UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

BOUNDED RATIONALITY IN DECISION MAKING: A MACHINE LEARNING APPROACH

TESIS PARA OPTAR AL GRADO DE
DOCTORA EN SISTEMAS DE INGENIERÍA

VERONICA CECILIA DIAZ GOMEZ

PROFESOR GUÍA:
RICARDO MONTOYA MOREIRA

MIEMBROS DE LA COMISIÓN:
CRISTIÁN GUEVARA CUE
SEBASTIÁN MALDONADO ALARCÓN
MARCELO OLIVARES ACUÑA

SANTIAGO DE CHILE
2020

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE DOCTORA EN SISTEMAS DE INGENIERÍA
POR: VERONICA CECILIA DIAZ GOMEZ
FECHA: 2020
PROF. GUÍA: RICARDO MONTOYA MOREIRA

BOUNDED RATIONALITY IN DECISION MAKING: A MACHINE LEARNING APPROACH

A key task for an effective marketing strategy is to understand how the consumers make choices. The way in which the consumers adopt, maintain or change their preferences is fundamental for the designing of new products, direct marketing campaigns, pricing or demand estimation of new products. This is not an easy task, because people are always influenced by numerous internal factors, such as emotions, or external factors, such as life events, which could affect their choices.

In this research, we focus on understanding the consequences of consumer behavior in two different circumstances, both under the bounded rationality framework. First, in a simulated discrete choice experiment context, in which consumers pay selective attention to the attributes of each profile. Second, in an empirical context, where consumers face a life event that makes them adjust their preferences.

In our first work, we propose the use of a machine learning approach based on Support Vector Machines (SVM), to identify the non-attendance of attributes at individual level and to predict the consumer choices in a conjoint experiment. We conduct an extensive simulation study to investigate the performance of the proposed approach. We compare the performance of our proposed approach to different benchmarks from the literature. Our results with simulated data show a better performance in terms of the identification of the non-attended attributes, that improves the predictive ability of the choices of consumers. Finally, we test our approach in two empirical applications previously reported in the literature. We demonstrate the superiority of our approach and the alternative insights derived from our method.

In our second work, we study how the consumption behavior of first-time parents is affected, both during the pregnancy period and after birth. We combine a unique dataset that identifies precisely the date of childbirth with a supermarket credit card data. We observe detailed supermarket transactions and aggregated purchases made at different external companies using the credit card, to investigate the relationship between pregnancy, childbirth and consumption. To examine the causal effect of pregnancy and childbirth on consumption, we use a causal random forest methodology. Our results show statistically significant impacts in 44% of the analyzed product categories during the pregnancy period, and in 48% of the product categories studied during the post-birth period.

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE DOCTORA EN SISTEMAS DE INGENIERÍA
POR: VERONICA CECILIA DIAZ GOMEZ
FECHA: 2020
PROF. GUÍA: RICARDO MONTOYA MOREIRA

BOUNDED RATIONALITY IN DECISION MAKING: A MACHINE LEARNING APPROACH

Una tarea clave para el desarrollo de una estrategia de marketing efectiva es entender cómo el consumidor hace elecciones. La manera en que el consumidor adopta sus preferencias, las cambia o las mantiene en el tiempo, resulta fundamental para, por ejemplo, el diseño de nuevos productos, el diseño de campañas de marketing directo, políticas de fijación de precios o la estimación de la demanda. Esta no es una tarea fácil, porque las personas siempre están influenciadas por numerosos factores internos, como las emociones, o externos, como vivir un evento importante de vida, que podrían afectar sus decisiones de elección.

En esta investigación, nos enfocamos en comprender las consecuencias del comportamiento del consumidor en dos circunstancias diferentes, ambas bajo el marco de racionalidad acotada. El primero está en un contexto de experimento de elección discreta simulado, donde los consumidores prestan atención selectiva a los atributos de cada perfil. El segundo está en un contexto empírico, donde los consumidores se enfrentan a un evento de la vida que les hace ajustar sus preferencias.

En nuestro primer trabajo, proponemos el uso de un enfoque de aprendizaje automático basado en Support Vector Machines (SVM), para identificar los atributos no atendidos a nivel individual y para predecir las elecciones del consumidor en un experimento de análisis conjunto. Nosotros llevamos a cabo un extenso estudio de simulación para investigar el desempeño del enfoque propuesto. Comparamos el desempeño de del enfoque propuesto con diferentes modelos de referencia propuestos previamente en la literatura. Nuestros resultados con datos simulados muestran un mejor desempeño en términos de identificación de los atributos no atendidos, que mejora la capacidad predictiva de las elecciones de los consumidores. Finalmente, probamos nuestro enfoque en dos aplicaciones empíricas previamente reportadas en la literatura. Demostramos la superioridad de nuestro enfoque y de las ideas alternativas derivadas de nuestro método.

En nuestro segundo trabajo, estudiamos cómo se ve afectado el comportamiento de consumo de los padres primerizos, tanto durante el embarazo como después del nacimiento del bebé. Combinamos un conjunto de datos único que identifica con precisión la fecha del parto con los datos de la tarjeta de crédito de un supermercado, donde observamos en detalle todas las transacciones realizadas tanto en el supermercado como las compras agregadas realizadas en distintos negocios externos a la compañía, para investigar la relación entre el embarazo y el parto y el consumo. Para examinar el efecto causal del embarazo y el parto en el consumo, utilizamos la metodología causal forest. Nuestros resultados muestran impactos estadísticamente significativos en el 44 % de las categorías de productos analizadas durante el período de embarazo, y en el 48 % de las categorías de productos estudiadas durante el posparto.

Acknowledgment

Hoy, terminando esta importante etapa de vida, me siento profundamente agradecida por todo el apoyo que he recibido en cada etapa del camino recorrido.

Agradezco a mi madre por su paciencia, amor y apoyo incondicional.

Agradezco a mis amigas y amigos por la motivación, la contención en los momentos difíciles y el cariño.

Agradezco a mi profesor tutor, Ricardo, por su dedicación y preocupación hacia nuestro trabajo. Por siempre tener un tiempo para resolver mis dudas y porque en los momentos más difíciles sus palabras de aliento siempre me transmitieron paz y confianza.

Agradezco a la Universidad de Antofagasta por la confianza que depositaron en mí, a la Universidad de Chile por la formación de calidad recibida y a la Agencia Nacional de Investigación y Desarrollo (ANID, ex CONICYT) por el apoyo financiero necesario para completar mis estudios de doctorado.

Contents

List of Tables

List of Figures

Introduction

Understanding how the consumers adopt their preferences and make choice decisions, is crucial for the development of an effective marketing strategy. For example, the designing of new products, direct marketing campaigns, pricing and demand estimation of new products. This is not an easy task, because people are influenced by numerous factors that could affect their choice decisions.

In quantitative marketing, most of the models used to explain consumer behavior are, in general, structural models. Those model estimates are based on the rational choice assumptions of the social and economic theory. According to this theory, the rational agent must always maximize her expected utility function, has well-defined and stable preferences over time, is selfish, and her preferences are based both on beliefs and all the available information. However, several studies over time have empirically demonstrated that most of these assumptions are not met, giving rise to negative consequences in the estimated parameters. Researchers are making efforts to handle this situation in classic choice model,s as well as they are proposing new methodologies that perform well under bounded rationality.

In this investigation, we focus on understanding the consequences of consumer behavior in two different circumstances, both under the bounded rationality framework. The first one is in a simulated discrete choice experiment context, in which consumers pay selective attribute attention to each profile. The second one is in an empirical context, where consumers are facing a life event that makes them adjust their preferences. Each one of these two behaviors is analyzed in detail in chapter 1 and chapter 2 of the thesis respectively.

In Chapter 1, we study the " limited attention " behavior, i.e., when people do not take into account all the available information before making a choice decision. In a marketing context, it was introduced as "Attribute Non-Attendance" (ANA). In this work, we conduct an extensive simulation study to test the performance of different Support Vector Machines approaches under the variation of ANA rates in the data. We also use empirical data to validate our results. This chapter is divided into an introduction, literature review, methodology, simulation study, results, and conclusions.

In Chapter 2, we discuss the consequences of facing one of the most relevant life events in the family cycle: pregnancy and the birth of the first child. We are interested in finding out the preference changes when consumers experience the event. This type of transcendental milestones in the cycle of life of people demolishes the assumption that preferences are stable over time. People must react in a short period of time, in order to face the change and the stress

that it means. We make a complete causal analysis study using observational data from records of purchases in a supermarket and other external business. We identify the effect of pregnancy and the birth of the first child in different product categories. The chapter is organized as introduction, literature review, methodology, observational data, methodology implementation procedure, results, and conclusions.

Chapter 1

Preferences Estimation Under Bounded Rationality: Identification of Attribute non Attendance using an SVM Approach.

1.1 Abstract

There is a growing interest in Economics and Marketing regarding the problem of estimating preferences of consumers, when they partially ignore the information provided in discrete choice experiments, a problem introduced as Attribute Non-Attendance. This line of research explores the consequences of assuming that consumers consider all available information concerning attributes to evaluate the product alternatives, when in fact they might completely ignore some attributes. Diverse choice models have been developed to accommodate non-attendance using choice data. For instance, in latent class models, each segment corresponds to a particular combination of relevant and irrelevant attributes. Due to the combinatorial nature of such approach, researchers typically explore a limited number of specifications. However, although diverse modeling approaches have been proposed to accommodate this behavior, there is no research investigating the capability of these approaches to correctly identify the true non-attended attributes.

In this work, we propose the use of a machine learning approach based on Support Vector Machines (SVM), to identify the non-attendance of attributes at individual level, and to predict the consumer choices in a conjoint experiment. We conduct an extensive simulation study to investigate the performance of the proposed approach. We compare the performance of our proposed approach to different benchmarks from the literature. Our results with simulated data show a better performance in terms of the identification of the non-attended attributes, that improves the predictive ability regarding of consumers choices. Finally, we test our approach in two empirical applications previously reported in the literature. We demonstrate the superiority of our approach and the alternative insights derived from our method.

1.2 Introduction

There is a growing interest in Economics and Marketing regarding the problem of estimating preferences of consumers when they partially ignore the information provided in discrete choice experiments, a problem introduced in the economic field as Attribute Non-Attendance (ANA)(Hensher et al., 2005; Scarpa et al., 2009). This line of research explores the consequences of assuming that consumers use all available information related to product attributes when they face different product alternatives, when instead these consumers might be ignoring some attributes.

To identify preferences of consumers about products, researchers typically design stated-choice experiments in which several product alternatives are presented sequentially to potential customers (survey respondents) (Hensher, 2006). Each alternative is characterized by a set of attributes. Based on the decisions made by those respondents, different econometric methods are used to uncover and quantify the importance that the respondents put on each product attribute. One of the main outputs of these experiments is the identification of the relevant attributes at the respondent level. The relevant attributes are those that the consumer considers when evaluating the product alternatives. Obviously, among those attributes there are some more important than others, but all are considered to some extent in order to perform the evaluation procedure. With the information of the most important aspects, a firm can make several marketing decisions such as product design, advertising or price promotions. The common approach used to determine the subset of attributes that are relevant to customers is the post-processing of the parameters of the estimated model. For instance, when using additive models, such as the mixed logit model, the relative range of the part-worths can be used to represent the attribute importance. Such a post-processing task assumes implicitly that consumers consider all attributes when conducting the evaluation decision.

Despite the wide use of this full-information assumption, past research has shown that respondents may ignore some characteristics of the product for several reasons, including (i) lack of knowledge or uncertainty regarding product attributes, (ii) use of a simplifying heuristic to deal with high complexity tasks, or (iii) the fact that the attribute is truly not relevant for evaluating the product in the choice task (Hensher et al., 2005).

Assuming that customers use all information, can introduce important biases on the estimated preferences (Hensher et al., 2005) with strong implications on, for instance, willingness-to-pay estimates (Hensher et al., 2012). As stated-choice experiments have been incorporating more complex products, meaning that they are characterized by a large number of attributes, it is expected that many consumers will be more selective regarding the attributes they really consider when evaluating the products. In those settings, the limited memory capacity of consumers may imply a more selective attention and therefore, some attributes that are not relevant to the decision are not considered by the respondent. Consequently, it is crucial to properly account for this phenomenon such that the preferences are correctly estimated.

Diverse econometric models have been developed to accommodate non-attendance using choice data. For instance, although mixed logit models do not infer any information processing strategies directly, it is possible to accommodate ANA by setting marginal utilities to zero when respondents declare to ignore an attribute (Hensher et al., 2005). Another approach

that incorporates more structure is based on a latent class specification, in which each segment corresponds to a particular combination of relevant and irrelevant attributes. Due to the combinatorial nature of such approach, researchers typically explore only a limited number of specifications (Campbell et al., 2011). More recently, hybrid approaches that combine random parameter and latent class specification, have been successfully introduced (Hole et al., 2013; Hess et al., 2013; Hess and Hensher, 2013; Hensher et al., 2013).

Interestingly, previous research has been silent regarding the capability of these approaches to correctly identify the true non-attended attributes. That is, past research has focused on accommodating this phenomenon but has not shown whether the relevant and irrelevant attributes can be successfully identified using only choice data or not.

In this research, we propose the use of a machine learning approach based on SVMs to account for ANA. Building on previous research, we conduct a thorough simulation experiment to analyze the capabilities of the proposed approach in recovering the true non-attended attributes and the predictive ability of customer choices. We compare the proposed approach to existing benchmarks. Therefore, we contribute to the ANA literature by (i) presenting a novel approach based on SVM to accommodate this phenomenon, (ii) studying the extend to which the non-attended attributes can be identified, and (iii) comparing the proposed approach to well established benchmarks.

1.3 Literature Review

Essentially, ignoring attributes correspond to a non-compensatory decision-making process, in which customers do not trade-off on non-relevant attributes (see e.g., Campbell et al., 2008; Collins and Hensher, 2015b). In addition, attribute non-attendance can be seen as a extreme case of low-importance attributes. That is, there are attributes that do not seem to be much relevant and their exclusion does not affect the characterization of the choices made by the consumer. Diverse non-compensatory models could accommodate this behavior partially, although from different perspectives. Lexicographic and Elimination-by-aspects (EBA) models do not allow consumers to make trade-offs since attributes are ordered by importance, and thus, non-relevant attributes would appear at the end of the ranking of importance (see e.g. Kohli and Jedidi, 2007; Hauser et al., 2010). However, the relevance of the attribute is context dependent, meaning that there are choice tasks in which attributes that have been irrelevant in previous tasks become relevant to break ties. Other non-compensatory decision rules, such as conjunctive and disjunctive rules, impose thresholds on relevant attributes only in the first stage. In the second stage, all attributes have some relevance as in any compensatory process (Gilbride et al., 2006). Therefore, these models allow for ANA only in the first stage. Typically, models based on Conjoint Analysis do not allow for discontinuous preferences implied by attribute non-attendance (Hensher et al., 2012; Campbell et al., 2008). This non-attendance to attributes occurs when customers completely neglect some attributes, and focus their attention on a reduced subset of product attributes. Indeed, compensatory approaches assume that consumers make trade-offs across all product features when evaluating product alternatives. As a consequence, this approach can experience problems in eliminating irrelevant attributes across consumers, especially when there is limited individual-level data.

Several researchers have studied the consequences of assuming full attendance when consumers might make choices under partial attendance (e.g., Hensher et al., 2005; Rose et al., 2005; Hensher, 2007; Hess and Hensher, 2010; Hensher et al., 2012; Collins, 2012). All these researchers identified biases in both parameter estimates and willingness to pay measures, especially if consumers do not attend the price (Hensher et al., 2012). Furthermore, it was found that the sensitivity to particular attributes may not be statistically significant in full-attention models (Rose et al., 2005), or that parameter estimates could have counter-intuitive signs when ANA is not considered (Hensher, 2007; Collins, 2012). However, the studies showed that when ANA information is incorporated into the models, those problems are solved. For instance, Hensher et al. (2005) investigated the bias on willingness to pay estimates when attribute non-attendance information was not considered in a stated-choice study of car commuting routes. When they compared the value of travel time savings distributions, before and after accounting for the attribute processing strategy of each individual, they found sizeable differences in the mean estimates. Similarly, Collins (2012) used simulated choice data with two alternatives, each one described by three attributes.They found that ANA leads to downward bias in the mean of the preference coefficients, upward bias in the extent of preference heterogeneity (especially when the true value is low), and an increase in implausibly signed coefficients, particularly when there is greater preference heterogeneity.

The attribute non-attendance problem has received limited attention in marketing literature. Some exceptions are Swait and Adamowicz (2001), Gilbride et al. (2006) and Yegoryan et al. (2019). Swait and Adamowicz (2001) analyze how the consumer decision process changes with task complexity, showing evidence of attribute non-attendance. Gilbride et al. (2006) extended the Bayesian variable selection procedure proposed by George and McCulloch (1993, 1997), incorporating the probability of attending an attribute at the individual and choice level, instead of at the aggregated level. Their method allows for bi-modal distributions using a Hierarchical Bayesian approach with mass around zero within the conjoint framework. Yegoryan et al. (2019) use a latent class model that simultaneously allows for ANA and preference heterogeneity. To understand and validate ANA in marketing contexts, two existing empirical applications involving coffee makers and laptops were evaluated. They found that the majority of respondents ignore some attributes, which has implications for willingness-to-pay estimates, segmentation, and targeting.

In Economics, this phenomenon is typically investigated in the context of stated-choice experiments, similar to conjoint[1] in Marketing. Several studies have been conducted in the context of transportation to investigate preferences of consumers regarding different methods and travel routes. For instance, Rose et al. (2005) study stated choices of respondents in the case of airline carrier for an interstate holiday in Australia. Hensher et al. (2012) investigated ANA in a stated choice experiment where respondents were asked to choose different car commutes routes based on travel times and toll costs. More recently, Balbontin et al. (2017) also used experimental data conducted to evaluate a new Metro system in Sydney in 2009. In the stated-choice experiment, respondents were asked to compare the available alternatives at the time together with a metro option. Each alternative was described by travel time, service cost, reliability and crowding. Furthermore, literature has also investigated ANA in landscape valuation and environmental damage contexts. Scarpa et al. (2009) focuses on valuating the landscape in

[1]For the remaining of this document, we use conjoint analysis and stated-choice experiments interchangeably

Ireland. Each alternative corresponded to a landscape improvement strategy: the protection of Mountain Land from overstocking, the enhancement of the visual aspect of Stonewalls, the maintenance of Farmyard Tidiness, and safeguarding of cultural heritage.

Similarly, Campbell et al. (2010) studied the benefits of landscape restoration to remedy the environmental damage generated by illegal landfills in the hills of Belfast. Each choice task contained one on-site restoration attribute (improvement at the dump sites), and three off-site restoration attributes (improvement to water quality, wildlife habitats, and outdoor recreation). For each restoration attribute, three possible levels of improvement were available. Campbell et al. (2012) studied the conservation of rare fish in Ireland. Each alternative described the conservation status of each fish species after the implementation of the conservation schemes. As we have seen, ANA is a phenomena of growing attention on diverse fields.

Methodologically, past research has proposed different ways of accommodating ANA. First, the identification of the relevant attributes can be made by asking respondents to self-report which attributes they attend to (Hensher, 2006), and even asking them to state the reasons why they do not attend some other attributes (Alemu et al., 2013). However, it has been demonstrated that this self-reporting is vulnerable to reporting error, and that models which incorporate this information implicitly perform poorly (Hess and Hensher, 2010; Collins and Hensher, 2015a). Another avenue for identifying relevant attributes is the use of eye-tracking devices (Yang et al., 2015; Meissner et al., 2016; Van Loo et al., 2018; Yegoryan et al., 2019). Typically the data collected in eye-tracking experiments in the context of conjoint analysis is comprised of eye fixations and fixation duration to particular areas of interest (AOI) on the screen. Indeed, this type of research reports evidence of customers completely ignoring some attributes when evaluating the products (Yang et al., 2015). Unfortunately, it has also been reported that respondents look at areas that are not relevant while making a decision, that is, irrelevant attributes also receive fixations, which makes the use of these eye-tracking data unproductive for identifying irrelevant attributes. Indeed, recent research shows that eye-tracking information does not provide further information for ANA purposes and does not improve predictive ability of the models (Van Loo et al., 2018; Yegoryan et al., 2019). A third methodological approach to address ANA is to infer relevant and irrelevant attributes using the stated choices of respondents. Research using this approach has been successful in describing and predicting those choices. In our study, we follow this line of research and focus on inferring ANA using only respondent choices.

Using choice data, diverse econometric models have been developed to accommodate ANA, and mitigate its unwanted effects. Although a mixed multinomial logit model (MML) does not directly incorporate any information processing strategy, it is possible to infer ANA behavior through the estimation of consumer preferences. For instance, Hensher et al. (2005) accommodated ANA by setting the marginal utilities to zero for those attributes that respondents declared to ignore in a car commuters stated-choice experiment. Train and Sonnier (2005) used a stated-choice analysis in which customers had to choose among gas, electric and hybrid vehicles, and proposed a transformation of normally distributed part-worths, where this transformation induces bounds to capture a mass of coefficients at zero. Hess and Hensher (2010), in a travel routes stated choice experiment context, attempted to infer attribute processing strategies through a posterior analysis, by conditioning a random parameter model to the observed

choices at the individual level. A simple assignment of a respondent based on the mean of the conditional distribution seemed inappropriate, the authors looked for a measure that was able to indicate when the mean of the conditional distribution was significantly equal to zero (in order to assume non-attendance). They conclude that the most appropriate was the variation coefficient, that is, the relationship between the standard deviation and the mean of the conditional distribution, obtaining significantly high values for respondents who had a conditional distribution mean close to zero.

Conversely to a fully compensatory linear specification (e.g., logit or mixed logit), a latent class multinomial logit (LCML) approach allows incorporating structures about the information processing strategies into the model. In this case, each class is defined as a combination of attended and non-attended attributes. Hess and Rose (2007) were the first in using this approach to formally handle ANA, in which only one non-attended attribute was modeled. Due to the combinatorial nature of this approach, the exploration of a higher number of attributes and thus class specifications, is a very difficult task (Campbell et al., 2011). Hensher and Greene (2010) used stated choice data in which car driving individuals choose between tolled and non-tolled routes. They defined classes based on rules that recognize the non-attendance of one or more attributes, as well as common metric attributes aggregation phenomenon [2].

To accommodate ANA, they constrain some parameters to zero for some classes. In a similar way, Scarpa et al. (2009) set the preference coefficients to zero for non-attended attributes, while the coefficients of the attended attributes took the same value in all classes. This is called equality-constrained latent class (ECLC) approach. Later, different studies (Campbell et al., 2010; Hensher et al., 2012) have generalized the latent class model allowing each possible combination of attended and non-attended attributes, rising to a total of 2^k segments, where k is the number of attributes. The hybrid approach, that combines random parameter and latent class models, has been predominant in the literature during last years. This approach allows adding another layer of preference heterogeneity within each class (Hensher et al., 2013). For instance, Hess et al. (2013) proposed an hybrid model to manage the preference heterogeneity of respondents who have low sensitivities to an attribute. In their approach, they specify two values for each preference coefficient. In one class the coefficient is constrained to zero representing non-attention whereas, in another class, the same coefficient follows a continuous log-normal distribution. Their results indicate that in the majority of the cases studied, the ANA rates recovered by other widely-used models, such as latent class, were greatly exaggerated.

Another important contribution to the random parameter specification of the logit model to handle ANA was made by Hole et al. (2013). They proposed a Mixed Endogenous Attribute Attendance (MEAA) model, that relaxes the assumption of Endogenous Attribute Attendance (EAA) model about identical preferences for attended attributes, allowing parameter variation across respondents. They used choice experiment data designed to establish the relative importance of different doctors criteria when prescribing medicines. They found that the MEAA model, which allows for both non-attendance and preference heterogeneity simultaneously, outperforms the EAA model in terms of goodness of fit and provides a richer picture of the decision-making behavior of respondents than either the EAA model or the standard mixed multinomial

[2]This heuristic is related to the way that common-metric attributes (e.g., partitions of travel time or cost) are jointly evaluated as either separate or combined attributes

logit.

Interestingly, previous research has been silent regarding the identification of true non-attended attributes. That is, despite the existence of different approaches to accommodate ANA in choice experiments, past research has not demonstrated its capability to recover ANA at the individual or aggregated level. For instance, Mariel et al. (2013) used simulated choice data to evaluate the performance in terms of choice predictions of three different approaches previously proposed on the literature to handle ANA (MML, S-ANA, I-ANA). They used a factorial design with three alternatives and four different attributes. They investigated the predictive performance of the existing methods regarding choices of respondents, but did not investigate whether those methods could correctly identify non-attended attributes or not.

Therefore there is a need to have more and better tools that allow incorporating this type of non-rational behavior, and mitigate its negative effects in the choice prediction process. In this document, we propose a machine learning approach, which due to its structural and non parametric advantages, has been successfully used for choice prediction and attribute selection purposes. The advantages of SVM as a predictive choice method, in contrast to traditional discrete choice models, is that the former does not need to assume a specific structure of the parameters to represent a particular information processing strategy. Evgeniou et al. (2005) were able to predict choice preferences in a conjoint analysis context by solving a SVM equivalent optimization problem.

SVMs were introduced by Vapnik and Chervonenkis (1971) in statistical learning theory field, and it was originally designed to solve the binary classification problem. Since then, SVMs have been successfully extended for different problems such as regression analysis, clustering or multiple classifications problems. Unlike the majority of the statistical learning methods that focus on empirical error minimization, SVMs are based on the structural risk minimization principle. That is, an objective function capable of making a trade-off between the complexity control of the model, and the empirical error minimization. The former is measured by the VC dimension (Vapnik and Chervonenkis, 1971) of the hypothesis space, and it is achieved by the separating hyper-plane margin maximization. The empirical error is a measure of the predictive quality of the model. It has been demonstrated that the structural risk minimization approach mitigates the over-fitting problem.

SVMs formulation is flexible enough to allow for sparse solutions. Sparseness inherently reduce the dimension of the problem, allowing to find a more compact set of informative attributes. Dimension reduction improves the understanding of the decision process by obtaining a more parsimonious and meaningful representation of customer preferences. Furthermore, this procedure could mitigate the effect of the curse of dimensionality, allowing to find significant results from small data sets at the individual level, as we can find in marketing applications. Relevant attributes selection also reduces storage and computational processing requirements, increasing the speed of estimation. This is ratified because SVMs has been widely and successfully used as a tool for the attribute selection process. Several works have highlighted the advantages of SVMs in this area (Bradley and Mangasarjan, 1998; Chapelle, 2002; Guyon et al., 2002; Miranda et al., 2005; Maldonado and Weber, 2009; Maldonado et al., 2011).

The attribute selection process with SVM has also been adapted to be used in conjoint

analysis and choice prediction data. For instance, Maldonado et al. (2015), were the first to present a SVM-based technique for selecting relevant attributes to the classification function that models consumer preferences. They proposed a backward elimination algorithm to select a subset of attributes. They demonstrated that a SVM with fewer attributes improves the predictive ability of the model.

Taking advantage of the SVM capability to select attributes, we propose to adapt the SVM formulation to identify non-attended attributes and to predict consumer choices. Similar to Maldonado et al. (2015), we propose a two-stage procedure. In the first stage, attended and non-attended attributes are selected for each respondent and in the second stage, preferences are estimated by pooling the information across respondents. In contrast to Maldonado et al. (2015), that use a backward elimination algorithm in the first stage, our proposal considers a single attribute contribution criteria, which indicates the relative importance of each attribute. The main added value of using a contribution criterion, instead of an elimination algorithm, is the simplification of the estimation procedure by reducing it to a single estimation at the individual level, consequently with a smaller calibration and processing time.

1.4 Proposed SVM-based Approach for ANA

In this section, we present a SVM approach to model consumer choices under attribute non-attendance.

SVM have been successfully used to model choices in the context of conjoint analysis (Evgeniou et al., 2005; Cui and Curry, 2005). One of the characteristics of SVM is that, based on the specific formulation, is able to select a reduced number of attributes. In the context of conjoint analysis, previous work has shown that SVM can reduce the complexity of the implied model without compromising its predictive capability (Maldonado et al., 2015, 2017). In this work, we propose to adapt the formulation of SVM, taking advantage of its selection capability to identify the non-attended attributes.

1.4.1 Formulation

Consider $i = 1, \ldots, N$ respondents evaluating $k = 1, \ldots, K$ randomly presented different product profiles in a choice-based conjoint context. Each consumer chooses one profile at each choice occasion $t = 1, \ldots, T$. Every product profile is described by $j = 1, \ldots, J$ attributes and each attribute is defined by n_j levels. For simplicity we assume a finite number of discrete levels.

The SVM formulation problem can be specified assuming an additive utility function for each consumer i of the form $u_i(\mathbf{x}) = \mathbf{w}_i{}^T\mathbf{x}$, which represents the utility that consumer i assigns to profile $\mathbf{x}$. The vector $\mathbf{w}_i$ indicates the relative importance of each component of the product. A non-linear formulation can also be used in a choice-based conjoint context. However, given that such formulation typically does not reduce the complexity of the implied model substantially, it would not work identifying non-attended attributes.

The proposed SVM formulation considers the following elements: (1) consumer decisions, (2) fit, (3) heterogeneity control, and (4) attribute non-attendance. We now proceed to describe

each of these components.

1. Consumer decisions. Consumer decisions can be modeled as tuples of the form $(\mathbf{x}_{it}, y_{it})$, where $\mathbf{x}_{it} = [\mathbf{x}_{it}^1, .., \mathbf{x}_{it}^K]$ represents the description of the product profile, with $\mathbf{x}_{it}^k \in R^J \ \forall \ k \in K$, and $y_{it} = k$ indicates that consumer i chooses product k at choice occasion t. Following previous research, after all the responses are collected, the information can be rearranged such that the k-th chosen profile at each occasion t is the 1-st profile in our formulation, i.e., $y_{it} = 1 \ \forall \ i = 1, \ldots, N \ ; \forall \ t = 1, \ldots, T$.

   We assume that the respondent chooses the profile that maximizes her utility. That is,

$$u_i(\mathbf{x}_{it}^1) \geq u_i(\mathbf{x}_{it}^b) \ \ \forall b \in \{2, \ldots, K\}$$

   which can be written as

$$\mathbf{w}_i^T \left(\mathbf{x}_{it}^1 - \mathbf{x}_{it}^b\right) \geq 0 \ \ \forall b \in \{2, \ldots, K\}$$

2. Model fit to consumer decisions. SVM for CA considers a soft-margin approach allowing for profile valuation error (Evgeniou et al., 2005; Cui and Curry, 2005). That is,

$$\mathbf{w}_i^T(\mathbf{x}_{it}^1 - \mathbf{x}_{it}^k) \geq 1 - \xi_{it}^k \qquad \forall \ i \in \{1, \ldots, N\} \ ; \forall \ t \in \{1, \ldots, T\} \ ; \forall \ k \in \{2, \ldots, K\}$$

   Then, the SVM model obtains a formulation by considering the minimization of the valuation error (ξ_{it}^k) in the objective function.

3. Heterogeneity control. Given that typical conjoint experiments collect a few responses for each consumer, it is advisable to pool individual models towards a population model. We follow Maldonado et al. (2017) and achieve this goal by adding a component $\|\mathbf{w}_i - \mathbf{w}_0\|_1$ in the objective function. Where $\mathbf{w}_i$ correspond to the parameters of individual i and $\mathbf{w}_0$ to the population parameters.

4. ANA component. Note that the elements of $\mathbf{w}_i$ represent the importance of each attribute level (typically called part-worths). Thus, we define the attribute j contribution (AC_j) as the difference between the highest and the lowest part-worths of the attribute j. Formally,

$$AC_j(\mathbf{w}_i^j) = max(\mathbf{w}_i^j) - min(\mathbf{w}_i^j) \text{ and}$$

$$\mathbf{w}_i^j = (w_{i1}^j, w_{i2}^j, \ldots, w_{in_j}^j).$$

   Therefore, we take advantage of the regularization goal of SVMs and select the attributes whose contribution help to characterize consumer decisions in a less complex way. We achieve this goal by including this attribute contribution in the objective function. Note that this AC component is equivalent to a modified L_∞-norm (which corresponds to the maximum element of a vector) where the minimum of the vector is subtracted. That is $AC_j(\mathbf{w}_i^j) = \|\mathbf{w}_i^j\|_\infty - min(\mathbf{w}_i^j)$

Considering all these elements, the proposed formulation can be written as

$$
\begin{aligned}
\min_{\mathbf{w}_i, \mathbf{w}_0, \xi_{it}^k, AC_j(\mathbf{w}_i^j)} \quad & \sum_{i=1}^{N} \sum_{j \in S_i} AC_j(\mathbf{w}_i^j) + \theta \sum_{i=1}^{N} \quad \|\mathbf{w}_i - \mathbf{w}_0\|_1 \; + \; C_2 \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{k=2}^{K} \xi_{it}^k \\
s.t. \quad & \\
& \mathbf{w}_i^T(\mathbf{x}_{it}^1 - \mathbf{x}_{it}^k) \geq 1 - \xi_{it}^k \quad \forall\, i \in \{1, ..., N\}\; ; \forall\, t \in \{1, ..., T\}\; ; \forall\, k \in \{2, ..., K\} \\
& \xi_{it}^k \geq 0 \quad \forall\, i \in \{1, ..., N\}\; ; \forall\, t \in \{1, ..., T\}\; ; \forall\, k \in \{2, ..., K\} \\
& \mathbf{w}_i^j = 0 \quad \forall\, i \in \{1, ..., N\}\; ; \forall j \notin S_i,
\end{aligned}
\tag{1.1}
$$

where $\|\mathbf{w}_i - \mathbf{w}_0\|_1 = \sum_{j \in S_i} |\mathbf{w}_i^j - \mathbf{w}_0^j| \quad \forall\; i \in \{1, ..., N\}$, and S_i indicates the set of attended attributes by individual i.

Parameters $\theta > 0$ and $C_2 > 0$ manage the relative importance of each element in the objective function. Note that in this formulation, we aim to achieve the goal of selecting the relevant attributes at the individual level and simultaneously pool the individual models towards a population model. This trade-off (at the individual level) cannot be solved in one step successfully using linear programming techniques, therefore we propose to solve it in two steps. In the first one we estimate individual level models with the goal of identifying the attended and non-attended attributes for each individual and in the second step we estimate the complete model considering as input the information of attended and non-attended attributes from the first-step. Note that solving (1.1) without performing the first step leads necessarily to a non-sparse solution, since the component $\|\mathbf{w}_i - \mathbf{w}_0\|_1$ pools the information across customers and would shift the part-worths associated to non-attended attributes to $\mathbf{w}_0$ instead of zero. We now describe the two-step approach.

1.4.2 Model estimation: A two-step approach

To obtain the attended and non-attended attributes we solve a simplified version of (1.1) for each customer i. Given that this formulation is solved at the individual level, we do not need to pool the models across individuals and thus we remove the heterogeneity control component of (1.1). The resulting formulation can be written as:

$$
\begin{aligned}
\min_{\mathbf{w}_i, \xi_{it}^k, AC_j(\mathbf{w}_i^j)} \quad & \sum_{j=1}^{J} AC_j(\mathbf{w}_i^j) + C_1 \sum_{t=1}^{T} \sum_{k=2}^{K} \xi_{it}^k \\
s.t. \quad & \\
& \mathbf{w}_i^T(\mathbf{x}_{it}^1 - \mathbf{x}_{it}^k) \geq 1 - \xi_{it}^k \quad \forall\, i \in \{1, \ldots, N\}\; ; \forall\, t \in \{1, \ldots, T\}\; ; \forall\, k \in \{2, \ldots, K\} \\
& \xi_{it}^k \geq 0 \quad \forall\, t \in \{1, \ldots, T\}; \forall\, k \in \{2, \ldots, K\}
\end{aligned}
\tag{1.2}
$$

The trade-off between the minimization of the model complexity (with ANA) and the choice inconsistencies is controlled by the parameter C_1, which is determined via a cross-validation procedure.

In contrast to the backward elimination procedure used in previous research (Maldonado et al., 2015, 2017), we propose to select the attributes based on a minimum contribution criterion

function in one-step. That is, an attribute j must satisfy $AC_j(\mathbf{w}_i^j) \geq \varepsilon$ to be considered as attended. If the attribute j is identified as a non attended, we do not consider it to be part of a population problem stage. The procedure is formally presented in algorithm 2.

Algorithm 1 Contribution Criterion Procedure for the Identification of Attended Attributes

Require: Full attributes set S, threshold ε
Ensure: Individual part-worth for attended attributes and 0 for non-attended attributes
  for all respondent $i = 1, \ldots, N$ **do**
    $S_i \longleftarrow S$
    $\mathbf{w}_i \longleftarrow$ from SVM formulation (1.2) using S_i
    if $AC_j(\mathbf{w}_i^j) \leq \varepsilon$ **then**
      $S_i \longleftarrow S_i \backslash \{j\} \quad \forall j$
      $\mathbf{w}_i^j = 0 \quad \forall j$
    end if
  end for

Parameter $\varepsilon \geq 0$ corresponds to a relevance threshold for the relative contribution of each attribute. This threshold needs to be sufficiently small to avoid eliminating attended attributes. The stopping criterion is reached when the contribution of all remaining attributes is above this threshold, or when only one attribute remains.

1.5 Simulation exercise

The goal of the simulation exercise is to evaluate the performance of the proposed approach in terms of the identification of attended and non-attended attributes, and compare such approach to other classic methodologies used in the literature. Indeed, as mentioned before, to the best of our knowledge, there is no research showing the attendance recovery capabilities of the approaches used to identify ANA. The existing methodologies typically conduct the model selection based only on the predictive ability of consumer decisions (which is the observed information).

1.5.1 Simulating preferences under non-attendance

Following previous research (Toubia et al., 2004; Evgeniou et al., 2005), we simulate consumer decisions as follows. We consider $N = 125$ customers choosing among $K = 4$ product profiles for each of $T = 20$ choice questions. The implied size of the dataset is common in different applications of conjoint analysis (Rao et al., 2014). Each product profile is described by $J = 10$ different attributes, and each attribute is characterized by $n_j = 4$ levels. This corresponds to choice sets of medium to high levels of complexity, that would require high cognitive effort and where we expect to find higher rates of ANA. Thus, this context would constitute a good setting to evaluate the performance of the different approaches. To design the questionnaire, we select the first profile of each question for each customer randomly, from all possible profiles. After that, we generate the next three profiles in each question using a cyclic approach.[3]

[3]For instance, suppose that the first attribute for the first profile is $x_1 = 2$, then for the next three profiles, such attribute takes the values $x_2 = 3, x_3 = 4, x_4 = 1$, for alternatives 2, 3, and 4, respectively. The same

To generate the choices, we assume a mixed logit specification for consumer utilities. That is, the utility for individual i, alternative j and choice occasion t is represented as $u_{ijt}(\mathbf{x_{ijt}}) = \beta_i^T \mathbf{x_{ijt}} + \varepsilon_{ijt}$. Where $\beta_i \sim N(\mu, \Sigma)$ with mean $\mu = (-\beta, \beta, -\beta, \beta)$, covariance matrix $\Sigma = \frac{\beta}{3} I$, and where I is the identity matrix. We use $\beta = 3$ in this simulation. The error terms ε_{ijt} are simulated independent and identically distributed by an Extreme Value Type I distribution.

To incorporate non-attended attributes for each individual, we randomly set to zero a subset of attributes. To study how the model perform under different ANA conditions, we varied the number of non-attended attributes considering 0%, 20%, 40%, 60% and 80% of the attributes. Note that within each condition, where all simulated respondents have the same number of attended and non-attended attributes, the individuals may vary in the specific attended and non-attended attributes.

1.5.2 SVM Model Specifications

As described previously, the goals of the proposed SVM are twofold: to identify attended and non-attended attributes, and to characterize and predict consumer decisions. We separated the estimation procedure of the model in two stages, and presented the formulation for each one in (1.1) and (1.2). These two formulations admit variations in the specification of the objective function, particularly regarding the regularization component. Accordingly, we tested three specifications of the proposed SVM regarding the regularization in the first and second stages: Attribute Contribution (AC), L_1-norm (N_1), and L_∞-norm (N_∞). As some of these versions have been used for feature selection purposes, we also investigated the contribution of each specification to the performance of the corresponding model. In addition, we investigated procedures to select the attended attributes: Backward elimination algorithm proposed by Maldonado et al. (2015, 2017), and our proposed contribution criterion. In the first approach, attributes are sequentially eliminated whereas in the second, attributes are selected in one step as described in algorithm 2.

Therefore, we estimate 18 different specifications of the SVM (see Table 1.1 for a summary of all SVM specifications analyzed). We compare these specifications in terms of their implied ability to identify ANA and to predict consumer decisions.

Other approaches

We consider the following benchmarks from the ANA literature.

1. Multinomial Logit (ML).
2. Latent Class Multinomial Logit (LCML). Each segment represents a particular combination of attended and non-attended attributes. The estimated specification is based on Erdem et al. (2013), accommodating ANA by attributes instead of attributes levels.
3. Mixed Multinomial Logit (MML). Although this model is not used for ANA, we include it to compare the performance of the investigated approaches regarding fit and predictive ability of consumer decisions.

procedure is used for all attributes.

Model	Stage 1 regularizer	ANA Identification Procedure	Stage 2 regularizer
AC - b.e - AC	AC	Backward elimination	AC
AC - b.e - N_1	AC	Backward elimination	N_1
AC - b.e - N_∞	AC	Backward elimination	N_∞
AC - $c.c$ - AC	AC	Contribution criterion	AC
AC - $c.c$ - N_1	AC	Contribution criterion	N_1
AC - $c.c$ - N_∞	AC	Contribution criterion	N_∞
N_1 - b.e - AC	N_1	Backward elimination	AC
N_1 - b.e - N_1	N_1	Backward elimination	N_1
N_1 - b.e - N_∞	N_1	Backward elimination	N_∞
N_1 - $c.c$ - AC	N_1	Contribution criterion	AC
N_1 - $c.c$ - N_1	N_1	Contribution criterion	N_1
N_1 - $c.c$ - N_∞	N_1	Contribution criterion	N_∞
N_∞ - b.e - AC	N_∞	Backward elimination	AC
N_∞ - b.e - N_1	N_∞	Backward elimination	N_1
N_∞ - b.e - N_∞	N_∞	Backward elimination	N_∞
N_∞ - $c.c$ - AC	N_∞	Contribution criterion	AC
N_∞ - $c.c$ - N_1	N_∞	Contribution criterion	N_1
N_∞ - $c.c$ - N_∞	N_∞	Contribution criterion	N_∞

Note: Attribute contribution (AC), L_1-norm (N_1), L_∞-norm (N_∞), backward elimination (b.e.) and contribution criterion (c.c.)

Table 1.1: SVM approaches notation.

4. Latent Class Mixed Multinomial Logit (LCMML). A natural extension of the deterministic latent class model is a random parameter latent class model ((Hensher et al., 2013)).

   We based our benchmark analysis on (Hole et al., 2013), who proposed a Mixed Endogenous Attribute Attendance (MEAA) model, relaxing the assumption of LCML model about identical preferences for attended attributes, allowing parameter variation across respondents.

1.5.3 Performance Metrics

The main outputs of each model are (1) attended and non-attended attributes at the individual level and (2) fit and predictive ability of the models to consumer decisions. We now describe the performance metrics derived from these results.

ANA identification

- Non-attended Attributes Rate (ANA): Percentage of non-attended attributes per individual.
- Accuracy ($Accu$): Percentage of correctly identified attributes as attended or non attended.
- Sensitivity ($Sens$): Percentage of true attended attributes correctly identified.
- Specificity ($Spec$): Percentage of true non-attended attributes correctly identified.

- Precision of attended attributes ($Prec_a$): Percentage of predicted attended attributes correctly identified.
- Precision of non-attended attributes ($Prec_{ana}$): Percentage of predicted non-attended attributes correctly identified.

Model prediction

- Hit Rate In: Rate of successful predictions of consumer choices in the training data.
- Hit Rate Test: Rate of successful predictions of consumer choices in the out of sample data.

1.5.4 Estimation Procedure

The SVM specifications can be solved as linear programming (LP) problems. We use concurrent optimization with Gurobi Interactive Shell 8.1.1 solver.

Parameters C_1 , ε , C_2, and θ were tuned using a cross-validation procedure. Specifically, we employed a Leave-One-Out Cross Validation (LOOCV) strategy to tune these parameters using only the calibration data set. In each iteration of the LOOCV procedure, we trained the model with a subset of 15 questions per individual from the calibration data set. The estimated individual part-worths were used to predict once choice of the validation data set. This procedure iterates until each question in the training data set has been part of the validation data set. Each of the tuning parameters was evaluated from a grid. Following previous research we used the grids as follow: C_1, C_2, $\theta \in (0.03125, 0.0625, 0.125, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 40, 48, 56, 64, 72, 80)$ and $\varepsilon \in (0, 0.1, 0.25, 0.5\ , 0.75, 1)$. Finally, the parameters were determined considering the best average hit rate in the validation data set.

After determining the parameters, the part-worths were estimated using the entire calibration data set. Finally, the hit rate performance metrics were calculated in the out of sample data, which remained unused during the calibration process.

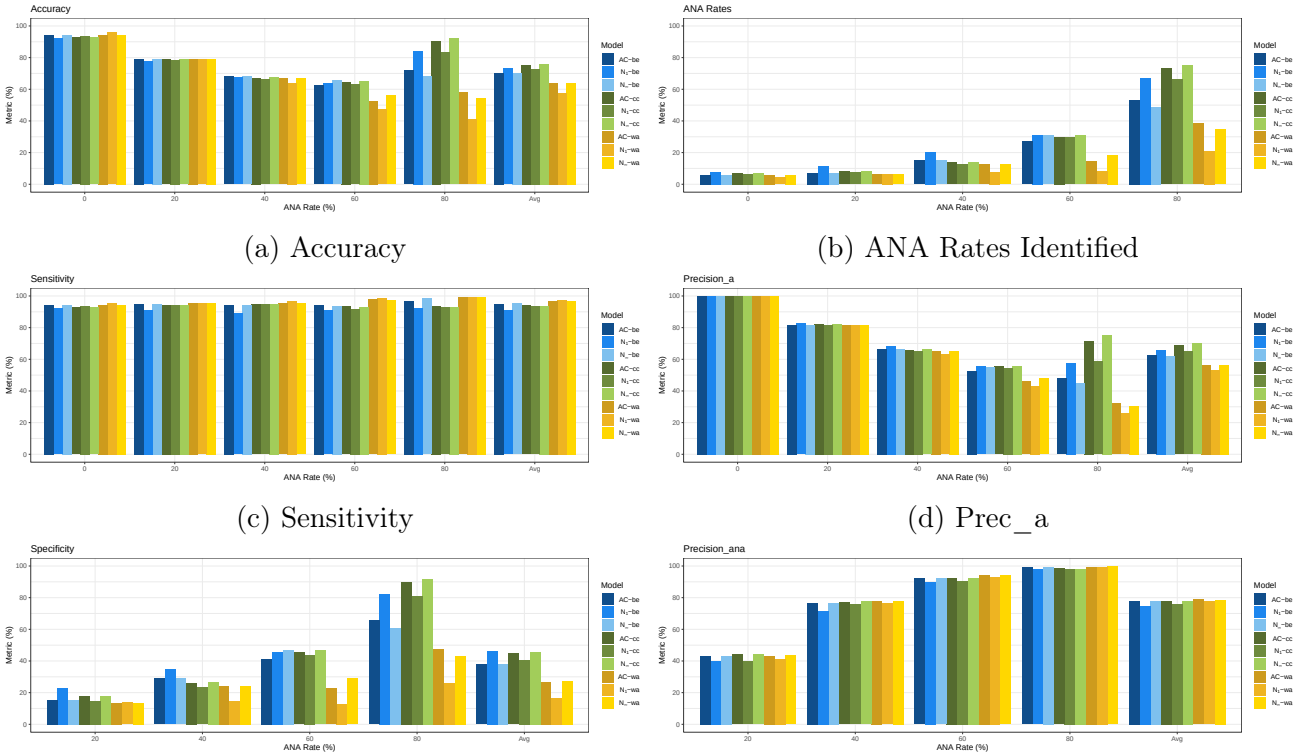
1.5.5 Results.

1.5.6 ANA Identification.

We tested all the SVM specifications mentioned in Table 1.1 using both, the backward elimination and the contribution criterion algorithms, to identify attended and non-attended attributes from the simulated choice data.

Tables 1.2 and 1.2 summarizes the results for all ANA identification metrics. We can note that the results are quite robusts to different SMV specifications. In figure 1.1 we can note that in general, accuracy and attended attributes precision tend to decrease when ANA Rates increase, except for an 80% of ANA, where both metrics improve. On the other hand, specificity and non-attended attributes precision tend to increase as ANA increases as well. As we can see, the predicted ANA rates for 20%, 40% and 60% instances are much less than what we actually simulated, this means that the SVM models underestimate the amount of non-attended

attributes when there is high or medium attendance. The consequences of this problem are that, although it has a good precision performance correctly identifying non-attended attributes, the underestimation of the real ANA rate makes the specificity to be lower. Then, the challenge here is to promote an even more sparse solution, in order to increase the number of correctly identified non-attended attributes, allowing to increase the predictive ability of the model, as we will analyze later. If specificity metrics increase, accuracy will also increase.



(a) Accuracy

(b) ANA Rates Identified

(c) Sensitivity

(d) Prec_a

(e) Specificity

(f) Prec_ana

Figure 1.1: ANA Identification Metrics

For a better performance evaluation of the different SVM elements, we conduct a regression analysis investigating the relationship between each component of a particular SVM specification and the performance metrics. The dependent variables are the ANA identification performance metrics as accuracy, sensitivity, specification, and precision rates. The independent variables are the SVM regularization components on the first stage problem (AC, N_1, and N_∞), the ANA identification algorithm (backward elimination or contribution criterion), and the simulated ANA rates (0%, 20%, 40%, 60%, and 80%). See Equation (1.3) for a particular instance of the regression analysis. The AC regularization component, no attribute elimination process (wa [4]), and full attendance instance were used as base levels, except for specification, where we used ANA_{20} rate as the base level.

$$Accu \sim \beta_0 + \beta_1 N_1^{1st} + \beta_2 N_\infty^{1st} + \beta_3 be + \beta_4 cc + \beta_5 ANA_{20} + \beta_6 ANA_{40} + \beta_7 ANA_{60} + \beta_8 ANA_{80} \quad (1.3)$$

As we can note on Table 1.3, there is no statistically significant difference between AC and

[4]Without Elimination Process

Ind	Model	Backward Elimination ANA Rate (%)					
		0	20	40	60	80	Avg
	AC	**94.24**	**79.03**	**68.09**	62.44	71.84	75.13
Accu	N_1	92.15	77.49	67.33	63.67	**84.23**	**76.97**
	N_∞	**94.24**	**79.03**	**68.09**	**65.59**	68.24	75.04
	AC	**94.24**	**94.98**	**94.2**	**94.23**	96.67	94.86
Sens	N_1	92.15	91.17	89.07	90.80	92.47	91.13
	N_∞	**94.24**	**94.98**	94.18	93.57	**98.33**	**95.06**
	AC	N/A	15.20	28.93	41.24	65.63	37.75
Spec	N_1	N/A	**22.8**	**34.73**	45.58	**82.17**	**46.32**
	N_∞	N/A	15.20	28.97	**46.93**	60.72	37.96
	AC	100	81.76	66.63	52.69	48.00	62.27
$\mathbf{Pr_a}$	N_1	100	**82.56**	**68.29**	**55.32**	**57.52**	**65.92**
	N_∞	100	81.76	66.64	54.89	44.63	61.98
	AC	N/A	**43.22**	**76.7**	**92.15**	98.96	77.76
$\mathbf{Pr_{ana}}$	N_1	N/A	40.09	71.43	89.74	97.80	74.77
	N_∞	N/A	**43.22**	76.66	92.05	**99.42**	**77.84**
	AC	5.76	7.05	15.05	27.05	53.17	
ANA	N_1	7.85	11.63	20.45	31.03	67.24	
	N_∞	5.76	7.05	15.08	30.73	48.91	

Ind	Model	Contribution Criteria ANA Rate (%)					
		0	20	40	60	80	Avg
	AC	92.73	**78.96**	67.28	64.55	90.60	78.82
Accu	N_1	**93.65**	78.48	66.11	62.96	83.35	76.91
	N_∞	92.75	78.95	**67.57**	**65.2**	**92.21**	**79.34**
	AC	92.73	94.22	94.73	**93.23**	**93.73**	**93.73**
Sens	N_1	**93.65**	**94.47**	94.67	91.83	92.67	93.46
	N_∞	92.75	94.22	**94.82**	93.03	93.13	93.59
	AC	N/A	**17.93**	26.10	45.42	89.82	44.82
Spec	N_1	N/A	14.53	23.27	43.71	81.02	40.63
	N_∞	N/A	17.87	**26.7**	**46.64**	**91.98**	**45.8**
	AC	100	**82.15**	65.89	55.48	71.49	68.75
$\mathbf{Pr_a}$	N_1	100	81.56	65.20	54.56	58.71	65.01
	N_∞	100	82.13	**66.1**	**55.92**	**75.39**	**69.89**
	AC	N/A	**44.15**	76.94	**92.57**	**98.32**	78.00
$\mathbf{Pr_{ana}}$	N_1	N/A	39.77	75.67	90.48	97.93	75.96
	N_∞	N/A	44.10	**77.64**	92.37	98.20	**78.08**
	AC	7.27	8.21	13.60	29.96	73.11	
ANA	N_1	6.35	7.33	12.51	29.49	66.28	
	N_∞	7.25	8.20	13.79	30.77	74.96	

Note: The average does not include 0 ANA instance.

Table 1.2: ANA identification metrics. SVM specifications.

N_∞, as regularization components of SVM in terms of the ANA identification metrics. In contrast, we note that AC performs better than N_1 in terms of sensitivity and ANA precision metrics. We can also note that the contribution criterion outperforms the backward elimination algorithm in terms of accuracy, specificity and attended attributes precision metrics. Furthermore, contribution criterion is capable of generating solutions with a higher ANA rate.

We also compared the performance of the SVM specifications regarding the identification of ANA with the LCML and LCMML approaches. Both latent class approaches allow us to compute a posterior probability of belonging to each class. This makes it possible to assign each respondent to a specific non-attention strategy defined by the segments of the model. Table 1.4 summarizes all ANA identification metrics.

Note that the data generation procedure for this simulation exercise follows closely a latent class mixed logit specification. As can be seen from Table 1.4, these models are successful in recovering the simulated attribute attendance structure. Consequently, we expect these models to outperform unspecified models (the proposed SVM specifications). However, we observe from Tables 1.2 and 1.2, that the SVM approach outperforms in terms of specificity, performs relatively close in terms of accuracy, and worse in the remaining metrics. This encourages us to continue investigating fairer procedures to simulate data in SVM choice models estimation context.

Because the first stage results indicate a better performance of the contribution criteria compared to the other analyzed alternatives, we also evaluate the contribution of each identification procedure to the final prediction ability of the model. These results are analyzed in the section 1.5.7.

1.5.7 Predictive Ability of ANA Models.

In terms of out-of-sample choice prediction, we notice from Table 1.5 that when the ANA rate increases, the performance of all SVM specifications remains stable without major variations. This demonstrates the robustness of this methodology under the presence of attribute non-attendance, compared to the benchmark models. Table 1.6 shows that the out-of-sample hit rate of ML, LCML, and MML approaches significantly decreases when the ANA rate increases. This phenomenon is not present in the LCMML approach, in which, similar to SVM models, it remains without significant variation as ANA increases.

The most competitive benchmark for our proposed model is the LCMML approach. Despite that the LCMML approach outperforms the contribution criterion SVMs model specification at the 60% ANA rate instance, we can observe that in the rest of the instances, as well as on average, the SVMs models, both with elimination algorithm and contribution criterion procedure, performs better.

As expected, in full attendance instances, all benchmark models predict better than SVM models. This is because the full-attendance data are the best scenario for these models that work under the rationality assumption. In fact, these benchmark models have been widely used in the literature to estimate preference coefficients in a rational conjoint analysis context. In contrast, the advantage of SVM is notorious in the presence of the non-attendance phenomenon.

	Dependent variable:					
	Accu	Sens	Spec	Prec_a	Prec_ana	ANA Rate
N_1^{1st}	−0.016	−0.011*	−0.019	−0.008	−0.017***	−0.062
	(0.014)	(0.005)	(0.026)	(0.014)	(0.004)	(0.161)
N_∞^{1st}	0.001	−0.0002	0.006	0.003	0.0004	0.018
	(0.014)	(0.005)	(0.026)	(0.014)	(0.004)	(0.161)
be	0.072***	−0.029***	0.170***	0.065***	−0.012**	1.010***
	(0.014)	(0.005)	(0.026)	(0.014)	(0.004)	(0.161)
cc	0.098***	−0.030***	0.201***	0.100***	−0.008·	1.245***
	(0.014)	(0.005)	(0.026)	(0.014)	(0.004)	(0.161)
ANA_{20}	−0.150***	0.007		−0.182***	0.425***	0.140
	(0.018)	(0.007)		(0.018)	(0.005)	(0.208)
ANA_{40}	−0.269***	0.006	0.097**	−0.342***	0.763***	0.745***
	(0.018)	(0.007)	(0.030)	(0.018)	(0.005)	(0.208)
ANA_{60}	−0.337***	0.007	0.211***	−0.483***	0.923***	1.823***
	(0.018)	(0.007)	(0.030)	(0.018)	(0.005)	(0.208)
ANA_{80}	−0.223***	0.023**	0.493***	−0.506***	0.988***	4.682***
	(0.018)	(0.007)	(0.030)	(0.018)	(0.005)	(0.208)
Constant	0.886***	0.961***	0.041	0.947***	0.012*	−0.114
	(0.017)	(0.007)	(0.030)	(0.017)	(0.005)	(0.197)
Observations	270	270	216	270	270	270
R^2	0.645	0.179	0.643	0.824	0.994	0.744
Adjusted R^2	0.634	0.154	0.631	0.818	0.994	0.736
Res. Std. Error	0.093	0.037	0.156	0.092	0.028	1.080
	(df=261)	(df=261)	(df=208)	(df=261)	(df=261)	(df=261)
F Statistic	59.275***	7.126***	53.520***	152.629***	5,630.976***	94.743***
	(df=8;261)	(df=8;261)	(df=7;208)	(df=8;261)	(df=8;261)	(df=8;261)

Note: . p<0.1; * p<0.05; ** p<0.01; *** p<0.001

Table 1.3: ANA Identification metrics. Regression analysis.

Ind	Model	ANA Rate (%)					
		0	20	40	60	80	Avg
	LCML	**99.88**	**88.83**	**86.63**	91.97	97.13	**92.89**
Accu	LCMML	99.79	87.92	85.85	**92.24**	**97.15**	92.59
	LCML	**99.88**	**96.57**	89.47	86.27	87.80	92.00
Sens	LCMML	99.79	96.38	**89.80**	**88.4**	**90.20**	**92.91**
	LCML	N/A	**57.87**	**82.37**	**95.78**	**99.47**	**83.87**
Spec	LCMML	N/A	54.07	79.93	94.80	98.88	81.92
	LCML	100	**90.19**	**88.38**	**93.19**	**97.62**	**92.35**
$\mathbf{Pr_a}$	LCMML	100	89.37	87.06	91.90	95.31	90.91
	LCML	N/A	**80.78**	**84.12**	91.29	97.03	88.31
$\mathbf{Pr_{ana}}$	LCMML	N/A	79.23	84.12	**92.47**	**97.58**	**88.35**
	LCML	0.12	14.32	39.27	62.96	82.01	
ANA	LCMML	0.21	13.71	38.09	61.52	81.07	

Note: The average does not include 0 ANA instance.

Table 1.4: ANA identification metrics. Benchmark models.

In the same line of the analysis results of ANA identification metrics, we conducted a regression analysis investigating the relationship between each component of each SVM specification and the performance metrics. The dependent variable was the choice prediction out of sample hit rate. The independent variables were the ANA identification procedure and the simulated ANA rates. See equation 1.4. No identification process (wa), and full attendance instance were the base levels respectively.

$$\begin{aligned} Hit\_rate_{test} \sim \beta_0 + \beta_1 be + \beta_2 cc + \beta_3 N_1^{2nd} + \beta_4 N_\infty^{2nd} \\ + \beta_5 ANA_{20} + \beta_6 ANA_{40} + \beta_7 ANA_{60} + \beta_8 ANA_{80} \end{aligned} \tag{1.4}$$

Table 1.7 shows that there are statistically significant differences in predictive ability between the different specifications of the SVM models. The contribution criterion for ANA identification procedure performs better than the backward elimination algorithm procedure. Similarly, in the population level problem, the AC regularizer component performs better than the L_1-norm specification. However, it is not statistically different from L_∞-norm.

1.6 Empirical Data

To evaluate the performance of the proposed approach, we used two choice-based conjoint experiments previously reported in the literature: Coffee Makers (Meissner et al., 2016) and Laptops (Yang et al., 2015). Note that in these two experiments eye-movement data were also collected. We do not use such information, because the eye-movement data do not provide further improvement neither to the identification of ANA, nor improvement the fit or predictions (Yegoryan et al., 2019).

Ind.	Model	Backward Elimination ANA Rate (%)					
		0	20	40	60	80	Avg
	AC-*AC*	91.61	93.16	95.47	98.82	92.72	94.36
	AC-N_1	90.46	92.26	95.75	97.96	91.70	93.63
Hit	*AC*- N_∞	86.68	**95.68**	95.58	97.16	91.95	93.41
Rate	N_1-*AC*	92.28	93.65	**96.05**	**98.86**	93.87	**94.94**
In	N_1-N_1	**92.34**	94.02	94.83	98.21	93.65	94.61
	N_1-N_∞	91.70	95.60	93.46	96.07	**94.09**	94.18
	N_∞-*AC*	91.61	93.16	95.49	98.04	93.90	94.44
	N_∞-N_1	90.46	92.26	95.73	98.21	93.42	94.02
	N_∞-N_∞	86.68	**95.68**	95.59	96.96	93.97	93.78
	AC-*AC*	75.48	66.00	62.21	61.89	62.60	65.64
	AC-N_1	**76.57**	**66.4**	61.90	60.33	61.73	65.39
Hit	*AC*- N_∞	75.78	66.20	62.07	**62.12**	61.34	65.50
Rate	N_1-*AC*	73.13	64.93	60.53	61.62	61.10	64.26
Test	N_1-N_1	73.60	64.80	59.83	59.75	61.36	63.87
	N_1-N_∞	73.47	64.93	60.60	60.88	61.52	64.28
	N_∞-*AC*	75.48	66.00	**62.28**	61.95	**63.26**	**65.79**
	N_∞-N_1	**76.57**	**66.4**	61.87	60.80	61.02	65.33
	N_∞-N_∞	75.78	66.13	62.17	61.49	61.06	65.33

Ind.	Model	Contribution Criteria ANA Rate (%)					
		0	20	40	60	80	Avg
	AC-*AC*	93.32	92.38	95.58	97.26	90.08	93.72
	AC-N_1	93.38	91.58	**95.93**	96.71	88.68	93.26
Hit	*AC*- N_∞	91.77	96.17	95.35	96.70	89.30	93.86
Rate	N_1-*AC*	94.32	91.13	93.71	**97.65**	**91.62**	93.69
In	N_1-N_1	93.84	89.13	92.85	97.11	91.13	92.81
	N_1-N_∞	**98.12**	94.98	92.70	95.42	90.46	**94.34**
	N_∞-*AC*	93.44	92.67	94.92	96.38	88.40	93.16
	N_∞-N_1	91.58	92.87	95.40	96.32	87.06	92.65
	N_∞-N_∞	85.64	**96.3**	94.59	96.59	87.29	92.08
	AC-*AC*	74.68	66.38	62.08	62.35	65.69	66.24
	AC-N_1	75.90	65.70	61.63	61.93	64.96	66.02
Hit	*AC*- N_∞	74.57	66.03	62.00	**62.89**	65.75	66.25
Rate	N_1-*AC*	74.25	66.50	62.15	60.66	61.99	65.11
Test	N_1-N_1	74.92	66.53	61.70	59.88	61.68	64.94
	N_1-N_∞	74.57	**66.67**	**62.7**	60.60	62.86	65.48
	N_∞-*AC*	74.90	66.53	62.64	62.66	65.64	**66.47**
	N_∞-N_1	**76.03**	65.47	61.50	62.21	65.29	66.10
	N_∞-N_∞	74.23	66.13	62.23	62.32	**65.79**	66.14

Table 1.5: Choice prediction success rates. SVM approaches with simulated data.

Ind	Model	ANA Rate (%)					
		0	20	40	60	80	Avg
Hit	MNL	80.33	69.24	59.91	51.13	40.85	60.29
Rate	LCML	80.65	80.64	81.25	78.78	68.18	77.90
In	MMNL	83.43	72.83	64.40	56.34	47.66	64.93
	MLCMNL	**83.9**	**84.12**	**85.28**	**83.88**	**75.04**	**82.44**
Hit	MNL	66.07	57.23	51.77	45.70	35.97	51.35
Rate	LCML	**66.33**	**61.17**	**61.13**	62.17	58.00	61.76
Out	MMNL	66.03	57.10	52.27	47.07	38.73	52.24
	MLCMNL	65.73	60.33	60.60	**63.27**	**61.17**	**62.22**

Table 1.6: Choice prediction success rates. Benchmark models with simulated data.

The Coffee Makers study sampled 59 regular coffee drinkers at a large European university. The analysis focuses on 12 conjoint choice tasks among three single-cup coffee brewers and a no-choice option. Each product is described by six attributes: brand (Braun, Krups, Philips, Severin), material (Stainless steel, Plastic, Brushed Aluminium), system (Pad, Capsule), design(A, B, C, D), price per cup (12 cents, 22 cents, 32 cents) and brewer price (99.99 €, 129.99 €, 159.99 €, 189.99 €). The choice experiment was designed orthogonal and level balanced.

The Laptops study sampled 70 respondents from a large European university. The respondents answered a survey with 20 conjoint choice questions in the behavioral lab of the university, using the online platform developed by Yang et al. (2015). Each product profile is described by six attributes: processor speed (1.6 GHz, 1.9 GHz, 2.7 GHz, 3.2 GHz), screen size (26 cm, 35.6 cm, 40 cm, 43 cm), hard drive capacity (160 GB, 320 GB, 500 GB, 750 GB), Dell support subscription (1 year, 2 years, 3 years, 4 years), McAfee antivirus subscription (30 days, 1 year, 2 years, 3 years), and price (350 €, 500 €, 650 €, 800 €). The questions were generated randomly (once for all participants, i.e., all participants saw the same set of questions).

1.6.1 Models and Estimation Procedure

After analyzing the simulated data results, we continue our empirical data study using the SVM specifications models that improve the performance in terms of ANA identification and choice prediction ability. That is, using AC and L_∞-norm as regularization component at the individual-level problem, contribution criterion as ANA identification procedure and also AC and L_∞-norm as regularization components at the population level. However, a complete analysis of all SVM specifications applied to empirical data can be found in Appendix A.

We used the same estimation procedure described before with the simulated data (see Section 1.5). First, we calibrated the parameters of the individual-level problem 1.2 and the contribution criterion algorithm 2 to ANA identification at the same time via LOOCV procedure. Then, we used the attended and non-attended attribute solution as input to the population-level problem. The second stage problem was also trained via LOOCV procedure. For the Coffee Makers dataset, we used the first 10 choice tasks as training sample and the last 2 choice tasks as holdout sample. In the case of the Laptops dataset, we used the first 16 choice tasks as

	Dependent variable:
	Hit.Rate.Test
N_1^{1st}	-0.012^{***}
	(0.003)
N_∞^{1st}	0.0004
	(0.003)
be	0.009^{*}
	(0.004)
cc	0.017^{***}
	(0.004)
N_1^{2nd}	-0.005^{*}
	(0.003)
N_∞^{2nd}	-0.004
	(0.003)
ANA_{20}	-0.095^{***}
	(0.003)
ANA_{40}	-0.140^{***}
	(0.003)
ANA_{60}	-0.148^{***}
	(0.003)
ANA_{80}	-0.134^{***}
	(0.003)
Constant	0.751^{***}
	(0.004)
Observations	630
R^2	0.810
Adjusted R^2	0.807
Residual Std. Error	0.027 (df = 619)
F Statistic	263.342^{***} (df = 10; 619)

Note: . p<0.1; * p<0.05; ** p<0.01; *** p<0.001

Table 1.7: Regression result of hit rate with test data.

training sample and the last 4 choice tasks as holdout sample. In both cases we tested C_1, ε, C_2 and θ parameters from the same grid used in the simulation exercise. The parameters were set on the combination that maximizes the average hit rate, computed using the validation question in the LOOCV procedure. After determining the parameters, the partworths of the utility function were estimated using the entire calibration data set. Finally, the performance metrics were calculated using the holdout sample.

All the SVM specifications were solved as linear programming (LP) problems, through concurrent optimization with Gurobi Interactive Shell 8.1.1 solver.

1.6.2 Results

As we can see on Table 1.8, SVM specifications outperform the benchmark models in terms of sample and out-of-sample choice prediction. Similarly to the simulated data, the LCMML model is the most competitive approach to account for ANA in both, Laptops and Coffee Makers datasets. These results agree with Yegoryan et al. (2019), who used the MEAA approach proposed by Hole et al. (2013), and compared their results with ML, MML, and EAA models. They found that the MEAA approach outperforms these benchmarks models in terms of in sample and out-of-sample prediction in the two empirical applications.

		Instance (%)		
Ind	**Model**	Laptops	Coffee	Avg
Hit Rate In	*AC-cc-AC*	86.61	95.48	91.05
	AC-cc-N_∞	86.61	**95.97**	**91.29**
	N_∞*-cc-AC*	86.79	88.87	87.83
	N_∞*-cc-*N_∞	**87.41**	92.90	90.16
	ML	55.71	53.54	54.63
	LCML	67.95	70.65	69.30
	MML	65.80	62.58	64.19
	LCMML	74.82	77.90	76.36
Hit Rate Test	*AC-cc-AC*	**72.08**	70.97	71.53
	AC-cc-N_∞	71.01	71.77	71.39
	N_∞*-cc-AC*	72.32	70.97	71.64
	N_∞*-cc-*N_∞	71.73	**74.60**	**73.16**
	ML	63.92	57.25	60.59
	LCML	60.36	61.29	60.82
	MML	50.71	56.45	53.58
	LCMML	70.71	68.54	69.63

Table 1.8: Choice prediction success rates with empirical data. SVM v/s benchmark models.

After estimating our proposed model, we are able to identify attended and non-attended attributes at the individual level. This is similar to the posterior membership analysis in LCML and LCMML models. Thus, we conducted a comparative analysis of the attribute attention probabilities for both empirical instances. The results are shown on figure 1.2.

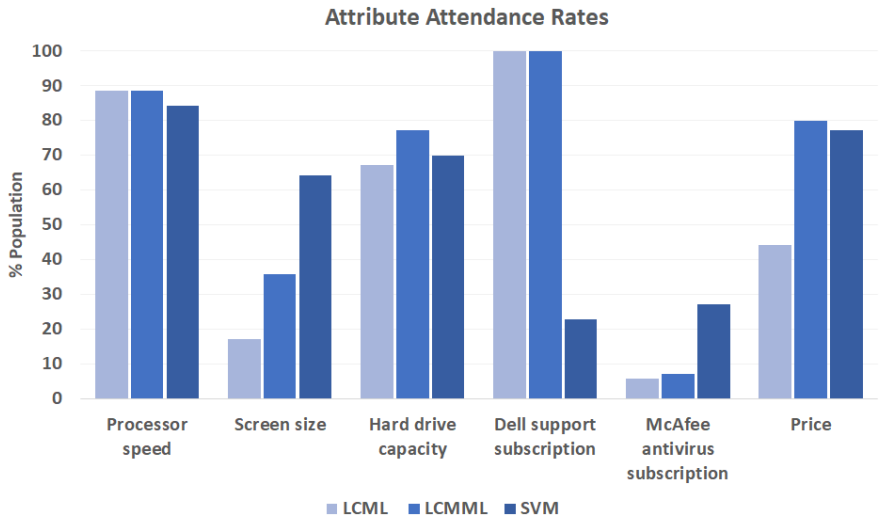
The first finding is that full attendance is not observed with the SVM approach, whereas, the LCML and LCMML models predict full attendance for Dell support subscription attribute in the case of Laptops, and for the brand and the design attributes for Coffee Makers. Moreover, it is possible to observe the non-attendance phenomenon present for most of the attributes in the two instances with the SVM approach. For Laptops instance, each attribute is attended by a 58% of the population on average, being Dell support subscription the least attended attribute for a 23% of the population and the processing speed the most attended one with a 84% of the population. In the case of Coffee Makers, the results are different as a higher attention rate is predicted. Each attribute is attended by a 74% of the population on average, being system the least attended attribute by a 59% of the population, and the brewer price the most attended one with an 84% of the population. These attribute attendance rates are consistent with the number of attended attributes per person analysis shown in the figure 1.3. We can observe there, in the Coffee Makers study, a shift of the probability distribution to the right, where the majority of respondent attended five and six attributes, instead of an evenly distributed data similarly to a bell curve for the Laptops instance, in which the majority of respondent attended three attributes.

We can also observe that models with greater preference heterogeneity tend to estimate higher attribute attention rates than the LCML model. This result also agrees with Yegoryan et al. (2019). An evident example is the screen size and the price attributes for Laptops instance, where SVM approach estimates a 74% and 43% higher rates respectively, compared with the LCML model. This phenomenon can be also observed in hard drive capacity and, antivirus subscription attributes for Laptops instance. In the case of Coffee Makers can be observed in material, system, price per cup, and brewer price attributes. It is also possible to account for some light exceptions, as the processor speed for Laptops or the brand and the design for Coffee Makers. However, we can also notice a strong exception in the Dell support subscription attribute for Laptops, where SVM account only for a 27% of attribute attention, instead of the full attendance estimated by LCML.

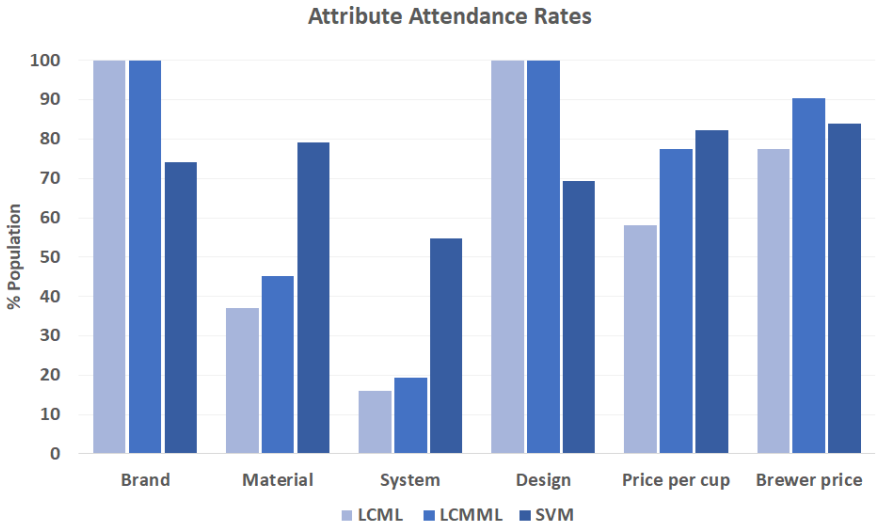
The results obtained with the SVM approach, both for Laptops and Coffee Makers, are quite intuitive. If the attributes were ordered from highest to lowest according to the attention rate identified, in the case of Laptops it would be plausible that the most important attributes to make a decision were the processor speed, the price, and the hard disk capacity. In the case of coffee makers the most important attributes would be the price of both brewer and cups.

1.7 Conclusions and Discussion

In this work, we conducted a complete analysis of the SVM approach capabilities to identify non-attended attributes, and to predict consumer choices in conjoint analysis data. Using simulated bounded rationality choice data, we tested a large set of different SVM specifications, modifying the regularization component at the individual and the population levels. Additionally, we tested two elimination algorithms to identify non-attended attributes: backward elimination proposed by Maldonado et al. (2015), and our proposed minimum contribution criterion. The two-stage model with the best performance, in both predictive capacity and processing time, was using indifferently attribute contribution (AC) or the L_∞-norm (N_∞) as a regularization components of the individual problem, minimum contribution criterion to identify non-attended attributes,

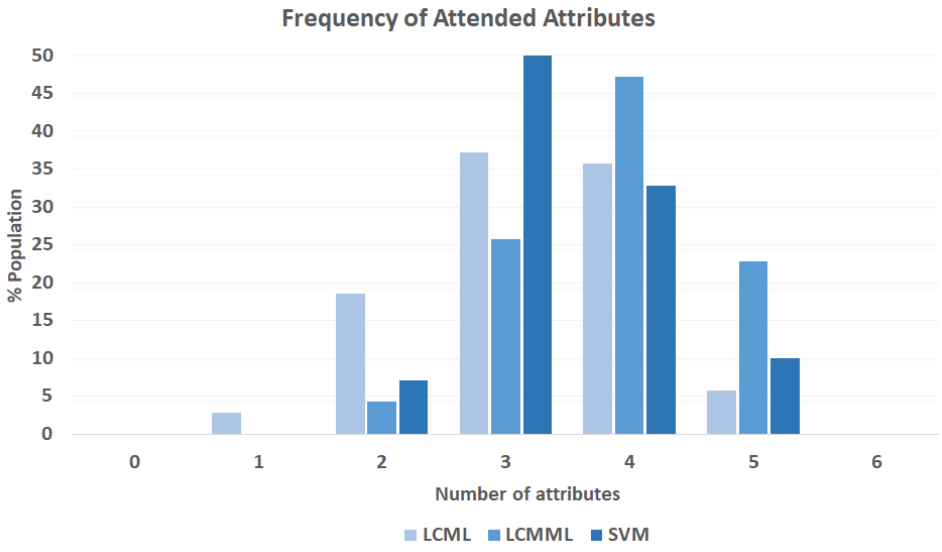


(a) Laptops

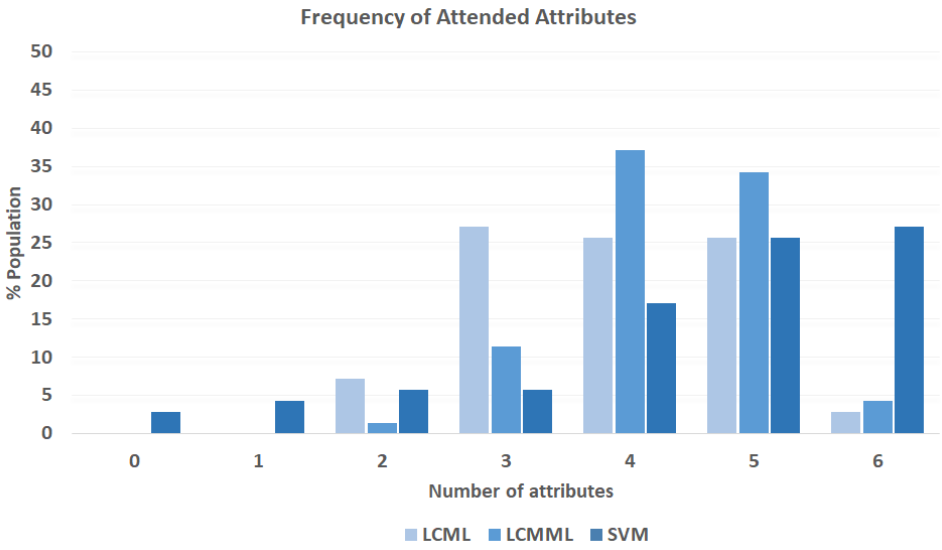


(b) Coffee Makers

Figure 1.2: Attribute attendance rates.



(a) Laptops



(b) Coffee Makers

Figure 1.3: Number of attended attributes per individual.

and indifferently the L_∞-norm (N_∞) or AC as a regularization component of the population problem.

The results with simulated data show that SVM exhibited to be a robust approach to the non-attendance rate variation. Unlike most Benchmark models, whose performance decreases as ANA increases. The predictive capacity of the SVM approach outperformed LC, LCML and MML models in all non-attendance instances, and in most of them to the LCMML approach proposed by Hole et al. (2013), except for 60% ANA instance.

Even though the SVM approach showed a good precision performance to correctly identifying non-attended attributes, the model underestimated the real ANA rate and this made the specificity to be lower. Therefore, the challenge here is to promote an even more sparse solution, in order to increase the number of correctly identified non-attended attributes. This would allow to increase the predictive ability of the model, because if specificity metrics increase, accuracy will also increase.

Although the results of the non-attended attributes identification process obtained with the LCML and LCMML models outperform the SVMs approach in all identification metrics, it does not translate into a better choice predictive capacity as it does with SVMs. In fact, the minimum contribution criteria procedure improved the out of sample choice prediction hit rate by 1.7% on average, compared to a population level model without individual-level information input. The impact of the attribute identification process is much greater as the ANA rate increases, reaching a 11% of improvement in the hit rate test using the minimum contribution criteria algorithm with 80% of ANA rate. This result can be explained because the data sets were simulated from a mixed logit model, considering a normal multivariate distribution for the parameters, randomly setting zero those coefficients of non-attended attributes, and an error that distributes Extreme Value Type I in the utility function. Therefore, latent class models were effectively able to properly recover the membership to each class defined by a specific combination of attended and non-attended attributes. A challenge for future work is to find a fair way to simulate a choice experiment data, that could be consistent with the SVM approach.

Once the SVM specifications that perform better with simulated data were selected, these models were used in two empirical data sets, both from conjoint analysis experiments in combination with eye-tracking tools. The first study, with Laptops choice data, has six attributes, each one with four levels. The second study, with Coffee Makers choice data, has also six attributes but different numbers of levels each, ranging from 2 to 4, in addition to a no-choice option.

As in the simulated data study, the specifications of the selected SVM approach perform better than all benchmark models in terms of out of sample choice prediction hit rate, in both empirical data set instances.

Because SVM performs better in empirical data than in simulated data at comparable ANA rates, we hypothesize that our prediction success estimates are underestimated. This is mainly due to the data generation process that actually follows a latent class mixed logit specification, instead of a SVM model. This encourages us to continue investigating fairer procedures to simulate data in SVMs choice models estimation context.

Another aspect of future work is related to the selection parameters criteria. We believe that the maximum average out of sample hit rate , as a criterion for setting parameters during the calibration process, could underestimate the predictive capacity of the model. Alternative forms related to the variance of posterior probabilities (Drechsler, 2010) have been suggested in the literature, and it could improve our results.

1.8 Appendix

1.8.1 Appendix A: Choice prediction success rates and regression analysis with empirical data.

Ind	Model	Backward Elimination Instance			Contribution Criteria Instance (%)		
		Laptops	Coffee	Avg	Laptops	Coffee	Avg
	AC-*AC*	89.38	93.39	91.38	86.61	95.48	91.05
	AC-N_1	89.20	93.87	91.53	86.52	94.68	90.60
Hit	*AC*- N_∞	88.48	93.39	90.93	86.61	95.97	91.29
Rate	N_1-*AC*	90.63	96.45	93.54	90.89	**97.26**	94.08
In	N_1-N_1	90.63	96.13	93.38	90.98	**97.26**	94.12
	N_1-N_∞	**91.16**	96.61	93.89	**91.61**	97.10	**94.35**
	N_∞-*AC*	90.00	**98.39**	94.19	86.79	88.87	87.83
	N_∞-N_1	90.80	94.03	92.42	87.05	89.52	88.28
	N_∞-N_∞	90.80	97.90	**94.35**	87.41	92.90	90.16
	AC-*AC*	73.39	72.58	**72.99**	72.08	70.97	71.53
	AC-N_1	73.39	67.74	70.57	72.26	71.77	72.02
Hit	*AC*- N_∞	73.04	70.97	72.00	71.01	71.77	71.39
Rate	N_1-*AC*	71.25	64.11	67.68	**73.39**	68.15	70.77
Test	N_1-N_1	71.61	66.53	69.07	71.43	69.76	70.59
	N_1-N_∞	71.25	64.92	68.08	72.14	64.92	68.53
	N_∞-*AC*	72.26	**73.39**	72.82	72.32	70.97	71.64
	N_∞-N_1	**75.12**	68.95	72.04	71.55	70.56	71.06
	N_∞-N_∞	74.05	71.37	72.71	71.73	**74.6**	**73.16**

Table 1.9: Choice prediction success rates with empirical data. All SVM approaches.

Table 1.10, similar to Table 1.3, shows a regression analysis. See equation 1.5 to evaluate the relationship between each component of each SVM specification, and the choice prediction hit rate using test empirical data. The independent variables were: SVM regularization components, both identification and choice prediction stages, the ANA identification procedure and the empirical data instance. AC regularization component, backward elimination procedure, and laptops instance was the base level fixed respectively in each case.

$$Hit\_rate_{test} \sim \beta_0 + \beta_1 N_1^{1st} + \beta_2 N_\infty^{1st} + \beta_3 cc + \beta_4 N_1^{2nd} + \beta_5 N_\infty^{2nd} + \beta_6 Instance \tag{1.5}$$

	Dependent variable:
	Hit Rate Test
Constant	0.732^{***}
	(0.009)
N_1^{1st}	-0.026^{***}
	(0.008)
N_∞^{1st}	0.005
	(0.008)
cc	0.003
	(0.006)
N_1^{2nd}	-0.003
	(0.008)
N_∞^{2nd}	-0.003
	(0.008)
Coffee Instance	-0.027^{***}
	(0.006)
Observations	36
R^2	0.557
Adjusted R^2	0.465
Residual Std. Error	0.019 (df = 29)
F Statistic	6.070^{***} (df = 6; 29)

Note: *p<0.1; **p<0.05; ***p<0.01

Table 1.10: Regression results, empirical data.

Chapter 2

Causal Analysis of Pregnancy and Child Birth on Consumption Behavior

2.1 Abstract

Major life transitions such as relocation, new job or pregnancy and birth of a child can have major implications on one's lifestyle and consumption patterns. In this research, we study how consumption behavior of first-time parents is affected, both during the pregnancy and after birth. We combine a unique dataset that identifies precisely the date of a childbirth with a supermarket credit card data, where we observe detailed supermarket transactions and aggregated purchases made at different external companies using the credit card to investigate the relationship between pregnancy and childbirth and consumption. To examine the causal effect of pregnancy and childbirth on consumption, we use a causal random forest methodology. Our results show statistically significant impacts in 44% of the analyzed product categories during the pregnancy period, and in 48% of the product categories studied during the post-birth period. The most affected categories by the first-child pregnancy were home improvement (+), travels (-), health (+) and entertainment services (-). On the other hand, the most affected categories by first-child birth were travels (-), health (+), restaurants (-), entertainment services (-) and pharmacy (+).

2.2 Introduction

A life event represents a break in the life path of a person, and split the time creating two periods: "before" and "after" (Sevin et al., 2008). Life events encourage people to carry out a large number of change adaptation activities in order to minimize the stress, and partially alleviate the work involved in facing those events. Examples of such life events are, for instance, having a child, getting married, home removals, children starting at school or facing the labor market by the first time. These adaptation activities can involve consumption activities or not.

Therefore, life events represent an opportunity for companies to offer timely and appropriate solutions to their clients under a life change process. If managers were able to identify when

customers are facing a crucial life event, they could offer products and services that might facilitate the adaptation to the process. This is not a trivial task, because there is a very thin line between the privacy vulnerability feeling, and the relief that is intended to generate. But, if consumers achieve to have a good consumption experience, then they will find themselves more receptive to being loyal to such companies.

The main contribution of our work to the marketing field is that, for the first time, a non-traditional causal analysis is carried out with observational data of real sales records, to estimate the consumption behavior effects of one of the most important life events in woman's life: motherhood. Specifically, we analyse the pregnancy and motherhood during the first nine months of the baby. Our findings are unique in this line of research.

Although this study was conducted with Chilean data, we believe that our results can be generalized to different geographical areas, because pregnancy and birth of a child is a life event that transcends cultures (Selin, 2009).

2.3 Literature Review

The life cycle of people is characterized by facing constant state changes, called "life events". A life event represents a break in the path of life, and create two periods: "before" and "after" (Sevin et al., 2008).

Several authors have developed the theory that a drastic change in the life course of people increase their stress levels, forcing them to create a generalized demand for readjustment and adaptation processes (Andreasen, 1984; Wheaton, 1990; Thoits, 1995; Mathur et al., 2008). Wheaton (1990) described the "life event form of stress" like a discrete and observable event that is thought to be threatening, because it represents a change.

Consumer behavior research in a marketing context has not been oblivious to the life events phenomenon. Andreasen (1984) introduced the concept of consumer "readiness-to-change". Behind this concept lies the idea that, if nothing changes in the consumers life, they tend to persist in their patterns of thought and consumption behavior to those which are already accustomed to. They hypothesized that a stressful life event may change the consumer attitudes, perceptions, and behavior with or without the intervention of a change agent. They show that the quantity of life status changes has a positive impact on brand preferences changes. In a similar line, Mathur et al. (2008) also suggests that stress is a mechanism that links life-changing events to changes in consumption patterns. They found a positive correlation between stressful life events, and consumption-coping behaviors.

There are other researches that do not necessarily link life events to stress as an engine for preferences changing, but rather explore different hypotheses about it. For instance, Koschate-Fischer et al. (2017) investigated some underlying factors that could influence changes in preferred brands of personal care products, such as, consumer innovativeness, variety seeking tendency and price consciousness. They show that people who experience a change in their life also have a higher predisposition to seek and try new products (consumer innovativeness) that,

in turn, also leads to a decrease in the share of wallet [1] for the preferred brand. Wilkes (1995) explored the stage of the household life cycle as an indicator of the allocation of expenditures. Their results show that as households make the transition from one stage to another, resources are from modestly to dramatically reallocated in order to accommodate the changed household circumstances and demands, depending on the stage and the product analyzed.

Therefore, events that change people's lives represent an opportunity for companies that offer products and services that can facilitate the adaptation to the change process. Consumers will be in a reassessing priorities process, and their consumption needs could be intensified (Mathur et al., 2008). If consumers achieve to have a good consumption experience, which partly relieves their anxiety about the change, then they will find themselves more receptive to being loyal to such companies. For this reason, it is crucial for marketing managers to be able to identify customers who are facing a life event, in order to offer timely and appropriate solutions to the change process.

The most relevant limitation of the reported studies here is that all change measures were based on self-reported data. For instance, Andreasen (1984) used cross-sectional data from an exploratory study carried out by telephone inquiries to 286 individuals from a large metropolitan area, Mathur et al. (2008) used retrospective and longitudinal data from surveys applied to 1442 household heads through mailing and postcard, and Koschate-Fischer et al. (2017) used information of 1473 German individuals from two combined data sets: first, an individual-level longitudinal data of self-reported purchase records for personal care products during three years, and second, demographic information from the panel members collected during the same period from a survey. This kind of data could potentially be subject to response biases (Andreasen, 1984), and in general, it is not large enough to perform a causal analysis. Instead, we are using data from real purchases records of customers in a major store in Chile. Using empirical data has the advantage of eliminating the possible bias generated by self-reported responses.

Most of studies in this line have analyzed the effects of several life events simultaneously. For instance, Andreasen (1984) studied the 23 most observed and objective life events from a list of 102 stressful life events proposed by Dohrenwend and Dohrenwend (1974). Koschate-Fischer et al. (2017) focused on a list of 10 recognized life events in the previous literature. They also differentiate between first time and repeated life events. Mathur et al. (2008), also following suggestions of previous researches, studied a list of 19 relevant life events to middle-aged and older people. Although all these works considered having a child as one of the life events analyzed in their lists, our study is specific and deep analyzing one of the most important milestones in a woman's life, the first experience of motherhood and its impact on consumer behavior. The pregnancy period and the birth of the first child are likely to be a key event in mothers' and fathers' lives. Several authors have considered it within the most important life events (Andreasen, 1984; Wilkes, 1995; Mathur et al., 2008; Selin, 2009; Koschate-Fischer et al., 2017).

The methodologies reviewed in this study used to uncover the effect of life events on consumer behavior have been based mostly on partial correlation analysis (Andreasen, 1984; Mathur et al.,

[1] Share of wallet is defined as "the percentage of money that a customer allocates to the preferred brand respect with the product category".

2008), and latent difference structural equation modeling (SEM) (Koschate-Fischer et al., 2017) approaches. These methodologies do not take care of the confounding problem (Rosenbaum and Rubin, 1983), therefore they cannot ensure that the estimated effect on consumer behavior effectively corresponds to the life event they are evaluating. In our case, we propose using a Causal Forest approach, which is a causal analysis methodology recently proposed in the literature, that allows us to quantify the true effect of the life event on consumer purchase behavior.

Causal analysis have been useful in many cases in which it is interesting to evaluate statistical inferences about causal effect of a specific treatment in a population. Fields such as healthcare, economics and education had been widely explored. Particularly in marketing, evaluating pricing policies, targeted advertising strategies or the impact of mass advertising campaigns, are frequent tasks that require estimations of the consumption behavior effects.

However, there are big challenges to generate valid statistical inference from observational data. First, because the treatment effect is the difference between two potential results, one of them is in practice impossible to observe, that is known as the potential outcomes framework (Rubin, 1974). Therefore, in order to evaluate causal effect in observational studies, the asymptotic theory is crucial. Second, because the information observed in observational data depends on variables which might also affect the outcome, that problem is known as confounding (Rosenbaum and Rubin, 1983).

Nowadays, it is possible to access a greater amount of data with information at the individual level, and hence estimating heterogeneous effects has become increasingly attractive. A classical approach used to estimate heterogeneous treatment effects has been nearest-neighbor matching.

There is a growing literature regarding the estimation of the effect of heterogeneous treatment, using machine learning tools. For instance, Tian et al. (2014) proposed a simple method of modifying covariates to use in a regression model, that allows to analyze potential interactions between treatment and a large set of covariates. Weisberg and Pontes (2015) proposed a cadit model, that can be useful in selecting from among a large number of potential predictors. They also introduced a new variable-selection algorithm that was applied in conjunction with cadit model, to identify and test individualized causal effects. They performed a successful treatment effect per groups analysis, despite that they found certain groups widely affected by the treatment. They suggest that predictive modeling should often replace classical subgroup analysis. Most recently, Taddy et al. (2016) presented a Bayesian nonparametric analysis to quantify the uncertainty associated with treatment effect measurement via both, linear projections and nonlinear regression trees (CART and random forests).

The main limitation is that those proposed methods were designed to analyze randomized clinical trials data, and may lose its causal interpretation when used in observational studies.

Shalit et al. (2017) focused on the problem of making causal effects predictions at the individual level based on observational data. They suggest a new type of regularization term in the generalization error, by learning representations with reduced IPM distance between treated and control, enabling a new type of bias-variance trade-off. Their method was tested using neural nets as representations and hypotheses. They applied this approach to both synthetic

and real data, showing that in every case their proposed method matches or outperforms the state-of-the-art.

Recently Wager and Athey (2018) developed a non-parametric causal forest method for estimating heterogeneous treatment effects, that extends the widely used random forest algorithm proposed by Breiman (2001). In the potential outcomes framework with unconfoundedness, they show that causal forests are point-wise consistent for the true treatment effect, and have an asymptotically Gaussian and centered sampling distribution. This work pioneer in proposing a method using the random forest to make statistical inference. Their results show that Causal Random Forest was substantially more powerful than classical methods based on nearest-neighbor matching, specially in the presence of irrelevant covariates.

Here lies the importance of our study. To our knowledge, our work is the first research being able to identify the causal effect of the first time pregnancy and childbirth on mothers' consumption changes. With our results, we can address the focus of the marketing force towards a mutual benefit that facilitates the adaptation process of the future mother, and in turn ensures their loyalty as a client. Although the study was conducted with Chilean data, we believe that our results can be generalized to different geographical areas, because pregnancy and the birth of a child is a life event that transcends cultures (Selin, 2009).

2.4 Methodology

2.4.1 Random Forest (RF)

The random forest is an ensemble method, which was originally designed for classification and regression.

A random forest is based on the decision tree algorithm for classification and regression (CART), proposed by Breiman et al. (1984). In the CART algorithm, a tree is constructed recursively dividing all the observations, and generating smaller branches. In this process, all covariates are evaluated in order to find the best candidate in each partition, aiming to maximize the improvement fit of the model. Therefore, observations within the same branch share similar values of covariates. Each branch of the tree ends in a node that is labeled according to the majority of votes, for categorical variables, or according to the average value of the observations, for continuous variables. The CART algorithm builds trees as large as possible, which could result in an unstable classification or prediction (Hastie et al., 2009). For this reason, random forests arise (Breiman, 2001).

In the random forest approach, trees are constructed using the CART algorithm based on the bootstrap samples of the original sample size. In this process, only a random subset of covariates in each partition of the tree is considered. In addition, the objective variable is predicted based on the average or a majority vote of the predictions on all trees. This generates a more accurate prediction compared to a single decision tree (Hastie et al., 2009).

2.4.2 Generalized Random Forest (GRF)

Unlike the classic random forest, generalized random forest (GRF), proposed by Athey et al. (2019), abandon the idea that the final estimation is obtained by averaging estimates from each member of an ensemble. Treating forests as a type of adaptive nearest neighbor estimator, is much more amenable to statistical extensions.

The main difference between GRF and the classic random forests approach to growing trees is in the quality measures of a split. GRF algorithm aims to maximize the heterogeneity in the quantity of interest across the child nodes, instead of the maximum improvement in MSE as the classic random forest does. Optimizing the heterogeneity criterion directly is very expensive to compute, and then, the algorithm optimizes only a linear approximation to the criterion, based on the gradient of the objective.

To predict, a test example is pushed down to determine what leaf it falls into in each tree of the forest. In the end, a weighted list of neighboring training examples is created according to how many times the example fell in the same leaf as in the test example. For regression forests, the prediction is the average outcome of the test example's neighbors. In causal prediction, the treatment effect is calculated using the outcomes and treatment status of the neighbor examples.

Another important difference between classic RF and GRF approach is in the way to perform the splits during training. In a classic random forest, a single sub-sample is used both to choose a split and to make predictions. In contrast, GRF randomly splits a sub-sample in half, and use only the first half when performing splitting and the second half to predict. That is known as honest forest. The motivation behind honesty is to reduce bias in tree predictions.

2.4.3 Causal Forest (CF)

The causal forest method proposed by Wager and Athey (2018) is a particular application of the generalized random forest algorithm, and uses the same general training and prediction framework described above.

It supposes that we have n independent and identically distributed feature vectors $X_i \in [0,1]^{\mathrm{d}}$ with $i = 1, ..., n$, an outcome variable $Y_i \in \mathbb{R}$, and a treatment indicator $W_i \in \{0, 1\}$.

$$Y_i = Y_i(W_i) = \begin{cases} Y_i(1) & \text{if } W_i{=}1 \\ Y_i(0) & \text{other case.} \end{cases} \tag{2.1}$$

Where $Y_i(1)$ and $Y_i(0)$ are the potential outcomes, $W = 1$ when the treatment is assigned, and $W = 0$ when is not. A necessary condition here is that treatment W_i was randomly assigned to the population.

We try to predict the average treatment effect (ATE):

$$E[\tau_i] = E[Y_i(1) - Y_i(0)]. \tag{2.2}$$

Or, if individuals have fixed attributes X_i, we could estimate the conditional average treatment effect (CATE):

$$E[\tau_i|X_i = x] = E[Y_i(1)|X_i = x] - E[Y_i(0)|X_i = x]. \tag{2.3}$$

The challenge here is that we do not see both outcomes at the same time. A standard way to make progress is to assume unconfoundedness (Rosenbaum and Rubin, 1983), that is, the treatment assignment W_i is independent of the potential outcomes for Y_i, conditional on X_i, i.e.

$$Y_i(0), Y_i(1) \perp W_i|X_i. \tag{2.4}$$

Assuming unconfoundedness, we can treat nearby observations in X-space as coming from a randomized experiment. In a decision tree, the closest points to X are those that fall in the same leaf L as it does. Then, we can estimate the treatment effect for any $X \in L$ as:

$$\hat{\tau}(x) = \frac{1}{|\{i : W_i = 1, X_i \in L\}|} \sum_{\{i:W_i=1, X_i \in L\}} Y_i.$$

$$- \frac{1}{|\{i : W_i = 0, X_i \in L\}|} \sum_{\{i:W_i=0, X_i \in L\}} Y_i. \tag{2.5}$$

Finally, given a procedure for generating a single causal tree, a causal forest generates an ensemble of B such trees, each of which outputs an estimate $\hat{\tau}_b(x)$. The forest then aggregates their predictions by averaging them: $\hat{\tau}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{\tau}_b(x)$. Wager and Athey (2018) show that causal forest method are point-wise consistent for the true treatment effect, and have an asymptotically Gaussian and centered sampling distribution. They proposed a method for constructing asymptotic confidence intervals for the true treatment effect, that are centered at the causal forest estimates. This results requires that the individual trees satisfy a strong condition: honesty. In a regression forest, a tree is honest if, for each training example i, it only uses the response Y_i to estimate the within-leaf treatment effect τ, using (2.5) or deciding where to place the splits, but not both (Wager and Athey, 2018). In a causal forest, a tree is not allowed to look at the responses Y_i when making splits, but can look at the treatment assignments W_i. Meanwhile, a regular causal tree must have at least k examples from both treatment classes in each leaf. Wager and Athey (2018) proposed two algorithms that satisfy this condition: Double Sample Trees and Propensity Trees. In this research, we are using Double Sample Trees algorithm, but we invite the reader to review the work of Wager and Athey (2018) for more details of the second one.

Algorithm 2 Double-Sample Trees for Causal Trees (Wager and Athey, 2018)

Double-sample trees split the available training data into two parts: one half for estimating the desired response inside each leaf, and another half for placing splits.

Require: n training examples of the form (X_i, Y_i, W_i), where X_i are features, Y_i is the response, and W_i is the treatment assignment. A minimum leaf size k.

1. Draw a random sub-sample of size s from $1, ..., n$ without replacement, and then divide it into two disjoint sets of size $|I| = s/2$ and $|J| = s/2$.
2. Grow a tree via recursive partitioning. The splits are chosen using any data from the J sample and X or W observations from the I sample, but without using Y observations from the I sample.
3. Estimate leaf-wise responses using only the I sample observations.

Double-sample causal trees estimate $\hat{(}\tau(x))$ using (2.5) on the I sample. The splits of the tree are chosen by maximizing the variance of $\hat{(}\tau(x))$ for $i \in J$. In addition, each leaf of the tree must contain k or more I sample observations of each treatment class.

2.5 Observational Data

Data comes from a Chilean multinational trade that owns department stores, supermarkets, home improvement stores, insurance agencies, banking, and travel agencies. In this study, we are using all the Chilean supermarket sales records from credit card holders customers, because these purchases are accurately registered. We select supermarket sales data because these transactions represent more frequent and regular purchases, allowing us to identify changes in consumption behavior better than with sporadic purchases, such as home store or department store purchases. We also use sales records for the the same supermarket customers, in different external business with their credit cards.

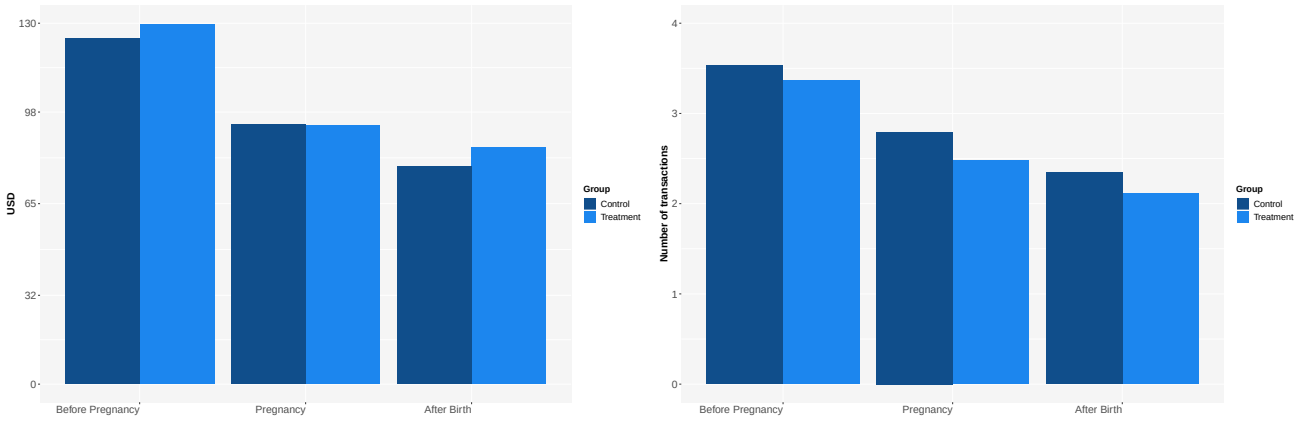
We follow the purchases of all those customers who were parents during 2015 and 2016. We assume that all cases correspond to term pregnancies, that is, 9 months pregnancy periods. We observe their purchase records during the 9 months previous pregnancy, the 9 pregnancy months, and the 9 after birth months, completing a 27 months period for each customer. Specifically, from July 2013 to September 2017 for all the customers identified in the data set.

Customers who had a children during the 2015 and 2016 periods, were identified from a civil records database that the company acquired from a public institution of the country. These records included an Id, date of birth, marital status, date of marriage, among other customers personal information.

Approximately 155 million sale records from 1.7 millions of customers were compiled. After conduct a data outlier elimination process, based on the number of transactions and the total expenditure during the analysis period, we could identify about 1.6 millions active customers. Among those, 122,096 had a child during 2015 or 2016, and 1,481,429 were not parents during the analysis period. We also made a previous filter data process that allowed us to measure the treatment effect on the target population of our study, that is, mothers who have their first child. In addition, we request a minimum of 1 purchase during the 9 months prior to

the pregnancy period. Finally, after these 3 filters, we identify 8,440 first-time mothers (final treatment group), and 126,429 women who were not mothers during the analysis period (final control group).

In Figure 2.1 we can inspect some behavior examples of both treatment and control groups, in equivalent periods of time (as we explain later). For example, figure 2.1a shows that in the pre-pregnancy period, the treatment group spend, on average, more than control group. This difference narrows during the pregnancy period, and increases during the after birth period. On the other hand, in figure 2.1b we can see that the control group makes, on average, more purchases than the treatment group during any period of time. It is interesting to note that, in general, women who have children make fewer purchases but spend more money on each of them. We could explain this behavior as a result of an optimization of the time dedicated to supermarket purchases.



(a) Total Expenditure.

(b) Total Transactions.

Figure 2.1: Monthly average information (unbalanced data).

One of the main issues that we faced with the data was related to defining the analysis time periods for the control group. The treatment group has fixed analysis time periods, determined by the birth date of the baby, which is a known fact. Assuming term pregnancies (9 months), we can clearly identify the 9 months before pregnancy, the 9 months of pregnancy, and the 9 months after birth. However, in the case of the control group, these periods of time are dynamic, because it depends on the treatment customer we are analyzing.

In order to solve this issue, we split the treatment group into 24 subgroups, depending on the month of birth of the baby. For example, the first group corresponds to all mothers who gave birth in January 2015, the second to those who did it in February 2015, the third in March, and so on. Thus, for each customer of the control group, we add their monthly consumption information in equivalent periods of time to the treatment groups. For example, in the first group, the aggregated information of the control group was calculated from January 2015, considering 9 months prior (April 2014 - December 2014) equivalent to the pregnancy period of the treatment group, the subsequent 9 months (February 2015 - October 2015) equivalent to the baby's post-birth period, and also the 9 months prior to equivalent pregnancy period (July 2013 - March 2014).

Finally, we form 24 groups of customers. Each group has different IDs of treatment, but the IDs of the control clients can be repeated in different groups. Each control customer has temporarily different aggregated information in each group, depending on the month of birth of the children of the respective treatment group. Table 2.1 shows the number of treatment and control population per analysis group.

Group	**Treatment**	**Control**	%	**Group**	**Treatment**	**Control**	%
1	342	56260	0.61	13	388	64579	0.60
2	297	57613	0.52	14	325	64154	0.51
3	361	58201	0.62	15	369	63769	0.58
4	347	58682	0.59	16	327	63816	0.51
5	322	59896	0.54	17	372	64244	0.58
6	318	61516	0.52	18	342	65035	0.53
7	357	60478	0.59	19	375	62982	0.60
8	356	61032	0.58	20	347	64054	0.54
9	366	61612	0.59	21	389	64983	0.60
10	363	63962	0.57	22	359	68083	0.53
11	353	63814	0.55	23	341	68808	0.50
12	368	63973	0.58	24	356	69191	0.51

Total Treatment=8440 Total Control=1510737 Percentage = 0.56

Table 2.1: Number of treatment and control individuals per group.

2.6 Procedure

As can we see in Table 2.1, our study is characterized by an important data imbalance between the control and the treatment group. When considering the 24 analysis groups, only a 0.56 % of the data belongs to the treatment group (1:178). This could represent a difficulty for learning algorithms, as they will be biased towards the majority group (Krawczyk, 2016), mainly because they aim to minimize the overall error rate instead of paying attention in positive examples class (Chen et al., 2004). To alleviate this problem, three main approaches have been proposed in the literature: data-level methods, that modify the examples to balance distributions; algorithms-level methods, which modify the existing learning algorithm to alleviate the bias; and hybrid methods that combine the two previous approaches (Krawczyk, 2016). Without a reasonable balance of treated and control examples, there will not be enough information in the node to obtain a good estimate of treatment effect. In the worst case, we could end up with nodes composed entirely of control (or treatment) examples. A recent research has suggested the estimation of the nuisance parameters, using Local Linear Forests to solve instances of unbalanced data and noise (Turjeman and Feinberg, 2019).

The GRF R-package includes a balance parameter denominated "min.node.size" which, as we will show later, is responsible for balancing the data of both groups on each node.

In order to measure the impact of the unbalanced data in the treatment effect estimations, we address the problem from two perspectives. First, we run the algorithm considering all

individuals in the control group controlling the imbalance problem through the tuning of balance parameters. Secondly, we applied a data-level approach to balance both groups. The strategy used in the second perspective was to make a matching propensity score in each of the 24 groups of individuals to choose a control group in a 1:1 ratio with respect to the treatment group.

2.6.1 Data Balancing

In order to balance the data in an equivalent ratio between the control and treatment groups, we applied a matching propensity score method.

We looked for a similar control group in terms of consumption and sociodemographic characteristics to the treatment group, during the pre-pregnancy period. This equivalent control group would be also valid to apply the differences in differences methodology that we use later to validate consistency of the method.

The propensity score, $\mathrm{e}(x)$, is the conditional probability of exposure given the covariates.

$$\mathrm{e}(x) = Pr(z = 1|x).$$

Where $z = 1$ if the individual was exposed to treatment, and $z = 0$ if was not.

We estimated the propensity score using a logistic regression model:

$$q(x) = \alpha + \beta f(x).$$

Where α and β are parameters to be estimated by maximum likelihood estimation (MLE), and $q(x)$ is the log odds against exposure. A linear function $f(x)$ was assumed in the covariates. We used 84 covariates in total: 3 sociodemographic and 81 transactional. Before matching, a covariates correlation analysis was made in order to avoid highly correlated variables in the matching. As a result, approximately 15% of initial covariates were excluded.

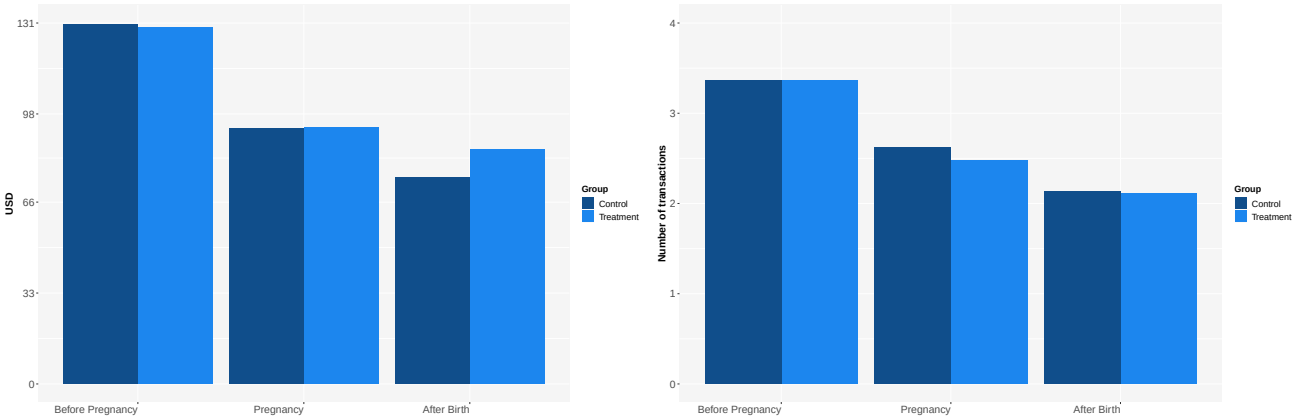
Propensity score matching is started with randomly ordering the subjects in the treatment group. After that, matches the first treated subject with an untreated subject which has the nearest neighbor linear propensity score. After matched, both of the subjects would be taken out of the pool and would not rejoin the process. The previous steps are repeated for all the treated subjects until all of them find the matched untreated subjects.

To make the analysis, we used the "MatchIt" R-package with default settings, that is, "distance = logistic regression", and "method = nearest neighbor matching".

We finally identified 8440 control customers with the nearest propensity score to the treatment group, of which, 4% is duplicated in some of the 24 analysis groups. Nevertheless, since they have chronologically different information, they were considered as independent subjects.

As we did in the previous analysis with the unbalanced data, we can inspect some behavior examples of both treatment and control groups, in equivalent periods of time. See figure 2.2. For example, figure 2.2a shows that in the pre-pregnancy period the matching control group spends, on average, the same as the treatment group. This occurs because this variable was

a covariate of the matching propensity score model. Instead, during the pregnancy and the afterbirth period, treatment group spends more than the matching control group. A similar patron is observable in 2.2b, where the control group makes, on average, more purchases than the treatment group during the pregnancy and the afterbirth period.



(a) Total Expenditure.

(b) Total Transactions.

Figure 2.2: Monthly average information (balanced data).

Tables 2.3 and 2.4 in Appendix A show a $\chi^2 - Test$ for categorical covariates and a $t - Test$ for numerical covariates, before and after matching respectively. Except for two and as expected, all the covariates do not have differences in means after the matching propensity score procedure.

2.6.2 Product Categories Selection

We selected 104 product categories to be analyzed. Table 2.6 shows, in their first column, all categories classified in 11 groups according to their type. A 16% (17 categories) corresponds to external business categories made with the holding credit card, labeled as (CC). A 7% (7 categories) corresponds to general consumption behavior, and a 77% (80 categories) to expenditure in supermarket categories exclusively.

2.6.3 Causal Forest Implementation

Causal Forest Algorithm was implemented using the "grf" R-package [2], and the "causal_forest" function. The most important parameters are the following:

- **num.trees:** Number of trees grown in the forest. We used 500 trees in all our experimental analysis.
- **min.node.size:** Each node must contain at least min.node.size of treated samples, and also at least min.node.size of control samples. To ensure this condition, the algorithm computes the average treatment values on the parent node. Then, the next split should consider each child node with "min.node.size" samples with treatment value less than the average, and at least the same number of samples with treatment value greater than or equal to the average. Default value is 5.

[2]https://github.com/grf-labs/grf

- **mtry:** The mtry parameter determines the number of variables considered during each split. By default, mtry is $min(\sqrt{p} + 20, p)$, where p is the number of variables in the dataset.
- **alpha:** The alpha parameter controls the maximum imbalance of a split. In particular, when splitting a parent node, the size of each child node is not allowed to be less than $size(parent) * alpha$. Its value must lie between (0, 0.25), and by default is 0.05.
- **imbalance.penalty:** The imbalance.penalty parameter controls how harshly imbalanced splits are penalized. When determining which variable to split on, each split is assigned a "goodness measure" related to how much it increases heterogeneity across the child nodes. The algorithm applies a penalty to this value in order to discourage child nodes from having very different sizes, specified by $imbalance.penalty * (1.0/size(left.child) + 1.0/size(right.child)$. This penalty can be seen as a complement to the hard restriction on splits provided by alpha.
- **Y.hat:** Is a nuisance parameter. Estimates of the expected responses $E[Y|X_i]$, marginalizing over treatment. If Y.hat = NULL, these are estimated using a separate regression forest. Default value is NULL. We used default values in every analysis.
- **W.hat:** Is a nuisance parameters. Estimates of the treatment propensities $E[W|X_i]$. If W.hat = NULL, these are estimated using a separate regression forest. Default is NULL. We used default values in every analysis.

Except for the number of trees and the nuisance parameters, we activated the tuning parameter option for balanced and unbalanced data analysis.

The tuning parameter option in the GRF R-package provides a cross-validation procedure to select the values of training parameters. The cross-validation procedure draws a number of random points in the space of possible parameter values. By default, 100 distinct sets of parameter values are chosen. For each set of parameter values, trains a forest with these values and compute the out-of-bag error. For tuning procedure to be computationally tractable, it only trains forest composed of 50 trees. In order to minimize the bias generated in the out-of-bag error, a simple variance decomposition is made. Finally, given the debiased error estimates for each set of parameters, the algorithm applies a smoothing function to determine the optimal parameter values. The optimal parameters are the ones minimizing the predicted smoothed error on a new random draw of possible parameter values.

For each of the 104 categories analyzed, a causal forest model was run considering the mentioned parameters. The covariates in all models corresponds to three types: the expenditure in certain product categories, some variables of global consumption behavior, and also sociodemographic characteristics of people, considering the 9 months prior to the pregnancy period. All covariates are described in the table 2.5, in the Appendix B.

In this work, we are estimating the heterogeneous treatment effect for the difference between the expenditure in a particular category during the treatment period, and the expenditure in the same category before the pregnancy period. What we are looking for is to compare our results with other methodologies like Differences in Differences.

In order to compare the treatment effect between both, the causal forest and the differences

in differences models, we estimate the average treatment effect across the population. In the causal forest model, we might average personalized treatment effects across training examples, however, a more accurate estimate can be obtained by plugging causal forest predictions into a doubly robust average treatment effect estimator. The GRF R-package provides the function "average_treatment_effect" to compute these estimations. This function implements two types of doubly robust average treatment effect estimations: augmented inverse-propensity weighting (Robins et al., 1994), and targeted maximum likelihood estimation (van der Laan and Rubin, 2006).

2.7 Results

To compare the results of our analysis, both with balanced and unbalanced data, we use a classical methodology in causal analysis: Differences in Differences.

For each of the 104 categories of products analyzed, we make a lineal regression of the form:

$$Y_i \sim \beta_0 + \beta_1 * parent_i + \beta_2 * period + \beta_3 * parent_i * period + \varepsilon \tag{2.6}$$

Where:
Y_i: Outcome variable.
$parent_i = \begin{cases} 1 & \text{if the individual had a child during the analysis period} \\ 0 & \text{other case} \end{cases}$
$period = \begin{cases} 1 & \text{if the variable } Y_i \text{ was measured during pregnancy/birth period} \\ 0 & \text{if the variable } Y_i \text{ was measured during the pre-pregnancy period} \end{cases}$
β_1: Treatment group specific effect.
β_2: Time trend, equal to control and treatment groups.
β_3: Treatment Effect.
β_3 is the differences in differences (DID) estimator.

The conventional DID estimator requires that, in absence of the treatment, the average outcomes for treated and controls would have followed parallel paths over time. This assumption may be implausible if pre-treatment characteristics, which are thought to be associated with the dynamics of the outcome variable, are unbalanced between the treated and the untreated group (Abadie, 2005). For this reason, we use the same control group identified with the propensity score for data balancing. This procedure ensures that the average value of the outcome variable is statistically equal between both groups during the 9 months prior to pregnancy.

Since we are working with observational data, we do not have a measure to evaluate the performance of these three models for the treatment effect prediction. Therefore, we focus our analysis on comparing the variance, statistical significance and the magnitude of the effects estimated by each one. The results of the three models, for the 104 categories analyzed, can be seen in table 2.6 in the Appendix C.

The average treatment effect of the five most affected categories by pregnancy and birth of first child are shown in figures 2.3a and 2.5a.

The figures 2.3a and 2.5a show that the DID estimator has the largest variance. This feature implies that this method predicts a smaller number of statistically significant categories affected by the treatment. On the other hand, the causal forest model with unbalanced data and calibrated tuning parameters have the lowest variance, which is why it is also the one that predicts a greater number of statistically significant categories affected by the treatment (see table 2.2). The bar graphs of the remaining categories, in which a statistically significant treatment effect was identified for at least one of the three models, can be found in Appendix C (see figures 2.7 and 2.8).

Model	**Treatment**	
	Birth	Pregnancy
DID	14%	27%
CF_PS	23%	31%
CF_ALL	44%	48%

Table 2.2: Percentage of statistically significant average treatment effect estimated

Regarding the magnitude of the effect, we performed a regression analysis to evaluate the contribution of the estimation model in the magnitude of the average treatment effect, controlling by the product category and the treatment. The results in table 2.7 in Appendix D show that no statistically significant differences exist between these three models.

Now, we analyze separately the most interesting changes in consumer behavior during the pregnancy and after birth period.

2.7.1 Treatment: Birth

There is a negative treatment effect on the number of purchases (see "total transaction" category on figure 2.3b). But there is also a positive treatment effect in the total expenditure category (see figure 2.3a). This could reflect an attempt by first-time mothers to minimize the time they spend on supermarket, by decreasing the frequency of purchases and increasing the basket size. This behavior is also supported by a negative effect on gasoline and transportation services, which could indicate that first-time mothers not only reduce the frequency of the supermarket visits, but also reduce their activities outside home (see figure 2.3b).

Another important behavior change that is worth noting is the negative treatment effect in all entertainment categories, such as travels, restaurants, entertainment [3], party stuff, alcoholic drinks, and cocktail food (see figure 2.3c). This reflects important priority changes in first-time mothers, as they significantly reduce their spending on leisure time and self-satisfaction activities, to give way to a significant increase on goods and services of common wellness for her and her child. For instance, there is a positive and significant effect in all baby products

[3]Entertainment external business category include: Computer and software stores, Bars/taverns/lounges/discos, News dealers/newsstands , Sporting/recreational camps, Beauty/barber shops, Motion picture theatres, Betting/track/casino/lottery and Recreation services

categories, with magnitudes from highest to lowest as follows: baby clothing, baby care [4], toys, baby food, baby shoes and baby bedroom (See figure 2.3d).

We can also observe a surprisingly positive treatment effect in photography and Christmas categories. This behavior could be due to the fact that the baby is the first child and the mothers are more willing to invest in family memories for later life, and also in photographs and celebrating Christmas with their own family for the first time.

The women's personal care after having their first child is another aspect that we would like to analyze. We were able to confirm that there is a negative effect on categories such as personal hygiene, [5] and care and beauty [6] products (see figure 2.4a). There is also a negative treatment effect in women's clothing categories, such as woman shoes, and underwear (see figure 2.4a). Although the supermarket is not the main source of clothing and footwear supply for customers, we can think that this behavior is widespread, because there is also a negative treatment effect on the department store spending made with the holding credit card (see figure 2.3a). Among those are precisely clothing stores, shoe stores, jewelry, and others. This behavior could indicate a postponement of the mother's care, beauty and personal presentation in pursuit of her new motherhood role. It would be interesting to compare this feminine behavior with the masculine one, in order to be able to draw more detailed conclusions and inquire whether it is due to individual postponement or simply to a family saving category.

Another aspect that we are interested in evaluating in this study is the preference change for brands in some crucial products. Our hypothesis is that mothers begin to try and incorporate better quality products into their usual consumption, an attribute that can be seen represented by the brand in most cases. Although the SKU-level study is proposed in future work, we have now a preliminary result evaluating the treatment effect in "store brand" products. The supermarket owns a store brand in most product categories analyzed, and the common characteristics between all of them are the lower prices and the standard quality within their category. We estimate the treatment effect in the percentage of store brand products with respect to the total products purchased monthly, and also, we measure the percentage of expenditure on store brand products with respect to the total monthly expenditure (see "% Trx Store Brands Products" and "% $ Store Brands Products" on figure 2.4b). The results indicate that there is a negative effect on both variables, that is, first-time mothers reduce spending and consumption of store brand products, which could indicate that they privilege quality instead of low prices. The magnitude of this effect could be attenuated by another type of behavior that mothers develop when a new member of the family arrives, and in most cases goes in opposition to maximizing quality: saving. The saving attempts can be verified since there is a positive treatment effect in the percentage of expenditure on discount products with respect to the total monthly expenditure (see "% $ Discount Products" in figure 2.4b). Therefore, mothers face a significant trade off: saving versus getting better quality products.

Categories evidently affected by the birth of a child are health services and medications.

[4]Baby care products: diapers, blowers, breast pumps, pacifiers, cotton swabs, shampoo, balm, talcum powder, creams, oils, colonies, wet towels, liquid soaps, bath gels

[5]Personal hygiene: feminine protection, soaps, shampoo, medical kit items, oral hygiene, conditioner, deodorants, talcum powder.

[6]Care and beauty: Hair removal, skincare, cosmetics, sun protection, colognes, perfumes, hair care

Figure 2.4c shows a significant positive treatment effect in health and pharmacies categories. Both correspond to the expenditure in medical establishments and drug stores made with the holding credit card. We have no information on whether the increase in spending in these categories corresponds to the medical care of the baby or the mother, but we can assume that the first months after the baby's born are the most critical and stressful for the new motherhood role, because for the very first time, she is absolutely in care of a person. For this reason, mothers do not skimp on expenses when facing a child's health problem. It would also be interesting to compare the magnitude of the effect with mothers having a second (or more) baby, to learn if there are significant differences between first-time motherhood and the next motherhood experiences.

In the same context of care and protection of the baby, we can highlight that there is also a positive treatment effect in the insurance category (see figure 2.4c). This behavior shows that first-time mothers have a special need for protection against eventualities, such as life, home or car insurance. A curious effect, but we think that it is also associated with the protection of the health and integrity of the baby, is that there is a negative treatment effect in pet products (see figure 2.4c), which could show that pets are considered a potential danger to the baby, and in many cases, are given up for adoption.

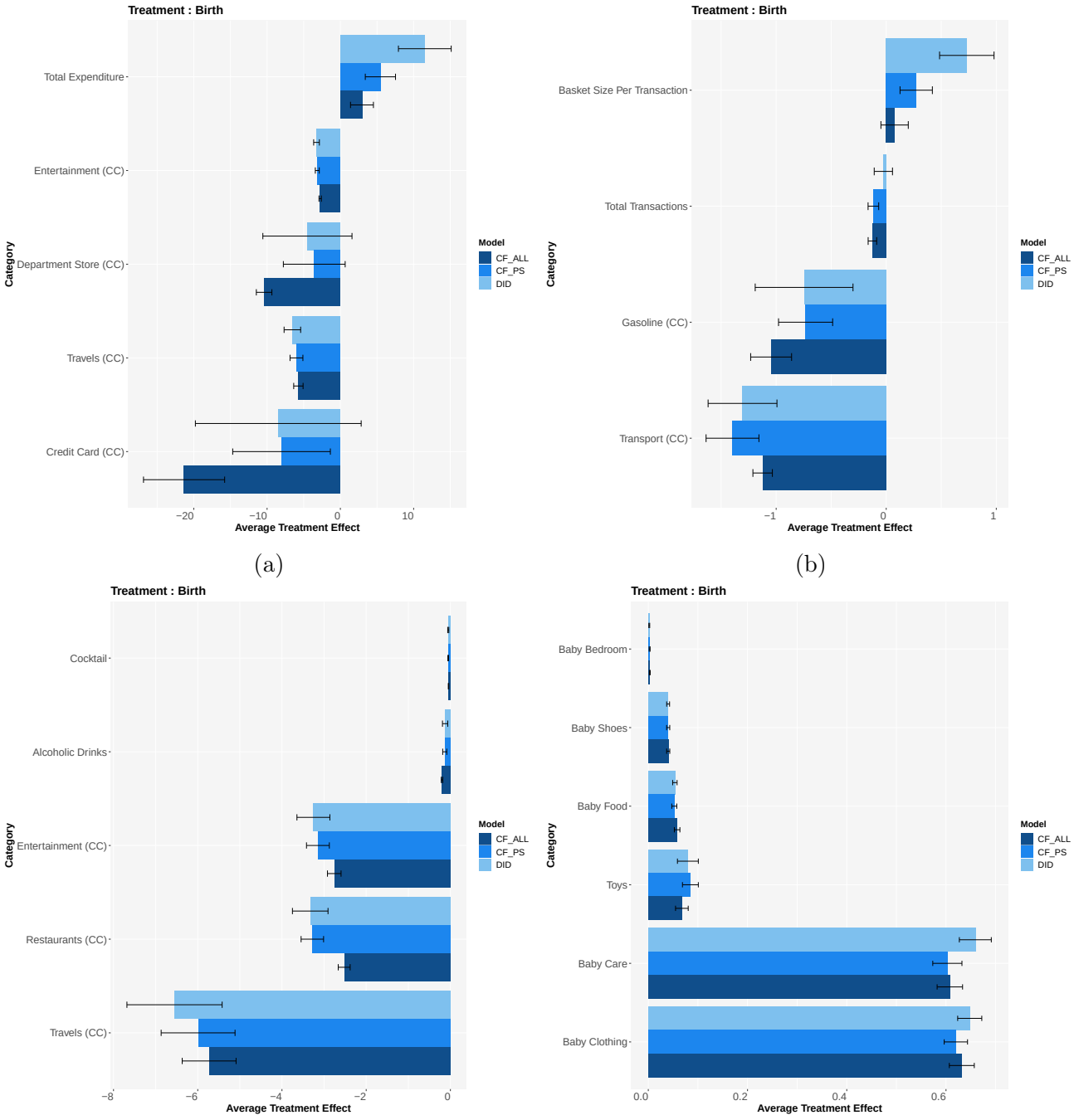
Finally, some surprising effects on food categories caught our attention. For example, there is a negative treatment effect in ready-to-eat products, canned food and refrigerated mass and pasta, the deli category and in cheese category (see figure 2.4d). This may be due to breastfeeding mothers taking care of their diet, preferring to prepare and eat fresh food instead of packaged or sausages. However, we are struck that there is no significant positive treatment effect in fresh and healthy foods categories, such as fruits and vegetables, chicken, turkey, fish, eggs or milk. We think that this behavior could be driven by the fact that mothers favor the purchase of fresh and healthy food in other establishments not documented in this study, such as vegetable markets, organic markets, butchers or farms.

2.7.2 Treatment: Pregnancy

Unlike the birth of the baby, during pregnancy, a negative treatment effect was estimated in the total monthly expenditure in the supermarket under study, as well as in the total monthly expenditure made with the holding credit card in the supermarket category (see "total expenditure" in the figure 2.5a). This total spending decrease could be caused by two reasons: the first one is that mothers delegate the shopping task on other people and that is why is not reflected in their transactions, and the second one is that, given the new family member arrival news, mothers begin to save.

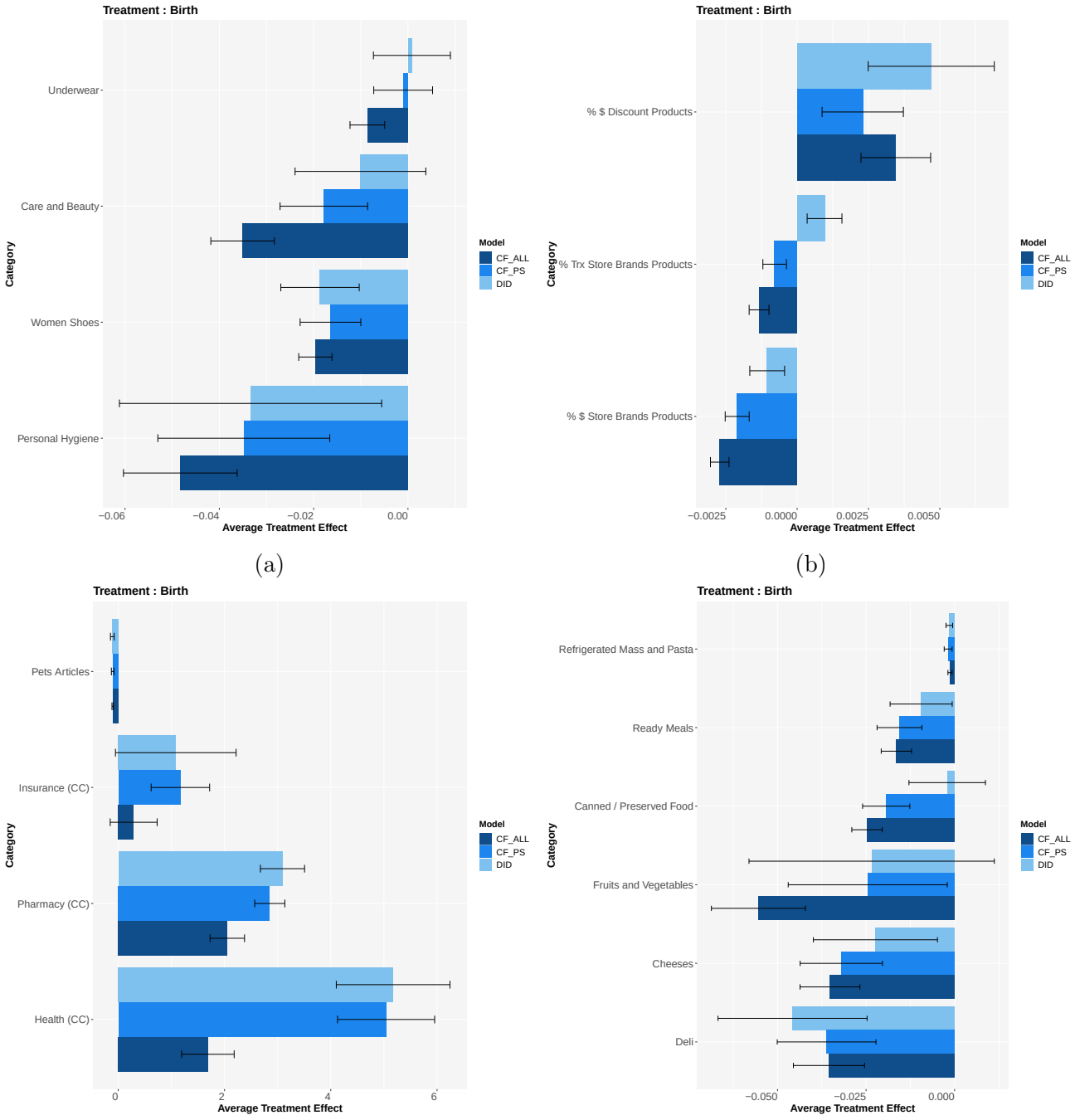
As in the birth of the baby, during pregnancy, there is also a negative treatment effect on the number of monthly purchases in the supermarket and a positive treatment effect on the basket size , which added to a negative treatment effect in transport and gasoline categories (see figure 2.5b), allow us to infer that there is an obvious stagnation of outside home activities of new mothers during pregnancy, including supermarket purchases.

Entertainment-related categories, such as travel, restaurants, entertainment, alcohol, and cocktails products, also have a negative treatment effect (see figure 2.5c), but the magnitude is



Note: The bar represents the magnitude and the line represents the standard deviation of the average treatment effect.

Figure 2.3: Average treatment effect of birth comparison.



Note: The bar represents the magnitude and the line represents the standard deviation of the average treatment effect.

Figure 2.4: Average treatment effect of birth comparison.

lower during pregnancy than the post-birth period. This could indicate that mothers continue having recreational and leisure activities during the first months of pregnancy, and then, the average monthly effect of the 9 months period of pregnancy is lower than the average monthly effect of the 9 months after the birth, where the cessation of leisure activities is not gradual. Similarly, all categories of baby products have a positive treatment effect (see figure 2.5d), but with smaller magnitudes during pregnancy than the post-birth period. This situation could be due to the fact that some mothers are preparing for this important life event, and they anticipate their purchases both in clothes, shoes, baby care, and baby bedroom products.

The women's personal care during the pregnancy period is also negatively affected by the treatment (see figure 2.6a). The magnitude of the treatment effect in the care and beauty category is lower during pregnancy than the baby's post-birth. Conversely, in personal hygiene category, the treatment effect is larger during pregnancy than the post-birth period. This is mainly explained because the sanitary napkins, which are part of this category, have an abrupt fall in sales during the pregnancy, and the purchases of this product increase again during the period after the baby is born, in which women recover their menstrual cycles. Also, there is a negative treatment effect in women's clothing category and in underwear products (see figure 2.6a). Although the supermarket is not the main source of clothing and footwear supply for customers, we can think that this behavior is widespread, because there is also a negative treatment effect on the department store spending made with the holding credit card (see figure 2.5a). Among those are precisely clothing stores, shoe stores, jewelry, and others. As in the previous case, the magnitude of this negative treatment effect is lower during pregnancy. This can also be explained in part because women do invest in maternal clothing and beauty products during the first months of pregnancy, which softens the average monthly effect of the period, compared to the period after the baby is born, in which there is a strong decrease in purchases of these products.

As we mentioned before, since we have the hypothesis that first-time mothers tend to substitute lower quality products for higher quality, an attribute that in many cases is directly linked with the brand, we are interested in evaluating changes in preferences for certain brands. At this point, we have a general approximation of this result, through the analysis of the "store brand products". The store brand products have common characteristics between all categories, such as lower prices and standard quality within their category.

There is a negative treatment effect in both the percentage of store brand products with respect to the monthly total products purchased, and the percentage of expenditure on store brand products with respect to the total monthly expenditure (see "% Trx Store Brands Products" and "% $ Store Brans Products" on figure 2.6b). Then, as during the post-birth period, during the pregnancy period first-time mothers also reduce spending and consumption of store brand products, which could indicate that they privilege quality instead of low prices. As we indicated earlier, the magnitude of this effect could be attenuated by the saving attempts of mothers. In this case, and unlike birth, there is no support for this hypothesis because there is a negative treatment effect also in the categories related to sales products purchases (see "% $ Discount Products" and "% Trx Discount Products" in figure 2.6b). We believe that this result is directly correlated with the general decrease in total expenditure on supermarket products, and is not necessarily due to a real tend to stop buying sale products.

A category evidently affected by the pregnancy is health services. Figure 2.6c shows a significant positive treatment effect in health category. This category correspond to the expenditure in medical establishments made with the holding credit card. This is not a surprising result, since it is evident that the need for medical attention increases during the period of pregnancy. Unlike the post-birth period, during pregnancy, there is a negative treatment effect in the pharmacy category. This can be explained in several ways: first of all, if the future mother is health ailing and requires medicines, she is very likely to delegate the purchase to other people, and secondly, there is also a significant decrease in the purchases of contraceptive products and hygiene feminine protection.

Finally, there are some differences between the birth and the pregnancy treatment effects in food product categories. For instance, there is a positive treatment effect in dairy products like milk, creams and yogurt (see figure 2.6d). This effect is not surprising, because in general during pregnancy it is recommended to increase the consumption of products rich in calcium. However, an unexpected effect is the positive treatment effect in the sweet desserts packaged category [7], mainly because, in general the recommendation is to reduce the consumption of sugar and carbohydrates during the pregnancy period. On the other hand, we can confirm that this behavior is not a trend, because a negative treatment effect was estimated in the categories of confectionery [8], and sweet breakfast [9].

As at birth, during pregnancy there is a negative treatment effect in ready-to-eat food, in refrigerated mass and pasta, and in cheese categories (see figure 2.6d). This may be due to future mothers taking care of their diet, preferring to prepare and eat fresh food, instead of packaged or sausages.

2.8 Conclusions and Discussion

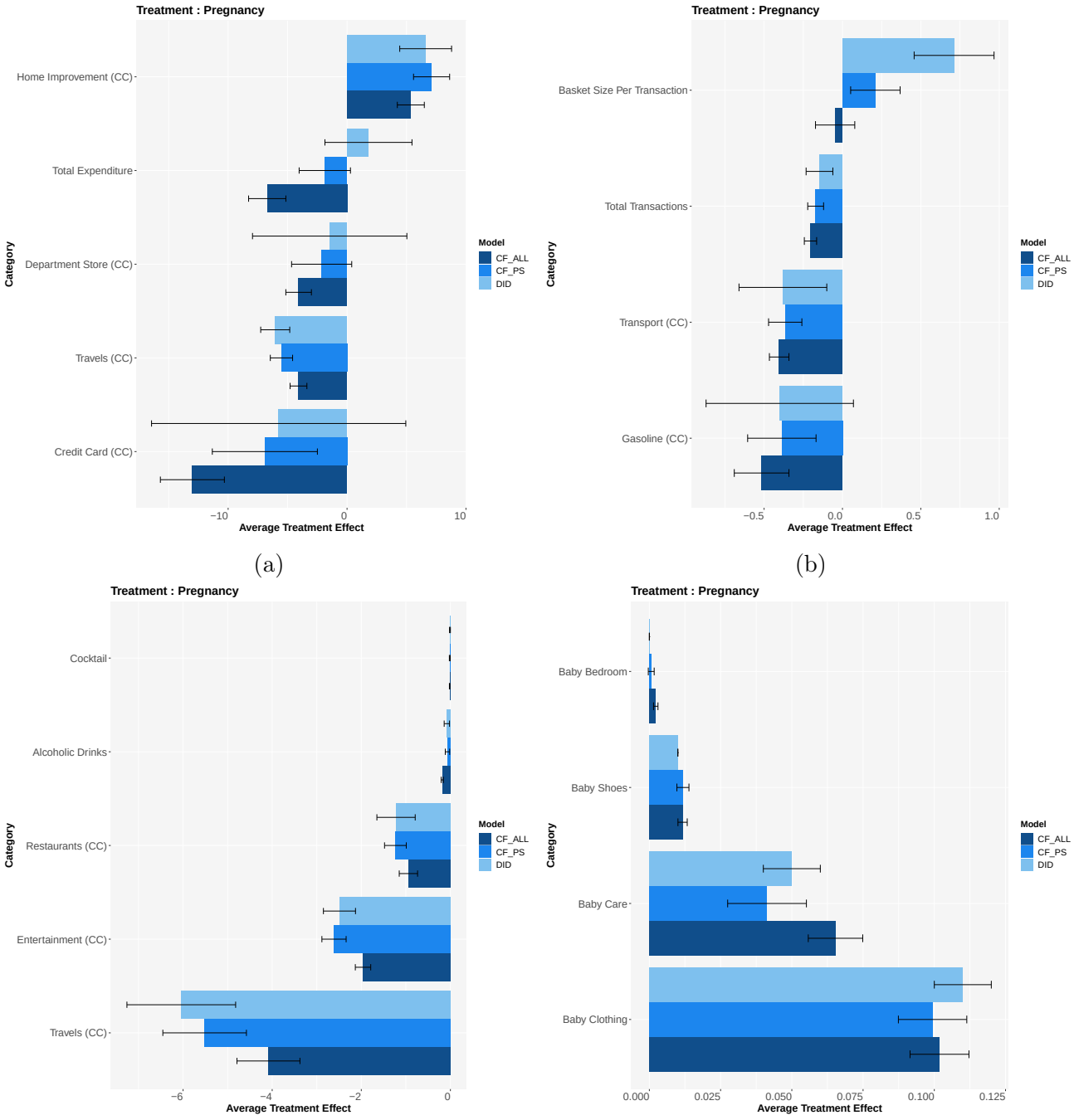
The first baby announcement, which involves the pregnancy and birth, is one of the most important life events, and therefore, it may encourage people to carry out a large number of change adaptation activities to minimize the stress, and partially alleviate the new tasks. These adaptation activities may include consumption activities. In this article, we focused on how the consumption behavior of the first-time mothers is affected both during the pregnancy and the afterbirth periods.

We used observational data from a supermarket that belongs to a Chilean multinational trade. We identified all purchases records of 8,440 first-time mothers (used as final treatment group), and 126,429 women who were not mothers during the analyzed period (used as final control group). We considered their purchases during the 9 months previous to pregnancy, the 9 pregnancy months (assuming full-term pregnancies), and the 9 months after birth; completing a 27 months period for each customer. We splitted the treatment group into 24 subgroups, depending on the month of birth of the baby (from January 2015 to December 2016), and then,

[7] Sweet desserts packaged category includes: *crème bavaroise*, *manjar* (a spread made of milk and caramelized sugar), semolina and milk, mousses, flan, dairy desserts, homemade *manjar*, diet *manjar*, jelly and fruits compote

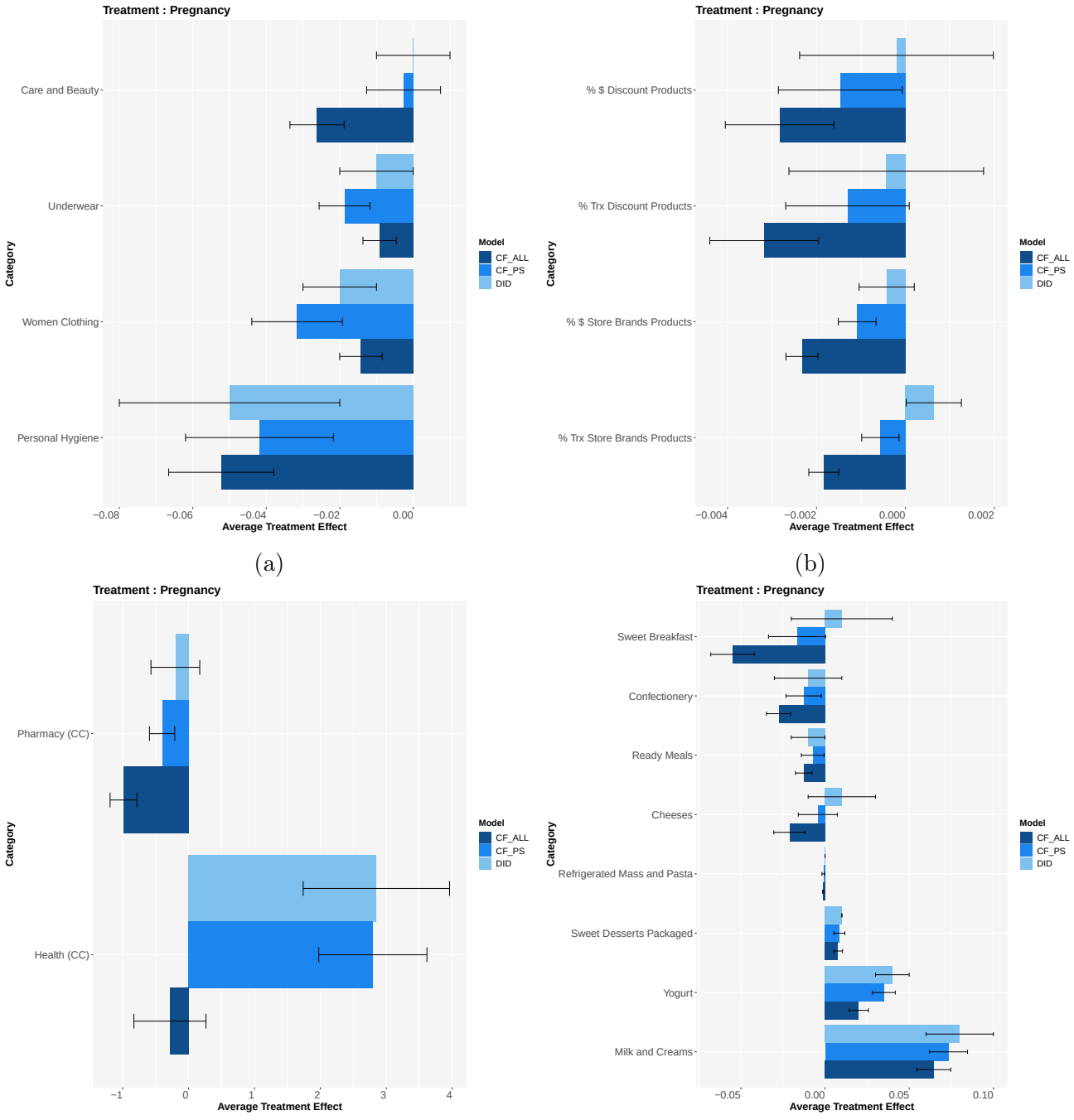
[8] Confectionery category includes: candies, gum, chocolate, cookies

[9] Sweet Breakfast category includes: sugar/substitutes, coffee, cereals, milk, creams, jams, honey, industrial bakery, biscuits, milk flavors, tea, herbs



Note: The bar represents the magnitude and the line represents the standard deviation of the average treatment effect.

Figure 2.5: Average treatment effect of pregnancy comparison.



Note: The bar represents the magnitude and the line represents the standard deviation of the average treatment effect.

Figure 2.6: Average treatment effect of pregnancy comparison.

for each customer in the control group, we monthly added their consumption information in equivalent periods of time to the treatment groups.

Our study was characterized by an important data imbalance between the control and the treatment group in a ratio of 1:178. In order to measure the impact of the unbalanced data in the treatment effect estimations, we addressed the problem using two approaches. First, we ran the causal forest algorithm considering all individuals in the control group (CF_ALL) managing the imbalance problem through the tuning of balance parameters, and secondly, we applied a matching propensity score (CF_PS) approach to balance both groups until a 1:1 ratio. Accordingly, we estimated 104 causal models, one for each analyzed product category. To compare the results of our analysis, both with balanced and unbalanced data, we used the classical Differences in Differences (DID) approach.

To evaluate the performance of these approaches, we compared the mean and the variance of the average treatment effect estimations. The magnitudes of the estimated effects showed to be no statistically different between this three approaches. However, differences in differences (DID) model estimations have the highest variance and the causal forest with unbalance data (CF_ALL) model estimations have the lowest. The CF_ALL approach estimated statistically significant treatment effects in 44 % of the categories during pregnancy, and 48 % of the categories during the post-birth period. Instead, DID approach estimated a statistically significant average treatment effect for only 14 % of the categories by pregnancy treatment, and 27 % of the categories by birth treatment. These results are explained by the high variance for the DID estimations, and the low variance for the CF_ALL estimations, as we indicated above.

When we compared the performance of the causal forest models in terms of resolution time, we obtained that the model with unbalanced data, in which the control group is 178 times larger than the treatment group, takes on average 66 times longer to run than the model with balanced data, where the control group is equal to the treatment group. Due to the strategic level of the decisions that could be made from the results of our study, the computing times are not a relevant variable to compare the performance of the models. For this reason and despite the increase in computing times, we conclude that the causal forest methodology performs well in highly large and unbalanced data scenarios, mainly based on the variance estimations results.

Interestingly, we identified more significant effects generated by the birth of the baby than those generated by the pregnancy. Similarly, the magnitude of these effects is also greater with birth condition. This behavior could be mainly due to the fact that the target population of this study, the mothers, delegate the shopping tasks to others during the pregnancy period, and we have gaps in the sales records.

Regarding the findings of the causal analysis, we showed that both, pregnancy and birth of the first child, motivate important changes in the consumption behavior of the new mothers. First, they postpone the consumption of goods and services oriented to their personal wellness and leisure, and replace it by goods and services for the well-being of the baby and the family. This is supported by the negative effects in most categories related to entertainment, personal care, beauty and women's clothing, and also by the positive effects in most product categories related to baby care, baby clothing and shoes, medical services, medications, and insurance.

We also found negative effects in categories related to the mobility of the mother outside the house, such as the number of monthly visits to the supermarket, as well as in the expenditure of transportation and gasoline services during the pregnancy and the post-birth period. This showed an important reduction of the daily activities of mothers.

Another interesting result is that we found evidence that mothers prefer higher quality products. This hypothesis is supported by a negative effect in the purchased quantity and in the expenditure related to store-brand products, characterized, in general, by having lower prices and standard quality compared with national products of the same category.

The magnitude of this effect could be attenuated by another type of behavior that mothers develop when a new member of the family arrives, and in most cases goes in opposition to maximizing quality: saving. The savings attempts, motivated by an increase in the total expenditure due to the arrival of a new member of the family, can be verified since there is a positive treatment effect in expenditure on discount products. Therefore, mothers face a significant challenge, save and at the same time to get better quality products, which generally have a higher price.

There are some surprising effects on food categories that caught our attention. For instance, we found a negative effect in ready-to-eat products, canned food and refrigerated mass and pasta, in the deli category and in cheese category. This may be due to mothers taking care of their diet, preferring to cook or to eat fresh food instead of packaged or sausages. However, we were surprised that there is no significant positive effect in fresh and healthy foods categories, such as fruits and vegetables, chicken, turkey, fish or eggs. We think that this behavior could be due to the fact that mothers favor the purchase of fresh and healthy food in other establishments not documented in this study, such as vegetable markets, organic markets, butchers or farms. It can also be a consequence of the delegation of the purchase tasks, keeping in mind that generally fruits, vegetables, and meat, are perishable products that must be purchased regularly and we already know that mothers decrease the frequency of their shopping.

2.8.1 Limitation and Future Work

One issue that we detected as a possible source of data bias was the gaps in the mother's sales records during the pregnancy period. Future work in this line is aimed at building household sales records, being able to complement the purchases made by the mother and also the father of the future baby. The problem is that this task is not trivial since the information of the parents is not directly contained in our database and must be inferred.

Our results could be more informative if the analysis were performed with monthly aggregated data, instead of 9 months period aggregated data. This would allow us to evaluate the temporal evolution of the parents behavior during pregnancy. Future work will be aimed at adding the millions of sales records on a monthly basis instead of a 9 month period. Our goal is to identify consumption needs in a timeline, from the first month of pregnancy to the 9 months after the birth. The main challenge here is in the management of large transactional databases.

We are also interested in studying some effects at the SKU level. For instance, to estimate brand preference changes for key products, that would allow us to identify interesting consumer

behaviors such as brand loyalty, willingness to innovate, price sensitivity or quality sensitivity. Therefore, future work is also oriented towards deepening the already carried out general analysis.

Finally, we will extend our analysis in several ways. First, it would be interesting to compare the magnitude of the effects in the case of mothers having a second (or more) baby. This would allow us to learn if there are significant differences between first-time and further motherhood experiences. Secondly, analyzing other life events, such as marriage, home moves, entry in the labor market, children starting at school, among others. The main challenge of this task is that there is no obvious record of the moment in which the life event happened. Therefore, it is necessary to figure out other ways to infer it from the data, or collecting stated information from customers.

2.9 Appendix

2.9.1 Appendix A: Covariates means comparison before and after matching.

Covariates	Value	Before Matching		After Matching	
		Control	Treatment	Control	Treatment
Zone of Chile	Tarapaca	3261	20	19	20
	Antofagasta	63995	351	402	351
	Atacama	13618	79	95	79
	Coquimbo	19261	104	104	104
	Valparaiso	161524	869	867	869
	B. OHiggins	74906	427	449	427
	Maule/Ñuble	60362	377	374	377
	Biobio	166811	977	991	977
	La Araucanía	8140	43	42	43
	Los Lagos	6378	45	35	45
	Aysén	669	2	4	2
	Magallanes	1394	9	6	9
	Metropolitana	968278	5474	5608	5474
	Los Rios	2454	22	20	22
	Arica	1698	5	6	5
	p_value	0.207		0.856	
Marital Status	Married	246293	2634	2688	2634
	Single	1305994	6170	6334	6170
	p_value	**0.000***		0.923	

Table 2.3: χ^2 Test for categorical covariates before and after matching

Covariates	Mean Before Matching			Mean After Matching		
	Control	Treat	p_value	Control	Treat	p_value
Age	37.63	28.60	**0.00***	28.37	28.60	**0.01***
Total Expenditure	124.54	129.42	**0.01***	130.64	129.42	0.65
Total Transactions	3.53	3.37	**0.00***	3.36	3.37	0.89
$ Store Brand Prod	0.02	0.02	0.92	0.02	0.02	0.37
Trx Store Brand Prod	0.02	0.02	**0.00***	-	-	-
$ Discount Prod	0.09	0.09	**0.00***	0.09	0.09	0.99
Trx Discount Prod	0.09	0.09	**0.00***	0.09	0.09	0.22
Basket Size Per Trx	11.46	12.77	**0.00***	12.97	12.77	0.30
Credit Card Expenditure (USD/month):						

Continue next page

Automotive (CC)	1.98	2.84	**0.00***	2.45	2.84	0.44
Gasoline (CC)	4.94	6.36	**0.00***	6.44	6.36	0.80
Entertainment (CC)	3.70	4.98	**0.00***	4.33	4.98	**0.02***
Pharmacy (CC)	7.28	5.68	**0.00***	5.50	5.68	0.48
Credit Card (CC)	189.57	238.18	**0.00***	-	-	-
Home Improvement (CC)	23.53	31.38	**0.00***	31.85	31.38	0.79
Restaurants (CC)	3.93	5.69	**0.00***	5.39	5.69	0.33
Health (CC)	6.40	7.93	**0.00***	8.36	7.93	0.58
Insurance (CC)	7.39	7.92	0.22	8.19	7.92	0.70
Supermarket (CC)	29.97	33.39	**0.00***	33.60	33.39	0.92
Department Store (CC)	75.64	98.15	**0.00***	105.36	98.15	0.16
Travels (CC)	6.66	10.40	**0.00***	9.39	10.40	0.27
Real States/Services (CC)	3.31	4.40	0.01*	3.45	4.40	0.07
Supermarket Categories Expenditure (USD/month):						
Alcoholic Drinks	0.87	1.09	**0.00***	1.12	1.09	0.65
Baby Care	0.16	0.10	**0.00***	0.09	0.10	0.85
Baby Clothing	0.10	0.09	**0.00***	0.09	0.09	0.61
Baby Food	0.01	0.01	**0.00***	0.01	0.01	0.24
Basic Foods	0.78	0.72	**0.00***	-	-	-
Beef Meat	0.87	0.97	**0.00***	1.00	0.97	0.57
Books,Magazines,Music	0.03	0.02	**0.00***	0.03	0.02	0.88
Boy Clothing	0.04	0.03	**0.01***	0.03	0.03	0.73
Butters	0.12	0.09	**0.00***	0.09	0.09	0.71
Canned/Preserved Food	0.26	0.23	**0.00***	0.23	0.23	0.51
Car	0.02	0.02	**0.01***	0.02	0.02	0.65
Care and Beauty	0.22	0.25	**0.01***	0.25	0.25	0.58
Cheeses	0.39	0.42	**0.00***	-	-	-
Cocktail	0.28	0.34	**0.00***	0.33	0.34	0.32
Confectionery	0.42	0.37	**0.00***	0.36	0.37	0.43
Deli	0.52	0.59	**0.00***	-	-	-
Eggs	0.07	0.08	**0.00***	0.08	0.08	0.92
Frozen Food	0.50	0.57	**0.00***	0.58	0.57	0.64
Fruits and Vegetables	0.52	0.60	**0.00***	0.61	0.60	0.74
Kitchenware	0.25	0.27	**0.03***	0.27	0.27	0.99
Lamb Meat	0.00	0.00	**0.00***	0.00	0.00	0.92
Men Shoes	0.01	0.02	**0.00***	0.02	0.02	0.98
Personal Hygiene	0.57	0.67	**0.00***	0.79	0.67	0.24
Pets Articles	0.47	0.42	**0.00***	0.42	0.42	0.89
Ready Meals	0.17	0.14	**0.00***	0.14	0.14	0.84
Soft Drinks	0.52	0.57	**0.00***	0.58	0.57	0.70
Sweet Breakfast	0.71	0.60	**0.00***	-	-	-
Turkey Meat	0.07	0.06	**0.03***	0.06	0.06	0.54
Audio Video	0.48	0.46	0.61	0.45	0.46	0.90
Baby Bedroom	0.00	0.00	0.21	0.00	0.00	0.45

Continue next page

Baby Carriage and Chairs	0.00	0.00	0.10	0.00	0.00	0.99
Baby Shoes	0.01	0.01	0.33	0.01	0.01	0.99
Bakery and Pastry	0.55	0.54	0.17	0.53	0.54	0.73
Battery and Rolling	0.01	0.01	0.59	0.00	0.01	0.15
Boy Shoes	0.01	0.00	0.06	0.01	0.00	0.33
Chicken Meat	0.47	0.49	0.29	0.53	0.49	0.39
Fish and Seafood	0.01	0.02	0.20	0.01	0.02	0.14
Girl Clothing	0.04	0.04	0.95	0.03	0.04	0.52
Girl Shoes	0.01	0.01	0.76	0.01	0.01	0.59
Large Size Appliances	0.14	0.13	0.62	0.15	0.13	0.39
Less Size Appliances	0.28	0.30	0.33	0.29	0.30	0.85
Materials	0.00	0.00	0.36	0.00	0.00	0.32
Men Clothing	0.03	0.04	0.13	0.03	0.04	0.31
Milk and Creams	0.26	0.25	0.31	0.24	0.25	0.34
Outdoor / Camping	0.01	0.01	0.10	0.01	0.01	0.14
Party Articles	0.04	0.04	0.84	0.04	0.04	0.66
Pork Meat	0.27	0.29	0.05	0.28	0.29	0.64
Desserts and Ice-Cream	0.05	0.05	0.44	0.05	0.05	0.69
Promotions	0.02	0.01	0.12	0.01	0.01	0.73
Refrigerated Mass and Pasta	0.00	0.00	0.12	0.00	0.00	0.88
Sports	0.03	0.04	0.26	0.04	0.04	0.97
Sweet Desserts Packaged	0.08	0.08	0.51	0.08	0.08	0.79
Toys	0.18	0.19	0.36	0.18	0.19	0.53
Washing	1.36	1.34	0.68	1.26	1.31	0.55
Women Clothing	0.10	0.11	0.21	0.11	0.11	0.93
Women Shoes	0.08	0.09	0.07	0.09	0.09	0.79
Yogurt	0.22	0.21	0.10	0.22	0.21	0.80

(-) means that this covariate was not part of the matching because it was highly correlated with others.

Table 2.4: t-Test for numerical covariates before and after matching

2.9.2 Appendix B: Causal forest covariates.

Sociodemographic
Age
Marital Status
Zone of residence
General Behaviour
% $ Discount Products
% $ Store Brands Products
% Trx Discount Products
% Trx Store Brands Products
Basket Size per Transaction
Number of Period Transactions
Period Supermarket Expenditure
Total Expenditure
Total Transactions
Expenditure on Supermarket
Alcoholic Drinks
Audio Video
Baby Bedroom
Baby Care
Baby Carriage and Chairs
Baby Clothing
Baby Food
Baby Shoes
Bags and Suitcases
Bakery and Pastry
Basic Foods
Bath
Battery and Rolling
Beach
Bedroom
Beef Meat
Books,Magazines,Music
Boy Clothing
Boy Shoes
Butters
Preserved Food
Car
Care and Beauty
Cheeses
Chicken Meat
Christmas
Clothing Accessories
Cocktail
Computing
Confectionery
Containers
Deli
Desserts and Ice-Cream
Dining Table Stuff
Eggs
Extended Warranty
Fish and Seafood
Frozen Food
Fruits and Vegetables
Girl Clothing
Girl Shoes
Hardware Store
Home Decoration
Kitchen
Kitchenware
Lamb Meat
Large Size Appliances
Less Size Appliances
Materials
Mattresses
Men Clothing
Men Shoes
Men Sports Clothing
Milk and Creams
Outdoor / Camping
Party Articles
Personal Hygiene
Pets Articles
Phones and Communications
Photography
Pork Meat
Promotions
Ready Meals
Refrigerated Mass and Pasta
School Clothing
Services and Donations
Soft Drinks
Sports
Stationer
Sweet Breakfast
Sweet Desserts Packaged
Terrace / Backyard
Toys
Turkey Meat
Underwear
Washing
Women Clothing
Women Shoes
Women Sports Clothing
Yogurt
Young Men
Expenditure on Credit Card
Automotive (CC)
Business (CC)
Communications (CC)
Credit Card (CC)
Department Store (CC)
Education (CC)
Entertainment (CC)
Gasoline (CC)
Health (CC)
Home Improvement (CC)
Insurance (CC)
Pharmacy (CC)
Real States/Basics Services (CC)
Restaurants (CC)
Supermarket (CC)
Transport (CC)
Travels (CC)
Web Payments (CC)

Table 2.5: Set of covariates for causal forest models.

2.9.3 Appendix C: Average treatment effects comparison.

Variable Description	Pregnancy			Birth		
	DID	**CF-PS**	**CF-ALL**	**DID**	**CF-PS**	**CF-ALL**
	β_3 *p-value* (*Sd*)	τ *p-value* (*Sd*)	τ *p-value* (*Sd*)	β_3 *p-value* (*Sd*)	τ *p-value* (*Sd*)	τ *p-value* (*Sd*)
Global Transactional Behaviour						
Total Expenditure	1.800 *0.620* *(3.660)*	-1.871 *0.274* *(2.157)*	-6.711 ***0.000**** ***(1.568)***	11.526 ***0.001**** *(3.601)*	5.464 ***0.012**** *(2.063)*	3.176 *0.062* *(1.647)*
Supermarket (CC)	-1.360 *0.260* *(1.210)*	-1.763 ***0.013**** *(0.675)*	-1.614 ***0.006**** *(0.554)*	3.138 ***0.014**** *(1.277)*	2.437 ***0.005**** *(0.822)*	1.790 ***0.009**** *(0.652)*
Basket Size Per Transaction	0.712 ***0.005**** *(0.255)*	0.210 *0.164* *(0.158)*	-0.047 *0.372* *(0.125)*	0.732 ***0.003**** *(0.247)*	0.272 *0.071* *(0.147)*	2.431 *0.354* *(4.998)*
% $ Discount Products	0.000 *0.926* *(0.002)*	-0.001 *0.230* *(0.001)*	-0.003 ***0.028**** *(0.001)*	0.005 ***0.034**** *(0.002)*	0.002 *0.109* *(0.001)*	0.003 ***0.007**** *(0.001)*
% $ Store Brands Products	0.000 *0.500* *(0.001)*	-0.001 ***0.015**** *(0.000)*	-0.002 ***0.000**** *(0.000)*	-0.001 *0.083* *(0.001)*	-0.002 ***0.000**** *(0.000)*	-0.003 ***0.000**** *(0.000)*
% Trx Store Brands Products	0.001 *0.302* *(0.001)*	-0.001 *0.163* *(0.000)*	-0.002 ***0.000**** *(0.000)*	0.001 *0.116* *(0.001)*	-0.001 *0.065* *(0.000)*	-0.001 ***0.000**** *(0.000)*
% Trx Discount Products	0.000 *0.843* *(0.002)*	-0.001 *0.257* *(0.001)*	-0.003 ***0.014**** *(0.001)*	0.003 *0.124* *(0.002)*	0.001 *0.257* *(0.001)*	0.002 *0.069* *(0.001)*
Credit Card (CC)	-5.750 *0.590* *(10.690)*	-6.912 *0.117* *(4.418)*	-13.006 ***0.000**** *(2.692)*	-8.475 *0.454* *(11.314)*	-8.029 *0.193* *(6.663)*	-20.688 ***0.000**** *(5.425)*
Total Transactions	-0.147 *0.086* *(0.085)*	-0.171 ***0.001**** *(0.051)*	-0.204 ***0.000**** *(0.039)*	-0.026 *0.754* *(0.083)*	-0.117 ***0.021**** *(0.048)*	-0.125 ***0.002**** *(0.039)*
Baby products						
Baby Care	0.050	0.041	0.065	0.659	0.603	0.611

Continue next page

	0.000*	***0.005****	***0.000****	***0.000****	***0.000****	***0.000****
	(0.010)	*(0.014)*	*(0.010)*	*(0.032)*	*(0.029)*	*(0.026)*
Baby Clothing	0.110	0.099	0.102	0.648	0.620	0.632
	0.000*	***0.000****	***0.000****	***0.000****	***0.000****	***0.000****
	(0.010)	*(0.012)*	*(0.010)*	*(0.024)*	*(0.024)*	*(0.025)*
Toys	-0.010	-0.004	-0.009	0.080	0.085	0.068
	0.510	*0.378*	*0.220*	***0.000****	***0.000****	***0.000****
	(0.020)	*(0.013)*	*(0.009)*	*(0.021)*	*(0.016)*	*(0.013)*
Baby Food	-0.010	-0.009	-0.006	0.054	0.052	0.059
	0.000*	***0.002****	***0.011****	***0.000****	***0.000****	***0.000****
	(0.000)	*(0.003)*	*(0.002)*	*(0.004)*	*(0.005)*	*(0.005)*
Baby Shoes	0.010	0.012	0.012	0.040	0.040	0.040
	0.000*	***0.000****	***0.000****	***0.000****	***0.000****	***0.000****
	(0.000)	*(0.002)*	*(0.002)*	*(0.003)*	*(0.003)*	*(0.003)*
Baby Bedroom	0.000	0.001	0.002	0.002	0.003	0.003
	0.500	*0.325*	***0.006****	***0.030****	***0.010****	***0.001****
	(0.000)	*(0.001)*	*(0.001)*	*(0.001)*	*(0.001)*	*(0.001)*
Baby Carriage and Chairs	0.000	0.001	0.001	0.001	0.001	0.001
	0.140	*0.268*	*0.098*	*0.317*	*0.269*	*0.194*
	(0.000)	*(0.001)*	*(0.001)*	*(0.001)*	*(0.001)*	*(0.001)*
Entertainment						
Travels (CC)	-6.040	-5.516	-4.084	-6.551	-5.992	-5.729
	0.000*	***0.000****	***0.000****	***0.000****	***0.000****	***0.000****
	(1.220)	*(0.938)*	*(0.707)*	*(1.131)*	*(0.877)*	*(0.640)*
Restaurants (CC)	-1.220	-1.234	-0.943	-3.329	-3.279	-2.522
	0.010*	***0.000****	***0.000****	***0.000****	***0.000****	***0.000****
	(0.430)	*(0.244)*	*(0.206)*	*(0.422)*	*(0.267)*	*(0.142)*
Entertainment (CC)	-2.490	-2.612	-1.961	-3.253	-3.146	-2.758
	0.000*	***0.000****	***0.000****	***0.000****	***0.000****	***0.000****
	(0.360)	*(0.272)*	*(0.173)*	*(0.391)*	*(0.271)*	*(0.161)*
Party Articles	0.000	-0.001	-0.001	-0.007	-0.006	-0.005
	0.710	*0.390*	*0.324*	***0.034****	***0.012****	***0.013****
	(0.000)	*(0.003)*	*(0.002)*	*(0.003)*	*(0.002)*	*(0.002)*
Outdoor / Camping	-0.010	-0.006	0.003	0.008	0.009	0.015

Continue next page

	0.080	*0.256*	*0.224*	*0.169*	*0.159*	***0.001****
	(0.010)	*(0.006)*	*(0.003)*	*(0.006)*	*(0.007)*	*(0.004)*
Beach	0.000	0.001	-0.001	-0.001	0.000	-0.001
	0.840	*0.214*	*0.206*	*0.525*	*0.398*	*0.195*
	(0.000)	*(0.001)*	*(0.001)*	*(0.001)*	*(0.001)*	*(0.001)*
Books,Magazines,Music	-0.010	-0.011	0.001	0.000	-0.001	0.000
	0.160	*0.052*	*0.395*	*0.942*	*0.394*	*0.397*
	(0.010)	*(0.005)*	*(0.004)*	*(0.005)*	*(0.004)*	*(0.002)*
Bags and Suitcases	0.000	-0.006	0.001	0.000	-0.002	0.003
	0.590	*0.091*	*0.369*	*0.984*	*0.360*	*0.304*
	(0.000)	*(0.004)*	*(0.003)*	*(0.004)*	*(0.003)*	*(0.004)*
Technology						
Phones and Communications	0.060	0.008	0.000	0.113	0.082	-0.015
	0.320	*0.395*	*0.399*	***0.024****	*0.121*	*0.321*
	(0.060)	*(0.058)*	*(0.026)*	*(0.050)*	*(0.053)*	*(0.022)*
Computing	0.050	0.043	0.001	0.056	0.011	-0.009
	0.010*	*0.067*	*0.398*	***0.005****	*0.288*	*0.304*
	(0.020)	*(0.023)*	*(0.014)*	*(0.020)*	*(0.013)*	*(0.013)*
Photography	0.000	0.000	0.002	0.011	0.010	0.005
	0.640	*0.399*	*0.213*	***0.012****	***0.048****	***0.016****
	(0.010)	*(0.008)*	*(0.002)*	*(0.004)*	*(0.005)*	*(0.002)*
Audio Video	0.000	-0.017	-0.020	0.063	0.015	0.021
	0.990	*0.383*	*0.350*	*0.355*	*0.387*	*0.347*
	(0.070)	*(0.060)*	*(0.038)*	*(0.068)*	*(0.060)*	*(0.040)*
Personal Care						
Care and Beauty	0.000	-0.003	-0.026	-0.010	-0.018	-0.038
	0.980	*0.386*	***0.001****	*0.467*	*0.063*	***0.000****
	(0.010)	*(0.010)*	*(0.007)*	*(0.014)*	*(0.009)*	*(0.006)*
Personal Hygiene	-0.050	-0.042	-0.052	-0.033	-0.035	-0.047
	0.100	***0.046****	***0.001****	*0.230*	*0.064*	***0.000****
	(0.030)	*(0.020)*	*(0.014)*	*(0.028)*	*(0.018)*	*(0.012)*
Healthy						
Milk and Creams	0.080	0.073	0.065	0.025	0.012	0.018
	0.000*	***0.000****	***0.000****	*0.076*	*0.187*	*0.118*
	(0.020)	*(0.011)*	*(0.010)*	*(0.014)*	*(0.010)*	*(0.012)*
Yogurt	0.040	0.035	0.020	-0.008	-0.013	-0.020

Continue next page

	0.000*	***0.000****	***0.001****	*0.401*	***0.021****	***0.000****
	(0.010)	*(0.007)*	*(0.006)*	*(0.009)*	*(0.005)*	*(0.004)*
Fish and Seafood	0.000	0.004	0.001	-0.007	-0.005	-0.002
	0.810	*0.179*	*0.376*	*0.059*	*0.131*	*0.294*
	(0.000)	*(0.003)*	*(0.003)*	*(0.004)*	*(0.004)*	*(0.003)*
Chicken Meat	0.060	0.071	-0.079	0.087	0.085	-0.022
	0.250	*0.120*	***0.035****	*0.090*	*0.091*	*0.262*
	(0.050)	*(0.046)*	*(0.036)*	*(0.051)*	*(0.049)*	*(0.024)*
Fruits and Vegetables	0.040	0.051	-0.001	-0.023	-0.025	-0.055
	0.280	***0.034****	*0.399*	*0.499*	*0.220*	***0.000****
	(0.040)	*(0.023)*	*(0.018)*	*(0.035)*	*(0.022)*	*(0.013)*
Turkey Meat	-0.010	-0.004	-0.008	0.003	0.003	0.004
	0.380	*0.316*	*0.063*	*0.655*	*0.324*	*0.193*
	(0.010)	*(0.006)*	*(0.004)*	*(0.007)*	*(0.005)*	*(0.003)*
Eggs	0.000	0.001	-0.002	-0.001	0.000	0.000
	0.900	*0.389*	*0.292*	*0.930*	*0.398*	*0.393*
	(0.010)	*(0.004)*	*(0.003)*	*(0.006)*	*(0.004)*	*(0.003)*
Unhealthy						
Alcoholic Drinks	-0.080	-0.066	-0.183	-0.126	-0.134	-0.202
	0.190	*0.166*	***0.000****	***0.043****	***0.009****	***0.000****
	(0.060)	*(0.050)*	*(0.024)*	*(0.062)*	*(0.049)*	*(0.018)*
Cocktail	-0.010	-0.014	-0.016	-0.051	-0.052	-0.048
	0.350	*0.127*	***0.012****	***0.000****	***0.000****	***0.000****
	(0.010)	*(0.010)*	*(0.006)*	*(0.013)*	*(0.009)*	*(0.005)*
Butters	0.010	0.004	0.001	0.011	0.007	0.005
	0.100	*0.185*	*0.387*	***0.026****	*0.081*	*0.141*
	(0.010)	*(0.004)*	*(0.003)*	*(0.005)*	*(0.004)*	*(0.004)*
Sweet Desserts Packaged	0.010	0.009	0.008	0.008	0.004	0.006
	0.010*	***0.013****	***0.004****	*0.073*	*0.181*	***0.027****
	(0.000)	*(0.003)*	*(0.003)*	*(0.005)*	*(0.003)*	*(0.003)*
Refrigerated Mass&Pasta	0.000	-0.001	-0.001	-0.001	-0.002	-0.001
	0.520	*0.207*	***0.000****	*0.110*	*0.104*	***0.032****
	(0.000)	*(0.001)*	*(0.000)*	*(0.001)*	*(0.001)*	*(0.001)*
Ready Meals	-0.010	-0.007	-0.013	-0.009	-0.016	-0.016
	0.470	*0.222*	***0.014****	*0.281*	***0.019****	***0.000****
	(0.010)	*(0.007)*	*(0.005)*	*(0.009)*	*(0.006)*	*(0.004)*

Continue next page

Sweet Breakfast	0.010	-0.017	-0.055	0.024	-0.007	-0.041
	0.820	*0.247*	***0.000****	*0.352*	*0.369*	***0.002****
	(0.030)	*(0.017)*	*(0.013)*	*(0.026)*	*(0.016)*	*(0.013)*
Confectionery	-0.010	-0.013	-0.028	0.012	0.008	0.000
	0.340	*0.192*	***0.000****	*0.420*	*0.288*	*0.399*
	(0.020)	*(0.010)*	*(0.007)*	*(0.015)*	*(0.011)*	*(0.008)*
Bakery and Pastry	0.030	0.012	-0.009	0.014	0.002	-0.019
	0.200	*0.251*	*0.283*	*0.478*	*0.396*	*0.056*
	(0.020)	*(0.013)*	*(0.011)*	*(0.019)*	*(0.013)*	*(0.010)*
Frozen Food	0.030	0.019	-0.018	0.011	-0.017	-0.023
	0.270	*0.220*	*0.099*	*0.675*	*0.222*	*0.057*
	(0.030)	*(0.017)*	*(0.011)*	*(0.026)*	*(0.016)*	*(0.012)*
Desserts and Ice-Cream	0.000	0.002	0.000	0.001	-0.002	-0.002
	0.310	*0.275*	*0.399*	*0.725*	*0.319*	*0.319*
	(0.000)	*(0.002)*	*(0.002)*	*(0.003)*	*(0.002)*	*(0.002)*
Canned/Preserved Food	0.010	-0.007	-0.010	-0.002	-0.019	-0.024
	0.590	*0.262*	*0.085*	*0.846*	***0.006****	***0.000****
	(0.010)	*(0.007)*	*(0.006)*	*(0.011)*	*(0.007)*	*(0.004)*
Clothes						
Women Shoes	0.000	0.000	-0.008	-0.019	-0.016	-0.020
	0.830	*0.398*	*0.055*	***0.025****	***0.015****	***0.000****
	(0.010)	*(0.006)*	*(0.004)*	*(0.008)*	*(0.006)*	*(0.004)*
Women Clothing	-0.020	-0.032	-0.014	-0.003	0.002	0.008
	0.050*	***0.015****	***0.019****	*0.768*	*0.385*	*0.356*
	(0.010)	*(0.012)*	*(0.006)*	*(0.011)*	*(0.008)*	*(0.017)*
School Clothing	0.000	-0.001	0.000	0.004	0.004	0.005
	0.790	*0.370*	*0.394*	*0.115*	*0.135*	***0.025****
	(0.000)	*(0.003)*	*(0.001)*	*(0.003)*	*(0.003)*	*(0.002)*
Girl Clothing	-0.010	-0.010	-0.010	-0.010	-0.010	0.000
	0.060	*0.127*	***0.002****	*0.194*	*0.127*	*0.399*
	(0.010)	*(0.007)*	*(0.003)*	*(0.007)*	*(0.006)*	*(0.003)*
Men Sports Clothing	0.000	0.003	0.002	-0.001	-0.001	0.000
	0.140	*0.148*	*0.153*	*0.419*	*0.366*	*0.398*
	(0.000)	*(0.002)*	*(0.002)*	*(0.002)*	*(0.002)*	*(0.001)*
Clothing Accessories	0.000	0.000	0.001	0.001	0.001	0.000

Continue next page

	0.900	*0.356*	*0.214*	*0.528*	*0.342*	*0.284*
	(0.000)	*(0.001)*	*(0.001)*	*(0.001)*	*(0.001)*	*(0.001)*
Sports	-0.010	-0.001	-0.006	-0.006	-0.003	-0.002
	0.240	*0.397*	*0.257*	*0.571*	*0.387*	*0.388*
	(0.010)	*(0.010)*	*(0.006)*	*(0.011)*	*(0.013)*	*(0.007)*
Boy Shoes	0.000	0.002	0.001	-0.001	-0.001	-0.003
	0.320	*0.215*	*0.275*	*0.611*	*0.324*	***0.003****
	(0.000)	*(0.002)*	*(0.001)*	*(0.001)*	*(0.002)*	*(0.001)*
Boy Clothing	-0.010	-0.008	-0.005	-0.003	-0.003	-0.003
	0.300	*0.241*	*0.196*	*0.613*	*0.353*	*0.286*
	(0.010)	*(0.008)*	*(0.004)*	*(0.007)*	*(0.006)*	*(0.003)*
Young Men	-0.010	-0.005	-0.005	-0.002	-0.004	0.001
	0.270	*0.279*	*0.181*	*0.733*	*0.312*	*0.373*
	(0.000)	*(0.006)*	*(0.004)*	*(0.005)*	*(0.005)*	*(0.004)*
Girl Shoes	0.000	0.001	0.000	-0.001	-0.001	-0.002
	0.900	*0.358*	*0.394*	*0.749*	*0.335*	*0.070*
	(0.000)	*(0.002)*	*(0.002)*	*(0.002)*	*(0.002)*	*(0.001)*
Men Clothing	-0.010	-0.005	0.001	-0.002	0.000	0.004
	0.390	*0.320*	*0.384*	*0.786*	*0.399*	*0.220*
	(0.010)	*(0.007)*	*(0.004)*	*(0.007)*	*(0.008)*	*(0.004)*
Women Sports Clothing	0.000	-0.001	-0.002	0.001	0.000	-0.001
	0.740	*0.389*	*0.123*	*0.800*	*0.396*	*0.154*
	(0.000)	*(0.003)*	*(0.002)*	*(0.002)*	*(0.003)*	*(0.001)*
Underwear	-0.010	-0.019	-0.009	0.001	-0.001	-0.009
	0.160	***0.010****	*0.054*	*0.918*	*0.394*	***0.026****
	(0.010)	*(0.007)*	*(0.005)*	*(0.008)*	*(0.006)*	*(0.004)*
Men Shoes	0.000	0.002	0.000	0.000	0.002	-0.001
	0.610	*0.346*	*0.399*	*0.931*	*0.339*	*0.301*
	(0.000)	*(0.004)*	*(0.002)*	*(0.003)*	*(0.003)*	*(0.002)*
Home						
Washing	0.110	0.047	0.020	0.120	0.054	0.039
	0.080	*0.208*	*0.310*	***0.044****	*0.152*	*0.140*
	(0.060)	*(0.041)*	*(0.028)*	*(0.059)*	*(0.039)*	*(0.027)*
Christmas	0.010	0.010	0.008	0.021	0.024	0.016
	0.240	*0.269*	*0.265*	*0.051*	***0.038****	*0.145*
	(0.010)	*(0.011)*	*(0.009)*	*(0.011)*	*(0.011)*	*(0.011)*

Continue next page

Home Improvement (CC)	6.610	7.109	5.362	2.754	4.209	0.528
	0.000*	***0.000****	***0.000****	*0.411*	*0.187*	*0.384*
	(2.180)	*(1.518)*	*(1.135)*	*(3.353)*	*(3.415)*	*(1.883)*
Hardware Store	0.000	-0.001	0.000	0.006	0.005	-0.003
	0.670	*0.394*	*0.399*	*0.062*	*0.129*	*0.188*
	(0.000)	*(0.004)*	*(0.005)*	*(0.003)*	*(0.003)*	*(0.002)*
Kitchen	0.000	-0.003	-0.001	-0.003	-0.003	-0.002
	0.060	*0.063*	*0.094*	*0.070*	*0.097*	***0.010****
	(0.000)	*(0.002)*	*(0.001)*	*(0.002)*	*(0.002)*	*(0.001)*
Containers	0.000	0.000	0.001	0.001	0.001	0.001
	0.550	*0.398*	***0.019****	*0.078*	*0.135*	*(0.000)*
	(0.000)	*(0.000)*	*(0.000)*	*(0.000)*	*(0.000)*	*0.135*
Large Size Appliances	0.050	0.016	-0.002	0.058	0.049	0.015
	0.220	*0.367*	*0.398*	*0.113*	*0.179*	*0.339*
	(0.040)	*(0.039)*	*(0.034)*	*(0.037)*	*(0.039)*	*(0.027)*
Home Decoration	0.000	-0.004	-0.090	-0.008	-0.006	-0.047
	0.800	*0.348*	*0.132*	*0.319*	*0.305*	*0.093*
	(0.010)	*(0.009)*	*(0.060)*	*(0.008)*	*(0.009)*	*(0.028)*
Bedroom	0.010	0.016	0.000	-0.008	-0.005	-0.004
	0.340	*0.064*	*0.398*	*0.395*	*0.333*	*0.342*
	(0.010)	*(0.008)*	*(0.007)*	*(0.009)*	*(0.009)*	*(0.007)*
Terrace / Backyard	0.000	0.018	0.008	-0.011	-0.018	0.002
	0.710	*0.140*	*0.232*	*0.414*	*0.182*	*0.388*
	(0.010)	*(0.012)*	*(0.007)*	*(0.013)*	*(0.014)*	*(0.010)*
Less Size Appliances	0.000	-0.006	-0.026	-0.014	-0.014	-0.043
	0.890	*0.377*	*0.062*	*0.592*	*0.290*	***0.011****
	(0.030)	*(0.018)*	*(0.014)*	*(0.025)*	*(0.017)*	*(0.016)*
Kitchenware	-0.020	-0.029	-0.029	0.005	-0.002	0.014
	0.230	***0.023****	***0.000****	*0.772*	*0.395*	*0.268*
	(0.020)	*(0.012)*	*(0.008)*	*(0.016)*	*(0.012)*	*(0.015)*
Mattresses	0.000	0.000	0.002	0.001	-0.001	0.007
	0.890	*0.399*	*0.293*	*0.846*	*0.388*	***0.025****
	(0.000)	*(0.005)*	*(0.002)*	*(0.004)*	*(0.005)*	*(0.003)*
Bath	0.000	0.004	0.001	0.000	0.000	0.000
	0.290	*0.103*	*0.314*	*0.898*	*0.395*	*0.399*

Continue next page

	(0.000)	*(0.003)*	*(0.002)*	*(0.004)*	*(0.003)*	*(0.002)*
Dining Table Stuff	0.000	0.001	0.000	0.000	-0.001	0.000
	0.290	*0.284*	*0.379*	*0.957*	*0.290*	*0.395*
	(0.000)	*(0.001)*	*(0.001)*	*(0.001)*	*(0.001)*	*(0.001)*
Stationer	0.000	-0.013	-0.014	0.000	-0.009	-0.016
	0.850	*0.054*	***0.000****	*0.977*	*0.126*	***0.000****
	(0.010)	*(0.006)*	*(0.004)*	*(0.008)*	*(0.006)*	*(0.003)*
Materials	0.000	0.000	0.000	0.000	0.000	0.000
	0.250	***0.034****	*0.389*	*0.993*	*0.372*	***0.008****
	(0.000)	*(0.000)*	*(0.000)*	*(0.000)*	*(0.000)*	*(0.000)*
Others Services						
Health (CC)	2.850	2.796	-0.282	5.173	5.041	1.687
	0.010*	***0.001****	*0.349*	***0.000****	***0.000****	***0.001****
	(1.110)	*(0.821)*	*(0.547)*	*(1.070)*	*(0.914)*	*(0.494)*
Pharmacy (CC)	-0.200	-0.401	-0.988	3.091	2.852	2.052
	0.590	***0.044****	***0.000****	***0.000****	***0.000****	***0.000****
	(0.370)	*(0.191)*	*(0.202)*	*(0.415)*	*(0.283)*	*(0.324)*
Transport (CC)	-0.380	-0.365	-0.404	-1.306	-1.397	-1.123
	0.170	***0.001****	***0.000****	***0.000****	***0.000****	***0.000****
	(0.280)	*(0.106)*	*(0.062)*	*(0.313)*	*(0.240)*	*(0.088)*
Gasoline (CC)	-0.400	-0.387	-0.516	-0.746	-0.733	-1.046
	0.400	*0.083*	***0.005****	*0.093*	***0.005****	***0.000****
	(0.470)	*(0.218)*	*(0.174)*	*(0.444)*	*(0.246)*	*(0.186)*
Web Payments (CC)	0.010	-0.056	-0.141	-0.478	-0.425	-0.090
	0.960	*0.391*	*0.161*	*0.144*	*0.197*	*0.297*
	(0.290)	*(0.282)*	*(0.104)*	*(0.327)*	*(0.357)*	*(0.118)*
Education (CC)	-0.210	-0.334	-0.844	-0.878	-1.518	-0.976
	0.740	*0.330*	***0.000****	*0.223*	***0.047****	***0.000****
	(0.640)	*(0.540)*	*(0.204)*	*(0.720)*	*(0.733)*	*(0.188)*
Communications (CC)	-0.160	-0.472	-0.313	-0.431	-0.630	-0.557
	0.650	***0.030****	***0.030****	*0.268*	***0.013****	***0.008****
	(0.350)	*(0.207)*	*(0.138)*	*(0.389)*	*(0.242)*	*(0.199)*
Insurance (CC)	0.660	0.420	0.079	1.083	1.169	0.290
	0.520	*0.212*	*0.373*	*0.340*	***0.042****	*0.321*
	(1.030)	*(0.374)*	*(0.214)*	*(1.135)*	*(0.550)*	*(0.442)*

Continue next page

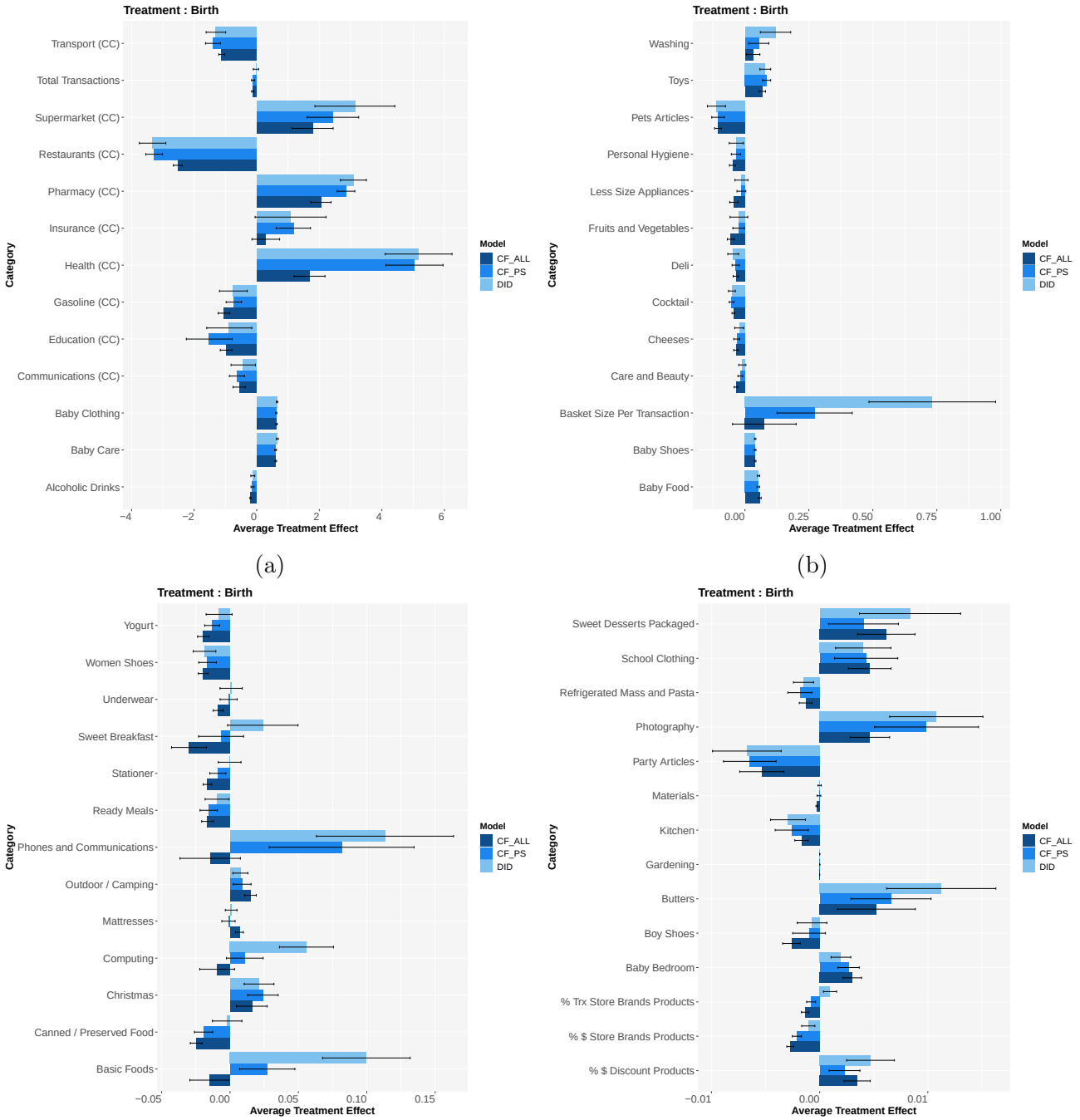
Department Store (CC)	-1.460	-2.135	-4.072	-4.495	-3.564	-10.406
	0.820	*0.279*	***0.000****	*0.460*	*0.279*	***0.000****
	(6.490)	*(2.528)*	*(1.076)*	*(6.083)*	*(4.219)*	*(1.061)*
Automotive (CC)	0.670	0.695	-0.236	-0.554	-0.595	-0.434
	0.530	*0.337*	*0.378*	*0.516*	*0.331*	*0.141*
	(1.050)	*(1.193)*	*(0.719)*	*(0.853)*	*(0.977)*	*(0.300)*
Real States (CC)	-0.630	-0.432	-0.633	0.549	0.169	-0.315
	0.410	*0.277*	***0.015****	*0.651*	*0.393*	*0.298*
	(0.760)	*(0.507)*	*(0.247)*	*(1.214)*	*(1.008)*	*(0.412)*
Business (CC)	0.190	0.229	0.173	0.162	0.049	0.074
	0.330	*0.089*	*0.070*	*0.383*	*0.376*	*0.318*
	(0.200)	*(0.133)*	*(0.093)*	*(0.186)*	*(0.144)*	*(0.109)*
Others Products						
Pets Articles	-0.040	-0.048	-0.045	-0.111	-0.105	-0.106
	0.200	*0.066*	***0.006****	***0.002****	***0.000****	***0.000****
	(0.030)	*(0.025)*	*(0.016)*	*(0.035)*	*(0.025)*	*(0.012)*
Deli	-0.020	-0.003	-0.021	-0.046	-0.036	-0.035
	0.420	*0.388*	*0.073*	***0.030****	***0.014****	***0.001****
	(0.020)	*(0.015)*	*(0.011)*	*(0.021)*	*(0.014)*	*(0.010)*
Promotions	0.000	0.002	0.003	-0.006	-0.004	-0.002
	0.770	*0.362*	*0.254*	*0.061*	*0.244*	*0.310*
	(0.000)	*(0.004)*	*(0.003)*	*(0.003)*	*(0.004)*	*(0.003)*
Beef Meat	0.030	0.015	-0.065	0.113	0.096	-0.010
	0.710	*0.379*	***0.009****	*0.088*	*0.075*	*0.366*
	(0.070)	*(0.047)*	*(0.024)*	*(0.066)*	*(0.053)*	*(0.025)*
Cheeses	0.010	-0.004	-0.021	-0.022	-0.032	-0.035
	0.780	*0.373*	***0.030****	*0.202*	***0.009****	***0.000****
	(0.020)	*(0.012)*	*(0.009)*	*(0.018)*	*(0.012)*	*(0.008)*
Car	0.000	0.002	0.002	-0.003	-0.001	-0.001
	0.760	*0.362*	*0.352*	*0.428*	*0.374*	*0.356*
	(0.000)	*(0.004)*	*(0.004)*	*(0.004)*	*(0.003)*	*(0.003)*
Soft Drinks	0.040	0.029	0.026	0.014	-0.001	-0.019
	0.070	*0.063*	***0.013****	*0.529*	*0.397*	***0.026****
	(0.020)	*(0.015)*	*(0.010)*	*(0.022)*	*(0.014)*	*(0.008)*
Lamb Meat	0.000	0.001	0.002	0.001	0.000	0.001
	0.240	*0.287*	*0.146*	*0.598*	*0.394*	*0.202*

Continue next page

	(0.000)	*(0.001)*	*(0.001)*	*(0.001)*	*(0.001)*	*(0.000)*
Pork Meat	0.000	0.012	-0.031	0.010	0.017	0.003
	0.960	*0.267*	***0.001****	*0.602*	*0.208*	*0.383*
	(0.020)	*(0.014)*	*(0.009)*	*(0.019)*	*(0.015)*	*(0.010)*
Battery and Rolling	-0.010	-0.003	0.000	-0.002	0.002	0.003
	0.080	*0.295*	*0.395*	*0.636*	*0.367*	*0.189*
	(0.000)	*(0.003)*	*(0.003)*	*(0.004)*	*(0.005)*	*(0.003)*
Basic Foods	0.070	0.021	-0.018	0.100	0.027	-0.020
	0.030*	*0.242*	*0.200*	***0.002****	*0.161*	*0.154*
	(0.030)	*(0.021)*	*(0.016)*	*(0.032)*	*(0.020)*	*(0.015)*

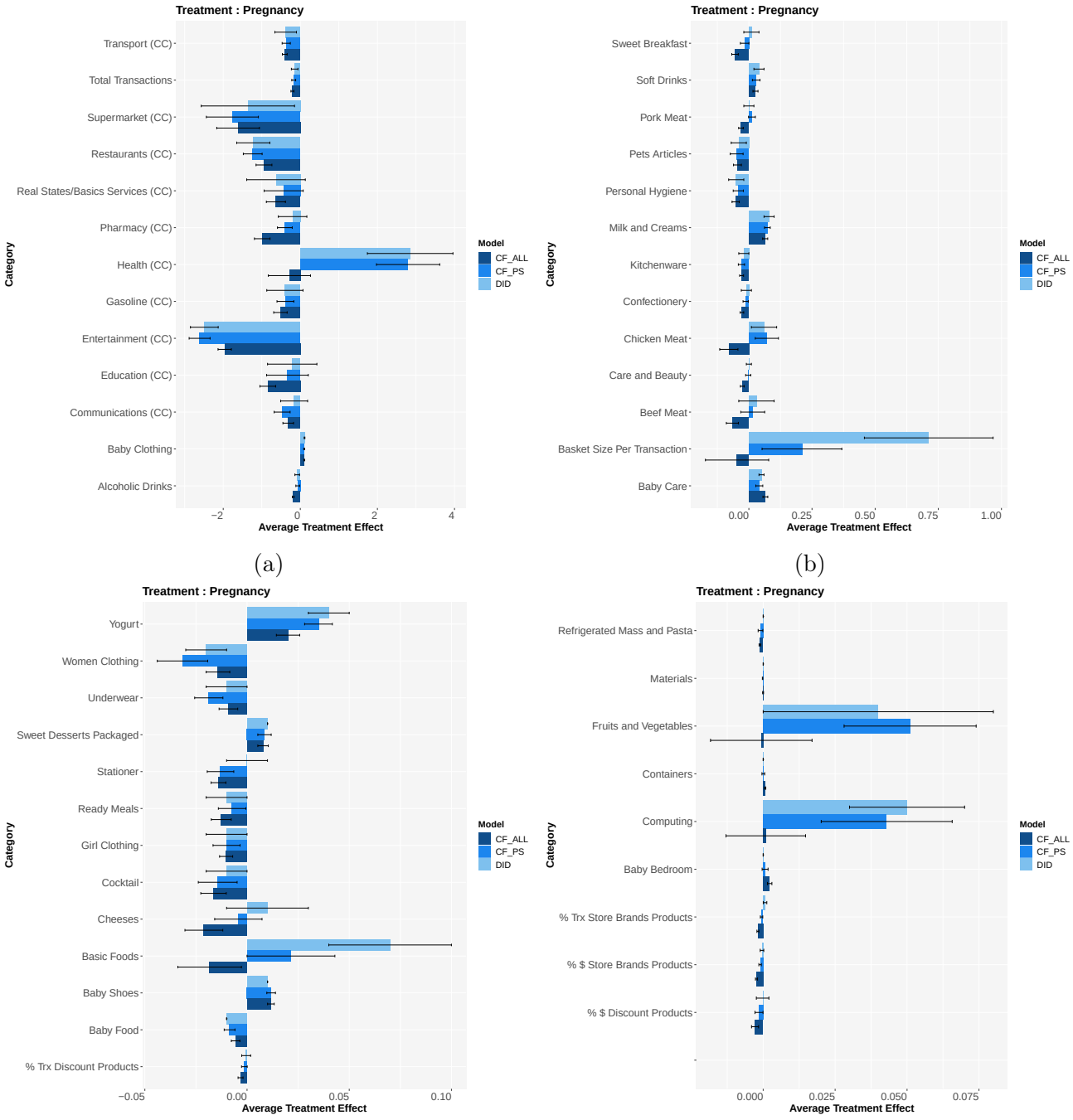
95% confidence intervals.

Table 2.6: Diff & Diff and Causal Forest results comparison



Note: The bar represents the magnitude and the line represents the standard deviation of the average treatment effect.

Figure 2.7: Average treatment effect of birth comparison.



Note: The bar represents the magnitude and the line represents the standard deviation of the average treatment effect.

Figure 2.8: Average treatment effect of pregnancy comparison.

2.9.4 Appendix D: Average treatment effect regression analysis.

	Dependent :
	ATE Magnitude
Model:CF_PS	−0.156
	(0.148)
Model: DID	−0.097
	(0.148)
Store Brands Products	−0.001
	(0.636)
% Trx Discount Products	0.284
	(0.782)
% Trx Store Brands Products	−0.001
	(0.636)
Alcoholic Drinks	0.129
	(0.636)
Baby Bedroom	−0.001
	(0.636)
Baby Care	0.335
	(0.636)
Baby Clothing	0.366
	(0.636)
Baby Food	0.029
	(0.636)
Baby Shoes	0.023
	(0.636)
Basic Foods	0.039
	(0.636)
Basket Size Per Transaction	0.339
	(0.636)
Beef Meat	0.319
	(0.782)
Boy Shoes	−0.286
	(0.782)
Butters	−0.280
	(0.782)
Canned / Preserved Food	−0.272
	(0.782)
Care and Beauty	0.013
	(0.636)
Cheeses	0.018
	(0.636)
Chicken Meat	0.352
	(0.782)
Christmas	−0.267
	(0.782)
Cocktail	0.029

Continue next page

	(0.636)
Communications (CC)	0.425
	(0.636)
Computing	0.026
	(0.636)
Confectionery	0.299
	(0.782)
Containers	0.283
	(0.782)
Credit Card (CC)	10.581***
	(0.636)
Deli	−0.248
	(0.782)
Department Store (CC)	4.353***
	(0.636)
Education (CC)	0.791
	(0.636)
Entertainment (CC)	2.701***
	(0.636)
Fruits and Vegetables	0.030
	(0.636)
Gardening	−0.287
	(0.782)
Gasoline (CC)	0.635
	(0.636)
Girl Clothing	0.292
	(0.782)
Health (CC)	2.969***
	(0.636)
Home Improvement (CC)	6.643***
	(0.782)
Insurance (CC)	0.560
	(0.782)
Kitchen	−0.285
	(0.782)
Kitchenware	0.308
	(0.782)
Less Size Appliances	−0.264
	(0.782)
Materials	−0.002
	(0.636)
Mattresses	−0.284
	(0.782)
Milk and Creams	0.355
	(0.782)
Outdoor / Camping	−0.277
	(0.782)

Continue next page

Party Articles	−0.281
	(0.782)
Personal Hygiene	0.041
	(0.636)
Pets Articles	0.073
	(0.636)
Pharmacy (CC)	1.595**
	(0.636)
Phones and Communications	−0.217
	(0.782)
Photography	−0.279
	(0.782)
Pork Meat	0.297
	(0.782)
Ready Meals	0.009
	(0.636)
Real States/Basics Services (CC)	0.847
	(0.782)
Refrigerated Mass and Pasta	−0.001
	(0.636)
Restaurants (CC)	2.085***
	(0.636)
School Clothing	−0.283
	(0.782)
Soft Drinks	0.314
	(0.782)
Stationer	0.006
	(0.636)
Supermarket (CC)	2.014***
	(0.636)
Sweet Breakfast	0.021
	(0.636)
Sweet Desserts Packaged	0.005
	(0.636)
Total Expenditure	5.051***
	(0.636)
Total Transactions	0.129
	(0.636)
Toys	−0.210
	(0.782)
Transport (CC)	0.826
	(0.636)
Travels (CC)	5.649***
	(0.636)
Underwear	0.006
	(0.636)
Washing	−0.219

Continue next page

	(0.782)
Women Clothing	0.304
	(0.782)
Women Shoes	−0.269
	(0.782)
Yogurt	0.020
	(0.636)
Treatment: Birth	0.570***
	(0.142)
Constant	−0.198
	(0.463)
Observations	333
R^2	0.796
Adjusted R^2	0.739
Residual Std. Error	1.101 (df = 259)
F Statistic	13.872*** (df = 73; 259)

Note: *p<0.1; **p<0.05; ***p<0.01

Regression: $|ATE| \sim \beta_0 + \beta_1 * Category + \beta_2 * Treatment + \beta_3 * Model + \varepsilon$

Table 2.7: Average treatment effect regression analysis.

Conclusions and Future Work

In this work we studied two types of consumer behavior under bounded rationality framework, the attribute non-attendance and the preferences changes under a life events.

Fist, in the attribute non attendance context, we studied the capability of the Support Vector Machines (SVM) approach to identify non-attended attributes, and to predict consumer choices in choice experiment data. We started by testing several specifications of the SVM model with selective attention simulated data, considering four different non-attention rates (20%,40%,60%,80%). The SVM specifications were formulated from three different regularization components (attribute contribution, one-norm, and infinite-norm), both at the individual level and at the population level. Additionally, we tested two elimination algorithms to identify non-attended attributes: backward elimination proposed by Maldonado et al. (2015), and our proposed minimum contribution criterion. The two-stage model with the best performance in predictive capacity was using indifferently attribute contribution (AC) or the infinite norm ($Norm_\infty$) as regularization component of the individual problem, minimum contribution criterion to identify non-attended attributes, and indifferently the $Norm_\infty$ or AC as regularization component of the population problem.

SVM showed to be a robust approach under different attribute non-attendance rates, unlike most Benchmark models, whose performance decreased as ANA rates increased. The predictive capacity of the SVM approach outperformed multinomial logit (LC), latent class multinomial logit(LCML) and mixed multinomial logit (MML) models in all non-attendance rates instances, and in the most of them to the latent class mixed multinomial logit (LCMML).

Once the SVM specifications that perform better with simulated data were selected, these models were used in two empirical data sets, both from conjoint analysis experiments in combination with eye-tracking tools. The first study, with Laptops choice data, had six attributes, each one with four levels. The second study, with Coffee Makers choice data, had also six attributes but different numbers of levels each, ranging from 2 to 4, in addition to a no-choice option. As in the simulated data study, the specifications of the selected SVM approach performed better than all benchmark models in terms of out of sample choice prediction hit rate, in both empirical data set instances.

In a second study, we focused on how the consumption behavior of the first-time mothers was affected both during the pregnancy and the afterbirth periods. We conducted a causal analysis using observational data from a credit card records. We observed detailed purchases records from a supermarket and aggregated purchases information from a different external business. Our study was characterized by an important data imbalance between the control and the treatment group in a ratio of 1:178. In order to measure the impact of the unbalanced data in the treatment effect estimations, we addressed the problem using two approaches. First, we ran the causal forest algorithm considering all individuals in the control group (CF_ALL), managing the imbalance problem through the tuning

of balance parameters; and secondly, we applied a matching propensity score (CF_PS) approach to balance both groups until a 1:1 ratio. We estimated 104 causal models, one for each analyzed product category.

To compare the results of our analysis, both with balanced and unbalanced data, we used the classical differences in differences (DID) approach. The magnitudes of the average treatment effects estimations showed to be no statistically different between these three approaches. However, DID model estimations had the highest variance, and the causal forest with unbalance data (CF_ALL) model estimations the lowest. Therefore, CF_ALL was also the model that predicted a greater number of categories significantly affected by treatment, both pregnancy and birth.

When we compared the performance of the causal forest models in terms of resolution time, we obtained that the model with unbalanced data, where the control group is 178 times larger than the treatment group, took on average 66 times longer to run than the model with balanced data, where the control group is equal to the treatment group. Due to the strategic level of the decisions that could be made from the results of our study, the computing times are not a relevant variable to compare the performance of the models. For this reason, and despite the increase in computing times, we conclude that the causal forest methodology performs well in highly large and unbalanced data scenarios, mainly based on the variance estimations results.

Regarding the findings of the causal analysis, we showed that both pregnancy and birth of the first child motivate important changes in the consumption behavior of the new mothers. First-time mothers postpone the consumption of goods and services oriented to their personal wellness and leisure (entertainment, personal care, beauty and women's clothing), and replace it by goods and services for the well-being of the baby and the family (baby products, medical services, medications and insurance). They also reduce their activity outside the house (transport, gasoline, number of visits to the supermarket). Related with food categories, we found a negative effect in ready-to-eat products, canned food, and refrigerated mass and pasta, in the deli category and in the cheese category. We also observed a surprisingly positive treatment effect in photography and Christmas categories, and an important negative treatment effect on pets articles.

We have important opportunities both to extend and to improve our current analysis in both projects.

In the attribute non-attendance context, we will continue investigating fairer procedures to simulate data in SVMs choice models, in order to avoid the underestimation of the success prediction measures. We are also working to find alternatives criteria to select parameters in the calibration process, in order to improve the choice prediction capability of the model.

In the consumer behavior changes triggered by a life event context, the future work is aimed at building household analysis, being able to complement the purchases made by the mother, and also the father of the future baby in order to avoid possible gaps and data bias in the mother's sales records.

Finally, we will extend our analysis in several ways. First, it would be interesting to compare the magnitude of the effects in the case of mothers having a second (or more) baby. This would allow us to learn if there are significant differences between first-time and further motherhood experiences. Secondly, analyzing other life events, such as marriage, home moves, an entry in the labor market, children starting at school, among others.

Bibliography

Abadie, A. (2005). Semiparametric Difference-in-Differences Estimators. *The Review of Economic Studies*, 72(1):1–19.

Alemu, M. H., Mørkbak, M. R., Olsen, S. B., and Jensen, C. L. (2013). Attending to the reasons for attribute non-attendance in choice experiments. *Environmental and resource economics*, 54(3):333–359.

Andreasen, A. R. (1984). Life Status Changes and Changes in Consumer Preferences and Satisfaction Life Status Changes and Changes in Consumer Preferences and Satisfaction. 11(3):784–794.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *Ann. Statist.*, 47(2):1148–1178.

Balbontin, C., Hensher, D. A., and Collins, A. T. (2017). Integrating attribute non-attendance and value learning with risk attitudes and perceptual conditioning. *Transportation Research Part E: Logistics and Transportation Review*, 97:172–191.

Bradley, P. S. and Mangasarjan, O. L. (1998). Feature Selection via Concave Minimization and Support Vector Machines. *Proceedings of the Fifteenth International Conference on Machine Learning (ICML '98)*, (6):82–90.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA.

Campbell, D., Hensher, D. A., and Scarpa, R. (2011). Non-attendance to attributes in environmental choice analysis: a latent class specification. *Journal of environmental planning and management*, 54(8):1061–1076.

Campbell, D., Hensher, D. A., and Scarpa, R. (2012). Cost thresholds, cut-offs and sensitivities in stated choice analysis: Identification and implications. *Resource and Energy Economics*, 34(3):396–411.

Campbell, D., Hutchinson, W. G., and Scarpa, R. (2008). Incorporating discontinuous preferences into the analysis of discrete choice experiments. *Environmental and resource economics*, 41(3):401–417.

Campbell, D., Lorimer, V., Aravena, C., and Hutchinson, G. (2010). Attribute processing in environmental choice analysis: implications for willingness to pay. *Agricultural Economics Society Annual Conference*, (January).

Chapelle, O. (2002). Choosing Multiple Parameters for Support Vector Machines. *Machine Learning*, (46):131–159.

Chen, C., Liaw, A., and Breiman, L. (2004). Using Random Forest to Learn Imbalanced Data. *Discovery*, (1999):1–12.

Collins, A. T. (2012). *Attribute nonattendance in discrete choice models: measurement of bias, and a model for the inference of both nonattendance and taste heterogeneity.* PhD thesis.

Collins, A. T. and Hensher, D. A. (2015a). The influence of varying information load on inferred attribute non-attendance. In *Bounded Rational Choice Behaviour: Applications in Transport*, pages 73–94. Emerald Group Publishing Limited.

Collins, A. T. and Hensher, D. A. (2015b). The influence of varying information load on inferred attribute non-attendance. In Rasauli, S. and Timmermans, H., editors, *Bounded Rational Choice Bahaviour*, chapter Chapter 4, pages 73–94. Emerald, UK, first edition.

Cui, D. and Curry, D. (2005). Prediction in Marketing Using the Support Vector Machine. *Marketing Science*, 24(4):595–615.

Dohrenwend, B. S. and Dohrenwend, B. P., editors (1974). *Stressful life events: Their nature and effects.* John Wiley & Sons, Oxford, England.

Drechsler, J. (2010). Using support vector machines for generating synthetic datasets. In Domingo-Ferrer, J. and Magkos, E., editors, *Privacy in Statistical Databases*, pages 148–161, Berlin, Heidelberg. Springer Berlin Heidelberg.

Erdem, S., Campbell, D., and Hole, A. R. (2013). Attribute-level non-attendance in a choice experiment investigating preferences for health service innovations. (44):1–26.

Evgeniou, T., Boussios, C., and Zacharia, G. (2005). Generalized Robust Conjoint Estimation. *Marketing Science*, 24(3):415–429.

George, E. I. and McCulloch, R. E. (1993). Variable Selection via Gibbs Sampling. *Journal of the American Statistical Association*, 88(423):881–889.

George, E. I. and McCulloch, R. E. (1997). APPROACHES FOR BAYESIAN VARIABLE SELECTION. *Statistica Sinica*, 7(2):339–373.

Gilbride, T. J., Allenby, G. M., and Brazell, J. D. (2006). Models for Heterogeneous Variable Selection. *Journal of Marketing Research (JMR)*, 43(3):420–430.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, 46(1):389–422.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction.* Springer, 2 edition.

Hauser, J. R., Toubia, O., Evgeniou, T., Befurt, R., and Dzyabura, D. (2010). Disjunctions of conjunctions, cognitive simplicity, and consideration sets. *Journal of Marketing Research*, 47(3):485–496.

Hensher, D. (2007). Reducing sign violation for vtts distributions through endogenous recognition of an Individual's attribute processing strategy. *International Journal of Transport Economics / Rivista internazionale di economia dei trasporti*, 34(3):333–349.

Hensher, D. A. (2006). How do respondents process stated choice experiments? attribute consideration under varying information load. *Journal of applied econometrics*, 21(6):861–878.

Hensher, D. A., Collins, A. T., and Greene, W. H. (2013). Accounting for attribute non-attendance and common-metric aggregation in a probabilistic decision process mixed multinomial logit model: A warning on potential confounding. *Transportation*, 40(5):1003–1020.

Hensher, D. A. and Greene, W. H. (2010). Non-attendance and dual processing of common-metric attributes in choice analysis: A latent class specification. *Empirical Economics*, 39(2):413–426.

Hensher, D. A., Rose, J., and Greene, W. H. (2005). The implications on willingness to pay of respondents ignoring specific attributes. *Transportation*, 32(3):203–222.

Hensher, D. A., Rose, J. M., and Greene, W. H. (2012). Inferring attribute non-attendance from stated choice data: Implications for willingness to pay estimates and a warning for stated choice experiment design. *Transportation*, 39(2):235–245.

Hess, S. and Hensher, D. A. (2010). Using conditioning on observed choices to retrieve individual-specific attribute processing strategies. *Transportation Research Part B: Methodological*, 44(6):781–790.

Hess, S. and Hensher, D. A. (2013). Making use of respondent reported processing information to understand attribute importance: A latent variable scaling approach. *Transportation*, 40(2):397–412.

Hess, S. and Rose, J. M. (2007). A latent class approach to recognising respondents' information processing strategies in SP studies. *Oslo Workshop on Valuation Methods in Transport Planning, Oslo*.

Hess, S., Stathopoulos, A., Campbell, D., O'Neill, V., and Caussade, S. (2013). It's not that I don't care, I just don't care very much: Confounding between attribute non-attendance and taste heterogeneity. *Transportation*, 40(3):583–607.

Hole, A. R., Kolstad, J. R., and Gyrd-Hansen, D. (2013). Inferred vs. stated attribute non-attendance in choice experiments: A study of doctors' prescription behaviour. *Journal of Economic Behavior and Organization*, 96:21–31.

Kohli, R. and Jedidi, K. (2007). Representation and inference of lexicographic preference models and their variants. *Marketing Science*, 26(3):380–399.

Koschate-Fischer, N., Hoyer, W. D., Stokburger-Sauer, N. E., and Engling, J. (2017). Do life events always lead to change in purchase? The mediating role of change in consumer innovativeness, the variety seeking tendency, and price consciousness. *Journal of the Academy of Marketing Science*, 46(3):516–536.

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232.

Maldonado, S., Montoya, R., and López, J. (2017). Embedded heterogeneous feature selection for conjoint analysis: A SVM approach using L1 penalty. *Applied Intelligence*, 46(4):775–787.

Maldonado, S., Montoya, R., and Weber, R. (2015). Advanced conjoint analysis using feature selection via support vector machines. *European Journal of Operational Research*, 241(2):564–574.

Maldonado, S. and Weber, R. (2009). A wrapper method for feature selection using Support Vector Machines. *Information Sciences*, 179(13):2208–2217.

Maldonado, S., Weber, R., and Basak, J. (2011). Simultaneous feature selection and classification using kernel-penalized support vector machines. *Information Sciences*, 181(1):115–128.

Mariel, P., Hoyos, D., and Meyerhoff, J. (2013). Stated or inferred attribute non-attendance? A simulation approach. *Economia Agraria y Recursos Naturales*, 13(1):51–67.

Mathur, A., Moschis, G. P., and Lee, E. (2008). A longitudinal study of the effects of life status changes on changes in consumer preferences. *Journal of the Academy of Marketing Science*, 36(2):234–246.

Meissner, M., Musalem, A., and Huber, J. (2016). Eye tracking reveals processes that enable conjoint choices to become increasingly efficient with practice. *Journal of Marketing Research*, 53(1):1–17.

Miranda, J., Montoya, R., and Weber, R. (2005). Linear Penalization Support Vector Machines for Feature Selection. In Pal, S. K., Bandyopadhyay, S., and Biswas, S., editors, *Pattern Recognition and Machine Intelligence*, pages 188–192, Berlin, Heidelberg. Springer Berlin Heidelberg.

Rao, V. R. et al. (2014). *Applied conjoint analysis*. Springer.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.

Rose, J. M., Hensher, D. A., and Greene, W. H. (2005). Recovering costs through price and service differentiation: Accounting for exogenous information on attribute processing strategies in airline choice. *Journal of Air Transport Management*, 11(6):400–407.

Rosenbaum, P. R. and Rubin, D. B. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45(2):212–218.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Scarpa, R., Gilbride, T. J., Campbell, D., and Hensher, D. (2009). Modelling attribute non-attendance in choice experiments for rural landscape valuation. *European Review of Agricultural Economics*, 36(2):151–174.

Selin, H. (2009). *Childbirth Across Cultures: Ideas and Practices of Pregnancy, Childbirth and the Postpartum*, volume 5.

Sevin, E., Ladwein, R., Sevin, E., and Soman, D. (2008). To Start Being ... . The Anticipation of a Social Role Through Consumption in Life Transition : The Case of the First-Time Pregnancy. 35:325–332.

Shalit, U., Johansson, F. D., and Sontag, D. (2017). Estimating individual treatment effect: Generalization bounds and algorithms. *34th International Conference on Machine Learning, ICML 2017*, 6:4709–4718.

Swait, J. and Adamowicz, W. (2001). The Influence of Task Complexity on Consumer Choice: A Latent Class Model of Decision Strategy Switching. *Journal of Consumer Research*, 28(1):135–148.

Taddy, M., Gardner, M., Chen, L., and Draper, D. (2016). A nonparametric bayesian analysis of heterogenous treatment effects in digital experimentation. *Journal of Business & Economic Statistics*, 34(4):661–672.

Thoits, P. A. (1995). Stress, coping, and social support processes: Where are we? what next? *Journal of Health and Social Behavior*, pages 53–79.

Tian, L., Alizadeh, A. A., Gentles, A. J., and Tibshirani, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532. PMID: 25729117.

Toubia, O., Hauser, J. R., and Simester, D. I. (2004). Polyhedral Methods for Adaptive Choice-Based Conjoint Analysis. *Journal of Marketing Research*, 41(1):116–131.

Train, K. and Sonnier, G. (2005). *Mixed Logit with Bounded Distributions of Correlated Partworths*, pages 117–134. Springer Netherlands, Dordrecht.

Turjeman, D. and Feinberg, F. M. (2019). When The Data Are Out : Measuring Behavioral Changes Following a Data Breach. (July).

van der Laan, M. J. and Rubin, D. (2006). Targeted maximum likelihood learning. *International Journal of Biostatistics*, 2(1).

Van Loo, E. J., Nayga Jr, R. M., Campbell, D., Seo, H.-S., and Verbeke, W. (2018). Using eye tracking to account for attribute non-attendance in choice experiments. *European Review of Agricultural Economics*, 45(3):333–365.

Vapnik, V. N. and Chervonenkis, A. Y. (1971). *On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities*, pages 11–30. Springer International Publishing, Cham.

Wager, S. and Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Weisberg, H. I. and Pontes, V. P. (2015). Post hoc subgroups in clinical trials: Anathema or analytics? *Clinical Trials*, 12(4):357–364.

Wheaton, B. (1990). Life transitions, role histories, and mental health. *American Sociological Review*, 55(2):209–223.

Wilkes, R. E. (1995). Household Life-Cycle Stages, Transitions, and Product Expenditures. *Journal of Consumer Research*, 22(1):27–42.

Yang, L., Toubia, O., and De Jong, M. G. (2015). A bounded rationality model of information search and choice in preference measurement. *Journal of Marketing Research*, 52(2):166–183.

Yegoryan, N., Guhl, D., and Klapper, D. (2019). Inferring attribute non-attendance using eye tracking in choice-based conjoint analysis. *Journal of Business Research*, (December 2018):1–15.