FUNDAÇÃO GETULIO VARGAS
ESCOLA DE ECONOMIA DE SÃO PAULO

MARIANA FURTADO BORELI

ESSAYS ON INTERNATIONAL TRADE AND INSURANCE MARKETS

SÃO PAULO
2021

MARIANA FURTADO BORELI

ESSAYS ON INTERNATIONAL TRADE AND INSURANCE MARKETS

Tese apresentada à Escola de Economia de São Paulo da Fundação Getulio Vargas como requisito para a obtenção do título de Doutora em Economia. Campo de conhecimento: Microeconomia Aplicada

Supervisor: João Paulo Cordeiro de Noronha Pessoa

Co-supervisor: Cristine Campos de Xavier Pinto

SÃO PAULO

2021

Boreli, Mariana Furtado.

Essays on international trade and insurance markets / Mariana Furtado Boreli. - 2021.

79 f.

Orientador: João Paulo Cordeiro de Noronha Pessoa.
Co-orientador: Cristine Pinto.

Tese (doutorado CDEE) – Fundação Getulio Vargas, Escola de Economia de São Paulo.

1. Comércio internacional. 2. Brasil - Relações econômicas exteriores - China. 3. China - Relações econômicas exteriores - Brasil. 4. Risco (Economia). 5. Seguro de automóveis. I. Pessoa, João Paulo Cordeiro de Noronha. II. Pinto, Cristine. III. Tese (doutorado) – Escola de Economia de São Paulo. IV. Fundação Getulio Vargas. V. Título.

CDU 339.5

Ficha Catalográfica elaborada por: Isabele Oliveira dos Santos Garcia CRB SP-010191/O
Biblioteca Karl A. Boedecker da Fundação Getulio Vargas - SP

MARIANA FURTADO BORELI

ESSAYS ON INTERNATIONAL TRADE AND INSURANCE MARKETS

Tese apresentada à Escola de Economia de São Paulo da Fundação Getulio Vargas como requisito para obtenção do título de Doutor em Economia de Empresas.

Campo de Conhecimento:
Microecoomia Aplicada

Data de Aprovação:
___/___/_____

Banca examinadora:

Prof. Dr. João Paulo C. de Noronha Pessoa
FGV-EESP

Profa. Dra. Cristine C. de Xavier Pinto
FGV-EESP

Prof. Dr. Marcelo Fernandes
FGV-EESP

Prof. Dr. Eduardo Fonseca Mendes
FGV-EMAp

Prof. Dr. Francisco J. Moreira da Costa
FGV-EPGE

Agradecimentos

Agradeço meus orientadores, João Paulo e Cristine, pelas inúmeras conversas e discussões. Também agradeço a Professora Élia Matsumoto que gentilmente me ajudou em diversas etapas da construção dos artigos com feedbacks e questionamentos importantes. Aos meus pais e meu irmão, Anísio, Angélica e Anisinho, muito obrigada pela paciência, amor e carinho.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Acknowledgements

I thank my advisors, João Paulo and Cristine, for the countless conversations and discussions. I also thank Professor Élia Matsumoto who kindly helped me in several stages of the construction of these papers with important feedback and questions. To my parents and my brother, Anísio, Angélica and Anisinho, thank you for your patience, love and affection.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Resumo

Esta tese é composta por três ensaios em microeconomia aplicada. O primeiro ensaio investiga o impacto do comércio entre o Brasil e a China nos consumidores Brasileiros. Os outros ensaios estão relacionados ao mercado de seguro de automóveis e compartilham a mesma base de dados.

No primeiro capítulo, mensuramos o impacto das importações Brasileiras da China ("Choque da China") entre 2002 e 2008. Para estimarmos esse impacto utilizamos como variável instrumental a razão entre a diferença das exportações da China para todos os países de baixa renda (exceto o Brasil) ao longo do tempo e o gasto total no Brasil em 2002. Os resultados mostram que quando comparamos um setor que teve a mediana do impacto das importações com um setor que não teve impacto, os preços no setor mediano cresceram menos 0.8 pontos percentuais entre 2002 e 2008.

No segundo capítulo, estimamos o limite inferior do coeficiente de aversão ao risco a partir dos dados fornecidos pela Superintendência de Seguros Privados. Essa estimação baseia-se apenas nos contratos de equilíbrio, assim o nosso *framework* pode ser replicado em um conjunto maior de dados. Além disso, o coeficiente de aversão ao risco que usamos não depende da forma funcional da função de utilidade e grande parte da sua estimação é feita com modelos flexíveis.

No terceiro capítulo, criamos um sistema de busca capaz de prever contratos de seguros de automóveis com base nas informações do segurado (idade, gênero) e do veículo (modelo, ano). Esse sistema é benéfico para os consumidores porque a partir da nossa previsão eles conseguem saber qual o contrato médio disponível no mercado. Isso também é útil para as seguradoras, pois a partir do nosso modelo empírico elas podem inferir se a estratégia de contratos que elas usam são consistentes com o mercado e podem verificar como a metodologia que elas utilizam para encontrar o contrato médio se com compara com a nossa.

Palavras-chaves: Comércio Internacional, Choque da China, Aversão ao Risco, Previsão de Contratos de Seguro.

Abstract

This thesis consists of three essays in applied microeconomics. The first essay analyzes the impact of trade between Brazil and China on Brazilian consumers. The other essays are related to the auto insurance market and share a single database.

In the first chapter, we measure the impact of Chinese imports to Brazil ("China Shock") on Brazilian consumers between 2002 and 2008. To estimate the impact, we use as an instrumental variable the ratio between the difference in total exports from China to all low-income countries (other than Brazil) over time and Brazilian total expenditure in 2002. Prices in a sector with the median impact on imports grew by 0.8 percentage point less between 2002 and 2008, compared to prices in a sector with no impact on imports.

In the second chapter, we investigate risk aversion in the Brazilian automobile insurance market. Using the data from the Superintendência de Seguros Privados, we estimate the lower bound of the risk aversion coefficient. We use only equilibrium contracts, in our estimate, so our framework can be replicated in a larger set of data. Additionally, our risk aversion coefficient does not depend on the functional form of the utility function, and most of its estimation is flexible.

In the third chapter, we create a quote request system capable of predicting auto insurance contracts based on information from the insured (age, gender) and the vehicle (model, year). This system benefits consumers, who can find out the average contract available on the market. It is also useful for insurance companies because they can infer from our empirical model whether their contract strategy is consistent with our prediction, and how their method to obtain the average contract compares to ours.

Key-words: International Trade, China Shock, Risk Aversion, Insurance Contract Forecasting.

List of Figures

List of Tables

Contents

1 What is the effect of imports from China on Brazilian consumers?

1.1 Introduction

Over the years, China has become one of the most powerful players in international trade. Additionally, trade between China and Brazil has increased rapidly in the last two decades. Despite this growth, few studies address the following question: What are the consequences of trade with China on prices paid by consumers in Brazil? An increase in Chinese imports to Brazil can potentially reduce domestic prices (a positive effect for consumers).

This question is relevant because of the main characteristic of the trade between these countries: the majority of Brazilian exports to China are commodities, and most Chinese exports to Brazil are manufactured goods. This feature is not limited to Brazil. Hence, studying the effects of the "China Shock" on Brazilian consumers is important not only because of the rapid growth of Brazilian imports from China, but also it allows us to investigate its impact on a developing country (trades commodities-for-manufactures). Most of the literature estimates the "China Shock" on developed countries.

In this paper, we estimate the impact of the "China Shock" on Brazilian consumer prices between 2002 and 2008. We take the theoretical framework developed by Arkolakis, Costinot, and Rodríguez-Clare (2012), and we associate price variations to changes in Brazilian domestic share expenditure in different sectors.

Additionally, we decompose the domestic share expenditure into an import effect (changes in total imports) and a local effect (variations in the amount spent on locally-produced goods). The idea is that more imports (greater competition and/or more products) potentially lead to price reductions.

To identify the effect of the rapid growth of Brazilian imports from China on domestic prices, we need an instrumental variable (IV) to isolate the part of the import effect associated with Chinese supply changes. We follow Autor, Dorn, and Hanson (2013), and construct our IV from the ratio between the change over time in exports from China to all low-income countries other than Brazil and Brazilian expenditure in 2002.

Our IV potentially identifies the "China Shock" because this instrument may be associated with the changes in Brazilian imports from China since all low-income countries probably faced a similar shock from Chinese exports. Moreover, this IV may not affect Brazilian prices directly, because domestic prices in Brazil are not determined by price

changes in other low-income countries.

Based on a two-stage least squares (2SLS) estimate of the regression of Brazilian domestic prices on import effect, using the IV described above, and covariates that capture supply and demand changes in Brazil, we find that prices in a sector with a median import effect grew by 0.8 percentage point less than a sector with no imports from China between 2002 and 2008.

We contribute to a growing literature on the implications of China's rapid growth. We estimate the impact of the China supply shock on Brazilian domestic prices between 2002 and 2008. Additionally, our discussion may be relevant for other developing countries as well.

This paper relates to the strand of literature that investigates the "China Shock" through the expenditure channel such as Amiti et al. (2020) and Jaravel and Sager (2018). Bai and Stumpner (2019) measure the impact of Chinese imports on US consumers between 2004 and 2015, and find a positive effect for consumers.

Our study is closely related to that of Bai and Stumpner (2019). We use an empirical strategy similar to theirs to estimate the consequences of the "China Shock" for Brazilian consumers. However, there are two main differences. First, we investigate the effect of imports from China on a developing country. Second, Bai and Stumpner (2019) focus on the relationship between price indices and variations in domestic share expenditure, whereas we decompose this share into changes in imports and expenditure on locally-produced goods, which allows us to investigate those effects separately.

Moreover, our paper relates to another strand of the literature that focuses on the distributional consequences of trade through price effects. Most papers focus on developed countries (FAJGEBAUM; KHANDELWAL, 2016; HOTTMAN; MONARCH, 2018; BORUSYAK; JARAVEL, 2018). As for developing countries, Porto (2006) analyzes the distributional effects of Mercosur, and Faber (2014) investigates the price effect of NAFTA on the cost of living in Mexico.

He (2018) generalizes the structural model proposed by Fajgelbaum and Khandelwal (2016). Also, the author investigate the effects of trade liberalization on wage inequality through expenditure and earnings channels. He (2019) examines the distributional effects of the "China Shock" on Mexico and Brazil. Unlike He (2019), we use reduced-form estimates to explore the relationship between price changes and import effect.

The main contribution of this paper is the estimate, based on reduced-form regressions, of the effect of the growth of Brazilian imports from China (commodities-for-manufactures trade) on Brazilian domestic prices between 2002 and 2008.

The remainder of the paper proceeds as follows. In Section 1.2, we provide the the-

oretical motivation. Section 1.3 contains a description of the data. Our empirical strategy is covered in Section 1.4. In Section 1.5, we present the results. Section 1.6 concludes.

1.2 Theoretical Motivation

Considering a large set of models, Arkolakis, Costinot, and Rodriguez-Clare (2012) investigate the domestic welfare effects of foreign shocks. They show that, for each country, these effects can be summarized by the change in the share of expenditure on domestically-produced goods (DSE). Similarly to Bai and Stumpner (2019), we explore this relationship at the sector level as follows:

$$\Delta \log(P_i) \propto \frac{1}{\theta} \Delta(DSE_i), \tag{1.1}$$

where P_i is the sector price, $DSE_i \equiv \dfrac{E_{it} - M_{it}}{E_{it}}$, E_{it} is total expenditure, M_{it} is total imports, and θ is trade elasticity.

Additionally, we decompose the domestic share expenditure to find that

$$\Delta log(P_i) \propto -\frac{1}{\theta} \overbrace{FSE_{it_0} \frac{\Delta M_i}{M_{it_0}}}^{\text{Import effect}}, \tag{1.2}$$

and

$$\Delta log(P_i) \propto \frac{1}{\theta} \underbrace{FSE_{it_0} \frac{\Delta Z_i}{Z_{it_0}}}_{\text{Local effect}}, \tag{1.3}$$

where $FSE_{it_0} \equiv 1 - DSE_{it_0}$ is the foreign share expenditure in the initial period, and $Z_{it} \equiv E_{it} - M_{it}$ is expenditure on locally-produced goods.[1]

This decomposition is relevant because it allows us to identify each effect separately. The idea behind Equation (1.2) is that, as imports rise, more products are available to consumers and/or domestic production faces more competition, so domestic prices decrease. Equation (1.3) shows that as consumers spend more on locally-produced goods, this greater demand drives higher prices.[2]

1.2.1 Category-level inflation rate

Following Bai and Stumpner (2019), we compute category-level inflation rates starting from a non-symmetric CES consumption function:

$$C_{it} = \left(\sum_k a_i^{k\,\frac{1}{\sigma}} c_{it}^{k\,\frac{\sigma-1}{\sigma}} \right)^{\frac{\sigma}{\sigma-1}},$$

[1] We can rewrite the import effect in Equation (1.2) as $\frac{\Delta M_i}{E_{it_0}}$.

[2] It may be the case that we have reverse causality: higher prices could increase expenditure on locally-produced goods.

where k is a product that belongs to i, c_{it}^k is the total consumption of k, a_i^k is the unobserved product quality (constant over time), and σ is the elasticity of substitution.

Additionally, this consumption bundle has the following ideal price index:

$$P_{it} = \left(\sum_k a_i^k p_{it}^{k\,1-\sigma} \right)^{\frac{1}{1-\sigma}}.$$

From Sato (1976) and Vartia (1976) the inflation for a constant basket of goods can be written as

$$\frac{P_{it}}{P_{it-1}} = \prod_k \left(\frac{p_{it}^k}{p_{it-1}^k} \right)^{\omega_{it}^k}, \tag{1.4}$$

where w_{it}^k is a variety-specific weight which is a function of expenditure shares.[3] Moreover, by taking the logarithm of both sides of Equation (1.4), we see that inflation for each sector is determined by a geometric sum of price changes of its products weighted by w_{it}^k. We use this calculation to create the price inflation measure used in this paper.

1.3 Data and Descriptive Statistics

1.3.1 Data Sources

Our empirical analysis requires four data sources: (i) the Pesquisa de Orçamento Familiar (POF), (ii) the Base pour l'Analyse du Commerce International (BACI); (iii) the Índice de Preços ao Consumidor-Semanal (IPC-S); and (iv) the Relação Anual de Informações Sociais (RAIS).

The POF is a national survey, from the Instituto Brasileiro de Geografia e Estatística (IBGE), conducted with Brazilian households to investigate their consumption across many items. The most recent rounds cover 1995-1996, 2002-2003, 2008-2009, and 2017-2018.[4] In this paper, we restrict our attention to the POFs from 2002-03 and 2008-09 because of their closer correspondence between product categories. Approximately 8,000 and 13,000 tradable goods are in each survey, respectively. The POF provides information about the total expenditure.

Our empirical strategy also requires data regarding trade among countries, which is available through the BACI managed by the Centre d'Études Prospectives et d'Informations

[3] Let $s_{it}^k \equiv \frac{p_{it}^k c_{it}^k}{\sum_k p_{it}^k c_{it}^k}$ be the expenditure share,

$$\omega_{it}^k \equiv \frac{\frac{s_{it}^k - s_{it-1}^k}{log(s_{it}^k) - log(s_{it-1}^k)}}{\sum_{k \in I_{it}} \frac{s_{it}^k - s_{it-1}^k}{log(s_{it}^k) - log(s_{it-1}^k)}}.$$

[4] The microdata from POF 2017-2018 is not available to the public.

Internationales (CEPII). The BACI provides information on imports and exports (values and quantities) for more than 200 countries based on the Harmonized System (HS) 6-digit classification codes of the International Trade Administration of the World Customs Organization. Also, it collects data from 1995 for approximately 5,000 products.

Another source of data is the IPC-S. It is an index used to analyze the purchasing power of consumers, managed by the Instituto Brasileiro de Economia (IBRE). Also, the IPC-S has seven macro-categories: food, housing, clothing, health and personal care, education and recreation, transport, and others. This database has the prices of products surveyed in São Paulo, Rio de Janeiro, Belo Horizonte, Salvador, Recife, Porto Alegre, and Brasília. The data runs from January 2003. The last period considered is March 2009, which coincides with the end of POF 2008-2009.

Two other sets of variables are also important: those that capture demand and supply shocks in Brazil. On the demand side, we use the information regarding the "chefe de família" from the POF 2002-2003. This is the person responsible for the family's expenditure and for whom socioeconomic data is available.

On the supply side, we use data regarding the labor market available from the RAIS. This is the most comprehensive Brazilian dataset regarding labor activities in the formal sector. It covers information about employees such as demographics, income, and occupation. Also, it provides variables associated with changes in Brazilian production. The data presented by the RAIS is aggregated at the Classificação Nacional de Atividades Econômicas (CNAE) level.

Due to their variety, we need to establish a common correspondence between all data sources. Therefore, we create a new classification such that for each category of the IPC-S, we find the corresponding items in the POF and the HS-6 codes. If any item belongs to more than one classification, we aggregate them, and if necessary, we also aggregate at the IPC-S level. As a result, we obtain 129 sectors.[5] Since the CNAE code is broader than our classification, we duplicate the data from RAIS for the sectors within each CNAE.

1.3.2 Descriptive Statistics

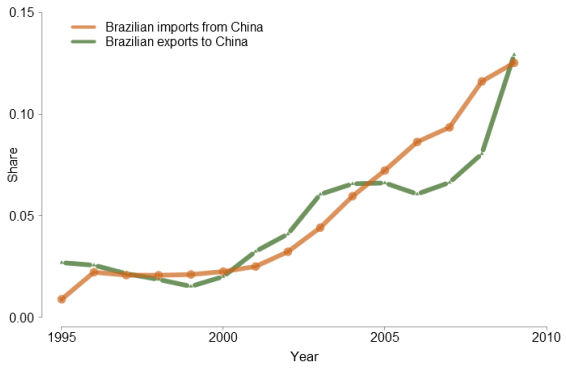
Trade between Brazil and China has grown considerably over the years. In fact, in 2002, 3% of Brazilian imports came from China, and 4% of its exports were sent to China. By 2008, those numbers changed to 12% and 8%, respectively. This fact is reflected in Panel (a) of Figure 1.1.

A key characteristic of the trade between these countries is that most of Brazilian

[5] A full list can be seen in the online appendix at <https://www.dropbox.com/scl/fi/w5pqa4vtw3ygltwzq536u/classification.xlsx?dl=0&rlkey=2298y2116hd64oh9dwhrcpiik>.

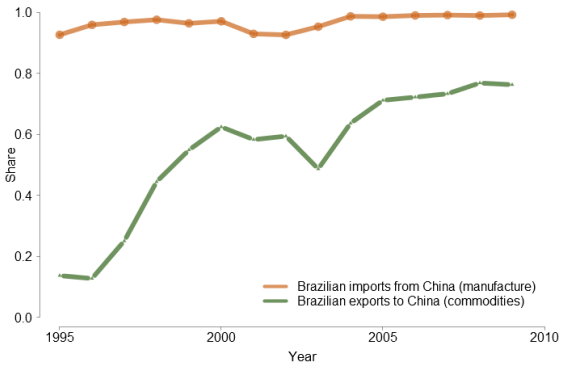
exports to China are commodities, most Brazilian imports from China are manufactured goods. According to Panel (b) of Figure 1.1, between 2002 and 2008 the proportion of manufactures among Brazilian imports from China increased from 93% to 99%. In the same time frame, the share of commodities in Brazilian exports to China grew from 59% to 77%.[6]

Figure 1.1 – Characteristics of the Trade between Brazil and China



Evolution of the ratio between Brazilian imports from China and total imports in Brazil; and the ratio of Brazilian exports to China to total Brazilian exports from 1995 to 2010. Source: BACI.

(a) Trade between Brazil and China

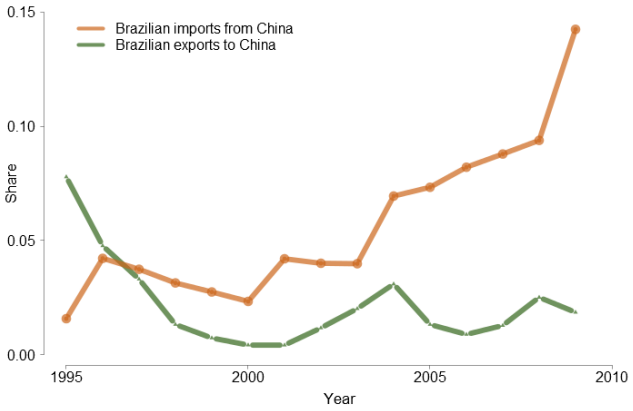


Evolution from 1995 to 2010 of the fraction of manufactured goods among Brazilian imports from China; and the share of Brazilian exports to China that are commodities. Source: BACI.

(b) Commodities-for-manufacture trade

Our analysis is not based only on imports and exports, but also takes into consideration expenditure and price indexes. Therefore, our data is restricted to sectors that are covered by all datasets described in the previous section. Figure 1.2 shows the share of Brazilian imports from (exports to) China presented in our data. From this figure, we notice a rapid growth of imports, but a smaller change in exports.

Figure 1.2 – Import and Export effects observed in our data



Evolution of the portion of Brazilian imports from China among total imports in Brazil, and the Brazilian exports to China among total exports from Brazil between 1995 and 2010. Source: BACI.

[6]Similar statistics are used by Costa et al. (2016).

The main drivers of the rapid growth of exports to China presented in Panel (a) of Figure 1.1 are the agricultural and extractivist sectors. These are led by soybeans and iron ore. These items are not observable in our data because the IPC-S does not compute their price. Thus, the export effect is different between Figures 1.1 and 1.2, and our sample complicates our efforts to identify the effect of the growth of Brazilian exports to China.[7] Therefore, our main analysis concentrates on the identification of China's supply shock.

An overview of the data is presented in Table 1.1. It shows the number of sectors, total Brazilian expenditure, total imports, total exports, imports from China, and exports to China in 2002-03 and 2008-09 for each of the seven macro-categories that compose the IPC-S. Over the years, almost every variable in every category has grown.

Table 1.1 – Descriptive statistics by IPC-S macro-categories

IPC-S macro-categories	Sectors	Expenditure		Imports		Exports		Imp. from China		Expo. to China	
		02/03	08/09	02/03	08/09	02/03	08/09	02/03	08/09	02/03	08/09
Food	69	71.74	104.90	2.40	4.83	23.62	46.75	0.05	0.31	0.42	1.69
Housing	28	23.16	43.98	0.62	1.81	2.67	3.10	0.12	0.97	0.002	0.0009
Clothing	14	23.86	49.36	0.34	1.52	4.67	3.47	0.13	0.93	0.0009	0.01
Health and Personal Care	9	12.35	24.12	0.31	0.60	0.26	0.73	0.004	0.04	0.0006	0.0001
Education and Recreation	2	2.27	4.85	0.07	0.17	0.01	0.02	0.03	0.10	0.000002	0.000004
Transport	3	5.01	12.22	0.25	0.70	1.16	5.60	0.001	0.13	0.001	0.01
Others	4	7.63	12.12	0.35	0.56	0.11	0.19	0.0001	0.002	0.00004	0.0002
Total	129	146.02	251.55	4.34	10.19	32.50	59.86	0.33	2.49	0.42	1.70

Values are expressed in billions and Brazilian Real. This table displays descriptive statistics of our data according to three sources: POF, BACI, IPC-S.

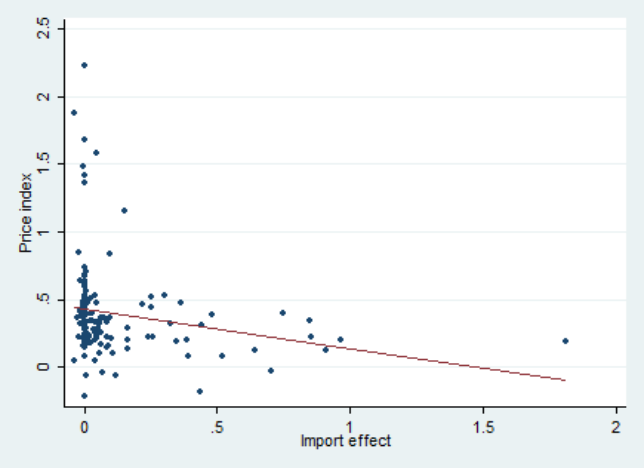
Table 1.1 shows that the macro-category Food has the highest number of sectors and the greatest share in expenditure, total imports and exports, and exports to China in both periods.[8] In contrast, most Brazilian imports from China are items in the Housing and Clothing categories. In both periods, they account for approximately 76% of the total imports from China. These statistics align with the data presented in Panel (b) of Figure 1.1.

Our goal is to discover the impact of the "China Supply Shock" on Brazilian prices. Following Arkolakis, Costinot, and Rodriguez-Clare (2012), we explore the relationship between price variations and changes in imports (import effect). Figure 1.3 suggests a negative association between these variables.

[7] The IPC-S contains soy oil and drinks made of soy.

[8] Bai and Stumpner (2019) also use a dataset which the food sector has the highest share of consumption. It accounts for around 50% of total expenditure. Their data is from Nielsen.

Figure 1.3 – Relationship between price index and import effect



The graph shows scatter plots of the import effect against price index. The line represents the estimate of a simple regression of price index on import effect. The coefficient is -0.29, the robust standard error is 0.09, and the t-statistic is -3.25.

Table 1.2 presents the five sectors with the highest growth according to price index, import and local effects. The "Tomatoes" sector has the highest price change, growing approximately 223% between 2002 and 2008. The greatest import effect is associated with the "Pans" sector, and the highest local effect is related to the sector "Fishes". Table A.1 presents summary statistics of all variables in the data. It shows that the average price index is 0.39, and the average import and local effects are 0.11.

Table 1.2 – Descriptive statistics by sectors

Sectors	Price variation	Sectors	Import effect	Sector	Local Effect
Tomatoes	2.23	Pans	1.81	Fishes	15.51
Fishes	1.88	Towels	0.96	Pans	2.24
Lemons	1.68	Pears	0.91	Pears	1.52
"Acerola", Cashew, Passion fruit	1.58	Olive oil	0.86	Chester	1.43
Potatoes	1.48	Canned olives	0.85	Fans	0.73

This table shows the name of the sectors that have the highest price variation, import effect and local effect, and the respective values. Sources: POF, CEPII BACI, IPC-S.

1.4 Empirical Strategy

Our main goal is to identify the effect of Brazil's imports from China on domestic prices. Our identification strategy follows that of Bai and Stumpner (2019). First, we specify a baseline model that relates price changes and import effect, as presented by Equation (1.2). Then, we use an instrumental variable to isolate the part of the import effect associated with China's supply shock.

1.4.1 Baseline Specifications and Controls

Our baseline specification considers the relationship between import effect and Brazilian price variations, as presented in Equation (1.2). Changes in imports and domestic prices are not led entirely by international trade. They are also influenced by variations in Brazilian supply and demand. For instance, a positive shock in wages can increase demand for imported goods. Therefore, to isolate the part of the change in imports specifically related to foreign shocks, we need to control for supply and demand shocks in Brazil. Otherwise, we would over- or underestimate the true effect. Therefore, a set of covariates at the sector level from 2002 is required: x_i. From the demand side, our covariates are average income and income growth. On the supply side, we consider age, hours worked, and wages as covariates. Hence, our baseline specification is:

$$\Delta log(P_i) = \alpha + \beta_1 \frac{\Delta M_i}{E_{it_0}} + x_i^{'}\gamma + \epsilon_i. \tag{1.5}$$

1.4.2 Instrumental Variable

Our goal is to identify the causal effect of the "China Shock" on Brazilian prices. Since the import effect is not entirely driven by China, an instrumental variable is required. We use an IV to isolate the part of the import effect that is associated with supply changes in China.

The main methodology to construct instrumental variables to identify the "China Shock" was proposed by Autor, Dorn, and Hanson (2013). Based on their method, our IV is the difference in Chinese exports to all low-income countries besides Brazil in sector i between 2002 and 2008, $\Delta\overline{X}_i$, divided by Brazilian expenditure in sector i at initial period t_0, as presented below:[9,10]

$$IV_{\text{(import effect)}} \equiv \frac{\Delta\overline{X}_i}{E_{it_0}}. \tag{1.6}$$

The "China Shock" did not occur only in Brazil; other countries have been affected by the growth of exports from China. These effects might be similar across countries. Therefore, changes in China's exports to other low-income countries and variations in Brazilian imports from China might be correlated. So, our IV and the import effect may be correlated. Since the variable presented in Equation (1.6) excludes Brazil, our IV captures the part of the variation in Brazilian imports that is related to supply changes in China, but may not correlate with Brazilian shocks.

Additionally, the IV should affect Brazilian prices solely through changes in Brazilian imports from China considering the sectors available in our data. For example, the

[9] Low-income countries follow the World Bank classification.

[10] We also use the instrumental variable proposed by Costa et al. (2016). This IV is interesting because it accounts for world price changes. The results are similar as presented in Appendix A.8

fact that Mexico increases its imports from China should not affect Brazilian consumers because local shocks in Mexico should not change prices globally or in other low-income countries. Summary statistics of the IV can be seen in Table A.2.

As explored by Bai and Stumpner (2019), the main potential problem with our identification strategy is that Chinese supply shocks may be correlated with those in the rest of the world (i.e. China and the rest of the world could experience productivity gains). If these shocks are positively (negatively) related, we could overestimate (underestimate) the "China Shock."

If these shocks are related, we would expect that supply changes in China would be accompanied by changes in Brazilian imports from the rest of the world (ROW), not only imports from China. Hence, to verify if Chinese and ROW shocks are correlated, we run each variable of the following decomposition on our instrument:

$$\frac{\Delta M_i}{E_{it_0}} = \frac{\Delta M_{i,CHN}}{E_{it_0}} + \frac{\Delta M_{i,ROW}}{E_{it_0}}. \tag{1.7}$$

Equation (1.7) shows that we can decompose the import effect into changes in Chinese imports to Brazil divided by total expenditure (China effect), $\frac{\Delta M_{i,CHN}}{E_{it_0}}$, and the difference in Brazilian imports from the rest of the world (ROW effect), divided by expenditure in the initial period, $\frac{\Delta M_{i,ROW}}{E_{it_0}}$. Therefore, we expect the IV for the import effect to only be strongly correlated with the China effect to identify the "China Shock".

1.5 Results

1.5.1 Main results

Our goal is to identify the Chinese supply shock on Brazilian prices. All results consider winsorized values of the variables of interest and IV.[11] Table 1.3 shows the results of sector-level regressions of the difference of the logarithm of prices between 2002 and 2008 on import effect, considering different specifications and estimation methods.

In Column (1), we run a regression of price index on import effect. From this estimate, we conclude that prices decreased in sectors with higher import penetration. Column (2) displays the same regression as Column (1), but it is estimated by 2SLS using the IV presented in Equation (1.6). As a result, we identify China's supply shock on Brazilian prices, and find that it has a positive effect on consumers (the coefficient is negative and statistically different from zero).

Columns (3) and (4) show 2SLS estimates with sector-level expenditure in the initial period as weight. Both columns also indicate a negative effect on prices of China's

[11] Winsorized values mean that we replace the lowest and highest values of the variable by the second-lowest and highest values, respectively. Non-winsorized in Appendix A present a weaker IV.

supply shock. Unlike the other columns, the last two columns consider covariates. Our preferred specification, Column (4), shows that prices in a sector with median import effect grew by approximately 0.8 ppt less than prices in a sector with no import effect, between 2002 and 2008. In this period, the median value of our price index is 33%. The F-statistic of the first-stage of the regression presented in Column (4) is 24.39.

Comparing the OLS and 2SLS estimates, we see that the Chinese supply shock had a greater impact on domestic prices than the changes in Brazilian imports from all countries. Also, Chinese imports to Brazil had a greater impact on the sectors with higher expenditure in 2002, as shown by the size of the coefficients associated with the import effect in Columns (3) and (4).

Table 1.3 – OLS and 2SLS estimates of the import effect, winsorized values

	(1) OLS	(2) 2SLS	(3) 2SLS	(4) 2SLS
Import effect	-0.362***	-0.625***	-0.771***	-0.802***
	(0.0917)	(0.217)	(0.261)	(0.263)
Weight	No	No	Yes	Yes
Controls	No	No	No	Yes
N	129	129	129	129
1st-stage F-stat		32.53	39.88	24.39

The dependent variable is the difference between the logarithm of prices in 2002 and 2008 across all specifications. Column (1) runs the dependent variable on the import effect. This regression is estimated by OLS and it doesn't consider controls and weight. The remaining specifications are estimated by 2SLS because we use the change in Chinese exports to low-income countries (besides Brazil) divided by the 2002 Brazilian expenditure as an instrument for the import effect. Columns (2) to (4) differ in terms of covariates and weight. Unless otherwise specified, regressions have as controls income and income growth in the demand side; age, worked hours and wages (supply-side); and are weighted by Brazilian total expenditure in 2002. The values of the import effect and IV are winsorized. Robust standard errors are in parentheses. $*p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

In Table 1.4, we present two regressions. Column (1) shows the results of running the China effect on the IV, and Column (2) presents the regression of ROW effect on the IV. Both models consider the covariates mentioned previously, with 2002 consumption as the weight. Table 1.4 provides evidence that China and ROW shocks are unrelated. Our IV is positively related to the China effect, with a statistically significant coefficient, but it is not associated with the ROW effect. Hence, we can identify China's supply shock.

Table 1.4 – Effect of the IV import effect on China and ROW effect, winsorized values

	(1) OLS	(2) OLS
	China effect	ROW effect
IV import effect	0.0375***	0.00640
	(0.00926)	(0.00464)
Weight	Yes	Yes
Controls	Yes	Yes
N	129	129

The IV for the import effect is the ratio between changes in Chinese exports to all low-income countries besides Brazil between 2002 and 2008, and the Brazilian expenditure in 2002. Column (1) shows the regression of China effect on IV, and Column (2) presents the regression of the ROW effect on IV. Both regressions are estimated by OLS and the covariates are income, income growth, age, hours worked, and wage. Also, all regressions are weighted by total consumption in 2002. Winsorized variables: IV import effect, China and ROW effect. Robust standard errors are in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

1.5.2 Additional results

1.5.2.1 China Effect

To further investigate China's supply shock, we analyze the direct relationship between imports from China and price changes in Brazil. We consider the following regression:

$$\Delta log(P_i) = \alpha + \beta_2 \frac{\Delta M_{i,CHN}}{E_{it_0}} + x_i^{'}\gamma + \varepsilon_i. \tag{1.8}$$

Table 1.5 shows the regressions of the price index on the China import effect. These regressions are similar to those in Table 1.3 in terms of controls and weight. For all specifications, the coefficient associated with the China import effect is negative and statistically significant. Also, based on our preferred specification, in Column (4), we conclude that when we compare the median $\frac{\Delta M_{i,CHN}}{E_{it_0}}$ with a case where its value is zero, prices in the median sector grew by 0.004 ppt less between 2002 and 2008.

Table 1.5 – Effect of imports from China on Brazilian domestic prices, winsorized values

	(1) OLS	(2) 2SLS	(3) 2SLS	(4) 2SLS
China import effect	-0.431***	-0.749**	-0.981***	-0.934***
	(0.111)	(0.303)	(0.370)	(0.332)
Weight	No	No	Yes	Yes
Controls	No	No	No	Yes
N	129	129	129	129
1st-stage F-stat		18.38	18.56	16.34

The dependent variable is the Brazilian price index for different sectors. The variable of interest is the China import effect (change in Brazilian imports from China divided by the 2002 consumption). Column (1) shows the OLS estimate of the regression of the price index on the China import effect, no covariates and weight. Additionally, Columns (2) to (4) show the results of a similar regression presented in Column (1), but we use the difference in China's exports to all low-income countries besides Brazil divided by the Brazilian expenditure in the initial period as an instrument for China import effect. Unless otherwise specified, the regressions have as controls income, income growth, age, hours worked and wages. Additionally, some regressions are weighted by Brazilian total expenditure in 2002 at the sector level. China import effect and IV are winsorized. Robust standard errors are in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

1.5.2.2 Local Effect

We could use our framework to identify China's demand shock on Brazilian prices using an IV for the local effect in the following regression:

$$\Delta log(P_i) = \alpha + \beta_3 FSE_{it_0} \frac{\Delta Z_i}{Z_{it_0}} + \gamma x_i + \epsilon_i. \tag{1.9}$$

The instrument for the local effect is $IV_{\text{(local effect)}} \equiv \frac{\Delta \overline{M}_i}{E_{it_0}}$, where $\overline{M}_i$ is the difference in Chinese imports from low-income countries besides Brazil between 2002 and 2008.

Two concerns with this identification strategy arise. First, the change in Brazilian exports to China observed in our data is small, as presented in Figure 1.2. Second, we would like to observe that the IV proposed above is related to the local effect, and that affects prices. However, it may be the case that the Chinese demand shock increases prices directly, and that raises expenditure on locally-produced goods.

We split our analysis into two parts: (i) we estimate local effect on prices using OLS (Panel A);[12] and (ii) we run the price index on the IV described previously to investigate if there is a direct effect, as mentioned above (Panel B). In Panel A, the first two columns do not consider controls; only the last two incorporate them. Also, Columns (2) and (4) use as weights the 2002 Brazilian expenditure. All four specifications of Panel A show that the local effect is not associated with price changes in Brazil. This may be a consequence of the absence of the sectors soybeans and iron ore in our data.

[12]The IV for the local effect is weak.

In Panel B of Table 1.6, we run the price index on the IV for the local effect. The weights and controls are similar to those used in Panel A. Panel B shows that the IV for the local effect is negatively correlated with changes in domestic prices for the sectors with the higher expenditure since Columns (2) and (4) present negative and statistically significant coefficients. We would expect a positive effect, since more demand (exports) would increase prices. However, it may be the case that firms become more productive, or that more productive firms enter the market in sectors with higher demand from China. Moreover, Column (4) of Panel B, our preferred specification, shows that when we compare a sector with the median value of the IV to one with no change, prices in the median sector grew by 0.11 ppt less during 2002 and 2008.

Table 1.6 – Effect of expenditure on locally produced goods and IV on prices, winsorized values

	(1) OLS	(2) OLS	(3) OLS	(4) OLS
Panel A. Local effect	0.0958	-0.0172	0.0367	-0.0247
	(0.181)	(0.0676)	(0.0995)	(0.0717)
Panel B. IV local effect	-0.0330	-0.399***	-0.0124	-0.381***
	(0.463)	(0.0947)	(0.430)	(0.116)
Weight	No	Yes	No	Yes
Controls	No	No	Yes	Yes
N	129	129	129	129

Panel A shows regressions of the variation in Brazilian domestic prices on the local effect considering 4 different specifications. Panel B presents 4 regressions of the price index on the IV for the local effect. The instrumental variable is the ratio between the change in China's imports from all low-income countries excluding Brazil and the consumption in Brazil in 2002. All regressions are estimated by OLS. Unless otherwise specified, the covariates are: income, income growth, age, hours worked and wages. Also, some regressions use the 2002 Brazilian expenditure as weight. Local effect and its IV are winsorized. Robust standard errors are in parentheses. $* p < 0.1$,$** p < 0.05$,$*** p < 0.01$.

We estimate the direct effect of changes in Brazilian exports to China, $\frac{\Delta X_{i,CHN}}{E_{it_0}}$, on domestic prices as follows

$$\Delta log(P_i) = \alpha + \beta_4 \frac{\Delta X_{i,CHN}}{E_{it_0}} + x_i^{'}\gamma + \varepsilon_i. \tag{1.10}$$

Also, we use the IV for the local effect to isolate the portion of changes in Brazilian exports to China related to the Chinese demand shock.

Similarly to Table 1.5, Table 1.7 presents the regressions of price index on China export effect, according to different specifications. Table 1.7 shows that for all specification, except Column (2), the coefficient of the China export effect is negative and statistically significant. This is evidence that the Chinese demand shock drove Brazilian prices down. As discussed previously, it may be the case that this higher demand resulted in productivity gains and/or attracted more productive firms. Additionally, the effect of

the median China export effect (zero) on the prices in the median sector was zero, based on Column (4).

Table 1.7 – China export effect on Brazilian domestic prices, winsorized values

	(1) OLS	(2) 2SLS	(3) 2SLS	(4) 2SLS
China export effect	-1.803**	-0.655	-2.224***	-2.150***
	(0.826)	(8.963)	(0.471)	(0.588)
Weight	No	No	Yes	Yes
Controls	No	No	No	Yes
N	129	129	129	129
1st-stage F-stat		1.415	40.27	39.01

The dependent variable is the Brazilian price index. The variable of interest is China's export effect. Column (1) shows a simple regression. Columns (2) to (4) show the results of a similar regression presented in Column (1), but we use the difference in China's imports from all low-income countries besides Brazil divided by the Brazilian expenditure in the initial period as an instrument for China's export effect. Unless otherwise specified, the regressions have as controls income, income growth, age, hours worked, and wages. Additionally, regressions are weighted by Brazilian total expenditure in 2002. Winsorized variables: IV for the export effect and China export effect. Robust standard errors are in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

1.6 Conclusion

Our results provide estimates of the "China Supply Shock" on Brazilian consumers between 2002 and 2008. Our analysis is based on reduced-form regressions. To our knowledge, we are the first to use sector-level data from all four of the sources we explore.

The methodology in this paper follows that of Bai and Stumpner (2019). We construct price indices using CES preferences. We use the ratio between the difference over time in China's exports to low-income countries (other than Brazil) and the total expenditure to isolate the portion of the import effect that is related to supply changes in China. Thus, we identify the "China Shock."

Our results indicate that the rapid growth of Brazilian imports from China resulted in gains for Brazilian consumers in the analyzed sectors in the form of lower price index. Additionally, when we compare the median import effect to one with no change, prices in the median sector grew by 0.8 ppt less during 2002 and 2008. This negative effect on Brazilian domestic prices is also found by He (2019). A similar effect is also observed in developed countries by Bai and Stumper (2019), and Amiti et al. (2020). Evidence for the export effect requires further investigation.

Our results should interest policy makers, since we show that trading with China was beneficial for Brazilian consumers. This trade stimulated competition and/or gave consumers a greater variety of products between 2002 and 2008.

2 Inferring risk aversion from equilibrium contracts

2.1 Introduction

Risk preferences have been explored in several fields including insurance, macroeconomics, and finance. Those are vastly explored on decision making under uncertainty which has a central role in insurance markets because people buy insurance to mitigate uncertainty.

Insurance markets present a suitable context to infer individuals' risk preferences. If an individual chooses a low (high) deductible value, she is more (less) risk-averse, because the insurance company indemnifies a greater (smaller) range of losses.[1,2] Hence, by observing deductible choices, we might determine policyholders' risk aversion (RA).

Researchers use insurance data to investigate risk aversion, and most empirical studies are based on the contract menu. However, in many situations, only equilibrium contracts are available. In fact, most insurance companies give only limited information on all contract options, to avoid disclosing their insurance policy design strategy. In this case, is it possible to obtain the coefficient of risk aversion? If so, can we do it for out-of-sample individuals?

In this paper, we address these questions by analyzing the Brazilian auto insurance market. We use an individual-level data set containing 125,000 policyholders (80% is used as training data and the remainder is the test set) who live in São Paulo and drive 1,000cc engine vehicles. This is the first paper to estimate the lower bound of the risk aversion coefficient proposed by Drèze (1981).[3] Additionally, we apply our framework to out-of-sample individuals to generate a RA estimate for them.

Our theoretical framework is based on that of Drèze (1981). The author gives us an analytical expression for the coefficient of risk aversion. This coefficient is a function of two variables: (i) how much insurance companies can charge above the expected indemnity; and (ii) deductible. The first variable is computed by dividing the premium by the expected indemnity.[4] However, expected indemnity is not observable. Thus, the first step of our empirical strategy is to estimate it, then do the calculation. The second variable is

[1] Deductible is the amount paid by a policyholder for an insured loss.

[2] Indemnity is the difference between the loss incurred by the policyholder and the deductible value.

[3] The only measure of risk preference we use in this paper is the lower bound of the coefficient of risk aversion determined by Drèze (1981). Therefore, we also refer to this metric as the coefficient of risk aversion.

[4] Premium is the price an individual pays for an insurance contract.

observed.

The expected indemnity is found by multiplying the probability of filing a claim by the expected claim cost.[5] We use Random Forests (RF) proposed by Breiman (2001) to find these values, and then estimate the expected indemnity. Next, we divide the premium by this estimated value. This tells us how much higher the price charged by the insurance company is, in comparison to the expected indemnity. Finally, we plug this value and the deductible into the expression in Drèze (1981) to calculate the lower bound of the coefficient of risk aversion.

To validate this estimate, we rely on two reduced-form regressions. First, we run the logarithm of the estimated RA on individual characteristics and car attributes to discover how they are related. Second, we run a logit model such that the dependent variable is a dummy that assumes 1 if the policyholder chooses a low deductible, and 0 otherwise. The regressors of the model are the observed characteristics of individuals and vehicles, and the logarithm of the estimated RA. Through this model, we investigate whether higher risk aversion is associated with low deductible choice, as described previously. Since we consider the estimate of the RA, not the true value, inference is based on bootstrapped standard errors.

We find that the average coefficient of risk aversion is 0.0011 with a standard deviation of 0.0004. From the reduced-form estimates, we can conclude that female and older drivers are more risk averse. Furthermore, more risk averse individuals are, on average, more likely to choose a lower deductible value. These results align with the existing literature, and appear both in- and out-of-sample.

Our main contribution is to estimate the lower bound of risk aversion proposed by Drèze (1981). This coefficient has two main advantages. It is based on equilibrium contracts rather than the contract menu (broader application), and it does not depend on a particular form of the utility function. Additionally, we propose a flexible empirical approach to estimate it.

Pioneers in empirical research regarding the estimation of RA using data from experiments include Yaari (1965) and Preston and Baratta (1948). Kachelmeier and Shehata (1992), and Smith and Walker (1993) focus on controlled experiments to measure risk aversion. Later researchers explore real market data starting with Cicchhetti and Dubind (1994). These authors estimate risk preferences based on individual-level information and telephone insurance purchases.

Regarding estimation of risk aversion in insurance markets, Cohen and Einav (2007) use deductible choices to estimate risk preferences. They use a structural econometric model that identifies the joint distribution of risk aversion and claim rate. Our study

[5] A claim is made when a policyholder informs the insurance company that a loss has happened.

is closely related to theirs. We both use theoretical models based on expected utility to derive an analytical expression for the coefficient of risk aversion which is not dependent on a functional form of the utility function. The main difference is that Drèze (1981) obtains an expression for risk aversion that is based on equilibrium contracts rather than the menu. Additionally, we expand our focus to an out-of-sample analysis.

Similar to the approach taken by Cohen and Einav (2007), Barseghyan, Prince, and Teitelbaum (2011) test the stability of risk preference over home and auto insurance markets. Barseghyan et al. (2013) add one more element to the discussion of the risk estimation from deductible choices by introducing a probability distortion function. Our work differs from theirs because the authors have no expression for the coefficient of risk aversion derived directly from the theoretical model, and they do not have an out-of-sample analysis.

Chetty (2006) proposes a new methodology that estimates the coefficient of RA based on data regarding labor supply behavior. Sydnor (2010) uses the framework proposed by Chetty (2006) to establish, using data from home insurance, a lower and upper bound for the coefficient of risk aversion. Unlike Drèze (1981), their bounds depend on the functional form of the utility function and the contract menu.

More recently Barseghyan et al. (2018) highlight the importance of risk aversion estimation, and survey the literature on estimates that use field data (individual and aggregate data). Meanwhile, Ledo and Lopes (2019) reproduce the model proposed by Cohen and Einav (2007) using data from the Brazilian automobile insurance market. Ledo and Lopes (2019) also investigate the effect of competition on risk and risk aversion (joint distribution).

Jaspersen et al. (2019) take a step further and instead of estimating the coefficient of risk aversion based on insurance decisions, the authors verify whether estimates of risk preferences from different structural models can predict insurance demand. They find poor predictive performance. Unlike Jaspersen et al. (2019), we use out-of-sample analysis to obtain an estimate of RA for individuals who are not in our data.

To our knowledge, we are the first to estimate the lower bound of the coefficient of risk aversion proposed by Drèze (1981). This estimation is important because it is more widely applicable given that it requires only equilibrium contracts, and its value is independent of the functional form of the utility function. Additionally, this paper deepens the discussion regarding the application of risk aversion estimation in an out-of-sample context.

The remainder of the paper proceeds as follows. In Section 2.2, we present the theoretical model developed by Drèze (1981) which gives us an analytical expression for the lower bound of the coefficient of risk aversion. Section 2.3 describes the data and shows

summary statistics. Section 2.4 discusses our empirical strategy, which consists of three parts: (i) estimation of risk preferences for both training and test sets; (ii) reduced-form regressions to validate these estimates; and (iii) bootstrap. Results are shown in Section 2.5 and Section 2.6 concludes.

2.2 Theoretical Motivation

The objective of the theoretical framework is to obtain an analytical expression for the lower bound of the risk aversion coefficient, following Drèze (1981). Denote by ω_0 individual's initial wealth who pays P (premium) for an insurance contract with deductible D. Loss L is a random variable with distribution $F(L)$, and $I(L)$ represents the value indemnified by the insurance company given a loss of magnitude L. An insurance company determines the premium as follows:

$$P = k\mathbb{E}[I(L)], \tag{2.1}$$

where k is a constant and $\mathbb{E}$ is the expectation operator.

Define by u a utility function which is twice differentiable and concave. From Arrow (1971), we have that if P is a function of $\mathbb{E}[I(L)]$, the optimal policy determines that:

$$I(L) = \max(L - D, 0).$$

Hence,

$$\mathbb{E}[I(L)] = \mathbb{E}[L - D|L \geq D] = \int_{L \geq D} (L - D)dF(L). \tag{2.2}$$

Since $\mathbb{E}[I(L)]$ is a function of D, then P is also a function of D: $P = P(D)$.

Let $W(L)$ represent the wealth of a policyholder who pays the premium and incurs loss L. She decides whether or not to make a claim based on the lesser of the loss and the deductible. Hence, her wealth is:

$$W(L) = w_0 - P(D) - \min(L, D).$$

Given a loss L, if $L \leq D$, she does not make a claim, and her utility is $u[\omega_0 - P(D) - L]$. Otherwise, if $L > D$, she files a claim, and her utility is $u[\omega_0 - P(D) - D]$. Therefore, her expected utility is:

$$\mathbb{E}[u(W(L))] = \int_{L \leq D} u[\omega_0 - P(D) - L]dF(L) + \int_{L > D} u[\omega_0 - P(D) - D]dF(L). \tag{2.3}$$

By maximizing Equation (2.3) with respect to D, we obtain the optimal deductible. From the first-order condition, and considering k is fixed, we obtain the following lower bound for the coefficient of risk aversion:

$$LB_{RA} \equiv \frac{k - 1}{kD}. \tag{2.4}$$

Equation (2.4) not only gives us an analytical expression for the coefficient of risk aversion, but suggests that low (high) deductible values are associated with higher (lower) risk aversion for a given k.

2.3 Data and Descriptive Statistics

2.3.1 Data Source

The database is recorded by the Superintendência de Seguros Privados (SUSEP) which supervises and controls the Brazilian insurance market. The variables are described in SUSEP Circular n. 197 of August 2002.

The data contains insurance contracts active for at least one day between June and December 2004. We limit our data to insurance policies that are relevant to our study. We select contracts that last for one year because they have the same exposure. These policies contain comprehensive coverage.[6] Also, we focus on policyholders from the São Paulo metropolitan area who drive nationally-produced cars with a 1,000cc engine. This is the most representative subsample since São Paulo has the highest number of policyholders and the majority of insured cars has 1,000cc engine. All contracts are personal with non-commercial vehicles. Endorsed or collective policies are disregarded.[7] This sample has 629,108 observations.

The theoretical model assumes that only one claim occurs. We follow this assumption, so we consider only the first claimed loss due to a collision. It helps us to simplify our empirical strategy. The advantage of this selection is that it does not change the identification of the policyholders who filed a claim. However, for people who filed more than one claim, we consider a smaller total loss. This may affect our estimate of the expected indemnity. In Section 2.5.2, we show that we obtain a reasonable prediction of the expected indemnity despite this assumption.

Originally, the database contains 3 deductible types: low, regular and high. We follow Cohen and Einav (2007) and categorize contracts with a high deductible as regular, because once an individual chooses high deductible, regular deductible would be chosen if only low and regular were available.[8] Hence, we have 2 deductible types namely high and low.

We select a random sample of 125,000 policyholders from the set described above such that the proportion of low (11%) and high deductibles is preserved. Additionally, the percentage of policyholders who filed at least one claim (1.8%) is maintained. Finally, the

[6] Comprehensive coverage offers coverage on collision, theft, and fire. Also, covers partial and total loss of the vehicle.

[7] An endorsement occurs when modifications are made in the insurance contract.

[8] Contracts with high deductible accounts for 0.46% of the sample, the loss of information is negligible.

proportion of individuals who purchased an insurance contract with low (high) deductible and also made a claim remains the same, at 2.8% (1.7%). The training data corresponds to 80% of the dataset, with the proportions mentioned above. The remaining data becomes the test set.

2.3.2 Descriptive Analysis

The variables in the database are divided into three groups: (i) covariates (individual and car characteristics); (ii) variables related to the insurance contract (premium, deductible value, and type); and (iii) indemnities and claim rate.[9] Table 2.1 shows summary statistics of the covariates in the training data. These variables are divided into two categories: individual characteristics (such as age and gender), and vehicle attributes (such as manufacturer and model year). This table shows that 49% of the policyholders are female. On average, the insured are 42 years old. Bonus, the percentage discounted off the premium according to the individual's claim history, is used as a proxy for driver's ability, and its average value is 18.55%.

Most insured vehicles are manufactured by General Motors (35%) and Fiat (27%). There are 16 car models and cars average four years of use. The oldest car dates from 1998 and the newest was manufactured in 2004. The average car value (the insured amount) is approximately BR 16,000. This variable is a proxy for income.[10]

Table 2.1 – Summary statistics - covariates

	Variables	Mean	Std. dev.	Min	Max
Individual	Female	0.49	0.5		
	Age	42.03	12.54	18	95
	Discount	18.55	12.97	0	45
Car attributes	Value	15929.9	4389.4	2500	55215
	Age	4	1.71	1	7
	General Motors	0.35	0.48		
	Fiat	0.27	0.44		
	Ford	0.16	0.37		
	Volkswagen	0.14	0.35		
	Renault	0.05	0.21		
	Hyundai	0.003	0.05		

This table displays summary statistics of the covarites. We divide the controls into two groups: individual characteristics and car attributes.

Table 2.2 shows summary statistics of the relationship between deductible type (low and high), premium, deductible value, and claims calculated from the data. Average premiums for low and high deductibles are close for the equilibrium contracts; the mean value for low (high) deductible choice is approximately BR 610 (BR 618).[11] The average

[9] Variables from the database are all described in Table B.1.

[10] A more comprehensive description of variables in Table B.2.

[11] Premium is the price paid to insure the vehicle.

low deductible, BR 490, is smaller than the average high deductible (BR 943), as expected. A low deductible is associated with a higher claim rate of approximately 3%. Indemnities for the high deductible are, on average, 55% smaller than those observed for low deductible policies. Indemnities are positive when losses are reported to the insurance company, otherwise, they are recorded as zero. Moreover, the average value of positive indemnity is approximately BR 1713 (BR 1947) for low (high) deductible choice.

Table 2.2 – Summary statistics - contract and claims

Variables	Deduc. type	Mean	Std. dev.	Min	Max	p10	p50	p90
Premium	Low	610.098	282.435	51	5861	326	564	952.1
	High	617.647	299.332	51	7540	308	582	975
Deductible Value	Low	490.27	98.61	315	2000	403	475	550
	High	943.42	131.12	180	3120	800	900	1040
Deductible Type	Low	0.11	0.31					
	High	0.89	0.31					
Claim rate	Low	0.03	0.16					
	High	0.02	0.13					
Indemnity	Low	47.96	386.63	0	12259	0	0	0
	High	33.1	326.23	0	12124	0	0	0
Pos. indemnities	Low	1712.97	1579.32	16	12259	393.5	1196	3284.6
	High	1946.93	1592.43	7	12124	454.2	1552	3914.8

This table shows summary statistics of the variables associated with the purchased insurance contract (premium, deductible value and type). It shows statistics of indemnities and claim rate.

2.4 Empirical Strategy

The goal of our empirical strategy is to estimate the risk aversion coefficient and run reduced-form regressions to validate it. This estimation is calculated from the training dataset, and then the framework is applied to out-of-sample data (the test set) to verify if it is applicable to new individuals. Additionally, inference on the reduced-form regressions are based on bootstrapped standard errors.

2.4.1 Coefficient of Risk Aversion

The estimation of the coefficient of risk aversion is based on Equation (2.4). In this expression, the only unknown parameter is k. This variable, as presented by Equation (2.1), is given by the ratio between premium and expected indemnity (the claim cost expected by insurance companies). We can not do this division because we do not observe the expected indemnity. Even though it is not observable, we do have information on claimed losses. Therefore, we can use them to estimate the insurance companies' expected claim cost, given a set of covariates x, $\mathbb{E}[I|x]$, using the following expression:

$$E[I|x] = \mathbb{P}(claim = 1|x)E[I|claim = 1, x], \tag{2.5}$$

where *claim* is a dummy variable which assumes 1 when a claim is made and 0 otherwise, and $E[I|claim = 1, x]$ represents the expected value of indemnities from claimed losses (the expected claim cost for insurers).[12] Thus, to find the expected indemnity, we estimate each term of the right-hand side of Equation (2.5). These estimates are based on Random Forests due to its predictive power and flexibility.

First, we predict the insurer's expected claim cost. Therefore, we create a Random Forest in the part of the training data that has positive claims. The target (dependent variable) is the indemnity; the features (covariates) are described in Table 2.1; and the loss function is the mean squared error (MSE). We evaluate the model's performance based on the root mean squared error (RMSE) calculated on the part of the test set that has positive claims.

We choose the RMSE to select the model with the best performance because this metric penalizes larger errors. If an insurance company predicts an indemnity of BR 100 and the observed value is BR 200, the insurance company has to pay twice as much for the filed claim. This directly influences profit. Hence, insurers try to avoid large errors.

Second, we obtain the probability of filing a claim. This prediction is divided into two parts: finding a classifier and a calibrator. We construct a Weighted Random Forest (WRF), proposed by Chen et al. (2004) because we have imbalanced classes (there are few claims). Our target is the variable *claim* described above. The features are presented in Table 2.1. We train the model using all of the training data. This classifier produces scores for each class, but these values can not be interpreted as the probability of filing or not filing a claim. Hence, to obtain the claim probability, we use a calibrator, g_i, which determines $\mathbb{P}(claim_i = 1|g_i)$. The chosen calibrator is a logistic model as proposed by Platt (1999).[13]

We choose the Area Under the Receiver Operating Characteristic Curve (AUROC) to evaluate the WRF performance because our data has few claims; we are interested in classifying both classes correctly, and we want to obtain the probability of filing a claim. The Receiver Operating Characteristic Curve (ROC) presents the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR) for each classification threshold.[14]

$$TPR = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}},$$
$$FPR = \frac{\text{False Positive}}{\text{False Positive} + \text{True Negative}}.$$

[12]To simplify our empirical strategy, we use the expression from Equation (2.5) instead of Equation (2.2). This change does not alter the lower bound of the coefficient of risk aversion in Equation (2.4).

[13]Based on Platt (1999), we model $\mathbb{P}(claim_i = 1|g_i)$ using the logistic function, $\sigma(Cg_i + D)$ where C and D are parameters determined by Maximum Likelihood.

[14]In this case, the threshold is a value such that, for all values greater than the threshold, we determine that a claim is filed.

To create a WRF we need to set its hyperparameters. The main hyperparameters are: (i) the number of decision trees; (ii) the maximum depth of trees; (iii) the maximum number of covariates considered at each split; and (iv) the minimum number of samples at a leaf node. The models are estimated over a grid of parameters such that the number of trees in the forest assumes values $(50, 100, 500, 1000)$, and the values for maximum depth of the tree are $(3, 4, 5, 6, 7, 8, 9, 10)$. All combinations between these variables are taken into consideration. Additionally, the maximum number of features in each split is given by the square root of the number of covariates, and the minimum number of samples at each leaf node is 5.[15] The last 2 hyperparameters are fixed. To find the model with the best performance, we use the out-of-bag estimate of the RMSE and AUROC.

After finding the probability of filing a claim and predicting the expected claim costs, we multiply them to compute $\widehat{E[I|X]}$ for each individual, as presented in Equation (2.5). Based on this value, we obtain k using $k = \frac{P}{\widehat{E[I|X]}}$. Next, we calculate the lower bound of the coefficient of risk aversion, $\widehat{LB}_{RA}$, given by Equation (2.4).

2.4.2 Reduced-form

To validate our estimate of the coefficient of risk aversion, we need to investigate: (i) how it relates to consumer and vehicle characteristics; and (ii) how this estimate relates to the choice of deductible. To address the first question, we run the following regression of the logarithm of the coefficient of risk aversion on the variables presented in Table 2.1: x_i.

$$log(\widehat{LB}_{RA,i}) = \beta_0 + x_i^{'}\beta + \epsilon_i.$$

Regarding the second question, we run a logit model. The dependent variable is a dummy, d_i, that assumes the value 1 if the person chooses a low deductible type and zero otherwise. The features are the variables described in Table 2.1 and the logarithm of the estimate of the coefficient of risk aversion. As predicted by Equation (2.4), we expect a positive sign for the coefficient associated with the logarithm of the lower bound of risk aversion, $\gamma_1 > 0$.

$$d_i = \gamma_0 + \gamma_1 log(\widehat{LB}_{RA,i}) + x_i^{'}\gamma + \mu_i, \tag{2.6}$$

$$d_i = \begin{cases} 1 & \text{if low deductible was chosen,} \\ 0 & \text{otherwise.} \end{cases}$$

2.4.3 Out-of-sample Analysis

Another important part of the empirical strategy is the out-of-sample analysis. We use what we learn from the training set to estimate the coefficient of risk aversion for new

[15] Hyperparameters values are based on Muller and Guido (2016) and the documentation from Scikit-learn created by Pedregosa et al. (2011).

individuals (those in the test set). The validation process is similar to the one described in the previous section.

2.4.4 Bootstrap

Inference in both regressions uses bootstrap standard errors, because our results are based on the estimated coefficient of risk aversion, not on the real value. The bootstrap contains 500 iterations and the random samples are generated keeping the proportion of claims constant at 1.8%. All steps in the bootstrap are described in Appendix B.13.

2.5 Results

2.5.1 Main Results

The first step of the empirical strategy is to find the RF with the best performance in predicting the insurer's claim cost. The out-of-bag error shows that, over the grid of parameters presented in Section 2.4.1, the best-performing model has 100 trees, the maximum depth of the tree is 3, and the RMSE is 1585.76.[16] Applying this model to the training data and test sets for the observations with positive claims, we have RMSEs of 1548.97 and 1765.88, respectively.

Similarly, to predict if a claim is made or not, the out-of-bag estimates show that the Weighted Random Forest with 500 trees and a maximum depth of 5 has the highest AUROC: approximately 0.57.[17] After calibrating this classifier, the AUROCs of this model in the training and test sets are 0.63 and 0.57, respectively.

Once we select the models, we compute the probability of filing a claim and the expected indemnity in both datasets. Table 2.3 displays the average probability score in these datasets: approximately 1.8%. The standard deviation for both is approximately 0.0095. Selecting only positive claims, Table 2.3 shows that insurance companies spend on average BR 1,884 in indemnities in both training and test sets.

From the predicted values of Table 2.3, the expected indemnity is calculated as presented by Equation (2.5). Additionally, k is computed as the ratio between the premium and the estimated expected indemnity, as in Equation (2.1). Next, we insert the estimated k and the deductible value into Equation (2.4) to determine the lower bound of the coefficient of risk aversion. These three estimates are presented in Table 2.4 for each dataset.

[16] The results of the out-of-bag error for all models can be seen in Table B.3.
[17] The out-of-bag estimates for the classification task can be seen in Table B.4.

Table 2.3 – Estimated probability and claim cost

Variables	Mean	Std. dev.	Min	Max	p10	p50	p90
			Train				
Predicted probability	0.0182	0.0094	0.01	0.1	0.01	0.02	0.03
Exp. Claim Cost	1884.31	171.031	1422.16	2664.37	1712.68	1877.74	2106.71
			Test				
Predicted probability	0.0183	0.0095	0.01	0.09	0.01	0.02	0.03
Exp. Claim Cost	1884.12	171.623	1417.19	2605.17	1712.1	1877.27	2107.61

This table displays two predictions: probability of filing a claim and indemnity. To obtain the probability of filing a claim we use a Weighted Random Forest classifier, then we use Pratt's calibrator to obtain the probability associated with each class (claim or no claim). Also, to predict indemnities, we use Random Forests. In both cases, the models are selected based on their out-of-bag error.

Table 2.4 shows that the average risk aversion is 0.0011 with a standard deviation of 0.0004 for both datasets. Insurance companies charge on average 20 times more than the expected claim cost. This cost averages approximately BR 35 in both datasets.[18]

Table 2.4 – Estimates of the parameters of the theoretical model

Variables	Mean	Std. dev.	Min	Max	p10	p50	p90
			Train				
Risk Aversion	0.0011	0.0004	0.00001	0.0051	0.0009	0.001	0.0017
k	20.33	9.96	1.01	51.56	9.15	18.98	33.69
$E[I\|X]$	34.88	19.73	8.53	205.61	17.43	29.38	59.6
			Test				
Risk Aversion	0.0011	0.0004	0.000004	0.0031	0.0009	0.001	0.0016
k	20.29	9.88	1.01	51.13	9.21	18.99	33.47
$E[I\|X]$	35.03	20.07	9.36	188.96	17.34	29.39	60.13

This table displays the estimates of the parameters of the theoretical model. Firstly, we estimate expected indemnity using Equation (2.5). Then, based on this value, we obtain k following Equation (2.1). The values of k greater than the 99 percentile are replaced by its threshold value. Also, in both datasets, there were values of k smaller than one and we replace them with the smallest positive value greater than one in the respective sets. 60 values were replaced in the train data, and 23 in the test set. Lastly, we use both estimates to compute the lower bound of the coefficient of risk aversion given by Equation (2.4).

Finally, to validate our estimate of the coefficient of risk aversion presented in Table 2.4, we run an OLS to identify how policyholder and car characteristics relate to the estimated coefficient of risk aversion. We use a logit regression to verify if this estimation is correlated with the deductible choice in the same way as the theoretical framework predicts: low deductible choice is associated with greater RA. Both models are presented in Table 2.5.

[18]Ledo and Lopes (2019) compute the point estimate of the coefficient of risk aversion proposed by Cohen and Einav (2007). The authors find that the average value of RA is 0.0006. They consider contracts purchased in São Paulo state in 2010 and cars with 1,000 and 2,000cc engines.

The results based on the OLS regression described above, presented in Column (1) of Table 2.5, are aligned with the Bayesian estimation of Ledo and Lopes (2019) and Cohen and Einav (2007). It shows that females, older drivers, and cars with higher values are associated with greater levels of risk aversion. The respective coefficients are statistically significant.

Additionally, the dependent variable of the logit model is a dummy that is equal to 1 if the policyholder chooses a low deductible, and 0 otherwise. The $log(\widehat{LB}_{RA,i})$ and the variables described in Table 2.1 are used as covariates. Column (2) shows that the coefficient of risk aversion is positive and statistically significant. Therefore, as predicted by the theoretical model, more risk-averse individuals are more likely to chose a low deductible.[19] All results described above are valid for both training and test set.

Table 2.5 – OLS and Logistic Regression

	OLS		Logit	
	Dep. var.: log(risk aversion)		Dep. var.: 1 if low deductible	
	(1)	Std. Dev.	(2)	Std. Dev.
		Train		
Log(risk aversion)	-	-	13.438***	(0.2434)
Age	-0.001	(0.0007)	-0.314***	(0.0107)
Age^2	0.047***	(0.0181)	6.627***	(0.2275)
Female	0.027***	(0.0011)	-0.084	(0.0852)
Log(Value)	0.094***	(0.0086)	4.748***	(0.1903)
Bonus	0.006***	(0.0002)	-0.038***	(0.0026)
		Test		
Log(risk aversion)	-	-	10.424***	(1.2083)
Age	0.004***	(0.0008)	0.1885***	(0.1046)
Age^2	-0.00002***	(0.00001)	-0.0018***	(0.0011)
Female	0.027***	(0.001)	1.1825	(0.5411)
Log(Value)	0.091***	(0.0097)	6.2573***	(0.9186)
Bonus	0.006***	(0.0002)	-0.0119	(0.0157)

This table displays 2 regressions. Column (1) shows the OLS estimate of the logarithm of the estimated risk aversion on the covariates described at Table 2.1. Column (2) shows a logistic regression such that the dependent variable assumes value 1 if low deductible is chosen and zero otherwise. The covariates in the logit are: variables described in Table 2.1 and the logarithm of the coefficient of risk aversion. Bootstrapped standard errors. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

2.5.2 Additional Results

2.5.2.1 Validation of Expected Indemnity

After predicting insurer's expected claim costs, it is important to validate this estimation. Since a premium is determined as a function of the expected claim costs, as presented in Equation (2.1), it is reasonable to assume that the estimate of the expected indemnity captures part of the premium's variability. Therefore, we analyze the following

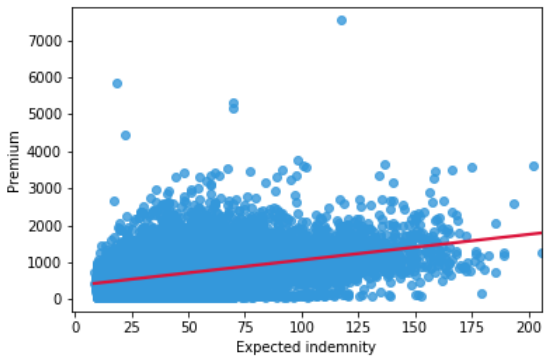
[19] Ledo and Lopes (2019) and Cohen and Einav (2007) find similar results.

regression between these variables:

$$P_i = \delta_0 + \delta_1 E[I_i|X_i] + \varepsilon_i. \tag{2.7}$$

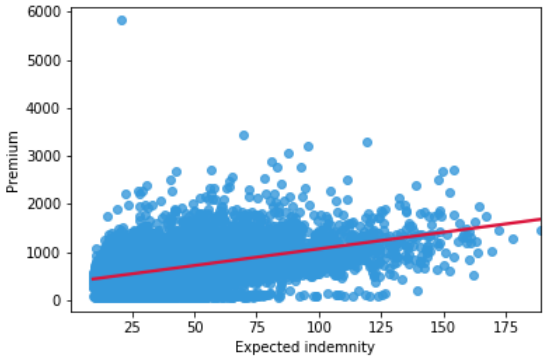
Figure 2.1 shows scatter plots of expected indemnity against premium. The line depicts the results of the simple regression in Equation (2.7). This figure shows a positive correlation between both variables. The coefficient associated with the expected indemnity is statistically significant in both datasets, as shown in Tables B.11 and B.12.

Figure 2.1 – Relationship between premium and expected indemnity



This graph presents scatter plots of expected indemnity against premium in Brazilian Real in the training data. The line depicts the results of a simple regression of premium on the expected indemnity.

(a) Train



This graph presents scatter plots of expected indemnity against premium in Brazilian Real in the test set. The line depicts the results of a simple regression of premium on the expected indemnity.

(b) Test

2.5.2.2 No Insurance Contract

To compute the risk aversion coefficient as described in Section 2.4.1, we need information regarding the values of the premium and the deductible for each policyholder. However, this information may not be available for some individuals. In fact, from the insurers' perspective, this happens on two occasions: (i) when we infer risk aversion for an individual who buys insurance for the first time; and (ii) when policyholders want to buy insurance from another company, since policyholders may not give information on their previous contracts. This exercise shows that our framework applies to these situations.

Additionally, this exercise is important from an econometric standpoint. As mentioned previously, we use the premium and deductible to compute the coefficient of risk aversion. Therefore, it is natural to expect this estimate to be related to the choice of deductible. After all, the insurance contract is composed of premium, deductible value, and choice. Perhaps the results presented in the previous section are entirely driven by this correlation. The following exercise shows that this is not true.

We compute k using the average premium in the training data, and the same estimate of the expected indemnity obtained previously. Next, we calculate the coefficient

of RA considering the average deductible in the training data. Both values are presented in Table 2.6.

Table 2.6 – Estimates of the parameters of the theoretical model given that insurance contracts are not observable

Variables	Mean	Std. dev.	Min	Max	p10	p50	p90
			Train				
Risk Aversion	0.0011	0.00004	0.0007	0.0011	0.001	0.0011	0.0011
k	22.35	9.97	3	52.37	10.35	20.99	35.4
			Test				
Risk Aversion	0.0011	0.00004	0.0008	0.0011	0.001	0.0011	0.0011
k	22.4	10.15	3.26	54.11	10.26	20.98	35.58

This table displays the estimation of the lower bound of the coefficient of risk aversion and k. These estimates consider the average contract observed in the train data.

From this table, we conclude that the average risk aversion is 0.0011, as obtained previously. The difference is that it has a smaller variance, because the only source of variability comes from the expected indemnity. Previously, there had been three sources: expected indemnity, premium, and deductible. Also, the average value of k is 22.

Based on the variables in Tables 2.6 and 2.1, we apply the same logistic regression described in Section 2.4 on training and test sets. As shown in Table 2.7, the results corroborate the estimates presented in Table 2.5: the coefficient of the log of risk aversion is positive and statistically significant, but it is smaller. This value is smaller because the individual-level information regarding premium and deductible are disregarded.

Table 2.7 – Logistic Regression

	Train		Test	
	Dep. var.: 1 if low deductible		Dep. var.: 1 if low deductible	
	(1)	Std. Dev.	(2)	Std. Dev.
Log(risk aversion)	0.662***	(0.0516)	0.181***	(0.0238)
Age	0.026***	(0.0007)	0.01528***	(0.0008)
Age^2	-0.394***	(0.0129)	-0.00007**	(0.00005)
Female	0.194***	(0.0098)	0.050***	(0.04067)
Log(Value)	0.382***	(0.0422)	-0.220***	(0.0170)
Bonus	0.038***	(0.0002)	0.038***	(0.00004)

This table displays 2 logistic regressions one for each data (train and test). The dependent variable is d_i and the covariates are the logarithm of the estimated coefficient of risk aversion and the variables described in Table 2.1. Column (1) shows the values of the coefficients in the train data. Similarly, Column (2) presents the coefficients in the test set. Additionally, RA is computed based on the average values of premium and deductible in the train data. Bootstrapped standard errors in parenthesis. Bootstrapped confidence intervals are shown in Tables B.9 and B.10. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

2.6 Conclusion

This paper lays out a framework to estimate a coefficient of risk aversion such that: (i) its value does not depend on the functional form of the utility function; and (ii) the value considers equilibrium contracts, rather than the menu. Therefore, it has a broader application since it is not necessary to observe the menu offered by insurance companies. Additionally, most of the empirical strategy is flexible, because we use Random Forests.

Furthermore, this study presents the first estimates of the lower bound of the coefficient of risk aversion presented by Drèze (1981). In this case, it has an average value of 0.0011. Our results corroborate the estimates made by Ledo and Lopes (2019) and Cohen and Einav (2007), because we also show that females and older drivers are more risk-averse, and more risk-averse individuals are more likely to choose lower deductible values.

We take one step further and show that our framework also applies to out-of-sample individuals into two scenarios. In the first scenario, the purchased insurance contract (premium and deductible), vehicle attributes, and individual characteristics are observable. In the second, information regarding the contract is not available.

Our framework may benefit insurance companies, because it allows them to assess risk aversion for their clients and new policyholders. Additionally, since our data contains all insurance companies, they can compare the measure presented in this study (market average) to risk measures they already have.

3 Quote request system for insurance contracts

3.1 Introduction

A common concern of individuals buying insurance contracts is what contract they should buy. They search on the internet, talk to insurance brokers, then make a decision. Discovering whether a consumer should buy a contract offered by an insurance company would be easier if there were a quote request system to *predict* the average insurance contract.

Some websites provide a quote request system: individuals supply the required information, and receive offers of contracts based on policies they want to commercialize and/or the contracts insurance companies allow them to sell. Therefore, not all contracts available on the market are provided. The displayed policies are not random and may be subject to changes in insurance companies' decisions. In fact, they could sell more contracts with high deductible to select individuals with potentially smaller risk. Consequently, it may be difficult for consumers to infer what the average contract is.

Unlike these websites, we take data on the purchased contracts from multiple insurance companies, learn from them, and predict the average contract. Even though the equilibrium contracts in our data are influenced by the insurance companies' decisions, our sample is composed by a random selection of purchased contracts from all insurance companies. Therefore, we can provide a more accurate estimate of the average contract.

Additionally, our system can be beneficial for both consumers and insurers. Based on our request quote system, consumers could insert their information and get a prediction of premium, deductible value and type (average contract).[1,2] Therefore, for any offer they receive, they could easily know how far it is from our prediction in terms of premium, deductible and risk (deductible choice). Insurance companies could also use our system to verify how their contract design strategy compares to our prediction, and how their methodology to find the average contract differs from ours.

Our quote request system uses data from the Brazilian auto-insurance market.

[1]Premium is the price paid for the insurance contract. Deductible is the value the policyholder pays when she files a claim.

[2]It is important to consider the deductible type because the policyholder chooses it. Additionally, when a policyholder chooses a low deductible instead of a higher deductible, the insurance company has to cover a bigger range of losses (the insurer faces a higher risk). Therefore, the deductible type is an important component of the insurance contract. We can not base our analysis only on the deductible value because it does not indicate the chosen deductible type.

We use a rich dataset containing 100,000 policyholders who live in the metropolitan region of São Paulo and drive cars with 1,000cc engines. Furthermore, insurance contract predictions come from the estimation of several Multi-Output Random Forests (MORF).

The insurance contract has multiple outcomes (premium, deductible value, and type). Premium and deductible are continuous variables. Deductible type is a dummy that assumes 1 if a low deductible is chosen, and zero otherwise. Hence, we face a prediction problem with joint classification-regression outputs. A suitable method to make this prediction is the MORF, because it simultaneously predicts these variables.[3]

As a result of implementing MORF, we find the root-mean-square-error in the test set for premium and deductible are approximately 42% and 34% of the respective mean values in the training data. The accuracy of the model in predicting the deductible type is 61% in the test set. Furthermore, this method reduces by 9% and 4% the root-mean-square-error for premium and deductible, when the average mean in the train data is used as a predictor.

This paper has three main contributions. First, it proposes a new service (quote request system) that predicts insurance contracts (premium, deductible value, and type) based on a random sample of equilibrium contracts from all Brazilian insurance companies. Second, it is the first study to make this prediction. Finally, it applies, for the first time, the Multi-Output Random Forest to insurance data.

Machine Learning (ML) applications have grown considerably across many fields, including insurance markets. Most studies that combine insurance data and ML methods focus on claim prediction/insurance pricing (WUTHRICH, BUSER, 2019; GUELMAN, 2012; CHAPADOS et al., 2002; FRANCIS, 2001) and fraud detection (VIAENE et al., 2005; VIAENE et al., 2002).

Claim prediction is closely related to the forecast of insurance premiums because the estimation of the *pure* premium is calculated from the expected claim cost. However, few studies model the price paid by consumers, which is our focus here.

Moreover, regarding the applications of Machine Learning methods to Brazilian auto insurance data, Freitas (2018) predicts premiums using the expected claim cost, then estimates the demand for insurance in Rio de Janeiro.

Our study differs from the literature in three aspects. First, it is the first to predict the insurance contract (premium, deductible value and type), not just the premium. Second, our analysis is based on the real price paid by consumers, rather than the expected costs (*pure* premium). Also, we are the first to apply the Multi-Output Random Forest, proposed by Linusson (2013), to insurance data.

The remainder of the paper proceeds as follows. In Section 3.2, we describe our

[3]Regression (classification) task maps a function to a continuous (categorical) variable - output.

data on the Brazilian auto-insurance market. Section 3.3 discusses our empirical strategy based on the Multi-Output Random Forest. Results are shown in Section 3.4, and we present our conclusions in Section 3.5.

3.2 Data and Descriptive Statistics

3.2.1 Data Source

Our dataset comes from the Superintendência de Seguros Privados (SUSEP). SUSEP has been responsible for regulating and supervising the Brazilian insurance market since 1999.[4] The data complies with Directive n. 197 of August 2002 established by SUSEP.

We limit our data to one-year policies with comprehensive coverage, containing all the contracts active (for at least one day) between June and December of 2004.[5] There are only personal policies with non-commercial vehicles. Endorsed or collective policies are disregarded.[6] Additionally, the analysis is restricted to policyholders from the São Paulo metropolitan area (the region with the highest number of policyholders in all of Brazil) who drive nationally-produced vehicles with 1,000cc engines. After imposing these filters, this sample has 629,108 observations and has the most relevant contracts for our study.

This data originally contains 3 deductible types: low, regular, and high. Cohen and Einav (2007), without generating bias, we record high deductible as regular. If a policyholder chooses a high deductible, she would choose regular if low and regular deductibles were the only options. We follow Cohen and Einav (2007), therefore our dataset has 2 deductible types namely low and high. The same approach is shared by Ledo and Lopes (2019).

To increase the speed of implementation, we consider a subsample (100,000 observations) of the sample described previously. Originally, approximately 11% of the individuals in the data choose a low deductible. If we keep this proportion in our subsample, we have imbalanced classes. Simply guessing that all individuals selected a high deductible, we obtain an accuracy of 89%. To avoid this problem, 50% of the policyholders in our training and test sets purchased an insurance contract with a low deductible. Additionally, we randomly draw our training data from the subsample to achieve 75,000 observations. The remaining contracts make up the test set.

[4]SUSEP does not supervise health insurance.

[5]Comprehensive coverage offers coverage on collision, theft, and fire. It also covers partial and total loss of the vehicle.

[6]A policy is endorsed when any modifications are made in the insurance contract.

3.2.2 Descriptive Analysis

The variables in the database are divided into two groups. The first variables are the features used in the empirical strategy: a collection of vehicles and attributes of the insured. Then we consider, variables related to the contract: premiums, deductible values, and type. These are the target variables.[7]

Table 3.1 shows summary statistics of the covariates in the training data in two categories: individual characteristics and vehicle attributes. Table 3.1 shows that 51% of the policyholders are female, on average they are 43 years old, and their Bonus is approximately 21%. Bonus is a discount the policyholder receives, based on her claim history. If she makes a lot of claims, the value of the Bonus is smaller. Hence, this variable is a *proxy* for the driver's ability.

Regarding car attributes, we analyze 16 Brazilian car models. Table 3.1 shows that the average car value (insured amount) is approximately BR 16,000. This variable is a *proxy* for income. Most vehicles are manufactured by General Motors, Fiat, and Ford which representing 36%, 26%, and 16% of the total, respectively. The oldest car was manufactured in 1998 and the newest dates from 2004. The average car age is 4 years.[8]

Table 3.1 – Summary statistics - covariates

	Variables	Mean	Std. dev.	Min	Max
Individual	Female	0.51	0.5		
	Age	42.74	12.59	18	90
	Bonus	20.91	13.03	0	45
Car attributes	Value	16113.5	4414.49	2300	64800
	Age	3.95	1.67	1	7
	General Motors	0.36	0.48		
	Fiat	0.26	0.44		
	Ford	0.16	0.37		
	Volkswagen	0.13	0.34		
	Renault	0.05	0.22		
	Pegeout	0.04	0.18		
	Hyundai	0.003	0.06		

This table displays summary statistics of the covariates used in the empirical strategy. This set of variables is divided into two groups: individual characteristics, and cars attributes.

Table 3.2 shows summary statistics of the target variables in the training data: equilibrium contracts (premium, deductible value and type).[9] The premium for a contract with a low deductible averages BR 610, and is 1.39% smaller then the premium for a contract with a high deductible. Also, the average high deductible is approximately 92% greater than the low deductible (BR 490). The proportion of low-deductible contracts is 50%, as explained in Section 3.2.1.

[7] Descriptions of all variables in the database are in Table C.1.
[8] A more comprehensive description of the data appears in Table C.2.
[9] Premium is the price charged to insure the vehicle.

Table 3.2 – Summary statistics - insurance contract

Variables	Deduc. type	Mean	Std. dev.	Min	Max	p10	p50	p90
Premium	Low	609.92	276.31	51	5821	327	563	951
	High	618.42	299.30	51	5454	306	582	977
Deductible Value	Low	490.43	97.75	280	2400	403	455	550
	High	942.13	132.91	270	3120	800	900	1040
Deductible Type	Low	0.5	0.5					
	High	0.5	0.5					

This table displays the summary statistics of the target variables in the training data: premium, deductible value and type.

3.3 Empirical Strategy

The goal of our empirical strategy is to build a research system for car insurance quotes so consumers can verify if the chosen/offered contract is consistent with the average in the market. Hence, our objective is to predict premium, deductible value, and insurance type based on the set of covariates presented in Table 3.1.

3.3.1 Baseline Model

Our baseline model is the simplest information that could be given to consumers: a sample average of the target variables in the training data. Therefore, the error in the test set is given by the difference between the real value of the target variables and their average from the training data.

3.3.2 Multi-Output Random Forest

The Multi-Output Random Forest is proposed by Linusson (2013). It is an extension of the Random Forest framework to solve decision problems with multiple outputs (regression and classification). Glocker et al. (2012) solve one classification and one regression problem, at the same time, based on a tree induction algorithm. Linusson (2013) generalizes the split function used by Glocker et al. (2012) to consider any number of classification and regression tasks based on the same dataset. Hence, the joint classification-regression Random Forest proposed by Linusson (2013) is ideal to simultaneously predict premium, deductible value and type.

3.3.3 Performance Evaluation

To evaluate the predictive performance of the models, we use accuracy (ACC) when the target variable is the deductible type. For regression tasks (premium and deductible

value) we consider the root mean square error (RMSE). Accuracy is defined as the fraction of correct classifications made on the test set, Z_{test}, as follows:

$$ACC(h) = \frac{|\{(x, y) \in Z_{test} | h(x) = y\}|}{|Z_{test}|}, \tag{3.1}$$

where x represents the vector of features, y is the dependent variable (deductible type), h represents a function (MORF), and $|.|$ is the number of observations. We chose accuracy because we have balanced classes and is equally important to correctly classify both classes since consumers can choose any of them.

RMSE is the square root of the average of the squared difference of the predicted and real values in the test set:

$$RMSE(h) = \sqrt{\frac{\sum_{(x,y)\in Z_{test}} (y - h(x))^2}{|Z_{test}|}}, \tag{3.2}$$

where, in this case, y is the premium or the deductible value. We select RMSE because prediction errors have a direct effect on consumers. Errors in prediction could lead them to pay more for an insurance contract and/or influence them to take more risk (choose a higher deductible value). Therefore, we use this metric because it penalizes larger errors.

Our evaluation is based on the 10-fold cross-validation. It randomly shuffles the data, then splits it into 10 datasets. For each one of them, use it as the test set and remaining datasets are the training data. Next, we fit the model on the training set and evaluate it on the test data, computing the metrics (ACC and RMSE). This process generates 10 scores for each metric. Then, we take the mean of those scores. We choose the model with the lowest RMSE and highest ACC.

The estimation and cross-validation are based on Morfist, which is the Python implementation of the Multi-Output Random Forest proposed by Linunson (2013).[10] The most important hyperparameters of the model are the number of trees; the maximum number of features when looking for the best split; the minimum number of samples at a leaf node; and the number of classification tasks.

We estimate the model over the following values for the number of trees $(10, 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000)$. The minimum sample size at a leaf node is 5. The maximum number of features is the square root of the number of covariates. The average of information gains are considered for the split, and there is one classification task (deductible type).

[10]The code is available at <https://github.com/donlnz/morfist/blob/master/README.md>.

3.4 Results

3.4.1 Main results

This section presents the results for the MORF and baseline model described in Section 3.3. Based on the cross-validation, the simplest model that has one of the best performances is the MORF with 100 trees.[11] Table 3.3 presents the RMSEs and accuracy of this model and the baseline results. This table also shows that the RMSEs of the baseline model are about 46% and 36% of the average premium and deductible values in the training data, respectively. Moreover, the root mean squared error and accuracy computed with the baseline model for premium, deductible value and type are BR 285.23, BR 254.89, and 50%, respectively.

Table 3.3 shows that the RMSE, obtained with MORF, for premium and deductible are approximately 9% and 4% less than the results obtained with the baseline model. Additionally, the accuracy of the prediction of the deductible type based on the Multi-Output Random Forest is 61%, which is 11 percentage points more than the value obtained in the baseline model.

Table 3.3 – Comparison of the performance between the baseline model and MORF in the test set

	Baseline Model (Mean)		
	Premium	Deductible value	Deductible type
RMSE	285.23	254.89	-
ACC	-	-	50%
	Multi-Output Random Forest		
	Premium	Deductible value	Deductible type
RMSE	260.41	245.02	-
ACC	-	-	61%

This table displays the RMSE and ACC computed in the test set based on the Multi-Output Random Forest and baseline model.

Hence, the MORF has better predictive performance overall, when compared to the baseline model. The MORF predicts individuals' deductible choice more accurately, and has smaller RMSE in predicting premium and deductible value.

3.4.2 Additional results

To expand on our analysis of the predictive performance of the baseline model and MORF, we find the absolute value of the difference between the observed and predicted values by percentile, as presented in Table 3.4. For most percentiles that we consider

[11] A more comprehensive description of the results of the 10-fold cross-validation method is in Table C.3.

the Multi-Output Random Forest produces a smaller absolute error for the premium and deductible.

Table 3.4 – Comparison of models' performance by percentiles

	Mean	Std dev.	Min.	Max.	P10	P50	P90
				Premium			
Baseline	211.48	191.4	0.2	5247.2	32.8	169.2	442.8
MORF	189.4	178.73	0.02	5275.54	28.35	147.9	403.37
				Deductible Value			
Baseline	234.71	99.39	3.12	2133.12	32.8	169.2	442.8
MORF	225.16	96.62	0.01	2065.97	115.87	227.19	324.34

This table displays the comparison, by percentile, of the predictive performance of the baseline model and MORF according to the absolute error. P10, P50, and P90 represent the 10th, 50th and 90th percentiles, respectively.

3.5 Conclusion

This study constructs a new quote request system that is useful for both consumers and insurance companies. The system predicts the *average equilibrium contract* (premium, deductible value, and type) based on individual and vehicle characteristics. Hence, consumers can easily find the average contract. With this information, they can find the differences between the offered contracts and the market average. Also, the system benefits insurance companies by allowing them to compare their contract design strategy to our prediction, and a similar system they have in place to ours.

We not only propose a new service that can potentially help consumers and insurance companies make better decisions, but also we are the first to predict the contract (premium, deductible value, and type) rather then only one of these variables. Additionally, we are the first to apply the Multi-Output Random Forest to insurance data.

As a result of MORF, the RMSEs of the prediction of premium and deductible are BR 260.41 and BR 245.02, respectively. Also, the accuracy of the Multi-Output Random Forest (61%) is 11 percentage points greater than that of the baseline model. Moreover, the RMSEs of the MORF for premium and deductible are approximately 42% and 34% of their average values in the training set.

Bibliography

AMITI, M. et al. How did china's wto entry affect us prices? *Journal of International Economics*, Elsevier, p. 103339, 2020.

ARKOLAKIS, C.; COSTINOT, A.; RODRIGUEZ-CLARE, A. New trade models, same old gains? *American Economic Review*, v. 102, n. 1, p. 94–130, 2012.

ARROW, K. J. The theory of risk aversion. *Essays in the theory of risk-bearing*, Markham Chicago, p. 90–120, 1971.

BAI, L.; STUMPNER, S. Estimating us consumer gains from chinese imports. *American Economic Review: Insights*, v. 1, n. 2, p. 209–24, 2019.

BARSEGHYAN, L. et al. The nature of risk preferences: Evidence from insurance choices. *American Economic Review*, v. 103, n. 6, p. 2499–2529, 2013.

BARSEGHYAN, L. et al. Estimating risk preferences in the field. *Journal of Economic Literature*, v. 56, n. 2, p. 501–64, 2018.

BARSEGHYAN, L.; PRINCE, J.; TEITELBAUM, J. C. Are risk preferences stable across contexts? evidence from insurance data. *American Economic Review*, v. 101, n. 2, p. 591–631, 2011.

BORUSYAK, K.; JARAVEL, X. *The Distributional Effects of Trade: Theory and Evidence from the United States*. 2018.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001.

CHAPADOS, N. et al. Estimating car insurance premia: A case study in high-dimensional data inference. In: *Advances in Neural Information Processing Systems*. [S.l.: s.n.], 2002. p. 1369–1376.

CHEN, C. et al. Using random forest to learn imbalanced data. *University of California, Berkeley*, v. 110, n. 1-12, p. 24, 2004.

CHETTY, R. A new method of estimating risk aversion. *American Economic Review*, v. 96, n. 5, p. 1821–1834, 2006.

CICCHETTI, C. J.; DUBIN, J. A. A microeconometric analysis of risk aversion and the decision to self-insure. *Journal of political Economy*, The University of Chicago Press, v. 102, n. 1, p. 169–186, 1994.

COHEN, A.; EINAV, L. Estimating risk preferences from deductible choice. *American economic review*, v. 97, n. 3, p. 745–788, 2007.

COSTA, F.; GARRED, J.; PESSOA, J. P. Winners and losers from a commodities-for-manufactures trade boom. *Journal of International Economics*, Elsevier, v. 102, p. 50–69, 2016.

DAVID, H.; DORN, D.; HANSON, G. H. The china syndrome: Local labor market effects of import competition in the united states. *American Economic Review*, v. 103, n. 6, p. 2121–68, 2013.

DREZE, J. H. Inferring risk tolerance from deductibles in insurance contracts. *Geneva Papers on Risk and Insurance*, JSTOR, p. 48–52, 1981.

FABER, B. Trade liberalization, the price of quality, and inequality: Evidence from mexican store prices. *UC-Berkeley Working Paper*, 2014.

FAJGELBAUM, P. D.; KHANDELWAL, A. K. Measuring the unequal gains from trade. *The Quarterly Journal of Economics*, Oxford University Press, v. 131, n. 3, p. 1113–1180, 2016.

FRANCIS, L. The basics of neural networks demystified. *Contingencies*, v. 11, n. 12, p. 56–61, 2001.

FREITAS, C. C. G. d. *Demanda por seguro de automóvel no Rio de Janeiro*. Tese (Doutorado), 2018.

GLOCKER, B. et al. Joint classification-regression forests for spatially structured multi-object segmentation. In: SPRINGER. *European conference on computer vision*. [S.l.], 2012. p. 870–881.

GUELMAN, L. Gradient boosting trees for auto insurance loss cost modeling and prediction. *Expert Systems with Applications*, Elsevier, v. 39, n. 3, p. 3659–3667, 2012.

GUIDO, S.; MÜLLER, A. *Introduction to machine learning with python*. [S.l.]: O'Reilly Media, 2016. v. 282.

HE, Z. Trade and real wages of the rich and poor: Cross-region evidence. SocArXiv, 2018.

HE, Z. *Trade and real wages of the rich and poor: Evidence from Brazil and Mexico*. [S.l.], 2019.

HOTTMAN, C.; MONARCH, R. Estimating unequal gains across us consumers with supplier trade data. 2018.

JARAVEL, X.; SAGER, E. What are the price effects of trade? evidence from the us and implications for quantitative trade models. 2018.

JASPERSEN, J. G.; RAGIN, M. A.; SYDNOR, J. R. *Predicting Insurance Demand from Risk Attitudes*. [S.l.], 2019.

KACHELMEIER, S. J.; SHEHATA, M. Examining risk preferences under high monetary incentives: Experimental evidence from the people's republic of china. *The American Economic Review*, JSTOR, p. 1120–1141, 1992.

LEDO, B. C. A.; LOPES, C. M. de A. Estimating risk and risk aversion in the automobile insurance market. *Brazilian Review of Econometrics*, v. 39, n. 1, 2019.

LINUSSON, H. *Multi-output random forests*. [S.l.]: University of Borås/School of Business and IT, 2013.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, v. 12, n. Oct, p. 2825–2830, 2011.

PLATT, J. et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, Cambridge, MA, v. 10, n. 3, p. 61–74, 1999.

PORTO, G. *Using survey data to assess the distributional effects of trade policy.* [S.l.: s.n.], 2006. v. 70. 140–160 p.

PRESTON, M. G.; BARATTA, P. An experimental study of the auction-value of an uncertain outcome. *The American journal of psychology*, JSTOR, v. 61, n. 2, p. 183–193, 1948.

SATO, K. The ideal log-change index number. *The Review of Economics and Statistics*, JSTOR, p. 223–228, 1976.

SMITH, V. L.; WALKER, J. M. Rewards, experience and decision costs in first price auctions. *Economic Inquiry*, Wiley Online Library, v. 31, n. 2, p. 237–244, 1993.

SYDNOR, J. (over) insuring modest risks. *American Economic Journal: Applied Economics*, v. 2, n. 4, p. 177–99, 2010.

VARTIA, Y. O. Ideal log-change index numbers. *scandinavian Journal of statistics*, JSTOR, p. 121–126, 1976.

VIAENE, S.; DEDENE, G.; DERRIG, R. A. Auto claim fraud detection using bayesian learning neural networks. *Expert Systems with Applications*, Elsevier, v. 29, n. 3, p. 653–666, 2005.

VIAENE, S. et al. A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection. *Journal of Risk and Insurance*, Wiley Online Library, v. 69, n. 3, p. 373–421, 2002.

WUTHRICH, M. V.; BUSER, C. Data analytics for non-life insurance pricing. *Swiss Finance Institute Research Paper*, n. 16-68, 2019.

YAARI, M. E. Convexity in the theory of choice under risk. *The Quarterly Journal of Economics*, Oxford University Press, v. 79, n. 2, p. 278–290, 1965.

Appendix

APPENDIX A - Appendix from First Chapter

A.1 Summary Statistics

Table A.1 – Summary statistics

	Variables	Mean	Std. Dev.	Min.	Max.	P10	P50	P90
Dependent Variable								
	Var. Price	0.39	0.37	-0.22	2.23	0.12	0.33	0.67
Variables of interest								
	Import Effect Winsor	0.11	0.22	-0.04	0.96	-0.003	0.01	0.39
	Local Effect Winsor	0.11	0.36	-0.60	2.24	0.000002	0.01	0.42
	China Import Effect Winsor	0.04	0.13	-0.0003	0.79	0	0.00004	0.13
	China Export Effect Winsor	0.002	0.02	-0.008	0.15	-0.0002	0	0.0005
Demand controls								
	Income	2903.81	1093.76	1255.41	8593.40	1973.41	2557.15	3992.49
	Income Growth	0.46	0.21	-0.72	1.17	0.25	0.48	0.64
Supply controls								
	Age	31.90	1.41	28.76	37.25	30.72	31.45	33.90
	Hours Worked	43.67	0.27	41.78	43.96	43.46	43.76	43.88
	Income	662.46	319.04	279.92	1925.05	347.28	584.99	1002.47

This table displays the summary statistics of all variables in our data. The 10th, 50th and 90th percentiles are described by columns P10, P50, and P90, respectively. Winsorized variables: import effect, local effect, China import and export effect. Sources: POF, CEPII BACI, IPC-S, and RAIS.

A.2 Summary Statistics - IV

Table A.2 – Summary statistics of instrumental variables

	Mean	Std. Dev.	Min	Max	P10	P50	P90
IV import effect Winsor	1.07	3.18	-2.36	22.17	0.0000006	0.10	2.62
IV local effect Winsor	0.05	0.25	-0.38	2.60	-0.002	0.003	0.10

This table displays the summary statistics of our instrumental variables. The first (second) row shows the IV for the import (local) effect which is the change in China's exports to (imports from) all low-income countries besides Brazil divided by the 2002 expenditure in Brazil. All variables are winsorized.

A.3 Import Effect

Table A.3 – OLS and 2SLS estimates of the import effect, no winsorization

	(1) OLS	(2) 2SLS	(3) 2SLS	(4) 2SLS
Import effect	-0.289***	-0.542**	-0.720***	-0.727**
	(0.0890)	(0.244)	(0.266)	(0.281)
Weight	No	No	Yes	Yes
Controls	No	No	No	Yes
N	129	129	129	129
1st-stage F-stat		8.037	13.54	9.996

The dependent variable of all specifications is the price index. Column (1) shows the OLS estimate of the price index on the import effect, no controls and weight. The remaining columns present 2SLS estimates. The IV for the import effect is the change in Chinese exports to all low-income countries (besides Brazil) divided by the Brazilian expenditure in 2002. Specifications presented in Columns (2) to (4) differ according to the presence of controls and weight. Unless otherwise specified, regressions have as controls import effect, income, income growth (demand side), age, hours worked, and wages (supply side). Additionally, regressions are weighted by Brazilian total expenditure in 2002. No winsorized variables. Robust standard errors are in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

A.4 China and ROW Effects

Table A.4 – Effect of the IV import effect on China and ROW effect, no winsorization

	(1) OLS	(2) OLS
	China effect	ROW effect
IV	0.0360**	0.00471
	(0.0138)	(0.00448)
Weight	Yes	Yes
Controls	Yes	Yes
N	129	129

The IV for the import effect which is the ratio between changes in Chinese exports to all low-income countries besides Brazil and the Brazilian expenditure in 2002. Column (1) shows the regression of the China effect on the IV. Similarly, Column (2) presents the regression of the ROW effect on the IV. Both regressions are estimated by OLS and the other covariates are income, income growth, age, worked hours, and wage. Also, all regressions are weighted by total consumption in 2002. No winsorized variables. Robust standard errors are in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

A.5 China Import Effect

Table A.5 – China import effect on Brazilian domestic prices, no winsorization

	(1) OLS	(2) 2SLS	(3) 2SLS	(4) 2SLS
China import effect	-0.273***	-0.635*	-0.886**	-0.823**
	(0.103)	(0.334)	(0.376)	(0.354)
Weight	No	No	Yes	Yes
Controls	No	No	No	Yes
N	129	129	129	129
1st-stage F-stat		5.424	7.795	6.772

All specifications have the same dependent variable: price index in Brazil. The variable of interest is the change in Brazilian imports from China divided by the Brazilian expenditure in 2002 (China import effect). The OLS estimate of the regression of the price index on the China import effect is shown in Column (1). Columns (2) to (4) present the 2SLS estimates of the regression in Column (1), but we use an IV for China import effect which is the difference in China's exports to all low-income countries besides Brazil divided by the Brazilian expenditure in 2002. Unless otherwise specified, regressions are weighted by 2002 Brazilian consumption, and have as controls: income, income growth, age, hours worked, and wages. No winsorized variables. Robust standard errors are in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

A.6 Local Effect

Table A.6 – Effect of expenditure on locally produced goods and IV on prices, no winsorization

	(1) OLS	(2) OLS	(3) OLS	(4) OLS
Panel B. Local effect	0.0858***	0.0533	0.0763***	0.0386
	(0.0119)	(0.0387)	(0.0149)	(0.0389)
Panel B. IV local effect	0.213	-0.205	0.226	-0.181
	(0.226)	(0.242)	(0.186)	(0.246)
Weight	No	Yes	No	Yes
Controls	No	No	Yes	Yes
N	129	129	129	129

All specifications are estimated by OLS. Panel A shows the relationship between the Brazilian price index and the local effect. Furthermore, Panel B displays the results of the regressions of the price index on the IV for the local effect which is the change in China's imports from all low-income countries (besides Brazil) divided by the total expenditure in Brazil during 2002. Unless otherwise specified, regressions are weighted by Brazilian consumption in 2002. Also, the controls are: income, income growth, age, hours worked, and wages. No winsorized variables. Robust standard errors are in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

A.7 China Export Effect

Table A.7 – China export effect on Brazilian domestic prices, no winsorization

	(1) OLS	(2) 2SLS	(3) 2SLS	(4) 2SLS
China export effect	-0.272	9.602	-0.608	-0.557
	(0.696)	(22.73)	(0.464)	(0.526)
Weight	No	No	Yes	Yes
Controls	No	No	No	Yes
N	129	129	129	129
1st stage F-stat		0.402	4.666	4.630

China export effect is the change in Brazilian exports to China divided by the expenditure in 2002. All 4 specifications the have Brazilian price index as the dependent variable. Column (1) shows the OLS estimate of the regression of the price index on the China export effect, no covariates and weight. Additionally, Columns (2) to (4) use the difference in China's imports from all low-income countries (besides Brazil) divided by Brazilian consumption in 2002 as an instrument for China export effect. Unless otherwise specified, our covariates are: income, income growth, age, hours worked and wages. Furthermore, Brazilian total expenditure in 2002 at the sector level is used as weight. No winsorized variables. Robust standard errors are in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

A.8 Alternative Instrumental Variable

We use the instrumental variable proposed by Costa et al. (2016).

Import Effect

Table A.8 – OLS and 2SLS estimates of the import effect, alternative IV and winsorized

	(1) OLS	(2) 2SLS	(3) 2SLS	(4) 2SLS
Import effect	-0.362***	-0.618***	-0.684**	-0.672**
	(0.0917)	(0.224)	(0.337)	(0.283)
Weight	No	No	Yes	Yes
Controls	No	No	No	Yes
N	129	129	129	129
1st stage F-stat		30.15	39.59	25.26

The dependent variable of all specifications is the price index. Column (1) shows the OLS estimate of the price index on the import effect, no controls and weight. The remaining columns present 2SLS estimates. The IV for the import effect is proposed by Costa et al. (2016). Specifications presented in Columns (2) to (4) differ according to the presence of controls and weight. Unless otherwise specified, regressions have as controls import effect, income, income growth (demand side), age, hours worked, and wages (supply side). Additionally, regressions are weighted by Brazilian total expenditure in 2002. IV and import effect are winsorized. Robust standard errors are in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.9 – OLS and 2SLS estimates of the import effect, alternative IV and no winsorization

	(1) OLS	(2) 2SLS	(3) 2SLS	(4) 2SLS
Import effect	-0.289***	-0.539**	-0.654*	-0.609**
	(0.0890)	(0.257)	(0.338)	(0.302)
Weight	No	No	Yes	Yes
Controls	No	No	No	Yes
N	129	129	129	129
1st stage F-stat		6.821	8.913	6.714

The dependent variable of all specifications is the price index. Column (1) shows the OLS estimate of the price index on the import effect, no controls and weight. The remaining columns present 2SLS estimates. The IV for the import effect is proposed by Costa et al. (2016). Specifications presented in Columns (2) to (4) differ according to the presence of controls and weight. Unless otherwise specified, regressions have as controls import effect, income, income growth (demand side), age, hours worked, and wages (supply side). Additionally, regressions are weighted by Brazilian total expenditure in 2002. No winsorized variables. Robust standard errors are in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

China and ROW Effects

Table A.10 – Effect of the alternative IV import effect on China and ROW effect, winsorized

	(1) OLS	(2) OLS
	China effect	RoW effect
IV	0.0388***	0.00566
	(0.00961)	(0.00495)
Weight	Yes	Yes
Controls	Yes	Yes
N	129	129

The IV for the import effect based on Costa et al. (2016). Column (1) shows the regression of the China effect on the IV. Similarly, Column (2) presents the regression of the ROW effect on the IV. Both regressions are estimated by OLS and the other covariates are income, income growth, age, worked hours, and wage. Also, all regressions are weighted by total consumption in 2002. IV, China and ROW effects are winsorized. Robust standard errors are in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.11 – Effect of the alternative IV for import effect on China and ROW effect, no winsorized

	(1) OLS	(2) OLS
	China effect	RoW effect
IV	0.0332**	0.00176
	(0.0132)	(0.00647)
Weight	Yes	Yes
Controls	Yes	Yes
N	129	129

The IV for the import effect based on Costa et al. (2016). Column (1) shows the regression of the China effect on the IV. Similarly, Column (2) presents the regression of the ROW effect on the IV. Both regressions are estimated by OLS and the other covariates are income, income growth, age, worked hours, and wage. Also, all regressions are weighted by total consumption in 2002. No variables are winsorized. Robust standard errors are in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

China Import Effect

Table A.12 – China import effect on Brazilian domestic prices, alternative IV and winsorized

	(1) OLS	(2) 2SLS	(3) 2SLS	(4) 2SLS
China import effect	-0.431***	-0.716**	-0.840*	-0.764**
	(0.111)	(0.299)	(0.425)	(0.338)
Weight	No	No	Yes	Yes
Controls	No	No	No	Yes
N	129	129	129	129
1st stage F-stat		18.15	18.88	16.32

All specifications have the same dependent variable: price index in Brazil. The variable of interest is the change in Brazilian imports from China divided by the Brazilian expenditure in 2002 (China import effect). The OLS estimate of the regression of the price index on the China import effect is shown in Column (1). Columns (2) to (4) present the 2SLS estimates of the regression in Column (1), but we use an IV for China import effect that is similar to the variable proposed by Costa et al. (2016). Unless otherwise specified, regressions are weighted by 2002 Brazilian consumption, and have as controls: income, income growth, age, hours worked, and wages. IV and China import effect are winsorized. Robust standard errors are in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.13 – China import effect on Brazilian domestic prices, alternative IV no winsorization

	(1) OLS	(2) 2SLS	(3) 2SLS	(4) 2SLS
China import effect	-0.273***	-0.609*	-0.748*	-0.641*
	(0.103)	(0.337)	(0.415)	(0.357)
Weight	No	No	Yes	Yes
Controls	No	No	No	Yes
N	129	129	129	129
1st stage F-stat		4.933	7.382	6.295

All specifications have the same dependent variable: price index in Brazil. The variable of interest is the change in Brazilian imports from China divided by the Brazilian expenditure in 2002 (China import effect). The OLS estimate of the regression of the price index on the China import effect is shown in Column (1). Columns (2) to (4) present the 2SLS estimates of the regression in Column (1), but we use an IV for China import effect that is similar to the variable proposed by Costa et al. (2016). Unless otherwise specified, regressions are weighted by 2002 Brazilian consumption, and have as controls: income, income growth, age, hours worked, and wages. No winsorized variables. Robust standard errors are in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Local Effect

Table A.14 – Effect of alternative IV for local effect on prices, winsorized

	(1) OLS	(2) OLS	(3) OLS	(4) OLS
IV local effect	0.216	-0.500*	0.201	-0.527*
	(0.264)	(0.288)	(0.207)	(0.314)
Weight	No	Yes	No	Yes
Controls	No	No	Yes	Yes
N	129	129	129	129

All specifications are estimated by OLS. The table displays the results of the regressions of the price index on the IV for the local effect based on Costa et al. (2016). Unless otherwise specified, regressions are weighted by Brazilian consumption in 2002. Also, the controls are: income, income growth, age, hours worked, and wages. IV is winsorized. Robust standard errors are in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.15 – Effect of alternative IV for local effect on prices, no winsorized

	(1) OLS	(2) OLS	(3) OLS	(4) OLS
IV local effect	0.0777**	-0.0309	0.0653**	-0.0300
	(0.0352)	(0.0896)	(0.0330)	(0.0852)
Weight	No	Yes	No	Yes
Controls	No	No	Yes	Yes
N	129	129	129	129

All specifications are estimated by OLS. The table displays the results of the regressions of the price index on the IV for the local effect based on Costa et al. (2016). Unless otherwise specified, regressions are weighted by Brazilian consumption in 2002. Also, the controls are: income, income growth, age, hours worked, and wages. No winsorized variables. Robust standard errors are in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

China Export Effect

Table A.16 – China export effect on Brazilian domestic prices, alternative IV and winsorized

	(1) OLS	(2) 2SLS	(3) 2SLS	(4) 2SLS
China export effect	-1.803**	18.08	-4.724	-4.946
	(0.826)	(33.24)	(3.852)	(4.044)
Weight	No	No	Yes	Yes
Controls	No	No	No	Yes
N	129	129	129	129
1st stage F-stat		1.022	4.417	4.767

China export effect is the change in Brazilian exports to China divided by the expenditure in 2002. All 4 specifications the have Brazilian price index as the dependent variable. Column (1) shows the OLS estimate of the regression of the price index on the China export effect, no covariates and weight. Additionally, Columns (2) to (4) use the IV proposed by Costa et al. (2016) as an instrument for the China export effect. Unless otherwise specified, our covariates are: income, income growth, age, hours worked and wages. Furthermore, Brazilian total expenditure in 2002 at the sector level is used as weight. IV and China export effect are winsorized. Robust standard errors are in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.17 – China export effect on Brazilian domestic prices, alternative IV and no winsorization

	(1) OLS	(2) 2SLS	(3) 2SLS	(4) 2SLS
China export effect	-0.272	44.57	-0.720	-0.810
	(0.696)	(70.19)	(1.547)	(1.730)
Weight	No	No	Yes	Yes
Controls	No	No	No	Yes
N	129	129	129	129
1st stage F-stat		0.403	0.699	0.697

China export effect is the change in Brazilian exports to China divided by the expenditure in 2002. All 4 specifications the have Brazilian price index as the dependent variable. Column (1) shows the OLS estimate of the regression of the price index on the China export effect, no covariates and weight. Additionally, Columns (2) to (4) use the IV proposed by Costa et al. (2016) as an instrument for the China export effect. Unless otherwise specified, our covariates are: income, income growth, age, hours worked and wages. Furthermore, Brazilian total expenditure in 2002 at the sector level is used as weight. No winsorized variables. Robust standard errors are in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

APPENDIX B - Appendix from Second Chapter

B.1 Summary Statistics

Table B.1 – Description of the variables in the data set

Variables	Description
Individuals	
Age	Corresponds to the age of the main driver of the vehicle.
Female	Dummy variable which assumes 1 if the policyholder is female.
Geographic region	Region where the main driver lives. We consider the metropolitan region of São Paulo: *Barueri, Caieiras, Cajamar, Carapicuíba, Cotia, Diadema, Embu, Embu-Guaçu, Francisco Morato, Franco da Rocha, Guarulhos, Itapecerica da Serra, Itapevi, Jandira, Juquitiba, Mairiporã, Mauá, Osasco, Pirapora do Bom Jesus, Ribeirão Pires, Rio Grande da Serra, Santana de Parnaíba, Santo André, São Bernardo do Campo, São Caetano do Sul, São Paulo, Taboão da Serra, Vargem Grande Paulista.*
Bonus	Percentage discount on the premium the policyholder has according to her claim history.
Vehicle	
Value	Insured value.
Model	Model of the vehicle insured. Cars with 1,000cc engines were considered: *Pegeout 206, Atos, Celta, Clio, Corsa, Fiesta, Fox, Gol, Ibiza, Ka, Kangoo, Palio, Parati, Polo, Siena, Twingo.* Dummy variables were created for the empirical models.
Year of the model	The year of the models range from 1998 to 2004.
Insurance Contract	
Premium	Price paid to the insurer for the insured vehicle.
Deductible value	Value established in the contract which there is no reimbursement for losses under this value.
Deductible type	There are 3 deductible types: low, regular and high. We record high deductible as regular and create a dummy variable which assume value 1 if the deductible is low, and zero otherwise.
Claims	
Indemnities	The value payed by the insurance company in case of a claim. It is the difference between the total repair cost and the deductible.
Claim	Dummy variable that assumes 1 if a claim was made and zero otherwise.

This table describes all the variables in the data. Source: SUSEP.

B.2 Summary Statistics - covariates

Table B.2 – Comprehensive summary statistics - covariates

Variables	Deduc. type	Mean	Std. dev.	Min	Max
Individual	Female	0.49	0.5		
	Age 18-25	0.07	0.26		
	Age 26-35	0.28	0.45		
	Age 36-45	0.28	0.45		
	Age 46-55	0.22	0.42		
	Age 56-65	0.1	0.3		
	Age more than 65	0.05	0.21		
	Discount	18.55	12.97	0	45
Car attributes	Value	15929.9	4389.4	2500	55215
	Corsa - General Motors	0.28	0.45		
	Palio - Fiat	0.24	0.43		
	Gol - Volkswagen	0.13	0.33		
	Fiesta - Ford	0.08	0.27		
	KA - Ford	0.08	0.27		
	Celta - General Motors	0.08	0.27		
	Clio - Renault	0.04	0.2		
	Siena - Fiat	0.03	0.17		
	Parati - Volkswagen	0.01	0.1		
	Atos - Hyundai	0.003	0.05		
	Kangoo - Renault	0.001	0.04		
	Twingo - Renault	0.001	0.03		
	Polo - Volkswagen	0.0006	0.02		
	Fox - Volkswagen	0.0002	0.02		
	Ibiza - Ford	0	0.01		
	Year 2004	0.06	0.23		
	Year 2003	0.16	0.37		
	Year 2002	0.19	0.39		
	Year 2001	0.23	0.42		
	Year 2000	0.13	0.34		
	Year 1999	0.12	0.32		
	Year 1998	0.1	0.31		

This table shows the summary statistics of all covariates. We divide the variables into two groups: individual characteristics, and car's attributes. Source: SUSEP.

B.3 Out-of-bag error - indemnity value

Table B.3 – Out-of-bag error for indemnity estimation

Number of trees	Maximum depth	Average RMSE
50	3	1587
50	4	1591
50	5	1593
50	6	1600
50	7	1602
50	8	1608
50	9	1610
50	10	1613
100	3	1586
100	4	1590
100	5	1592
100	6	1597
100	7	1598
100	8	1604
100	9	1607
100	10	1610
500	3	1586
500	4	1589
500	5	1592
500	6	1595
500	7	1597
500	8	1602
500	9	1604
500	10	1605
1000	3	1587
1000	4	1589
1000	5	1591
1000	6	1594
1000	7	1597
1000	8	1601
1000	9	1603
1000	10	1605

To select the parameters of the Random Forest that best predicts indemnities, we use the out-of-bag error, and we evaluate model's performance based on the RMSE. Also, we consider the following grid of hyperparameters: number of trees assume values 10, 50, 100, 500 and 1000; and, the maximum depth of the trees assume values 3, 4, 5, 6, 7, 8, 9, and 10.

B.4 Out-of-bag error - claims

Table B.4 – Out-of-bag error for predicting if a claim is made

Number of trees	Maximum depth	AUROC
50	3	0.5611
50	4	0.5637
50	5	0.5667
50	6	0.5566
50	7	0.5522
50	8	0.5502
50	9	0.5444
50	10	0.5451
100	3	0.5646
100	4	0.5660
100	5	0.5672
100	6	0.5637
100	7	0.5571
100	8	0.5517
100	9	0.5493
100	10	0.5496
500	3	0.5672
500	4	0.5669
500	5	0.5681
500	6	0.5661
500	7	0.5630
500	8	0.5572
500	9	0.5549
500	10	0.5495
1000	3	0.5668
1000	4	0.5672
1000	5	0.5675
1000	6	0.5663
1000	7	0.5632
1000	8	0.5585
1000	9	0.5554
1000	10	0.5487

Out-of-bag error is used to select the hyperparameters of a Weighted Random Forest that best predicts if a claim is made or not. We consider the following grid of parameters: number of trees assume values 50, 100, 500 and 1000; and, the maximum depth of the trees assume values 3, 4, 5, 6, 7, 8, 9, and 10. The model selection is based on the highest AU ROC.

B.5 OLS - train data - risk aversion

Table B.5 – OLS regression in the train dataset, the log of risk aversion is the dependent variable.

Variable	coef.	std. err	[0.025 0.095]
Intercept	-8.353***	0.068	[-8.5951 -8.3311]
Age	-0.001	0.0007	[-0.002 0.0004]
Age^2	0.047***	0.0181	[0.0218 0.0864]
Female	0.027***	0.0011	[0.0259 0.0303]
Log(Value)	0.094***	0.0086	[0.085 0.117]
Bonus	0.006***	0.0002	[0.0055 0.0062]
Atos	-0.163***	0.0103	[-0.1715 -0.1337]
Celta	0.165***	0.0075	[0.1314 0.1653]
Clio	-0.026*	0.0073	[-0.0226 0.0042]
Corsa	0.091***	0.0099	[0.0842 0.1254]
Fiesta	0.088***	0.0116	[0.0819 0.1296]
Fox	0.174***	0.0143	[0.1606 0.2166]
Gol	0.083***	0.0082	[0.0817 0.1141]
Ibiza	-0.208***	0.0081	[-0.2144 -0.1843]
Kangoo	-0.249***	0.0126	[-0.2582 -0.2083]
Ka	0.103***	0.0089	[0.0943 0.1289]
Palio	0.072***	0.0061	[0.0683 0.0915]
Parati	-0.011	0.0112	[-0.0207 0.0218]
Polo	-0.071*	0.0216	[-0.0819 0.0023]
Siena	0.066***	0.0085	[0.0617 0.0936]
Twingo	0.228***	0.0095	[0.2182 0.2563]
Year 1998	0.114***	0.0056	[0.1199 0.1413]
Year 1999	0.115***	0.003	[0.1182 0.1304]
Year 2000	0.103***	0.0032	[0.106 0.1187]
Year 2001	0.092***	0.0029	[0.0967 0.1086]
Year 2002	0.076***	0.0022	[0.0778 0.0867]
Year 2003	0.038***	0.0036	[0.031 0.0444]

This table displays the OLS estimate of the regression of the log of risk aversion on the covariates described at Table 2.1. We use the train data, $N = 100,000$. The second column shows the values of the coefficients; third column presents the bootstrapped standard errors as described in Appendix B.13. Additionally, the last column displays the bootstrapped confidence interval for 5% significance level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

B.6 OLS - test data - risk aversion

Table B.6 – OLS regression in the test dataset, the log of risk aversion as the dependent variable.

Variable	coef.	std. err	[0.025 0.095]
Intercept	-8.132***	0.0684	[-8.1764 -7.8969]
Age	0.004***	0.0008	[0.0035 0.0066]
Age^2	-0.00002***	0.00001	[-0.00004 -0.00004]
Female	0.027***	0.001	[0.0262 0.0299]
Log(Value)	0.091***	0.0097	[0.0581 0.0951]
Bonus	0.006***	0.0002	[0.0055 0.0063]
Atos	-0.146***	0.0111	[-0.152 -0.1114]
Celta	0.159***	0.0085	[0.1511 0.1852]
Clio	-0.006	0.0072	[-0.0127 0.0148]
Corsa	0.104***	0.0093	[0.1026 0.1403]
Fiesta	0.098***	0.0106	[0.095 0.1384]
Fox	-0.009	0.0154	[-0.0166 0.0414]
Gol	0.094	0.0074	[0.0929 0.121]
Ibiza	-0.262***	0.0074	[-0.2648 -0.2349]
Kangoo	-0.289***	0.0122	[-0.2973 -0.2469]
Ka	0.097***	0.0083	[0.0894 0.1251]
Palio	0.08***	0.0054	[0.0768 0.0995]
Parati	-0.001	0.0109	[-0.0035 0.0356]
Polo	-0.167***	0.0201	[-0.1759 -0.0983]
Siena	0.07***	0.0063	[0.0594 0.0889]
Twingo	0.179***	0.0097	[0.1745 0.213]
Year 1998	0.108***	0.0062	[0.1063 0.1304]
Year 1999	0.109***	0.0039	[0.1077 0.1225]
Year 2000	0.098***	0.0038	[0.0977 0.1114]
Year 2001	0.082***	0.0035	[0.0813 0.0946]
Year 2002	0.065***	0.0025	[0.0629 0.0731]
Year 2003	0.028***	0.0053	[0.0074 0.0288]

This table displays regression of the log of risk aversion on the covariates described at Table 2.1 estimated by OLS. We use the test set that has 25,000 observations. The second column shows the values of the coefficients; third column presents the bootstrapped standard errors as described in Appendix B.13. Additionally, the last column displays the bootstrapped confidence interval for 5% significance level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

B.7 Logit - train data - risk aversion

Table B.7 – Logit regression in the train dataset with the deductible choice as the dependent variable.

Variable	coef.	std. err	[0.025 0.095]
Intercept	2.158***	0.1946	[1.527 2.316]
Log(risk aversion)	13.438***	0.2434	[13.194 14.212]
Age	-0.314***	0.0107	[-0.322 -0.279]
Age^2	6.627***	0.2275	[5.893 6.828]
Female	-0.084	0.0852	[-0.176 0.154]
Log(Value)	4.748***	0.1903	[4.781 5.569]
Bonus	-0.038***	0.0026	[-0.044 -0.033]
Atos	0.66***	0.1022	[0.336 0.755]
Celta	1.072***	0.0716	[1.064 1.326]
Clio	1.268***	0.1025	[0.807 1.247]
Corsa	0.109	0.0522	[-0.11 0.113]
Fiesta	0.324***	0.0763	[0.094 0.369]
Fox	-0.019***	0.0029	[-0.022 -0.01]
Gol	-0.131**	0.0605	[-0.3 -0.044]
Ibiza	-0.002***	0.00005	[-0.003 -0.002]
Kangoo	0.33***	0.0521	[0.165 0.377]
Ka	0.757***	0.0631	[0.713 0.987]
Palio	0.275**	0.0861	[0.116 0.43]
Parati	-0.998***	0.0756	[-1.056 -0.733]
Polo	0.154***	0.0268	[0.078 0.18]
Siena	-1.596***	0.0807	[-1.661 -1.299]
Twingo	-0.346***	0.0527	[-0.396 -0.178]
Year 1998	3.058***	0.1504	[2.517 3.138]
Year 1999	2.34***	0.1756	[1.799 2.475]
Year 2000	2.126***	0.2772	[1.324 2.297]
Year 2001	1.467***	0.1735	[0.77 1.581]
Year 2002	1.053***	0.2184	[0.213 1.126]
Year 2003	0.356	0.2655	[-0.62 0.435]

The dependent variable is a dummy which assumes value 1 if the low deductible type was chosen, and zero otherwise. The controls are the variables described in Table 2.1 and the log of the estimated risk aversion. We run a logit in the train data that has 100,000 observations. The second column shows the estimated values of each coefficient; third column presents the bootstrapped standard errors; the last column displays the bootstrapped confidence interval for a 5% significance level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

B.8 Logit - test data - risk aversion

Table B.8 – Logit regression in the test set with the deductible choice as the dependent variable.

Variable	coef.	std. err	[0.025 0.095]
Intercept	0.5661***	0.2071	[0.465 0.985]
Log(risk aversion)	10.424***	1.2083	[8.368 12.647]
Age	0.1885***	0.1046	[-0.029 0.396]
Age^2	-0.0018***	0.0011	[-0.004 0.0]
Female	1.1825	0.5411	[-0.135 1.183]
Log(Value)	6.2573***	0.9186	[4.913 7.928]
Bonus	-0.0119	0.0157	[-0.04 0.026]
Atos	0.0533***	0.0145	[0.044 0.082]
Celta	0.4579***	0.0945	[0.205 0.464]
Clio	0.6738***	0.0727	[0.446 0.687]
Corsa	0.0494***	0.0657	[0.015 0.192]
Fiesta	0.1231	0.1098	[-0.153 0.125]
Fox	-0.0037***	0.0014	[-0.007 -0.003]
Gol	-0.6324***	0.0897	[-0.79 -0.508]
Ibiza	-0.0022***	0.0002	[-0.003 -0.002]
Kangoo	-0.0026	0.0054	[-0.003 0.01]
Ka	-0.0481	0.2529	[-0.08 0.52]
Palio	-0.3508	0.5068	[-0.385 0.812]
Parati	0.0159	0.1086	[-0.23 0.027]
Polo	0.0002	0.0017	[-0.004 0.0003]
Siena	-0.0908***	0.1369	[-0.391 -0.065]
Twingo	-0.0286***	0.0169	[-0.064 -0.024]
Year 1998	-0.464	0.5291	[-0.51 0.753]
Year 1999	-0.3034	0.4691	[-0.346 0.766]
Year 2000	0.1552***	0.4894	[0.075 1.228]
Year 2001	0.239***	0.1999	[0.187 0.667]
Year 2002	0.4904	0.3821	[-0.408 0.518]
Year 2003	0.344	0.7948	[-1.452 0.424]

The dependent variable is a dummy which assumes value 1 if the low deductible type was chosen, and zero otherwise. The controls are the variables described in Table 2.1 and the log of the estimated risk aversion. We run a logit in the test data that has 25,000 observations. The second column shows the estimated values of each coefficient; third column presents the bootstrapped standard errors; the last column displays the bootstrapped confidence interval for a 5% significance level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

B.9 Logit - train data - risk aversion, no contract

Table B.9 – Logit regression in the train set with the deductible choice as the dependent variable.

Variable	coef.	std. err	[0.025 0.095]
Intercept	-0.101***	0.0086	[-0.114 -0.081]
Log(risk aversion)	0.662***	0.0516	[0.492 0.696]
Age	0.026***	0.0007	[0.023 0.026]
Age^2	-0.394***	0.0129	[-0.402 -0.35]
Female	0.194***	0.0098	[0.186 0.225]
Log(Value)	0.382***	0.0422	[0.233 0.41]
Bonus	0.038***	0.0002	[0.038 0.039]
Atos	0.075***	0.006	[0.056 0.081]
Celta	0.324***	0.01	[0.296 0.336]
Clio	0.122**	0.0379	[0.036 0.184]
Corsa	-0.194***	0.0104	[-0.213 -0.174]
Fiesta	-0.034***	0.0101	[-0.05 -0.012]
Fox	0.036***	0.0054	[0.024 0.045]
Gol	-0.426***	0.0197	[-0.439 -0.362]
Ibiza	-0.012***	0.002	[-0.015 -0.008]
Kangoo	0.021***	0.0014	[0.016 0.022]
Ka	-0.048***	0.0067	[-0.055 -0.028]
Palio	-0.218***	0.0129	[-0.244 -0.198]
Parati	-0.137***	0.0332	[-0.196 -0.074]
Polo	0.07***	0.0099	[0.047 0.085]
Siena	-0.61***	0.0303	[-0.602 -0.476]
Twingo	-0.019***	0.0038	[-0.025 -0.011]
Year 1998	-0.142***	0.0246	[-0.196 -0.099]
Year 1999	0.041***	0.0166	[0.008 0.066]
Year 2000	0.058***	0.0199	[0.015 0.09]
Year 2001	0.072***	0.0106	[0.049 0.09]
Year 2002	0.077***	0.0153	[0.062 0.115]
Year 2003	-0.092***	0.0098	[-0.096 -0.058]

The dependent variable is a dummy which assumes value 1 if the low deductible type was chosen, and zero otherwise. The controls are the variables described in Table 2.1 and the log of the estimated risk aversion. We run a logit in the train data that has 100,000 observations. The second column shows the estimated values of each coefficient; third column presents the bootstrapped standard errors; the last column displays the bootstrapped confidence interval for a 5% significance level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

B.10 Logit - test data - risk aversion, no contract

Table B.10 – Logit regression in the test set with the deductible choice as the dependent variable.

Variable	coef.	std. err	[0.025 0.095]
Intercept	-0.026***	0.0044	[-0.043 -0.026]
Log(risk aversion)	0.181***	0.0238	[0.18 0.278]
Age	0.01528***	0.0008	[0.015 0.015]
Age^2	-0.00007**	0.00005	[-0.00007 -0.00005]
Female	0.050***	0.04067	[0.05 0.23]
Log(Value)	-0.220***	0.0170	[-0.22 -0.151]
Bonus	0.038***	0.00004	[0.038 0.038]
Atos	0.001***	0.0025	[0.001 0.01]
Celta	0.015***	0.0226	[0.015 0.121]
Clio	0.023***	0.0339	[0.023 0.156]
Corsa	-0.005***	0.0063	[-0.009 -0.005]
Fiesta	0.004***	0.0048	[0.004 0.029]
Fox	-0.0002***	0.0004	[-0.001 -0.0]
Gol	-0.030***	0.0381	[-0.179 -0.029]
Ibiza	-0.0001***	0.0002	[-0.001 -0.0001]
Kangoo	-0.0005***	0.0013	[-0.004 -0.001]
Ka	-0.008***	0.0105	[-0.046 -0.008]
Palio	-0.030***	0.0232	[-0.168 -0.03]
Parati	-0.001**	0.0004	[-0.001 -0.001]
Polo	-0.0002***	0.0003	[-0.001 -0.0002]
Siena	-0.002***	0.0054	[-0.014 -0.002]
Twingo	-0.0008***	0.0018	[-0.006 -0.001]
Year 1998	-0.027***	0.0340	[-0.183 -0.027]
Year 1999	-0.018***	0.0192	[-0.118 -0.018]
Year 2000	0.0008***	0.0051	[0.001 0.005]
Year 2001	-0.0003*	0.0049	[-0.0003 0.02]
Year 2002	0.012***	0.0189	[0.012 0.117]
Year 2003	0.008***	0.0156	[0.008 0.097]

The dependent variable is a dummy which assumes value 1 if the low deductible type was chosen, and zero otherwise. The controls are the variables described in Table 2.1 and the log of the estimated risk aversion. We run a logit in the test data that has 25,000 observations. The second column shows the estimated values of each coefficient; third column presents the bootstrapped standard errors; the last column displays the bootstrapped confidence interval for a 5% significance level.* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

B.11 OLS indemnity - train data

Table B.11 – Regression of premium on predicted indemnity, train data.

Variable	coef.	std. err	[0.025 0.095]
Intercept	375.44	41.21	[266.41 432.78]
Indemnity	6.92	1.2	[5.37 10.23]

This table displays an OLS estimate of premium on the estimated indemnity in the train data. Bootstrapped standard errors and confidence interval. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

B.12 OLS indemnity - test data

Table B.12 – Regression of premium on predicted indemnity, train data.

Variable	coef.	std. err	[0.025 0.095]
Intercept	373.75	39.69	[269.29 430.1]
Indemnity	6.92	1.15	[5.3 9.97]

This table displays an OLS estimate of premium on the estimated indemnity in the test set. Bootstrapped standard errors and confidence interval. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

B.13 Bootstrap

In the results presented at Section 2.4, the variable of interest in the reduced-form regressions is the true lower bound of the coefficient of RA. But, we only have its estimate, and this may affect inference. Therefore, we use bootstrap to obtain the standard errors and derive the confidence intervals. Our bootstrap follows the steps bellow:

1. Generate a stratified random sample (hold constant the number of claims, 1.8%) with replacement from the train and test sets. These samples have the same number of observations as the original sets.

2. Find the best-performing models over the grid of parameters described in Section 2.4.

3. Run those models in the training data and predict the insurer's expected claim cost and the probability to make a claim (using the calibrator) in both datasets (train and test). Use these values to compute the expected indemnity for each individual and follow Equation (2.1) to obtain k. Use this estimate and deductible value to compute the lower bound of the coefficient of risk aversion, as presented in Equation (2.4), in the train and test sets.

4. Run the OLS and logistic models and keep the values of the coefficient of each feature in each dataset.

5. After doing steps 1-4 500 times, compute the confidence interval using the estimated coefficients.

APPENDIX C - Appendix from Third Chapter

C.1 Description of the variables

Table C.1 – Description of the variables in the data set

Variables	Description
Individuals	
Age	Corresponds to the age of the main driver of the vehicle.
Female	Dummy variable which assumes 1 if the policyholder is female.
Geographic region	Region where the main driver lives. We consider the metropolitan region of São Paulo: *Barueri, Caieiras, Cajamar, Carapicuíba, Cotia, Diadema, Embu, Embu-Guaçu, Francisco Morato, Franco da Rocha, Guarulhos, Itapecerica da Serra, Itapevi, Jandira, Juquitiba, Mairiporã, Mauá, Osasco, Pirapora do Bom Jesus, Ribeirão Pires, Rio Grande da Serra, Santana de Parnaíba, Santo André, São Bernardo do Campo, São Caetano do Sul, São Paulo, Taboão da Serra, Vargem Grande Paulista.*
Bonus	Percentage discount on the premium the policyholder has according to her claim history.
Vehicle	
Value	Insured value.
Model	Model of the vehicle insured. Cars with 1,000cc engines were considered: *Pegeout 206, Atos, Celta, Clio, Corsa, Fiesta, Fox, Gol, Ibiza, Ka, Kangoo, Palio, Parati, Polo, Siena, Twingo.* Dummy variables were created for the empirical models.
Year of the model	The year of the models range from 1998 to 2004.
Insurance Contract	
Premium	Price paid to the insurer for the insured vehicle.
Deductible value	Value established in the contract which there is no reimbursement for losses under this value.
Deductible type	There are 3 deductible types: low, regular and high. We record high deductible as regular and create a dummy variable which assume value 1 if the deductible is low, and zero otherwise.

This table displays a comprehensive description the variables in the data. Source: SUSEP.

C.2 Summary statistics - covariates

Table C.2 – Comprehensive summary statistics - covariates

Variables	Deduc. type	Mean	Std. dev.	Min	Max
Individual	Female	0.51	0.5		
	Age 18-25	0.06	0.24		
	Age 26-35	0.26	0.44		
	Age 36-45	0.28	0.45		
	Age 46-55	0.23	0.42		
	Age 56-65	0.11	0.31		
	Age more than 65	0.05	0.22		
	Bonus	20.91	13.03	0	45
Car attributes	Value	16113.5	4414.49	2300	64800
	Corsa - General Motors	0.27	0.45		
	Palio - Fiat	0.23	0.42		
	Gol - Volkswagen	0.12	0.32		
	Fiesta - Ford	0.08	0.28		
	KA - Ford	0.08	0.27		
	Celta - General Motors	0.08	0.28		
	Clio - Renault	0.05	0.22		
	206 - Pegeout	0.04	0.18		
	Siena - Fiat	0.03	0.17		
	Parati - Volkswagen	0.01	0.11		
	Atos - Hyundai	0.003	0.06		
	Kangoo - Renault	0.001	0.04		
	Twingo - Renault	0.001	0.03		
	Polo - Volkswagen	0.0006	0.02		
	Fox - Volkswagen	0.0001	0.01		
	Ibiza - Ford	0.0001	0.01		
	Year 2004	0.05	0.23		
	Year 2003	0.16	0.37		
	Year 2002	0.2	0.4		
	Year 2001	0.24	0.42		
	Year 2000	0.14	0.34		
	Year 1999	0.11	0.31		
	Year 1998	0.09	0.29		

This table displays a comprehensive description of our covariates followed by their summary statistics. The last 4 columns of the table shows the mean, standard deviation, the minimum and maximum values of each variable. Also, we split our covariates into two groups: individuals' characteristics and cars' attributes.

C.3 Cross-validation MORF

Table C.3 – 10-fold cross-validation MORF

Num. of tress	ACC	RMSE - Deduc.	RMSE - Prem.
10	0.5876	245.846	264.412
50	0.6005	245.035	263.652
100	0.6031	244.575	262.776
200	0.6045	244.995	263.761
300	0.6041	244.799	263.514
400	0.6039	244.707	263.064
500	0.6042	244.656	262.834
600	0.6052	244.73	262.946
700	0.604	244.64	263.102
800	0.6044	244.724	263.202
900	0.6048	244.83	263.376
1000	0.6034	244.742	263.181

The table displays the results of the 10-fold cross-validation for the Multi-Output Random Forest. The model is run over the following values for the number of trees: 10, 50, 100, 200, 300, 400, 500, 600, 700, 800, 900 and 1000. The metric to evaluate the classification task (prediction of the deductible type) is accuracy, and the root mean squared error is used for both regression tasks (premium prediction and deductible value). The second column shows the accuracy; the third and forth columns present the RMSE for the deductible value and the premium, respectively.