ESSAYS ON METHODS OF DEMAND AND PRODUCTION FUNCTION ESTIMATION

Ruizhi Ma

A DISSERTATION

in

Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2021

Supervisor of Dissertation

_____

Aviv Nevo

Professor of Economics

Graduate Group Chairperson

_____

Jesús Fernández-Villaverde

Professor of Economics

Dissertation Committee

Xu Cheng, Associate Professor of Economics

Jose Miguel Abito, Assistant Professor of Business Economics & Public Policy

ESSAYS ON METHODS OF DEMAND AND PRODUCTION FUNCTION ESTIMATION

To my parents, my parents-in-law and my wife.

ACKNOWLEDGMENT

I am deeply indebted to my Advisor, Professor Aviv Nevo, and my other committee members Xu Cheng and Jose Miguel Abito, for their invaluable guidance, support, patience, and encouragement. Without your help, this dissertation would not be possible. I would like to especially express my gratitude to my Advisor, Professor Nevo, who has spent an enormous amount of time and energy to help me learn how to research economics throughout the course of my graduate study. Thanks to my coauthor and committee member, Professor Abito, for guiding me to learn so much about production function estimation. I also would like to thank Professor Karun Adusumilli and Juan Camilo Castillo, and seminar participants in the Penn empirical micro lunch for useful comments in developing this dissertation.

I also want to thank my friends at Penn for their friendship and moral support: Alejandro Sanchez Becerra, Gorkem Bostanci, Brian Collopy, Joao Granja, Michal Hodor, Omer Faruk Koru, Tomas Larroucau, Desen Lin, Jinfeng Luo, Gokhan Oz, Yu Sun, Gabrielle Vasey, Sergio Villalvazo, and Betty Xiao Wang. I thank Professor Jingbo Cui, who is my undergraduate Advisor, for his continuous support.

Special thanks to my parents and my wife for their love and my parents-in-law for their encouragement and support. This dissertation is dedicated to them.

ABSTRACT

ESSAYS ON METHODS OF DEMAND AND PRODUCTION FUNCTION ESTIMATION

Ruizhi Ma

Aviv Nevo

Estimating consumer demand is fundamental to analyzing pricing decisions, welfare gains from new products, and changes in market structure. The first two chapters of this dissertation analyze econometric methods that allow researchers to estimate richer distributions of heterogeneity, and therefore more flexible demand models.

The first chapter compares several newly developed methods for estimating individual heterogeneity by Monte Carlo simulations. I use them to estimate a simplified mixed logit model without price endogeneity. I find the method from Malone et al. (2019) achieves good performances with a significantly lower computational cost, and the method from Cheng et al. (2019) is well-suited for scenarios with sparse type interactions. I provide some recommendations on how an empirical researcher could use these methods in practice.

The second chapter extends the first chapter by adapting the fixed-grid likelihood method (from Malone et al. 2019, henceforth FG) and a clustering method (from Cheng et al. 2019, henceforth CSS) to the full mixed logit model with price endogeneity and unobserved consumer demographics. I compare FG, CSS, and a conventional parametric MLE procedure using real panel data. I compare their predictions on welfare estimates in three hypothetical scenarios: a new product, a merger, and a divestiture. I find that the parametric approach distorts the welfare predictions.

The third chapter estimates how weather affects Chinese manufacturers' productivity during 1998-2007 and predicts how climate change would affect their productivity by 2040-2042. We use Abito (2020)'s production function estimation method, allowing firm-specific fixed effects in productivity. We find most industries in

our sample exhibit persistent differences in firm-level productivity, and that weather significantly affects a firm's productivity. We use the estimated effects to predict the mean productivity level in 2040-2042 for each firm. Comparing to the firms' historical mean productivity levels, we find the productivity will be lowered by about -4% by 2040-2042 on average across firms. We also find several industries where Ackerberg et al. (2015)'s productivity estimates induce significantly biased climate predictions. These industries amount to a large portion of the real value-added output, and are where the productivity heterogeneity is the largest and most persistent.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1

COMPARING RECENT METHODS TO MODEL INDIVIDUAL HETEROGENEITY

BY RUIZHI MA

1.1 Introduction

Uncovering latent individual heterogeneity is relevant to various policy- and business-oriented problems. Prominent examples include estimating substitution patterns among differentiated products, evaluating welfare after a change in market structure, predicting the profit of a new product, and simply classifying agents (consumers/firms/workers) into different groups. However, estimating such heterogeneity is often challenging for empirical researchers in real-world settings. Sometimes researchers don't have many observations for each individual. This is the case for some firm production data used in the production function estimation literature (see the empirical application in Cheng et al. 2019 and Abito 2020, and also the data in the third chapter of this dissertation). Besides, there could be multiple dimensions of heterogeneity, making it harder to estimate the distribution of heterogeneity flexibly and identify each individual's type accurately (see Cheng et al. 2019 for discussions on the case of sparse-type interactions).

The recent developments in the economic literature provide various estimators aimed at estimating multidimensional heterogeneity with short panel data (Malone et al. 2019, Cheng et al. 2019, Bonhomme et al. 2019, Bonhomme and Manresa 2015, Fox et al. 2016, Ackerberg 2009). No one estimator can be the one-size-fits-all solution for all empirical scenarios. Researchers might be interested in different aspects of the heterogeneity (e.g., overall distribution versus types of certain individuals) or might estimate the heterogeneity for different purposes (e.g., to better estimate a common parameter or make better welfare evaluations). A method might be better suited for certain purposes but not for others. Also, these methods

model heterogeneity differently and have different tuning parameters, like the number of latent types or the support of the distribution of the heterogeneity. However, it might not be clear how to choose these tuning parameters, as the heterogeneity is unobserved. Utilizing multiple methods could help choose the tuning parameters properly.

Therefore, this chapter aims to compare some of these methods in Monte Carlo experiments and offer some recommendations on how empirical researchers could use these methods in practice. I compare 5 methods: the fixed-grid likelihood (henceforth FG) estimator in Malone et al. (2019), the k-mean pre-clustering method in Bonhomme et al. (2019), the multi-dimensional clustering method from Cheng et al. (2019) (henceforth CSS), the "structural" k-means from Bonhomme and Manresa (2015) (the name of the method will become clearer in the method description section below), and the Heckman-Singer from Heckman and Singer (1984) and Train (2008) (henceforth HS).

I conduct 5 sets of Monte Carlo experiments to compare these methods in settings where individual consumers repeatedly choose from horizontally differentiated products. The performance measures will be (1) welfare gains from introducing a new product and (2) distance between the true and the estimated types of each individual. The 5 experiments differ in their true distributions of unobserved heterogeneity. The first 3 experiments are designed to compare the performances in relatively ideal settings where their assumptions are met, and tuning parameters are correctly specified. The last 2 experiments have more realistic/complicated taste distributions.

I find that no one method clearly dominates all other methods in the Monte Carlo experiments. Focusing on the last experiment, where the unobserved heterogeneity follows a mixture of normal distributions, I find that FG is speedy and has the lowest MSE in welfare gains. FG's median run time is 3 minutes with a decently dense grid, which is 4 times faster than the second-fastest method, and 24 times faster than the slowest method (HS). On the other hand, CSS produces the smallest MSE in welfare gains among the clustering methods (i.e., CSS, k-means pre-clustering, and the "structural" k-means). Overall, CSS is usually at least as good as other clustering methods. I also find the "structural" k-means is really good at estimating the common parameter, but not as good at estimating the heterogeneous parameters and the welfare gains.

The rest of this chapter is organized as follows. Section 1.2 describes the model and the estimation methods. Section 1.3 describes the simulation designs and reports the results. Section 1.4 provides some discussions on how an empirical researcher could use these methods in practice. Section 1.5 concludes.

1.2 Model and methods

The setting is a short balanced panel of I consumers, where each consumer is observed T periods (or in T markets), with $T \ll I$. Each period the consumer makes a purchase from a range of horizontally differentiated products. Product attributes vary exogenously over time. Consumers have the same form of (logit) utility function, but differ in their taste β over the attributes of the product j at time t, x_{jt}. The utility of consumer i consuming product j in period t is

$$u_{ijt} = x'_{jt}\beta_i - \alpha p_{jt} + \epsilon_{ijt}$$

where ϵ_{ijt} is an i.i.d. (across i, j, t) type-1 extreme-value random variable. Price vector in time t, $\boldsymbol{p_t}$, is the equilibrium outcome of a static Bertrand-Nash price-setting game. Each firm has one product in each market. At the beginning of time t, firms set their prices according to the expected market share of its product, and equilibrium prices are computed. Then each consumer randomly chooses a product given her choice probabilities.

This model is different from the "standard" random coefficient logit model in three ways. First, the price sensitivity coefficient (α) usually varies across individual consumers in the "standard" model. It is assumed to be common here because some of the methods to be discussed (the clustering methods) are designed to tackle situations with both common and heterogeneous parameters. Second, β_i is not a function of consumer demographic variables in this model. In the second chapter, I will allow β_i to be a function of both observed and unobserved demographics. Third, there is no unobserved product attribute in this model, so the error term is not correlated with price or features. In the "standard" random coefficient logit model there is an unobserved product feature (usually denoted "ξ_{jt}"). Since by sellers' price-setting behavior the price is correlated with the unobserved product feature, the price is correlated with the error term. Here the price is exogenous because the focus of this chapter is to estimate heterogeneity, which is a somewhat orthogonal

issue to price endogeneity. The next chapter will estimate a more realistic random coefficient logit model where the price is endogenous.

1.2.1 Method descriptions

1.2.1.1 Bonhomme et al. (2019): k-means pre-clustering

Bonhomme et al. (2019) utilize informative moments of the underlying heterogeneity to cluster individuals into different groups with a k-means algorithm. In the context of consumer choice, the k-means algorithm assumes that there are K groups among the consumers. There is preference heterogeneity across groups, but consumer preference is the same within each group. The k-means pre-clustering approach first estimates the group memberships of each consumer (i.e., classifying consumers) with only the choice data. It then treats the estimated memberships as given and uses them to estimate group-specific preference parameters.

Let c_{it} be the observed product choice of consumer i in market t, and let $\boldsymbol{c_i} = (c_{i1}, c_{i2}, \dots, c_{iT})$. The objective function of the k-means clustering here is

$$\min_{k(1),\dots,k(I),H_1,\dots,H_K} \sum_{i=1}^{I} \left(\boldsymbol{c_i} - H_{k(i)}\right)^2$$

where $k(i)$ is the group membership of firm $i, and\ H_k$ (a vector of length T) is cluster center of the cluster k. The simple k-means algorithm looks like this:

0, Initialize with a guess of group memberships and group centers.

At iteration n,

1, (assignment) Given initial values $\{k(i)^{(n-1)}\}_{i=1}^{I}\ and\ \{H_k^{(n-1)}\}_{k=1}^{K}$, assign each individual i to the group k where $\left(\boldsymbol{c_i} - H_k^{(n-1)}\right)^2$ is the smallest.

2, (update) Given $\{k(i)^{(n)}\}_{i=1}^{I}$, compute the new group centers for each group j: $H_j^{(n)} = \sum_{i\ s.t.\ k(i)=j} \boldsymbol{c}_i/n_j$, where n_j is the number of consumers currently in group j.

3, If $k(i)^{(n)} = k(i)^{(n-1)}$ for all i, stop; else repeat 1-3.

I use the R canned command *kmeans* in my estimations. It carries out the estimation using a modified version of the above algorithm, which is the algorithm of Hartigan and Wong (1979).

I petition individuals into groups using k-means, and treat individuals within a group as having the same preference parameters. Suppose now I make correct structural assumptions on the utility function and the error term. By doing so, I get individual i's log-likelihood function given data and parameters α, β_i:

$$log\left(L\left(\{c_{it}\}_{t=1,\dots,T}\middle|\{x_t, \boldsymbol{p_t}\}_{t=1,\dots,T}, \alpha, \beta_i\right)\right) \quad (1)$$

where c_{it} is the observed choice of consumer i in market t, and x_t are product features. I can also compute the overall total log-likelihood:

$$\sum_{i=1}^{I} log\left(L\left(\{c_{it}\}_{t=1,\dots,T}\middle|\{x_t, \boldsymbol{p_t}\}_{t=1,\dots,T}, \alpha, \beta_{k(i)}\right)\right) \quad (2)$$

and the likelihood for each group j:

$$\sum_{i\ s.t.\ k(i)=j} log\left(L\left(\{c_{it}\}_{t=1,\dots,T}\middle|\{x_t, \boldsymbol{p_t}\}_{t=1,\dots,T}, \alpha, \beta_{k(i)}\right)\right) \quad (3)$$

Given these notations, the algorithm I use to estimate the model with k-means pre-clustering is as follows:

0, (pre-clustering) Determine group memberships using k-means with the choice data, as described in the algorithm above. In the following the memberships are treated as given.

1, Initialize with a guess of the common parameter α.

At iteration n:

2, Given the guess of $\alpha^{(n-1)}$, estimate group-specific $\beta^{(n)}$ separately in each group by maximizing the equation (3).

3, Given $\beta^{(n)}$, find $\alpha^{(n)}$ by maximizing equation (2).

4, Evaluate convergence: stop if $|\alpha^{(n)} - \alpha^{(n-1)}| < \epsilon$ for a sufficiently small ϵ.

The advantage of such a procedure is that it only requires one run of the classification algorithm, is really fast, and does not rely on structural assumptions in the classification step. However it is less efficient than methods that make use of structural assumptions. For example, here market 1 and market 2 have different

product assortments, but the k-means procedure completely ignores such useful variation in classifying consumers into groups.

1.2.1.2 Bonhomme and Manresa (2015): "structural" k-means

The estimation method from Bonhomme and Manresa (2015) also models the heterogeneity in the form of "grouped fixed effects". Again the heterogeneous parameter β_i vary across groups, but are assumed to be the same within each group of individuals. This approach's advantage is similar to that of simple k-mean in that it pools together observationally similar individuals. Therefore, it can estimate heterogeneity even with a short panel. The "structural" k-means from Bonhomme and Manresa (2015) differs from the k-means pre-clustering algorithm in how to measure the distance between two individuals. In the k-means pre-clustering method above, the two individuals are in the same group if their observed choices are close in a simple Euclidean sense. For "structural" k-means, an individual will be classified into one group if that group's taste parameter values give this individual the highest likelihood value. By using structural assumptions on the utility function and the decision-making process, this method is more efficient than the k-means pre-clustering at classifying individuals into different groups. One cost of this advantage is that, instead of running the classification only once, the assignment of group memberships needs to be done many times in the estimation, each time with a updated value of the common parameter.

Let K be the number of groups. Again denote k_i as the group membership of individual i, β_k as the group-specific parameters of group k. The objective of the "structural" k-means from Bonhomme and Manresa (2015) is:

$$\max_{k(1),\dots,k(I),\beta_1,\dots,\beta_K,\alpha} \sum_{i=1}^{I} log\left(L\left(\{c_{it}\}_{t=1,\dots,T}\middle|\{x_t\}_{t=1,\dots,T},\alpha,\beta_{k(i)}\right)\right)$$

Therefore, this method thus estimates the group-membership and group-specific tastes recursively based on likelihood functions. For related theoretical results and examples, please see Bonhomme and Manresa (2015)[1].

[1] Bonhomme and Manresa (2015) provides consistency and asymptotic results for linear model with time-varying heterogeneous intercept.

The estimation algorithm of this method is as follows:

0, Initialize with a guess of group memberships and parameters.

At iteration n:

1, Assignment: given a guess of $\alpha^{(n-1)}$, for each i, compute Equation (1) with each β_k, and assign i to the group that gives highest log-likelihood.

2, Update β_k: given $\alpha^{(n-1)}$ and updated group memberships, for each group k, choose β_k that maximize equation (3).

3, Update α: given the updated β_k and updated memberships, find $\alpha^{(n)}$ by maximizing equation (2).

4, Repeat above until convergence.

1.2.1.3 Cheng et al. (2019): "structural" k-means with a plus

Cheng et al. (2019) refine the "structural" k-means above in multi-dimensional settings by giving multiple group memberships to an individual, one for each dimension of the heterogeneity. For example, if the consumers have different tastes over two features of a product, then CSS estimates two different memberships for a single individual, one for each feature preference. In contrast, the Bonhomme and Manresa (2015) method would allow each individual only to have one group membership, and the two group-specific parameters within that group are jointly updated using only individuals in that group.

By design, CSS has the advantage of using data from more individuals to update group-specific parameters in each dimension. Thus the corresponding estimates are more precise than those produced by the "structural" k-means in Bonhomme and Manresa (2015). This advantage is more evident when feature tastes have sparse interactions. The cost is that CSS is more computationally intense than Bonhomme and Manresa (2015).

Suppose $\beta_k \equiv (\beta_{k1}, \beta_{k2})$. Individual i has two memberships: $k1(i)$ and $k2(i)$. The estimation algorithm of this method is as follows:

0, Initialize with a guess of group memberships and parameters.

At iteration n,

1, Given $\alpha^{(n-1)}$ and $\{k1(i)^{(n-1)}, k2(i)^{(n-1)}\}_{i=1}^{I}$, update $(\beta_{k1}, \beta_{k2})_{k=1}^{K}$ jointly using equation (2).

2, Update α: given the updated β_k and updated memberships, find a new α that maximizes equation (2).

3, Update $k1(i)$: for each i, compute equation (1) with each β_{k1}, and assign i to the group with highest log-likelihood, to get $\{k1(i)^{(n)}\}_{i=1}^{I}$

4, Repeat step 2-3 to update again $(\beta_{k1}, \beta_{k2})_{k=1}^{K}$ and α, to find $\left(\beta_{k1}^{(n)}, \beta_{k2}^{(n)}\right)_{k=1}^{K}$ and $\alpha^{(n)}$.

5, Update $k2(i)$: for each i, compute equation (1) with each β_{k2}, and assign i to the group with highest log-likelihood, to get $\{k2(i)^{(n)}\}_{i=1}^{I}$

6, Compute objective value and assess convergence.

1.2.1.4 Malone et al. (2019): Fixed-grid Likelihood (FG)

The fixed-grid likelihood estimator (FG) from Malone et al. (2019) estimates multi-dimensional sprobabilities of belonging to each type. By design, it does not need to estimate group-specific parameters (i.e., type values). Therefore, it is crucial for this method to properly specify the type grids: the grids need to be wide enough to cover the whole support of the unobserved heterogeneity and dense enough within the support.

The method starts by creating grids in each dimension of the heterogeneity, followed by simulating each type of consumer's behaviors on the grid. It then computes the likelihoods of observing each consumer's actual behaviors under each type and uses these likelihoods to compute posterior type probabilities for each consumer, with a uniform prior.

As mentioned above, the best practice for using FG is to "saturate" the parameter space by making the grid for type values cover enough range and making the grid fine enough along each dimension of heterogeneity. When there is no homogeneous parameters, FG is most advantageous since the evaluation only needs to be done once for each type in the estimation. Unfortunately, this is not the case here. In my estimation algorithm,

the posterior type probabilities are computed for each iteration of the homogeneous parameters in the optimization problem. However even in this case, FG still shows a considerable computational advantage over other methods in terms of estimation time.

To be specific, let $v_1, \ldots, v_M$ be the fixed grid for unobserved heterogeneity parameter β_i. Again let $L\left(\{c_{it}\}_{t=1,\ldots,T} \middle| \{x_t, \boldsymbol{p_t}\}_{t=1,\ldots,T}, \alpha, \mathrm{v}_m\right)$ be the likelihood of observing individual i's outcome $\boldsymbol{c_i}$ (choices), given the observed features and prices of all available products, a guess of the common parameter, and the individual's type v_m for the unobserved demographics. Assuming a uniform prior $\{\pi_m\}_{m=1}^M$ common to all individuals, the probabilities of the individual i belonging to type v_m is given by the Bayes rule:

$$P_v\left(v_m \middle| \{c_{it}, x_t, \boldsymbol{p_t}\}_{t=1,\ldots,T}; \alpha, v_1, \ldots, v_M\right) = \frac{L\left(\{c_{it}\}_{t=1,\ldots,T} \middle| \{x_t, \boldsymbol{p_t}\}_{t=1,\ldots,T}, \alpha, \mathrm{v}_m\right)\pi_m}{P(\{c_{it}, x_t, \boldsymbol{p_t}\}_{t=1,\ldots,T} | \alpha)}$$

$$= \frac{L(\{c_{it}\}_{t=1,\ldots,T} | \{x_t, \boldsymbol{p_t}\}_{t=1,\ldots,T}, \alpha, \mathrm{v}_m)\pi_m}{\sum_{m=1}^M L(\{c_{it}\}_{t=1,\ldots,T} | \{x_t, \boldsymbol{p_t}\}_{t=1,\ldots,T}, \alpha, \mathrm{v}_m)\pi_m} \quad (4)$$

Let y_{ijt} be an indicator variable for consumer i choosing product j in market t. The objective of FG is thus

$$\alpha = argmax \sum_{i,j,t} y_{ijt} \log\left(\sum_{m=1}^M \frac{e^{U_{ijt,m}}}{\sum_{k \in C_t} e^{U_{ikt,m}}} P_v\left(v_m \middle| \{c_{it}, x_t, \boldsymbol{p_t}\}_{t=1,\ldots,T}; \alpha, \boldsymbol{v}\right)\right) \quad (5)$$

where

$$U_{ijt,m} = x_{jt}' \mathrm{v}_m - \alpha p_{jt}$$

The algorithm of FG is as follows:

0, Fix $v_1, \ldots, v_M$. Initialize with a guess of parameters $\alpha^{(0)}$.

At iteration n,

1, Given the current values $\alpha^{(n-1)}$, for each consumer i, compute the posterior type probabilities using (4).

2, Find $\alpha^{(n)}$ by equation (5), using the computed posterior type probabilities.

3, Evaluate convergence. If convergence is not achieved, repeat 1-3.

1.2.1.5 Heckman-Singer

Heckman-Singer assumes the unobserved heterogeneity has a discrete distribution. The version of Heckman-Singer used here is from Train (2008), which aims at estimating both the type weights (i.e., shares of population that are of each type) and type values (i.e., the values of the discrete points as the support of the distribution). To be specific, let there be M discrete points for the distribution of the unobserved heterogeneity: $v_1, \ldots, v_M$. Let s_m be the share of type m in the population. Given these notations, HS solves the following problem:

$$(\alpha, v_1, \ldots, v_M, s_1, \ldots, s_M) = argmax \sum_i \log\left(\sum_{m=1}^{M} \prod_{jt}\left(\frac{e^{U_{ijt,m}}}{\sum_{k\in C_t} e^{U_{ikt,m}}}\right)^{y_{ijt}} s_m\right) \quad (6)$$

where

$$U_{ijt,m} = x'_{jt}\mathrm{v}_m - \alpha p_{jt}$$

Given a guess of α, I use the expectation-maximization (EM) algorithm described in Train (2008) to carry out the estimation. Given α, the EM recursion solves

$$(v_1, \ldots, v_M, s_1, \ldots, s_M) = argmax \sum_i \sum_{m=1}^{M} P_{im} \log\left(\prod_{jt}\left(\frac{e^{U_{ijt,m}}}{\sum_{k\in C_t} e^{U_{ikt,m}}}\right)^{y_{ijt}} s_m\right)$$

where P_{im} is posterior probability of consumer i belonging to type m:

$$P_{im} = P_v\left(v_m \middle| \{c_{it}, x_t, \boldsymbol{p_t}\}_{t=1,\ldots,T}; \alpha, v_1, \ldots, v_M, s_1, \ldots, s_M\right) = \frac{L\left(\{c_{it}\}_{t=1,\ldots,T} \middle| \{x_t, \boldsymbol{p_t}\}_{t=1,\ldots,T}, \alpha, \mathrm{v}_m\right) s_m}{\sum_{m=1}^{M} L\left(\{c_{it}\}_{t=1,\ldots,T} \middle| \{x_t, \boldsymbol{p_t}\}_{t=1,\ldots,T}, \alpha, \mathrm{v}_m\right) s_m} \quad (7)$$

Note that the prior here is the estimated population shares, $s_1, \ldots, s_M$, instead of a uniform prior as in FG.

The EM algorithm is as follows[2]:

0, Choose initial values $\left(v_1^{(0)}, \ldots, v_M^{(0)}, s_1^{(0)}, \ldots, s_M^{(0)}\right)$.

At iteration n,

1, Compute posterior probabilities for each consumer i, $P_{im}^{(n)}$, using equation 7.

[2] See page 46-47 of Train (2008).

2, Update the type shares for each type m: $s_m^{(n)} = \frac{\sum_i P_{im}^{(n)}}{\sum_{m'} \sum_i P_{im'}^{(n)}}$.

3, Update type values for each type m: $v_m^{(n)}$ is computed by estimating a simple logit using the full sample, but with each individual's likelihood weighted by the individual's probability of belonging to type m, $P_{im}^{(n)}$.

4, Evaluate convergence. If not, repeat 1-4.

In my case here, there is also a common parameter α. I estimate the parameter α in an outer loop by maximizing (6). In the inner loop, given a value of α, I estimate the heterogeneous parameters using the EM algorithm I just described. In practice, I run a simple grid-search (the grid consists of 11 points tightly around the true value) for the common parameter in the outer loop, instead of a full maximization. The reason for this is simply that it would be too time-consuming to do a full maximization for the common parameter for Heckman-Singer. This choice of estimation procedure makes HS performs better than it would be in a more realistic setting where the true value of the homogeneous parameter is unknown.

Heckman-Singer quickly becomes infeasible when there are more dimensions (even a small number), as there will be too many types and thus too many parameters to estimate. Other methods above are more feasible than Heckman-Singer in higher-dimensions.

1.3 Monte Carlo Simulation

I present 5 Monte Carlo experiments. In the first 3 experiments, the true distribution of heterogeneous parameters is discrete, in the 4th one, the true distribution is a Gaussian distribution, and in the last one, the distribution is a mixture of three Gaussian distributions. In particular, experiment 1 sets the baseline (best possible) performance of these 5 methods when all tuning parameters (number of clusters and/or grid points) of each method are set correctly; experiment 2 shows that CSS is better than k-means methods if there are groups with only a few members (i.e., if there is sparse type intersection); experiment 3 shows that FG may perform better than CSS when its total number of groups (i.e. the number of groups in each dimension

multiplied by the number of dimensions) is comparable with the number of individuals. The last 2 experiments compare the performances of these methods in more realistic settings.

1.3.1 Simulation design

As a quick reminder, the model is a short balanced panel of I consumers in a market, where each consumer is observed T periods ($T \ll I$). Each period the consumer makes a purchase from a range of horizontally differentiated products. Product attributes vary exogenously over time. Consumers have the same form of (logit) utility function, but differ in their taste β over the attributes of the product j at time t, x_{jt}. The utility of consumer i consuming product j in period t is

$$u_{ijt} = x'_{jt}\beta_i - \alpha p_{jt} + \epsilon_{ijt}$$

where ϵ_{ijt} is an i.i.d. (across i, j, t) type-1 extreme-value random variable. There is no unobserved product features, and prices and features are not correlated with the error term.

Let the dimension of non-price attributes be K. I choose $K = 2,\ I = 1000$ and $T = 50$. Also, the number of product in the market $J = 5$. In terms of individual heterogeneity, all consumers have the same price coefficient $\alpha = 2$. β will have several different distributions in different experiments, the details of which will be provided in the follow subsections for each specific experiments.

On the supply side, the two product attributes are drawn from two correlated standard normal distribution, with a covariance of 0.5. Marginal cost is the sum of deterministic cost $x'_{jt}c$ plus a disturbance w, which is drawn i.i.d from a standard normal distribution. The parameter $c = (0.5,0.5)'$. Price vector in time t, p_t, is the equilibrium outcome of a static Bertrand-Nash price-setting game. Each firm has one product in each market. At the beginning of time t, firms set their prices according to the expected market share of its product, and equilibrium prices are computed. Then each consumer randomly chooses a product given her choice probabilities.

I compare these methods by their (1) price coefficient estimates, (2) estimates of the first non-price coefficient, and (3) estimates of the welfare gains of the new product with attributes $\tilde{\boldsymbol{x}}_{\boldsymbol{jt}} = (-0.5,1)'$. In each of the experiments, 100 runs are conducted for each method. The welfare statistics are based on the middle 90% (i.e., 5% to 95%) of the simulation results because welfare estimates tend to have extreme outliers.

1.3.2 Experiment 1: baseline (best) performances

Consumers have heterogeneous preferences over the two non-price features of the product. There are four distinctive types of $\beta \equiv (\beta_1, \beta_2)$: (2,0), (2, -2), (0, 0), (0, -2), and their values and fractions in the population is given in Table 1.1 below. All methods have the correct tuning parameters (i.e., grid points and/or the number of groups).

Table 1.1: Experiment 1 type distribution

		β_1	
		2	0
β_2	0	0.2	0.3
	-2	0.4	0.1

In this experiment, I expect to see slightly better performance FG than the other 3 methods because the latter 3 have to estimate the grid points, which is taken as given correctly for the former 2 methods in this experiment. Besides, among the latter 3 (clustering-based) methods, I expect to see CSS to be at least as good as OB-based k-means, as the former is designed to be a refinement of the latter.

Table 1.2 reports the experiment results. The first thing to notice is that FG indeed has the best performance across all three performance measures in MSE. It is especially noticeable that it gets the heterogeneous parameters exactly right, which is not surprising given that it has the correct grid choices. The second best estimator here is CSS, which out-performs the structural" k-means, as expected. Also, the "structural" k-means out-performs k-means pre-clustering in measuring both the first non-price coefficient and the welfare gains, which is not surprising given that the structural assumptions of the "structural" k-means are in fact

correct here, giving it higher efficiency compared to the k-means pre-clustering. In sum, this experiment confirms the basic theoretical predictions regarding the performances of these methods.

Table 1.2: Experiment 1 results

	Price coefficient			
	Mean Bias	Median Bias	S. D.	MSE
Heckman-Singer	-0.0020	0.0000	0.0100	0.0001
FG	-0.0001	-0.0015	0.0083	0.0001
"Structural" k-means	0.0245	0.0136	0.0365	0.0019
K-means pre-clustering	0.0130	0.0038	0.0286	0.0009
CSS	0.0015	0.0011	0.0100	0.0001
	First Non-price coefficient			
	Mean Bias	Median Bias	S. D.	MSE
Heckman-Singer	0.0038	0.0012	0.0039	0.0015
FG	0.0000	0.0000	0.0000	0.0000
"Structural" k-means	-0.0165	-0.0153	0.0833	0.0072
K-means pre-clustering	0.0035	-0.0004	0.0908	0.0082
CSS	0.0017	0.0025	0.0164	0.0003
	Welfare gain			
	Mean Bias	Median Bias	S. D.	MSE
Heckman-Singer	-0.0210	-0.0110	0.0914	0.0088
FG	-0.0007	-0.0007	0.0033	0.0000
"Structural" k-means	0.0094	0.0104	0.0387	0.0016
K-means pre-clustering	-0.0163	-0.0051	0.0493	0.0027
CSS	-0.0028	-0.0010	0.0143	0.0002

1.3.3 Experiment 2: sparse type intersections

In the second experiment, there are again four distinctive types of $\beta \equiv (\beta_1, \beta_2)$: $(2,0), (2,-2), (0,0), (0,-2)$, and their values and fractions in the population is given in the Table 1.3 below.

Table 1.3: Experiment 2 type distribution

		β_1 = 2	β_1 = 0
β_2	0	0.2	0.31
	-2	0.48	0.01

In particular, note that only 1% of the population is of the last type (-2, 0). This will give the "structural" k-means and k-means pre-clustering a hard time estimating that type, and as a consequence, will affect the estimation of price coefficient. However, CSS will still fair well by design. On the other hand, FG is still expected to be at least as good as CSS, as they are given the correct values for grid points, which are to be estimated for the HS, k-means, and CSS methods.

Table 1.4: Experiment 2 results

	Price coefficient			
	Mean Bias	Median Bias	S. D.	MSE
Heckman-Singer	-0.0039	0.0000	0.0119	0.0002
FG	-0.0027	-0.0032	0.0084	0.0001
"Structural" k-means	0.0116	0.0107	0.0158	0.0004
K-means pre-clustering	0.0163	0.0122	0.0270	0.0010
CSS	-0.0004	-0.0001	0.0104	0.0001
	First Non-price coefficient			
	Mean Bias	Median Bias	S. D.	MSE
Heckman-Singer	0.0015	0.0016	0.049	0.0024
FG	0.0000	0.0000	0.0000	0.0000
"Structural" k-means	-0.0114	-0.0242	0.2092	0.0439
K-means pre-clustering	-0.0019	-0.0117	0.2777	0.0771
CSS	0.0027	0.0002	0.0168	0.0003
	Welfare gain			
	Mean Bias	Median Bias	S. D.	MSE
Heckman-Singer	0.0025	0.0000	0.0429	0.0018
FG	-0.0016	-0.0015	0.0035	0.0000
"Structural" k-means	0.0010	0.0046	0.0293	0.0009
K-means pre-clustering	-0.0061	-0.0071	0.0230	0.0006
CSS	0.0007	0.0032	0.0159	0.0003

Table 1.4 reports the experiment results. FG still has the best performance across all three performance measures in terms of MSE. However, CSS seems to have less biased estimates for both the price coefficient and the welfare predictions. The second best estimator in terms of MSE here is again CSS. The CSS significantly out-performs the structural" k-means in all measures of performance, which is exactly as expected. Comparing Table 1.4 and Table 1.2, I find that the two k-means methods' performance significantly deteriorated due to the sparse type interactions. On the other hand, the "structural" k-means out-performs k-means pre-clustering in the MSE of measuring the coefficients, but not the MSE of welfare gains. The latter is actually because "structural" k-means give a slightly larger standard deviation. However, "structural" k-means produces a much smaller mean bias compared to the simple k-means.

1.3.4 Experiment 3: more types

This experiment is designed to compare FG and CSS further when there are more types (so more estimation burden for CSS). In this sub-experiment, there are 10 types in each dimension of preference for non-price attributes, thus 100 types. The de-meaned values of these taste parameters are uniformly distributed on [−5, 5] in each dimension. In this experiment, FG has the correct grid choices, and CSS has the correct number of groups.

Table 1.5 reports the experiment 3 results. The FG method out-performs CSS in almost all measures, except for the standard deviation of the price coefficient, which is very close across the two methods. Interestingly, although the mean biases of the coefficients are relatively close for the two methods, FG actually performs much better for the bias of the welfare gains. On the other hand, although there are 100 types for the 1000 consumers, which in the first glance seems challenging for the CSS to estimate group-specific values, it is actually not. This is because CSS assigns different group memberships for an individual in different dimensions of heterogeneity. This effectively increases the number of individuals within each group from a magnitude of 10 to 100.

Table 1.5: Experiment 3 results

	Price coefficient			
	Mean Bias	Median Bias	S. D.	MSE
Heckman-Singer	-0.0023	0.0000	0.0185	0.0003
FG	0.0552	0.0444	0.0198	0.0034
"Structural" k-means	-0.0114	-0.0102	0.0168	0.0004
K-means pre-clustering	0.0674	0.0677	0.017	0.0048
CSS	-0.0614	-0.0615	0.0177	0.0041
	First Non-price coefficient			
	Mean Bias	Median Bias	S. D.	MSE
Heckman-Singer	0.0842	0.1258	0.6676	0.4528
FG	-0.0741	0.1111	0.6415	0.4170
"Structural" k-means	0.0385	-0.015	0.8216	0.6765
K-means pre-clustering	0.1013	0.0474	0.9387	0.8914
CSS	0.0870	0.1928	0.7000	0.4976
	Welfare gain			
	Mean Bias	Median Bias	S. D.	MSE
Heckman-Singer	-0.0438	-0.0628	1.3241	1.7552
FG	-0.0989	-0.2959	1.3691	1.8842
"Structural" k-means	-0.0078	-0.0718	1.8938	3.5865
K-means pre-clustering	-0.5508	-0.3648	1.5545	2.7199
CSS	-0.1989	-0.3271	1.4416	2.1178

1.3.5 Experiment 4: Gaussian distributed tastes

While the first 3 experiments compare these methods in relatively ideal settings (i.e., when their assumptions are satisfied), experiment 4 and experiment 5 are designed to mimic more realistic settings. In these settings, for any of these methods, getting good performance hinges on either a dense grid (HS and FG) or a proper number of groups (k-means methods and CSS). However, for all methods here except FG, it quickly becomes very computationally costly to do so. Therefore, in section 1.3.7, I compare the median estimation time of these methods in experiments 4 and 5.

In experiment 4, β follows a bivariate Normal distribution $N(\mu, \Sigma)$, with $\mu = (1, -1)'$, and $\Sigma = (0.5, 0.2; 0.2, 0.5)$. The FG has 225 types (15 in each dimension), and the CSS has 15 groups in each dimension. I experimented on increasing FG's grid to 1225 types (35 in each dimension), and the results are essentially identical.

Table 1.6 reports the experiment results. It should be noted that the good performance of Heckman-Singer here is partially a result of my choice on its estimation procedure: the price coefficient is estimated with a simple grid search over 11 values tightly around the true value. HS should perform worse and take much more time to estimate in a real application where a full optimization routine is used for estimating the price coefficient. Even in this simplified procedure, the median run time for this experiment for HS is still about 45 minutes, which is quite long and longest among all methods (see Table 1.8). Based on the fact that 11 evaluations in the outer loop (where the price coefficient is estimated) takes 45 minutes, on average, one evaluation takes about 4 minutes. It would not be surprising in a real application that convergence would require much more evaluations than that.

In Table 1.6, among the rest 4 methods, not one method is uniformly best across the different performance measures. FG seems to produce the least biased estimates if focusing on the welfare gains, while CSS has the smallest MSE; FG would not have such large MSE only if its standard deviations were smaller. On the other hand, if one also cares about the non-price coefficient, FG actually has the smallest MSE among the methods excluding HS. In fact, FG has both the smallest mean bias and standard deviation in this case. Interestingly, the structural" k-means seems to perform very well in terms of both bias and variance (especially the bias) for the price coefficient.-based k-means seems to perform very well in terms of both bias and variance (especially the bias).

Table 1.6: Experiment 4 results

	Price coefficient			
	Mean Bias	Median Bias	S. D.	MSE
Heckman-Singer	0.0023	0.0000	0.0185	0.0003
FG	-0.0873	-0.0874	0.0126	0.0078
"Structural" k-means	-0.0001	0.0012	0.0116	0.0001
K-means pre-clustering	0.0488	0.0477	0.0131	0.0026
CSS	-0.0216	-0.0209	0.0116	0.0006
	First Non-price coefficient			
	Mean Bias	Median Bias	S. D.	MSE
Heckman-Singer	-0.0411	-0.0031	0.279	0.0795
FG	-0.0205	-0.0219	0.3696	0.1370
"Structural" k-means	-0.0434	-0.0564	0.4200	0.1783
K-means pre-clustering	-0.0493	-0.0473	0.4143	0.1741
CSS	-0.0754	-0.0687	0.3764	0.1474
	Welfare gain			
	Mean Bias	Median Bias	S. D.	MSE
Heckman-Singer	-0.0294	-0.0076	0.1014	0.0111
FG	-0.0142	0.0028	0.2509	0.0632
"Structural" k-means	0.0129	0.0303	0.1348	0.0183
K-means pre-clustering	-0.0275	0.0189	0.1977	0.0398
CSS	0.0840	0.0675	0.0898	0.0151

1.3.6 Experiment 5: Irregularly distributed tastes

In this section, β follows a bivariate distribution that is a mixture of three bivariate Normal distributions plus a location shifter (i.e. mean) of (1,-1)'. To be specific, let β_i be the random vector for taste of consumer i. Then β_i has equal probability of drawing from the following three distributions:

1, $N(\mu_1, \Sigma_1), with\ \mu_1 = (-2,1)', and\ \Sigma_1 = (0.5,0.2; 0.2,0.5)$.

2, $N(\mu_2, \Sigma_2), with\ \mu_2 = (3,-1)', and\ \Sigma_2 = (2,0.8; 0.8,1)$.

3, $N(\mu_3, \Sigma_3), with\ \mu_3 = (-1,0)', and\ \Sigma_3 = (1,0.1; 0.1,2)$.

Table 1.7: Experiment 5 results

	Price coefficient			
	Mean Bias	Median Bias	S. D.	MSE
Heckman-Singer	0.0147	0.0200	0.0229	0.0007
FG (225 types)	-0.1024	-0.1101	0.0303	0.0114
FG (1225 types)	-0.0951	-0.1028	0.0278	0.0098
"Structural" k-means	0.0135	0.0120	0.0118	0.0003
K-means pre-clustering	0.0828	0.0828	0.0133	0.0070
CSS	-0.0321	-0.0336	0.0119	0.0012
	First Non-price coefficient			
	Mean Bias	Median Bias	S. D.	MSE
Heckman-Singer	-0.0645	-0.0633	0.6667	0.4486
FG (225 types)	0.0869	0.0974	0.6793	0.4690
FG (1225 types)	0.0936	0.1050	0.5375	0.2977
"Structural" k-means	0.0562	0.0777	0.6857	0.4733
K-means pre-clustering	-0.0768	-0.0117	0.8865	0.7918
CSS	0.0626	0.0923	0.6056	0.3707
	Welfare gain			
	Mean Bias	Median Bias	S. D.	MSE
Heckman-Singer	-0.2293	-0.0235	0.6926	0.5323
FG (225 types)	-0.0919	-0.0174	0.5561	0.3177
FG (1225 types)	-0.0906	-0.0281	0.5613	0.3233
"Structural" k-means	-0.0019	-0.0068	0.7944	0.6311
K-means pre-clustering	-0.1827	-0.0287	0.5796	0.3693
CSS	-0.0464	-0.0047	0.6600	0.4378

In this experiment, the FG has two sets of runs, one having 225 types (15 in each dimension), and the other having 1225 types (35 in each dimension). The CSS has 15 groups in each dimension. It is not clear changing from a regular normal distribution to a mixture of normal distributions would favor which estimation method. None of these methods relies on parametric assumptions.

Table 1.7 reports the results. Focusing on the welfare gains estimates. FG has the second-best performance in bias and the best standard deviation and MSE. Interestingly, the k-means pre-clustering method has the second smallest MSE, which is actually largely because of its relatively small standard deviation. In fact, looking at the median bias alone, the CSS has the best performance, and the k-means pre-clustering has the worst. On the other hand, even with grid search over a tight grid, HS takes a lot of time to estimate and does

not perform well. “Structural” k-means performs quite well in terms of bias, though it has a relatively large standard deviation.

Increasing the grid from 225 points to 1225 points does not change the FG’s performance in its welfare gain estimates. However, it does significantly improve its performance in the coefficient estimates. Both bias and variance are smaller for the price coefficient estimates, and the variance is smaller in its first non-price coefficient.

1.3.7 Run time

An important consideration in choosing an estimation method in real settings is the computational (time) cost. Table 1.8 reports the median run time of the estimation methods in the previous two experiments. Even in the experimental setting here, some of these estimation methods could be quite costly. The most time-consuming method is the Heckman-Singer, and the least time-consuming one is the FG with a moderate number of grid points. The k-means methods are fast in simple settings but quickly become much more time-consuming in slightly complicated situations: from normal to mixture distribution, the median run time for the “structural” k-means increases by more than 3 times, and the median run time for the k-means pre-clustering increases by about 2 times. Such an increase is about 60% for Heckman-Singer and CSS. However, FG does not seem to suffer from increased complexity in the heterogeneity here.

Table 1.8: Median run time (in mins)

	Normal	Mixture
Heckman-Singer	45.435	73.620
FG (225 types)	2.695	2.975
FG (1225 types)	23.320	23.185
“Structural” k-means	5.400	17.825
K-means pre-clustering	6.245	12.940
CSS	40.800	66.165

FG needs to have dense enough grid points to have a good performance. Looking at Table 1.8, it seems that by increasing the grid points by about 5.4 folds, the run time for FG increases by about 7.6 folds. However, 1225 types seem more than enough in the last two experiments. In fact, it seems that 225 points are already quite enough in getting a good welfare gains measure, and further gains by increasing the number of points seem marginal. I thus conclude that FG has a large computational advantage over other methods, and its performance is among the best in many cases above. Some of the other methods have better performance than FG in certain cases, and the trade-off is in computational time and additional gains in either bias or standard deviation.

Note that the comparisons here are all constrained to two dimensions of heterogeneity. With higher dimensions, some of these methods will quickly become computationally infeasible (e.g., Heckman-Singer) or data demanding ("structural" k-means). FG and CSS suffer less from such shortcomings, and they will show a further advantage in such settings.

1.4 Recommendations to empirical researchers

In practice, all the methods being studied in this chapter require some prior knowledge on the distribution of unobserved heterogeneity: FG requires the choice of the fixed grid, the clustering methods and the HS method require the number of clusters/types. However, for an empirical researcher, it might be difficult to start using the FG and the clustering methods from Cheng et al. (2019), Bonhomme et al. (2019), and Bonhomme and Manresa (2015), precisely because of the lack of such prior knowledge. A simple way to get a rough idea of the unobserved heterogeneity is to start with the k-means pre-clustering method. One should run the k-means multiple times with different tuning parameters (i.e., number of clusters) to determine a proper number of clusters. In this way, one can quickly get an idea of the support of the unobserved heterogeneity in each dimension and whether some areas of the support are more densely populated than the others.

Given the preliminary knowledge of the unobserved heterogeneity, the researcher can pick some reasonable tuning parameters and experiment with one of the other 4 methods. Again there is no one-size-fits-all solution. With multi-dimensional models, if the researcher is time-constrained, FG is a good choice due to

its computational advantage. The best practice for using FG is to "saturate" the parameter space by making the grid for type values cover enough range and making the grid fine enough along each dimension of heterogeneity. Sometimes it might happen to be the case that there exist "gaps" in the support of the heterogeneous parameters. For example, the k-means pre-clustering might indicate the taste coefficient lies between [0, 9], and there seems to be no consumer in the [3, 6] region. In this case, the researcher would want to reduce the number of points covering [3, 6] and increase the number of points in [0, 3] and [6, 9]. This could be especially helpful in saving computation time in multi-dimensional cases, as the wasted points in [3, 6] interact with points from all other dimensions. On the other hand, another situation where FG is clearly desirable is when the model only has heterogeneous parameters in a multi-dimensional setting (as is the case in Malone et al. 2019). In this case, the model only has to be solved *once*, instead of many times with other methods.

If there is only one dimension of heterogeneity, the "structural" k-means and the CSS are the same. If the researcher finds it tricky to fine-tuning the grid for FG, in this case, it might be easier to use the "structural" k-means/CSS. Cheng et al. (2019) provide guidance on how to pick the number of clusters with an information criterion properly.

Besides the issues discussed above, empiricists might want to know which methods are better suited for their research demands. For example, if the researcher cares about accurately estimating the common parameter, based on the Monte Carlo results here, it seems the "structural" k-means from Bonhomme and Manresa (2015) is a good choice. If the researcher really wants to know about all possible types of individuals out there and is having trouble fine-tuning the grid for FG in a multi-dimensional case with irregular distributions, it is worth going for CSS.

Finally, it might be worth trying to combine some of these methods in certain applications to achieve higher accuracy without adding much computational cost. For example, in my last Monte Carlo experiment, the "structural" k-means gives the best estimate for the common parameter. Using this estimate as a fixed input for the FG method will improve the FG's performance and make FG much faster, since now it is the case for FG that there is no longer a common parameter, and the model only needs to be solved once to get the distribution of unobserved heterogeneity.

1.5 Conclusions

This chapter compares 5 recently-developed methods in estimating multi-dimensional heterogeneity with a short panel of repeated choices by a set of consumers. I compare their performances in estimating the common parameter, the heterogeneous parameter, and welfare gains of a new product in 5 sets of Monte Carlo experiments. I find that no one method clearly dominates all others. FG achieves good performance in terms of both bias and variance in many performance measures with significantly lower computational costs than other methods. I also confirm CSS performs better than the other clustering methods when there are sparse type interactions. Both FG and CSS should show further advantages with larger dimensions of heterogeneity, as they are less constrained by computational cost and data requirements than other methods. On the other hand, I find that the "structural" k-means method accurately estimates the common parameters in my Monte Carlo experiments.

Based on my analysis, I provide a list of recommendations to empirical researchers who would like to apply these methods with real data. No one method would fit all the different empirical scenarios. It is recommended that the researchers start with k-means pre-clustering to get a rough idea of the underlying heterogeneity before choosing the tuning parameters. Researchers must pick the proper tuning parameters for these methods. Cheng et al. (2019) provide guidance on choosing the number of clusters with an information criterion. FG needs to have a wide and fine enough grid to provide adequate coverage of the support of the unobserved distribution. Finally, I provide some examples of how the researcher could use some of these methods if the researcher cares about different aspects or implications of the heterogeneity.

References

Ackerberg, D. A New Use of Importance Sampling to Reduce Computational Burden in Simulation Estimation. Quantitative Marketing and Economics, 7(4): 343-376, 2009.

Manuel Arellano and Stéphane Bonhomme. Robust priors in nonlinear panel data models. Econometrica, 77(2):489–536, 2009.

Stéphane Bonhomme and Elena Manresa. Grouped patterns of heterogeneity in panel data. Econometrica, 83(3):1147–1184, 2015.

Stéphane Bonhomme, Thibaut Lamadon, and Elena Manresa. A distributional framework for matched employer employee data. Econometrica, 87(3):699–739, 2019.

Xu Cheng, Frank Schorfheide, and Peng Shao. Clustering for multi-dimensional heterogeneity. Working paper, 2019.

Jeremy Fox, Kyoo Kim, and Chenyu Yang. A simple nonparametric approach to estimating the distribution of random coefficients in structural models. Journal of Econometrics, 195(2):236–254, 2016.

J. A. Hartigan and M. A. Wong. Algorithm AS 136: A K-Means Clustering Algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics) 28(1): 100-108, 1979.

J. Heckman and B. Singer. A method for minimizing the impact of distributional assumptions in econometric models for duration data. Econometrica, 52(2): 271-320, 1984.

Jacob Malone, Aviv Nevo, and Jonathan Williams. The tragedy of the last mile: Economic solutions to congestion in broadband networks. Working paper, 2019.

Aviv Nevo, John Turner, and Jonathan Williams. Usage-based pricing and demand for residential broadband. Econometrica, 84(2):411–443, 2016.

Kenneth E. Train. EM Algorithms for nonparametric estimation of mixing distributions. Journal of Choice Modelling, 1(1): 40-69, 2008.

CHAPTER 2

ESTIMATING FLEXIBLE DISTRIBUTIONS FOR THE RANDOM COEFFICIENT LOGIT MODEL WITH FIXED-GRID LIKELIHOOD AND CLUSTERING METHODS

BY RUIZHI MA

2.1 Introduction

Consumer demand estimation is a key component in many policy- and business-relevant analyses. Prominent examples include merger simulations, evaluating potential damages of certain anti-competitive conducts, and predicting market outcomes of new products. Several demand estimation methods in the market of differentiated products have been proposed in the past 20 years. In particular, the random coefficient logit model, one of the most widely utilized models of consumer demand, has had several estimation approaches proposed for it since Berry et al. (1995). In practice, however, the estimation of the model often relies on strong restrictions on the distribution of the unobserved component of consumer demographics, which could produce potentially seriously biased results.

This chapter provides two approaches to flexibly estimate the random coefficient logit model of demand when there is microdata in addition to data on market shares. The new approaches allow for a fully flexible distribution of the unobserved consumer demographics *for each consumer* (up to tuning parameters related to the distribution). The first approach adapts the "fixed-grid likelihood" (FG) method utilized in Malone et al. (2019), and the second approach adapts the clustering method from Cheng et al. (2019) (CSS). Both approaches assume the unobserved heterogeneity takes values from a finite number of types. For each individual, given a type, these methods produce a likelihood of observing the individual's outcomes under

the type. FG uses these likelihoods across types to update each individual's type distribution using Bayes rule, while CSS assigns the type that gives the highest likelihood to the individual.

Estimating mixed logit with both approaches requires market shares and some forms of data on the individual level. An ideal example of such data is a panel of repeated purchases by a representative sample of consumers, with some variables on their demographics. This chapter utilizes such a dataset from Nielsen panel in a U.S. city's cat wet food market in a recent year. However, the panel data utilized here contain many repeated purchases for each consumer, which is more than enough for the two approaches. Since FG incorporates a Bayesian approach to estimate the individual type distribution, in principle, a cross-section of individual purchase records, or a cross-section of certain kinds of survey responses (for example, the one in Berry et al. 2004), in addition to market shares, would suffice for FG to produce some information on unobserved demographics. More information at the individual level, for example, a more extended panel, will help to make type distribution more precise for each individual. As the panel data here are very long, the estimated type distributions by FG are mostly degenerate to one type for each individual. On the other hand, CSS does require panel data, but when the dimension of unobserved heterogeneity is low, a short panel will suffice. Asking data to determine the grid values for latent types (CSS) or integrating over types (FG) can be time-consuming. Thus one trade-off in the choice between the two methods is estimation time in different scenarios. To provide some reference, I report estimation time for these methods in my application.

This chapter contributes to the empirical Industrial Organization literature on demand estimation by allowing correlations between observed and unobserved demographics in the widely used random coefficient logit model of demand. The random coefficient logit model of demand can be estimated using market-level data, as is demonstrated in Berry et al. (1995), Nevo (2001), Petrin (2002), among others. Berry et al. (2004) demonstrate how to use data on second choices, in addition to market shares, to estimate the model with both micro- and macro-moments in a minimum-distance approach. In these studies, the unobserved demographics are assumed to be independent of the observed demographics.

This chapter also contributes to how to utilize both macro (market-level) data and micro (for example, purchase records, survey responses) data in a unified way to estimated demand. Perhaps most

related, Goolsbee and Petrin (2004) uses microdata on almost 30000 households in 317 markets to estimate a demand system of satellite, various cable, and antenna services. This paper is similar to Goolsbee and Petrin (2004) in how the estimation of product-market fixed effects is nested in a maximum likelihood estimation procedure. Goolsbee and Petrin (2004) allow different individuals' utilities to correlate differently across choices by (1) interacting features with observed demographics and (2) specifying a multivariate Normal distribution for the individual error term that is fully flexible in its product-level covariance matrix. However, this flexibility cannot capture correlations between observed demographics and the error term, as the error term is by assumption still independent of the observed variables. The approaches in this chapter allow for such correlations between observed demographics and unobserved demographics in the error term. This chapter also follows the insights from both Berry et al. (2004) and Goolsbee and Petrin (2004) in that: (1) the macro data are useful to uncover product-market fixed effects (so-called "mean utilities"), and (2) the microdata are useful to identify how tastes change with observed demographics. Another advantage of the chapter's methods is that they are inherently nonparametric, thus more flexible than conventional approaches that impose parametric distributional assumptions on unobserved heterogeneity.

One major focus in estimating the mixed logit demand model is achieving causal inference via quasi-experimental methods like instrumental variables. This is, however, an orthogonal topic to the estimation of latent heterogeneity, which in some sense is to provide adequate control variables to estimate the parameter of interest. Therefore in this chapter, I follow the conventional ways (as is presented in Berry et al. 2004) to model the price endogeneity. There are some recent developments on constructing better instruments using market-level data, for example, Gandhi and Houde (2019) and Petrin and Seo (2019). On the other hand, microdata's availability opens the question of whether individual level price endogeneity is empirically important. However, in that case, allowing for both flexible heterogeneity and more flexible endogeneity is a challenging task, especially in structured models with certain functional assumptions on agent behaviors, which could make the endogeneity problem non-separable.

The rest of the chapter is structured as follows. The second section is the model, estimation, and inference procedures of the mixed logit demand model with the proposed approaches. The third section is a simple example comparing FG and conventional parametric approaches without price endogeneity. The fourth section is an application to the cat wet food market data in a U.S. city in a recent year. The fifth section concludes.

2.2 Model and Estimation

2.2.1 Overview

Assume the utility of consumer i from consuming product j in market t with product features $\boldsymbol{x_{jt}}$ (including price) is

$$u_{ijt} = \boldsymbol{x_{jt}'\beta_i} + \xi_{jt} + \epsilon_{ijt} \tag{1}$$

where ξ_{jt} is the so-called ``structural error term" that captures the effect of unobserved quality of the product, ϵ_{ijt} is the i.i.d. logit error term, and $\boldsymbol{\beta_i}$ are functions of observable demographics $\boldsymbol{D_i}$ and unobserved demographics of consumer i:

$$\beta_{i,k} = \boldsymbol{D_i'\theta_k} + \gamma_k + v_{i,k}, \quad k = 1,2,\dots,K \tag{2}$$

Here, γ_k is the mean of the effect of the unobserved demographics on coefficient k, and $v_{i,k}$ measures how the effect of individual i's unobserved demographics deviates from the mean effect on the coefficient $\beta_{i,k}$. In a parametric approach, the $v_{i,k}$ can be assumed to follow the same mean-zero distribution, for example, $N(0, \sigma_k^2)$, for all i. However, FG can estimate a different posterior distribution of $\boldsymbol{v_i}$ for each consumer. These posterior distributions are not necessarily mean-zero. CSS treats $\boldsymbol{v_i}$ as a fixed value for each individual i, and backs out that value based on data on i.

Substituting Equation (2) into Equation (1), the utility can be written as

$$u_{ijt} = \sum_{k=1}^{K} \gamma_k\, x_{jt,k} + \sum_{k=1}^{K}\sum_{l=1}^{L} \theta_{k,l} x_{jt,k} D_{i,l} + \sum_{k=1}^{K} x_{jt,k} v_{i,k} + \xi_{jt} + \epsilon_{ijt}$$

where L is number of observed demographic variables. The coefficients in this equation are thus $\boldsymbol{\gamma}$, $\boldsymbol{\theta}$ and $\boldsymbol{v}$. One of the most critical aspects of the random coefficient logit model is its ability to estimate $\boldsymbol{\gamma}$ consistently,

in the presence of price endogeneity arising from the unobserved product quality term ξ_{jt}. This is done by utilizing the moment conditions $E(\delta_{jt} - \boldsymbol{x}'_{\boldsymbol{jt}}\boldsymbol{\gamma})Z_{jt} = 0$, where Z_{jt} is a set of instruments, and $\delta_{jt} = \boldsymbol{x}'_{\boldsymbol{jt}}\boldsymbol{\gamma} + \xi_{jt}$. δ_{jt} is the so-called mean utility, or the product-market fixed effects.

To utilize these moment conditions, one has first to estimate δ_{jt}. Substituting $\delta_{jt} = \boldsymbol{x}'_{\boldsymbol{jt}}\boldsymbol{\gamma} + \xi_{jt}$ into the utility expression, I get

$$u_{ijt} = \sum_{k=1}^{K}\sum_{l=1}^{L}\theta_{k,l}x_{jt,k}D_{i,l} + \sum_{k=1}^{K}x_{jt,k}v_{i,k} + \delta_{jt} + \epsilon_{ijt} \quad (3)$$

where now the coefficients are $\boldsymbol{v}, \boldsymbol{\theta}$ *and* $\boldsymbol{\delta}$, and there is no further endogeneity issue (assuming exogeneity of both product features and demographics).

Therefore, all the methods utilized in this paper follow a two-step estimation procedure, where in the first step, I estimate $\boldsymbol{v}$, $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$, and in the second step, I regress estimated $\boldsymbol{\delta}$ on product features using two-stage-least-squares (2SLS).

The Outside Option I assume each consumer may choose to not purchase any product in the market at the time. The utility of choosing this outside option is normalized to 0 plus an i.i.d. logit error term for each consumer:

$$u_{i0t} = 0 + \epsilon_{i0t}$$

Market Definition and Choice Set I treat each week as a market. I assume the choice set of a consumer in a week is just the set of all products ever purchased by any consumer in that week, plus the outside option. This assumption on choice set is necessary for the estimation of nontrivial values for product-market fixed effects. Otherwise, those products that do not have a single purchase record in a market will have a product-market fixed effect of negative infinity.

Household Choice Problem As is standard in applications with discrete choice models, I treat a single record of the multiple-unit purchase in the raw data as repeated choices, where within each choice scenario, the

household chooses 1 unit of product from the choice set in that week. Since the outside option is defined as not buying a wet food product, I assume the household makes the same number of choices each week, which is equal to the maximum number of cans the household ever purchased in a week.[3]

2.2.2 Two-step Estimation Procedure

2.2.2.1 First Step Estimation Algorithms

In the first step estimation, if we have a long panel of individual consumer's purchase records, a straight-forward way to estimate (θ, v, δ) is by MLE:

$$(\theta, \mathrm{v}, \delta) = \operatorname{argmax} \sum_{\mathrm{i,j,t}} y_{ijt} \log\left(\frac{e^{U_{ijt}}}{\sum_{k \in C_t} e^{U_{ikt}}}\right)$$

where,

$$U_{ijt} = \sum_{k=1}^{K} \sum_{l=1}^{L} \theta_{k,l} x_{jt,k} D_{i,l} + \sum_{k=1}^{K} x_{jt,k} v_{i,k} + \delta_{jt}$$

However, I do not take this approach to directly estimate the parameters, for 2 reasons. First, there are too many parameters in δ and v for an optimizer to handle appropriately. For example, in my application, there are about 10000 parameters in δ, because there are 52 markets (weeks), and in each market, there are around 200 products, resulting in about 10000 product-market fixed effects. Second, unless a long panel data of individual purchases is available (which is rare), there are not enough observations to separately estimate $\{v_{i,k}\}_{k=1}^{K}$ for each individual i.

To address the issue that there are many parameters in δ, I incorporate the contraction mapping algorithm from Berry et al. (1995). To address the second issue and that there are also many parameters in v, I adapt FG and CSS, which classify individuals into discretized types. This makes it feasible to identify the

[3] While this assumption would have an impact on the estimation result, such impact should be more or less the same to all these estimation methods here (as they are estimating the same model). The focus is to compare these methods using the same data, rather than discussing the validity of such assumptions.

distribution of $\boldsymbol{v}$ even without a long panel. Both methods will work with only a short panel on individual purchases. Technically, FG only requires a cross-section of individual purchase records.

To further illustrate the first step estimation procedure, I start by describing a simple parametric MLE approach that nests the contraction mapping algorithm. This "conventional" approach will be used as a benchmark for comparison with FG- and CSS- incorporated approaches. Both FG- and CSS- modified MLE can be viewed as direct relaxations of this "conventional" approach. As an example of such a parametric MLE approach, let me start by assuming the unobserved demographics $\boldsymbol{v}$ follows a Gaussian distribution $N(0, \Sigma)$. The $\boldsymbol{v}$ will be integrated out. In practice, the integration can be carried out numerically by discretizing the space of $\boldsymbol{v}$. This can be done by first fixing M draws from the Standard Gaussian distribution, $\{\tilde{v}_m\}_{m=1}^{M}$, and then expressing $\boldsymbol{v}$ as a linear transformation of the Standard Gaussian random vector $\tilde{v}$. Let $v = \Gamma\tilde{v}$, then $\Sigma = \Gamma\Gamma'$, and the optimization problem is the following:

$$(\boldsymbol{\theta}, \Gamma) = argmax \sum_{i,j,t} y_{ijt} \log\left(\frac{1}{M}\sum_{m=1}^{M}\frac{e^{U_{ijt,m}}}{\sum_{k\in C_t} e^{U_{ikt,m}}}\right) \tag{4}$$

where,

$$U_{ijt,m} = \sum_{k=1}^{K}\sum_{l=1}^{L} \theta_{k,l} x_{jt,k} D_{i,l} + x'_{jt}\Gamma\tilde{v} + \widehat{\delta_{jt}}(\theta, \Gamma)$$

$$\hat{\delta}(\theta, \Gamma) = \widehat{S^{-1}}(s|\theta, \Gamma)$$

The last equation signifies that, given a value of θ and Γ, the δ can be obtained by inverting the 1-to-1 mapping $\hat{S}$ from δ to observed market shares, s. Berry et al. (1995) took advantage of the fact that the market shares are monotonic in δ to construct a contraction mapping that can quickly convert observed market shares to δ. Here the contraction mapping for solving δ is thus nested within an algorithm to solve for θ and Γ. Formally, Algorithm 1 below described how the estimation is carried out, with an optimizer:

Algorithm 1 (Parametric MLE)

Start with stopping criterion $\epsilon_c$and initial guess $\left(\theta^{(0)}, \Gamma^{(0)}\right)$.

1, At iteration n, given the current values $\left(\theta^{(n)}, \Gamma^{(n)}\right)$, compute $\delta^{(n)}$ with contraction mapping.

2, Compute the objective value with $\theta^{(n)}, \Gamma^{(n)}, \delta^{(n)}$, and evaluate convergence. If convergence is not achieved, obtain $\theta^{(n+1)}, \Gamma^{(n+1)}$ using the method provided by the optimizer, and repeat 1-2.

Implicit in the above approach is the assumption that each individual has the same de-meaned distribution of unobserved demographics $F(v)$, regardless of the individual's observed demographics. The FG-incorporated MLE approach relaxes this assumption in the above ``conventional" approach by letting each individual have a different discretized distribution over a grid of $v, v_1, \ldots, v_M$. $P_v(v_m)$, the probabilities of an individual belonging to type v_m,is implied by the individual's own likelihoods with each type.

To be specific, let $L(Y_i, D_i, X|v_m; \theta)$ be the likelihood of observing individual i's outcome Y_i (choices), given the observed demographics D_i, features of all available products X, a value of θ, and the individual's type v_m for the unobserved demographics. Assuming a uniform prior $\{\pi_m\}_{m=1}^M$ common to all individuals, the probabilities of an individual belonging to type v_m is given by the Bayes rule:

$$P_v(v_m|D_i, Y_i, X; \theta) = \frac{L(Y_i, D_i, X|v_m; \theta)\pi_m}{P(X_i, Y_i|\theta)}$$

$$= \frac{L(\boldsymbol{Y_i}, \boldsymbol{D_i}, \boldsymbol{X}|\boldsymbol{v_m}; \boldsymbol{\theta})\pi_m}{\sum_{m=1}^M L(\boldsymbol{Y_i}, \boldsymbol{D_i}, \boldsymbol{X}|\boldsymbol{v_m}; \boldsymbol{\theta})\pi_m}$$

Therefore, the FG-incorporated MLE solves the following problem:

$$\theta = argmax \sum_{i,j,t} y_{ijt} \log\left(\sum_{m=1}^{M} \frac{e^{U_{ijt,m}}}{\sum_{k\in C_t} e^{U_{ikt,m}}} P_v(v_m|X_i, Y_i, X; \theta)\right) \quad (5)$$

where,

$$U_{ijt,m} = \sum_{k=1}^{K}\sum_{l=1}^{L} \theta_{k,l} x_{jt,k} D_{i,l} + \sum_{k=1}^{K} x_{jt,k} v_{m,k} + \widehat{\delta_{jt}}(\theta)$$

$$\hat{\delta}(\theta) = \widehat{S^{-1}}(s|\theta)$$

$$P_v(v_m|D_i, Y_i, X; \theta) = \frac{L(Y_i, D_i, X|v_m; \theta)\pi_m}{\sum_{m=1}^M L(Y_i, D_i, X|v_m; \theta)\pi_m}$$

The second to last equation above again is solved by a nested contraction mapping. Note that computing market shares predicted by the model requires the posterior probabilities for each individual, which in turn changes with δ. Thus for FG, the Bayes updating (last equation above) is also embedded in the contraction mapping: for each θ, the data implies a posterior type distribution for each individual. Given these posteriors, the predicted market shares can be computed, and $\hat{S}$converted to obtain $\hat{\delta}(\theta)$.

Formally, Algorithm 1 below described how the estimation is carried out, with an optimizer:

Algorithm 2 (FG)

Start with stopping criterion ϵ_c and initial guess $\theta^{(0)}$.

1, At iteration n, given the current values $\theta^{(n)}$,compute $\delta^{(n)}$ with contraction mapping. As the posterior type distributions change with δ, the Bayes updating is nested in the contraction mapping algorithm.

2, Compute the posterior type distributions implied by $\theta^{(n)}$, $\delta^{(n)}$.

3, Compute the objective value with $\theta^{(n)}$, $\delta^{(n)}$ and the implied posterior type distributions.

4, Evaluate convergence. If convergence is not achieved, obtain $\theta^{(n+1)}$ using the method provided by the optimizer, and repeat 1-3.

Similar to the parametric and FG approach, CSS-incorporated MLE also discretizes the value space for the unobserved demographics. Unlike FG, which allows a distribution over fixed types for each individual, CSS assigns a deterministic type to each individual and also estimates the type value for each group of individuals who are of the same type. CSS thus aims to estimate both grid values of the types $\{v_k\}_{k=1}^{K}$, and the group memberships of each individual $\{B_i\}_{i=1}^{I}$, in addition to the homogeneous parameters θ.[4]

Assume there are K_u dimensions of unobserved heterogeneity.[5] For simplicity, assume the number of types in each dimension is the same, M. Then for each k, $v_k = (v_{1k}, \dots, v_{Mk})$. The key advantage of CSS over other

[4] These parameters are called "homogeneous" simply because they are common to all individuals.
[5] If the unobserved demographics interact with all product features, then $K_u = K$, where K is the number of product features.

clustering methods is that, instead of assigning a single group membership to each individual, CSS assigns a membership for the individual *in each dimension*. Thus each individuals will have K_u memberships, denoted as $B_i = (B_{i1}, \dots, B_{iK_u})$.

With these notations in hand, the CSS-incorporated MLE solves the following problem:

$$(\theta, \{v_k\}_{k=1}^K, \{B_i\}_{i=1}^I) = argmax \sum_{i,j,t} y_{ijt} \log\left(\frac{e^{U_{ijt}}}{\sum_{k \in C_t} e^{U_{ikt}}}\right)$$

where,

$$U_{ijt} = \sum_{k=1}^{K}\sum_{l=1}^{L} \theta_{k,l} x_{jt,k} D_{i,l} + \sum_{k=1}^{K} x_{jt,k} v_{B_{ik},k} + \widehat{\delta_{jt}}(\theta, \{v_k\}_{k=1}^K, \{B_i\}_{i=1}^I)$$

$$\hat{\delta}(\theta, \{v_k\}_{k=1}^K, \{B_i\}_{i=1}^I) = \widehat{S^{-1}}(s|\theta, \{v_k\}_{k=1}^K, \{B_i\}_{i=1}^I)$$

Unlike the previous two problems, for CSS I do not carry out the estimation by asking an optimizer to directly search for the parameters, $(\theta, \{v_k\}_{k=1}^K, \{B_i\}_{i=1}^I)$. This is simply because the group memberships $\{B_i\}_{i=1}^I$ take discrete values. Instead, the CSS is carried out by a modified version of an expectation-maximization algorithm, where in each iteration, the classification is carried out first to update the distribution of types (expectation step, updating $\{\boldsymbol{B_i}\}_{i=1}^I$), the homogeneous parameters (θ *and* δ) are then updated with MLE with nested contraction mapping (maximization step), and finally the grid for the discrete types ($\{v_k\}_{k=1}^K$) are updated. Denote

$$\Lambda(\theta, \{v_k\}_{k=1}^K, \{B_i\}_{i=1}^I) = \sum_{i,j,t} y_{ijt} \log\left(\frac{e^{U_{ijt}}}{\sum_{k \in C_t} e^{U_{ikt}}}\right)$$

The algorithm to carry out CSS-incorporated MLE is the following:

Algorithm 3 (CSS)

Iteration 0: Set convergence tolerance ϵ_c. Start with initial guess $\theta^{(0)}$, $\{v_k^{(0)}\}_{k=1}^K$, $\{B_i^{(0)}\}_{i=1}^I$, and the implied $\delta^{(0)}$. Compute Λ with the initial guess, denoted $\Lambda^{(0)}$.

At iteration n:

1, Given $\theta^{(n-1)}$, $\{v_k^{(n-1)}\}_{k=1}^K$, $\{B_i^{(n-1)}\}_{i=1}^I$, $\delta^{(n-1)}$, find $\{B_i^{(n)}\}_{i=1}^I$: for each i, choose B_{ik} that maximizes the individual likelihood of i for $k = 1, \dots, K_u$:

$$B_{ik}^{(n)} = argmax_{B_{ik}\in\{1,\dots,M\}} L\left(X_i, Y_i \middle| B_{ik}, \{B_{ir}^{(n-1)}\}_{r\neq k}, \theta^{(n-1)}, \{v_k^{(n-1)}\}_{k=1}^K\right)$$

2, Given $\{v_k^{(n-1)}\}_{k=1}^K$, $\{B_i^{(n)}\}_{i=1}^I$, find $\theta^{(n)}$ and $\delta^{(n)}$ by first solving

$$\theta^{(n)} = argmax \sum_{i,j,t} y_{ijt} \log\left(\frac{e^{U_{ijt}}}{\sum_{k\in C_t} e^{U_{ikt}}}\right)$$

where,

$$U_{ijt} = \sum_{k=1}^{K}\sum_{l=1}^{L} \theta_{k,l} x_{jt,k} D_{i,l} + \sum_{k=1}^{K} x_{jt,k} v_{B_{ik}^{(n)},k}^{(n-1)} + \widehat{\delta_{jt}}\left(\theta, \{v_k^{(n-1)}\}_{k=1}^K, \{B_i^{(n)}\}_{i=1}^I\right)$$

$$\hat{\delta}\left(\theta, \{v_k^{(n-1)}\}_{k=1}^K, \{B_i^{(n)}\}_{i=1}^I\right) = \widehat{S^{-1}}\left(s \middle| \theta, \{v_k^{(n-1)}\}_{k=1}^K, \{B_i^{(n)}\}_{i=1}^I\right)$$

The $\delta^{(n)}$ is then the δ implied by $\theta^{(n)}, \{v_k^{(n-1)}\}_{k=1}^K, \{B_i^{(n)}\}_{i=1}^I$.

3, Given $\boldsymbol{\theta^{(n)}}, \boldsymbol{\delta^{(n)}}, \{\boldsymbol{B_i^{(n)}}\}_{i=1}^I$, find $\{\boldsymbol{v_k^{(n)}}\}_{k=1}^K$ by solving

$$v^{(n)} = argmax \sum_{i,j,t} y_{ijt} \log\left(\frac{e^{U_{ijt}}}{\sum_{k\in C_t} e^{U_{ikt}}}\right)$$

where,

$$U_{ijt} = \sum_{k=1}^{K}\sum_{l=1}^{L} \theta_{k,l}^{(n)} x_{jt,k} D_{i,l} + \sum_{k=1}^{K} x_{jt,k} v_{B_{ik}^{(n)},k} + \widehat{\delta_{jt}^{(n)}}$$

4, Compute $\Lambda^{(n)}$. If $\left|\Lambda^{(n)} - \Lambda^{(n-1)}\right| < \epsilon_c$, stop and report convergence; otherwise, repeat steps 1-3.

It should be emphasized that in the above algorithm, $\{\boldsymbol{v_k^{(n)}}\}_{k=1}^K$ should be jointly estimated. This is crucial in obtaining the efficiency gains of CSS in cases of multidimensional heterogeneity.

2.2.2.2 Second step estimation

In the second step, I regress estimated δ on own-product features in a 2SLS. Therefore I am estimating the following equation with 2SLS:

$$\delta_{jt} = x'_{jt}\gamma + \xi_{jt}$$

In addition to own-product features, I use two sets of instruments for price: first, the average price of the same product in other geographic markets at the same time (the so-called "Hausman-type IV"), and, second, the (distribution of) features of other products available in the same geographic market at the same time.[6] I use both sets of IVs in my application, and the use of these IVs drives the estimated price coefficient more negative, which is the usually expected direction if unobserved demand shocks (that are observable and are responded upon by sellers) are indeed the source of price endogeneity.

The modeling of endogeneity In this model, a product's prices are assumed to be the same for all consumers in the same market. This is the same assumption made in Berry et al. (2004) and makes the price endogeneity problems here practically the same as one in models for market-level data. To be specific, here the price endogeneity problem arises as price (one of the $\boldsymbol{x'_{jt}}$) and the residual ξ_{jt} in the equation $\delta_{jt} = x'_{jt}\gamma + \xi_{jt}$ are potentially correlated, due to both unobserved product features and/or unobserved product-market specific demand shocks (for example, temporary promotions). Whereas in market-level data, it is natural to use a single price for a product, in microdata with individual purchase records, there could be considerable variations across individuals in the same product's prices in the same market (defined as a single area in a specific time). For example, in my application, such variations are comparable to both cross-product price variation and the same product's cross-time price variation. Such variation thus is a potential additional useful source of identification for price coefficients. On the other hand, to instrument for price coefficient in that setting is also more challenging, as there could be unobserved individual-specific, individual-product-

[6] In my application, I focus on only one geographic market in different weeks, thus the description of the IVs here cater specifically to my setting.

specific, individual-market specific, and even individual-product-market-specific factors that are potentially correlated to price faced by an individual. The model and estimation procedure also need to be modified to accommodate such a setting, and it might be the case that the endogeneity becomes non-separable, and certain control function approaches are more feasible than IV approach. Though exciting and potentially important in empirical applications with microdata, this issue is not the focus of this paper. Thus, I follow the previous micro BLP papers in assuming a single price of a product in a market.

2.2.3 Statistical Inference

For the parametric MLE approach, asymptotic standard errors of estimated $\boldsymbol{\theta}$ and $\boldsymbol{\Gamma}$ can be obtained by inverting the Fisher Information matrix. Since $\boldsymbol{\delta}$ is a function of $\boldsymbol{\theta}$ and $\boldsymbol{\Gamma}$ that does not have analytical expression, the derivative of $\boldsymbol{\delta}$ with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\Gamma}$ can be obtained with the help of Implicit Function Theorem. For the FG approach, asymptotic standard error of $\boldsymbol{\theta}$ can be obtained in the same way as that in the parametric MLE approach, except that the derivative of $\boldsymbol{\delta}$ with respect to $\boldsymbol{\theta}$ is more complicated, as the Bayes updating is nested in the contraction mapping. For the CSS approach, I fix the estimated group memberships as given, and compute standard errors by treating CSS as if it was solving for $(\boldsymbol{\theta}, \boldsymbol{\nu})$ using MLE. For a detailed explanation of the standard error formula, please see the Appendix.

2.3 A simple example

In this section, I illustrate how conventional parametric approach could yield biased results via a simple example. I start by estimating the following simplified version of the mixed logit model using my data:

$$u_{ijt} = \gamma_1 Price_{jt} + \gamma_2 Size_{jt} + \gamma_3 Chicken_{jt} + \gamma_4 Tuna_{jt} + \theta Tuna_{jt} \times Family\ size_i + v_i Tuna_{jt} + \epsilon_{ijt}$$

where a cat food product j in week t is characterized by its price, size of can, whether it is made of chicken, and whether it is made of tuna. The family size of household i is simply the number of people living in the household, which is the observed demographics variable. The structural error term is assumed away for ease of exposition. The price endogeneity from the structural error term will be treated in the full model.

I estimate this simplified model with the help of the fixed-grid likelihood (FG) method. The data is on purchasing records of canned cat wet food by 50 randomly selected households. Figure 2.1 plots the estimated unobserved demographics v against family size, where each dot is a household. There is a strong correlation between family size and estimated unobserved demographics. Moreover, the correlation is positive in this case. This correlation cannot be captured by conventional parametric approach, since it assumes independence between observed and unobserved demographics.

As the conventional mixed logit estimation assumes independence of v from family size, I expect it to overestimate θ following the logic of omitted variable bias. This is indeed the case, as shown in Table 2.1. Table 2.1 presents estimation results of two models. The results from the mixed logit model with FG are in column (2). In column (1), I present the results from a conventional maximum likelihood estimation of the model, where I assume v follows a Gaussian distribution independent of other variables. The "Standard Deviation" in Table 1 refers to the standard deviation of the distribution of v. The estimated θ from FG in column (2) is -2.09, which is much smaller and of opposite sign compared to 0.28, the estimated θ from column (1).2 The FG also achieves lower objective value (negative log-likelihood) at optimum, which is also expected, since FG is otherwise the same as the conventional MLE, except that it allows the weights on the discretized types of v to be more flexibly determined by data.

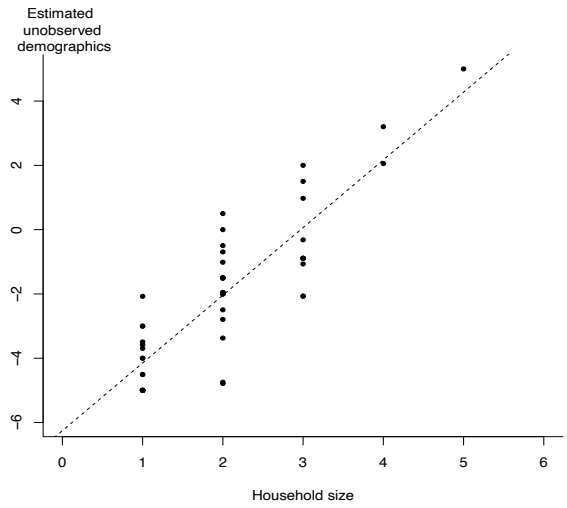

Figure 2.1: Correlation between observed and unobserved demographics

Table 2.1: Estimates of conventional MLE and FG in simplified model

	(1)	(2)
Imposing Independence	Yes	No
Tuna * Household Size	0.2799	-2.0920
	(0.0216)	(0.0097)
Price	-9.5910	-9.3354
	(0.0220)	(0.0011)
Size	-0.5350	-0.5612
	(0.0026)	(0.0098)
Chicken	-0.5617	-0.5616
	(0.0129)	(0.0482)
Tuna	-4.5621	3.6235
	(0.4829)	(0.0211)
Standard Deviation	2.1670	2.3224
	(0.3436)	(0.0035)
Value of objective	122.7843	122.0429

Note: Standard errors in parentheses.

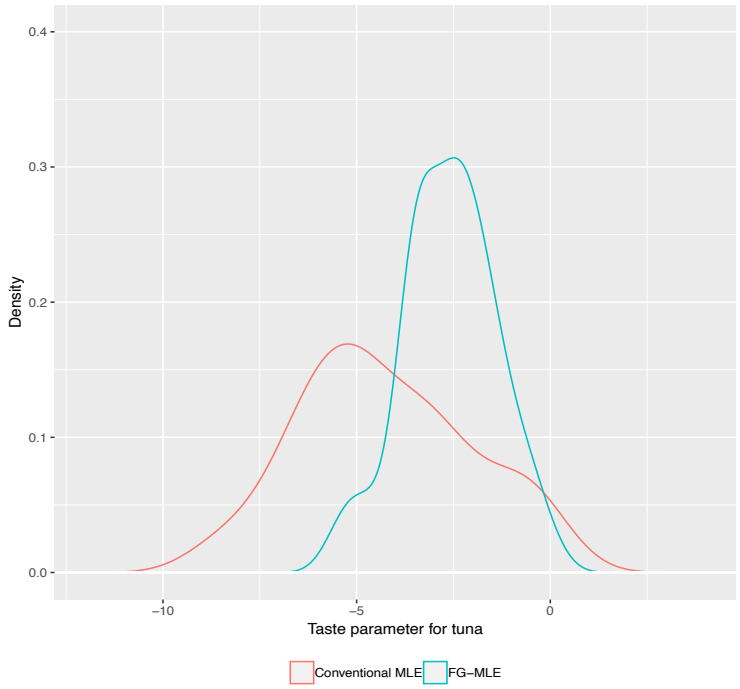

Figure 2.2: Estimated distribution of random coefficient on tuna in simplified model

The correlation between unobserved and observed demographics further propagates into the estimated taste distribution for tuna, and elasticity estimates related to tuna products. Figure 2.2 compares the estimated taste distributions for the random coefficient on tuna by conventional MLE and FG. The random coefficient here, $\beta_{i,4}$, is simply defined as $\beta_{i,4} = \gamma_4 + \theta \times \text{Family size} + v_i$, which is meant to capture the taste towards tuna for household i. FG estimates a much narrower variation in such taste across households. This is not surprising: FG uncovers a strong *positive* correlation between unobserved and observed demographics, and the estimated coefficient θ for the observed demographic variable is *negative* (-2). Therefore, in determining the random coefficient of tuna, v and D cancel out each other's variations to some extent.

Finally, a narrower taste variation makes FG predict a substitution pattern where the difference between the cross-price elasticities from a tuna product towards tuna- and non-tuna- products is not as drastic as what the conventional method will predict. This is the case in Table 2.2. Table 2.2 shows elasticities of 4 products from a week of the year, with respect to the price increase of product 1, which is a tuna product. These four products have the same size and very similar prices. Both methods predict the price increase of product 1

will benefit more of the other tuna product (product 2) rather than the chicken (product 3) or non-tuna-non-chicken product (product 4). But, as the conventional method suggests that people have more diverse tuna taste compared to FG, the extent to which existing tuna buyers love tuna more than non-buyers is larger under the conventional method than that under FG. Therefore, the conventional method overestimates the cross-price elasticities for product 2 (by about 75%). Moreover, it underestimates the cross-price elasticities for products 3 and 4. The cross-price elasticities for products 3 and 4 are virtually identical within a column, because the coefficient in front of the chicken dummy variable is assumed to be not random in this simplified model.

Table 2.2: Elasticities with respect to price change of product 1

Product	Chicken	Tuna	EM	FG
1	0	1	-4.21840	-4.10661
2	0	1	0.00166	0.00095
3	1	0	0.00031	0.00037
4	0	0	0.00031	0.00037

To summarize, through the example above, I show that the conventional parametric approach produces bias in the coefficients related to observed demographic variables. The bias further distorts estimated taste distribution and elasticity estimates.

2.4 Application

In this section, I apply the three methods to a dataset from the Nielsen Homescan database. In this application, there are 50 randomly selected households from the full sample, with two observed demographic variables ($L = 2$, household income, and household size), choosing from more than 200 differentiated products in each of the 52 weeks. For each product, four features are observed ($K = 4$): price, size, chicken flavor dummy, and tuna flavor dummy). For ease of exposition and consistency with the simple example in the previous section, I interact all features with observed demographics, but only allow tuna to have unobserved individual fixed effects. Therefore the model in this section is one with only one dimension of unobserved heterogeneity:

$$u_{ijt} = \sum_{k=1}^{4} \gamma_k x_{jt,k} + \sum_{k=1}^{4}\sum_{l=1}^{2} \theta_{k,l} x_{jt,k} D_{i,l} + x_{jt,4} v_{i,4} + \xi_{jt} + \epsilon_{ijt}$$

where $u_{ijt}$is consumer i's utility from product j in market t, $x_{jt,k}$ is the kth feature of product j in market t, $D_{i,l}$ is the lth *observed* demographic variable of consumer i, $v_{i,k}$ is the kth *unobserved* demographic variable of consumer i (a feature-individual fixed effect), ξ_{jt} is the so-called "structural error" term that is assumed to be potentially correlated with price, and ϵ_{ijt} is the logit error term (that is often assumed to follow i.i.d. type-1 extreme value distribution). This leads to 8 parameters in $\boldsymbol{\theta}$, 50 "unknowns" in $\boldsymbol{v}$ (the same number as number of households), and about 10000 parameters in $\boldsymbol{\delta}$. The algorithms here however can be used directly to estimate a model with multi-dimensional unobserved heterogeneity.

2.4.1 Data

The data is the purchase records on wet cat food from Nielsen Panel households living in a U.S. city in a recent year.[7] The data consists of three parts: household demographic information, household purchase records (date and choice), and information on all available products. The last part contains more products than what is purchased by households in the panel. As described in the model section, I focus on only those products ever purchased by households in my sample. Also, I keep only those records that are from households that have at least 10 purchase records on wet cat food. The above two restrictions cater to the FG- and CSS- MLE approach: products with no purchase records will lead to a product-market fixed effect of negative infinity from the MLE (so I drop them), and the preference of a household with too few purchase records cannot be identified. These two data selection constraints are not restrictive at all, as households

[7] I thank the Kilts Center for Marketing at the University of Chicago Booth School of Business for providing access to the data. The following disclaimers apply: "Researcher(s) own analyses calculated (or derived) based in part on data from Nielsen Consumer LLC and marketing databases provided through the NielsenIQ Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the NielsenIQ data are those of the researcher(s) and do not reflect the views of Nielsen. Nielsen is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein."

dropped only contribute to a tiny amount of purchases, and products not picked up by households in the sample most likely only have a minimal market presence.

On average, each consumer has about 70 purchase records in a year, and the median number of purchases is about 40. Though small, this is enough to identify the grouped random-effects or grouped fixed effects of consumer preferences as each product only has 4 features in my application. For product features, I have the price (dollars per ounce), size of can, and two dummies for chicken and tuna flavors. For demographics, I have two variables: income and size of household. On the other hand, there are a few hundred different products (200-300) purchased in the data in each week. In sum, there are about 10000 unique product-market interactions.

When it comes to constructing the two sets of IVs, for the Hausman-type IV, I compute the average price of a product in a week in areas other than the city in my sample as an instrument of its price in the city in my sample. For the features of other products in the market, since there are always many products in a given week, in practice, I use quantiles of the feature distribution in that week as instruments.

2.4.2 Assumptions on unobserved demographics

For the parametric approach, I assume the unobserved demographics v follow a standard Gaussian distribution $N(0, \sigma^2)$. This distribution is independent of the observed demographics. For FG, I assume v takes one of the 21 values evenly distributed between [-5,5]. For CSS, I assume there could be up to 21 different groups in terms of the unobserved demographics.

2.4.3 Estimation results

2.4.3.1 Convergence time and convergence evaluations

I begin by comparing the estimation time of FG, CSS, and the parametric approach: 2.43 hours, 2.47 hours, and 12 minutes. Compared to FG, CSS has this additional step to update grid point values within each

iteration. Compared to CSS, FG has to integrate over types for each individual in the contraction mapping algorithm. Therefore it is not necessarily the case that one will be faster than the other.[8]

These time records' absolute levels are also subject to the hardware, software, coding practices, and convergence criteria. All the estimations here are done on a cloud server with a Unix system on top of a high-performance computing cluster. A typical node I work on has 24 Intel Xeon CPUs (with specifications E5-2643 v2 @ 3.50GHz). I use the open-source software R to implement the estimation. I use the optimizers in R package *nloptr*. For the BLP contraction mapping, I use R's implementation of *SQUAREM* to speed it up. *SQUAREM* initially proposed by Varadhan and Roland (2008) to speed up any EM algorithm, and is then used and recommended by Conlon and Gortmaker (2020) to speed up the BLP contraction mapping.

When it comes to tolerance levels, there is also a choice for the tolerance level for the convergence of the BLP fixed-point algorithm inside the optimization step of the common parameters. As is discussed extensively in Dubé et al. (2012), a low tolerance level (e.g., 10^{-4}) for the contraction mapping may lead to results far away from the actual optimum. I set it to be 10^{-8}, as further tightening the tolerance (up to 10^{-12}) seems to make little difference besides increasing time in my case.

For the convergence criterion of the outer loops: for parametric and FG approach, I set it to stop when the objective value changes less than 10^{-6} in relative terms; for the CSS approach, I set it to stop when the objective value changes less than 10^{-8} in absolute terms (about 10^{-10} in relative terms). It took FG about 2.4 hours to converge, which, given the complexity of the algorithm, I think, is rather fast. CSS converged after about 2.4 hours.

[8] Note that the results reported here are from the estimation where I directly estimate the coefficients of interactions between features and demographic variables. If instead, the random coefficients are regressed on the demographic variables post-estimation, the common parameters would only include the product-market fixed effects, and estimation time could be much less. However, in that case, more burden would be on the correct choice of the grid points for the random coefficients, and the coefficients in front of the demographic variables might be less accurate due to the discreteness of the random coefficients that act as dependent variables.

2.4.3.2 Estimates of random coefficients

Figure 2.3 below shows the correlations between the FG-estimated unobserved demographics and observed household income and household size variables. In Figure 2.3 each point is a household. Note that FG generates a distribution of unobserved demographics for each household. The value of the estimated unobserved demographics for a household in the Figure is just the corresponding distribution's mean value. In my case, since there are many observations on each household, the estimated type distributions are degenerate on one point for about 80% of the households. The correlations are both negative in this case (CSS yields similar results).

Since household income and household size interact with all four features, these correlations will potentially bias all the estimated coefficients for these interaction terms in the parametric approach, where I assume these correlations are zero. Such bias could be especially evident for the parameters related to taste for tuna. However, unlike the simple example in the previous section, it is not straightforward to predict the direction of bias in this case. This is because by using contraction mapping, $\boldsymbol{\delta}$ is now a function of the coefficients of the feature-demographics interaction terms, which makes the utility a non-linear function of these coefficients.

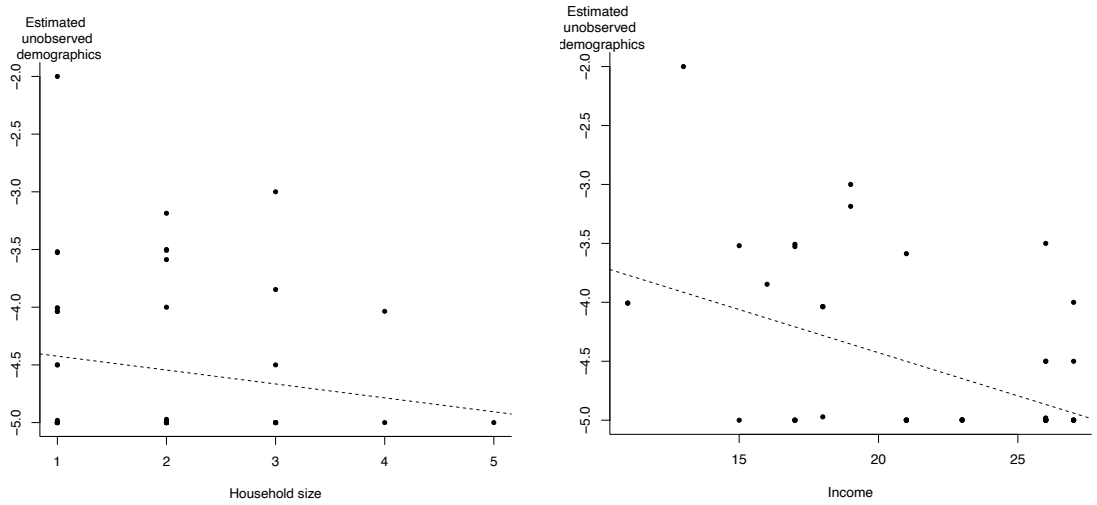

Figure 2.3: Correlation between Household Size and unobserved demographics

Table 2.3 below reports the estimation results of the random coefficients for all three methods. The price unit is dollars per can. The household's annual income is measured in tiers of 5,000 dollar increments (so, for example, 10 means at least 50,000 dollars and less than 55,000 dollars). The household size is simply the number of people in the household.

As is predicted above, the estimated coefficients of the tuna-demographics interactions are quite different between the parametric approach and the other two approaches. For example, FG and CSS suggest that households with more people like tuna more (with estimated coefficients around 0.08-0.24) but the parametric approach suggests the opposite (with estimated coefficient at about -0.02). The parametric approach also overestimates the effect of income on tuna taste, compared to FG and CSS. The results again suggest the conventional parametric approach underestimates the mean price coefficients for tuna, consistent with the findings in the simple example in the Introduction. For the estimated coefficients of the other interaction terms (for chicken, price, and size), in most cases, FG and CSS also produce very different estimates to those by the parametric approach.

Besides these anticipated differences between the conventional approach and the new approaches, it is worth noting that FG and CSS produce somewhat different estimates. For example, FG suggests higher-income households like chicken less, while CSS suggests the opposite; FG suggests larger households are less price-sensitive, while CSS predicts the opposite. These differences are the result of different estimates for the unobserved demographics for the tuna taste. The reason for this could be twofold. First, FG assumes random-effect at individual level, while CSS assumes fixed effect. Second, CSS allows more flexible grid choice, but could also introduce classification error in assigning households into different groups.

On the other hand, qualitatively, these methods mostly yield similar reasonable indications. For example, mean coefficients of price are negative; price sensitivity decreases with household income; larger families prefer a larger size of can. Specifically, except for Tuna, the means of these random coefficients are essentially the same across different estimation methods. This is again consistent with the findings in my earlier simple example. (It seems that in my case, the tuna-demographic interaction correlates little with the

feature variables Price, Size and Chicken, see Table 2.6 in the Appendix A). An interesting but somewhat less relevant observation from Table 2.3 is that the coefficients of household income and household size are negatively correlated: whenever one increases, the other decreases. This is simply because household income and household size are positively correlated, as is shown in Figure 2.7 in the Appendix B.

Table 2.3: Random coefficient estimates

		Mean	Interaction with Demographic Variables	
			Income	Household Size
Variable				
Tuna	Parametric	-3.2409	0.1267	-0.0157
		(0.0007)	(3.36e-05)	(4.13e-06)
	FG	-1.4760	0.0885	0.0834
		(0.0003)	(1.42e-05)	(1.34e-06)
	CSS	-1.4105	0.0647	0.2449
		(0.0001)	(1.19e-05)	(1.22e-04)
Chicken	Parametric	-1.0459	-0.0018	-0.3534
		(0.0002)	(7.62e-06)	(9.88e-07)
	FG	-1.0878	-0.0231	-0.0109
		(0.0002)	(8.38e-06)	(7.95e-07)
	CSS	-1.2034	0.0758	-0.9164
		(0.0001)	(2.20e-06)	(2.61e-05)
Price	Parametric	-5.502	0.3160	-0.0331
		(0.0011)	(5.98e-06)	(1.29e-06)
	FG	-5.127	0.2176	0.0649
		(0.0013)	(2.26e-06)	(5.80e-07)
	CSS	-5.464	0.3536	-0.3440
		(0.0006)	(1.73e-05)	(2.21e-04)
Size of can	Parametric	-0.5464	-0.0239	0.0443
		(0.0024)	(5.79e-06)	(5.39e-05)
	FG	-0.5794	-0.0148	0.0271
		(0.0029)	(7.34e-07)	(9.83e-07)
	CSS	-0.5759	-0.0330	0.1044
		(0.0012)	(2.36e-06)	(1.43e-05)

Note: standard errors in parentheses. Please see Appendix on how the standard errors are computed.

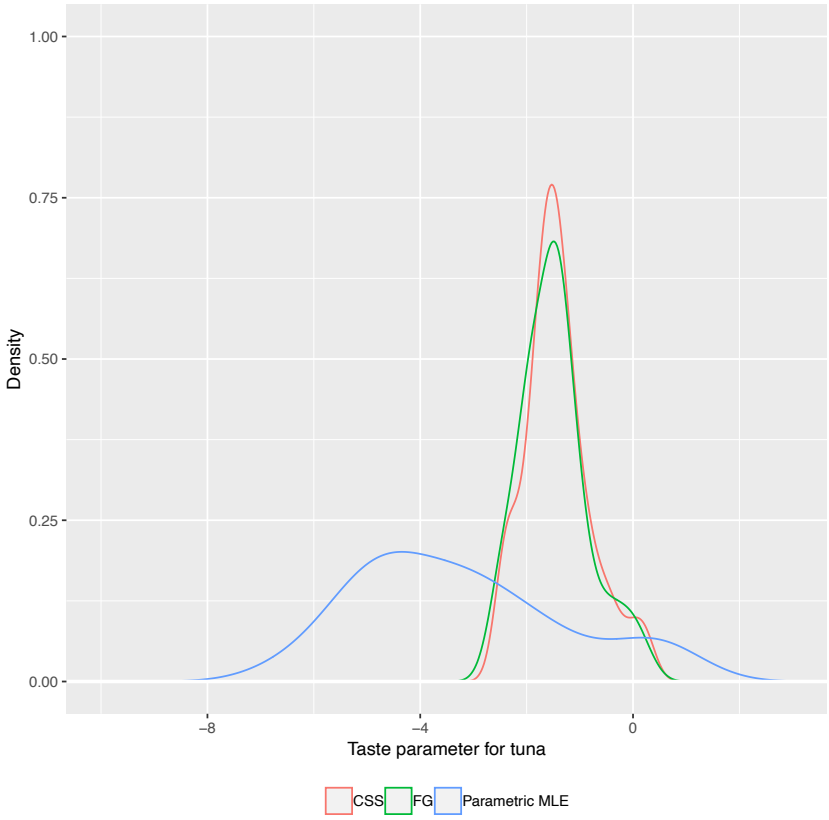

Figure 2.4: Estimated random coefficient distributions of tuna

To give a full picture of random coefficient estimates, I report the estimated standard deviation of the unobserved demographics for tuna: 1.7713 for the parametric approach, 0.7085 for the FG approach, and 0.6644 for the CSS approach. Moreover, Figure 2.4 shows the smoothed distributions of the random coefficients for tuna. Figure 2.8 in the Appendix B shows the smoothed distributions of the random coefficients for chicken, price and size of can.

Focusing on tuna. In Figure 2.4, both FG and CSS produce much narrower taste distribution for tuna. Three factors mainly drive this. First, as is mentioned above, FG and CSS indicate a smaller variation in the unobserved demographics, compared to the parametric approach. Second, as is shown in Figure 2.3, both household size and household income are negatively correlated with unobserved demographics. Third, as is shown in Table 2.3, FG and CSS both produce positive estimates for the two tuna-demographics interaction

terms. The second and third factors together dictate the variations of observed and unobserved demographics will cancel out each other to some extent in determining the variation of the random coefficient for tuna.

One of the noticeable results from my estimation here is that the unobserved demographics plays a very important role in deciding the taste distribution. Focusing on tuna again, Figure 2.9 in the Appendix presents the (de-meaned) estimated distribution of the unobserved demographics for tuna for all three methods. This distribution closely resembles the overall tuna taste distribution in Figure 2.4. One way to measure the contribution of unobserved demographics to the taste dispersion is to compare the variance of the unobserved demographics to the variance of the taste distribution. For the parametric approach, it is 1.7713 versus 1.9489: a ratio of 0.91, which is consistent with the finding in previous works with micro data that the unobserved demographics plays a substantial part. For the FG approach, it is 0.7085 versus 0.6372, and for the CSS approach, it is 0.6644 versus 0.6284. The finding that the variance of the unobserved demographics is actually larger than the variance of the overall taste distribution is new in the literature. This is only possible because my approaches here allows for correlations between observed and unobserved demographics. It turns out in my case that the correlation is negative, explaining this finding. Previous estimation methods, even nonparametric ones, cannot allow such findings. This again highlights the importance of having the flexibility of my approach compared to previous methods.

2.4.3.3 Elasticities

Table 2.4 reports a sample of price elasticities estimated by the three methods. The products reported here are 5 products in a certain week. These products are of similar size and price, so I omit price and size of can in the feature columns. The number at the intersection of row j and column k is the elasticity of product j with respect to the price of product k.

Table 2.4: Elasticity estimates

	Features		1	2	3	4	5
Product Number	Chicken	Tuna					
Parametric 1	1	0	-2.4601	0.0100	0.0045	0.0014	0.0051
2	1	0	0.0126	-2.4626	0.0045	0.0014	0.0051
3	0	1	0.0087	0.0070	-1.9224	0.0070	0.0037
4	0	1	0.0087	0.0070	0.0231	-1.9385	0.0037
5	0	0	0.0116	0.0092	0.0044	0.0013	-2.3732
FG 1	1	0	-2.2487	0.0089	0.0057	0.0017	0.0048
2	1	0	0.0112	-2.2510	0.0057	0.0017	0.0048
3	0	1	0.0110	0.0088	-2.2116	0.0028	0.0048
4	0	1	0.0110	0.0088	0.0091	-2.2179	0.0048
5	0	0	0.0110	0.0088	0.0056	0.0017	-2.2058
CSS 1	1	0	-2.3466	0.0102	0.0049	0.0015	0.0045
2	1	0	0.0127	-2.3492	0.0049	0.0015	0.0045
3	0	1	0.0095	0.0075	-2.4672	0.0037	0.0066
4	0	1	0.0095	0.0075	0.0121	-2.4757	0.0066
5	0	0	0.0102	0.0081	0.0078	0.0024	-2.4353

The substitution patterns observed in Table 2.4 are highly consistent with the estimated taste distributions in Figure 2.4 and Figure 2.8. First, when random coefficients are of variance zero (as in a simple logit model), cross-elasticities within the same column should be the same. This is essentially the case here for chicken products under FG (first two columns in the middle), which is expected as the FG-estimated chicken taste distribution is almost a point mass in Figure 2.8. Second, for products with features that have non-degenerate taste distributions, these products are expected to substitute more with alternatives closer to themselves in the product space. This is the case for all other cross-price elasticity estimates in Table 2.4 other than the FG estimates in the first two columns. For example, looking at the third column of the FG results, since product 3 and 4 have the same features (they are both tuna products) that distinguish them from the other products, the cross-product elasticity from product 4 to the price of product 3 is much larger than other products to the price of product 3. In addition, since the FG and CSS approaches estimate narrower tuna taste distributions,

I expect the cross-price elasticities to vary less within a column for tuna products, compared to the results of the parametric approach. This is the case in column 3 and 4.

As is pointed out by Nevo (2004), how far away the cross-elasticities are from identical within a column can be used to check if the model has overcome the logit restrictions for specific products. If applying this test to results in Table 2.4, it seems that allowing for unobserved demographics is vital for flexible substitution patterns: the differences among cross-price elasticities for tuna products (columns 4 and 5) are significantly larger than those for chicken products (columns 1 and 2). In column 3, the ratio of the largest cross-price elasticity to the smallest one within a column is around 5, while the same ratio in column 1 is only around 1.5.

To further understand substitution patterns towards different alternatives, Figure 2.5 and Figure 2.6 show joint taste distributions for tuna and chicken by these three methods. For example, in Figure 2.5 it seems that tuna lovers generally also like chicken more than a random consumer, while most chicken lovers are those who fall on the left of tuna taste distribution. This explains why in Table 2.4 the parametric approach predicts higher cross-price elasticities towards the non-tuna-non-chicken product (product 5) than towards the tuna products in the first two columns, and why it predicts higher cross-price elasticities towards the chicken products (product 1 and 2) than towards the non-tuna-non-chicken product in columns 3 and 4.

Now compare Figure 2.5 and the right figure in Figure 2.6. A significant difference between the joint chicken-tuna taste distributions estimated by the parametric and CSS approaches is that the CSS approach suggests that tuna lovers do not seem to love chicken significantly more than a random consumer. This could help explain why for the tuna products (columns 3 and 4), CSS indicates the substitutions are more towards the non-tuna-non-chicken products, rather than the chicken products.

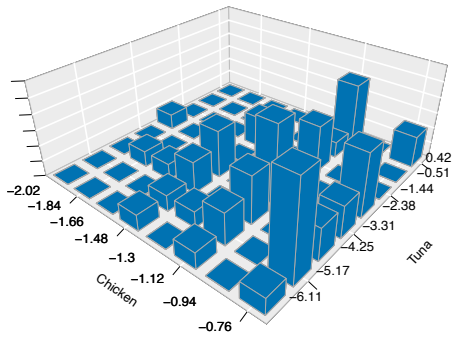

Figure 2.5: Parametric approach: joint taste distribution of tuna and chicken flavor

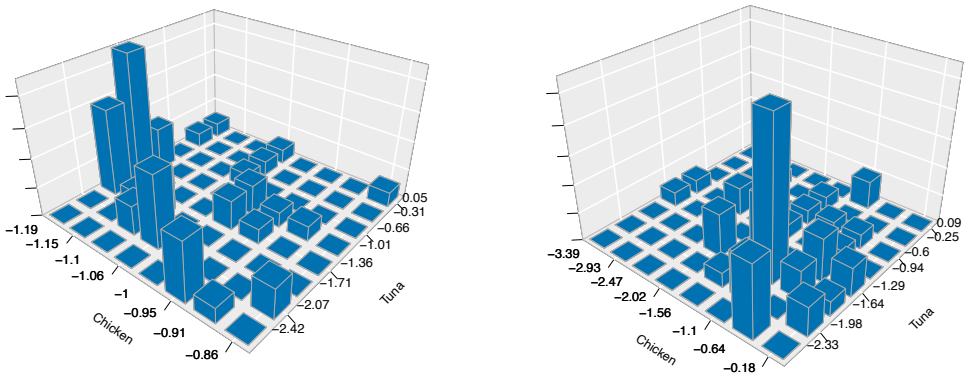

Figure 2.6: Grouped-effect approach: joint taste distribution of tuna and chicken flavor (left: FG, right: CSS)

2.4.3.4 Welfare estimates

To further compare these methods, I conduct separate welfare evaluations in three hypothetical scenarios: a new product, a merger, and a divestiture. I simulate prices and compute welfare changes in two hypothetical scenarios in a week of the year with the following alternative market structures: (1) a merger of 4 firms, and (2) a divestiture of a firm into two separate firms, where the two firms split the (pre-divestiture) market share

of the original firm by an approximate 5:2 ratio. For the "new product" scenario, I consider a new chicken flavor product of a brand A, with a marginal cost equal to the mean of all existing chicken flavor wet food.

Table 2.5: Welfare estimates

		Parametric	FG	CSS
New Product	Δ Profit (%)	+0.3175	+0.2617	+0.2341
	Δ Consumer Surplus (%)	+0.3031	+0.2432	+0.2273
Merger	Δ Profit (%)	+0.0562	+0.0610	+0.0642
	Δ Consumer Surplus (%)	-0.2743	-0.3022	-0.2692
Divestiture	Δ Profit (%)	-0.5287	-0.5718	-0.6665
	Δ Consumer Surplus (%)	+3.9513	+4.6306	+4.139

Table 2.5 reports the welfare evaluations of these hypothetical scenarios. First, for the introduction of the new product, change in welfare is tiny, as the market is already filled with a large number of similar products. All three methods predict an increase in consumer surplus and an increase in overall profit by all firms. To understand why this is the case: holding the prices of existing products unchanged, the immediate effect of the new product tends to (mechanically) increase the consumer surplus, as now consumers simply have more choices. However, the prices will adjust. On the one hand, overall, the prices tend to decrease. On the other hand, as the new product belongs to brand A, brand A products tend to have slightly higher prices, since brand A has more market power in this market with the addition of the product. The total effect, in this case, is a net increase in the consumer surplus, as the benefit of more choices seems to slightly overweight the effect of increasing brand A's market power. Despite qualitatively similar predictions, FG and CSS predict a relatively smaller increase in consumer surplus and profit.

Second, for the merger and divestiture, all three methods are making the qualitatively correct predictions. However, the quantities are again slightly different. For example, in the merger case, the parametric approach predicts a 0.0562% increase in profit, while both FG and CSS predict a larger increase of around 0.0610% to 0.0642%. These differences are not only the result of level differences in their mean price sensitivity but also are affected by differences in random coefficient distributions. For example, FG-estimated taste distributions are narrower than those estimated by the parametric approach. Thus the market is more competitive under

FG than the parametric approach. Gains from merger (and losses from divestiture) for the firms tend to be larger in a more competitive market. This potentially explains why compared to the parametric approach, the FG has larger numbers (in absolute values) for the merger and divestiture.

2.5 Conclusion

This chapter shows how to flexibly estimate a random coefficient logit model of demand using microdata. I provide estimation strategies of two approaches, the first adapting the fixed-grid likelihood method from Malone et al. (2019), and the other adapting a multi-dimensional clustering method from Cheng et al. (2019). By utilizing the microdata at the individual level, one can avoid making independent or parametric assumptions on the unobserved type distributions. With real microdata from the Nielsen Homescan database, I show that the conventional approach with such assumptions produces biased estimates on preferences, elasticities, and welfare measures. In particular, in my application to the cat wet food market, I show that the more flexible approaches uncover negative correlations between observed and unobserved demographics. Such correlations decrease the variations of the corresponding random coefficient distributions. The more flexible estimation methods also yield different predictions on changes in profit and consumer surplus in the hypothetical scenarios of a new product, a merger, and a divestiture. In particular, the conventional parametric approach over-estimates a new product's welfare gains by 20% - 38%. In the hypothetical merger/divestiture cases, the conventional parametric approach underestimates the magnitudes of profit gains and consumer welfare gains by around 10% - 20%.

The estimated models in this paper allow for unobserved demographics in tuna-specific taste. However, the algorithms in this paper can be directly used for multi-dimensional unobserved demographics, if needed. For example, allowing unobserved taste types for the price could have changed the estimated average price sensitivity of the consumers, which, in turn, could have even larger impacts on elasticity and welfare estimates.

One potentially interesting extension in estimating random coefficient models with microdata is to allow for more flexible forms of price endogeneity, as the variations of the price of the same product across individuals

within a market could be considerable. It would be interesting to examine how to modify the current estimation procedures in such scenarios.

References

Manuel Arellano and Stéphane Bonhomme. Robust priors in nonlinear panel data models. Econometrica, 77(2):489–536, 2009.

Steven Berry, James Levinsohn, and Ariel Pakes. Automobile prices in market equilibrium. Econometrica, 63(4):841–890, 1995.

Steven Berry, James Levinsohn, and Ariel Pakes. Differentiated products demand systems from a combination of micro and macro data: The new car market. Journal of Political Economy, 112(1):68–105, 2004.

Stéphane Bonhomme and Elena Manresa. Grouped patterns of heterogeneity in panel data. Econometrica, 83(3):1147–1184, 2015.

Stéphane Bonhomme, Thibaut Lamadon, and Elena Manresa. A distributional framework for matched employer employee data. Econometrica, 87(3):699–739, 2019.

Xu Cheng, Frank Schorfheide, and Peng Shao. Clustering for multi-dimensional heterogeneity. Working paper, 2019.

Christopher Conlon and Jeff Gortmaker. Best practices for differentiated products demand estimation with pyblp. RAND Journal of Economics, 51(4):1108-1161, 2020.

Jean-Pierre Dubé, Jeremy T. Fox, and Che-Lin Su. Improving the numerical performance of static and dynamic aggregate discrete choice random coefficients demand estimation. Econometrica, 80(5):2231–2267, 2012.

Jeremy Fox, Kyoo Kim, and Chenyu Yang. A simple nonparametric approach to estimating the distribution of random coefficients in structural models. Journal of Econometrics, 195(2):236–254, 2016.

Amit Gandhi and Jean-François Houde. Measuring substitution patterns in differentiated products industries. NBER Working Paper, 2019.

Austan Goolsbee and Amil Petrin. The consumer gains from direct broadcast satellites and the competition with cable TV. Econometrica, 72(2):351–381, 2004.

J. Heckman and B. Singer. A method for minimizing the impact of distributional assumptions in econometricmodels for duration data. Econometrica, 52(2): 271-320, 1984.

Jacob Malone, Aviv Nevo, and Jonathan Williams. The tragedy of the last mile: Economic solutions to congestion in broadband networks. Working paper, 2019.

Aviv Nevo. Measuring market power in the ready-to-eat cereal industry. Econometrica, 69(2): 307–342, 2001.

Aviv Nevo. A practitioner's guide to estimation of random-coefficients logit models of demand. Journal of Economics and Management Strategy, 9(4):513–548, 2004.

Aviv Nevo, John Turner, and Jonathan Williams. Usage-based pricing and demand for residential broadband. Econometrica, 84(2):411–443, 2016.

Amil Petrin. Quantifying the benefits of new products: The case of the minivan. Journal of Political Economy, 110(4):705–729, 2002.

Amil Petrin and Boyoung Seo. Identification and estimation of discrete choice demand models when observed and unobserved characteristics are correlated. Working Paper, 2019.

Liangjun Su, Zhentao Shi, and Peter Phillips. Identifying latent structures in panel data. Econometrica, 84(6):2215–2264, 2016.

Ravi Varadhan and Christohpe Roland. Simple and globally convergent methods for accelerating the convergence of any EM algorithm. Scandinavian Journal of Statistics, 35(2):335-353, 2008.

Appendix

Appendix A Further analysis for the simple example without endogeneity

This section presents an ad-hoc regression analysis for the effect of "omitting" the correct unobserved demographics variable in the simple example presented in the Introduction section. It is ad-hoc because of two reasons: first, the estimation conducted in the example is not linear regression; second, technically in the conventional parametric approach, the unobserved demographics is not missing, but assumed to be independent of all other regressors. Nevertheless, I compute the interaction of tuna dummy and FG-estimated unobserved demographics, and regress this interaction term on the rest of the regressors: interaction of tuna and household size (observed demographics), price, size of can, chicken dummy, and tuna dummy. Table 2.6 reports the regression results. The coefficients on the interaction term are 2.35, which predicts the conventional method should be overestimating θ by around 2.35, which is the case. The coefficients on the interaction term are 2.35, and the coefficients on the tuna variable is −5.86. If this is technically indeed a classic omitted variable bias case, this predicts the conventional method should overestimate θ by around 2.35 and it should underestimate γ_4 by around 5.86, which is the case. The correlations between the "omitted variable" and other three regressors seem to be tiny, which is consistent with the fact that estimated γ_1 to γ_3 essentially do not change across estimation methods in Table 2.1.

Table 2.6: "Omitted variable" analysis

	Dependent variable:
	Tuna × Unobserved demographics
Tuna x Household Size	2.350***
	(0.002)
Price	0.031***
	(0.002)
Size of can	- 0.003***
	(0.0002)
Chicken	- 0.013***
	(0.001)
Tuna	- 5.864***
	(0.004)

Note: *p<0.1; **p<0.05; ***p<0.01. Standard errors in parentheses.

Appendix B Additional results for the elasticity estimates of the full model

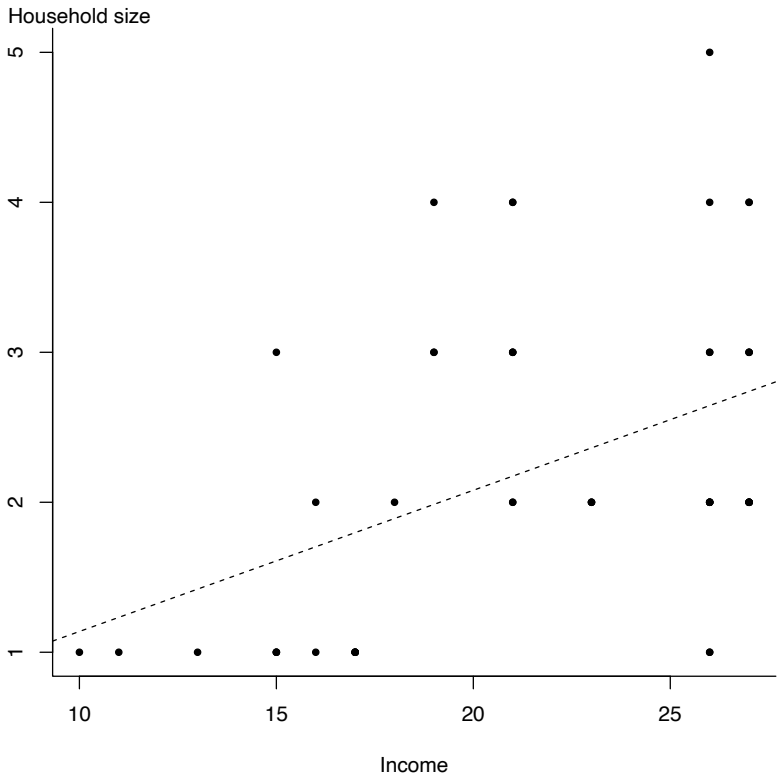

Figure 2.7: Correlation between Household Size and Income

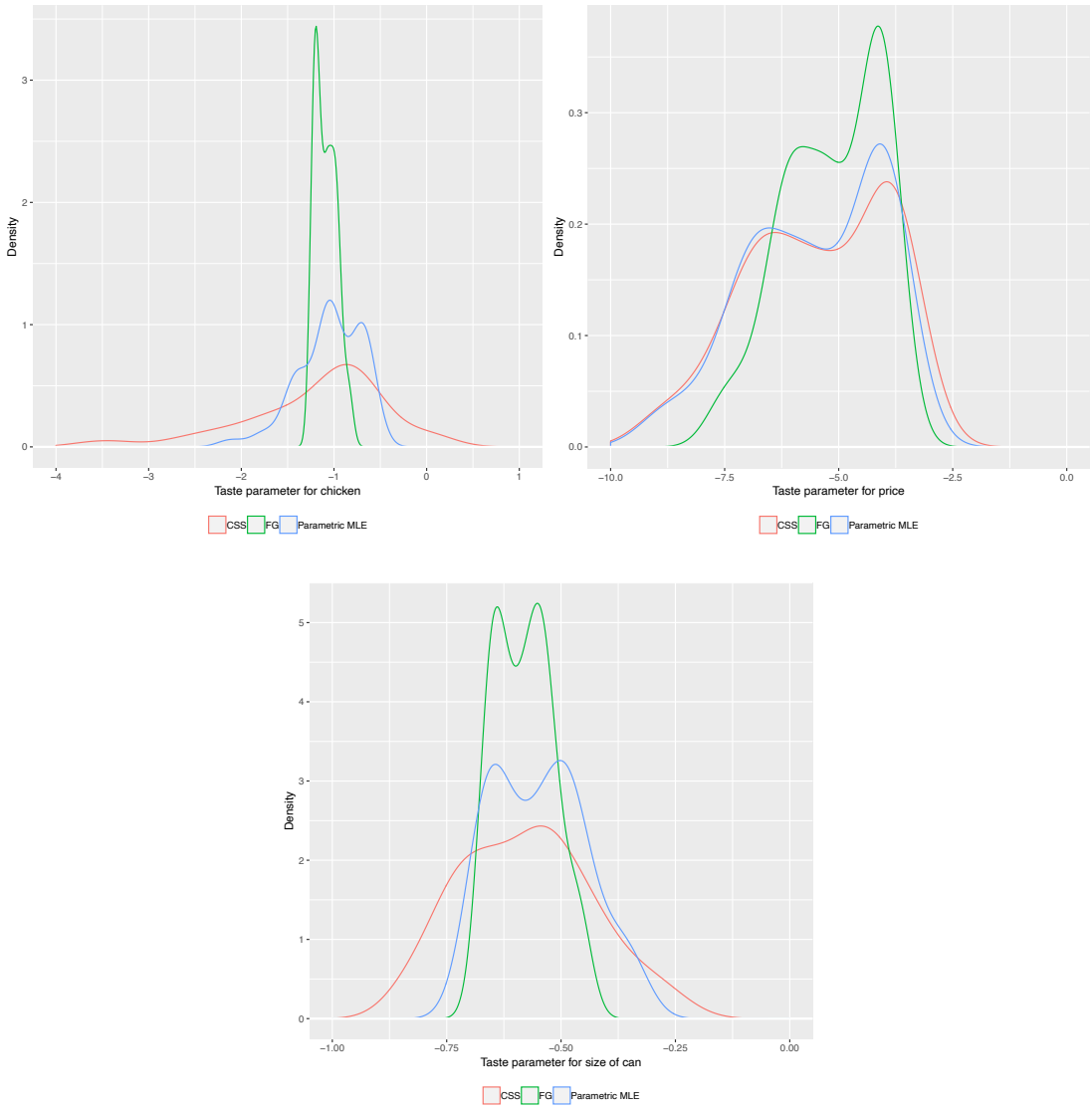

Figure 2.8: Estimated random coefficient distributions of chicken flavor, price and size of can

Looking at the chicken taste distribution, the CSS-produced distribution is the one that spread out most, followed by the parametric approach. The distribution by FG is so narrow that it is not far away from a degenerated mass around -1. Unlike the case for tuna, the distribution for chicken is solely determined by the observed demographics, thus the differences across results by different methods can be understood by looking at estimates in Table 2.3 (This is also true for the estimated distributions for price and size of can

For example, looking at the coefficients of chicken-demographics interactions, CSS estimates (0.076, -0.916) are way much larger than those estimated by FG (-0.023, -0.011) in terms of absolute values, which explains the discrepancies in their estimated distributions. Across all three methods, FG always produces the most narrow distribution for each of these four features. CSS estimates an essentially identical distribution for the price, wider distributions for chicken and size of can, and narrower distribution for tuna compared to the parametric approach. Therefore, among all three hypothetical worlds described by these methods, the one narrated by FG is the most competitive, where household tastes are least diverse.

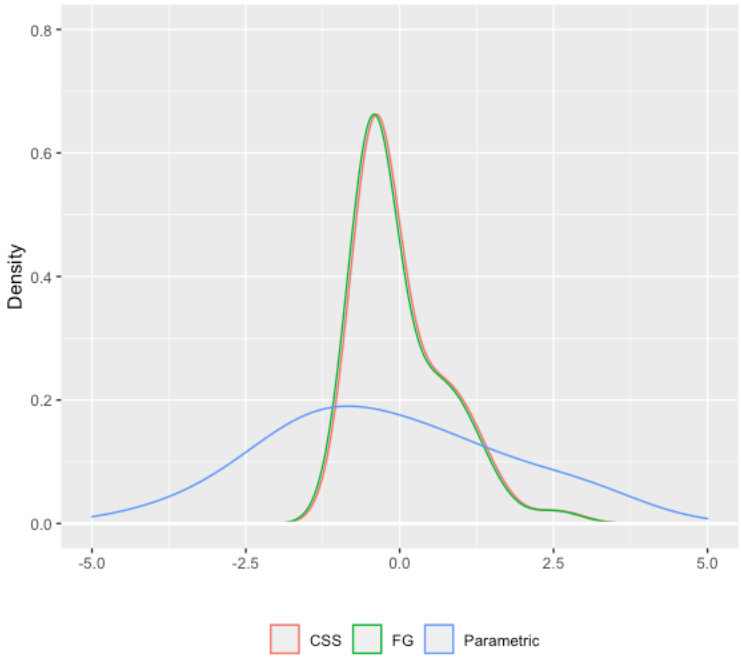

Figure 2.9: Distribution of unobserved demographics for tuna taste

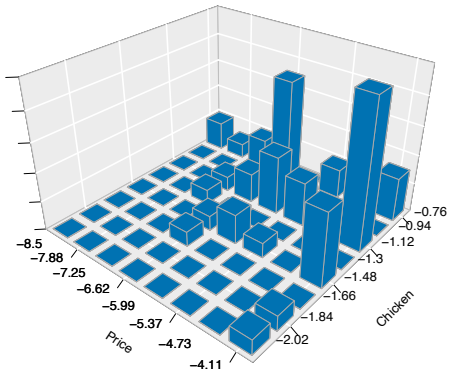

Figure 2.10: Parametric approach: joint taste distribution of price and chicken flavor

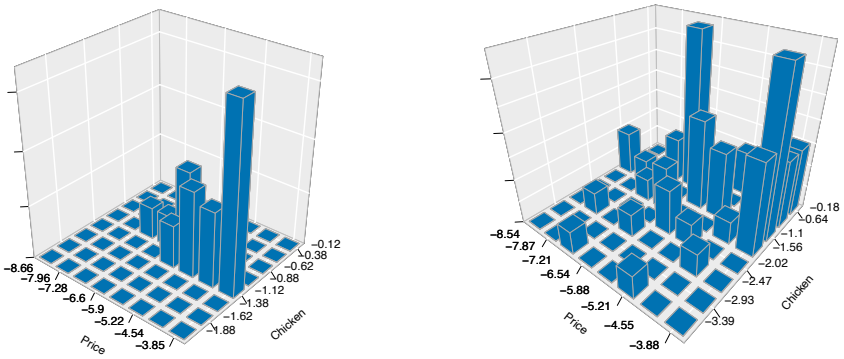

Figure 2.11: Grouped-effect approach: joint taste distribution of price and chicken flavor

Note: FG on the left, and CSS on the right

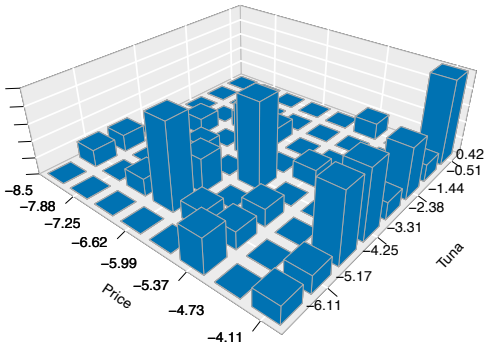

Figure 2.12: Parametric approach: joint taste distribution of price and tuna flavor

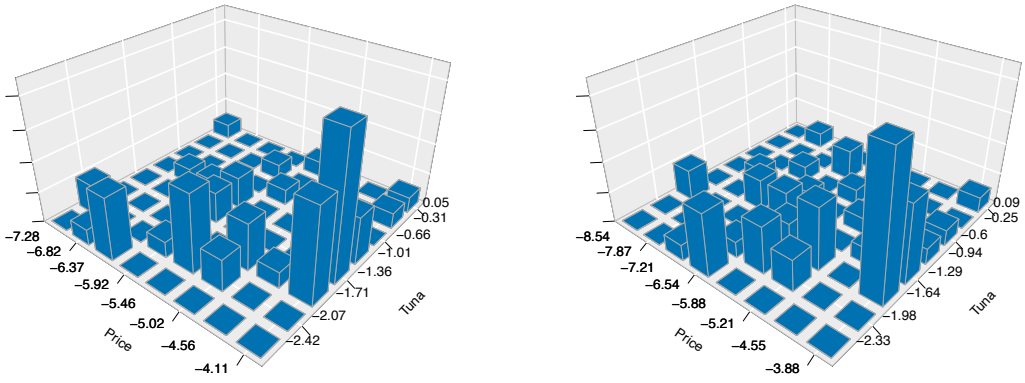

Figure 2.13: Grouped-effect approach: joint taste distribution of price and tuna

Note: FG on the left, and CSS on the right

Appendix C details on statistical inference

C.1 Parametric approach

The parametric approach solves the following problem:

$$(\boldsymbol{\theta}, \Gamma) = argmax \sum_{i,j,t} y_{ijt} \log\left(\frac{1}{M}\sum_{m=1}^{M} \frac{e^{U_{ijt,m}}}{\sum_{k\in C_t} e^{U_{ikt,m}}}\right) \quad (4)$$

where,

$$U_{ijt,m} = \sum_{k=1}^{K}\sum_{l=1}^{L} \theta_{k,l} x_{jt,k} D_{i,l} + x'_{jt}\Gamma\tilde{v} + \widehat{\delta_{jt}}(\theta, \Gamma)$$

$$\hat{\delta}(\theta, \Gamma) = \widehat{S^{-1}}(s|\theta, \Gamma).$$

Denote $\eta = (\boldsymbol{\theta}, \Gamma)$, and

$$l_{ijt} = \log\left(\frac{1}{M}\sum_{m=1}^{M} \frac{e^{U_{ijt,m}}}{\sum_{k\in C_t} e^{U_{ikt,m}}}\right)$$

Let N be the number of observations (this should be the total number of decisions of all consumers in all markets). The asymptotic variance-covariance matrix of $(\boldsymbol{\theta}, \Gamma)$ is given by

$$V = \frac{I^{-1}}{N}$$

where I is the Fisher Information matrix. The element at a-th row and b-th column of I is

$$I_{ab} = \sum_{ijt \text{ s.t. } y_{ijt}=1} \frac{\partial}{\partial \eta_a} l_{ijt} \frac{\partial}{\partial \eta_b} l_{ijt} / N$$

Thus the goal is to derive $\frac{\partial}{\partial \eta_a} l_{ijt}$ for all the parameters in η. The first parameter in η, θ_{11} is the coefficient in front of the first feature-observed demographics interaction term. Let M be the number of types for the unobserved demographics, s_{ijtm} be the choice probability of individual I choosing product j in market t if the individual is of type m, s_{ijt} be the expected choice probability of individual i choosing product j in market t (expectation taking over M types):

$$\frac{\partial l_{ijt}}{\partial \theta_{11}} = y_{ijt} \frac{1}{M} \sum_{m=1}^{M} \left(\frac{\partial U_{ijtm}}{\partial \theta_{11}} - \sum_{k=1}^{J+1} s_{iktm} \frac{\partial U_{iktm}}{\partial \theta_{11}} \right) s_{ijtm} / s_{ijt}$$

$$\frac{\partial U_{ijtm}}{\partial \theta_{11}} = x_{jt,1} D_{i,1} + \frac{\partial \widehat{\delta_{jt}}}{\partial \theta_{11}} \tag{6}$$

The derivative for Γ can be derived similarly, with:

$$\frac{\partial U_{ijtm}}{\partial \Gamma} = x_{jt,4} v_m + \frac{\partial \widehat{\delta_{jt}}}{\partial \Gamma} \tag{7}$$

Now I compute $\frac{\partial \widehat{\delta_{jt}}}{\partial \theta_{11}}$ and $\frac{\partial \widehat{\delta_{jt}}}{\partial \Gamma}$. Let s_{jt} be the market share of product j in market t, and δ_t be the vector of product-market fixed effects in market t. By Implicit Function theorem,

$$\frac{\partial \widehat{\delta_t}}{\partial \theta_{11}} = \begin{bmatrix} \frac{\partial s_{1t}}{\partial \delta_{1t}} & \cdots & \frac{\partial s_{1t}}{\partial \delta_{Jt}} \\ \vdots & \ddots & \vdots \\ \frac{\partial s_{Jt}}{\partial \delta_{1t}} & \cdots & \frac{\partial s_{Jt}}{\partial \delta_{Jt}} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial s_{1t}}{\partial \theta_{11}} \\ \vdots \\ \frac{\partial s_{Jt}}{\partial \theta_{11}} \end{bmatrix}$$

where

$$\frac{\partial s_{jt}}{\partial \delta_{kt}} = \begin{cases} s_{jt} - \sum_i \frac{1}{M} \sum_m s_{ijtm} s_{ijtm} & if\ k = j \\ 0 - \sum_i \frac{1}{M} \sum_m s_{ijtm} s_{iktm} & if\ k \neq j \end{cases}$$

$$\frac{\partial s_{jt}}{\partial \theta_{11}} = \sum_i \frac{1}{M} \sum_m s_{ijtm} \left(\frac{\partial U_{ijtm}(\boldsymbol{\theta}, \Gamma, \delta)}{\partial \theta_{11}} - \sum_k s_{iktm} \frac{\partial U_{iktm}(\boldsymbol{\theta}, \Gamma, \delta)}{\partial \theta_{11}} \right)$$

Note that here the partial derivatives of U_{ijtm} is assuming it is a function of $(\boldsymbol{\theta}, \Gamma)$ **and** δ, so that there won't be $\frac{\partial \widehat{\delta_{jt}}}{\partial \theta_{11}}$ term in these partial derivatives. $\frac{\partial \widehat{\delta_{jt}}}{\partial \Gamma}$ can be derived in a similar fashion.

C.2 FG approach

The FG approach solves the following problem:

$$\boldsymbol{\theta} = argmax \sum_{i,j,t} y_{ijt} \log \left(\sum_{m=1}^{M} \frac{e^{U_{ijt,m}}}{\sum_{k \in C_t} e^{U_{ikt,m}}} P_v(\boldsymbol{v_m} | \boldsymbol{X_i}, \boldsymbol{Y_i}, \boldsymbol{X}; \boldsymbol{\theta}) \right) \tag{5}$$

where,

$$U_{ijt,m} = \sum_{k=1}^{K}\sum_{l=1}^{L}\theta_{k,l}x_{jt,k}D_{i,l} + \sum_{k=1}^{K}x_{jt,k}v_{m,k} + \widehat{\delta_{jt}}(\boldsymbol{\theta})$$

$$\widehat{\boldsymbol{\delta}}(\boldsymbol{\theta}) = \widehat{S^{-1}}(\boldsymbol{s}|\boldsymbol{\theta})$$

$$P_v(\boldsymbol{v_m}|\boldsymbol{D_i}, \boldsymbol{Y_i}, \boldsymbol{X}; \boldsymbol{\theta}) = \frac{L(\boldsymbol{Y_i}, \boldsymbol{D_i}, \boldsymbol{X}|\boldsymbol{v_m}; \boldsymbol{\theta})\pi_m}{\sum_{m=1}^{M} L(\boldsymbol{Y_i}, \boldsymbol{D_i}, \boldsymbol{X}|\boldsymbol{v_m}; \boldsymbol{\theta})\pi_m}$$

Denote

$$l_{ijt} = \log\left(\frac{1}{M}\sum_{m=1}^{M}\frac{e^{U_{ijt,m}}}{\sum_{k\in C_t} e^{U_{ikt,m}}}\right)$$

$$P_{im} = P_v(\boldsymbol{v_m}|\boldsymbol{D_i}, \boldsymbol{Y_i}, \boldsymbol{X}; \boldsymbol{\theta})$$

then

$$\frac{\partial l_{ijt}}{\partial \theta_{11}} = y_{ijt}\sum_{m=1}^{M}\left(\frac{\partial U_{ijtm}}{\partial \theta_{11}} - \sum_{k=1}^{J+1} s_{iktm}\ \frac{\partial U_{iktm}}{\partial \theta_{11}}\right)s_{ijtm}P_{im}/s_{ijt} + y_{ijt}\sum_{m=1}^{M} s_{ijtm}\frac{\partial P_{im}}{\partial \theta_{11}}/s_{ijt}$$

where

$$\frac{\partial P_{im}}{\partial \theta_{11}} = P_{im}\left(A_{im} - \sum_{m'} A_{im'}\,P_{im'}\right) \tag{8}$$

$$A_{im} = \sum_{jt} y_{ijt}\frac{\partial s_{ijtm}}{\partial \theta_{11}}\frac{1}{s_{ijtm}} \tag{9}$$

$$\frac{\partial s_{ijtm}}{\partial \theta_{11}} = \left(\frac{\partial U_{ijtm}}{\partial \theta_{11}} - \sum_{k=1}^{J+1} s_{iktm}\ \frac{\partial U_{iktm}}{\partial \theta_{11}}\right)s_{ijtm} \tag{10}$$

The terms $\frac{\partial U_{ijtm}}{\partial \theta_{11}}$ are given in Equations (6), where $\frac{\partial \widehat{\delta_{jt}}}{\partial \theta_{11}}$ we now derive, again using the Implicit Function theorem.

$$\frac{\partial s_{jt}}{\partial \delta_{kt}} = \begin{cases} s_{jt} - \sum_i \frac{1}{M}\sum_m s_{ijtm}s_{ijtm} + \sum_{im} s_{ijtm}\frac{\partial P_{im}(\boldsymbol{\theta}, \delta)}{\partial \delta_{kt}} & if\ k = j \\ 0 - \sum_i \frac{1}{M}\sum_m s_{ijtm}s_{iktm} + \sum_{im} s_{ijtm}\frac{\partial P_{im}(\boldsymbol{\theta}, \delta)}{\partial \delta_{kt}} & if\ k \neq j \end{cases}$$

$$\frac{\partial P_{im}(\boldsymbol{\theta}, \delta)}{\partial \delta_{kt}} = P_{im}\left(B_{imkt} - \sum_{m'} B_{im'kt}\,P_{im'}\right)$$

$$B_{imkt} = y_{ikt} - \sum_{jt} y_{ijt}\, s_{iktm}$$

Now, for the terms $\frac{\partial s_{jt}}{\partial \theta_{11}}$,

$$\frac{\partial s_{jt}}{\partial \theta_{11}} = \sum_i \sum_m \left(\frac{\partial s_{ijtm}(\boldsymbol{\theta}, \delta)}{\partial \theta_{11}} + \sum_k s_{ijtm} \frac{\partial P_{im}(\boldsymbol{\theta}, \delta)}{\partial \theta_{11}} \right)$$

Note here that P_{im} is again a function of $\boldsymbol{\theta}$ **and** δ, so that when using Equations (6) and (8) – (10) in the above expression, one should omit the terms $\frac{\partial \widehat{\delta_{jt}}}{\partial \theta_{11}}$.

C.3 CSS approach

The CSS approach solves the following problem:

$$(\boldsymbol{\theta}, \{\boldsymbol{v_k}\}_{k=1}^K, \{\boldsymbol{B_i}\}_{i=1}^I) = argmax \sum_{i,j,t} y_{ijt} \log \left(\frac{e^{U_{ijt}}}{\sum_{k \in C_t} e^{U_{ikt}}} \right)$$

where,

$$U_{ijt} = \sum_{k=1}^K \sum_{l=1}^L \theta_{k,l} x_{jt,k} D_{i,l} + \sum_{k=1}^K x_{jt,k} v_{B_{ik},k} + \widehat{\delta_{jt}}(\boldsymbol{\theta}, \{\boldsymbol{v_k}\}_{k=1}^K, \{\boldsymbol{B_i}\}_{i=1}^I)$$

$$\widehat{\boldsymbol{\delta}}(\boldsymbol{\theta}, \{\boldsymbol{v_k}\}_{k=1}^K, \{\boldsymbol{B_i}\}_{i=1}^I) = \widehat{S^{-1}}(\boldsymbol{s}|\boldsymbol{\theta}, \{\boldsymbol{v_k}\}_{k=1}^K, \{\boldsymbol{B_i}\}_{i=1}^I)$$

Denote

$$l_{ijt} = \log \left(\frac{1}{M} \sum_{m=1}^M \frac{e^{U_{ijt,m}}}{\sum_{k \in C_t} e^{U_{ikt,m}}} \right)$$

I compute the standard errors of $(\boldsymbol{\theta}, \{\boldsymbol{v_k}\}_{k=1}^K)$, taking the group memberships $\{\boldsymbol{B_i}\}_{i=1}^I$ as given. The CSS approach assign a single type for each individual, which makes the derivations easier, compared with the previous two approaches. I have

$$\frac{\partial l_{ijt}}{\partial \theta_{11}} = y_{ijt} \left(\frac{\partial U_{ijt}}{\partial \theta_{11}} - \sum_{k=1}^{J+1} s_{ikt} \frac{\partial U_{ikt}}{\partial \theta_{11}} \right)$$

where $\frac{\partial U_{ijt}}{\partial \theta_{11}}$ is again given by Equation (6). For the group-specific parameters $\{\boldsymbol{v_k}\}_{k=1}^K$, I have

$$\frac{\partial U_{ijt}}{\partial v_{km}} = \begin{cases} x_{jt,k} + \frac{\partial \widehat{\delta_{jt}}}{\partial v_{km}} & if B_{ik} = m \\ \frac{\partial \widehat{\delta_{jt}}}{\partial v_{km}} & if B_{ik} \neq m \end{cases} \quad (11)$$

To calculate $\frac{\partial \widehat{\delta_{jt}}}{\partial \theta_{11}}$ *and* $\frac{\partial \widehat{\delta_{jt}}}{\partial v_{km}}$ with the Implicit Function theorem, I have

$$\frac{\partial s_{jt}}{\partial \delta_{kt}} = \begin{cases} s_{jt} - \sum_i s_{ijt} s_{ijt} & if\ k = j \\ 0 - \sum_i s_{ijt} s_{ikt} & if\ k \neq j \end{cases}$$

$$\frac{\partial s_{jt}}{\partial \theta_{11}} = \sum_i s_{ijt} \left(\frac{\partial U_{ijt}(\boldsymbol{\theta}, \{\boldsymbol{v_k}\}_{k=1}^K, \delta)}{\partial \theta_{11}} - \sum_k s_{ikt} \frac{\partial U_{ikt}(\boldsymbol{\theta}, \{\boldsymbol{v_k}\}_{k=1}^K, \delta)}{\partial \theta_{11}} \right)$$

$$\frac{\partial s_{jt}}{\partial v_{km}} = \sum_i s_{ijt} \left(\frac{\partial U_{ijt}(\boldsymbol{\theta}, \{\boldsymbol{v_k}\}_{k=1}^K, \delta)}{\partial v_{km}} - \sum_k s_{ikt} \frac{\partial U_{ikt}(\boldsymbol{\theta}, \{\boldsymbol{v_k}\}_{k=1}^K, \delta)}{\partial v_{km}} \right)$$

Note here that U_{ijt} is again a function of $\boldsymbol{\theta}, \{\boldsymbol{v_k}\}_{k=1}^K, \delta$, so that when using Equations (6) and (11) in the above expression, one should omit the terms $\frac{\partial \widehat{\delta_{jt}}}{\partial \theta_{11}}$ and $\frac{\partial \widehat{\delta_{jt}}}{\partial v_{km}}$.

C.4 Implied estimates

Some of the estimates (e.g. mean of a random coefficient) are functions of both first and second step estimates. To correctly compute their standard errors using the delta method, one needs to take into account the fact that the dependent variable in the second step, δ, is itself estimated in the first step. Since I have obtained the derivatives of δ and P (in the case of FG) with respect to the first step parameters, I use them directly together with the delta method to compute the standard errors for these calculated estimates.

As an example, here I describe how to compute the standard errors of the mean price coefficients for the three approaches (the standard errors are reported in Table 2.3). For the parametric approach, the mean price coefficient, $\overline{\beta_p}$, is calculated by

$$\overline{\beta_p} = \widehat{\gamma_p} + \widehat{\theta_{11}}\overline{Income} + \widehat{\theta_{12}}\overline{HS}$$

where $\widehat{\gamma_p}$ is estimated price coefficient from the second step IV regression, $\widehat{\theta_{11}}$ is estimated coefficient for the price-income interaction and $\widehat{\theta_{12}}$ is estimated coefficient for the price-household size interaction. The latter two coefficients are estimated in the first step estimation. $\overline{Income}$ stands for mean income in the sample, and $\overline{HS}$ stands for mean household size. By the Delta method, the variance of $\overline{\beta_p}$, is given by the first diagonal element of the following matrix $V_{\overline{\beta_p}}$:

$$V_{\overline{\beta_p}} = \nabla \Omega \nabla'$$

where Ω is the variance-covariance matrix of $\eta = (\boldsymbol{\theta}, \Gamma)$, and

$$\nabla = P_{zw} \frac{\partial \delta}{\partial \eta} + E$$

$$P_{zw} = (Z'W(W'W)^{-1}W'Z)^{-1}Z'W(W'W)^{-1}W'$$

$$\frac{\partial \boldsymbol{\delta}}{\partial \eta} = \left(\frac{\partial \boldsymbol{\delta}}{\partial \theta_{11}}, \ldots, \frac{\partial \boldsymbol{\delta}}{\partial \theta_{42}}, \frac{\partial \boldsymbol{\delta}}{\partial \Gamma}\right)$$

W is the second step instrument matrix, Z is the second step regressor matrix, and E is given by

$$E = \begin{pmatrix} \overline{Income}, \overline{HS}, 0,0,0,0,0,0,0 \\ 0,0, \overline{Income}, \overline{HS}, 0,0,0,0,0 \\ 0,0,0,0, \overline{Income}, \overline{HS}, 0,0,0 \\ 0,0,0,0,0,0, \overline{Income}, \overline{HS}, 0 \end{pmatrix}$$

For FG, the mean price coefficient is[9]

$$\overline{\beta_p} = \widehat{\gamma_p} + \widehat{\theta_{11}}\overline{Income} + \widehat{\theta_{12}}\overline{HS} + \frac{1}{IM}\sum_{i,m} v_{m,1} P_{im}$$

Let $v_m = (v_{m,1}, \ldots, v_{m,4})'$. I have

$$\nabla = P_{zw} \frac{\partial \boldsymbol{\delta}}{\partial \boldsymbol{\theta}} + E + \frac{1}{IM}\sum_{i,m} v_m \otimes \frac{\partial P_{im}}{\partial \theta}$$

$$\frac{\partial \boldsymbol{\delta}}{\partial \eta} = \left(\frac{\partial \boldsymbol{\delta}}{\partial \theta_{11}}, \ldots, \frac{\partial \boldsymbol{\delta}}{\partial \theta_{42}}\right)$$

[9] In my application, I don't allow for unobserved demographics in price taste, in which case there is no last term in the equation. This applies to both FG and CSS.

$$\frac{\partial P_{im}}{\partial \boldsymbol{\theta}} = \left(\frac{\partial P_{im}}{\partial \theta_{11}}, \ldots, \frac{\partial P_{im}}{\partial \theta_{42}}\right)$$

For CSS, the mean price coefficient is

$$\overline{\beta_p} = \widehat{\gamma_p} + \widehat{\theta_{11}}\overline{Income} + \widehat{\theta_{12}}\overline{HS} + \frac{1}{I}\sum_i v_{B_{i1},1}$$

where B_{i1} is the membership of individual i in the first dimension (price taste). Let $\eta = (\boldsymbol{\theta}, \{\boldsymbol{v_k}\}_{k=1}^{K})$, I have

$$\nabla = P_{zw}\frac{\partial \boldsymbol{\delta}}{\partial \boldsymbol{\eta}} + E + \frac{1}{I}\sum_i \frac{\partial v_{B_i}}{\partial \boldsymbol{\eta}}$$

$$\frac{\partial \boldsymbol{\delta}}{\partial \boldsymbol{\eta}} = \left(\frac{\partial \boldsymbol{\delta}}{\partial \theta_{11}}, \ldots, \frac{\partial \boldsymbol{\delta}}{\partial \theta_{42}}, \frac{\partial \boldsymbol{\delta}}{\partial v_{11}}, \ldots, \frac{\partial \boldsymbol{\delta}}{\partial v_{KM}}\right)$$

$$\frac{\partial v_{B_{i1}}}{\partial \boldsymbol{\eta}} = \left(\frac{\partial v_{B_{i1},1}}{\partial \boldsymbol{\eta}}, \frac{\partial v_{B_{i2},2}}{\partial \boldsymbol{\eta}}, \frac{\partial v_{B_{i3},3}}{\partial \boldsymbol{\eta}}, \frac{\partial v_{B_{i4},4}}{\partial \boldsymbol{\eta}}\right)$$

$$\frac{\partial v_{B_{i1},1}}{\partial \boldsymbol{\eta}} = \left(\underbrace{0, \ldots, 0}_{8}, \kappa_{ik}, \underbrace{0, \ldots, 0}_{(K-1)M}\right)$$

$$\frac{\partial v_{B_{i2},2}}{\partial \boldsymbol{\eta}} = \left(\underbrace{0, \ldots, 0}_{8+M}, \kappa_{ik}, \underbrace{0, \ldots, 0}_{(K-2)M}\right)$$

$$\frac{\partial v_{B_{i3},3}}{\partial \boldsymbol{\eta}} = \left(\underbrace{0, \ldots, 0}_{8+2M}, \kappa_{ik}, \underbrace{0, \ldots, 0}_{(K-3)M}\right)$$

$$\frac{\partial v_{B_{i4},4}}{\partial \boldsymbol{\eta}} = \left(\underbrace{0, \ldots, 0}_{8+3M}, \kappa_{ik}\right)$$

Here, K is the number of dimensions (where unobserved demographics are allowed, assumed to be 4 in the equations above), M is the number of groups in each dimension (assuming all dimensions have the same number), and $\boldsymbol{\kappa_{ik}}$ is a M-by-1 matrix, where the mth element equals 1 if $B_{ik} = m$, and 0 otherwise.

CHAPTER 3

REVISITING THE EFFECT OF CLIMATE ON PRODUCTIVITY OF CHINESE MANUFACTURING FIRMS

BY JOSE MIGUEL ABITO and RUIZHI MA

3.1 Introduction

Climate change is predicted to permanently alter the current temperature level, potentially impacting many economic activities in the future. Zhang et al. (2018) quantify the effect of changing temperature on the productivity of Chinese manufacturing firms in the 2040-2059 horizon, using the production function estimation method from Olley and Pakes (1996) (henceforth OP). However, OP assumes the unobserved productivity follows a Markov process, and does not allow for firm-level fixed effects in productivity. Abito (2020) proposes a new production function estimation method allowing for a firm fixed effect component in unobserved productivity. Abito (2020) shows the new method works better than existing ones in scenarios where the simple Markov assumption fails.

Allowing for a permanent component in productivity in production function seems to be important in the case of Chinese manufacturing firms. For example, during the sample period of Zhang et al. (2018), many Chinese firms are owned by the state, while many others are privately-owned, and such ownership difference could be associated with a significant performance gap.

This chapter aims to estimate the production functions of Chinese firms with the same sample in Zhang et al. (2018), but with the method from Abito (2020), and predict the effect of temperature on the newly estimated productivity. We also compare our production function estimates and climate change predictions with those produced by OP and Ackerberg et al. (2015) (henceforth ACF). We choose OP because it is the method used in Zhang et al. (2018). We choose ACF because its assumptions are identical to that of Abito (2020), except

that the latter allows for firm fixed effects in productivity. Therefore, the comparison between ACF and Abito (2020) will highlight the potential bias from ignoring firm fixed effects in production function estimations.

With Abito (2020)'s productivity, we first show that most of the 36 industries in our sample have persistent differences in firm-specific productivity. The heterogeneity in unobserved productivity varies a lot across industries. We also find in 24 of the 36 industries, ACF and Abito (2020) produce statistically different input elasticity estimates, and in many of the 24 industries, the differences are also economically considerable. In industries where the firm-specific productivity differences are largest and persistent, the differences between ACF and Abito (2020)'s elasticity estimates are the largest.

We then examine the effect of weather on productivity, and use the estimates to predict the mean level of productivity in 2040-2042 for each firm in our sample. Comparing to the firms' historical mean levels of productivity during 1998-2007, we find the productivity will be lowered by about -4.07% in the future on average across firms, with the 95% confidence interval [-7.01, -1.13]. With ACF's productivity, the prediction is a 4.65% decrease, with the 95% confidence interval [-7.20, -2.10]. Looking at individual industries, we find 3 industries (communication equipment, general machinery, and special machinery) where Abito (2020) and ACF produce significantly different predictions on the effect of climate change on productivity. These industries amount to about 17% of the real value-added output in our sample and are those with the largest and persistent difference in firm-specific productivity. Therefore we conclude that it is crucial to allow firm fixed effects in production function estimation to get credible predictions in our analysis.

The rest of the chapter is organized as follows. Section 3.2 described the model used in our analysis. Section 3.3 explains our data sources. Section 3.4 illustrates summary statistics of the cleaned and merged dataset. Section 3.5 presents reduced-form evidence of the productivity dynamics, and relates them to elasticity estimates. Section 3.6 is the results of our analysis. Section 3.7 concludes, followed by an appendix for details on data cleaning and quality checks.

3.2 Model and estimation

Let y_{it}, l_{it}, k_{it} be the logarithm of output, labor input and capital of firm i in period t. The production function to be estimated is

$$y_{it} = \beta_l l_{it} + \beta_k k_{it} + w_{it} + \epsilon_{it}$$

where w_{it} is observed by firm at period t, but unobserved by the econometrician, and ϵ_{it} is the output shock unexpected to the firm. The sum of the two parts, $w_{it} + \epsilon_{it}$, is what usually been called total factor productivity (TFP). The term w_{it} is potentially correlated with inputs, causing endogeneity problem. To address this, the standard approach is to assume another variable, like investment (OP) or intermediate/material input (LP and ACF), is strictly increasing in w_{it}, thus can be used to control for w_{it}. This other variable is called a proxy. Specifically, consider the case of material input, which is the choice of proxy in our analysis. Following ACF, the logarithm of material input, m_{it}, is assumed to be a strictly increasing function of w_{it}, given k_{it} and l_{it}:

$$m_{it} = f(w_{it}, k_{it}, l_{it})$$

Due to strict monotonicity, this relationship can be inverted to express w_{it} as a function of m_{it}, l_{it} and k_{it}:

$$w_{it} = f^{-1}(m_{it}, k_{it}, l_{it})$$

Plugging this proxy equation back into production function, one can estimate $\widehat{y_{it}} = \beta_l l_{it} + \beta_k k_{it} + w_{it}$, as a polynomial function of m_{it}, l_{it} and k_{it}:

$$\widehat{y_{it}} = \beta_l l_{it} + \beta_k k_{it} + w_{it} = \beta_l l_{it} + \beta_k k_{it} + f^{-1}(m_{it}, k_{it}, l_{it}) = y(m_{it}, k_{it}, l_{it}) \quad (1)$$

In LP, labor input is assumed to be a static decision, and l_{it} does not enter the proxy equation. Therefore in LP, β_l is already identified at this stage, but not β_k. In ACF, neither can be identified at this stage, as they are not separated from the correlations between the inputs and w_{it} via the proxy equation.

To further identify β_l and β_k, ACF assumes that w_{it} follows a Markov process. This is where Abito (2020) is different from previous methods. Here we follow Abito (2020) to assume w_{it} has a firm fixed effect component:

$$w_{it} = r_{it} + a_i$$

where r_{it} is assumed to follow a Markov process:

$$\mathrm{r}_{it} = E[\mathrm{r}_{it}|\mathcal{I}_{it-1}] + \xi_{it} = g(\mathrm{r}_{it-1}) + \xi_{it}$$

where $\mathcal{I}_{it-1}$ is the information set of firm i up to time $t-1$. Plugging the above two equations into $\widehat{y_{it}} = \beta_l l_{it} + \beta_k k_{it} + w_{it}$, we get

$$\widehat{y_{it}} = \beta_l l_{it} + \beta_k k_{it} + w_{it} = \beta_l l_{it} + \beta_k k_{it} + g(\mathrm{r}_{it-1}) + \xi_{it} + a_i$$

Assuming $g(\mathrm{r}_{it-1}) = \gamma_0 + \gamma_1 \mathrm{r}_{it-1} + \gamma_2\, r_{it-1}^2 + \cdots + \gamma_n r_{it-1}^n$, we have

$$\widehat{y_{it}} = \beta_l l_{it} + \beta_k k_{it} + \gamma_0 + \gamma_1 \mathrm{r}_{it-1} + \gamma_2\, r_{it-1}^2 + \cdots + \gamma_n r_{it-1}^n + \xi_{it} + a_i \quad (2)$$

Here we have two issues. First, l_{it} and k_{it} are potentially correlated with a_i. Second, we do not observe r_{it}. Instead, given a guess of β_l and β_k, we can get an estimate of w_{it} by subtracting $\beta_l l_{it} + \beta_k k_{it}$ from $\widehat{y_{it}}$. However, using w_{it} in place for r_{it} creates endogeneity problem for the γs due to measurement error of the variable r_{it}, with the measurement error being the a_i term.

Abito (2020) proposes a set of instruments to be used in a procedure similar to the two-stage-least-squares that can solve both issues together. Specifically, Abito (2020) proposes using changes in unobserved productivity

$$\Delta \mathrm{w}_{it-j} = \Delta h(x_{it-j}) = h(x_{it-j}) - h(x_{it-j-1})$$

for j ≥1 as instruments in a two-stage estimation procedure, where

$$h(x_{it-j}) = \widehat{y_{i-j}} - \beta_l l_{i-j} - \beta_k k_{i-j} = \mathrm{w}_{it-j}.$$

and x_{it} stands for the collection of variables l_{it}, k_{it}, m_{it}. Two additional assumptions are necessary for the instruments to work:

$$E\left[\xi_{it}\middle|k_{it-j}, l_{it-1-j}\right] = 0 \text{ for all j} \geq 0$$

$$E[a_i|\xi_{it}] = 0 \text{ for all t}$$

The first assumption is conventional in the production function estimation literature and comes from the timing of input decisions: k_{it-j}, l_{it-1-j} for all j ≥0 are decided before the innovation ξ_{it} is realized. Specifically, ACF also makes this assumption. In cases where there is no firm fixed effect component in productivity, this assumption enables one to use past inputs as instruments to identify the input elasticities.

Denote the set of instruments as z. Crucially, by the Markov assumption and the assumption that $E[a_i|\xi_{it}] = 0$ for all t, $E[a_i|z] = 0$, thus

$$E(r_{it-1}|z) = E(w_{it-1} - a_i|z) = E(w_{it-1}|z)$$

In the first stage of the estimation, we regress l_{it}, k_{it}, and r_{it-1} on polynomials of z:

$$l_{it} = E(l_{it}|z) + \eta_l = \widehat{l_{it}} + \eta_{lit}$$

$$k_{it} = E(k_{it}|z) + \eta_k = \widehat{k_{it}} + \eta_{kit}$$

$$r_{it-1} = E(r_{it-1}|z) + v_{it-1} = E(w_{it-1}|z) + v_{it-1} = E(h(x_{it-1})|z) + v_{it} = \widehat{r_{it-1}} + v_{it-1}$$

Note that although the last regression is infeasible due to r_{it-1} being unobserved, we can still regress w_{it-1} on z to get $\widehat{r_{it-1}}$, which is used in the second stage of the estimation. Substituting the above three equations into (2), we have the second stage equation:

$$\widehat{y_{it}} = \beta_l \widehat{l_{it}} + \beta_k \widehat{k_{it}} + \sum_{j=0}^{n} \alpha_j \widehat{r_{it-1}^j} + e_{it}$$

where

$$\alpha_j \equiv \left(\sum_{q=j}^{n} \binom{q}{j} \gamma_q E\left(v_{it-1}^{q-j}\right) \right)$$

$$e_{it} \equiv \sum_{j=0}^{n} \widehat{r_{it-1}^j} \left(\sum_{q=j}^{n} \binom{q}{j} \gamma_q \left[v_{it-1}^{q-j} - E\left(v_{it-1}^{q-j}\right) \right] \right) + (\beta_l \eta_{lit} + \beta_k \eta_{kit} + \xi_{it} + a_i)$$

As long as the instrument vector z satisfies $E[e_{it}|z] = 0$, then

$$E\left[e_{it}|\widehat{l_{it}}, \widehat{k_{it}}, \widehat{r_{it-1}}\right] = 0$$

and we can estimate β_l, β_k and $\alpha_j's$ using ordinary least squares. The exogeneity requirement $E[e_{it}|z] = 0$ is satisfied under the assumptions made above, plus a strong instrument assumption requiring the instruments z to be independent of v_{it-1}. To verify this, first note that $E\left[\eta_{ijt}|z\right] = E[\eta_{ikt}|z] = 0$ are satisfied by construction, $E[a_i|z] = 0$ is satisfied due to Markov assumption and the assumption that $E[a_i|\xi_{it}] = 0$ for all t, and $E[\xi_{it}|z] = 0$ is satisfied by the Markov assumption (or the assumption on timing of input decisions). At last, $E\left(v_{it-1}^q|z\right) = E\left(v_{it-1}^q\right)$ for all $1 \leq q \leq n$ would require more than mean independence between z and v_{it-1} for any $n > 1$. A sufficient condition is z independent of v_{it-1}. Such an instrument is

called a Berkson-type instrument (Chen, Hong and Nekipelov 2011, Schennach 2007) in the context of the nonlinear error-in-variable model.

Therefore, we use the following algorithm to estimate the production function:

Algorithm (Abito 2020)

0, regress y_{it} on a polynomial of l_{it}, k_{it}, m_{it}, and store the predicted values $\widehat{y_{it}}$,

1, in iteration n, given an initial guess of $\beta_l^{(n-1)}$ and $\beta_k^{(n-1)}$, compute $h(x_{it-j}) = \widehat{y_{it-j}} - \beta_l^{(n-1)} l_{it-j} - \beta_k^{(n-1)} k_{it-j}$, and then the instruments $\Delta h(x_{it-j})$ (at least 3 lags of $\Delta h(x_{it-j})$ for exact identification).

2, estimate conditional expectations, $\widehat{l_{it}} = E(l_{it}|z)$, $\widehat{k_{it}} = E(k_{it}|z)$, and $\widehat{r_{it-1}} = E(h(x_{it-1})|z)$

3, OLS regression of $\widehat{y_{it}}$ on $\widehat{l_{it}}, \widehat{k_{it}}$ and powers of $\widehat{r_{it-1}}$ (we set n=3). This gives $\beta_l^{(n)}$ and $\beta_k^{(n)}$. Evaluate convergence and repeat until $\beta_l^{(n)} = \beta_l^{(n-1)}$ and $\beta_k^{(n)} = \beta_k^{(n-1)}$.

After we estimate the production function, we examine the effect of temperature on productivity with the following empirical specification for firm i in postal code c at time t:

$$TFP_{ict} = \sum_m \beta_m T_{ct}^m + \boldsymbol{\delta}' \boldsymbol{W_{ct}} + \boldsymbol{\theta}' \boldsymbol{Z_{ict}} + \epsilon_{ict}$$

where TFP_{ict} is total factor productivity (i.e., $TFP_{ict} = w_{ict} + \epsilon_{ict}$), T_{ct}^m is the number of days the daily temperature falls into the mth bin for temperature (which will be clear in the next section), $\boldsymbol{W_{ct}}$ are weather controls including variables related to precipitation, relative humidity, wind speed, and visibility, and $\boldsymbol{Z_{ict}}$ is a set of semi-parametric controls, including firm fixed effects, year-county fixed effects and year-sector fixed effects.

3.3 Data

Firm data on annual manufacturing outputs and inputs come from the annual surveys conducted by the National Bureau of Statistics (NBS) in China. The data cover all state-owned industrial firms and non-state firms with annual sales over nominal CNY 5 million (USD 0.66 million) from 1998 to 2007. Industries include mining, manufacturing, and public utilities, of which manufacturing represents about 94% of the total

observations. Several well-known data issues can be dealt with following instructions in Brandt et al. (2012). Our main analysis uses the deflators from Yang (2015) to compute real value-added output and real input. In our robustness checks, we use alternative deflators from Brandt et al. (2012).

Weather data come from the National Climatic Data Center (NCDC) at the National Oceanic and Atmospheric Administration (NOAA). Zhang et al. (2018) utilize data from this source. Figure 3.9 in the appendix shows the weather stations in China where our weather data are observed. The weather data contain temperature, precipitation, dew point temperature, visibility, and wind speed. Relative humidity can be constructed using a standard meteorological formula with temperature and dew point temperature. Also, following Zhang et al. (2018), we obtain predictions for future weather conditions from the HadCM3 model data. In our main analysis, we assume firms stay in their first observed postal code in the future. In our robustness checks, we assume a firm could be at one of the multiple locations, with the possibility of at one location the fraction of time it was observed staying in that location during years 1998-2007.

The final step in data preparation is to combine firm and weather data. Weather data are daily from individual weather stations, while firm production data are on an annual basis. For each firm, the weather in the postal code the firm resides in will be used. Therefore, we first interpolate each postal code's daily weather using the data observed in nearby weather stations. After this is done for each postal code, we will transform daily weather data into annual variables. For temperature, this is done by counting the number of days a weather variable falls into a certain bin within a year. A simple annual mean is computed for other weather variables, except for precipitation, for which an annual sum is computed (these go into the controls $\boldsymbol{W_{ct}}$ in the TFP regression). For a detailed description of our data quality check and cleaning process, please check the Appendix.

3.4 Summary Statistics

Table 3.1 reports the summary statistics for the final cleaned dataset, with total factor productivity computed using OP, ACF, and Abito (2020). This is a set of unbalanced firm panel data, with annual observations for 509,671 firms. The data contain 38 2-digit industries. We can see that, on average, a firm in our sample is relatively large, with about 212 employees. The three production function estimation methods produce quite

different mean levels of productivity (total factor productivity, TFP) estimates. OP yields the highest value for mean TFP, while Abito (2020) yields the lowest. For the weather variables, it is clear that temperature will be higher in the future. Interestingly, precipitation will also be significantly higher in the future on average according to the HadCM3 model, while average relative humidity and average wind speed stay relatively the same.

Table 3.1 is directly comparable to the summary statistics table in Zhang et al. (2018). Our table has two major discrepancies compared to theirs. The first one is the values for OP-estimated productivity. We think this might be due to different data construction choices for the firm production variables, for example, which variable(s) in the raw data used for constructing value-added output and choice of deflators. The second one is the average temperature. We provide a straightforward explanation on why we think our temperature measures are valid in the Appendix.

To get a clearer idea on shifts in temperature, Figure 3.1 reports the temperature distribution within a year during 1998-2007 and during 2040-2042. There are more days in the hottest two temperature bins in 2040-2042, and less days in the mild temperature bins. This is the climate change we focus on.

Table 3.1 Summary Statistics of Cleaned Data

Variable	# of Obs	Mean	Std. Dev.	Min	Max
Labor (person)	1,807,577	211.93	303.09	11	2786
Capital stock (thousand USD)	1,807,577	2221.97	5188.83	8.62	66491.95
Investment (thousand USD)	1,523,362	385.17	3305.93	0.00	1615473.00
Log TFP (OP)	1,807,577	4.02	1.35	-4.56	10.05
Log TFP (ACF)	1,807,577	1.87	1.32	-7.42	7.90
Log TFP (Abito 2020)	1,807,577	-1.25	2.57	-18.76	8.68
Mean daily temperature, 1998-2007 (F)	1,807,577	61.79	7.23	24.28	80.45
Mean daily temperature, 2040-2042 (F)	1,807,577	62.89	9.10	24.90	85.42
Mean annual precipitation, 1998-2007 (inch)	1,807,577	33.94	26.04	0.00	119.91
Mean annual precipitation, 2040-2042 (inch)	1,807,577	47.31	19.34	3.81	94.27
Mean daily relative humidity, 1998-2007 (%)	1,807,577	69.41	7.39	24.46	85.24
Mean daily relative humidity, 2040-2042 (%)	1,807,577	69.18	6.74	30.48	81.03
Mean daily wind speed, 1998-2007 (mile/h)	1,807,577	5.23	1.30	0.59	13.29
Mean daily wind speed, 2040-2042 (mile/h)	1,807,577	5.44	1.67	0.91	12.73
Mean daily visibility, 1998-2007 (mile)	1,807,577	7.75	1.90	2.90	18.60

Note: All monetary values are in 2007 USD. This is an unbalanced panel for 509,671 firms with their production and weather information from 1998 to 2007. An observation is a firm-year. Future daily weather predictions in 2040-2042 are first aggregated to a single average for each variable and each postal code, and then merged with historical data by postal code. Firms are assumed to remain in their first observed historical location in the future. The raw data for firm production, which originate from the annual survey by the Chinese National Bureau of Statistics, cover all industrial firms that are identified as "either state-owned, or are non-state firms with sales above 5 million RMB". The industries covered include mining, manufacturing and public utilities. The raw data for historical weather come from National Centers for Environmental Information Global Surface Summary of Day data, which is aggregated from hourly raw observations by the National Centers for Environmental Information (NCEI) at the National Oceanic and Atmospheric Administration (NOAA). The raw data for future (2040-2042) daily weather come from the Centre for Environmental Data Analysis in UK (HadCM3 A1FI run, stored as part of Met Office data from the Climate Impacts Link Project). The industry-level input and output deflators for monetary values come from Yang (2015). All weather data are linked to firm data at postal code level. Postal codes are geocoded using Google geocoding API. HadCM3 data do not provide predictions for visibility. Historical values for relative humidity are computed using formula in Zhang et al. (2018). Investment is only observable from 1998-2006. Of the 1,523,362 investment observations, 468,250, or 30.74% are zero, which will become missing once logged in using OP for production function estimation.

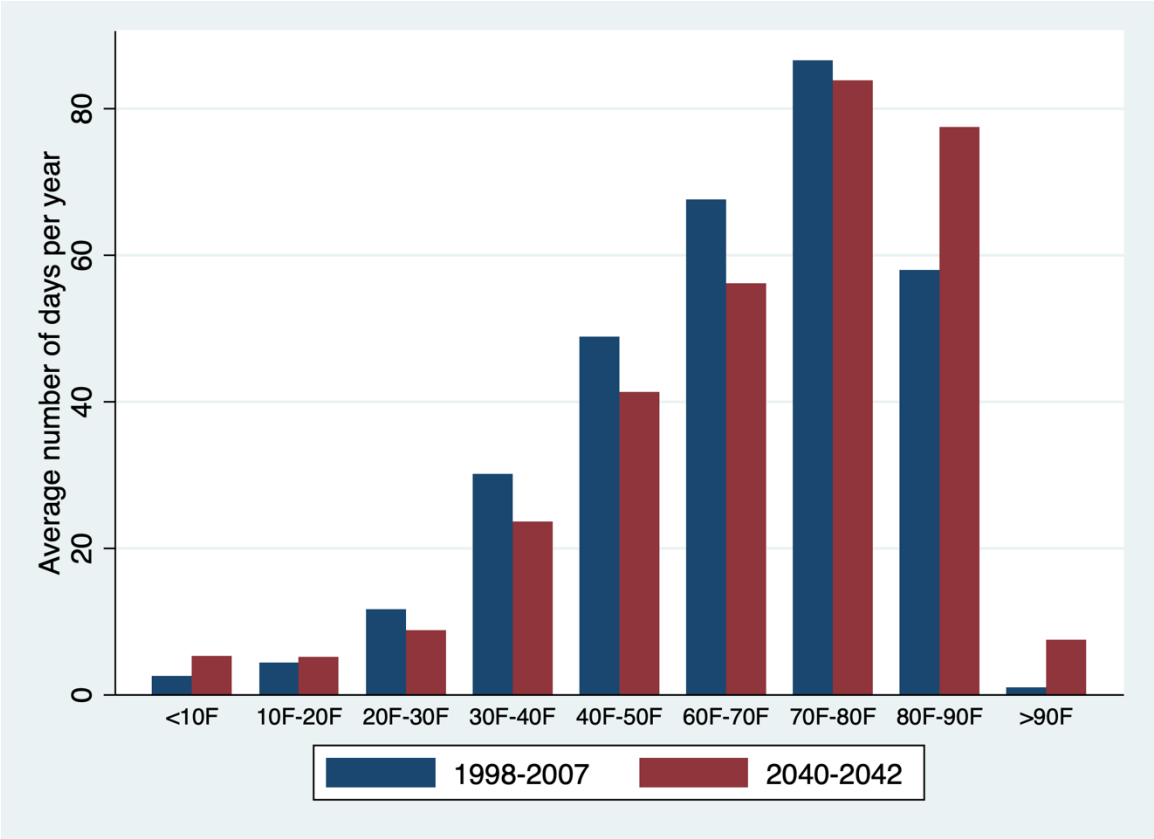

Figure 3.1: Temperature distributions

Note: Temperature distribution within a year experienced by an average firm in the sample, 1998-2007 v. s. 2040-2042. Firms are assumed to remain in their first observed historical location in the future.

3.5 Productivity dynamics

The ACF and Abito (2020) production function estimation methods differ on assumptions of productivity dynamics. ACF assumes the unobserved productivity follows a Markov process. In contrast, Abito (2020) assumes firm-specific fixed effects in the productivity process, and the remaining part of productivity again follows a Markov process.

To learn about the size of productivity heterogeneity and examine whether there are persistent differences in productivity, in this section we first analyze productivity dynamics by using only the balanced sub-panel of firms (i.e., firms with 10 years of observations). We start by grouping them by their productivity ranking in 1998 (first year in the sample). This is done separately for each industry, using production function estimates by Abito (2020). The four groups are simply the top 25%, 50-75%, 25-50%, and the bottom 25%. We compute the median TFP of each group for each year. If TFP differences are persistent for all firms, we expect the medians to stay parallel over the years. If not, the medians will converge.

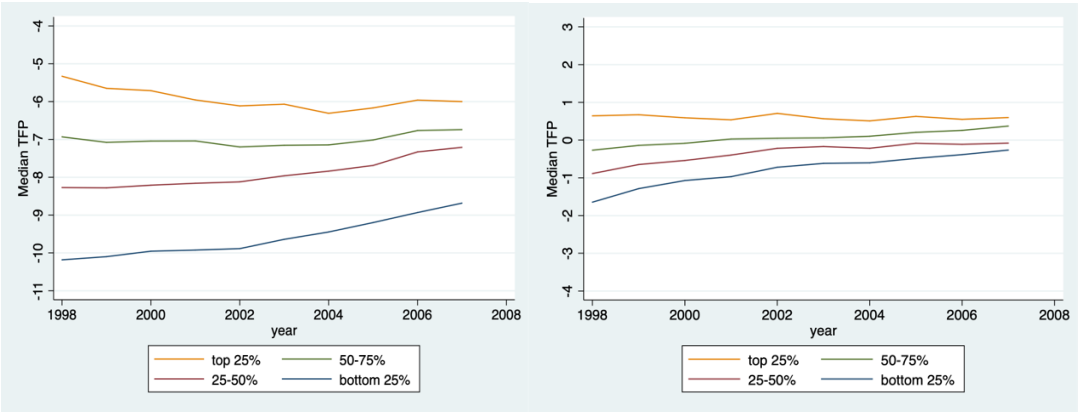

Figure 3.2: Examples of industries with persistent productivity differences

Note: Left: industry 35 (general machinery), right: industry 34 (metal). Industry numbers are their 2-digit code in the raw data according to post-2003 China industry indexing standard (see Table 3.9 for details).

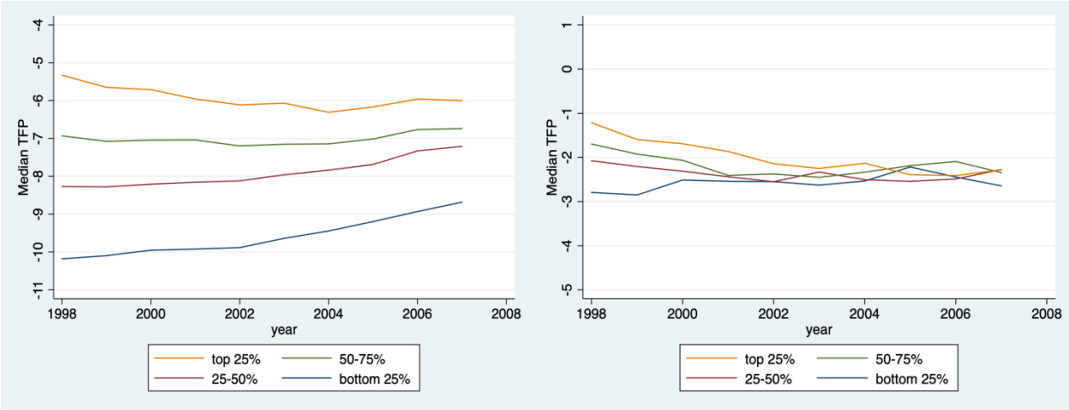

Figure 3.3: Industries with different productivity persistency

Note: left: industry 35 (general machinery), right: industry 46 (water). Industry numbers are their 2-digit code in the raw data according to post-2003 China industry indexing standard (see Table 3.9 for details).

We find persistent productivity differences across firms in about 24 of the 36 industries in the balanced subsample. Across these industries, the size of productivity heterogeneity differs significantly. Figure 3.2 gives two examples of industries where there are persistent productivity differences. On the left of Figure 3.2 is the industry with the largest productivity heterogeneity (general machinery); on the right of Figure 3.2 is an industry with smaller yet still persistent productivity heterogeneity (metal). Among the 24 industries, 4 industries have the largest persistent productivity heterogeneity similar to that of the general machinery industry (i.e., the initial distance between the median TFPs of the top versus the bottom 25% is about 4-6), and about 20 industries have more moderate yet persistent heterogeneity like the metal industry. On the other hand, there are about 5 industries where the productivity differences are clearly not persistent. The right panel of Figure 3.3 shows such an industry (water supply). The initial productivity differences in the first year are also relatively small in these industries.

Is the difference between Abito (2020) and ACF's elasticity estimates larger in industries with more persistent and/or larger productivity differences? Table 3.2 reports the elasticity estimates of ACF and Abito (2020) for each industry in our full sample (i.e., not the balanced subsample), together with the distances between their productivity estimates. The distance is defined as the sum of the squared difference between the labor input

Table 3.2: Production Function Elasticity Estimates

Industry	ACF				Abito (2020)				Difference	
	Labor	S.D.	Capital	S.D.	Labor	S.D.	Capital	S.D.	Distance	S.D.
35	0.87	(0.04)	0.17	(0.02)	0.65	(0.04)	1.52	(0.04)	1.85	(0.05)
36	0.36	(0.04)	0.34	(0.02)	0.59	(0.06)	1.61	(0.07)	1.67	(0.09)
10	0.48	(0.07)	0.30	(0.04)	0.24	(0.10)	1.47	(0.14)	1.42	(0.18)
40	0.56	(0.04)	0.28	(0.02)	0.04	(0.06)	1.25	(0.06)	1.21	(0.05)
33	0.92	(0.06)	0.06	(0.04)	0.17	(0.08)	0.79	(0.07)	1.09	(0.09)
26	0.43	(0.02)	0.40	(0.02)	0.00	(0.04)	1.25	(0.04)	0.91	(0.04)
30	0.77	(0.04)	0.23	(0.02)	0.10	(0.04)	0.89	(0.05)	0.90	(0.05)
27	0.60	(0.04)	0.38	(0.03)	0.00	(0.08)	1.06	(0.09)	0.83	(0.07)
37	0.49	(0.05)	0.34	(0.03)	0.00	(0.06)	1.08	(0.06)	0.78	(0.06)
20	0.98	(0.07)	0.08	(0.03)	0.42	(0.06)	0.74	(0.07)	0.75	(0.12)
13	0.88	(0.04)	0.09	(0.02)	0.26	(0.04)	0.69	(0.03)	0.75	(0.04)
6	0.48	(0.05)	0.31	(0.03)	0.16	(0.05)	1.05	(0.07)	0.65	(0.07)
44	0.73	(0.05)	0.49	(0.03)	0.01	(0.08)	0.83	(0.12)	0.64	(0.12)
31	0.56	(0.03)	0.30	(0.02)	0.37	(0.04)	1.05	(0.04)	0.60	(0.04)
22	0.50	(0.05)	0.36	(0.03)	0.07	(0.05)	1.00	(0.07)	0.59	(0.07)
34	0.49	(0.03)	0.33	(0.02)	0.09	(0.05)	0.97	(0.05)	0.56	(0.06)
41	0.51	(0.06)	0.29	(0.04)	0.17	(0.10)	0.91	(0.11)	0.50	(0.15)
19	0.70	(0.05)	0.23	(0.03)	0.32	(0.07)	0.78	(0.07)	0.44	(0.07)
32	0.50	(0.06)	0.38	(0.04)	0.16	(0.07)	0.87	(0.07)	0.36	(0.06)
23	0.40	(0.07)	0.43	(0.04)	0.33	(0.09)	1.01	(0.11)	0.34	(0.19)
15	0.46	(0.05)	0.35	(0.04)	0.20	(0.10)	0.88	(0.08)	0.34	(0.12)
14	0.24	(0.08)	0.34	(0.03)	0.78	(0.07)	0.49	(0.08)	0.32	(0.10)
45	0.79	(0.24)	0.39	(0.11)	0.32	(0.24)	0.09	(0.28)	0.31	(0.24)
18	0.95	(0.05)	0.08	(0.02)	0.66	(0.05)	0.53	(0.05)	0.28	(0.08)
39	0.46	(0.03)	0.39	(0.02)	0.25	(0.04)	0.87	(0.05)	0.27	(0.04)
25	0.12	(0.12)	0.22	(0.08)	0.39	(0.25)	0.63	(0.34)	0.25	(0.30)
29	0.49	(0.07)	0.35	(0.05)	0.52	(0.11)	0.81	(0.11)	0.22	(0.14)
21	0.49	(0.10)	0.11	(0.04)	0.71	(0.08)	0.52	(0.09)	0.21	(0.11)
16	-0.16	(0.48)	0.63	(0.17)	0.28	(0.32)	0.59	(0.32)	0.20	(0.53)
17	0.42	(0.02)	0.34	(0.02)	0.31	(0.04)	0.76	(0.04)	0.18	(0.03)
42	0.32	(0.06)	0.32	(0.03)	0.41	(0.06)	0.72	(0.07)	0.17	(0.06)
24	0.60	(0.07)	0.16	(0.04)	0.67	(0.10)	0.54	(0.10)	0.15	(0.09)
9	0.46	(0.10)	0.29	(0.06)	0.53	(0.14)	0.61	(0.13)	0.10	(0.14)
8	0.35	(0.11)	0.45	(0.07)	0.40	(0.13)	0.62	(0.10)	0.03	(0.13)
28	0.35	(0.10)	0.46	(0.09)	0.50	(0.15)	0.55	(0.15)	0.03	(0.15)
46	0.50	(0.05)	0.55	(0.02)	0.60	(0.09)	0.62	(0.10)	0.01	(0.06)

Note: Industries are ordered by distance in elasticity estimates between ACF and Abito (2020), from largest to smallest. The distance is defined as the sum of the squared difference between labor input elasticities and the squared difference between capital input elasticities. Bootstrap standard errors reported in parentheses. For meanings of industry code, please refer to Table 3.9 (industry codes post-2003). There are not enough observations for industries 7 and 43 (not reported here).

elasticities and the squared difference between capital input elasticities. Industries are ordered by distance in elasticity estimates between ACF and Abito (2020), from largest to smallest. Estimations are carried out by an R program, with bootstrapped standard errors.[10]

Among the 36 industries reported in Table 3.2 (there are two industries not having enough observations, which are not reported in this table), The difference between Abito (2020) and ACF's elasticity estimates is statistically different in 24 industries (at 99.7% confidence level). The differences are especially large in the first four industries reported in Table 3.2, and they are indeed the four industries with the largest persistent productivity differences as illustrated in the left panel of Figure 3.2. These four industries are general machinery (number 35), special machinery (number 36), nonmetal mining (number 10) and communication equipment (number 40). Industries 35, 36 and 40 are among the largest manufacturing sectors and their combined production amounts to about 17.27% of the total real value-added output in our sample. Also, given the nature of these 4 industries, it is not surprising to find large capital input elasticity estimates. On the other hand, the industries with the smallest differences between ACF and Abito (2020)'s elasticity estimates coincide with those without persistent productivity differences or those with smallest productivity differences. To be specific, industries 46, 28, 8, 16, and 25 have productivity dynamics illustrated in the right panel of Figure 3.3; in industries 9, 24, 42, 17, and 21, productivity differences do not shrink over time as much as that shown in the right panel of Figure 3.3, but the difference is small and quickly becomes smaller (the initial differences between the median TFPs of top and bottom 25% are usually about 2 in these industries). At last, industries in the middle of Table 3.2 are those with moderate and persistent productivity differences, as illustrated on the right panel of Figure 3.2.

To sum, in this section, we find that the difference between Abito (2020) and ACF's elasticity estimates is larger in industries with more persistent and larger productivity differences. This is consistent with our expectations, as the main difference between Abito (2020) and ACF is that Abito (2020) allows for firm-specific fixed effects. Thus one would expect ACF to produce biased elasticity estimates in cases where the size of such fixed effects varies significantly across firms. In our sample, most of the industries have

[10] For elasticity estimates using OP, please check Table 3.10 in the appendix.

persistent productivity differences, and in some of the largest industries, such differences are large, leading to significantly biased elasticity estimates by ACF. The evidence supports our decision to use Abito (2020) as our production function estimation method.

3.6 Results

As a reminder, we examine the effect of temperature on productivity with the following empirical specification for firm i in postal code c at time t:

$$TFP_{ict} = \sum_{m} \beta_m T_{ct}^{m} + \boldsymbol{\delta}' \boldsymbol{W_{ct}} + \boldsymbol{\theta}' \boldsymbol{Z_{ict}} + \epsilon_{ict}$$

where TFP_{ict} is the total factor productivity (i.e., $TFP_{ict} = w_{ict} + \epsilon_{ict}$), T_{ct}^{m} is the number of days the daily temperature falls into the mth bin for temperature, $\boldsymbol{W_{ct}}$ are weather controls including variables related to precipitation, relative humidity, wind speed, and visibility, and $\boldsymbol{Z_{ict}}$ is a set of semi-parametric controls, including firm fixed effects, year-county fixed effects and year-sector fixed effects.

3.6.1 Main results

We first run the above regression with our full sample using Abito (2020)'s TFP.[11] Figure 3.4 shows the estimated coefficients of the temperature bin variables, with the bin 10-C16C as the reference group. We find that having more days with hot or cold temperatures generally has adverse effects on a firm's TFP, compared to having more days in the milder temperature range. However, the effect of the hottest days is not statistically significant and not significantly larger than the effect of the number of days in the 27C-32C bin. This is different from what is found in Zhang et al. (2018), where the number of hottest days in a year has a much larger adverse effect on a firm's TFP.

[11] Industry 7 and 43 do not have enough observations for production function estimations. We use industry 6 and industry 44's elasticity estimates for these two industries.

We then use our regression results to predict the effect of climate change on an average firm's TFP in our sample by 2040-2042. We also repeat the analysis using ACF's TFP. With Abito (2020)'s TFP, it is predicted that the climate change would lead to a 4.07% decrease in productivity on average across firms, with 95% confidence interval [-7.01, -1.13]. With ACF's TFP, the prediction is a 4.65% decrease in productivity on average, with 95% confidence interval [-7.20, -2.10].

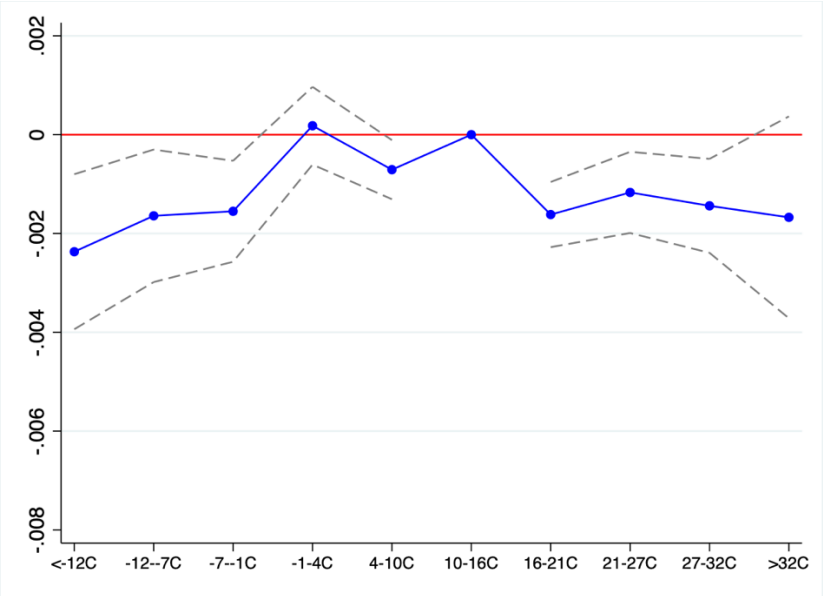

Figure 3.4: Effect of temperature on TFP

Note: X-axis: temperature bins. Y-axis: coefficient in front of the variable for a temperature bin. Data constructed with Yang (2015) deflators. Bins are 95% confidence intervals. Standard errors are clustered by firm and year-county.

We compute the predictions for each industry with Abito (2020)'s TFP, reported in Figure 3.5. We run separate regressions for each industry and use these results to make predictions. Among the 33 industries reported here (these are the ones reported in Zhang et al. 2018), the average firm's TFP in 9 industries is predicted to be significantly affected by climate change. We repeat the industry-specific analysis with ACF's TFP and compare the predictions with those in Figure 3.5. We find that the predictions with Abito (2020) are significantly different (i.e. the point predictions are outside or on the edge of ACF's 95% CI) in 3 industries, including communication equipment, general machinery and special machinery. For example, in the special

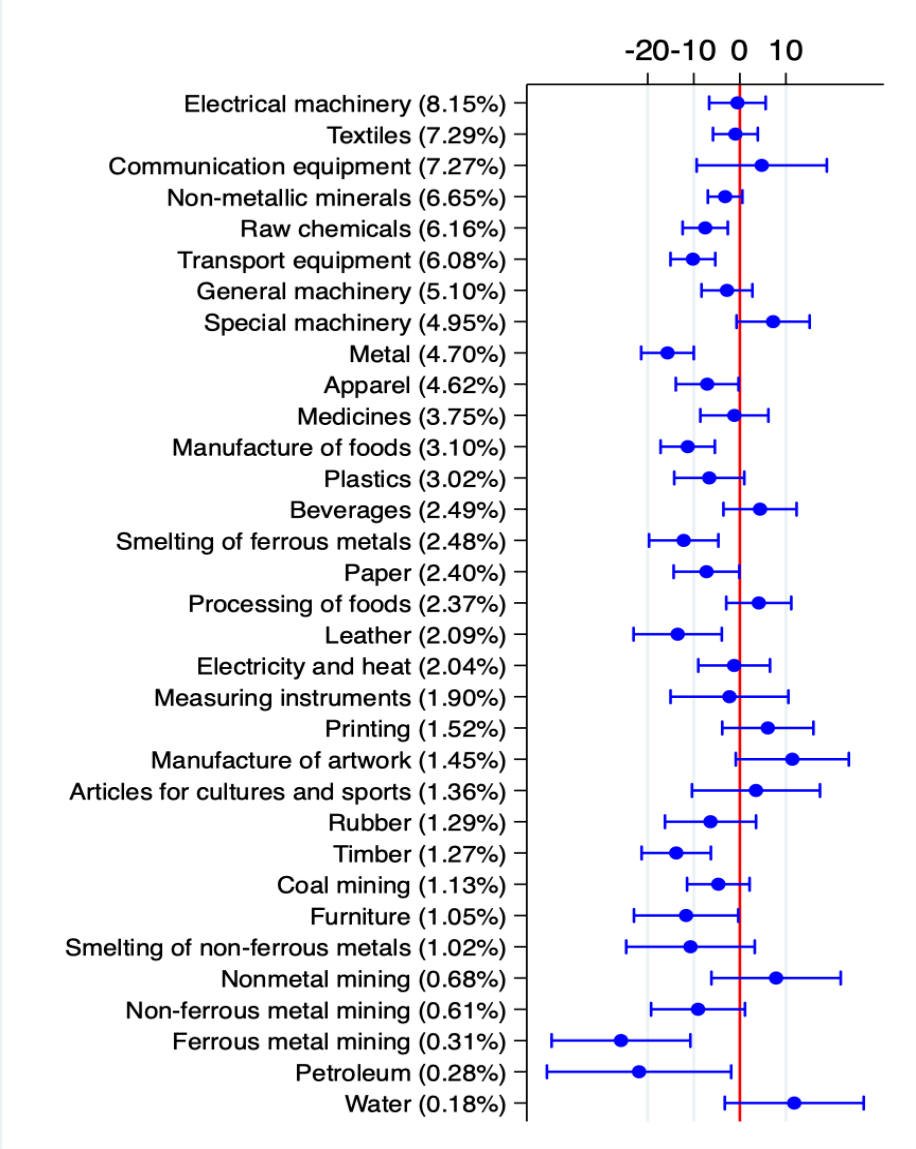

Figure 3.5: Predicted effect of climate change on TFP by industry

Note: X-axis: industries. Y-axis: predicted impact of climate change in percentage. Data constructed with Yang (2015) deflators. Bins are 95% confidence intervals. The numbers in parentheses are the shares in real value-added output.

machinery industry, Abito (2020) predicts a productivity decrease of about +9%, while ACF predicts zero effect. These are the 3 of the 4 industries with the largest differences in elasticity estimates in Table 3.2, and are the industries having the largest persistent productivity differences. In addition, a number of industries have somewhat different (but small compared to CI) predictions between ACF and Abito (2020), including printing, nonmetal mining, nonmetallic minerals, raw chemicals, measuring instruments, manufacturing of artwork, rubber, timber, and furniture.

We expect the differences in climate change prediction to be larger in industries with larger differences between ACF and Abito (2020)'s elasticity estimates. This is the case, as shown in Figure 3.6. In Figure 3.6, the horizontal axis is the squared distance between elasticities (the same distances in Table 3.2). The vertical axis is the absolute difference between ACF and Abito (2020)'s predictions in percentage points. In Figure 3.6, the 4 industries on the top right are the 4 industries with the largest differences in elasticity estimates: general machinery (number 35), special machinery (number 36), nonmetal mining (number 10) and communication equipment (number 40). For the industries with the smallest differences in elasticities, the differences in predictions are also the smallest. This suggests that biases in elasticity estimates will translate into biases in making predictions on the effect of climate change on TFP. Given that the industries having the most biased predictions (general machinery, special machinery, communication equipment, etc.) have large shares in the output, it is important to allow for persistent differences in firm-specific productivity in the production function estimation.

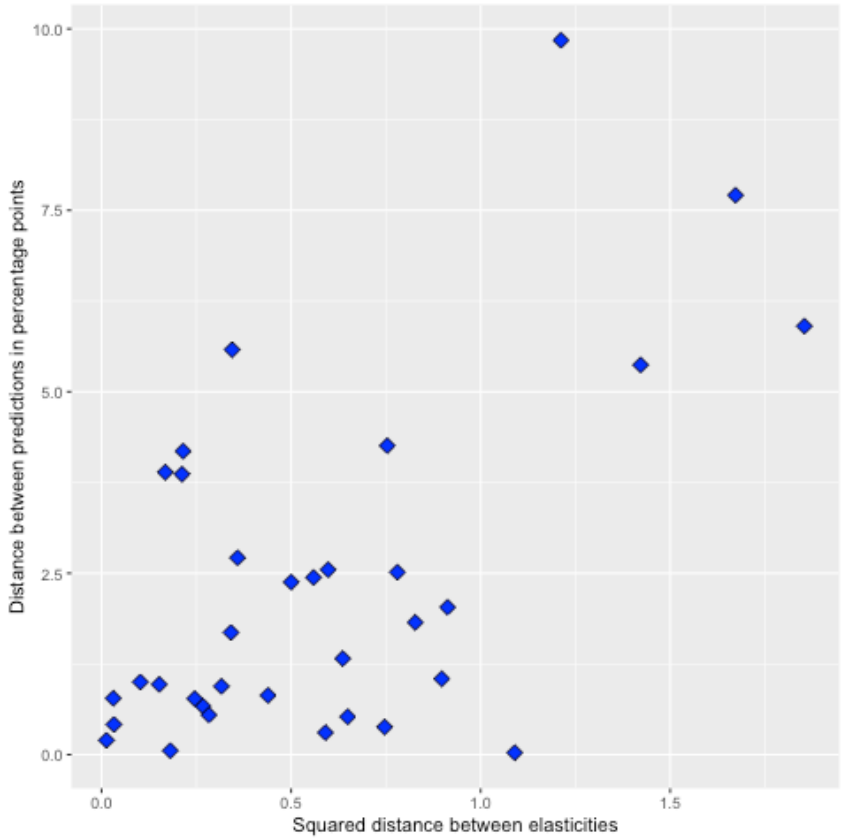

Figure 3.6: Difference in elasticity estimates versus difference in prediction

Note: each point is an industry. The horizontal axis is the squared distance between elasticities (the same distances in Table 3.2, defined as the the sum of the squared difference between labor input elasticities and the squared difference between capital input elasticities), and the vertical axis is the absolute difference between ACF and Abito (2020)'s predictions in percentage points.

3.6.2 Robustness checks

3.6.2.1 Alternative deflators

In this section, we conduct production function estimation, regressions and predictions using the data constructed with industry-level input and output deflators from Brandt et al. (2012). We think the Yang (2015) deflators are better than Brandt et al. (2012) deflators (see Appendix for details) and use the former in our main analysis above. Since Zhang et al. (2018) seem to be using the Brandt et al. (2012) deflators, it

is worth checking if the findings on the effect of climate change in our main analysis would change under these deflators.

Figure 3.7 shows the estimated regression coefficients for temperature variables with Abito (2020)-estimated productivity. The estimated coefficients are very close to those reported in Figure 3.4. Abito (2020) predicted that climate change would lead to a 3.65% decrease in productivity on average, with 95% confidence interval [−6.31,−0.99]. These results show that overall the effect of climate change on TFP is negative, consistent with the conclusion in our main analysis.

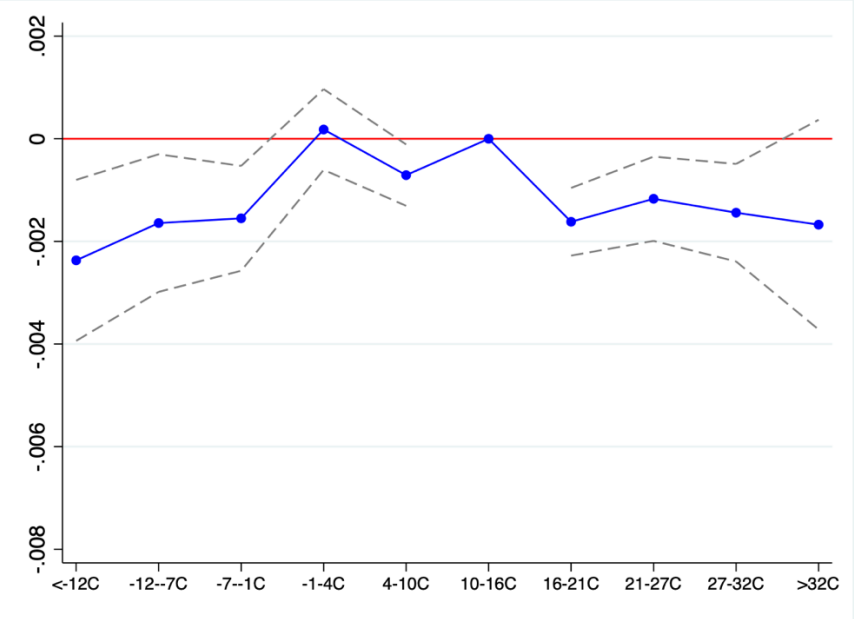

Figure 3.7: Effect of temperature on TFP: robustness check

Note: X-axis: temperature bins. Y-axis: coefficient in front of the variable for a temperature bin. constructed with Brandt et al. (2012) deflators. Bins are 95% confidence intervals.

We again compute the predictions for each industry using Abito (2020)'s TFP, reported in Figure 3.8. Here we find 12 industries where the adverse effects of climate change on TFP are statistically significant at the 95% level. The point predictions are mostly very close to those in Figure 3.5 (i.e., differences are minimal compared to the standard errors), except for 2 industries, electronic machinery (+6% with Brandt et al. 2012 deflators, but 0% with Yang 2015 deflators) and special machinery (+1% with Brandt et al. 2012 deflators, but +9% with Yang 2015 deflators). These two industries happen to be large in terms of their shares in real value-added output. Thus the researchers do need to be careful in choosing the proper deflators.

3.6.2.2 Alternative assumption on future firm locations

In our main analysis, firms are assumed to remain in their first observed historical location in the future. Now we assume a firm could be at one of the multiple locations, with the possibility of at one location the fraction of time it was observed staying in that location during years 1998-2007. With this alternative assumption, under Yang (2015) deflators and Abito (2020) productivity, we find the overall negative effect of climate change at −2.73%, with 95% CI [−5.26%, -0.20%]. This is much smaller than the previous value −4.07%. This indicates firms are, on average, moving to regions that have preferable climate conditions to their productivity.

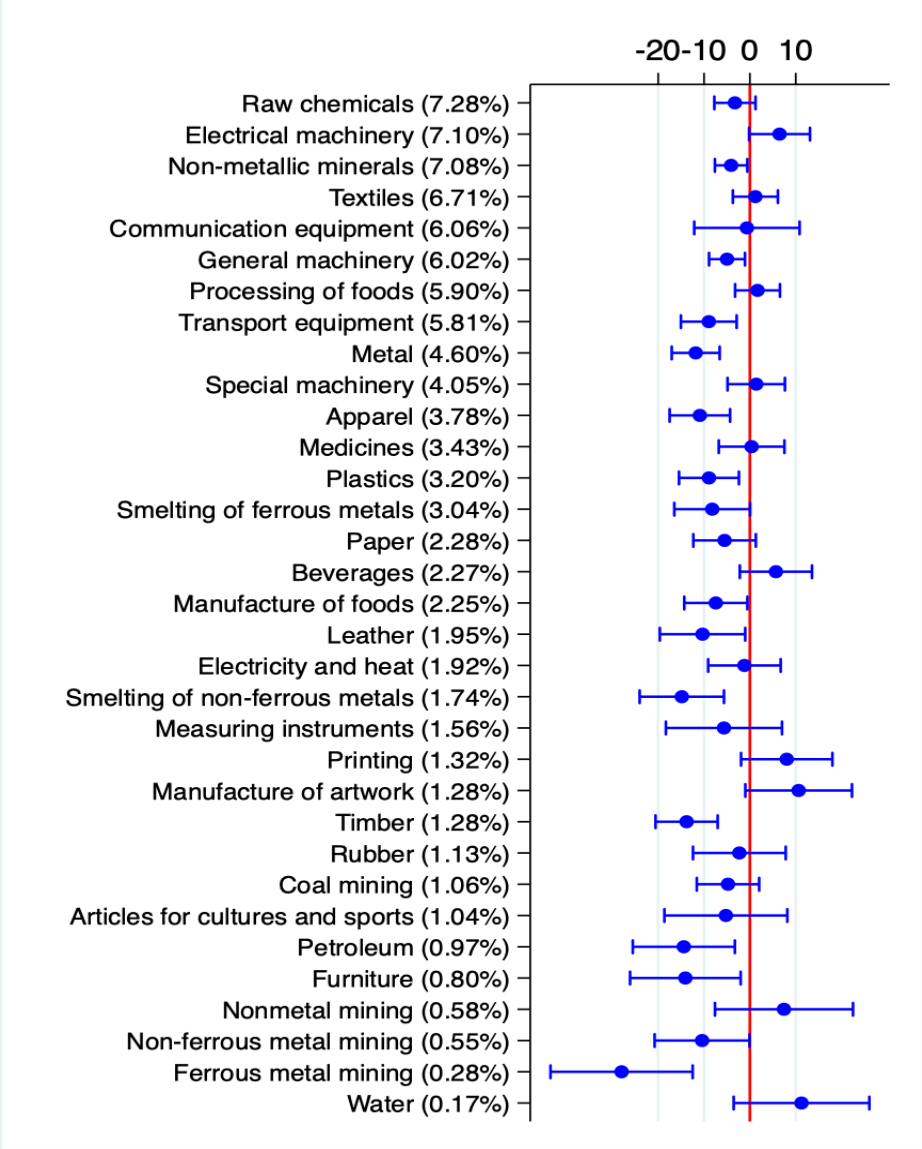

Figure 3.8: Predicted effect of climate change on TFP by industry: robustness check

Note: X-axis: industries. Y-axis: predicted impact of climate change in percentage. Data constructed with Brandt et al. (2012) deflators. Bins are 95% confidence intervals. The numbers in parentheses are the shares in real value-added output.

3.7 Conclusions

This paper estimates how weather affects the productivity of Chinese manufacturing firms surveyed by the National Bureau of Statistics during 1998-2007 and predicts how climate change would affect their productivity by 2040-2042. A notable previous study on this topic, Zhang et al. (2018), assumes no firm fixed effects in productivity when estimating the production functions. This is because traditional production function estimation methods do not allow for such fixed effects. We show that many industries in our sample exhibit considerable and persistent differences in firm-specific productivity. In industries with larger productivity heterogeneity, ACF and Abito (2020)'s elasticity estimates deviate more from each other.

With Abito (2020)'s TFP estimates, we predict the TFP will be lowered by about −4% for an average firm in our sample by 2040-2042, compared to 1998-2007. Predictions using productivity estimates by Olley and Pakes (1996) (OP) or Ackerberg et al. (2015) (ACF) put the effect at around −4.9% or −4.7%. An earlier version of Zhang et al. (2018), which studies the same question with OP but focuses on the 2020-2040 horizon, puts the effect at −5.7%. Thus using OP- and ACF-estimated productivity overstated the average adverse effect of climate change on productivity in our sample. We also examine the effects by industry and find a number of cases where using ACF-estimated TFP induces significantly biased elasticity estimates and predictions on the effects of climate change. In particular, the industries with the most biased predictions are those with the largest and persistent differences in firm-specific productivity. The prediction bias generally increases with the bias in elasticity estimates. We thus conclude that to get accurate predictions, it is important to allow for firm-specific productivity fixed effects in production function estimations.

References

Jose Miguel Abito. Estimating production functions with fixed effects. Working paper, 2020.

Daniel A. Ackerberg, Kevin Caves, and Garth Frazer. Identification properties of recent production function estimators. Econometrica, 83(6):2411–2451, 2015.

Berkson, J., Are There Two Regressions? Journal of the American Statistical Association, 45(250):164-180, 1950.

Loren Brandt, Johannes Van Biesebroeck, and Yifan Zhang. Creative accounting or creative destruction? firm-level productivity growth in Chinese manufacturing. Journal of Development Economics, 97(2):339–351, 2012.

Chen, X., H. Hong and D. Nekipelov, Nonlinear Models of Measurement Errors. Journal of Economic Literature, 49(4): 901-937, 2011.

Mert Demirer. Production function estimation with factor-augmenting technology: An application to markups. Working paper, 2020.

James Levinsohn and Amil Petrin. Estimating production functions using inputs to control for unobservables. The Review of Economic Studies, 70(2):317–341, 2003.

Jacob Malone, Aviv Nevo, and Jonathan Williams. The tragedy of the last mile: Economic solutions to congestion in broadband networks. Working paper, 2019.

Schennach, S. M., Instrumental Variable Estimation of Nonlinear Errors-in-Variables Models. Econometrica, 75(1): 201-239, 2007.

G. Steven Olley and Ariel Pakes. The dynamics of productivity in the telecommunications equipment industry. Econometrica, 64(6):1263–1297, 1996.

Rudai Yang. Study on the total factor productivity of Chinese manufacturing enterprises. Economic Research Journal (in Chinese), 50(2):61–74, 2015.

Peng Zhang, Olivier Deschenes, Kyle Meng, and Junjie Zhang. Temperature effects on productivity and factor reallocation: Evidence from a half million Chinese manufacturing plants. Journal of Environmental Economics and Management, 88:1–17, 2018.

Appendix

Appendix A: Details on data cleaning

A.1 Firm data quality check

Our firm panel data are the underlying dataset used for China's official annual statistics report. Therefore our summary statistics are directly comparable to those in the China Statistics Yearbook (Chinese pinyin "Zhongguo Tongji Nianjian"). Table 3.5 reports summary statistics of the raw data, which covers all state-owned firms and firms with annual sales larger than 5 million RMB. The numbers are almost always either identical or extremely close to those reported in Table A.1 of Brandt et al. (2012), which used a slight variant of the same dataset. For a few cases where our numbers are not so close to theirs, our numbers are at times much closer to those reported in China Statistics Yearbook. One can check this by examining the tables following Table A.1 in the Appendix of Brandt et al. (2012).

Table 3.3: Raw firm data summary statistics

Year	# of firms	Value added	Sales	Output	Labor	Export	Net value of fixed assets	
								(Original value)
1998	165,118	1.94	6.54	6.77	56.44	1.08	4.41	(6.48)
1999	162,033	2.16	7.06	7.27	58.05	1.15	4.73	(7.18)
2000	162,883	2.54	8.37	8.57	53.68	1.46	5.18	(7.86)
2001	171,240	2.83	9.32	9.54	54.41	1.62	5.52	(8.60)
2002	181,557	3.30	10.86	11.08	55.21	2.01	5.95	(9.39)
2003	196,222	4.20	13.95	14.23	57.49	2.69	6.61	(10.55)
2004	274,763	5.68	19.70	20.09	66.56	4.04	7.92	(12.49)
2005	271,835	7.21	24.69	25.16	69.31	4.77	8.95	(14.31)
2006	301,961	9.10	31.08	31.66	73.49	6.05	10.58	(16.88)
2007	336,768	11.69	39.76	40.51	79.26	7.34	12.34	(19.87)

Note: Table replicating Table A.1, panel (a) in Brandt et al. (2012). All monetary values are nominal, in trillion RMB. Labor is in billions. Values are simple sums of all firms in each year. We also produce 2-digit industry-level summary statistics for value added output and total output (or the ratio of the two) for each year, which we find are either identical or very close to those reported in China Statistics Yearbooks published by the Chinese National Bureau of Statistics.

To further check the quality of our data, we produce industry-by-year summary statistics for the number of firms, nominal gross output, nominal value-added output, and the ratio of the former two, in Tables 3.4, 3.5, 3.6, and 3.7. We compare the numbers in these tables with those reported in the China Statistics Yearbooks and again find that they are almost always identical or extremely close. Therefore we are confident our raw dataset is of very high quality.

A.2 Firm data cleaning

We closely follow, word-by-word, both Brandt et al. (2012) and Zhang et al. (2018) in cleaning the firm data whenever possible. Zhang et al. (2018) does not clearly indicate their choice of investment deflators or industry-level input and output deflators. Nor did they document their exact way of construct the real value-added output, real investment, and real capital stock. We outline our choices and procedures for these matters below.

We first manually interpolate the 6-digit county code of 122 observations with missing or non-numeric county code, based on their postal code. We also drop observations that violate any of the following accounting rules: liquid asset larger than total asset; fixed asset larger than total asset; net value of fixed asset larger than total asset; original value of fixed asset larger than total asset; intangible asset larger than total asset; liquid debts or long-term debts larger than total debts; current depreciation larger than cumulative depreciation.

Then, for each year, we rename the variables according to those used in Brandt et al. (2012)'s Stata program for matching firms over time. For a detailed variable name crosswalk and variables for identifying invalid observations, please refer to Table 3.8. At this point, we produce 10 data files, one for each year, which are used as input in Brandt et al. (2012)'s Stata program for matching firms over time.

Table 3.4: Number of firms in each industry each year in raw data

Industry	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
6	3202	2795	2666	2602	2812	3139	5065	5787	6797	7537
7	76	75	82	90	84	112	183	174	175	184
8	576	578	598	646	696	913	1650	2087	2495	2899
9	1416	1428	1439	1309	1291	1276	1364	1529	1862	2183
10	1851	1817	1770	1748	1711	1827	2168	2242	2601	3004
11	25	19	20	20	17	13	12	14	16	24
12	634	564	498	431	383					
13	11907	11231	10676	10380	10413	11192	14023	14575	16356	18140
14	5373	4963	4691	4563	4615	4636	5498	5553	6056	6644
15	3809	3579	3409	3307	3287	3194	3441	3519	3914	4422
16	351	352	343	320	287	255	205	190	179	150
17	11287	10981	10968	12065	13248	14863	24152	22569	25345	27914
18	6772	6611	7064	8037	9061	9717	12025	11865	13072	14770
19	3313	3192	3164	3538	3932	4518	6386	6227	6859	7452
20	2481	2420	2552	2808	3033	3501	4926	5397	6374	7852
21	1470	1473	1498	1625	1767	2046	3016	3074	3603	4110
22	4766	4657	4672	5027	5285	5570	7443	7461	7892	8376
23	3862	3824	3701	3691	3806	4084	5114	4826	5029	5083
24	1786	1807	1879	2024	2327	2516	3379	3378	3633	4087
25	1052	988	993	1027	1144	1323	2011	1990	2160	2149
26	11303	11337	11430	12031	12637	13803	18651	18716	20715	22981
27	3280	3272	3301	3487	3681	4063	4681	4971	5368	5748
28	803	803	834	885	909	937	1534	1306	1402	1556
29	1785	1805	1783	1777	1822	2016	3159	3034	3353	3695
30	6016	6047	6230	6883	7665	8382	12249	12041	13504	15376
31	14496	14366	14540	14706	15305	16245	19780	20111	21936	24278
32	3260	3042	2997	3175	3333	4119	6953	6649	6999	7161
33	2406	2426	2538	2823	2942	3367	5169	5163	5863	6701
34	8135	8176	8376	9273	10039	9746	14100	13802	15573	18008
35	9289	9160	9338	10027	10767	12546	20519	19981	22905	26757
36	6644	6470	6406	6390	6546	7129	10889	10260	11615	13409
37	6782	6701	6850	7111	7470	8281	11791	11315	12586	14091
39	136	134	127	123	114	10400	16123	15366	16905	19322
40	7550	7624	7845	8675	9385	5856	9150	8868	9709	11220
41	4175	4289	4459	4824	5320	2515	3913	3723	4084	4526
42	1821	1817	1860	2018	2146	4259	5119	5131	5764	6416
43	3583	3569	3753	4185	4582	107	385	438	529	652
44	4992	4941	4825	4871	4946	4998	5392	5527	5731	5565
45	291	295	300	320	329	352	496	484	526	591
46	2362	2405	2408	2398	2420	2406	2649	2492	2476	1735

Note: Industry code is year-specific. To cross-validate with China Statistics Yearbook, refer to Table 13-2 in 2008, Table 14-2 in 2007, Table 14-4 in 2006. Previous years don't report this statistic. For industry code meanings, see Table 3.9 (post-2003 codes).

Table 3.5: Nominal total output in each industry each year in raw data

Industry	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
6	130	124	128	152	198	246	403	572	721	920
7	180	208	313	278	276	348	460	628	772	829
8	15	15	17	19	23	35	72	99	139	213
9	34	36	41	42	46	57	78	114	167	229
10	33	34	36	37	42	49	59	76	103	137
11	0.44	0.28	0.36	0.36	0.20	0.75	0.58	0.86	0.52	1.10
12	16	14	12	11	11					
13	352	352	372	410	478	615	833	1,060	1,300	1,750
14	121	126	144	163	197	229	289	378	471	607
15	157	166	175	182	200	223	243	309	390	508
16	137	139	145	169	204	224	255	284	321	378
17	438	453	515	562	637	773	1,030	1,270	1,530	1,870
18	202	204	229	260	291	343	400	497	616	760
19	119	120	135	157	180	227	276	346	415	515
20	49	56	66	74	83	99	137	183	243	352
21	30	32	37	44	52	72	115	143	188	242
22	124	133	159	180	208	253	335	416	503	633
23	54	58	62	73	83	103	120	144	171	212
24	55	56	62	68	78	97	122	148	176	210
25	233	271	443	459	478	624	892	1,200	1,510	1,790
26	463	492	575	630	722	924	1,290	1,640	2,040	2,680
27	137	150	178	204	238	289	323	425	502	636
28	83	98	124	102	112	145	195	261	321	412
29	77	78	81	89	106	131	182	220	273	346
30	150	162	190	214	249	306	419	507	638	812
31	320	339	369	403	456	565	742	920	1,170	1,560
32	388	410	473	571	649	1,000	1,670	2,150	2,540	3,370
33	163	179	218	237	260	356	591	794	1,290	1,800
34	215	222	254	285	329	386	515	656	853	1,140
35	259	269	305	351	425	571	852	1,060	1,370	1,840
36	192	198	219	235	282	383	506	609	795	1,060
37	421	466	536	647	836	1,120	1,380	1,570	2,040	2,710
39	23	26	22	23	26	792	1,120	1,390	1,820	2,400
40	363	402	483	548	614	1,580	2,230	2,700	3,310	3,920
41	489	583	755	899	1,130	164	219	278	354	431
42	69	71	87	94	109	131	164	204	253	339
43	82	86	96	109	122	5	20	29	42	68
44	362	400	461	509	589	686	1,450	1,780	2,150	2,650
45	10	13	17	19	23	27	42	52	73	99
46	27	32	33	34	38	43	51	58	72	80

Note: Industry code is year-specific. Unit is billions RMB. To cross-validate with China Statistics Yearbook, refer to Table 13-2 in 2008, Table 14-2 in 2007, Table 14-4 in 2006. Previous years don't report this statistic. For industry code meanings, see Table 3.9 (post-2003 codes).

Table 3.6: Nominal value-added total output in each industry each year in raw data

Industry	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
6	60	57	58	70	92	115	198	289	359	470
7	119	144	221	202	194	239	350	481	599	644
8	5	5	6	7	9	15	33	43	59	93
9	11	13	14	14	15	18	27	43	68	97
10	11	12	12	13	14	16	21	28	38	52
11	0.17	0.08	0.11	0.10	0.05	0.24	0.19	0.27	0.18	0.33
12	8.24	6.65	6.15	5.54	5.57					
13	68	76	84	94	111	146	195	274	349	463
14	33	34	42	45	55	67	87	117	147	186
15	54	59	62	64	71	80	91	116	144	188
16	88	89	94	109	136	157	185	206	238	292
17	102	112	127	139	157	191	253	324	396	491
18	48	51	59	69	75	92	112	142	183	226
19	27	28	32	39	46	59	72	94	117	148
20	11	13	16	19	21	27	37	51	69	103
21	8	8	9	12	14	18	29	38	50	64
22	32	36	41	48	57	68	87	115	139	174
23	18	20	20	24	28	33	40	46	56	69
24	14	14	16	18	20	25	30	38	46	55
25	53	59	79	88	100	129	170	198	231	310
26	110	122	142	160	186	246	357	439	540	734
27	43	52	63	72	83	102	117	153	181	229
28	19	25	30	22	25	30	36	49	60	81
29	20	20	22	25	29	37	49	60	72	96
30	35	39	46	54	65	76	101	127	167	214
31	91	100	113	121	137	175	228	281	366	485
32	98	108	130	153	180	282	460	578	700	900
33	33	41	51	59	63	90	139	193	320	447
34	50	54	61	71	84	97	131	169	222	300
35	70	74	84	97	115	159	231	296	380	510
36	49	52	58	64	78	101	139	168	229	306
37	108	119	132	163	218	289	341	382	492	695
39	3.58	4.91	4.58	4.31	5.55	202	279	357	461	604
40	88	100	123	138	158	347	447	571	706	787
41	112	135	182	203	251	44	57	73	97	116
42	17	18	21	24	27	35	44	57	71	92
43	22	23	26	30	33	1.07	3.78	5.99	9.47	16.20
44	188	216	233	270	317	361	471	572	691	883
45	1.41	3.66	3.47	4.61	5.31	7.53	11.90	13.40	19.10	30.60
46	12	15	15	16	17	19	24	26	32	37

Note: Industry code is year-specific. Unit is billions RMB. To cross-validate with China Statistics Yearbook, refer to Table 13-2 in 2008, Table 14-2 in 2007, Table 14-4 in 2006. Previous years don't report this statistic. For industry code meanings, see Table 3.9 (post-2003 codes).

Table 3.7: Nominal value-added output as percentage points of total output in each industry each year in raw data

Industry	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
6	46.29	45.71	45.67	45.61	46.37	46.84	49.13	50.47	49.77	51.04
7	66.05	68.99	70.57	72.62	70.27	68.65	76.16	76.56	77.55	77.69
8	35.89	36.05	37.79	37.83	38.26	41.65	45.18	43.10	42.36	43.59
9	32.79	34.87	34.48	33.80	32.50	30.99	34.46	37.49	40.53	42.52
10	33.74	34.64	34.36	33.54	33.99	33.44	35.30	37.08	36.73	37.88
11	38.97	27.51	29.71	29.25	26.22	31.59	33.46	31.46	34.85	29.78
12	51.07	48.80	50.77	49.42	49.91					
13	19.37	21.64	22.43	23.04	23.26	23.77	23.45	25.80	26.86	26.48
14	26.76	27.29	28.82	27.75	28.10	29.12	30.09	30.90	31.09	30.65
15	34.32	35.31	35.30	35.21	35.54	35.63	37.23	37.68	36.88	37.04
16	64.04	64.12	64.48	64.50	66.73	70.38	72.71	72.52	74.04	77.29
17	23.23	24.65	24.70	24.64	24.61	24.66	24.48	25.55	25.85	26.21
18	23.90	24.81	25.83	26.47	25.58	26.72	28.06	28.51	29.73	29.77
19	22.92	23.66	24.05	24.88	25.40	25.97	25.98	27.25	28.23	28.71
20	22.84	23.70	23.98	26.00	25.82	26.73	26.79	27.91	28.21	29.26
21	26.00	24.48	25.62	27.01	26.53	25.29	25.31	26.87	26.47	26.56
22	25.59	26.77	25.94	26.31	27.42	26.97	26.08	27.54	27.52	27.55
23	33.57	34.20	32.65	33.57	33.85	32.55	32.94	32.08	32.66	32.67
24	25.54	25.22	25.12	26.36	26.13	25.82	24.89	25.54	26.37	26.34
25	22.69	21.82	17.79	19.25	20.98	20.65	19.01	16.51	15.27	17.34
26	23.84	24.70	24.62	25.39	25.79	26.65	27.61	26.83	26.38	27.38
27	31.53	34.38	35.58	35.39	35.08	35.45	36.15	35.98	36.01	35.93
28	22.34	25.89	23.79	21.71	22.18	20.38	18.44	18.60	18.84	19.64
29	26.53	25.96	26.94	27.78	27.48	28.16	26.72	27.08	26.16	27.68
30	23.64	23.88	24.44	25.34	25.98	24.88	24.13	25.07	26.12	26.30
31	28.33	29.59	30.51	30.09	29.95	30.93	30.73	30.53	31.18	31.16
32	25.30	26.38	27.45	26.81	27.71	28.21	27.48	26.90	27.57	26.70
33	20.39	22.58	23.51	24.94	24.07	25.28	23.46	24.28	24.70	24.82
34	23.43	24.40	23.98	24.98	25.51	25.15	25.38	25.80	26.06	26.24
35	27.01	27.59	27.58	27.71	27.13	27.82	27.12	27.93	27.64	27.68
36	25.21	26.02	26.49	27.04	27.72	26.29	27.44	27.60	28.85	28.92
37	25.62	25.59	24.65	25.21	26.02	25.81	24.81	24.34	24.16	25.61
39	15.48	18.95	20.63	19.04	21.66	25.53	24.85	25.69	25.40	25.16
40	24.24	24.92	25.47	25.13	25.78	21.90	20.08	21.14	21.35	20.08
41	22.90	23.08	24.13	22.55	22.27	27.15	26.14	26.33	27.32	26.93
42	24.30	25.57	24.69	25.34	24.60	26.57	26.88	28.01	27.82	27.07
43	26.37	26.43	26.88	27.16	27.33	21.37	18.63	20.45	22.55	23.80
44	51.99	54.08	50.50	52.99	53.75	52.57	32.44	32.15	32.08	33.39
45	13.68	27.91	20.38	24.95	23.64	27.63	28.33	26.08	26.14	30.95
46	45.93	46.46	46.35	47.01	45.28	44.24	46.03	45.19	44.09	45.91

Note: Industry code is year-specific. To cross-validate with China Statistics Yearbook, refer to Table 14-4 in 2004, Table 13-6 in 2003, 2002 and 2001, Table 13-7 in 2000, and Table 13-9 in 1999. This statistic is not reported in Yearbooks since 2005. For industry code meanings, see Table 3.9 (post-2003 codes).

In matching firms, we directly use the aforementioned program as it is, except that we deleted the variables "street", "town" and "village" whenever they appear in the original program. This is simply because our data don't have these variables. This is a minor issue. The matching program first attempt to match by the ID of the firm's legal person, the to match the unmatched firms by firm name, then by legal person name, then by phone number plus county code, then by founding year plus county code plus industry code plus the name of the main product. The firm's legal person ID already links the vast majority of firms in the first step.

We feed the matched dataset into another Stata program provided by Brandt et al. (2012) to compute real capital stock for each firm in 1998-2007. Our only major modification to the original program is to add the year 2007, which the original program does not include. The variable picked up as the nominal capital stock is called "fa_original" (see Table 3.8). The main idea of the program is to first back-out the nominal capital stock at the first year of operation for each firm and then compute forward real capital stock year-by-year using Brandt-Rawski investment deflator (provided inside the program). It is assumed that the yearly depreciation rate is 0.09.

We then compute real value-added output by subtracting real input from the real output. This is the way it is constructed in Brandt et al. (2012). Real input and real output are computed by separately deflating raw input and raw output using industry-level (2-digit code industries) deflators. For our choice of raw input and output, please see Table 3.8. Our main analysis uses the industry-level deflators constructed by Yang (2015), which comes together with our dataset. Since Yang (2015) is published in Chinese, we translate his description of how these deflators are constructed below:

"This paper differs notably from Brandt et al. (2012) in how (annual 2-digit industry level) output and input deflators are constructed. First, the year-on-year by-industry output price indices come from China City Price Yearbook 2011. In Brandt et al. (2012), such price indices during 1998-2003 are computed as weighted averages based on real output and nominal output in the given data. Second, when computing input deflators, Brandt et al. (2012) only utilize the Input-Output table from one period, which does not consider structural changes over the years. This paper handles this issue more properly by using the 1997 124-industry Input-Output table for weights in years 1998-2000, using the 2002 122-industry Input-Output table for weights in years 2001-2005, and using the 2007 135-industry Input-Output table for weights in years 2006-2009."

Table 3.8: Variable name crosswalk: from raw data to Brandt et al. (2012) programs

Interim variable	Name in raw data	Meaning
	Key variables	
output	gyzczxjxgd	Nominal value of production (post value-added tax)
input	zjtrhj	Sum of intermediate inputs
employment	cyrs	Number of employees
fa_original	gdzcyjhj	Fixed assets at original value
cic	hylb	4-digit industry code
zip	yzbm	6-digit postal code
dq	dqdm	6-digit county code
	Variables used in data quality check	
fa_net	gdzcjznpjye	Net value of fixed asset
sales	gyxsczxjxgd	Sales
export	ckjhz	Value of export products
va_tax	bnyjzzs	Value-added tax
	Variables used only for matching firms over time	
id	frdm	ID of legal person
legal_person	frdbxm	Name of legal person
name	qymc	Name of the firm
phone	dhhm	Phone number
product1_	cp1	Product 1
bdat	kysjn	Start year
	Variables used only for identifying invalid observations	
	zczj	Sum of all assets
	ldzchj	Sum of liquid assets
	gdzchj	Sum of fixed assets
	wxzc	intangible assets
	fzhj	Sum of all debts
	ldfzhj	Sum of liquid debts
	cqfzhj	Sum of long-term debts
c_dep	ljzj	Cumulative depreciation
a_dep	bnzj	Current depreciation

Note: The variable "va" is computed using the formula va=output−input+va_tax. The names in raw data are abbreviations of the variable names in Chinese pinyin. The interim names are those used in Brandt et al. (2012) programs (available on their websites).

In our robustness checks, we use the deflators constructed by Brandt et al. (2012) (available on their website). They do not provide deflators for mining and utility industries, for which we use the Yang (2015) deflators. The industry codes changed in 2003 (see Table 3.9). We make the codes consistent over time by using the crosswalk provided by Brandt et al. (2012). It is essentially coding industry using post-2003 codes. The

county code also went through extensive revisions over the years. This is relevant for our analysis because we will be clustering standard errors along the county dimension in some of the regressions. We thus created a county code crosswalk table, where each county has a time-invariant code.

In the final steps of firm data cleaning, we drop observations with employment less than 10 or missing, with missing or non-positive real value-added or real capital stock, or with values outside the 0.5 to 99.5 percentile range for real value-added, employment, and real capital stock. We then translate the real values (now in 1998 RMB) into 1998 USD using the 1998 International Monetary Fund annual average middle exchange rate for US dollar to Chinese yuan (8.2790). Finally, we translate them into 2007 USD using U.S. Bureau of Economic Analysis, Gross Domestic Product: Implicit Price Deflator [GDPDEF] (retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/GDPDEF, October 8, 2020. deflator values are q1/1998 and q1/2007: 74.933, 91.708).

For a more detailed description of the matching algorithm and the way the real capital stock and deflators are constructed, please refer to the Appendix of Brandt et al. (2012).

A.3 Historical weather data quality

The raw data for historical weather come from National Centers for Environmental Information Global Surface Summary of Day data, which is aggregated from hourly raw observations by the National Centers for Environmental Information (NCEI) at the National Oceanic and Atmospheric Administration (NOAA). Zhang et al. (2018) start with the raw hourly observations data and aggregate to daily by themselves. We instead choose to directly use the version that is already aggregated by NCEI to the daily level, for three reasons. First, it saves us time and energy. Second, it is apparent from the documentation of hourly data that different observations are of different levels of reliability. Third, it is hard for us to properly aggregate the observations on precipitation in the hourly dataset because, although the observations are once every three hours, the precipitation could be reported as cumulative value of the previous 3, 6, 12, or 24 hours, and the report of precipitation does not occur evenly across time. Moreover, precipitation reports with overlapping time intervals happen all over the place. According to its documentation, the daily dataset that NCEI creates seems to have dealt with the above second and third issues.

Table 3.9: Industry codes

Industry code post-2003	Industry code pre-2002	Name
6	6	Coal mining
7	7	Oil and natural gas mining
8	8	Ferrous metal mining
9	9	Non-ferrous metal mining
10	10	Nonmetal mining
11	11	Other mining
	12	Logging
13	13	Processing of foods
14	14	Manufacture of foods
15	15	Beverages
16	16	Tobacco
17	17	Textiles
18	18	Apparel
19	19	Leather
20	20	Timber
21	21	Furniture
22	22	Paper
23	23	Printing
24	24	Articles for cultures and sports
25	25	Petroleum
26	26	Raw chemicals
27	27	Medicines
28	28	Chemical fiber
29	29	Rubber
30	30	Plastics
31	31	Non-metallic minerals
32	32	Smelting of ferrous metals
33	33	Smelting of non-ferrous metals
34	34	Metal
35	35	General machinery
36	36	Special machinery
37	37	Transport equipment
39	40	Electrical machinery
40	41	Communication equipment
41	42	Measuring instruments
42	43	Manufacture of artwork
43	6290	Recycling
44	44	Electricity and heat
45	45	Gas production
46	46	Water

Comparing our summary statistics in Table 3.1 with Table 1 in Zhang et al. (2018), we find our weather numbers are overall close to each other, except for the numbers for precipitation. Our number is 33.94 inches (per year), while theirs is 73.17 inches (per year). A simple cross-check with any precipitation map of China available would reveal that the value 73.17 is way too high. A possible reason for this is that Zhang et al. (2018) might sum up the precipitation observations in the raw hourly dataset without realizing that the observation intervals overlap.

A.4 Historical weather data cleaning

We start by taking out observations from weather stations in Mainland China and Taiwan from the Global Surface Summary of Day data. We keep those stations that are present in all years 1998-2007. At this point, we have 377 stations, close to the number reported in Zhang et al. (2018). We properly identify missing values based on their respective codes in the raw data, for temperature, precipitation, wind speed, visibility, and dew point temperature. We compute the relative humidity according to the formula provided in Zhang et al. (2018). We then keep weather stations with valid temperature data for at least 364 days in any year during 1998-2007. By doing so, 136 stations are dropped. We then drop 2 stations that have 2 days missing in 2000. We then interpolate the missing day with previous and subsequent days for the remaining 239 stations. The geographic distribution of the 239 stations is shown in Figure 3.9.

We use the data from these 239 stations to construct daily weather for all postal codes ever-present in the firm dataset. We first find the geographic coordinates for each postal code using the Google Geocoding API service. (An address with postal code as its only component is fed into the API in the process) For each postal code, we then follow Zhang et al. (2018) to compute daily values for temperature, precipitation, wind speed, visibility, and relative humidity, as a weighted average of the values from stations within 200 KM radius, with the weight as the simple reciprocal of the distance between them. The distance is computed using the Haversine formula. The daily data are then collapsed into yearly data in the same way as in Zhang et al. (2018). The yearly historical weather data are then merged with the cleaned firm panel by year and postal code. The merged dataset is the one underlying the summary statistics in Table 3.1.

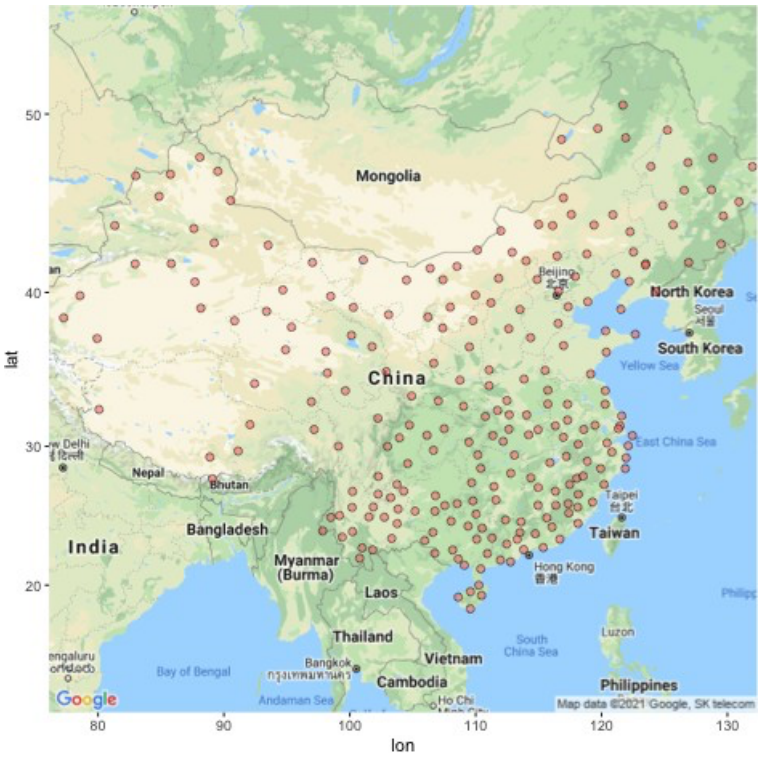

Figure 3.9: Map of weather stations used for historical weather observations

A.5 Future weather data

The raw data for future (2040-2042) daily weather come from the Centre for Environmental Data Analysis in the UK (HadCM3 A1FI run, stored as part of Met Office data from the Climate Impacts Link Project). We thank the Centre for Environmental Data Analysis for granting us access. The data is stored in PP binary files, which is an uncommon format. Citing CEDA: "PP-format is a record-based binary format used in a number of datasets archived in the CEDA archives. It is a Met Office proprietary format mainly associated with Met Office products, though not exclusively." Each day and each variable is stored in a separate pp file. We downloaded the data for the years 1998-2007 (for cross-validation purposes) and the years 2040-2042. We could only see data for part of 2045, and none of 2044 or 2046-2049 is available.

We start by processing and combining the PP files into files with Python module IRIS. (Please check CEDA documentations) We then keep those coordinates that are between latitude $8 - 54$ and longitude $73 - 135$. This covers all of China. We then construct postal code level daily temperature, precipitation, wind speed,

and relative humidity the same way we construct them with the historical weather from NOAA, except that the inclusion radius is now 300 KM.

The final step for cleaning HadCM3 weather data is cross-validating with historical actual weather. Following Zhang et al. (2018), we compute the average difference between HadCM3 predictions and actual observations across years for each day, each postal code, and each variable in 1998-2007. These average differences are then added to HadCM3 future daily predictions as corrections. A notable feature of HadCM3 data is that in their world, a month always has 30 days. Thus a year always has 360 days. Thus, the above cross-validation is done for those 358-359 days that actually exist, and the final cleaned daily future weather data for each postal code thus have only 358-359 days in a year for the years 2040-2042. Finally, these daily data are aggregated into yearly observations for each postal code in the same way as in Zhang et al. (2018).

Appendix B: Results with OP

Table 3.10 reports elasticity estimates using OP, with investment as the proxy variable.

Table 3.10: Production Function Elasticity Estimates with OP

Industry	Labor	S.D.	Capital	S.D.
6	0.22	(0.01)	0.14	(0.02)
7	-0.07	(0.12)	0.66	(0.35)
8	0.57	(0.03)	0.31	(0.07)
9	0.30	(0.03)	0.27	(0.08)
10	0.30	(0.02)	0.07	(0.15)
13	0.37	(0.01)	0.12	(0.02)
14	0.36	(0.01)	0.09	(0.08)
15	0.33	(0.02)	0.47	(0.20)
16	0.54	(0.08)	0.23	(0.17)
17	0.37	(0.01)	0.12	(0.01)
18	0.45	(0.01)	0.12	(0.01)
19	0.41	(0.01)	0.12	(0.01)
20	0.39	(0.01)	0.35	(0.14)
21	0.53	(0.02)	0.09	(0.04)
22	0.34	(0.01)	0.12	(0.01)
23	0.25	(0.02)	0.60	(0.09)
24	0.41	(0.01)	0.09	(0.02)
25	0.21	(0.03)	0.38	(0.14)
26	0.25	(0.01)	0.14	(0.12)
27	0.31	(0.01)	0.13	(0.18)
28	0.39	(0.02)	0.17	(0.15)
29	0.34	(0.01)	0.13	(0.07)
30	0.39	(0.01)	0.12	(0.01)
31	0.24	(0.01)	0.11	(0.01)
32	0.37	(0.01)	0.17	(0.02)
33	0.31	(0.02)	0.13	(0.03)
34	0.33	(0.01)	0.13	(0.01)
35	0.30	(0.01)	0.13	(0.01)
36	0.22	(0.01)	0.07	(0.11)
37	0.38	(0.01)	0.11	(0.17)
39	0.34	(0.01)	0.12	(0.01)
40	0.39	(0.01)	0.11	(0.01)
41	0.27	(0.02)	0.03	(0.13)
42	0.39	(0.01)	0.12	(0.01)
43	-0.16	(0.17)	0.33	(0.59)
44	0.35	(0.02)	0.09	(0.02)
45	-0.07	(0.06)	0.14	(0.17)
46	0.28	(0.02)	0.07	(0.05)

Note: We use Stata canned command *levpet* for OP estimation. Standard errors in parentheses. Yang (2015) deflators are used for the construction of data that is used here.

We compare predictions on the effect of climate change on TFP between OP and Abito (2020). We find that OP has significant prediction bias (i.e., the point predictions are outside or on the edge of each other's 95% CI) in about 7 industries, including communication equipment, non-metallic minerals, transportation equipment, general machinery, special machinery, metal, and non-metal mining. The two have somewhat different (small compared to CIs) predictions in about 9 industries like raw chemicals, paper, leather, electricity and heating, printing, manufacture of artwork, rubber, furniture, and smelting of non-ferrous metal. Overall, with OP-estimated productivity, the effect of climate change on productivity is -4.90% on average, with 95% CI [-7.25%, -2.55%]. The numbers are −4.07% and [−7.01%, −1.13%] respectively for Abito (2020)-estimated productivity. Using OP-estimated productivity would overestimate the magnitude of the overall adverse impact of climate change on TFP by about 20%.