ABSTRACT

Title of dissertation: PARAMETRIC AND NON-PARAMETRIC APPROACHES FOR THE PREDICTION OF THE DIFFUSION OF THE ELECTRIC VEHICLE

Javier Bas Vicente

Dissertation directed by: Prof. Cinzia Cirillo
Dept. of Civil
& Environmental Engineering
University of Maryland

Prof. José Luis Zofío Prieto
Dept. of Economic Theory
Universidad Autónoma de Madrid

Driven by environmental awareness and new regulations for fuel efficiency, electric vehicles (EVs) have significantly evolved in the last decade, yet their market share is still much lower than expected. In addition to understanding the reasons for this slow market penetration, it is crucial to have appropriate tools to correctly predict the diffusion of this innovative product. Recent works in forecasting the EV market combine substitution and diffusion models, where discrete choice specifications are used to address the former, and Bass-type to account for the latter.

However, these methodologies are not dynamic and do not consider the fact that innovation occurs through social channels among members of a social system.

This research presents two advanced methodologies that make use of real data to evaluate the adoption of the EVs in the State of Maryland. The first consists of a disaggregated substitution model that considers social influence and social conformity, which is then embedded in a diffusion model to predict electric vehicle sales. The second, in contrast, relies on non-parametric machine learning techniques for the classification of potential EV purchasers. Both make use of data collected through a stated choice experiment specifically designed to capture the inclination of users towards EVs.

PARAMETRIC AND NON-PARAMETRIC APPROACHES FOR THE PREDICTION OF THE DIFFUSION OF THE ELECTRIC VEHICLE

by

Javier Bas Vicente

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2020

Advisory Committee:
Professor Cinzia Cirillo, Chair/Co-advisor
Professor Jose Luis Zofio Prieto, Co-advisor
Professor Paul Schonfeld
Professor Vanessa Frias-Martinez
Professor Eric Battistin, Dean's Representative

Dedication

To my dearest wife, Vanesa, and my
beloved children, Valeria and Diego,
who give meaning to my life.

Acknowledgments

I owe my gratitude to all the people who have made this thesis possible and because of whom my graduate experience has been one that I will remember forever. I would like to thank my co-advisors Drs. Cinzia Cirillo and Jose Luis Zofio for their support and guidance. I would like to extend this gratitude to Dr. Elisabetta Cherchi, who has almost become an advisor for me, and who I admire and respect.

I would like to dedicate special thanks to Greg Birmingham and his family, who have become my own American-Thai family over the years. I would like to thank as well Alexi Sánchez de Boado and Veronica Duque, whose generous friendship I will always treasure.

Finally, I would be remiss if I did not recognize the coffees, trips and interesting conversations brought by people who crossed my path at some point over the years. Darshan, Santiago, Nicholas, and Riccardo, you made the journey easier and even more valuable.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ANN	Artificial Neural Networks
AVC	Asymptotic Variance-Covariance
AQ	Acquaintances
BEV	Battery Electric Vehicles
DCM	Discrete Choice Models
DT	Decission Tree
CR	Close Relatives
EV	Electric Vehicle
FR	Friends
GHG	Green House Gas Emissions
ICV	Internal Combustion Vehicle
KNN	K-Nearest Neighbors
ML	Machine Learning
MNL	Multinomial Logit
MML	Multinomial Mixed Logit
NCR	Non-Close Relatives
PHEV	Plug-in Hybrid Electric Vehicle
RF	Random Forest
RT	Random Tree
R&D	Research and Development
SC	Stated Choice

SoC	Social Conformity
SVM	Support Vector Machine
YoY	Year over Year
ZEV	Zero Emission Vehicle

1. Introduction

The number of cities that have implemented traffic restrictions due to pollution is high. Moreover, in many cases, these policies are accompanied by other pro-EV measures. It is a common practice in many cities of the U.S. and Europe to allow EVs to park in regulated areas without paying for their cost, to drive on High Occupancy Lanes, or to access cities when other more polluting vehicles are not allowed to. These incentives for the use of EVs are accompanied by an interest that emanates from the demand. Users, increasingly concerned about the environment, are more inclined to adopt this technology. The prospect of savings in the medium term presents also an economic incentive, given the current costs of energy. This attention has sparked, in turn, the interest of the industry, which is consistently working on the development of new generations of more efficient high-performance vehicles. In the last decade EVs have evolved significantly, progressively closing the usability gap. They currently are a realistic option for many users, and have opened new service-based business opportunities. Finally, the public administration is making its own contribution to the state of the matter, too, offering stimulating tax deductions when purchasing a vehicle of these characteristics.

These factors, altogether, place the market in a situation that, although unique and interesting, is fraught with uncertainty. Consequently, the need for reliable information about the future of the EV is greater than ever. A solid prediction might be the basis for industry strategic decisions, as well as for public administration to regulate in favor of EVs. However, all attempts in this regard have been unsuccessful. The forecasts published so far, in academia or other spheres, have substantially differed from the actual evolution that has taken place thereafter; falling short in some cases and being too optimistic in others. There are different reasons for this, and the following subsections shed some light on these aspects. Namely, regulation, Demand and Supply conditioning factors, and state-of-the-art of EV market share forecasts. The aim is to provide a more comprehensive view of the key elements in the development of the EV.

1.1 Rules and regulation

In 2018, 38 states in the U.S. made 102 modifications to the rules and regulations governing EVs. They are shown in Figure 1.1. Most of them were related to fees, public utilities, and electric bicycles.

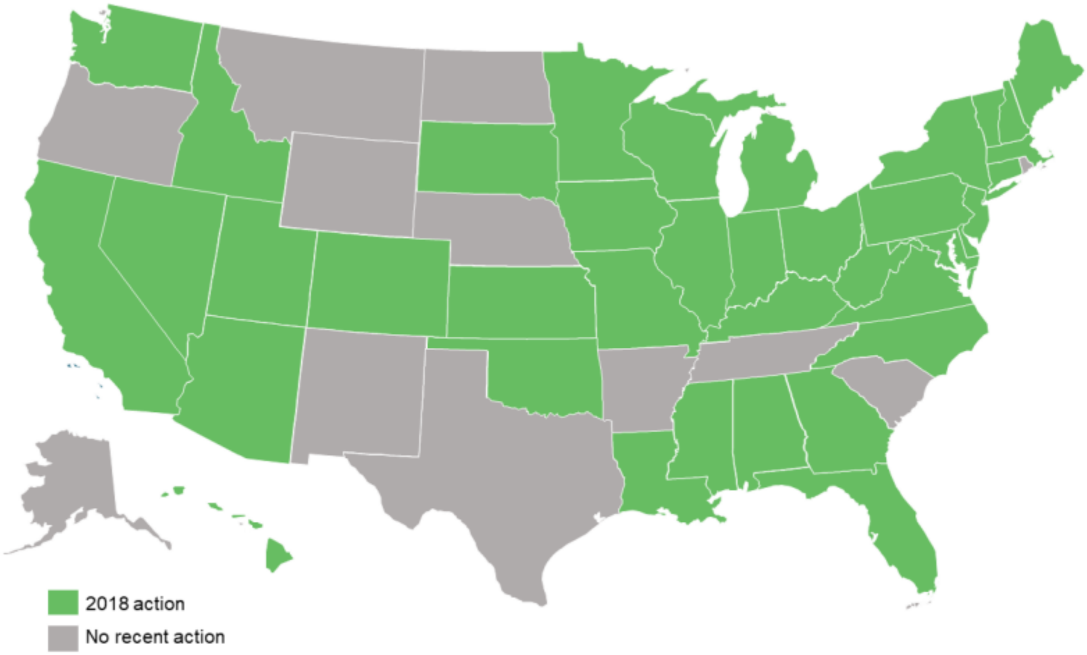
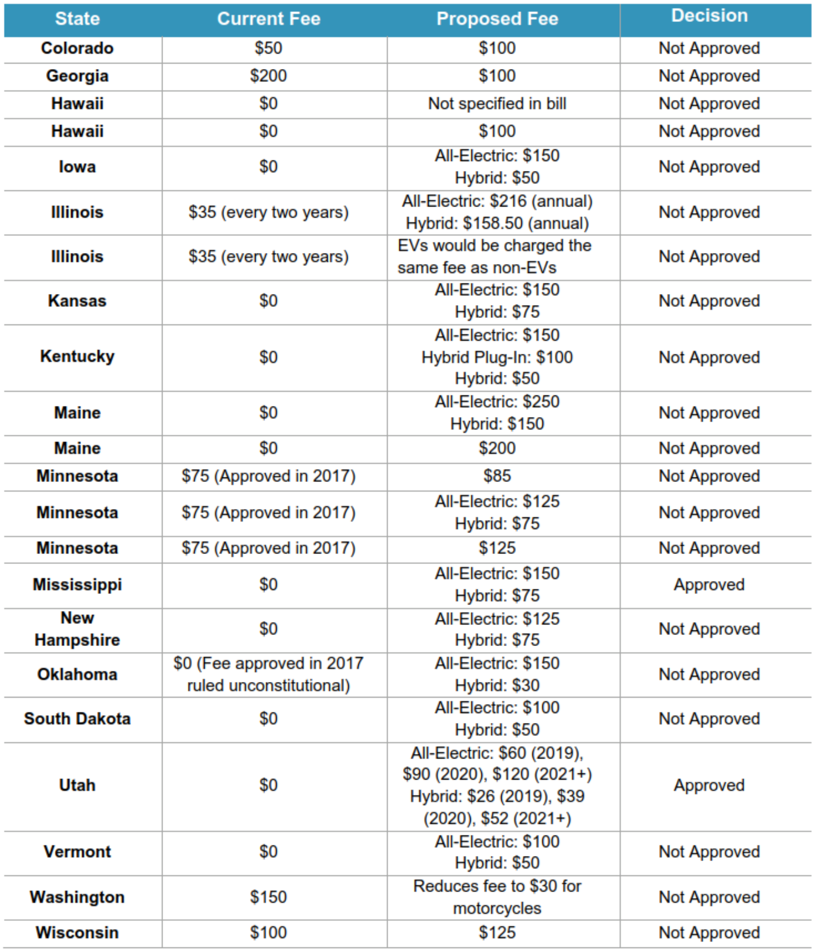


Figure 1.1: States that regulated EVs in 2018. Source: [17].

Fuel taxes are considered as the major and most important source of infrastructure funding in the U.S. As such, the wide use of EVs may result in a reduction of fuel tax revenue, and increase the infrastructure funding shortages. Many states, as well as the federal Highway Trust Fund, have already experienced funding issues due to the improvements in fuel efficiency and to the fact that most states do not adjust their fuel taxes based on inflation. Therefore, policymakers are considering additional fees to be placed on EVs, to make up parts of these shortages, or at least reduce their impact. In this respect, some states have proposed a flat annual fee, while others advocate lower fees for hybrid vehicles ([17]). Table 1.1 summarizes the registration fees in some states for EVs.

Table 1.1: State actions on EV regulation, 2018. Source: [17].

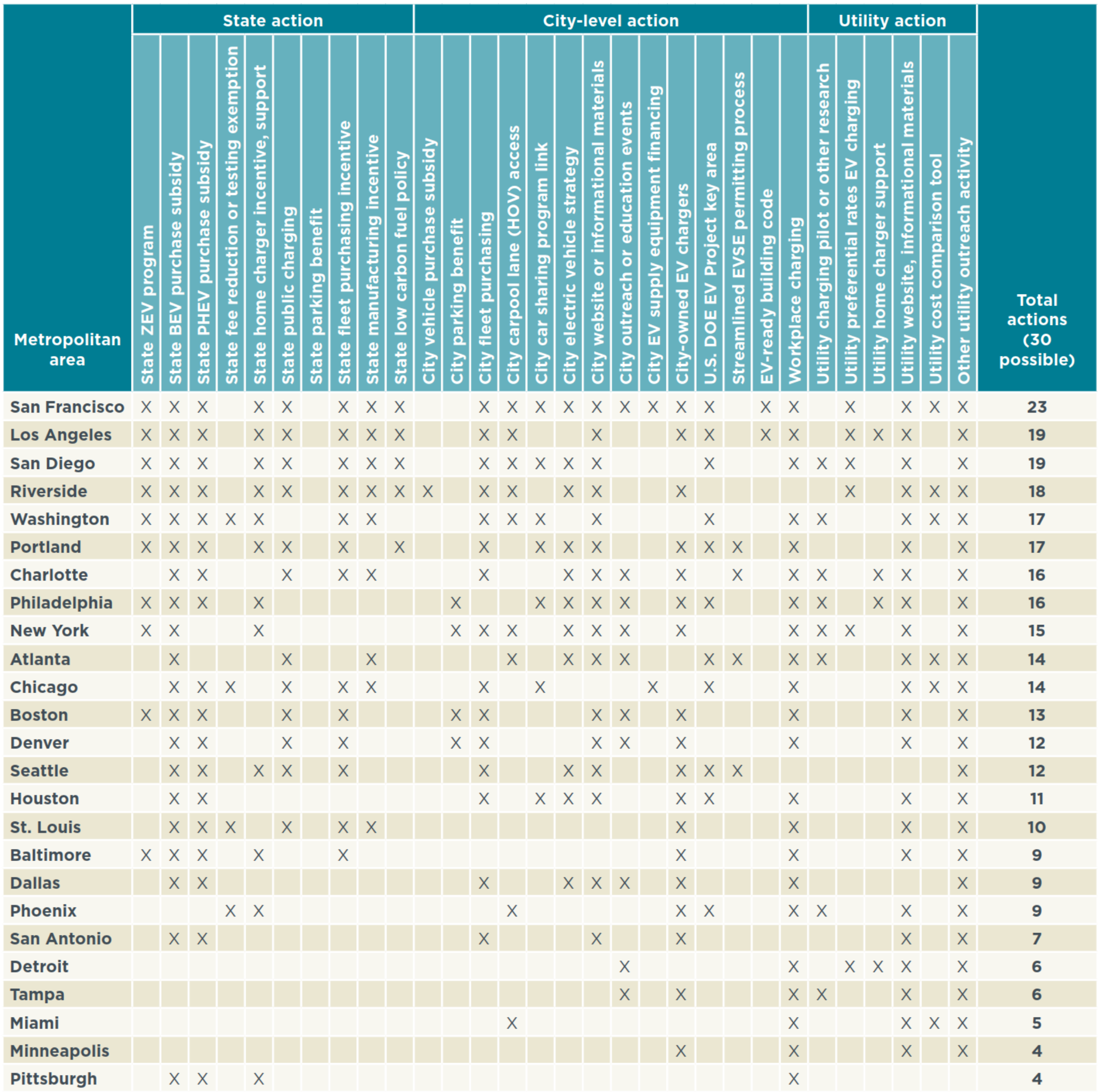


State	Current Fee	Proposed Fee	Decision
Colorado	$50	$100	Not Approved
Georgia	$200	$100	Not Approved
Hawaii	$0	Not specified in bill	Not Approved
Hawaii	$0	$100	Not Approved
Iowa	$0	All-Electric: $150 Hybrid: $50	Not Approved
Illinois	$35 (every two years)	All-Electric: $216 (annual) Hybrid: $158.50 (annual)	Not Approved
Illinois	$35 (every two years)	EVs would be charged the same fee as non-EVs	Not Approved
Kansas	$0	All-Electric: $150 Hybrid: $75	Not Approved
Kentucky	$0	All-Electric: $150 Hybrid Plug-In: $100 Hybrid: $50	Not Approved
Maine	$0	All-Electric: $250 Hybrid: $150	Not Approved
Maine	$0	$200	Not Approved
Minnesota	$75 (Approved in 2017)	$85	Not Approved
Minnesota	$75 (Approved in 2017)	All-Electric: $125 Hybrid: $75	Not Approved
Minnesota	$75 (Approved in 2017)	$125	Not Approved
Mississippi	$0	All-Electric: $150 Hybrid: $75	Approved
New Hampshire	$0	All-Electric: $125 Hybrid: $75	Not Approved
Oklahoma	$0 (Fee approved in 2017 ruled unconstitutional)	All-Electric: $150 Hybrid: $30	Not Approved
South Dakota	$0	All-Electric: $100 Hybrid: $50	Not Approved
Utah	$0	All-Electric: $60 (2019), $90 (2020), $120 (2021+) Hybrid: $26 (2019), $39 (2020), $52 (2021+)	Approved
Vermont	$0	All-Electric: $100 Hybrid: $50	Not Approved
Washington	$150	Reduces fee to $30 for motorcycles	Not Approved
Wisconsin	$100	$125	Not Approved

There are other important questions related to EVs that are a source of regulation. One is to address whether EV charging stations are considered public utilities and whether they can resell the electricity. Equally important are; the ability of homeowner associations to restrict charging infrastructure; penalties for non-EVs parking in dedicated charging spaces; or regulations concerning electric bicycles. In this sense, California is a particularly interesting case, as it is one of the states more concerned about Green House Gas (GHG) emissions. The state has passed several bills and implemented programs to reduce GHG emissions, such as the Alternative and Renewable Fuel and Vehicle Technology Program, the Cap-and-Trade Program, and the California Global Warming Solution Act. This has turned California into one of the most predominant advocates of EVs in the U.S. In 2010, the state required all car manufactures to sell an increasing share of Zero Emission Vehicles (ZEVs). In 2013, the state allowed EVs to access carpool lanes, a measure that came with a federal income tax deduction of between $2,500 and $7,500, depending on the vehicle's weight. The EV Everywhere Grand Challenge of 2012 focused on cutting the battery cost, drive system cost and vehicle weight, and provided funding to increase the charging infrastructure [5].

Table 1.2 summarizes the actions taken in different cities in the U.S. at state and city level, as well as utility actions. San Francisco is the leading city in promoting the EV, with 23 actions, followed by Los Angeles and San Diego. Most of the cities seem to be taking actions on *utility outreach activities* and *workplace charging activities*. *Fleet purchasing*, and *website development* are among the most important actions taken at the city level. State level actions are mostly focused towards Battery Electrict Vehicle (BEV) and Plug-in Hybrid Electric Vehicle (PHEV) purchase subsidies.

Table 1.2: Summary of EV promotion actions across major cities. Source: [92].



Metropolitan area	State action										City-level action														Utility action						Total actions (30 possible)
	State ZEV program	State BEV purchase subsidy	State PHEV purchase subsidy	State fee reduction or testing exemption	State home charger incentive, support	State public charging	State parking benefit	State fleet purchasing incentive	State manufacturing incentive	State low carbon fuel policy	City vehicle purchase subsidy	City parking benefit	City fleet purchasing	City carpool lane (HOV) access	City car sharing program link	City electric vehicle strategy	City website or informational materials	City outreach or education events	City EV supply equipment financing	City-owned EV chargers	U.S. DOE EV Project key area	Streamlined EVSE permitting process	EV-ready building code	Workplace charging	Utility charging pilot or other research	Utility preferential rates EV charging	Utility home charger support	Utility website, informational materials	Utility cost comparison tool	Other utility outreach activity	
San Francisco	X	X	X		X	X		X	X	X			X	X	X	X	X	X	X	X	X		X	X		X		X	X	X	23
Los Angeles	X	X	X		X	X		X	X	X			X	X			X			X	X		X	X		X	X	X		X	19
San Diego	X	X	X		X	X		X	X	X			X	X	X	X	X				X			X	X	X		X		X	19
Riverside	X	X	X		X	X		X	X	X	X		X	X		X	X			X						X		X	X	X	18
Washington	X	X	X	X	X			X	X				X	X	X		X				X			X	X			X	X	X	17
Portland	X	X	X		X	X		X		X			X		X	X	X			X	X	X		X				X		X	17
Charlotte		X	X			X		X	X				X			X	X	X		X		X		X	X		X	X		X	16
Philadelphia	X	X	X		X							X			X	X	X	X		X	X			X	X		X	X		X	16
New York	X	X			X							X	X	X		X	X	X		X				X	X	X		X		X	15
Atlanta		X				X			X					X		X	X	X			X	X		X	X			X	X	X	14
Chicago		X	X	X		X		X	X				X		X				X		X			X				X	X	X	14
Boston	X	X	X			X		X				X	X				X	X		X				X				X		X	13
Denver		X	X			X		X				X	X				X	X		X				X				X		X	12
Seattle		X	X		X	X		X					X			X	X			X	X	X								X	12
Houston		X	X										X		X	X	X			X	X			X				X		X	11
St. Louis		X	X	X		X		X	X											X				X				X		X	10
Baltimore	X	X	X		X			X												X				X				X		X	9
Dallas		X	X										X			X	X	X		X				X						X	9
Phoenix				X	X									X						X	X			X	X			X		X	9
San Antonio		X	X										X				X			X								X		X	7
Detroit																		X						X		X	X	X		X	6
Tampa																		X		X				X	X			X		X	6
Miami														X										X				X	X	X	5
Minneapolis																				X				X				X		X	4
Pittsburgh		X	X		X																			X							4

"X" denotes given electric deployment action is in place in the metropolitan area in 2014
ZEV = Zero Emission Vehicle; BEV = Battery electric vehicle; PHEV = Plug-in hybrid electric vehicle; HOV = high-occupancy vehicle lane

From an international perspective, France has taken notable steps towards the use of EVs. It has placed incentives to increase the number of charging stations up to seven million in 2030, as well as establishing the requirement that half of the government vehicles be low emission [5]. Another notable policy implemented, in a similar line to the U.S., is a rebate scheme on vehicle registration. The registration cost ranges from €150 to €8,000, and the rebate ranges from €150 to €6,300. The reduction is higher for BEVs and lower for PHEVs. The annual vehicle ownership tax is based on CO2 emissions, which exempts EVs from the annual vehicle registration taxes. Finally, if the owners of an EV scrap diesel cars registered before 2001 they can get up to €10,000 benefit per car [5].

Germany has also promoted the use of EVs through the National Electromobility Development Plan passed in 2009 [5]. EV owners are exempted from annual vehicle registration taxes for 5 to 10 years (depending on the license date) in addition to receiving a 50% purchase tax deduction, as well as other incentives (€2,000 for full EVs and €1,500 for hybrid). Moreover, the electricity that is used for public transport is subsidized, and the operators of charging points are treated as final consumers and are not subject to energy suppliers tax obligations. Finally, municipalities across Germany are authorized to provide parking and bus lane privileges to EVs.

Figure 1.2 plots, for several European countries and regions, fiscal incentive as a percentage of the cost of a comparable gasoline car and the number of charging points per 1,000 registered cars. Norway and its two cities (purple circles) have the highest fiscal incentive, and accumulate a 53.7% of the market share of all new vehicles in 2014. In addition, Oslo and Bergen provide also the highest density of charging points, followed by Amsterdam.

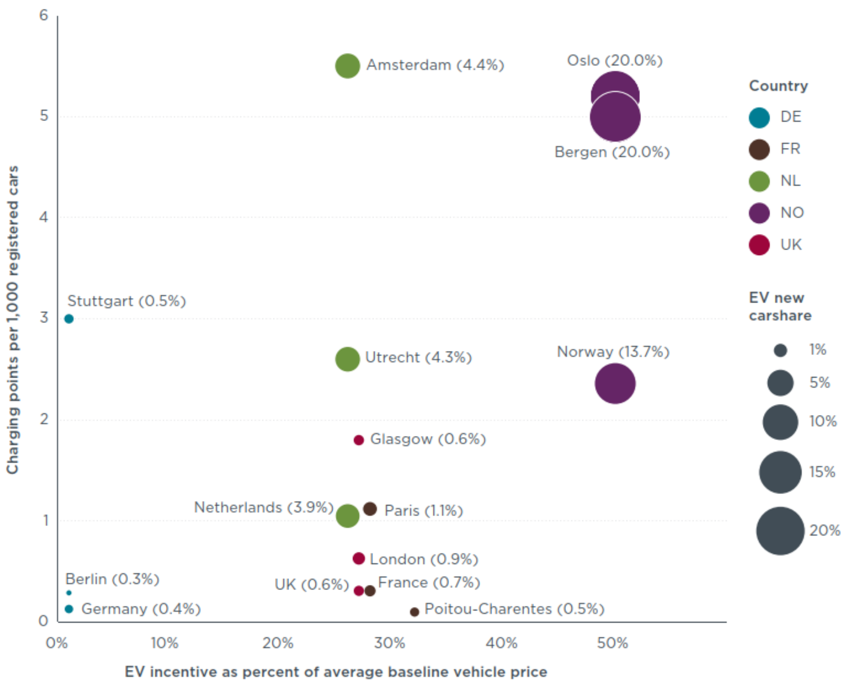
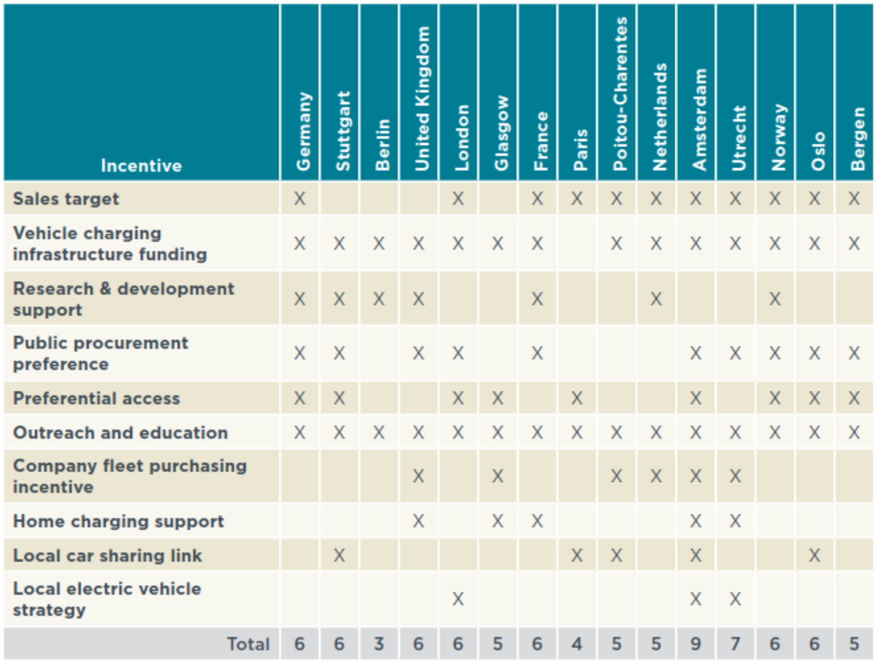


Figure 1.2: Fiscal incentives, market share, and charging point density for European countries and regions. Source: [?]

Similarly, Table 1.3 shows non-fiscal incentives, such as charging infrastructure funding, research support, car sharing link, or local EV strategy definition. Amsterdam has taken the highest number of these actions among the regions surveyed, followed by Utrecht; both cities of the Netherlands. *Outreach and educational programs* and *Vehicle charging infrastructure funding* were among the most widely used strategies to promote EVs.

Table 1.3: Comparison of non-fiscal incentives for European countries and regions. Source: [135].



Incentive	Germany	Stuttgart	Berlin	United Kingdom	London	Glasgow	France	Paris	Poitou-Charentes	Netherlands	Amsterdam	Utrecht	Norway	Oslo	Bergen
Sales target	X				X		X	X	X	X	X	X	X	X	X
Vehicle charging infrastructure funding	X	X	X	X	X	X	X		X	X	X	X	X	X	X
Research & development support	X	X	X	X			X			X			X		
Public procurement preference	X	X		X	X		X				X	X	X	X	X
Preferential access	X	X			X	X		X			X		X	X	X
Outreach and education	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Company fleet purchasing incentive				X		X			X	X	X	X			
Home charging support				X		X	X				X	X			
Local car sharing link		X						X	X		X			X	
Local electric vehicle strategy					X						X	X			
Total	6	6	3	6	6	5	6	4	5	5	9	7	6	6	5

In summary, as far as public regulation is concerned, it seems that focusing on vehicle emissions is not enough to encourage users to purchase EVs. Policy in this spirit should be coupled with other financial and operational incentives, of which the most important ones appear to be:

- Fiscal incentives to reduce initial vehicle purchase, registration cost, and ownership costs.

- Engagement with electricity utilities.

- Deployment of public and workplace charging networks.

- Information and awareness actions.

- Operational incentives such as toll and parking cost exemption, and use of bus and carpool lanes.
- Implementation of tighter CO2 emission standards.

1.2 Demand and Supply

China has the largest EV market, followed by Europe and the U.S. Nearly one million EVs were sold in China in 2018, which was slightly less than the total sales in the U.S. and Europe combined. Figure 1.3 shows the evolution of sales in different countries. In terms of inventory, China again tops the list, with two million. Interestingly, half of this inventory was sold in 2018, which was actually doubled with respect to 2017. However, putting these figures in relation to population change this picture. In that case, Norway leads the ranking with 55.9 EVs per one thousand population, followed by the Netherlands, Sweden, and with 8.7, 7.8, and 3.4, respectively. In contrast, China has 1.6 EVs per one thousand population. In global terms, there were 5,127,297 EVs with a vehicle-to-thousand population ratio of 0.7 in 2018.

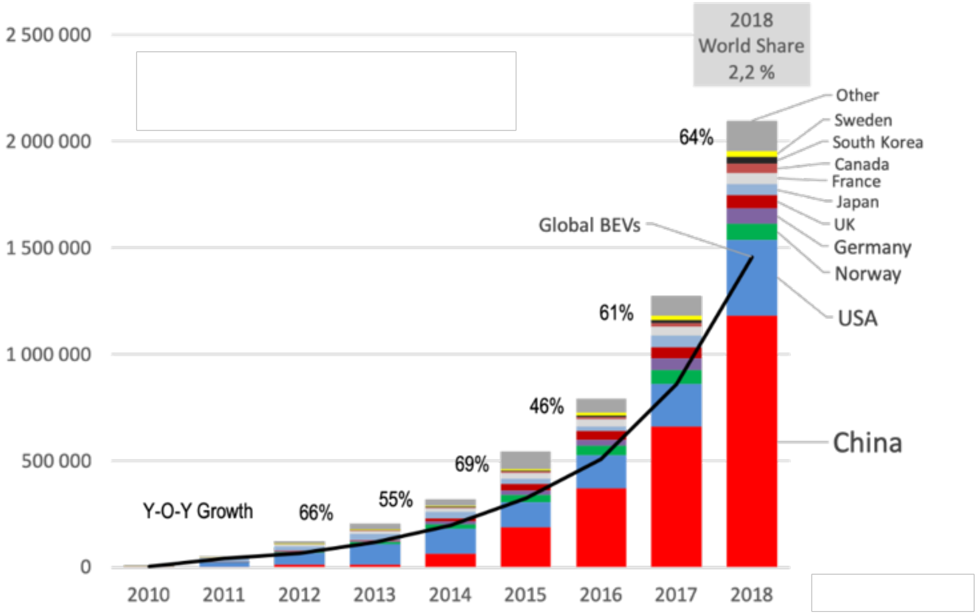


Figure 1.3: Global trend in EV sales by country, 2010-2018. Source [60].

In the U.S., the east coast states had the highest share of EVs in 2018 due to the incentives that they offered, as previously mentioned. Nationwide, nearly 360,800 PHEVs were sold in 2018, meaning an 81% growth in the EV market. BEVs had the highest share, with 66% of the total, in contrast to the 53% of 2017. Total electric light vehicle sales reached 2.1% over the entire year, topping at 3% in November and December. Figure 1.4 shows these market trends.

The European EV market has also experienced growth. With more than one million PHEVs reached in 2018, the Year over Year (YoY) growth is nearly 42% [32]. In the first half of 2017 and 2018, Norway had the highest number of EV sales in Europe, closely followed by Germany and the UK [6]. The sales in some European countries nearly doubled in the first half of 2018 – i.e. Denmark, Finland, Portugal, Netherlands, and Spain. This growth continued in the first month of 2019, in which

EV sales jumped 67% YoY.

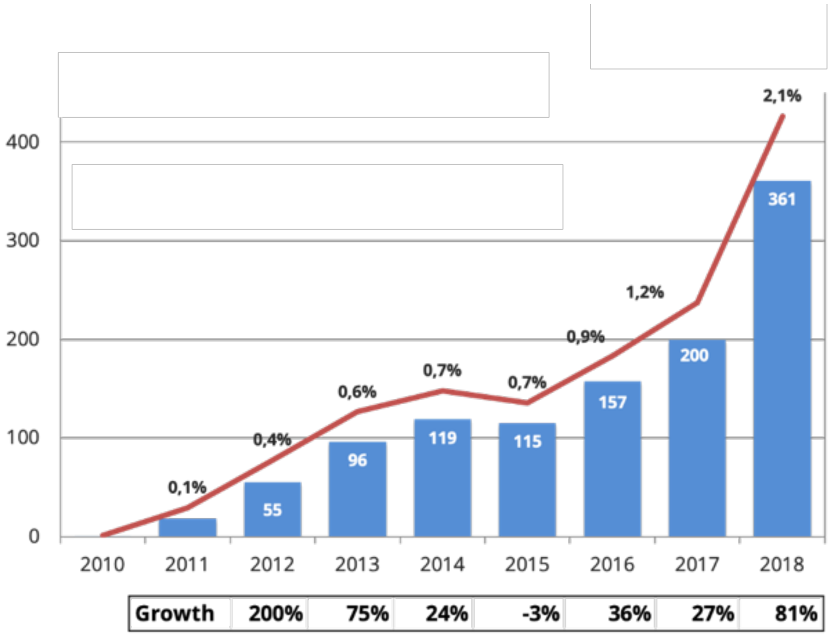


Figure 1.4: EV sales trend in the U.S. Source [61].

Backed by this growing attention on the part of users and public institutions, the interest of the automobile industry in promoting the EV technology has increased likewise, especially recently. The environmental concerns of both public administrations and users discussed above are probably the main driver for manufacturers, which are doing important investments in improving the models that they already have in their catalogs, as well as extending the number of them, with the aim of reaching a wider range of consumer profiles. Nevertheless, although the industry is willing to introduce EVs in the market, this objective depends on the combination of many factors to become effective. For instance, the sales strategy and the availability of the models have not always been favorable. Especially in the early stages of the introduction of EVs, car dealers lacked the experience needed to provide adequate and convincing information about the electric models. In addition, there were frequently no test vehicles of this class at the dealership. That caused important misinformation about costs and benefits, as well as about other relevant factors to consider when making the purchasing decision. Advertising has not always been used optimally as well. It has been observed ([35]) that the publicity of EVs was significantly lower than to conventional vehicles. Therefore, there has been a space to fill with respect to the strategies deployed to convince potential users, beyond the mere improvement of the EV characteristics.

In regard to these attributes, the superior price of EVs in comparison to gasoline vehicles still is, undoubtedly, an important factor that is limiting its diffusion, even though this gap is closing. Figure 1.5 shows the high prices of representative EV models with respect to those of internal combustion. One of the reasons behind elevated prices is the significant cost of battery production, as well as the R&D associated with it. The current generation of batteries has low energy density and provides limited driving range; meaning long charging times (even in fast charging mode) to achieve relatively short driving autonomy. Although numerous studies refute the feared *range anxiety*, equating the ranges of electric and gasoline vehicles is a capital objective for manufacturers.

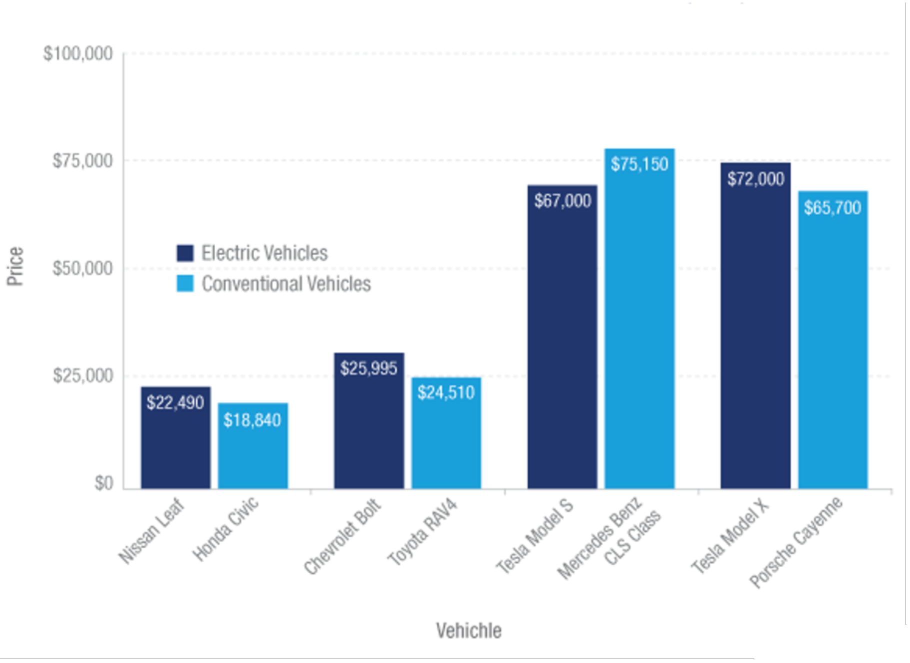


Figure 1.5: Vehicle prices for representative models in the U.S., electric vs. conventional. Year 2016. Source [97].

Despite these aspects, there has been a steady increase in the number of models available for sale in the U.S during the last years, as Figure 1.6 shows, giving users the possibility to choose according to their tastes and preferences.

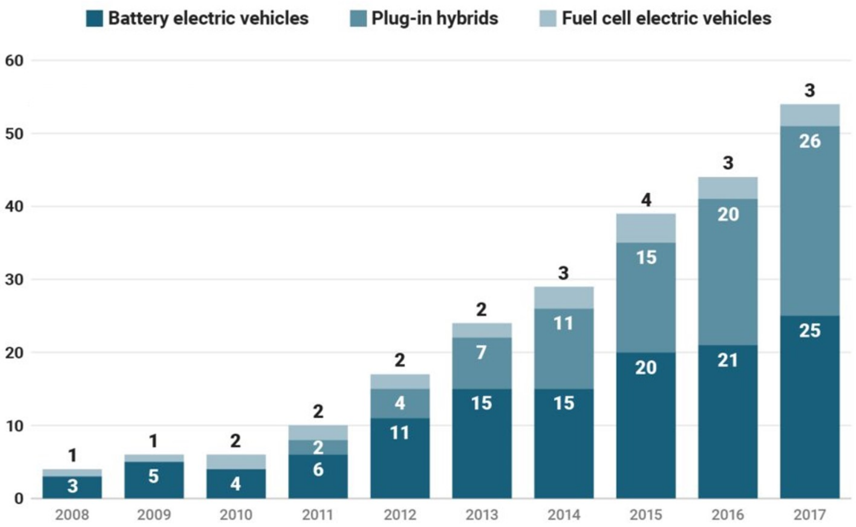


Figure 1.6: Number of EV models in the U.S. Source [130].

Regarding the evolution of the automotive industry, although producers have been experimenting with prototypes over some time, scale production began at the end of 90s. The Nissan Leaf was the first model to be available in the U.S, in 2010, followed by models by Ford, Toyota, and Honda in 2011 and 2012. Nissan sold 8,720 units of the Leaf in its first 11 months ([91]). In comparison, the Tesla Model S, a luxury BEV introduced in 2012 (but rescheduled to 2013), reached around 20,000 units sold in its first year. Chevy Volt is another emblematic EV of the American market. Until August 2012, 13,479 Volts were sold ([142]), and it is currently still a popular model. It passed the Consumer Reports Owner Satisfaction Survey for 2011 and 2012, with 92% of owners saying they would make the same purchase again. Figure 1.7 presents the market shares of the above models, and others. Among the models that dominate the European market are the Renault Zoe, Mitsubishi Outlander and Nissan Lift; only the Lift has a significant market in the U.S.

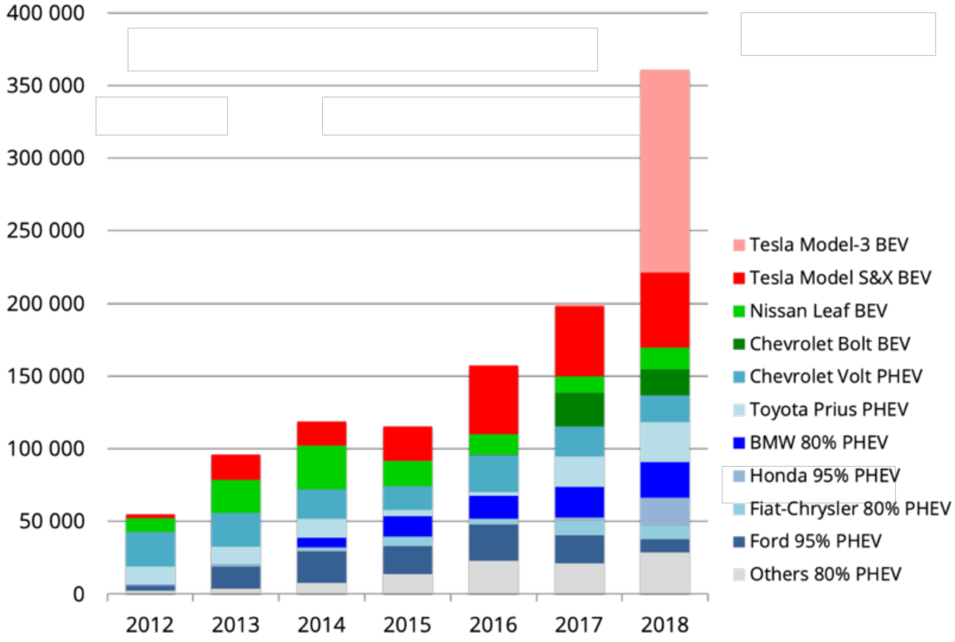


Figure 1.7: Market shares for different makes and models in the U.S. Source [61]

PHEVs do not have the same mileage range limitations as fully battery-powered vehicles although, in some cases, they tend to be more expensive than battery and gasoline cars since they incorporate both kind of engines. Non-pluggable hybrids have been mass-produced in the USA since 2000, with the Honda Insight being the first hybrid available in the United States. Since then most producers have hybrid models in their catalogues. The best-selling hybrid currently on the road is the well-known Toyota Prius, which sold almost one million units between 2000 and 2010 ([141]). Figure 1.8 shows the percentage of hybrid retail registrations by state. California, Vermont, the District of Columbia, Oregon, Arizona, and Washington have the highest penetration in the hybrid market and are also leading markets for the initial deployment of PHEVs ([20]).

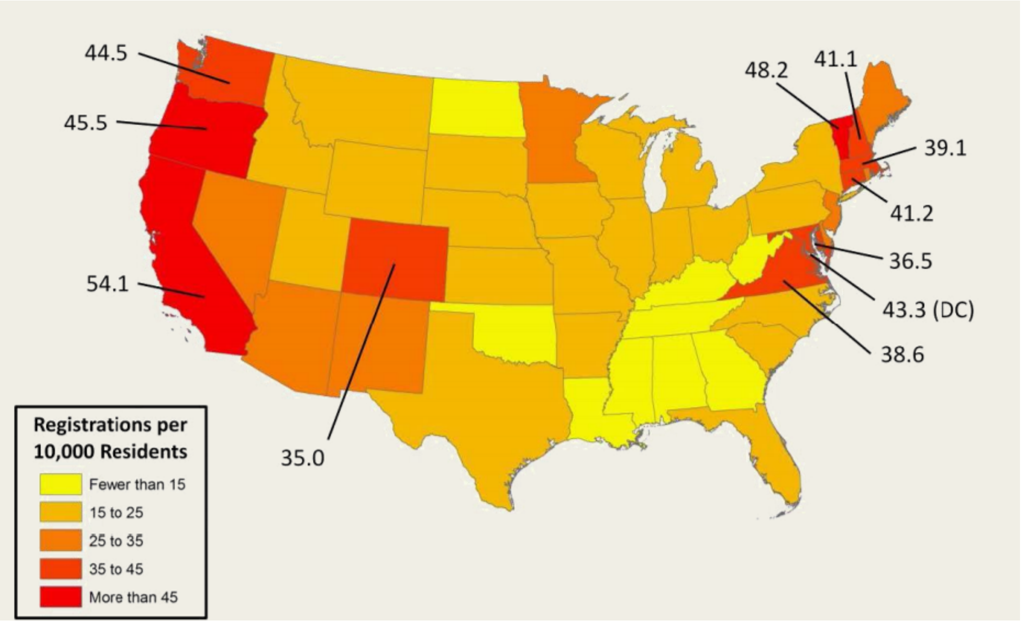


Figure 1.8: Percentage of hybrid retail registrations by state. Source [20].

The performance of the hybrid market can give an idea of how the BEV one will mature. ([103]) foresee that by 2017, PHEVs will account for 1.2% of car sales in the U.S. However, although the PHEV business is more developed and may share elements with the BEV, it is difficult to accurately predict how quickly the latter will expand. In any case, the variety of models described has been accompanied by a clear trend of substantial growing sales, as Figure 1.9 shows, encouraging the hope of the definitive arrival of BEVs, and arising the old question of whether Demand conditions Supply or vice versa.

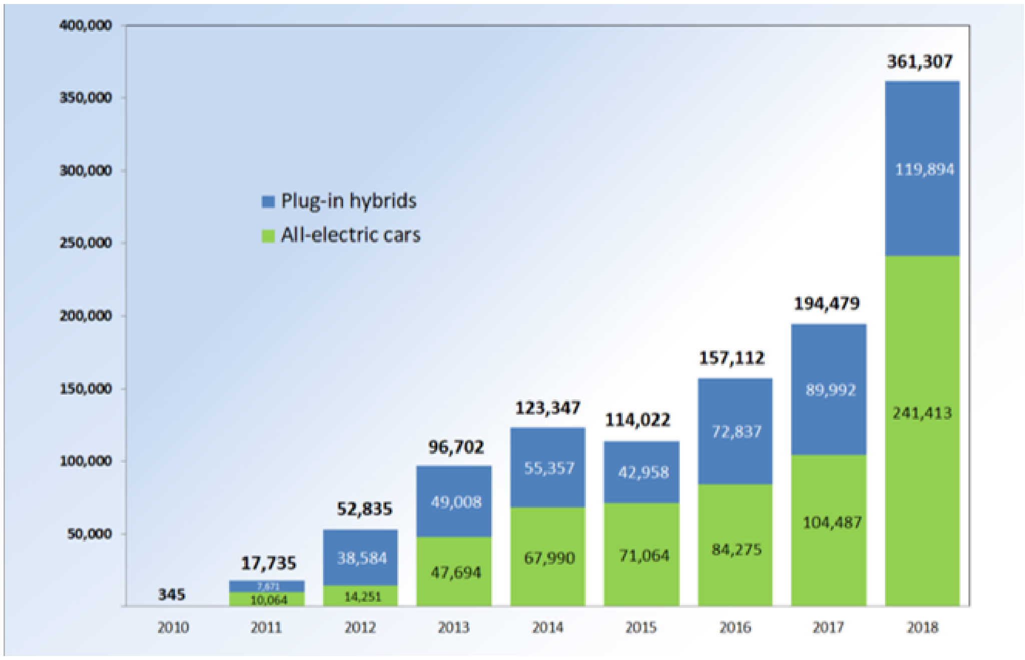


Figure 1.9: Annual sales of EVs in the U.S. 2010 - 2018. Source: [40].

1.3 State of the art of EV market share forecasts

1.3.1 Substitution and diffusion models

The factors described above, altogether, place the market in a situation that, although unique and interesting, is fraught with uncertainty. The need for reliable information about the future of the EV is greater than ever. A solid prediction might be the basis for industry strategic decisions as well as for public administrations to regulate in favor of EVs. However, all attempts in this regard have been unsuccessful. The forecasts published so far —in academia or other spheres— have differed substantially from the actual evolution that has taken place thereafter; falling short in some cases and being too optimistic in others. There are different reasons for this, some of them methodological and others due to the approach to the subject adopted.

For instance, Discrete Choice Models (DCMs) are a popular tool for predicting demand. However, DCMs rely on the responses provided in hypothetical scenarios, i.e. Stated Preferences (SP) data. Thus, when used for prediction, the alternative specific constants of these models need to be calibrated to reflect the unobserved factors present in real market situations. These disaggregated demand models also present a limitation of special importance in the case of innovative products; they are suitable for predicting the demand for stable markets but are inadequate for disruptive goods whose sales behave peculiarly.

Not considering the evolution of demand over time is another drawback of the predictions performed so far. Innovations spread at a slow pace and need long periods to reach significant market shares. Sales growth curves barely increase in the introduction period, to exponentially grow once critical mass is reached. This behavior cannot be represented by demand models based on DCMs, and new approaches are required. One of them is the use of classic diffusion models, such as Bass ([6]), Gompertz ([49]), or extensions of these two, in order to incorporate different aspects such as competitive products or different generations of the good at hand.

Another reason that may explain why attempts to predict the spread of EVs have failed is the absence of social elements in the models. Although diffusion is a process that occurs through social channels [112], the methodologies that explicitly account for them are reduced.

Hence, for a prediction to be accurate and reliable, it should include three elements:

- Substitution: Which may happen by replacing an Internal Combustion Vehicle (ICV) – gasoline or any other engine – by an EV, by purchasing an additional EV for the household, by upgrading to a new generation of EV, or even by selling an EV back. In any case, there is clearly a substitution component in the diffusion of EVs since it is a good that individuals consume in order to substitute another mean of transportation, especially a car with another type of engine.

- Diffusion: The approach in the classic Bass models and its extensions is that a new technology is first adopted by a subgroup of *innovators*, who are followed by *imitators*. This process simply happens through a progressive increase in sales; i.e. in a classic Bass model, the accumulated sales increase, which, in turn, contribute to increasing the future sales. However, it is known that this process is eminently social, as Rogers points out ([112]). Therefore, it is absolutely necessary to include social elements in any prediction of diffusion of technology.

- Dynamics: Diffusion not only depends on the context at a particular moment but also on what occurred in the previous periods. Models that aim to predict sales should include variables referring to previous periods of time.

The second of these elements is of special interest for this research . Roger's assertion that diffusion occurs through social channels ([112]) can be understood from a double perspective. On the one hand, people around us influence our behavior. Whether they are family, friends, or even unrelated people, their conduct and their own decision-making condition ours. Somehow, all of us, give in to peer pressure to some extent, and act in accordance with the position of the majority. This phenomenon is known as Social Conformity (SoC) and has been extensively studied, as will be detailed in the next chapter. SoC is a type of social influence, which involves changes in one's attitudes, beliefs and behavior in order to fit into a group ([30]). There are also several categories of social influence; one is *descriptive* normative beliefs, which refer to what an individual thinks others do in a particular situation; another is *injunctive* normative beliefs, which reflect beliefs about what others approve or disapprove of. On the other hand, attitudes and perceptions also play a role in diffusion since they affect individual behavior, too. Research conducted in psychology suggests that preferences are formed by experience. Therefore, either having direct experience with EVs or receiving feedback from significant ones on their experience with EVs may influence preferences and, consequently, impact market penetration. Another interesting reflection is the influence, not of the social elements themselves, but of the size of the social network from which they proceed. Relevant questions are whether the size of an individual's network is important, or whether the nature of their connections matters more. This research aims to address these questions. In order to do so, the approach developed parts from the intuitive idea that the persons who form the social network of an individual pertain to two groups:

- **Inner circle**: People trusted and/or with which identification is felt. It may be formed by family and friends.

- **Outer circle**: People with whom one has direct contact, but who are not especially relevant in terms of affection or trust, such as non-close friends or acquaintances.

The number of members of the inner circle may change little since some of them are 'replaced' by others as time goes by. Family members die while others grow up and 'occupy' their place in the network in terms of confidence and influence. Likewise, close friends sometimes stop being so or move far away, while one makes one or two of them with the same probability.

Considering this analysis, it can be concluded that there is a range of improvement in the development of substitution/diffusion models to predict the growth of EV sales, which will allow for a better understanding of the full impact in the economy and in the transportation sector, in particular. An interesting research question is the impact of EVs in macroeconomic key indicators. Some studies ([7]) foresee improvements in the trade balance, positive net creation of jobs if a domestic manufacturing industry is developed, significant healthcare cost savings, and substantial reduction of GHG emissions. Although this research is developed at a microeconomic level, it may eventually support studies of macroeconomic magnitude.

1.3.2 Machine Learning methods

The number of studies that have explored EV adoption is large, either taking the agent's perspective, or trying to predict penetration through more macroeconomic approaches. However, although these studies point together in the same direction, they offer very different EV market evolution in terms of time and magnitude. This causes a lack of reliability from which it is difficult to make strategic decisions, either by the industry or the public sector. Therefore, new methodological perspectives are required, Machine Learning (ML) being one of them.

ML techniques are currently applied to an enormous variety of topics such as fraud detection, robotics, spam filtering, translation services, preventive health care, and computer vision, as well as transportation. This has been possible thanks to the exponential growth of information brought about by electronic devices; an amount that will continue to expand due to the Internet of Things. In the case of transportation, the smart use of the data generated by on-road vehicles presents an extraordinary opportunity to improve transportation systems. However, this task overcomes the capabilities of traditional data analysis and clearly points to ML as a solution. Congestion reduction, safety improvement, environmental impact mitigation and energy consumption optimization are examples of the most common lines of research in which ML techniques have been applied. However, there are other less explored fields of application, such as the classification of potential consumers into adopters/non-adopters. This is a topic that presents interesting challenges. Adoption is demand-driven, and Demand roots into purchasers' behavior, beliefs and attitudes; elements that are intrinsically difficult to define and gather. Even if reliable information on these aspects is available, it is unlikely to be in large quantities, so certain methodologies cannot be used as they only perform well in large data sets.

ML algorithms can be categorized into three types; supervised learning, used both in classification and regression tasks; unsupervised learning, used for clustering and dimensionality reduction; and reinforcement learning, based on reward maximization. In supervised learning observations are labelled, i.e. each one has a class assigned, an associated response. The algorithm processes the data on a training subset to generate label predictions that are validated in a different testing subset. The error resulting from this comparison is used to fit the model as much as possible to the data. An advantage of most ML methods is that, unlike ordinary regression models, they are not parametric. They do not rely on assumptions about the relationships between variables present in the model in order to minimize the error (cost function). In contrast, when the cost function becomes more complex, with more parameters and high dimensionality, minimizing it becomes a difficult task. Another difficulty is the enormous diversity of algorithms that can be applied to the same problem. Although there may be some guidelines on which one should be applied to each case, the truth is that different approaches may lead to significant deviations of the level of performance. In addition, each technique has specific hyperparameters with no reliable predefined values. This complexity implies that, when facing a project —either classification or regression, it is first necessary to assess which method may be appropriate for it and then to carry out an iterative process of tuning. This process must also include a resampling procedure to avoid any possible bias when splitting the data into training/testing subsamples. In this case *K-Fold Cross Validation* is employed, which randomly divides the sample into k groups, or folds. The model is estimated on the first fold and validated in the remaining $k - 1$. The process is repeated k times; each time a different subset is treated as the validation one. The final accuracy measure is computed as the average of those obtained in all folds. These procedures are highly compute-intensive, especially as the number of data points and dimensionality grows. Therefore, it is common to conduct a feature selection process, which consists on a pre-filtering of the most important variables in order to reduce the dimensionality.

1.4 Problem statement

The problem statement is summarized by the need for reliable information about the future of the EV, which is greater than ever. However, all attempts in this regard have been unsuccessful. The sales forecasts published so far in the academia or other spheres have differed substantially from the actual evolution that has taken place outer; falling short in some cases and being too optimistic in others. The reasons for this are different; i) lack of suitable DCMs for predicting the demand of a disruptive technology; ii) disregard of the dynamism of the demand; and iii) absence of social elements in the quantification of the diffusion process. Although existing works try to deal with these aspects separately, there is no single model that covers all of them. Whether using parametric or non-parametric specifications, joining all these aspects providing reliable information is the first challenge of this research since a solid prediction might be the basis for industry strategic decisions as well as for public administration to regulate in favor of EVs.

1.5 Research objective

The research that sustains this dissertation primarily deals with finding both parametric and non-parametric models that properly gathers the social components of the diffusion of EV and yields accurate, reliable, sales forecast. These methodologies must be scalable, to be adapted to different scenarios; namely, geographic regions with different population levels or user profiles. They must be also flexible, to provide the possibility to adapt it to more vehicle alternatives or to a natural combination of EVs and services, such as "electric vehicle plus electric plan", or "electric vehicle in a Mobility as a Service context". This may be of the interest of private companies and public administrations in need to implement a solution in which the prediction of the demand is the ultimate output, or an intermediate input in their procedures.

A clarification on the terminology should be made at this point. In the field of transportation engineering, the act of projecting explanatory variables in the future and introducing them into a model to obtain its output is called a *forecast*. However, in economics this term is generally linked only to the analysis of time series and an exercise of the kind developed in this research is denominated a *counterfactual analysis*. Since this dissertation lays in the middle of both areas of knowledge, it has been decided to use the engineering acceptation.

1.6 Contributions

The contributions of this research can be divided in those purely oriented to research and academia, and those oriented to the private and the public sector. For the former, this is the only work (to the best of this researcher's understanding) that combines substitution and diffusion altogether dynamically and accounting for social effects. In addition, one of the main ideas developed, the existence of an *Inner* and an *Outer* social circle is novel, as well as how its particularities have been designed and modelled. This also applies to the non-parametric techniques utilized in this research. There are clear applications of this methodology in the interest of private companies and the public sector. Besides vehicles manufacturers, power companies, for instance, are interested in knowing where the adoption of EVs will take place, in order to launch commercial campaigns to those potential purchasers. On the other hand, municipalities and states are also attentive to the market penetration of EVs in order to plan charging infrastructure.

1.7 Dissertation organization

After the introduction, Section 2 presents a literature review for each of the main elements that composes this research: Social influence, Diffusion-Substitution models, Machine Learning, and Data and Survey Design. Section 3 presents two case studies for the use of Substitution-Diffusion models for market share prediction. The first is an initial approach to social influence in the context of the Danish market for EVs. The second presents a more developed methodology and was carried out in the American market. Section 4 exhibits a comparison of ML techniques for the classification of potential Ev purchasers. Lastly, Section 6 presents the most relevant findings of this research as well as a discussion on the principal topics of this dissertation.

2. Literature review

2.1 Diffusion and social influence

As expressed by Rogers [112], the diffusion process is defined as that by which an innovation is communicated through certain channels over time among members of a social system. The behavioral assumption is that an innovation is first adopted by a small segment of innovators, and then later adopted by an increasing number of customers, the imitators, who are influenced by the number of adoptions that have already occurred. This conception corresponds to what in the psychological literature is defined as Social Conformity (SoC). Friends, family and acquaintances influence our behavior. Even people who we do not know personally may indirectly influence our decisions. [36] pointed out that we tend to give in to group pressures and act according to the majority, and [123] and [1] proved that individuals may be influenced to make a wrong choice when they are insecure and look to others. SoC can occur because one wants to be accepted by the members of a certain group (Normative conformity) or because they want to act like is supposedly right. In the last case, people consult members of their group to obtain information (Informational conformity). Therefore, conformity is a type of social influence involving a change in attitudes, beliefs, and behaviors in order to fit in a group, matching the group's norms and beliefs ([30]).

SoC has been extensively studied in psychology ([31], [72], [27], [28], [29], [30], [112], [48], [118]). However, there are also some applications in other fields, as economics. [4] advocates for considering the incidence of social influence in the decision-making process, and not exclusively the rational approach, since economics is actually closely linked to social and psychological aspects. In this regard, as a social science, economics is naturally influenced by human behavior, in both a small and a large scale. Nowadays, the way to influence other is more direct, considering that people interact and express themselves more and more frequently through social media. Companies wishing to attract consumers devise marketing strategies taking advantage of that use ([131]). For their part, [58] reached a similar conclusion, considering that nowadays any marketing campaign should find the relevance of social networks in social behavior, and thus be able to conceive useful strategies that generate the expected behavior among consumers. [23] were interested in studying what people consider when making small investments in prosocial projects since better understanding of this process would lead to more effective crowdfunding.

Another aspect concerning social influence that also applies to economic matters is the *herd behavior* of masses. This is a phenomenon by which individuals act as part of a group, making decisions that they would not make as an individual. This conduct has been reviewed by [12] who have made a study of the behavior of people in the stock market, where quick decisions are made, sometimes led by what the majority does.

Regarding the field of transportation, [102] conducted an investigation with data obtained through an online survey made to students of the University of California, to test the hypothesis that certain factors condition social influence in the adoption of bicycles. They concluded that, although the social influence of acquaintances exists, its degree of leverage depends on certain factors, such as the characteristics of the daily trip or the journey distance. Other authors have also studied which factors affect the adoption of a bicycle as a means of transportation, such as [125], who sought to determine how social influence affected the decision to ride or not a bike in England. They observed that, among other factors, social influence played a vital role in promoting cycling.

On the other hand, other studies focused on an environmental approach; like the one presented by [33], which investigated how the importance that people give to the environmental impact of using private vehicles can make them switch to public transportation. A similar investigation was carried out by [150] who developed a survey in Shanghai to determine how the awareness about of the problems caused by the excessive use of private transport influences the decision to opt for public transportation. Related to that line of research, the work of [37] conducted a study to analyze the kind of influence exerted by the social network and the strength of the bonds of the person who is deciding to opt for public transportation. Finally, [127] sought to determine the factors that influence people adopting EVs. The author conducted a survey of 173 people of the Netherlands, with emphasis on pro-environmental aspects that are related to this technology.

Focusing on works that relate the social network and the electric vehicle, [47] studied the differences between the choices of an individual and the choice of others, in the context of vehicle access. [80] evaluated the impact of the opinion of others in a person's choice of private vehicle use. [2] used the personal network and the experiences of individuals with a hybrid vehicle to study the effect of interpersonal influence in the adoption of this technology. [140] provided a description of EV buyers and their relationships with EV communities, finding that early drivers used forums to find more information about the EV characteristics. [146] studied how social networks affect the decision to opt for EVs, and considered that governments can adjust their policies to reach more effectively the public that is inclined to adopt EVs. Two works directly measure the effect of social influence as an attribute in the stated preference experiment. The first is [82], who included the market share of EV. The second, [106], extended this methodology, dividing the market share of EV in four reference groups. They also added an attribute measuring a positive/negative overview of the EV.

With respect to how social influence, in general, and SoC, in particular, affect Diffusion, there are remarkable contributions. [104] focused on the role of the social network in the diffusion process; in particular, in the imitation component. They proposed an extension of the Bass model to make the social network a function of the number of customers who adopted the product. However, their model only accounted for the rate of adoption and did not account for the substitution effect. [128] looked at the role of social identity and how the diffusion of a product in the society is affected by identity signaling. They extended a Bass model assuming that the probability of adoption is influenced by three factors: a background ratio of spontaneous adoption, social influence from one's group members, and social influence from members of the outgroup. However, the model assumptions made it unsatisfying, especially with regard to the well-mixed population implicit in the Bass model; i.e. members of different social groups interact with one another, but individuals often preferentially interact with the members of their own social groups. [99] found that information about a large number of previous adopters positively influenced adoption only if those previous adopters were described as similar to the potential ones.

There are other alternative approaches, as [79], who modelled the diffusion of technologies based on decision making with the underlying assumption that users make rational choices aiming to maximize their utility. This decision-based proposition reflects the heterogeneity of potential adopters; i.e. adopters differ in their characteristics, which results in different utilities. [42] proposed a social simulation that aims to validate a psychological theory that include goals, deliberative decisions, status quo bias, social environment, communication over personal networks, and sensitivity towards external events such as price changes or messages from the media. Finally, the work of [25] has been closely followed in this research. The author considers both Informational and Normative conformity; the latter being subdivided into Descriptive norms, Injunctive norms and Social signaling.

Regarding one of the contributions of this research, the consideration of an inner circle and an outer circle of social connections, some authors have pointed in a similar direction. However, this idea has not been developed in the explicit manner that is developed in this research. Most of the works in the field, in general terms, either study the nature of the ties between the *ego* and the *alter*, or the evolution of the network ([94], [19], [24] or [119], [120], [121]). In [94], authors understood that different processes occur in different social structures, considering implicitly a level of closeness. From other fields, is worth to mention the work of [22], who provided an overview of the approaches used to measure *social capital*. There is no unified definition of social capital, but it certainly comprises individuals' network connection. One of the differentiations of social capital is that of structural versus cognitive. The structural component describes properties of the networks, relationships and institutions that bring people and groups together; while the cognitive dimension is derived from mental processes and reflects people's perceptions of the level of trust, confidence, shared values, norms, and reciprocity. Although this paper is written from a Public Health perspective, they mentioned, at an individual level, mechanisms through which social capital may influence health, which in this case, would correspond to diffusion. These mechanisms are: social support, social influence through shared norms or social control, social engagement and social participation, physical person-to-person contacts. These elements are especially relevant in health –infectious diseases– but not that much in the diffusion of a technology.

Another necessity related to the idea of social circles is to quantify their size. In this regard, there are comprehensive tools like Name, Position, and Resource Generators. Name Generators aim to identify the network structure and content. Respondents are asked to write down names of people they know. Then, information about the relation with those individuals is collected. Position Generators look for social resources that an individual has access to through connections to somebody that, due to his or her occupation, can provide wealth, power and prestige. A

Position Generator usually provides a list of occupations and asks if the interviewee knows someone dedicated to each occupation. Then it is asked whether these people are family, friends, or acquaintances. Finally, Resource Generators measure access to specific resources, previously defined by the researcher, that are relevant to the outcome.

Some valuable findings have been discovered using these methodologies. [98] calculated (in specific Dutch networks) that the average number of people one would ask for help with small jobs in and around the house and/or with whom one would discuss important personal matters (Name and Position Generators) increased from 3.58 in 2000 to 4.15 in 2007. The average number of confidants for discussion of important personal matters was, on the contrary, remarkably stable; 2.31 in 2000 and 2.41 in 2007. [145], for their part, conducted a study on network change with data from 1968/1978. The data showed that 845 Torontonians reported 3,930 socially close 'intimate' ties, which is around 4.5 per respondent; interestingly close to [98] results despite the time difference between the two studies.

2.2 Discrete choice and Diffusion models

Great effort has been made by some authors in predicting the growth path of EVs in the last decade, but this is a challenging task. The most recent and advanced works in predicting the market share of the EV combine substitution and diffusion models. Typically, DCM are used for the first, providing a quantification of the willingness to purchase an EV instead of an ICV; whereas the diffusion effect is based on Bass ([6]) models, or its multiple extensions. For instance, [41] used an agent-based choice model where the purchasing decision of customers is affected by media coverage and social interactions. However, it was only used to explore potential nonlinear interactions between several elements of influence. [53] developed a diffusion specification that incorporated multicriteria analysis and choice models, they focused on the geographical uptake of EVs and the effect of policy incentives. [122] and [132] proposed a simulation system that integrated disaggregate demand and system dynamic models, including the diffusion effect. However, the parameters were exogenously defined rather than estimated. [77] set up a simulation system where a richer disaggregate demand model (estimated separately in previous research) was integrated into a dynamic simulation system. This approach allowed modeling the important feedbacks in demand as a result of the dynamic evolution of EVs and their charging characteristics. However, they did not account for the diffusion effect. More recently, [67] suggested a method that combines diffusion, as typically estimated in the marketing literature, with advanced DCM. All the parameters of the joint substitution/diffusion model are estimated jointly and the disaggregate model estimated with SP data is adjusted to the real market endogenously in the diffusion process. However, this extension only included innovation through one only term, which effect on the probability of choosing EV varies over time linearly. Moreover, imitation was left aside since it was dependent on the number of individuals that had adopted the product already.

[3] proposed in their work to evaluate the interaction between accumulated sales of EVs and the battery price using the Generalized Bass diffusion. Among

its main results, the author shows that the Moroccan market reaches the maximum sales of EVs after 14 years, and that the cost of the battery has a significant effect on the market, accelerating the diffusion of the EV.

[100] made a forecast on sales of EVs in Spain until the year 2040. Taking into account the different scenarios, the authors used a developed version of the Bass model, known as BB-04X that allows contemplating the different generations of the product, as well as the 'jump' phenomenon (switch to a different generation). The data used in this research was obtained from a survey developed by a company in different countries of the European Union in 2012.

[95] studied the potential of Bass models to evaluate the policies to promote the diffusion of the EV market in Germany. The authors suggest that researchers may have problems to choose appropriate values for the parameters of the model due to the high variation in the number of them. Therefore, the varying values found in the literature appear debatable. Similarly, [7] adopted a Bass model to forecast, using as inputs; market size of the new technology; a parameter that captured the percentage of buyers whose purchase decision was not influenced by the purchasing behavior of others; and a metric that captured the likelihood that additional consumers adopted the technology in response to the buying experience of others.

With a similar interest in knowing what drives the EV market, [45] conducted a thorough analysis of 40 research articles that developed diffusion models as well as other approaches to the adoption of electric cars. The authors aim to study the similarities among the models to offer recommendations for their implementation.

On the other hand, according to [78], the Bass model is not easily parameterized when there is no available market data. Hence, radically new products that imply changes in consumer behavior, such as the case of EVs, restrict its use. [114], in his research, presented a Bass model and a DCM with a dynamic perspective to assess how consumer preferences and social forces influence the introduction of EVs on the market. The model offered certain advantages; on the one hand, both time and market share could be estimated jointly; second, the model was flexible concerning the number of products and attributes; and, finally, it was easily parameterized through joint analysis without the need of market data.

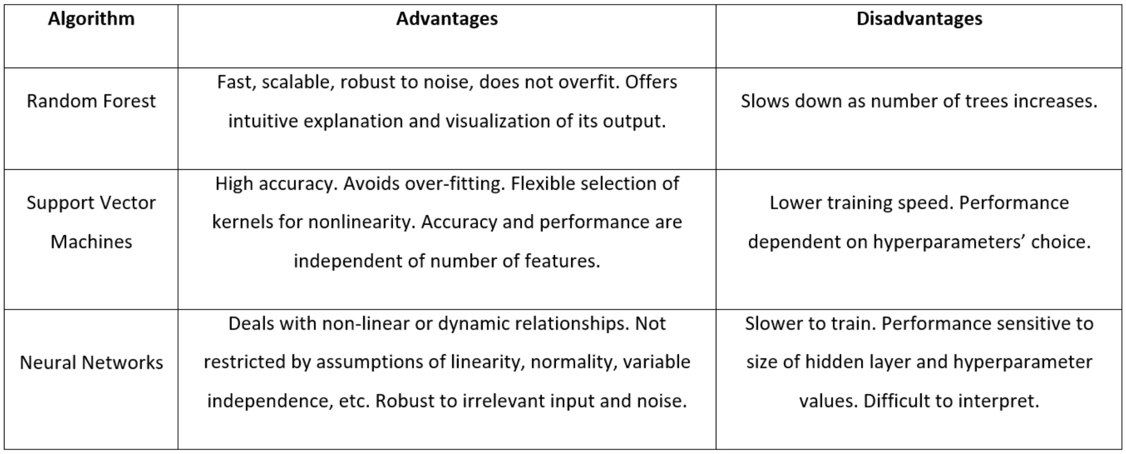
Finally, [69] proposed a Generalized Bass Model, which offers better overall performance than other specifications, both in terms of fit and forecast performance. The analysis also shows that it is also suitable to explain the multigenerational diffusion process, and that the effect of marketing mix variables can be incorporated to model the pace of the diffusion.

2.3 Machine Learning

Alternative fuel vehicles have been subject of several ML applications, especially in topics such as battery estimation, energy consumption, or range estimation. ANN [151] and SVM [124] have been used to estimate the state of health or the state of charge of batteries; as well as other less-known approaches such as fuzzy c-means clustering with backpropagation [55]. More recently, [43] proposed the use of energy consumption predictive models to forecast the energy consumption of new EVs in absence of training data. To estimate vehicle's range, [148] utilized an ANN with one hidden layer of 60 neurons in conjunction with a Decision Tree (DT) to estimate the road type when it is not known. Stop delivery times prediction [59], traffic flow estimation [86], driving behavior recognition [149], or parking occupancy prediction [147] are other specific transportation topics to which ML techniques have been applied.

There exist several works that have carried out comparison of algorithms. [63] used data from cellphones' accelerometers and gyroscopes to predict transportation mode, comparing the prediction accuracy of SVM, DT methods and K-Nearest Neighbors (KNN). Results showed that RF and SVM had best performance, although they have difficulties in differentiating between car mode and bus mode. [56] discriminated driving conditions using speed and acceleration data, comparing the prediction throughputs of SVM, ANN, linear and quadratic classifiers, and K-means clustering. A similar work is that of [143], who applied similar techniques to driving-style classification. One especially comprehensive work is that of [133] who compared the results of Multinomial Logistic Regression (MLR), Classification and Regression Trees (CART), and Gradient Boosting Decision Trees (GBDT) for the prediction of electrical vehicle range. Results showed that GBDT could optimize predictions and reduce error better than the other two techniques. Another exhaustive comparative study is the one carried out by [46] who estimated utility factor (i.e. ratio of miles travelled with electric energy over the total number of miles travelled) in hybrid vehicles. Four different approaches were compared: Regression Tree (RT), RF, SVM and ANN, concluding that SVM and ANN gave the best estimation accuracy. More in line with the spirit of the present work are the studies of [38] and [68]. The first uses K-means clustering to create six consumer segments around EV adoption. The second compares five ML techniques in the context of alternative fuel vehicles. [85] also presents an interesting exercise that combines a Bass model with ML algorithms to explain the diffusion process of pre-launched products. Finally, there exist two general reviews of classification techniques. [81] provided a score on relevant aspects to several methods. RF excels at speed of classification, handling all kinds of attributes (discrete/continuous) and explanation ability, although accuracy is not one of its strengths. On the contrary, SVM are very accurate and fast, with high tolerance to irrelevant attributes, although its results are difficult to explain and its speed of learning decreases significantly as the number of attributes grows. Finally, the performance of the ANN seems to be somewhere in between, with a dangerous tendency to overfitting. More recently, [126] carried out a similar exercise in terms of pros and cons that is shown in Table 2.1. Both works seem to coincide in their conclusions.

Table 2.1: Pros and cons of supervised machine learning classification algorithms. Edited from [126].



Algorithm	Advantages	Disadvantages
Random Forest	Fast, scalable, robust to noise, does not overfit. Offers intuitive explanation and visualization of its output.	Slows down as number of trees increases.
Support Vector Machines	High accuracy. Avoids over-fitting. Flexible selection of kernels for nonlinearity. Accuracy and performance are independent of number of features.	Lower training speed. Performance dependent on hyperparameters' choice.
Neural Networks	Deals with non-linear or dynamic relationships. Not restricted by assumptions of linearity, normality, variable independence, etc. Robust to irrelevant input and noise.	Slower to train. Performance sensitive to size of hidden layer and hyperparameter values. Difficult to interpret.

Thus, only the study of Jia ([68]) is similar in nature to our work, although with notable differences: it was not specific to EVs, it only considered the newest vehicle of the household, the observations with missing information were removed from the dataset, and it did not take into account the social component involved in the adoption of a new technology. This work, on the contrary, was specifically designed to gather individuals' willingness to purchase an EV; performs an advanced process for imputing unknown information; includes social elements involved in the decision-making process; and makes use of neural networks to predict adoption.

2.4 Survey design and data availability

SC experiments, as proposed by [89] and [88], have been widely applied in many different disciplines. However, it is fair to say that for a long time the research in decision experiments has probably been led mainly by the field of marketing, while transportation focused more on advances in discrete choice modelling. [8], [51], [90], [137] are good examples of this; however, none of these works address the issue of SC experiment design in depth, and it is difficult for transportation practitioners to find background material that can provide a complete understanding of this discipline.

At a conceptual level, a design consists of a series of values to be mapped onto each choice situation –also referred as *choice task*. This design is of substantial mathematical complexity and the use of specific software is needed. In this research Ngene ([26]), has been used. The specifics of the design will be fully covered in sections 3.2.1 and 3.2.1. For now, it is sufficient to indicate that the output yielded by Ngene consists of a matrix of values in which each column represents a variable, while each row represents a choice situation. This is how some authors present their designs ([15], [110]). Others, on the contrary, use rows for alternatives and columns for their attributes ([21], [57], [75], [76], [113], [115]). In this case, multiple rows are grouped to form a choice situation.

One of the approaches in experimental design to populate the choice tasks has been the orthogonal design ([90]). More recently, however, some researchers have begun to question it, such as [116]. They argue that the property of orthogonality –all between-attribute correlations are zero– is in conflict to other desirable properties of the models used to analyze Stated Choice (SC) data. It is true that orthogonality is an important criterion to determine independent effects in linear models, but DCM are not linear. In theses types of models, the correlation structure between the attributes is not of importance, but the correlations among the differences in the attributes. [57] showed that designs that relax orthogonality and attempt to reduce the asymptotic standard errors of the parameter estimates generally improve the reliability of the parameters. Ultimately, the attempts to reduce the asymptotic standard errors of the parameter estimates resulted in a class of designs known as *efficient* or *optimal*. These designs require measures of their degree of efficiency, which are derived from the Asymptotic Variance-Covariance (AVC) matrix. Several efficiency criteria have been proposed ([13], [15], [73], [74], [76], [117]), the A-error, D-error and C-error being the most common. The D-error is the approach selected for this research, as it will be explained in its corresponding section.

The purpose of designing a survey is, of course, the collection of data. However, there are studies that use data collected for other works. This is the case of [77], who did not design a specific survey for the development of their model, but they adapted a proposal made by [144]. This approach modeled explicitly the interdependencies between consumer choice, consumer characteristics, evolution of propulsion technologies, and availability of service stations, at the macro level. [7] also did not conduct an survey exclusive for their study, as well as [3], due to the scarcity of data in Morocco concerning the adoption of EV. The latter adopted a method called *analogy prediction* to Liquefied Petroleum Gas vehicles. Other examples are [45] and [114].

On the other hand, [100] did use data obtained from a survey conducted in 2012 in different countries of the European Union (Spain, France, the United Kingdom, Italy, Germany and Poland). ([134]), for their part, performed a study founded by the Institute for Energy and Transport –one of the seven Institutes of the Joint Research Centre, a Directorate General of the European Commission– for which they collected data. A selection of participants had to fill in a trip diary. Then the respondent was presented with information associated to a conventional ICV in comparison to an EV (time to charge, purchase price range, cost of use, and emissions), to finally be asked to evaluate on a scale of 0 to 100 the possibility of buying the EV. Similarly, [95] used existing German car market data to investigate how practitioners could choose adequate values for the parameters of a Bass model in order to forecast the diffusion of EV.

3. Substitution-Diffusion models for the prediction of EV market share

3.1 Electric vehicle adoption: the case of the Danish market

As pointed out previously, forecasting the demand for EVs is a difficult task, and the predictions published so far either haven fallen short or have been too optimistic. Although EV penetration has been slow so far, substantial increases in sales are beginning to be seen, encouraging the hope of its arrival. This is particularly true in the U.S., where cumulative sales reached 1.44 million units sold in January 2019 ([87]). Therefore, reliable predictions are needed, which could be the basis of both strategic decisions of private companies, and public policies in favor of EVs. This chapter presents an extension of a work already done, and which constitutes the basis of the more developed and ambitious methodology described in Section 3.2. In brief, it consists of improving the formulation found in [67] by including in the model specific variables to account for both intrinsic innovation and imitation. In doing so a dynamic model is built, making the demand of EVs in time t dependent on the number of EVs sold in time $t-1$. Hence, the methodology adopted is based on three pillars:

- Substitution: Which may happen by replacing an Internal Combustion Vehicle (ICV) – gasoline or any other engine – by an EV, by purchasing an additional EV for the household, by upgrading to a new generation of EV, or even by selling an EV back. In any case, there is clearly a substitution component in the diffusion of EVs since it is a good that individuals consume in order to substitute another means of transportation, especially a car with a different type of engine.

- Diffusion: The approach in the classic Bass models and its extensions is that a new technology is first adopted by a subgroup of *innovators*, who are followed by *imitators*. This process simply happens through a progressive increase in sales; i.e. in a classic Bass model, the accumulated sales increase, which in turn contributes to increasing the future sales. However, it is known that this process is eminently social, as Rogers points out ([112]). Therefore, it is absolutely necessary to include social elements in any prediction of diffusion of technology.

- Dynamics: Diffusion not only depends on the context at a particular moment but also on what occurred in the previous periods. Models that aim to predict sales should include variables referring to previous periods of time.

Considering these three capital elements in predicting the diffusion of a new technology, the methodology followed makes use of real data from the Danish vehicle market to estimate a disaggregated substitution model that accounts for social influence and social conformity. Its output feeds an extended Bass diffusion model, which finally yields total sales in each period using projected variables. This procedure can be summarized as follows:

1. Treatment of the Danish EV market data. The structure of these data is similar to [67], but in this case they have been updated by the Danish Energy Agency until the year 2018.

2. Estimation of a disaggregated substitution model with an extended formulation that includes social influence and social conformity as described below. Its output model is used in the estimation of a diffusion model.

3. Projection of the variables into the future and use the estimated diffusion-substitution model to forecast EV sales in Denmark.

3.1.1 Data

The data used in this research comes from different sources. The SoC coefficients are part of the work of [25], while the ones related to vehicle attributes proceed from [67]. The vehicle characteristics, as well as the Danish EV monthly sales, have been computed by the Danish Energy Agency until the year 2018. The projection of the attributes is based on the following sources: European Centre for Mobility Documentation, Danish EV Alliance, Norwegian Information Council for Road Traffic, Clever A/S.

As for the social conformity parameters, they are derived from a survey performed in Denmark in the period between December 2014 and January 2015. The survey was built specifically to study the effect of parking policies on the choice of

EVs versus ICVs, as well as the role played by social conformity on this choice. It consisted of five sections:

1. Detailed information about the last parking activity and information on household vehicle ownership and use, definitions on the most likely future vehicle purchase and information on whether a new EV car would replace an existing one or if it would be an additional one in the household. Users were also asked to indicate the degree of influence that they had in the decision about the type of car.

2. A Stated Choice (SC) experiment, pivoted around the values collected in the first section. The SC included attributes related to the car characteristics and to the parking options, plus attributes that allow measuring the effect of conformity.

3. The third section was dedicated to gathering socioeconomic and residential information.

4. Individuals' attitude and perception towards several aspects related to EVs, injunctive social norms, affections, and values in life. Injunctive norms define when the individual's behavior is affected by what other people think of them doing something. In this case, the norm is measured asking respondents about the level of agreement to the following statements:

   a. *People who are important to me (friends, family) would approve of me using an electric vehicle instead of my conventional car.*

b. *People who are important to me (friends, family) think that using an EV instead of my conventional car is appropriate.*

c. *People who are important to me (friends, family) think that more people should use an EV instead of their conventional car.*

5. Finally, information about personal and family income was asked.

For further information about the SC experiment, see [25]. The sample was gathered from a list of individuals who had signed up to participate in a real-life experiment in 2010 in which they could use an EV for three months. 39% of the participants had already heard and been informed about EVs. Thus, including people who have tried an EV in real life, allowed the authors for some consideration about the impact of experience in the diffusion process. Table 3.1 reports a summary of the information about the recruitment process.

Table 3.1: Recruitment process. Source: [25].

	TOTAL	% with respect to		
		Contacts	Replied	Eligible
Contacted	17,299	100%		
Replied (i.e. at least entered the survey or sent an email)	5369	31.04%	100%	
Asked explicitly to be removed	68	0.39%	1.27%	
Have changed email or do not live in Denmark anymore	48	0.28%	0.89%	
Already have an EV	3	0.02%	0.06%	
Do not have a car, have not driven or did not experience parking problems in the last 2 months	108	0.62%	2.01%	
Eligible	5142	29.72%	95.77%	100%
Only entered the survey	2270	13.12%	42.28%	44.15%
Dropped after parking information (Section 1)	216	1.25%	4.02%	4.20%
Dropped after the SC experiment (Section 2)	37	0.21%	0.69%	0.72%
Dropped after SE characteristics (Section 3)	24	0.14%	0.45%	0.47%
Dropped after Attitudes statement (Section 4)	10	0.06%	0.19%	0.19%
Dropped after Affects statement (Section 4)	100	0.58%	1.86%	1.94%
Dropped after Values statement (Section 4)	82	0.47%	1.53%	1.59%
Completed the full survey	2403	13.89%	44.76%	46.73%
Completed the SC experiment	2656	15.35%	49.47%	51.65%

The majority of the individuals are male (73%) and employed (78%), while only half of them (42%) have a job with a fixed number of hours. On average, respondents are 47 years old and live in households with 3.12 members and 1.52 cars. Table B.1 in Appnedix B illustrates other characteristics of the 2,363 respondents of the sample.

Finally, for the forecasting part, only one scenario has been designed, in which the EVs experience gradual improvements in their characteristics thanks to technological progress. This information is based on the following sources: European Centre for Mobility Documentation, Danish EV Alliance, Norwegian Information Council for Road Traffic, Clever A/S).

3.1.2 Model structure

The model of reference is that of [67], which in turn, is based on the one of [71], which is an extension of a basic diffusion model that accounts for substitution effects. [71] included the diffusion effect into the utility of purchasing a new technology k at time t.

$$V_t^{(i,k)} = q^{(i,k)}(t - \tau^k + 1) + \beta^{(i,k)} x^{(i,k)} \tag{3.1}$$

where $x^{(i,k)}$ is a vector of the technology attributes, $\beta^{(i,k)}$ its corresponding coefficients, $q^{(i,k)}$ the time-dependent diffusion effect, and τ^k is the period of the introduction of this technology in the market. The superindex (i,k) refers to the case of an individual owning a technology i who switches to k. The probability of purchasing a product of generation k is

$$P_t^{i,k} = \frac{exp(V_t^{i,k})}{exp(c) + \sum_j exp(V_t^{i,j})}, \qquad k \geq i, j \geq k \tag{3.2}$$

Considering M_t the potential market at time t, and Y_{t-1} the total number of units of product at time $t-1$, the number of sales in each period is

$$\begin{aligned} S_t^k &= (M_t - Y_{t-1}) \cdot P_t^k && (3.3) \\ &= (M_t - Y_{t-1}) \cdot \frac{\exp(V_t^{i,k})}{\exp(c) + \sum_j \exp(V_t^{i,j})} && (3.4) \end{aligned}$$

From here, [67] defined their model as

$$\begin{aligned} S_t^{EV} &= (M^{EV} - Y_{t-1}^{EV}) \cdot Pr(EV_t) \\ &= (M^{EV} - Y_{t-1}^{EV}) && (3.5) \\ &\cdot \frac{exp(ASC^{EV} + q^{EV}(t - \tau^{EV} + 1) + \lambda(\widehat{\beta}^{EV} x_t^{EV}))}{exp(\lambda(\widehat{\beta}^{ICV} x_t^{ICV})) + exp(ASC^{EV} + q^{EV}(t - \tau^{EV} + 1) + \lambda(\widehat{\beta}^{EV} x_t^{EV}))} \end{aligned}$$

where $\widehat{\beta}^{EV}$ and $\widehat{\beta}^{ICV}$ were estimated using SP data and fixed in the diffusion process. The three parameters estimated are ASC^{EV}, q and λ, which represent the alternative specific constant, the *diffusion* parameter, and a scale coefficient, respectively.

That being said, the research presented here brings a several improvements to this methodology. In order to consider the effect of social influence on the individual choices, two elements are included; the number of EV sold in the previous period $t-1$, which makes the model dynamic; and the information that the potential customer receive about specific characteristics of EV. Regarding the latter, [25] showed that only negative feedback is significant, due to the *negativity bias* effect. This refers to the understanding that "negative information tends to influence evaluation stronger than comparably extreme positive information" ([62]). The specific information that other users report on are parking spaces reserved to EV, range, and the need to change activities. They all are compiled in an unique dummy variable named $Info$, which is 1 (negative information received by the potential purchase), for all time periods until the charging variables reach 33% of presence and the EV range also reaches 33% of the ICV range. At that point, it is assumed that the negative feedback about parking spaces, range and need to change activities becomes positive and, therefore, $Info$ takes the value 0 onwards.

On the other hand, the preliminary data analysis showed a peak in sales in December 2015. This was provoked by the Danish government announcement that the registration tax for EV would be increased. Instead of considering this information as an outlier, a dummy variable was defined to model the anticipation of this policy.

Considering all these aspects, the utility function that is base of our formulation was:

$$V_t^{EV} = ASC + q(t-\tau+1) + \lambda(\widehat{\beta}\ln(N_{t-1}^{EV sold}) + \beta \cdot Info_{t-1} + \widehat{\beta} X_t^{EV}) + \beta_1 \cdot Antic \quad (3.6)$$

The elements common to [67] maintain their meaning, $Info$ stands for the concept discussed above, and $Antic$ for the aforementioned anticipation to the registration tax policy. λ is the substitution parameter, which reflects the effect of the attributes and information received about the EV. It is worth mentioning that the number of EV sold in the previous period is considered in logarithms since the relation of this variable with its lags is clearly not linear. Equation 3.6 leads to a number of sales in each period equal to:

$$\begin{aligned} S_t^{EV} &= (M^{EV} - Y_{t-1}^{EV}) \cdot Pr(EV_t) \\ &= (M^{EV} - Y_{t-1}^{EV}) \cdot \frac{exp(V_t^{EV})}{exp(V_t^{ICV}) + exp(V_t^{EV}))} \end{aligned} \tag{3.7}$$

The next section provides the results of this formulation and their discussion, as well as a comparison with those of [67]. Special emphasis is placed on sales forecasting, and on the impact that the new elements considered have on it.

3.1.3 Results

To forecast the diffusion of a new technology it is necessary to make some assumptions about the development of its attributes. In this case, one only scenario has been designed, in which the EV experiences gradual improvements in its characteristics thanks to technological progress. Table 3.2 provides the values assumed in the forecast, which are realistic, but also allow for some substantial advances. This information is based on the following sources: European Centre for Mobility Documentation, Danish EV Alliance, Norwegian Information Council for Road Traffic, Clever A/S).

Table 3.2: Forecasting scenario 2018 - 2050.

MONTH	YEAR	CTY_SL	SHO_SL	CTY_FA	SHO_FA	PP_EV	PP_GAS	FU_EV	FU_GAS	RA_EV	RA_GAS	CO2_EV	CO2_GAS
3	2018	0.2	0.2	0.2	0.2	262,032.1	212,573.2	0.3	0.7	220.5	852.8	41.9	140.9
4	2018	0.2	0.2	0.2	0.2	262,032.1	212,573.2	0.3	0.7	220.5	852.8	41.9	140.9
5	2018	0.2	0.2	0.2	0.2	262,032.1	212,573.2	0.3	0.7	220.5	852.8	41.9	140.9
6	2018	0.2	0.2	0.2	0.2	262,032.1	212,573.2	0.3	0.7	220.5	852.8	41.9	140.9
7	2018	0.2	0.2	0.2	0.2	262,032.1	212,573.2	0.3	0.7	220.5	852.8	41.9	140.9
8	2018	0.2	0.2	0.2	0.2	262,032.1	212,573.2	0.3	0.7	220.5	852.8	41.9	140.9
9	2018	0.2	0.2	0.2	0.2	262,032.1	212,573.2	0.3	0.7	220.5	852.8	41.9	140.9
…	…	…	…	…	…	…	…	…	…	…	…	…	…
5	2050	1	1	1	1	1417,96.5	194,178.3	0.3	0.6	373.7	852.8	0	108.8
6	2050	1	1	1	1	141,796.5	194,178.3	0.3	0.6	373.7	852.8	0	108.8
7	2050	1	1	1	1	141,796.5	194,178.3	0.3	0.6	373.7	852.8	0	108.8
8	2050	1	1	1	1	141,796.5	194,178.3	0.3	0.6	373.7	852.8	0	108.8
9	2050	1	1	1	1	141,796.5	194,178.3	0.3	0.6	373.7	852.8	0	108.8
10	2050	1	1	1	1	141,796.5	194,178.3	0.3	0.6	373.7	852.8	0	108.8
11	2050	1	1	1	1	141,796.5	194,178.3	0.3	0.6	373.7	852.8	0	108.8
12	2050	1	1	1	1	141,796.5	194,178.3	0.3	0.6	373.7	852.8	0	108.8

For the model defined in Equations (3.6) and (3.7), the potential market, M is defined as 877,000. This is half of the car-owning families in Denmark. More complex assumptions might be made, but they would have been difficult to validate, while this one seemed simple yet good enough.

Table 3.3 shows the results for three different specifications. Model 1 is the original specification from Jensen et al. ([67]), estimated with the new data. Model 2 is the model proposed in (3.6) and (3.7), but in which the variable $Info$ (feedback provided by others) has not been included. Finally, Model 3 is the full model proposed in (3.6) and (3.7). In other words, Model 1 does not include any social element, while Model 2 includes SoC (EV sales) and Model 3 includes SoC and social influence ($Info$). This distinction is made in order to quantify separately the impact of these elements in the prediction.

Table 3.3: Estimation results.

	Model 1		**Model 2**		**Model 3**	
	Value	**p-value**	**Value**	**p-value**	**Value**	**p-value**
`ASC_EV`	-13.26	0.0001	-12.78	0.0001	-11.38	0.0001
q	0.94	0.03	0.83	0.06	0.58	0.05
λ	0.05	0.81	0.14	0.52	0.12	0.59
$Antic$	2.64	0.0001	2.61	0.0001	2.59	0.0001
$R2$	0.707		0.707		0.711	
$Pr > F$	0.0001		0.0001		0.0001	

For the full model, the values of both ASC and q are in line with the findings of [67], although in the lowest value of the confidence interval. The diffusion parameter is significant at 95% level, which shows that considering the effect of diffusion allows a more realistic forecast of the EV spread. On the other hand, λ is not significant, reflecting that substitution plays a minor role in choice. The variable that gathers the effect of December 2015 is highly significant.

The most interesting result of Model 2 is that the value of the substitution parameter increases and gains significance if the information offered by a close friend is not included. On the other hand, Model 1, being simpler than the others, offers a higher value for the parameter q, but significant in turn. A possible reason for that may be that the elimination of both the social conformity and social influence variables causes the entire explanatory power to fall on the remaining variables associated with q.

Figure 3.1 shows, for the three models, the forecast obtained for the period 2018 — 2050. Namely, EV monthly sales and the cumulative number of EVs sold expressed as a percentage of the total.

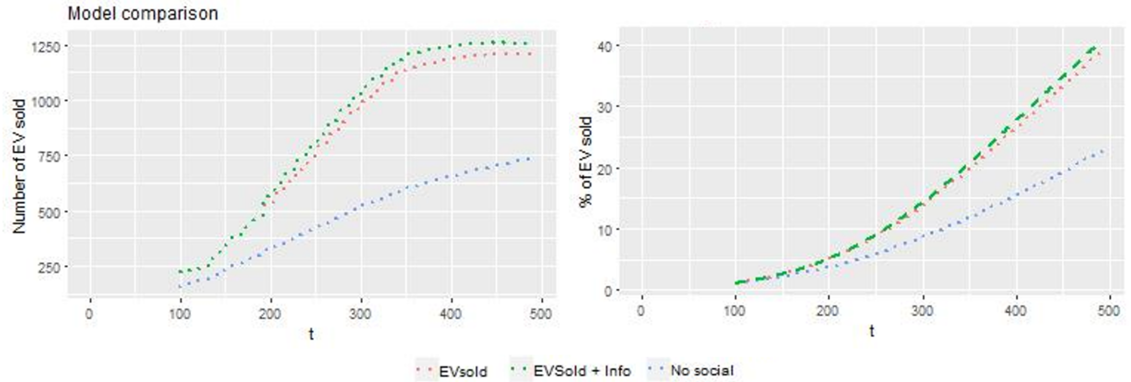


Figure 3.1: Actual, fitted and forecast sales. Cumulative number and percentage of EVs sold.

The results illustrate the classic S-shape, with low market penetration in the early stages, and later progressive increase once the product is more present in the market. For Model 2 and Model 3 the share of EVs evolves from just 2% at the beginning of 2018 to around 40% by 2050. However, this evolution is more progressive in the model that includes no social elements, Model 1, reaching a more conservative final market share. That might be the most important contribution of this research; the substantial differences in the forecast made when including the social elements.

3.2 Electric vehicle adoption under the effect of Social Conformity: the case of the American market

As previously stated, to achieve an accurate prediction of the diffusion of the EV, the methodology carried out in the Danish case has been significantly improved. It was also applied to new data, of better quality and specific for the state of Maryland. Both aspects are described in detail in the following subsections. Nevertheless, this approach is based on the three pillars indicated above: substitution, diffusion, and dynamism. For the first two, an initial demand model will be estimated using advanced discrete choice techniques. It will explicitly include variables related to SoC and to the social network of the respondent. This model will gather the subjacent inclinations of individuals to substitute ICVs vehicle by EVs. Then, it will be integrated into a Bass diffusion model which will finally yield future market shares, i.e. the spread of this technology over time among society. The dynamic aspect is provided by time-dependent variables, present in both the utility functions and the Bass model, because diffusion not only depends on the context at a particular moment but also on what occurred in previous periods.

As detailed above, SoC is a relevant aspect in diffusion because people around us, such as family members, friends, colleagues, or even people that we do not know, influence our behavior and decisions, directly or indirectly. According to [36] we all have some tendency to either yield to group pressures or to agree to the majority, which can happen because of the desire of being accepted, or because of the desire to do the *right thing*. Either way, individuals tend to turn to members of their own group in order to gather information, which may involve a change in attitudes, beliefs or behavior. The way in which this concept is brought into this research is by providing to the individuals feedback from one person they know. This person belongs to what I have defined as the *Inner* or *Outer* circles, which existence is probably the main conceptual body of this research. As reported in Section 2, although this notion resides in some way in several papers that stand out in the field of Transportation ([94], [19], [24], [119], [120], [121]), this idea has not been developed in the explicit manner that is carried out in this research. The purpose here is to explore explicitly the size, nature, and impact of these relationships. This required the development of several elements, the main one being a new SC experiment that gathers data about both the individuals' preferences for EVs and the aspects derived from the Inner/Outer circle idea.

3.2.1 Stated Choice experiments

The purpose of stated choice experiments is to determine the influence of the characteristics of a set of alternatives on the probability of choosing them. A study of this type normally consists of an individual making a choice in a hypothetical scenario in which different levels of attributes related to the set of alternatives are presented. Given the difficulty (economic or technical) of counting with a high number of individuals, it is common practice to make the respondent face a number of these choice situations and pool the responses.

Since attributes levels are the main decision element, a careful predefinition of them is important. Frequently, they come from; i) some other previous experiment; ii) real information obtained from the market; iii) existing literature. Nevertheless, the reality is that, in most cases, this information comes from all these sources and is slightly modified to overcome potential problems such as lexicographic biases.

Conceptually, a design consists of a series of values to be displayed in each scenario. Each column represents a variable, while each row represents a choice situation. This is how some authors represent their designs. Others, in contrast, use rows to represent alternatives and columns to represent their attributes. In these cases, multiple rows are grouped together to form a choice situation. They are simply different ways of representing the same information. In this study the first approach is taken, since it is the one used by the software Ngene ([26]), utilized for the design.

Therefore, the fundamental question is how to distribute the levels of the attributes throughout all the situations of choice that will appear in the questionnaire. This is not a trivial matter, and requires a great deal of preliminary work, as the number of attributes, their levels, and the number of alternatives exponentially increase the combinations needed for a correct design. In addition, other complexities such as the type of design and the underlying discrete choice model also come into play. Regarding the former, there are three main approaches: Full factorial, Orthogonal, and Efficient designs. Full factorial designs are unfeasible, except in the case of a small number of alternatives, attributes, and attribute levels, as this type of designs considers all possible attribute combinations. In the case, for example, of three attributes with 2, 2, and 3 levels, there would exist 12 choice situations, as shown in Table 3.4.

Table 3.4: Example of full factorial design.

s	A	B	C
1	-1	-1	-1
2	-1	-1	0
3	-1	-1	1
4	-1	1	-1
5	-1	1	0
6	-1	1	1
7	1	-1	-1
8	1	-1	0
9	1	-1	1
10	1	1	-1
11	1	1	0
12	1	1	1

In general, if there are J alternatives, each with K_j attributes, where attribute $k \in K_j$ has l_{jk} levels, the total number of choice situations in a full factorial design is

$$S^{ff} = \prod_{j=1}^{J} \prod_{k=1}^{K_j} l_{jk} \tag{3.8}$$

For two alternatives with 3 attributes with 4 levels each, the combinations are $(4 \times 4 \times 4) \times (4 \times 4 \times 4) = 4096$. It is easy to understand why the practical application of the full factorial is almost non-existent.

Orthogonal designs, widely used for many years, are another option to populate choice situations. However, there are arguments against its use since orthogonality does not meet some desired properties of the econometric models estimated afterwards. Orthogonality means, in statistical terms, that the attribute levels of the designed structure are not correlated. While this is a very desirable property in linear models, DCM are non-linear and therefore it is less relevant. In fact, what is important is the correlation of the differences in attributes. Therefore, *Efficient* designs are positioned in contrast to the orthogonal ones. This methodology tries to minimize the standard error of the estimated parameters using prior information about them (estimations available in the literature, or in previous studies, for instance). These standard errors can be predicted by determining the AVC matrix based on the underlying experiment and information about the parameter estimates obtained beforehand, technically called *priors.*

Various measures have been proposed in the literature in order to assess the efficiency of a design. They are usually expressed as an error; thus, the objective is to minimize it. The most used is the *D-error*, which is derived from the determinant of the AVC matrix. Practically, it is very difficult to find a design with the lowest D-error, and the researcher is usually satisfied if it is small enough. Different D-errors have been defined in the literature, depending on the available information on the prior of the parameters. Three different cases are distinguished:

a. There is no information about the parameters, not even their sign. Then the prior is set to $\tilde{\beta} = 0$. The D-error is then called $D_z - error$.

b. There is accurate information about β. The priors are fixed based on these values that are known, and the D-error is called $D_p - error$.

c. There is some information about β, but it is uncertain. Then, instead of assuming fixed priors, they are defined as random, following a probability distribution which expresses the uncertainty about the true value of the parameter. This is the Bayesian approach, and the error is called $D_b - error$.

There are other measures besides the D-error, such as the A-error. However, in this work the Bayesian approach is adopted, as will be detailed below in the section that fully describes the questionnaire design. It is worth noting that following this approach to construct SC experiments requires that the efficiency of a design be evaluated over numerous draws taken from the prior parameter distributions. The efficiency is then calculated as the expected value of the measure of efficiency over all the draws taken. The Bayesian approach, therefore, necessitates the use of simulation methods (Pseudo-random draws/quasi-random Monte Carlo draws) to approximate the expectations for differing designs. The potential to provide better coverage of the distribution space for each prior parameter distribution should theoretically result in a lower approximation error in calculating the simulated choice probabilities for a given design. This, in turn, results in greater precision in generating the AVC matrix.

Summarizing, the problem of generating an efficient design may be stated as:

Given feasible attribute levels Λ_{jk} for all J alternatives, K attributes, and choice situations S, and given the prior parameter values $\tilde{\beta}$ (or probability distributions of $\tilde{\beta}$), determine a level balanced design X with $x_{jks} \in \Lambda_{jk}$ that minimizes the efficiency error.

For this purpose, there are row-based algorithms and column-based algorithms, which follow different procedures to find an efficient design. The Modified Federov algorithm ([34]), illustrated in Figure 3.2 is adopted in this research.

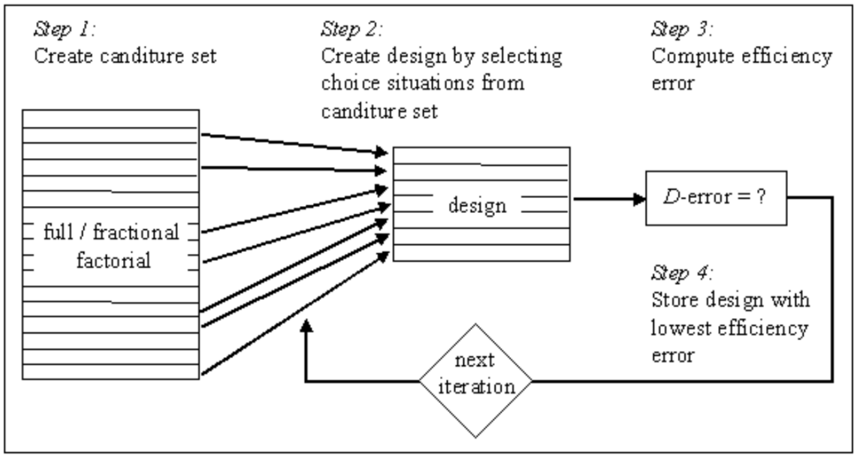

Step 1:
Create canditure set
Step 2:
Create design by selecting choice situations from canditure set
Step 3:
Compute efficiency error
full / fractional factorial
design
D-error = ?
Step 4:
Store design with lowest efficiency error
next iteration

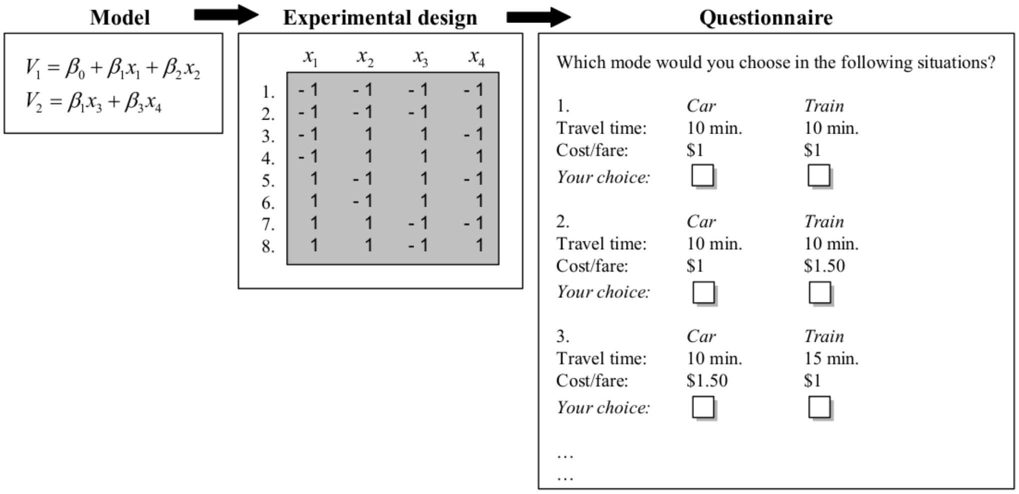
Figure 3.2: Modified Federov algorithm. Source: [26].

The algorithm starts by selecting a candidate design that can be the full factorial or a fractional factorial. Then, a new design is created by selecting choice situations from the candidate set and the efficiency measure is computed. If it is lower than the efficiency measure of the candidate, the new design is kept as the most efficient so far, and continues with the next iteration, repeating the process. The algorithm terminates if all possible combinations of choice situations have been evaluated (which in general is an enormous number of situations) or after a predefined number of iterations.

One last consideration refers to the number of choice situations. It does not seem to have an important impact on the efficiency of the design if the number of choice situations is not smaller than $K/(J-1)$. Obviously, the more scenarios are presented to the respondent, the more data available. However, too many choice situations may lead to other problems such as inaccurate or incoherent answers due to user fatigue. Therefore, it is important to find a balance between data amount, efficiency, and fidelity in responses. In general, the number of choice tasks depends on the intuition of the researcher about how many of them the user can handle.

3.2.2 Experiment design for the State of Maryland

This section describes the particularities of the SC experiment specifically designed to collect data in the State of Maryland with the aim of estimating EV adoption. It particularizes the elements indicated in the previous section for this specific case. For the sake of clarity, the process is described in three major steps: Model specification, Experimental design, and Questionnaire, illustrated in Figure 3.3. The Ngnene syntax provided in Appendix A can be of help in understanding Steps 1 and 2.



$$V_1 = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$
$$V_2 = \beta_1 x_3 + \beta_3 x_4$$

	x_1	x_2	x_3	x_4
1.	- 1	- 1	- 1	- 1
2.	- 1	- 1	- 1	1
3.	- 1	1	1	- 1
4.	- 1	1	1	1
5.	1	- 1	1	- 1
6.	1	- 1	1	1
7.	1	1	- 1	- 1
8.	1	1	- 1	1

Which mode would you choose in the following situations?

1.	*Car*	*Train*
Travel time:	10 min.	10 min.
Cost/fare:	$1	$1
Your choice:	□	□

2.	*Car*	*Train*
Travel time:	10 min.	10 min.
Cost/fare:	$1	$1.50
Your choice:	□	□

3.	*Car*	*Train*
Travel time:	10 min.	15 min.
Cost/fare:	$1.50	$1
Your choice:	□	□

…
…

Figure 3.3: Process of designing a stated choice experiment. Source: [26].

3.2.2.1 Model specification

For the specification of a discrete choice model, it is necessary to define the alternatives, the attributes, and the model type (Multinomial Logit, Mixed Multinomial Logit, Nested Logit...). This information is summarized in Table 3.5.

Table 3.5: Model specification.

Alternatives	Gasoline Electric None
Attributes	Price Propulsion cost Range Fast Charging time* / Refueling Tax deduction* Number EV sold*
Model type	Multinomial Logit, 3 categories

* EV specific attributes

The alternatives considered are *Gasoline*, *EV*, and the opt-out *Other*. By EV it is meant Battery Electric Vehicle (BEV). Thus, the EV attributes, as well as the number of EVs sold and income tax deduction, refer to BEVs. Although the number of EV sold and the tax deduction are not strictly alternative attributes, in this context attribute means any element to be shown in the scenario that has different levels. Therefore, they must be included in the programming of the design.

As for the initial model on which the design is calculated, a Multinomial Logit has been chosen due to its simplicity. Nevertheless, this does not imply that other models cannot be estimated afterwards with the data obtained from the experimental design. Additionally, three different segments of vehicles *Compact*, *Mid-size* and *Large* are considered. Therefore, nine utility functions are used to compute the parameters. It is worth noting that the absence of utility function for the *Other* alternative in the script presented in Appendix A is not an error. The opt-out alternative does not need to be specified in the Ngene syntax. Finally, different weights to the three vehicle categories were assigned; 0.31, 0.62, and 0.07, respectively, in order to form the Fisher information matrix ([137]) needed for the design. I relied for these weights on the work of [67].

3.2.2.2 Experimental design

As described above, an experimental design is basically a table of numbers in which each row represents a choice situation. The numbers correspond to the attribute levels, which are placed in the questionnaire. Therefore, the first step (after deciding the model specification) is to define the attribute levels. This is a delicate task. On the one hand, the number of levels considered must be the minimum that provides a reasonable variability of the attribute. On the other hand, although the values have to be realistic, it is not desirable that they are too much. They are the input for the user to make a choice; but it is also important to avoid lexicographic bias as well as allow the researcher to present specific situations of interest, such as a hypothetical scenario in which gasoline and electricity have the same cost. In this research, values obtained from vehicles of reference were used for starters. However, there were significant differences in several of the attributes that could persistently lead the user towards one of the alternatives. That was the case of Price, Range, and, especially, Propulsion Cost. Thus, the levels of those attributes were modified so they did not diverge in excess, but respecting the evident differences of both technologies with regard to those characteristics. Table 3.6 lists the levels of the EV and Gas attributes, for each of the three categories of vehicles.

Table 3.6: Level of attributes of the experimental design.

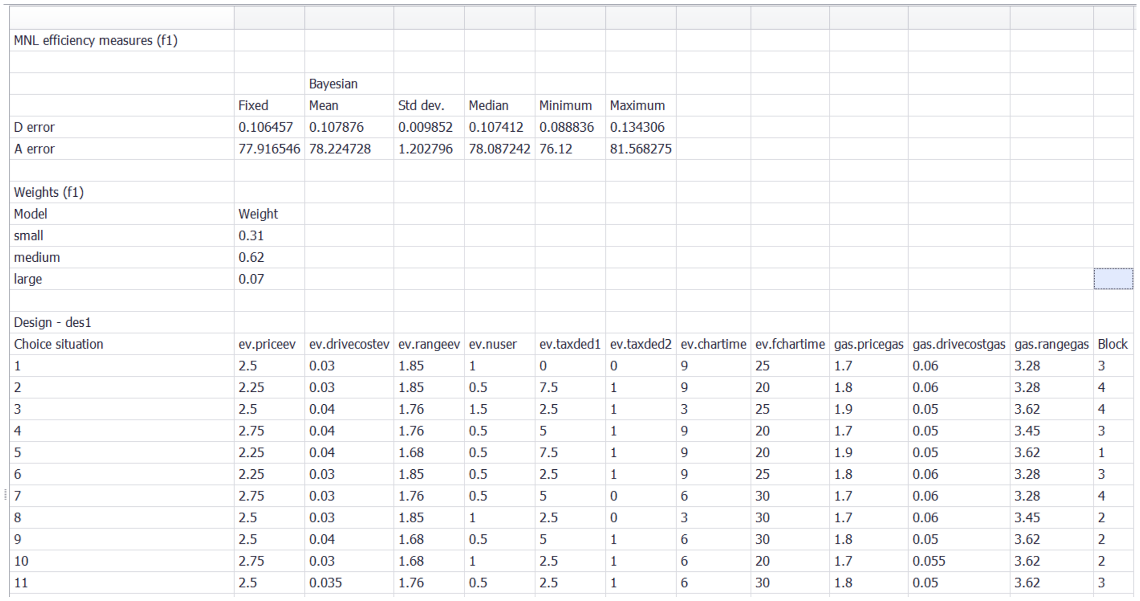
Alternative	**Attribute**	**Compact**	**Mid-size**	**Large**
Gas	Price ($10,000)	1.7 / 1.9 / 2.1	2.6 / 2.8 / 3	4.4 / 4.6 / 5
	Propulsion cost ($)	0.06 / 0.08 / 0.1	0.08 / 0.1 / 0.12	0.12 / 0.14 / 0.16
	Range (miles)	325 / 350 / 375	425 / 450 / 475	400 / 425 / 450
	Refueling time (minutes)		5	
EV	Price ($10,000)	2 / 2.2 / 2.4	2.9 / 3.1 / 3.3	4.8 / 5.2 / 5.4
	Propulsion cost ($/mile)	0.015 / 0.03 / 0.05	0.03 / 0.05 / 0.07	0.05 / 0.07 / 0.09
	Range (miles)	175 / 225 / 295	275 / 325 / 395	250 / 300 / 375
	Fast charging time (minutes)	15 / 22 / 30	25 / 35 / 45	40 / 50 / 60
	Tax deduction amount ($)	0 / 1,500 / 2,500 / 4,000	0 / 2,000 / 3,000 / 5,000	0 / 2,000 / 4,0000 / 6,000
	Number of EV sold		500 / 1,500 / 3,000	

Since an efficient design is followed in this study due to the advantages discussed above, the preliminary information about the values of the coefficients to be estimated was obtained from [67] and [25]. However, these priors were not assumed fixed, but were defined as random, to consider the uncertainty of their true value. Namely,the Uniform distributions shown in Table 3.7 were assumed.

Table 3.7: Distribution of priors.

Attribute	**Distribution of priors**
Price (EV)	$U(-0.447, -0.26)$
Propulsion cost (EV)	$U(-1.14, -0.95)$
Range (EV)	$U(0.16, 0.23)$
Number of EV sold	$U(0.01, 0.067)$
Fast charging time	$U(-0.04, -0.02)$

Regarding the choice situations, 24 of them were defined, divided into 4 blocks, terminating the algorithm when reaching 3,500 iterations. That allowed to have attribute level balance, which ensures that the parameters can be correctly estimated on the whole range of levels. Figure 3.4 shows a few choice situations of the final design of the Compact segment. Efficiency measures can be observed in at the top of the image, along with the weights of each category in the Fisher matrix.



MNL efficiency measures (f1)												
		Bayesian										
	Fixed	Mean	Std dev.	Median	Minimum	Maximum						
D error	0.106457	0.107876	0.009852	0.107412	0.088836	0.134306						
A error	77.916546	78.224728	1.202796	78.087242	76.12	81.568275						
Weights (f1)												
Model	Weight											
small	0.31											
medium	0.62											
large	0.07											
Design - des1												
Choice situation	ev.priceev	ev.drivecostev	ev.rangeev	ev.nuser	ev.taxded1	ev.taxded2	ev.chartime	ev.fchartime	gas.pricegas	gas.drivecostgas	gas.rangegas	Block
1	2.5	0.03	1.85	1	0	0	9	25	1.7	0.06	3.28	3
2	2.25	0.03	1.85	0.5	7.5	1	9	20	1.8	0.06	3.28	4
3	2.5	0.04	1.76	1.5	2.5	1	3	25	1.9	0.05	3.62	4
4	2.75	0.04	1.76	0.5	5	1	9	20	1.7	0.05	3.45	3
5	2.25	0.04	1.68	0.5	7.5	1	9	20	1.9	0.05	3.62	1
6	2.25	0.03	1.85	0.5	2.5	1	9	25	1.8	0.06	3.28	3
7	2.75	0.03	1.76	0.5	5	0	6	30	1.7	0.06	3.28	4
8	2.5	0.03	1.85	1	2.5	0	3	30	1.7	0.06	3.45	2
9	2.5	0.04	1.68	0.5	5	1	6	30	1.8	0.05	3.62	2
10	2.75	0.03	1.68	1	2.5	1	6	20	1.7	0.055	3.62	2
11	2.5	0.035	1.76	0.5	2.5	1	6	30	1.8	0.05	3.62	3

Figure 3.4: Choice situations for the Compact class.

3.2.2.3 Questionnaire

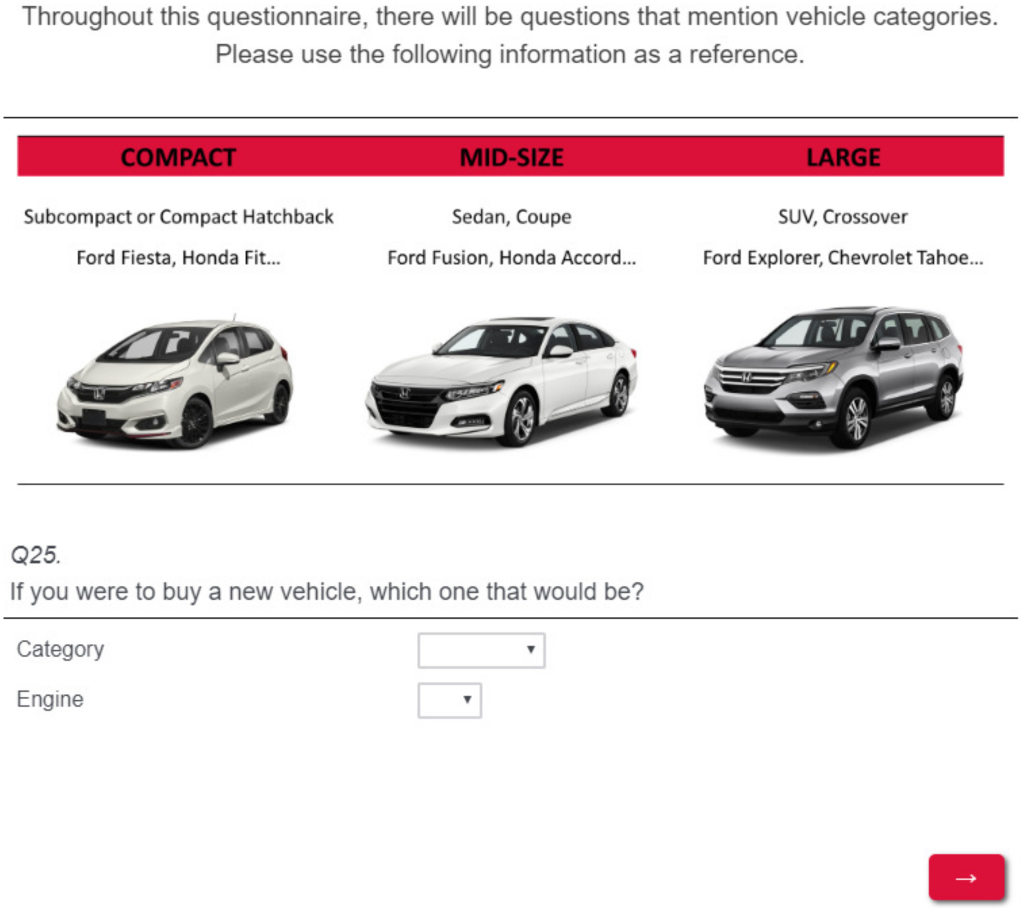
The questionnaire was built using the experimental design discussed above, and consisted of nine sections, each of them containing a number of questions that serves a specific purpose. The main part is *Stated Preference Experiment* since it contains the choice tasks. The questions have been coded using Qualtrics Research Core ([105]). Qualtrics is a survey platform that includes features suitable for generating a study of this complexity. It is a powerful tool with advanced capabilities such as logic, randomization, scoring or multitude of question types, as well as suitable management of responses and data. The different parts of the questionnaire and the survey flow are detailed below. Unfortunately, it is not possible to attach a copy of the survey as an appendix to this document. Its complexity required an intricate logical system, producing a large number of survey branches that occupy dozens of pages when printed. Nevertheless, it is important to note that the users would not face this problem, as they would only answer a small number of questions (the estimated completion time is 15-20 minutes). In other words, the survey contains all the paths that can be taken, but the user only goes through one of them. The population of interest is made up of individuals from the State of Maryland older than 18, with a driver's license. Finally, this study is not limited to first purchases, so individuals that already have an EV are not discarded.

Purpose and consent

This section consists of an introduction to the experiment, in which it is explained that the objective is to study the effect of social aspects in the willingness to purchase an EV. It also serves as a waiver of consent, complying with the ethical requirements of a study of this type.

Preliminary questions

There are two main elements in this section: next most likely purchase, and social information. For the former, the respondent indicates the category of vehicle (Compact, Mid-size, Large) they are most likely to purchase at their next acquisition. The scenarios that they will face later in the survey will be specific for that category of car. That is, they will contain specific attributes of a Compact, Mid-size or Large vehicle, as described above in the design section. In terms of the survey flow, a different branch has been programmed for each of those three categories, and the user will be conducted to the one selected. This feature differentiates this work from other studies; the elections are more realistic since the user evaluates characteristics in levels that are expected.



Throughout this questionnaire, there will be questions that mention vehicle categories. Please use the following information as a reference.

COMPACT	MID-SIZE	LARGE
Subcompact or Compact Hatchback	Sedan, Coupe	SUV, Crossover
Ford Fiesta, Honda Fit...	Ford Fusion, Honda Accord...	Ford Explorer, Chevrolet Tahoe...

Q25.
If you were to buy a new vehicle, which one that would be?

Category	▾
Engine	▾

Figure 3.5: Vehicle category choice.

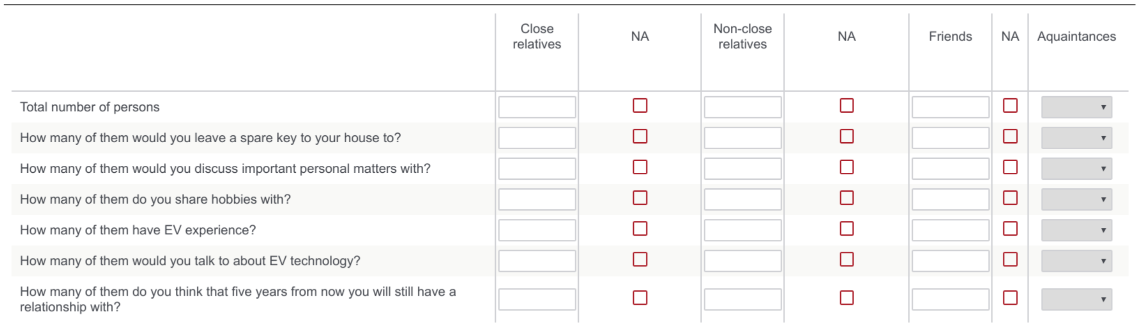
On the other hand, the social information provided prior to the experiment is one of the principal elements of this work. It is necessary in order to define the size and structure of the social network of the respondent. This information asked corresponds to four different groups:

- **Close relatives**: Parents, spouse, sibling, and children. It also includes co-habitants in the household.

- **Non-close relatives**: People of the family like grandparents, uncles and aunts, cousins, nephews, etc.

- **Friends**: Someone one would tell a personal problem or to whom one would turn for help with an important issue.

- **Acquaintances**: Someone one does not know much about, and could not call a friend. One would not tell them about a personal problem or ask for help.



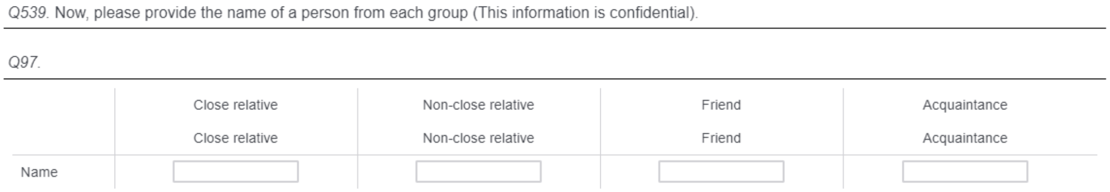
	Close relatives	NA	Non-close relatives	NA	Friends	NA	Aquaintances
Total number of persons		□		□		□	▾
How many of them would you leave a spare key to your house to?		□		□		□	▾
How many of them would you discuss important personal matters with?		□		□		□	▾
How many of them do you share hobbies with?		□		□		□	▾
How many of them have EV experience?		□		□		□	▾
How many of them would you talk to about EV technology?		□		□		□	▾
How many of them do you think that five years from now you will still have a relationship with?		□		□		□	▾

Figure 3.6: Social network information.

These questions aim to identify the size of the inner and outer circles, as well as their nature. Close and Non-close relatives, as well as Friends are considered to pertain to the Inner circle, while Acquaintances define the Outer circle. In order to reinforce the nature of these relationships, other questions are asked from which concepts such as trust and identification are inferred:

- *How many of them would you leave a spare key to your house to?*
- *How many of them would you discuss important personal matters with?*
- *How many of them do you share hobbies with?*
- *How many of them have EV experience?*
- *How many of them would you talk to about EV technology?*
- *How many of them do you think that five years from now you will still have relationship with?*

After providing this quantitative information, the user is asked to provide the name of one person pertaining to each group. This is an innovative feature of this survey. One of these names will be randomly chosen to either provide feedback later in the choice tasks, or to manifest hypothetical approval of the purchase of an EV. The details are described below, in the *Choice Tasks* section.



Q539. Now, please provide the name of a person from each group (This information is confidential).

Q97.

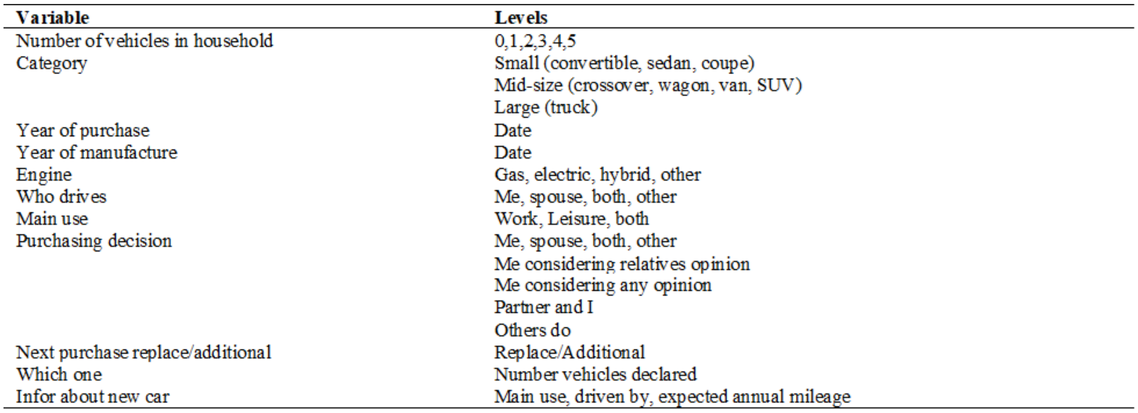
	Close relative	Non-close relative	Friend	Acquaintance
	Close relative	Non-close relative	Friend	Acquaintance
Name				

Figure 3.7: Name one person from each social group.

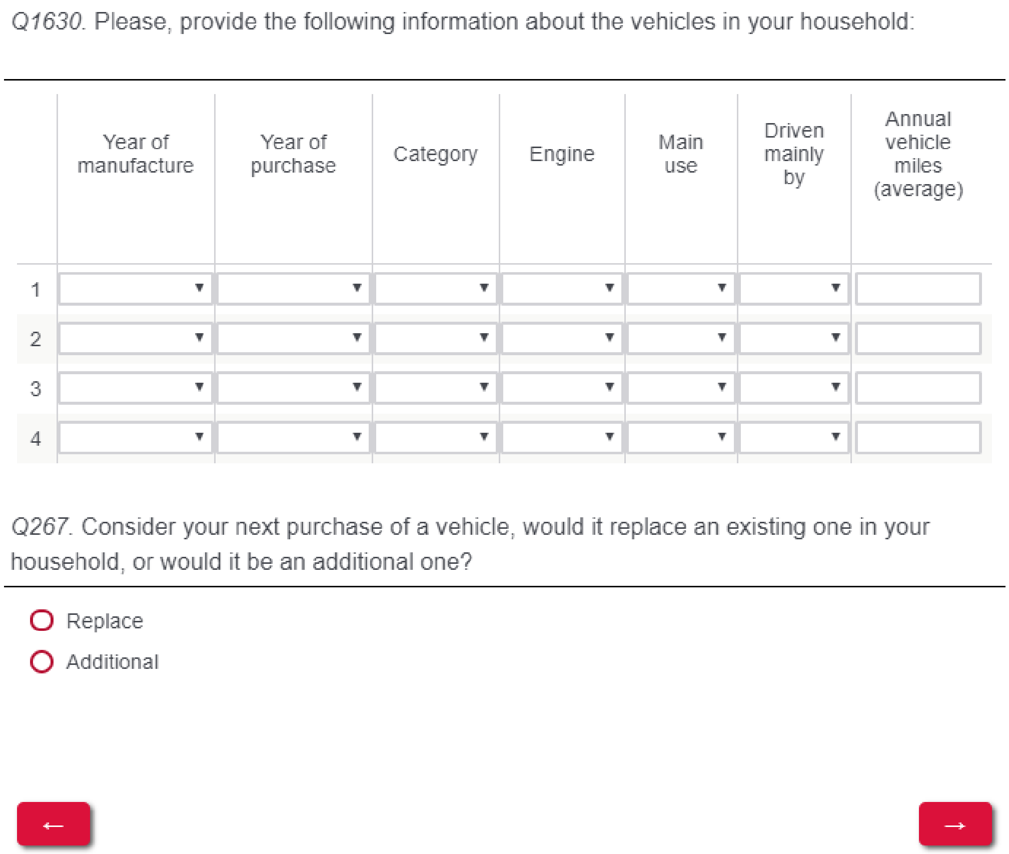
Car ownership

This set of questions aim to identify the vehicles owned in the household, and if the next purchase will be an additional one or will be replacing one of them. The purpose is to understand whether the potential EV choices made later in the scenarios would serve to rotate the vehicle structure existing in the household. Information about the new vehicle is also asked since the propulsion cost shown in the choice tasks are calculated using the mileage introduced here. Table 3.8 shows the information extracted from the questions, as well as their levels. Although it is a comprehensive list, the information is gathered in an easy-to-fill matrix, as Figure 3.8 shows.

Table 3.8: Car ownership variables and their levels.



Variable	Levels
Number of vehicles in household	0,1,2,3,4,5
Category	Small (convertible, sedan, coupe) Mid-size (crossover, wagon, van, SUV) Large (truck)
Year of purchase	Date
Year of manufacture	Date
Engine	Gas, electric, hybrid, other
Who drives	Me, spouse, both, other
Main use	Work, Leisure, both
Purchasing decision	Me, spouse, both, other Me considering relatives opinion Me considering any opinion Partner and I Others do
Next purchase replace/additional	Replace/Additional
Which one	Number vehicles declared
Infor about new car	Main use, driven by, expected annual mileage



Q1630. Please, provide the following information about the vehicles in your household:

	Year of manufacture	Year of purchase	Category	Engine	Main use	Driven mainly by	Annual vehicle miles (average)
1	▼	▼	▼	▼	▼	▼	
2	▼	▼	▼	▼	▼	▼	
3	▼	▼	▼	▼	▼	▼	
4	▼	▼	▼	▼	▼	▼	

Q267. Consider your next purchase of a vehicle, would it replace an existing one in your household, or would it be an additional one?

○ Replace

○ Additional

Figure 3.8: Car ownership matrix.

Info about EV

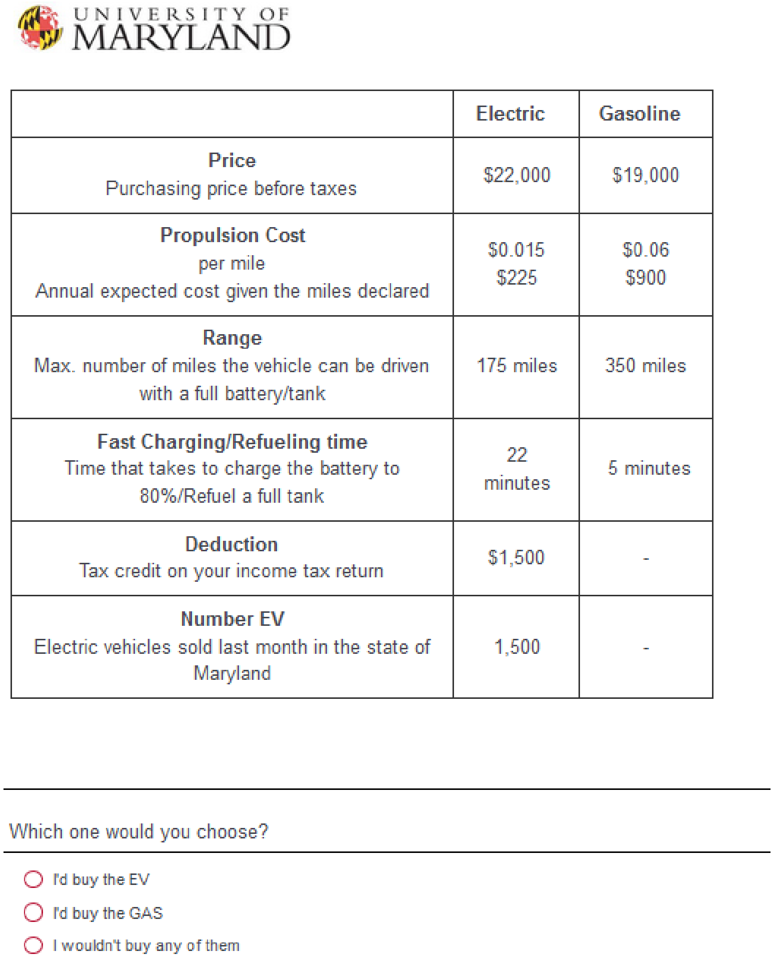
This is the section immediately before the choice tasks. It provides general information about different aspects of EVs, in order to give to the respondent a general overview on which to make an informed decision. It comes from official sources (the U.S. Department of Energy, and the Office of Energy Efficiency & Renewable Energy) and includes energy efficiency, environmental friendliness, performance benefits, energy dependence, driving range, recharging time, infrastructure availability, and tax credit. They all are directly or indirectly related to the information that will be provided in the choice tasks. Special care has been taken in the figures provided here since they cannot conflict with the data in the scenarios. It has been tried as much as possible to provide useful information without giving concrete figures, to prevent these numbers from remaining in the users' mind when they move on to the next section and give their answers.

Two elements to know if the interviewee made informed choices have also been included. It is first asked if they think that the information provided was enough to have a general idea of what an EV is and how the technology works. It is also registered if the user clicked the links provided to expand the reading material. The combination of these two elements will provide interesting insights.

Choice tasks

The choice tasks, together with the section referring to the social network, are the most important element of this study. Basically, a decision-making task consists, as Figure 3.9 shows, of providing a series of characteristics for several alternatives that the users evaluate, and which lead them to choose one of these alternatives. In this case, after the block with information about the EV, a user is redirected to the survey branch corresponding to the vehicle category selected previously. Then, a scenario in which the specific level of attributes generated in the mathematical design is shown. The respondent evaluates the information and makes a choice. This process is repeated six times, facing a different choice task each time. The alternatives between which to decide are Gasoline, EV and the opt-out Other. By EV it is meant Battery Electric Vehicles (BEV). Although hybrids are not considered explicitly as an alternative, if the respondent selects Other, two additional questions are asked to know whether she or he would purchase a vehicle with a different engine (hybrid, natural gas. . .).

In terms of design, four different blocks containing six different scenarios each have been created. When the users move on from the Info EV section to the choice tasks, they are randomly assigned to one of these blocks. Therefore, 4 categories × 6 tasks × 3 categories yield 72 different scenarios designed, similiar the one shown in Figure 3.9.



UNIVERSITY OF MARYLAND

	Electric	Gasoline
Price Purchasing price before taxes	$22,000	$19,000
Propulsion Cost per mile Annual expected cost given the miles declared	$0.015 $225	$0.06 $900
Range Max. number of miles the vehicle can be driven with a full battery/tank	175 miles	350 miles
Fast Charging/Refueling time Time that takes to charge the battery to 80%/Refuel a full tank	22 minutes	5 minutes
Deduction Tax credit on your income tax return	$1,500	-
Number EV Electric vehicles sold last month in the state of Maryland	1,500	-

Which one would you choose?

○ I'd buy the EV
○ I'd buy the GAS
○ I wouldn't buy any of them

Figure 3.9: Choice task for the Compact category.

Now, we must not forget, that the main objective of this study is to evaluate the influence of the social aspects involved in the decision to purchase an EV. Specifically, the effect of Informational and Injunctive conformity that proceeds from people of the inner and outer circles. Therefore, it is necessary to incorporate somehow these elements explicitly in the scenarios. Much effort has been devoted to this task, deeply analyzing the limited literature in the field and trying, in turn, to improve it. After several approaches, the following methodology was defined.

After the six choice tasks, the respondent answers again three of them (not knowing that they are repeated scenarios), chosen randomly. However, a new piece of information is provided along with the level of the attributes; a phrase that expresses one of the following:

- <u>Feedback about charging aspects</u>: Positive or negative information from someone of your inner or outer circle regarding the necessity of charging the EV.

  Peter thinks that having to watch out for charging your EV causes almost no concern.

- <u>Feedback about changing activities</u>: Positive or negative information from someone of your inner or outer circle regarding the necessity of modifying your schedule due to driving an EV.

  Peter thinks that having to change your activities because of driving an EV is annoying.

- <u>Approval</u>: Approval or disapproval of your choice of an EV from someone of the inner or outer circle.

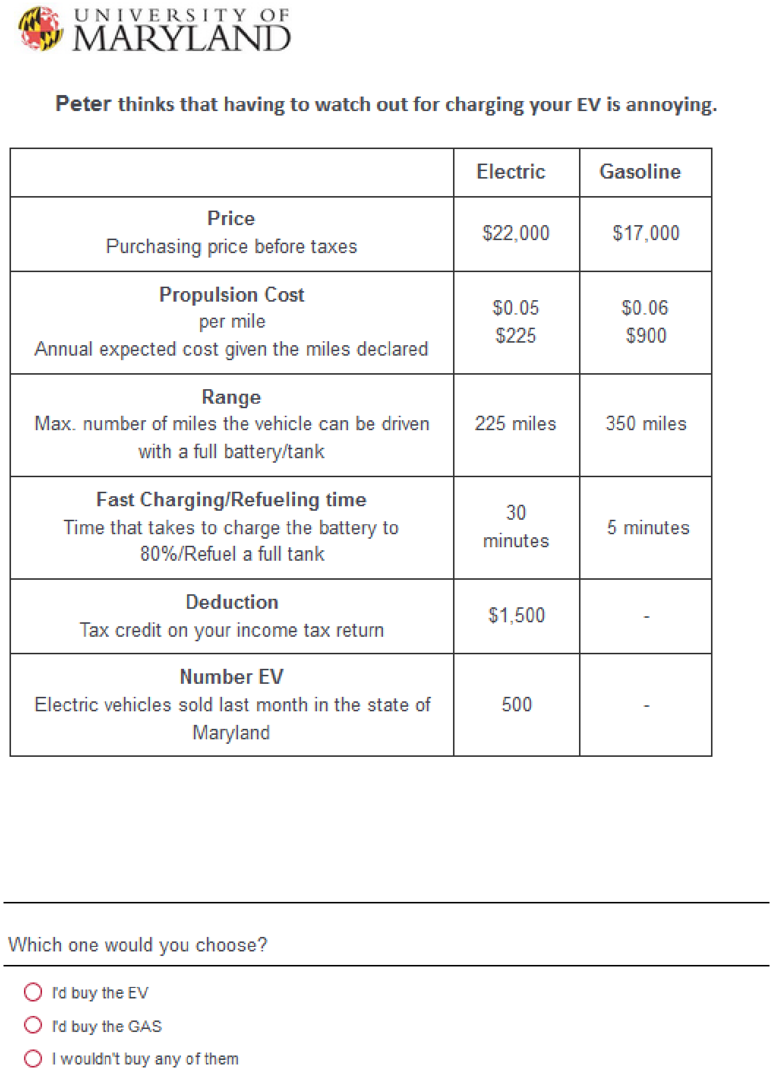
Peter does not approve of you using an EV.

The name appearing in the sentence –Peter in this case, for illustration purposes– is taken from the preliminary question in which the name of a person belonging to each social group was asked. Remember that Close, Non-close and Friends form the Inner circle while the Acquaintances form the outer circle. In survey flow terms, a combination of inner/outer, positive/negative, and type of feedback is randomly assigned, with equal probability, to the user.

The purpose of this procedure is to check whether the user changes their choice by receiving feedback. The second three scenarios have already been answered, and the information that they show is the same as in the previous round. The only difference is the feedback. Type of feedback, if it is positive or negative, and if it comes from a person of the inner or outer circle will be categorical variables to be included in the model.

Finally, in order to have more information about the bond to the person providing the feedback, or to what extent this person is trusted, the respondent is asked to respond, in a Likert scale:

- *I trust this person, in general terms.*
- *I trust this person in matters related to vehicles.*
- *I am close to this person.*



UNIVERSITY OF MARYLAND

Peter thinks that having to watch out for charging your EV is annoying.

	Electric	Gasoline
Price Purchasing price before taxes	$22,000	$17,000
Propulsion Cost per mile Annual expected cost given the miles declared	$0.05 $225	$0.06 $900
Range Max. number of miles the vehicle can be driven with a full battery/tank	225 miles	350 miles
Fast Charging/Refueling time Time that takes to charge the battery to 80%/Refuel a full tank	30 minutes	5 minutes
Deduction Tax credit on your income tax return	$1,500	-
Number EV Electric vehicles sold last month in the state of Maryland	500	-

Which one would you choose?

○ I'd buy the EV

○ I'd buy the GAS

○ I wouldn't buy any of them

Figure 3.10: Choice task including approval of a person of one of the social circles.

Trips

This block of questions collects information about the trips made by the respondent, to have additional information about the possible use of the EV.

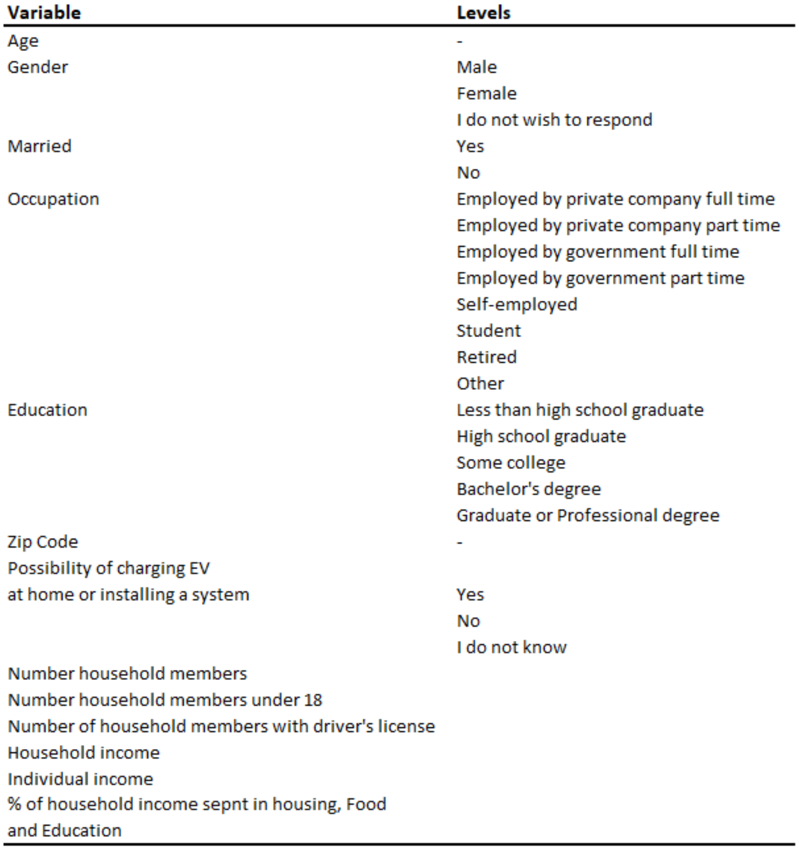
Car sharing

Three questions to know the patterns of use of car sharing and rideshare apps.

Sociodemographics and attitudes

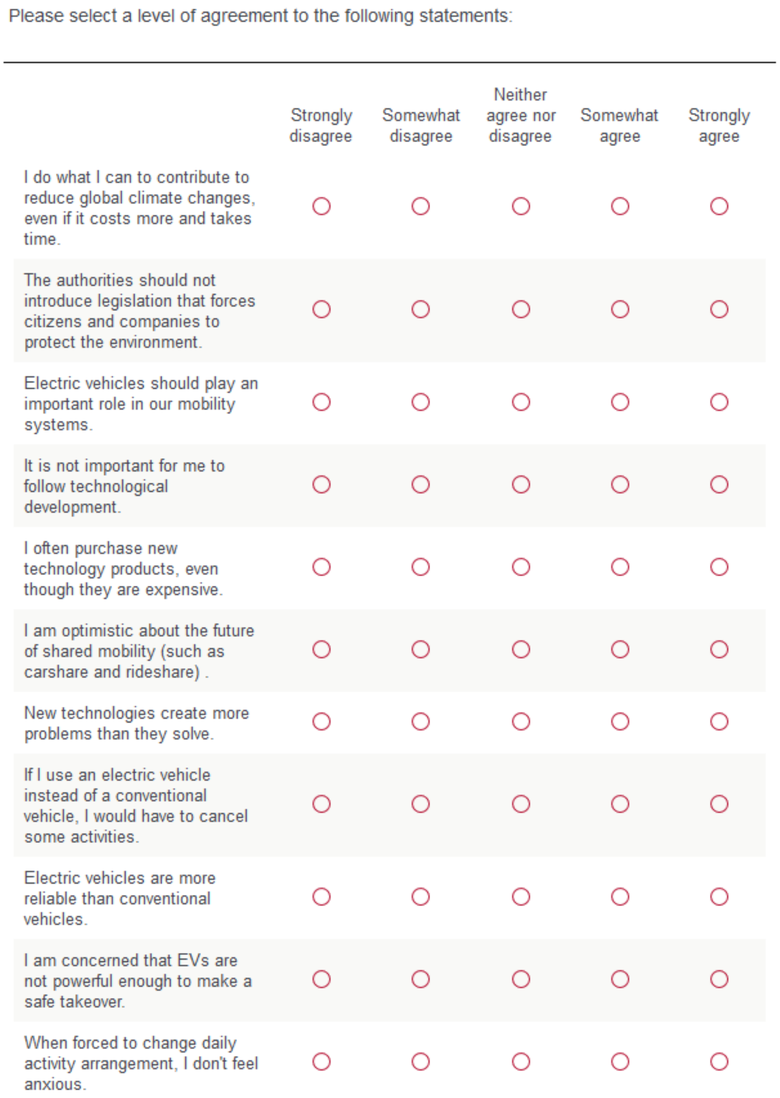
This section has two parts. The first gathers the general sociodemographic information shown in Table 3.9.

Table 3.9: Sociodemographic variables and their levels.



Variable	Levels
Age	-
Gender	Male
	Female
	I do not wish to respond
Married	Yes
	No
Occupation	Employed by private company full time
	Employed by private company part time
	Employed by government full time
	Employed by government part time
	Self-employed
	Student
	Retired
	Other
Education	Less than high school graduate
	High school graduate
	Some college
	Bachelor's degree
	Graduate or Professional degree
Zip Code	-
Possibility of charging EV at home or installing a system	Yes
	No
	I do not know
Number household members	
Number household members under 18	
Number of household members with driver's license	
Household income	
Individual income	
% of household income sepnt in housing, Food and Education	

The second part presents a question in which it is necessary to select a level of agreement to attitudinal statements. These statements pertain to three different categories (unknown to the respondent): Environmental concern, Technology innovator, and Pro-EV. The aim is to use this information to better understand the choices made by individuals. This question is shown in Figure 3.11.



Please select a level of agreement to the following statements:

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
I do what I can to contribute to reduce global climate changes, even if it costs more and takes time.	○	○	○	○	○
The authorities should not introduce legislation that forces citizens and companies to protect the environment.	○	○	○	○	○
Electric vehicles should play an important role in our mobility systems.	○	○	○	○	○
It is not important for me to follow technological development.	○	○	○	○	○
I often purchase new technology products, even though they are expensive.	○	○	○	○	○
I am optimistic about the future of shared mobility (such as carshare and rideshare) .	○	○	○	○	○
New technologies create more problems than they solve.	○	○	○	○	○
If I use an electric vehicle instead of a conventional vehicle, I would have to cancel some activities.	○	○	○	○	○
Electric vehicles are more reliable than conventional vehicles.	○	○	○	○	○
I am concerned that EVs are not powerful enough to make a safe takeover.	○	○	○	○	○
When forced to change daily activity arrangement, I don't feel anxious.	○	○	○	○	○

Figure 3.11: Attitudinal questions.

Contact info

Finally, there is the possibility to provide contact information for a possible follow-up.

3.2.3 Data Analysis

The data collection for this study has been made in three phases: *Pre-pilot, Pilot, and Release*. The purpose of the Pre-pilot (5% of the sample size) was to identify issues in the questionnaire. This was certainly a useful exercise, especially to test the functioning of the Social Network question. This question contained a significant number of cross-validations and, although tests were performed before publishing the survey, user behavior when filling it helped to identify cases that still had to be coded. On the other hand, the aim of the Pilot (50% of the final sample size) was to perform preliminary estimations of the models described in section 3.2.4. The Release data (100% of the sample size) has been used to estimate the final version of the model, on which the forecast will be carried out.

A descriptive analysis of the data has been developed and is presented in this section. First, information about the vehicle ownership of each household is shown. Then, several graphs help in understanding the structure of the social network of the interviewees. Finally, the data concerning their attitudes towards several aspects is plotted. It is worth mentioning that the analysis below is not comprehensive since approximately 200 variables were gathered in the survey. Plotting such amount of information would make the reader to lose focus on the information presented and would not convey a direct message. Therefore, only the findings that have been considered most relevant are shown. Additionally, the following conventions will be followed for the sake of brevity; households with 1, 2, 3, 4, and 5+ vehicles will be referred to as HH1, HH2, HH3, HH4, and HH5, and this dimension will be generically named *household size*.

3.2.3.1 Vehicle ownership

This section presents an analysis of the household vehicle composition, although HH5 statistics are not presented in Figures 3.14 - 3.17 since the number of vehicles in some of this households is actually very high. That would have made representation difficult, and they actually represent just 2% of the sample.

Most of the family units, 88%, own 1 or 2 vehicles (Figure 3.12) that were bought recently, as shown by the increasing slope of the lines in Figure 3.13. In the case of HH1, 81.4% of them were purchased in the last 5 years. Moreover, this is a general trend except for HH4 and HH5, where more variability can be observed.

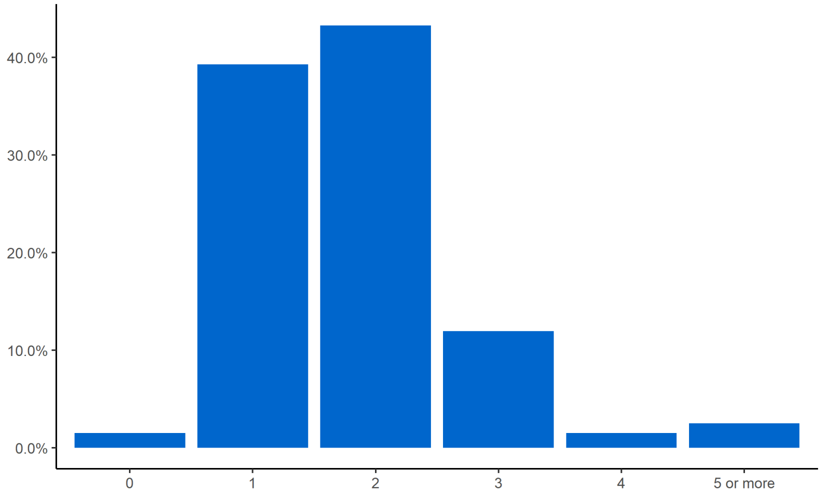


Figure 3.12: Number of vehicles.

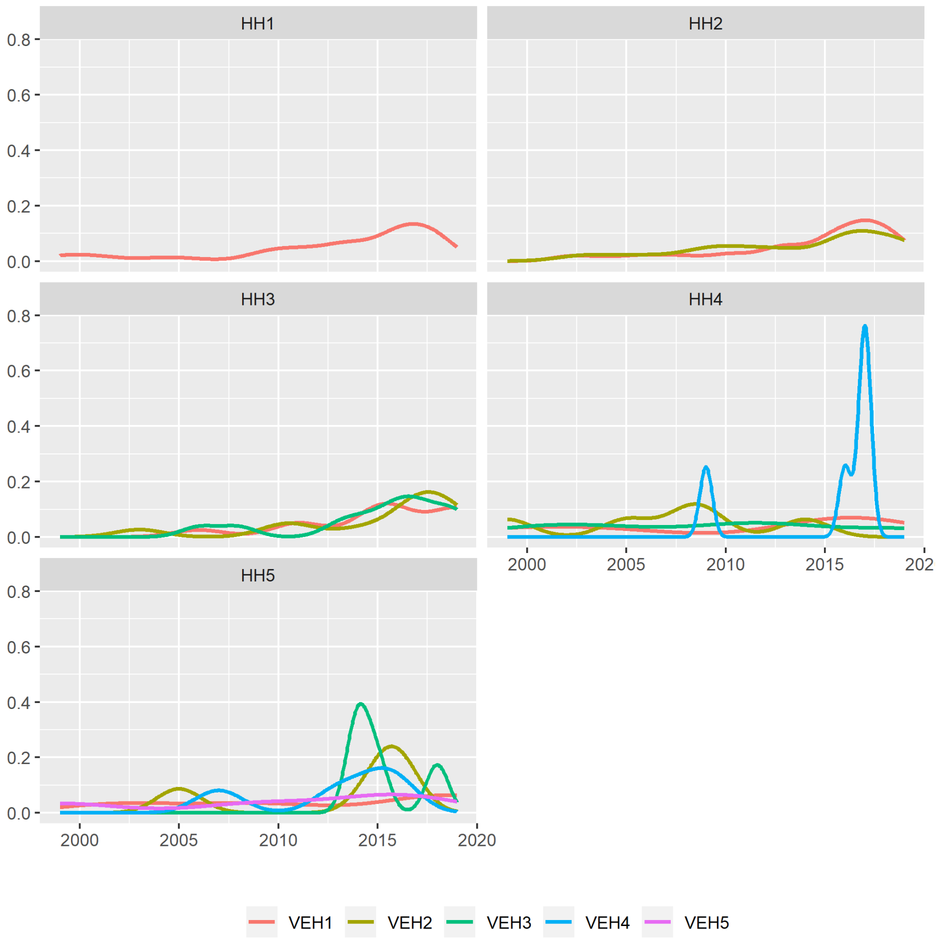


Figure 3.13: Year of purchase.

Regarding the vehicle category, Mid-Size seems to be the predominant class, followed by Large, no matter if the vehicle is the main, second, or third one (Figure 3.14). This pattern is not maintained in the households with four cars where the third is commonly Large and the fourth, Compact. On the other hand, almost all the vehicles reported were Gasoline, with reduced percentages for Hybrid and residuals for Electric.

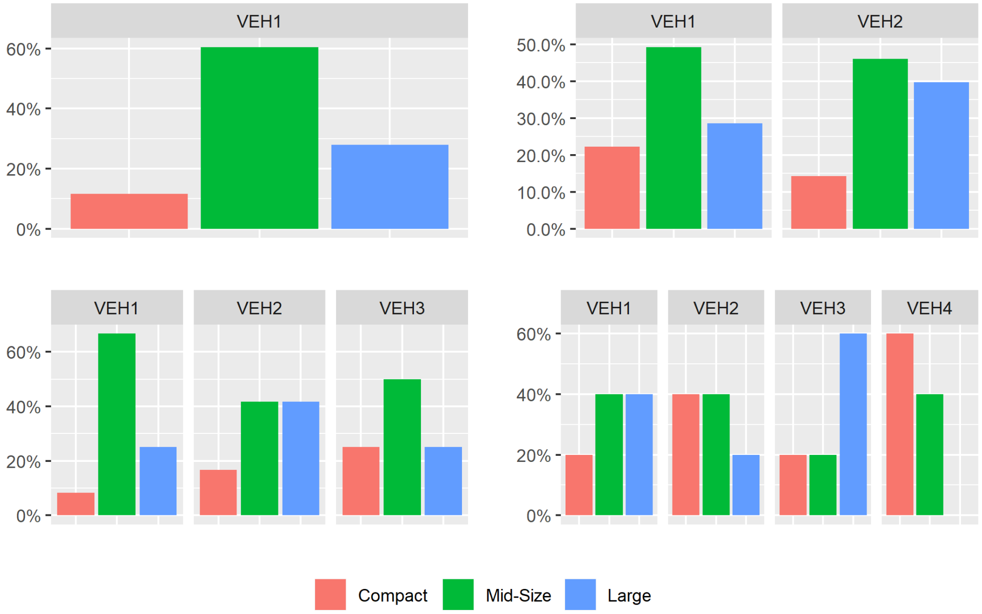


Figure 3.14: Vehicle category by household size.

Curiously, the person that mainly drives the first vehicle in the house is who takes the survey. The second vehicle is predominantly used by the spouse, and the third by the spouse, both, or other person (Figure 3.15). Figure 3.16 shows that the purpose of this driving is predominantly both work and leisure in all cases, except in HH4, which is eminently work.

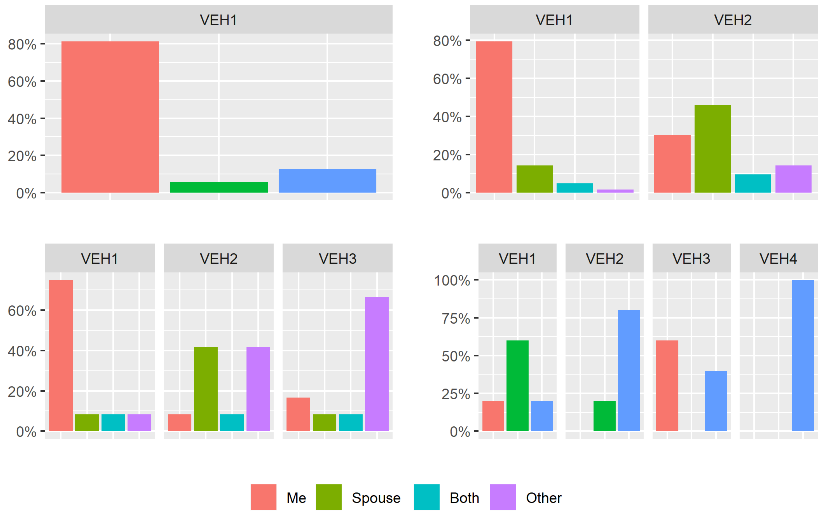


Figure 3.15: Who uses the vehicle by household size.

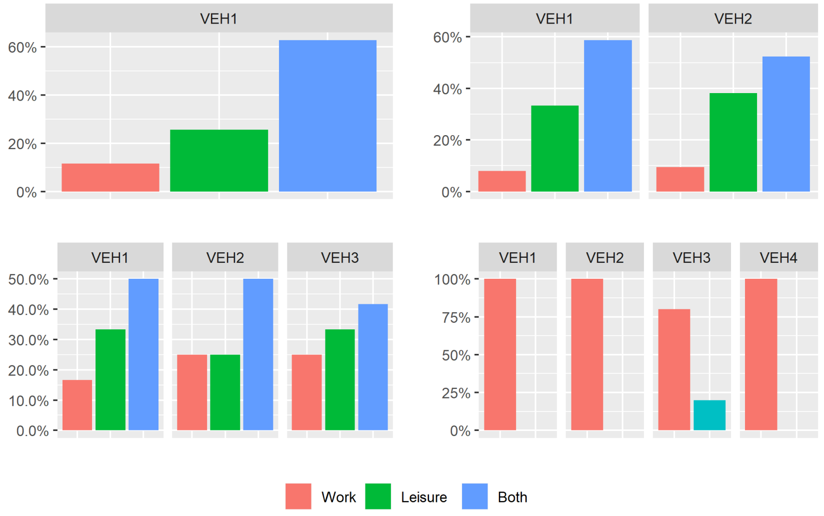


Figure 3.16: Vehicle main use by household size.

Figure 3.17 shows the average annual mileage of all vehicles in a household, by household size. The median is similar across households and vehicles, although the interquartile range is wider for HH1. The average annual mileage is between 17,000 and 36,000.

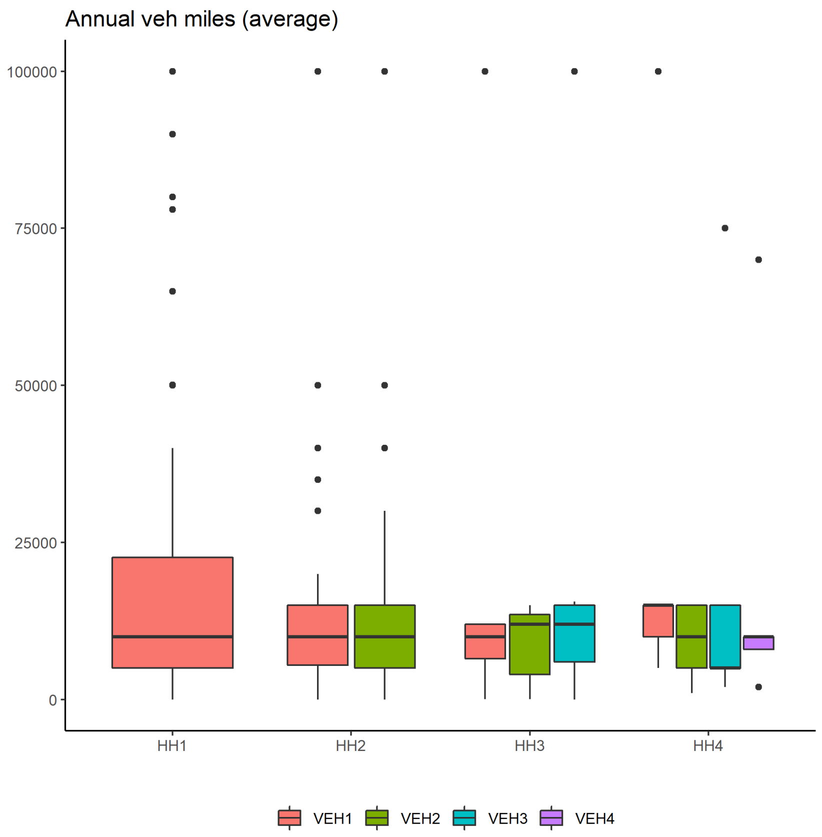


Figure 3.17: Annual vehicle mileage (average).

This section of the questionnaire also asks for information about the next most likely purchase of a vehicle. As can be seen in Figure 3.18, almost 60% of respondents declared that they would buy a Mid-Size car. In addition, Gasoline is the most common engine, to be driven mainly by the survey respondent, and the main purpose would be both Work and Leisure. This new vehicle would replace an existing one in the household in 80% of the cases.

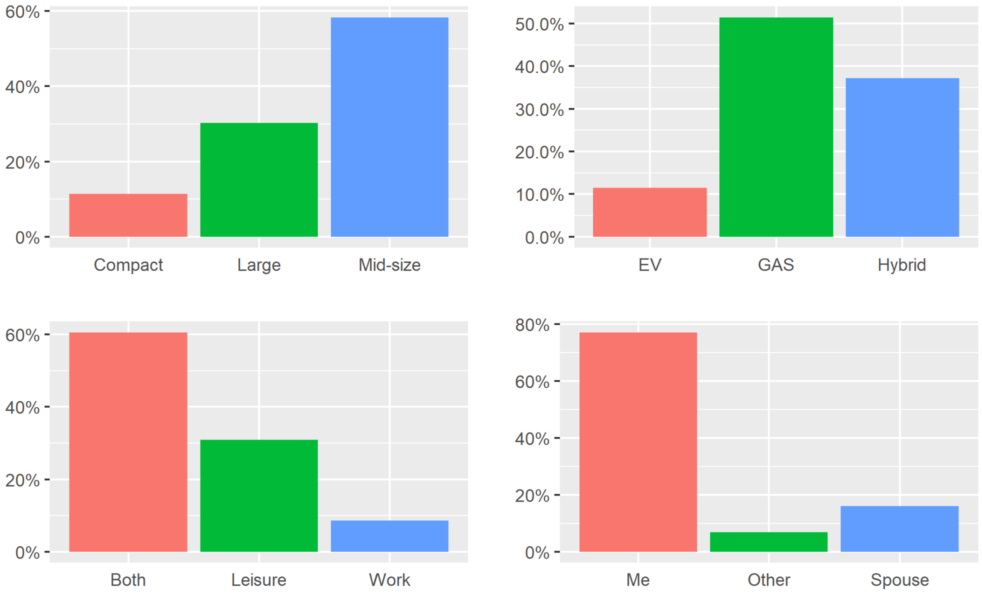


Figure 3.18: Next vehicle characteristics.

The last question of this section of the questionnaire was who makes the purchasing decision. In about 50% of the cases the respondent is this person, while in 30% of the cases are both the respondent and their partner. These shares are followed by about 15% of interviewees who responded *I do, although I take into account the opinion of my relatives.* Surprisingly, the percentage of users that stated *Other do* is superior than *I do, although I take into account any opinion that comes to my ears.*

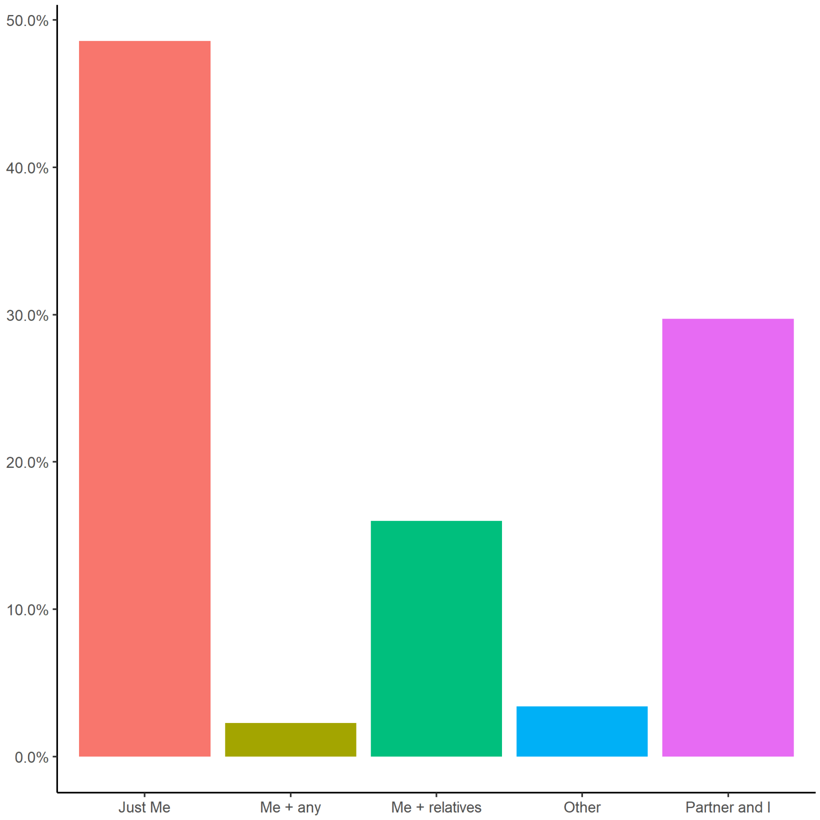


Figure 3.19: who makes the purchasing decision.

3.2.3.2 Social network

The question referring to the social network has several particularities. Since its purpose is to identify the size and the structure of the people connected to the respondent, it was necessary to define many specific aspects that led to certain complexity in the design. It is worth remembering that the four different groups are: *Close relatives (CR), Non-Close relatives(NCR), Friends (FR), and Acquaintances (AQ)*; the first three forming the Inner circle, and the fourth, the Outer circle.

The user always had the option of responding *NA* since they may not know the number of persons pertaining to a group. These values were imputed following the procedure described in section 4.2.1. Figure 3.20 shows the differences in the spread of the values, which are clearly skewed right.

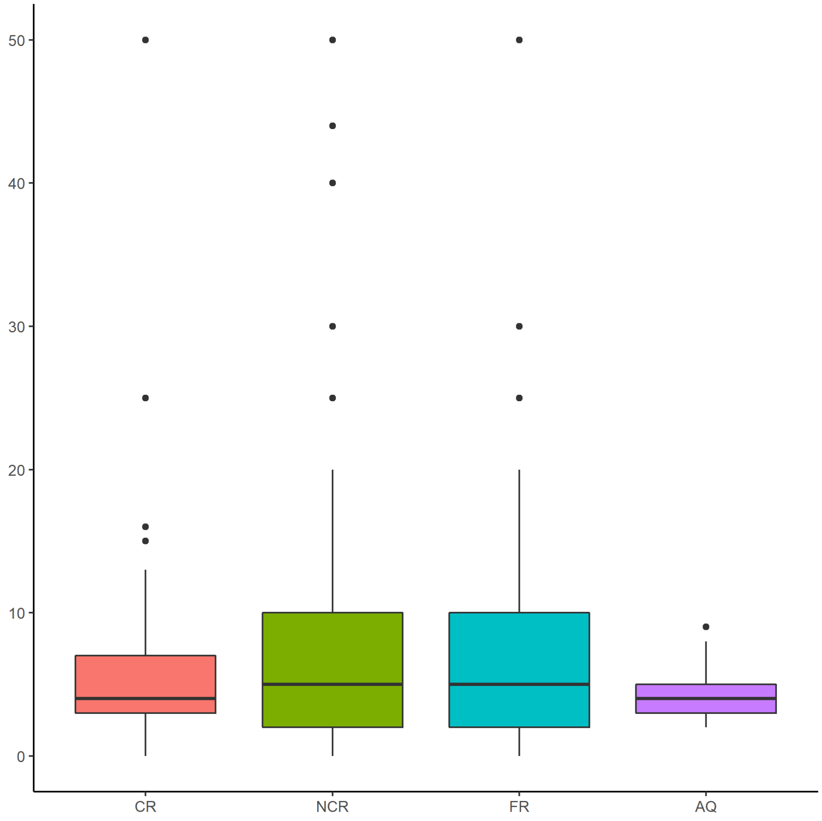


Figure 3.20: Number of Close relatives, Non-Close relatives, and Friends.

The average number of each group is; close relatives, 2; non-close relatives, 3; friends, 12; and acquaintances 4. It is surprising the reduced number of acquaintances declared, a consistent fact among the pilots and the final survey. It was presumed that the number of them would be large since the number of co-workers, fellow hobbyists, gym mates, neighbors or any other person of such category should be higher than that of friends and relatives.

In addition to the group size, it was of special interest to identify the nature of the groups. To do so, the respondent was asked to introduce the number of persons (from those declared before) in questions regarding several aspects that denote trust or identification (see 3.2.2.3 for the details). In general, the distribution is skewed right, with a few large values (see Figure 3.21). This means responses for which the size of the subgroup is equal to the number of the whole group. In other words, users that declare that they would leave a spare key to all their friends, for instance. For the "key" and the "matters discussion" questions, the mean of the group of close relatives is higher than the one of the non-close, which is higher than the mean of both the friends and acquaintances groups; an obvious behavior. Values are high for the "5 years" question, evidencing certain optimism of individuals regarding the future of their social relations. For the questions related to EVs, the average number of people with experience driving them is low, as expected.

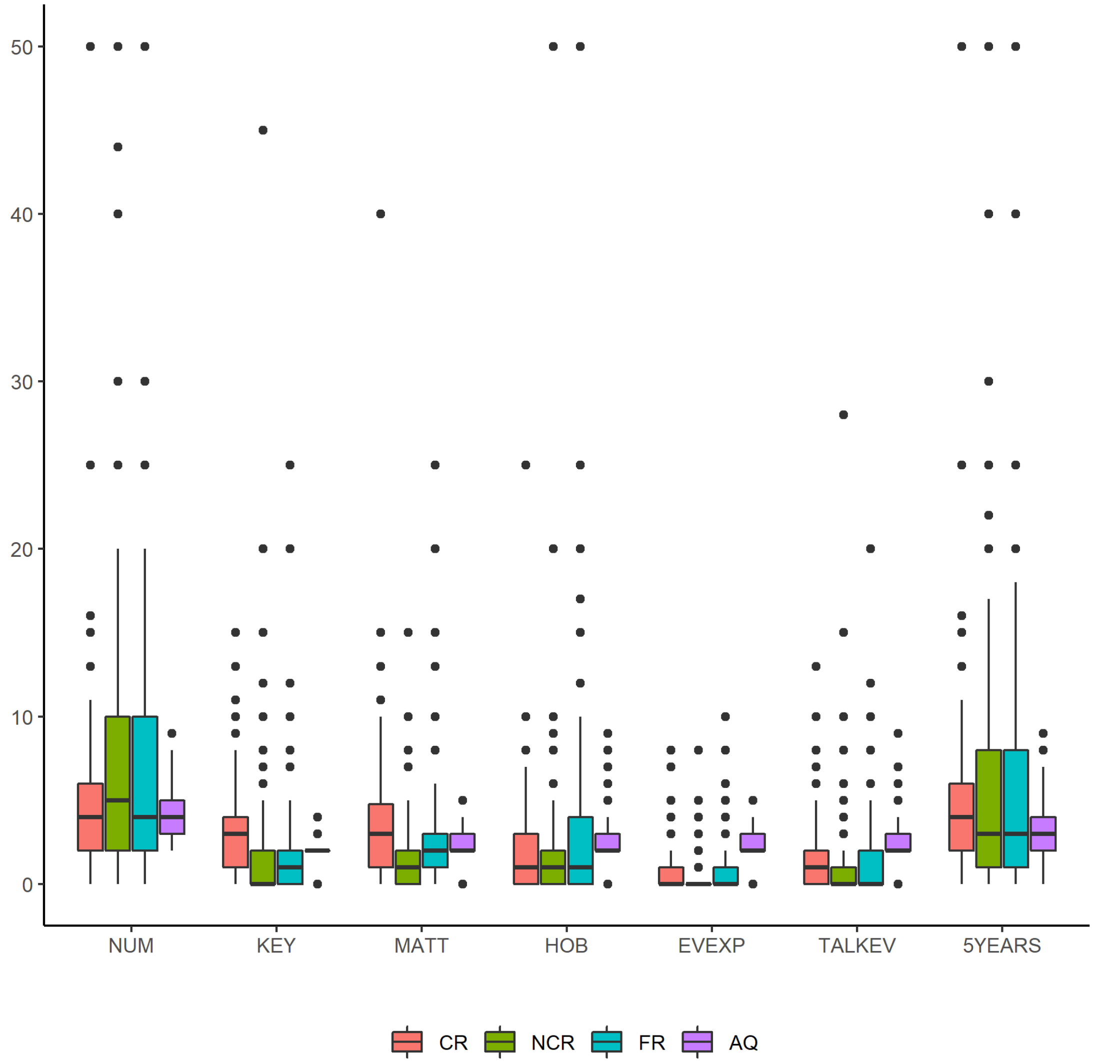


Figure 3.21: Number of people for each specific Social Network question.

With respect to the average frequency of contact with the members of each group (left plot of Figure 3.22), in the case of CR *Once a week* is surprisingly more common than *Every day*. The figure also shows that there is more contact with friends than with Non-Close relatives. The frequency for the group of acquaintances is more disperse, which is natural since it is gathering people with whom one has very different kinds of relationships. One can see one's co-workers every day, a gym partner a couple of times a week, and the mechanic of the car workshop a couple of times a year.

More important, however, is the frequency of contact with the person indicated in the question (again, see section 3.2.2.3 for more detail), shown in the right plot of the figure. Interestingly, the behavior seems to be opposite than for the general subgroups. Users declare little frequency of contact with the person indicated as close relative, and more frequency with friends and non-close relatives. This is probably because interviewees introduced the name of parents and siblings, which sometimes one see less than friends that live in the same neighborhood or co-workers.

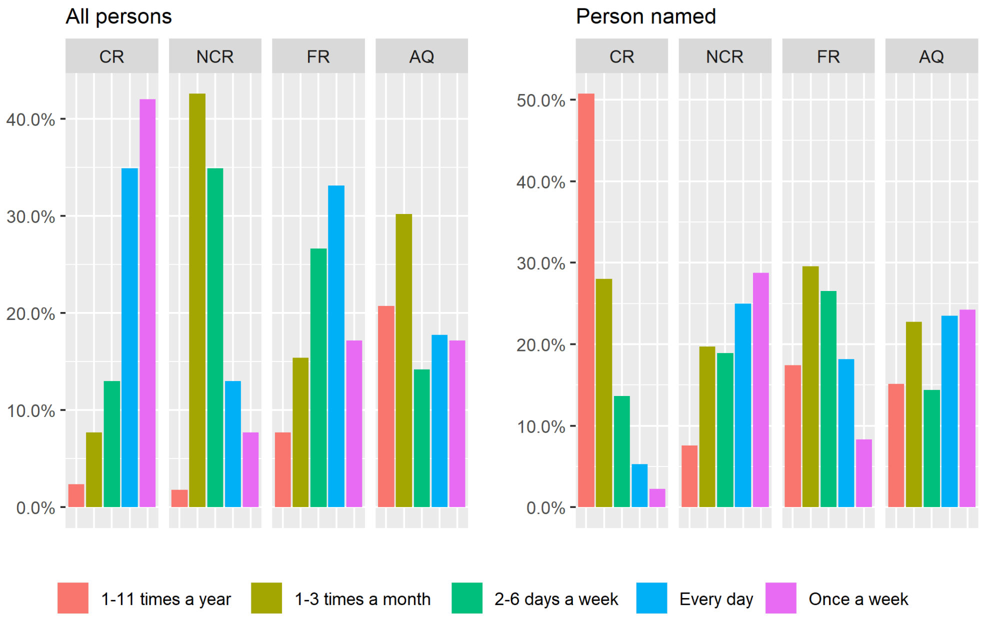


Figure 3.22: Frequency of contact by social group.

3.2.3.3 Sociodemographics and Attitudes

Tables B.2 and B.3 in Appendix B shows some sociodemographic characteristics of the sample. It can be seen that while the sample is well distributed in terms of age, women are slightly over-represented. *Private full-time* is the most common employment status as well as *Bachelor's* and *Some college* are the most common educational degree. In terms of income, it seems that this sample is distributed as expected, with an average individual income of around $54,000 and a household income of around $86,000. One interesting aspect that is asked in the survey is the percentage of the household income allocated to housing, healthcare, insurance, food and education expenses. This figure provides a proxy of the disposable income to spend in transportation. The mean of this share is 60.7%. Interestingly, the share of individuals that do not know if they are able to charge an EV at home (or able to install a system to do so) is high, almost a third of the sample. This may indicate a lack of knowledge about this technology, even though information about it had been provided earlier in the survey , and no respondent answered *No* to the question *Do you think the information provided in the reading material is sufficient for you to get a general idea of what an EV is and how the technology works?*.

Besides the socioeconomic information, an important part of this subsection was the identification of the attitudes and perceptions of the respondents. Although the details can be consulted in 3.2.2.3, it is worth to remember that the user selected a level of agreement to sentences related to their environmental concerns, technology adoption, and inclination towards EV. A value from -2 to 2 was assigned to each response (since they were expressed in a Likert scale) and then added up for each category in order to compute a representative score. Their distributions are plotted in Figure 3.23, where the dashed lines indicate the average value.

The *Technology Inclined* attitude scores the highest on average, above 4 out of a maximum of 10 (and a minimum of -10), meaning that respondents might be early adopters, or at least, that they have interest in new technologies. *Pro-EV* and *Environmental Concern* have a lower average score. In addition, the three distributions are reasonably symmetric, with the Pro-EV one having long tails, representing more dispersion of the sample in this matter. Therefore, it is possible to conclude that this are individuals inclined to technology, with no special interest in the environment, and equally in favor and against EVs.

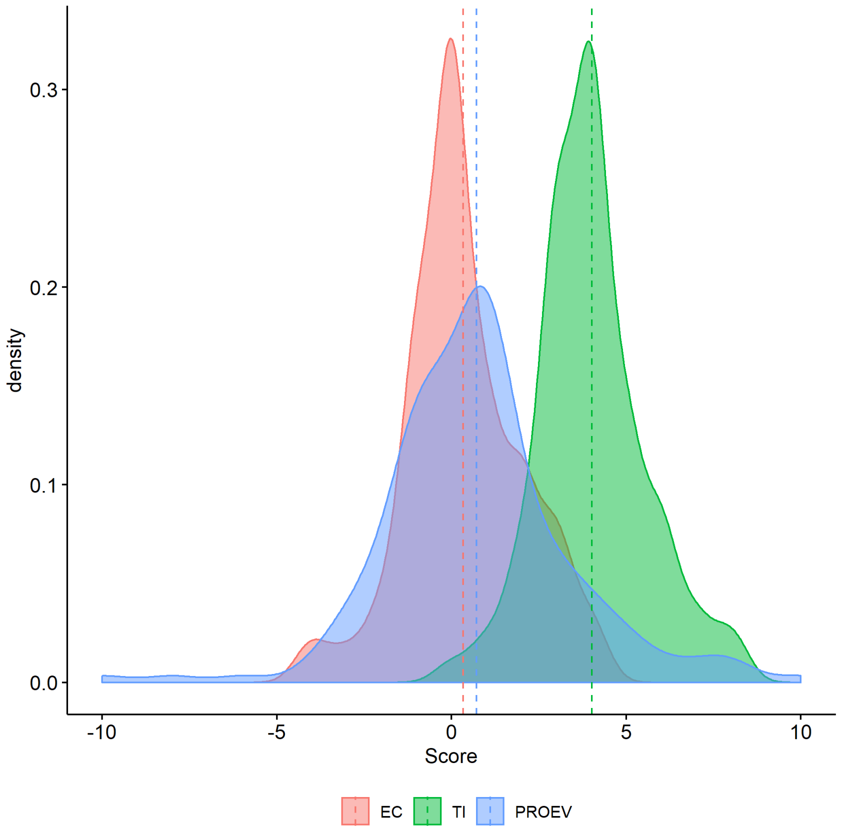


Figure 3.23: Attitudes score distribution.

3.2.4 Model structure

The methodology for the application of the substitution-diffusion model comprises of two steps. First, the estimation of a disaggregated DCM that includes the attributes of the alternatives and the social components. Second, the use of the resulting estimated parameters to construct a new utility that will feed the Bass model, which will yield the diffusion and scale coefficients. This section explains this process.

With respect to the first of these steps, the estimation of the DCM is similar to that implemented in the Danish case. However, it is necessary to make modifications to include the conceptual improvements related above, as well as to incorporate the particularities that affect the case of the state of Maryland, eliminating those of Denmark. The utility functions that underly the DCM can be expressed in compact form as:

$$V = ASC + (\beta X + \alpha SC^{EV}) \tag{3.9}$$

where X is the vector of attributes shown in the choice tasks, and β its corresponding coefficients. On the other hand, SC is a function that aggregates the Social Conformity indicators shown in the second round of scenarios that the respondent faces. Since the goal is to measure the effect of the SC elements in the choice of the EV, these indicators are only present in its utility.

Therefore, the part of the function related to the EV attributes takes the form:

$$\begin{split}\beta X^{EV} = \beta_1 Price + \beta_2 PropulsionCost + \beta_3 Range + \beta_4 FastCharTime +\\ + \beta_5 TaxDeduction\end{split} \tag{3.10}$$

while in the Gasoline case:

$$\beta X^{Gas} = \beta_1 Price + \beta_2 PropulsionCost + \beta_3 Range \tag{3.11}$$

On the other hand, the part related to the SC takes the form:

$$\begin{aligned}\alpha SC = \alpha_1 NumberEV + \alpha_2 InnSize + \alpha_3 OutSize + \\ + ((\alpha_4 D_{Char} + \alpha_5 D_{Act} + \alpha_6 D_{App})(D_{Inn} + D_{Out}))D_{Rep} \quad (3.12)\end{aligned}$$

where $NumberEV$ corresponds to the attribute described in section 3.2.2.2, $InnSize$ and $OutSize$ are the sizes of the inner (close relatives, Non-close relatives, and Friends) and outer (Acquaintances) groups, and $D_{Char}, D_{Act}, D_{App}, D_{App} and D_{Out}$ are dummy variables that represent, respectively: whether the information provided was about charging the EV, the need to change activities, if it conveyed a sentence of approval, and whether the information was provided by a member of the Inner or Outer circle. Whatever the feedback, it is only important in the three repeated scenarios as only in those it was provided. This is represented by the dummy variable D^{Rep}.

Therefore, the complete utilities can be expressed as:

$$\begin{aligned} V^{EV} &= ASC + \beta_1 Price + \beta_2 PropulsionCost + \beta_3 Range + \\ &+ \beta_4 FastCharTime + \beta_5 TaxDeduction + \\ &+ \alpha_1 NumberEV + \alpha_2 InnSize + \alpha_3 OutSize + \\ &+ ((\alpha_4 D_{Char} + \alpha_5 D_{Act} + \alpha_6 D_{App})(D_{Inn} + D_{Out}))D_{Rep} \quad (3.13)\end{aligned}$$

$$V^{Gas} = \beta_1 Price + \beta_2 PropulsionCost + \beta_3 Range \quad (3.14)$$

where the parameters to be estimated are ASC, β_i, and α_i following a Mixed Multinomial Logit (MML) specification. MMLs are flexible models that circumvent the limitations of the MNL, allowing random variation of preferences, substitution patterns, and correlation between unobserved factors. In the case of the MMNL, the probabilities are different from those indicated in Equation (3.5), and are given by:

$$P_{ni} = \int L_{ni}(\beta) f(\beta) d\beta \tag{3.15}$$

where $L_{ni}(\beta)$ is the probability evaluated in β:

$$L_{ni}(\beta) = \frac{e^{V_{ni}(\beta)}}{\sum_{j=1}^{J} e^{V_{nj}(\beta)}} \tag{3.16}$$

$f(\beta)$ is a density function and $V_{nj}(\beta)$ the observable part of the utility. Therefore, the MMNL probability is a weighted average of the logit formula, evaluated for different values of β, where the weights are given by the density $f(\beta)$. This density may be, for instance, a Normal distribution of mean b and covariance W. Then, the choice probability becomes:

$$P_{ni} = \int \left(\frac{e^{\beta' x_{ni}}}{\sum_{j=1}^{J} e^{\beta' x_{ni}}} \right) \phi(\beta \vee b, W) d\beta \tag{3.17}$$

Other distributions as Log-normal, Uniform, Triangular, or Gamma may be used in the same way. Log-normal is useful when it is known that the coefficient has the same sign for all decision makers, as is the case of the coefficient of cost, which is negative for all individuals. In most of the existent cases of use of this model ([111]; [96], [9]), $f(\beta)$ has been specified as Normal or Log-normal. Nevertheless, [107], [50], and [136] have experimented with triangular and uniform distributions.

Another possibility is to use the MMNL in an interpretation of error components that creates correlations between the utilities of different alternatives. In this case, the utility is specified as:

$$U_{nj} = \alpha' x_{nj} + \mu_n^{'} z_{nj} + \epsilon_{nj} \tag{3.18}$$

where x_{nj} and z_{nj} are vectors containing the observed variables of alternative j, α is a vector of fixed coefficients, μ is a vector of random coefficients with zero mean, and ϵ_{nj} is extreme iid. The z_{nj} terms are error components that, along with ϵ_{nj}, define the stochastic part of the utility.

Another advantage of the MML models is the possibility of using panel data. This specification also allows to consider repeated choices of the same individual. The simplest way to do this is through coefficients that enter the utility as parameters that vary between individuals but are constant between choice situations. The utility of alternative j in a choice situation t by person n is $U_{njt} = \beta_n x_{njt} + \epsilon_{njt}$, with ϵ_{njt} extreme iid over time, people and alternatives. Considering a sequence of alternatives over time, the probability that an individual will make that sequence of choices, conditioned to β, is:

$$L_{ni}(\beta) = \prod_{t=1}^{T} \left[\frac{e^{\beta_n^{'} x_{nit}}}{\sum_{j=1}^{J} e^{'_n x_{nit}}} \right] \tag{3.19}$$

The second part of the modeling phase is the integration of the choice model into the diffusion model. The procedure is similar to the one described in section 3.1.2. According to Equation 3.4, the probability obtained from the choice model is used to compute the sales of each time period, considering the potential market share at time t, M_t, and the cumulative sales at time $t-1$, Y_{t-1}. Therefore, the number of sales in each period, S_t, is, on the basis of the general MML panel data specification:

$$\begin{aligned} S_t &= (M_t - Y_{t-1}) \cdot P_t \\ &= (M_t - Y_{t-1}) \cdot \frac{e^{\beta'_n x_{nit}}}{\sum_{j=1}^{J} e^{\beta'_n x_{nit}}} \end{aligned} \tag{3.20}$$

Finally, for a complete description of the model, joining Equations (3.10), (3.13), and (3.20), and following the same approximation as in the Danish case (Equation (3.5)), the diffusion over time will be given, expressed in reduced form, by:

$$S_t = (M_t - Y_{t-1}) \cdot \frac{exp(q(t-\tau+1)+\lambda V^{EV})}{\lambda V^{Gas} + exp(q(t-\tau+1)+\lambda V^{EV}) + exp(\widehat{ASC^{None}})} \tag{3.21}$$

Note that the elements contained in V^j, $(\widehat{\beta}X + \widehat{\alpha}SC)$ are directly calculated form the data and the coefficients estimated in the choice model. Thus, the parameters to be estimated are λ and q, the substitution and diffusion parameters.

In order to test the validity of the model, a resampling technique known as *bootstrapping* has been applied. A number of subsamples containing a large proportion of the data are randomly generated, estimating a model for each of them. Then, the mean and standard deviation of the coefficients estimated across all subsamples are compared with the ones obtained from the full sample using a t-test.

3.2.5 Variable projection

Once the diffusion model is estimated, the prediction exercise requires a projection of the variables. According to a particular scenario, the attributes of the vehicles as well as the variables related to the SoC are modified, under certain assumptions, to show an evolution in the following periods. The new set of data and the estimated diffusion model are used to predict. The time horizon of the study described in Section 3.1 was the year 2050. In this case, it has been decided to shorten this period, and take the forecast to a closer moment in time since, paraphrasing John M. Keynes, *in the long run we are all dead*[1]. This brilliant statement means, in a way, that predictions too far in time are meaningless because uncertainty grows until they become completely unreliable. On the other hand, due to this same uncertainty, the future scenario proposed is ‘neutral’, in the sense that it is not especially favorable or detrimental to the EV. Macroeconomic factors have been kept unchanged, such as legal regulation or economic and transport policy related to EV (parking management, toll, infrastructure, etc). Specifically, the evolution of the variables has been defined as:

- **Price**: The price of the gasoline vehicle is barely decreased over the period considered, since this is a very mature market with optimized production processes and no significant improvements that would bring down the price (in

[1] Keynes wrote this in one of his earlier works, *The Tract on Monetary Reform*, in 1923.

nominal terms) are expected. On the contrary, the EV production processes have to be greatly improved, and important technological enhancements that will lower the price of the vehicle itself (and especially the batteries) are expected. Thus, the price of the EV is progressively reduced.

- **Range**: Mainly because of mainly technological reasons, the evolution of the autonomy of gasoline and EVs has been defined as similar to that of price.

- **Propulsion costs**: Although this is probably the aspect subject to most uncertainty due to the strategic and volatile nature of its production (and highly influenced by external shocks), it has been decided to very slightly and gradually decrease the price per mile of gasoline and electricity.

- **Tax deduction amount**: Based on the belief that subsidies for the purchase of electric vehicles will be withdrawn as they are more widely adopted, the income tax deduction is progressively reduced over the period 2026 - 2033, and completely removed in years 2034 and 2035.

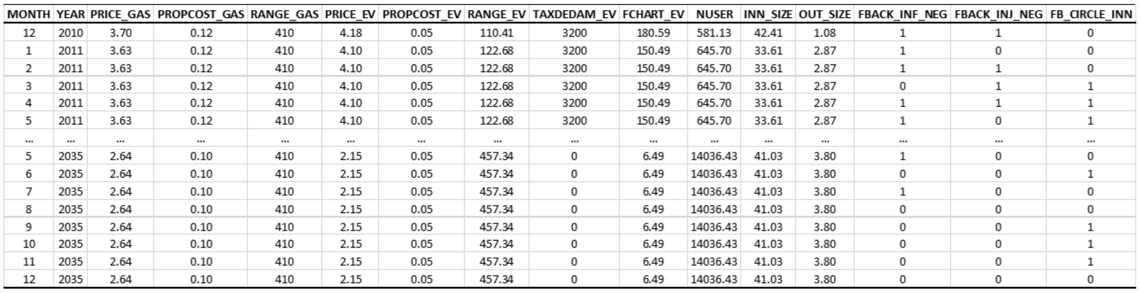
- **Fast charging time**: Again, due to technological advances, it is significantly reduced over the time horizon.

- **Inner and Outer circle size**: The number of members of a social group does not vary much over time (especially in the inner circle case), as mentioned in Section 2.1. Therefore, groups sizes are randomly assigned according to a Normal distribution characterized by the sample mean and the sample standard deviation.

- **Feedback**: A probability of obtaining negative informational and injunctive feedback is defined, depending on the moment in time. The chances of obtaining a negative opinion about charging times, need of changing activities, or receiving the disapproval of a member of one's community on buying an EV is 50% between 2020 and 2024, 30% between 2025 and 2029, and 10% from the year 2030 on. The reason is that, as the characteristics of the EV are improved and it is more extended and more used, the prevailing opinion in the public will be more favorable. Moreover, it has been decided to follow the distribution of cases in the sample where the feedback is given by someone from the inner (and outer) circle, as well as the distribution of the vehicle categories. This, allows for the projection of the variables referring to the feedback that are present in the model below.

Table 3.10 shows a glimpse of the projection of the values of these variables.

Table 3.10: Forecasting scenario 2020 - 2035.



MONTH	YEAR	PRICE_GAS	PROPCOST_GAS	RANGE_GAS	PRICE_EV	PROPCOST_EV	RANGE_EV	TAXDEDAM_EV	FCHART_EV	NUSER	INN_SIZE	OUT_SIZE	FBACK_INF_NEG	FBACK_INJ_NEG	FB_CIRCLE_INN
12	2010	3.70	0.12	410	4.18	0.05	110.41	3200	180.59	581.13	42.41	1.08	1	1	0
1	2011	3.63	0.12	410	4.10	0.05	122.68	3200	150.49	645.70	33.61	2.87	1	0	0
2	2011	3.63	0.12	410	4.10	0.05	122.68	3200	150.49	645.70	33.61	2.87	1	1	0
3	2011	3.63	0.12	410	4.10	0.05	122.68	3200	150.49	645.70	33.61	2.87	0	1	1
4	2011	3.63	0.12	410	4.10	0.05	122.68	3200	150.49	645.70	33.61	2.87	1	1	1
5	2011	3.63	0.12	410	4.10	0.05	122.68	3200	150.49	645.70	33.61	2.87	1	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5	2035	2.64	0.10	410	2.15	0.05	457.34	0	6.49	14036.43	41.03	3.80	1	0	0
6	2035	2.64	0.10	410	2.15	0.05	457.34	0	6.49	14036.43	41.03	3.80	0	0	1
7	2035	2.64	0.10	410	2.15	0.05	457.34	0	6.49	14036.43	41.03	3.80	1	0	0
8	2035	2.64	0.10	410	2.15	0.05	457.34	0	6.49	14036.43	41.03	3.80	0	0	0
9	2035	2.64	0.10	410	2.15	0.05	457.34	0	6.49	14036.43	41.03	3.80	0	0	1
10	2035	2.64	0.10	410	2.15	0.05	457.34	0	6.49	14036.43	41.03	3.80	0	0	1
11	2035	2.64	0.10	410	2.15	0.05	457.34	0	6.49	14036.43	41.03	3.80	0	0	1
12	2035	2.64	0.10	410	2.15	0.05	457.34	0	6.49	14036.43	41.03	3.80	0	0	0

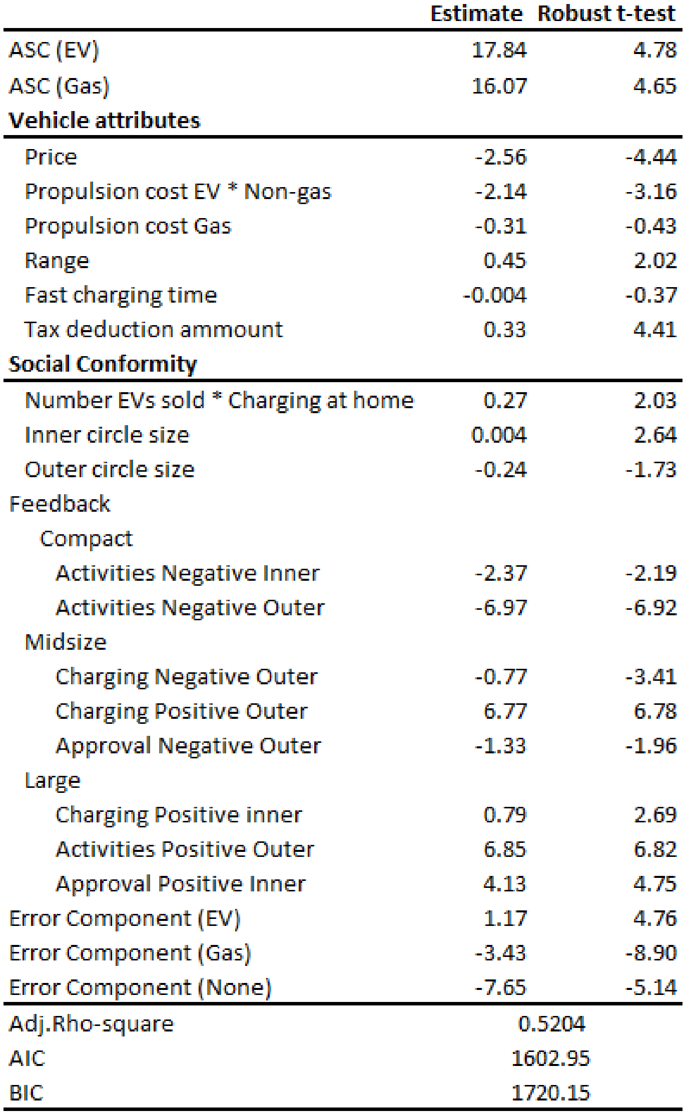
3.2.6 Results

3.2.6.1 Substitution model

The estimation of the substitution model required of an iterative process to find the best specification for the data at hand. The procedure usually starts by estimating an MNL model that serves as a baseline. Normally, its results are not satisfactory due to the limitations of MNL. However, as pointed out above, MML overcomes these limitations providing three key methodological improvements: panel data (several observations for each individual), random parameters (distribution of the values of the coefficients) and error component (substitutability among alternatives). For a discussion on these technical aspects, see [137].

In this case, this first approach yielded inconsistent coefficients, clearly failing in capturing the underlying behavior of the users' decision-making process. However, the adoption of an MML specification allowed to find parameters of the correct sign and reasonable magnitude. It also improved dramatically the performance of the estimation. Although the long process of finding the best fit involves a plethora of specifications, only the final results are presented here. Table 3.11 shows them as well as some statistics to measure goodness-of-fit. It is worth noting that *bootstrapping* has been used for the estimation. Bootstrapping consists of generating a number of training/testing subsets of data (10, in this case) by re-sampling with replacement the observations of the original dataset. Then parameters are estimated for each of these new samples in the training subset, validated in the testing one, and averaged.

Table 3.11: Estimation results, substitution model.



	Estimate	Robust t-test
ASC (EV)	17.84	4.78
ASC (Gas)	16.07	4.65
Vehicle attributes		
Price	-2.56	-4.44
Propulsion cost EV * Non-gas	-2.14	-3.16
Propulsion cost Gas	-0.31	-0.43
Range	0.45	2.02
Fast charging time	-0.004	-0.37
Tax deduction ammount	0.33	4.41
Social Conformity		
Number EVs sold * Charging at home	0.27	2.03
Inner circle size	0.004	2.64
Outer circle size	-0.24	-1.73
Feedback		
Compact		
Activities Negative Inner	-2.37	-2.19
Activities Negative Outer	-6.97	-6.92
Midsize		
Charging Negative Outer	-0.77	-3.41
Charging Positive Outer	6.77	6.78
Approval Negative Outer	-1.33	-1.96
Large		
Charging Positive inner	0.79	2.69
Activities Positive Outer	6.85	6.82
Approval Positive Inner	4.13	4.75
Error Component (EV)	1.17	4.76
Error Component (Gas)	-3.43	-8.90
Error Component (None)	-7.65	-5.14
Adj.Rho-square	0.5204	
AIC	1602.95	
BIC	1720.15	

All the vehicle attribute coefficients present the correct sign. Naturally, the higher the price, the cost of the energy, and the charging time, the less probable is that the vehicle is chosen. In contrast, a better range or a larger tax deduction increases the probabilities. These attributes are significant except for *Propulsion Cost* in the gasoline alternative, and *Fast charging time* in the EV one. Regarding the per-mile costs, this was a factor to which the interviewees did not seem to give importance, except for those who had stated that they would purchase a non-gasoline vehicle (EV, hybrid, or other). For them, the per-mile cost of the electricity was very significant, as the results show. With respect to the fast charging time, it is worth mentioning that the pilot survey, which showed both regular charging time and fast charging time, yielded inconclusive information on how users perceived them . Therefore, at that point, with the aim of reducing the informational burden in the scenarios, it was decided to keep only the fast charging time since it would be more comparable to refueling times of a gasoline vehicle. Interestingly, in light of these results, it turns out that the differences in the fast charge with respect to the refueling time of an ICV are not significant. This is true for individuals who are able to charge the car at home, as well as those who are not. Finally, the two aspects related to the direct cost of the vehicle, i.e. *Price* and *Tax deduction* seem to be of special importance. It is also worth mentioning that the significance of *Propulsion cost* for the *electric* alternative and *Number EVs sold* is conditional to individuals that wish to buy a non-gasoline car, and that have the posibility of charging the EV at home, respectively.

On the other hand, the results corresponding to the variables related to the SoC (which are only present in the EV alternative) are very illuminating and confirm some of the hypothesis of this dissertation. Firstly, the effect of the number of EVs sold in the previous month, a measure of the tendency to conform to others, is not only positive and significant, but also close in magnitude to that of the tax deduction. Secondly, the size of the inner circle plays a role in the choice of the EV; the larger the group, the more probable is to choose it. This would confirm the idea that more connections lead to more information received on EVs, and that it has a significant effect, positive in this case. In contrast, the sign of the size of the outer circle is negative, something consistent in all models estimated, although it was never significant. In all fairness, this might be due to the inaccurate information collected at this respect. As depicted in Figure 3.20, the number of members of the group *Acquaintances* seems abnormally small.

The feedback received in the repeated choice tasks does not have a homogenous effect among the different car categories and among all the types of feedback. For the users that declared that they would buy a *Compact* vehicle, only negative feedback on the need to change activities matters, whether it comes from a person in the inner or outer circle. However, for the users that declared that they would buy a *Midsize* one, it is more important the opinion received on charging it. As pointed out in the survey description, a *Midsize* vehicle refers to a family car. Thus, it is reasonable to think that people who wish to purchase a vehicle of this category are probably parents, who are probably concerned about being able to complete a tour of the *home-school-work-shopping* kind. Therefore, the need for sufficient autonomy is relevant. Approval from others is also important for these individuals, although only that of members of the outer circle, such as neighbors or coworkers. Finally, all aspects – charging, activities, and approval – are significant for potential purchasers of *Large* vehicles, but only if it is expressed in positive terms. In summary, the results evidence that not all kinds of feedback are relevant, and that its negativity/positivity, as well as its source matters.

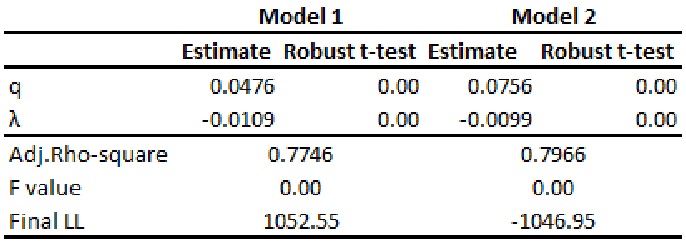
Other specifications that included random parameters were tested, specifically different combinations of *Price*, *Range*, and *Tax Deduction*. However, none of these models performed better than the one showed in Table 3.11. They either yielded a lower fit (*Adjusted rho-squared*, *AIC*, and *BIC*) or failed in finding the coefficients. This is not uncommon in complex models in which accurately identifying the distribution functions of the random parameters is difficult. In this case *Normal* and *Lognormal* were tested, common distributions used for this purpose, with no clear benefits, as mentioned.

3.2.6.2 Diffusion model

Once the coefficients of the DCM are obtained, they are to be introduced in Equation (3.21), multiplying the remaining potential market in each period, $M_t - Y_{t-1}$, to forecast sales. M is assumed to be equal to $2,192,518$, the number of households in the State of Maryland according to the U.S. Census ([18]). The formulation is a nonlinear regression model in which q and λ are estimated. The first corresponds to the time-dependent diffusion effect, while the second is a scale parameter of the coefficients of the dissaggregated model, that can also be interpreted as the effect of the substitution among alternatives that occurs in the spread of the technology.

For comparative purposes, the estimates of two models, as well as some goodness-of-fit measures, are presented in Table 3.12. Model 1 does not include any of the SoC elements, i.e. it only considers that the diffusion of the EV is based on the vehicle attributes. On the contrary, Model 2 includes all SoC factors.

Table 3.12: Estimation results, diffusion model.



	Model 1		Model 2	
	Estimate	Robust t-test	Estimate	Robust t-test
q	0.0476	0.00	0.0756	0.00
λ	-0.0109	0.00	-0.0099	0.00
Adj.Rho-square	0.7746		0.7966	
F value	0.00		0.00	
Final LL	1052.55		-1046.95	

In both cases q and λ are significant. Although the substitution parameter is of similar magnitude in both models, the diffusion parameter is about 60% larger in the model that includes SoC (Model 2). However, this does not necessarily mean a consistently higher rate of sales since the spread of the technology depends on the evolution of the vehicle attributes over time, and in some periods some aspects may dominate over others. To illustrate that, Figure 3.24 (a) presents the progression of sales over the period 2020 - 2035. The plot shows the actual sales observed between December 2010 and December 2019 (month 109), as well as the forecast number of EV sold until 2035; Model 1 in green, Model 2 in blue. It is possible to appreciate how the sales are superior in the model that does not contemplate SoC during the first 4 years of forecast, approximately. In that moment, Model 2 predicts higher monthly sales.

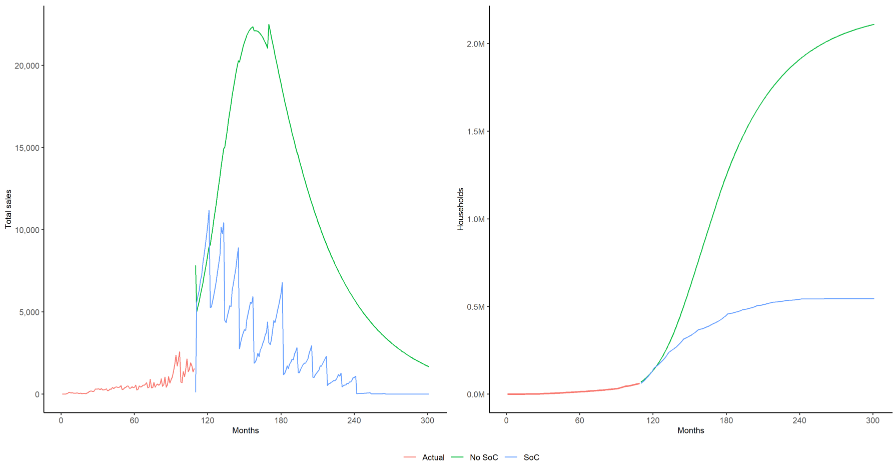


Figure 3.24: Actual and forecast sales (a). Forecast cumulative sales (b)

On the other hand Figure 3.24 (b) illustrates the cumulative sales, which depicts the classic S-shape of disruptive goods. That is, in a first period of introduction, sales grow slowly until they reach a critical point in which they explode. In a second phase, the growth is faster, to slow down again in a third stage in which the remaining of the potential market is captured. In this case, the first period was probably achieved around 2018 – 2019, close to the end of the red line, where the inflexion point is clear. Then, according to these forecasts, sales start to accelerate. Model 1 predicts that the top of the market (2.2 million households) is reached by year 2035 (month 300). However, Model 2, which considers SoC, presents a more progressive evolution and the potential market does not empty by the end of the period of analysis. That is to say, if only the attributes of the vehicles (price, range, etc.) are accounted for the spread of the technology occurs more rapidly than if the attributes plus the social aspects inherent in the diffusion process are considered. In other words, the characteristics of the electric vehicle matter, but the feedback that people obtain about them, which can be negative, matters as well. In fact, the combination of these two aspects slows down the diffusion and delays its full implementation by five years. Therefore, the results of both substitution and diffusion models justify the idea that the information we receive from our social network impacts our choices and, consequently, the diffusion process itself, counteracting the effect of technological improvements or cost-benefit analysis.

3.2.7 Conclusions

The methodology developed in this part of the research is an extension of that of the Danish case presented in Section 3.1. Moreover, it was applied to new data of better quality and specific for the State of Maryland. The approach is based on three pillars: substitution, diffusion, and dynamism. For the first two, an initial disaggregated demand model was estimated using advanced discrete choice techniques. It included variables related to SoC and to the social network of the respondent to capture the tendency of individuals to turn to members of their own group when they make choices. This concept is brought into this study by providing to the individuals feedback from one person they know. This person belongs to either the *Inner* or *Outer* circles, which existence is probably the main conceptual body of this research. This substitution sub-model gathered the subjacent inclination to substitute ICVs by EVs. Then, its results were integrated into an extended Bass diffusion model which finally yielded future sales, i.e. the spread of this technology over time among society. The dynamic aspect is provided by time-dependent variables, present in both the utility functions and the Bass model.

The results of the substitution sub-model (Table 3.11) show that, most vehicle attributes are significant, with the exception of the *Fast charging time* and the *Propulsion costs* of the *gasoline* alternative. Concretely, the aspects related to the direct cost of the vehicle, i.e. *Price* and *Tax deduction*, seem to be of special importance. With respect to the SoC variables, the effect *Number of EVs sold* – a measure of the tendency to conform to others – is not only positive and significant, but also close in magnitude to that of *Tax deduction*. Although only for those individuals who declared that they are currently able to charge an EV at home. Secondly, the size of the inner circle plays a role in the choice of the EV; the larger the group, the more probable is to choose it. This would confirm the idea that more connections lead to more information received on EVs, and that it has a significant effect, positive in this case. In contrast, the sign of the size of the outer circle is negative. Nevertheless, this result may not be completely reliable in light of the information collected on this aspect.

The effect of the feedback received in some of the scenarios varies depending on its type and on the category of car the user is willing to purchase. For those that declared that they would buy a *Compact* vehicle, only negative feedback on the need to change activities matters, whether it comes from a person of the inner or the outer circle. However, for the interviewees that declared that they would buy a *Midsize* one, it is more important the opinion received on charging it. As pointed out in the survey description, a *Midsize* vehicle refers to a family car. Thus, it is reasonable to think that people who wish to purchase a vehicle of this category are parents, who are probably concerned about being able to complete a tour of the *home-school-work-*

shopping kind. Therefore, the need for sufficient autonomy is relevant. Approval from others is also important for these individuals, although only that of members of the outer circle, such as neighbours or coworkers. Finally, all aspects, charging, activities, and approval, are significant for potential purchasers of *Large* vehicles, but only if it is expressed in positive terms. In summary, the results evidence that not all kinds of feedback are relevant, and that its negativity/positivity, as well as its source, are relevant.

On the other hand, the results of the diffusion model that takes into account SoC present a more moderate sales evolution than the model that does not. The pace of the spread of this technology is slower and, therefore, the market is emptied later. Taking as a reference the number of households in the State of Maryland, this would occur around the year 2035, five years later than the model predictions that only consider vehicle attributes. This leads to the conclusion that although the characteristics of the vehicle are relevant, so is the number of social connections of the individual, as well as the information obtained from them. The combination of this two spheres modifies the decision-making process at a disaggregated level, consequently impacting the diffusion of the electric vehicle in aggregate terms.

4. Machine Learning methods for the classification of potential EV purchasers

The number of studies that have explored EV adoption is large, either taking the agent's perspective, or trying to predict penetration through more macroeconomic approaches, as is done in the research presented in the previous section. However, although these studies point together in the same direction, they offer very different EV market evolution in terms of time and magnitude. This causes a lack of reliability on which it is difficult to make strategic decisions, either by the industry or the public sector. Therefore, new methodological perspectives are required, Machine Learning (ML) being one of them.

ML techniques are currently applied to an enormous variety of topics such as fraud detection, robotics, spam filtering, translation services, preventive health care, computer vision, as well as transportation. This has been possible thanks to the exponential growth of information brought about by electronic devices; an amount that will continue to expand due to the Internet of Things. In the case of transportation, the smart use of the data generated by on-road vehicles presents an extraordinary opportunity to improve transportation systems. However, this task overcomes the capabilities of traditional data analysis and clearly points to ML as a solution. Congestion reduction, safety improvement, environmental impact mitigation and energy consumption optimization are examples of the most common lines of research in which ML has been applied. However, there are other less explored fields of application, such as the classification of potential consumers into adopters/non-adopters. This is a topic that presents interesting challenges. Adoption is demand-driven, and Demand roots into purchasers' behavior, beliefs and attitudes – elements that are intrinsically difficult to define and gather. Even if reliable information on these aspects is available, it is unlikely to be in large quantities, so certain methodologies cannot be used as they only perform well in large data sets. The aim of this research is, precisely, to compare the throughput of several supervised ML techniques when applied to the classification of individuals into EV adopters. Namely, three methods of different nature are applied, Random Forests (RF), Support Vector Machines (SVM), and Artificial Neural Networks (ANN), to the survey data previously described in Section 3.2.3.

4.1 Supervised Machine Learning techniques for classification

ML algorithms can be categorized into three types; supervised learning, used both in classification and regression tasks; unsupervised learning, used for clustering and dimensionality reduction; and reinforcement learning, based on reward maximization. In supervised learning observations are labelled, i.e. each one has a class assigned, an associated response. The algorithm processes the data on a training subset to generate label predictions that are validated in a different testing subset. The error resulting from this comparison is used to fit the model as much as possible to the data. An advantage of most ML methods is that, unlike ordinary regression models, they are not parametric. They do not rely on assumptions about the relationships between variables present in the model in order to minimize the error (cost function). In contrast, when the cost function becomes more complex, with more parameters and high dimensionality, minimizing it becomes a difficult task. Another difficulty is the enormous diversity of algorithms that can be applied to the same problem. Although there may be some guidelines on which one should be applied to each case, the truth is that different approaches may lead to significant deviations of the level of performance. In addition, each technique has specific hyperparameters with no reliable predefined values. This complexity implies that, when facing a project (either classification or regression), it is first necessary to assess what method may be appropriate for it and then to carry out an iterative process of tuning. This process must also include a resampling procedure to avoid any possible bias when splitting the data into training/testing subsamples. In this case *k-Fold Cross Validation* is used, which randomly divides the sample into k groups, or folds. The model is validated on the first fold and fit in the remaining k-1. The process is repeated k times; each time a different subset is treated as the validation one. The final accuracy measure is computed as the average of those obtained in all folds. These procedures are highly compute-intensive, especially as the number of data points and dimensionality grows. Therefore, it is common to conduct a feature selection process, which consists on a pre-filtering of the most important variables in order to reduce the dimensionality. For this work, various ML algorithms were considered, and finally it was decided to compare the performance of Support Vector Machines (radial and polynomial kernel), Random Forest, and Neural Networks. The reason is two-fold. First, as stated in Section 1, they maintain a good balance between flexibility and overfitting. This is, they can fit the data well, while maintaining the ability to reasonably predict new observations for which their classes are unknown. Secondly, these are models that are not restricted by assumptions of linearity, normality, or variable independence. The following subsections detail the methodology conducted.

4.1.1 Support Vector Machines

SVM is an extension of the Support Vector Classifier (SVC). SVCs rely on the concept of an optimal separating hyperplane, which consists in finding a $p-1$ dimensional space, where p is the number of features, that perfectly separates the training observations according to two classes. If such hyperplane exists, then a straightforward classifier may be constructed; a test observation x^* is categorized based on the sign of:

$$f(x^*) = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + ... + \beta_p x_p^* \tag{4.1}$$

where the right hand of the equation is the general expression of a hyperplane.

However, it could be case that a classifier based on a separating hyperplane is not desired. This happens when it assigns classes to the training observations extremely well, which may be a sign of overfitting. In this situation, a procedure that does not perfectly separate the classes for the training observations may produce a better fit on the testing ones. Following this rationale, the SVC finds a hyperplane that correctly separates most of the training observations, but misclassify a few of them, as the solution of the following optimization problem:

$$\begin{aligned} & \mathit{maximize} \quad M \\ & \mathit{subject\ to} \quad \sum_{j=1}^{p} \beta_j^2 = 1 \\ & \qquad\qquad y_i \left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} \right) > M \left(1 - \epsilon_i \right), \\ & \qquad\qquad \epsilon_i \geq 0, \ \sum_{i=1}^{n} \epsilon_i \leq C \end{aligned} \tag{4.2}$$

where M represents a margin of the hyperplane that is sought to make as large as possible, and $\epsilon_i, ..., \epsilon_n$ are variables that allow the observations –which are, again, classified by determining on which side of the hyperplane lies– to be on the wrong side of the margin or the hyperplane (when $\epsilon_i > 1$). C is a non-negative tuning parameter that bounds the sum of the ϵ_i – i.e. it controls for the number of margin violations. As C increases, the margin widens and the classifier becomes more tolerant. The observations lying in the margin or that violate the margin are called support vectors.

However, the SVC performs poorly if the actual boundary between classes is not linear, which most of the times is the case. One way to overcome this problem is by enlarging the feature space using quadratic, cubic or other polynomial functions of the predictors, so a non-linear boundary between classes may be accommodated.

In order to do so, 4.1 may be transformed into:

$$f(x^*) = \beta_0 + \sum_{i \in S} \alpha_i K(x, x_i) \tag{4.3}$$

a new classifier that uses some function K, referred as *kernel*, in order to produce non-linear boundaries that may better gather the training observations into the classes that they pertain to. When a kernel function is involved, the algorithm is called a Support Vector Machine (SVM). However, finding the appropriate boundary contours is an exploratory task. In this study, polynomial and radial kernels are used to train the model and then classify the testing observations. Polynomial kernel of degree d has the form:

$$f\left(x^*\right) = \beta_0 + \sum_{i \in S} \alpha_i \left(1 + \sum_{j=1}^{P} x_i x_{i'j}\right)^d \tag{4.4}$$

which produces non-linear boundaries which shape depends on d. The radial kernel, from its part, is specified as:

$$f\left(x^*\right) = \beta_0 + \sum_{i \in S} \alpha_i exp(-\gamma \sum_{j=1}^{p} \left(x_i - x_{i'i}\right)^2) \tag{4.5}$$

where γ is a positive constant. The use of a radial kernel makes the classifier to work more 'locally' because the training observations far from x^* play little role in its classification. Finally, it is worth to mention that SVM are naturally suited for two classes, due to the concept of separating hyperplanes it is based on. However, it is possible to apply it K-classes using different methods. The one followed in this work is the *one-versus-one* method, which constructs $\binom{K}{2}$ SVM, each of which compares a pair of classes. The algorithm counts the number of times that the test observation is assigned to each of the K classes. The final classification is that in which the test observation is most frequently assigned.

4.1.2 Random Forest

The so-called Decision Trees (DT) recursively stratify the predictor space into distinct non-overlapping regions, making a prediction for each observation falling into a region. This structure is calculated with the objective of minimizing a measure of prediction error. In the case of classification problems with a large feature space the Gini index is one of those. It represents the total variance across classes:

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1-\hat{p}_{mk}) \tag{4.6}$$

where $\hat{p}_{mk}$ stands for the proportion of training observations that do not belong to the most common class observed in the region. A small value evidences that a region contains predominantly observations from a single class, which is not desirable. However, one of the main drawbacks of this simple approach is that trees are very sensitive to changes in the data; a different training set may lead to important differences in the final tree estimated. RFs solve this issue in several ways. First it uses bootstrap aggregation to reduce the variance resulting from splitting the feature sample space, providing robustness and improving performance. Concretely, *Bootstraping* is a cross-validation technique that randomly takes different samples from the training set, estimating a prediction model for each of them, and averaging their predictions. Interestingly, although each tree has a high variance, their combination has a low one, yet maintaining the accuracy. Another improvement brought by RFs is that each time a split is considered, a reduced random sample of predictors is chosen as split candidates, and the split can use only one of those predictors. Although counter-intuitive, this is a valuable procedure in avoiding the dominance of features. In this way, not all the trees explored will contain the dominant variable and, therefore, will not be similar to each other, avoiding correlations. Therefore, RF can be considered an ensemble learning method that overcome the limitations of simple decision trees.

4.1.3 Artificial Neural Networks

An ANN consists of a group of interconnected nodes that process input information to make predictions. Data enter the network by the input layer, is processed by hidden layers, and finally an output (a prediction) is provided. Every node in a layer is connected to every node in its following layer by a weight; and each node has an activation bias, a constant that sums up when activity reaches the node. The ANN learns by updating these weights and biases to reduce the error in predictions, according to the following equation:

$$y_i = \sum_{q=1}^{h} w_q^{(2)} g\left(\sum_{p=1}^{d} w_{qp}^{(1)} x_{ip}\right) for\ i = 1, 2, ..., n \tag{4.7}$$

where w_q^h,w_{qp}^d and $g()$ are the weights connecting input and output nodes, and the activation function. One approach to implement this training process is the back-propagation algorithm with an underlying gradient descent formula. This optimization process updates connection weights and biases by the product of a learning rate value and the partial derivative of the error over the connection weight or bias (Equation 4.8), stopping when a pre-defined threshold is reached.

$$\Delta w_{ij} = -\alpha \frac{\partial \varepsilon}{\partial \omega_{ij}} \tag{4.8}$$

Among the different versions of backpropagation available, I chose a resilient backpropagation algorithm with weight backtracking [109] due to its efficiency in terms of prediction accuracy and time consumption.

Although ANN are very effective, the number of hyperparameters to be tuned is high, which makes them very intensive in computing power and time. Additionally, neural networks are kind of black-boxes in the sense that is not possible to provide explicit meaning to intermediate output besides its computational role. When estimating an ANN, the researcher basically feeds the network in order to train it and obtains an output. But there are not equations, coefficients or p-values that define a relationship.

4.2 Data

As described in Section 3.2.5, the *Release* data was used to estimate the substitution-diffusion model. The reason for this is that after analyzing the information collected in the pilot phase, it was found that the levels defined in the statistical design did not provide clear trade-offs to make the estimation of the DCM efficient. However, ML methods are less constrained by these circumstances. In addition, the accuracy of these techniques increases substantially as the number of observations grows. Therefore, to carry out this part of the study it was decided to join the *Pilot* and *Release* data sets. Tables B.4 and B.5 in Appendix B shows its descriptives statistics, which do not differ much from those corresponding to the *Release* only (B.2 and B.3 in Appendix B).

4.2.1 NA imputation and feature selection

The social question included in the first section of this survey is a distinguishing feature of this data collection. It will help to identify whether the social network structure of the individuals is significant in adopting EVs. However, its design involved a particularly inconvenient casuistry; i.e. the interviewee may not know the number of individuals in a group. That is, one may genuinely not know how many acquaintances has or how many friends can talk to about EV technology, for instance. Therefore, it was necessary to offer an *I don't know option*, which meant a missing value when selected. Since the ML techniques to be applied cannot handle missing values, it was necessary to impute them. For this task I relied on the Multiple Imputation by Chained Equations (MICE) method. The MICE algorithm regresses a variable with missing values, x_i, on other selected features. The *NA* are then replaced by simulated draws form the predictive distribution of x_i.

To provide coherent imputations, the process is repeated several times producing a single imputed dataset. Moreover, the entire procedure is performed M times producing M imputed datasets, which are ultimately combined. In our case, the number of members of the main social groups, where missing, were imputed using all the other information in the data. Then, the social subgroups were imputed in a second round using the same information plus the recently imputed one.

On the other hand, counting with a large set of predictors may actually be a drawback in the analysis. As their number grows, they are more likely to be correlated. Figure 4.1 shows a map of the correlations of our dataset. Some 'correlation clusters' can be identified, but they mostly respond to the social network variables, which columns are located next to each other for each subgroup. For instance, all the columns storing the information regarding the Friends group, are placed together and, obviously, the number of friends one shares hobbies with, talks about personal matters, etc., is correlated to the total number of friends that one has.

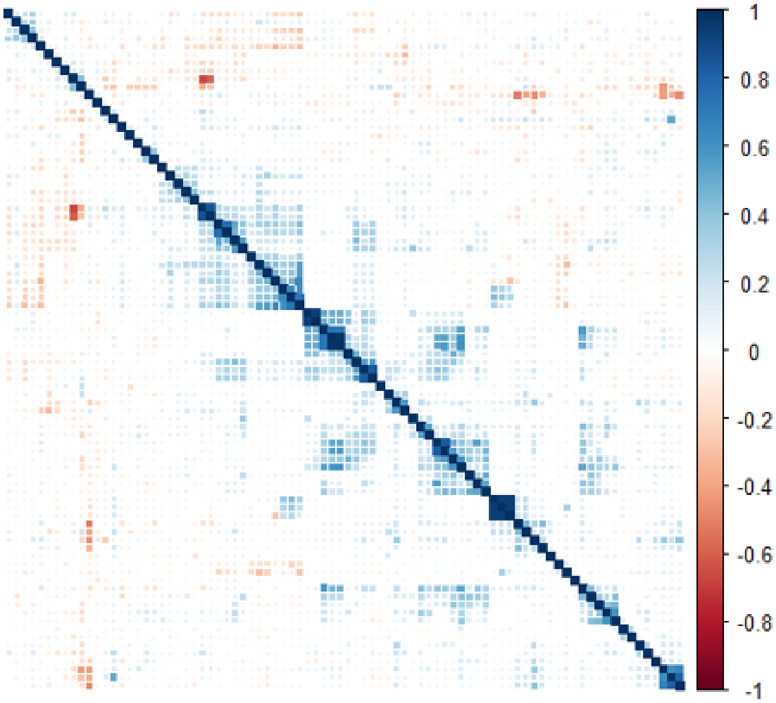


Figure 4.1: Feature correlation map.

Non-parametric classifiers, like the ones applied in this study, are not very sensitive to correlation. However, the inclusion of unnecessary variables leads to the *curse of dimensionality*; the larger the feature space, the sparser the data becomes. In other words, the amount of training data that is required to ensure that there are several samples for each combination of feature values becomes insufficient. On the other hand, a large number of features also increases the complexity of the models, which become prone to overfitting; they will fit the training data so well that they will not be able to correctly predict the classes of new observations. Fortunately, these issues may be overcome through dimension reduction techniques that reduce the number of variables, yet preserving, to a reasonable extent, the information that they keep. The approach followed in this study is the application of a preliminary Random Forest (RF) in order to identify the importance of each variable in the data set. Then, the top variables in terms of importance will be incorporated into the models. I chose this procedure over others, such as the widely used Principal Component Analysis since it keeps variables in its original form, instead of building new constructs that are difficult to interpret. Figure 4.2.a shows the 30 most important variables (over a total of 84) when choosing the type of vehicle after running a RF composed of 500 trees on the original dataset, number of trees for which the error rate stabilizes (Figure 4.2.b).

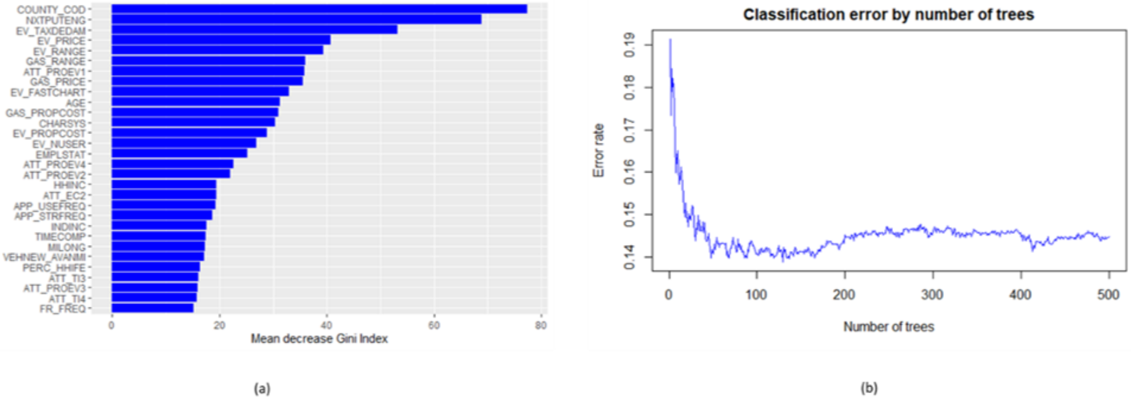


Figure 4.2: Variable importance (a) and Classification error by number of trees (b)

The most important features in this case are the county in which the user resides and the engine of the next purchase (*/Electric, Gasoline, Hybrid* or *other*). For the former, some of the counties in the State of Maryland are among the richest in the U.S. Thus, this variable may actually be reflecting a geographical high-income distribution. They are followed by: the amount of the income tax deduction associated to the EV purchase, its price and range; the range of the gasoline vehicle; and ATT_PROEV1, which reflects the respondents' level of agreement to the sentence *Electric vehicles should play an important role in our mobility systems.* The rest of the top ten inputs are the price of the gasoline vehicle, the time of fast charging of the electric one, and the age of the respondent. EV_NUSER, which measures the effect of social conformity, is also ranked high, even above household income. Also, some of the variables that provide information on the individuals' attitudes towards environment (ATT_EC2), EVs (ATT_PROEV, ATT_PROEV2), and technological progress (ATT_TI3, ATT_TI4). It is encouraging to confirm that the number of members of some social groups is also important (FR_FREQ). On the opposite side, not shown in the figure, are; other sociodemographic variables such as gender or marital status; the size of the next vehicle to be purchased, who will drive it or for what purpose; and the structure of the outermost social group (Acquaintances).

Finally, as implied above, the use of a greater number of variables does not necessarily yield better results. Although the performance of a technique first increases as the number of dimensions grows, it decreases if the feature space continues to enlarge, resulting in the known *peaking paradox*. Therefore, it is necessary to combine the tuning and validation process previously mentioned with the search for the optimal number of variables to include. In order to do so, the estimation of each method is iterated over the 5, 10, 15, 20 and 25 top variables. In other words, for each of this number top of variables, the best combination of hyperparameters is repeatedly found and cross-validated, leading to the estimation of thousands of models, namely: 2,500 SVM radial kernel, 2,400 SVM polynomial kernel, 725 RF, and 2500 ANN. Among those, the best SVM, RF, and ANN that can be fitted to this data were finally selected.

4.3 Results

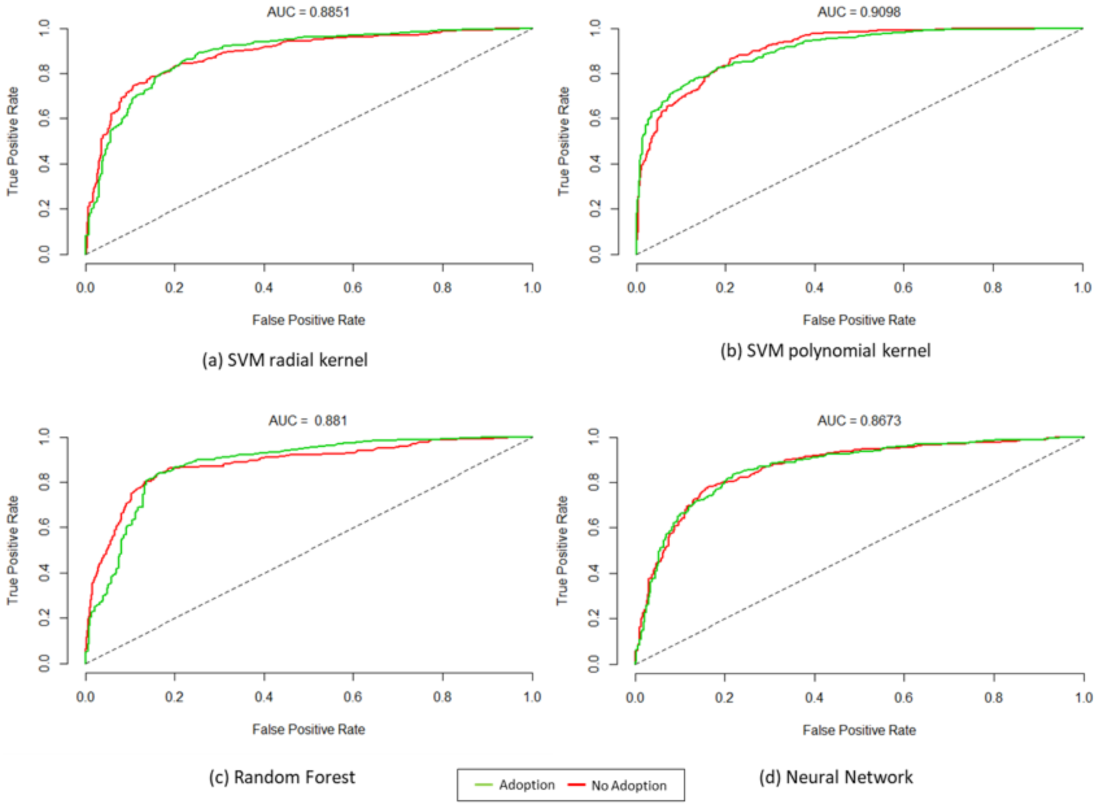
The results of the analysis described above are reported in Table 4.1. It shows the confusion matrices, as well as the optimal number of top variables and the optimal values of the hyperparameters resulting from the tuning process. Although the averaged accuracy over all k-folds is the main performance statistic used to assess model performance, *Sensitivity* and *Specificity* are as well of special importance. They provide the proportion of true positives (an adopter classified as such) and true negatives (a non-adopter classified as such). *Cohen's Kappa statistic* and *p-value(NIR)* are as well useful to evaluate model performance. Cohen's Kappa measures the agreement between raters who classify N items into C mutually exclusive categories. Following [83], a zero value means an agreement equivalent to chance, 0-0.2 slight agreement, 0.2-0.4 fair agreement, 0.4-0.6 is a moderate agreement, 0.6-0.8 substantial, and 10.8-1 almost perfect. The p-value (NIR) indicates if the predictions are accurate over no information rate, i.e. random level.

Table 4.1: Models performance.

	SVM Radial		SVM Polynomial		Random Forest		Neural Network	
	Adopting	No Adopting	Adopting	No Adopting	Adopting	No Adopting	Adopting	No Adopting
Adopting	280	63	302	80	292	77	292	94
No adopting	95	515	73	498	83	501	83	484
Accuracy	0.8342		0.8395		0.8321		0.8143	
Confidence Interval	(0.8090 – 0.8573)		(0.8146 – 0.8622)		(0.7848 – 0.8355)		(0.7881 – 0.8385)	
p-value (NIR)	0.00		0.00		0.00		0.00	
Sensitivity	0.7467		0.8053		0.7413		0.7787	
Specificity	0.891		0.8616		0.8564		0.8374	
Cohen's Kappa	0.6474		0.6647		0.6017		0.6129	
Hyperparameters								
Number variables	25		25		25		20	
C	100		0.1		-		-	
γ	0.01		0.5		-		-	
d	-		5		-		-	
α	-		2		-		-	
Trees	-		-		1,000		-	
Variables at split	-		-		3		-	
Hidden neurons							17	
Threshold	-		-		-		0.1	

The accuracy of all methods is very similar; not lower than 0.81, with polynomial SVM standing out slightly. The p-value (NIR) under 0.05 indicates that the predictions are accurate over random level. Therefore, we can safely affirm that more than 81% of the choices made by individuals were predicted correctly, no matter the technique used. In this regard, the confusion matrices at the top of the table present the actual choices (row) and the predictions (column). The values in the diagonals correspond to correct predictions, which are homogeneous among the methods. Sensitivity and Specificity are calculated from these figures, and are satisfyingly high. On the other hand, since the kappa statistic is not lower than 0.6 in all four cases, we can conclude that there is substantial agreement.

Sensitivity and Specificity lead to the popular ROC curves. These are historic graphs that display type I and II errors for all possible threshold values for the posterior probability that is used to perform the class assignment. The overall performance of a classifier, summarized over all possible thresholds, is given by the area under the curve (AUC), which is maximum when it reaches the top left corner. Figure 4.3 shows the ROC curves and their respective averaged AUC for all four techniques.



(a) SVM radial kernel

(b) SVM polynomial kernel

(c) Random Forest

(d) Neural Network

Figure 4.3: ROC curves and corresponding AUC.

Therefore, attending to the statistics described and the AUC, it can be concluded that SVM with polynomial kernel seems to have the highest capabilities in predicting the adoption of EVs when compared to the other algorithms considered. For illustrative purposes, Figure 4.4 evidences the similarity of the pattern of the actual classes and the classes predicted by this method, plotted by two of the most relevant variables found in the preliminary analysis.

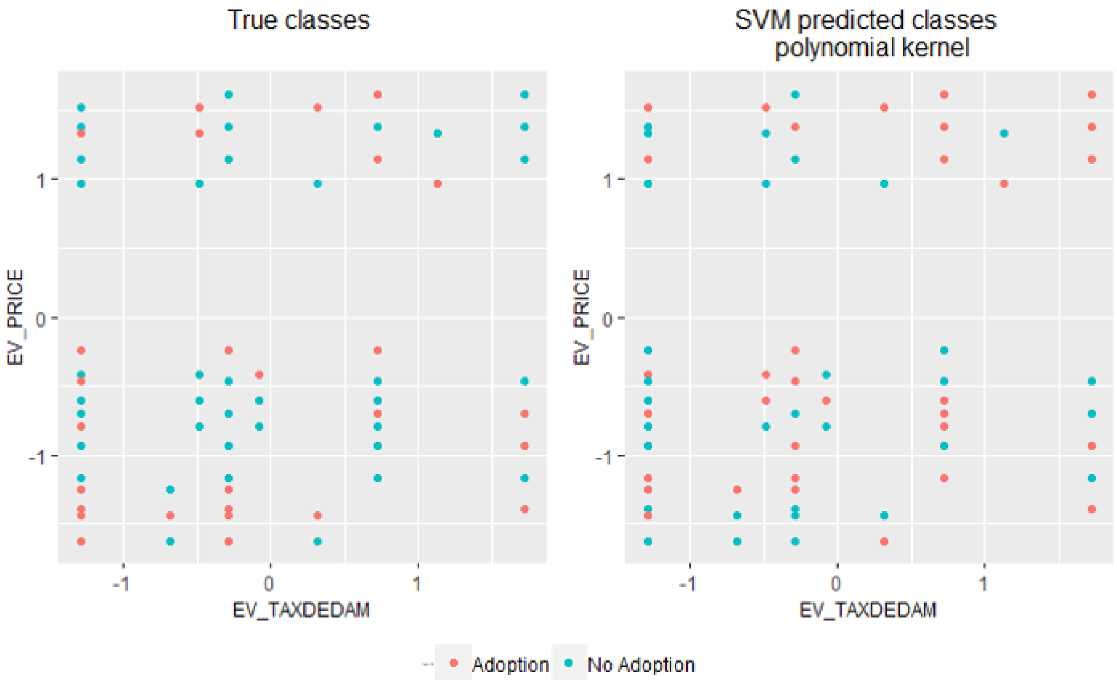
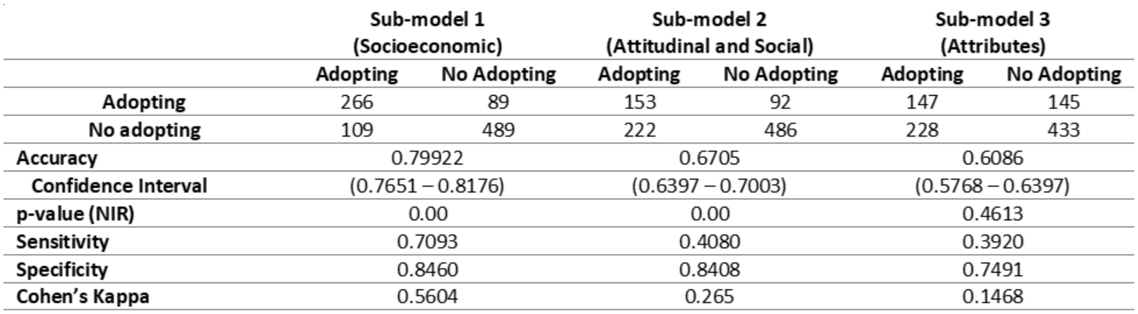


Figure 4.4: Actual and predicted classes by Tax deduction amount and EV price.

Now, the top variables depicted in 4.2 are of different nature, and they can be grouped into three well-differentiated areas; socioeconomic; attitudinal and social; and (vehicle) attributes-related. Table 3 shows this classification. It is worth remembering that the attitudinal variables correspond to several indicators unveiling the inclination of individuals towards the environment, technology, and EV.

Considering this very differentiated groups, an interesting question is which of them represents the bulk of the predictive power. To answer this question, we estimate again the best model found (Support Vector Machine with polynomial kernel), for each group of variables. Their results are shown in 4.5.



	Sub-model 1 (Socioeconomic)		Sub-model 2 (Attitudinal and Social)		Sub-model 3 (Attributes)	
	Adopting	No Adopting	Adopting	No Adopting	Adopting	No Adopting
Adopting	266	89	153	92	147	145
No adopting	109	489	222	486	228	433
Accuracy	0.79922		0.6705		0.6086	
Confidence Interval	(0.7651 – 0.8176)		(0.6397 – 0.7003)		(0.5768 – 0.6397)	
p-value (NIR)	0.00		0.00		0.4613	
Sensitivity	0.7093		0.4080		0.3920	
Specificity	0.8460		0.8408		0.7491	
Cohen's Kappa	0.5604		0.265		0.1468	

Figure 4.5: SVM with polynomial kernel performance by group of top variables.

As expected, the accuracy with respect to the general model decreases in all cases. However, in sub-models 2 and 3 the fall is dramatic; 17 and 24 percentage points are lost, respectively. Moreover, the decrease in the Sensitivity in these sub-models is especially notorious; it drops from 0.7467 in the general model to just 0.408 and 0.392. That is, if only attitudinal or only attributes-related variables are used, most of the *adopters* will be misclassified as *non-adopters*. In fact, sub-model 3 is not statistically significantly different from assigning randomly the classes, as the p-value (NIR) above 0.05 evidence. Finally, the Cohen's Kapp statistics are significantly reduced, too, clearly limiting the validity of these specifications.

4.3.1 Misclassified observations

The best model (SVM with polynomial kernel) does not correctly classify about 16% of the observations. An interesting question is whether these individuals share characteristics that make the algorithm fail when classifying them. In order to reveal these traits, we first carried out a cluster analysis of the misclassified observations to identify, if they existed, groups of individuals. Then, we performed an exploratory data analysis on all the variables incorporated to the model estimation. Cluster analysis is a term that covers several procedures for finding subgroups of observations that are similar to each other in a data set. These subgroups may exist or not, therefore, the first step is to assess if the data is clusterable. In order to do so, the Hopkins' statistic ([84]) is calculated. It measures the probability that a given set of data is generated by a uniform distribution. In other words, it tests the randomness of the information. Specifically, if the observations are uniformly distributed the statistic would be 0.5. However, if clusters are present, the value is higher. A result above 0.75 indicates a clustering tendency at the 90% confidence level. In the case of our misclassified observations, the Hopkins' statistic is 0.799, therefore, this group of individuals is clusterable. Visual assessment is also possible relying in the algorithm of ([11]), which computes the dissimilarities between the observations of the data set and displays them in an image. Figure 4.6 illustrates this visualization for our case. White or red points represent low dissimilarity between two observations. Therefore, the whiter or redder the image, the more clusterable the data set is. Attending to both Hopkins' statistic and the visual assessment, we can conclude that our misclassified individuals are subject to clustering.

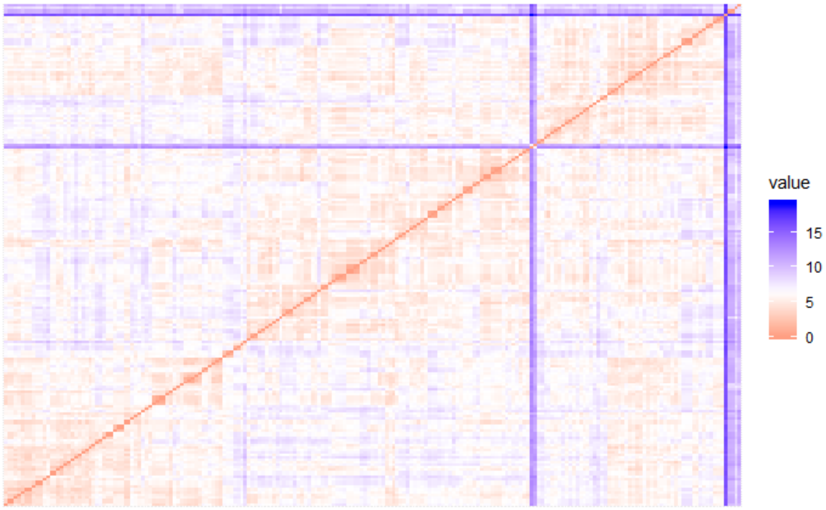


Figure 4.6: Clustering tendency of the misclassified observations.

The second step is to find out how many clusters the data should be divided into, since this is not known in advance. One approach to identify the groups is Hierarchical clustering, which provides a tree-based representation of the observations called *dendogram*. This technique starts by treating each of the n observations as its own cluster, then the two that are most similar to each other are merged, leaving $n - 1$ clusters. Next, the two most similar clusters are merged, leaving $n - 2$ clusters, and so on (for a complete description of this algorithm, we refer the reader to [64]. The dendogram is a representation of this process. Observations that merge at the bottom are very similar, while observations that fuse close to the top are different. The number of branches in which the dendogram splits at the top of the tree indicate the optimal number of clusters the data may be split into. Figure 4.7 shows the dendogram of the misclassified observations, which evidence two clusters.

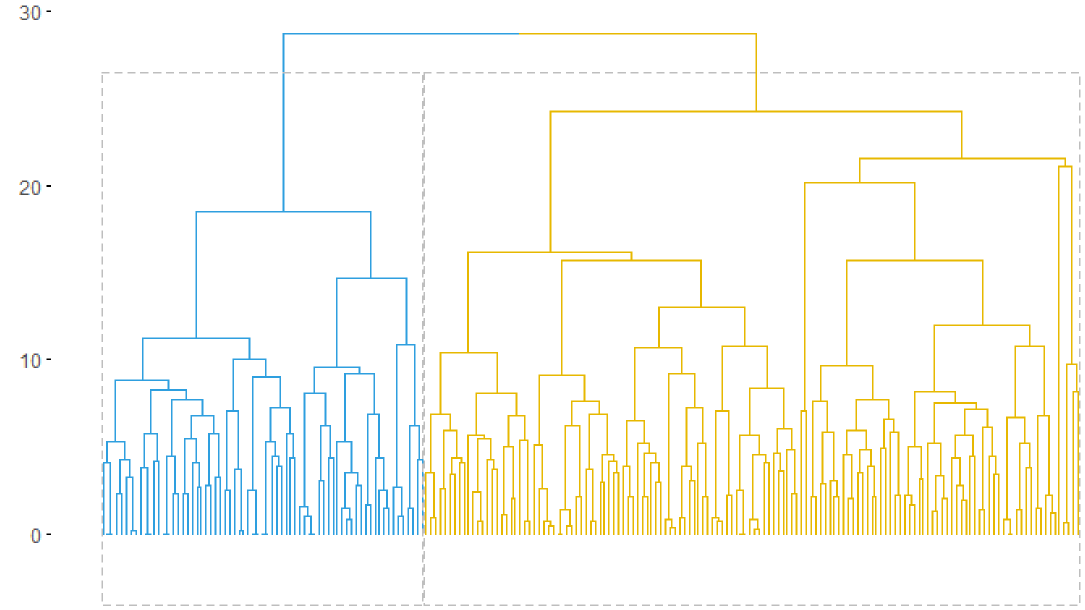


Figure 4.7: Hierarchical clustering of the misclassified observations.

After identifying that the data is clusterable into two subgroups, the classification is performed. To do so, we opted for the K-means algorithm [93], which partitions the data set into K distinct, non-overlapping clusters seeking the smallest *within-cluster variation*. Formally, the problem to solve is:

$$minimize C_1, ..., C_k \left\{ \sum_{k=1}^{K} W\left(C_k\right) \right\} \tag{4.9}$$

where $W\left(C_k\right)$ represents a measure of the amount by which the observations belonging to a cluster differ from each other. The most common measure is the squared Euclidean distance:

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i\prime \in C_k} \sum_{j=1}^{P} (x_{ij} - x_{i\prime j})^2 \tag{4.10}$$

were $|C_k|$ denotes the number of observations in the kth cluster. Thus, the within-cluster variation of a cluster is the sum of all the pairwise squared Euclidean distances between the observations in the cluster, divided by the total number of observations in the cluster. Consequently, Equation 4.9 becomes:

$$minimizeC_1, ..., C_k \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i\prime \in C_k} \sum_{j=1}^{P} (x_{ij} - x_{i\prime j})^2 \right\} \tag{4.11}$$

It is possible to visualize the partitioning results for the chosen number of clusters (two, attending to the preliminary analysis) drawing a scatter plot of data points colored by cluster. Since the data set contains more than two variables, a Principal Component Analysis has been performed to reduce the dimensionality (for a comprehensive description of this method see [70]).

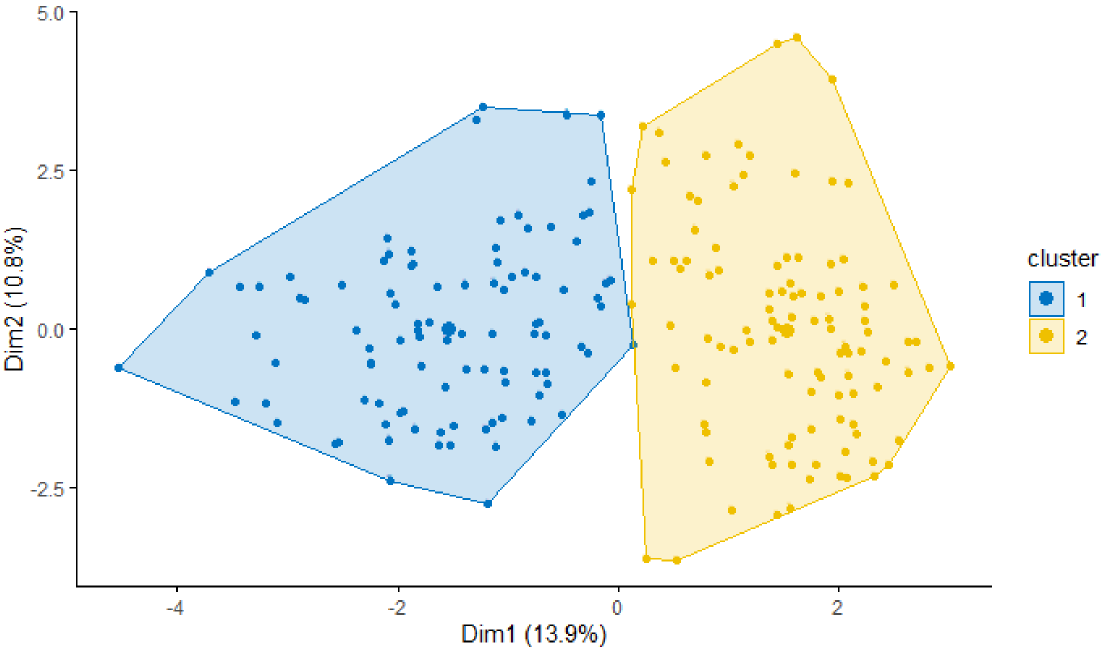
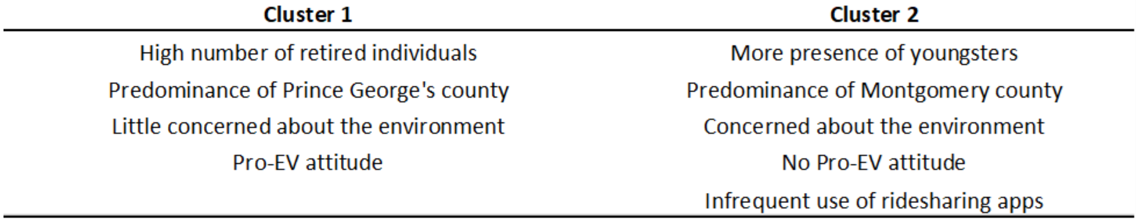


Figure 4.8: Clustering results, two main Principal Components.

Now, in order to find the characteristics common to the members of each cluster, and the differences in-between clusters, an exploratory data analysis has been carried out. This required of the examination of the main statistics of each variable as well as of their distribution. The results are summarized in Table 4.9.



Cluster 1	Cluster 2
High number of retired individuals	More presence of youngsters
Predominance of Prince George's county	Predominance of Montgomery county
Little concerned about the environment	Concerned about the environment
Pro-EV attitude	No Pro-EV attitude
	Infrequent use of ridesharing apps

Figure 4.9: Clusters characteristics.

A high share of the individuals belonging to Cluster 1 are retired and live in the Prince George's county in Maryland (a low-income one, in comparison to the other counties), while those belonging to Cluster 2, are younger and live, predominantly in the Montgomery county (a high-income one). Cluster 2 seems to present an interesting infrequent use of ridesharing apps as well, either riding alone or with strangers. However, the most enlightening characteristic of the first group of potential adopters may be the fact that they are little environmental concerned although they scored high in the pro-EV attitude evaluation. Moreover, the opposite occurs in the second group, where individuals scored high in environmental concern, but low in Pro-EV attitude. This is to some extent contradictory since EVs contribute positively to reduce climate change, so a person that is environmental concerned usually has also a pro-EV attitude and frequently chooses EV. We think that this contradiction is precisely what may be behind the misclassification; the algorithm might find trouble in labelling a person that states that cares about the environment but does not have an inclination for EVs, and vice versa.

4.4 Conclusions

The aim of this research is two-fold; first, to explore the most influencing factors in the adoption of the electric vehicle; second, to carry out a comparison of machine learning methods in order to find out which one leads to better predictions. To do so, we use data collected through a stated choice experiment specifically designed to gather the inclination of individuals towards EVs and the role that their social structure plays in their choices. SVMs, RFs, and ANNs were estimated to classify individuals in adopters and non-adopters, examining their respective throughputs.

With respect to the first objective, the most relevant features when classifying the individuals of this dataset were the county in which they live, the type of engine of the next vehicle to be acquired, some vehicle characteristics, and several attitudinal variables. Since there are not special differences among the counties of the State of Maryland in terms of power grid or charging infrastructure, the county variable may actually hide an income effect. Some of these counties are among the richest in the U.S. and they evidence a clear geographical income distribution. Among the vehicle characteristics, the most relevant seems to be the income tax deduction that the U.S. government provides when buying an EV. Considering that the fourth most important variable is the price of the type of vehicle, we can conclude that all the elements that gravitate around the purchase cost are fundamental in the individual's inclination to adopt this technology. However, other vehicle attributes are also capital, such as the range and the time of fast charging, as well as the existence of charging infrastructure in the household. This is valuable information for the automotive and power industry since these are precisely some of the barriers highlighted by users and researchers for a wide-scale implementation of this technology ([10], [139], [16], [39]). Some of these findings are also important for the public administration, beyond the economic incentives. The attitude of individuals towards the environment and towards technology is relevant, too. Although these aspects are obviously inherent in each person, fostering social awareness of environmental care would also boost the EV market.

On the other and since the top variables were clearly different in nature (socioeconomic, attitudinal and social, and attributes-related) three sub-models were re-estimated using one of these groups at a time with the objective of identifying their importance. The results confirm that counting only with limited information dramatically decreases the general accuracy as well as the rate of true positives, especially if only the attitudinal variables or those related to vehicle attributes were included. Therefore, it becomes clear that the synergy of all the variables is what produces a better classification, and that it is not convenient to rely only on aspects specific to the individual under analysis.

Regarding the second objective of this study, the four ML methods analyzed show a similar performance, with SVM with polynomial kernel slightly standing out from the others. It provides better accuracy, highest Sensitivity, Specificity, and Cohen's Kappa statistic, as well as the highest average area under the ROC curves.

However, it is fair to mention that the biggest advantage of these predictive models can also be the biggest drawback. These techniques are flexible enough to capture complexity in datasets avoiding introducing model assumptions since they are completely data driven, but this precisely makes them very dependent on the input information. Sampling bias was avoided, or at least mitigate by implementing k-fold cross validation. RF predictions are very robust in this regard, thanks to the aggregation and decorrelation of trees that is performed in its estimation. Moreover, the computing time of RF is usually inferior to the other techniques, which is something to take into account when using large datasets. Therefore, it is not unreasonable, at least for this case, to consider this algorithm as the most convenient, although its performance is below that of SVM. In the same vein, ANN were much more complex to code and estimate, taking significantly longer computing time, which did not translate in the end into better results.

Finally, we tried to identify characteristics common to the misclassified individuals. To do so, we carry out a cluster analysis followed by an exploratory data analysis. The results show that the observations incorrectly predicted belong to two well-differentiated groups. The first is characterized by retired persons that live in a low-income county and that do not care much about the environment but have a pro-EV attitude. The second cluster, in contrast, is composed by young potential customers that live in a high-income county, and that care about the environment although do not show special interest for the EV. This apparent contradiction might be the reason because the algorithm fails in classifying them. In any case, this topic requires of further research, counting on more data and inspecting other approaches that can better fit the case at hand, since ML practice has a great exploratory component. Nevertheless, this work may enlighten the application of these promising methodologies to the adoption of new transportation technologies. Since its potential lies in knowing if a person is prone to adopt, it may become an assistance for multiple agents, especially private companies that are looking for a reliable procedure to classify potential clients for whom they have only limited information.

5. Future Work

Future work planned for completion involves taking the methodologies applied in this research further by focusing on the areas of improvement detected. This task begins by carrying out a third survey design leading to a third wave of data collection. For example, it has been detected that the results corresponding to the role of propulsion costs and, especially, of the outer circle are not completely coherent. For the former, work is already underway on a design with improved attribute levels that can more clearly highlight the differences between the electric/gasoline vehicle capabilities. For the second, the way to improve the question that collects information about the individual's social network is also under study –it may be sufficient to offer clearer instructions for its completion. Finally, since attitudinal aspects have proved to be of great importance, more attention will be paid to this section of the questionnaire, which will be expanded.

With these new data, there are two straightforward lines of research. Firstly, to replicate the substitution-diffusion structure, trying to find a diffusion function that better adjust the data and, therefore, lead to better predictions. This is a path that offers many possibilities given the variety of possible approaches, which is why it is interesting. Secondly, an improvement of the non-parametric approach is also direct. Given the spectacular advance of the machine learning field, new tools have emerged since that part of the research was carried out, allowing the exploration of literally dozens of models in a much more simplified and inexpensive way.

The third line is more ambitious and will probably lead to future research for the author in the next few years. If both above-mentioned investigations are successful, then it would make sense to modify the substitution-diffusion paradigm, replacing discrete choice models with machine learning methodologies. This is a completely unexplored idea that responds to a great challenge: to bring together disaggregated and aggregated perspectives efficiently that leads to the best possible results.

Finally, perceptions and attitudes have been shown to be drivers of adoption, and deserve more attention. Hybrid choice models are best suited to bring out and explain these aspects, so further exploration seems a reasonable idea. However, perhaps the best use is to use advanced classification methodologies that connect in some way these underlying individual principles with purchasing desires; not from a marketing perspective, but rather a study of consumer behavior in the purest microeconomic sense.

6. Conclusions

This research presents two advanced methodologies that make use of real data to evaluate the adoption of the EVs in the State of Maryland. The first consists of a disaggregated substitution model that considers social influence and social conformity, which is then embedded in a diffusion model to predict EVs sales. The second, in contrast, relies on non-parametric ML techniques for the classification of potential EV purchasers. Both make use of data collected through a SC experiment specifically designed to capture the inclination of users towards EVs.

The first of these two approaches is based on the three pillars: substitution, diffusion, and dynamism. In order to consider substitution, an initial demand model including variables related to SoC and to the social network of the respondent has been estimated. It gathered the subjacent inclinations of individuals to substitute ICVs vehicle by EVs. Then, it has been integrated into a Bass diffusion model which finally yielded future market shares, i.e the spread of this technology over time among society. The dynamic aspect is provided by time-dependent variables, present in both the utility functions and the Bass model.

SoC is a relevant aspect in diffusion because people around us, such as family members, friends, colleagues, or even people that we do not know, influence our behavior and decisions, directly or indirectly. This concept is brought into this research by providing to the individuals feedback from one person they know. This person belongs to either an *Inner* or an *Outer*circle. The purpose is to explore explicitly the impact of the size and the nature of these relationships. This required the development of several elements, the main one being a new SC experiment to collect real data.

Regarding the second approach of this research, it makes use of the same data structure to predict the adoption of the EV, yet abandoning the agent's perspective and making use of non-parametric ML techniques. It aims to classify potential EV purchasers into *Adopters/Non adopters*. Namely, the performance of three methods of different nature is compared: Random Forests, Support Vector Machines, and Artificial Neural Networks. Secondary objectives are to identify the key factors in the adoption of the EV, as well as to determine the common characteristics of the individuals missclasified. For the former, a preliminary Random Forest is performed. For the latter, a hierarchical cluster analysis is applied, followed by a *K-means* classification and, finally, exploratory data analysis of the resulting groups. This part of the research makes use of the same data structure than the previous one, although incorporating the Pilot data.

6.1 Summary of conclusions

The results of the substitution sub-model (Table 3.11) show that most vehicle attributes are significant, with the exception of the *Fast charging time* and the *Propulsion costs* of the gasoline alternative. Also that the aspects related to the direct cost of the vehicle, i.e. *Price* and *Tax deduction* seem to be of special importance. With respect to the SoC variables, the effect *Number of EVs sold*, a measure of the tendency to conform to others, is not only positive and significant, but also close in magnitude to that of the tax deduction. Although only for those individuals who declared that they are currently able to charge an EV at home. Secondly, the size of the inner circle plays a role in the choice of the EV; the larger the group, the more probable is to choose it. This would confirm the idea that more connections lead to more information received on EVs, and that it has a significant effect, positive in this case. In contrast, the sign of the size of the outer circle is negative. Nevertheless, this result may not be completely reliable in light of the information collected on this aspect.

The effect of the feedback received in some of the scenarios is heterogeneous and varies depending on its type and on the category of car the user is willing to purchase. For those that declared that they would buy a *Compact* vehicle, only negative feedback on the need to change activities matters, whether it comes from a person of the inner or the outer circle. However, for the interviewees that declared that they would buy a *Midsize* one, it is more important the opinion received on charging it. As pointed out in the survey description, a *Midsize* vehicle refers to a family car. Thus, it is reasonable to think that people who wish to purchase a vehicle of this category are probably parents, who are probably concerned about being able to complete a tour of the *home-school-work-shopping* kind. Therefore, the need for sufficient autonomy is relevant. Approval from others is also important for these individuals, although only that of members of the outer circle, such as neighbours or coworkers. Finally, all aspects, charging, activities, and approval, are significant for potential purchasers of *Large* vehicles, but only if it is expressed in positive terms. In summary, the results evidence that not all kinds of feedback are of importance, and that its negativity/positivity, as well as its source, are relevant.

On the other hand, the results of the diffusion model that takes into account SoC present a more moderate sales evolution than the model that does not. The pace of the spread of this technology is slower and, therefore, the market is emptied later. Taking as a reference the number of households in the State of Maryland, this occurs around the year 2035, five years later than the model predictions that only consider vehicle attributes. This leads to the conclusion that although the characteristics of the vehicle are relevant, so is the number of social connections of the individual, as well as the information obtained from them. The combination of this two spheres modifies the decision-making process at a disaggregated level, consequently impacting the diffusion of the electric vehicle in aggregate terms.

As for the results obtained in the research carried out under a non-parametric approach, they show that the most relevant features are the county in which the users live, the type of engine (electric or not) of the next vehicle to be acquired, some vehicle characteristics, and several attitudinal variables. Since there are not special differences among the counties of the State of Maryland in terms of power grid or charging infrastructure, the effect of the county may hide an income effect. Some of these counties are among the richest in the U.S. and they evidence a clear geographical income distribution. Among the vehicle characteristics, the most relevant seems to be the income tax deduction that the U.S. government provides when buying an EV. Considering that the fourth most important variable is the price of the vehicle, it can be concluded that all the elements that gravitate around the purchase cost are fundamental in the inclination of individuals to adopt this technology. However, other vehicle attributes are also capital, such as the range and the time of fast charging, as well as the existence of charging infrastructure in the household. This is valuable information for the automotive and power industry since these are precisely some of the barriers highlighted by users and researchers for a wide-scale implementation of this technology. Some of these findings are also important for the public administration, beyond economic incentives. The attitude of individuals towards the environment and towards technology is relevant, too. Although these aspects are obviously inherent in each person, fostering social awareness of environmental care would boost the EV market.

Interestingly, these top variables were clearly different in nature, i.e. socioeconomic, attitudinal and social, and attributes-related. Thus, three sub-models were retrained using one of these groups at a time, with the objective of identifying the importance of each cluster. The results confirm that counting only with limited information dramatically decreases the general accuracy, as well as the rate of true positives, especially if only the attitudinal variables or those related to vehicle attributes are included. Therefore, it becomes clear that the synergy of all the variables is what produces a better classification, and that it is not convenient to rely only on aspects specific to the individual.

On the other hand, there are no marked differences in the performance of the models estimated, although SVM with polynomial kernel slightly stands out from the others. It provides better accuracy, highest Sensitivity, Specificity, and Cohen's Kappa statistic, as well as the highest average area under the ROC curves. However, RF was the most robust method thanks to the aggregation and decorrelation of trees that is performed in its estimation. In addition, the computing time of RF was significantly inferior to the other techniques. In the same vein, ANN were much more complex to code and estimate, and took significantly longer computing time, which did not translate in the end into better results.

6.2 Main findings and hypotheses

In summary, the main findings related to the first of the approaches, the substitution-diffusion model, are:

- The size of the Inner circle is significant, as well as the feedback about aspects related to the EV or the approval of others. Moreover, the effect of the information received is heterogeneous, and depends on the type of vehicle, whether it is positive or negative, and whether it comes from the inner or outer social circle. The existence of these groups and their influence on the decision-making process was the main hypothesis of this research.
- The forecast rate of adoption is reduced when social aspects come into play. The top of the market is reached five years later than predicted by the model that does not consider SoC. Therefore, the spread of the EV technology is

influenced by the presence of social factors, which was the second hypothesis.

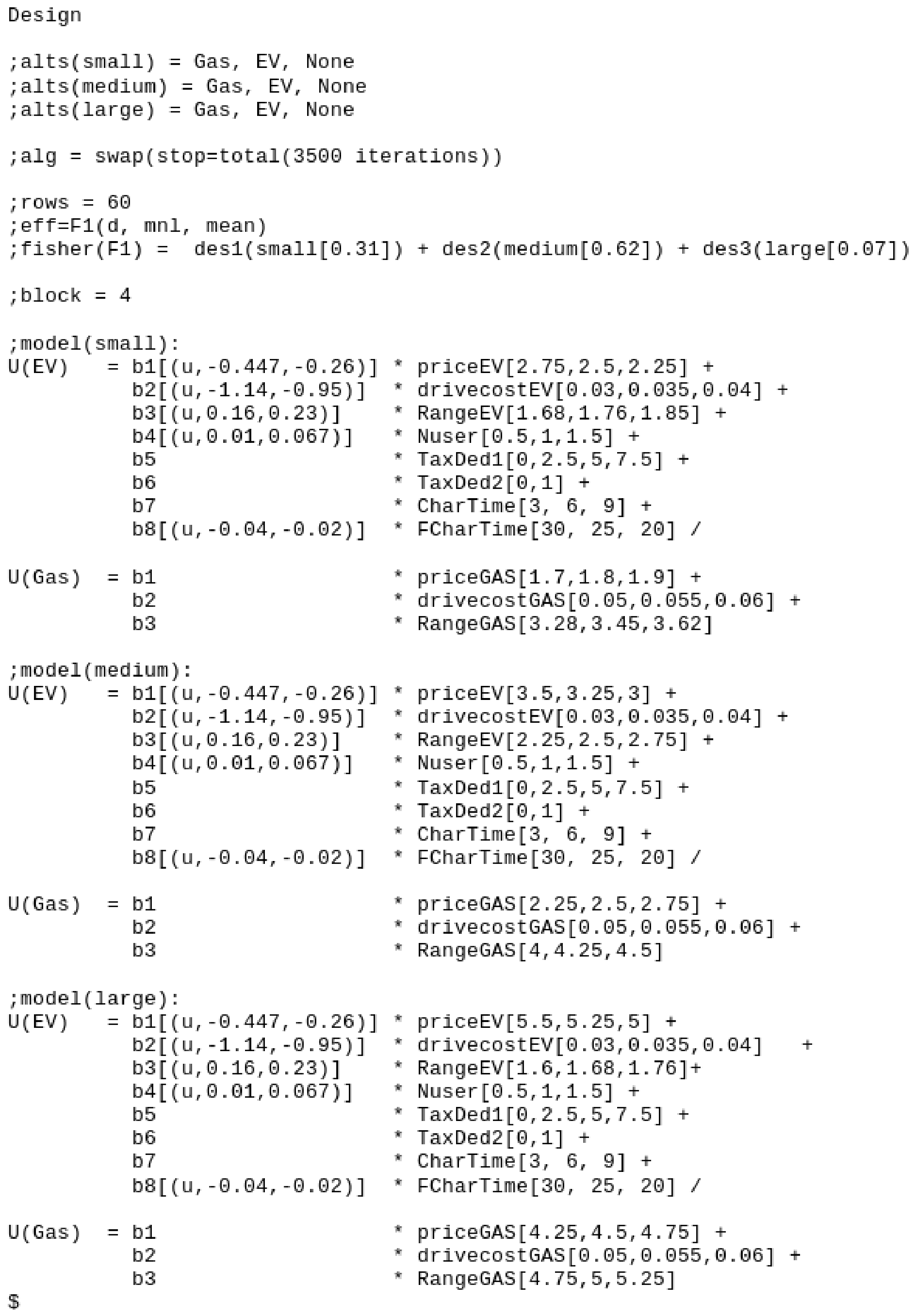
On the other hand, the main findings regarding the use of ML techniques to classify potential EV purchasers are:

- Cost-related and attribute-related variables are fundamental in the individuals' inclination to adopt the EV, but attitudes play an important role as well.

- Counting only with limited information dramatically decreases the prediction power. It is the synergy of all the elements what produces a better classification.

However, if there was only one aspect of this research that stands out with respect to the prediction of the EV diffusion (and probably of any technology), this would be that it is erroneous to rely solely on an economic rationale. The results of this research show that pure cost-benefit analysis does not serve to explain the actual adoption of the EV. Attitudes and social relations are so relevant in the decision-making process that they overcome the traditional *Homo Economicus* approach. Diffusion is heavily characterized by elements that are intrinsic to us: our beliefs and our attitudes towards society, technology, and environment. And this set of elements is, in turn, clearly influenced by the experiences, beliefs and opinions of those with whom we are connected.

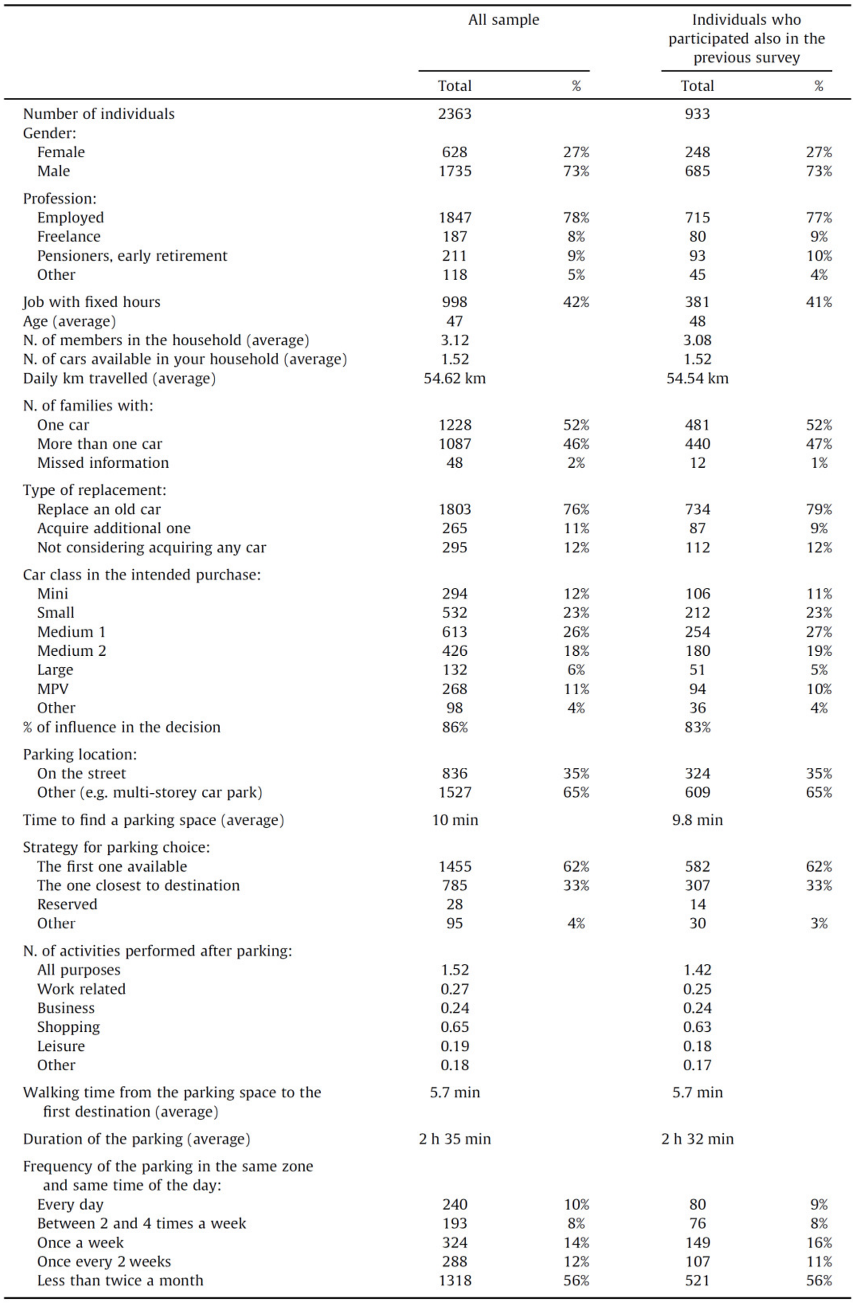
Appendix A: Ngene syntax



```
Design

;alts(small) = Gas, EV, None
;alts(medium) = Gas, EV, None
;alts(large) = Gas, EV, None

;alg = swap(stop=total(3500 iterations))

;rows = 60
;eff=F1(d, mnl, mean)
;fisher(F1) =  des1(small[0.31]) + des2(medium[0.62]) + des3(large[0.07])

;block = 4

;model(small):
U(EV)    = b1[(u,-0.447,-0.26)] * priceEV[2.75,2.5,2.25] +
           b2[(u,-1.14,-0.95)]  * drivecostEV[0.03,0.035,0.04] +
           b3[(u,0.16,0.23)]    * RangeEV[1.68,1.76,1.85] +
           b4[(u,0.01,0.067)]   * Nuser[0.5,1,1.5] +
           b5                   * TaxDed1[0,2.5,5,7.5] +
           b6                   * TaxDed2[0,1] +
           b7                   * CharTime[3, 6, 9] +
           b8[(u,-0.04,-0.02)]  * FCharTime[30, 25, 20] /

U(Gas)   = b1                   * priceGAS[1.7,1.8,1.9] +
           b2                   * drivecostGAS[0.05,0.055,0.06] +
           b3                   * RangeGAS[3.28,3.45,3.62]

;model(medium):
U(EV)    = b1[(u,-0.447,-0.26)] * priceEV[3.5,3.25,3] +
           b2[(u,-1.14,-0.95)]  * drivecostEV[0.03,0.035,0.04] +
           b3[(u,0.16,0.23)]    * RangeEV[2.25,2.5,2.75] +
           b4[(u,0.01,0.067)]   * Nuser[0.5,1,1.5] +
           b5                   * TaxDed1[0,2.5,5,7.5] +
           b6                   * TaxDed2[0,1] +
           b7                   * CharTime[3, 6, 9] +
           b8[(u,-0.04,-0.02)]  * FCharTime[30, 25, 20] /

U(Gas)   = b1                   * priceGAS[2.25,2.5,2.75] +
           b2                   * drivecostGAS[0.05,0.055,0.06] +
           b3                   * RangeGAS[4,4.25,4.5]

;model(large):
U(EV)    = b1[(u,-0.447,-0.26)] * priceEV[5.5,5.25,5] +
           b2[(u,-1.14,-0.95)]  * drivecostEV[0.03,0.035,0.04]   +
           b3[(u,0.16,0.23)]    * RangeEV[1.6,1.68,1.76]+
           b4[(u,0.01,0.067)]   * Nuser[0.5,1,1.5] +
           b5                   * TaxDed1[0,2.5,5,7.5] +
           b6                   * TaxDed2[0,1] +
           b7                   * CharTime[3, 6, 9] +
           b8[(u,-0.04,-0.02)]  * FCharTime[30, 25, 20] /

U(Gas)   = b1                   * priceGAS[4.25,4.5,4.75] +
           b2                   * drivecostGAS[0.05,0.055,0.06] +
           b3                   * RangeGAS[4.75,5,5.25]
$
```

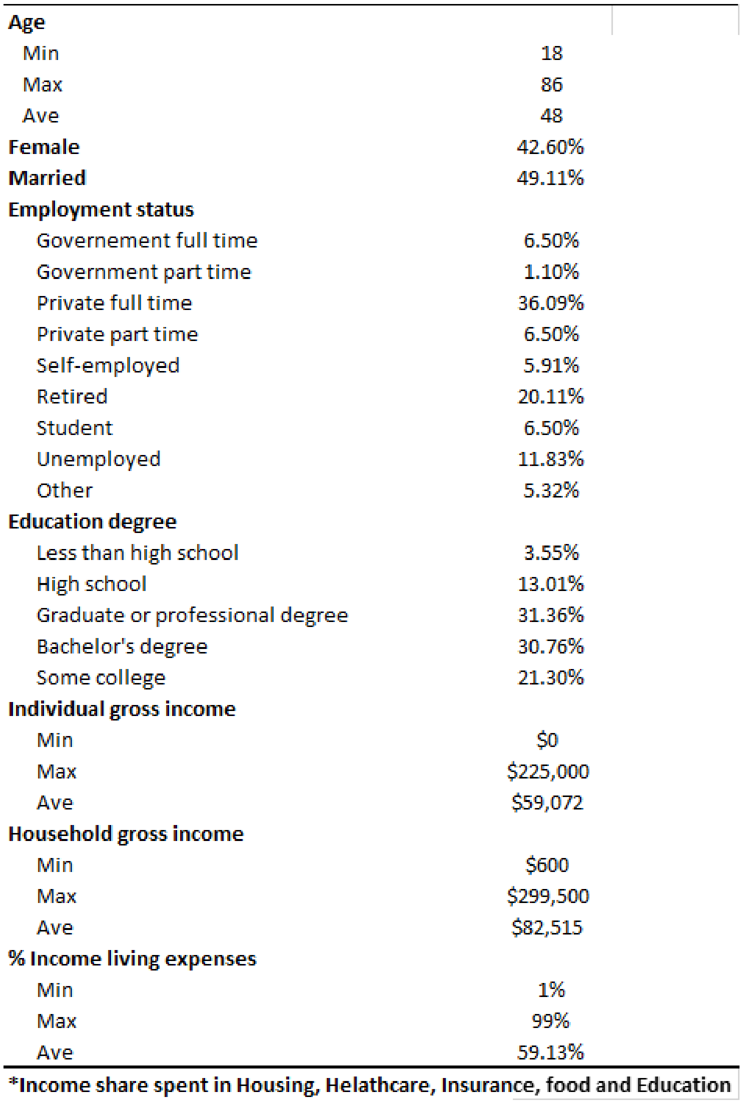
Appendix B: Sample sociodemographics

Table B.1: Denmark study. Source [25]



	All sample		Individuals who participated also in the previous survey	
	Total	%	Total	%
Number of individuals	2363		933	
Gender:				
Female	628	27%	248	27%
Male	1735	73%	685	73%
Profession:				
Employed	1847	78%	715	77%
Freelance	187	8%	80	9%
Pensioners, early retirement	211	9%	93	10%
Other	118	5%	45	4%
Job with fixed hours	998	42%	381	41%
Age (average)	47		48	
N. of members in the household (average)	3.12		3.08	
N. of cars available in your household (average)	1.52		1.52	
Daily km travelled (average)	54.62 km		54.54 km	
N. of families with:				
One car	1228	52%	481	52%
More than one car	1087	46%	440	47%
Missed information	48	2%	12	1%
Type of replacement:				
Replace an old car	1803	76%	734	79%
Acquire additional one	265	11%	87	9%
Not considering acquiring any car	295	12%	112	12%
Car class in the intended purchase:				
Mini	294	12%	106	11%
Small	532	23%	212	23%
Medium 1	613	26%	254	27%
Medium 2	426	18%	180	19%
Large	132	6%	51	5%
MPV	268	11%	94	10%
Other	98	4%	36	4%
% of influence in the decision	86%		83%	
Parking location:				
On the street	836	35%	324	35%
Other (e.g. multi-storey car park)	1527	65%	609	65%
Time to find a parking space (average)	10 min		9.8 min	
Strategy for parking choice:				
The first one available	1455	62%	582	62%
The one closest to destination	785	33%	307	33%
Reserved	28		14	
Other	95	4%	30	3%
N. of activities performed after parking:				
All purposes	1.52		1.42	
Work related	0.27		0.25	
Business	0.24		0.24	
Shopping	0.65		0.63	
Leisure	0.19		0.18	
Other	0.18		0.17	
Walking time from the parking space to the first destination (average)	5.7 min		5.7 min	
Duration of the parking (average)	2 h 35 min		2 h 32 min	
Frequency of the parking in the same zone and same time of the day:				
Every day	240	10%	80	9%
Between 2 and 4 times a week	193	8%	76	8%
Once a week	324	14%	149	16%
Once every 2 weeks	288	12%	107	11%
Less than twice a month	1318	56%	521	56%

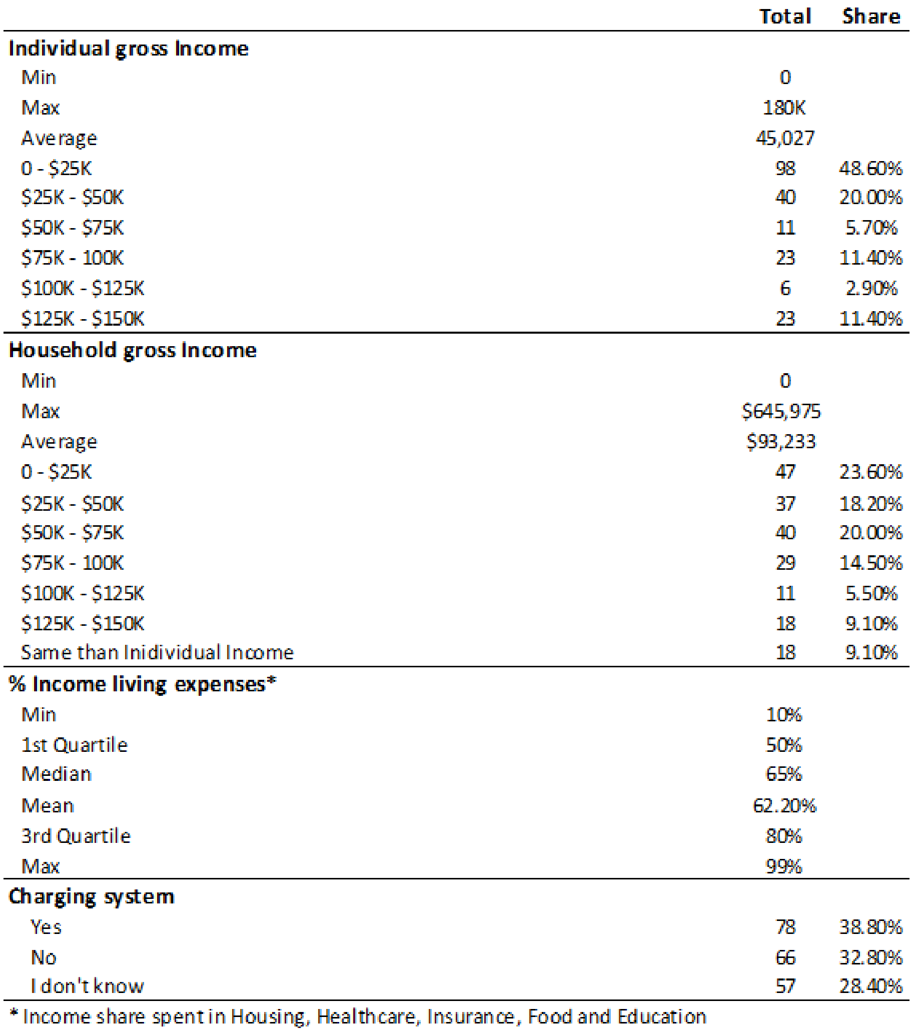
Table B.2: U.S. study, *Release* data. Part I



Age	
Min	18
Max	86
Ave	48
Female	42.60%
Married	49.11%
Employment status	
Governement full time	6.50%
Government part time	1.10%
Private full time	36.09%
Private part time	6.50%
Self-employed	5.91%
Retired	20.11%
Student	6.50%
Unemployed	11.83%
Other	5.32%
Education degree	
Less than high school	3.55%
High school	13.01%
Graduate or professional degree	31.36%
Bachelor's degree	30.76%
Some college	21.30%
Individual gross income	
Min	$0
Max	$225,000
Ave	$59,072
Household gross income	
Min	$600
Max	$299,500
Ave	$82,515
% Income living expenses	
Min	1%
Max	99%
Ave	59.13%

***Income share spent in Housing, Helathcare, Insurance, food and Education**

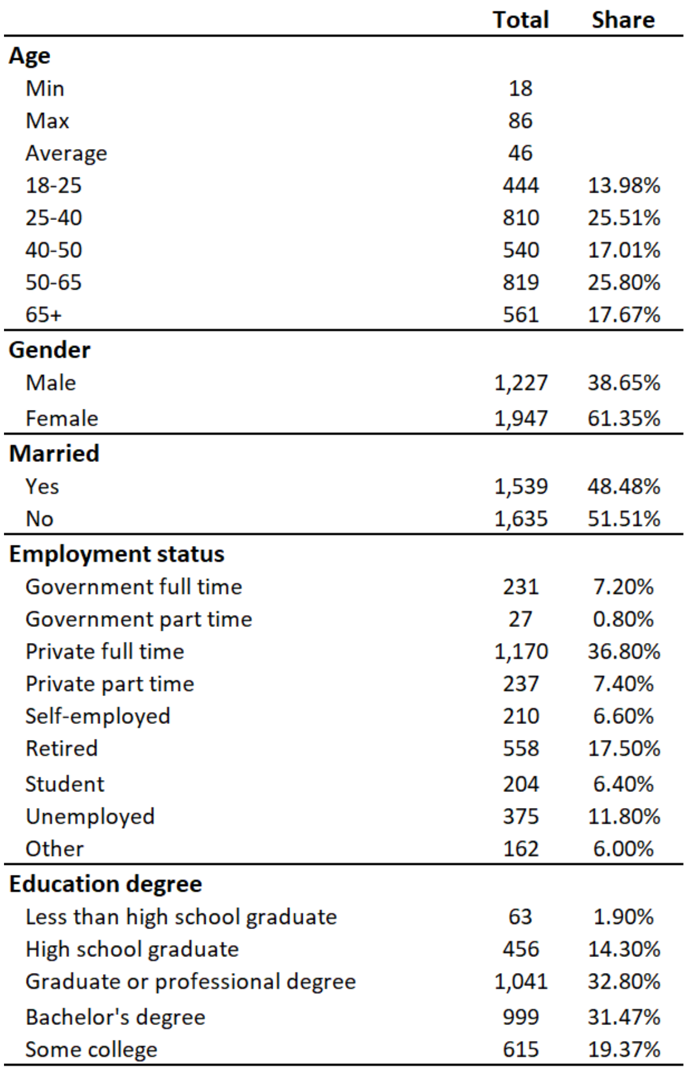
Table B.3: U.S. study, *Release* data. Part II



	Total	Share
Individual gross Income		
Min	0	
Max	180K	
Average	45,027	
0 - $25K	98	48.60%
$25K - $50K	40	20.00%
$50K - $75K	11	5.70%
$75K - 100K	23	11.40%
$100K - $125K	6	2.90%
$125K - $150K	23	11.40%
Household gross Income		
Min	0	
Max	$645,975	
Average	$93,233	
0 - $25K	47	23.60%
$25K - $50K	37	18.20%
$50K - $75K	40	20.00%
$75K - 100K	29	14.50%
$100K - $125K	11	5.50%
$125K - $150K	18	9.10%
Same than Inidividual Income	18	9.10%
% Income living expenses*		
Min	10%	
1st Quartile	50%	
Median	65%	
Mean	62.20%	
3rd Quartile	80%	
Max	99%	
Charging system		
Yes	78	38.80%
No	66	32.80%
I don't know	57	28.40%

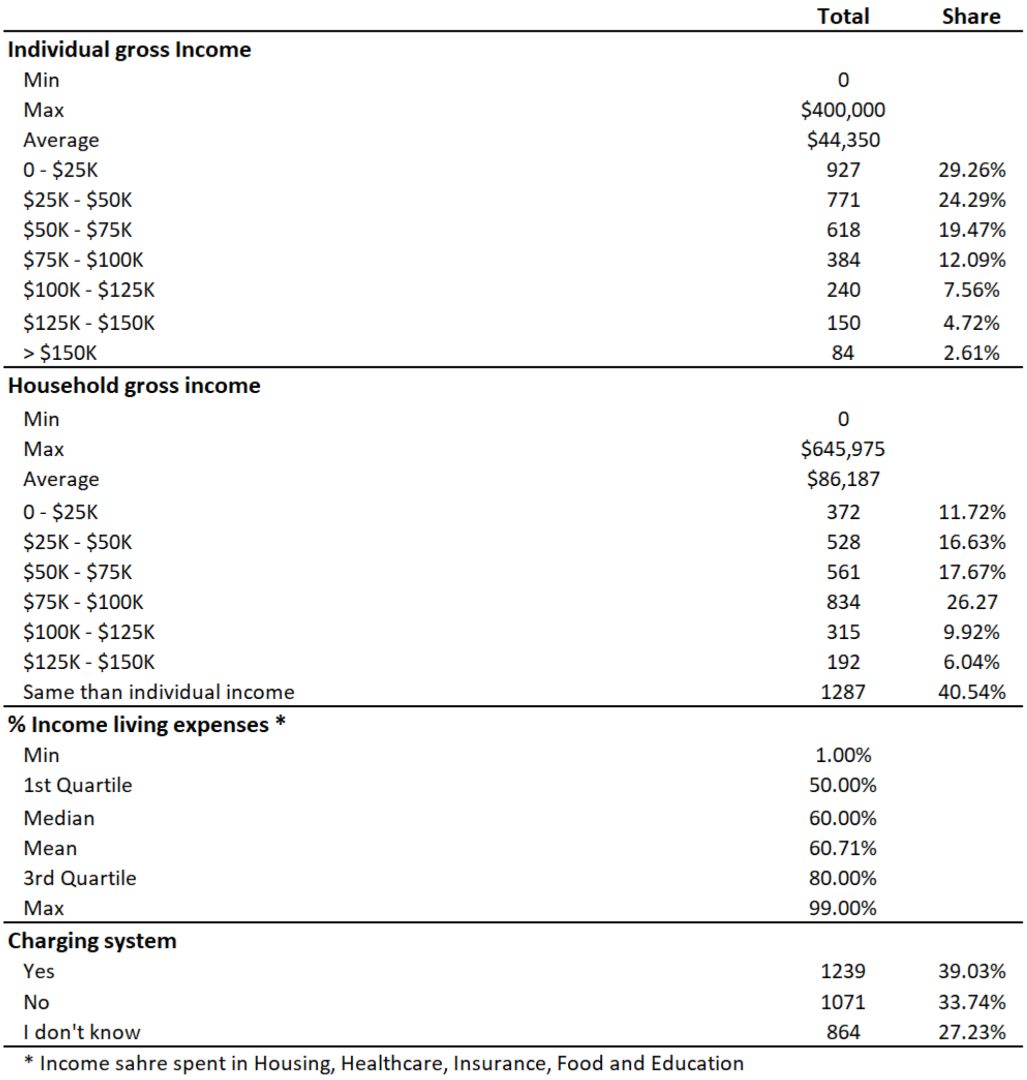
* Income share spent in Housing, Healthcare, Insurance, Food and Education

Table B.4: U.S. study, *Pilot* and *Release* data.



	Total	Share
Age		
Min	18	
Max	86	
Average	46	
18-25	444	13.98%
25-40	810	25.51%
40-50	540	17.01%
50-65	819	25.80%
65+	561	17.67%
Gender		
Male	1,227	38.65%
Female	1,947	61.35%
Married		
Yes	1,539	48.48%
No	1,635	51.51%
Employment status		
Government full time	231	7.20%
Government part time	27	0.80%
Private full time	1,170	36.80%
Private part time	237	7.40%
Self-employed	210	6.60%
Retired	558	17.50%
Student	204	6.40%
Unemployed	375	11.80%
Other	162	6.00%
Education degree		
Less than high school graduate	63	1.90%
High school graduate	456	14.30%
Graduate or professional degree	1,041	32.80%
Bachelor's degree	999	31.47%
Some college	615	19.37%

Table B.5: U.S. study, *Pilot* and *Release* data.



	Total	Share
Individual gross Income		
Min	0	
Max	$400,000	
Average	$44,350	
0 - $25K	927	29.26%
$25K - $50K	771	24.29%
$50K - $75K	618	19.47%
$75K - $100K	384	12.09%
$100K - $125K	240	7.56%
$125K - $150K	150	4.72%
> $150K	84	2.61%
Household gross income		
Min	0	
Max	$645,975	
Average	$86,187	
0 - $25K	372	11.72%
$25K - $50K	528	16.63%
$50K - $75K	561	17.67%
$75K - $100K	834	26.27
$100K - $125K	315	9.92%
$125K - $150K	192	6.04%
Same than individual income	1287	40.54%
% Income living expenses *		
Min	1.00%	
1st Quartile	50.00%	
Median	60.00%	
Mean	60.71%	
3rd Quartile	80.00%	
Max	99.00%	
Charging system		
Yes	1239	39.03%
No	1071	33.74%
I don't know	864	27.23%

* Income sahre spent in Housing, Healthcare, Insurance, Food and Education

Bibliography

[1] Asch, S. E., & Guetzkow, H. (1951). Effects of group pressure upon the modification and distortion of judgments. *Documents of gestalt psychology*, 222-236.

[2] Axsen, J., & Kurani, K. S. (2012). Interpersonal influence within car buyers' social networks: applying five perspectives to plug-in hybrid vehicle drivers. *Environment and Planning A, 44(5)*, 1047-1065.

[3] Ayyadi, S., & Maaroufi, M. (2018, October). Diffusion Models For Predicting Electric Vehicles Market in Morocco. *In 2018 International Conference and Exposition on Electrical And Power Engineering (EPE)* (pp. 0046-0051). IEEE.

[4] Baddeley, M. (2010). Herding, social influence and economic decision-making: socio-psychological and neuroscientific analyses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1538), 281-290.

[5] Barton, B., & Schütte, P. (2017). Electric vehicle law and policy: a comparative analysis. *Journal of Energy & Natural Resources Law, 35(2)*, 147-170.

[6] Bass, F. M. (1969). A new product growth for model consumer durables. *Management science, 15(5)*, 215-227.

[7] Becker, T. A., Sidhu, I., & Tenderich, B. (2009). Electric vehicles in the United States: a new model with forecasts to 2030. *Center for Entrepreneurship and Technology, University of California, Berkeley*, 24.

[8] Ben-Akiva, M. E., Lerman, S. R., & Lerman, S. R. (1985). *Discrete choice analysis: theory and application to travel demand (Vol. 9)*. MIT press.

[9] Ben-Akiva, M., Bolduc, D., & Walker, J. (2001). Specification, identification and estimation of the logit kernel (or continuous mixed logit) model.

[10] Berkeley, N., Bailey, D., Jones, A., & Jarvis, D. (2017). Assessing the transition towards Battery Electric Vehicles: A Multi-Level Perspective on drivers of, and barriers to, take up. *Transportation Research part A: policy and practice, 106*, 320-332.

[11] Bezdek, James & Hathaway, R.J.. (2002). VAT: A tool for visual assessment of (cluster) tendency. Proceedings of the International Joint Conference on Neural Networks. 3. 2225 - 2230. 10.1109/IJCNN.2002.1007487.

[12] Biel, A., Andersson, M., Hedesström, M., Jansson, M., Sundblad, E. L., & Gärling, T. (2010). *Social Influence in Stockmarkets: A Conceptual Analysis of Social Influence Processes in Stock Markets* (No. 2010/13). Sustainable Investment Research Platform.

[13] Bliemer, M. C., & Rose, J. M. (2005). Efficiency and sample size requirements for stated choice studies.

[14] Bliemer, M. C., & Rose, J. M. (2005). Efficient designs for alternative specific choice experiments.

[15] Bliemer, M. C., & Rose, J. M. (2009). Efficiency and sample size requirements for stated choice experiments (No. 09-2421).

[16] III, H. A., & Lusk, A. C. (2016). Addressing electric vehicle (EV) sales and range anxiety through parking layout, policy and regulation. Transportation Research Part A: Policy and Practice, 83, 63-73.

[17] Heather Brutz, Allison Carr, Brian Lips, Autumn Proudlove, David Sarkisian, & Achyut Shrestha. (Februrary 2019) *50 States of Electric Vehicles.* NC Clean Energy Technology Center.

[18] Bureau, U. C. (n.d.). Historical Households Tables. The United States Census Bureau. Retrieved March 2, 2020, from https://www.census.gov/data/tables/time-series/demo/families/households.html

[19] Calastri, C., Hess, S., Daly, A., Carrasco, J. A., & Choudhury, C. (2018). Modelling the loss and retention of contacts in social networks: The role of dyad-level heterogeneity and tie strength. *Journal of choice modelling*, 29, 63-77.

[20] Center for Automotive Research. (2011, January). Deployment Rollout Estimate of Electric Vehicles 2011-2015. Ann Arbor: Hill, K. & Cregger, J. Retrieved from http://www.cargroup.org/pdfs/deployment.pdf

[21] Carlsson, F., & Martinsson, P. (2003). Design techniques for stated preference methods in health economics. *Health economics, 12(4)*, 281-294.

[22] Carrillo Álvarez, E., & Riera Romaní, J. (2017). Measuring social capital: further insights. Gaceta sanitaria, 31, 57-61.

[23] Cecere, G., Le Guel, F., & Rochelandet, F. (2017). Crowdfunding and social influence: an empirical investigation. *Applied economics, 49(57)*, 5802-5813.

[24] Chávez, Ó., Carrasco, J. A., & Tudela, A. (2018). Social activity-travel dynamics with core contacts: evidence from a two-wave personal network data. *Transportation Letters, 10(6)*, 333-342.

[25] Cherchi E (2017). A stated choice experiment to measure the effect of informational and normative conformity in the preference for electric vehicles. *Transportation Research Part A: Policy and Practice 100*, 88-104

[26] ChoiceMetrics (2014) Ngene 1.1.2 *User Manual & Reference Guide*, Australia.

[27] Cialdini, R. B. (2009). *Influence: Science and practice* (Vol. 4). Boston: Pearson education.

[28] Cialdini, R. B. (2005). Basic social influence is underestimated. *Psychological inquiry, 16(4)* 158-161.

[29] Cialdini, R. B. (2007). Descriptive social norms as underappreciated sources of social control. *Psychometrika, 72(2)*, 263.

[30] Cialdini RB , Goldstein NJ (2004). Social influence: compliance and conformity. *Annual Review of Psychology 55*, 294-310.

[31] Cialdini, R. B., & Trost, M. R. (1998). Social influence: Social norms, conformity and compliance.

[32] Pontes, J. (2019, March 4). Electric Vehicle Sales Jump 67% In Europe — CleanTechnica EV Sales Report. Retrieved from https://cleantechnica.com/2019/03/04/electric-vehicle-sales-jump-67-in-europe-cleantechnicas-europe-ev-sales-report/

[33] Collins, C., Chambers, S.(2005) Psychological and Situational Influences on Commuter-Transport-Mode Choice. *Environment and Behavior, 37(5)*, 640–661.

[34] Cook, R. D., & Nachtrheim, C. J. (1980). A comparison of algorithms for constructing exact D-optimal designs. *Technometrics, 22(3)*, 315-324.

[35] Coplon-Newfield, G. (2018, August). Automakers Are Still Not advertising Electric Cars. Retrieved from: https://www.sierraclub.org/compass/2018/08/automakers-are-still-not-advertising-electric-cars

[36] Crutchfield, R. S. (1955). Conformity and character. *American Psychologist, 10(5)*, 191.

[37] de Kleijn, M. S. (2015). *The influences of an individual's social network on the choice of travelling by public transport* (Bachelor's thesis).

[38] de Rubens, G. Z. (2019). Who will buy electric vehicles after early adopters? Using machine learning to identify the electric vehicle mainstream market. *Energy, 172*, 243-254.

[39] Dimitropoulos, A., Rietveld, P., & Van Ommeren, J. N. (2013). Consumer valuation of changes in driving range: A meta-analysis. *Transportation Research Part A: Policy and Practice, 55*, 27-45.

[40] Duran, M. (2019) Historical trend of annual sales of plug-in electric passenger cars in the U.S. by type of powertrain (2010-2018). Retrieved from: https://commons.wikimedia.org/wiki/File: US_PEV_Sales_2010_by_PHEV_vs_BEV.png

[41] Eppstein, M. J., Grover, D. K., Marshall, J. S., & Rizzo, D. M. (2011). An agent-based model to study market penetration of plug-in hybrid electric vehicles. *Energy Policy, 39(6)*, 3789-3802.

[42] Ernst, A., & Briegel, R. (2017). A dynamic and spatially explicit psychological model of the diffusion of green electricity across Germany. *Journal of Environmental Psychology, 52*, 183-193.

[43] Fukushima, A., Yano, T., Imahara, S., Aisu, H., Shimokawa, Y., & Shibata, Y. (2018). Prediction of energy consumption for new electric vehicle models by machine learning. *IET Intelligent Transport Systems, 12(9)*, 1174–1180. https://doi.org/10.1049/iet-its.2018.5169

[44] Glerum, A., Stankovikj, L., Thémans, M., & Bierlaire, M. (2013). Forecasting the demand for electric vehicles: accounting for attitudes and perceptions. *Transportation Science, 48(4)*, 483-499.

[45] Gnann, T., Stephens, T. S., Lin, Z., Plötz, P., Liu, C., & Brokate, J. (2018). What drives the market for plug-in electric vehicles?-A review of international PEV market diffusion models. *Renewable and Sustainable Energy Reviews, 93*, 158-164.

[46] , D., & Plötz, P. (2019). Machine learning estimates of plug-in hybrid electric vehicle utility factors. Transportation Research Part D: Transport and Environment, 72(April), 36–46. https://doi.org/10.1016/j.trd.2019.04.008

[47] Goetzke, F., & Rave, T. (2015). Automobile access, peer effects and happiness. *Transportation, 42(5)*, 791-805.

[48] Goldstein, N. J., & Cialdini, R. B. (2007). Using social norms as a lever of social influence. *The science of social influence: Advances and future progress*, 167-192.

[49] Gompertz, B. (1825). XXIV. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. In a letter to Francis Baily, Esq. FRS &c. *Philosophical transactions of the Royal Society of London, (115)*, 513-583.

[50] Hensher, D. A., & Greene, W. H. (2001). The mixed logit model: The state of practice and warnings for the unwary. *Institute of Transport Studies*, 1-39.

[51] Hensher, D.A., Rose, J.M. & Greene, W.H. (2005). *Applied Choice Analysis: A Primer*, Cambridge University Press, UK.

[52] Hensher, D. A., & Button, K. J. (Eds.). (2007). *Handbook of transport modelling.* Emerald Group Publishing Limited.

[53] Higgins, A., Paevere, P., Gardner, J., & Quezada, G. (2012). Combining choice modelling and multi-criteria analysis for technology diffusion: An application to the uptake of electric vehicles. *Technological Forecasting and Social Change, 79(8)*, 1399-1412.

[54] Hoen, A., & Koetse, M. J. (2014). A choice experiment on alternative fuel vehicle preferences of private car owners in the Netherlands. *Transportation Research Part A: Policy and Practice, 61*, 199-215.

[55] Hu, X., Li, S. E., & Yang, Y. (2016). Advanced Machine Learning Approach for Lithium-Ion Battery State Estimation in Electric Vehicles. *IEEE Transactions on Transportation Electrification, 2(2)*, 140–149. https://doi.org/10.1109/TTE.2015.2512237

[56] Huang, X., Tan, Y., & He, X. (2011). An intelligent multifeature statistical approach for the discrimination of driving conditions of a hybrid electric vehicle. *IEEE Transactions on Intelligent Transportation Systems, 12(2)*, 453–465. https://doi.org/10.1109/TITS.2010.2093129

[57] Huber, J. & K. Zwerina (1996.) The Importance of Utility Balance in Efficient Choice Designs. *Journal of Marketing Research 33*, 307-317.

[58] Trong Hue, D., Nguyen, L., Honf Tien, V., & Nguyen, T. (2018). Online Social Influence and Social Marketing. *Australia and New Zealand Marketing Academy Conference (ANZMAC)* At: University of Adelaide, Australia. September 2018.

[59] Hughes, S., Moreno, S., Yushimito, W. F., & Huerta-Cánepa, G. (2019). Evaluation of machine learning methodologies to predict stop delivery times from GPS data. *Transportation Research Part C: Emerging Technologies, 109*, 289-304.

[60] Irle, R. (2019, July). *Global EV Sales for 2018 – Final Results*. Retrieved from http://www.ev-volumes.com/country/total-world-plug-in-vehicle-volumes/

[61] Irle, R. (2019, July). *USA Plug-in Sales for the First Half of 2019*. Retrieved from http://www.ev-volumes.com/country/usa/

[62] Ito, T. A., Larsen, J. T., Smith, N. K., & Cacioppo, J. T. (1998). Negative information weighs more heavily on the brain: The negativity bias in evaluative categorizations. *Journal of personality and social psychology, 75(4)*, 887.

[63] Jahangiri, A., & Rakha, H. A. (2015). Applying Machine Learning Techniques to Transportation Mode Recognition Using Mobile Phone Sensor Data. *IEEE Transactions on Intelligent Transportation Systems, 16(5)*, 2406–2417. https://doi.org/10.1109/TITS.2015.2405759

[64] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, pp. 3-7). New York: springer.

[65] Jensen, A. F., Cherchi, E., & Mabit, S. L. (2013). On the stability of preferences and attitudes before and after experiencing an electric vehicle. *Transportation Research Part D: Transport and Environment, 25*, 24-32.

[66] Jensen, A. F., Cherchi, E., & de Dios Ortúzar, J. (2014). A long panel survey to elicit variation in preferences and attitudes in the choice of electric vehicles. *Transportation, 41(5)*, 973-993.

[67] Jensen, A., Cherchi, E., Mabit, S. & Ortúzar, J. de D. (2016) Predicting the potential market of electric vehicles. *Transportation Science 51*, 427-440.

[68] Jia, J. (2019). Analysis of Alternative Fuel Vehicle (AFV) Adoption Utilizing Different Machine Learning Methods: A Case Study of 2017 NHTS. *IEEE Access, 7*, 112726-112735.

[69] Jiang, Z., & Jain, D.(2014). A Generalized Norton–Bass Model for Multigeneration Diffusion. *Management Science 2012 58:10*, 1887-1897

[70] Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202.

[71] Jun, D. B., & Park, Y. S. (1999). A choice-based diffusion model for multiple generations of products. *Technological Forecasting and Social Change, 61(1)*, 45-58.

[72] Kallgren, C. A., Reno, R. R., & Cialdini, R. B. (2000). A focus theory of normative conduct: When norms do and do not affect behavior. *Personality and social psychology bulletin, 26(8)*, 1002-1012.

[73] Kanninen, B. J. (1993). Optimal experimental design for double-bounded dichotomous choice contingent valuation. *Land Economics, 69(2).*

[74] Kanninen, B.J. (1993b). Design of sequential experiments for CV studies, *Journal of Environmental Economics and Management, 25*, 1–11

[75] Kanninen, B.J. (2002). Optimal Design for Multinomial Choice Experiments. *Journal of Marketing Research 39*, 214-217.

[76] Kessels, R., Goos, P., & Vandebroek, M. (2006). A comparison of criteria to design efficient choice experiments. *Journal of Marketing Research, 43(3)*, 409-419.

[77] Kieckhäfer K, Volling T, & Spengler TS (2014). A hybrid simulation approach for estimating the market share evolution of electric vehicles. *Transportation Sci. 48(4)*, 651–670.

[78] Klasen, J. R., & Neumann, D. (2011). An agent-based method for planning innovations. *International Journal of Innovation and Sustainable Development*, 5(2-3), 159-184.

[79] Klinger, T. (2017). Moving from monomodality to multimodality? Changes in mode choice of new residents. *Transportation Research Part A: Policy and Practice, 104*, 221-237.

[80] Kormos, C., Gifford, R., & Brown, E. (2015). The influence of descriptive social norm information on sustainable transportation behavior: A field experiment. *Environment and Behavior, 47(5)*, 479-501.

[81] Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering, 160*, 3-24.

[82] Kuwano, M., Tsukai, M., & Matsubara, T. (2012, July). Analysis on promoting factors of electric vehicles with social conformity. *In 13th International Conference on Travel Behaviour Research* (pp. 15-20).

[83] Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *International Biometric Society, 33(1)*, 159–174.

[84] Lawson, Richard G., and Peter C. Jurs (2002). New Index for Clustering Tendency and Its Application to Chemical Problems. Research-article. World. May 1, 2002. https://pubs.acs.org/doi/pdf/10.1021/ci00065a010.

[85] Lee, H., Kim, S. G., Park, H. W., & Kang, P. (2014). Pre-launch new product demand forecasting using the Bass model: A statistical and machine learning-based approach. *Technological Forecasting and Social Change, 86*, 49-64.

[86] Liu, Z., Liu, Y., Meng, Q., & Cheng, Q. (2019). A tailored machine learning approach for urban transport network flow estimation. *Transportation Research Part C: Emerging Technologies, 108*, 130-150.

[87] "Light Duty Electric Drive Vehicles Monthly Sales Updates — Argonne National Laboratory." n.d. Accessed March 6, 2020. https://www.anl.gov/es/light-duty-electric-drive-vehicles-monthly-sales-updates.

[88] Louviere, J. J., & Hensher, D. A. (1983). Using discrete choice models with experimental design data to forecast consumer demand for a unique cultural event. *Journal of Consumer research, 10(3)*, 348-361.

[89] Louviere, J.J., & G. Woodworth (1983). Design and analysis of simulated consumer choice or allocation experiments: an approach based on aggregated data. *Journal of Marketing Research 20*, 350-367.

[90] Louviere, J.J., D.A. Hensher, & J.D. Swait (2000). *Stated Choice Methods—Analysis and Application*. Cambridge University Press, UK.

[91] Loveday, E. (2011, December 1). *Chevy Volt Sales Trump Nissan LEAF in November 2011. PlugInCars.* Retrieved from http://www.plugincars.com/chevy-volt-sales-nissan-leaf-november-2011-110656.html

[92] Lutsey, N., Searle, S., Chambliss, S., & Bandivadekar, A. (2015). Assessment of leading electric vehicle promotion activities in United States cities. *International Council on Clean Transportation.*

[93] MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297).

[94] Maness, M., Cirillo, C., & Dugundji, E. R. (2015). Generalized behavioral framework for choice models of social influence: Behavioral and data concerns in travel behavior. *Journal of transport geography*, 46, 137-150.

[95] Massiani, J., & Gohs, A. (2015). The choice of Bass model coefficients to forecast diffusion for innovative products: An empirical investigation for new automotive technologies. *Research in Transportation Economics*, 50, 17-28.

[96] Mehndiratta, S. (1996). *'Time-of-day effects in inter-city business travel'*, PhD Thesis, University of California, Berkeley.

[97] Meyer, Gereon & Dokic, Jadranka & Jürgens, Heike & Tobias, Diana. (2017). *Hybrid and Electric Vehicles - The Electric Drive Chauffeurs.* 10.13140/RG.2.2.30491.98084.

[98] Mollenhorst, G., Volker, B., & Flap, H. (2014). Changes in personal relationships: How social contexts affect the emergence and discontinuation of relationships. *Social Networks*, 37, 65-80.

[99] Morvinski, C., Amir, O., & Muller, E. (2017). "Ten Million Readers Can't Be Wrong!," or Can They? On the Role of Information About Adoption Stock in New Product Trial. *Marketing Science, 36(2)*, 290-300.

[100] Redondo, A. N., & Cagigas, A. P. (2015). Sales Forecast of Electric Vehicles. *Journal of Engineering and Architecture, 3(1)*, 79-88.

[101] Nealer, R. Reichmuth (2015). *Cleaner Cars from Cradle to Grave. How Electric Cars Beat Gasoline Cars on Lifetime Global Warming Emissions.* Union of Concerned Scientists. Online Report. November 2015

[102] Pike, S., & Lubell, M. (2018). The conditional effects of social influence in transportation mode choice. *Research in transportation economics, 68*, 2-10.

[103] Pike Research. (2011). *Electric Vehicle Geographic Forecasts.* Retrieved from http://www.pikeresearch.com/research/electric-vehicle-geographic-forecasts

[104] Pyo, T. H., Gruca, T. S., & Russell, G. J. (2017). *A New Bass Model Utilizing Social Network Data.*

[105] Qualtrics Research Core (Qualtrics, Provo, UT).

[106] Rasouli, S., & Timmermans, H. (2013). Incorporating Mechanisms of Social Adoption in Design and Analysis of Stated-Choice Experiments: Illustration and Application to Choice of Electric Cars. *Transportation Research Record, 2344(1)*, 10-19.

[107] Revelt, D., & Train, K. (2000). Customer-specific taste parameters and mixed logit: Households' choice of electricity supplier.

[108] Reynolds, K. J., Subašić, E., & Tindall, K. (2015). The problem of behaviour change: From social norms to an ingroup focus. *Social and Personality Psychology Compass, 9(1)*, 45-56.

[109] Riedmiller, M., & Braun, H. (1993). A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm. *IEEE International Conference on Neural Networks.*

[110] Rose, J. M., & Bliemer, M. C. (2007). Stated preference experimental design strategies. In *Handbook of Transport Modelling: 2nd Edition* (pp. 151-180). Emerald Group Publishing Limited.

[111] Revelt, D., & Train, K. (1998). Mixed logit with repeated choices: households' choices of appliance efficiency level. *Review of economics and statistics, 80(4)*, 647-657.

[112] Rogers, E. M. (2010). *Diffusion of innovations.* Simon and Schuster.

[113] Sándor, Z., & M. Wedel (2001) Designing Conjoint Choice Experiments Using Managers' Prior Beliefs. *Journal of Marketing Research 38*, 430-444.

[114] Santa Eulalia, L. A., Neumann, D., & Klasen, J. (2011, October). A simulation-based innovation forecasting approach combining the bass diffusion model, the discrete choice model and system dynamics-an application in the german market for electric cars. In *The third international conference on advances in system simulation.*

[115] Sándor, Z., & M. Wedel (2002). Profile Construction in Experimental Choice Designs for MixedLogit Models, *Marketing Science 21(4)*, 455-475.

[116] Sándor, Z., & M. Wedel (2005). Heterogeneous conjoint choice designs. *Journal of Marketing Research 42*, 210-218.

[117] Scarpa, R., & Rose, J. M. (2008). Design efficiency for non-market valuation with choice modelling: how to measure it, what to report and why. *Australian journal of agricultural and resource economics, 52(3)*, 253-282.

[118] Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J., & Griskevicius, V. (2007). The constructive, destructive, and reconstructive power of social norms. *Psychological science, 18(5)*, 429-434.

[119] Sharmeen, F., Arentze, T., & Timmermans, H. (2014). An analysis of the dynamics of activity and travel needs in response to social network evolution and life-cycle events: a structural equation model. *Transportation Research Part A: Policy and Practice, 59*, 159-171.

[120] Sharmeen, F., Arentze, T., & Timmermans, H. (2015). Predicting the evolution of social networks with life cycle events. *Transportation, 42(5)*, 733-751.

[121] Sharmeen, F., Chávez, O., Carrasco, J. A., Arentze, T., & Tudela, A. (2016). A Modelling Population-Wide Personal Network Dynamics Using a Two-Wave Data Collection Method and An Origin-Destination Survey.

[122] Shepherd S, Bonsall P, & Harrison G (2012). Factors affecting future demand for electric vehicles: A model based study. *Transport Policy 20*, 62–74.

[123] Sherif, M. (1935). A study of some social factors in perception. *Archives of Psychology (Columbia University).*

[124] Sheng, H., & Xiao, J. (2015). Electric vehicle state of charge estimation: Nonlinear correlation and fuzzy support vector machine. *Journal of Power Sources, 281*, 131–137. https://doi.org/10.1016/j.jpowsour.2015.01.145

[125] Sherwin, H., Chatterjee, K., & Jain, J. (2014). An exploration of the importance of social influence in the decision to start bicycling in England. *Transportation Research Part A: Policy and Practice, 68*, 32-45.

[126] Singh, A., Thakur, N., & Sharma, A. (2016, March). A review of supervised machine learning algorithms. In 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 1310-1315). IEEE.

[127] Slot, R., (2017) *Factors Influencing the Adoption of Electric Vehicles in the Netherlands*. Master: Business Information Management. Rotterdam School of Management, Erasmus University. Master Thesis

[128] Smaldino, P., Janssen, M., Hillis, V., & Bednar, J. (2017). Adoption as a Social Marker: The Diffusion of Products in a Multigroup Environment. *Journal of Mathematical Sociology, 41(1)*, 26-45.

[129] Smith, J. R., Louis, W. R., & Schultz, P. W. (2011) *Introduction: Social influence in action* . Group Processes & Intergroup Relations, 2011; pp. 599–603

[130] Richter, F. (2019) *Electric Vehicle Buyers Have the Agony of Choice*. Retrieved from: https://www.statista.com/chart/13465/electric-vehicle-models-available-in-north-america/

[131] Stibe, A. (2014). Socially influencing systems: persuading people to engage with publicly displayed Twitter-based systems. *Acta Universitatis Ouluensis.*

[132] Struben J, & Sterman J (2008). Transition challenges for alternative fuel vehicle and transportation systems. *Environment Planning B: Planning Design 35(6)*, 1070-1097.

[133] Sun, S., Zhang, J., Bi, J., Wang, Y., & Moghaddam, M. H. Y. (2019). A Machine Learning Method for Predicting Driving Range of Battery Electric Vehicles. *Journal of Advanced Transportation, 2019.* https://doi.org/10.1155/2019/4109148

[134] Thiel, C., Alemanno, A., Scarcella, G., Zubaryeva, A., & Pasaoglu, G. (2012). Attitude of European car drivers towards electric vehicles: a survey. *JRC report.*

[135] Tietge, U., Mock, P., Lutsey, N., & Campestrini, A. (2016). Comparison of leading electric vehicle policy and deployment in Europe. *Int. Council Clean Transp, 49*, 847129-102.

[136] Train, K. (2001). A comparison of hierarchical Bayes and maximum simulated likelihood for mixed logit. *University of California, Berkeley*, 1-13.

[137] Train, K. (2003). Discrete Choice Methods with Simulation. *Cambridge University Press*, UK.

[138] Train, K. E. (2008). EM algorithms for nonparametric estimation of mixing distributions. *Journal of Choice Modelling, 1(1)*, 40-69.

[139] , M., Banister, D., Bishop, J. D., & McCulloch, M. D. (2012). Realizing the electric-vehicle revolution. *Nature climate change, 2(5)*, 328-333.

[140] TyreeHageman, J., Kurani, K. S., & Caperello, N. (2014). What does community and social media use look like among early PEV drivers? Exploring how drivers build an online resource through community relations and social media tools. *Transportation Research Part D: Transport and Environment, 33*, 125-134.

[141] U.S. Department of Energy. (2012, February 7). *Data, Analysis & Trends. Alternative Fuels & Advanced Vehicles Data Center*. Retrieved from http://www.afdc.energy.gov/afdc/data/vehicles.html

[142] Voelcker, John. (2012, September 4). *August Plug-In Electric Car Sales: Volt Surges, Leaf Static*. Retrieved from: https://www.greencarreports.com/news/1078919_august-plug-in-electric-car-sales-volt-surges-leaf-static

[143] Wang, W., Xi, J., Chong, A., & Li, L. (2017). Driving Style Classification Using a Semisupervised Support Vector Machine. *IEEE Transactions on Human-Machine Systems, 47(5)*, 650–660. https://doi.org/10.1109/THMS.2017.2736948

[144] Walther, G., Wansart, J., Kieckhäfer, K., Schnieder, E., & Spengler, T. S. (2010). Impact assessment in the automotive industry: mandatory market introduction of alternative powertrain technologies. *System Dynamics Review, 26(3)*, 239-261.

[145] Wellman, B., Wong, R. Y. L., Tindall, D., & Nazer, N. (1997). A decade of network change: Turnover, persistence and stability in personal communities. *Social networks, 19(1)*, 27-50.

[146] Yang, P. (2018) *Adoption of Electric Cars in Different Neighborhoods in Amsterdam – The Role of Social Networks and Social Influences.* Environmental and Resource Management. IVM Institute for Environmental Studies. Master Thesis.

[147] Yang, S., Ma, W., Pi, X., & Qian, S. (2019). A deep learning approach to real-time parking occupancy prediction in spatio-termporal networks incorporating multiple spatio-temporal data sources. arXiv preprint arXiv:1901.06758.

[148] Yavasoglu, H. A., Tetik, Y. E., & Gokce, K. (2019). Implementation of machine learning based real time range estimation method without destination knowledge for BEVs. *Energy, 172*, 1179–1186. https://doi.org/10.1016/j.energy.2019.02.032.

[149] Yi, D., Su, J., Liu, C., Quddus, M., & Chen, W. H. (2019). A machine learning based personalized system for driving state recognition. *Transportation Research Part C: Emerging Technologies, 105*, 241-261.

[150] Zhang, D., Schmöcker, J. D., Fujii, S., & Yang, X. (2016). Social norms and public transport usage: empirical study from Shanghai. *Transportation, 43(5)*, 869-888.

[151] Zahid, T., Xu, K., Li, W., Li, C., & Li, H. (2018). State of charge estimation for electric vehicle power battery using advanced machine learning algorithm under diversified drive cycles. *Energy, 162*, 871–882. https://doi.org/10.1016/j.energy.2018.08.071