and Jaravel, 2021). Agribalyse data was designed to be calibrated to the French food market. Therefore, it cannot be directly extrapolated to other countries, and within each food category, the data is representative of the average consumption. For instance, the CO2 equivalent emissions of one kilogram of *prepackaged pizza regina* (one of the food categories) is based on the typical pizza dough (mode of the pizza dough distribution) and the average tomato found on the French market (mean carbon emissions of imported and domestic tomatoes), see ADEME (2020b)) These $2,500$ food categories are the most disaggregated elements of the Ciqual classification, originally developed by the French Agency for Food, Environmental and Occupational Health & Safety (ANSES) to structure nutritional information on food consumed in France. Hence, at each different level, this classification groups together products that are dietary similar and it is therefore sensible to use this classification to estimate a nested CES preference structure. In the final data sets, there are 11 product departments, 62 product groups, 114 product subgroups and 961 product modules (see section 1.2.2)

In this paper, I only use the CO2 emissions data from Agribalyse, aggregating over all life stages of the product. However, I hope that this paper will contribute to research exploiting more fully the opportunities offered by Agribalyse. Interestingly, note that farming and transportation account for 72% and 10% of the overall carbon emission from an average consumer basket(ADEME, 2020b).

Private retailer scanner data. I have access to scanner data from a large private retailer in France. The data covers all transactions from about $800,000$ loyalty cards between 2017 and 2019, which is a 10% random sample all existing loyalty cards issued by the retailer. For this retailer, sales from registered loyalty cards account for 70 % of overall sales. Product level information include overall description, brand, an indicator for being organic and its nutriscore, a widely available label ranking food products into 5 categories according to their nutritional quality[4]. Consumer level information is coarse. The only information available is whether the household is a family, a young household (18-35 with no kids), a middle aged household (36-60 with no kids) or a senior household (61+ with no kids). I also have access the full transaction data of this retailer between January and August 2019, that is including non-loyalty card transactions.

This data is similar in nature to the Nielsen Homescan data, which has been used extensively in the literature (for a description of the Nielsen data, see among others Broda and Weinstein (2010); Allcott et al. (2019a); Jaravel (2019)). Like the Nielsen data, price and quantity are separately available for each transaction and food consumption is recorded at the bar code level, which ensures that I can keep track of quality improvement over time as retailers change bar codes only when meaningful characteristics of the product are changed (Broda and Weinstein, 2010). It also suffers the same drawbacks, as it does not record food purchased in restaurants or similar establishments.

Contrary to the Nielsen Homescan data however, households in the raw data are not weighted to be nationally representative. In the final data sets however, they are reweighted so as to fit the expenditure pattern of the Household Budget Survey from the French statistical agency (Insee), see section 1.2.2. In addition, this data only comes from a unique retailer so food purchases from other retailers are not recorded. Since the final data set consists of regular shoppers from this private retailer, it is likely that this bias is less concerning than in the raw data. In addition, my data is likely to contain fewer scanning errors than in the Nielsen household panel, as product scanning is recorded at the point of sale and is not conducted by the household themselves.

The fact that this paper is mostly about carbon intensity also reduces concerns about non-representativeness of the data. I am certainly not capturing all take-home food expenditures from any household, including the loyal customers. Similarly, sales of one product of the raw data is not necessarily representative of the national average, so making statements about absolute the absolute level of carbon emissions can be more hazardous. However, I can still document carbon intensity patterns for these products and households. For the results of this paper to be externaly valid, I only require that loyal consumers at this particular retailer are representative of the overall set of loyal customers of the French population, and that product assortment at this retail chain is representative of product availability nationally, which seems a much more reasonable assumption.

OpenFoodFact data. I use the OpenFoodFact[5] database to merge the data from Agribalyse with the one from the private retailer. OpenFoodFact is an non-for-profit organization and collaborative project gathering data on food products in France and the rest of the world.

[4] For more information, see https://www.beuc.eu/publications/beuc-x-2019-051_nutri-score_factsheet.pdf; accessed 1st October 2021

[5] https://fr.openfoodfacts.org/

Data is supplied both by producers and by a network of over $25,000$ active contributors who send photos of the product labels and barcodes. Artificial intelligence and optical character recognition are then used to extract and clean the data by standardizing data fields and removing inconsistencies. Overall, information is available for over 1.9 million products worldwide, including approximately $800,000$ products in France. The data contains detailed information on nutritional ingredients, food labels including Ecoscore, an environmental label developed by ADEME, product weight, and Ciqual classification for every barcode, and hand checks suggest that the data is very accurate. The main disadvantage in using this data is that it is a selected sample of all available products based on popularity and on contributor and producer interest. For instance, the average OpenFoodFact contributor is likely to be more conscious about the nutritional quality of her food consumption than the average French consumer, so the Open Food Fact database is likely to be skewed towards products that have either high or low nutritional quality. However, OpenFoodFact management team reports that in 2020 in France, $91,2\%$ of all user-scanned products are already in their database and according to their internal calculations, OpenFoodFact covers 95% of the top 10 000 products sold in France. Furthermore, the match performance with the universe of products from the private retailed data is good, suggesting that most sales patterns are captured, see section 1.2.2.

1.2.2 Data construction

I construct three data sets for analysis. The first data set is a product-level data set using the universe of transaction of this retailer between January and August 2019 (henceforth 'Product data'). The second data set is a household-level data set, consisting of regular shoppers only (henceforth 'Household data'). It is arguable that for these households, who are loyal to this retailer, I capture most of the take-home food expenditures. Descriptive analysis is based on these first two data sets. The third data set is comes from the Household data data but is aggregated at the product by quarter by geography level and imposes more stringent availability restrictions. The empirical estimation of elasticities of substitution in section 1.5 is conducted on this data set (henceforth 'Analysis data').Only observations from metropolitan France are used, and all prices are deflated to January 2017 euros using the Insee mass retail price index.

I first clean the OpenFoodFact data by deleting and consolidating duplicate observations (products having been scanned more than once with different inputted data in the OpenFoodFact database. Less than 0.03%% of the $999,245$ products have conflicting Ciqual codes or quantity numbers. When there are conflicting observations for a given barcode, I keep the most frequent Ciqual code and quantity number, with ties being randomly broken. The resulting data set is then matched to the Agribalyse data and to the products characteristics dataset from the retailer. Only food products are kept, and animal food is discarded. I then compute carbon emissions per barcode, and trim the top and bottom 0.1% of the carbon emission distribution to discard outliers. Appendix table 1.B.1 shows the number of product groups, sub-groups, modules and products by product department from the Product data set.

This data is then matched to the private retailer data, for which I initially dropped all products that are sold less than 10 times a month on average. As can be seen from appendix table 1.B.2, matched products represents over 60% of sales in all three data sets. Overall the Product data set captures annual emissions equivalent to 5.5 Mt of CO2, which represents 1.3% of French green-house gas emissions in 2018[6]. This is roughly coherent with the share of food and agriculture in total GHG emissions, the market share of the retailer and the share of retail take-home expenditures in the total food expenditure.

For the Product data set, expenditure and quantity purchased are weighted so as to match the national spending patterns retrieved from the Household Budget Survey constructed by Insee. Specifically, it matches the nationally representative budget shares for food products, aggregated at the third level of the Classification of Individual Consumption by Purpose (Coicop), which represents 11 categories. Appendix table 1.B.3 displays summary statistics on carbon intensity and carbon emissions per products for all product modules in the 'Fats' department.

To create the Household data set, I select households for which I observe at least 50 euros of expenditures on matched products for at least 50% of the months available in my sample, and drop the top 0.1% of households with the largest average monthly expenditure. This ensures that I only keep the most regular and loyal shoppers. Overall, 166,662 households are selected, spending on average 125 € every month, and who are present on average in 93 % of the months. Appendix table 1.B.4 reports statistics on the average monthly expenditure of these households, as well as the number of months they are observed. Non-family households spend marginally less that the average. Further, households with high carbon intensity tend to spend slightly less than households with lower carbon intensity, even though the difference is small. For the Household data set, expenditure and quantity purchased are weighted so as to match the national budget shares for food products by type of household (family or not family).

To construct the Analysis data set, I start from the Household data and aggregate observations at the product by quarter by geographic department level. Further, I impose additional restrictions on product availability following DellaVigna and Gentzkow (2019). I request products to be sold in at least 50% of the quarter by geographic department observations. Further, all product modules with less than four products are dropped. Overall, the Analysis data set contains 8,652 products from 435 product modules.

[6] Total French GHG emissions in 2018 are estimated to be 419 MtCO2eq; see Key Climate Figures - France, Europe and World, 2021 edition available on https://www.statistiques.developpement-durable.gouv.fr; accessed 1st October 2021.

1.3 Stylized facts

This section presents stylized facts on the distribution of carbon emissions and carbon intensity across products and across consumers.

1.3.1 Product-level stylized facts

Table 1.1 correlates carbon intensity with product characteristics across all three data sets. The first point to note is that broad product departments are not clearly predictive of carbon intensity: out of the 11 product departments, only 4 have significantly different carbon intensity than the baseline (Cereal products). As could be expected, a higher Ecoscore grade is correlated with lower carbon intensity, which is sensible as it is based on same data. Products with organic label also tend to have lower carbon intensity. Interestingly, products with higher nutritional score tend to have higher carbon intensity, suggesting more broadly that there might be a conflict between healthy and eco-friendly eating behaviors. Last, results are very similar across datasets, suggesting that the selection procedure highlighted above the does not harm the representativeness of the data on the product dimension.

Stylized fact 1: Food consumption is skewed towards high-carbon intensity products. Indeed, high-carbon intensity products tend to have higher market share than low carbon intensity products. This holds both at the aggregate level, as well as within narrow product category. Overall, this suggests that there is room for consumption-shifting to reduce carbon emissions even within narrow product categories.

Figure 1.1 shows that products with 10% higher expenditure have on average a 0.004 higher carbon intensity, or 1% of the average carbon intensity. This relationship is not driven by cheaper products being both more consumed and less carbon intensive, as panel (a) demonstrates. Panels (b) and (c) show that the positive relationship also holds when considering carbon emissions per product (the numerator of carbon intensity) and when considering quantities instead of expenditure, respectively. Last, panel (d) shows that the positive relationship between carbon intensity and expenditure is not driven by capacity effects. For example, the difference in carbon emissions between larger and smaller capacity products, say $1.5L$ and $0.5L$ mineral water of the same brand are likely to be close to proportional, whereas their price difference are arguably less than proportional. If, in addition, larger capacity products are more popular than their small capacity counterpart, it could create a spurious correlation between carbon intensity and expenditure. In non-reported results, I show that these results also hold for the Household dataset.

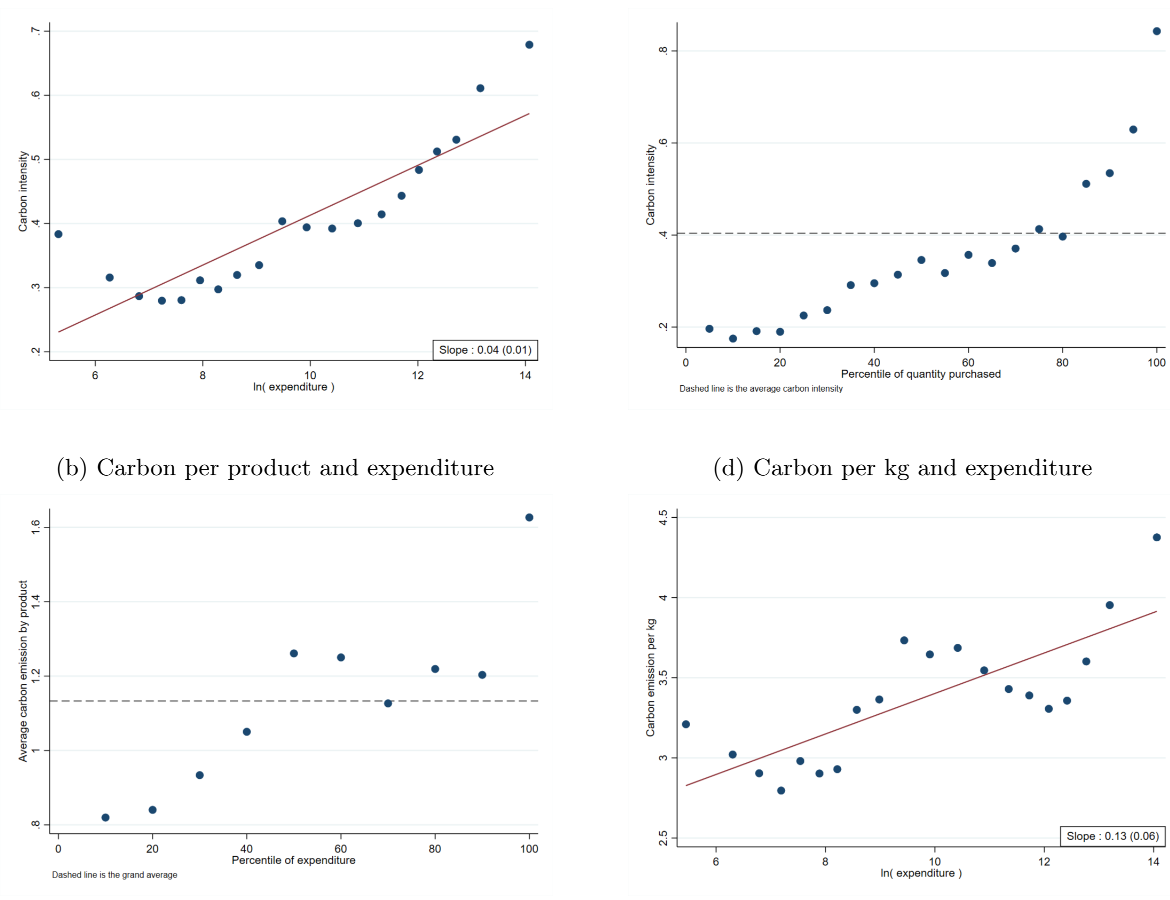
As can be seen from appendix table 1.B.5, this patterns also holds within narrow product category. The relationship between carbon intensity and market share is strong and

Table 1.1: Predictors of carbon intensity - Product level

	Product data		Household data	
Product departments:				
Prepared dishes	0.145	(0.139)	0.154	(0.153)
Vegetables and fruits	-0.229***	(0.059)	-0.254***	(0.065)
Cereals	0.000	(.)	0.000	(.)
Meat, fish and eggs	-0.084	(0.084)	-0.105	(0.091)
Dairy	0.219***	(0.074)	0.260***	(0.076)
Beverages	-0.183**	(0.075)	-0.182**	(0.071)
Sweets	-0.041	(0.073)	0.056	(0.083)
Iced creams	-0.087	(0.101)	-0.025	(0.074)
Fats	0.391	(0.243)	0.547*	(0.295)
Other products	-0.071	(0.079)	0.000	(0.101)
Infant food	0.051	(0.070)	0.009	(0.071)
1(Organic)	-0.094***	(0.016)	-0.086***	(0.021)
Nutriscore:				
A	0.224***	(0.047)	0.248***	(0.052)
B	0.170***	(0.056)	0.168***	(0.064)
C	0.000	(.)	0.000	(.)
D	-0.033	(0.034)	-0.063*	(0.037)
E	0.100*	(0.057)	0.105*	(0.062)
N/A	-0.055	(0.033)	-0.049	(0.036)
Ecoscore:				
A	-0.205***	(0.035)	-0.242***	(0.049)
B	-0.123***	(0.024)	-0.126***	(0.031)
C	0.000	(.)	0.000	(.)
D	0.108***	(0.031)	0.137***	(0.040)
E	0.419***	(0.092)	0.457***	(0.112)
N/A	0.279**	(0.112)	0.265***	(0.093)
Constant	0.440***	(0.063)	0.464***	(0.068)
R^2	0.203		0.210	
N	27,553		19,068	

Notes: This table presents predictors of carbon intensity for the Product and Household data set. Predictors are indicators for product departments, for being organic, for Nutriscore category, a dietary label, and for Ecoscore, a environmental label. For both label, A indicates the highest nutritional and environmental quality, and N/A indicates that the label is unknown. This represents 23% and 3% of products for Nutriscore and Ecoscore in the Product dataset respectively. The regressions are not weighted by quantity in order to be representative of the choice set rather than the actual consummer choices. Standard errors in parentheses are clustered at the product module level. Base level for product departments is Cereals, and is score C for Ecoscore and Nutriscore.

Figure 1.1: Carbon emissions and expenditure



(a) Controlling for price

(c) Carbon intensity and quantity

(b) Carbon per product and expenditure

(d) Carbon per kg and expenditure

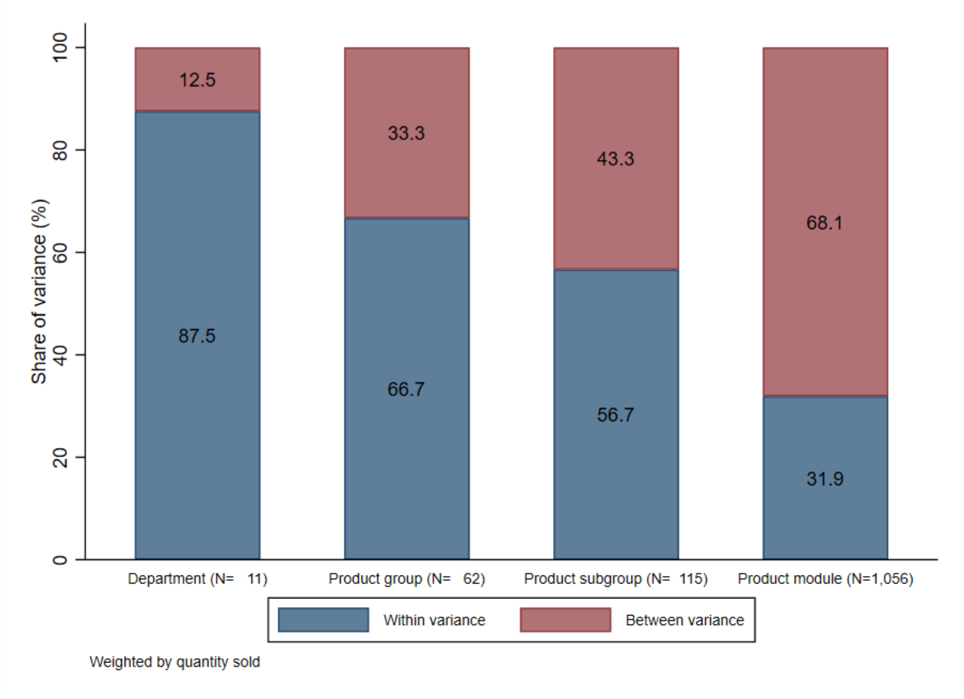
Notes: This figure presents four binned scatter plots documenting the relationship between carbon emissions and expenditure from the Product data set. Panel (a) relates carbon intensity and log expenditure, controlling for product price. Panel (b) relates the carbon emissions per product (in eqCO2kg) and log expenditure. Panel (c) relates carbon intensity and percentile of quantity product purchased. Panel (d) relation carbon emission per kg of products and log expenditure. In panel (b), dashed line indicates the unweighted average of the variable on the y-axis. Where a slope is indicated, standard errors are clustered at the product module level.

statistically significant within department, product groups, product sub-groups and product modules. The size of the coefficient is only 20% lower at the the product module level (1000 categories) than at the product department level (11 categories). This suggests that positive relationship between carbon intensity and product expenditure established at the aggregate level is not, or not simply, driven by comparing types of products with low degree of substitutability, and that the relationship is just as strong within more substitutable product categories: within product module, a 10% increase in market share is associated with a 0.0025 increase in carbon intensity, that is a 1.2 % of the average standard deviation within module. This results also does not follow from the fact that more carbon intensive products might also be cheaper products: in appendix table 1.B.6 while the coefficient on market share drop by 35% from 0.047 to 0.031, the positive relation between carbon emission per product and product market share remains very strong. Within product module, a 10% increase in market share is associated with a 0.0031 increase in carbon intensity, that is a 1.6 % of the average standard deviation within module. Appendix table 1.B.7 shows as a robustness check that controlling for price has little quantitative implications. Various reasons could explain why carbon-intensive products are more popular. They could be of actual or perceived higher quality, taste better, the packaging could be nicer, average position in the stores could be different, therefore making it more visible to consumers, or consumer could be entrenched in habits consuming higher carbon intensity products.

Stylized fact 2: Even within detailed product category, within-group variation in carbon intensity explains a significant share of the overall variation. This suggests that focussing on detailed product categories to incentivize consumption shifting can be impactful.

To show this, I decompose the variance of carbon intensity across products as the sum of a between-product category and a within-product category component. For any variable y_j indexed over products $j = 1, ..., J$, and for any partition of products into G product category where each category $g = 1, ..., G$ contains J_g products, we have that $V(y_j) = \sum_{g=1}^{G} \frac{J_g}{J}(\bar{y}_g - \bar{y})^2 + \sum_{g=1}^{G} \frac{J_g}{J}\frac{1}{J_g}\sum_{j=1}^{J_g}(y_j - \bar{y}_g)^2$, so that variance of y is the sum of the (weighted) variance across groups of the group averages $\bar{y}_g$ and the weighted average of within-product group variance. A similar formula can be derived when products are weighted. Figure 1.2 plots this decomposition for carbon intensity across the different product categories. It is striking to see that, even within detailed product categories, within group variation explains a significant share of overall variation. Specifically, 57% of the overall variance in carbon intensity comes from within product sub-group variance, and 32% comes from within-product modules variance. Panel (a) of appendix figure 1.C.2 highlights that the share of the within-category component is even higher when products are not weighted by quantity purchased, although within the same order of magnitude. Panel (b) shows that this effect is not simply driven by price variation within and across product

Figure 1.2: Between-within decomposition of carbon intensity by product category



Notes: This figure presents decomposition of the variance of carbon intensity into between group variation and within group variation, for different product category. Data comes from the Product data.

categories as the same pattern holds for the variance of carbon emitted per product. Last, panel (c) shows that almost all of the variation in overall carbon emissions is driven by within product module variation.

1.3.2 Consumer-level stylized fact

Stylized fact 3: Difference in spending patterns within detailed product categories explains an important share in overall difference in carbon intensity across households. Differences in carbon intensity across households can be explained by three different factors. First, households can spend their food expenditure on different product categories with different average carbon intensity. For instance, some households can be vegetarian and spend a disproportionate amount of their expenditure on vegetables and cereals, or tend to consume a lot of dairy products and spend more than average on milk and cheese. Second, within product categories, say cheese, households can have varying tastes for slightly different products, for instance between hard and soft cheese, both with different carbon intensities. Third, households can have exactly the same expenditure shares on all products, but some households are facing higher average prices, reducing their average carbon intensity.

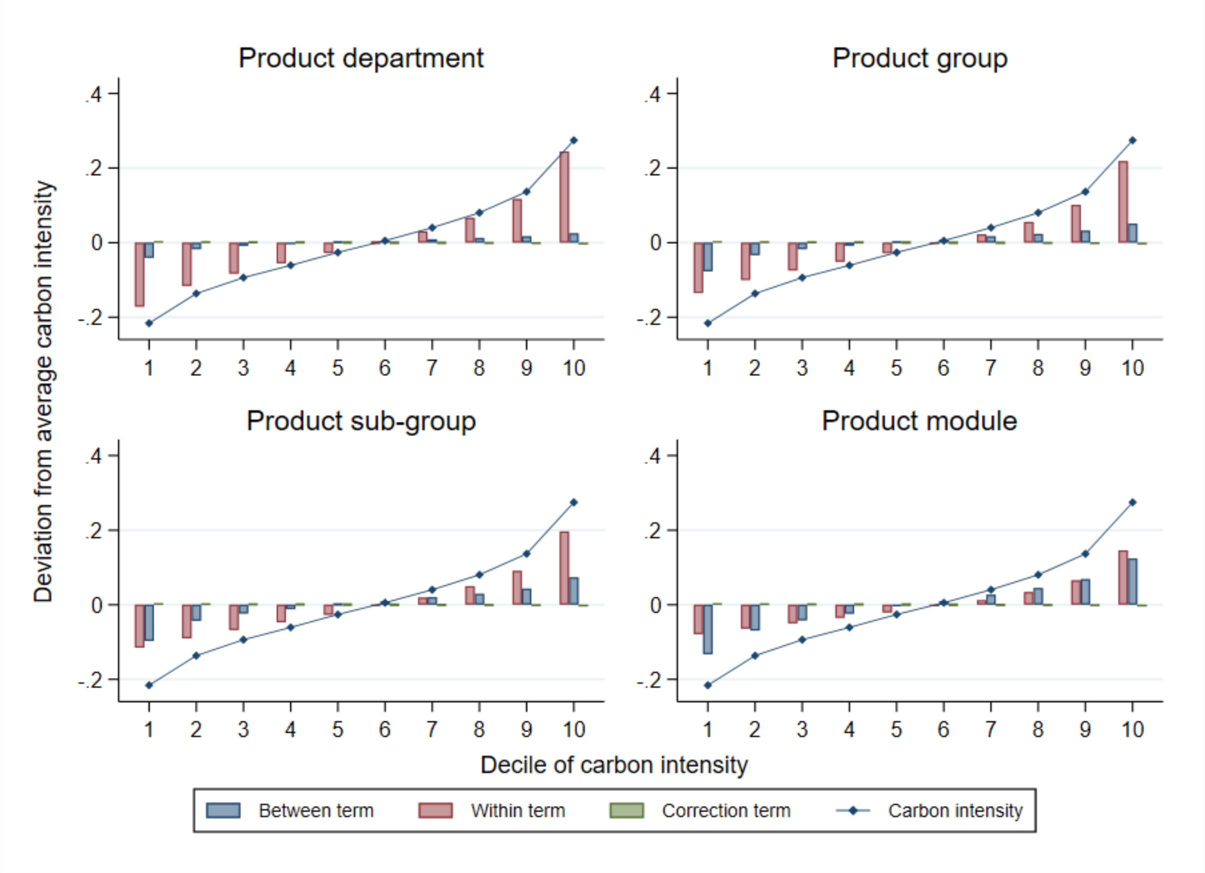
More precisely, denote e_r the carbon intensity of consumer group $r$and $\bar{e}$ the economy-wide carbon intensity. Then, the difference in carbon intensity can be written as

$$e_r - \bar{e} = \underbrace{cov^g\left(\omega_r^g - \bar{\omega}^g, \bar{e}^g\right)}_{\text{between product categories}} + \underbrace{\sum_{g=1}^{G} \omega_r^g \cdot cov^{i\in I_g}\left(\omega_{ir}^g - \bar{\omega}_i^g, \bar{e}_i - \bar{e}^g\right)}_{\text{within product categories}} - \underbrace{\sum_{i=1}^{I} \omega_{ir} \cdot \delta_{ir}}_{\text{price correction}} \tag{1.1}$$

where g indexes product categories, i indexes products, ω_r^g, $\bar{\omega}^g$ are the expenditure shares on product group g of, respectively, consumer group $r$and the average consumer, ω_{ir}^g and $\bar{\omega}_i^g$ are the spending shares of product i within product group g of, respectively, consumer group $r$and the average consumer. The first term in (1.1), the "between product categories" term will be high if consumer group r spends a disproportionally high share of its expenditure on carbon intensive product categories, evaluating the carbon intensity of product category g, $\bar{e}^g$, at national prices. The second term, the "within product categories" term will be high if, within product group g, consumer group r is spending relatively more on more carbon intensive products than what the average consumer does. Here, carbon intensity is evaluated relative to the average carbon intensity of the product group, at national prices. The last terms, the "price correction" term, corrects for differential price faced by consumer group $r$relative to national prices. Here, $-\delta_{ir} \equiv e_{ir} - \bar{e}_i$, the difference in carbon intensity of a given product bought by consumer group r and of the same product bought by the average consumer. δ_{ir} will be positive if the price face by consumer group r for product i, p_{ir}, is higher than $\bar{p}_i$, the national price for this product. Overall, this term will be high if ω_{ir} and δ_{ir} are positively correlated, that is if for instance consumer group r buys products that are priced more highly in their area. The precise derivation of (1.1) can be found in appendix 1.D.

Figure 1.3 shows this decomposition when households are grouped into decile of carbon intensity. Four points are worth highlighting. First, the difference in carbon intensity between the first and last decile is high, at about 0.45 relative to an average carbon intensity of 0.65. Second, the correction term is negligible, for all decile and all product categories, suggesting that price variations is not driving differences in carbon intensity. Third, at the product department, group and sub-group level, the within term drives the difference in carbon intensity between deciles. Fourth, even at the product module category level, the within term explains a sizable share of the overall difference. More precisely, the within term accounts for 46 % of the overall difference in carbon intensity between the first and last decile at the product module level (85 % at the product department level). Appendix figure 1.C.3 confirms that the pattern is similar when grouping households by expenditure

Figure 1.3: Decomposition of difference in carbon intensity



Notes: This figure presents decomposition of the difference in carbon intensity relative the grand average following (1.1). Carbon intensity is defined at the household level and data comes from the Household data.

decile. However, the difference in average carbon intensity between the first and last decile is now about 0.06, an order of magnitude lower than in figure 1.3. While now the between term is driving the difference in carbon intensity, the within term still accounts for 21 % of the difference in carbon intensity between the first and last decile at the product module level, and 28 % at the product department level.

One way to get a sense of the economic importance of the within term is to ask by how much would carbon emissions decrease if households were to adopt the spending patterns of the best performing decile. Suppose that all households allocate their spending between product departments in the same way (same amount spent on fats, on cereals, on vegetables, etc.), but that within each product department, all households allocate their spending in the same way households in the bottom decile of the carbon intensity distribution would. More precisely, I compute a simulated carbon intensity for decile r, $e_r^{sim} = \sum_{g=1}^{G} \omega_r^g \cdot \sum_{i=1}^{I_g} \omega_{i1}^g \cdot \bar{e}_i$

where ω_{i1}^{g} is the spending share of decile 1 on product i relative to its spending in product category g. Note that actual carbon intensity is $e_r^{actual} = \sum_{g=1}^{G} \omega_r^g \cdot \sum_{i=1}^{I_g} \omega_{ir}^g \cdot e_{ir}$. I then multiply e_r^{sim} by the household total expenditure to arrive to a counterfactual carbon emissions. Doing so would reduce carbon emissions by 25 %. One can argue however that changing spending patterns within product department, a broad product category, involves major changes. For instance, it could involve switching from meat to eggs or fish, within the meat, fish and eggs product department. Repeating the same exercise at the product module level, so that households change their spending patterns only within about a thousand product categories, and not across them, would still reduce carbon emissions by 11 %. This would still be a significant reduction in carbon emissions considering that it implies much smaller consumption changes. Repeating the same exercise when ordering consumers by expenditure decile would imply a decrease in carbon emissions of only 0.65 % when at the product department level, and of 0.43 % when at the product module level.

Appendix 1.A provides additional stylized facts on products and households. In particular, it documents that (i) carbon intensity and carbon emissions of food choices are very heterogenous, even within narrow product categories; (ii) that targeting a small number of detailed product categories has the potential to achieve significant reduction in carbon emissions; (iii) that there is a very modest negative relationship between carbon intensity and expenditure across households; and (iv) that while CO2-expenditure elasticity estimate is in line with previous literature, this hides important heterogeneity across products.

1.4 Theoretical model

Taken together, the stylized facts discussed in the above section as well as the various back of the enveloppe calculations show that there is significant heterogeneity in the carbon intensity of products even within narrow product categories so that policies targeting carbon emissions at a detailed product level could have a meaningful impact. However, these calculations do not consider the cost to the household of making those changes: while chocolate based products are very carbon-intensive, it might not be optimal to focus on this product category if consumers have no acceptable substitutes to it, so that the utility cost of switching is high, or if substitute products, say coffee based products, are even more carbon-intensive.

In this section, I develop a theoretical framework to address this issue and to think about the private utility cost and the social welfare impact of a change in consumption pattern. First, an optimal taxation framework is built in order to highlight under what conditions a social planner willing to set up a tax on a particular good should care about the carbon content of its substitutes. Additional structure is then added to the model to discuss the impact of tax changes on private utility and carbon emissions. This setting can also be used to derive formulas quantifying the relative efficiency of policy targeting. Last, I discuss how we can calibrate these formulas to compute overall welfare gains of tax reforms.

1.4.1 Model set-up

A representative household maximizes private utility $U(x_0, ..., x_N)$ over $N+1$ goods, subject to the budget constraint $\sum_i q_i x_i = Z + T$, where Z is non wage income and T is a transfer. q_i are tax inclusive prices, that is $q_i = (1+t_i)p_i$. I choose to focus on ad-valorem taxes so that they can be then meaningfully compared across different product category levels. This differs from usual optimal taxation frameworks, in which taxes are often expressed per units. However, the choice between the per-unit or ad-valorem taxes is purely cosmetic as it depends only on whether quantities are normalized or not. The social planner rebates all revenues to consumers as a lump tax subsidy: $T = \sum_i t_i p_i x_i$. To rule-out lump-sum taxes, I assume that good 0 is not taxed: $t_0 = 0$. In addition, good 0 is the numeraire so that $q_0 = p_0 = 1$. The first order condition for the household's optimization problem is then $U_i = \alpha q_i$, where α is the Lagrangian on the household's budget constraint, as well as the marginal value of income $M = Z + T$ to the household. Optimization yields Marshallian demand $x_i(\mathbf{q}; M)$ and indirect utility function $V(\mathbf{q}; M)$, where $\mathbf{q}$ denotes the vector of prices.

The social planner's objective is to set a schedule of prices so as to maximize household's utility, taking as given the household's behaviour and its budget constraint. In addition, the social planner is subject to a total carbon budget: $\sum_i E_i \cdot x_i \leq \bar{E}$, where E_i is the amount of carbon emitted by good i. This way to set-up the externality problem as a fixed constraint seems more relevant from a policy and natural science perspective than the "welfare damage approach" taken by most recent papers on optimal taxation in this context which express the externality as a cost to social welfare (see for instance Allcott et al. (2019b); Jacobsen et al. (2020); O'connell and Smith (2020)). Again, this is purely cosmetic: as shown below, both approaches are equivalent. I abstract away from supply-side considerations by assuming perfect competition, marginal cost pricing and no profits, so that the supply side does not enter the social planner's problem. The social planner's Lagrangian is the $\mathcal{L} = V(\mathbf{q}; Z+T) + \lambda \left[\bar{E} - \sum_j E_j \cdot x_j\right]$ and the first order condition with respect to t_j is:

$$\frac{\partial V}{\partial t_j} + \frac{\partial V}{\partial M}\left[p_j x_j + \sum_k t_k p_k \frac{\partial x_k}{\partial t_j}\right] - \lambda\left[\sum_k E_k \cdot \frac{\partial x_k}{\partial t_j}\right] = 0 \tag{1.2}$$

This first order condition has a very conventional interpretation: the first two terms represent the impact of the tax change on the household's private utility: it impacts both the relative prices faced by the household, but also its income following the change in transfers. The third term represents the value to the government of the change in carbon emissions, as it is the product of the marginal value to the government of an additional unit of carbon emitted times the total change in carbon emissions following the tax change. At the optimal tax level, these three terms sum up to zero. The first order condition can then be reformulated (see appendix 1.D) as:

$$\forall j \geq 1, \sum_{k=1}^{J} \left[t_k - \frac{\lambda}{\alpha} \frac{E_k}{p_k} \right] \cdot s_k \cdot \epsilon_{k,j} = 0 \tag{1.3}$$

where $s_k = p_k x_k / \sum_i p_i x_i$ is the market share of product k and $\epsilon_{k,j}$ is the elasticity of product k with respect to the price of product j.

$t_k^* = \frac{\lambda}{\alpha} \frac{E_k}{p_k}$ for all products k is a solution to these first order conditions and is exactly the optimal Pigouvian tax. Note that $E_k / p_k = e_k$, so that the optimal tax level is proportional to the carbon intensity of product k, as announced in the previous sections. More precisely, the multiplicative factor λ/α is the social cost of carbon: it expresses the value of an additional unit of carbon to the society expressed in monetary units and not in utility units. This approach is equivalent to the "welfare damage approach" as $\frac{\lambda}{\alpha}$ is the cost of an additional unit of carbon to the social planner. Henceforth, I define $\phi_k \equiv \frac{\lambda}{\alpha} \frac{E_k}{p_k}$ as the externality cost of good k as a percentage of its private (marginal) cost.

1.4.2 Characterizing optimal taxes

In order to provide intuitive optimal tax formulas for the case when Pigouvian taxation is not feasible, I impose an additional assumption.

Assumption 1. $\forall j$, $cov_{k \neq j}(\phi_k - t_k, \epsilon_{k,j}) = 0$. *That is, there is no (market share-weighted) correlation between the (uncorrected) externality of a good k and its cross-price elasticity relative to good j.*

Note that if U has a CES or nested CES form, this assumption is satisfied as the cross price elasticities depend on product j but not on product k. This assumption can be relaxed, as is done in the mathematical appendix, even though it brings little additional insight.

Product-level optimal taxes. Under assumption 1, we can express the optimal tax as:

$$t_j = \underbrace{\phi_j}_{\text{direct emission targeting}} - \underbrace{\frac{\epsilon_{-j,j}}{-\epsilon_{j,j}} \cdot \frac{1 - s_j}{s_j} \cdot \phi_{-j}}_{\text{indirect emission targeting}} \tag{1.4}$$

In this formula, $\epsilon_{-j,j} = \frac{1}{1-s_j} \sum_{k \neq j} s_k \cdot \epsilon_{k,j}$ is the market share-weighted average cross-elasticity for good j. Intuitively, when p_j goes up by 1%, quantity demanded for good $j' \neq j$ goes up by $\epsilon_{-j,j}$ % on average. Further, $\phi_{-j} = \sum_{k \neq j} \frac{s_k}{1-s_j} \cdot (\phi_k - t_k)$ is the market share-weighted average uncorrected externality and $\epsilon_{j,j} < 0$. Since I assume that at least one good is not taxed, this term is never zero.

The first term in equation (1.4) is the classical Pigouvian term: polluting goods should be taxed to the extent of their marginal externalities. Note that the first-best setting of a uniform price of carbon, in which all prices adjust accordingly, would be the same as the optimal product-specific tax in this framework. The novelty of this formula comes from the indirect emission targeting term, which makes it clear that an optimal tax system should also consider the carbon emissions arising from substitution patterns away from good j. Specifically, the balance between the direct and indirect emission targeting in an optimal commodity tax system is driven by four terms.

First, it depends on the amount of externality under- (or over-) correction of the other goods in the economy (as long as $\epsilon_{k,j} \neq 0$ and $s_k > 0$). Indeed, if the average uncorrected carbon externality of other goods, ϕ_{-j}, is sufficiently large, then subsidizing good j becomes optimal, even though it might be (very) polluting. On the contrary, if other goods are overtaxed, in the sense that $\phi_k < t_k$, then t_j should be higher than in the classical Pigouvian situation. Second, the relative importance of indirect emission targeting depends on the average cross elasticity of good j, $\epsilon_{-j,j}$. If it is very easy to substitute away from good j, then targeting indirect emissions will be key and in the limit, it is optimal to subsidize good j and to favour indirect over direct carbon reduction efforts. On the other hand, if it is hard to substitute away from good j, then it is more efficient to ignore substitution patterns and to focus on direct emission targeting. If good j is a complement to the other (uncorrected) goods in the economy, $\epsilon_{-j,j} < 0$, then the direct and indirect emission targeting terms go in the same direction: taxing good j reduces both carbon emission from good j and from all the complementary products. In practice, this could mean for instance that policy makers should take into account the amount of untaxed pollution from car manufacturing when setting gasoline taxes. Third, the indirect term depends on the own-price elasticity of good j. If $\mid -\epsilon_{j,j} \mid$ is very large, then demand for good j is very sensitive to price and it makes sense to favour direct carbon reduction over indirect carbon emissions. This result is the exact opposite of the inverse elasticity rule in a Ramsay framework of commodity taxation: the objective is different but the mechanism is the same. In a typical Ramsay framework, higher price elasticity increases the deadweight loss of taxation so the optimal tax on this good should be lower for efficiency reasons. Here, higher price-elasticity means that the demand response of this good will be high, so the optimal tax on this good should be higher for efficiency reasons because it will lead to a large decrease in carbon emissions. Last, the extend of indirect emission targeting depends on the relative importance of good j. The higher the market share of good j, the lower the term $\frac{1-s_j}{s_j}$, so the optimal tax will favor direct carbon reduction because good j will be responsible for most of the social cost of carbon. It is important to note that like most optimal tax formula, equation (1.4) is endogenous, as for instance the optimal tax depends heavily on the market share of good j, which itself depends on the overall tax schedule.

Carbon Border tax. As an illustration of equation (1.4), suppose that goods $2, ...J_0$ are imported goods, therefore cannot be taxed. Suppose further that all other goods $J_0 + 1, ..., I$ are already taxed at their optimal Pigouvian level ϕ_j. Then, the optimal tax on good $j = 1$ is:

$$t_j = \phi_j - \frac{\bar{\epsilon}_{imp,j}}{-\epsilon_{j,j}} \cdot \frac{s_{imp}}{s_j} \sum_{k=2}^{J_0} s_k^{imp} \cdot \phi_k$$

where $s_{imp} = \sum_{k=2}^{J_0} s_k$ is the import share, s_k^{imp} is the market-share of good k among imported goods and $\bar{\epsilon}_{imp,j} = \frac{1}{1-s_M} \sum_{k=2}^{J_0} s_k^{imp} \cdot \epsilon_{k,j}$ is the average cross-elasticity between good j and imported goods. Note that a variant of assumption (1) is needed, namely that $cov_{k \in M}(\phi_k, \epsilon_{k,j}) = 0$, that is, across all imported goods, there should be no (expenditure-weighted) correlation between the externality of imported good k and its cross-price elasticity relative to good j. Intuitively, the tax on a home good j should consider carbon-leakage to other imported, (partially) untaxable goods to the extend that import share is high, and that good j is tradable (high elasticity of substitution with imported goods).

Product category level optimal tax. Suppose that goods are partitioned into $g \leq G$ categories, with I_g products per category. Suppose further that the social planner can only set one tax rate per product category, denoted t_g. For instance, product level information on carbon emissions could be unavailable or too costly to collect. In this case, equation (1.4) carries over to product categories, with appropriately defined elasticities and externalities. All derivations can be found in appendix 1.D.

$$t_g = \tilde{\phi}_{g,g} - \frac{\bar{\epsilon}_{-g,g}}{-\bar{\epsilon}_{g,g}} \cdot \frac{1 - s_g}{s_g} \cdot \tilde{\phi}_{-g} \tag{1.5}$$

To understand equation (1.5), let us define a number of objects. First, $\bar{\epsilon}_{k',g} = \sum_{l \in g} \epsilon_{k',l}$ is the sum of all cross elasticity between good k' and goods in product category g. Intuitively, $\bar{\epsilon}_{k',g}$ is the percentage change in demand of product k' following a uniform, marginal increase in price of all products in group g. Note that when $k' \in g$,$\bar{\epsilon}_{k',g}$ incorporates both the effect of the own-price elasticity and the within-product category cross-price elasticity. Second, $\bar{\epsilon}_{g',g} = \sum_{k' \in g'} \bar{\epsilon}_{k',g} \cdot s_{k'}^{g'}$ is the expenditure-weighted average cross-elasticity between product category g and g', where $s_{k'}^{g'}$ is the expenditure share of product k' in category g', $s_{k'} = s_{g'} \cdot s_{k'}^{g'}$ and $s_{g'} = \sum_{k' \in g'} s_{k'}$ is the expenditure share of product category g'. Intuitively, $\bar{\epsilon}_{g',g}$ is the average percentage change in demand from a product in category g' following a uniform, marginal increase in price of all products of category g. Again, note that $\bar{\epsilon}_{g,g} = \sum_{k \in g} s_k^g \cdot \bar{\epsilon}_{k,g}$ is category g's expenditure-share weighted average price elasticity, accounting for all own-price elasticities and within-product category cross-elasticities. Third, let us also define the contribution of good $k' \in g'$ to the average cross-elasticity of category g' with category g as $w_{k'}^{g',g}$

$$w_{k'}^{g',g} = \frac{\bar{\epsilon}_{k',g} \cdot s_{k'}^{g'}}{\sum_{m \in g'} \bar{\epsilon}_{m,g} \cdot s_m^{g'}} = \frac{\bar{\epsilon}_{k',g} \cdot s_{k'}^{g'}}{\bar{\epsilon}_{g',g}}$$

Fourth, the elasticity-weighted average carbon externality of category g' can be defined as $\tilde{\phi}_{g',g}$, where

$$\tilde{\phi}_{g',g} = \sum_{k' \in g'} \phi_{k'} \cdot w_{k'}^{g',g}$$

Intuitively, $\tilde{\phi}_{g',g}$ is the percentage change in the social value of carbon emissions from product category g' following a marginal, uniform increase in prices of good in product category g. Following Bernheim and Taubinsky (2018); Allcott et al. (2019b), it can also be interpreted as the average marginal externality of product category g' with respect to product category g. Hence, $t_g = \tilde{\phi}_{g,g}$ is the constrained optimal Pigouvian tax when only g tax rates are available. Fifth, let $\tilde{\phi}_{-g} = \sum_{g' \neq g} \frac{s_{g'}}{1-s_g} \cdot \left(\tilde{\phi}_{g',g} - t_{g'}\right)$ be the market share-weighted average marginal uncorrected externality at the product category level. Last, $\bar{\epsilon}_{-g,g} = \frac{1}{1-s_g} \sum_{g' \neq g} s_{g'} \bar{\epsilon}_{g',g}$ is the market share-weighted average cross-elasticity between product category g and all other categories. The only difference with the intuition from the product-level optimal tax is that now the relevant externality concepts over product categories, and that category level elasticities include all within-category cross-elasticities. Formula (1.5) also necessitate an assumption on the covariance between category-level average marginal externality and average cross-elasticity, as discussed in appendix 1.D.

The constrained-optimal Pigouvian tax is now $t_g = \tilde{\phi}_{g,g}$. If U is a nested CES function, the optimal tax is proportional to the category-level average carbon intensity:

$$t_g = \sum_{k \in g} \phi_k \cdot s_k^g = \frac{\lambda}{\alpha} \sum_{k \in g} s_k^g \cdot \frac{E_k}{p_k} = \frac{\lambda}{\alpha} \frac{\sum_{k \in g} E_k \cdot x_k}{\sum_{k \in g} p_k \cdot x_k} \tag{1.6}$$

1.4.3 Deriving change in utility and carbon emissions following a change in tax schedule

Counterfactual utility change. I now derive a formula for the impact of a change in the tax schedule on the overall welfare. Reduced form or sufficient statistics approaches resulting from applying Harberger's (1964) insight would not work in this setting for two reasons. First, the most general formulation cannot be brought to the data, as I would need to compute unrestricted cross-elasticities across all goods in the economy. Second, I want to consider potentially large changes in tax schedule, affecting multiple taxes, so that the assumptions needed to derive empirically implementable reduced form formula are not justified in this context. Instead, I assume a specific functional form for the household's utility U, which enables me to express the change in utility as a function of an appropriately defined household-specific price index. In particular, I specify utility function U as a nested CES following Broda and Weinstein (2006, 2010) because it is theoretically tractable, it has desirable aggregation properties which enables me to exploit the structure of this data in a theoretically grounded manner, and because this functional form is empirically implementable. In this case, contrary to the optimal tax framework above, I assume that the tax revenue is not rebated to the household so that total household income stays constant.

I assume that goods $x_1, ..., x_N$ are food products, and that good x_0 represents all other goods. Food and non-food products are related by a Cobb-Douglas function $U(x_0, ..., x_N) = x_0^{1-\beta} \cdot U_F(x_1, ..., x_N)^{\beta}$, so that households spend a constant fraction β of their expenditure on food products. In turn, U_F is a four-level nested CES function: food product j is nested in product module m, which is nested in product group g, itself nested in food department d. At the lowest level, we have:

$$C_{mt} = \left(\sum_{j \in \Omega_{mt}} (d_j \cdot x_{jt})^{\frac{\sigma_m - 1}{\sigma_m}} \right)^{\frac{\sigma_m}{\sigma_m - 1}}$$

where C_{mgdt} is the composite consumption index of product module m at time t, Ω_{mt} is the set of product in module m at time t, d_j is a time-invariant unobservable quality component (which can also be interpreted as an unobservable taste), x_{jt} is quantity purchased of good j at time t, and $\sigma_m > 1$ is the elasticity of substitution between products, within module m. Composite consumption indices at higher levels are defined in the same manner:

$$C_{gt} = \left(\sum\nolimits_{m \in \Omega_g} C_{mt}^{\frac{\sigma_g - 1}{\sigma_g}} \right)^{\frac{\sigma_g}{\sigma_g - 1}} \quad C_{dt} = \left(\sum\nolimits_{g \in \Omega_d} C_{mt}^{\frac{\sigma_d - 1}{\sigma_d}} \right)^{\frac{\sigma_d}{\sigma_d - 1}}$$

$$U_F = \left(\sum_{d \in \Omega} C_{dt}^{\frac{\sigma - 1}{\sigma}} \right)^{\frac{\sigma}{\sigma - 1}}$$

where σ_g, σ_d and σ are the elasticities of substitution within product groups, within department and across departments respectively. Ω_g, Ω_d and Ω are the set of product modules in product group g, the set of product groups in department d and the set of product departments.

In this setting, the impact of a change in the price schedule $d\mathbf{q}$ following a tax change on household utility can be expressed as:

$$\frac{d\ln U}{d\mathbf{q}} = -\beta \cdot d\ln P_F \tag{1.7}$$

where P_F is the CES ideal price index for food, and $d\ln P_F = \ln P_{F1} - \ln P_{F0}$ is the inflation rate between the counterfactual and the initial situations. More generally, let us denote y_0 and y_1 the values of variable y before and after the tax change respectively. Using the insights from Sato (1976) and Vartia (1976), we can express $d\ln P_F$ as a function of the counterfactual change in prices and market shares only, see appendix 1.D. Furthermore, under assumption 2, appendix 1.D shows that all that is needed to compute $d\ln U/d\mathbf{q}$ are the counterfactual changes in prices, the elasticities of substitution and the initial market shares.

Assumption 2. *Within product-modules, changes in market shares following a change in tax schedule are small, so that* $\frac{s^m_{j1}-s^m_{j0}}{s^m_{j0}} \approx d\ln s^m_{j0}$.

This assumption is less restrictive than the usual assumption of small tax changes. Indeed, this assumption places no restriction on the size of tax changes, nor on the overall change in market share of good j, as $s_j = s^m_j \cdot s_m$. Under assumption 2, counterfactual market shares can be expressed as:

$$d\ln s^m_j \approx (1-\sigma_m) \cdot (d\ln p_j - \sum_{j\in\Omega_m} s^m_{j0} \cdot d\ln p_j)$$

Note that if assumption 2 is violated, equation (1.7) can be interpreted as a first order approximation of private utility change.

Counterfactual carbon emissions. I now show how to recover counterfactual carbon emissions in this framework. We are interested in

$$\Delta E = \sum_{j=1}^{I} E_j \cdot \Delta x_j$$

So that only counterfactual quantities demanded after the change are needed. Using the nested CES structure, one can show that

$$\begin{aligned} d\ln x_j =& -\sigma_m d\ln p_j + (\sigma_m - \sigma_g) d\ln P_m + (\sigma_g - \sigma_d) d\ln P_g \\ &+ (\sigma_d - \sigma) d\ln P_d + (\sigma - 1) d\ln P_F \\ =& -\sigma_m(d\ln p_j - d\ln P_m) - \sigma_g(d\ln P_m - d\ln P_g) \\ &- \sigma_d(d\ln P_g - d\ln P_d) - \sigma(d\ln P_d - d\ln P_F) - d\ln P_F \end{aligned} \tag{1.8}$$

So that we have: $\Delta x_j = \exp\left(d\ln x_j + \ln x_{j0}\right) - x_{j0}$. Equation (1.8) highlights the role of taxation at different category level on quantity demanded. Intuitively, a proportional tax on every products of module m implies that $d\ln P_m = d\ln p_j\ \forall j$, so that the elasticity of substitution between products within module m, σ_m, does not impact the quantity demanded. Further, in such a case, the change in quantity demanded is the same across products of the same module, so a product module level tax does not leverage the within-product category heterogeneity in carbon intensity to reduce emissions.

1.4.4 Effectiveness of policy targeting.

Carbon reduction and policy targeting This framework can also be used to understand what determines the effectiveness of a given tax schedule. Let M_F be the total expenditure on food and e_{i0} the pre-tax carbon intensity of good i, so that following a change in tax schedule, $\Delta E = \sum_i E_i \cdot \Delta x_i \approx \sum_i E_i \cdot x_i \cdot d\ln x_i = M_F \cdot \sum_i e_i \cdot s_i \cdot d\ln x_i$, and $E = M_F \cdot \bar{e}$.

Then, I can express the change in carbon emissions as

$$\frac{\Delta E}{E} = -d\ln P_F - \frac{\sigma}{\bar{e}}\mathcal{T} - \frac{1}{\bar{e}}\sum_d s_d \cdot \sigma_d \mathcal{T}_d - \frac{1}{\bar{e}}\sum_g s_g \cdot \sigma_g \mathcal{T}_g - \frac{1}{\bar{e}}\sum_m s_m \cdot \sigma_m \mathcal{T}_m \tag{1.9}$$

where $\mathcal{T}_m \equiv cov^{i\in m}\left(e_i - \bar{e}_m, d\ln p_i - d\ln P_m\right) = \sum_{i\in m} s_i^m \cdot (e_i - \bar{e}_m)(d\ln p_i - d\ln P_m)$ is the expenditure-share weighted covariance between excess carbon intensity relative to the module average and differential price change relative to the module average.

Similarly, $\mathcal{T}_g \equiv cov^{m\in g}\left(\bar{e}_m - \bar{e}_g, d\ln P_m - d\ln P_g\right)$, $\mathcal{T}_d \equiv cov^{g\in d}\left(\bar{e}_g - \bar{e}_d, d\ln P_g - d\ln P_d\right)$ and $\mathcal{T} \equiv cov^d\left(\bar{e}_d - \bar{e}, d\ln P_d - d\ln P_F\right)$.

$\mathcal{T}_m$ can be interpreted as a measure of quality of policy targeting within module m: the higher the covariance, the more targeted towards high carbon intensity products the tax schedule is. $\mathcal{T}_g$, $\mathcal{T}_d$ and $\mathcal{T}$ can similarly be interpreted as the quality of policy targeting within product group g, product department d and across product departments respectively.

Equation (1.9) delivers four insights. First, at each product category level, the effectiveness of a change in tax schedule $d\mathbf{q}$ is the product of the quality of policy targeting and the substitution capacity. For instance, the effectiveness of $d\mathbf{q}$ on reducing carbon emissions from module m depends whether the policy targets relatively more carbon intensive products within this module (weighted by their market-share), but also whether substitution within this module is easy, as quantified by σ_m. Effective policies need to combine both aspects. Second, optimal Pigouvian taxes maximize policy-targeting in the following sense. For any optimal tax schedule of a given variance $var(t_j)$, we have that $\mathcal{T}_m = p^{CO_2} \cdot var^{i\in m}\left(e_i - \bar{e}_m\right)$ with $p^{CO_2} \equiv \lambda/\alpha$ being the social cost of carbon. This follows from $t_i = \phi_i = p^{CO_2} \cdot e_i$, $d\ln p_i = p^{CO_2} e_i$ and $d\ln P_m = p^{CO_2}\bar{e}_g$. Causchy-Schwartz inequality implies that it is greater than any other tax schedule with the same variance. The result also holds at other product category level in case we are considering constrained optimal Pigouvian taxes. Third, equation (1.9) quantifies the additional effectiveness from using tax instruments at a finer level. At indeed, suppose that a proportional tax on all goods j is introduced. Then, $d\ln p_j = d\ln P_m = d\ln P_g = d\ln P_d = d\ln P_F \; \forall j, m, g, d$, so that $\mathcal{T}_m = \mathcal{T}_g = \mathcal{T}_d = \mathcal{T} = 0$, as the uniform tax does not target more carbon intensive goods. Following a uniform carbon tax, $\Delta E = -d\ln P_F \cdot \bar{e} \cdot M_F = -d\ln P_F \cdot E$ so that carbon reduction simply comes from decreased overall consumption due to higher prices. $M_F \cdot \sigma\mathcal{T}$ is then the additional carbon saved from the availability of tax rate at the product department level, and $M_F \cdot \sum_d s_d \cdot \sigma_d \mathcal{T}_d$ is the additional carbon saved from the availability of tax rate at the product group level, etc. Fourth, combining equations (1.7) and (1.9) imply that when we allow for subsidies in addition to taxes, there exists tax schedules that do not change overall price level hence private utility (so that $d\ln P_F = 0$), while reducing carbon emissions as long as $d\mathbf{q}$ is a mean-preserving spread around zero that is positively correlated with $\bar{e}_i$, so that $corr(\bar{e}_i, dq_i) > 0$. It is unclear however whether these changes can be budget-balanced or not.

Utility cost of reaching a specific carbon emission target. I now reverse the problem and show that for a given target in emission reduction $\overline{\Delta E}$, taxation at a finer level is always more efficient, that is achieves smaller private utility loss and social welfare. The extent of this increased efficiency is driven by the product of the variance of the within category carbon intensity and the associated elasticity of substitution. For a given target emission $\overline{\Delta E}$, we have the following results:

$$d\ln U^{unif} = \beta \frac{\overline{\Delta E}}{M_F \cdot \bar{e}} = \beta \frac{\overline{\Delta E}}{E} \qquad p^{CO2,unif} = \frac{1}{\bar{e}} \cdot \frac{-\overline{\Delta E}}{E} \tag{1.10}$$

$$\frac{d\ln U^{dep}}{d\ln U^{unif}} = \frac{p^{CO2,dep}}{p^{CO2,unif}} = \frac{\bar{e}}{\bar{e} + \sigma \cdot var\left(\bar{e}_d - \bar{e}\right)} \equiv \frac{\bar{e}}{\bar{e} + \mathcal{E}^{dep}} \leq 1 \tag{1.11}$$

$$\frac{d\ln U^{group}}{d\ln U^{unif}} = \frac{p^{CO2,group}}{p^{CO2,unif}} = \frac{\bar{e}}{\bar{e} + \mathcal{E}^{dep} + \sum_d s_d \cdot \sigma_d \cdot var^{g\in d}\left(\bar{e}_g - \bar{e}_d\right)} \equiv \frac{\bar{e}}{\bar{e} + \mathcal{E}^{dep} + \mathcal{E}^{group}} \leq 1$$

$$\frac{d\ln U^{mod}}{d\ln U^{unif}} = \frac{p^{CO2,mod}}{p^{CO2,unif}} = \frac{\bar{e}}{\bar{e} + \mathcal{E}^{dep} + \mathcal{E}^{group} + \sum_g s_g \cdot \sigma_g \cdot var^{m\in g}\left(\bar{e}_m - \bar{e}_g\right)} \leq 1$$

$$\frac{d\ln U^{prod}}{dd\ln U^{unif}} = \frac{p^{CO2,prod}}{p^{CO2,unif}} = \frac{\bar{e}}{\bar{e} + \mathcal{E}^{dep} + \mathcal{E}^{group} + \mathcal{E}^{mod} + \sum_m s_m \cdot \sigma_m \cdot var^{j\in m}\left(e_j - \bar{e}_m\right)} \leq 1$$

Where $d\ln U^{unif}$ is the log change in household's utility following a optimal uniform tax necessary to achieve $\overline{\Delta E}$, $d\ln U^{dep}$ is the log change in household's utility following a optimal department-level Pigouvian tax necessary to achieve $\overline{\Delta E}$, etc. $p^{CO2,unif}$ is the implicit social cost of carbon given the constraint and the available tax instruments. Equation (1.10) is intuitive. Given the unit price-elasticity of food from the Cobb-Douglas form of the higher nest of U, reducing carbon emissions by $x = \frac{\overline{\Delta E}}{E}$ % using a uniform tax on food implies reducing uniformly quantity of food products by $x\%$ from a $x\%$ increase in prices, implying a utility loss of $\beta x\%$.

When targeting finer product categories is feasible, the welfare cost of achieving a $x\%$ reduction in carbon emissions is strictly smaller as with the uniform tax as long as carbon intensity varies across categories. Intuitively, this allows to focus carbon emission reduction efforts by setting higher taxes on more carbon intensive categories, so that fewer goods see a price increase. The efficiency gains relative to the uniform tax situation will be higher the higher the elasticity of substitution, that is the lower the utility cost of substituting across products. In particular, the social cost of carbon implicitly set by the reduction constraint is actually lower the more flexible policy instruments are.

1.4.5 Calibrating the impact on social welfare of a change in tax schedule

I am now interested in expressing the impact of a change in tax schedule $d\mathbf{q}$ on overall social welfare expressed in monetary units per household: $\frac{1}{\alpha}W = \frac{1}{\alpha}U - \frac{\lambda}{\alpha}E + T$, where T is the tax revenue and $\frac{\lambda}{\alpha}\Delta E$ is the monetary value of the change in carbon emissions to the social planner . The marginal cost of public funds is assumed to be one here, as W is devided by the household's marginal utility of income. Further, I do not consider potential uses of the tax revenue, nor the associated carbon emissions (these could be positive, if for instance the revenue is rebated lump sum to the household who uses it to increase its carbon emissions, or negative, if it is used to fund carbon negative investments). In particular, I am interested in $\Delta W/\alpha = \frac{\Delta U}{U} \cdot U/\alpha - \frac{\lambda}{\alpha}\Delta E + \Delta T$.

To estimate $\Delta W/\alpha$, we can recover ΔE and ΔT from the data, $\frac{\Delta U}{U}$ can be recovered from the data and by calibrating $\beta = 11.3\%$ using the 2018 share of household's final consumption on take-home food and drinks[7]. Further, I calibrate U/α to $\frac{55{,}780}{6{,}303/1{,}510} \approx 13,363$ €, which is the average household's final consumption in 2018[8], scaled to account for the fact that only a share of all food expenditure is observed. Specifically, national accounts suggest that the average French household spend 6,303€ on food and non-alcoholic beverages, whereas in this data, the average household spends only 1,510€ a year on the selected products for the analysis. Last, the social cost of carbon $p^{CO_2} = \frac{\lambda}{\alpha}$ can be set by the researcher or the policy maker.

[7] Expenditure on food and non-alcoholic beverages represents 10% of final consumption and expenditure on alcoholic beverage is 1.3%, see Figure 1 of https://www.insee.fr/fr/statistiques/4277709; accessed 1st October 2021.

[8] Equal to 1628.8 bn€, see https://www.insee.fr/en/statistiques/4132094, divided by 29.2 millions, the number of households in France in 2016, the latest available year, see https://www.insee.fr/fr/statistiques/4277630; accessed 1st October 2021.

Two other statistics are useful to compare different policies. First, $\frac{\Delta U/U}{\Delta E} \approx \frac{-\beta d \ln P_F}{\Delta E}$ is the percentage reduction in private utility per unit of carbon reduced, not taking into account tax rebate. Second, adding another layer of calibration, $\frac{\Delta W/\alpha}{-\Delta E} = p^{CO_2} + \frac{\frac{\Delta U}{U} \cdot \frac{U}{\alpha} + \Delta T}{-\Delta E}$ is the change in welfare expressed in euros per unit of carbon saved by a change $d\mathbf{q}$.

1.5 Empirical model and estimation

The previous section established that elasticities of substitution are a key ingredient to design optimal policies and to measure their effectiveness. In this section, I estimate elasticities of substitution from the data. Due to the nature of the data in which we observe purchases from one retail chain only, the estimated elasticities are likely to be lower bounds on true elasticity of substitution. However, as I focus on loyal customers only, the magnitude of the bias is likely to be attenuated.

1.5.1 Empirical model and instrument

Empirical model.

Four sets of elasticities are estimated: within product module, within product group, within department and across departments. To estimate product module level elasticities, I exploit the relationship between market shares and prices and their geographic variation across French geographic departments (not to be confused with product departments). There are 96 geographic departments in metropolitan France, so this is a geographic unit between the US states and counties. The CES structure of the utility function implies that:

$$\ln s_{jct}^{m} = (1 - \sigma_m) \ln p_{jct} + (\sigma_m - 1) \ln d_{jc} + (\sigma_g - 1) \ln P_{mct}$$

where $\ln s_{jct}^{m}$ is the log market-share of product j within module m in geographic department c in quarter t, σ_m is the constant elasticity of substitution within product module m, p_{jct} is the log price, d_{jc} is the time-invariant unobservable quality and P_{mct} is the price index of module m in geographic department c in quarter t. I difference out the unobserved quality by estimating

$$\Delta \ln s_{jct}^{m} = (1 - \sigma_m) \Delta \ln p_{jct} + \delta_{mct} + X_{jct} + \epsilon_{jct}$$

where δ_{mct} is a product module by time by geography fixed effect, X_{jct} is a vector of controls and ϵ_{jct} is a mean-zero disturbance. This specification is preferred to a one with $\ln q_{jct}$ are dependent variable and $\ln p_{jct}$ as independant variable as one cannot aggregate product quantities across categories. These elasticities are estimated for the overall population of households, but also by quartile of the expenditure distribution and quartile of the carbon intensity distribution. In all the regressions for this section, standard errors are clustered at the geographic department levels. Observations are weighted by number of transactions and these regressions are run separately for each product departments.

The vector of controls X_{jct} includes a constant and the share of sales of product j in quarter $t$and geographic department c coming from promotions. In the data, a transaction is flagged as a promotion if one or more of the following is true: the product has a reduced price, is part of a 'bulk' promotion such as two-for-one promotions, or has been put in a particular display, such at the end of an aisle. Controlling for X_{jct} eliminates variation in prices and market shares due to short-term promotions which is important since we want to capture long-term elasticities of substitution.

Instrumental variable approach.

Consistent OLS estimation requires that $\mathbb{E}\left[\Delta \ln p_{jct} \cdot \epsilon_{jct}\right] = 0$ conditional on the controls. This is likely not to be the case. Indeed, within a quarter, product module and geographic department, the change in local prices could be correlated to a variety of local time-varying demand shifters captured by ϵ_{jct}, for instance if stores are increasing prices in times of higher localdemand. To address this endogeneity bias, I instrument the change in log prices with the average change in log prices for the same product in other geographic departments. This instrument is a version of the one introduced by Nevo (2001) and Hausman and Bresnahan (2008), and studies using scanner data regularly build on this strategy (DellaVigna and Gentzkow, 2019; Allcott et al., 2019b,a). The IV instrument is:

$$\overline{\Delta \ln p_{jt,-c}} = \frac{1}{C-1} \sum_{c' \neq c} \Delta \ln p_{jct}$$

where C is the total number of geographic departments. For the instrument to be valid and the estimation to be consistent, two conditions must be satisfied. First, the instrument must be relevant, that is it must be correlated with the endogenous variable: $\mathbb{E}\left[\overline{\Delta \ln p_{jt,-c}} \cdot \Delta \ln p_{jct}\right] \neq 0$ conditional on the controls. Second, the instrument must satisfy the exclusion restriction that $\mathbb{E}\left[\overline{\Delta \ln p_{jt,-c}} \cdot \epsilon_{jct}\right] = 0$. Relevance can be directly checked from the data from the first-stage. Panel (a) of figure 1.4 shows that the first stage is extremely strong with a coefficient close to 1, which is very similar to what DellaVigna and Gentzkow (2019) obtain using scanner data in the United States. Results from the reduced form regression is plotted in panel (b) of figure 1.4.

The identifying assumption behind the exclusion restriction is that within a quarter, product module and geographic department, the local, time-varying demand shocks ϵ_{jct} that are not driven by promotions are not related with national level changes in baseline prices. While this assumption cannot be checked in practice, the detailed set of time by geography by product module fixed effects, as well as controlling for promotions, makes it more likely to hold. For instance, a nationwide advertising campaign for a specific product, which would impact its national-level prices and correlate with an unobserved demand shock, would be absorbed by the controls X_{jct}. Further, any demand shock which would violate the exclusion restriction would need to impact differentially the products within the same product module, within the same time period and within the same geographic department. In my view, the granular nature of product modules makes such demand shock implausible.

Following standard practice in the literature (Broda and Weinstein, 2010; DellaVigna and Gentzkow, 2019; Jaravel, 2019, etc.), I impose restrictions on the estimated elasticities. In particular, I require elasticities to be greater than 1.05 and to have standard errors in the range of $[0.01, 1.25]$. When this is not the case, I impute product module elasticity using the median value of correctly estimated elasticities within same the product group.

Estimating higher level elasticities. In order to estimate elasticities at the product group and product department model, as well as overall elasticity across product department I follow Broda and Weinstein (2010) and reproduce my estimation procedure at a higher level. Specifically, the market shares are recovered from the data and prices are aggregated using the exact CES price index deriving in appendix section 1.D.

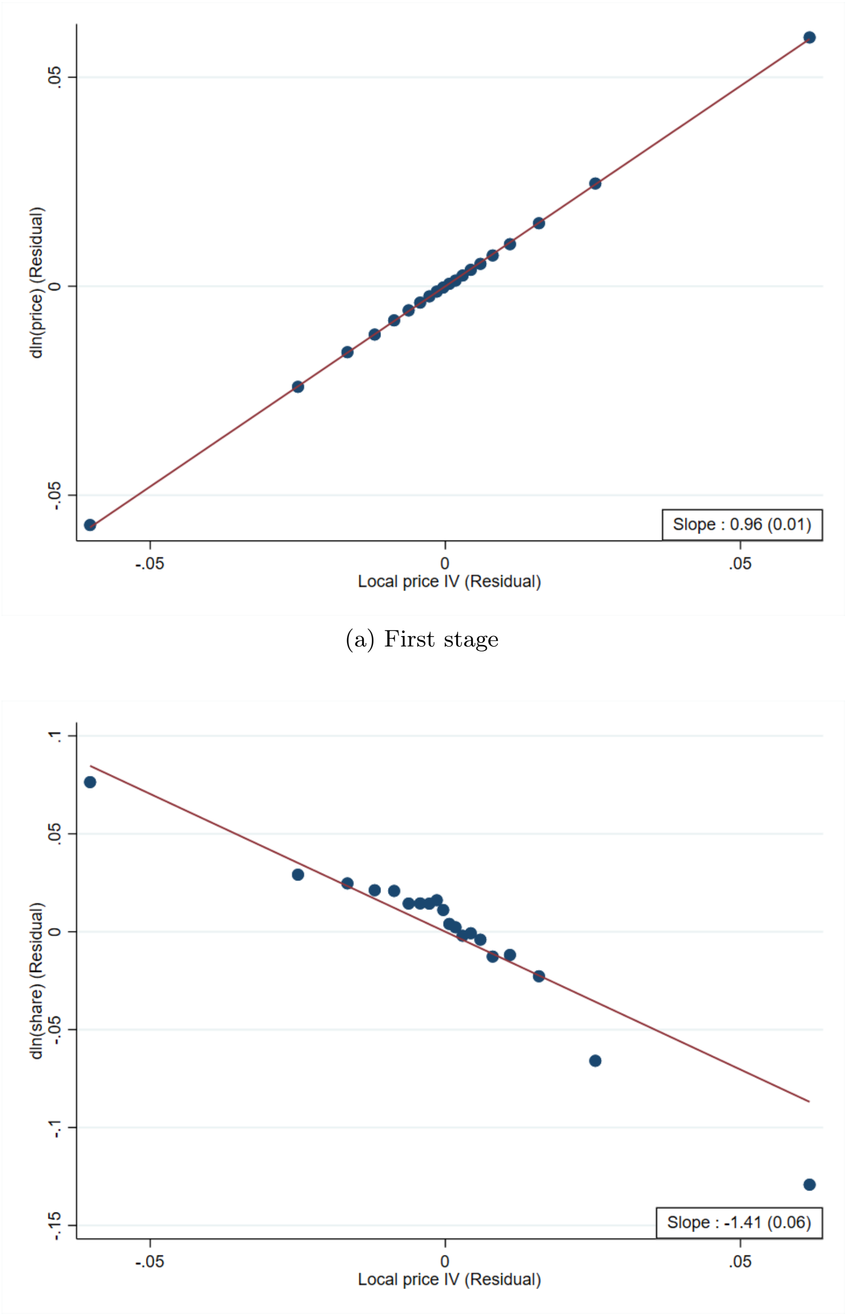
1.5.2 Estimation results

Figure 1.5 plots the distribution of estimated elasticities of substitution at the product module level. Most of the distribution mass within 1 and 5, which falls within the generally accepted ballpark estimates for this type of elasticities. By comparison, DellaVigna and Gentzkow (2019) find median product price elasticities of 2.5-3 but focus on 40 product modules only. This can roughly compared with my estimates since under CES utility, own-price elasticity is also equal to the elasticity of substitution. Allcott et al. (2019a) find a price elasticity of sugar sweetened beverages of 1.48[9], and out of the 6 product modules associated with sugar sweetened beverages, the median estimated elasticity of substitution within module is 1.34 in this data. By contrast, Broda and Weinstein (2010) estimate elasticities of substitution for 122 product groups with a median elasticity of 11.5 but using a more structural estimation approach.

Table 1.2 plots summary statistics of the distribution of estimated elasticities at within product modules, group, department and across departments (column overall). The table is consistent with the theory as elasticities of substitution are smaller at a higher level since we are comparing less similar products.

[9] Column 3 of table III.

Figure 1.4: Carbon - expenditure elasticity by product category

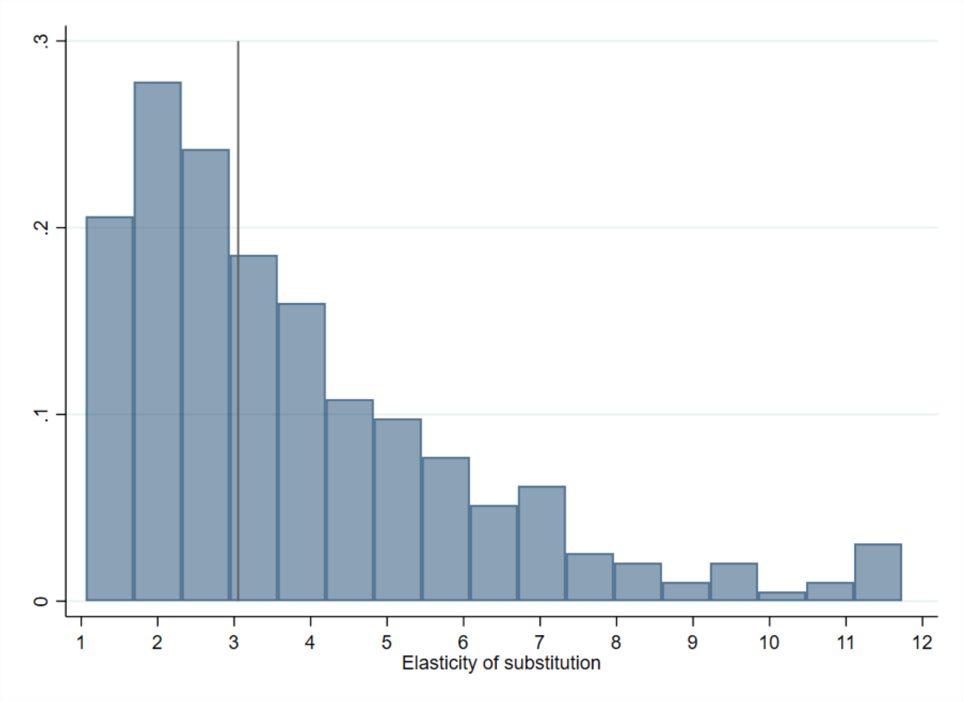


(a) First stage

(b) Reduced form

Notes: This figure plots the first-stage (panel (a)) and reduced form estimates (panel (b)) of the instrumental variable regression discussed in section 1.5. Variables are residualized on a product module by quarter by geographic department fixed effects. Standard errors are clustered at the geographic department level.

Figure 1.5: Histogram of estimated product module elasticities



Notes: This figure plots the histogram of the 309 product module elasticities of substitution estimated as described in section 1.5. Vertical line indicates the median elasticity. Figure is winsorized at the 99-th percentile.

Table 1.2: Estimated elasticities of substitution

	Module	Group	Department	Overall
5-th percentile	1.34	1.47	0.79	1.28
25-th percentile	2.21	1.68	1.15	1.28
Median	3.04	2.81	1.79	1.28
75-th percentile	4.50	3.43	2.81	1.28
95-th percentile	7.76	6.57	4.45	1.28
N	435	51	10	1

Notes: Overall elasticity of substitution denotes elasticity across product departments, other elasticities are within the given product category.

Appendix table 1.B.8 compares the estimated product module elasticities across a range of different samples. Focusing the distribution of households ranked by monthly expenditure, we can see that the median product module elasticity is slightly higher for the bottom 25% than for the top 25%, consistent with poorer households being more price-sensitive. Interestingly, the 25% of households with highest carbon intensity have a much larger estimated median elasticity than the 25% of households with the smallest carbon intensity. Overall, the fraction of imputed elasticities is of the same order of magnitude as the fraction of elasticities estimated by grid search in Jaravel (2019) and Broda and Weinstein (2010).

As robustness checks, panel (a) of appendix figure 1.C.4 plots the histogram of the baseline IV and OLS estimates. Consistent with the assumption that unobservable demand shocks are positively correlated with prices and the fact that $\sigma_m = 1 - \hat{\beta}$, with $\hat{\beta}$ being the estimate, elasticities obtained from OLS are smaller than from IV. Panel (b) plots the histogram of the baseline IV estimate as well as estimates obtained from IV without controlling for promotions. Controlling for promotion does indeed shrinks the value of estimated elasticities, as could be expected from long-run elasticities being smaller than short-run elasticities.

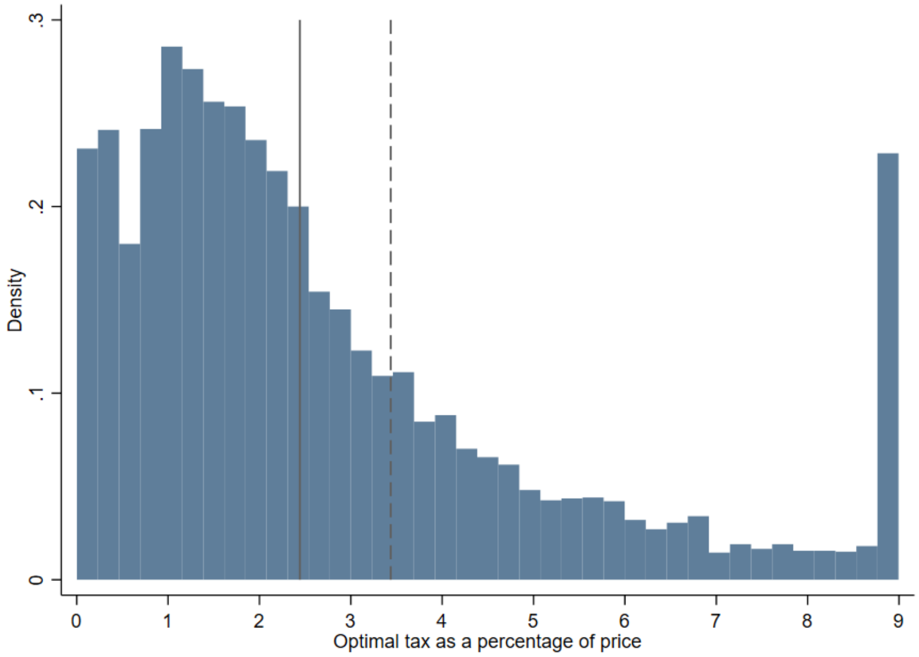
1.6 Results

This section analyses various policies aimed at reducing carbon emissions and quantifies the different drivers of their effectiveness. Unless indicated otherwise, optimal taxes are based on a social price of carbon of $p^{CO_2} = 50$ € per ton. This choice can be justified by the fact that it is close to the prevailing estimates used for policy decisions for instance in the US (2021), although many academic and policy publications suggest this value is seriously underestimated (Stiglitz et al., 2017; Weitzman, 2014; Stern and Stiglitz, 2021). As of September 2021, futures for a ton of carbon have been trading at between 50€ to 60€ over the last five months on the European ETS market.

1.6.1 Optimal and constrained optimal taxes

Figure 1.6 plots the histogram of product-level optimal Pigouvian taxes $t_j^* = p^{CO_2} \cdot e_j$, expressed as a percentage of pre-tax prices. The median and mean tax rates are modest, at 2.4 % and 3.4 % respectively. At a carbon price of 150€ per ton, they would amount to 7.3 % and 10.3 % respectively. While the average tax rate is modest, 5% of the products face an optimal tax rate of 9% or higher. Table 1.3 shows how, under CES utility, constrained optimal Pigouvian taxes would look like when they are set a the product module, group or department level. As established in section 1.4, the market-share weighted average tax rate is constant, at $p^{CO_2} \cdot \bar{e} = 3.4\%$.

Figure 1.6: Product level optimal taxes



Notes: This figure plots the histogram of the product level optimal Pigouvian tax based on a carbon price of 50€ per ton. Full and dashed lines indicate the (pre-tax) market-share weighted median and mean of the distribution, respectively. Top 5% of the distribution is winsorized.

Table 1.3: Optimal tax rates

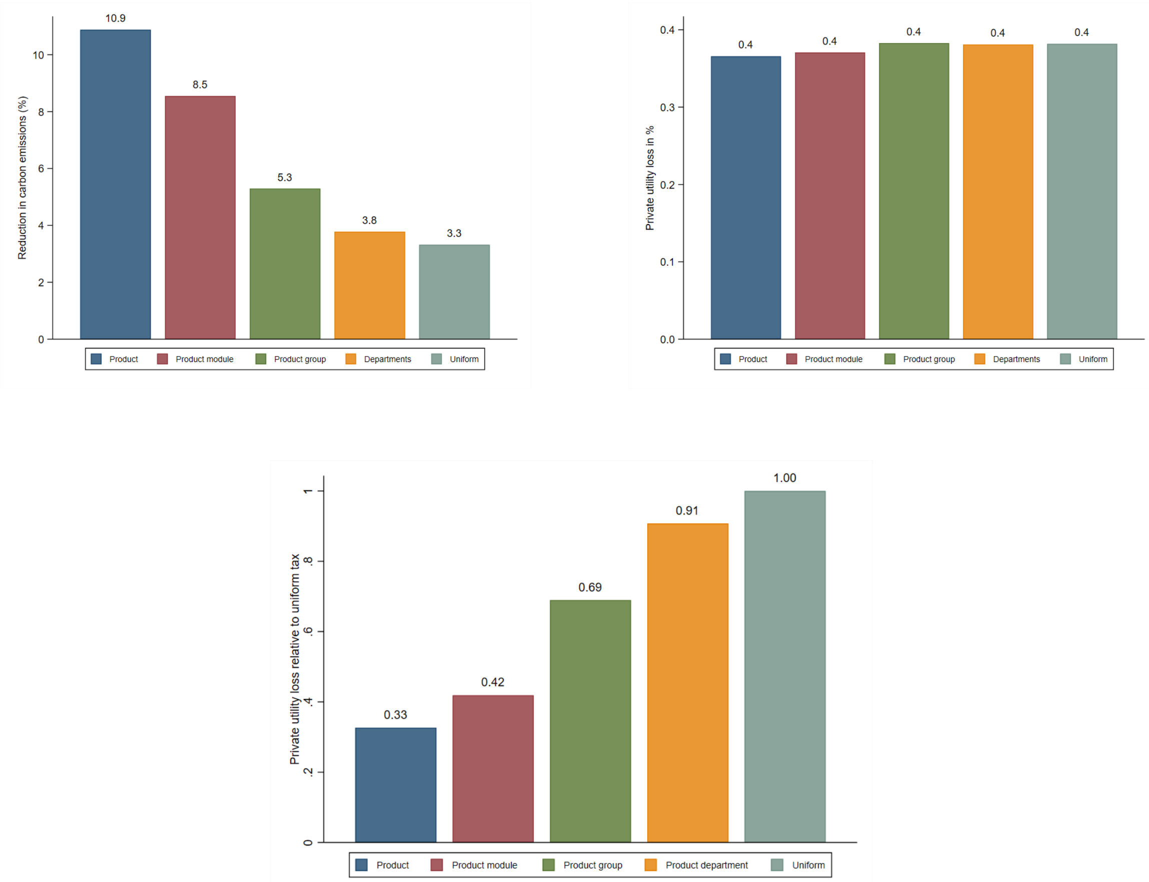
	Product	Module	Group	Department	Uniform
Mean	3.4	3.4	3.4	3.4	3.4
Std. dev.	3.9	3.3	2.1	1.2	0.0
p1	0.1	0.1	0.1	1.7	3.4
p10	0.5	0.6	1.3	1.7	3.4
p25	1.2	1.5	2.1	2.3	3.4
Median	2.4	2.7	3.1	3.3	3.4
p75	4.2	4.3	4.2	3.8	3.4
p90	7.3	6.4	6.6	4.9	3.4
p99	17.1	14.6	8.4	7.3	3.4
# tax rates	8,650	435	51	10	1
# products	8,652	8,652	8,652	8,652	8,652

Notes: Optimal taxes are expressed as a percentage of pre-tax price, and are based of a price of carbon of 50€ per ton. Observations are weighted by the (pre-tax) market-share.

Panels (a) and (b) of figure 1.7 plots results from implementing equations (1.7) and (1.8) at different levels of the product taxonomy. For reference, this figure includes the impact of a uniform tax on all food products. Comparing this tax policy to policies set at finer level illustrates what part of reduction in carbon emissions is due to substitution between more or less carbon intensive products, and what part is due to a pure negative income effect following the reduction of the consumer's budget set (see equation (1.9)). Panel (a) summarizes one of the key messages of the paper: implementing a carbon tax, or another policy aimed at reducing carbon emissions, at a granular level is much more efficient than the same policy implemented at a coarser level because the former leverages the high elasticities of substitution between closely substitutable products. Specifically, in my data and given the estimated elasticities, setting a product-level optimal carbon tax would reduce carbon emissions by 10.2% and be nearly three times as efficient in this regards than setting an optimal carbon tax at the product department level. Even smaller scale, and perhaps more implementable, changes in the level at which policy is set would be meaningful: as going from a department level carbon tax to a product group level carbon tax means a 40% higher reduction in carbon emissions, from 3.8% to 5.3%. To put these numbers into perspective, a reduction of 5.3% in carbon emissions from food represents 5% of the effort needed to achieve France's objective to reduce carbon emissions by 26% by 2030, assuming food represents 26% of all emissions. Overall, this is a small but non-negligible number. Further, panel (b) of figure 1.7 shows that the difference in carbon emission reduction between coarse and fine-grained taxes does not translate into differential change in private utility loss. Indeed, under CES preferences, optimal and constrained optimal all lead to the same average price increase, driving the overall loss in utility from equation (1.7).

Panels (a) and (c) of appendix figure 1.C.5 reproduce these results for a carbon price of 150€ a ton. At this price, an optimal tax set at the department level reduces carbon emissions by 16.2% whereas a tax set at the product module level reduces emissions by 27.6%, roughly at the same private utility cost. Panel (b) and (d) plot the change in carbon emissions and private utility for a range of carbon prices. Panel (b) shows that the relation between in carbon emissions and carbon prices is close to linear for uniform carbon taxes and has a concave shape for taxes at finer levels. Panel (b) suggests that a price of carbon between 75€ and 250€ a ton maximizes the relative efficiency of finer taxes. Intuitively this comes from the fact that carbon emission reduction is bounded by 100%. Panel (d) shows that the estimated private utility loss varies depending on the product category at which taxes are set, in apparent contradiction with the results established in section 1.4 that private utility loss is the same across all taxation level. This is because for large values of p^{CO_2}, assumption 2 that the change is market share within each product category is small is violated. At high carbon prices, utility loss from coarser taxation is much higher than from finer taxation because substitution effects are very strong in the latter, which impacts the effective tax rates and price increase faced by households. It is noteworthy that this results

Figure 1.7: Impact of optimal and constrained optimal taxes on carbon emissions and private utility



(a) Percentage reduction in carbon emission

(b) Percentage reduction in private utility

(c) Relative utility cost of achieving a specific CO2 reduction target

Notes: This figure plots the estimated percentage reduction in carbon emissions (panel (a)) and private utility (panel (b)) following optimally set taxes at different levels of the product category classification as discussed in section 1.4. Panel (c) plots the private utility loss from achieving a specific carbon reduction objective with a given policy instrument relative to uniform taxation. Numbers are based on a price of carbon of 50€ per ton.

holds remarkably well for small values of p^{CO_2}, while the violation of the assumption can already be seen in panel (c).

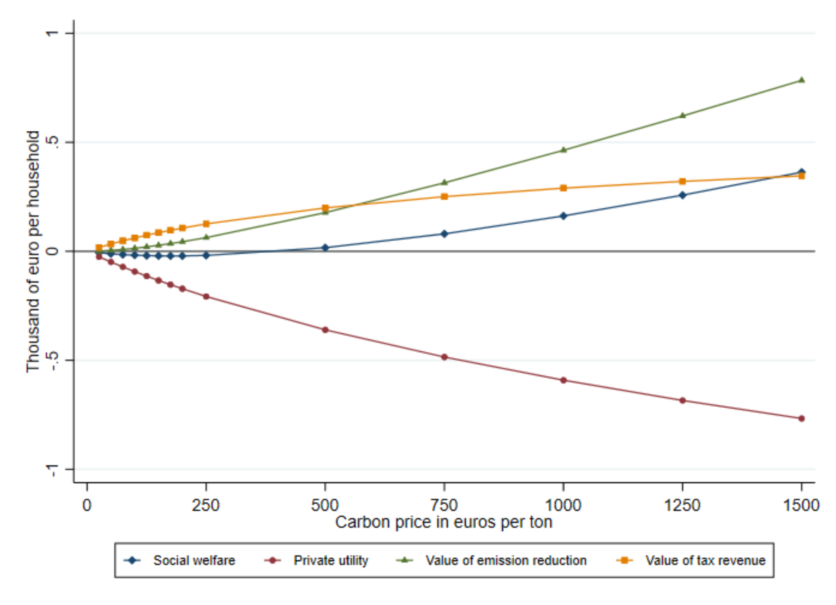
While the magnitude of change in private utility seems much smaller than the change in carbon emissions (-0.4% relative to -11% in the baseline estimate of product level taxes), comparing the two requires particular assumption on the respective baseline level. Appendix figure 1.C.6 plots the components of social welfare across the different tax policies, expressed in euro per household using the calibration assumption discussed in section 1.4.5. The key takeaway of this graph is that under current assumptions on the baseline level of household utility and value of carbon emissions, the difference in efficiency of carbon emissions reduction across tax policies translates into much smaller changes in social welfare. The overall social welfare loss implementing a product module level optimal tax amounts to 11€ per household, whereas it is 18% higher for department level optimal taxes, at 13€. The loss in private utility is nearly constant across tax policies, the difference being driven from deviations from assumption 2 as discussed above. Similarly, the value to the social planner of tax revenue varies between 34€ and 36€ per household, the difference being driven by the strength of substitution effects at different levels of the product taxonomy, reducing the effective tax burden faced by consumers. While the difference in social value of carbon emissions reduction is stark across tax policies, being four times more important for a product module level tax than for a product department level tax, in absolute value these differences are small, so that the difference between the private utility change and the social welfare change is mostly driven by tax revenue rather than the value of carbon emissions avoided. Appendix figure 1.C.7 plots another complementary but useful way to highlight the relative inefficiency for coarser tax policy by asking how much social welfare or private utility is lost for any ton of avoided carbon emission.

Naturally, these results depends on the social cost of carbon. Given a high enough social cost of carbon, any reduction in carbon emissions will lead to a positive welfare change. Figure 1.8 shows that a social cost of carbon of about 350€ per ton is necessary for a product-level optimal tax to be welfare improving. This minimum social cost of carbon necessary to justify such optimal tax from a welfare perspective is well-above the current ballpark estimates of what a social cost of carbon is. However, it highlights the discrepancy between these estimates of social cost of carbon and the social cost of carbon implicitly set by policy objectives such as the Paris agreement, as is developed below.

1.6.2 Heterogeneity across expenditure and carbon intensity levels

I now turn to the distributional impact of optimal and constrained optimal carbon taxes. In theory, heterogenous impact of such taxes could arise either because of different demand

Figure 1.8: Components of social welfare by price of carbon

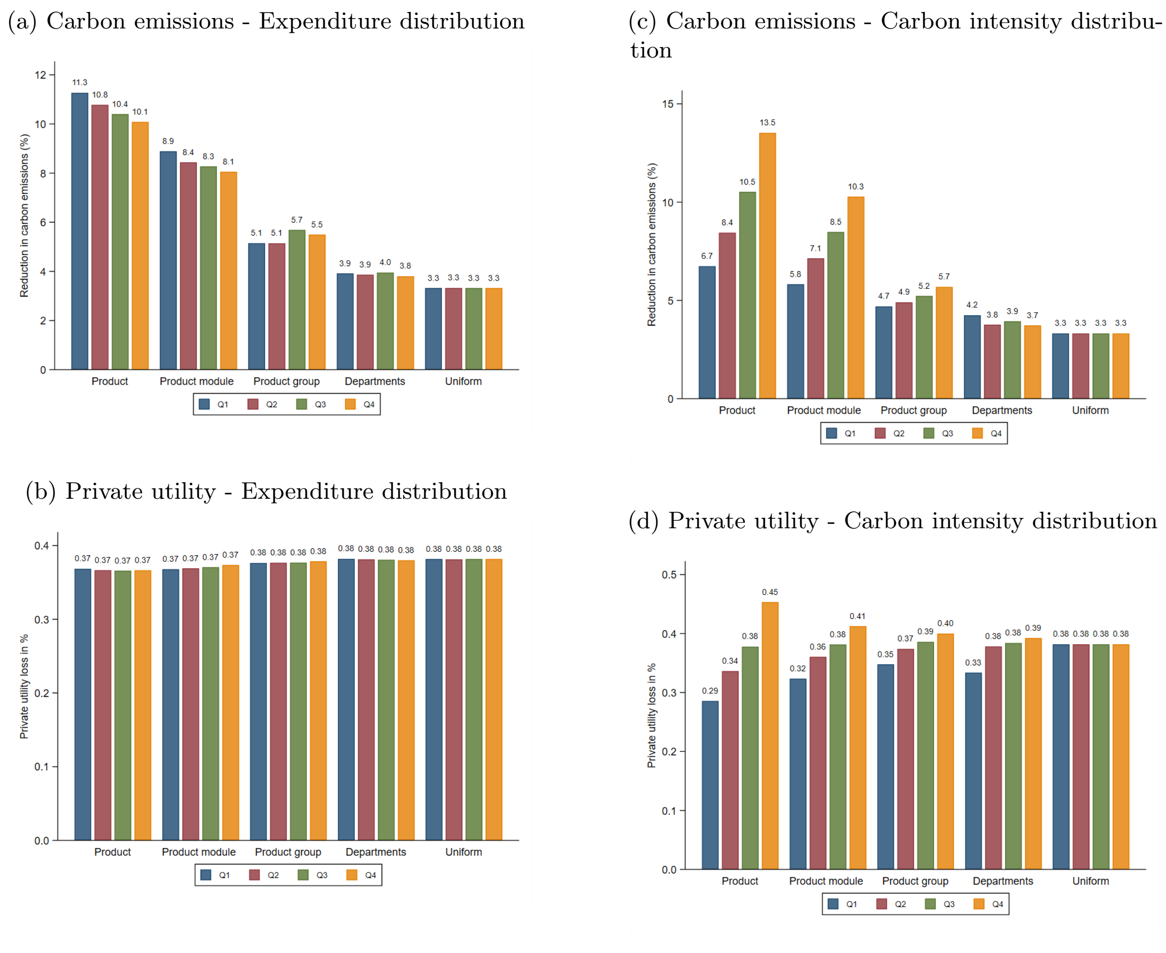


Notes: This figure plots the components of social welfare for a product level optimal tax.

structure (different σ_m, σ_g, σ_d and σ) or because of different price levels faced by the households.

Heterogeneity across expenditure distribution. Figure 1.9 plots the estimated impact of different tax policies on quartiles of the expenditure distribution for a cost of carbon of 50€ per ton. Panel (a) shows that there indeed exists a difference in the percentage reduction in carbon emissions between high and low-expenditure households when taxes are set at a fine level. Product level optimal taxes would lead to a reduction of 11.3% in CO2 from households in the bottom quartile and of 10.1% from households in the top quartile. This is driven by the fact that at fine category levels such as within product groups or product modules, low-expenditure households tend to have higher elasticities of substitution than high-expenditure households. As could be expected, the percentage reduction in CO2 is the same across quartiles when faced with a uniform tax, as this policy leave no room for differentiated substitution patterns. To a lesser extent, this insight carries over to product department level taxes: since elasticities of substitution between broad product categories such as between fats, cereals and dairy, are likely to be similar across the expenditure distribution, as is confirmed by the data, setting department-specific carbon taxes will have a very similar impact on all households. Panel (b) shows however that any of the tax policies considered here have a similar impact on private utility across the expenditure distribution. In practice, this shows that all quartiles of the distribution are facing the same expenditure-weighted aggregate price changes, whatever the tax policy considered.

Figure 1.9: Heterogeneity by quartile of different distributions



(a) Carbon emissions - Expenditure distribution

(b) Private utility - Expenditure distribution

(c) Carbon emissions - Carbon intensity distribution

(d) Private utility - Carbon intensity distribution

Notes: This figure plots the estimated percentage reduction in carbon emissions (panels (a) and (c)) and private utility (panels (b) and (d)) following optimally set taxes at different levels of the product category classification, for different quartiles of the expenditure distribution or the carbon intensity distribution. Numbers are based on a price of carbon of 50€ per ton.

Relatedly, panels (a) and (b) of appendix figure 1.C.8 confirms that these results are mostly driven by different demand structure across the distribution and not different prices leading to initial different initial expenditure patterns: the same exercise as in figure 1.9 has been reproduced assuming equal elasticities of substitution across quartiles, and under this assumption, most of the difference in CO2 reduction between high- and low-expenditure households vanishes. Panel (c) of figure 1.C.8 shows that higher expenditure households contribute disproportionally to the overall tax revenue raise and to the overall carbon reduction, purely due to their higher expenditure level. This is even sufficient to make the social welfare benefit of taxing the top quartile positive.

Heterogeneity across carbon intensity distribution. The above section shows, in consistency with the stylized facts section, that there is little quantitative difference in the impact of different tax policies across the expenditure distribution as regards changes in utility and carbon emissions. However, meaningful dimensions of heterogeneity might not be captured by expenditure only. Accordingly, I reproduce the same exercise by estimating different elasticities of substitution for quartiles of the carbon intensity distribution. Panels (c) and (d) of figure 1.9 shows that the distributional impact of different tax policies across the carbon intensity distribution varies widely. When optimal taxes are set at the product level, the bottom quartile reduces carbon emissions by 6.7%, against 13.5% for the top quartile, a 101% difference. Even when optimal taxes are constrained to be set at the product group level, the percentage reduction in carbon emission of high carbon intensity households is 21% higher than low carbon intensity households. This effect is mediated through both different elasticities of substitution and through different spending patterns. Panel (a) of figure 1.C.9 shows that when I restrict households to have the same substitution elasticities, initial differences in expenditure also matter to explain the difference in carbon emission reduction across quartile of carbon intensity. Comparing the panel (c) of figure 1.9 and panel (a) of 1.C.9, we can see that the difference in elasticities accounts for roughly 60% of the difference in carbon reduction across the top and bottom quartile, when focussing on product level or product module level taxes. Panel (d) of figure 1.9 confirms that low and high carbon intensity households are exposed to significantly different average tax rates leading to differentiated impact on private utility. Product level optimal taxes would reduce private utility of low carbon intensity households by 0.29%, against 0.45% for high carbon intensity households, a meaningful 50% larger cost.

It is important to combine these results with the stylized fact that much of the variation in carbon intensity comes from within bins of the expenditure distribution, rather than from difference across bins. Taken together, they confirm that the more targeted the tax policy, the more important the variability of its impact (CO2 reduction or utility loss) on households. However, these results also show that the impact variability happens mostly across households of similar expenditure level.

1.6.3 The utility cost of reaching a specific carbon reduction target

To shed light on the efficiency of policy targeting, I look at the same problem from a different perspective by asking what is the utility cost of different tax policies required to achieve a specific target in emission reduction. It is important to note that this does not fundamentally alter the nature of the problem at hand, since the social planner's Lagrangian $\mathcal{L} = V(\mathbf{q}; Z + T) + \lambda \left[\bar{E} - \sum_j E_j \cdot x_j\right]$ makes it clear that the relationship between the overall carbon constraint $\bar{E}$ and the social cost of carbon $p^{CO2} = \lambda/\alpha$ is mediated through $\partial\mathcal{L}/\partial\bar{E} = \lambda$. However, this perspective enables me to specify the relative efficiency of fine-tuned policy targeting, explore its drivers, and to highlight how the social cost of carbon implicitly set by hard constraints depends crucially on the available policy instruments and the preference structure of the economy.

In this section, I consider a specific reduction target of $\overline{\Delta E}/E = -0.26$, which corresponds to the overall reduction in CO2 necessary to meet France's 2030 objectives and assumes that the same level of efforts should be demanded for the food sector than the rest of the economy. Further, in order to be able to compare similar tax policies with varying degree of targeting, I compare optimal and constrained optimal tax set at each product category level.

Panel (c) of figure 1.7 plots the private utility loss, or equivalently, the implicit relative social cost of carbon from achieving a specific carbon reduction objective with a given policy tool relative to uniform taxation. The efficiency gains from having access to finer policy instrument is striking. A product specific optimal carbon carbon tax would be 67% less costly in terms of private utility than a uniform tax on food products. Given the target of 26% reduction, this suggests that a private utility loss from a uniform tax would be $0.113 \cdot 0.26 \approx 2.9\%$, whereas it would only be 1.0% for product-level taxation. Even small changes are meaningful: for a given reduction target, a product group level carbon tax is $1 - \frac{0.69}{0.91} \approx 25\%$ more effective than a department level tax. As discussed in section 1.4, this increased efficiency is driven by the small social cost of carbon when policy instruments are more flexible.

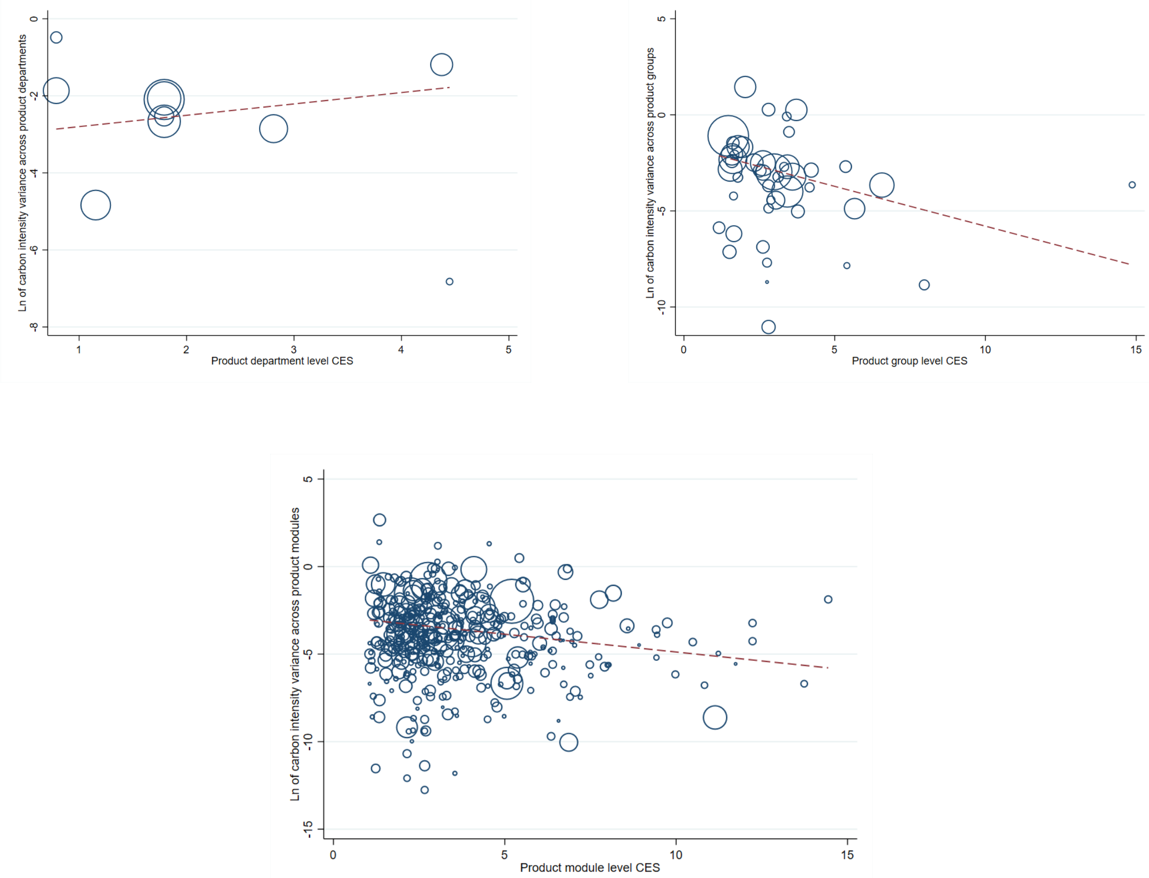
As highlighted by equations (1.11) and below, the relative efficiency of more flexible policy instruments depends on the expenditure-weighted covariance between substitution elasticities and variance of carbon intensity. In an ideal world, within a product category, we would like to have both high variance of carbon intensity as well as high substitutability across products, so that a product-specific tax within this category would be very efficient. Figure 1.10 shows that, at least in the case of food products, we have such positive relationship only at the product department level. At the product group and product module level, higher substitution elasticities are generally associated with slightly smaller variance in carbon intensity.

Under the set of assumptions laid out in this paper, and in particular perfect competition (that the products' prices reflect their marginal private cost), variance of carbon intensity across product category and elasticity of substitutions are structural parameters of the economy and are not something that a social planner can act on. Nonetheless, the quantities $\mathcal{E}^{dep}, \mathcal{E}^{group}, \mathcal{E}^{mod}$ derived in equation (1.11) are useful to inform the policy makers of the potential efficiency gains to have access to more flexible policy instruments and can be compared with the associated administrative cost of more flexible instruments. Further, equation (1.11) and figure 1.10 provide guidance as to the potential benefits of interaction between taxes and other other types of policy instruments. For instance, information-based policies aimed at reducing consumer's biases and hence increasing their elasticity of substitution should be best targeted towards markets with high variance in carbon intensity across products. Similarly, investments in carbon reducing technologies, leading to new, less carbon intensive products, should be targeted towards markets with high substitution

Figure 1.10: Relationship CES and carbon intensity

(a) Product department level

(b) Product group level



(c) Product module level

Notes: Panel (a), (b) and (c) plot the relationship between estimated CES at the product department, group and module level, respectively, and the associated within category variance in product intensity. The size of circle is proportional to the category's market share. The dashed line is the market-share weighted linear fit.

elasticity.

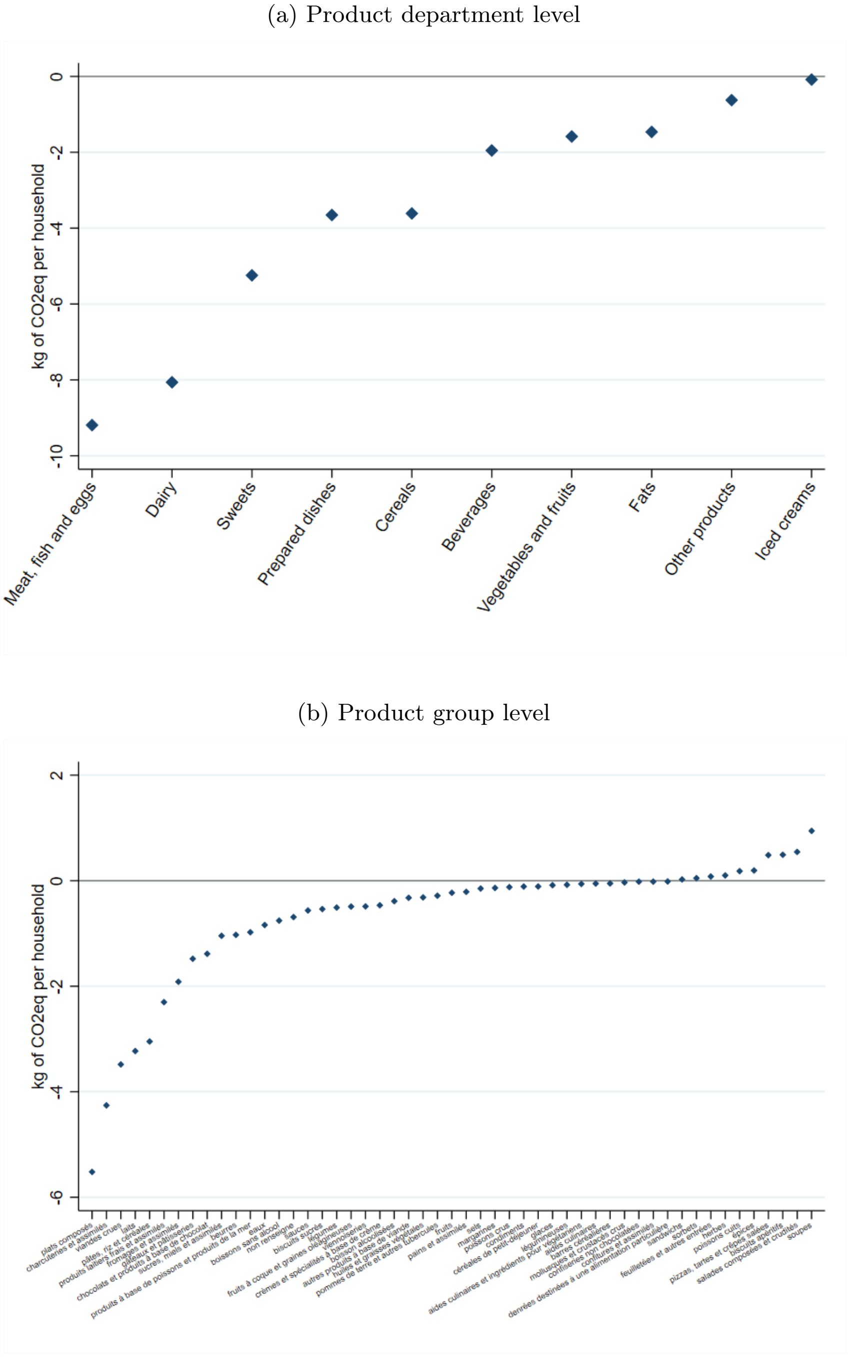
Overall, panel (c) of figure 1.7 and figure 1.10 show that the adaptation cost to a specific carbon emission constraint depends crucially on the availability of flexible policy instruments, and on the structural parameters of the economy.

1.6.4 The welfare cost of reducing carbon emissions

This section moves away from optimal taxation, and asks whether focusing on some specific product categories are more efficient in reducing carbon emissions than others. Efficiency is defined here as the change in social welfare (or private utility) per kilogram of carbon avoided, expressed in euros. More precisely, I consider the impact of a 5% price increase in a single product module, product group or product department, on carbon emissions, private utility and social welfare. This number has been chosen to be approximately the 75-th percentile of the distribution of the product-level Pigouvian taxes. However, since most results are expressed in units of welfare per kilogram of carbon avoided, the precise size of this price increase does not really matter. Unreported results show that the results are qualitatively the same with a 2.5% and a 10% price hike. This price increase could conceptually come from a variety of reasons: an environmental value-added tax, the inclusion of a specific product category in a carbon trading system, but also the money-metric equivalent of a non-price intervention (e.g. a public information campaign, increase in search costs). The results of this section can be interpreted as the general equilibrium impact of a price increase in a unique product category, as the exercise incorporates all the substitution patterns with other products, within and between product categories (under the maintained assumption of perfect competition, so that there is no supply-side response). Figure 1.12, which summarizes the results of this section, can be considered as a demand-side version of McKinsey's (2009) technology marginal abatement cost curve, where the cost comes from imperfect substitution opportunities.

First, panel (a) of figure 1.11 displays the amount of carbon emissions avoided from a 5% price increase in a single product department. The result is quite intuitive. A uniform price increase on all 'meat, fish and egg' products would reduce household carbon emissions by 9 kgCO2eq (for reference, the average annual carbon emission of a household in this data is 753 kgCO2eq), which is much more than an equivalent price increase on products from the 'vegetable and fruits' or the 'fats' product department (≈ -1.8 kgCO2eq per household). This difference is driven both by the carbon intensity of the different departments as well as their different market shares. Panel (b) reproduces the same exercise at the product group level. It shows that setting a 5% tax on some categories is counter-productive and raises carbon emissions, because close substitutes have higher carbon intensity. Naturally, such results would not be possible under optimal or constrained optimal taxation, as the

Figure 1.11: The impact of a 5% tax rate on specific product categories



(a) Product department level

(b) Product group level

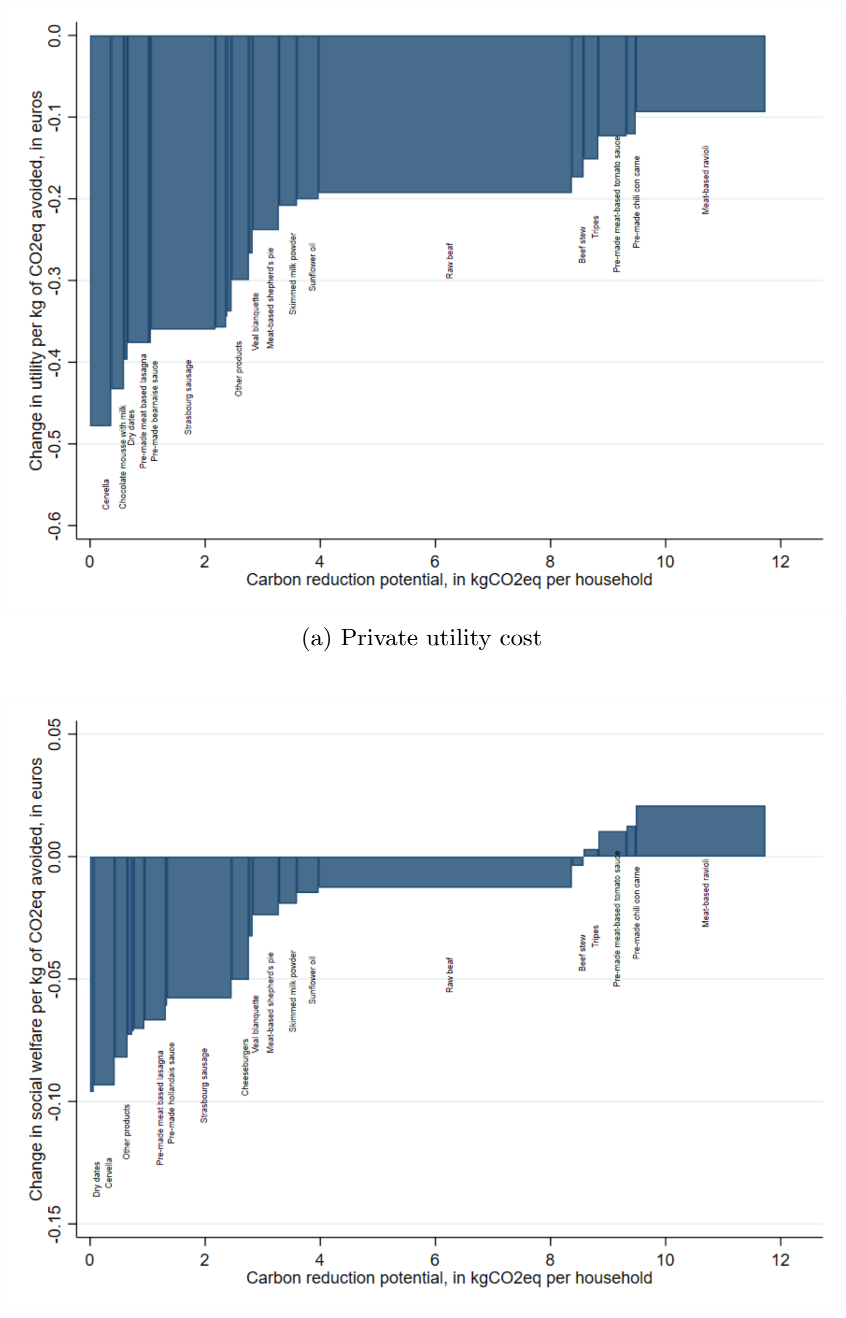
Notes: This figure plots the impact of a 5% price increase on carbon emissions on a unique product department (panel (a)) and product group (panel (b)).

optimality condition from equation (1.3) requires that high carbon intensity goods are always taxed at a higher rate than low carbon intensity goods.

Figure 1.12 plots the relative efficiency of a 5% price increase at the product module level. Product modules are ranked by how efficiently they respond to a price increase, measured as change of utility (panel (a)) or social welfare (panel (b)) in euros per kilogram of carbon emissions avoided, using the calibration discussed in section 1.4.5. The width of the bars is proportional to their carbon reduction potential, measured in total carbon emissions, and thus accounts for their relative importance. In addition, only the top 20 product modules according to this measure are shown. Panel (a) of figure 1.12 shows that all taxes considered incur some private utility loss, but some taxes more than others. Some product modules, such as 'cervella' or 'dry dates' respond relatively inefficiently: they represent a small fraction of overall carbon reduction potential, and are relatively more costly in terms of utility per kilogram of carbon avoided. In contrast, assuming no interaction between price hikes, a 5% price incrase for raw beef and meat-based raviolis would reduce carbon emissions by 7.5 kgCO2eq per household, or slightly less that 1% of their overall carbon emissions. Further, it would be achieved at a utility cost of 0.1 to 0.2€ per kilogram of carbon avoided, whereas a set of optimal tax rates set at the same product category level has a utility cost of 0.77 euro per kgCO2eq (see figure 1.C.7). Focusing on welfare cost of targeted price increase, panel (b) reveals that the qualitative results remain similar. Since the welfare measure also accounts for the social value of carbon emissions avoided and for the value of the price increase (implicitly assuming it is transfered to the government or to firms), the efficiency cost is even smaller than when measuring it private utility. Note that the ranking of product efficiency differs slightly between the two measures of efficiency (for instance 'cheeseburgers' are within the top 20 of product modules with highest efficiency when measured with social welfare, but drop out the top 20 when measured with private utility). This is due to the fact that the value of carbon emissions avoided and the value of price transfers depend on the elasticity of substitution between specific product modules, and on the relative carbon intensity of substitutes products, similar to the intuition developped in equation (1.9). All else equal, higher substitution elasticity reduces direct tax revenues and increases carbon emissions avoided (the net balance between the two depending on the social cost of carbon). At a social cost of carbon of 50€ per ton, a price increase of meat-based ravioli, pre-made chili con carne, pre-made meat-based tomato sauce and tripes becomes welfare improving.

Appendix figure 1.C.10 looks at broader product categories and plots the relative efficiency of a price increase at the product department level. Panel (a) shows that some vegetables and fruits are respond inefficiently. In contrast, prepared dishes, dairy, and meat, fish and eggs are relatively more efficient according to this measure, and account for a large part of the overall carbon reduction potential. Note however that as established above, a price

Figure 1.12: Cost of reducing carbon emissions by product module



(a) Private utility cost

(b) Social welfare cost

Notes: This figure ranks product modules by the efficiency measure developped in section 1.D.3, expressed as the change in utility in euro per kilogram of carbon emissions avoided (panel (a)) and the change in social welfare in euro per kilogram of carbon emissions avoided (panel (b)). Computations use the calibration of utility and social welfare discussed in section 1.4.5 and a social cost of carbon of 50€ per ton. Only the first 20 product modules according to this ranking are shown in this graph. Bar width is proportional to their carbon reduction potential, measured as the total carbon emissions, and thus accounts for their relative importance.

increase of 5% coming from e.g. a tax increase is not optimal at this broad level, since it does not leverage substitution potential across products within product departments. Looking at panel (b) and product groups, a 5% price hike on prepared dishes, chocolate-based products, butter, raw meats and milk would lead to a decrease of 14 kilograms of CO2eq, or 2% of a household's annual average carbon emissions in the data. This would be achieved at a utility cost of about 0.66 euros per kilogram of carbon, or 9€ per household, which is equivalent to 0.6% of their annual expenditure on food. Appendix figure 1.C.11 reproduce the same exercise for welfare.

Overall, these results suggest that the welfare impact of price increases is also very heterogenous across product categories, and that targeting a handful of product categories is a relevant way to minimize utility loss. A careful design of taxes leveraging this newly documented heterogeneity in carbon intensity across and within product categories, and accounting for substitution patterns across products paves the way for meaningful carbon emission reduction while improving overall welfare.

1.7 Conclusion

This paper documents large heterogeneity in carbon intensity across households and products, even within detailed product categories. Combining this stylized fact with the fact that substitution elasticities are higher between close subsitutes of a detailed product category than between different product modules, it provides a new justification for well-targeted environmental policies. In particular, it quantifies the welfare gains from granular environmental taxes or, equivalently, carbon markets over coarser tax systems.

These findings are based on a data set covering food products only, and one can wonder whether some results might change if the whole consumption basket is considered. While the main stylized facts that products' carbon intensity is heterogeneous even within detailed product categories is likely to hold for most parts of the consumption baskets (think for instance about the variation in lifetime carbon emissions across electric versus diesel sport utility vehicules), I expect the finding of a modest carbon intensity - expenditure gradient across households to be more specific to food and beverages (think for instance of different leisure choices across households).

This paper also illustrates that at least in some settings, exposure to carbon taxes can vary widely across households with similar expenditure levels, suggesting that income might not be the ideal tagging device and that schemes aimed to compensate households for the income loss from environmental policies should be more sophisticated. In particular, one can see the 2018 French Yellow Vest protest movement as a reaction to poorly targeted and insufficiently compensated carbon taxes

Appendices

1.A Additional stylized facts

Additional stylized fact 1: Carbon intensity and carbon emissions of food choices are very heterogenous, even within narrow product categories. I show this using various metrics of dispersion. Overall, this suggests that it is possible for consumers, given the current supply-side constraints (technology, market structure, etc.), to choose products with lower carbon intensity.

As a general overview, figure 1.C.12 plots the histogram of (demeaned) carbon intensity in the raw data and controlling for the different product categories. While the distribution of carbon intensity tends to shrink as we control for finer categories, sizable variation remains. In particular, standard deviation decreases from 0.58 in the raw data to 0.47 (18% decrease) and 0.36 (36% decrease) when controlling for product sub-groups and product modules respectively, which is not much considering the number of different product categories at these finer levels. Furthermore, figure 1.C.13 plots the coefficient of variation at the product module level for carbon intensity and carbon emission per product. As can be seen from these graphs, nearly 90% of product modules have a coefficient of variation of 100% or higher, which indicates particularly high variation in carbon intensity relative to the mean carbon intensity in each category. This cannot be solely explained by price differences since the pattern is similar for carbon emissions per product. This figure also highlights the relatively high variation in within-module heterogeneity across product categories. Figures 1.C.14 and 1.C.15 show that the pattern is similar at all category level, both for carbon intensity and carbon emission per product. Interestingly, the median coefficient of variation of carbon intensity across product category increases as product categories become finer, from a median of 95 % for product department to a median of 169 % and 342 % for product sub-groups and modules respectively. This is likely driven by the fact that standard deviation only decreases moderately while average carbon intensity changes more. All these results weight every product equally in order to be representative of the choice set rather than the actual consumer choices but the results are qualitatively unchanged when weighting by quantity sold.

Additional stylized fact 2: Targeting a small number of detailed product categories has the potential to achieve significant reduction in carbon emissions I go beyond panel (c) of figure 1.C.2 and ask whether it would make sense for policy makers to focus on a few product category to achieve significant reduction in carbon emissions using, as an illustration, the Ecoscore label. As discussed in section 2.2, Ecoscore is an environmental label developed by ADEME and was released publicly in late 2020, so that consumers in my data set did not observe this label while making their purchases. Nonetheless, I use it as a useful broad categorization of a product's environmental quality: figure 1.C.16 and table 1.1 show that products with better Ecoscore grade also have lower carbon intensity. I ask two questions. First, by how much would carbon emissions be reduced if, within product sub-groups, demand shifted to products with Ecoscore grade A and B only? Second, how much of the change can be accounted for by the largest 10 product categories? Exploiting this environmental label shows that if consumers shifted their demand from products with Ecoscore C and above to products within the same product sub-group, but with Ecoscore A and B, overall carbon emissions would be reduced by 28.6 %, (1019 of tCO2eq in my data). Further more, the top 10 product sub-groups with highest impact account for 64.1 % of the overall change. Figure 1.C.17 highlights these results and suggests, for instance, that a policy focusing solely on 'pasta based prepared dishes' and 'sausage and related products', and which would incentivize consumers to shift to products with Ecoscore A or B within the same category, would reduce carbon emissions from food by close to 6% (20% of 28.6 %).

Repeating the same exercise at the broader product department level, I find that overall carbon emissions would instead be reduced by 39.4 %, and that the largest product department, meat, fish and eggs, accounts for 46 % of the overall change (cf. figure 1.C.18). The higher numbers do make sense as it would involve shift between less similar products.

Additional stylized fact 3: There is a very modest negative relationship between carbon intensity and expenditure across households. Panel (a) of figure 1.C.19 shows that households with higher average expenditure have a lower average carbon intensity. We would ideally also like to analyze the relationship between carbon intensity and household income, but this data is not available. Instead, I focus for this stylized fact on non-family households (results are qualitatively unchanged when considering families as well) and rely on the assumption that household size is homogenous and the fact that food is a normal good to conclude that the relationship between carbon intensity and household income is likely to be negative as well. As an imperfect robustness check, I proxy household income with median available income of the employment zone the household shops in (as in, among others, Chetty et al. (2020)). Figure 1.C.20 confirms that the pattern remains similar.

The overall gradient in panel (a) of figure 1.C.19 is small: a 10% increase in monthly expenditure is associated with a decrease of 0.003 points of carbon intensity. However, the slope is still statistically significant and economically meaningful. To see this, I ask by how much would carbon emissions decrease if households in the bottom half of the income distribution had an average carbon intensity equal to the grand average of carbon intensity, keeping all other behaviour constant. Overall, starting from a grand average carbon intensity of 0.63 for non-family households, such change in carbon intensity would lead to a decrease in carbon emissions of 38.1 %. The high heterogeneity in carbon intensity within percentile of the expenditure distribution is likely to explain this results.

Additional stylized fact 4: While CO2-expenditure elasticity estimate is in line with previous literature, this hides important heterogeneity across products. It is important to note that additional stylized fact 3 does not at all imply that households with higher food expenditures (hence, likely wealthier households) emit less carbon. On the contrary, panel (b) of figure 1.C.19 shows that the CO2 - expenditure elasticity is 0.95. This number is in line with previous literature on the topic; for instance Chancel et al. (2015) discuss past work and a carbon - expenditure elasticity of 1 is their preferred estimate. In addition, this data can go one step deeper and look at the CO2 - expenditure elasticity within finer product categories. As figure 1.C.21 highlights, carbon-expenditure elasticity is indeed centered around 1, but almost half of the product categories (43%; 49 out of 115 product sub-groups and 27 out of 62 product groups) have elasticities significantly greater than 1. Among these products, meat products and alcohols seem overrepresented, although the pattern is not very clear.

1.B Additional tables

Table 1.B.1: Products, modules, subgroups and groups by product departments

	# groups	# sub groups	# modules	# products
Prepared dishes	6	13	137	1,709
Vegetables and fruits	5	12	172	2,613
Cereals	4	7	86	2,272
Meat, fish and eggs	11	25	176	3,396
Dairy	4	14	118	3,321
Beverages	3	12	74	4,775
Sweets	9	9	143	6,056
Iced creams	3	3	9	587
Fats	4	4	17	585
Other products	9	12	117	2,129
Infant food	4	4	7	110
Total	62	115	1,056	27,553

Notes: Table built from the Product data set.

Table 1.B.2: Summary statistic on matching

	Product data	Household data
Unit of observation	Product	Product × client
# products	120,574	43,873
# products matched	27,553	19,068
Share product matched (%)	22.9	43.5
Sales (mE)	9,004	1,090
Sales matched (mE)	5,534	679
Share sales matched (%)	61.5	62.3
Carbon emissions (mtCO2eq)	5.528	0.158
N	27,553	92,099,172

Table 1.B.4: Summary statistics on selected households

	Monthly expenditure				# months in data set			
	Mean	25th	Med.	75th	Mean	25th	Med.	75th
Overall	125	81	109	153	33	33	36	36
By expenditure:								
Q1	68	62	68	75	34	33	35	36
Q2	94	87	94	101	33	31	35	36
Q3	129	118	127	139	33	32	36	36
Q4	210	170	194	233	33	33	36	36
By carbon intensity								
Q1	130	83	113	159	33	32	36	36
Q2	126	82	110	154	33	32	36	36
Q3	125	81	109	153	34	33	36	36
Q4	119	77	103	145	34	33	36	36
Non-family households	116	77	102	139	34	33	36	36

Notes: This table displays statistics on the average monthly expenditure and the number months present in the dataset for households from the Household data set. The different rows of the table show different split of the households. For instance, the median monthly expenditure of households in the first quartile of the distribution of households ranked by carbon intensity is 113€. Household can be observed for a maximum of 36 months.

Table 1.B.3: Summary statistics for product department: Fat

	# products	CO2 intensity	Std. dev. CO2 intensity	Mean CO2 / prod.	Std. dev. CO2 / prod.
Butter, salted	6	0.52	0.40	3.87	0.00
Butter, semi-salted	111	0.52	0.35	4.53	1.66
Butter, unsalted	132	0.49	0.37	4.40	1.54
Colza oil	16	1.02	0.33	1.61	0.43
Duck fat	13	7.55	0.03	0.46	0.02
Goose fat	8	10.31	0.02	0.46	0.02
Grape seed oil	13	2.28	0.11	1.59	0.53
Groundnut oil	2	0.72	0.21	4.71	0.00
Hazelnut oil	31	2.95	0.15	1.86	1.00
Lard	4	3.31	0.09	0.44	0.19
Linseed oil	3	5.16	0.02	0.73	0.00
Margarine-like fat, unsalted	85	3.53	0.16	0.62	0.24
Mixed milk and vegetable fat	1	1.28	0.00	2.67	0.00
Mixed oil oil	19	0.90	0.32	4.06	3.34
Sesame oil	8	6.98	0.05	0.60	0.38
Sunflower oil	19	0.38	0.95	4.54	3.48
Virgin olive oil	114	4.01	0.07	1.34	0.57

Notes: This table displays a number of summary statistics for all 17 product modules in the product department “Fats”. Standard deviation of carbon emission per kg of product is zero for all modules. CO2 per product is in kgCO2eq per product. Table built from the Product data set. For the Analysis data set, product modules “Groundnut oil”, “Linseed oil” and “Mixed milk and vegetable fat” would be discarded as they include less than four products.

Table 1.B.5: Carbon intensity and market share at different product category level

	(1) CO2 intensity	(2) CO2 intensity	(3) CO2 intensity	(4) CO2 intensity
ln (dep market share)	0.032^{***}			
	(0.005)			
ln(group market share)		0.030^{***}		
		(0.005)		
ln(sub-group market share)			0.029^{***}	
			(0.004)	
ln(module market share)				0.025^{***}
				(0.004)
Constant	0.734^{***}	0.655^{***}	0.633^{***}	0.551^{***}
	(0.059)	(0.050)	(0.039)	(0.024)
Product department FE	Yes			
Product group FE		Yes		
Product sub-group FE			Yes	
Product module FE				Yes
R^2	0.123	0.238	0.367	0.648
N	27,553	27,553	27,553	27,553

Notes: This table presents regression results linking carbon intensity and product market share at different level of the product classification from the Product dataset. Carbon intensity is the ratio of CO2 to price, defined at the product (barcode) level. Independent variables are the market share of a given product computed at the product department, group, subgroup and module level respectively. Standard errors in parentheses are clustered at the product module level.

Table 1.B.6: Carbon emissions per product and market share at different product category level

	(1) CO2 / product	(2) CO2 / product	(3) CO2 / product	(4) CO2 / product
ln (dep market share)	0.047^{***}			
	(0.013)			
ln(group market share)		0.041^{***}		
		(0.011)		
ln(sub-group market share)			0.037^{***}	
			(0.007)	
ln(module market share)				0.031^{***}
				(0.006)
Constant	1.616^{***}	1.478^{***}	1.424^{***}	1.320^{***}
	(0.161)	(0.118)	(0.066)	(0.035)
Product department FE	Yes			
Product group FE		Yes		
Product sub-group FE			Yes	
Product module FE				Yes
R^2	0.132	0.227	0.405	0.635
N	27,553	27,553	27,553	27,553

Notes: This table presents regression results linking carbon emissions per product and product market share at different level of the product classification from the Product dataset. Carbon emission per product is in eqCO2kg. Independent variables are the market share of a given product computed at the product department, group, subgroup and module level respectively. Standard errors in parentheses are clustered at the product module level.

Table 1.B.7: Carbon intensity and market share - Robustness

	(1)	(2)	(3)	(4)
	CO2 intensity	CO2 intensity	CO2 intensity	CO2 intensity
ln(sub-group market share)	0.029***	0.024***		
	(0.004)	(0.004)		
price		-0.034***		-0.028***
		(0.006)		(0.007)
ln(module market share)			0.025***	0.023***
			(0.004)	(0.004)
Constant	0.633***	0.719***	0.551***	0.638***
	(0.039)	(0.044)	(0.024)	(0.039)
Product sub-group FE	Yes	Yes		
Product module FE			Yes	Yes
R^2	0.367	0.387	0.648	0.657
N	27,553	27,553	27,553	27,553

(a) Carbon intensity

	(1)	(2)	(3)	(4)
	CO2 / product	CO2 / product	CO2 / product	CO2 / product
ln(sub-group market share)	0.037***	0.053***		
	(0.007)	(0.008)		
price		0.117***		0.164***
		(0.029)		(0.045)
ln(module market share)			0.031***	0.045***
			(0.006)	(0.007)
Constant	1.424***	1.126***	1.320***	0.806***
	(0.066)	(0.098)	(0.035)	(0.143)
Product sub-group FE	Yes	Yes		
Product module FE			Yes	Yes
R^2	0.405	0.431	0.635	0.667
N	27,553	27,553	27,553	27,553

(b) Carbon emissions per product

Notes: This table presents regression results linking carbon emissions per product and product market share at the product sub-group and product module level from the Product dataset. In both panels, columns (1) and (3) are the same as in figure 1.B.5 and 1.B.6. Independent variables are the market share of a given product computed at the product subgroup and module level respectively. Standard errors in parentheses are clustered at the product module level.

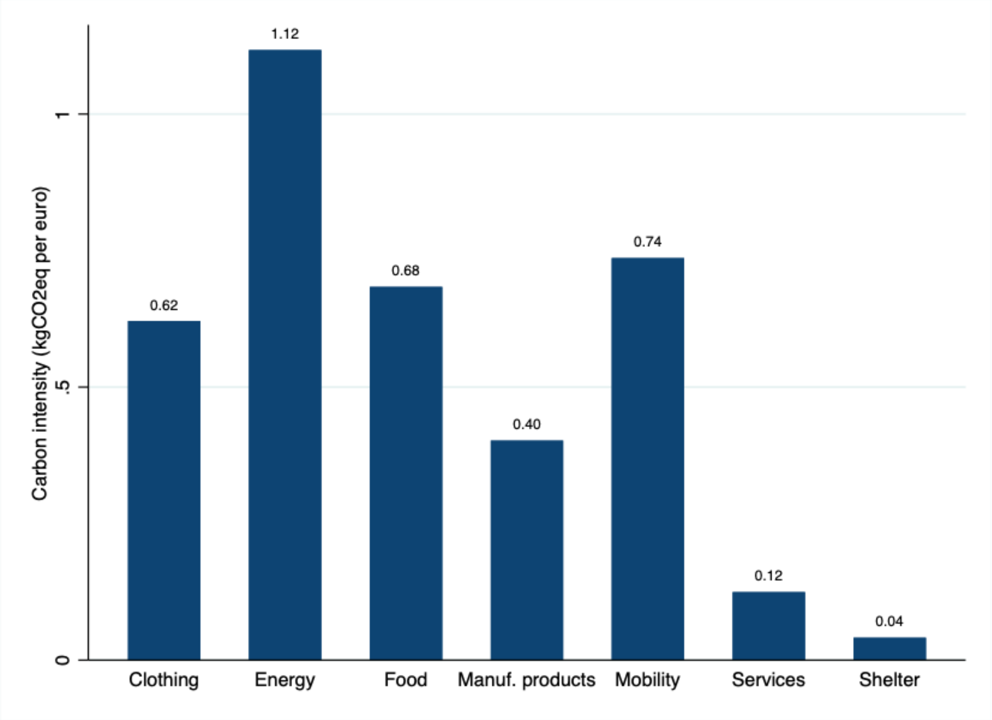
Table 1.B.8: Product module elasticities across different samples

	All households	By expenditure		By carbon intensity		Non families
		Bottom 25%	Top 25%	Bottom 25%	Top 25%	
5-th percentile	1.34	1.48	1.50	1.22	1.42	1.32
25-th percentile	2.21	2.17	1.98	1.57	2.42	2.06
Median	3.04	2.72	2.56	2.08	3.50	2.58
75-th percentile	4.50	3.84	3.79	2.83	4.85	4.05
95-th percentile	7.76	6.05	5.95	5.10	7.44	7.61
N	435	435	435	435	435	435
Fraction imputed (%)	29.4	36.6	38.8	51.9	30.4	32.5

Notes: Product modules are restricted to those common to all different samples.

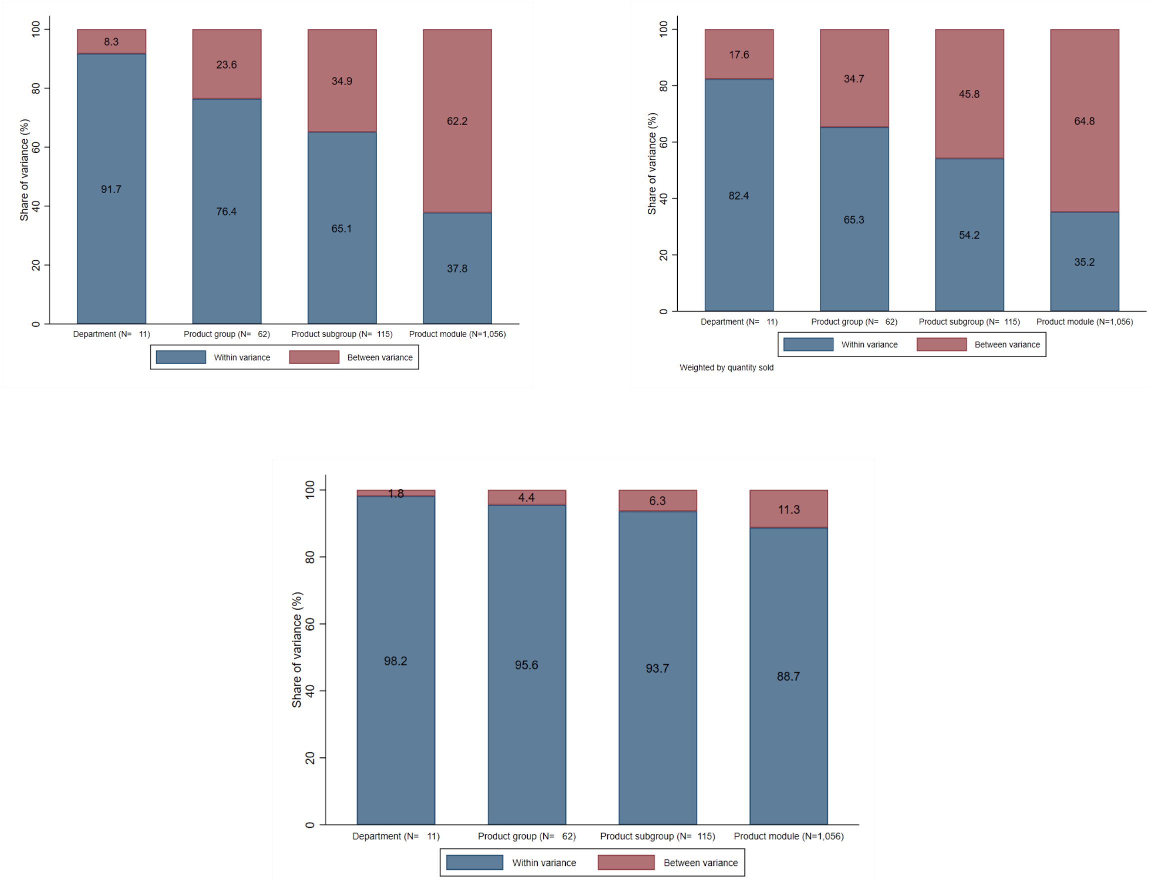
1.C Additional figures

Figure 1.C.1: Carbon intensity for various sectors, Exiobase



Notes: This figure displays the carbon intensity of various sectors from the Exiobase data set. Carbon intensity is expressed in kgCO2eq per real euro spent. Euro is deflated to 2017 level and carbon emissions includes direct and indirect emissions following the inversion of the input-output matrix.

Figure 1.C.2: Coefficient of variation of carbon intensity



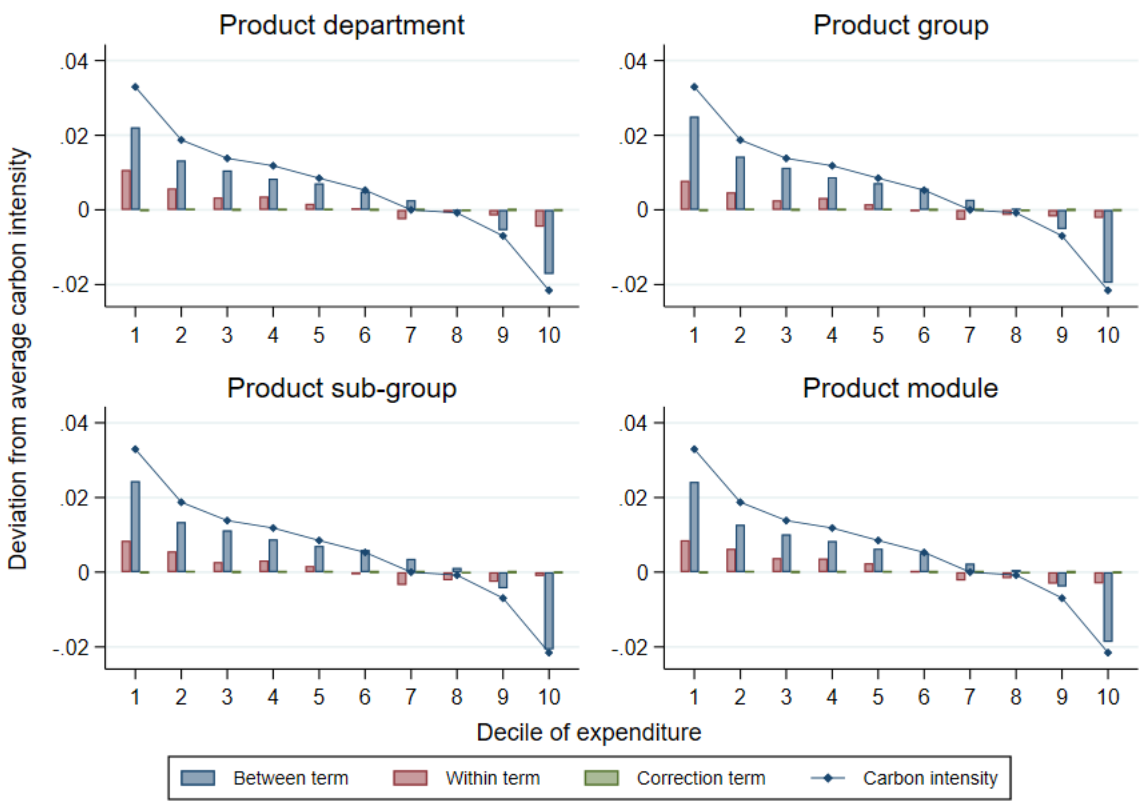
(a) Carbon intensity, unweighted

(b) Carbon emission per product

(c) Total carbon emission

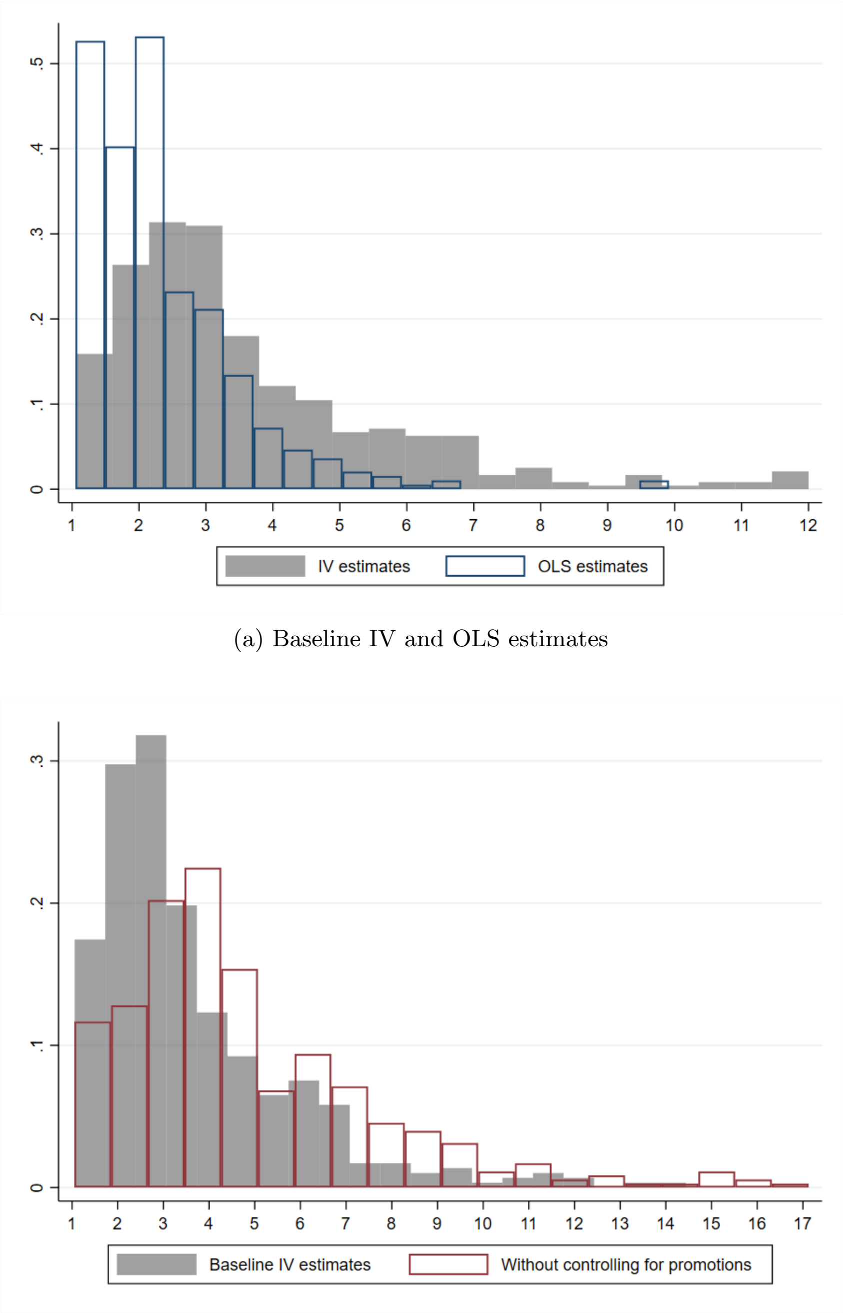
Notes: This figure presents decomposition of the variance of carbon intensity into between group variation and within group variation, for different product category. Data comes from the Product data.

Figure 1.C.3: Decomposition of difference in carbon intensity



Notes: This figure presents decomposition of the difference in carbon intensity relative the grand average following (1.1). Expenditure is defined as the average monthly expenditure of a given household.

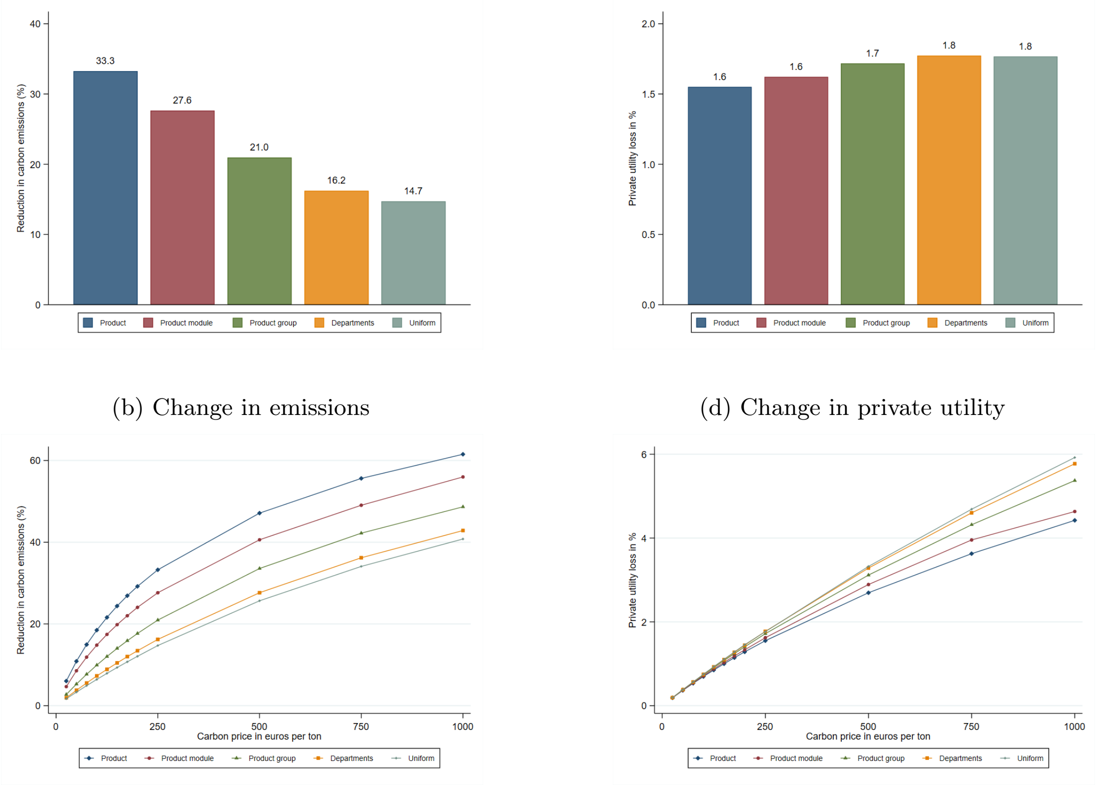
Figure 1.C.4: Histogram of various elasticities estimates



(a) Baseline IV and OLS estimates

(b) Baseline IV and IV without promotions

Figure 1.C.5: Carbon emission and utility change by carbon price



(a) Change in emissions, 150€ per tCO2eq

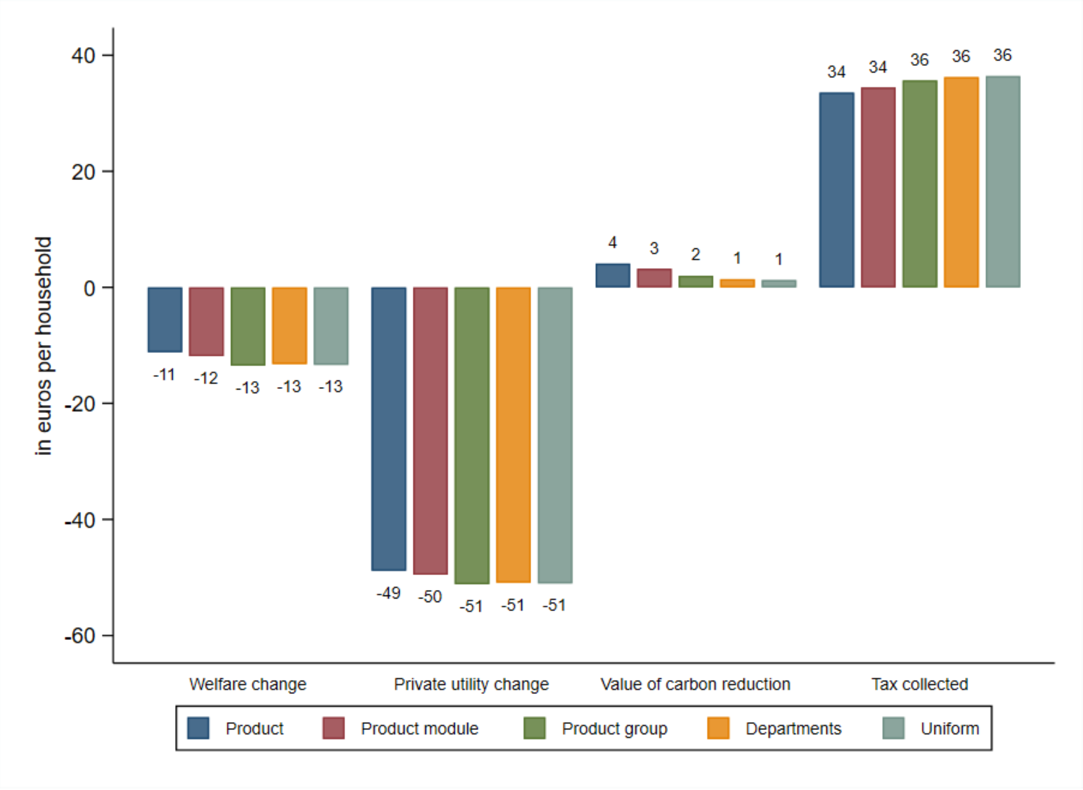
(c) Change in private utility, 150€ per tCO2eq

(b) Change in emissions

(d) Change in private utility

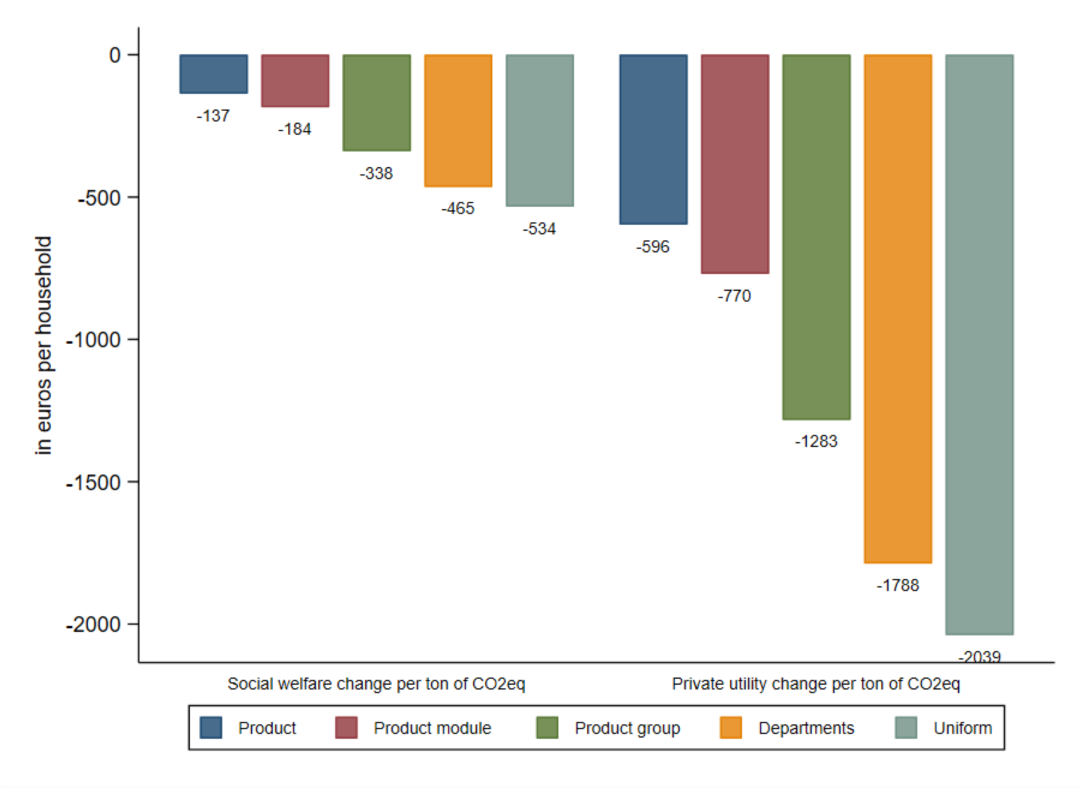
Notes: This figure plots the estimated percentage reduction in carbon emissions (panel (a)) and private utility (panel (c)) following optimally set taxes at different levels of the product category classification as discussed in section 1.4 and a price of carbon of 150€ per ton. Panels (b) and (d) plot the change in carbon emissions and private utility across a range of values of the price of carbon.

Figure 1.C.6: Components of social welfare



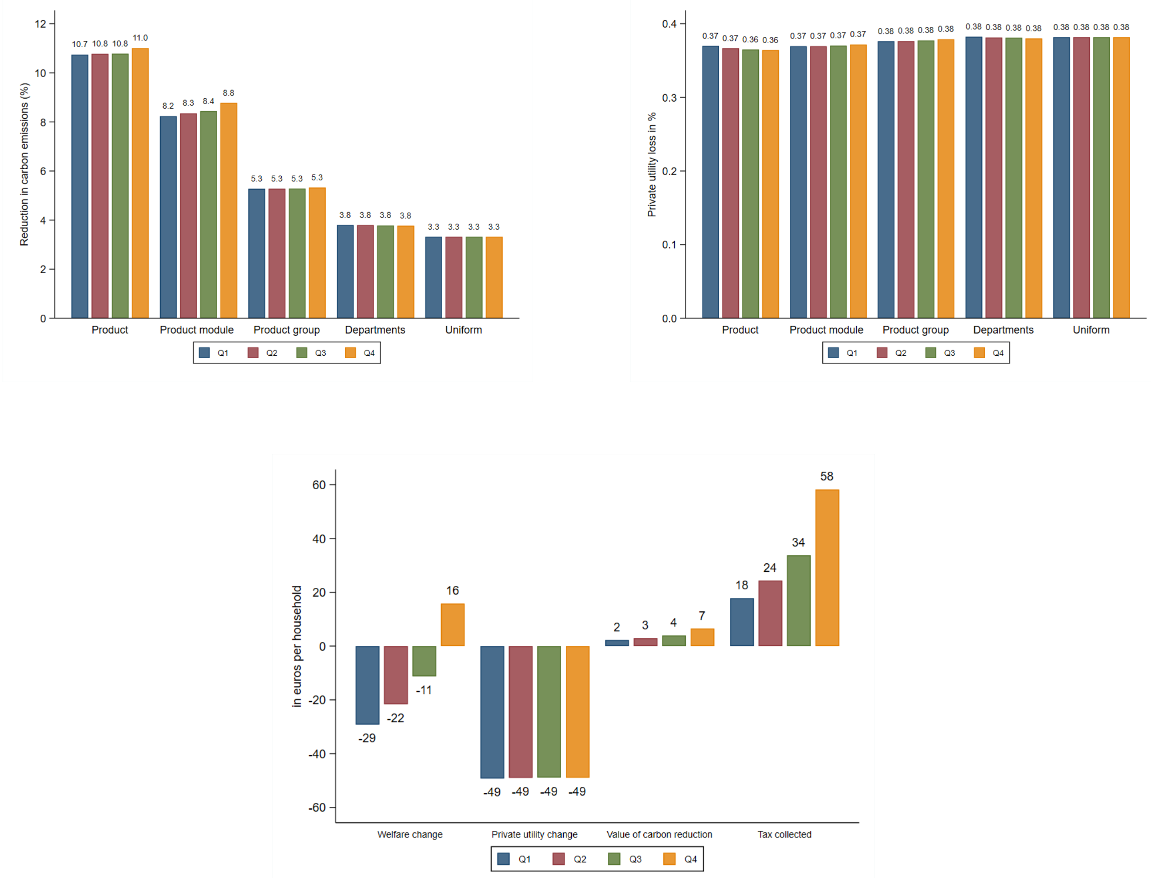
Notes: This figure uses the calibration assumptions discussed in section1.4.5. Numbers are based on a product level taxes carbon price of 50€ per ton.

Figure 1.C.7: Efficiency measures of tax policies



Notes: Numbers are based on a product level taxes carbon price of 50€ per ton.

Figure 1.C.8: Heterogeneity by expenditure - Additional results



(a) Reduction in carbon emissions, common CES

(b) Reduction in private utility, common CES

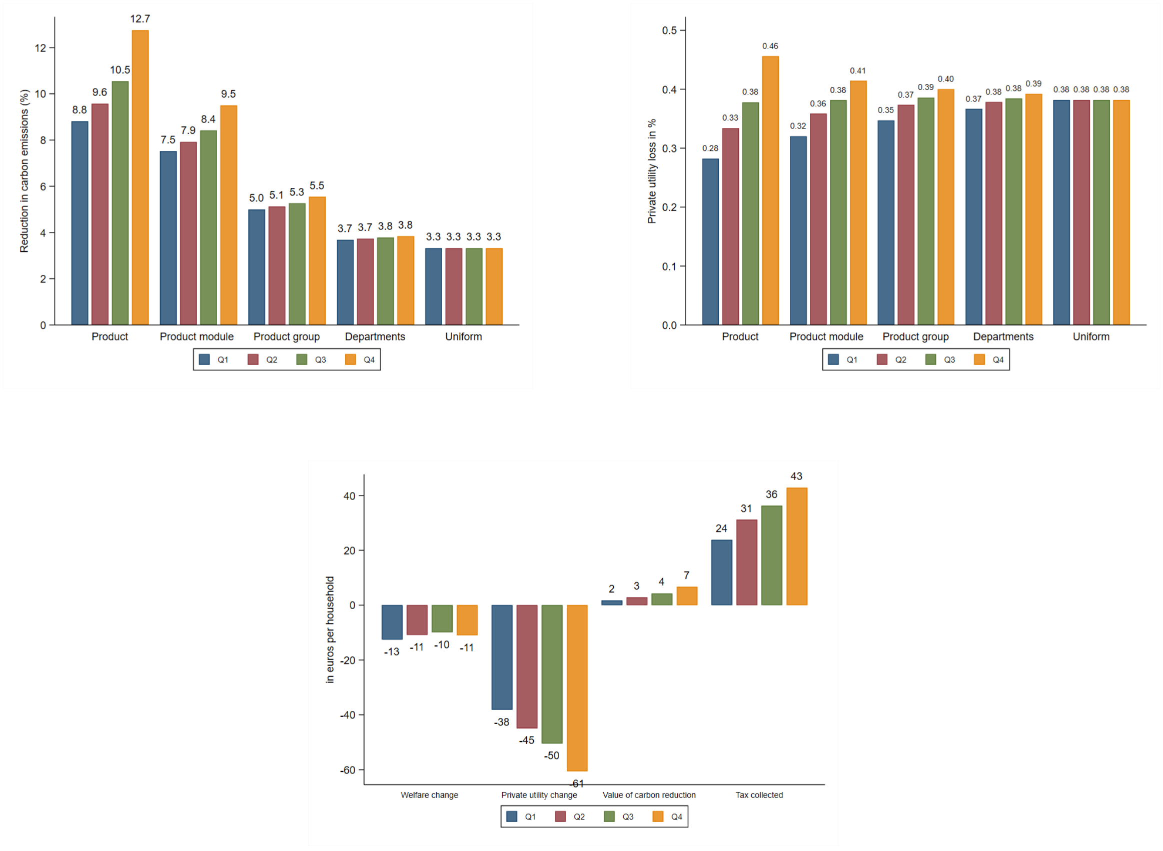
(c) Components of social welfare

Notes: Panel (a) and (b) displays estimates of carbon emission reduction and private utility loss under the assumption that all households share the same preference structure and the same elasticities of substitution. Panel (c) plots the components of social welfare under the assumption of different preference across households and under an optimal product-level tax with a social cost of carbon of 50 euros per ton. Numbers are based on a carbon price of 50€ per ton.

Figure 1.C.9: Heterogeneity by carbon intensity - Additional results

(a) Reduction in carbon emissions, common CES

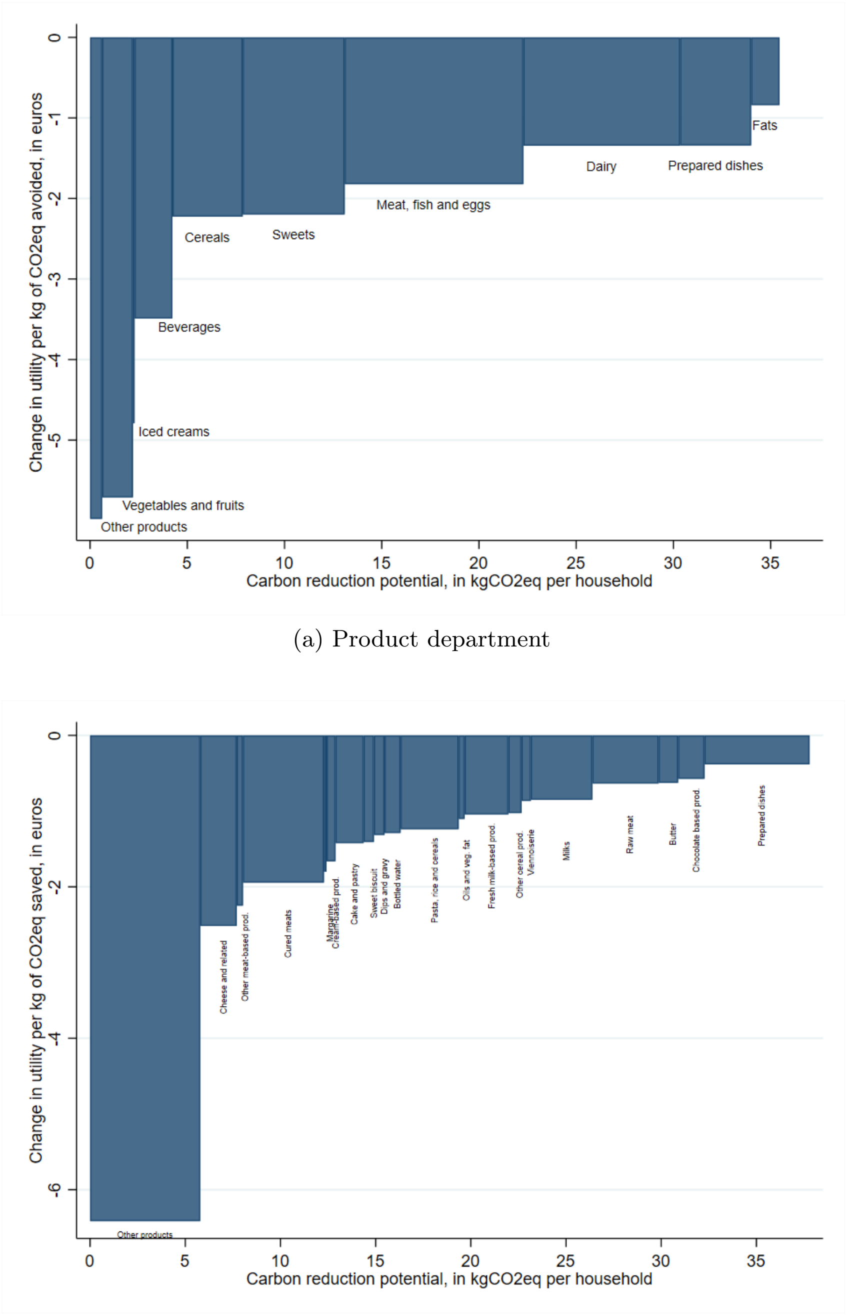
(b) Reduction in private utility, common CES



(c) Components of social welfare

Notes: Panel (a) and (b) displays estimates of carbon emission reduction and private utility loss under the assumption that all households share the same preference structure and the same elasticities of substitution. Panel (c) plots the components of social welfare under the assumption of different preference across households and under an optimal product-level tax with a social cost of carbon of 50 euros per ton. Numbers are based on a carbon price of 50€ per ton.

Figure 1.C.10: Utility cost of reducing carbon emissions at different product category levels

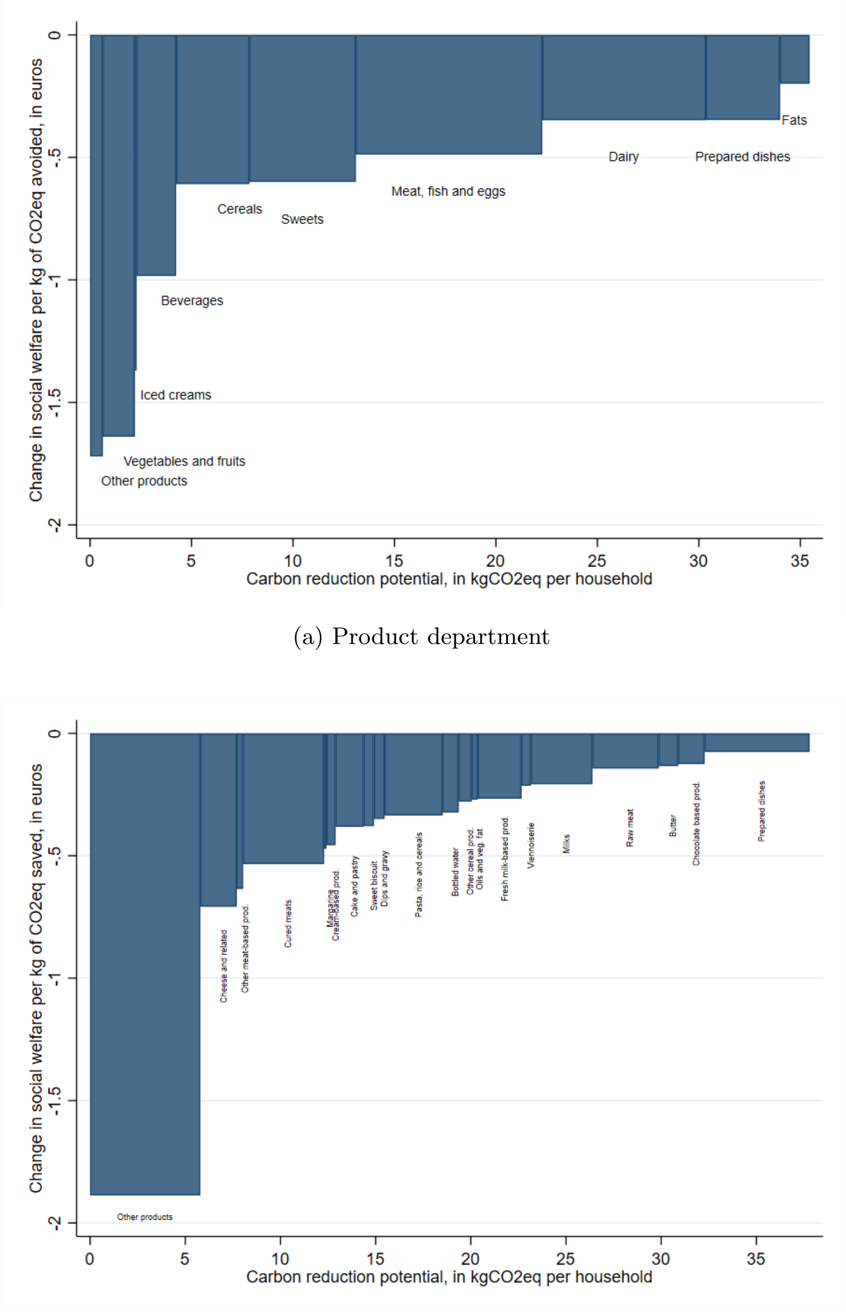


(a) Product department

(b) Product group

Notes: This figure plots the relative efficiency of a 5% price increase at the product department (panel (a)) and product group (panel (b)) level. Product categories are ranked by the efficiency of this tax, expressed as the change in utility in euro per kilogram of carbon emissions avoided. Computations use the calibration of utility and social welfare discussed in section 1.4.5 and a social cost of carbon of 50€ per ton. For panel (b), only the first 19 product groups, according to this ranking, are shown individually. The other product groups are grouped into a single category, for which the efficiency measure is defined as the unweighted mean change in social welfare per kg of CO2 avoided as the efficiency measure. The width of the bars is proportional to their carbon reduction potential, measured as the total carbon emissions, and thus accounts for their relative importance.

Figure 1.C.11: Welfare cost of reducing carbon emissions at different product category levels

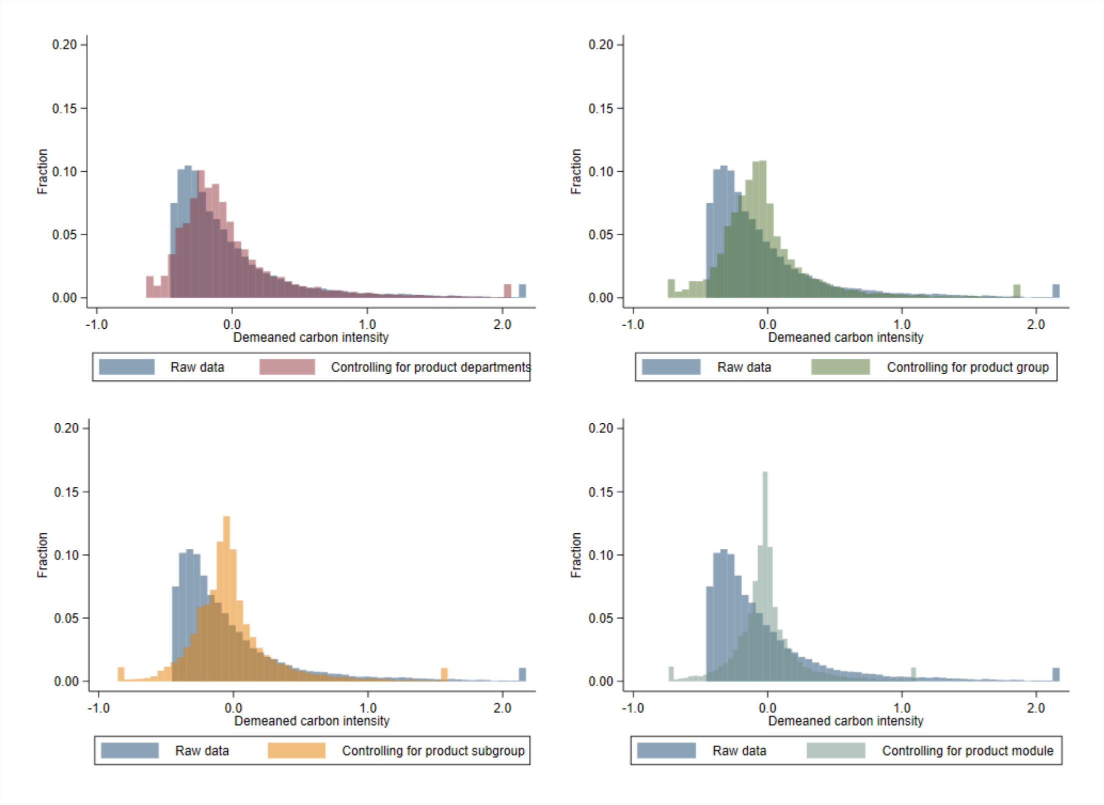


(a) Product department

(b) Product group

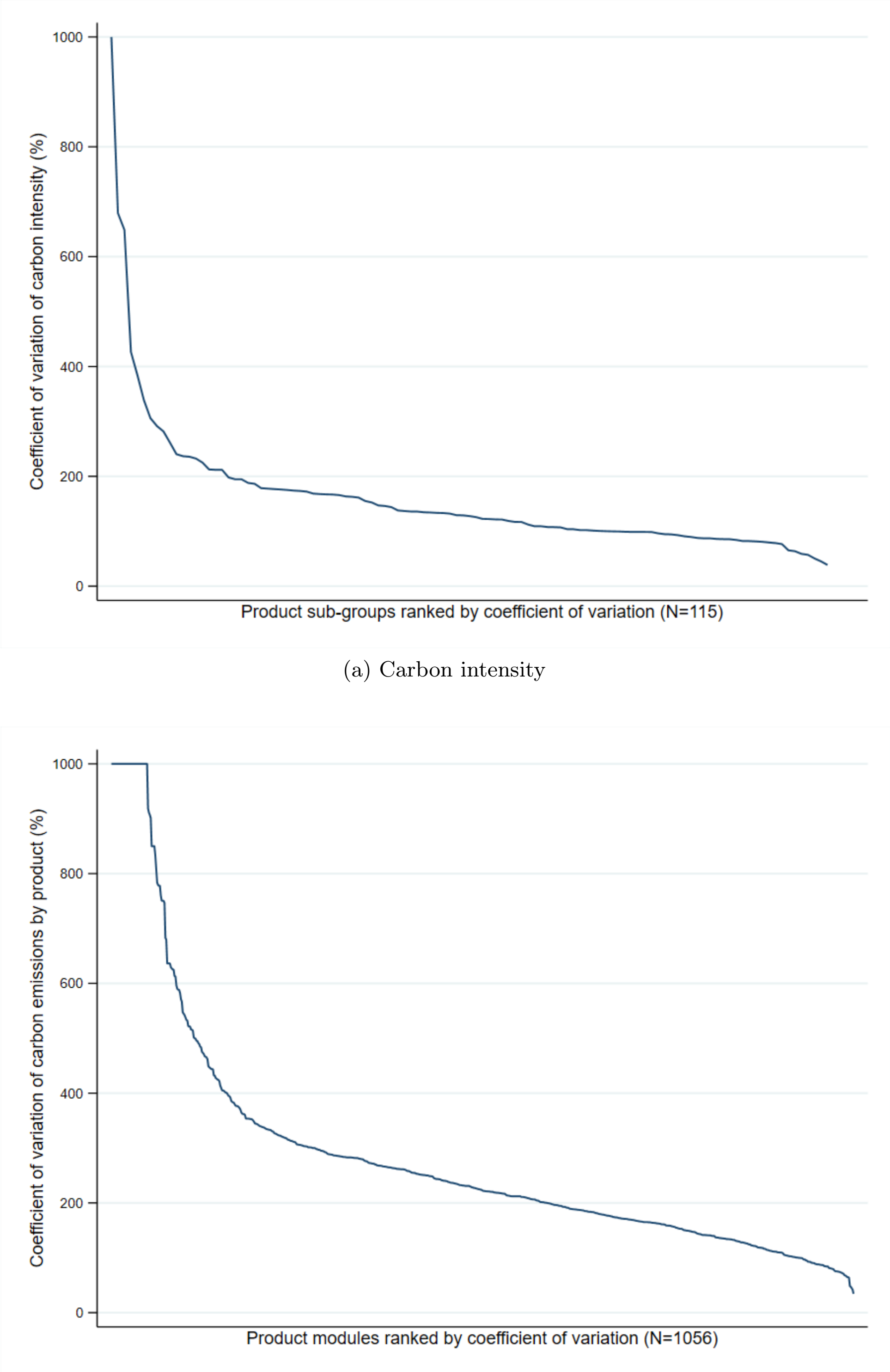
Notes: This figure plots the relative efficiency of a 5% price increase at the product department (panel (a)) and product group (panel (b)) level. Product categories are ranked by this efficiency measure, expressed as the change in social welfare in euro per kilogram of carbon emissions avoided. Computations use the calibration of utility and social welfare discussed in section 1.4.5 and a social cost of carbon of 50€ per ton. For panel (b), only the first 19 product groups, according to this ranking, are shown individually. The other product groups are grouped into a single category, for which the efficiency measure is defined as the unweighted mean change in social welfare per kg of CO2 avoided as the efficiency measure. The width of the bars is proportional to their carbon reduction potential, measured as the total carbon emissions, and thus accounts for their relative importance.

Figure 1.C.12: Histogram of carbon intensity



Notes: This figure presents the histogram of carbon intensity in the raw data and controlling for product departments, groups, sub groups and module respectively. Data is winsorized at the 1th and 99th percentile.

Figure 1.C.13: Coefficient of variation by product modules

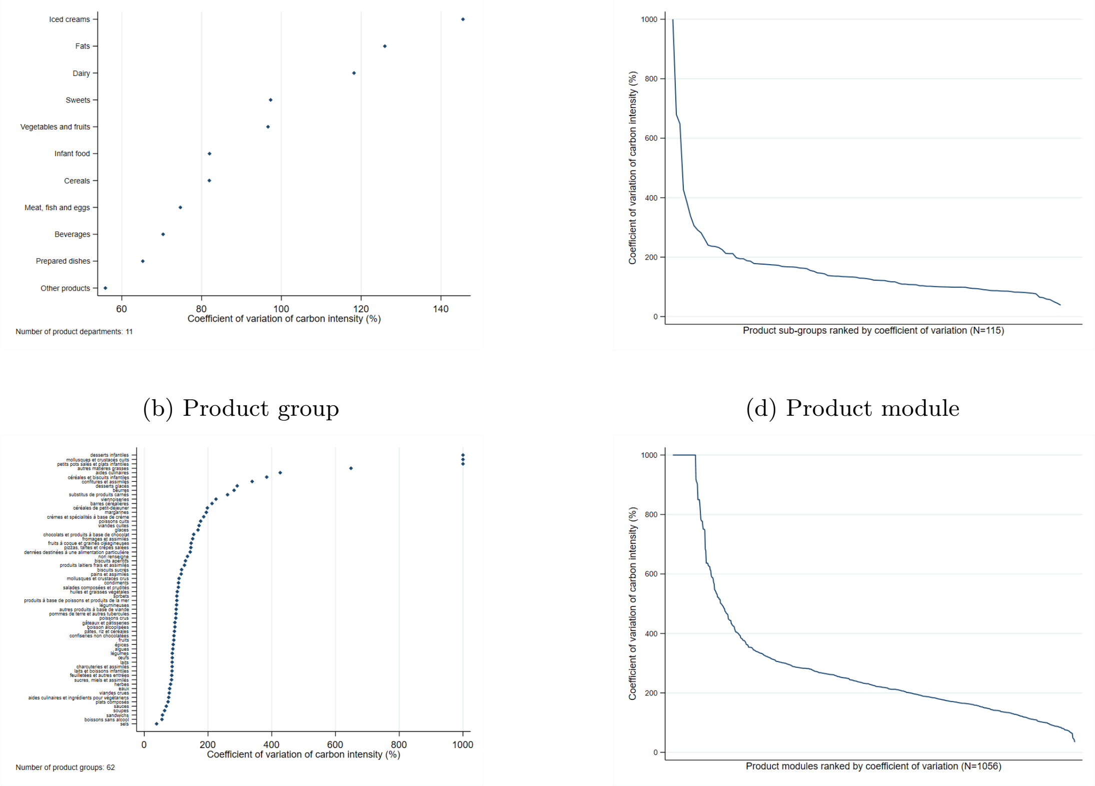


(a) Carbon intensity

(b) Carbon emission per product

Notes: This figure presents the coefficient of variation of carbon intensity (panel (a)) and carbon emission by product (panel (b)), by product modules. Data comes from the Product data. The coefficient of variation for a variable is the ratio of its standard error to its the mean and is expressed in this figure in percent. To ease exposition, coefficient of variation is winsorized at 1000%.

Figure 1.C.14: Coefficient of variation of carbon intensity


(a) Product department

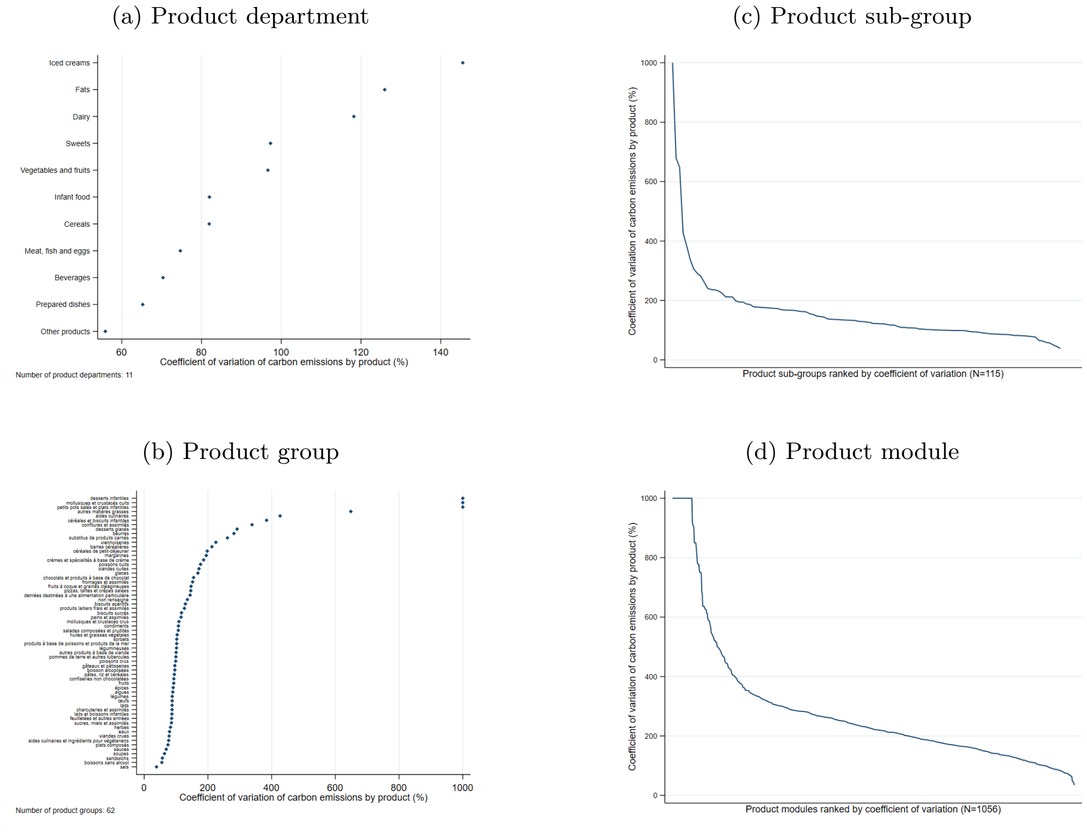
(c) Product sub-group

(b) Product group

(d) Product module

Notes: This figure presents the coefficient of variation of carbon intensity for different product levels. Data comes from the Product data. The coefficient of variation for a variable is the ratio of its standard error to its the mean and is expressed in this figure in percent. To ease exposition, coefficient of variation is winsorized at 1000%. Note that for product department and product groups, the axis are reversed.

Figure 1.C.15: Coefficient of variation of carbon intensity



(a) Product department

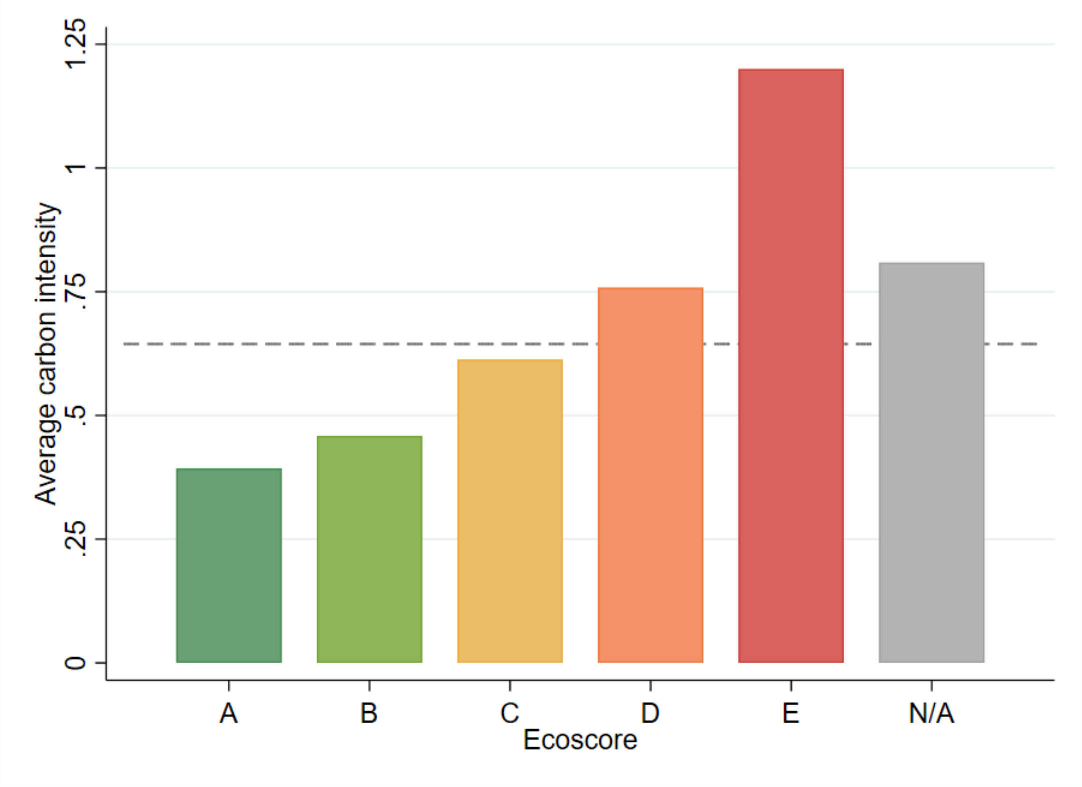
(c) Product sub-group

(b) Product group

(d) Product module

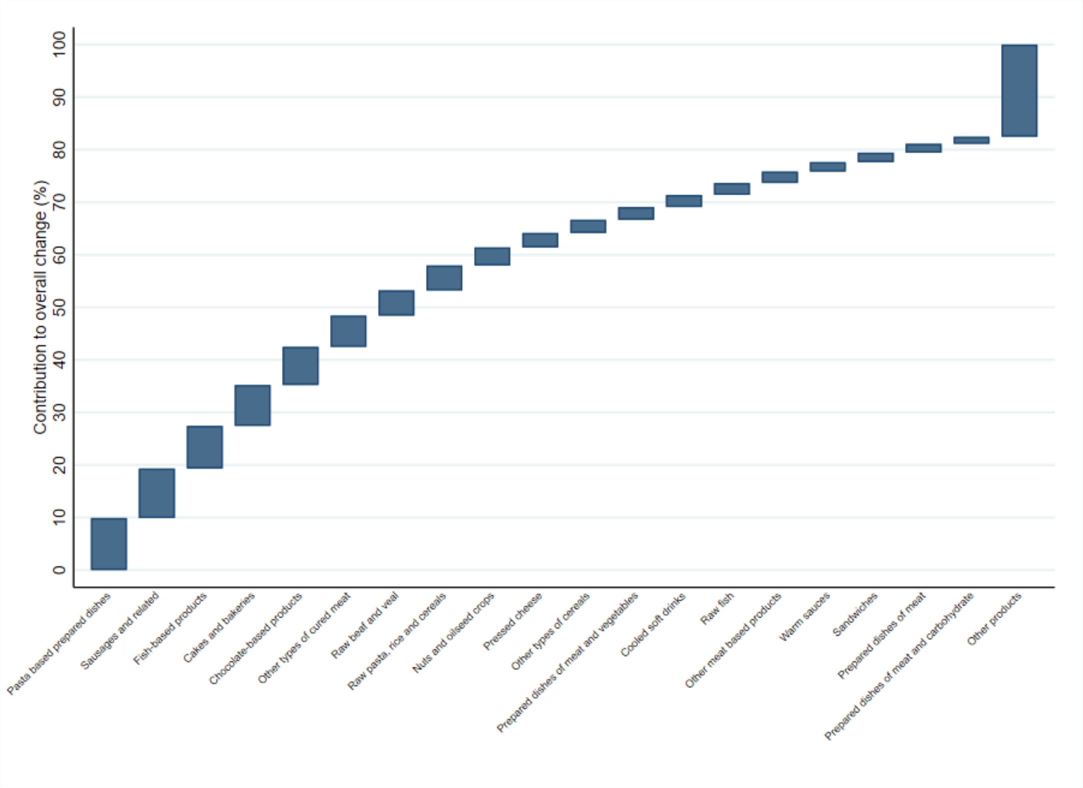
Notes: This figure presents the coefficient of variation of carbon intensity for different product levels. Data comes from the Product data. The coefficient of variation for a variable is the ratio of its standard error to its the mean and is expressed in this figure in percent. To ease exposition, coefficient of variation is winsorized at 1000%. Note that for product department and product groups, the axis are reversed.

Figure 1.C.16: Carbon intensity by Ecoscore



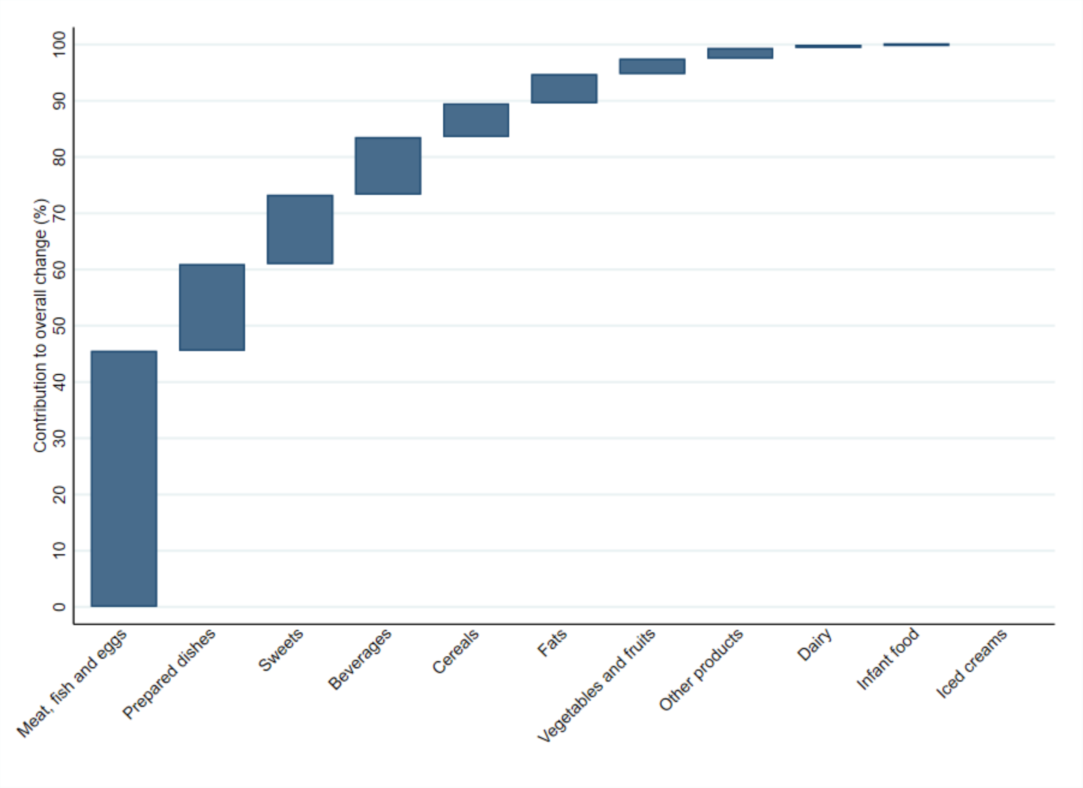
Notes: Dashed line represent the carbon intensity grand average. Data comes from the Product data set.

Figure 1.C.17: Largest contributors following switch to Ecoscore A and B, by product sub-groups



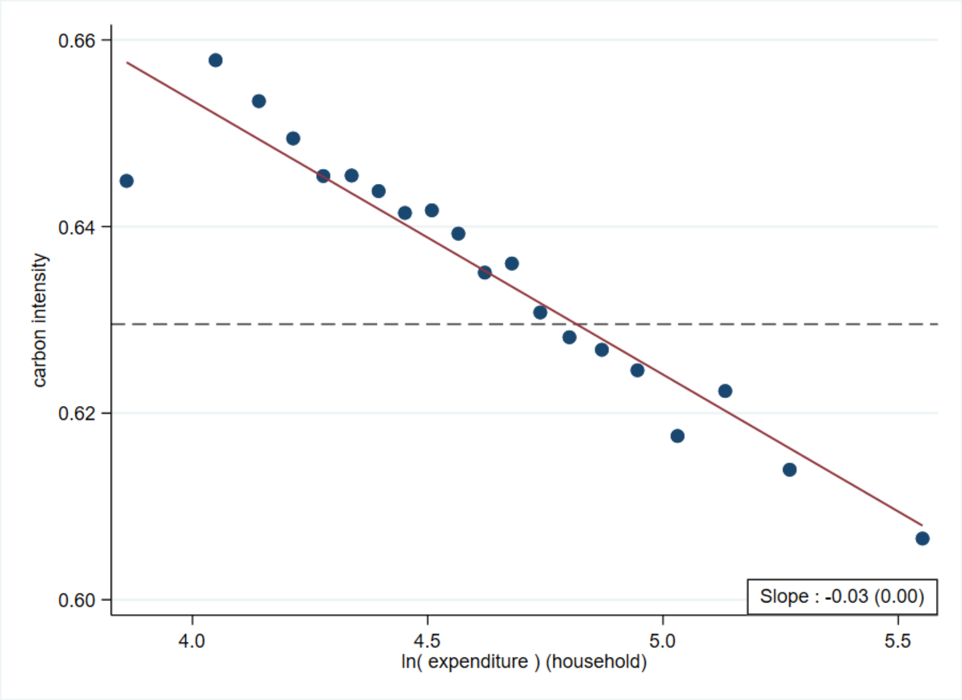
Notes: This figure plots contribution of the top 20 product sub-groups to the overall decrease in carbon emission assuming a shift in consumer demand from product with Ecoscore C and above to products within the same category with Ecoscore A or B. Data comes from the Product data set.

Figure 1.C.18: Largest contributors following switch to Ecoscore A and B, by product departments

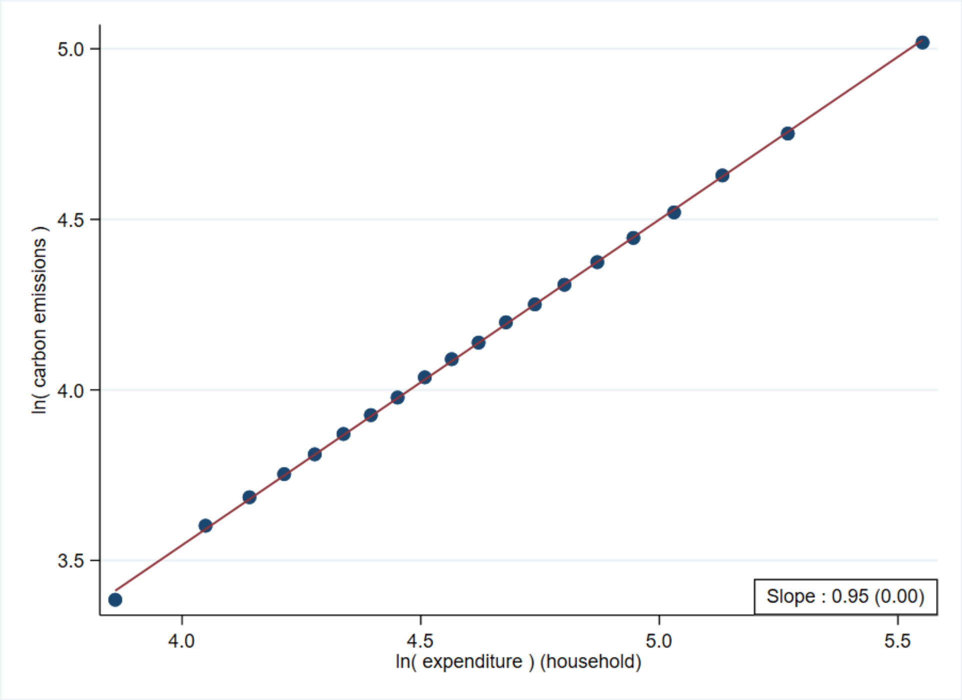


Notes: This figure plots contribution of each product department to the overall decrease in carbon emission assuming a shift in consumer demand from product with Ecoscore C and above to products within the same category with Ecoscore A or B. Data comes from the Product data set.

Figure 1.C.19: Carbon intensity, carbon emissions and expenditures



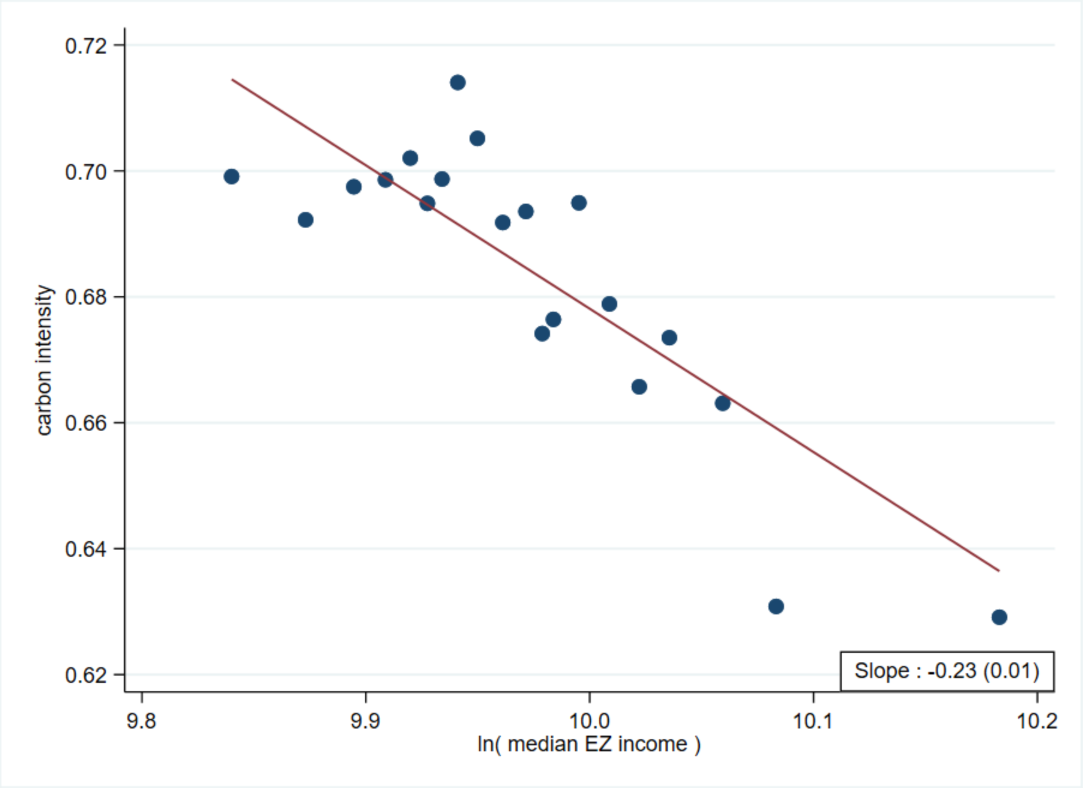
(a) Carbon intensity



(b) Carbon emissions

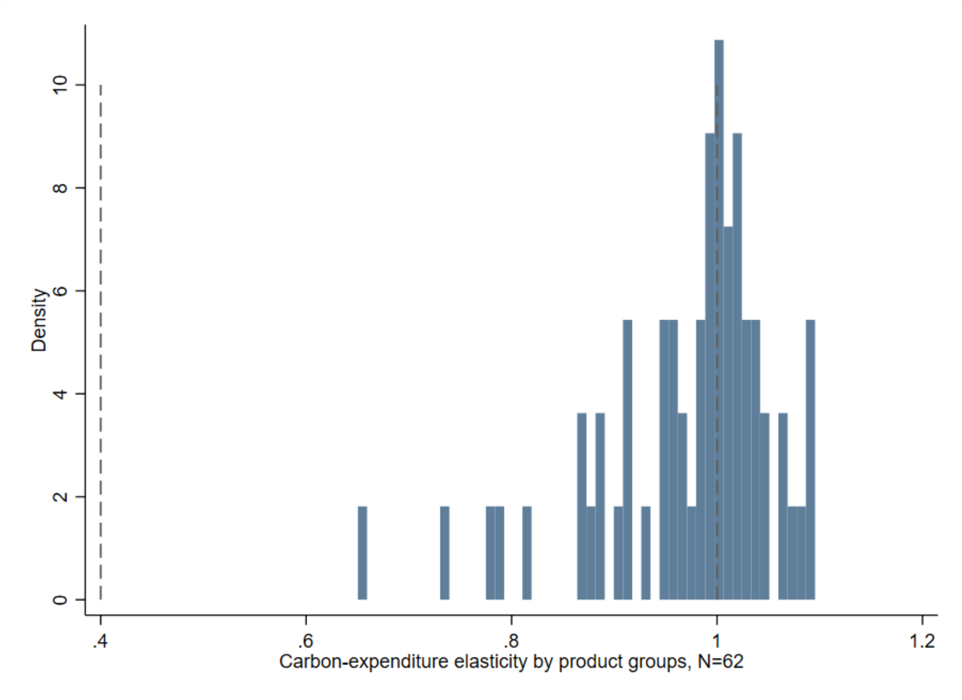
Notes: This figure displays the relationship between carbon intensity and log expenditure (panel (a)) and the carbon-expenditure elasticity (panel (b)) across 116, 816 non-family households from the Household data. Expenditure is defined as the average monthly expenditure of a given household. Dashed line represent the grand average of the y-axis variable. Heteroskedasticity consistent standard errors are reported.

Figure 1.C.20: Carbon intensity and median available income

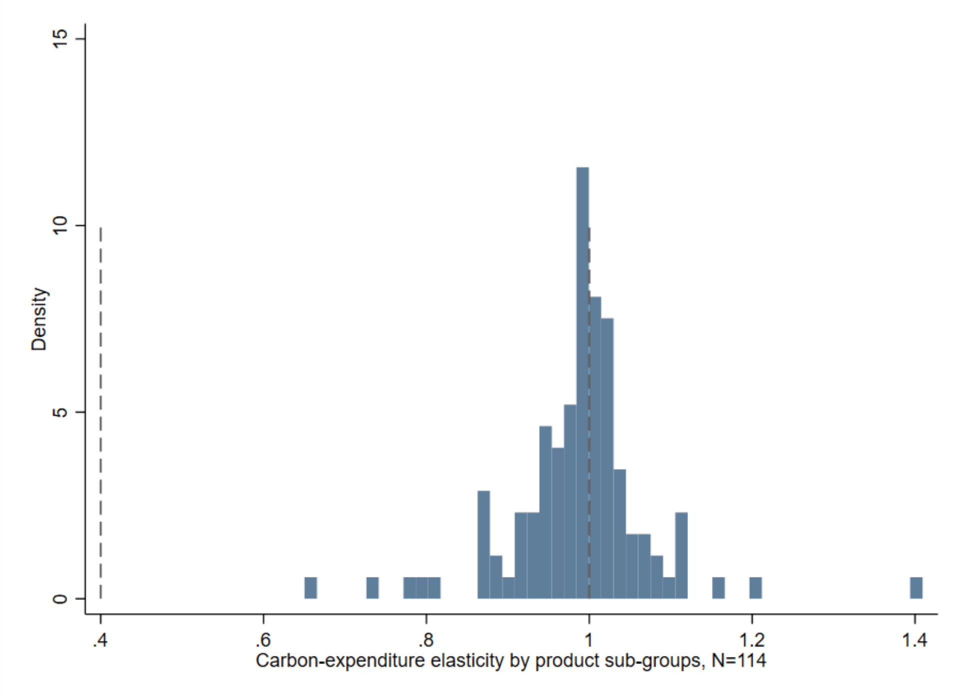


Notes: This figure displays the relationship between carbon intensity and log median available income of the employment zone households shop in, across 116, 816 non-family households from the Household data. Dashed line represent the grand average of the y-axis variable. Heteroskedasticity consistent standard errors are reported.

Figure 1.C.21: Carbon - expenditure elasticity by product category



(a) Product group



(b) Product sub-group

Notes: This figure plots the histogram of estimates of carbon-expenditure elasticity obtained from running a regression of log carbon emissions on log household expenditure for each product group (panel (a)) and for each product subgroup (panel (b)). Data comes from the Household data set. Dashed lines give the Chancel et al. (2015) bounds on carbon-expenditure estimates.

1.D Mathematical appendix

1.D.1 Decomposing carbon intensity

This section derives equation (1.1) decomposing carbon intensity across consumer groups as the sum of three terms. There are I products i in G product categories, with I_g products in every category g. There are R consumer groups indexed by r. Every product i emits E_i units of CO_2, irrespective of where it is sold. Sales of product i to consumer group r are denoted $s_{ir} = p_{ir} \cdot q_{ir}$ where p_{ir} and q_{ir} are the price and quantity of good i faced by consumer group r respectively. Carbon intensity is $e_{ir} = \frac{E_i}{p_{ir}} = \frac{E_i \cdot q_{ir}}{s_{ir}}$. Note that aggregate carbon intensity is always a market share weighted average of carbon intensity at a lower level. For instance, average carbon intensity for consumer group $r$is $e_r = \frac{\sum_i E_i \cdot q_{ir}}{\sum_i s_{ir}} = \sum_i \omega_{ir} \cdot e_{ir}$ where $\omega_{ir} = \frac{p_{ir} \cdot q_{ir}}{s_{ir}}$ is the market share of product i in consumer group r. In what follows, $\bar{x}$ is the grand average value of variable x, across consumer groups.

$$\begin{aligned} e_r - \bar{e} &= \sum_{g=1}^{G} \omega_r^g \cdot e_r^g - \sum_{g=1}^{G} \bar{\omega}^g \cdot \bar{e}^g \pm \sum_{g=1}^{G} \omega_r^g \cdot \bar{e}^g \\ &= \sum_{g=1}^{G} (\omega_r^g - \bar{\omega}^g) \cdot \bar{e}^g + \sum_{g=1}^{G} \omega_r^g \cdot (e_r^g - \bar{e}^g) \end{aligned}$$

Note that

$$\begin{aligned} e_r^g - \bar{e}^g &= \sum_{i=1}^{I_g} e_{ir} \cdot \omega_{ir}^g - \sum_{i=1}^{I_g} \bar{e}_i \cdot \bar{\omega}_i^g \pm \sum_{i=1}^{I_g} \bar{e}_i \cdot \omega_{ir}^g \\ &= \sum_{i=1}^{I_g} (\omega_{ir}^g - \bar{\omega}_i^g) \cdot \bar{e}_i + \sum_{i=1}^{I_g} \omega_{ir}^g \cdot (e_{ir} - \bar{e}_i) \end{aligned}$$

And that

$$\sum_{g=1}^{G} \omega_r^g \cdot \sum_{i=1}^{I_g} \omega_{ir}^g \cdot (e_{ir} - \bar{e}_i) = \sum_{i=1}^{I} \omega_{ir} \cdot (e_{ir} - \bar{e}_i) \equiv - \sum_{i=1}^{I} \omega_{ir} \cdot \delta_{ir}$$

So that, using $\sum_{i=1}^{I_g} (\omega_{ir}^g - \bar{\omega}_i^g) = 0$,

$$\begin{aligned} e_r - \bar{e} &= \sum_{g=1}^{G} (\omega_r^g - \bar{\omega}^g) \cdot \bar{e}^g + \sum_{g=1}^{G} \omega_r^g \cdot \sum_{i=1}^{I_g} (\omega_{ir}^g - \bar{\omega}_i^g) \cdot (\bar{e}_i - \bar{e}^g) - \sum_{i=1}^{I} \omega_{ir} \cdot \delta_{ir} \\ &= cov^g \left(\omega_r^g - \bar{\omega}^g, \bar{e}^g\right) + \sum_{g=1}^{G} \omega_r^g \cdot cov^{i \in I_g} \left(\omega_{ir}^g - \bar{\omega}_i^g, \bar{e}_i - \bar{e}^g\right) - \sum_{i=1}^{I} \omega_{ir} \cdot \delta_{ir} \end{aligned}$$

1.D.2 First order conditions and optimal taxes

Starting from (1.2), using Roy's identity $\frac{\partial V}{\partial(1+t_j)p_j} = -\alpha x_j$, and the fact $\partial V/\partial M = \alpha$ leads to

$$\begin{aligned}
0 &= \sum_k \left[t_k p_k - \frac{\lambda}{\alpha}E_k\right]\frac{\partial x_k}{\partial t_j} \\
&= \sum_{k=1}^{J}\left[t_k - \frac{\lambda}{\alpha}\frac{E_k}{p_k}\right]\cdot p_k \cdot \frac{\partial x_k}{\partial(1+t_j)p_j}\cdot\frac{(1+t_j)p_j}{x_k}\cdot\frac{x_k}{(1+t_j)p_j}\cdot\frac{\partial(1+t_j)p_j}{\partial t_j} \\
&= \sum_{k=1}^{J}\left[t_k - \frac{\lambda}{\alpha}\frac{E_k}{p_k}\right]\cdot p_k \cdot x_k \cdot \frac{\epsilon_{k,j}}{1+t_j}
\end{aligned}$$

where the third line assumes that $\partial p_j/\partial t_j = 0$. Multiplying both sides of the equation by $(1+t_j)/\sum_k p_k \cdot x_k$ gives (1.3).

To derive (1.4), note that

$$\begin{aligned}
t_j - \phi_j =& \frac{1-s_j}{-\epsilon_{j,j}\cdot s_j}\sum_{k\neq j}\frac{s_k}{1-s_j}\cdot(t_k-\phi_k)\cdot\epsilon_{k,j} \\
=& \frac{1-s_j}{-\epsilon_{j,j}\cdot s_j}\mathbb{E}_{k\neq j}\left[(t_k-\phi_k)\cdot\epsilon_{k,j}\right] \\
=& \frac{1}{-\epsilon_{j,j}}\cdot\frac{1-s_j}{s_j}\left(\mathbb{E}_{k\neq j}\left[t_k-\phi_k\right]\mathbb{E}_{k\neq j}\left[\epsilon_{k,j}\right] + cov_{k\neq j}\left[t_k-\phi_k,\epsilon_{k,j}\right]\right) \\
=& \frac{1}{-\epsilon_{j,j}}\cdot\frac{1-s_j}{s_j}\left(\left[\sum_{k\neq j}\frac{s_k}{1-s_j}\cdot(t_k-\phi_k)\right]\cdot\left[\sum_{k\neq j}\frac{s_k}{1-s_j}\cdot\epsilon_{k,j}\right] - cov_{k\neq j}\left[\phi_k-t_k,\epsilon_{k,j}\right]\right) \\
=& -\frac{\epsilon_{-j,j}}{-\epsilon_{j,j}}\cdot\frac{1-s_j}{s_j}\cdot\left(\phi_{-j} + cov_{k\neq j}\left[\phi_k-t_k,\epsilon_{k,j}\right]\right)
\end{aligned}$$

Applying assumption1 delivers the required results. Note that in a case where $cov_{k\neq j}\left[\phi_k - t_k, \epsilon_{k,j}\right] > 0$, that there is an even higher incentive to subsidize good j.

To derive (1.5), note that we can express Marshallian demand for good k as

$x_k = x_k\left((1+t_1)p_1, ..., (1+t_1)p_{I_1}, (1+t_2)p_{I_1+1}, ...; Z+T\right)$. Then, using the quantities defined in section 1.4, we have:

$$\begin{aligned}
\forall g, 0 &= \frac{\partial V}{\partial t_g} + \frac{\partial V}{\partial M}\left[\sum_{j\in g} p_j x_j + \sum_{g'=1}^{G}\sum_{k'\in g'} t_{g'} p_{k'} \frac{dx_{k'}}{dt_g}\right] - \lambda\left[\sum_{g'=1}^{G}\sum_{k'\in g'} E_{k'}\cdot\frac{dx_{k'}}{dt_g}\right]\\
&= \sum_{g'=1}^{G}\sum_{k'\in g'}\left[t_{g'}p_{k'} - \frac{\lambda}{\alpha}E_k\right]\frac{dx_{k'}}{dt_g}\\
&= \sum_{g'=1}^{G}\sum_{k'\in g'}\left[t_{g'} - \phi_{k'}\right]\cdot p_{k'}\cdot\sum_{k\in g}\frac{\partial x_{k'}}{\partial(1+t_g)p_k}\cdot\frac{\partial(1+t_g)p_k}{\partial t_g}\\
&= \sum_{g'=1}^{G}\sum_{k'\in g'}\left[t_{g'} - \phi_{k'}\right]\cdot p_{k'}\cdot\sum_{k\in g}\epsilon_{k',k}\cdot\frac{x_{k'}}{(1+t_g)p_k}\cdot p_k\\
&= \sum_{g'=1}^{G}\sum_{k'\in g'}\left[t_{g'} - \phi_{k'}\right]\cdot\frac{p_{k'}\cdot x_{k'}}{(1+t_g)}\cdot\sum_{k\in g}\epsilon_{k',k}
\end{aligned}$$

$$\begin{aligned}
0 &= \sum_{g'=1}^{G}\sum_{k'\in g'}\left[t_{g'} - \phi_{k'}\right]\cdot s_{k'}\cdot\bar{\epsilon}_{k',g}\\
&= \sum_{g'\in G}\left[t_{g'}\sum_{k'\in g'} s_{k'}\cdot\bar{\epsilon}_{k',g} - \sum_{k'\in g'} s_{k'}\cdot\bar{\epsilon}_{k',g}\cdot\phi_{k'}\right]\\
&= \sum_{g'\in G}\left[t_{g'}s_{g'}\sum_{k'\in g'} s_{k'}^{g'}\cdot\bar{\epsilon}_{k',g} - s_{g'}\sum_{k'\in g'} s_{k'}^{g'}\cdot\bar{\epsilon}_{k',g}\cdot\phi_{k'}\right]\\
&= \sum_{g'\in G} s_{g'}\left[t_{g'}\bar{\epsilon}_{g',g} - \sum_{k'\in g'} s_{k'}^{g'}\cdot\bar{\epsilon}_{k',g}\cdot\phi_{k'}\right]\\
\Rightarrow 0 &= \sum_{g'\in G} s_{g'}\bar{\epsilon}_{g',g}\left[t_{g'} - \tilde{\phi}_{g',g}\right]
\end{aligned}$$

Where the second line uses a modified Roy's identity. Denoting $\tilde{\phi}_{-g} = \sum_{g'\neq g}\frac{s_{g'}}{1-s_g}\cdot\left(\tilde{\phi}_{g',g} - t_{g'}\right)$ and $\bar{\epsilon}_{-g,g} = \frac{1}{1-s_g}\sum_{g'\neq g} s_{g'}\bar{\epsilon}_{g',g}$, we have

$$t_g = \tilde{\phi}_{g,g} - \frac{\bar{\epsilon}_{-g,g}}{-\bar{\epsilon}_{g,g}}\cdot\frac{1-s_g}{s_g}\cdot\left(\tilde{\phi}_{-g,g} + cov_{g'\neq g}\left[\tilde{\phi}_{g',g} - t_{g'}, \bar{\epsilon}_{g',g}\right]\right)$$

Under the assumption that $\forall g$, $cov_{g'\neq g}\left[\tilde{\phi}_{g',g} - t_{g'}, \bar{\epsilon}_{g',g}\right]$, that is, there is no (market share-weighted) correlation between the average marginal (uncorrected) externality of a product category g' and its average cross-price elasticity relative to category g, we find (1.5).

To prove (1.6), all we need to show is that under nested CES, we have $\tilde{\phi}_{g,g} = \sum_{k\in g}\phi_k\cdot s_k^g$ as the rest follows from the definition of carbon intensity. Let us denote σ_g and σ the constant elasticity of substitution at the lower and higher nest respectively. Under nested CES, when k and k' belong to the same category g, then $\epsilon_{k,k'} = \sigma^H \cdot s^g_{k'} - (1-s_g)\cdot\sigma^L \cdot s^g_{k'} - s^g_{k'} \cdot s_g$ for $k \neq k'$ and $\epsilon_{k',k'} = -(1-s^g_{k'})\cdot \sigma^H - (1-s_g)\cdot\sigma^L\cdot s^g_{k'} - s^g_{k'}\cdot s_g$. Hence, we have that

$$\begin{array}{rcccc}
\bar{\epsilon}_{k,g} = \sum_{k'\neq k} & \sigma^H \cdot s^g_{k'} & -(1-s_g)\cdot\sigma^L\cdot s^g_{k'} & -s^g_{k'}\cdot s_g \\
+ & -(1-s^g_k)\cdot\sigma^H & -(1-s_g)\cdot\sigma^L\cdot s^g_k & -s^g_k\cdot s_g \\
= & 0 & -(1-s_g)\cdot\sigma^L & -s_g
\end{array}$$

Then,

$$\begin{aligned}
\bar{\epsilon}_{g,g} =& \sum_{k\in g} \bar{\epsilon}_{k,g}\cdot s^g_k \\
=& \sum_{k\in g}\left[-(1-s_g)\cdot\sigma^L - s_g\right]\cdot s^g_k \\
=& -(1-s_g)\cdot\sigma^L\sum_{k\in g} s^g_k - s_g\sum_{k\in g} s^g_k \\
=& -(1-s_g)\cdot\sigma^L - s_g \\
=& \bar{\epsilon}_{k,g}
\end{aligned}$$

so that $w^{g,g}_k = \frac{\bar{\epsilon}_{k,g}\cdot s^g_k}{\bar{\epsilon}_{g,g}} = s^g_k$ and $\tilde{\phi}_{g,g} = \sum_{k\in g}\phi_k\cdot s^g_k$

1.D.3 Welfare change

Counterfactual utility change CES-ideal aggregate price indices are defined as:

$$\begin{array}{rr}
P_{mt} = \left(\sum_{j\in\Omega_{mt}}(p_{jt}/d_j)^{1-\sigma_m}\right)^{\frac{1}{1-\sigma_m}} & P_{gt} = \left(\sum_{m\in\Omega_g} P^{1-\sigma_g}_{mt}\right)^{\frac{1}{1-\sigma_g}} \\
P_{dt} = \left(\sum_{g\in\Omega_d} P^{1-\sigma_d}_{gt}\right)^{\frac{1}{1-\sigma_d}} & P_{Ft} = \left(\sum_{d\in\Omega} P^{1-\sigma}_{dt}\right)^{\frac{1}{1-\sigma}}
\end{array}$$

We first show how we can recover $d\ln P_{mt} = \ln P_{mt} - \ln P_{mt-1}$ from knowing the substitution elasticities and market shares only. This derivation follows closely Broda and Weinstein (2010). First, household optimization implies that we can express the market share of product j in module m at time t as: $s^m_{jt} = \left(\frac{p_{jt}/d_j}{P_{mt}}\right)^{1-\sigma_m}$, or equivalently that $P_{mt} = s^m_{jt}\cdot(p_{jt}/d_j)$ $\forall k \in \Omega_{mt}$. Let the set of goods j available both in time t and $t-1$ in module m be Ω^*_{mt}. Then, using the fact that $p_{jt}\cdot x_{jt} = (p_{jt}/d_j)^{1-\sigma_m}\cdot P^{-\sigma_g}_{mt}$ the share of common varieties in period t is:

$$
\begin{aligned}
\lambda_t =& \frac{\sum_{j\in\Omega^*_{mt}} p_{jt}x_{jt}}{\sum_{j\in\Omega_{mt}} p_{jt}x_{jt}} \\
=& \frac{\sum_{j\in\Omega^*_{mt}} (p_{jt}/d_j)^{1-\sigma_m}}{\sum_{j\in\Omega_{mt}} (p_{jt}/d_j)^{1-\sigma_m}}
\end{aligned}
$$

Hence, we have that :

$$
\begin{aligned}
s^m_{jt} &= \frac{(p_{jt}/d_j)^{1-\sigma_m}}{P_{mt}^{1-\sigma_m}} \\
&= \frac{(p_{jt}/d_j)^{1-\sigma_m}}{\sum_{j\in\Omega^*_{jt}} (p_j/d_j)^{1-\sigma_m}} \cdot \lambda_t \\
&= s^{m*}_{jt} \cdot \lambda_t \qquad \forall j \in \Omega^*_{mt}
\end{aligned}
$$

where s^{m*}_{jt} is the market share of product j among all products of module m in period t that are also available in period $t-1$. We can then express the change in aggregate price index as:

$$
\begin{aligned}
\frac{P_{mt}}{P_{mt-1}} &= \frac{p_{jt}}{p_{jt-1}} \cdot (\lambda_t/\lambda_{t-1})^{\frac{1}{\sigma_m-1}} \cdot \left(s^{m*}_{jt}/s^{m*}_{jt-1}\right)^{\frac{1}{\sigma_m-1}} \\
\ln\frac{P_{mt}}{P_{mt-1}} &= \frac{1}{\sigma_m-1}\ln\left(\frac{\lambda_t}{\lambda_{t-1}}\right) + \underbrace{\ln(p_{jt}/p_{jt-1}) + \frac{1}{\sigma_m-1}\ln\left(s^{m*}_{jt}/s^{m*}_{jt-1}\right)}_{\equiv d\ln P^*}
\end{aligned}
$$

Note that we can reorder $d\ln P^*$ as

$$
\frac{1}{\sigma_m-1} = \frac{d\ln P^* - \ln(p_{jt}/p_{jt-1})}{\ln\left(s^{m*}_{jt}/s^{m*}_{jt-1}\right)}
$$

Multiplying both sides by $(s^{m*}_{jt} - s^{m*}_{jt-1})$ and summing over j, gives:

$$
\begin{aligned}
0 &= \sum_{j\in\Omega^*_{mt}} (s^{m*}_{jt} - s^{m*}_{jt-1}) \frac{d\ln P^* - \ln(p_{jt}/p_{jt-1})}{\ln\left(s^{m*}_{jt}/s^{m*}_{jt-1}\right)} \\
\sum_{k\in K_m} \frac{s^{m*}_{jt} - s^{m*}_{jt-1}}{\ln s^{m*}_{jt} - \ln s^{m*}_{jt-1}} d\ln P^* &= \sum_{j\in\Omega^*_{mt}} \frac{s^{m*}_{jt} - s^{m*}_{jt-1}}{\ln s^{m*}_{jt} - \ln s^{m*}_{jt-1}} \ln(p_{jt}/p_{jt-1}) \\
\Rightarrow d\ln P^* &= \sum_{j\in\Omega^*_{mt}} \omega^{SV}_{kt} \ln(p_{jt}/p_{jt-1}) \\
\text{where } \omega^{SV}_{jt} &= \frac{\frac{s^{m*}_{jt} - s^{m*}_{jt-1}}{\ln s^{m*}_{jt} - \ln s^{m*}_{jt-1}}}{\sum_{l\in\Omega^*_{mt}} \frac{s^{m*}_{lt} - s^{m*}_{lt-1}}{\ln s^{m*}_{lt} - \ln s^{m*}_{lt-1}}}
\end{aligned}
$$

are the Sato-Varia weights. Bringing everything together, we have that

$$d\ln P_{mt} = \frac{1}{\sigma_m - 1}\ln\left(\frac{\lambda_t}{\lambda_{t-1}}\right) + \sum_{j\in\Omega^*_{mt}} \omega^{SV}_{kt}\ln\left(p_{jt}/p_{jt-1}\right)$$

This shows that in a counterfactual analysis, where the set of available goods does not change, so that $\lambda_t = \lambda_{t-1}$, where t and $t-1$ denote the counterfactual and initial situation respectively, we can recover the change in aggregate price index using only counterfactual changes in prices and market shares.

Using assumption 2, we have $s^m_{jt} - s^m_{jt-1} = \frac{s^m_{jt} - s^m_{jt-1}}{s^m_{jt-1}} \cdot s^m_{jt-1} \approx d\ln s^m_{jt} \cdot s^m_{jt-1}$ so that we can reformulate the Sato-Varia weights as:

$$\begin{aligned}
\omega^{SV}_{jt} &= \frac{\frac{s^{m*}_{jt} - s^{m*}_{jt-1}}{\ln s^{m*}_{jt} - \ln s^{m*}_{jt-1}}}{\sum_{l\in\Omega^*_{mt}} \frac{s^{m*}_{lt} - s^{m*}_{lt-1}}{\ln s^{m*}_{lt} - \ln s^{m*}_{lt-1}}} \\
&\approx \frac{\frac{d\ln s^m_{jt} \cdot s^m_{jt-1}}{d\ln s^m_{jt}}}{\sum_{l\in\Omega^*_{mt}} \frac{d\ln s^m_{lt} \cdot s^m_{lt-1}}{d\ln s^m_{lt}}} \\
&= \frac{s^m_{jt-1}}{\sum_{l\in\Omega^*_{mt}} s^m_{lt-1}} \\
&= s^m_{jt-1}
\end{aligned}$$

Hence, we can use the fact that $s^m_{jt} = d_j^{\sigma_m - 1} \cdot \left(\frac{p_{jt}}{P_{mt}}\right)^{1-\sigma_m}$, to show that:

$$\begin{aligned}
d\ln s^m_{jt} &= (1-\sigma_m)\cdot(d\ln p_{jt} - d\ln P_{mt}) \\
&= (1-\sigma_m)\cdot(d\ln p_{jt} - \sum_{j\in\Omega^*_{mt}} \omega^{SV}_{jt} \cdot d\ln p_{jt}) \\
&\approx (1-\sigma_m)\cdot(d\ln p_{jt} - \sum_{j\in\Omega^*_{mt}} s^m_{jt-1} \cdot d\ln p_{jt})
\end{aligned}$$

It is then easy to recover $d\ln P_g$, $d\ln P_d$ and $d\ln P_F$ using the same process. To arrive to equation (1.7), note that the nested CES structure imply that $U_F = \frac{\beta M}{P_F}$, where βM is the (constant) amount of expenditure on food from the household so that $d\ln U_F = -d\ln P_F$. From then, we simply use the fact that the indirect utility function is $U = (1-\beta)^{1-\beta}\cdot\beta^\beta\cdot p_0^{-(1-\beta)}\cdot P_F^{-\beta}\cdot M$.

Counterfactual quantity changes. The nested structure of the CES model implies:

$$\begin{aligned} x_j = d_j^{\sigma_m - 1} \cdot (p_j/P_m)^{-\sigma_m} \cdot C_m \quad & C_m = (P_m/P_g)^{-\sigma_g} \cdot C_g \\ C_g = (P_g/P_d)^{-\sigma_d} \cdot C_d \quad & C_d = (P_d/P_F)^{-\sigma} \cdot U_F \end{aligned}$$

Taking the log difference between the counterfactual and the initial situation and substituting in implies

$$\begin{aligned} d \ln x_j =& - \sigma_m d \ln p_j + (\sigma_m - \sigma_g) d \ln P_m + (\sigma_g - \sigma_d) d \ln P_g \\ & + (\sigma_d - \sigma) d \ln P_d + \sigma d \ln P_F + d \ln U_F \\ =& \sigma_m (d \ln P_m - d \ln p_j) + \sigma_g (d \ln P_g - d \ln P_m) + \sigma_d (d \ln P_d - d \ln P_m) \\ & + \sigma (d \ln P_F - d \ln P_d) + d \ln U_F \end{aligned}$$

From there, using $d \ln U_F = -d \ln P_F$ gives equation (1.8).

To derive, equation (1.9), we first ease the exposition by assuming that the household's utility for food U_F is nested CES with only two levels, and let σ_g be the constant elasticity of substitution of products within category g and σ be the constant elasticity of substitution across categories. The proof for more nests follows the same idea. We have that $d \ln x_i = -d \ln P_F - \sigma(d \ln P_g - d \ln P_F) - \sigma_g(d \ln p_j - d \ln P_F)$. We have then:

$$\begin{aligned} & \sum_i e_i \cdot s_i \cdot \sigma (d \ln P_g - d \ln P_F) \\ =& \sum_g s_g \cdot \sigma (d \ln P_g - d \ln P_F) \sum_{i \in g} e_i \cdot s_i^g \\ =& \sigma \sum_g s_g \cdot \bar{e}_g (d \ln P_g - d \ln P_F) \\ =& \sigma \sum_g s_g \cdot (\bar{e}_g - \bar{e}) \cdot (d \ln P_g - d \ln P_F) \\ =& \sigma \cdot cov^g \left(\bar{e}_g - \bar{e}, d \ln P_g - d \ln P_F \right) \end{aligned}$$

where the third line comes from the fact that $\sum_g s_g(d\ln P_g - d\ln P_F) = 0$. Similarly

$$\begin{aligned}
&\sum_i e_i \cdot s_i \cdot \sigma_g(d\ln p_i - d\ln P_g) \\
=&\sum_g s_g \cdot \sigma_g \sum_{i\in g} s_i^g \cdot e_i \cdot (d\ln P_g - d\ln P_F) \\
=&\sum_g s_g \cdot \sigma_g cov^{i\in g}\left(e_i, d\ln P_g - d\ln P_F\right) \\
=&\sum_g s_g \cdot \sigma_g cov^{i\in g}\left(e_i - \bar{e}_g, d\ln P_g - d\ln P_F\right) \\
=&\sum_g s_g \cdot \sigma_g \mathfrak{T}_g
\end{aligned}$$

so that equation (1.9) follows.

Carbon emission target. To prove equation (1.10), we start from the fact that at a uniform Pigouvian tax has the form $t^{unif} = p^{CO_2} \cdot \bar{e}$, where $\bar{e}$ is the average carbon intensity and $p^{CO_2} = \lambda/\alpha$ is the social cost of carbon. The key of the argument is to realize that from the social planner Lagrangian, we know that p^{CO_2} is a function of the total constraint, hence of $\overline{\Delta E}$. Therefore, using (1.9), we can solve for p^{CO_2}. First, we have that $d\ln P_F \approx \sum_i s_i d\ln p_i = \sum_i s_i t^{unif} = p^{CO_2} \cdot \bar{e}$, so that $\overline{\Delta E}/E = -\bar{e} \cdot p^{CO_2}$. Solving for p^{CO_2} and plugging into the formula for the log utility change (1.7) gives equation (1.10).

Proof for equation (1.11) follows the same idea. First, a product department level has the form $t_d = p^{CO_2} \cdot \bar{e}_d$, where $\bar{e}_d$ is the average carbon intensity of product department d. Further, $d\ln P_d \approx \sum_{i\in d} s_i^d d\ln p_i = p^{CO_2} \cdot \bar{e}_d$ and $d\ln P_F \approx \sum_d s_d d\ln P_d = p^{CO_2} \cdot \sum_d s_d \bar{e}_d = p^{CO_2} \cdot \bar{e}$. Last, note that $\mathfrak{T} = cov(\bar{e}_d - \bar{e}, d\ln P_d - d\ln P_F) = p^{CO_2} \cdot var(\bar{e}_d - \bar{e})$. Thus,

$$p^{CO2} = \frac{-\overline{\Delta E}/E}{\bar{e} + \sigma \cdot var(\bar{e}_d - \bar{e})}$$

Plugging into $d\ln U^{dep} = -\beta d\ln P_F = -\beta\bar{e} \cdot p^{CO2}$ and dividing by $d\ln U^{unif}$ gives (1.11). Proof for the other equations follow the same idea.

Chapter 2

Inflation dynamics of fast-moving consumer goods during lockdown in France

Abstract

This paper uses scanner data to document the inflation dynamics of fast-moving consumer goods during and after the first lockdown in France (March - May 2020). I find that the lockdown lead to an important, generalized but transitory inflation spike of 2.3% at the highest in April 2020. Most of the inflation is accounted for by a price increase from national brands, rather than from retailer's owned brands. Contrary to what has been found in other countries, the role of promotions and net entry of products did not significantly change during this period. Further, this inflation shock was very asymmetric: 9.4 % of households experienced an inflation rate of 5% or higher, nineteen times higher than in 2019. Importantly, this asymmetry persisted beyond the lockdown. Overall, this paper illustrates how scanner-level data can be useful in conducting quasi-real time analysis of inflation in time of crisis, especially in understanding the heterogenous impacts of exogenous shocks.

2.1 Introduction

Inflation matters as a measure of well-being and cost of living, but also as a tool for economic diagnosis. In this regard, the recent Covid-19 pandemic has been particularly problematic for two reasons. Firstly, it disrupted the construction of important economic indicators used for economic policy, including inflation. For instance, in France, the price data collection process from the national statistical agency Insee was stopped during the first lockdown, leading to lower quality inflation data during these months[1]. Secondly, it unfolded at an extreme pace, making traditional economic indicators, which are published by official agencies with a lag of at least several weeks, much less relevant for policy making and for designing adequate policy responses.

This paper uses real-time scanner data from a large French retailer to document the inflation dynamics of fast-moving consumer goods during the first lockdown in France, which lasted 55 days between the 17th of March 2020 and the 10th of May 2020. In addition to providing stylized facts and aggregate measures of inflation, it explores the underlying mechanisms as well as inflation heterogeneity across products, cities and households. By doing so, it also illustrates how statistical agencies could use real-time scanner data to gain new insights on the dynamics and heterogeneity of inflation beyond aggregate statistics, which has important implications for economic diagnosis and efficient policy targeting.

This paper contains four sets of results. First, it documents an important inflation spike during lockdown compared to the same period in 2019, at 2.3% in April 2020 whereas inflation was only 0.9% in April 2019. This inflation spike was transitory, as it returns to 2019 levels at the end of the lockdown. Second, the usual inflation indices were no more subject to the well-known product entry and subsitution biases in the exceptionnal circumstances of the lockdown than in the previous year. Contrary to what has been found in the United-Kingdom (Jaravel and O'Connell, 2020b), product entry and exit did not significantly impact the measurement of inflation during lockdown. Further, while changes in consumption patterns were significantly more important in 2020 than in 2019, they did not translate to higher substitution bias. Third, the vast majority of the inflation spike is driven by a price increase in products from national brands, rather than by a price increase from private label brands or a composition effects between brand types with different inflation trajectories. Inflation from national brands, which accounted for 77 % of sales in 2019, was 2.9% in April 2020, whereas private label brand inflation was 1.2% only. Discussions with the private retailer's management and pricing team suggest that this has been driving by competition for goods across retailers during lockdown in a context of high tensions on the supply side. Similarly, I document significantly higher inflation from store types from predominently urban areas, which is not explained by composition effects. Promotions or

[1] see for instance: https://www.insee.fr/fr/statistiques/serie/001759971; accessed 15 November 2021.

differential inflation between in-store and online sales were not a major driver of inflation during lockdown. Fourth, the inflation shock during lockdown was very asymmetric, with a small number of products modules, households and to a lesser extent, cities, facing significant inflation and the bulk of the distribution facing only moderate inflation. For example, the share of households experiencing an inflation rate of 5% or higher in April 2020 was 9.4 , nineteen times higher than in 2019, where the same proportion was 0.5 . Importantly, the right-tail of the inflation distribution across products, cities and households generated by the lockdown shock reduces post-lockdown, but does not disappear nor returns to 2019 levels even though the average inflation does. This suggests that longer term impacts for the most exposed product modules, cities or households are likely, and that going beyond the first moment of the distribution matters.

This paper relates first and foremost to the literature on the impact of Covid-19 on economic activity. Several papers have used credit and debit card transaction data to study the evolution of consumption during the pandemic, both in the United-States (Baker et al., 2020b) as well as in other countries (e.g., Andersen et al. (2020) in Danemark, Chen et al. (2021) in China). Others have looked at firm activity (Bartik et al., 2020), employment (Forsythe et al., 2020), uncertainty (Baker et al., 2020a), or a combination of these (Chetty et al., 2020). To the best of my knowledge, the only studies on inflation dynamics during the Covid lockdowns using real-time data where conducted in the United-Kingdom(Jaravel and O'Connell, 2020b,a), in the United-States (Cavallo, 2020) and Switzerland (Seiler, 2020). This paper contributes to this literature by providing, to the best of my knowledge, the first evidence on inflation dynamics and distribution across cities, products and households during the first lockdown in France. This paper also relates to an older literature on measuring inflation using scanner data (Broda and Weinstein, 2006, 2010; Jaravel, 2019). It illustrates how this type of data be used to inform policy making in quasi-real time in times of crisis, but also how higher moments of the inflation distribution which are invisible in aggregate statistics can deliver new and important insights.

The rest of the paper is organized as follows. Section 2.2 presents the data, section 2.3 analyzes the inflation dynamics as well as potential sources of bias, section 2.4 discusses different mechanisms that might explain the aggregate findings, and section 2.5 explores inflation heterogeneity across products, cities and households. Section 3.7 concludes.

2.2 Data

In this section, I present the data and key stylized facts on spending, prices, promotions and product variety.

2.2.1 Data

I use scanner data from a large private retailer chain in France. The data covers the universe of fast-moving consumer goods transactions from this retailer between January and August 2020, as well as between January and August 2019. These fast-moving consumer goods include food, beverages, cleaning products, pet food and cosmetics. Product level information includes overall description, brand, whether it has been sold online and whether it has been sold in promotion. The data set contains 1.4 million universal product codes (UPCs), nested in 843 different product modules, which are in turn nested in 299 product groups, themselves nested in 27 product departments. This retailer operates as of 2021 approximatly $2,000$ stores in metropolitan France. I also use publicly available data from the French statistical agency to enrich the private retailer data with city-level information on average income and socio-demographic composition[2]. Furthermore, each transaction contains a loyalty card identifier, which is used to construct a panel of household-level transactions across the period and is further discussed in section 2.5.3. Sales from loyalty card represent 70 % of overall sales. Throughout the paper, I assume that a loyalty card is equivalent to a household. While this is certainly not always true, discussions with the private retailer's management suggest this is largely true in practice. Other than a unique identifier, household-level information is very scarce. The only other information available is whether the household is a family, a young household (18-35 with no kids), a middle aged household (36-60 with no kids) or a senior household (61+ with no kids).

This data is similar in nature to the Nielsen Homescan data, which has been used extensively in the literature (for a description of the Nielsen data, see among others Broda and Weinstein (2010); Allcott et al. (2019a); Jaravel (2019)). Like the Nielsen data, price and quantity are separately available for each transaction and consumption is recorded at the UPC level, which ensures that we can keep track of quality improvement over time, as retailers change barcodes when meaningful characteristics of the product are changed (Broda and Weinstein, 2010). Contrary to the Nielsen Homescan data however, households in the raw data are not weighted to be nationally representative. I follow most of the literature using private data to inform economic activity (e.g., Chetty et al. 2020; Baker et al. 2020b; Andersen et al. 2020) and do not attempt to reweight the data to make it representative. Rather, I instead compare in section 2.2.2 the inflation dynamics in this data with the closest publicly available inflation series.

2.2.2 Comparison with Insee

The French statistical agency Insee publishes inflation data from the mass retail sector, which is the closest publicly available data to compare the private retailer's data with. To

[2] In particular, I use the Filosophie data base available at https://www.insee.fr/fr/metadonnees/source/serie/s1172; accessed 15 November 2021.

make both data sets comparable, I construct a mass retail price index following as closely as possible their technical reports (Caillaud, 1998; Insee, 2016a,b). First, follow the Insee and I ignore fresh products (seafood, fish, flowers, etc.). Second, Insee builds its aggregate index by constructing a weighted average of elementary price indices for every geography by product variety cell. To define these cells, Insee uses 13 geographic zones and $1,100$ product varieties; but the data used in this paper allows me only to use 8 geographic zones and 650 product variety. Third, Insee uses the cell's previous year expenditure share as weights, which are updated every year in January. To follow this set-up and construct the 2018 expenditure weights for every cell, I use another data set from the same retailer covering all transactions from about $800,000$ loyalty cards between 2017 and 2019, which is a 10% random sample all existing loyalty cards issued by the retailer. Fourth, the monthly price index within a cell depends on the product variety. For heterogenous product varieties, Insee computes a Jevon index, an unweighted geometric average of the price ratios between periods t and $t-1$, defined as $J_{c,t} = \Pi_{i=1}^{N} \frac{p_{i,t}^{1/N}}{p_{i,t-1}^{1/N}}$, where $p_{i,t}$ is the price of product i from product variety c in period t. For homogenous product varieties, Insee computes a Dutot index, a ratio of the unweighted arithmetic average of prices between periods t and $t-1$, defined as $D_{c,t} = \frac{N^{-1} \cdot \sum_{i=1}^{N} p_{i,t}}{N^{-1} \cdot \sum_{i=1}^{N} p_{i,t-1}}$. Note that both $J_{c,t}$ and $D_{c,t}$ are defined for continuing products only. Because Insee does not specify which variety is considered as homogenous and heterogenous, I report two series, one constructed with Jevons indices only and the other one with Dutot indices only.

The result of this comparison is presented in appendix figure 2.B.1 for the year 2019. Overall, both inflation series constructed following the process described above closely track Insee's mass retail inflation index. The difference between Jevons-based and Dutot-based price indices using the private retailer's data is relatively small compared to overall inflation and seems constant across months. Overall, I conclude that the private retailer's data is roughly representative of the mass retail sector in France, even though it might be less representative of the overall French population.

2.2.3 Stylized facts

Table 2.1 presents the share of 2019 and 2020 sales made on different types of stores, online versus in-store sales, and sales on items in promotion. The split by store types is relatively constant between 2019 and 2020 in aggregate, that is accounting for lockdown as well as non-lockdown periods in 2020. About three quarters of overall sales are made on 'Super' stores, a mid-sized type of stores. Larges 'Hyper' stores represent approximately 12% of sales, smaller 'Contact' and 'Express' stores account for about 6% and 3% respectively. While online sales represented only 2.6% of overall sales in 2019, it surged to 4.2% in 2020. Promotions represent about 6.9% of sales in 2019 and this proportion does not change in 2020.

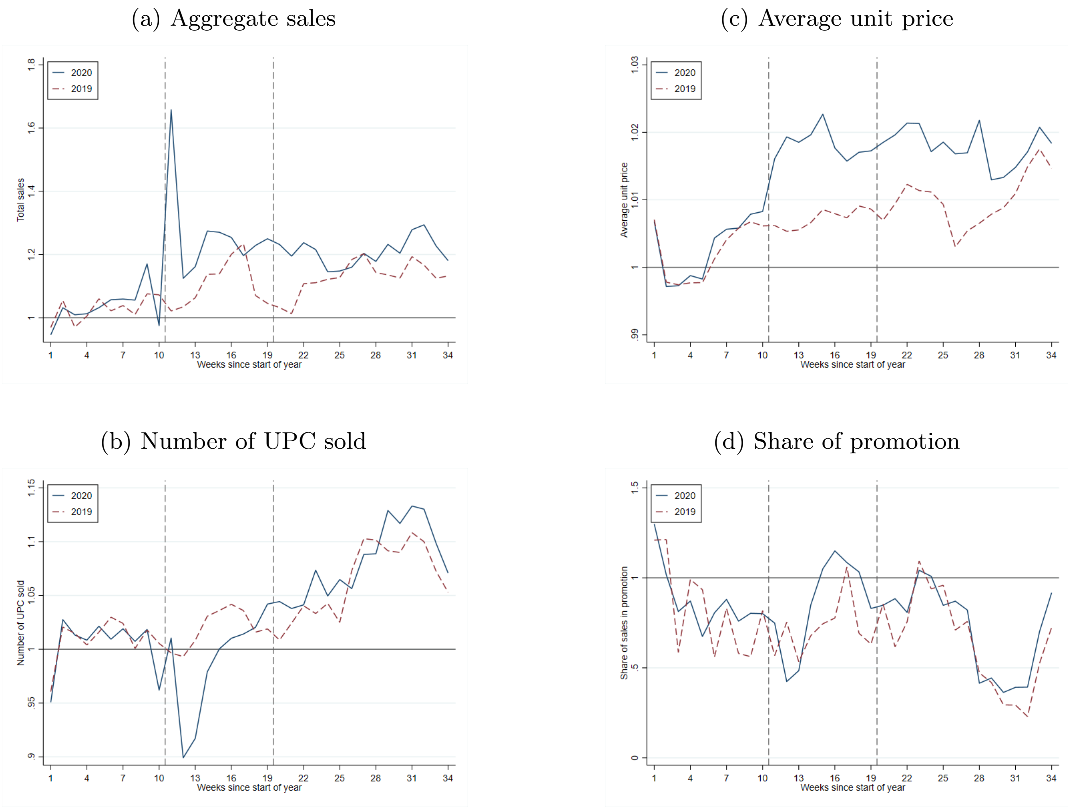
Table 2.1: Sales by store type, drive status and promotions in 2019, in % of total

	2019	2020
Store type		
Contact stores	5.7	5.9
Express stores	2.8	3.0
Super stores	78.9	78.7
Hyper stores	12.6	12.4
Online	2.6	4.2
Promotions	6.9	6.9

Notes: Contact stores are between 500 and 1,200 square meters, primarily in rural areas; Express stores are between 300 and 1,200 square meters, primarily in urban areas; Super stores are between 1,200 and 3,500 square meters and Hyper stores are between 3,500 and 6,000 square meters.

Figure 2.1 reports descriptive evidence on the dynamics of overall sales, number of UPCs sold, average unit price and share of sales in promotion in 2019 and 2020. For all panels, values are expressed relative to the average of the first four weeks in every year. Panel (a) shows that the 2020 aggregate sales followed closely the 2019 until the beginning of the lockdown, during the 11th week of the year. Aggregate sales during the first week of lockdown increased by 60% relative to the month of January, suggesting an important initial shock of lockdown. From week 12 onwards, sales remained roughly constant and approximately 20% higher than in January, even after the end of lockdown on week 19. The 2020 aggregate sales trend starts again to track the 2019 trend from week 25 onwards, approximately a month after the end of lockdown. This return to normal coïncides with the partial reopening of restaurants on 2nd of June 2020. Panel (b) documents the evolution of the number of UPCs sold, which is an important metric as a change in product variety impacts consumer welfare even if prices do not change. Before lockdown, the number of UPCs sold is relatively constant across weeks for both years, both for 2019 and 2020. While in 2019 the number of UPCs sold increased slightly and relatively constantly from week 12 onwards, there is an important fall of close to 10% in the number of UPCs purchased right after the beginning of lockdown. In contrast to what was reported for the United-Kingdom however (Jaravel and O'Connell, 2020b), this drop is only transitory and lasts a few weeks only: by week 15 onwards, the trend becomes similar to 2019 again. While the data cannot distinguish whether this fall is due to UPCs not being offered or not being chosen, this particular pandemic setting, as well as the fact that the drop is temporary and resumes before the end of lockdown points towards a temporary shortage rather than, for instance, a temporary change in shoping behaviours. Panel (c) presents the evolution of the average unit price at this retailer in 2019 and 2020. The average unit price is constructed for every week by taking the sales-weighted average of unit price of available UPCs. The

Figure 2.1: Stylized facts



(a) Aggregate sales

(c) Average unit price

(b) Number of UPC sold

(d) Share of promotion

Notes: This figure plots total sales (panel (a)), number of UPC solds (panel (b)), the average unit price (panel (c)) and the share of transactions in promotion (panel (d)). Panel (c) uses sales-weighted average unit price. Values are normalized to the average of the first four weeks in every year. Dashed lines denotes the start and end of the French first lockdown.

evolution of the average unit price in both years closely tracks each other during the first two months. Starting from the lockdown week, the 2020 average unit price jumps by more than 1 percentage point, whereas the 2019 average unit price is relatively flat. This 2020 average unit price remains higher than 2019 during lockdown and beyond: only from week 30 onwards do the 2019 and 2020 trends track each other closely again. Overall, this provides simple evidence that prices did increase during lockdown in France. Panel (d) plots the weekly evolution of the share of sales in promotion during both years. Overall, the 2020 and 2019 trends are very similar, and contrary to what was found for the United-Kingdom (Jaravel and O'Connell, 2020b), the drop in promotions does not seem to be related with the increase in unit prices.

2.3 Inflation dynamics during French lockdown

This section first presents the aggregate price indices used in the main analysis, justifies their usage. It then presents results on aggregate inflation, and considers whether well-known biases, namely the product entry bias and the substitution bias, had an impact on the measurement of inflation during lockdown.

2.3.1 Aggregate price indices

Following Jaravel and O'Connell (2020b), this paper uses the following prices indices:

$$\text{Laspeyres: } 1+\pi_t^L = \sum_i s_{i,t-1} \cdot \frac{p_{i,t}}{p_{i,t-1}} \tag{2.1}$$

$$\text{Paasche:} 1+\pi_t^P = \sum_i s_{i,t} \cdot \frac{p_{i,t}}{p_{i,t-1}} \tag{2.2}$$

$$\text{Fisher: } 1+\pi_t^F = \left(1+\pi_t^L\right)^{1/2} \cdot \left(1+\pi_t^P\right)^{1/2} \tag{2.3}$$

$$\text{Tornqvist:} 1+\pi_t^T = \Pi_i \left(\frac{p_{i,t}}{p_{i,t-1}}\right)^{\frac{s_{i,t}+s_{i,t+1}}{2}} \tag{2.4}$$

$$\text{CES: } 1+\pi_t^{CES} = \Pi_i \left(\frac{p_{i,t}}{p_{i,t-1}}\right)^{\omega_{i,t}} \tag{2.5}$$

$$\text{with } \omega_{i,t} = \frac{(s_{i,t}-s_{i,t-1})/(\ln s_{i,t} - \ln s_{i,t-1})}{\sum_j (s_{j,t}-s_{j,t-1})/(\ln s_{j,t} - \ln s_{j,t-1})}$$

$$\text{Fixed weight Laspeyres:} 1+\pi_t^{L,fixed} = \sum_i s_{i,1} \cdot \frac{p_{i,t}}{p_{i,t-1}} \tag{2.6}$$

$$\text{Fixed weight Paasche:} 1+\pi_t^{P,fixed} = \sum_i s_{i,T} \cdot \frac{p_{i,t}}{p_{i,t-1}} \tag{2.7}$$

$$\text{Fixed weight Fisher} 1+\pi_t^{F,fixed} = \left(1+\pi_t^{L,fixed}\right)^{1/2} \cdot \left(1+\pi_t^{P,fixed}\right)^{1/2} \tag{2.8}$$

where $p_{i,t}$ is the price of good i in period t, $s_{i,t}$ is the expenditure share on good i in period t, and T is the last time period considered in a sample. The most intuitive price index is arguably the Laspeyres index, which is a weighted average of the price ratios of all goods between two consecutive time periods, with the weights being the expenditure shares of the goods in the previous period. While intuitive, this price index tends to overstate the true change in cost of living. This is because all else equal, as long as goods display downward sloping demand curves, increase in prices in period t will lead to lower expenditure share on this good in period t, so that the weight set in $t-1$, $s_{i,t-1}$ should actually be lower. Therefore, π_t^L overstates the true cost of living and is subject to substitution bias. The Paasche price index is similar to the Laspeyres index but uses contemporeanous expenditure shares as weights. Relatedly, it suffers from the opposite problem as it tends to understate the true change in cost of living, since its weights already incorporate the substitution of consumers towards less expensive goods. The Fisher index is the geometric average of the Laspeyres and the Paasche price indices. This index is termed an ideal price index as it satisfies all of tests for desirable proporties (homogeneity, invariance, symmetry, monotonicity, etc.)

laid out in the literature on axiomatic price theory (see for instance ILO et al. (2004)). In addition, the Fisher index, just as the Tornqvist index are superlative indices, in the sense that they are a second-order approximation to any true cost of living $\frac{e(\bar{u},p_1)}{e(\bar{u},p_0)}$ index for specific families of utility functions, where $e(\bar{u}, p_1)$ is the minimum expenditure necessary for a consumer to achieve utility level $\bar{u}$ under price vector p_1 (Diewert, 1976). Similarly, the constant elasticity of substitution (CES) price index exactly coïncides with a consumer's true cost of living when her preference is of the CES form. All the price indices exposed so far are chained price indices, in the sense that it compares price changes within two consecutive periods and that the weights $s_{i,t}$ are updated every period. While this is desirable because it accounts for the substitution patterns between two time periods, chained indices are also subject to "chain drift", the fact that when prices fluctuate between several time periods, but come back to their original level, chained price indices will indicate a positive inflation, which is an undesirable property. Chain drift is particularly important with high frequency observations. This is why this paper also uses fixed weight indices, with weights set at a baseline period - the first period for the fixed Laspeyres index, the last available period for the fixed Paasche index. Fixed base indices do not suffer from chain drift (ILO et al., 2004), at the cost of using less representative expenditure weights, the more so when the base period is far away from the current period.

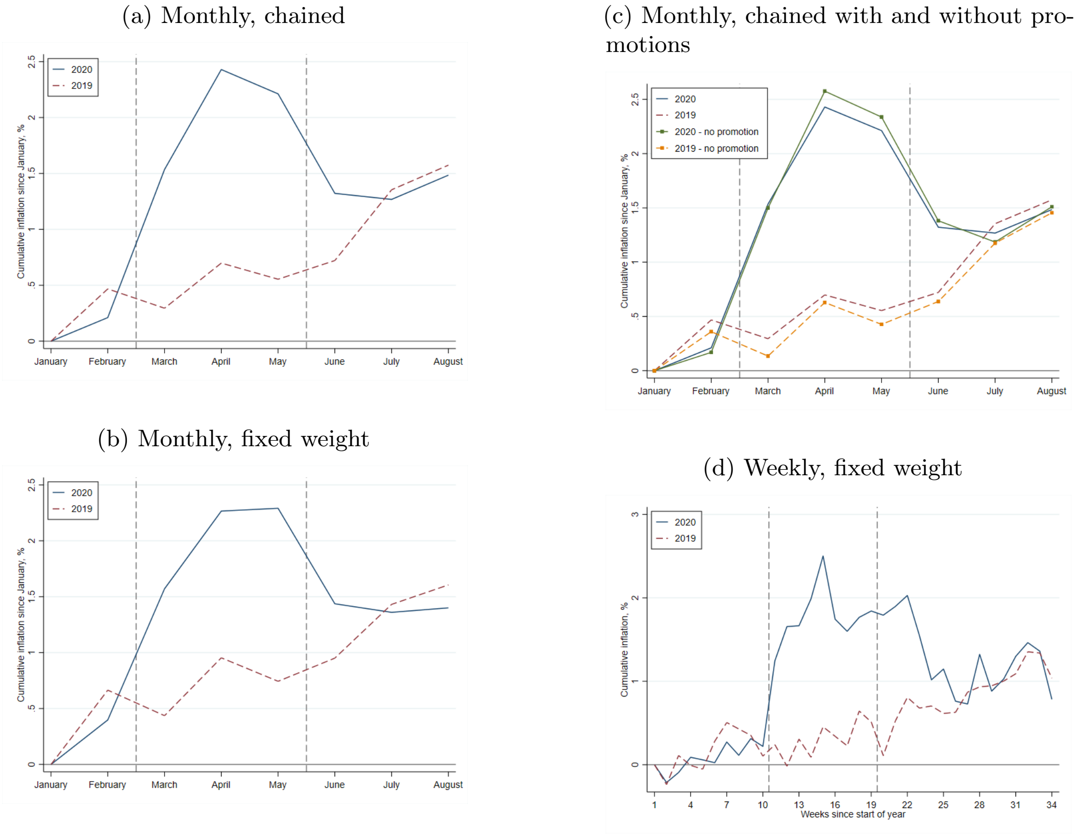
In what follows, most of the results will be presented using chained and fixed base Fisher price indices, but all results remain qualitatively similar using the other price indices exposed above. Last, it is important to note that these price indices only focus on continuing goods - goods that are present in two consecutive periods for chained indices, or in all the studied time period for fixed base indices. Accounting for entry and exit of goods is the focus of section 2.3.3.

2.3.2 Inflation for continuing products

Panel (a) of figure 2.2 displays the evolution of inflation for the first eight months of 2019 and 2020 for a chained Fisher index. In February, inflation is similar between both years. Inflation spikes from March 2020 onwards, culminating at about 2.5% in April 2020 relative to January 2020, against 0.6% in 2019. However, this inflation spike is transitory: 2020 inflation returns to 2019 levels by the month of July, at about 1.3%. Overall, inflation was higher than the previous year between March and May, mirroring the results on average unit price. Panel (b) shows that the results are quantitatively and qualitatively similar using a fixed base rather than chained Fisher index. 2020 cumulative inflation spikes at 2.3% and is back to 2019 levels from July 2020 onwards, at 1.4%. Panel (c) reproduces panel (a) but overlays a price index computed excluding all transactions with promotions. It appears that changes in promotions patterns is not a driver of the inflation hike in 2020, as both series closely tracks each other. This is one significant difference between the French and British situation, as promotions were a key driver of the inflation dynamics in the United-Kingdom

(Jaravel and O'Connell, 2020b). Interestingly, the difference between the main inflation series and the one without promotions is higher in 2019. Further, the 2019 inflation is lower when promotions are excluded, suggesting that promotions are targeted at products with smaller inflation dynamics. A higher frequency inflation series is depicted in panel (d). At a weekly frequency, one can see that inflation is very similar in 2019 and 2020 for the first 10 weeks, and then abruptly jumps to 2.5% within four weeks.

Figure 2.2: Aggregate Fisher inflation index



(a) Monthly, chained

(c) Monthly, chained with and without promotions

(b) Monthly, fixed weight

(d) Weekly, fixed weight

Notes: This figure plots different aggregate Fisher inflation indices for 2019 and 2020. Panel (a) plots a monthly chained index; panel (b) plots a monthly index with fixed weights; panel (c) plots monthly chained indices, with the "no promotion" series being constructed excluding all transactions with promotions, panel (d) plots a weekly index with fixed weights, conditioning on products present in all 34 weeks. These products represent 89 % of all sales. Dashed lines denotes the start and end of the French first lockdown.

Appendix figure 2.B.2 reproduces these results with other price indices. Panel (a) compares the CES, Tornqvist and chained Fisher indices, which all closely track each other. Panels (b) and (c) plot the fixed weight Laspeyres, Paache and Fisher indices. Consistent with theory, the Fisher index is always bounded above by the Laspeyres index and below by the Paasche index. While all three indices follow the same pattern exposed above, their levels are quite different. By the end of the sample, the Laspeyres index is 1.5 percentage point higher than the Fisher index, which is itself about 1.5 percentage point higher than the Paasche index. This difference suggests that substitution patterns are quite substantial, as equations (2.1) to (2.3) imply that all three indices are equal when the expenditure shares do not change across periods. This is further developped in section 2.3.4.

2.3.3 Accounting for product entry and exit

The patterns described above do not account for product entry and exit, as they are defined for continuing products only. Intuitively, new product entry reduces the cost of living as consumers have a love for variety, have access to a greater number of products, and can substitute towards them even if the price of continuing goods does not change. The reverse is true when products exit the market. Inflation indices not accounting for entry and exit are therefore biased. It can be shown (see for instance Broda and Weinstein (2010); Jaravel (2019)) that under CES preferences, the true price index $\tilde{\pi}_t^{CES}$, accounting for product entry and exit, is related to π_t^{CES} from equation (2.5) as follows:

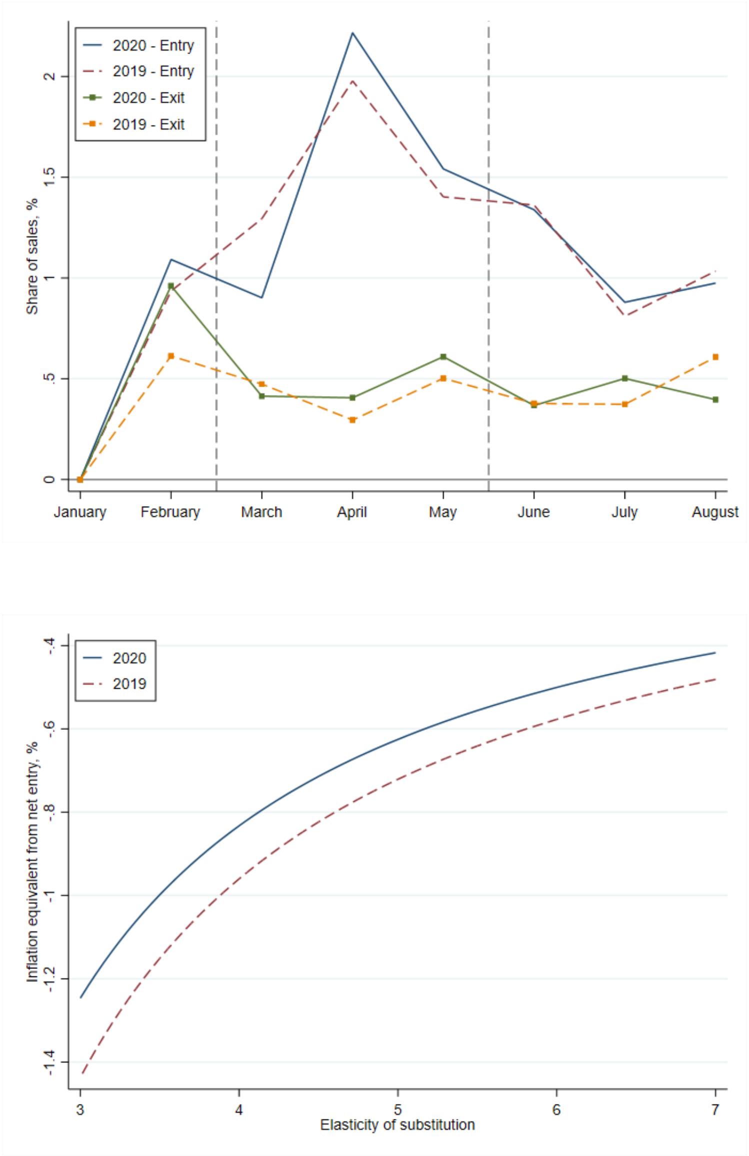
$$1 + \tilde{\pi}_t^{CES} = \left(1 + \pi_t^{CES}\right) \cdot \underbrace{\left(\frac{1 - ne_t}{1 - nx_{t-1}}\right)^{\frac{1}{\sigma - 1}}}_{Inflation\, correction\, term} \tag{2.9}$$

where ne_t is the share of expenditure in period t spent on entering goods, that are present in period t but not in period $t-1$, nx_{t-1} is the share of expenditure in period $t-1$ spent on exiting goods, that were present in period $t-1$ but are not present in period t anymore, and $\sigma > 1$ is the elasticity of substitution between goods. Intuitively, when the entry share is larger than the exit share, the inflation correction term is smaller than one, so that the true cost of living is smaller than the inflation accounting for continuing products only. Similarly, then σ is high, the inflation correction term tends towards one and $\tilde{\pi}_t^{CES} \approx \pi_t^{CES}$: this is because goods are very good substitutes to one another, so consummers can already easily substitute expensive goods with cheaper goods and entry or exit of goods is not useful in this regard.

Panel (a) of figure 2.3 reveals that the patterns of entry and exit of products in 2019 and 2020 are very similar. The share of sales from entering products varies from 1% to 2% of monthly sales both in 2019 and 2020, while the share of sales from exiting products is stable at around 0.5% of sales across months. Interestingly, the exit share of products in February 2020 is quite a bit larger than in February 2019, at 1% against 0.5%. According to discussion with managers at the private retailer, this is arguably unrelated to the pandemic as February is the month where the annual negociations with the suppliers take place and where a unusual entry or exit figures can be seen depending on the years. Panel (b) plots the inflation correction term from equation (2.9), expressed as percentage point of inflation equivalent for different values of the substitution elasticity σ, within the generally accepted range of $[3; 7]$ (see DellaVigna and Gentzkow (2019); Broda and Weinstein (2010)). The

Figure 2.3: Net entry inflation correction

(a) Entry and exit



(b) Inflation correction term

Notes: This figure provides information on net entry of products and its impact on inflation. Panel (a) plots the entry and exit of products in 2019 and 2020, expressed as a share of sales in a given month. Panel (b) plots the inflation correction term for different values of the elasticity of substitution, as exposed in section 2.3.3.

inflation correction term implies a lower inflation than using π_t^{CES} because net entry of goods is positive both in 2020 and 2019. Also, the correction term is smaller in absolute value the higher the elasticity of substitution, as could be expected from equation (2.9). The 2020 inflation correction is between 0.2 and 0.1 percentage point smaller than the 2019 inflation correction term, depending on the elasticity value. While this suggests that net entry was smaller in 2020 than in 2019, the difference is twice as small as in the United-Kingdom

(Jaravel and O'Connell, 2020b). Overall, this figure implies that product entry and exit did not significantly impact the inflation measurement from continuing products during lockdown.

2.3.4 Change in expenditure patterns and substitution bias

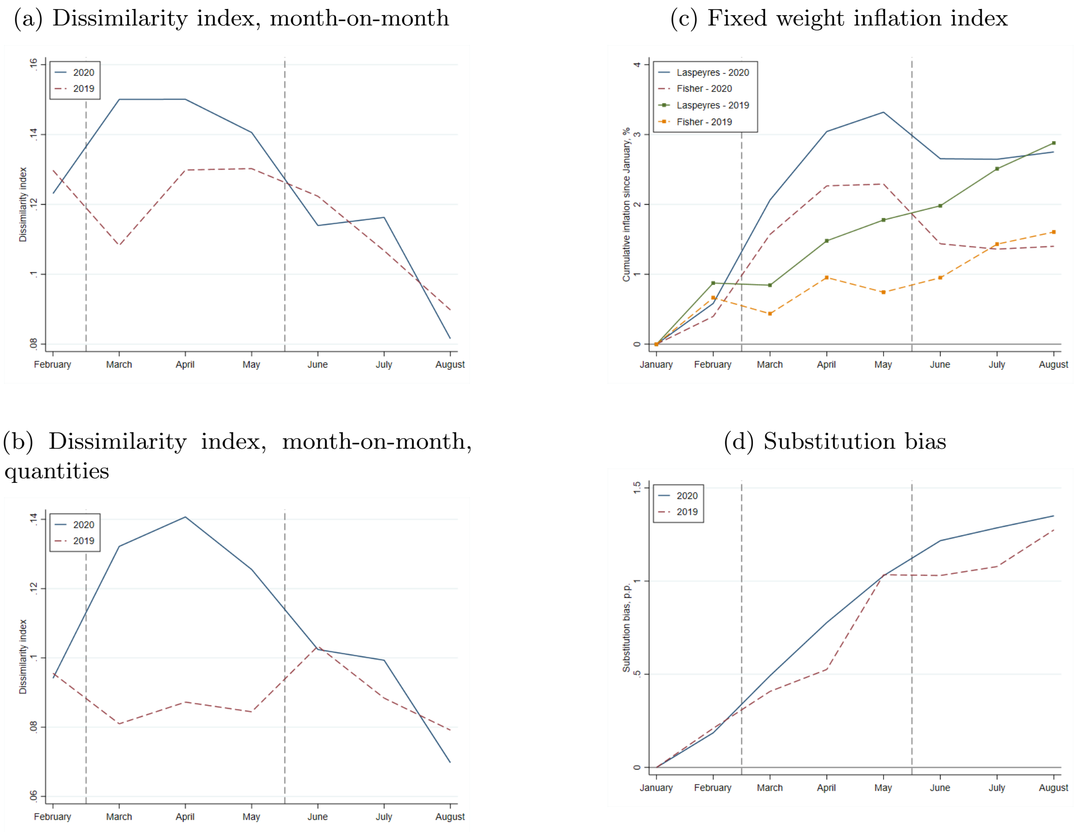
Entry and exit of goods is not the only source of bias in inflation figures reported by statistical agencies. In rapidly evolving situations such as the Covid-19 pandemic and lockdown, spending patterns are likely to change drastically. Because statistical agencies usually update their weights once every year only, the reported figure can be subject to substitution bias as exposed in section 2.3.1, where inflation figures over- or under-state inflation because the expenditure weights are out of date. Figure 2.4 explores this hypothesis. Panel (a) and (b) explore how consumption patterns changed over 2019 and 2020 by plotting a UPC-level dissimilarity index:

$$D_{t,t-1} = \frac{1}{2} \sum_{i=1} |s_{i,t} - s_{i,t-1}| \tag{2.10}$$

where $s_{i,t}$ is the expenditure (or quantity) share of UPC i in period t. $D_{t,t-1}$ has an intutitive interpretation: it is the percentage of expenditure one should reallocate across UPCs in period t in order to match the expenditure distribution of UPCs in period $t-1$. Panel (a) plots month-over-month D_t for 2019 and 2020. In February of both years, between 12.5% and 13% of sales had to be reallocated across UPCs so as to match the expenditure distribution across UPCs from the previous month. This number jumps by 25%, to 15%, in March 2020 and remains at this level in April 2020, whereas it was between 11% and 13% in 2019. This suggests higher-than-usual changes in expenditures due to the exceptionnal aspect of the lockdown. From the month of May onwards, the gap between the 2019 and 2020 dissimilarity indices narrows progressively. Panel (b) plots the same dissimilarity index, but computed for quantities rather than expenditures in order to control for the fact that a high dissimilarity index could be in part caused by differential inflation across goods, even if quantities sold do not change. The quantity dissimilarity index is similar in February of both years, at approximately 9.5%. However, it jumps at 13.5% and 14% in March and April 2020 respectively, whereas it stayed at around 8.5% in 2019. Both the jump between February and April 2020, as well as the percentage point difference between March - May 2020 and March - May 2019 are more significant for the quantity dissimilarity index rather than the expenditure dissimilarity index. One possible explanation, which is not necessarily the only one, is that the goods experiencing important decrease in quantity demanded are also the ones experiencing higher inflation so that the overall change in market share expenditure on these goods is smaller than the quantity market share.

Panels (c) and (d) test whether this higher-than-usual changes in consumption patterns

Figure 2.4: Dissimilarity indices and inflation bias


(a) Dissimilarity index, month-on-month

(c) Fixed weight inflation index

(b) Dissimilarity index, month-on-month, quantities

(d) Substitution bias

Notes: Dissimilarity index is computed over products present in all months, following the formula in equation 2.10. Dashed lines denotes the start and end of the French first lockdown.

translates into increased substitution bias. Following Jaravel and O'Connell (2020b), substitution bias is expressed as the percentage point difference between a fixed-weight Laspeyres index and a fixed-weight Fisher index. Equations (2.1) to (2.3) imply that this bias is zero when $s_{i,1} = s_{i,T} \, \forall i$. Panel (c) plots the fixed base inflation indices for 2020 and 2019, whereas panel (d) plots the substitution bias for both years. It is clear that the substitution bias increases over time, as the $s_{i,1}$ weights used in the fixed base Laspeyres index become further away from the true expenditure-weights $s_{i,t}$ as t grows, while at the same time $s_{i,T}$ become closer to $s_{i,t}$. By the month of August, the substitution bias amounts to about 1.3 percentage points in both years. However, there is no obvious difference in the dynamics of substitution bias between 2020 and 2019, as both series closely track each other. Overall, I conclude that while changes in consumption patterns are significantly more important in the midst of the lockdown than in the equivalent period in 2019, they do not translate to increased substitution bias, a result that has also been found in the United-Kingdom (Jaravel and O'Connell, 2020b).

2.4 Mechanisms behind the inflation spike

This section explores three potential sources for inflation: differential inflation dynamics by type of brand, by store types, or by in-store versus online sale status.

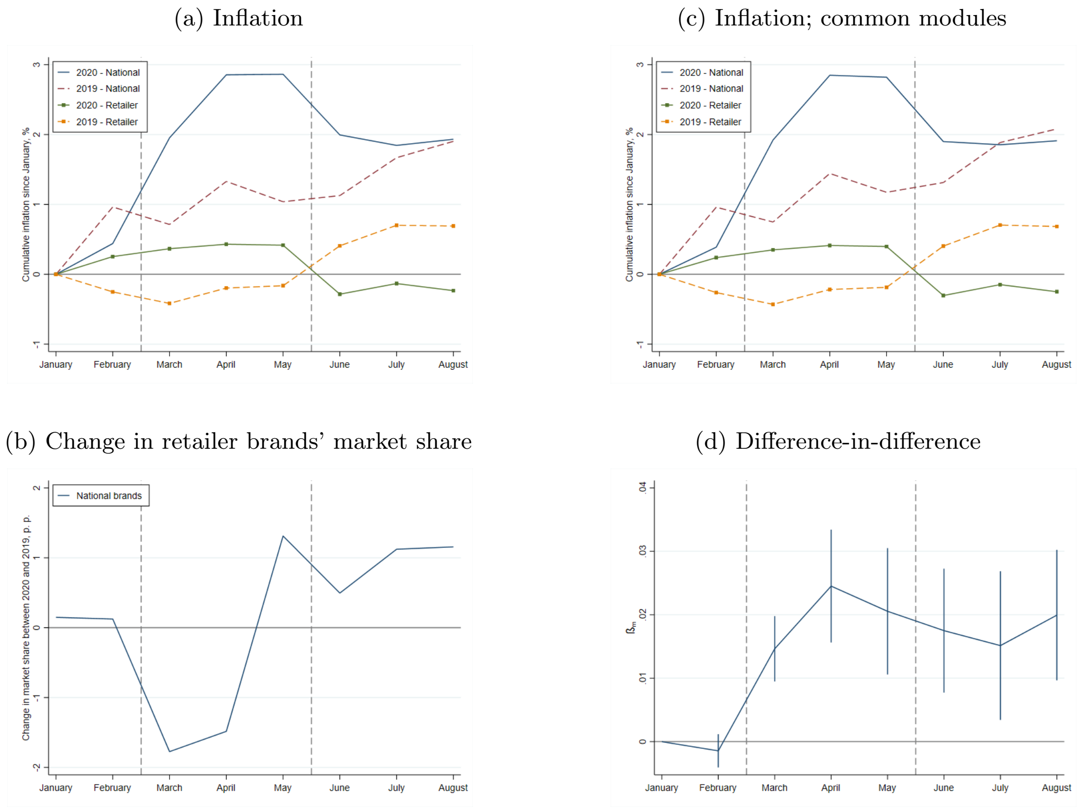
2.4.1 Brand types

This data set can distinguish between two types of brands under which UPCs are sold. National brands, which account for 77 % of overall sales in 2019, are independent brands which can be distributed both in the retailer's stores for which I have data, but also in stores of other retailers. These are often brands well known the the average consumer and are owned by large consumer packaged good companies (e.g. Procter & Gamble, Nestlé, Unilever, etc.). In contrast, private label brands are brands owned or managed by the private retailer. Therefore, the key difference with national brand is that the retailer has a greater control over the supply chain and pricing decisions for these products, as they do not depend on contractual agreement with other companies for distribution. In 2019, private label brands accounted for 23 % of overall sales in the data set.

Panel (a) of figure 2.5 plots the fixed-base Fisher inflation index for both national and private label brands in 2019 and 2020. In both years, cumulative inflation from January was relatively constant for private label brands, oscillating between -0.3% and 0.7%. In particular, there is no distinguishable spike in inflation during the lockdown, and one can even note a small deflation from June 2020 onwards relative to January 2020, whereas inflation during those months was positive in 2019. By contrast, an inflation spike is clearly visible for national brands during lockdown. While national and private label brand inflation track each other in February 2020, national brand inflation jumps to 2% in March 2002, and 2.9% in April and May 2020. For comparison, national brand inflation was approximately flat at 1% during those months in 2019, and private label brand inflation was approximately flat at 0.5% during those months in 2020. National brand inflation goes down to 2% in June 2020, and goes back to 2019 levels starting July 2020. Panel (b) plots the change in national brand's market share between 2020 and 2019. National brands' market share in the first two months of 2020 is very similar to the 2019 figure. In March and April 2020 however, it drops by 1.8 and 1.6 percentage points relative to 2019 indicating a substitution away from national brands. This also suggests that brand inflation as depicted in panel (a) is slightly overstated because of substitution bias. Nonetheless, the bias is likely small given the high baseline market share in 2019. Between May and August 2020, national brands' market share is about 1 percentage point higher than in 2019, possibly indicating a return to pre-lockdown consumption habits.

Why did national brands' prices increase more than private label brands during lockdown? One possibility is a composition effect: national brands product might come from different

Figure 2.5: Inflation dynamics by brand type


(a) Inflation

(c) Inflation; common modules

(b) Change in retailer brands' market share

(d) Difference-in-difference

Notes: National and private label, retailer brands accounted for 77 % and 23 % of overall sales in 2019 respectively. Panels (a) and (c) display inflation from a fixed-base Fisher index. Panel (c) plots inflation by brand type including observations only from product modules with both national and private label brands. Panel (d) plots the coefficient β_m from equation (2.11). Vertical bars indicate the 95% confidence interval. Standard errors are clustered at the UPC level. Only UPCs observed all months in 2020 are included. Observations are weighted by expenditures. Dashed lines denotes the start and end of the French first lockdown.

product groups than private label brands, which might be on different inflation trajectories before and during lockdown. Panel (c) of figure 2.5 suggests that this is not the case. Inflation dynamics across brand types are virtually unchanged when we exclude product modules with only national or only private label brands.

$$\Delta \ln p_{it} = \alpha + \delta_{subm(i)} + \mu_t + \sum_{m=1}^{8} \beta_m \cdot NatBrand_i \cdot \mathbb{I}\left[t = m\right] + \epsilon_{it} \qquad (2.11)$$

Panel (d) exploits a difference-in-difference setting to compare more closely the price dynamics of national and private label brands. It plots the coefficient β_m from equation (2.11), where $\Delta \ln p_{it}$ is the difference in log price of product i between time t and January 2020, $\delta_{subm(i)}$ is a sub-module fixed effect, μ_t is a month fixed effect, $NatBrand_i$ is an indicator function equals to one if product i is part of a national brand and ϵ_{it} is an unobserved disturbance. β_m is as the log difference in price change for month m between national and private label brands products. For this to be the causal impact of lockdown on the price difference between both types of brands, the identification assumption that $\mathbb{E}\left(NatBrand_i \cdot \mathbb{I}\left[t = m\right] \cdot \epsilon_{it} \mid \delta_{subm(i)}, \mu_t\right) = 0$ needs to hold for $m \in [3;5]$. Intuitively, the timing of lockdown needs to be uncorrelated with unobserved determinants of price, conditional on the fixed effects, and national and private label brand should have been on parallel trends regarding price changes had the lockdown not happened. Timing of lockdown can arguably be considered as an unexpected, exogenous shock: the development of the Covid-19 pandemic in early 2020 was extremely quick, and nation-wide lockdown is an extremely exceptional measure unheard of in recent decades. The coefficient β_m for February 2020 is statistically insignificant, suggesting that the parallel trends assumption holds for the month prior the lockdown. To gain additional insight, appendix figure 2.B.3 plots the regression result of equation (2.11) for 2019. Importantly, between February and August 2019, the coefficients β_m are statistically indistinguishable from each other. This suggests that from February 2019 onwards, the parallel trend assumption holds relatively well. It is also important to note that the statistically significant coefficient for February 2019 is due to a very specific event: as of 1st of February 2019, retailers had to apply a minimum gross margin of 10% to all their food products. This measure has been voted in late 2018 as part of French bill 'Egalim' in an attempt to regulate the price on loss-leader products and to increase farmers' income[3]. Overall, panel (d) shows that in March 2020, prices of products from national brands grew on average 1.5% faster than private label brands within the same product submodule. This number was 2.4% in April 2020 and oscillated between 1.6% and 2% between May and August 2020.

To sum up, figure 2.5 suggests that most of the excess inflation during lockdown was driven higher inflation from national brands products, and not from composition effects between different brand types. Discussions with managers from the private retailer suggest two driving forces behind this result. First, an exceptional surge in demand for products over first few weeks of lockdown, leading to heightened competition for supplies between distributors in order to reduce the prevalence of shortages. This, in turn, lead to higher demand and prices in the wholesale market. Second, there was a clear willingness from the private retailer, out of reputation concerns, not to increase the price of private label brands, at a time of high public scrutiny where retailers were at the forefront of the economic response to the lockdown.

2.4.2 Store types

Panel (a) of figure 2.6 plots the inflation dynamics in 2020 by store type. While cumulative inflation for all store types follow the same pattern - a sharp increase in March and April,

[3] see https://agriculture.gouv.fr/egalim-comprendre-le-seuil-de-revente-perte-et-lencadrement-des-promotions, accessed 10th June 2022

then a decrease in May and June, followed by a stabilisation in July and August, the difference in level is quite sizeable. In April 2020, inflation was 2.5% for small Contact stores, versus 1.4% only for large, Hyper stores. In addition, inflation stabilized post-lockdown to between 1% and 1.5% for all store types, except for Express stores, which falls back to close to 0% in July-August. Panel (b) reproduces this analysis by focusing only on common UPCs, observed at least once in every store type in 2020. This confirms that heterogeneity of inflation across store types is driven by differences in prices for similar products, rather than by different products being sold. Panel (c) plots the change in market share between 2019 and 2020 for each store types. For all store types, the change is quite flat both before and after lockdown, indicating that any change in shopping behaviour across store types reverted back after lockdown. Small (Contact and Express) gain about 0.6 percentage point market share, from a baseline of 5.7% and 2.8% respectively. This comes primarily at the expense of large, Hyper stores which loose 0.9 percentage point market share in March-April 2020 relative to 2019. Interestingly, the correlation between inflation and market shares does not follow the standard theoretical intuition. Market share in Contact stores increased even though it is the store type with highest inflation, and the reverse is true for Hyper stores. This suggests that changes in shopping patterns across stores during lockdown where dictated by non-price factors, such as movement restrictions. Overall, this suggests that the inflation shock, while positive, varies substantially across geographies as different types of stores are mostly specialized by geographies, with Contact and Super stores being found in primarily urban areas, while Express and Hyper store are rather found in primarily suburban and rural areas (see section 2.2). Heterogeneity by location is further explored in section 2.5.2.

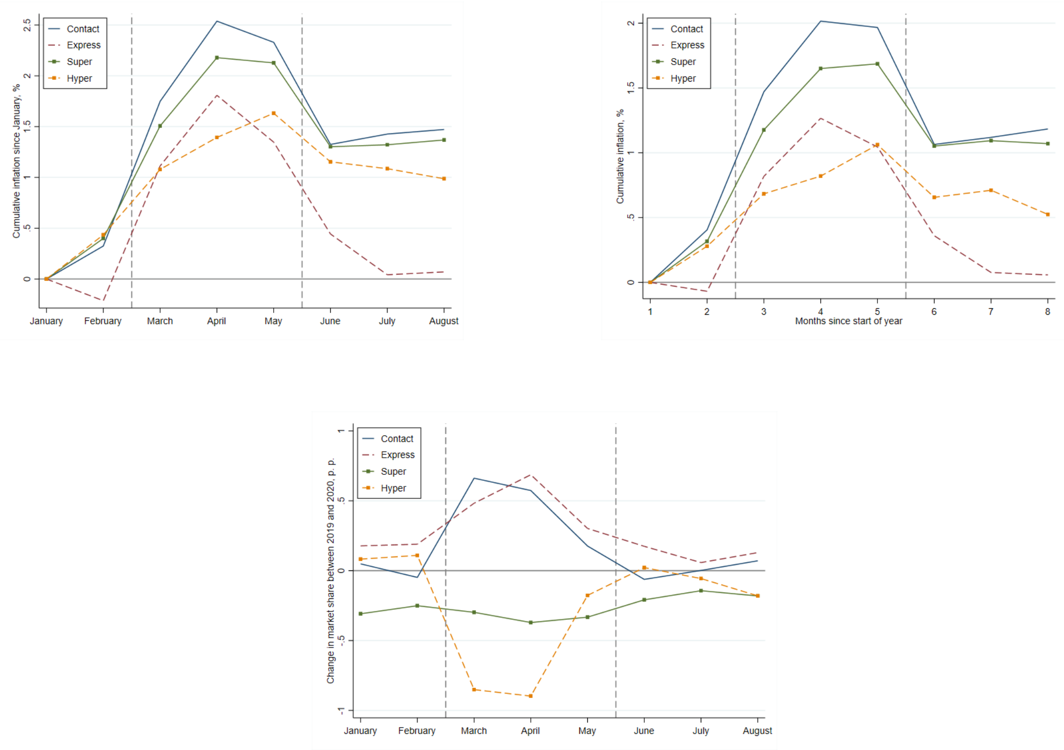
2.4.3 In-store versus online status

Panel (a) of figure 2.7 plots the inflation dynamics in 2020 by online status. In-store and online inflation numbers are similar in February, but in-store inflation spikes to 2.3% in April, much more than online inflation which increase to about 1%. However, this difference is entirely driven by a composition effect. Panel (b) shows that when considering only on UPCs being sold both online and in-store, online inflation is actually higher than in-store inflation during lockdown. In addition, the overall inflation is much lower, culminating at 0.6% for online inflation and 0.4% for in-store inflation. Finally, panel (c) plots the change in market share between 2019 and 2020 for online inflation. It documents the stark increase in online market share during lockdown relative to the same period in 2019 (+3.5 percentage points, from a baseline of 2.6%), at a time where overall spending increased dramatically as well. Overall, figure 2.7 as well as the small market size of online sales, as evidenced from table 2.1, suggest that online sale status did not play a major role in the inflation spike during lockdown.

Figure 2.6: Inflation dynamics by store type

(a) Inflation in 2020

(b) Inflation in 2020, common UPCs



(c) Market share dynamics

Notes: This figure provides information on inflation dynamics by store type in 2020. Panel (a) plot inflation from a fixed base Fisher price index. Panel (b) plots the same index but restricting the data to common UPCs across store types. Panel (c) plots for each month the difference of a given store type's market share in 2020 and 2019.

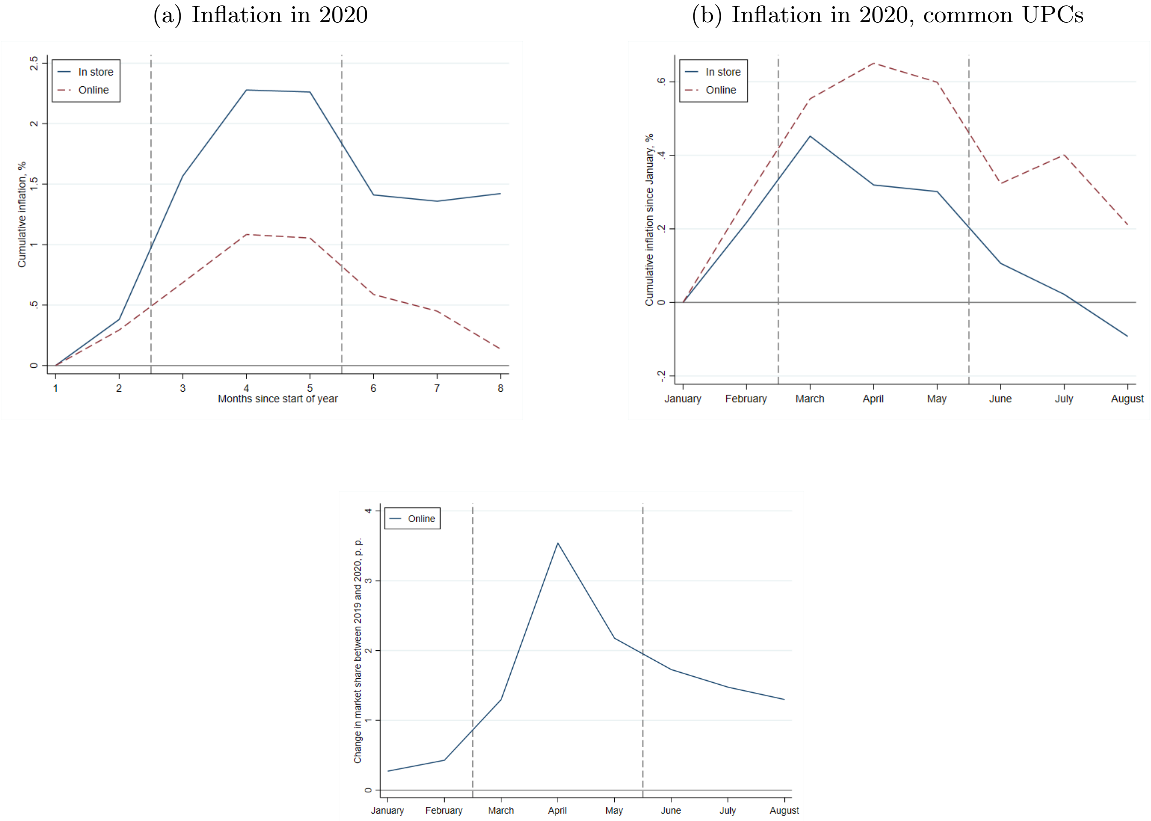
2.5 Inflation rates heterogeneity

Products, cities or households can experience drastically different inflation levels, which has important implications for economic diagnosis and for designing optimal policy to preserve purchasing power. This section therefore explores inflation heterogeneity during lockdown across these three dimensions.

2.5.1 Heterogeneity across products

Figure 2.8 provides evidence on inflation heterogeneity across product groups. Panel (a) plots the histogram of product group level inflation in April 2020 and 2019, in the midst of lockdown. The market-share weighted average inflation across product categories is 2.3 % in April 2020 and 0.7 % in April 2019, a three-fold increase. It makes clear that the magnitude of the shock is very important. However, this hides important disparities across product groups. The median inflation across product groups is much more similar in 2020 and 2019, at 0.4 % and 0.6 % respectively.

Figure 2.7: Inflation dynamics in-store and online



(a) Inflation in 2020

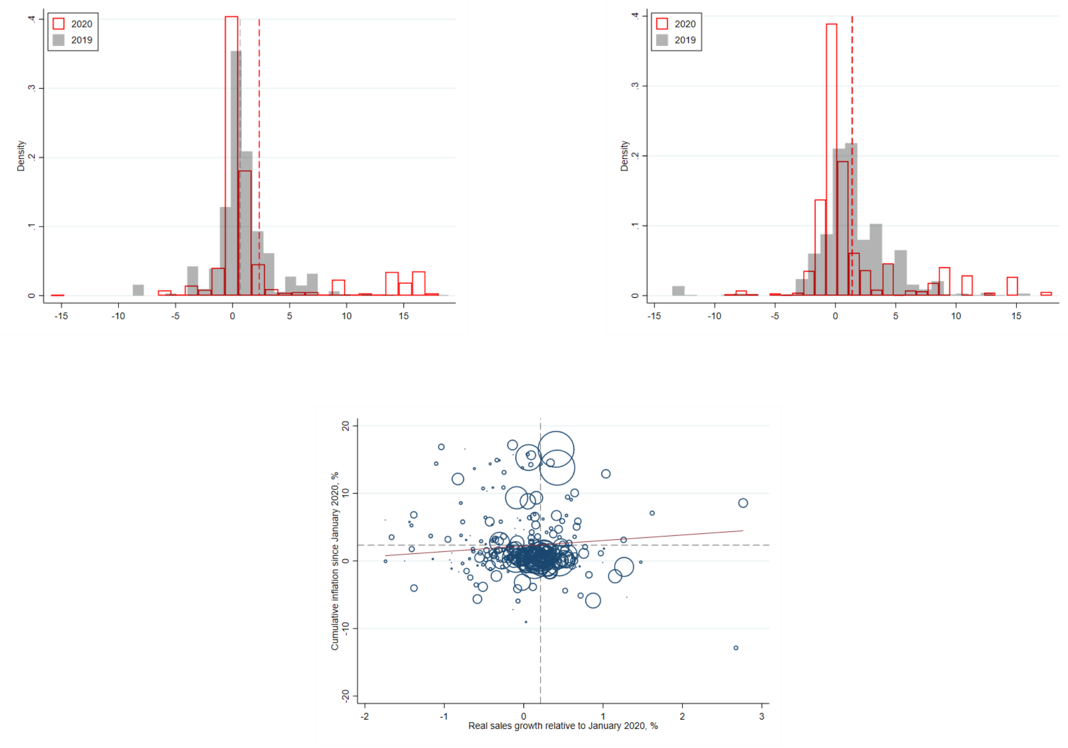
(b) Inflation in 2020, common UPCs

(c) Market share

Note: This figure provides information on inflation dynamics by online status in 2020. Panel (a) plot inflation from a fixed base Fisher price index. Panel (b) plots the same index but restricting the data to common UPCs across online status. Panel (c) plots the difference of online sales' market share in 2020 and 2019.

Panel (a) also shows that the modal inflation in 2020 is even slightly smaller than a year before. Most of the inflation is concentrated in a few categories, as can be seen from the long right tail of the inflation distribution. Strikingly, there is a mass of product groups facing inflation of 12.5% or above, and very few product categories with inflation between 2.5% and 7.5%, as opposed to 2019. More precisely, 11.8 % of product modules experience an inflation higher than 7.5% in April 2020, against 8.1 % in April 2019. In addition, looking at the third moment of the distribution enables us to quantify in a more aggregate manner the extent of asymmetry. Skewness was 1.8 in April 2020, against -3.7 in 2019, suggesting it has a significant right-tail in 2020 and a left-tail in 2019. Panel (b) plots the histogram of product group level inflation in August 2020 and 2019. Mean inflation in 2020 was 1.4 %, and has returned of 2019 level (1.3 %). However, the right-tail of the 2020 distribution, even though it has reduced, is still very much visible, and skewness in August 2020 was 1.8 , unchanged from April. Furthermore, there are fewer product categories with inflation of 12.5% and above, but more categories in the [2.5%; 7.5%] range for 2020. Panel (c) plots the

Figure 2.8: Inflation heterogeneity by product groups



(a) Cumulative inflation in April, %

(b) Cumulative inflation in August, %

(c) Inflation and real sale growth, April 2020

Notes: For panels (a) and (b), each of the 322 product groups is weighted by its market share year in the given year. Inflation is computed using a fixed base Fisher price index. Top and bottom 1% of observations are winsorised. Dashed red and gray lines denote the market-share weighted average inflation in 2020 and 2019 respectively. Panel (c) is a scatter plot of inflation and real sale growth across product categories in April 2020. Inflation and real sale growth are computed using a fixed base Fisher price index. Top and bottom 1% of observations are excluded. Dashed vertical and horizontal lines denote the market-share weighted average inflation and real spending growth. The size of the circles is proportional to the product category's market share in 2019. The red line plots the linear fit between the two variables, and the relationship is insignificant.

correlation between inflation and real sales growth in April 2020. The figure does not display any significant relationship between these two variables. The bulk of product categories display moderate sales growth and inflation. While the majority of product categories displaying significant inflation also experienced negative real sales growth, their overall market share is quite small and hence does not drive a negative relationship between the two variables. Finally, appendix tables 2.A.1 and 2.A.2 list the 30 product groups with highest and lowest inflation respectively. Most of the top 30 product categories prices increased by 10% or more, for instance fresh vegetables (+16.5%), beef (+15.3%), fresh fruits (+13.8%), but also women clothes (+12.14%). Similarly, all products with lowest inflation experienced negative inflation, for instance garden plants (−12.9%), make-up (−3.85%) or champagne (−2.22%).

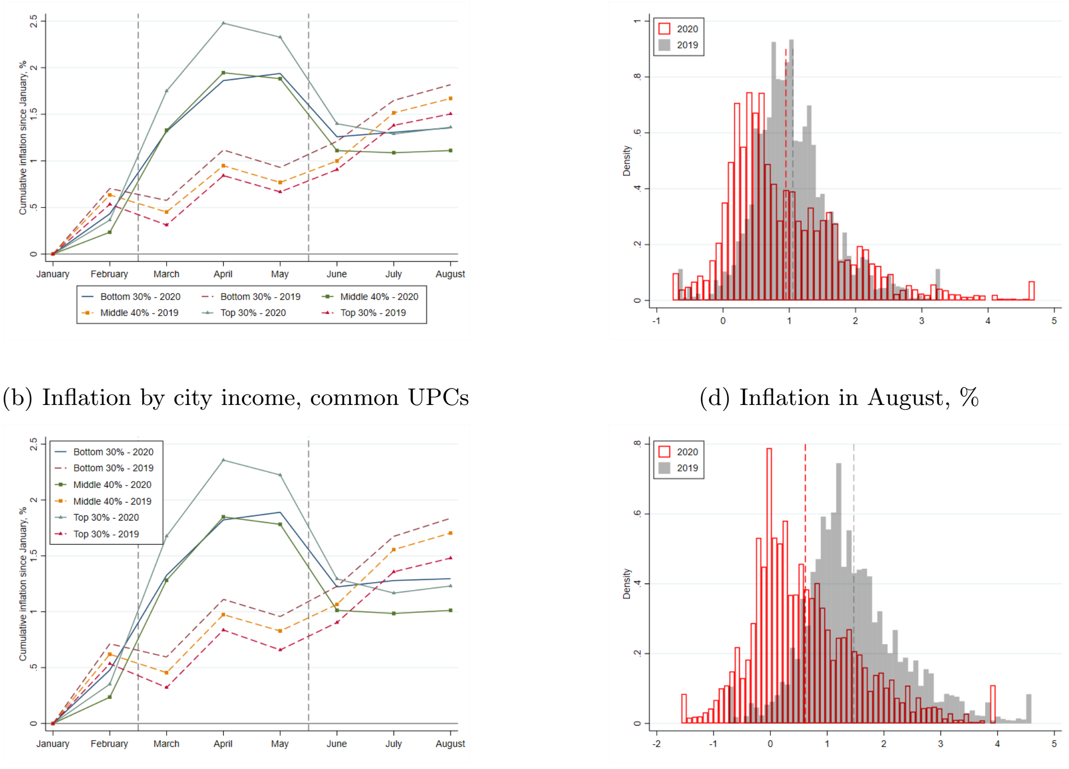
Overall, figure 2.8 suggests that the long right-tail of the distribution has been driving the bulk of the inflation during lockdown, but that the post-lockdown situation, even though similar on average to the year before, still exhibits strong asymmetry in inflation rates. This paints a different picture from what was reported in the United-Kingdom during lockdown (Jaravel and O'Connell, 2020b), where the distribution of product-level inflation shifted entirely, with almost no product categories displaying negative inflation.

2.5.2 Heterogeneity across cities

I now turn to the analysis of heterogeneous inflation across cities. Panel (a) of figure 2.9 ranks cities according to their median average taxable income and plots the dynamics of inflation in 2020 and 2019 by city terciles. The overall pattern remains consistent with what was established in section 2.3: inflation dynamics between 2019 and 2020 was similar in February, and all city terciles experienced an inflation spike in March-April 2020 where inflation was between 1.9% and 2.5%, 1 percentage point higher than the 2019 average for these months. Further, inflation returns to the 2019 levels by the end of the lockdown in June. Focusing on the differences across cities, we see that the top 30% of richest cities experienced higher inflation during lockdown than the bottom 30% or the middle 40%, which experienced similar patterns throughout 2020. The difference is sizeable: in April 2020, inflation in richer cities was 0.6 percentage points higher than in other cities, at 2.5% against 1.9%, respectively. The gap between the top and bottom 30% of cities halved after the end of the lockdown, to 0.3 percentage points. This is interesting, since in 2019, poorer cities experienced a systematically higher inflation than richer cities. The gap between the bottom and top 30% of cities was constant at 0.3 percentage points throughout 2019. Panel (b) reproduce this figure by selecting common UPCs suggesting that these results do not come from richer and poorer cities consuming different types of products with different inflation trajectories. Overall, panels (a) and (b) suggest that richer cities were on average more exposed to the inflation shock during lockdown. This finding relates to e.g. Chetty et al. (2020), which documented that richer neighborhood had a relatively higher drop in overall expenditures than poorer ones.

Panel (c), just like the histograms in figure 2.8, shows that the distribution of inflation across cities in April 2020 displayed a much longer right tail than in April 2019. More specifically, 4.9 % of cities experienced an inflation of 3% or higher in April 2020, against only 1.4 % at the same time in 2019. Skewness of the distribution in April 2020 was 1.1 , against 0.6 in April 2019. Interestingly, the center of the distribution in 2020 is shifted to the left compared to 2019, suggesting that inflation is concentrated in a small numbers of geographic areas only. This is reflected in the sales-weighted average inflation in 2020 of 0.9 %, lower than the one of 2019, at 1.1 %. How can these average inflation numbers be reconciled with the high inflation spike documented at the aggregate level? This is because a weighted average of several Fisher price indices across a given dimension e.g.

Figure 2.9: Inflation heterogeneity by city



(a) Inflation by city income

(c) Inflation in April, %

(b) Inflation by city income, common UPCs

(d) Inflation in August, %

Notes: Inflation in all figures is computed using a fixed based Fisher price index. For panels (a) and (b), city income is computed as the median average taxable income in each city. There are 1,678 different cities in the data set. Dashed lines denotes the start and end of the French first lockdown. For panels (c) and (d), each city is weighted by its share of sales in the given year. Weighting by population instead gives qualitatively similar results. Top and bottom 1% of observations are winsorised. Dashed red and gray lines denote the market-share weighted average inflation in 2020 and 2019 respectively.

a city does not aggregate to the aggregate Fisher price index. In particular, it has been shown that superlative index numbers, including the Fisher index, are not consistent in aggregation, especially so when one uses a fixed base rather than a chained price index (Diewert, 1978; ILO et al., 2004). Moreover, the property of "approximate consistency in aggregation" established in Diewert (1978) only applies around a point where the vector of prices and quantities are equal across time period (ILO et al., 2004), an assumption that is arguably largely not satisfied when comparing time periods before and during lockdown. Panel (d) plots the same histogram as panel (c), but for the months of August. We can note that the 2020 distribution shifted left compared to April 2020. However, the skewness of the August 2020 distribution remains the same as the April 2020 distribution, at 1.1 , suggesting that despite the average decrease in inflation, inflation heterogeneity across cities remained high after lockdown and higher than 2019. This is similar to what is observed for the distribution of inflation across product categories.

Finally, appendix tables 2.A.3 and 2.A.4 list the 30 cities with highest and lowest inflation in April 2020, respectively. There does not seem to be a particular geographic pattern: the city with highest inflation is Villebon-sur-Yvette (+6.17%), south of Paris, the second is Bouzonville (+5.81%) in east of France, and the third highest is Jullian (+5.63%), in south of France. More systematic analysis is conducted in the next section.

2.5.3 Heterogeneity across households

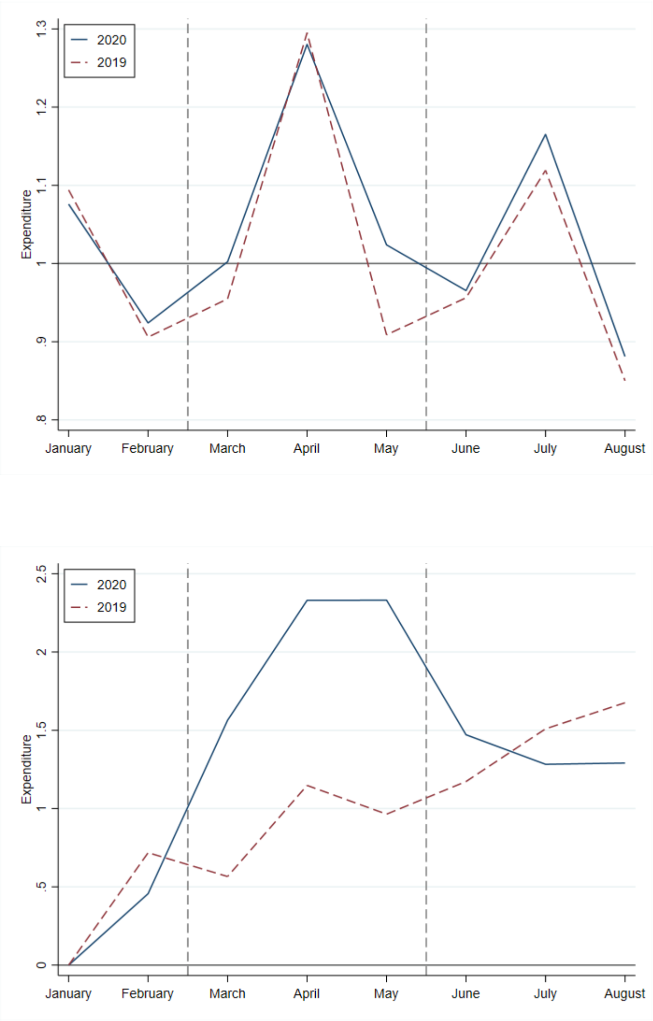
Data selection and price index construction. The particular structure of the data enables the construction of a panel of household-level transactions across the time period studied. In order to select household which are hopefully loyal customers so that I can capture a significant share of their groceries expenditures, only households spending at least 100€ per month for every 16 months available in the data are kept. In addition, the top and bottom 0.5% of households satisfying this restriction are excluded, in order to filter out outliers. The final panel consists 1,088,628 of households.

Panel (a) of figure 2.10 plots the average expenditure for the selected panel across 2020 and 2019. Contrary to panel (a) of figure 2.1, there is no evident spike in expenditure during lockdown, and expenditure patterns for both years closely track each other, suggesting that the selected household did not significantly change their expenditure level during the lockdown, nor did they significantly switched away from this retailer. Similarly, panel (b) reproduces the inflation patterns established in figure 2.2 for the main data set: an inflation spike at 2.4% during lockdown, much higher than the previous year, and a convergence of inflation rates between both years starting June. Further, appendix table 2.A.5 provides summary statistics on the average monthly expenditure of the households selected in the panel, by household type. Households spend on average 457€ every month, with families spending up to 514€ and 18-35 with no kids spending on average 402€ per month. Overall, this suggests that the household panel can be thought of as representative of the overall data set.

I compute household-level inflation rates using a fixed base Fisher price index with household-specific expenditure weights, but common prices, as in Jaravel and O'Connell (2020b). Common prices are defined as the average unit price at the aggregate level for a particular product in a given time period. Using common prices, as opposed to the price actually paid by the household avoids the need to condition on goods purchased by the household in every period, which is usually a small share of the overall basket. The drawback of this method is that heterogeneous inflation rates only come from differences in consumption basket and does not consider differences in actual prices paid. Nonetheless, consumption basket is arguably the most important driver of heterogeneous inflation rates: for instance, DellaVigna and Gentzkow (2019) document in the US context the close to uniform pricing

Figure 2.10: Stylised facts of the household panel

(a) Expenditures over time



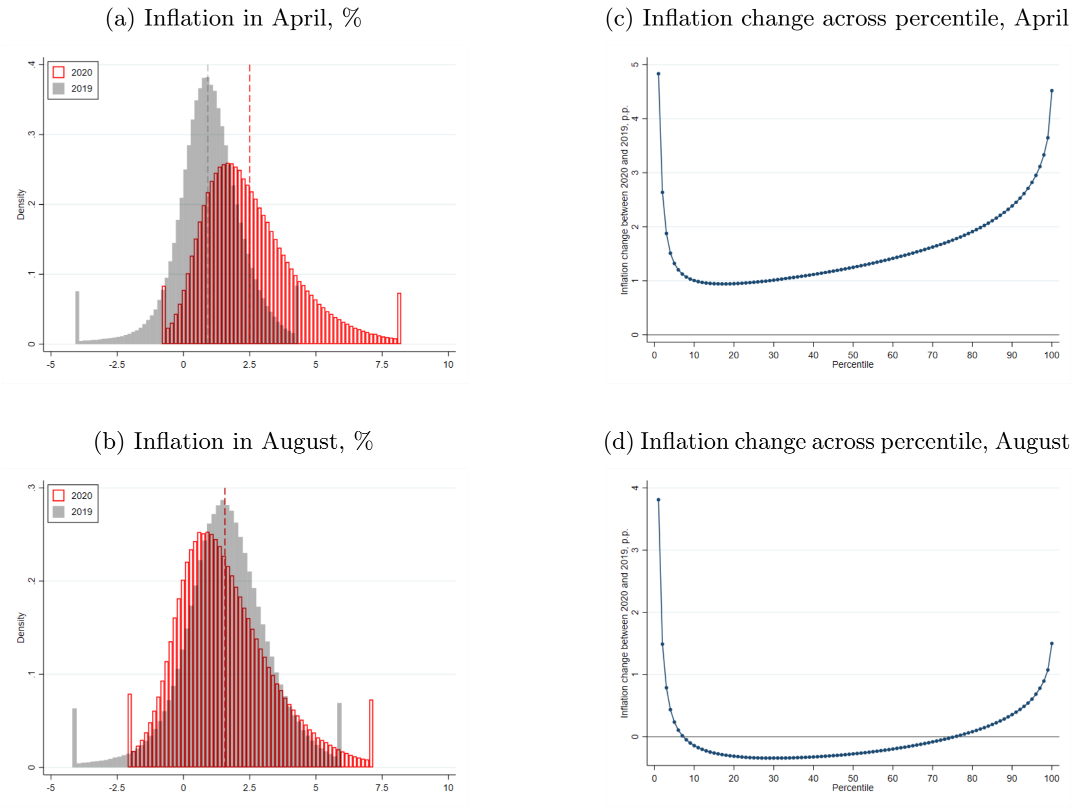
(b) Inflation

Notes: Panel (a) documents expenditures of households selected in the panel, spending more than 100€ per month every month, as described in section 2.5.3. Values are normalized to the first two months of every year. Panel (b) plots the aggregate inflation rate based on the underlying data from the selected panel. Inflation is computed using a fixed based Fisher price index. Dashed lines denotes the start and end of the French first lockdown.

of goods across stores of a given retailer.

Results. Panel (a) of figure 2.11 plots the distribution of these inflation rates for April 2020 and 2019 and clearly shows the shift to the right of the distribution in 2020, arguably caused by the lockdown. Mean inflation rate in 2020, at 2.5 %, is more than twice as high as its 2019 counterpart (0.9 %). Going beyond the first moment of the distribution, one can note the important right-tail in 2020 (skewness of 1.0), which in not present in 2019 (skewness of -1.8). Furthermore, a significant number of households experienced very high inflation rates: in April 2020, the share of households experiencing an inflation rate of 5%

Figure 2.11: Inflation heterogeneity by household


(a) Inflation in April, %

(c) Inflation change across percentile, April

(b) Inflation in August, %

(d) Inflation change across percentile, August

Notes: Inflation in all figures is computed using a fixed based Fisher price index. For panels (a) and (b), top and bottom 1% of observations are winsorised. There are 1,088,628 selected households. Dashed red and gray lines denote the average inflation in 2020 and 2019 respectively. Panels (c) and (c) plot the inflation change in percentage points across percentile between 2020 and 2019, for the months of April and August respectively.

or higher was 9.4 %, whereas this number was only 0.5 % in 2019. Similarly, the share of households experiencing negative inflation rate dropped to 4.0 % in 2020, down from 18.3 % in 2019. Appendix figure 2.B.4 reproduces panel (a) for different household types and shows that the overall qualitative results remain unchanged, even though inflation distributions for 61+ and 36-60 with no kids households display the longest right tails. Differential inflation by household type is further explored in table 2.2.

Panel (b) plots the same histogram for the month of August. The 2020 inflation distribution shifted back to the left, close to its 2019 level. Mean inflation across households is now the same in 2020 and 2019, at 1.6 %. In the same vein, 18.3 % of households faced negative inflation in August 2020, a fourfold increase compared to April 2020, and an even larger share than in 2019 (13.6 %). Importantly, the right-tail of the distribution reduced but did not disappear. Skewness is at almost the same level as in April, at 0.9 . In addition, the share of individuals experiencing inflation rates of at least 5% in August is still almost twice as high as in 2019, at 4.9 % against 2.6 %.

Panels (c) and (d) document the change in the inflation distribution from another angle. For each percentile, they plot for percentage point difference in mean inflation rates between years for the months of April and August respectively. Unreported results suggest that the results are unchanged when ploting the change in median inflation rates within percentiles. Panel (c) confirms the important overall shift in the inflation distribution in April 2020 relative to April 2019. Across all percentiles, inflation is at least one percentage point higher in April 2020 than in April 2019. Furthermore, it sheds light the compressed left tail of the distribution, as the first four percentiles experienced a higher change in inflation that the mean change of $2.5 - 0.9 = 1.6$ percentage points. Similarly, if confirms the significant increase in the left tail of the distribution: the top 30 percentiles experienced a higher change in inflation that the mean change. Interestingly, panel (d) shows that the bulk of the distribution (situated in the middle of the distribution, between percentiles 7 and 76) faced lower inflation than in August 2020 than in August 2019. By contrast, the left tail of the distribution remains compressed compared to 2019, as the first seven percentiles experienced higher inflation change than the mean inflation change in August of $1.6 - 1.6 = 0$. Last, it also confirms that the inflation distribution across households in August 2020 distribution is still much more skewed to the right than in 2019, as the top 24 percentiles experienced positive change a higher inflation than in 2019 and higher inflation change than the mean change.

All in all, what factors predict being exposed to high inflation rates in 2020? To answer this question, table 2.2 presents the results of an OLS regression of household-level inflation rate in April 2020 on a number of explanatory variables. 2019 exposure to future high inflation products, computed as the share of expenditure of the 2019 expenditures spent on the top 20% of product modules with highest inflation in April 2020, is a strong predictor of future inflation. A one standard deviation increase in this exposure measure (≈ 10 p.p.) is associated with an increase in 2020 inflation rate of 0.9 percentage points, or 36% of the mean inflation rate in April 2020. Interestingly, a measure of the geographic wealth, the log of average taxable income of the household's city of residence, is not significantly related with higher inflation rates, contrary to what has been established at the city level. Unreported results show that this is uniquely due to the household-level exposure measure being included in the regression. Rank in the 2019 expenditure distribution and inflation in 2020 are significantly negatively related, even controlling for household status. Even though expenditure is a poor proxy for household income, this relates to Chetty et al. (2020) finding that high-income consumers were most hit by the lockdown shock. However, the size of relationship relatively modest with the top quartile of households having on average a 0.11 percentage point lower inflation rate than the bottom quartile. Households having experienced higher inflation in 2019 are also the ones experiencing higher inflation in 2020;

Table 2.2: Predictors of inflation in April 2020

	Infl. April 2020	
Exposure in 2019 to high inflation products	0.093***	(0.001)
ln average taxable income	0.057	(0.066)
Quartile of expenditure 2019:		
- Q1	0.000	(.)
- Q2	-0.033***	(0.005)
- Q3	-0.052***	(0.006)
- Q4	-0.101***	(0.007)
Quartile of inflation 2019:		
- Q1	0.000	(.)
- Q2	0.070***	(0.007)
- Q3	0.198***	(0.010)
- Q4	0.434***	(0.016)
Household type:		
- Families	-0.057***	(0.005)
- 18-35 no kids	-0.177***	(0.011)
- 36-60 no kids	0.000	(.)
- 61+	0.144***	(0.007)
Store type:		
- Contact	0.166***	(0.045)
- Express	-0.223***	(0.066)
- Super	0.000	(.)
- Hyper	-0.351***	(0.045)
Region:		
- Ile De France	0.000	(.)
- Centre Val De Loire	-0.347***	(0.092)
- Bourgogne Franche Comte	-0.165*	(0.087)
- Normandie	-0.159*	(0.083)
- Hauts De France	-0.177**	(0.081)
- Grand Est	-0.085	(0.087)
- Pays De La Loire	-0.467***	(0.093)
- Bretagne	-0.357***	(0.074)
- Nouvelle Aquitaine	-0.410***	(0.073)
- Occitanie	-0.146**	(0.074)
- Auvergne Rhone Alpes	-0.171**	(0.070)
- Provence Alpes Cote D Azur	-0.093	(0.078)
Constant	-0.732	(0.686)
R^2	0.308	
N	1,088,319	

Notes: This table presents the regression result of inflation in April 2020 on a number of explanatory variables at the household level. Exposure in 2019 to high inflation product is constructed for each household as the 2019 share of expenditure on the top 20% of product modules with highest inflation in April 2020. Household's city is determined as the city from which households spend most in 2019, which is assumed to be the city of residence. Average taxable income refers to the average taxable income of the city in which a given household lives. Quartiles of inflation in 2019 is computed using the April 2019 inflation number. All inflation figures are computed using a fixed base Fisher price index using the process exposed in section 2.5.3. Variables with no standard errors are reference categories. Standard errors are clustered at the city level.

the top 25% of households having experienced most inflation in 2019 also faced a 0.43 percentage point higher inflation in 2020 than the bottom quartile, or 17% of the mean inflation. In addition, younger households, as well as families, tend to face lower inflation in 2020 than older households. In accordance to what was established in section 2.4, inflation differs by store type. Households shopping in stores located in primarily rural or suburban areas ('Express' and 'Hyper' stores) faced significantly higher inflation than individuals shopping in stores located primarily in urban areas ('Contact' and 'Super' stores). The inflation differential between an average shoper in a Contact store and an average shoper in a Hyper store is a meaningful 0.52 percentage points, or 20% of the mean inflation rate. Last, there is no specific difference between regions, other than the Greater Paris area ("Ile De France" region) facing a significantly higher inflation than most other regions.

2.6 Conclusion

This paper uses quasi-real time data on fast-moving consumer goods to document and analyze inflation during the first French lockdown. I find that there is an important yet transitory inflation spike during the months of March and April 2020, driven mostly by national brand products. This inflation shock is asymmetric, with a small number of products, cities and households experiencing significant inflation rates. Importantly, while average inflation returns rapidly to the 2019 levels, the long tail in inflation distribution created by the lockdown shock persists at least until August 2020.

Even though this paper focuses on a very specific period, I believe that two broader lessons can be learned, especially in a context where inflation becomes a renewed concern both in the United-States and in Europe. First, it is important to delve deeper than aggregate inflation numbers and to carefully analyse inflation heterogeneity along a number of dimensions. As it has been shown, inflation shocks can be very heterogeneous, and this heterogeneity does not transpire in the usually reported statistics. Accounting for this heterogeneity is key to understand how inflation is perceived and experienced by different population groups, how inflation expectations can change in response, and what are the most cost-effective mitigation policies. Second, there is a lot to be gained from a more systematic use of quasi-real time data from private organisations. In periods where usual survey methods are unavailable, or when policy making requires updated information with a time lag of a couple of days, such data sources can be extremely useful. During more normal times, this type of data can effectively complement administrative data and economic surveys, to uncover new facts relevant for policy making.

Appendices

2.A Additional tables

Table 2.A.1: Top 30 product categories with highest inflation in April 2020 (in French)

	Share sales 2019 (%)	Infl. 2019 (%)	Infl. 2020 (%)
Confection Femme Grande Taille	0.00	14.04	18.51
Viande Ovine Trad	0.06	19.65	18.08
Viande Chevaline Trad	0.33	6.62	17.17
Confection Homme	0.09	14.01	16.89
Animation Textile	0.00	-1.61	16.57
Legumes Frais	3.79	-3.77	16.53
Vrac Epicerie Sucree	0.02	-0.80	15.80
Alimentation Pour Animaux	0.00	23.31	15.73
Boucherie Tr/Fe Pdv	0.24	8.57	15.66
Viande Bovine Trad	2.10	7.32	15.32
Confec Layette Fille / Mixte	0.03	8.68	14.93
Destockage Bazar	0.00	4.50	14.90
F/L Frais Emballe	0.17	3.96	14.54
Non Affecte	0.00	-19.06	14.50
Confection Enfant Fille/Mixte	0.02	22.82	14.41
Confection Layette Garcon	0.01	15.20	14.38
Gibier Boucherie Tr	0.01	10.99	14.35
Vrac Fruits Et Legumes	0.03	-1.37	14.27
Fruits Frais	3.60	-0.85	13.82
Vrac Epicerie Salee	0.01	0.16	13.80
Confection Junior Garcon	0.00	12.63	13.67
Ameublement Et Mobilier	0.04	1.18	13.12
Produits Elabores Fe Pdv	0.16	11.79	12.90
Confection Femme	0.45	6.77	12.14
Triperie / Abats Trad	0.02	9.06	10.88
Confection Junior Fille/Mixte	0.01	6.92	10.87
Lavage	0.01	11.88	10.74
Animation Sec Ls	0.00	3.26	10.33
Produits Elabores Tr	0.13	9.85	10.07
Image	0.02	4.62	9.46

Notes: Inflation is computed using a fixed base Fisher price index. Inflation in both years is measured in April and is relative to the month of January

Table 2.A.2: Bottom 30 product categories with lowest inflation in April 2020 (in French)

	Share sales 2019 (%)	Infl. 2019 (%)	Infl. 2020 (%)
Fleurs Coupees	0.16	-2.33	-15.92
Plantes D Exterieur	0.02	-2.53	-12.87
Patisserie Fraiche Trad	0.26	-2.64	-5.63
Bio Parfumerie	0.00	0.00	-5.12
Plantes D Interieur	0.07	-1.71	-4.39
Pains Ls Pdv	0.22	-2.59	-4.10
Patisserie Gel Trad	0.14	-8.51	-4.01
Maquillage	0.18	-1.81	-3.85
Viennoiserie Trad	0.28	-5.01	-3.82
Patisserie Fraiche Emb Ls Pdv	0.05	0.00	-3.54
Pains Blancs / Boulangerie Tr	0.89	-1.22	-3.15
Viennoiserie Ls Pdv	0.08	-3.04	-2.45
Parapharmacie	0.21	-0.64	-2.26
Champagnes	0.34	1.78	-2.22
Rasage Feminin Et Depilatoire	0.11	-2.08	-2.04
Piles	0.22	1.00	-1.39
Saucissons Saucisses Seches	0.54	0.51	-1.23
Cd	0.03	-0.44	-1.22
Surgeles Sucre	1.04	-0.63	-0.88
Traiteur Frais Emballe Pdv	0.03	1.35	-0.87
Dentaire	0.54	-1.11	-0.79
Sauce Salade/Jus De Citron	0.10	-1.00	-0.77
Jus Et Boissons Frais	0.32	-0.01	-0.76
Volaille Fe Ind	0.00	-2.77	-0.74
Aperitifs Sans Alcool	0.04	0.77	-0.73
Appareils Et Accessoires Photo	0.00	7.48	-0.66
Traiteur Chaud	0.02	3.75	-0.55
Chips	0.34	-0.58	-0.52
Patisserie Gel Ls Pdv	0.02	-1.55	-0.37
Confiserie De Sucre	0.44	0.39	-0.27

Notes: Inflation is computed using a fixed base Fisher price index. Inflation in both years is measured in April and is relative to the month of January

Table 2.A.3: Top 30 cities with highest inflation in April 2020

	Share sales 2019 (%)	Infl. 2019 (%)	Infl. 2020 (%)
Villebon Sur Yvette	0.06	0.55	6.17
Bouzonville	0.03	1.60	5.81
Juillan	0.09	1.56	5.63
Drulingen	0.02	1.99	5.39
Wimereux	0.02	3.19	5.38
Vieux Thann	0.04	2.17	5.37
Ensisheim	0.03	2.23	5.25
Sains En Gohelle	0.01	0.67	5.20
Fontvieille	0.02	1.90	5.13
Chambon Sur Voueize	0.02	2.10	5.06
Mery Sur Oise	0.07	2.48	4.95
La Ferte Gaucher	0.06	1.03	4.92
Lorquin	0.02	1.76	4.91
Beaulieu Sur Dordogne	0.05	2.18	4.79
Salignac Eyvigues	0.02	0.85	4.73
Lamure Sur Azergues	0.05	2.73	4.72
La Bazoche Gouet	0.02	2.57	4.70
Samer	0.01	4.33	4.55
Cherisy	0.04	1.87	4.54
Bartenheim	0.03	2.19	4.52
Massy	0.00	0.00	4.39
Ancizan	0.04	1.45	4.35
Chateauvillain	0.03	1.76	4.28
Ste Croix Aux Mines	0.02	0.55	4.27
Le Malesherbois	0.06	1.82	4.18
Nangis	0.09	1.95	4.11
Coulommiers	0.03	1.58	4.09
Tullins	0.06	2.23	4.09
Lemberg	0.02	1.48	4.07
Meximieux	0.07	1.32	3.96

Notes: Inflation is computed using a fixed base Fisher price index. Inflation in both years is measured in April and is relative to the month of January

Table 2.A.4: Bottom 30 cities with lowest inflation in April 2020

	Share sales 2019 (%)	Infl. 2019 (%)	Infl. 2020 (%)
Chazelles Sur Lyon	0.04	0.52	-6.31
La Chapelle D Abondance	0.07	0.10	-2.94
Piegut Pluviers	0.02	2.04	-1.38
Longueau	0.02	0.09	-1.29
Beon	0.04	1.34	-1.17
Colombey Les Belles	0.02	0.82	-1.15
Fontaine La Guyon	0.02	1.93	-1.05
Mirecourt	0.06	2.13	-1.05
St Lyphard	0.04	2.31	-1.03
Jouy Le Moutier	0.00	0.00	-0.98
Grigny	0.10	0.64	-0.98
St Ouen L Aumone	0.00	0.00	-0.95
Privas	0.06	0.30	-0.93
Montbrison	0.10	0.63	-0.91
St Pierre Du Perray	0.07	1.53	-0.82
Lugrin	0.10	1.37	-0.81
Boulogne Sur Gesse	0.04	0.60	-0.76
Girancourt	0.01	4.41	-0.74
Pont St Esprit	0.05	0.59	-0.69
Saujon	0.06	-1.69	-0.68
Bourg En Bresse	0.10	0.48	-0.66
St Varent	0.03	1.66	-0.62
Brou	0.06	1.63	-0.60
Canet En Roussillon	0.03	0.73	-0.60
Poitiers	0.10	1.97	-0.60
Bayonne	0.04	0.68	-0.59
Auxon	0.01	0.21	-0.58
Noyal Pontivy	0.01	0.33	-0.56
Aigurande	0.05	1.63	-0.56
Villard Sur Doron	0.03	-0.49	-0.54

Notes: Inflation is computed using a fixed base Fisher price index. Inflation in both years is measured in April and is relative to the month of January

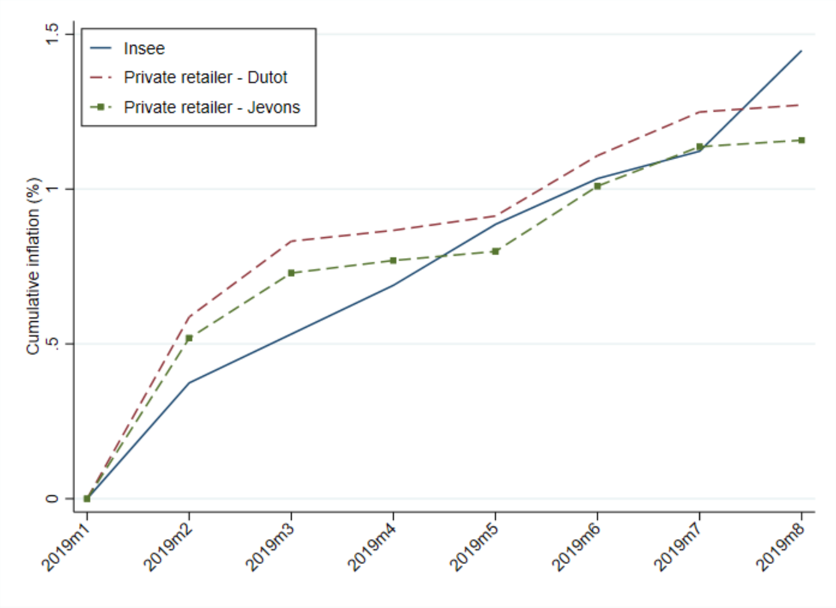
Table 2.A.5: Descriptive statistics for selected households

	Overall	Families	18-35 no kids	36-60 no kids	61+
Mean	457	514	402	460	413
Std. dev.	173	179	144	170	156
p1	198	221	194	202	191
p10	264	307	250	270	244
p25	328	380	301	334	297
Median	425	487	374	429	381
p75	552	620	471	554	493
p90	695	761	589	694	623
p99	985	1,031	898	982	915
N	1,088,628	344,887	20,920	283,156	439,621

Notes: This table presents descriptive statistics on the average monthly expenditure for the 1,088,628 households present all 16 months, and spending at least 100€ every month. All units are in euros.

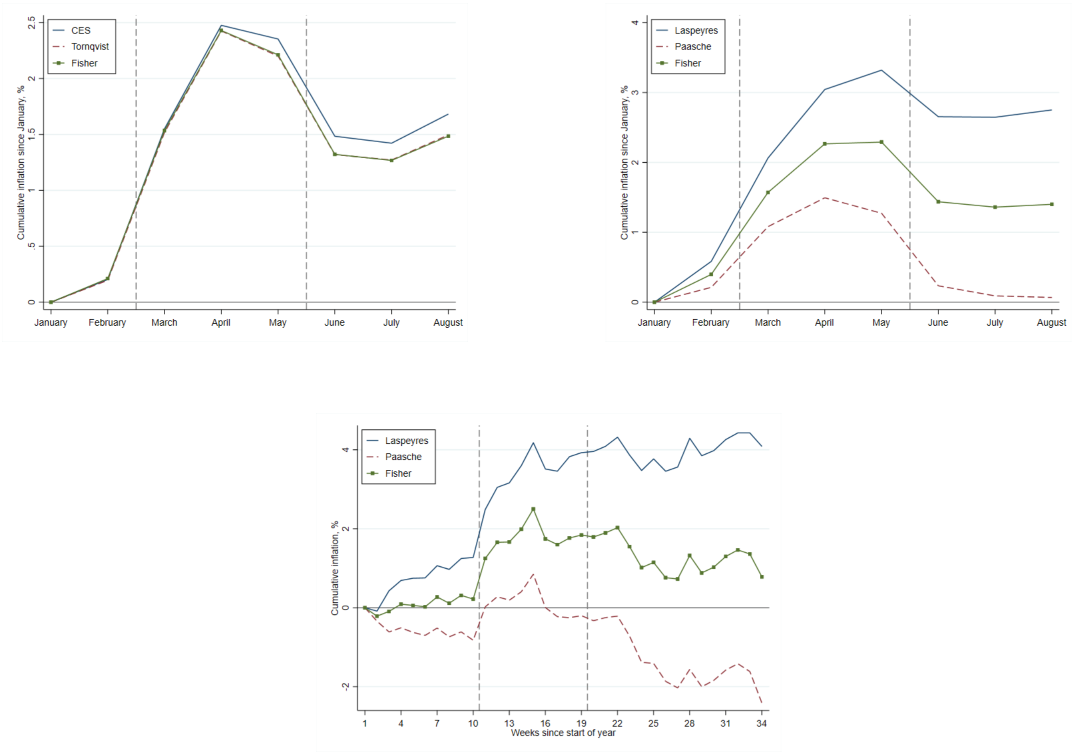
2.B Additional figures

Figure 2.B.1: Comparison between Insee's mass retail price index and private retailer data



Notes: This figure compares the publicly available mass retail price index from Insee with two similar price indices constructed from the private retailer's data and whose construction is described in section 2.2.2. The "Private retailer - Dutot" series only uses Dutot indices as elementary price indices, while the "Private retailer - Jevons" series only uses Jevons indices as elementary price indices

Figure 2.B.2: Additional inflation indices for 2020



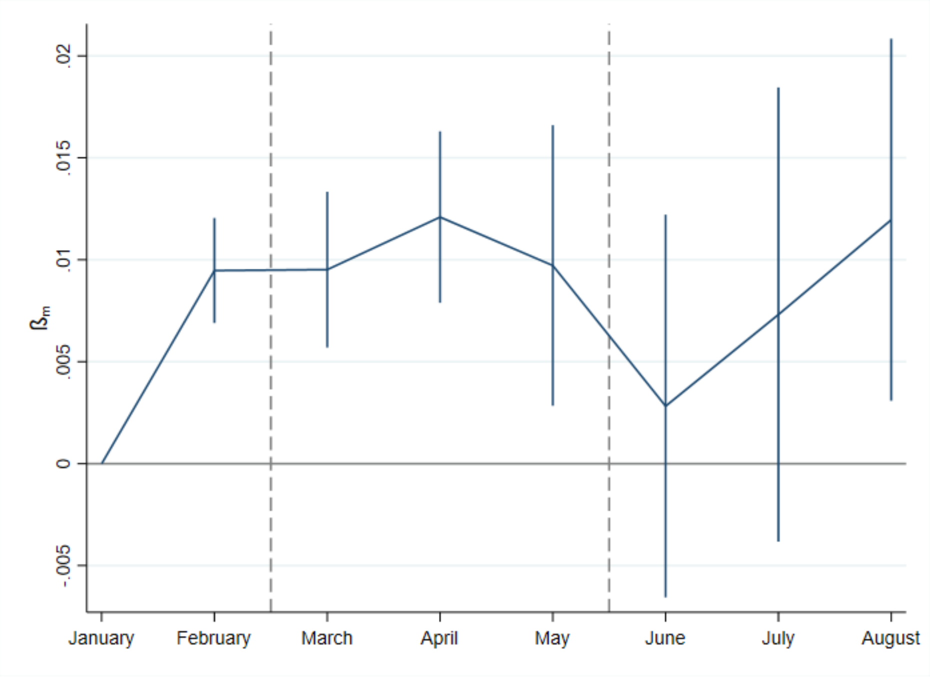
(a) Monthly chained indices

(b) Monthly fixed weight indices

(c) Weekly fixed weight indices

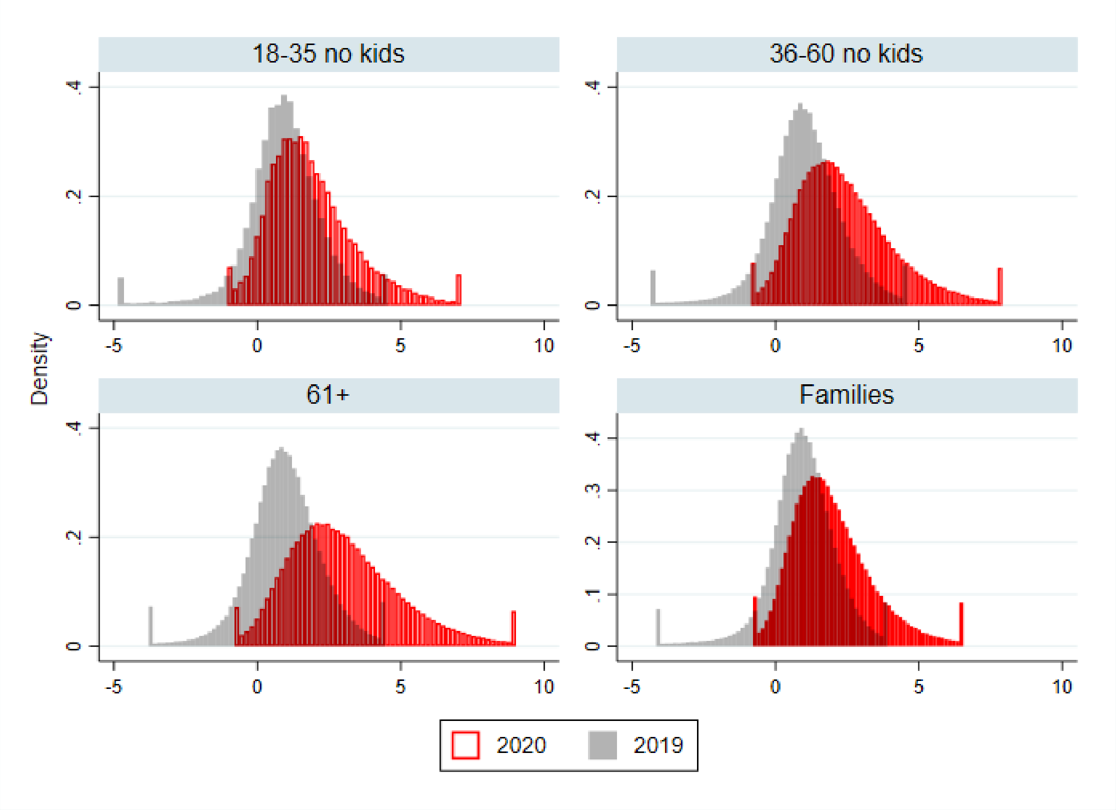
Notes: This figure plots different aggregate inflation indices 2020. Panel (a) plots monthly CES, Tornqvist and Fisher indices; panel (b) plots monthly fixed weights Laspeyres, Paasche and Fisher indices; panel (c) plots weekly fixed weights Laspeyres, Paasche and Fisher indices. Fixed weight indices condition on products present in all 34 weeks. These products represent 89 % of all sales. Dashed lines denotes the start and end of the French first lockdown.

Figure 2.B.3: β_m coefficients for 2019



Notes: This figure plots the coefficient β_m from equation (2.11) using observations from 2019 only. Vertical bars indicate the 95% confidence interval. Standard errors are clustered at the UPC level. Only UPCs observed all months in 2019 are included. Observations are weighted by expenditure shares. Dashed lines highlight the period between March and May, during which the lockdown took place in 2020.

Figure 2.B.4: Histogram of inflation in April 2020, by household type



Notes: Inflation in all figures is computed using a fixed based Fisher price index. Top and bottom 1% of observations are winsorised.

Chapter 3

Distinguishing between signal and noise in the measurement of the firm wage premium

Abstract

There is a growing interest about firm-side drivers of wage differentials, as different studies show that this component is driving the increase in inequality in many developed countries. In this paper, I contribute to this literature in three respects. First, I reconsider the widely used model from Abowd, Kramarz and Margolis used to decompose the respective contributions of firm and individual heterogeneity. I suggest an easily applicable split-sample procedure to uncover the extent of overfitting in this model. Using French administrative data, I find evidence of sizeable overfitting: conservative estimates suggest that the contribution of firm heterogeneity to wage inequality is overestimated by at least 25%. Second, I provide a simple procedure to recover the correct signal variance of firm effects and the covariance between individual and firm effects. Third, I show how to recover better prediction of the firm effects using shrinkage estimators. This matters quantitatively: due to shrinkage, half of the firm effects are shrunk by 38% or more, and 40% of firms end up in different deciles when ranked according to their firm effects.

3.1 Introduction

The increase in within-group inequalities (Juhn et al., 1993; Lemieux, 2006) over the past decade, mirrored by the increasing disparities between and within firms, first documented by Davis et al. (1991), triggered a growing interest in firm-side determinants of wages. In line with this body of work, recent evidence suggest that the employer is an increasingly important determinant of an individual's wage (Barth et al., 2016; Gruetter and Lalive, 2009; Song et al., 2019; Card et al., 2013). For instance, Song et al. (2019) estimate that about 40% of earning inequality in the United-States is accounted for by between-firms variations, and that this component explains two-thirds of the rise in earnings inequality over the past three decades.

In this context, researchers are often interested in disentangling sorting into firms from true wage premium, as these mechanisms correspond to substantially different explanations of between-firm earning inequality. Most of the empirical research focusing on these issues uses the two-way fixed effects regression model first introduced by Abowd et al. (1999). A simple variant of this model decomposes the natural logarithm of wage y_{it} of worker i at time t into time-varying observables, $x'_{it}\beta$, unobservable individual heterogeneity θ_i, unobservable firm heterogeneity ψ_j and a residual r_{it} :

$$y_{it} = x'_{it}\beta + \theta_i + \psi_{j(i,t)} + r_{it} \tag{3.1}$$

Where $j(i,t)$ refers to the firm $j = j(i,t)$ in which individual i works at time t. Under suitable assumptions discussed below, this model can be estimated by OLS and delivers unbiased estimates for θ_i and ψ_j. These can be then used to provide more detailed insight into the variance of wages. For instance the ratio $Var(\psi_j)/Var(y_{it})$ would give the share of the variance in wages explained by the firm wage premiums only.

However, the main parameters of interests, $\hat{\theta}_i$ and $\hat{\psi}_j$, are fixed effects, which are unbiased but can be poorly estimated if the sample size does not grow asymptotically in the required dimensions. In finite sample, the effective number of observations used to estimate the average firm or individual effect is quite low, which leads to "overfitting". This term originates from the statistical learning literature and describes the extent to which a low-bias but high-variance estimator tends to mistakenly interpret random noise for meaningful signal, and hence has poor out-of-sample predictive power. This overfitting raises two issues, first mentioned and analyzed by Abowd et al. (2004) and Andrews et al. (2008). First, by modeling the raw estimate $\hat{\psi}_j$ as the sum of the true signal ψ_j and an independent noise ν_j, it can be seen that the variance of the wage premium σ^2_ψ[1], is overestimated when the

[1] Unless otherwise indicated, the terms "variance of wage premium", σ^2_ψ and "variance of firm effects", $\widehat{Var}(\hat{\psi}_j)$ in this paper refer to the observation-weighted sample dispersion of the true or estimated firm

noise is not negligible and when the wage premium is measured by $s^2_\psi \equiv \widehat{Var}(\hat{\psi}_j) = \sigma^2_\psi + \sigma^2_\nu$. Additionally, this creates a downward bias in the estimate of sorting of workers in firms, as measured by $Cov(\hat{\theta}_i, \hat{\psi}_j)$. Indeed, $\hat{\theta}_i$ is estimated as the mean difference between y_{it} and the fitted value $x'_{it}\hat{\beta} + \hat{\psi}_j$, inducing a mechanical negative correlation between $\hat{\theta}_i$ and$\hat{\psi}_j$. Abowd et al. (2004) and Andrews et al. (2008) label these problems "limited mobility bias". The bias in the variance and covariance terms come from the non-vanishing squared estimation error of the fixed effects, which is itself due to the limited mobility of individuals. This paper will rather use the term overfitting, because it primarily refers to the fixed effects and thus encompasses both the bias in the variance terms and the poor precision of the fixed-effect estimators.

This paper suggests an easy way to recover an unbiased estimate of the variance of the firm's wage premium, and proposes a simple method to derive better prediction of firm effects using shrinkage estimators. It finds that the variance of French wage premiums is at least 25% lower than what would be estimated using simple plug-in estimators, and that half of the firms in the sample have a predicted estimate that is at least 38% lower than the raw estimate. This difference in predicted estimates matters quantitatively. Ranking firms according to their estimated wage premium, a researcher focusing only on raw estimate would conclude that randomly moving a worker from a firm in the first quartile to a firm in the fourth quartile would increase her yearly wage by about $9,000$ euros on average, assuming the estimates are causal. Another researcher using shrunk effects to measure predicted wage premium would give a 35% lower number.

More specifically, this paper makes three main contributions. In the first part of the analysis, it suggests an easily applicable split-sample procedure, uncovering robust evidence of overfitting using French administrative data. Within each firm-year cell, every worker is randomly allocated to one of two samples, making sure that every worker appears in only one sub-sample. This ensures that every firm is present in both samples, while leaving intact information relative to the workers' employment history. This step is crucial as firm effects are identified from moves between jobs, and random splitting at the dataset level would break the pattern of moves for each workers. With these two samples, one can get two different estimates $\hat{\psi}_{1j}$ and $\hat{\psi}_{2j}$ for every firm j. If the effect is precisely estimated, these two sets of estimates should be similar. This paper also documents that overfitting is stable over time and seems to primarily come from top and bottom of the within-firm wage distribution. Secondly, this paper provides a simple procedure to recover the correct signal variance of firm effects and the covariance between individual and firm effects: corrected estimates suggest that the variance of firm effects is 25% lower than previously estimated, and the covariance is 25% higher. While the sample variance of $\hat{\psi}_j$ is biased due to the squared error

effects respectively, that is: $\frac{1}{N-1}\sum_{i,t}\left(\psi_{j(i,t)} - \bar{\psi}_{j(i,t)}\right)^2$ and $\frac{1}{N-1}\sum_{i,t}\left(\hat{\psi}_{j(i,t)} - \bar{\hat{\psi}}_{j(i,t)}\right)^2$, where N is the sample size.

term, ν_j^2, having non-zero expectation, the sample covariance between $\hat{\psi}_{1j}$ and $\hat{\psi}_{2j}$ recovers the correct magnitude of the variance of the firm estimates, because each estimate has been computed using different data and thus the interaction of the two estimation noises, $\nu_{1j} \cdot \nu_{2j}$, has expectation of zero. Similarly, the sample covariance between individual effects from sample 1, $\hat{\theta}_{1i}$ and firm effects from sample 2, $\hat{\psi}_{2j}$ is an unbiased estimator of the sorting estimate $Cov(\theta_i, \psi_j)$ because the data used to estimate $\hat{\theta}_{1i}$ was not used to estimate $\hat{\psi}_{2j(i,t)}$. Using split-sample to recover unbiased estimates comes at some efficiency cost, because one relies on half as many observations only to estimate every firm effect. Based on a simplified statistical model, this paper argues that this trade-off - as measured by the mean squared error of the estimators - is most likely to be favorable to a split-sample solution. Thirdly, this paper shows how to recover better prediction of the firm effects by shrinking each raw estimate by its signal to noise ratio in order to minimize the prediction error. Shrinkage estimators have been introduced by James and Stein (1961) and Efron and Morris (1973), and have been used for instance in the teacher value-added literature (Chetty et al., 2014a,b). This paper documents substantial shrinkage for a significant fraction of the data set. For instance, half of the firm effects are shrunk by at least 38%, and 40% of firms end up in different decile of the observation-weighted distribution of firm effects after the shrinkage procedure has been carried out. To achieve this, this paper optimally trades off bias and variance so that shrunk firm estimates are now slightly biased but have much lower variance, resulting in lower mean squared error. In particular, this paper approximates the variance of the noise ν_j as the ratio of a common component on the number of moves from and to firm j, as motivated by Jochmans and Weidner (2019), in order to recover a different signal-noise ratio for every firm.

This paper relates to several strands of the literature. First and foremost, it relates to several studies quantifying the separate contribution of sorting and the firm's wage premium to wage inequalities. This literature was initiated by Card et al. (2013), who first applied two-way fixed effect regressions to this problem, followed by, among others, Song et al. (2019) and Gruetter and Lalive (2009). Bloom et al. (2018) and Colonnelli et al. (2018) apply the same framework but focus on the large-firm wage premium. Card et al. (2013) find that the share of the variance of wage that can be attributed to firm-specific premium has been increasing over time. As their primary focus is thus not on the absolute share of the variance explained by firm-effects, but rather on its evolution across time periods, this paper does not contradicts their results because it does not find any evidence that the extent of the overfitting problem has been evolving over time. Rather, this paper fits in a growing body of evidence suggesting that wage premiums are less important than previously thought in explaining the overall variance of wage. For instance, Kline et al. (2020) also focus on the bias in variance components and reach the same conclusion when looking at a set of Italian province. Relatedly, Song et al. (2019) establish that the "non-parametric" between-firm component of wage variance in the US is mostly and increasingly explained by sorting rather than wage premium. Similarly, Bloom et al. (2018) and Colonnelli et al. (2018) apply this same framework to study the large-firm wage premium.

This paper also builds on a number of econometric studies analyzing the limitations of the two-way regression model pioneered by Abowd et al. (1999), and tries to bridge it with the literature on wage inequalities by providing an easily implementable solution to the known overfitting problem. The negative bias in the estimate of the covariance between individual and firm effect was first discussed by Abowd et al. (2004), who labeled the problem "limited mobility bias", as it arises from the low number of workers' moves in the data. This perspective was enriched by Andrews et al. (2008) and Andrews et al. (2012), who formally discuss the bias in the variance and covariance terms, and suggests a correction assuming independent and homoskedastic errors. Bonhomme et al. (2019) take a radically different approach to the same problem, by first allocating ex-ante workers and firms to a small number of groups and estimating and estimating a model using these groups. While this approach is promising, it is also further away from the standard toolbox of the applied researcher. In addition, they also show in a related paper (Bonhomme et al., 2022) that the classification error in their first step can create a bias of the same order than the one from traditional two-way fixed-effect regressions. Jochmans and Weidner (2019) build on Andrews et al. (2008) but take a network approach, and provide bounds on the variance of the fixed-effects depending on the characteristics of the graph generated by the data. Closest to this paper is the one from Kline et al. (2020). They introduce “leave-one out” unbiased estimators of quadratic forms of OLS estimates, of which our parameters of interest are a particular case. The intuition behind Kline et al. (2020) and this paper is essentially the same: the bias in estimates of variance of fixed-effects comes from the squared, non-negligible error term ν_j. Split-sample estimates (or leave-one-out estimates, which are a particular form of sample-splitting), provide an intuitive solution to eliminate of this squared error term, by instead multiplying independent errors. Relative to Kline et al. (2020), this paper has two advantages and two drawbacks. First, it shows how to recover more precise estimates of firm fixed effects. This is important as it can potentially enable researchers to use two-way fixed effect regression estimates with more confidence. Second, this paper's split-sample procedure is relatively simpler and quicker to use: in addition to the OLS regression on the whole initial sample, only need two more regressions are needed, one for each sample. This is useful as given the usual size of the data set – a whole country's population-, a single OLS regression can take up to several days on standard computers. On the other hand, Kline et al. (2020) estimates are more general as they can also be applied to the unbiased estimation of the variance of individual fixed effects, which is not possible to do in the context of this paper. This would be an important area for further improvement. Last, Kline et al.'s estimation procedure is less sensible to smaller samples. To see this, note that firm fixed-effects are jointly identified only in sets where firms are connected by worker's moves between them. Their leave-one out procedure requires that the connected set of firms does not change when any worker is removed, while this paper requires that the connected set does not to change when each firm's labour force is randomly split in two groups and hence that there is "enough" workers in each firm. In this sense, this paper's procedure is more data-intensive.

Finally, this paper can be seen as bridging a methodological gap by introducing in the study of wage inequalities new tools to draw economic conclusions from fixed effects estimates. The value-added literature, whether it is of teachers (Chetty et al., 2014a,b; Angrist et al., 2017), of neighbourhood effects (Chetty and Hendren, 2018) or of patent examiners (Feng and Jaravel, 2020) regularly use fixed effects as the main parameters of interest. They highlight the necessity to take into account the fixed effects' estimation error in order to work with plausible estimates; and introduce shrinkage procedures to this end. To my knowledge, these tools have not been used to study wage inequality so far.

The reminder of paper is organized as follows. Section 3.2 develops a simplified statistical model to help build intuition about the sample-splitting approach. Section 3.3 describes the data and discusses the challenges arising when implementing empirically the two-way fixed effect regressions. Section 3.4 presents the sample-splitting procedure and evidence of overfitting. Sections 3.5 and 3.6 discuss how to recover the right (co)variance and better firm effects respectively. Section 3.7 concludes.

3.2 Statistical model

This section develops a simplified statistical model to help build intuition about the two-way regression model, overfitting, and the split-sample approach.

3.2.1 Set-up

Consider a simplified version of equation (3.1) where we abstract from time-varying observables x_{it} (or alternatively, where x_{it} have already been partialed out).

$$y_{it} = \theta_i + \psi_j + r_{it} \tag{3.2}$$

Where ψ_j refers implicitly to $\psi_{j(i,t)}$, the firm j at which individual i works at time t. In the remainder of the paper, this notation is used when no confusion is possible.

In matrix form, we have $y = D\theta + F\psi + r$, and we assume that $\mathbb{E}(r) = 0$, $\mathbb{E}(D'r) = 0$, $\mathbb{E}(F'r) = 0$, $\mathbb{E}(r_{it}r_{i't'}) = 0$ for all $i \neq i'$ and for all t, t', so that the error term is assumed to be uncorrelated across individuals but can be arbitrarily correlated within individuals. Further, errors r_{it} are homoskedastic: $\mathbb{E}(r_{it}^2) = \sigma_r^2$, but all results below can be extended to the heteroskedastic case. There are I individuals, each of them being observed T_i times, J

firms, T time periods, and thus $N \equiv \sum_{i=1}^{I} T_i$ observations.

We are primarily interested in the variance of the firm fixed effects and the covariance between individual and firm fixed effects.

$$\sigma^2_\psi \equiv \frac{1}{N-1}\sum_{i,t}(\psi_{j(i,t)} - \overline{\psi})^2$$
$$\sigma_{\psi\theta} \equiv \frac{1}{N-1}\sum_{i,t}(\psi_{j(i,t)} - \overline{\psi})(\theta_i - \overline{\theta})$$

Where $\bar{x}$ denotes the sample mean of variable x. In this simplified model, $T = T_i = 2$ for all i. Further, I assume a star economy in which there is one large firm and J small firms. In period 1, all individuals work in the large firm, and in period 2, all of them move to the small firms. Specifically, there are M_j individuals moving to firm j in the second period.

Last, the effect of the large firm is normalized to zero, so that $\psi_0 = 0$ and we assume that $\mathbb{E}(\psi_j) = 0$ for all small firms j. The aim of these assumptions is three fold. First, it abstracts away from the need to keep track of the origin of every worker, so that the wage difference $y_{i2} - y_{i1}$ will always be relative to the large firm. In addition, connectedness, that is the fact that firms are connected to each other through workers move is not a problem, as all small firms are connected to the big firm. Second, assuming that workers work in the big firm in period 1 and then move to a small firm in period 2 is innocuous and simplifies exposition. Appendix 3.C.1 shows that if there are no period effects (or if they are partialed out), the direction of the move does not matter. Third, as discussed below, firm effects are only identified relative to their average, which can be set to zero. Assuming that the effect of the large firm is 0 is therefore without loss of generality.

3.2.2 Two-way regression and overfitting

Appendix 3.C.1 shows that in this simple example, the OLS solution of the two-way fixed effect regression model is:

$$\hat{\psi}_j = \frac{1}{M_j}\sum_{i=1}^{M_j}(y_{i2} - y_{i1}) = \psi_j + \frac{1}{M_j}\sum_{i=1}^{M_j} u_i \equiv \psi_j + \nu_j \tag{3.3}$$
$$\hat{\theta}_i = \frac{1}{T_i}\sum_{t=1}^{T_i}(y_{it} - \hat{\psi}_j) = \theta_i + \frac{r_{i1} + r_{i2}}{2} - \frac{1}{2}\nu_j \tag{3.4}$$

Where $u_i = r_{i2} - r_{i1}$ has mean zero and finite variance σ^2_u. ν_j is independent of ψ_j, has also mean zero and variance $\sigma^2_j = \sigma^2_u / M_j$. $Var\left[\psi^2_j\right]$ and $Var\left[u^2_i\right]$ are assumed to be finite and constant for all j, and i. This amounts to assuming finite fourth moments and ensures convergence of estimators. These assumptions are standard and ease the notation, but can also be relaxed.

Consider the plug-in estimator for σ_ψ^2:

$$\tilde{\sigma}_\psi^2 = \frac{1}{I}\sum_i \hat{\psi}_{j(i)}^2 = \frac{1}{I}\sum_j M_j \cdot \hat{\psi}_{j(i)}^2 \tag{3.5}$$

Where $j(i)$ refers to the firm j in which i works in period 2. This is the correct plug-in estimate of σ_ψ^2 for this simplified model. The sum consists of I terms and not $N = 2I$. By doing so, we focus only on the variance of firm effects in period 2. This makes sense as period 1 variance of firm effects is zero (every one works in the large firm). Further, one can easily take period 1 variance into account by dividing all the results by 2. In addition, the deviation of the fixed effects from their mean is not accounted for because the mean of the firm effects is normalized to zero. Overall, $\tilde{\sigma}_\psi^2$ of this simplified model recovers all important features from the general setting while easing the exposition.

Appendix 3.C.1 shows that the bias and variance of this estimator are:

$$\mathbb{E}\left[\tilde{\sigma}_\psi^2\right] - \sigma_\psi^2 = \frac{1}{I}\sum_j \sigma_j^2 = \sigma_u^2 \cdot \frac{1}{I/J} > 0 \tag{3.6}$$

$$Var\left[\tilde{\sigma}_\psi^2\right] = Var\left[\psi_j^2\right] \cdot \frac{1}{I^2}\sum_j M_j^2 + \frac{4}{I} \cdot \sigma_\psi^2 \cdot \sigma_u^2 + Var\left[u_i^2\right] \cdot \frac{1}{I^2} \cdot \sum_j \frac{1}{M_j} \tag{3.7}$$

Where I/J is the average number of observations by firm or the number of movers per firm if $\forall j$, $M_j = M$ so that $I = M \cdot J$. These formulas are easy to interpret. Equation (3.6) makes it clear that the plug-in estimate $\tilde{\sigma}_\psi^2$ can be strongly biased if the average number of observations per firm (or in this case, of moves) is small. Further, the variance of $\tilde{\sigma}_\psi^2$ is the sum of three components. The first term is irreducible and is due to the heterogeneity of the firm effects themselves. When the M_j's are constant across j, this term boils down to $Var\left[\psi_j^2\right]/J$, which tends to zero as the number of firms increases. The second term comes from the fact that $\hat{\psi}_j$ is the sum of two random variables, and the third component relates to the squared estimation error term.

Similarly, one can define the following plug-in covariance estimator for $\sigma_{\psi\theta}$:

$$\tilde{\sigma}_{\psi\theta} = \frac{1}{I}\sum_i \hat{\psi}_{j(i)} \cdot \hat{\theta}_i \tag{3.8}$$

Appendix 3.C.1 shows that the bias of this estimate equals $-\frac{J}{2I}\left(\mathbb{E}(r_{i1}^2) - Cov(r_{i1}, r_{i2})\right)$ and is indeed negative in this example when we assume homoskedasticity for r_{i1} and r_{i2} or when we assume that $Cov(r_{i1}, r_{i2})$ is not too high. The bias is driven by the fact that the estimation error ν_j enters positively in $\hat{\psi}_{j(i)}$ and negatively in $\hat{\theta}_i$, as discussed in Abowd et al. (2004).

3.2.3 Split-sample approach

This subsection shows that one can recover an unbiased estimate of σ^2_ψ and $\sigma_{\psi\theta}$ through sample splitting, and that the mean squared error of these estimators is lower than the one of the plug-in estimate $\tilde{\sigma}^2_\psi$ so that the trade-off between lower bias, but higher variance arising through the split-sample approach is advantageous.

Suppose that for each firm j, M_j is equally split into sample 1 and 2, so that we have $M_{1j} = M_{2j} = M_j/2$. For simplicity, assume that M_j is even. Appendix 3.C.1 shows that splitting samples equally minimizes the variance. Similarly to the section above, we have :

$$\hat{\psi}_{kj} = \psi_j + \frac{1}{M_{kj}} \sum_{i=1}^{M_{kj}} u_i \equiv \psi_j + \nu_{kj} \tag{3.9}$$

$$\hat{\theta}_{ki} = \frac{1}{T_i} \sum_{t=1}^{T_i} (y_{it} - \hat{\psi}_{kj}) = \theta_i + \frac{r_{i1} + r_{i2}}{2} - \frac{1}{2}\nu_{kj} \tag{3.10}$$

With $k \in \{1, 2\}$ denoting the k-th (sub)sample. The error term ν_{kj} is specific to the subsample and thus is independent across both k and j. The only difference between $\hat{\theta}_{ki}$ and its counterpart in (3.3) is the error term ν_{kj}, because every individual is either in sample 1 or sample 2. For the same reason, we only observe $\hat{\theta}_{1i}$ or $\hat{\theta}_{2i}$ for every i, but never both.

3.2.3.1 Unbiased estimate of σ^2_ψ

The split-sample estimator for σ^2_ψ is:

$$\hat{\sigma}^2 = \frac{1}{I} \sum_i \hat{\psi}_{1j} \cdot \hat{\psi}_{2j} = \frac{1}{I} \sum_j M_j \cdot \hat{\psi}_{1j} \cdot \hat{\psi}_{2j} \tag{3.11}$$

Appendix 3.C.1 shows that this term is unbiased, because the error terms are independent of each other. The same appendix also shows that:

$$Var\left[\hat{\sigma}^2\right] = Var\left[\psi_j^2\right] \cdot \frac{1}{I^2} \sum_j M_j^2 + \frac{4}{I} \cdot \sigma^2_\psi \cdot \sigma^2_u + 4 \cdot \left(\sigma^2_u\right)^2 \frac{J}{I^2} \tag{3.12}$$

It is interesting to note that only the third term in $Var\left[\hat{\sigma}^2\right]$ differs from $Var\left[\tilde{\sigma}^2\right]$, so that we have:

$$Var\left[\hat{\sigma}^2\right] - Var\left[\tilde{\sigma}^2\right] = 4 \cdot \left(\sigma_u^2\right)^2 \frac{J}{I^2} - Var\left[u_i^2\right] \cdot \frac{1}{I^2} \cdot \sum_j \frac{1}{M_j} \tag{3.13}$$

So that we can have $Var\left[\hat{\sigma}^2\right] > Var\left[\tilde{\sigma}^2\right]$ when M_j are large without changing the ratio of J/I^2 , that is when the number of movers per firm is high. Intuitively, this is the cost paid for unbiasedness through sample-splitting. Because each firm effect is estimated with half the number of observations, the precision of every estimate is lower.

The split-sample estimator $\hat{\sigma}^2$ of the variance of the firm effects is thus unbiased but has higher variance than the plug-in estimator. Is the trade-off worth it? To answer this question, one can compare the mean squared error of both estimators, as it is the most common way to measure the trade-off between bias and variance. As the mean squared error is the sum of the squared bias and the variance, we have:

$$\begin{aligned} MSE\left[\tilde{\sigma}^2\right] = \mathbb{E}\left[\left(\tilde{\sigma} - \sigma_\psi^2\right)^2\right] &= Var\left[\psi_j^2\right] \cdot \frac{1}{I^2} \sum_j M_j^2 + \frac{4}{I} \cdot \sigma_\psi^2 \cdot \sigma_u^2 \\ &+ Var\left[u_i^2\right] \cdot \frac{1}{I^2} \cdot \sum_j \frac{1}{M_j} + \left(\sigma_u^2\right)^2 \frac{J^2}{I^2} \\ MSE\left[\hat{\sigma}^2\right] = Var\left[\hat{\sigma}^2\right] &= Var\left[\psi_j^2\right] \cdot \frac{1}{I^2} \sum_j M_j^2 + \frac{4}{I} \cdot \sigma_\psi^2 \cdot \sigma_u^2 \\ &+ 4 \cdot \left(\sigma_u^2\right)^2 \frac{J}{I^2} \end{aligned} \tag{3.14}$$

In equations (3.14), the last term is the squared bias, and the terms on the first row are the terms in common between $\hat{\sigma}^2$ and $\tilde{\sigma}^2$. Hence, a sufficient condition for $MSE\left[\hat{\sigma}^2\right] < MSE\left[\tilde{\sigma}^2\right]$ is that $J > 4$, which is always satisfied in this type of data. The last term in $MSE\left[\hat{\sigma}^2\right]$ comes from variance of $\hat{\sigma}^2$ and is the only term not in common with the variance of $\tilde{\sigma}^2$. This sufficient condition thus really compares the increased bias in $\tilde{\sigma}^2$ from the increase variance in $\hat{\sigma}^2$.

Can we further improve on $Var\left[\hat{\sigma}^2\right]$? Motivated by Chernozhukov et al. (2018), appendix 3.C.1 explores the possibility to equally split the original samples in $K \geq 2$ subsamples, where by construction, K must be so that $K \leq \min_j M_j$. In this case, we would have K different estimates for each firm effect, as well as $\binom{K}{2}$ possible estimates of σ_ψ^2, $\hat{\sigma}_{kk'}^2$. By taking the simple average of these estimates, $\hat{\hat{\sigma}}^2 = \binom{K}{2}^{-1} \sum_{k,k'} \hat{\sigma}_{kk'}^2$ is an unbiased estimate of σ_ψ^2 with variance

$$Var\left[\hat{\hat{\sigma}}^2\right] = Var\left[\psi_j^2\right] \cdot \frac{1}{I^2} \sum_j M_j^2 + \frac{4}{I} \cdot \sigma_\psi^2 \cdot \sigma_u^2 + \left(\sigma_u^2\right)^2 \cdot \frac{2J}{I^2} \frac{K}{K-1} \tag{3.15}$$

Which boils down to equation (3.12) when $K = 2$. The empirical application in this paper focuses on the case where $K = 2$, but in unreported result, I find that this procedure with $K = 3$ does not significantly change the estimates of σ^2_ψ.

Overall, this simple example shows that even though unbiasedness comes at a cost, the increased variance in the estimator one has to accept in return is comparatively low, so that the mean squared error of the split-sample estimator is lower than the mean squared error of the plug-in estimator.

3.2.3.2 Unbiased estimate of $\sigma_{\psi\theta}$

Following the same intuition as for σ^2_ψ, an unbiased estimator of $\sigma_{\psi\theta}$ is one where the estimated firm effect of firm j from sample 1 covaries with the estimated individual effects for individuals who belong to the same firm but are in sample 2:

$$\hat{\sigma}_{\psi\theta} = \frac{1}{I}\sum_i \hat{\psi}_{1j(i)} \cdot \hat{\theta}_{2i} \tag{3.16}$$

Appendix 3.C.1, shows that this estimator is indeed unbiased. In the empirical application, based on the intuition developed in the multi-split case for σ^2_ψ and further analyzed in Chernozhukov et al. (2018), I get a second estimate of the covariance by inverting the role of sample 1 and 2, and then take the simple average of these two estimators to reduce the variance.

3.3 Data and empirical implementation of two-way fixed effect regressions

This section describes the data, discusses the two-way fixed effect regression and the challenges arising from its empirical implementation.

3.3.1 Data

3.3.1.1 Data source and sample

This paper uses the Décaration Annuelle de Données Sociale (DADS), a French matched employer-employee administrative data. Every year, employers send to the Social Security administration information about the pay of each of their employees, along with information on job duration, occupation, industry, hours worked and place of work. The French statistical agency (INSEE) then constructs a panel of individuals who can be followed over time, covering one twelth of the population. More specifically, all individuals born in October enter this data set called Panel DADS. Before 2002, the panel was half as big, as

Table 3.1: Summary statistics of different samples

	Baseline	Baseline	Analysis	Analysis sample 1	Analysis sample 2
		largest	largest	largest	largest
	(1)	(2)	(3)	(4)	(5)
Panel A: Individuals					
No. Individuals	2,427,892	2,231,920	1,687,078	839,610	839,335
Mean Wage	10.264	10.269	10.309	10.309	10.310
Q1 Wage	9.962	9.967	9.999	9.999	9.999
Median Wage	10.183	10.189	10.237	10.237	10.238
Q4 Wage	10.487	10.494	10.549	10.549	10.550
% Men	0.629	0.632	0.628	0.628	0.628
% living in IDF	0.255	0.258	0.276	0.276	0.277
Mean Age	38.6	38.4	38.9	38.9	38.9
Panel B: Firms					
No. Firms	988,428	784,112	41,975	41,018	41,010
Mean Obs/Year / Firm	2.1	2.3	23.3	11.8	11.8
Mean Obs / Firm	16.6	19.7	219.7	112.2	112.1
Mean Moves / Firm	6.0	7.4	61.8	31.6	31.6
Q1 Moves / Firm	1.0	1.0	8.0	4.0	4.0
Median Moves / Firm	2.0	2.0	17.0	9.0	9.0
Q3 Moves / Firm	3.0	4.0	40.0	21.0	20.0
Observations	16,426,490	15,479,838	9,220,625	4,600,336	4,599,142

Notes: Wage data is in log gross real annual format. The Analysis data set is restricted to firm-year cells with at least 8 observations. Data in columns (2) to (5) come from the largest connected set of the respective samples. Moves refers to the number of moves to or from a given firm.

only individuals born in October in an even year were included. Observed firms are those at which individuals from the panel work.

This paper focuses on full-time employees aged 20-60 in the private sector between 2002 and 2015. For every spell of employment, real gross wage data is observed. Following Card et al. (2013), if individuals have more than one spell of employment per year, I select a unique individual-year observation as follows. INSEE constructs a variable indicating whether a spell can be considered as a "side" job. If the individual has only one non-side job, I select this spell as the main job for that year. If the individual has more than one non-side job, I select as the main job the job at which the individual earned most. If the individual has only side jobs, I discriminate in a similar manner. Once a unique observation for every individual-year cell is constructed, I compute a yearly "full-time equivalent" log real gross wage. The baseline sample thus consists of about 2.5 million individuals and 1 million firms.

Column (1) and Panel A of table 3.1 provide some summary statistics for individuals in this sample. Slightly less than two thirds of them are men, and about 25% live in the Ile-de-France region around Paris.

3.3.1.2 Wage Inequality in France

Figure 3.1 provides some basic insight into the level and change of wage inequality in over the past fifteen years. Panel (a) shows that along a number of inequality measures summarizing the central part of the distribution, wage inequalities in France have been roughly constant between 2002 and 2015. This is very different from the dynamics of inequality in the US (Song et al., 2019), the UK (Van Reenen, 2011) or Germany (Card et al., 2013). In addition, the economic crisis of 2008-2009 does not seem to have significantly impacted these inequality measures. Panel (b) shows the evolution of the variance of wage between 2002 and 2015 and its non-parametric decomposition between a between-firm and a within-firm component. For every year, we have: $\sum_i (y_i - \overline{y})^2 = \sum_i (y_i - \overline{y}_{j(i)})^2 + \sum_i (\overline{y}_{j(i)} - \overline{y})^2$, where $\overline{y}$ is the average wage. The first component on the right-hand side is the within-firm wage variation whereas the second is the between-firm variation. The share of total wage variance coming from the within-firm component varies between 55% at the end of the sample and about two-thirds at the beginning. This is of a similar order of magnitude as in Song et al. (2019).

Overall, these two figures are consistent with previous findings on the distribution of French gross wage. For instance, Verdugo (2014) find that the French wage distribution compressed between 1969 and 2008.

3.3.2 Fitting two-way fixed effects regressions to the data

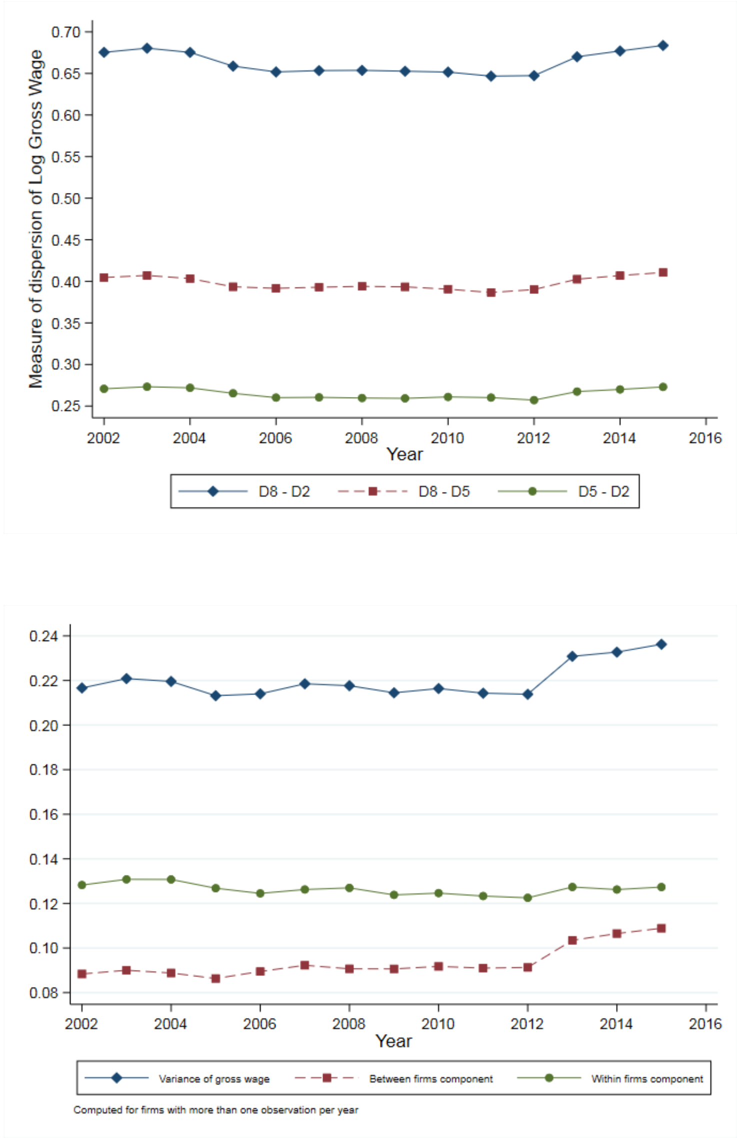
This subsection exposes the identifying assumptions made in most of the literature since Card et al. (2013), discusses the specificities of the empirical implementation of the two-way regression model, and examines its application to French data.

3.3.2.1 Identifying assumptions

Equation (3.1) allows for unbiased estimation of parameters β, θ_i and ψ_j under the identifying assumptions that $\mathbb{E}(X'r) = 0$, $\mathbb{E}(D'r) = 0$ and $\mathbb{E}(F'r) = 0$ where r is the stacked vector of error terms r_{it}, X is the matrix of time-varying observable covariates, and D and F are the design matrices of worker and firm indicators respectively. As Card et al. (2013) note, the only non-standard assumption is the last one. A sufficient assumption for it to hold is conditional exogenous mobility: the probability that individual i works at firm j in period t can depend in an unrestricted way on the individual heterogeneity θ_i, the different wage premiums prevailing in the economy $\{\psi_1, \ldots, \psi_J\}$, but cannot depend on the error term r. While recent evidence suggest this might be a too strong assumption, at least in the US framework (Abowd et al., 2018), this paper follows the literature by instead providing graphical evidence that this assumption approximately holds in the data.

Figure 3.1: Evolution of wage inequality and its determinants, 2002-2015

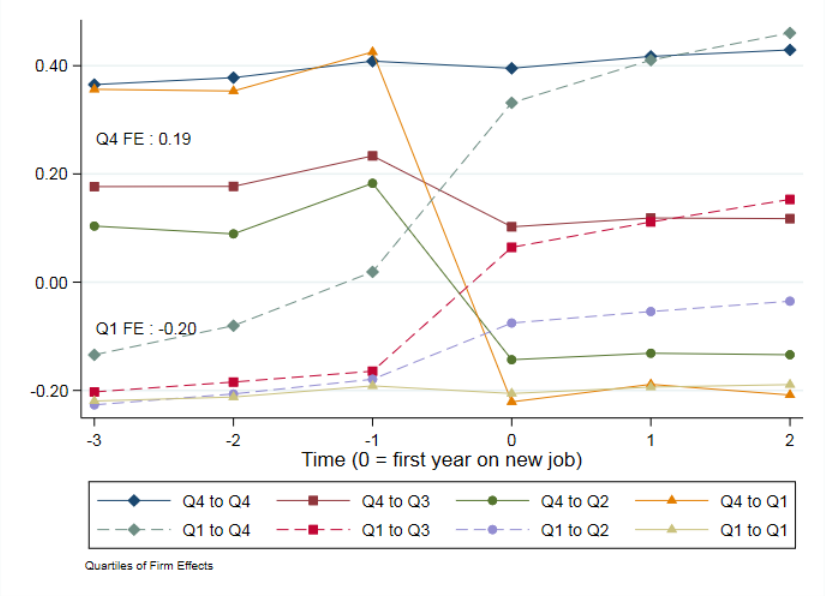
(a) Measures of wage inequality in France



(b) Non-parametric wage decomposition

Notes: Data comes from the baseline sample

Figure 3.2: Mean Wage of Job Changers



Notes: This graph plots the mean wage of job switchers before and after the switch, depending on the quartile of firm effects to which they belong following Card et al. (2013). Selected individuals are observed at least six times in a row and only change job once.

Card et al. (2013) model the error term as the sum of a match effect, a unit-root term, and an idiosyncratic shock. Following their notation: $r_{it} = \eta_{ij(i,t)} + \zeta_{it} + \epsilon_{it}$. Conditional exogenous mobility implies then that individuals moving from high-paying firms to low-paying firms experience a wage decrease of similar amplitude than the wage increase of individuals moving from low-paying to high-paying firms ($\mathbb{E}(F'\eta) = 0$) and that there are no pre-trends before a move from a high-paying firm to a low-paying firm or vice-versa ($\mathbb{E}(F'\zeta) = 0$). Figure 3.2 follows Card et al. (2013) and plots the average wage of job-switchers in high- and low-paying firms around the time of their move. Firms are classified as high or low paying firms depending on the quartile of firm effects to which they belong. The different curves draw the mean wage of job movers depending on the quartile of their origin and destination firm. Even though this paper does not intend to discuss in detail the identifying assumptions of two-way regression models, one of the main results suggests that one should treat these type of graphical evidence with caution. Indeed, I show in section 3.6 that 15% of firms switch quartiles when estimates are corrected. Figure 3.2 suggests that the identifying assumption seems relatively well satisfied for most groups of movers. Except for the group of movers from firms from the bottom quartile to the top quartile, no pre-trends in wages during the move is distinguishable.

Another assumption relates to the additive structure of equation (3.1), which assumes that the wage premium in firm j is the same for every individual working at that firm. While this is certainly not true in reality, the estimates $\hat{\psi}_j$ can be seen are unbiased approximation of the mean wage premium in the firm. In addition, both Card et al. (2013) and Bonhomme et al. (2019) conclude from different approaches that the additive structure of equation (3.1) approximates well true wage process, in the sense that allowing for match-specific fixed effects does not sensibly improves the estimation fit. In unreported results, I also allow for match-specific effects and find that the R^2 increases only slightly.

3.3.2.2 Empirical Implementation

The implementation of two-way fixed effect regression models in this matched employer-employee framework differs in two main respects from the usual OLS regressions used by applied researchers. First, workers and firm fixed-effects are separately identified only within a connected set. To see this, consider a oversimplified data set of two periods and no time-varying observables, in which there is only one worker staying in the unique firm for both periods. In this case, it is impossible to distinguish whether the wage level is due to the firm's wage premium or the individual's unobservable characteristics. If instead the individual changes firms between these two periods, one can identify the difference in wage premiums between firms as the wage difference between period 1 and period 2. More generally, Abowd et al. (2002) show that workers and firm fixed-effects are separately identified only within a connected set, and provide an algorithm to compute these sets. Intuitively, two firms are in the same connected set if they can be linked by worker moves. In practice, researchers constrain their analysis to the largest connected set, which usually encompasses above 90% of the observations. This paper follows the literature in this respect. Table 3.1 displays summary statistics of the baseline sample (column (1)) and of the largest connected set in this sample (column (2)). Panel A confirms that individual in the largest connected set are very comparable to the ones in the baseline sample, and that summary statistics of the baseline sample and its largest connect set are extremely similar. Panel B suggests that most of the observations outside the connected set come from small firms. Overall, the largest connected set in my data recovers 94% of the baseline sample. In addition, it is important to note that within each connected set, each firm effect is identify only relative to a normalized value. In the previous oversimplified example, one could only identify the difference of the firm effects. Setting the average of fixed effects to zero enables separate identification of the two effects. In practice, either one particular firm effect or the mean of the firm effects are normalised to zero. This paper sets the sample mean of the firm effects to be zero, so that $\hat{\psi}_j$ can thus be conceptually interpreted as the wage premium given by firm j to its employees above their market wage. If all firms increase their wage premiums by δ, this becomes part of the market wage. Naturally, this normalization does not change the variance of the firm effects. Further, one can note that the fixed effects of non-movers in the largest connected set are also identified, because the fixed effect of their firm is already identify by other moving employees.

A second empirical challenge is the size of the data set. It is now common for researchers to have access to administrative data comprising of millions of individuals and several hundreds of thousands of firms. As a result, the matrix $Z'Z$ is a square matrix of size $K+N+J$, where $Z=[X,D,F]$. Practically, such a matrix is very hard to store on a standard computer (Andrews et al., 2006), and is impossible to invert with standard computer configuration. This has two consequences. First, as the $Z'Z$ matrix cannot be inverted, one cannot solve the normal equations to compute the OLS estimates. Instead, researchers rely on iterative algorithms. This paper uses the `-reghdfe-` Stata command from Correia (2016). Second, and more importantly, it also implies that one cannot recover the standard errors on the different fixed-effect coefficients.

3.3.2.3 Variance decomposition

Following Song et al. (2019), this paper decomposes of variance of wages as:

$$\begin{aligned} Var\left(y\right) =& Var\left(\theta-\bar{\theta^j}\right)+Var\left(X\beta-\bar{X^j}\beta\right)+2\cdot Cov\left(\theta-\bar{\theta^j},X_i\beta-\bar{X^j}\beta\right)+Var\left(r\right) \qquad (3.17)\\ &+Var\left(\psi_j\right)+Var\left(\bar{\theta^j}\right)+Var\left(\bar{X^j}\beta\right)\\ &+2\cdot Cov\left(\bar{\theta^j},\psi_j\right)+2\cdot Cov\left(\bar{\theta^j},\bar{X^j}\beta\right)+2\cdot Cov\left(\psi_j,\bar{X^j}\beta\right)\end{aligned}$$

where variable $\bar{\theta^j}=\frac{1}{\sum_{i,t}\delta_{ijt}}\sum_{i,t}\delta_{ijt}\cdot\theta_i$, $\delta_{i,j,t}$ is an indicator variable equals to one if worker i works in firm j at time t, where the other variables $\bar{\cdot^j}$ are defined accordingly. Thus, $\bar{\theta^j}$ denotes the average worker effect for workers in firm j throughout the period. Further, the variance and covariance terms from the between-firm component are taken over all firms, and weighted by the number of observations. Equation (3.17) decomposes the variance of wage into several components. The terms on the first row together form the within-firm component, whereas the terms on the second and third row are the between-firm component of the variance of wages. Crucially, we are interested in this paper on the variance of the firm wage premiums, $Var(\psi_j)$, as well as the measure of worker sorting into firms, $Cov\left(\bar{\theta^j},\psi_j\right)$. Table 3.2 reports the decomposition of the variance of wage into between- and within- firm components in the baseline sample. As has been reported by others(Card et al., 2013; Song et al., 2019), most of the variance in wages comes from individual heterogeneity, either through variation within firms (41.1%), or through variation between firms (27.5%). Further, in the baseline dataset, the variance of firm effects accounts for about 15% of the total variation in wage, which is on the lower end of usual values from other studies. As a comparison, Card et al. (2013) finds that variance of establishment effects explains between 18.5% and 21% of wage variance. The estimated covariance between individual and firm fixed effect is slightly negative, as in Abowd et al. (2004). Appendix table 3.A.1 shows the same regression results from the Analysis sample (see section 3.4.1, when only firm-year

Table 3.2: Decomposition of the variance of wages between and within firms

Decomposition a la Song et al. (2019)	Var. Component	Share of total
Total variance	0.215	1.000
Between firms	**0.095**	**0.44**
Var. of $\bar{\theta}^j$	0.059	0.275
Var. of Firm Effect ψ_j	0.032	0.147
Var. of $\bar{X}^j\beta$	0.009	0.041
$2\,Cov(\bar{\theta}^j, \psi_j)$	-0.006	-0.027
$2\,Cov(\bar{\theta}^j, \bar{X}^j\beta)$	-0.001	-0.005
$2\,Cov(\psi_j, \bar{X}^j\beta)$	0.002	0.011
Within firms	**0.120**	**0.56**
Var. of $\theta - \bar{\theta}^j$	0.088	0.411
Var. of $X\beta - \bar{X}^j\beta$	0.024	0.111
Var. of Residual	0.029	0.134
$2\,Cov(\theta - \bar{\theta}^j, X\beta - \bar{X}^j\beta)$	-0.021	-0.099
Number of Person effects	2,231,920	
Number of Firm Effects	784,112	
Number of Different Spells	5,149,360	
Sample size	15,479,838	

Notes: $\bar{Y}^j$ refers to the mean of variable Y taken over individuals working at firm j. The variance and covariance terms are weighted by the number of observations. Data comes from the baseline sample.

cells with more than 8 observations are considered). In this case, the covariance term is positive, at 0.008. This suggests that part of the negative covariance is indeed driven by negative bias, which is partly attenuated when one considers bigger firms. For the same reason, the variance of firm effects is also much lower, at 5%. Overall, the between-firm share of wage variation is about 45% in total, in line with evidence from panel (b) of figure 3.1.

3.4 Evidence of overfitting

This section describes the sample-splitting procedure and presents evidence that overfitting is important in the French data. In this data, overfitting seems constant over time and is primarily driven by the top and bottom of the within-firm wage distribution.

3.4.1 Sample-Splitting Procedure

This subsection shows that the proposed sample-splitting procedure delivers balanced subsamples and allows to recover two independent estimates for every firm for most of the baseline sample.

Algorithm. Usual split-sample procedures randomly allocating observations into two different sub-samples at the data set level would not work. Due to the panel and network structure of the data, it would lead to three impractical problems. First, while most firms would be present in both samples, the number of observations available for every firm might differ greatly, inducing an imbalance. More importantly, a given individual i would be observed in both samples. It is common to assume that even controlling for the unobserved heterogeneity θ_i, the error term r_{it} is correlated within i across t. Observations from the same individual will thus be used in both samples to compute the firms effects and $\hat{\psi}_{1j}$ would not be independent of $\hat{\psi}_{2j}$ (conditional on ψ_j). $Cov(\hat{\psi}_{1j}, \hat{\psi}_{2j})$ would then not be an unbiased estimate of σ^2_{ψ}. In addition, the discussion above on the empirical challenges of two-way regressions made clear that the driver of the connectedness of a given data set is the pattern of mobility of workers across various firms. Randomly sampling at the data level could break this pattern, which could then result in very different largest connected sets in both sub-samples.

To circumvent these problems, I build on the dense sampling algorithm of Woodcock (2008) and adapt it to the purpose of this paper. Workers are randomly allocated to a sub-sample within each firm-year cell, so that every worker appears in only one sub-sample. This ensures that every firm is present in both samples, while leaving intact the pattern of mobility. The precise algorithm works as follows.

We start with the first year $t = 1$. Within each firm observed in $t = 1$, we divide randomly into two sub-samples (sample 1 and sample 2) all observations which have not been previously allocated to a sub-sample. At $t = 1$, these are all observations. At the end of this step, for a firm having N_j employees at time $t = 1$, $N_j/2$ of them will be allocated to sample 1 and $N_j/2$ will be allocated to sample 2. In case of an odd number of employees N_j, we make sure the additional employee is not systematically allocated to a particular sample. Because this step is reproduced every year, each of these observations represent one unique individual. In a second step, we then allocate for every individual all their employment history to the same sub-sample they have been allocated to. Specifically, in $t = 1$, if individual i has been allocated to sample 1 and is observed in $t' > t$, the observation corresponding to individual i in t' will be allocated in sample 1 as well. In a third step, we then move to $t = 2$. Some individual will enter the panel in this period and thus will not have been previously allocated. Thus, within each firm observed in $t = 2$, we divide randomly into the two samples all observations which have not been previously allocated. This randomly allocates all individuals appearing for the first time in the panel in $t = 2$. We then assign all their employment history (that is, for all years $t' \in \{3, \ldots, T\}$) to their sub-sample. Last, we repeat this step 3 until all years have been exhausted.

Sample restriction. The sample needs to be restricted in order to have a sufficient number of observations per firm. In the baseline sample, about $750,000$ different firms can be identified, for about 15.5 millions observations in total. The mean number of observations per firm is low, at about 20 and the number of moves per firm, which is the main number of interest when it comes to effective sample size, is even lower at about 7.5 (see table 3.1). This is due both to the skewed nature of the firm size distribution, and to the characteristic of the data: only firms at which individuals from the panel work are observed. However, this sample-splitting procedure is data intensive in the sense that if we are to split every firm into two different samples and estimate a firm effect out of these, we would like to have firms with many employees, to be able to recover two precisely estimated estimates. In this paper, all firm-year cells with less than eight observations are droped. This naturally shifts the focus of the analysis to bigger firms. While this is a very rough requirement, it seems to work well in practice. The resulting data set is henceforth called the Analysis dataset. Columns (2) and (3) of table 3.1 display summary statistics of the baseline and analysis data set resulting from the ≥ 8 observations per firm-year cell restriction. First of all, the new sample recovers about 60% of the original observations, and 75% of the individuals. On the other hand, the number of firms dramatically shrinks from about $750,000$ to about $42,000$. As expected, the mean number of observations and moves per firm dramatically increases. In addition, this sample restriction barely modifies the summary statistics of the individuals. Both Kline et al. (2020) and Woodcock (2008) impose a similar restriction. In the application of his dense sample algorithm, Woodcock (2008) looks at firms with more than five employees per year. Kline et al. (2020) impose the slightly more complex but finer requirement that the largest connected set remains connected if one drops out any one worker. In doing so, they only drop about 15% of their sample.

Empirical Implementation. This sample-splitting procedure works remarkably well in practice. At the end of the procedure, the only possibility for a firm not be be in both sub-samples is for it to appear in year $t' > 1$ and to have all its employees, already assigned to the same sample. In my data, this happens a negligible number of times (about 150 observations out of several millions). In addition, it ensures as expected high connectedness in both samples. The largest connected sets of both samples recovers more than 99.8% of the observations. As a consequence, virtually all firms present in the original sample end up in both largest connected sets, and one can thus have two estimates of their firm effect. The last three columns of table 3.1 show that the sample splitting manages to recover balanced samples. Panel A demonstrates that the characteristics of individuals in both sub-samples are the same. Similarly, panel B shows in the last two columns that firms' summary statistics are also extremely similar. In particular, the main moments of the distribution of the number of moves per firms is close to identical.

3.4.2 Evidence of overfitting

This subsection shows that estimating two-way fixed effect regressions with my data leads to very imprecise estimates of fixed-effects, and that this stylised fact is likely to be present in other data sets as well.

Strategy. To test for overfitting, I follow the insight from Feng and Jaravel (2020) that the estimated firm effects $\hat{\psi}_j$ should not pick-up random noise. In other words, the estimation error should be negligible. To implement this intuition, I regress $\hat{\psi}_{2j}$ on $\hat{\psi}_{1j}$. Appendix 3.C.2 shows that the coefficient on $\hat{\psi}_{1j}$ tend towards the signal-noise ratio.

$$\hat{\gamma} \equiv \frac{\sum_{i,t} \hat{\psi}_{1j(i,t)} \cdot \hat{\psi}_{2j(i,t)}}{\sum_{i,t} \hat{\psi}^2_{1j(i,t)}} \xrightarrow[J\to\infty]{p} \frac{\sigma^2_\psi}{\sigma^2_\psi + \bar{\sigma}^2_1} \tag{3.18}$$

Where $\bar{\sigma}^2_1 = \sum_j \omega_j \sigma^2_j$, $\omega_j = M_j / \sum_{j'} M_{j'}$, and where the demeaning of firm effects has been omitted for clarity. Equation (3.18) uses the term $\hat{\psi}_{1j(i,t)}$, implicitly assuming that for every individual i at time t, one has two firm effects estimated even though individual i is only in one out of two subsamples. The correct interpretation of $\hat{\psi}_{1j(i,t)}$ is the firm effect estimate from subsample 1 for the firm $j = j(i,t)$, irrespective whether individual i is in sample 1 or 2.

The coefficient on $\hat{\psi}_{1j}$ is thus indicative of the amount of estimation noise. Regressing $\hat{\psi}_{2j}$ on $\hat{\psi}_{1j}$ gives a regression coefficient of one if the noise $\bar{\sigma}^2_1$ is negligible, and a coefficient between zero and one otherwise. If there is no noise, , and $\hat{\gamma}$ will be close to one. The larger the overfitting, the further away from one will $\hat{\gamma}$ be. In this sense, the slope coefficient $\hat{\gamma}$ provides a simple and easily interpretable metric of the extent of overfitting.

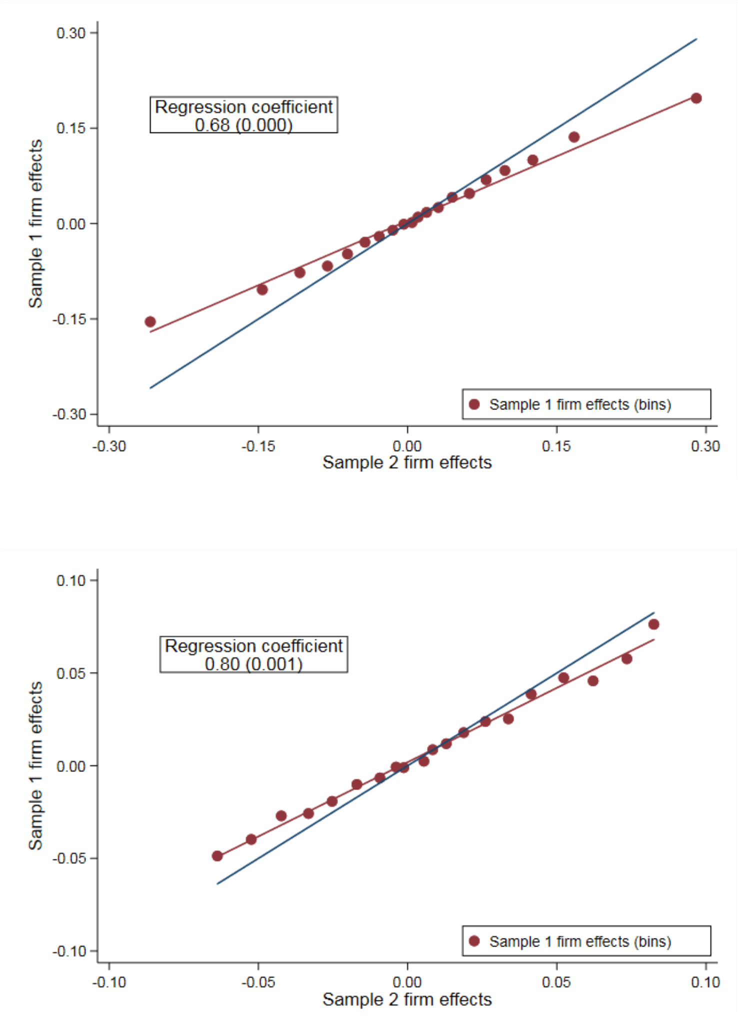
Results. Figure 3.3 provides the main result of this section. Panel (a) of figure 3.3 is a binscatter plot of firm effects from sample 1 on firm effects from sample 2. The blue line is the 45 degree line and the red line is the fit from a regression of ψ_{1j} on ψ_{2j}. Panel (a) shows that in this data set, there is clear evidence of overfitting of the firm fixed effects, as the slope coefficient is well below one at 0.68. Interestingly, the rank correlation between ψ_{1j} and ψ_{2j} is even lower, at 0.47. This suggest that the two samples not only recover firm effects of different magnitude, but also their rank in their respective sample's distribution is also very different. Most of the firm effect distribution is located between -0.10 and 0.10, with both extremity of the firm effect distribution far away from zero, suggesting that a few firms have important wage-premium - or are poorly estimated. The binned scatters are all relatively close to their fit, indicating that the overfitting does not come from a few, poorly estimated firm effects. This is supported by panel (b) of figure 3.3. Panel (b) depicts exactly the same relationship, with the exception that the top and bottom 20% of the observation-weighted distribution of firm effects are excluded. While the slope coefficient does increase, suggesting that firm effects at the extremity of the distribution are more poorly estimated than those in the middle, the slope remains well below 1, at 0.80, suggesting that overfitting is not simply an issue of extreme values of firm effects.

Appendix figure 3.B.1 provides additional insight into the extent of overfitting. Panel (a) plots firm effects from both samples when only looking at firms with more than 70 observations per year only. Just as in panel (b) of figure 3.3, the slope coefficient increases, but does not become close to one. This result is very stable across a wide range of observation restrictions, and does not depend on the precise number 70. For instance, the slope coefficient when only firms with more than 200 observations per year are considered is 0.82. Panel (b) repeats the same procedure with unweighted effects. The slope coefficient is lower than in figure 3.3, indicating that some firm effects with a small number of observations are very poorly estimated.

To investigate whether overfitting evolves over time, I divide the original sample ranging from 2002 to 2015 into two equally long time periods, 2002-2008 and 2009-2015, baring in mind that the 2008 economic crisis mostly impacts the second interval. The same sample-splitting procedure is then reproduced in both intervals, resulting in figure 3.4. The level of overfitting is constant over those two time intervals, depicted in panels (a) and (b). This is important, as it suggests that even though overfitting is sizable and leads to overestimating the contribution of firms' wage premium to the overall variance of wages, the main conclusions of Card et al. (2013) and Song et al. (2019) are likely to remain valid, as they both focus on the evolution over time of the different components of wage inequality. Strikingly, the slope coefficient decreases from 0.68 in panel (a) of figure 3.3 to 0.46 in figure 3.4. This is probably due to the number of observations being about twice as low in each sample.

Figure 3.5 provide suggestive evidence that the origin of overfitting lies in within-firm outliers. To construct this figure, the Analysis sample is first restricted to firm-year cells with more than 10 observations. Then, within every cell, the top and the bottom 20% of the wage observations is removed, such that every cell has at least 6 observations. The same sample-splitting procedure as indicated in the subsection above is then applied. Last, a two-way regression model is estimated in each subsample. Appendix table 3.A.2 provides summary statistics of this sample split and ensures that the balance of characteristics is respected. The strinking feature of this figure is that the slope coefficient is much closer to 1 than previously, at 0.90. This indicates that most of the overfitting is driven by the top and bottom of the within-firm wage distribution. This provides an interesting explanation of the source of overfitting. When high- or low-wage individuals change firms, the OLS estimators

Figure 3.3: Overfitting in the Analysis data set

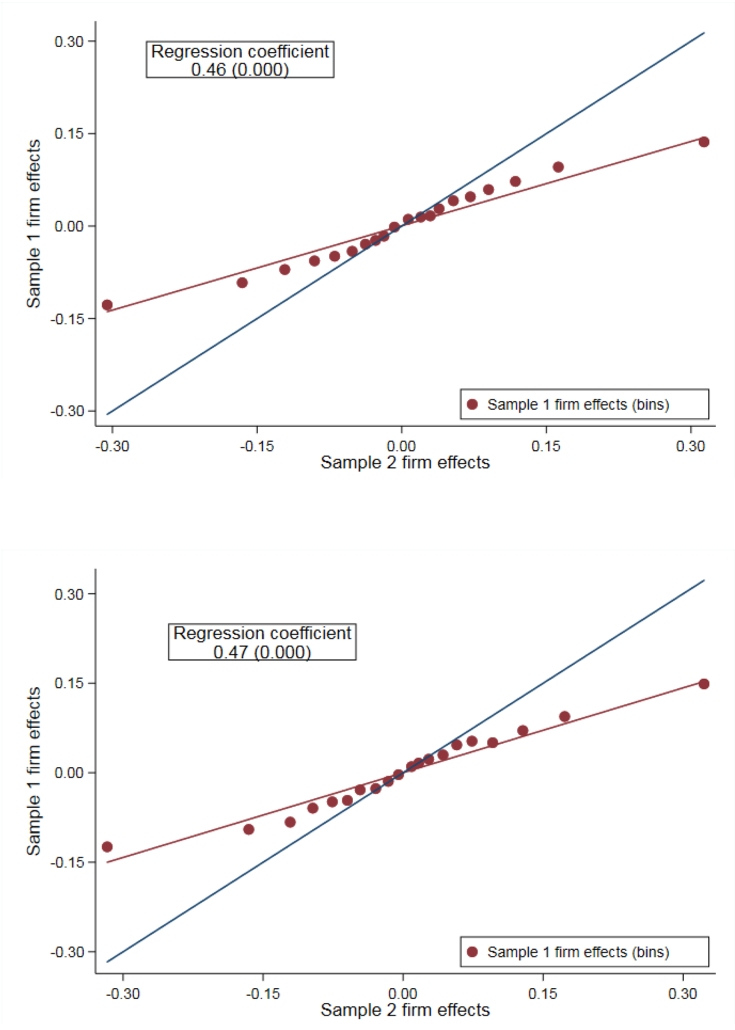


(a) Overall sample

(b) Top and Bottom 20% of firm effects excluded

Notes: Panel (a) plots a binned scatter plot of a regression of firm effects from sample 1 on firm effects from sample 2. The blue line indicates the 45 degree line and the red line is the fit from a regression of $\hat{\psi}_{1j}$ on $\hat{\psi}_{2j}$, with the regression coefficient indicated in the box. Panel (b) reproduces panel (a) but excludes the top and bottom 20% of firm fixed effects. Regressions are weighted by the number of observations. Heteroskedasticity-robust standard errors are reported. Data comes from the Analysis data set.

Figure 3.4: Overfitting over time

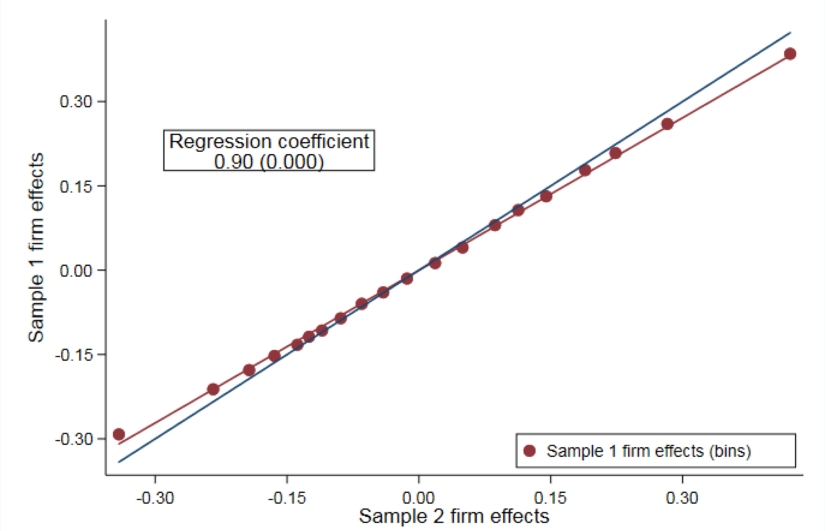


(a) 2002-2008 sample

(b) 2009-2015 sample

Notes: This figure plots a binned scatter plot of a regression of firm effects from sample 1 on firm effects from sample 2 for the years 2002-2008 (panel (a)) and 2009-2015 (panel (b)). The blue line indicates the 45 degree line and the red line is the fit from a regression of $\hat{\psi}_{1j}$ on $\hat{\psi}_{2j}$, with the regression coefficient indicated in the box. Regressions are weighted by the number of observations. Heteroskedasticity-robust standard errors are reported. Data comes from the Analysis sample, but is restricted to firm-year cells with more than 10 observations.

Figure 3.5: Overfit in the Analysis sample - Without within-firms outliers



Notes: This figure plots a binned scatter plot of a regression of firm effects from sample 1 on firm effects from sample 2 excluding within-firms outliers. The blue line indicates the 45 degree line and the red line is the fit from a regression of $\hat{\psi}_{1j}$ on $\hat{\psi}_{2j}$, with the regression coefficient indicated in the box. Regression is weighted by the number of observations. Heteroskedasticity-robust standard errors are reported.

Table 3.3: Summary of overfitting by sample

	2002-2015	2002-2008	2009-2015	No Within Firm Outliers	No Within Firm Top Outliers
Analysis sample	0.675	0.456	0.473	0.905	0.822
Top & Bottom 20% of firm effects excluded	0.802	0.726	0.728	0.936	0.878
Firms with ≥70 obs. per year	0.734	0.530	0.528	0.923	0.855
Unweighted	0.439	0.297	0.309	0.767	0.618

Notes: Regressions are weighted by the number of observations unless otherwise indicated.

incorrectly interprets the wage difference as a difference in firm effects, whereas in reality it might be only an idiosyncratic shock. Note that this interpretation does not mean that the conditional exogenous mobility assumption would be violated, as this assumption is a population, economy-wide level which does not need to hold exactly in finite sample. Table 3.3 summarizes the results. The first column shows the robustness of the overfitting in the main sample, across a range of sample restrictions. The first row shows that overfitting is constant over time and comes from the top and bottom of the within-firm salary distribution (column labeled "No Within-Firm Outliers"). The last column reruns the same analysis, but only when the top of the within-firm distribution is excluded. Interestingly, the results are almost exactly in between the first and the fourth column, suggesting that overfitting comes equally from both sides of the distribution.

External Validity. One possible concern could be that by splitting the data set into two samples, the number of observations is dramatically reduced, so that the result on overfitting here might not apply to other studies, simply because of the different size of the data sets. This seems not to be the case. In each of the subsamples, there are about 4.6 million observations and about $40,000$ firms, so a ratio of approximately 112 observations per firm fixed-effect to estimate. This ratio is only about 10 in Abowd et al. (1999), while Card et al. (2013) and Song et al. (2019) use on average 70 and 78 observations per firm fixed effect to estimate. I compare the average number of observations per firm effect to estimate instead of the more relevant ratio of number of moves on the number of firm effects simply because the above mentioned studies do not report the total number of moves.

Another threat to external validity is that workers' mobility in France is lower that in other countries, so that even if the ratio of observations per firm is higher than for other studies on different countries, the number of moves per firm might be lower. Appendix figure 3.B.2 plots the average tenure length for a number of OECD countries for the latestest available year. The average tenure length in France in 2020 was 11.1 years, higher than in some other developed countries such as Denmark, the United-Kingdom, Sweden, or the OECD average (10.1 years), but of the same order of magnitude than Germany, Spain and the European Union average. Overall, this suggests that even though worker's average mobility displays important variation between countries, France is not an outlier.

A further threat to external validity is that this paper estimates firm fixed effects (the most granular data available for a panel in France), while other studies such as Card et al. (2013) estimate establishment fixed effects. One possibility, which cannot be tested with the current available data, is that there is much less wage dispersion at the establishment level, such that for a given number of observation, an establishment-level estimate is more precise than a firm-level estimate.

3.5 Recovering signal variance and covariance

This section turns to the next part of the analysis and shows how to recover the signal variance and covariance from the sample-split.

3.5.1 Variance and covariance estimates

The strategy to recover the correct variance estimate has been exposed in section 3.2 and is discussed in more details in appendix 3.C.1. Intuitively, the variance of the firm effect is biased upward because the squared estimation error term ν_j^2 has non-zero expectation. Computing the variance of the firm effects as in (3.11) means that the ν_j^2 term is replaced

Table 3.4: Estimates of signal variance

	Weighted		Unweighted
	Variance component	Share of total	
	(1)	(2)	(3)
Analysis sample			
$\geq$ 8 obs. /year			
Raw variance $\tilde{\sigma}^2$	0.012	5.4%	0.028
Signal variance $\hat{\sigma}^2$	0.009	4%	0.017
Ratio	75%	75%	61%
Extrapolation Baseline sample			
Raw variance $\tilde{\sigma}^2$	0.032	15%	0.104
Signal variance $\hat{\sigma}^2$	0.024	11,2%	0.063
Ratio	75%	75%	61%

Notes: Weighted and unweighted refer to weights on firm effect estimates. In the first two columns, firm effects are weighted by the number of observations.

by $\nu_{1j} \cdot \nu_{2j}$, which has expectation zero. $\hat{\sigma}^2$ hence recovers the signal variance σ^2_ψ whereas $\tilde{\sigma}^2$ estimate a raw, upward-biased term.

The first panel of table 3.4 shows that in the Analysis data set, the signal variance is 25% lower that the original estimate. Just as in the previous section, this demonstrate that the extent of the bias is substantial. As a comparison, column (3) gives the raw and signal variance when firms are not weighted by the number of observations. In this case, as expected, the ratio is even lower.

Given this results on the Analysis sample, which restricts firm-year cells to have more than 8 observations, it would be interesting to see how this results extends to the baseline sample. I take the simple approach to assume that the ratio of $\hat{\sigma}^2$ to $\tilde{\sigma}^2$, which is really the signal-noise ratio, is the same in the overfit and in the baseline sample. By doing so, the signal-noise ratio in the baseline sample is likely to be underestimated. Indeed, the baseline sample include many firms with less than eight observations per year, whose effects are likely to be more poorly estimated than firms in the Analysis sample. For this reason, the raw variance estimated by $\tilde{\sigma}^2$ is likely to be higher in the baseline than in the Analysis sample. The second panel of table 3.4 provides results of this extrapolation. By assumption, the ratio is 75%, and the signal variance of firm effects now explains about 11% of the variance of wages.

Similarly, an unbiased estimate of the covariance between firm and individual effects can also be recovered using the split-sample estimates. The intuition is that $\hat{\psi}_{1j(i)} \cdot \hat{\theta}_i$ is an

Table 3.5: Estimates of signal covariance

$Cov\left(\hat{\psi}_{j(i)}, \hat{\theta}_i\right)$	0.008
$Cov\left(\hat{\psi}_{1j(i)}, \hat{\theta}_{2i}\right)$	0.010
$Cov\left(\hat{\psi}_{2j(i)}, \hat{\theta}_{1i}\right)$	0.010
$\frac{1}{2}\,Cov\left(\hat{\psi}_{1j(i)}, \hat{\theta}_{2i}\right) + \frac{1}{2}Cov\left(\hat{\psi}_{2j(i)}, \hat{\theta}_{1i}\right)$	0.010

Notes: Weighted and unweighted refer to weights on firm effect estimates. In the first column, firm effects are weighted by the number of observations. Data comes from the Analysis data set.

unbiased estimate of $\sigma_{\psi\theta}$, where $\hat{\psi}_{1j(i)}$ is the estimate of firm j from subsample 1 and $\hat{\theta}_i$ is the estimate of individual effect, who is in subsample 2. On the other hand, $\hat{\psi}_{j(i)} \cdot \hat{\theta}_i$ is biased, as the same information is used to compute both $\hat{\psi}_{j(i)}$ and $\hat{\theta}_i$. To increase efficiency, I follow Chernozhukov et al. (2018) and use cross-fitting and get two estimates of the covariance term by inverting the role of the two samples and then by averaging the estimates. Table 3.5 shows the different estimates. As mentioned in section 3.3, the covariance in the Analysis sample is positive, at 0.008. The next two rows show the split-sample unbiased estimate $\hat{\sigma}_{\psi\theta}$. Rows 2 and 3 demonstrate not difference in estimate depending on the choice of the subsample for $\hat{\psi}$ or $\hat{\theta}$. The result suggest a 25% higher covariance between individuals and firms than previously estimated. The last row demonstrates that averaging the two covariance estimates does not change the results

These results on the variance and covariance suggest that the contribution of the firm's wage premium to variation in wage is at least 25% less important than previously estimated. On the other hand, assortative matching between high-wage workers and high-paying firms is proportionally more important. Overall, it indicates that the between-firm component is more driven assortative matching than by true wage premium. Extrapolating both results to the baseline sample, this implies that variance of wage premium explains 11% of wage variation, instead of 15%; and 25% of the between-firm component, rather than 33%. This relates to the findings in Song et al. (2019), who show that most of the increase in earning inequality in the US is due to assortative matching rather than wage premiums.

3.5.2 Firm-year shocks

Firm-year shocks could contaminate the estimate $\hat{\sigma}^2$ and prevent it to perfectly recover σ^2_{ψ}. Intuitively, if all individuals in firm j at time t face a firm-year specific wage shock, $\hat{\psi}_{1j}$ and $\hat{\psi}_{2j}$ would be correlated beyond σ^2_{ψ}. Using notation of the toy model from section 3.2, we have now $u_{ki} = \epsilon_{i2} - \epsilon_{i1} + s_{j2} - s_{j1}$, where s_{jt} is the firm-year shock, and ϵ is the idiosyncratic shock, so that $\mathbb{E}\left[\hat{\psi}_{1j} \cdot \hat{\psi}_{2j}\right] \neq \sigma^2_{\psi}$ because of the squared term $(s_{j2} - s_{j1})^2$ that will appear. If firm-year shocks are indeed important, then $\hat{\sigma}^2$ will overestimate the true variance of wage premium, and hence be an upper bound. However, all other bias corrections I am aware of suffer from the same problem. Specifically, variance estimates proposed by both Kline et al. (2020) and Andrews et al. (2008) require independence between observations or between clusters of observations, and hence between individuals. Assuming firm-year shocks violates this requirement and would leads wages of different individuals within a same firm to be systematically correlated. Hence, while imperfect, the split-sample variance estimator is not less robust to firm-year shocks than other solutions suggested in the literature.

3.6 Recovering better firm effect estimates

This section shows how to recover more precise firm effect estimate from the baseline AKM regression, presents results for this data and runs robustness and validation tests.

3.6.1 Getting more precise estimate through shrinkage

The main application of two-way fixed effect regressions so far has been to decompose the variance of the dependent variable - the log gross wage - into different components which can be interpreted as reduced form parameters for wage premium, assortative matching, etc. More generally though, researchers might want to use fixed-effect estimates in another way, for instance by conducting counterfactual analysis, assuming the identification assumption holds and the estimates are causal: by how much would the wage of individual i change if a policy maker would force firm j to hire her rather than firm j', controlling for assortative matching and individual heterogeneity? In this case, researchers or policy makers would be interested in knowing with high degree of confidence the firm effects of firms j and j'. However, as shown in the previous sections, estimates like $\hat{\psi}_j$, though unbiased, are poorly estimated in the sense that they have a very high variance. Such estimate have poor predictive power and would not be of great help to the researcher. In such cases, we might be willing to accept biased estimates if the reduction in variance is sufficiently important. To achieve this, the OLS estimate $\hat{\psi}_j$ are shrunk toward zero:

$$\tilde{\psi}_j = \frac{\sigma_\psi^2}{\sigma_\psi^2 + \sigma_j^2} \cdot \hat{\psi}_j \tag{3.19}$$

Equation (3.19) can be interpreted in several different ways. First, as shown in appendix 3.C.3, $\tilde{\psi}_j$ is the best linear predictor of ψ_j given $\hat{\psi}_j$, as it minimizes the mean squared error function. As discussed in Chetty et al. (2014a), albeit in a different context, $\sigma_\psi^2/(\sigma_\psi^2 + \sigma_j^2)$ is the coefficient on $\hat{\psi}_j$ from an hypothetical regression of ψ_j on $\hat{\psi}_j$. Furthermore, (3.19) has an empirical Bayes interpretation. $\hat{\psi}_j$ can be seen as a noisy estimate of ψ_j. More specifically, assuming normality, we have $\hat{\psi}_j \mid \psi_j \sim \mathcal{N}(\psi_j, \sigma_j^2)$, $\psi_j \sim \mathcal{N}(0, \sigma_\psi^2)$, and hence $\hat{\psi}_j \sim \mathcal{N}(0, \sigma_j^2 + \sigma_\psi^2)$. Using the properties of jointly distributed normal distribution, it can be shown that $\tilde{\psi}_j$ is the posterior mean of ψ_j given prior information $\hat{\psi}_j$. The empirical

Bayes perspective follows from the fact that no prior distribution is assumed to derive the shrinkage factor, but rather it is estimated from the data. In the spirit of Efron and Morris (1973), $\tilde{\psi}_j$ can be seen as the optimal linear combination of a low-bias, high-variance estimator, $\hat{\psi}_j$, and a high-bias, low-variance estimator, the grand mean of the firm effects, normalized to zero: $\tilde{\psi}_j = (1-c) \cdot 0 + c \cdot \hat{\psi}_j$ is

In order to empirically implement equation (3.19) , σ^2_ψ and σ^2_j need to be estimated separately. In the previous sections, it is shown that σ^2_ψ can be recovered as the empirical counterpart of $Cov\left(\hat{\psi}_{1j} \cdot \hat{\psi}_{2j}\right)$, and that the plug-in variance estimate recovers $\sigma^2_\psi + \bar{\sigma}^2$, where $\bar{\sigma}^2$ is a weighted average of the firm-specific variance terms.While shrinking each OLS estimate by a constant equal to $\sigma^2_\psi/(\sigma^2_\psi + \bar{\sigma}^2)$ would still improve the mean-squared errors of the estimates, we would not be able to differentially shrink every estimate $\hat{\psi}_j$ and would thus not be able to discriminate between an already precisely estimated firm effect that does not require to be shrunk, and an poorly estimated effect, for which a high shrinkage factor drastically improves the mean-squared error.

In order to separately estimate σ^2_j from σ^2_ψ, I assume that σ^2_j can be decomposed as the product of a common variance term and a idiosyncratic, observable component. Specifically, motivated by Jochmans and Weidner (2019), I assume that:

$$\sigma^2_j = \frac{1}{M_j} \sigma^2_\nu \tag{3.20}$$

Where σ^2_ν is the common component and M_j is the number of moves from or to firm j. In the particular case of the simplified model of section 3.2, appendix 3.C.3 shows that equation (3.20) holds perfectly, but this does not need to be true in general. Section 3.6.3 provides evidence that assumption(3.20) holds very well in practice and is a very close approximation of the true firm-specific variance.

Under this assumption we have that $s^2 \equiv \sigma^2_\psi + \bar{\sigma}^2 = \sigma^2_\psi + \bar{M} \cdot \sigma^2_\nu$, where $\bar{M} = \frac{1}{N}\sum_{i,t} \frac{1}{M_{J(i,t)}}$. As $\bar{M}$ is observed, one can then estimate $\hat{\sigma}^2_\nu$ as:

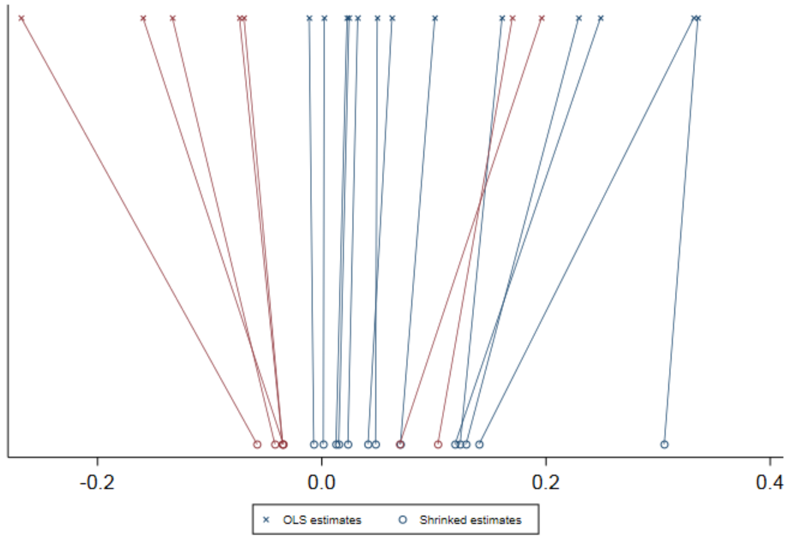
$$\hat{\sigma}^2_\nu = \bar{M}^{-1} \cdot \left(\hat{s}^2 - \hat{\sigma}^2_\psi\right) \tag{3.21}$$

One can then recover an estimate of σ^2_j as $\hat{\sigma}^2_j = \frac{1}{M_j}\hat{\sigma}^2_\nu$.

3.6.2 Shrinkage results

Figure 3.6 shows graphically the extend of shrinkage from a random subset of firms from the Analysis data set. Raw firm estimates are ordered on the top of the figure, while shrunk estimates are displayed at the bottom. The slope of the lines connecting raw and shrunk estimates varies across firms, graphically showing the differential amount of shrinkage. Figure

Figure 3.6: Random sample of firm effects before and after shrinkage



Notes: This figures plots a random sample of firm effects before (top of figure) and after shrinkage (bottom of figure). X-axis is the size of the firm effects. Red lines highlights firms classified in different deciles of the firm effect distribution when their firm effect estimate is shrunk. Data comes from the Analysis sample.

3.6 reveals the the overall significant amount of shrinkage. The mean (observation-weighted) shrinkage factor is 0.87. The figure also suggests that the larger the raw estimate in absolute value, the more important the shrinkage is likely to be. This indicates that very important firm estimates are much more likely to come from poor estimation rather than true wage premium. Appendix table 3.A.3 provides more detailed statistics about the extend of heterogeneity in the shrinkage factor. Half of the firm effects are shrunk by 38% or more. Further, the observation-weighted standard deviation is very high, at 0.16, whereas the interquartile range is of similar importance.

Shrinking estimates are not only useful to get better firm-level information, but also because it enables researchers to retrieve a better ranking of wage premiums. In figure 3.6, firms in red are those with a fixed effects from a different decile before and after shrinkage. Because some firm estimates are more shrunk than others, this changes the ranking. To get a sense of the quantitative importance of this change, table 3.6 provides information of the firm effects before and after shrinkage for firms in the Analysis dataset. Out of 40, 164 firms, 16, 506 of them, or 41% change deciles of the observation-weighted distribution of firm effects due to shrinkage. About 5, 927 of them, or 15% end up in different quartiles.

In addition, the change in mean firm effect by quartile before and after shrinkage displayed

Table 3.6: Summary Statistics of Firm Effects Before and After Shrinkage

	Raw Firm Effect	Shrunk Firm Effect
Mean FE in Q1	-0.158	-0.101
Mean FE in Q2	-0.031	-0.025
Mean FE in Q3	0.025	0.020
Mean FE in Q4	0.171	0.113
No of firms changing quartiles	5,927	
in % of firms	0.148	
in % of obs	0.056	
No of firms changing deciles	16,506	
in % of firms	0.411	
in % of obs	0.183	

Notes: All statistics are from observation-weighted distributions. Numbers reported are for firms in the Analysis sample.

in this table is substantial. As a simple example, consider an individual with an annual, real log gross wage of 10.237 before accounting for a firm's wage premium, which is the median wage in the data. This individual is randomly picked from a firm in the first quartile of the distribution, and transferred in a random firm in the fourth quartile. Assuming firm effects are unbaised, a researcher focusing on raw estimates would predict a wage increase coming from different wage premiums of about $\exp(10.237 + 0.171) - \exp(10.237 - 0.158) = 9,286$ euros. Using shrunk estimates predict a 35% lower wage differential of $6,021$ euros. Overall, it shows that the amount of shrinkage is important and matters quantitatively.

3.6.3 Validity of the assumption on the firm-specific error term

I test the validity of assumption (3.20) using out-of sample analysis. First, the original data is split using the procedure exposed in section 3.3 in three subsamples, to get three different estimates $\hat{\psi}_{1j}$, $\hat{\psi}_{2j}$ and $\hat{\psi}_{3j}$ for every firm. I then use $\hat{\psi}_{1j}$ and $\hat{\psi}_{2j}$ in order to estimate $\hat{\sigma}^2_{\psi}$ and $\hat{\sigma}^2_{\nu}$ and hence to recover shrunk estimates $\tilde{\psi}_{1j}$ and $\tilde{\psi}_{2j}$. To check the validity of the assumption, the third subsample is used as an out-of-sample dataset where the raw estimate $\hat{\psi}_{3j}$ is regressed on the shrunk estimate from the first or second subsample, say $\tilde{\psi}_{2j}$. This is because, as no data from subsample 3 have been used to estimate $\hat{\sigma}^2_{\psi}$ and $\hat{\sigma}^2_{\nu}$, these estimates are independent of $\hat{\psi}_{3j}$. Moreover, if equation (3.20) holds exactly, then the coefficient of a regression of $\hat{\psi}_{3j}$ on $\tilde{\psi}_{2j}$ is exactly one. On the other hand, if the assumption does not hold, the coefficient will recover $(\sigma^2_{\psi} + \sigma^2_j)/Var(\hat{\psi}_{2j})$, which need not be one. The closer to one the regression coefficient is, the better assumption (3.20) is.

Table 3.7 shows that equation (3.20) almost holds. The first row of both panels shows that overfitting is of the same order of magnitude as for the Analysis sample, at about 0.64. Regressing the raw firm estimate from the left-out sample on shrunk estimates from the

Table 3.7: Validity check for σ_j^2

	Sample 3 Raw estimate $\hat{\psi}_{3j}$	
	Coefficient	Standard error
Panel A: Sample 1		
Raw estimate $\hat{\psi}_{1j}$	0.637	0.000
Shrunk estimate $\tilde{\psi}_{1j}$	1.032	0.000
Panel B: Sample 2		
Raw estimate $\hat{\psi}_{2j}$	0.662	0.000
Shrunk estimate $\tilde{\psi}_{2j}$	1.038	0.000

Notes: This table summarizes the regression results of sample 3 raw estimate $\hat{\psi}_{3j}$ on raw and shrunk estimates from sample 1 (panel A) and 2 (panel B). All regressions are weighted by the number of observations. The last column displays heteroskedasticity-robust standard errors. Data comes from the Analysis sample, but is restricted to firm-year cells with at least 12 observations.

first two subsamples gives a coefficient very sligthly bigger than one and thus suggests that while the maintained assumption on σ_j^2 is not entirely true, as is to be expected, it remains accurate and a good approximation. Appendix figure 3.B.3 displays this result graphically.

3.7 Conclusion

In this paper shows that, as in other datasets (Andrews et al., 2008; Kline et al., 2020), French employer-employee data is subject to important overfitting and that the firm wage premiums are on average sizeably lower than what has been previouslt estimated. These results are established by introducing a simple split-sample procedure which allows a researcher to measure the precision of the estimation and to recover the true variance of wage premium. This procedure estimates that the contribution of firm heterogeneity to wage inequality is overestimated by at least 25%, and that the contribution of worker's sorting into firm to wage inequalities is underestimated by the same amount. This procedure also allows for a better prediction of firm effects by shrinking the original OLS estimates by a factor equal their signal to noise ratio, thus lowering the mean squared error. This paper finds that shrinkage is substantial for a significant fraction of firms, suggesting that the raw firm effect estimates are not only improper to estimate the variance of wage premium, but also to draw conclusions about firm-specific wage premiums.

Overall, these results are in line with recent evidence (Song et al., 2019; Kline et al., 2020) suggesting that, at least for some countries, the contribution of firm-specific wage premiums to inequality is lower than previously thought (Card et al., 2013), and that sorting of workers in firms plays a relatively more important role explaining this wage inequality. This paper also introduce shrinkage methods in the litterature on firm wage premium, which can be useful for future work aiming to draw economic conclusions from estimated firm fixed-effects.

Appendices

3.A Additional tables

Table 3.A.1: Decomposition of the variance of wages - Analysis sample

Decomposition *a la* Song et al. (2019)	Var. Component	Share of total
Total variance	0.225	1.000
Between firms	**0.085**	**0.38**
Var. of $\bar{\theta}^j$	0.044	0.195
Var. of Firm Effect ψ_j	0.012	0.054
Var. of $\bar{X}^j\beta$	0.005	0.022
$2\,Cov(\bar{\theta}^j, \psi_j)$	0.016	0.071
$2\,Cov(\bar{\theta}^j, \bar{X}^j\beta)$	0.005	0.021
$2\,Cov(\psi_j, \bar{X}^j\beta)$	0.003	0.011
Within firms	**0.140**	**0.62**
Var. of $\theta - \bar{\theta}^j$	0.111	0.495
Var. of $X\beta - \bar{X}^j\beta$	0.028	0.125
Var. of Residual	0.028	0.124
$2\,Cov(\theta - \bar{\theta}^j, X\beta - \bar{X}^j\beta)$	-0.027	-0.118
Number of Person effects	1,687,078	
Number of Firm Effects	41,975	
Number of Different Spells	2,983,709	
Sample size	9,220,625	

Notes: $\bar{Y^j}$ refers to the mean of variable Y taken over individuals working at firm j. The variance and covariance terms are weighted by the number of observations. Data comes from the Analysis sample.

Table 3.A.2: Summary statistics of different samples - Excluding Within-Firms Outliers

	≥10 obs./year	
	Sample 1	Sample 2
	(1)	(2)
Panel A: Individuals		
No. Individuals	644,431	644,417
Mean Wage	10.275	10.276
Q1 Wage	10.039	10.039
Median Wage	10.229	10.229
Q4 Wage	10.471	10.471
% Men	0.626	0.627
% living in IDF	0.270	0.270
Mean Age	39.0	39.0
Panel B: Firms		
No. Firms	29,688	29,693
Mean Obs/Year / Firm	9.4	9.4
Mean Obs / Firm	90.0	90.0
Mean Moves / Firm	21.7	21.7
Q1 Moves / Firm	3.0	3.0
Median Moves / Firm	7.0	7.0
Q3 Moves / Firm	15.0	15.0
Observations	2,671,489	2,673,834

Notes: Wage data is expressed in log gross real annual format. Data come from the largest connected set of the respective samples. Moves refers to the number of moves to or from a given firm. Data comes from the Analysis sample with the additional restriction that all firm-year cells have at least 10 observations.

Table 3.A.3: Summary statistics of shrinkage factor

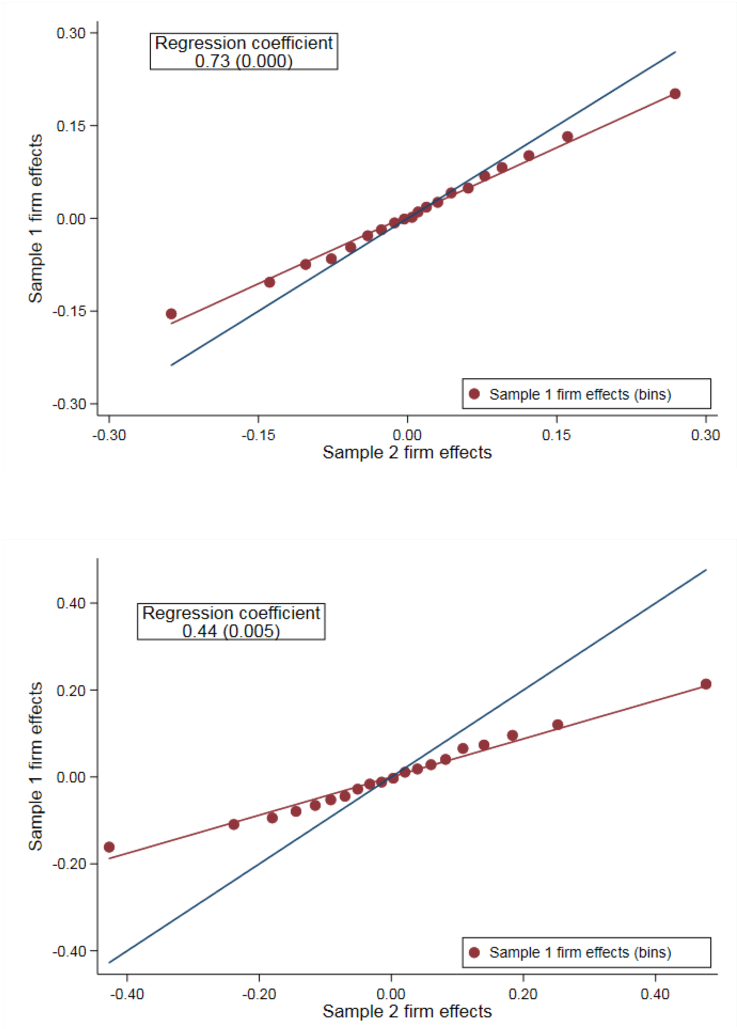
	Shrinkage Factor	
	Weighted	Unweighted
Mean	0.87	0.62
Std. Dev.	0.16	0.21
Q1	0.80	0.45
Q2	0.93	0.62
Q3	0.98	0.79
Minimum	0.15	0.15
Maximum	0.99	0.99

Notes: Weights are observations weights. Numbers reported are for the shrinkage factors of firms in the Analysis sample.

3.B Additional figures

Figure 3.B.1: Overfit in Analysis sample

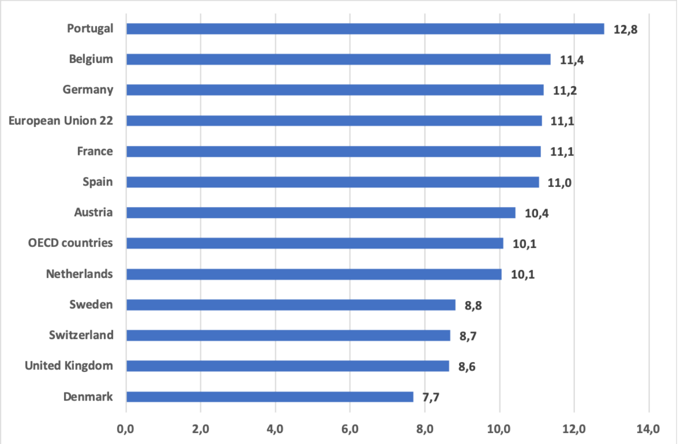
(a) Firms with more than 70 observations



(b) Unweighted firm effects

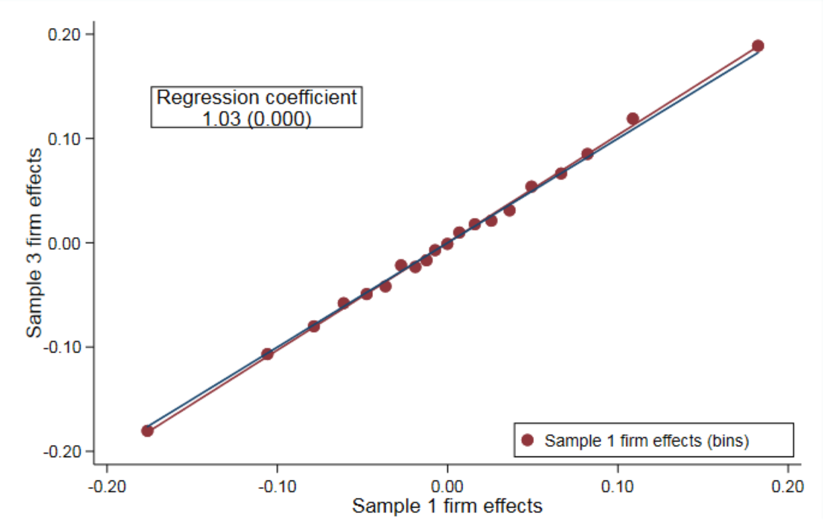
Notes: These figures plot a binned scatter plot of a regression of firm effects from sample 1 on firm effects from sample 2, as exposed in section 3.4.2. The blue line indicates the 45 degree line and the red line is the regression fit, with the regression coefficient indicated in the box. Data comes from the Analysis sample. Panel (a) is constructed using firms with at least 70 observations and regression is weighted by the number of observations, while panel (b) does not weight the firm effects by the number of observations. Heteroskedasticity-robust standard errors are reported.

Figure 3.B.2: Average tenure length for a selected OECD countries, 2020



Notes: Data comes from OECD's website - Series "Employment by job tenure intervals - average tenure", see here. Accessed 28 June 2022.

Figure 3.B.3: Regression of $\hat{\psi}_{3j}$ on $\tilde{\psi}_{2j}$



Notes: This figure plots a binned scatter plot of a regression of firm effects from sample 3 on firm effects from sample 1, as exposed in section 3.6.3. The blue line indicates the 45 degree line and the red line is regression fit, with the regression coefficient indicated in the box. Firm effects are weighted by the number of observations. Data comes from the Analysis sample, but is restricted to firm-year cells with at least 12 observations. Heteroskedasticity-robust standard errors are reported.

3.C Technical appendix

3.C.1 Statistical model

This section provides the proofs of the simplified model presented in section 3.2.

3.C.1.1 OLS estimates

Set-up and interpretation. Consider the simplified two-way fixed effects model from equation (3.2) where the time-varying individual covariates are omitted for simplicity.

$$Y = D\theta + F\psi + r$$

For instance, Y can be seen as being the log wage net of age, year, gender and experience effects. θ is the vector of individual fixed effects, ψ is the vector of firm fixed effects, D and F are the design matrices for the corresponding individual and firm effects, and r is an error term, respecting assumptions made in section 3.2.

There are I individuals indexed by i. Each individual is observed T_i times, indexed by t, so there are $N \equiv \sum_{i=1}^{I} T_i$ observations in total, indexed by k. Hence, for every observation k, there is a unique couple (i, t). In what follows, k and its corresponding individual-year equivalent are used interchangeably. There as J firms indexed by j. Using the Frisch-Waugh-Lowell theorem, we have:

$$\hat{\psi} = (\tilde{F}'\tilde{F})^{-1}\tilde{F}'\tilde{Y} \tag{3.22}$$

where $\tilde{F} = M_D F$, $\tilde{Y} = M_D Y$, and $M_D = I - D(D'D)^{-1}D'$. $\tilde{Y}$ is a vector of size N where the k-th observation is equal to the demeaned log wage with respect to the individual : $y_{i,t} - \bar{y}_{i,\cdot}$. $\tilde{F}$ is a $N \times J$ matrix whose $[k, j]$-th element is equal to

$$\begin{aligned}[\tilde{F}]_{k,j} \equiv \tilde{f}_{kj} &= f_{kj} - \frac{1}{T_i}\sum_{n=1}^{N} d_{ni} f_{nj} \\ &= f_{kj} - \bar{f}_{i,j,\cdot}\end{aligned}$$

where f_{kj} is an indicator variable equals to 1 if individual corresponding to observation k is working in firm j at the time of the observation, 0 else; d_{nj} is an indicator variable equals to 1 if observation n corresponds to individual i, 0 else. Because $T_i = \sum_{n=1}^{N} d_{ni}$, one can interpret $\bar{f}_{i,j,\cdot}$ as the fraction of the time individual i spent working at firm j. $\tilde{f}_{kj}$ takes values strictly between 1 and -1. It takes positive values when individual corresponding to observation k is working in firm j at that time, negative values else. When individual i spends most of its time in firm j, $\bar{f}_{i,j,\cdot}$ is close to one, so $\tilde{f}_{kj}$ will be close to 0 when individual works in j, and close to -1 when individual works in $j' \neq j$. Note that $\tilde{f}_{kj}$ is equal to 0 only if individual i corresponding to observation k never works for firm j, or is always observed working for firm j. Hence, $\tilde{f}_{kj} = 0$ indicates non-movers.

Case where $J = 2$. First, we focus on the case when there are only two firms. In period 1, all individuals are in the large firm, and then M_j of them move in period 2 to the firm j. The normal equations solved by the OLS estimator from (3.22) give:

$$\begin{pmatrix} \sum_k \tilde{f}_{k1}^2 & \sum_k \tilde{f}_{k1} \cdot \tilde{f}_{k2} \\ \sum_k \tilde{f}_{k1} \cdot \tilde{f}_{k2} & \sum_k \tilde{f}_{k2}^2 \end{pmatrix} \cdot \begin{pmatrix} \hat{\psi}_1 \\ \hat{\psi}_2 \end{pmatrix} = \begin{pmatrix} \sum_k \tilde{f}_{k1} \cdot y_k \\ \sum_k \tilde{f}_{k2} \cdot y_k \end{pmatrix}$$

Where we can arbitrarily label ψ_1 as the firm effect of the small firm and ψ_2 as the effect of the large firm.

$$\sum_k \tilde{f}_{k1}^2 \cdot \hat{\psi}_1 = \sum_k \tilde{f}_{k1} y_k - \hat{\psi}_2 \sum_k \tilde{f}_{k1} \cdot \tilde{f}_{k2} \tag{3.23}$$

Because we know that the firm effects are identified only relative to one particular firm in the connected set, we can normalize, one of the effect to 0 so that $\hat{\psi}_2 = 0$, as indicated in section 3.2. We thus have:

$$\hat{\psi}_1 = \frac{\sum_k \tilde{f}_{k1} \cdot y_k}{\sum_k \tilde{f}_{k1}^2} \tag{3.24}$$

For individuals who do not move to j, $f_{kj} = 0$ for both periods, hence and $\tilde{f}_{kj} = \bar{f}_{i,j,\cdot} = 0$. For individuals who move to j, we have: $f_{(i,1)j} = 0$ and $f_{(i,2)j} = 1$, so that$\tilde{f}_{(i,1)j} = -1/2$ and $\tilde{f}_{(i,2)j} = 1/2$. Plugging in (3.24), this gives :

$$\hat{\psi}_1 = \frac{\sum_{i=1}^{M_j} (-1/2)\, y_{i1} + \sum_{i=1}^{M_j} (1/2)\, y_{i2}}{\sum_{i=1}^{M_j} (-1/2)^2 + \sum_{i=1}^{M_j} (1/2)^2} = \frac{1}{M_j} \sum_{i=1}^{M_j} (y_{i2} - y_{i1}) \tag{3.25}$$

Extending to star network. The OLS estimate of ψ_j are the same in the star network as in the case $J = 2$ above as the big firm effect is normalized to zero and as every individual i moves to only one firm j. The star network example in section 3.2 is equivalent to J separate estimations of $\hat{\psi}_j$.

OLS estimate of $\hat{\theta}_i$. It follows from the OLS normal equations, noting that: $\hat{\theta} = (D'D)^{-1}D'(Y - F\hat{\psi})$. Direction of moves does not matter when there are no year effects, or that it is controlled for. When individuals move from the small firm j in period 1 to the big firm in period 2, $\hat{\psi}_j = M_j^{-1}\sum_{i=1}^{M_j}(y_{i1} - y_{i2})$. This can be seen by plugging formulas for $\tilde{f}_{kj}$. Without loss of generality, one can thus assume that moves go in only one direction.

3.C.1.2 Sample-splitting and recovering variance estimates

Equal-size subsample minimizes variance. Consider a within-firm sample-split where $M_{1j} = c{\cdot}M_j$ and $M_{2j} = (1{-}c){\cdot}M_j$. I want to choose c so as to minimize $0.5\left(Var(\hat{\psi}_{1j}) + Var(\hat{\psi}_{2j})\right) = \sigma_\psi^2 + \frac{1}{2}\frac{\sigma_u^2}{M_j}\left(\frac{1}{c} + \frac{1}{1-c}\right)$. Setting the first order condition to 0 gives $c = 1/2$.

Bias of $\tilde{\sigma}^2$ and $\hat{\sigma}^2$. First, note that we have:

$$\begin{aligned}
\mathbb{E}\left[\hat{\psi}_j^2\right] &= \mathbb{E}\left[\psi_j^2 + \nu_j^2 + 2\cdot\psi_j\cdot\nu_j\right] = \sigma_\psi^2 + \sigma_j^2\\
\mathbb{E}\left[\hat{\psi}_{1,j}\cdot\hat{\psi}_{2,j}\right] &= \mathbb{E}\left[\psi_j^2 + \nu_{1,j}\cdot\nu_{2,j} + \psi_j\cdot\nu_{1,j} + \psi_j\cdot\nu_{2,j}\right] = \sigma_\psi^2
\end{aligned}$$

So that:

$$\begin{aligned}
&\mathbb{E}\left[\tilde{\sigma}^2\right] - \sigma_\psi^2 = \frac{1}{I}\sum_j M_j(\sigma_\psi^2 + \sigma_j^2) - \sigma_\psi^2 = \frac{1}{I}\sum_j \sigma_u^2 = \frac{J}{I}\sigma_u^2\\
&\mathbb{E}\left[\widehat{\sigma^2}\right] - \sigma_\psi^2 = 0
\end{aligned}$$

Where on the first line I use the fact that $\sum_j M_j = I$ and $\sigma_u^2 = M_j\cdot\sigma_j^2$.

Variance of $\tilde{\sigma}^2$ and $\hat{\sigma}^2$. First, note we have:

$$\begin{aligned}
\mathbb{E}\left[\nu_j^4\right] &= \mathbb{E}\left[M_j^{-4}\left(\sum_{i=1}^{M_j} u_i\right)^4\right]\\
&= M_j^{-4}\sum_{i=1}^{M_j}\mathbb{E}\left[u_i^4\right] + 6\,M_j^{-4}\sum_{i=1}^{M_j}\sum_{i'\neq i}\mathbb{E}\left[u_i^2\,u_{i'}^2\right]\\
&= M_j^{-3}\mathbb{E}\left[u_i^4\right] + 6\,M_j^{-3}\left(M_j - 1\right)\left(\sigma_u^2\right)^2
\end{aligned}$$

where the second line comes from the multinomial theorem and keeping only terms whose expectation is not zero, and the third line comes from (i) the fact that u_i is independent of $u_{i'}$ when $i \neq i'$; and (ii) the assumption that the fourth moment of u_i is finite and constant across i.

Second, note that when $M_{1,j} = M_{2,j} = \frac{M_j}{2}$,

$$\begin{aligned}
\sigma_{1j}^2\,\sigma_{2j}^2 &= \frac{1}{M_{1j}\,M_{2j}} \cdot \left(\sigma_u^2\right)^2 \\
&= 4\,M_j^{-1} \cdot \left(\sigma_u^2\right)^2 \\
\sigma_{1j}^2 + \sigma_{2j}^2 &= \frac{M_{1j} + M_{2j}}{M_{1j}\,M_{2j}} \cdot \sigma_u^2 \\
&= 4\,M_j^{-1} \cdot \sigma_u^2
\end{aligned}$$

So that we can express the variance of $\hat{\psi}_j^2$ and $\hat{\psi}_{1j} \cdot \hat{\psi}_{1j}$ as:

$$\begin{aligned}
Var\left[\hat{\psi}_j^2\right] &= \mathbb{E}\left[\hat{\psi}_j^4\right] - \mathbb{E}\left[\hat{\psi}_j^2\right]^2 \\
&= \mathbb{E}\left[\psi_j^4 + 4\,\psi_j^3\,\nu_j + 6\,\psi_j^2\,\nu_j^2 + 4\,\psi_j\,\nu_j^3 + \nu_j^4\right] - \left(\sigma_\psi^2 + \sigma_j^2\right)^2 \\
&= \mathbb{E}\left[\psi_j^4 + 6\,\psi_j^2\,\nu_j^2 + \nu_j^4\right] - \left(\sigma_\psi^2 + \sigma_j^2\right)^2 \\
&= \mathbb{E}\left[\psi_j^4\right] + \mathbb{E}\left[\nu_j^4\right] + 6\,\sigma_\psi^2\,\sigma_j^2 - \left(\sigma_\psi^2 + \sigma_j^2\right)^2 \\
&= \mathbb{E}\left[\psi_j^4\right] - \left(\sigma_\psi^2\right)^2 + \mathbb{E}\left[\nu_j^4\right] - \left(\sigma_j^2\right)^2 + 4\,\sigma_\psi^2\,\sigma_j^2 \\
&= \mathbb{E}\left[\psi_j^4\right] - \left(\sigma_\psi^2\right)^2 + M_j^{-3}\,\mathbb{E}\left[u_i^4\right] + M_j^{-3}\,(M_j - 1)\left(\sigma_u^2\right)^2 - M_j^{-2}\left(\sigma_u^2\right)^2 + 4\,M_j^{-1}\,\sigma_\psi^2\,\sigma_u^2 \\
&= Var\left[\psi_j^2\right] + 4\,M_j^{-1}\,\sigma_\psi^2\,\sigma_u^2 + M_j^{-3}\,Var\left[u_i^2\right]
\end{aligned}$$

where the third line comes from the fact that terms with odd exponents have expectation zero as ψ_j and ν_j are independent and both with mean zero, the sixth line comes from the decompositions above and the fact that $\sigma_j^2 = M_j^{-1} \cdot \sigma_u^2$.

Similarly, we have:

$$\begin{aligned}
Var\left[\hat{\psi}_{1j} \cdot \hat{\psi}_{2j}\right] &= \mathbb{E}\left[\hat{\psi}_{1j}^2 \cdot \hat{\psi}_{2j}^2\right] - \mathbb{E}\left[\hat{\psi}_{1j} \cdot \hat{\psi}_{2j}\right]^2 \\
&= \mathbb{E}\left[\left(\psi_j^2 + \nu_{1,j}^2 + 2\,\psi_j\nu_{1j}\right) \cdot \left(\psi_j^2 + \nu_{2j}^2 + 2\,\psi_j\nu_{2j}\right)\right] - \left(\sigma_\psi^2\right)^2 \\
&= \mathbb{E}\left[\psi_j^4\right] + \mathbb{E}\left[\nu_{1j}^2 \cdot \nu_{2j}^2\right] + \mathbb{E}\left[\psi_j^2\nu_{1j}^2\right] + \mathbb{E}\left[\psi_j^2\nu_{2j}^2\right] - \left(\sigma_\psi^2\right)^2 \\
&= \mathbb{E}\left[\psi_j^4\right] - \left(\sigma_\psi^2\right)^2 + \sigma_{1j}^2 \cdot \sigma_{2j}^2 + \sigma_\psi^2\left(\sigma_{1j}^2 + \sigma_{2j}^2\right) \\
&= Var\left[\psi_j^2\right] + 4\,M_j^{-1}\sigma_\psi^2\,\sigma_u^2 + 4\,M_j^{-1}\,\left(\sigma_u^2\right)^2
\end{aligned}$$

where the third line comes from the fact that ψ_j, ν_{1j} and ν_{2j} are all independant with mean zero,

We can then compute the variance of the estimators:

$$\begin{aligned}
Var\left[\tilde{\sigma}^2\right] &= \frac{1}{I^2}\sum_j M_j^2\, Var\left[\hat{\psi}_j^2\right] \\
&= \frac{1}{I^2}\sum_j M_j^2 Var\left[\psi_j^2\right] + \frac{4}{I^2}\sum_j M_j\,\sigma_\psi^2\,\sigma_u^2 + \frac{1}{I^2}\sum_j M_j^{-1}\,Var\left[u_i^2\right] \\
&= Var\left[\psi_j^2\right]\,\frac{1}{I^2}\sum_j M_j^2 + \frac{4}{I}\,\sigma_\psi^2\,\sigma_u^2 + Var\left[u_i^2\right]\frac{1}{I^2}\sum_j M_j^{-1}
\end{aligned}$$

And:

$$\begin{aligned}
Var\left[\hat{\sigma}\right] &= \frac{1}{I^2}\sum_j M_j^2\, Var\left[\hat{\psi}_{1j}\cdot\hat{\psi}_{2j}\right] \\
&= \frac{1}{I^2}\sum_j M_j^2 Var\left[\psi_j^2\right] + \frac{4}{I^2}\sum_j M_j\,\sigma_\psi^2\,\sigma_u^2 + \frac{1}{I^2}4\,\left(\sigma_u^2\right)^2\sum_j 1 \\
&= Var\left[\psi_j^2\right]\,\frac{1}{I^2}\sum_j M_j^2 + \frac{4}{I}\,\sigma_\psi^2\,\sigma_u^2 + 4\,\left(\sigma_u^2\right)^2\,\frac{J}{I^2}
\end{aligned}$$

Note that $Var\left[\hat{\sigma}^2\right] > Var\left[\tilde{\sigma}^2\right]$ when M_j is large as the latter term of $Var\left[\tilde{\sigma}^2\right]$ tends towards zero, whereas the last term of $Var\left[\hat{\sigma}^2\right]$ does not. Intuitively, this increased variance is the cost paid for unbiasedness through sample-splitting.

Mean Squared Error of $\tilde{\sigma}^2$ and $\hat{\sigma}^2$. First, note that the squared bias of $\tilde{\sigma}$ is:

$$\left(\mathbb{E}\left[\tilde{\sigma}^2\right] - \sigma_\psi^2\right)^2 = \left(\sigma_u^2\right)^2\,\frac{J^2}{I^2}$$

Using the fact that the mean squared error is the sum of the variance and the squared bias, we have:

$$\begin{aligned} MSE\left(\tilde{\sigma}^2\right) &= Var\left[\psi_j^2\right] \frac{1}{I^2} \sum_j M_j^2 + \frac{4}{I} \sigma_\psi^2 \sigma_u^2 \\ &+ Var\left[u_i^2\right] \frac{1}{I^2} \sum_j M_j^{-1} + \left(\sigma_u^2\right)^2 \frac{J^2}{I^2} \end{aligned}$$

$$\begin{aligned} MSE\left(\hat{\sigma}^2\right) &= Var\left[\psi_j^2\right] \frac{1}{I^2} \sum_j M_j^2 + \frac{4}{I} \sigma_\psi^2 \sigma_u^2 \\ &+ 4\left(\sigma_u^2\right)^2 \frac{J}{I^2} \end{aligned}$$

Where the first line of both formula show the common component. It can be easily seen that a sufficient condition for $MSE\left(\tilde{\sigma}^2\right) > MSE\left(\hat{\sigma}^2\right)$ is $J > 4$, which is always satisfied in the type of data set used.

3.C.1.3 Extension to multisplits

This part generalizes the results from the previous subsection by introducing K splits where $K \geq 2$.

Set-up. Note that as we allocate randomly every mover within each firm to a sample $k \in \{1, ..., K\}$, hence we have that $K \leq \min_j M_j$.

$$\begin{aligned} \hat{\psi}_j &= \psi_j + \frac{1}{M_j} \sum_{i=1}^{M_j} u_i \equiv \psi_j + \nu_j \\ \hat{\psi}_{k,j} &= \psi_j + \frac{1}{M_{kj}} \sum_{i=1}^{M_{kj}} u_i \equiv \psi_j + \nu_{kj} \end{aligned}$$

With : $M_j = \sum_k M_{kj}$ and $M_{kj} = M_{k'j}$, $\nu_{k,j} \sim (0, \sigma_{kj}^2)$, $\psi_j \sim (0, \sigma_\psi^2)$, $u_i \sim (0, \sigma_u^2)$, ψ_j and ν_{kj} are independent for all k, as well as observations across k and across j. Following estimators are considered :

$$
\begin{aligned}
\tilde{\sigma}^2 &= \frac{1}{I}\sum_i \hat{\psi}_j^2 = \frac{1}{I}\sum_j M_j\,\hat{\psi}_j^2 \\
\hat{\sigma}_{kk'} &= \frac{1}{I}\sum_{i=1} \hat{\psi}_{kj}\cdot\hat{\psi}_{k'j} \\
\hat{\hat{\sigma}}^2 &= \frac{2}{K(K-1)}\sum_{k'\neq k}\hat{\sigma}_{kk'}
\end{aligned}
$$

Where $\binom{K}{2} = \frac{K(K-1)}{2}$ is the number of different variance estimates we have and $\hat{\psi}_j$ are defined as in equation (3.3). Hence, $\hat{\hat{\sigma}}^2$ is simple the average of all possible $\hat{\sigma}^2_{kk'}$. Because $\hat{\sigma}^2_{kk'}$ is unbiased from the previous subsection, $\hat{\hat{\sigma}}^2$ is also unbiased. However, the variance of $\hat{\hat{\sigma}}^2$ is more complex.

$$
Var\left[\hat{\hat{\sigma}}^2\right] = \frac{4}{K^2(K-1)^2}\left[\sum_{k\neq k'} Var\left(\hat{\sigma}^2_{kk'}\right) + \sum_{k\neq k'}\sum_{\substack{k\neq k'\\(kk')\neq(k'''k'')}} Cov\left(\hat{\sigma}^2_{kk'},\hat{\sigma}^2_{k''k'''}\right)\right]
$$

Computing $Var\left[\hat{\sigma}_{kk'}\right]$. First, we have that:

$$
\begin{aligned}
\sigma^2_{kj}\cdot\sigma^2_{k'j} &= \left(\sigma^2_u\right)^2\frac{1}{M_{kj}\,M_{k'j}} = \left(\sigma^2_u\right)^2\frac{K^2}{M_j^2} \\
\sigma^2_{kj}+\sigma^2_{k'j} &= \sigma^2_u\left(\frac{1}{M_{kj}}+\frac{1}{M_{k'j}}\right) = \sigma^2_u\,\frac{2\,K}{M_j}
\end{aligned}
$$

For $k\neq k'$, so that

$$
\begin{aligned}
Var\left[\hat{\psi}_{kj}\cdot\hat{\psi}_{k'j}\right] &= \mathbb{E}\left[\hat{\psi}^2_{kj}\cdot\hat{\psi}^2_{k'j}\right] - \mathbb{E}\left[\hat{\psi}_{kj}\cdot\hat{\psi}_{k'j}\right]^2 \\
&= \mathbb{E}\left[\psi_j^4\right] - \left(\sigma^2_\psi\right)^2 + \sigma^2_{kj}\cdot\sigma^2_{k'j} + \sigma^2_\psi\left(\sigma^2_{kj}+\sigma^2_{k'j}\right) \\
&= Var\left[\psi_j^2\right] + \left(\sigma^2_u\right)^2\frac{K^2}{M_j^2} + \sigma^2_\psi\,\sigma^2_u\,\frac{2\,K}{M_j}
\end{aligned}
$$

Hence:

$$Var\left[\hat{\sigma}_{kk'}\right] = Var\left[\psi_j^2\right]\,\frac{1}{I^2}\sum_j M_j^2 + 2K\,\frac{1}{I}\,\sigma_\psi^2\,\sigma_u^2 + K^2\,\frac{J}{I^2}\,\left(\sigma_u^2\right)^2$$

Computing $Cov\left(\hat{\sigma}_{kk'}^2, \hat{\sigma}_{k'''k''}^2\right)$. Suppose k,k', k'', k''' are all different integers between 1 and K. For simplicity, they are denoted 1 to 4. Then we have that:

$$\begin{aligned} Cov\left(\hat{\psi}_{1j}\hat{\psi}_{2j}, \hat{\psi}_{3j}\hat{\psi}_{4j}\right) &= Cov\left(\psi_j^2 + \nu_{1j}\,\nu_{2j} + \psi_j\,\nu_{1j} + \psi_j\,\nu_{2j}, \psi_j^2 + \nu_{3j}\,\nu_{2j} + \psi_j\,\nu_{3j} + \psi_j\,\nu_{4j}\right) \\ &= Var\left[\psi_j^2\right] \end{aligned}$$

Because of the joint independence of ψ_j and ν_{kj}.

So that,

$$Cov\left(\hat{\sigma}_{kk'}^2, \hat{\sigma}_{k'''k''}^2\right) = Var\left[\psi_j^2\right]\,\frac{1}{I^2}\sum_j M_j^2$$

Computing $Cov\left(\hat{\sigma}_{kk'}^2, \hat{\sigma}_{kk''}^2\right)$. Suppose k, k', k'', are all different integers between 1 and K. For simplicity, they are denoted 1 to 3. Then we have that:

$$\begin{aligned} Cov\left(\hat{\psi}_{1j}\hat{\psi}_{2j}, \hat{\psi}_{1j}\hat{\psi}_{3j}\right) &= Cov\left(\psi_j^2 + \nu_{1j}\,\nu_{2j} + \psi_j\,\nu_{1j} + \psi_j\,\nu_{2j}, \psi_j^2 + \nu_{1j}\,\nu_{3j} + \psi_j\,\nu_{1j} + \psi_j\,\nu_{3j}\right) \\ &= Var\left[\psi_j^2\right] + Var\left[\psi_j\,\nu_{1j}\right] \\ &= Var\left[\psi_j^2\right] + \sigma_\psi^2\,\sigma_{1j}^2 \end{aligned}$$

Because of the joint independence of ψ_j and ν_{kj}, so that,

$$\begin{aligned} Cov\left(\hat{\sigma}_{kk'}^2, \hat{\sigma}_{kk''}^2\right) &= Var\left[\psi_j^2\right]\,\frac{1}{I^2}\sum_j M_j^2 + \sigma_\psi^2\,\frac{1}{I^2}\sum_j M_j^2\,M_j^{-1}\,\sigma_u^2 \\ &= Var\left[\psi_j^2\right]\,\frac{1}{I^2}\sum_j M_j^2 + K\,\frac{1}{I}\sigma_\psi^2\,\sigma_u^2 \end{aligned}$$

Computing number of occurrences. Out of K samples, there are $\binom{K}{2}$ possible variance estimates $\hat{\sigma}_{kk'}^2$. For every possible estimates, it can covary with $\binom{K}{2} - 1$ other estimates, so there is overall $\binom{K}{2}\left(\binom{K}{2} - 1\right) = \frac{K(K-1)}{2} \cdot \frac{K(K-1)-2}{2}$ covariance terms, each of them appearing twice. Note that from all $\binom{K}{2}$ possible estimators, sample $\tilde{k}$ appears $K-1$ times. Taking any one variance estimate $\hat{\sigma}_{\tilde{k}\tilde{k}}^2$, let us focus on the $\binom{K}{2} - 1$ other possible terms it can covary with. Every other variance estimates have either 0 or 1 sample $k \in \{1, ..., K\}$ in common. Out of the $\binom{K}{2} - 1$ other possible terms, $\tilde{k}$ appears $(K-1) - 1$ times, and $\tilde{\tilde{k}}$ also appears $(K-1)-1$ times, so that there is a total of $2(K-2)$ other variance estimators which share either $\tilde{k}$ or $\tilde{\tilde{k}}$. This being true for all $\hat{\sigma}_{\tilde{k}\tilde{\tilde{k}}}$, out of the $\binom{K}{2}\left(\binom{K}{2} - 1\right)$ covariance terms, $\binom{K}{2}2(K-2)$ share one sample together (every one appearing twice) and the remaining covariance terms share no samples.

Bringing everything together. We have:

$$
\begin{aligned}
Var\left[\hat{\hat{\sigma}}^2\right] &= \frac{4}{K^2(K-1)^2}\binom{K}{2} Var\left[\hat{\sigma}^2_{kk'}\right] \\
&+ \frac{4}{K^2(K-1)^2}\binom{K}{2}\, 2 \cdot (K-2) Cov\left(\hat{\sigma}^2_{kk'}, \hat{\sigma}_{kk''}\right) \\
&+ \frac{4}{K^2(K-1)^2}\binom{K}{2}\left[\binom{K}{2} - 1 - 2(K-2)\right] Cov\left(\hat{\sigma}^2_{kk'}, \hat{\sigma}^2_{k'''k''}\right)
\end{aligned}
$$

$$
\begin{aligned}
Var\left[\hat{\hat{\sigma}}^2\right] &= \frac{2}{K(K-1)}\left[Var\left[\psi_j^2\right] \frac{1}{I^2}\sum_j M_j^2 + 2K\,\frac{1}{I}\,\sigma_\psi^2\,\sigma_u^2 + K^2\,\frac{J}{I^2}\,\left(\sigma_u^2\right)^2\right] \\
&+ \frac{4(K-2)}{K(K-1)}\left[Var\left[\psi_j^2\right] \frac{1}{I^2}\sum_j M_j^2 + K\,\frac{1}{I}\sigma_\psi^2\,\sigma_u^2\right] \\
&+ \frac{(K-2)(K-3)}{K(K-1)}\left[Var\left[\psi_j^2\right] \frac{1}{I^2}\sum_j M_j^2\right]
\end{aligned}
$$

$$
\begin{aligned}
Var\left[\hat{\hat{\sigma}}^2\right] &= Var\left[\psi_j^2\right] \frac{1}{I^2}\sum_j M_j^2 \\
&+ \left(\sigma_u^2\right)^2 \frac{2J}{I^2}\,\frac{K}{K-1} \\
&+ 4\,\sigma_u^2\,\sigma_\psi^2\,\frac{1}{I}
\end{aligned}
$$

Where the first line in the last equation pulls together the common term in $Var\left[\psi_j^2\right]$, the second line comes from the variance term and the third line comes from the covariance terms with one sample k in common and the variance term.

3.C.1.4 Bias of $\tilde{\sigma}_{\psi\theta}$ and correct covariance estimate

This section computes the bias of $\tilde{\sigma}_{\psi\theta}$ as indicated in section (3.2), and shows that $\hat{\sigma}_{\psi\theta}$ is unbiased.

Bias of $\tilde{\sigma}_{\psi\theta}$. First, note that:

$$\begin{aligned}
\mathbb{E}\left[(r_{i1}+r_{i2})\cdot\nu_j\right] &= \mathbb{E}\left[(r_{i1}+r_{i2})\cdot\frac{1}{M_j}\sum_{i'}(r_{i'2}-r_{i'2})\right]\\
&= \frac{1}{M_j}\mathbb{E}\left[(r_{i1}+r_{i2})(r_{i2}-r_{i1})\right]\\
&= \frac{1}{M_j}\mathbb{E}\left[r_{i2}^2-r_{i1}^2\right]\\
&\equiv \frac{1}{M_j}\sigma_{\tilde{r}}^2
\end{aligned}$$

Where the second line comes from the independence across i, and $\sigma_{\tilde{r}}^2 \neq 0$ as long as we do not assume homoskedasticity. We have then:

$$\begin{aligned}
\mathbb{E}\left[\hat{\theta}_i\cdot\hat{\psi}_{j(i)}\right] &= \mathbb{E}\left[\left(\theta_i+\frac{r_{i1}+r_{i2}}{2}-\frac{1}{2}\nu_j\right)\left(\psi_{j(i)}+\nu_j\right)\right]\\
&= \mathbb{E}\left[\theta_i\psi_{j(i)}+\theta_i\,\nu_{j(i)}+\frac{r_{i1}+r_{i2}}{2}\,\psi_{j(i)}+\frac{r_{i1}+r_{i2}}{2}\,\nu_{j(i)}-\frac{1}{2}\,\nu_{j(i)}\,\psi_{j(i)}-\frac{1}{2}\,\nu_{j(i)}^2\right]\\
&= \sigma_{\psi\theta}-\frac{1}{2}\sigma_{j(i)}^2+\frac{1}{2\,M_j}\sigma_{\tilde{r}}^2
\end{aligned}$$

Where the result comes from the independence between the signals ψ_j and θ_i and the errors r_i and ν_j . As a result,

$$\begin{aligned}
\mathbb{E}\left[\tilde{\sigma}_{\psi\theta}-\sigma_{\psi\theta}\right] &= -\frac{1}{2I}\sum_j\left[M_j\,\frac{\sigma_u^2}{M_j}+M_j\,\frac{\sigma_{\tilde{r}}^2}{M_j}\right]\\
&= \frac{J}{2I}\left(\sigma_{\tilde{r}}^2-\sigma_u^2\right)\\
&= -\frac{J}{I}\left(\mathbb{E}(r_{i1}^2)-Cov(r_{i1},r_{i2})\right)
\end{aligned}$$

Where the last line come from expanding the formula for $\sigma_{\tilde{r}}^2$ and σ_{u}^2. Under homoskedasticity, $\sigma_{\tilde{r}}^2$ is 0 and the bias is obviously negative.

$\hat{\sigma}_{\psi\theta}$ recovers the correct covariance. Following the same steps as before, note that:

$$\begin{aligned}
\mathbb{E}\left[\hat{\theta}_{k'i} \cdot \hat{\psi}_{kj(i)}\right] &= \mathbb{E}\left[\left(\theta_i + \frac{r_{i1}+r_{i2}}{2} - \frac{1}{2}\nu_{k'j}\right)\left(\psi_{j(i)} + \nu_{kj}\right)\right] \\
&= \mathbb{E}\left[\theta_i \psi_{j(i)} + \theta_i\, \nu_{kj(i)} + \frac{r_{i1}+r_{i2}}{2}\, \psi_{j(i)} + \frac{r_{i1}+r_{i2}}{2}\, \nu_{kj(i)}\right] \\
&\quad + \mathbb{E}\left[-\frac{1}{2}\, \nu_{kj(i)}\, \psi_{j(i)} - \frac{1}{2}\, \nu_{kj(i)} \cdot \nu_{k'j(i)}\right] \\
&= \sigma_{\psi\theta}
\end{aligned}$$

Where k, $k' \in \{1,2\}$ denote the sub-sample from which ψ and θ are estimated. Hence:

$$\mathbb{E}\left[\hat{\sigma}_{\psi\theta} - \sigma_{\psi\theta}\right] = 0$$

3.C.2 Asymptotics

This section shows that the coefficient on $\hat{\psi}_{1j}$ from a regression of $\hat{\psi}_{2j}$ on $\hat{\psi}_{1j}$ tends towards the signal - average noise ratio. I first reexpress the denominator of equation (3.18) as a weighted mean.

$$\frac{1}{N}\sum_{i,t} \hat{\psi}^2_{1j(i,t)} = \sum_j \omega_j \hat{\psi}^2_{1j}$$

where $\omega_j = M_j/N = M_j/\sum_j M_j$. To prove that it converges towards the signal-average noise ratio, I use results from the Law of Large Numbers for weighted average from Eremin (1999). Specifically, it assumes that $\hat{\psi}^2_{1j}$ are independent with finite variance, and that, for some $q > 1$,

$$\max_j\{|\omega_j|\} - \min_j\{|\omega_j|\} \leq \left(J\,(J-1)^{1-2q}\right)^{1/2q}$$

Independence and finite variance are relatively standard assumptions. While the first one holds in section 3.2 because every individual participates in the estimation of only one firm, it might not be exactly the case in reality, as an individual moving twice over a give time period participates to the estimation of the effect of two firms. However, this assumption will approximately hold given the large number of firms in an economy, as an individual would move between a very small amount of firms relative to the total number of firms in the economy. The second assumption amounts to assuming finite fourth moment for ψ_j and ν_j, and is standard for convergence proofs with unequal variance. The last assumption poses a condition on how big any given firm can be. Intuitively, the weighted sum of variances cannot be driven by only a few firms. Under those conditions, Eremin (1999) shows that

$$\sum_j \omega_j \hat{\psi}_{1j}^2 - \mathbb{E}\left[\sum_j \omega_j \hat{\psi}_{1j}^2\right] = \sum_j \omega_j \hat{\psi}_{1j}^2 - (\sigma_\psi^2 + \bar{\sigma}_1^2) \xrightarrow[J\to\infty]{p} 0$$

where $\bar{\sigma}_1^2 = \sum_j \omega_j \sigma_j^2$. The same argument demonstrates that

$$\frac{1}{N}\sum_{i,t} \hat{\psi}_{1j(i,t)} \cdot \hat{\psi}_{2j(i,t)} = \sum_j \omega_j \hat{\psi}_{1j(i,t)} \cdot \hat{\psi}_{2j(i,t)} \xrightarrow[J\to\infty]{p} \sigma_\psi^2$$

Applying Slutsky's theorem to these two results finished the proof.

3.C.3 Shrinkage

This appendix section derives claims made in section 3.6.

3.C.3.1 Interpretation of $\tilde{\psi}_j$

Minimum mean squared error. $\hat{\psi}_j$, $j \in 1, \dots, J$ estimates with error the firm effect ψ_j. More specifically, we have that $\hat{\psi}_j \mid \psi_j \sim (\psi_j, \sigma_j^2)$, $\psi_j \sim (0, \sigma_\psi^2)$, and $\hat{\psi}_j \sim (0, \sigma_j^2 + \sigma_\psi^2)$. We can thus express $\hat{\psi}_j$ and ψ_j in an additive way. $\hat{\psi}_j = \psi_j + \nu_j$, $\nu_j \sim (0, \sigma_j^2)$ and is independent from all other variables. Let $\hat{\psi} = (\hat{\psi}_1, \hat{\psi}_2, \dots, \hat{\psi}_J)$ and $\psi = (\psi_1, \psi_2, \dots, \psi_J)$. We want to find an estimator $m_j(\hat{\psi})$ of ψ, which minimizes the expected prediction error

$$MSE = \mathbb{E}\left[\sum_j \left(m_j(\hat{\psi}) - \psi_j\right)^2\right]$$

We consider the best estimator of the form $m_j(\hat{\psi}) = \alpha + \beta_j \hat{\psi}_j$. We have:

$$\begin{aligned}\mathbb{E}\left[\left(m_j(\hat{\psi}) - \psi_j\right)^2\right] &= \mathbb{E}\left[\alpha^2 + (\beta_j \hat{\psi}_j)^2 + 2\,\alpha\,\beta_j\,\hat{\psi}_j + \psi_j^2 - 2\,\alpha\,\psi_j - 2\,\beta_j\,\hat{\psi}_j \psi_j\right] \\ &= \alpha^2 + \beta_j^2\,\sigma_j^2 + (\beta_j - 1)^2\,\sigma_\psi^2\end{aligned}$$

The first order condition for α gives $\alpha^* = 0$ (because $\mathbb{E}[\psi_j] = \mathbb{E}[\hat{\psi}_j] = 0$) and the first order condition for β_j gives

$$\beta_j^* = \frac{\sigma_\psi^2}{\sigma_\psi^2 + \sigma_j^2}$$

Which is simply the signal / (signal + average noise) ratio. Hence, $\tilde{\psi}_j = m_j(\hat{\psi}) = \frac{\sigma_\psi^2}{\sigma_\psi^2 + \sigma_j^2}\,\hat{\psi}_j$ is the best-linear predictor of ψ_j given $\hat{\psi}_j$. Further, if we constrain β_j to be equal across j, we have that β^* and α^* solve $\min_{\alpha,\beta}\sum_j\left(\alpha+\beta\,\hat{\psi}_j-\psi_j\right)^2$, so they are really the solution of a hypothetical regression of ψ_j on $\hat{\psi}_j$.

Empirical Bayes. We have $\hat{\psi}_j \mid \psi_j \sim \mathcal{N}(\psi_j,\sigma_j^2)$, $\psi_j \sim \mathcal{N}(0,\sigma_\psi^2)$, and hence $\hat{\psi}_j \sim \mathcal{N}(0,\sigma_j^2+\sigma_\psi^2)$. Properties of the normal distribution imply that if x_1 and x_2 are jointly normal with mean μ_1 and μ_2, with variance σ_1^2 and σ_2^2 and covariance σ_{12}, we have $\mathbb{E}(x_2 \mid x_1)=\mu_2+\frac{\sigma_{12}}{\sigma_1^2}\left(x_1-\mu_1\right)$. Plugging-in gives the required result. In a fully Baysian framework, one would assume a distribution for the parameters μ and σ, and estimate them through maximum-likelihood. The empirical strategy in this paper can be seen as following an empirical Bayes framework, as the parameters are estimated from the data.

3.C.3.2 Rationale for the validity checks

Assumption on σ_j^2 . Consider three estimates of the same firms from different samples, $\hat{\psi}_{kj}=\psi_j+\nu_{kj}$ for $k\in\{1,2,3\}$, $\nu_{kj}\sim(0,\sigma_{kj}^2)$ and is independent of other variables. For $k'\in\{1,2\}$, consider the shrunk estimate $\tilde{\psi}_{k'j}=\frac{\sigma_\psi^2}{\sigma_\psi^2+\tilde{s}_{k'j}^2}\hat{\psi}_{k'j}$, where $\tilde{s}_{k'j}^2$ is constructed as mentioned in section 3.6. Because the shrinkage factor is constructed from sample 1 and 2 only, it is independent of the error term in sample 3. Under assumption stated in equation (3.20),$\tilde{s}_{k'j}^2=\sigma_{k'j}^2$ recovers the correct variance of $\nu_{k'j}$. We thus have that

$$\begin{aligned}
Cov\left(\hat{\psi}_{3j},\tilde{\psi}_{2j}\right) &= \frac{\sigma_\psi^2}{\sigma_\psi^2+\sigma_{2j}^2}\,Cov\left(\psi_j+\nu_{3j},\psi_j+\nu_{2j}\right)\\
&= \frac{\left(\sigma_\psi^2\right)^2}{\sigma_\psi^2+\sigma_{2j}^2}=\frac{\left(\sigma_\psi^2\right)^2}{\left(\sigma_\psi^2+\sigma_{2j}^2\right)^2}\,Var\left[\hat{\psi}_{2j}\right]\\
&= Var\left[\tilde{\psi}_{2j}\right]
\end{aligned}$$

So that the coefficient on $\tilde{\psi}_{2j}$ from a regression of $\hat{\psi}_{3j}$ on $\tilde{\psi}_{2j}$ is one. On the other hand, should we have $\tilde{s}_{k'j}^2\neq\sigma_{k'j}^2$, the coefficient of the same regression would be

$$\begin{aligned}
\frac{Cov\left(\hat{\psi}_{3j},\tilde{\psi}_{2j}\right)}{Var\left[\tilde{\psi}_{2j}\right]} &= \frac{\sigma_{\psi}^{2}}{\sigma_{\psi}^{2}+\tilde{s}_{2j}^{2}}\frac{Cov\left(\psi_{j}+\nu_{3j},\psi_{j}+\nu_{2j}\right)}{Var\left[\tilde{\psi}_{2j}\right]} \\
&= \frac{\left(\sigma_{\psi}^{2}\right)^{2}}{\sigma_{\psi}^{2}+\tilde{s}_{2j}^{2}}\frac{\left(\sigma_{\psi}^{2}+\tilde{s}_{2j}^{2}\right)^{2}}{\left(\sigma_{\psi}^{2}\right)^{2}}\frac{1}{Var\left[\hat{\psi}_{2j}\right]} \\
&= \frac{\sigma_{\psi}^{2}+\tilde{s}_{2j}^{2}}{\sigma_{\psi}^{2}+\sigma_{2j}^{2}} \neq 1
\end{aligned}$$

Thus, the closest to one this coefficient is, the more likely to hold is assumption stated in equation (3.20).

Bibliography

Abowd, John, Francis Kramarz, Paul Lengermann, and Sébastien Pérez-Duarte (2004) "Are good workers employed by good firms? A test of a simple assortative matching model for France and the United States," *Unpublished Manuscript*.

Abowd, John M, Robert H Creecy, and Francis Kramarz (2002) "Computing person and firm effects using linked longitudinal employer-employee data,"Technical report, Center for Economic Studies, US Census Bureau.

Abowd, John M, Francis Kramarz, and David N Margolis (1999) "High wage workers and high wage firms," *Econometrica*, Vol. 67, No. 2, pp. 251–333.

Abowd, John M, Kevin L McKinney, and Ian M Schmutte (2018) "Modeling endogenous mobility in earnings determination," *Journal of Business & Economic Statistics*, pp. 1–14.

Acemoglu, Daron (2002) "Directed technical change," *The Review of Economic Studies*, Vol. 69, No. 4, pp. 781–809.

Acemoglu, Daron, Philippe Aghion, Lint Barrage, and David Hemous (2019) "Climate change, directed innovation, and energy transition: The long-run consequences of the shale gas revolution," in *2019 Meeting Papers*, No. 1302, Society for Economic Dynamics.

Acemoglu, Daron, Philippe Aghion, Leonardo Bursztyn, and David Hemous (2012) "The environment and directed technical change," *American Economic Review*, Vol. 102, No. 1, pp. 131–166.

ADEME (2020a) "Documentation - AGRIBALYSE® documentation (EN)," URL: https://doc.agribalyse.fr/documentation-en/agribalyse-data/documentation.

——— (2020b) "Guide utilisateur - Documentation AGRIBALYSE® - Français," URL: https://doc.agribalyse.fr/documentation/acces-donnees.

Aghion, P, R Benabou, R Martin, and A Roulet (2021) "Environmental Preferences and Technological Choices: Is Market Competition Clean or Dirty?."

Allcott, Hunt (2011) "Consumers' perceptions and misperceptions of energy costs," *American Economic Review*, Vol. 101, No. 3, pp. 98–104.

Allcott, Hunt, Rebecca Diamond, Jean-Pierre Dubé, Jessie Handbury, Ilya Rahkovsky, and Molly Schnell (2019a) "Food Deserts and the Causes of Nutritional Inequality," *The Quarterly Journal of Economics*, Vol. 134, No. 4, pp. 1793–1844.

Allcott, Hunt and Michael Greenstone (2012) "Is there an energy efficiency gap?," *Journal of Economic Perspectives*, Vol. 26, No. 1, pp. 3–28.

Allcott, Hunt, Benjamin B Lockwood, and Dmitry Taubinsky (2019b) "Regressive Sin Taxes, with an Application to the Optimal Soda Tax," *The Quarterly Journal of Economics*, Vol. 134, No. 3, pp. 1557–1626.

Allcott, Hunt, Sendhil Mullainathan, and Dmitry Taubinsky (2014) "Energy policy with externalities and internalities," *Journal of Public Economics*, Vol. 112, pp. 72—-88.

Allcott, Hunt and Dmitry Taubinsky (2015) "Evaluating behaviorally motivated policy: Experimental evidence from the lightbulb market," *American Economic Review*, Vol. 105, No. 8, pp. 2501–2538.

Andersen, Asger Lau, Emil Toft Hansen, Niels Johannesen, and Adam Sheridan (2020) "Consumer responses to the COVID-19 crisis: Evidence from bank account transaction data," *Available at SSRN 3609814*.

Andrews, Martyn J, Len Gill, Thorsten Schank, and Richard Upward (2008) "High wage workers and low wage firms: negative assortative matching or limited mobility bias?," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, Vol. 171, No. 3, pp. 673–697.

Andrews, Martyn J, Leonard Gill, Thorsten Schank, and Richard Upward (2012) "High wage workers match with high wage firms: Clear evidence of the effects of limited mobility bias," *Economics Letters*, Vol. 117, No. 3, pp. 824–827.

Andrews, Martyn, Thorsten Schank, Richard Upward, and Others (2006) "Practical fixed-effects estimation methods for the three-way error-components model," *Stata Journal*, Vol. 6, No. 4, p. 461.

Angrist, Joshua D, Peter D Hull, Parag A Pathak, and Christopher R Walters (2017) "Leveraging lotteries for school value-added: Testing and estimation," *The Quarterly Journal of Economics*, Vol. 132, No. 2, pp. 871–919.

Baker, Scott R, Nicholas Bloom, Steven J Davis, and Stephen J Terry (2020a) "Covid-induced economic uncertainty,"Technical report, National Bureau of Economic Research.

Baker, Scott R, Robert A Farrokhnia, Steffen Meyer, Michaela Pagel, and Constantine Yannelis (2020b) "How does household spending respond to an epidemic? Consumption during the 2020 COVID-19 pandemic," *The Review of Asset Pricing Studies*, Vol. 10, No. 4, pp. 834–862.

Barth, Erling, Alex Bryson, James C Davis, and Richard Freeman (2016) "It's Where You Work: Increases in the Dispersion of Earnings across Establishments and Individuals in the United States," *Journal of Labor Economics*, Vol. 34, No. S2, pp. S67–S97.

Bartik, Alexander W, Marianne Bertrand, Zoë B Cullen, Edward L Glaeser, Michael Luca, and Christopher T Stanton (2020) "How are small businesses adjusting to COVID-19? Early evidence from a survey,"Technical report, National Bureau of Economic Research.

Beck, Günter and Xavier Jaravel (2021) "Prices and Global Inequality: New Evidence from Worldwide Scanner Data," *Available at SSRN 3671980*.

Bernheim, B Douglas and Dmitry Taubinsky (2018) "Behavioral public economics," *Handbook of Behavioral Economics: Applications and Foundations 1*, Vol. 1, pp. 381–516.

Besley, Timothy and Torsten Persson (2020) "Escaping the Climate Trap? Values, Technologies, and Politics," *Unpublished paper*.

Bloom, Nicholas, Fatih Guvenen, Benjamin S Smith, Jae Song, and Till von Wachter (2018) "The Disappearing Large-Firm Wage Premium," *AEA Papers and Proceedings*, Vol. 108, pp. 317–322.

Bonhomme, Stéphane, Thibaut Lamadon, and Elena Manresa (2019) "A distributional framework for matched employer employee data," *Econometrica*, Vol. 87, No. 3, pp. 699–739.

——— (2022) "Discretizing unobserved heterogeneity," *Econometrica*, Vol. 90, No. 2, pp. 625–643.

Bovenberg, A. Lans and Lawrence H. Goulder (2002) "Environmental Taxation and Regulation," *Handbook of Public Economics*, Vol. 3, pp. 1471–1545, DOI: 10.1016/S1573-4420(02)80027-1.

Broda, Christian and David E. Weinstein (2006) "Globalization and the Gains From Variety," *The Quarterly Journal of Economics*, Vol. 121, No. 2, pp. 541–585, URL: `https://academic.oup.com/qje/article/121/2/541/1884019`, DOI: 10.1162/QJEC.2006.121.2.541.

——— (2010) "Product Creation and Destruction: Evidence and Price Implications," *American Economic Review*, Vol. 100, No. 3, pp. 691–723, DOI: 10.1257/AER.100.3.691.

Caillaud, A (Insee) (1998) "Pour comprendre l'indice des prix," *Insee Méthodes*, Vol. 81-82, URL: `https://www.insee.fr/fr/metadonnees/source/fichier/Indice{_}des{_}prix.pdf`.

Card, David, Jörg Heining, and Patrick Kline (2013) "Workplace heterogeneity and the rise of West German wage inequality," *The Quarterly journal of economics*, Vol. 128, No. 3, pp. 967–1015.

Cavallo, Alberto (2020) "Inflation with Covid consumption baskets."

Chancel, Lucas, Thomas Piketty, Branko Milanovic, Christopher Lakner, Paul Segal, Sudhir Anand, Glenn Peters, and Robbie Andrews (2015) "Carbon and inequality: from Kyoto to Paris Trends in the global inequality of carbon emissions (1998-2013) & prospects for an equitable adaptation fund."

Chen, Haiqiang, Wenlan Qian, and Qiang Wen (2021) "The impact of the COVID-19 pandemic on consumption: Learning from high-frequency transaction data," in *AEA Papers and Proceedings*, Vol. 111, pp. 307–311.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins (2018) "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, Vol. 21, No. 1, pp. C1–C68.

Chetty, Raj, John N Friedman, Nathaniel Hendren, Michael Stepner, and Opportunity Insight Team (2020) "The economic impacts of COVID-19: Evidence from a new public database built using private sector data,"Technical report, National Bureau of Economic Research.

Chetty, Raj, John N Friedman, and Jonah E Rockoff (2014a) "Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates," *American Economic Review*, Vol. 104, No. 9, pp. 2593–2632.

——— (2014b) "Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood," *American Economic Review*, Vol. 104, No. 9, pp. 2633–2679.

Chetty, Raj and Nathaniel Hendren (2018) "The impacts of neighborhoods on intergenerational mobility I : Childhood exposure effects," *The Quarterly Journal of Economics*, Vol. 133, No. 3, pp. 1107—-1162.

Colonnelli, Emanuele, Joacim Tåg, Michael Webb, and Stefanie Wolter (2018) "A Cross-Country Comparison of Dynamics in the Large Firm Wage Premium," *AEA Papers and Proceedings*, Vol. 108, pp. 323–327.

Correia, Sergio (2016) "Linear Models with High-Dimensional Fixed Effects: An Efficient and Feasible Estimator,"Technical report, Duke University.

Davis, Steve J, John Haltiwanger, Lawrence F Katz, and Robert Topel (1991) "Wage Dispersion between and within U.S. Manufacturing Plants, 1963-86," *Brookings Papers on Economic Activity. Microeconomics*, Vol. 1991, pp. 115–200.

DellaVigna, Stefano and Matthew Gentzkow (2019) “Uniform Pricing in U.S. Retail Chains,” *The Quarterly Journal of Economics*, Vol. 134, No. 4, pp. 2011–2084, URL: https://academic.oup.com/qje/article/134/4/2011/5523148, DOI: 10.1093/QJE/QJZ019.

Diewert, W. E. (1976) “Exact and superlative index numbers,” *Journal of Econometrics*, Vol. 4, No. 2, pp. 115–145, DOI: 10.1016/0304-4076(76)90009-9.

Diewert, W E (1978) “Superlative Index Numbers and Consistency in Aggregation,” *Econometrica*, Vol. 46, No. 4, pp. 883–900, URL: https://about.jstor.org/terms.

Efron, Bradley and Carl Morris (1973) “Stein’s estimation rule and its competitors. An empirical Bayes approach,” *Journal of the American Statistical Association*, Vol. 68, No. 341, pp. 117–130.

Eremin, E V (1999) “Law of large numbers for the weighted mean,” *Measurement Techniques*, Vol. 42, No. 7, pp. 635–642.

Feng, Josh and Xavier Jaravel (2020) “Crafting Intellectual Property Rights: Implications for Patent Assertion Entities, Litigation, and Innovation,” *American Economic Journal: Applied Economics*, Vol. 12, No. 1, pp. 140–81.

Forsythe, Eliza, Lisa B. Kahn, Fabian Lange, and David Wiczer (2020) “Labor demand in the time of COVID-19: Evidence from vacancy postings and UI claims,” *Journal of Public Economics*, Vol. 189, p. 104238, DOI: 10.1016/J.JPUBECO.2020.104238.

Ganapati, Sharat, Joseph S. Shapiro, and Reed Walker (2020) “Energy Cost Pass-Through in US Manufacturing: Estimates and Implications for Carbon Taxes,” *American Economic Journal: Applied Economics*, Vol. 12, No. 2, pp. 303–42, DOI: 10.1257/APP.20180474.

Granade, Hannah Choi, Jon Creyts, Anton Derkach, Philip Farese, Scott Nyquist, and Ken Ostrowski (2009) “Unlocking energy efficiency in the US economy,” *McKinsey & Company*, URL: https://www.mckinsey.com/{~}/media/mckinsey/dotcom/client{_}service/epng/pdfs/unlockingenergyefficiency/us{_}energy{_}efficiency{_}exc{_}summary.ashx.

Gruetter, Max and Rafael Lalive (2009) “The importance of firms in wage determination,” *Labour Economics*, Vol. 16, No. 2, pp. 149–160.

Harberger, Arnold C (1964) “The measurement of waste,” *The American Economic Review*, Vol. 54, No. 3, pp. 58–76.

Hausman, Jerry A and Timothy F Bresnahan (2008) *Valuation of New Goods under Perfect and Imperfect Competition*: University of Chicago Press.

Holland, Stephen P., Erin T. Mansur, Nicholas Z. Muller, and Andrew J. Yates (2016) "Are There Environmental Benefits from Driving Electric Vehicles? The Importance of Local Factors," *American Economic Review*, Vol. 106, No. 12, pp. 3700–3729, DOI: 10.1257/AER.20150897.

ILO, IMF, OECD, UNECE, Eurostat, and World Bank (2004) *Consumer price index manual: Theory and practice*: Geneva: International Labour Office, peter hill edition, URL: `https://www.ilo.org/global/statistics-and-databases/WCMS{_}331153/lang--en/index.htm`.

Insee (2016a) "Note méthodologique - Indice des prix à la consommation,"Technical report.

——— (2016b) "Note méthodologique - Indice des prix dans la grande distribution,"Technical report.

Interagency Working Group on Social Cost of Greenhouse Gases, IWGSCGHG (2021) "Technical support document: Social cost of carbon, methane, and nitrous oxide, interim estimates under executive order 13990,"Technical report, Tech. rep., White House.

Jacobsen, Mark R, Christopher R Knittel, James M Sallee, Arthur A van Benthem, Eduardo Azevedo, Richard Blundell, Meghan Busse, Don Fullerton, Alex Gelber, Jeff Grogger, and Jean-François Houde (2020) "The Use of Regression Statistics to Analyze Imperfect Pricing Policies," *Journal of Political Economy*, Vol. 128, No. 5.

James, William and Charles Stein (1961) "Estimation with quadratic loss," in *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, Vol. 1, pp. 361–379.

Jaravel, Xavier (2019) "The Unequal Gains from Product Innovations: Evidence from the U.S. Retail Sector," *The Quarterly Journal of Economics*, Vol. 134, No. 2, pp. 715–783, URL: `https://academic.oup.com/qje/article/134/2/715/5230867`, DOI: 10.1093/QJE/QJY031.

Jaravel, Xavier and Martin O'Connell (2020a) "High-Frequency Changes in Shopping Behaviours, Promotions and the Measurement of Inflation: Evidence from the Great Lockdown," *Fiscal studies*, Vol. 41, No. 3, pp. 733–755, DOI: 10.1111/1475-5890.12241.

——— (2020b) "Real-time price indices: Inflation spike and falling product variety during the Great Lockdown," *Journal of Public Economics*, Vol. 191, p. 104270, DOI: 10.1016/J.JPUBECO.2020.104270.

Jochmans, Koen and Martin Weidner (2019) "Fixed-Effect Regressions on Network Data," *Econometrica*, Vol. 87, No. 5, pp. 1543–1560.

Juhn, Chinhui, Kevin M Murphy, and Brooks Pierce (1993) "Wage inequality and the rise in returns to skill," *Journal of Political Economy*, Vol. 101, No. 3, pp. 410–442.

Kline, Patrick, Nber Raffaele Saggio, and Mikkel Sølvsten (2020) "Leave-Out Estimation of Variance Components," *Econometrica*, Vol. 88, No. 5, pp. 1859–1898.

Lemieux, Thomas (2006) "Increasing Residual Wage Inequality: Composition Effects, Noisy Data, or Rising Demand for Skill?," *American Economic Review*, Vol. 96, No. 3, pp. 461–498.

Nevo, Aviv (2001) "Measuring market power in the ready-to-eat cereal industry," *Econometrica*, Vol. 69, No. 2, pp. 307–342.

O'connell, Martin and Kate Smith (2020) "Corrective Tax Design and Market Power," apr.

Pigou, Arthur Cecil (1920) *The economics of welfare*: London:MacMillan.

Poore, J. and T. Nemecek (2018) "Reducing food's environmental impacts through producers and consumers," *Science*, Vol. 360, No. 6392, pp. 987–992, DOI: 10.1126/SCIENCE.AAQ0216.

Sato, Kazuo (1976) "The ideal log-change index number," *The Review of Economics and Statistics*, pp. 223–228.

Seiler, Pascal (2020) "Weighting bias and inflation in the time of COVID-19: evidence from Swiss transaction data," *Swiss Journal of Economics and Statistics*, Vol. 156, No. 1, pp. 1–11.

Shapiro, Joseph S (2021) "The Environmental Bias of Trade Policy," *The Quarterly Journal of Economics*, Vol. 136, No. 2, pp. 831–886, URL: `https://academic.oup.com/qje/article/136/2/831/6039348`, DOI: 10.1093/QJE/QJAA042.

Song, Jae, David J. Price, Fatih Guvenen, Nicholas Bloom, and Till Von Wachter (2019) "Firming Up Inequality," *The Quarterly Journal of Economics*, Vol. 134, No. 1, pp. 1–50.

Stadler, Konstantin, Richard Wood, Tatyana Bulavskaya, Carl-Johan Södersten, Moana Simas, Sarah Schmidt, Arkaitz Usubiaga, José Acosta-Fernández, Jeroen Kuenen, Martin Bruckner, Stefan Giljum, Stephan Lutter, Stefano Merciai, Jannick H. Schmidt, Michaela C. Theurl, Christoph Plutzar, Thomas Kastner, Nina Eisenmenger, Karl-Heinz Erb, Arjan de Koning, and Arnold Tukker (2018) "EXIOBASE 3: Developing a Time Series of Detailed Environmentally Extended Multi-Regional Input-Output Tables," *Journal of Industrial Ecology*, Vol. 22, No. 3, pp. 502–515, URL: `https://onlinelibrary.wiley.com/doi/full/10.1111/jiec.12715https://onlinelibrary.wiley.com/doi/abs/10.1111/jiec.12715https://onlinelibrary.wiley.com/doi/10.1111/jiec.12715`, DOI: 10.1111/JIEC.12715.

Stern, Nicholas and Joseph E Stiglitz (2021) "The social cost of carbon, risk, distribution, market failures: An alternative approach."

Stiglitz, Joseph E, Nicholas Stern, Maosheng Duan, Ottmar Edenhofer, Gaël Giraud, Geoffrey M Heal, Emilio Lèbre La Rovere, Adele Morris, Elisabeth Moyer, Mari Pangestu, and Others (2017) "Report of the high-level commission on carbon prices."

Van Reenen, John (2011) "Wage inequality, technology and trade: 21st century evidence," *Labour Economics*, Vol. 18, No. 6, pp. 730–741.

Vartia, Yrjö O (1976) "Ideal log-change index numbers," *Scandinavian Journal of Statistics*, pp. 121–126.

Verdugo, Gregory (2014) "The great compression of the French wage structure, 1969–2008," *Labour Economics*, Vol. 28, pp. 131–144.

Weitzman, Martin L (2014) "Fat tails and the social cost of carbon," *American Economic Review*, Vol. 104, No. 5, pp. 544–546.

Woodcock, Simon D (2008) "Match effects,"Technical report, Simon Fraser University.