Determinants of the tobacco industry: a cross-market analysis to support strategy definition

Ana Catarina Ribeiro Coelho

Master's Dissertation

Supervisor: Prof. Manuel Augusto de Pina Marques

U. PORTO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Mestrado Integrado em Engenharia e Gestão Industrial

2020-07-24

Abstract

The demand for tobacco has been changing over the past few years. Consumers are becoming more and more aware of the effects of tobacco and are choosing to either stop smoking or, with the recent introduction of reduced-risk products by tobacco companies, to consume these products considered less harmful. Reduced-risk products or RRP are considered to contain less damaging compounds and some of them do not contain tobacco at all.

This dissertation emerges in the context of a consulting project for a leading international tobacco company. This company intends to empower the available information to support strategic decisions, through three workstreams.

The first workstream refers to an analysis tool for the consolidation of national tobacco consumption forecast, at the market level, combined with the possibility of analyzing the sensitivity of consumption to changes in price and regulations in place. In this workstream, the focus was the scenarios analysis. Information was initially collected and the potentially explanatory variables for tobacco consumption were selected. Subsequently, different causal models were developed to project the impact of input variables fluctuations. From this analysis, it was found that, for all segments of analysis, the panel regression presents the best performance. It was verified that macroeconomic factors, such as personal disposable income and unemployment rate, allied to price are considered the most predominant factors in consumer behaviour.

The second workstream emerged already during the development of the project due to the urgent need of perceiving the impact of Covid-19 in tobacco consumption for the years of 2020 and 2021. This exceptional circumstance has affected greatly the tobacco sales and, therefore, this study provides a realignment of the forecast, obtained in the first workstream, to be readjusted to the new reality of 2020. To develop this analysis, the effects of Covid-19 on tobacco consumption were listed, including, for example, effects related to the borders closure and the lockdown period. Each of these effects was analysed based on the information collected. This analysis was done for each country selected by the client company taking into account two possible scenarios: a more optimistic scenario, in which a single wave of proliferation and a slight decline in the economy are expected, and another scenario considering the possibility of a second wave of proliferation at the end of 2020 and a more drastic decline in the economy. According to the results obtained, tobacco consumption will undergo a greater decline in the second quarter of 2020 after which, for the most optimistic scenario, there will be a gradual recovery to pre-Covid consumption levels.

While the first two workstreams were focused on the tobacco market evolution, the third workstream represents another way to leverage information by monitoring performance indicators regarding an important process in any industry: the intermediate chain between the company and end customer. With this purpose, a tool for consolidating performance indicators was developed. Allied to this tool, the user has the possibility to redefine the data (as, for example, volume sold) and recalculate the metrics. The inference of variables is simpler, assuming that the indicators per unit sold remain constant.

Resumo

A procura de tabaco tem vindo a sofrer sucessivas mudanças ao longo dos últimos anos. Cada vez mais os consumidores estão cientes dos efeitos do tabaco e optam por deixar de fumar ou migrar para produtos de menor risco recentemente introduzidos no mercado pelas empresas tabaqueiras.

Esta dissertação surge no contexto de um projeto de consultoria para uma das empresas tabaqueiras líder do mercado de tabaco. Esta empresa pretende potenciar a informação de que dispõe para suporte de decisões estratégias, através de três frentes de ação.

A primeira refere-se a uma ferramenta de análise para consolidação de previsão do consumo nacional de tabaco, ao nível do mercado, aliado à possibilidade de analisar a sensibilidade do consumo face a mudanças de variáveis como preço e regulações em vigor. Para este segmento de análise, cujo foco centrou-se na análise de cenários alternativos, foi inicialmente recolhida informação e selecionadas as variáveis potencialmente explicativas do consumo de tabaco. Posteriormente, utilizaram-se diferentes modelos causais para projetar o impacto das variações das variáveis de entrada. Desta análise, verificou-se que, para todos os segmentos de análise, a regressão com dados em painel apresentou a melhor performance. Esta análise permitiu verificar que fatores macroeconónimos, como o rendimento disponível e a taxa de desemprego, aliados ao preço são os fatores mais preponderantes no comportamento dos consumidores.

A segunda frente de ação surgiu já durante o desenvolvimento do projeto em resultado da necessidade que a empresa sentiu de obter urgentemente uma previsão do impacto do Covid-19 no consumo de tabaco a nível nacional, para os anos de 2020 e 2021. Esta circunstância excepcional afectou consideravelmente as vendas de tabaco e, por conseguinte, este estudo proporciona um realinhamento das previsões de consumo de forma a reajustar os valores à nova realidade de 2020. Para o desenvolvimento desta análise, os efeitos do Covid-19 foram listados, incluindo, por exemplo, efeitos relacionados com o fecho de fronteiras e o período de confinamento. Cada um destes efeitos foi analisado com base na informação recolhida. Esta análise foi feita para cada país selecionado pela empresa cliente, tendo em conta dois possíveis cenários: um cenário mais otimista, em que é esperado uma única onda de proliferação e um queda mais ligeira da Economia; e um outro, considerando a possibilidade de uma segunda onda de proliferação no final do ano de 2020 e uma queda mais drástica na Economia. Pelos resultados obtidos, o consumo de tabaco apresentará uma queda maior no segundo trimestre de 2020 a partir do qual, para o cenário mais otimista, se verifica uma recuperação gradual rumo aos níveis de consumo anteriormente verificados num período pré-Covid.

Enquanto os dois primeiros fluxos de ação se focam na evolução do mercado de tabaco, o terceiro representa uma outra forma de beneficiar da informação disponível, através do controlo de indicadores de desempenho. Neste caso, de um processo particularmente importante: a cadeia intermédia entre a empresa e o consumidor. Desta forma, foi desenvolvida uma ferramenta para consolidação de métricas de desempenho. Adicionalmente, o utilizador pode alterar e recalcular as métricas. A inferência das variáveis é mais simples, assumindo que os indicadores por unidade se mantêm.

Acknowledgements

This dissertation represents a remarkable milestone - the transition between the academic journey to the working life. There are certainly a number of people who have accompanied me on this journey and deserve my appreciation.

First of all, a sincere thank you to Professor Manuel Pina Marques for supporting me in writing this dissertation, for your availability, guidance and careful insights that allow me to improve my work.

A thank you to LTP for the way it welcomed me with open arms, for the interesting challenges and for the opportunity it gave me to meet such great people. A special thank you to Paulo Sousa for his admirable personality and for all the incredible advice that helped me to grow as a person.

To my friends who accompanied me during these past 5 years, thank you. My academic journey became more fun with you by my side.

To you, Mom, thank you for teaching me to be persistent. Dad, thank you for teaching me to be curious and always want to learn more. Miguel and Tomás, I have to thank you for supporting me all the time and being the most amazing brothers.

Lastly, the greatest thank you goes to you, Diogo, for all the unconditional support, love and trust you place in me. I am so grateful to have you in my life.

It was a long journey to get where I am today, but all the support I got helped me grow personally and professionally. Therefore, thank you.

Ana Coelho

"The goal is to turn data into information, and information into insight."

Carly Fiorina

Contents

Acronyms and Symbols

ADC	Average Daily Consumption
CMI	Cross Market Insights
FCT	Fine Cut Tobacco
GLM	Generalized Linear Model
HDI	Human Development Index
IQR	Interquartile Range
KPI	Key Performance Indicator
MAE	Mean Absolute Error
OTP	Other Tobacco Products
PDI	Personal Disposable Income
RMC	Ready Made Cigarettes
RRP	Reduced-Risk Products
SVM	Support Vector Machine
SVR	Support Vector Regression
WAP	Weighted Average Price
WHO	World Health Organization

List of Figures

List of Tables

Chapter 1

Introduction

In the world that we live nowadays, the competition within each industry is increasing with the features that mould the global society. Those, specially promoted by the evolution of technologies during the last century, led to globalization. If, on the one hand, it was an opportunity for companies to grow and increase their customer pool, on the other hand, the competitors are now more and more aggressive. A competitive environment, such as the one found in many industries nowadays, is distinguished by an eager for companies to become more efficient and effective in their processes and decisions.

The tobacco industry is no different in this concern. Even more with the increasing public awareness of harmful effects, the tobacco industry is fighting to keep up with the numbers of consumers it once had, specially in more developed countries (Murphy, 2019).

Being considered as one of few “unhealthy” industries, there is great monitoring of governments in order to control the tobacco consumption through regulations, additional taxes over the products and different campaigns to create awareness among the population about the damaging impact of smoking.

Therefore, tobacco companies have been reinventing themselves by introducing alternative products considered to pose less risk to people than traditional products (Bialous and Glantz, 2018).

Thus, given the required efficiency to keep up with the competitors and the instability of the market itself, the project arises from a consulting request from one of the tobacco companies with the purpose of creating a solid basis for scenario analysis and perceiving how the effect of relevant variables, such as price changes and application of new regulation by the government, may impact the future of this industry.

Additionally, with the emergence of Covid-19 and respective impact on the world’s economy, uncertainty about tobacco consumption has increased. The study of Covid-19’s impact on consumption was crucial to the company’s decision-making in order to better react to this outstanding situation.

1.1 Project description

Given the current prospect of the cigarette industry, with several markets in decline and the industry itself creating innovative alternatives to traditional cigarettes, the main workstream of this dissertation is intended to model tobacco consumption at the national level.

The first major focus of this first workstream is to create reliable forecasts from 2020 to 2030 for the volume expected to be sell and value generated to the different product categories. Especially what regards innovative products which, given their recent introduction in the market, have less historical data and are therefore more difficult to predict volume and value. They are, however, the products of greatest interest since it is in this direction that the tobacco industry is heading.

The second focus of the workstream, and in which this dissertation will go into more detail, is a complementary feature to the forecast. This feature allows to study alternative scenarios on top of the baseline forecast by changing parameters, for instance a certain product price, to perceive the expected impact on tobacco consumption. Through a collective analysis of data from different countries, it is possible to infer the effects that a certain variable will have in a certain country by learning from what happened in other countries, for example, the application of a new regulation. Therefore, this scenario analysis is also designated as cross-market insights.

Complementary to the workstream described above, a study of the impact of Covid-19 on tobacco consumption was developed. The main purpose was to anticipate the consumer's behaviour due to changes in family income and changes in people's daily routine given the restrictions imposed to prevent proliferation of Covid-19. And, therefore, readjust the forecasts developed to the new reality allowing the company to have a solid basis for decision making in order to mitigate the effect of the pandemic on the company's sales.

Another topic of this dissertation is focused on an important process for the company, namely, the intermediate chain between the company itself and the final customer. The company has a set of key performance indicators (KPIs) in order to monitor and control operational effectiveness and efficiency. The aim of this analysis is to consolidate the indicators in a single interface and additionally have a what if analysis feature.

1.2 Objectives and research questions

This dissertation can be categorized in three main workstreams: the cross-market insights, the analysis of Covid-19's impact on tobacco consumption and the operational monitoring regarding route to customer process.

Regards cross market insights, the main objectives for this stream are:

(i) Identification of variables subject to change and with possible impact on tobacco consumption;

(ii) Information collection and data processing;

(iii) Selection and development of causal models.

Having the goals defined, the main questions to be answered are:

- Which variables have the greatest impact on tobacco consumption?
- When changing the variables above identified, what is the impact on tobacco consumption?
- What is the best approach to model the determinants of the tobacco industry?

As far as concerns the analysis of Covid-19's impact on tobacco consumption, the goals are defined as:

(i) Identification of the effects of Covid-19 in tobacco consumption;

(ii) Information collection;

(iii) Computation of the impacts of the outlined effects for each country under study.

The third and last workstream of this dissertation aims to develop an interface to present the indicators of volume, value and cost regarding the operation of delivery the product from the factory to the client.

1.3 Thesis outline

The thesis is divided in 6 chapters. In chapter 1 the problem to be addressed is briefly presented. It includes a contextualization of the project, what are the motivations and expected objectives.

Chapter 2 purpose is to provide a broad background regarding predictive models and a literature review of the driving factors regarding smoker behaviour.

Chapter 3 is dedicated to detail the different components of project and what it is expected in each workstream.

In chapter 4, it is explained the approach used in the different workstreams with major focus on scenario analysis.

Chapter 5 presents the results obtained in the several workstreams while chapter 6 is used to summarize and reflect about the work developed.

Chapter 2

Theoretical Background

This chapter is devoted to the literature review regarding the theoretical concepts useful for the elaboration of this project and also the relevant factors to the smoker's behaviour.

In the section 2.1 is presented an overview of the data preparation that must be done to guarantee a clean set of data as well as the predictive models that can be used.

In the section 2.2, the significant factors that have a role on tobacco consumption are identified.

2.1 Predictive models

Forecasting is described as a technique used to anticipate what it is expected to happen in the future. It can be divided into two categories: qualitative and quantitative methods. Taking into account the objectives of this project, the quantitative methods, the ones that use the historical data and statistical techniques to develop a forecast, that are relevant to address. These can be divided into time series and causal models (Brillio, 2018).

Time series techniques use a set of chronologically data points in order to identify trends and cyclical patterns while causal techniques rely on cause-effect relationships, based on one or more independent variables, in order to predict the output (Ray, 2015).

Regarding the first goal of the main workstream, related with development of a baseline forecast, the effort is on time series. On the other hand, regarding the scenario analysis (the main focus of this thesis), since the goal is to understand the impact of a certain variable in the dependent variable, the causal techniques are the models that present a relevant role for this section.

Considering causal models, those can be divided into classification and regression models. According to Garbade (2018), the main difference between them is that the output variable for classification is categorical while for regression is numerical. In this project, since the output is related with tobacco consumption and therefore it has a numerical nature, the focus will be on regression models.

According to Ray (2015), there are some key factors that should be considered to select the right regression model:

1. Data exploration;

2. Metrics of accuracy to compare models;

3. Cross-validation;

4. Feature selection;

5. The objective of the problem should be considered. Sometimes a less powerful model is easier to implement comparing to highly complex models.

These topics listed above are covered in the subsections below.

2.1.1 Data exploration

When exploring data, it is important to guarantee a quality dataset. One of the biggest problems when dealing with new data is the existence of unusual values (or outliers). Outliers can be problematic since they can cause statistical analysis to miss significant findings or distort results (Frost, 2019).

According to Frost, there are no strict rules that allow for a clear identification of the outliers. It depends of the context of the problem. There are three main causes for outliers: data entry or measurement errors, sampling problems and unusual conditions and, lastly, natural variation.

When an outlier is due to an error, it should be corrected or, if not possible, removed from the analysis since it is an incorrect value. However, when dealing with outliers related to natural variation, even though it is an unusual value, since it is in fact a real observation, it should not be removed from the dataset.

There are five methods to identify outliers. Starting from those with visual assessments and moving towards more analytical methods, the first one is sorting the datasheet and have a look over the unusually high and low values; other way is to display the data in a graph (e.g. a boxplot or a scatterplot) and identify potentially values that differ from the typical data; using z-score to detect points beyond a cut-off value, generally three standard deviations, from the mean value; other way is to use interquartile range by considering boundaries above the 25th percentile and 75th percentile and, lastly, using hypothesis tests in which the null hypothesis assumes that all data follow the same normal distribution while the alternative hypothesis considers that at least one value do not belong to it (Frost, 2019).

Frost (2019) considers that z-score and the hypothesis are not the best options to use since they assume the data follows a normal distribution, what it is not always the case. And ironically, they are sensitive to the presence of outliers.

2.1.2 Metrics of accuracy

According to Walther (2005), when accessing the performance of an estimator both bias and precision should be taken into account. The more biased and less precise an estimator is, the worse the ability to predict accurately.

The metrics can then be classified as bias, precision or accuracy measures. An unbiased prediction would represent a distribution of under and over-estimates whose overall value is zero, while a precise prediction represents a distribution of estimates that shows little variation and, lastly, an accurate measure represents how close the predictions are to the true value (Walther et al., 2005).

The most common bias measures used are: Mean Error (ME) and Mean Percentage Error (MPE). Regarding precision measures, the most common are: variance and standard deviation.

What regards accuracy measures, the most used metrics are: Mean Squared Error (MSE), MAE (Mean Absolute Error) and Mean Absolute Percentage Error (MAPE).

2.1.3 Cross validation

Predictive models are used to forecast accurately unseen data. Therefore, a common practice is to divide the dataset into training and test data in order to use the first dataset to, as the name indicates, train the model and the second to realize its ability to predict on unseen date. However, from the same dataset, different training and test datasets can arise and, consequently, may result in different accuracies for the same model dealing with the same dataset.

A better method to access the accuracy of a model is therefore cross-validation. The main idea behind is to divide the dataset into k groups, of approximately equal size. Then the model runs k times and, in each time, a different group is used as test dataset and the remaining k-1 groups to train the model. The final accuracy is given by the mean of k model scores. This way, the final accuracy has less bias (Brownlee, 2018a).

2.1.4 Feature selection

Feature selection is intended to reduce the number of input variables to be considered in a predictive model (Brownlee, 2019). The aim is to remove non-informative or redundant variables since the presence of those variables can add uncertainty to the model and reduce its performance (Kuhn and Johnson, 2013).

There are two methods to perform feature selection: supervised and unsupervised. The difference between those is whether variables are selected based on the output variable or not. Supervised methods are the ones that do consider the output while unsupervised do not.

The goal of using unsupervised methods is to remove redundant variables. The best way to verify potential redundancies between variables is to use correlations coefficients such as Pearson's correlation coefficient and Spearman's rank coefficient.

On the other hand, supervised methods serve a different purpose. They are used to remove irrelevant variables from the model. As represented in the Figure 2.1, supervised methods can be categorized into three different types.

Wrapper feature selection creates several models with different subsets of inputs and then selects the features that result in the model with the best performance. There is also the filter feature selection that relies on statistical techniques to assess the relationship between each input variable and the output variable. Those scores are the basis to filter the input variables that will

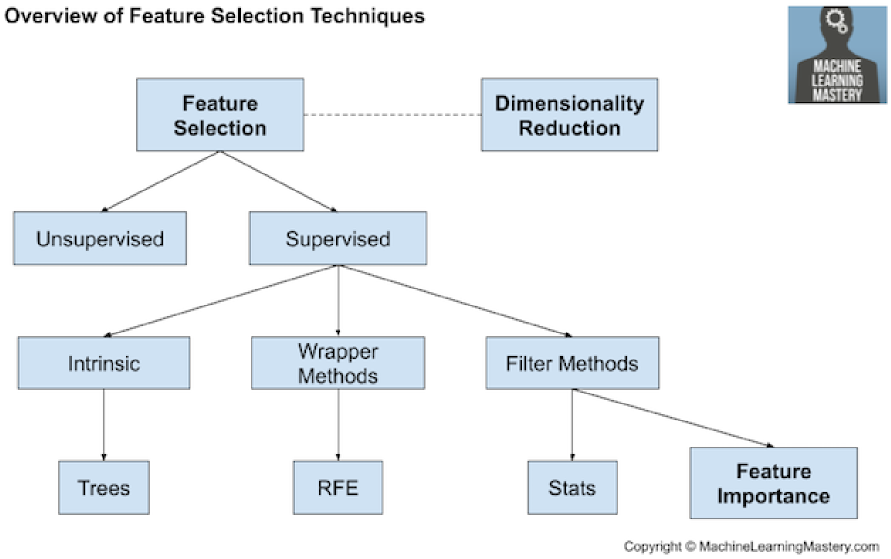


Figure 2.1: Overview over feature selection techniques (source: Brownlee (2019))

be consider in the model. And, lastly, there are intrinsic feature selection techniques, referring to some machine algorithms that already perform a feature selection automatically. It is the case, for example, of regression models and decision trees (Brownlee, 2019).

2.1.5 Regressive models

Linear Multiple Regression

Linear multiple regression is the most widely known prediction model. This model establishes a relationship between the output variable (Y) and more than one input variables (X) using the best fit straight line. A linear regression can be written as:

$$E(Y_i) = \sum_{1}^{p} x_{ij}\beta_j; \qquad i = 1, \ldots, n \tag{2.1}$$

where p represents the number of input variables in the model, n represents the number of Y observations and β_j are the unknown coefficients that have to be estimated.

To have a good performance, the relationship between input and output variables should be linear. When this is not the case, one option may be transforming the input variable to increase the linear relationship with the output variable. The most common transformations are done using exponential, quadratic, logarithmic or power concepts (DEI, 2019).

Linear multiple regression is very affected by outliers. As well as multicollinearity since it increases the variance of coefficient estimates making them very sensitive to changes and therefore results in unstable coefficients.

Panel Regression

Panel regression is considered to be an econometric model. According to Economist's "Dictionary of Economics", econometrics refers to "the setting up of mathematical models describing mathematical models describing economic relationships (such as that the quantity demanded of a good is dependent positively on income and negatively on price), testing the validity of such hypotheses and estimating the parameters in order to obtain a measure of the strengths of the influences of the different independent variables." (Moffatt, 2018).

A panel data, also known as longitudinal or cross-sectional time series data, refers to a multidimensional data that respects three properties: the observations take place at multiple points in time; at each point in time, the individuals studied are the same and several variables are collected for those individuals. Examples of individuals can be: countries and companies.

When dealing with panel data, from an econometric perspective, there are three different approaches. Pooled OLS (Ordinary Least Square) model is considered to be similar to an ordinary linear regression since it does not consider time and individual dimensions. On the other side, the other two approaches, fixed effect model and random effect model, are intended to capture the effect of unobserved or unmeasured variables as, for example, cultural factors or different practices across companies, that otherwise would not be considered in the model. Fixed effect model should be used when the goal is to capture all variables that do no change with time such as the personal peculiarities of each individual, while random effect model pretends to capture unobserved effects that can vary between entities but also vary over time (Alam, 2020).

Random Forest

Before introduce the random forest algorithm, the concept of decision tree must be explained. Decision trees, also designated as Classification and Regression Trees (CART), are a predictive modelling approach. This model resembles a tree structure, where, at each node, an input variable is tested (e.g. whether price is above 5) and branches are developed representing the outcome of this test (e.g. one branch for data with price above 5 and other for the remaining).

The random forest model is an ensemble model. Ensemble model refers to an algorithm that is able to combine several machine learning models into one in order to decrease variance (in the case of bagging), bias (when talking about boosting) or improve predictions (stacking) (Smolyakov, 2017). Random Forest is a case of bagging, also known as bootstrap aggregation, where models, in this case decision trees, are created in parallel to encourage exploration of independence and minimize the error by averaging the results from those different models.

The reason why a random forest may be used instead of a simple decision tree is that a decision tree is by itself prone to fit the training data very closely and therefore it is likely to overfit. However, since random forest builds several random decision trees, this effect may be mitigated.

The term "random" comes from two concepts used in the model: bootstrapping and feature randomness. Bootstrapping refers to the fact that, for building each of the decision trees, a random sampling with replacement of data is used. While the second term refers to the fact that, in the nodes of the decision trees, only a few random subset of features is considered (Koehrsen, 2018a).

Therefore, the random forest allows to develop an uncorrelated forest of trees that, consequently, are able to outperform any single decision tree because, when dealing with several decision trees, the individual errors of each tree did not impact the final result (Yiu, 2019).

Support Vector Machine

Considered as a black box, Support Vector Machine (SVM) is a linear model for either classification or regression problems. SVM can deal with both linear and non-linear problems (Pupale, 2018).

In a simplified way and taking as example the Figure 2.2, the main idea behind this algorithm is to find the line (or hyperplane, in a case of a four or higher dimension) that best divides the different data classes. cThe points, of different classes, that are closest to each other are called support vectors. The distance between each of these points and the line/hyperplane is called margin. The optimal line/ hyperplane is the one that maximizes the margin since it is in this point that the probability of a correct classification is higher (Pupale, 2018; Yadav, 2018).

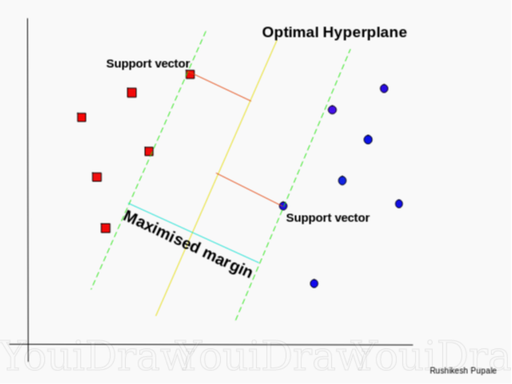


Figure 2.2: Exemplification of how SVM works (source: Pupale (2018))

In more complex cases, where the separation cannot be defined as a linear line/ hyperplane, the SVM transforms the data in order to convert it to a linearly separable data in a higher dimension space and then applying a similar approach as explained before (Pupale, 2018; Awad and Khanna, 2015).

SVM generalization to Support Vector Regression (SVR) consists in introducing an ε-insensitive region around the function, designated as the ε-tube, and the optimization problem is to, instead of finding the maximum margin separating the hyperplane, find the tube that best approximates the numerical function or, in other words, the flattest tube that includes most of the training instances (Awad and Khanna, 2015).

SVR is considered to have excellent generalization capacity aligned with high prediction accuracy (Pupale, 2018).

Interpretability and flexibility trade-off

Comparing multiple linear regression with random forest and SVM algorithms, the regression is considered to be more restrictive or, in other words, to present more difficulties in adapting to more complex relationships that may exist between the input and output variables (James et al., 2013).

However, it should not be discarded from the analysis as the context of the problem should be considered. There are two main goals when developing predictive models. On one hand, the objective of the analysis may be obtain highly accurate predictions or, on the other hand, infer variables' impact (James et al., 2013).

In the latter case, the models' interpretability is more relevant than its flexibility and, in this case, the use of simpler models, such as linear regression, is of greater interest.

In the algorithms described in section 2.1.5, there is a trade-off between interpretability and flexibility. Random forest, like SVM, do not provide interpretable insights but they tend to outperform regression models. However, not always. Even when the goal of the analysis is to develop the best performance model, more flexible models may present worse accuracy. Since they have a greater capacity to adapt to data, the greater the risk of adapting too much and not being able to extrapolate information to unseen data and, therefore, overfit the data (James et al., 2013).

2.2 Determinants of smoking consumption

At an early stage, it is necessary to understand the behaviour of consumers towards the product in question - tobacco. In a study conducted by Northwestern University, in which it was intended to study in detail the driving factor of tobacco consumption, it was considered that there are two types of smokers: those who smoke out of necessity (addiction) and who are therefore influenced mostly by an internal factor; and those who smoke because of the intervention of external factors, for example, do so because other people around are smoking. The smokers who are more easily subject to internal factors, due to the development of an addiction associated with the additive components of cigarettes, are called heavy smokers since these would potentially be the smokers who consume more tobacco. The second group they nominated them as the light smokers (Herman, 1974).

The conclusions of the study suggest that the predominant factor is actually the internal factor, in the case of heavy smokers, and the external factor, in the case of light smokers, however, both groups of smokers are not indifferent to the other factor (Herman, 1974).

Those external factors are the ones that can be more easily regulated by the government. According to Borland (2003), there are four main tasks that must be considered to control the tobacco market. Those are:

1. Discourage people from smoking tobacco;

2. Encourage current smokers to quit;

3. Protect non-smokers from the smoke of others;

4. Reduce exposure to toxins for regular smokers.

Increasing smoking taxation is considered one of the most effective measures to reduce the number of smokers and prevent others from starting this habit. Faced with a 10% price increase for the tobacco pack, demand is expected to decrease by 4% in richer countries and 8% in lower income countries. This measure is particularly effective for controlling tobacco consumption in adolescents, as they are the most price sensitive group (World Bank, 1999).

Another measure adopted by the government to discourage people from smoking is the application of regulations. According to Borland (2003), in order to prevent the use of tobacco, the tobacco industry's activities that are somehow making tobacco more attractive to the consumer should be eliminated, such as product promotion (e.g., packaging and branding that adds extrinsic value to the product) and the addition of ingredients to the tobacco, that increase the perception of value and contribute to customer engagement. Regulations regarding plain packaging and flavor bans, as well as regulations that avoid smoking at work and in public places are others measures taken by the government, that aligned with application of health warning labels and dissemination of health consequences in what concerns smoking, are considered to be effective in reducing tobacco consumption (Borland, 2003; Scollo et al., 2018).

Banning smoking in public places, restaurants and bars is especially efficient since smokers are lead to search for alternative products to cigarettes and makes smoking less acceptable and, therefore, protects non-smokers from undesirable smoke (Garrett et al., 2015; Wyckham, 1999).

Regarding the reduction of exposure of toxins for regular smokers, Borland (2003) mentions three ways to reduce the harmful effect of tobacco: make it less toxic (which is what tobacco companies currently intend to do when introducing risk reduced products), make it less additive, and make it less tasty. The first form of reduction directly reduces the harmful effect while the two later ones reduce the motivation to consume. These can be controlled through regulations.

Even though all countries intend to control the tobacco consumption, the policies implemented by each country vary from region of the world and country to country. In general, more robust strategies are implemented in more developed countries (Hawkins et al., 2018).

Besides the government control over the taxation applied and the regulations in force, there are other factors, e.g. macroeconomics and demographic, that also influence tobacco consumption. In a study carried out in Poland, a variety of factors were studied. The data collection took place in two different years (2003 and 2012) and the main objective was to understand the prevalence of tobacco consumption among adults, at different age groups and different locations (one considered rural and another more urbanized) taking into account social and respiratory health determinants (Sozańska et al., 2016).

There was a reduction in tobacco consumption in this country, due to measures such as increasing awareness of the harmful effects of tobacco, as well as public campaigns and economic decisions such as increasing taxes on cigarettes to make them less affordable. According to the GATS study, the increase in cigarette prices was the main factor for smokers to stop smoking.

Regarding social determinants, one of the main conclusions of this study was that smoking was higher among men than women. However, men with better education were more willing to quit smoking.

The decline in tobacco consumption was mainly reflected in the city under study, with greater dropout among the population with higher education. Several other studies, referring to the housing location of the population - city or village - found adverse results. In Germany, a study using national censuses concluded that the population of the city was more likely to be a smoker. Studies in Canada and USA showed opposite results. In the study itself, the results clashed with those obtained by Global Adult Tobacco Survey (GATS), which indicate that smokers living in urban areas smoke more than smokers in rural areas (Sozańska et al., 2016).

Regarding the unemployment factor, even though this factor can affect the affordability of a person to buy any product, there is an opposite effect to that too. According to this study, many of those who became unemployed are more likely to live a more stressful life and it is shown that stressful events increase the consumption of tobacco.

Another study, considering GATS data from thirteen low-middle income countries, corroborates several previous conclusions mentioned above. The tobacco use was significantly higher for males in all countries as well as for urban regions in most of them. Also social determinants related with inequality, such as education and income, were considered to have an increasing tobacco use effect (Palipudi et al., 2012; Garrett et al., 2015). This study lead to conclude that increasing levels of education are related with decreasing number of smokers. (Palipudi et al., 2012). According to Subanti et al. (2019), in a study conducted in Indonesia, the gender and price of the product not only impacts the prevalence in smoking but also the consumption of tobacco by a smoker. Female smokers tend to consume less tobacco than male smokers, and price has as well a significant impact since the increase in price leads to a decrease in consumption (Subanti et al., 2019; Garrett et al., 2015).

As mentioned before, the tobacco consumption has declined and led to the tobacco companies to reinvent themselves with the development of E-vapor and heated tobacco products, considered to be safer than traditional cigarettes[1]. Tobacco industries are convincing the regulators that these products should not be subject to the same regulatory restrictions as conventional products since it is considered a healthier product. In WHO Framework Convention on Tobacco Control (WHO FCTC), a framework created in 2003 with the goal of encouraging countries to take a series of measures regarding tobacco products, heated tobacco is not explicit seen as all other tobacco products (Bialous and Glantz, 2018). Therefore, some countries may apply to RRP products the same restrict regulations already prevailing to conventional products, while others may decide to impose less constraints to these alternative products.

[1] However, there are still little evidence that those products are less harmful than conventional products (Whiteside, 2019).

Chapter 3

Problem Description

The first section of this chapter presents the whole overlying project initially requested by the client company, which includes the development of a baseline forecast complemented with a tool for analysing different scenarios. The dimensions and granularity of the analysis will also be presented.

With the emergence of a global pandemic, the baseline forecast in section 3.1 must adjust to the new reality. Section 3.2 is focused on the Covid-19 impact on tobacco consumption with the purpose of complementing the project above mentioned.

Section 3.3 details the workstream related to the operating process between company and end client.

3.1 Strategic cross-market insights

The introduction of reduced-risk products into to the tobacco market has also introduced uncertainty in predicting smokers' behaviour. While conventional products, already well established in the market, follow a usual trend of consumption, with the entry of alternative products and consequently the adoption of these products by smokers, it is necessary to account for cannibalization effects and potential smokers who in a first stage would tend to give up, but with the existence of an alternative option, now choose to consume risk-reduced products (RRP).

The tobacco company in question is present in more than 100 countries (also designated as markets). In line with the above mentioned, a gap was found in the organizational structure of the company. The markets act as individual units and each unit is responsible for developing, based on different criteria, its own forecast of volume sold and generated market value as well as other additional analyses such as price sensitivity.

Thus, this project emerged with the purpose of develop a uniform forecast of expected national tobacco consumption for each country, complemented with the ability to customize smoker's behaviour relevant variables, with emphasis on those not controllable by the companies in the sector. As example, we have the application of a new regulation or tax increase, which, in case these circumstances occur in the future need to be reflected in the baseline forecast in order to be in line with the reality. For this last tool, the aggregation of information from different countries allowed a solid base to enable cross market insights.

The added value is to have, in a single integrated tool, the possibility of selecting the market and be able to have a robust forecast displayed. On top of that, it is possible to readjust the forecast if changes in the market happen over time or even create hypothetical scenarios for analysis and, consequently, understand which are the more relevant determinants to the smoker behaviour. In addition, the markets benefit from these cross-market insights, since otherwise they might not be able to access information to perform their own analysis.

Another relevant point is the decision support that this tool brings with the consolidation of information from different markets and the potential adoption of new products in each market, since it allows the identification of the most interesting markets to introduce a new product.

Considering the high amount of users expected to use this tool, a MS Excel interface was chosen given the greater familiarity of the users pool with this program.

The tool incorporates a 10-year baseline forecast, from 2020 to 2030, on an annual basis, for each country. On top of that the user is able to change the variables in order to readjust the forecast at the granularity and dimensions presented below.

Granularity

In terms of granularity, the analysis was performed on the category and subcategory level. In Figure 3.1, the product structure is schematized.

Regarding conventional products, Ready Made Cigarettes (or RMC) represent the most common traditional cigarettes, Fine Cut Tobacco (or FCT) stands for any rolling tobacco while OTP stands for Other Tobacco Products and includes products as cigarillos, cigars, shisha and pipe. Regarding the more recent products in the market, E-vapor is known as electronic cigarettes and heated tobacco is the designation given to a product that contains tobacco but instead of being burned it is heated. Finally, oral products represent tobacco that is specifically designed to be placed in the mouth for oral ingestion.

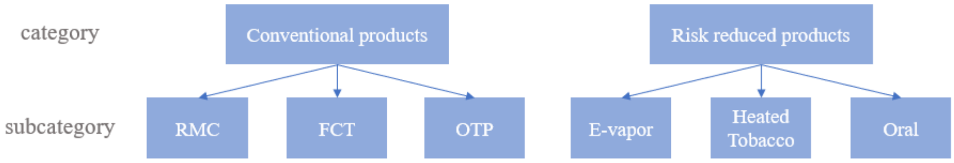


Figure 3.1: Product structure in the tobacco industry

Dimensions

For the markets, the most relevant dimensions to be determined for each category are: the expected volume to be sold, the market value generated and the number of consumers.

As regards the volume (or, in other words, the quantity of sticks sold), given the difference between the different products on offer by the tobacco companies and the fact that some products are not actually a stick, a metric called 'equivalent sticks' makes it possible to convert the units of each category into a single unit in order for the volumes of the categories to be comparable. With regard to "value" or market value, likewise, the unit used is generalized to all markets and it is the $US. Exchange rates were used to convert the value to dollars. Finally, the number of consumers for a category is given by the number of individuals who regularly consume that product.

The focus of this workstream is on the subsequent readjustment of the forecast baseline, also designated scenario analysis or cross market insights.

In the development of this project, other needs regarding readjustments in the forecast baseline have emerged.

3.2 Analysis of Covid-19's impact

Around the world, a global pandemic has swallowed the planet in a few months and isolated billions of people at home. As expected, most companies were impacted: some benefited as health-related industries; others were damaged as touristic companies.

Similarly, the tobacco industry was also affected. The need to analyze the effects of Covid-19 on tobacco consumption patterns was a sudden but urgent topic to be covered.

The impact of Covid-19 on tobacco industry was analyzed to 57 countries, some of them not yet incorporated in the workstream discussed in section 3.1. This analysis was performed country by country, under the same conditions of dimensions and granularity discussed for Cross Market Insights: an analysis of volume, market value and consumers in the market at the subcategory level (RMC, FCT, OTP, e-vapor, heated tobacco and oral). Regarding the time horizon and regularity of the analysis, 2020 and 2021 were analysed quarter by quarter.

According to Euromonitor[1] information, there are potentially 3 scenarios outlined for the impact of Covid on the economy, one that portrays a more optimistic scenario in view of Covid-19 (henceforth referred to as the optimistic Covid scenario), one slightly pessimistic (referred to as the pessimistic Covid scenario) and the other one considerably pessimistic. In the same sequence they are arranged by the highest to the lowest likelihood of occurrence. Given the low probability of an extreme pessimist scenario occurring, the third scenario delineated by Euromonitor was not considered to the analysis.

The main purpose of this analysis was to anticipate the consumption patterns, for each country, for the two selected scenarios. Subsequently, and after defining clusters of countries given macroeconomic and tobacco market similarities, the firm intends to define lines of action, for each cluster, to mitigate possible damages.

Regarding the impact of Covid-19 on consumption patterns, there are several effects that were considered:

- Macroeconomic changes: Due to the economic downturn, several industries have seen demand for their products fall, with many companies cutting their investment plans and/or

[1] Euromonitor is an independent provider of market research

laying off their workforce. This whole scenario contributes to an increase in the unemployment rate and decrease in income (PDI per capita) and, consequently, a decrease in affordability;

- Downtrading: With the decline in family income, not only is the volume of tobacco consumed subject to change but also the profit from a packet of tobacco sold, since the tendency is for premium consumers to migrate to less expensive tobacco products;

- Travel restrictions: During lockdown time, as governments close borders to curb the spread of the virus and discourage travel, cross-border tobacco also decreases. The main reason for tobacco inflow and outflow is directly related to the tobacco price in border countries compared to the price carried out in the country itself;

- Labour migrations: After the announcement of lockdown, there was a temporary return of emigrant workers to their countries of birth. Since there are countries where the proportion of emigrant and immigrant workers is quite distinct, with the outflow of emigrants and the influx of immigrants, the number of consumers in each country changes;

- Behaviour shifts: Measures to prevent the proliferation of Covid-19 implied a period of mandatory lockdown, followed by a period of reintegration of normal routines. These measures also impacted the consumption pattern of smokers due to the change of usual routines: increased level of stress, isolation at home and the need to wear masks in public places.

3.3 Operational monitoring

Although the tool described in the section 3.1 allows each market to conduct a nationwide study of the tobacco consumption and extrapolate the volume sold by the company itself through the application of share of market (and the same rationale applies to value), this analysis does not consider a particularly important process in the company - the intermediary chain that connects the tobacco company with the ultimate customer - constituted by distributors, retailers and wholesalers.

Therefore, the analysis of the route to consumer process arises, with focus on the flow between the tobacco company and end-client.

Currently the company itself has a set of Key Performance Indicators (KPIs) related to this subject, used to understand the efficiency of each market to obtain the proposed goals, year by year. These KPIs are detailed at the level of channels (physical and/or online) and three product categories: conventional, heated tobacco and e-vapor.

In order to calculate the KPIs, it is important to have into account the national and company's own tobacco volume sold, as well as the value generated and how it is distributed among the different intermediaries of the network (tobacco company, distributor, retail and wholesaler). It is also important to consider the related costs: distribution, sales team and investments. Another important indicator is the total number of tobacco sales outlets in the market and the proportion of those visited by the company's sales team. The indicators can be evaluated in absolute terms, by unit sold and/or net sales ratio.

Therefore, the aim of this request is to build an approach to consolidate data related to the route to consumer process, as well as the possibility of iterating on variables in order to check the impact on the KPIs. This last complementary feature is mainly useful to adjust the KPIs when dealing with forecasted values, that, as usual, have some uncertainty associated.

Data from markets, ranging from 2016 to 2020 (where values from 2020 are predictions), was collected in a template. This way, the user, when using the tool can have access to historical data for benchmarking and check the evolution of market data and indicators.

Chapter 4

Solution Approach

This chapter addresses the approach used to develop the several analyses proposed. In section 4.1 the focus is on scenario analysis that goes on top of the baseline forecast, designated as cross-market insights. Several steps of the process will be presented: methodology, the variables selected for the models, how the data was collected and prepared and then the last subsection presents the algorithms used as well as the metrics chosen to compare models.

Section 4.2 details the methodology used to analyse the impact of Covid-19 on tobacco consumption and section 4.3 presents the approach used to monitor and create an what if analysis regarding the route to consumer process.

4.1 Strategic cross-market insights

As mentioned in the problem description, the project had initially two main goals. One refers to the construction of baseline forecast, country by country, making use of diffusion curves to capture the characteristic shape of the curve of adoption of new products and time series forecast to capture the trend observed in historical data for the remaining products. The second objective, which is the focus of this dissertation, is to develop methods that allow the user to readjust and create alternative scenarios on top of the baseline.

Since the aim is to work with explanatory variables, that when changed, result in a new readjusted forecast, we are in the domain of causal models. For this analysis, it is intended to cross insights from the different markets.

The explanatory variables are described in subsection 4.1.2 while the predictive variables are presented in subsection 4.1.1.

4.1.1 Methodology

Given the relevant dimensions to determine: volume sold, value generated and number of consumers, for the granularity defined in chapter 3, and the direct interdependence between them, it was decided to actually have as predictive variables the country's percentage of smokers, designated as incidence, and the average daily consumption of each smoker in sticks (ADC). From these indicators, it is possible to reach all the desired dimensions, as expressed in formulas 4.1, 4.2 and 4.3.

$$Number\ of\ consumers = Market\ incidence\ x\ Adult\ population \tag{4.1}$$

$$Volume = Number\ of\ consumers\ x\ ADC \tag{4.2}$$

$$Value = Volume\ x\ WAP \tag{4.3}$$

where WAP represents the category (or subcategory) Weighted Average Price.

Note that the volume and value dimensions need only to be calculated at subcategory level, since at category level the volume and value correspond to the sum of the volume and value respectively of the dependent subcategories. The case of the number of consumers is a peculiar one and will be dealt with in the following subsection.

Incidence

Incidence is given by the number of smokers among the total population of the country.

When dealing with incidence, there one particular detail that must be taken into consideration. Let us look at the case of the conventional category. The sum of the incidences of the respective subcategories (RMC, FCT and OTP) could lead us to believe that it should equal the incidence of the parental category. But this may not be the case.

For this analysis, two types of smokers were considered: solos and dualists. Solos represent smokers who consume only one category of product, while dualists consume more than one. Due to the existence of dualists, the sum of the incidence of the subcategories associated with a category may total more than the incidence of the category itself (since, in the sum of the incidences, dualistic smokers may be considered multiple times). Thus, unlike volume and value, to obtain the total number of consumers as well as the consumers from each category the solution is not to add up the values obtained from the respective subcategories but to develop a specific model for each one.

Thereby, it would constitute an incidence for the total market, as well as for each category and each adjacent subcategory.

However, given the dependency between the incidences, where knowing, for example, that if it is a conventional smoker can then be a RMC, FCT or OTP consumer, it was decided to build a pyramid framework as represented in Figure 4.1.

There is a model focused on the total market incidence and, subsequently, there are models that represent the balance of how many of those actually belong to the conventional comparing to how many of those smoke RRP (represented in block 1). And the same applies to the subcategories for conventional and RRP (blocks 2 and 3).

Although, due to the existence of dualists, the balance cannot be calculated as the incidence of the category (or subcategory) divided by the total incidence (or incidence of the parental category) in order to totalize 100%. It needs to be readjusted by the formulas 4.4 and 4.5, where *N* represents

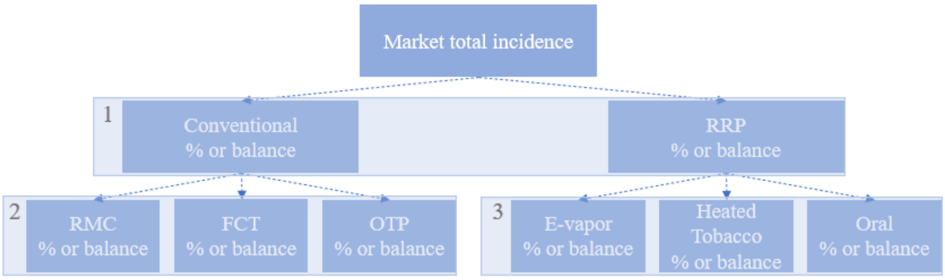


Figure 4.1: Pyramid framework based on total market incidence and balances.

the total number of categories and *M* the number of subcategories within the paternal category. *C* and *S* represent, respectively, the several categories and subcategories.

$$Balance_{category\ i} = \frac{Incidence_{category\ i}}{\sum_{j=1}^{N} Incidence_{category\ j}} \qquad \forall\ i\ in\ C \tag{4.4}$$

$$Balance_{subcategory\ k} = \frac{Incidence_{subcategory\ k}}{\sum_{j=1}^{M} Incidence_{subcategory\ j}} \qquad \forall\ k\ in\ S \tag{4.5}$$

Instead of using the incidence of the category respectively above, the sum of the incidences of the same category is calculated.

Given the dependence between the balances of the same block and knowing that the sum totals 100%, a potential advantage of this methodology would be the reduction in the number of models to be studied. For a category with k subcategories, only k - 1 models would need to be developed since the balance for the kth subcategory would be given by the difference of 100 and the sum of balances of all other subcategories. However, it was opted for redundancy and to develop a model for each subcategory in order to overcome possible errors in the models and that would, otherwise, transpose from one balance to another.

This way it totals 9 analysis segments in this dimension, as indicated in Figure 4.1. In Figure 4.2, it follows a schematization of how the incidences, for a certain country, can be obtained in an alternative scenario. The blue cells represent the baseline data (without any variation in the input variables). The orange is the new values after changing the input variables.

The total market incidence changes directly by adding to the total market incidence baseline (obtained by the time series forecast) the delta of difference between the result obtained by the Cross Market Insights (CMI) for the alternative scenario and the baseline scenario according to the CMI.

In the context of the project, since conventional products tend to follow a stable consumption trend, while RRP products, when introduced in the market, follow a typical adoption curve, the time series forecast, used for the design of the baseline forecast, is more adequate and accurate than a causal model in this circumstance. This is the reason why the baseline dependent variable given by the time forecast is always considered and a delta of difference is applied on top of that instead of considering right from the start the incidence value suggested by the causal model.

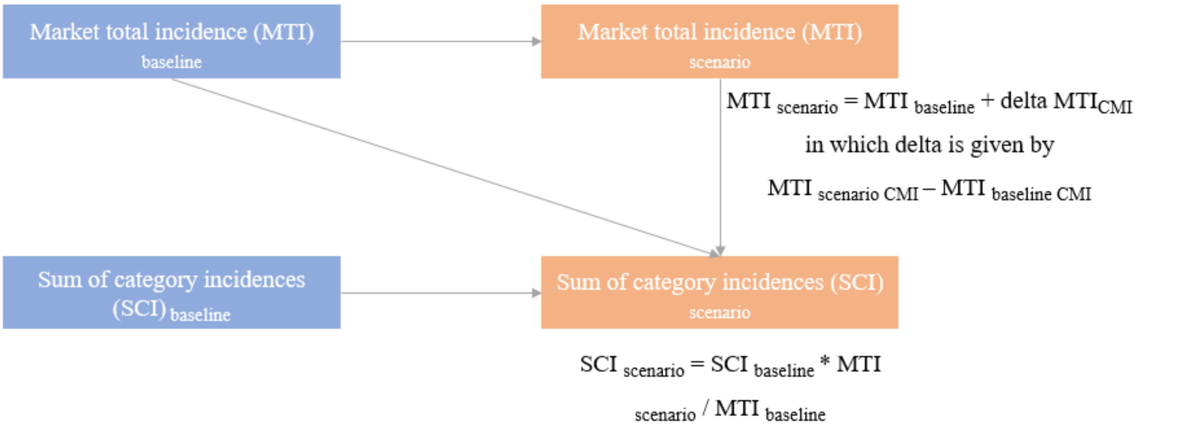

Market total incidence (MTI) baseline
Market total incidence (MTI) scenario
$\text{MTI}_{\text{scenario}} = \text{MTI}_{\text{baseline}} + \text{delta MTI}_{\text{CMI}}$
in which delta is given by
$\text{MTI}_{\text{scenario CMI}} - \text{MTI}_{\text{baseline CMI}}$
Sum of category incidences (SCI) baseline
Sum of category incidences (SCI) scenario
$\text{SCI}_{\text{scenario}} = \text{SCI}_{\text{baseline}} * \text{MTI}_{\text{scenario}} / \text{MTI}_{\text{baseline}}$

Figure 4.2: Methodology used to obtain the market total incidence in an alternative scenario

Regarding the category and subcategory incidence calculations, the process is slightly different as schematized in Figure 4.3. Let us look at the case of the categories.

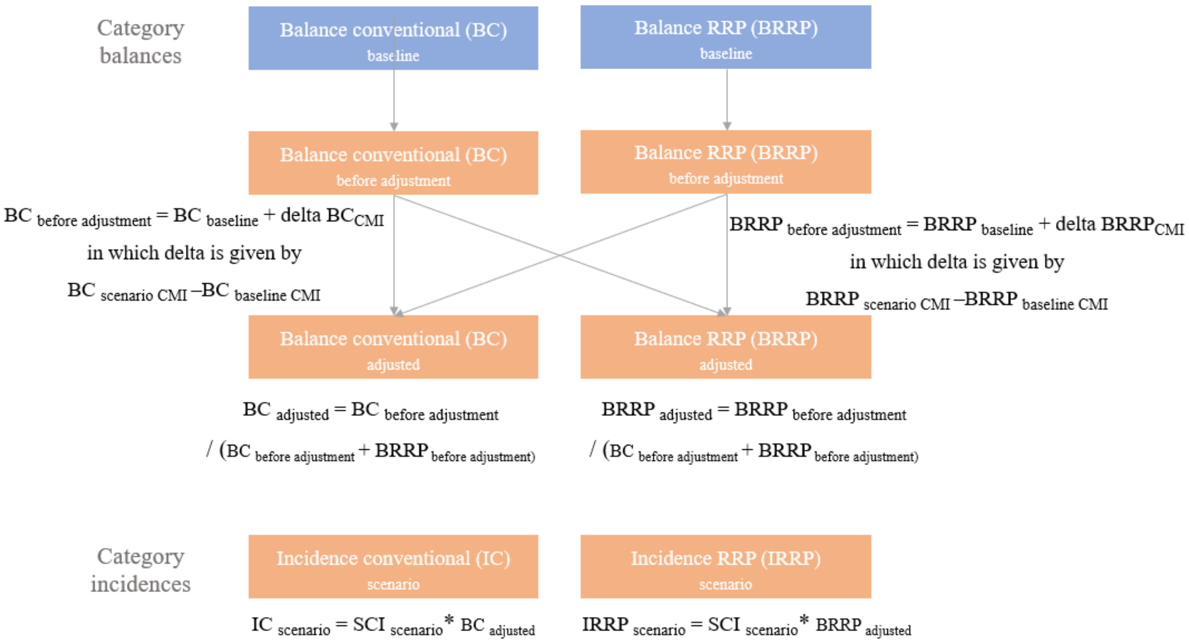

Category balances
Balance conventional (BC) baseline
Balance RRP (BRRP) baseline
Balance conventional (BC) before adjustment
Balance RRP (BRRP) before adjustment
$\text{BC}_{\text{before adjustment}} = \text{BC}_{\text{baseline}} + \text{delta BC}_{\text{CMI}}$
in which delta is given by
$\text{BC}_{\text{scenario CMI}} - \text{BC}_{\text{baseline CMI}}$
$\text{BRRP}_{\text{before adjustment}} = \text{BRRP}_{\text{baseline}} + \text{delta BRRP}_{\text{CMI}}$
in which delta is given by
$\text{BRRP}_{\text{scenario CMI}} - \text{BRRP}_{\text{baseline CMI}}$
Balance conventional (BC) adjusted
Balance RRP (BRRP) adjusted
$\text{BC}_{\text{adjusted}} = \text{BC}_{\text{before adjustment}} / (\text{BC}_{\text{before adjustment}} + \text{BRRP}_{\text{before adjustment}})$
$\text{BRRP}_{\text{adjusted}} = \text{BRRP}_{\text{before adjustment}} / (\text{BC}_{\text{before adjustment}} + \text{BRRP}_{\text{before adjustment}})$
Category incidences
Incidence conventional (IC) scenario
Incidence RRP (IRRP) scenario
$\text{IC}_{\text{scenario}} = \text{SCI}_{\text{scenario}} * \text{BC}_{\text{adjusted}}$
$\text{IRRP}_{\text{scenario}} = \text{SCI}_{\text{scenario}} * \text{BRRP}_{\text{adjusted}}$

Figure 4.3: Methodology used to obtain the category incidences in an alternative scenario

In a first stage, the conventional and RRP balance values are given by a similar formula used in the total market incidence - the delta obtained by the CMI model is added to the baseline balance.

Then, to address possible errors in the models that may lead to the sum of the conventional and RRP balances not totalling 100%, the balances are readjusted to ensure this condition.

Having the conventional and RRP balances, the incidences of the categories are obtained by multiplying the balances by the sum of the incidences - the reverse operation expressed in formula 4.4.

The best estimate of the sum of the incidences of the categories for the alternative scenario, as shown in Figure 4.2, considers that the proportional growth/decrease in the sum of the incidences of the categories is the same as in the total market incidence.

This way, the category incidences are obtained. A similar process is applied to obtain the subcategories' incidences.

Average Daily Consumption

In terms of daily consumption, on the other hand, it is a simpler variable. Since it only impacts the volume and value variables, it only needs to be developed at the subcategory level.

Resulting in a total of 6 analysis segments: RMC, FCT, OTP, E-vapor, heated tobacco and oral ADC.

In the same way as with total market incidence, when subject to a variation of the input variables, a delta is added to the ADC baseline.

4.1.2 Variables formulation

According to the studies presented in the literature review, there are three major groups of variables that influence tobacco consumption: macroeconomic, demographic and market variables.

In this respect, macroeconomic variables associated with the country development were included in the model, namely Gini coefficient[1] and Human Development Index (HDI), as well as variables associated with income: disposable income per capita (PDI per capita) and unemployment rate.

Given the demographic variables, two indicators were selected: percentage of male population and percentage of urban population, since, according to the studies reviewed in the literature, there is statistical evidence that men have a greater tendency to smoke and, among smokers, they also tend to consume more than women smokers. On the other hand, several studies have shown differences between consumption in urban and rural settings, yet the studies' results are not unanimous regarding which has the highest consumption.

Considering market variables, there are two major factors: price and regulation. In order to compare the price of a pack of tobacco between different countries, it was actually chosen to include a variable called affordability, which represents the financial effort for a smoker to buy a pack of tobacco[2] given the tobacco price and the average daily income per capita in that country, as expressed in formula 4.6.

$$Affordability = \frac{WAP\ x\ 20\ sticks}{Annual\ PDI\ per\ capita\ /\ 360\ days} \tag{4.6}$$

[1] statistical measure of income distribution used to evaluate a country's economic inequality

[2] to this analysis, a tobacco pack is considered to have 20 sticks

On the other hand, for the analysis of balance models, it is important to also consider the cannibalization due to prices. When a subcategory's price becomes relatively lower than another, consumers are encouraged to switch subcategories (or categories). Therefore, the gap between subcategories (and categories) affordabilities was taken into account as a variable.

Regarding the regulations applied for tobacco control, with more than 50 different regulations, a first selection was made to identify the most relevant regulations [appendix A]. Several variables were defined in order to understand how restricted regulation is in each country, in general terms and what regards the main types of regulations. Therefore resulting in the following variables:

- Total regulation;
- Consumption regulation[3];
- Flavours in RRP products regulation;
- Menthol flavour in RMC products regulation;
- Packaging in RMC products regulation[4];
- RRP online selling regulation;

Each of these 6 variables represents a factor ranging from 0, an extremely restricted country, to 1, a country with a lot of freedom in relation to the tobacco market.

Even though the variables mentioned above are the most interesting to consider in the scenario analysis, other explanatory variables should also be included in the models (even if they cannot later be subject to variation in the sensitivity analysis) in order to capture all the explanatory effects of tobacco consumption and, therefore, the models be able to consider the correct effect for each of the explanatory variables.

Thus, for the models associated with incidences, it was considered the amount of RRP subcategories already introduced in each market, since it is directly related to the number of alternatives that the smoker has when he wants to pursue a healthier option instead of giving up smoking tobacco. As well as the number of years since RRP was introduced into the market, designated as RRP maturity, since, with time, it is expected that more and more smokers move from conventional to RRP products.

On the other hand, for the subcategories balances, as the name suggests, we are dealing with balances and therefore the existence of more subcategories in the market is expected to lead to a smaller value of balance for each subcategory. Thus, it was considered whether or not the other subcategories of the parental category exist in the market.

Regarding ADC models, other relevant variable to consider, designated as solos, is the percentage of smokers of a certain subcategory who purely consume that product. Since, it is expected that

[3] Consumption regulation defines the public places where smoking is allowed.

[4] If this regulation is in place, the RMC' packages are standardized to all brands, with no possibility to include any branding, logos or promotional elements in the package.

the higher the proportion of solos consumers, the higher the average subcategory's consumption because those smokers do not split the consumption among other subcategories.

After the generalization of the most relevant variables, summarized in Table 4.1, it is necessary to take into account that for each analysis segment of the 15 under analysis, the choice of variables was tailored according to what was more coherent from the business perspective. In appendix B, the variables selected for each segment of analysis are listed.

Table 4.1: Variables selection

Variable group	Variables
Macroeconomic variables	Gini coefficient Human Development Index PDI per capita Unemployment rate
Demographic variables	Percentage of male population Percentage of urban population
Market related variables	Affordability Affordability gaps Total regulation Consumption regulation Flavours in RRP products regulation Menthol flavour in RMC products regulation Packaging in RMC products regulation RRP online selling regulation RRP number of subcategories RRP maturity Subcategory existence Solos

4.1.3 Data collection

Described the methodology and the ouput and input variables, it is now necessary to discuss how the required information was collected for the development of this workstream.

Initially a template was designed to be filled by each market with values of volume, value and incidence regarding national tobacco consumption, for four historical years (2016 to 2019), with split between solos and dualists, at the level of granularity defined in chapter 3. Other additional tables, related to the tobacco market, were also made available by the client company. An example is an aggregated table of information regarding the regulations in place in each market.

Subsequently, information regarding demographic and macroeconomic factors for model training was synthesized from public sources of information for the same historical horizon. Table 4.2 illustrates the sources used.

Table 4.2: Sources used to collect the necessary data

Content	Source(s)
Macroeconomic data	World Bank, The Economist, United Nations
Demografic data	World Bank
Market related data	Internal source

Regarding the study variables - total incidence, balances and ADC -, for historical years, the first one was obtained by summing up the incidences of solos and dualists smokers, obtained from the templates filled by the markets; the balances, also based on the incidences of the templates were obtained applying the formulas 4.4 and 4.5. Lastly, ADC results from the reverse formula 4.2.

Also, the price per subcategory is obtained through the template data by a simple division of the value by volume, while the category price is given by the weighted average price of the underlying subcategories.

As mentioned above, the client company is present in more than 100 markets and the historical data needs to be filled by each one of them. Therefore, it was decided to integrate the markets in a phased process. Initially by the countries of the Western European region, which are the richest countries in terms of information namely with the highest adoption of RRP products.

4.1.4 Data preparation

In an initial step, after collecting the data, it is important to explore it in order to identify data problems as well as methodologies to mitigate them.

Outliers detection

As mentioned in the literature review, a very important step is to ensure the quality of the data to be introduced into the analysis.

Given that macroeconomic and demographic data came from reliable sources such as the World Bank, the focus of this analysis was on the templates. Each one was filled in by a department of a different country, and consequently, even though it is a template with well-defined guidelines, because of difficulties in obtaining data on the country's national consumption or because of the lack of knowledge on the split between incidence of solos and dualists, the data filled in are not always the most correct.

As mentioned above, the information in the template serves as a basis for collecting, for the years 2016 to 2019, the volume sold, the value generated, the incidence of each category and subcategory and, consequently, consumption and price (which when applied in the model, translates into affordability) for each of the subcategories. It is these two variables, consumption and affordability, that are mostly subject to outliers.

For detection and removal of outliers of consumption and affordability, it was decided to use the interquartile range (IQR). Figure 4.4 allows a better understanding of how the detection is performed.

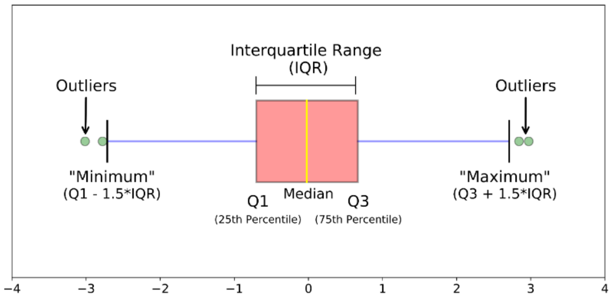


Figure 4.4: Visual demonstration of how outliers are detected according to IQR (source: Galarnyk (2017))

For each variable, the Interquartile Range is given by the difference between the third (Q3) and first (Q1) quartile. The boundaries, above which the observations will be considered as outliers, are given by formulas 4.7 and 4.8.

$$Upper\ Limit = Q3 + IQR\ x\ k \tag{4.7}$$

$$Lower\ Limit = Q1 - IQR\ x\ k \tag{4.8}$$

According to Brownlee (2018b), the usual values for k are 1.5 or 3 in cases where the user pretends to identify mainly extreme outliers. In this project, a value of 3 was chosen to perform a less severe outliers removal.

In the case of incidences, potential errors were detected in template filling, namely there were cases of countries that did not have solos and dualists split and, consequently, the market total incidence estimation would not be correct. Therefore, these observations were not considered.

As the study involved 20 countries and the historical information was collected for 4 years, it would be expected to have 80 observations. However due to the listed points below, the number of observations, for each model, is actually lower.

- Removal of outliers;
- In market total incidence model, it was only considered countries where the split between solos and dualists consumers exists;
- For each subcategory balance model, it was only consider countries where the corresponding subcategory is present and, additionally, where the subcategory is not the only one of the parental category in the market. Since, for this two cases, the subcategory balance would always be 0 and 1, respectively;
- For each subcategory ADC model, only countries where the corresponding subcategory is available in the market were considered.

Therefore resulting in the number of observations presented in Table 4.3.

Table 4.3: Number of observations per segment of analysis

Incidence variable	Num. observations	ADC variable	Num. observations
Market total incidence	43	RMC ADC	80
Conventional balance	60	FCT ADC	75
RRP balance	60	OTP ADC	39
RMC balance	67	E-vapor ADC	58
FCT balance	73	Heated Tobacco ADC	40
OTP balance	41	Oral ADC	16
E-vapor balance	40		
Heated Tobacco balance	32		
Oral balance	16		

Consequently, the number of observations for some segments of analysis are quite scarce, specially Oral subcategory that is present in a very small number of countries.

Variables transformation

Since linear models will be used for analysis of the balance and consumption segments, an important precaution is to check that the relationship between the input and output variables is linear. In cases where this linearity is not evident, it was chosen to transform the input variable.

From the scatter plots analysis, it was found that there were two potential variables that subject to transformation would have a more linear relationship with the several output variables: PDI per capita and RRP maturity. Due to the existence of countries like Switzerland, where the PDI is substantially higher than in other countries without, however, the same impact been observed on the output variables, the PDI per capita variable was logarithmized, since logarithmization transforms a highly skewed variable distribution into a less skewed shape (Roka, 2019).

The scatter plots represented in Figure 4.5 exemplify the before and after transformation, using as example the conventional balance dependent variable.

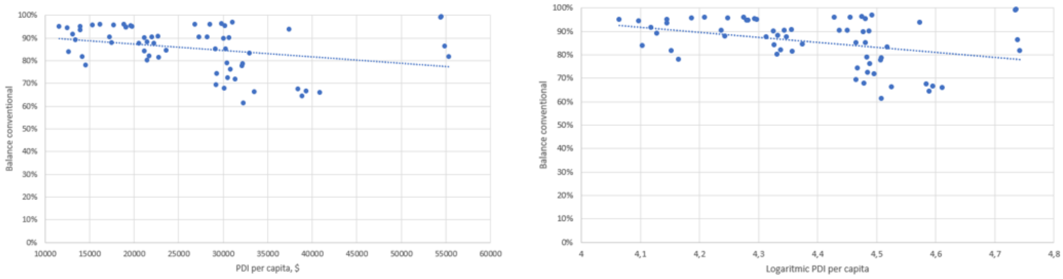


Figure 4.5: Difference in PDI variable pattern before and after logarithmization

Regarding RRP maturity, it is expected that during the first years of RRP in the market, more and more people tend to migrate from conventional to RRP products. However, as soon as RRP is well established in the market, the balance between conventional and RRP tends to an equilibrium, as exemplified in Figure 4.6 using data from 4 countries and comparing the conventional balance as RRP maturity increases. Conventional consumers in countries 1 and 2 are migrating to RRP while in countries 3 and 4, the balance remains roughly the same.

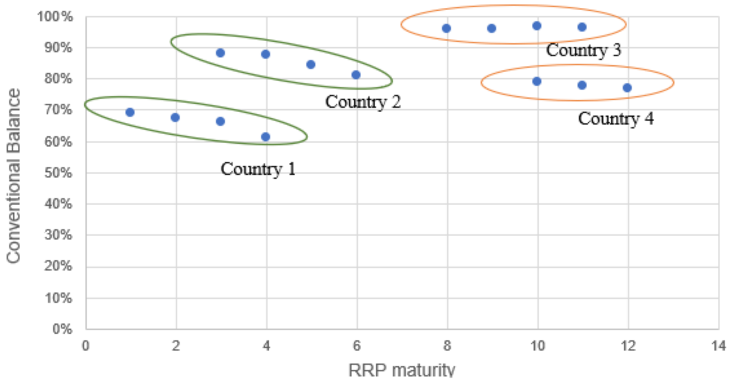


Figure 4.6: Migration from conventional to RRP products taken as example 4 countries

Thus, based on the scatter plot, it was chosen to consider that the RRP maturity variable had a maximum 6-year value, since from this value on, the RRP products are considered to have stabilized in the market.

Feature selection

According to Kuhn and Johnson (2013), the presence of redundant or meaningless variables in the model causes uncertainty and reduces its performance. Therefore, a preliminary analysis must be done to remove these two types of variables.

Regarding the correlated variables, according to Brownlee (2019), the best way to check redundancies between input variables is through the analysis of the correlation matrix. Since the number of distinct input variables considered in the 15 analysis models is considerably high (more than 40), it was decided to choose groups of variables that could potentially be related for analysis. Those groups were: macroeconomic variables and variables related with regulations.

Regarding the possibility of having variables that do not significantly explain the output variables, since the development of these models aims to supply a tool for scenario analysis the objective is not focused on finding the selection of variables that optimizes the accuracy of the model, but on inferring which variables have a significant impact and should therefore be considered. Thus, for the linear regressions under analysis, the selection of explanatory variables was done through the evaluation of p-values retrieved by an initial model that includes all variables in the analysis. The traditional threshold of 5% was considered, above which the variable is considered statistically significant. Regarding the Random Forest model, the feature importance matrix was used to identify the most important input variables to be considered in the final model.

4.1.5 Models formulation

After the complete process of collecting, processing the information and selecting uncorrelated variables, the dataset is considered ready to be modelled. The purpose of this section is to present the causal models used. The algorithms were carefully chosen according to the project's requirements, in terms of interpretability, and limitations (limited data available).

The models developed are listed below.

- Generalized Linear Models (GLM): Two models were developed regarding GLM - delta and absolute.

  Recalling the methodology mentioned above, the purpose of the CMI is, given a variation in the input variables, to obtain a delta in the output. The most straightforward way to do it would be to train the model directly with changes in input variables and respective output variations (designated, in this dissertation, as the delta model).

  However, this approach has a drawback. Considering that, for each country, there are 4 historical years, computing the differences between the figures for consecutive years, it lowers the number of observations from 4 to 3 for each country.

  On the other hand, the absolute GLM approach enables the use of a larger number of observations. The absolute model, instead of training with variables variations, is trained with the absolute values of input and output variables. Afterwards, the delta in the output variable can be calculated based on the absolute prediction difference between the alternative and baseline scenarios.

- Panel regression: The data collected is structured in a panel data format. As described in section 2.1.5, it is possible in an analysis like this one extract fixed effects and test how significant those are. This additional effect is the main difference between GLM and panel regression.

  In this case, countries, given the inherently different culture and habits, may by themselves differ in tobacco consumption patterns, so isolating the effect of each country may allow to extrapolate, for the remaining variables, a more generalized and less biased impact towards some more specific patterns of certain countries.

- Random Forest: This algorithm emerges as a trade-off between interpretability and flexibility. Random Forest compared with linear multiple regression is less interpretive since, while in regression models, it is possible to extract the variables importance as well as their impact, the random forest only retrieves a feature importance matrix. It is, however, a more flexible model to more complex (and non linear[5]) data structures since the basis of this algorithm are the decision trees, which in turn are nonlinear models built from linear limits (Koehrsen, 2018a). Due to the way the decision tree is built, this model has also the ability to capture variables interactions besides individual variable impact.

It should be taken into account that only the first model was created directly based on deltas and the remaining ones are considered absolute models.

[5] unlike linear models, input variables can be processed without any transformation to linearize them

All the algorithms described are available in the Statsmodels package in Python, except random forest from which a different package, h2o, was used.

Evaluation and parameter tuning

For this project, the metric chosen to compare models was the Mean Absolute Error. Comparing with other accuracy measures, MAE is easier to interpret than MSE and, unlike MAPE, when dealing with actual values that are close to zero (e.g. balances of products recently introduced to the market) this metric do not tend to infinite.

To obtain a more reliable model performance, as mentioned in the section 2.1.3, it was chosen to use cross validation since this method is not biased by the choice of the training and test dataset since all observations are used as test data.

Due to the small number of observations in each segment of analysis, the models are prone to overfitting. The detection of overfitting is done by evaluating the model's performance for the training set and test set. If the performance is considerably better in the training dataset, it means that there has been overfitting.

According to Elite (2019), there are several ways to control overfitting, among which it is mentioned: enriching the dataset with more observations (which would not be possible at this stage of the project), reducing the number of noisy input variables or opting for early stopping in the model.

Concerning early stopping, decision trees are models that easily overfit the data. The use of ensemble models, such as random forest, may mitigate this effect, however, the risk still exists when the dataset is very limited. One of the decision tree parameters that can be modified to control tree growth is 'maximum depth' or, by other words, the maximum levels that a decision tree may have (Koehrsen, 2018a). Therefore, this variable was considered as an important parameter to be optimize.

Since the pool of data is small, the number of trees created by random forest was also considered as a parameter to be studied. The values for maximum depth ranged from 2 to 10 and the number of trees from 10 to 100, with 10-in-10 increments.

The hyperparameter tuning can be developed in two ways: Grid Search or Random Grid Search. The Grid Search consists of an exhaustive search of the best performed model as every parameter combination is tested. Another option consists of Random Grid Search, in which random parameter combinations are selected for testing. The advantage of this method is that it allows a search on a wider range of parameters (Koehrsen, 2018b).

In this project, it was decided to use the Random Grid Search.

4.2 Analysis of Covid-19's impact

Given the extreme uncertainty regarding the smokers' behaviour during the pandemic period, the analysis of Covid-19 on the tobacco industry has proven to be of critical importance for the company's strategical planning.

A set of 57 countries were studied, per quarter, for the years 2020 to 2021, regarding the impact of Covid-19 on the tobacco volume, value and number of smokers. As in the CMI, the incidence and ADC of tobacco for each country were considered, as well as the relationships defined in formulas 4.1, 4.2 and 4.3.

Given the fact that this analysis covers countries not yet within the scope of the first section, sources for tobacco incidence and consumption for the various countries needed to be modified. The expected tobacco forecast in a non-Covid situation was obtained through Euromonitor.

In order to analyse the impact on tobacco consumption, two aspects needed to be taken into account: the two possible Covid scenarios and the various effects on tobacco market related to the appearance of this pandemic.

As mentioned in section 3.2, there are two scenarios outlined by Euromonitor for which tobacco consumption was evaluated. These scenarios foresee the decline in the economy in terms of household income and unemployment rate as well as the expected travel reduction, for the year 2020 and 2021. The pessimistic Covid scenario is characterized by a more severe collapse of the economy, it also considers the possibility of a second proliferation wave of Covid by the end of 2020.

In view of the effects referred to in Chapter 3, the approaches followed for each of these are listed below:

- Macroeconomic changes: Based on Euromonitor data on expected income and unemployment for each of the scenarios and combined with the insights obtained by the CMI, the impact of macroeconomic changes has translated into a change in the incidence and consumption of tobacco in the various subcategories;

- Downtrading: Due to the panorama faced, there was a migration of consumption from premium to non-premium products, so the average price of each category, weighted by the volume of each segment, decreased. Given the impossibility of accessing volume data on premium and non-premium products, the best proxy was based on a price reduction proportional to the reduction in the PDI per capita indicator;

- Travel restrictions: Based on KPMG pre-covid data on tobacco inflow and outflow in European countries combined with the expected reduction in traveler arrivals and departures, it was possible to obtain an estimate of the reduction in cross-border tobacco sales for the lockdown months;

- Labour migrations: Based on Eurostat data on the number of emigrants and immigrants in each country, it was considered, during the lockdown months according to the different scenarios, that a part of the emigrants returned to their origin countries. In some countries due to a very high outflow of immigrants, and in other cases of emigrants inflow, the change in the smoking population in some countries has changed considerably;

- Behaviour shifts: Taken into account a study done in China regarding the consumption patterns during and after Covid lockdown as a proxy, it was considered that there are three main stages.

  The first one happens during the initial three weeks of lockdown, since people are a bit more stressed about the situation, the tobacco consumption goes up. The second stage takes place when people get used to the lockdown routine and, consequently, the consumption goes to the regular values. The third and last stage happens when the desconfinement starts. Due to the need to use a mask in public places, consumers are less likely to smoke and tobacco consumption decreases drastically during the following 8 weeks.

  Therefore, in order to calculate the pandemic consumption pattern regarding lockdown and disconfinement periods, adjustments to the consumption were performed, proportional to what had occurred in China.

In terms of effects, macroeconomic changes, downtrading, labour migrations and behaviour shifts directly affect the 4.1, 4.2 and 4.3 expressions.

Regarding travel restrictions, after calculating the expected volume, per quarter, for each of the scenarios, a readjustment was made to the volume. For some countries, the volume decreases because part of the tobacco considered represents cross-border tobacco that is no longer sold. For other countries, where the price of tobacco tends to be higher than the price practiced by border countries, there is a positive increase in volume since citizens are no longer able to buy tobacco in neighboring countries and start to consume domestic tobacco.

4.3 Operational monitoring

From an operational perspective, one of the most relevant processes for the company is the intermediate chain between the firm and the end consumer. This analysis emerges with the purpose of, in a single interface, consolidate all the KPIs combined with the ability to infer alternative scenarios if the initial input values turn out to be different from those projected.

The KPIs were grouped by classes as presented below.

- Industry and company's volume consumption;
- Industry and company's market value and value distribution by the intermediaries (retailers, wholesalers and distributors);
- Distribution, sales and investment costs;
- Outlets (total number in the market and percentage visited by the company's sales force).

Accordingly, the MS Excel interface was designed as following:

- An initial sheet for market selection and an index with a link to the remaining sheets;
- A page for consulting the baseline values;

- The third sheet allows the user to modify the variables as desired;
- A summary sheet of the alternative scenario in comparison with the baseline figures.

Contrary to what occurs in section 4.1, the scenario analysis in this case is based on a simple variable inference, assuming that, for each year, the price per unit remains the same, as well as other indicators per unit sold. If this assumption is not verified, the user has always the possibility to override any value in order to recalculate the performance metrics.

Additionally, the user has the possibility to fill the industry volume expected for the years of 2021 and 2022 and the remaining KPIs are pre-filled based on the trend calculated from historical years.

Chapter 5

Results

This chapter presents the results obtained when applying the methodology detailed in chapter 4. Section 5.1 focus on cross-market insights' results regarding feature selection as well as the accuracy performance of the models developed and the main insights extracted.

Section 5.2 portrays the impact of Covid-19 on tobacco consumption among the selected countries and details the impact of each effect on tobacco demand.

Lastly, section 5.3 is dedicated to describe the interface developed regarding route to customer process.

5.1 Strategic cross-market insights

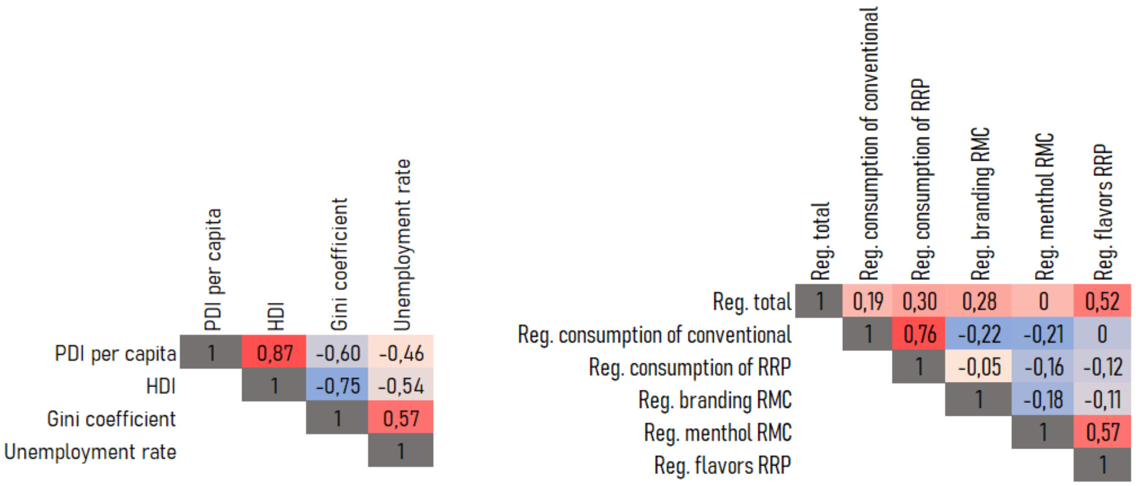
5.1.1 Feature selection

Taking into account the approach explained in chapter 4, there are several models to be developed for the prediction of the dimensions at the desired granularity. As a first step, for each analysis segment, the variables were critically selected for each model by combining the business perspective allied with insights from the literature review regarding the predominant factors in tobacco consumption.

However, a more detailed analysis for feature selection was preceded given the possibility of correlated input variables as well as variables with no significant explanatory value for the output variable.

Therefore, a matrix of correlations to groups of potentially similar variables was developed in order to identify highly correlated variables. At an early stage of the project, correlation matrices were developed for each model, however, for reporting purposes, it was decided to identify the groups of variables that proved to be most correlated. Those were: macroeconomic variables and variables related to tobacco consumption regulations. The correlation coefficients are presented in Figure 5.1.

It was used for analysis the absolute value of 0.7 as the correlation coefficient above which the variables are considered as highly correlated (Mindrila and Balentyne, 2017). By this criterion,



	PDI per capita	HDI	Gini coefficient	Unemployment rate
PDI per capita	1	0,87	-0,60	-0,46
HDI		1	-0,75	-0,54
Gini coefficient			1	0,57
Unemployment rate				1

	Reg. total	Reg. consumption of conventional	Reg. consumption of RRP	Reg. branding RMC	Reg. menthol RMC	Reg. flavors RRP
Reg. total	1	0,19	0,30	0,28	0	0,52
Reg. consumption of conventional		1	0,76	-0,22	-0,21	0
Reg. consumption of RRP			1	-0,05	-0,16	-0,12
Reg. branding RMC				1	-0,18	-0,11
Reg. menthol RMC					1	0,57
Reg. flavors RRP						1

Figure 5.1: Correlation coefficients [a] among macroeconomic variables (on the left) and regulation related data (on the right)

[a] The color tone is directly associated with the sign and magnitude of the coefficient. The more red, the higher the positive correlation between variables, the more blue, the more negative the correlation.

2 pairs of macroeconomic variables are identified: PDI per capita/ HDI and HDI/ Gini coefficient. Regarding the regulation variables, the following pair of variables stand out: conventional consumption regulation / RRP product consumption regulation.

In order to minimize the redundancy of the models, it was decided to remove the HDI and Gini coefficient variables since these, compared to the PDI per capita, are of less interest at the level of scenario analysis since they are subject to less variation. Regarding the variables related to the regulations applied to the consumption of conventional products and RRP, in the models where both variables were present, it was chosen to maintain the former since it is the conventional products that are most subject to restrictions.

For the non-significant variables of the model that simply produce noise, as mentioned in section 4.1.4, a two-step approach was applied for each model. In the first step, the significant variables for the model were identified. For linear regression, this analysis was done based on the p-values[1]. For random forest, the feature selection was based on feature importance matrix. Subsequently, with a smaller number of variables, the final models were developed.

5.1.2 Models accuracy

This subsection is intended to present the accuracy results of the different models. Table 5.1 presents the MAE for each segment of analysis and corresponding algorithm used.

The results for the delta GLM models are not presented in Table 5.1, since performance measures are not comparable between delta and absolute models as the magnitude of the dependent

[1] Particular cases of input variables presenting opposite impacts to those expected from the business point of view (for example, a price increase implying an increase in consumption) were removed from the model, even if significant.

variable is different. In addition, this model faced some difficulties in identifying significant effects since the variations that occurred from year to year are not very significant.

Table 5.1: MAE metric[a] for each segment of analysis and algorithm

	GLM Absolute	Panel Regression	Random Forest
Total market incidence	3,9%	1,7%	2,7%
Convencional balance	4,4%	3,4%	4,6%
RRP balance	4,8%	3,6%	5,1%
RMC balance	8,0%	3,9%	6,0%
FCT balance	6,9%	2,9%	6,2%
OTP balance	3,3%	1,7%	4,2%
E-vapor balance	9,3%	6,9%	12,2%
Heated tobacco balance	10,5%	8,5%	13,3%
Oral balance	14,1%	3,9%	8,3%
RMC ADC	2,4	0,8	2,4
FCT ADC	4,4	1,1	3,2
OTP ADC	3,0	1,4	2,1
E-vapor ADC	3,8	1,7	2,9
Heated tobacco ADC	4,9	3,3	3,4
Oral ADC	15,5	4,9	6,6

[a] Total market incidence and balance models' metric is given in percentage while ADC metric is given in sticks.

According to the results, the MAE is minimized for the models obtained by the panel regression. Compared to absolute GLM, the panel regression presents an additional effect - the country itself. In each model, each country is represented by a binary variable and therefore the resulting coefficient values indicate what is the baseline for each country in each analysis segment. This way, the impact of remain input variables are expurgated from potential effects of countries with particular characteristics.

Although, in the absolute GLM, the macroeconomic and demographic variables already aimed to distinguish countries with different levels of economic development and demographic characteristics, it can be concluded, from the difference in the results between GLM and panel regression, that there is an extra relevant factor that is not associated with only those variables of the country but with the customs of each culture already intrinsic in each country.

On the other hand, regarding random forest, which compared to the panel regression always presents a worse performance, also comparing with the absolute GLM, in several segments of analysis, presents a lower performance. This effect is because random forest models have adapted too much to the training date and are not able to generalize when applied on the test date. This effect is called overfitting. Although a feature selection was performed and the growth of decision trees were controlled by the tuning of the maximum depth, there are still signals of overfitting due to data limitation, especially for recently introduced products models.

5.1.3 Main outcomes

Due to the length required to present the entire process for the various segments of analysis in each algorithm, it was chosen to focus only on the panel regression results since it was the one with best performance.

The variables considered most relevant for the different analysis segments are presented below. Due to the confidentiality required regarding the insights obtained in this project, the values of the coefficients were changed in order to not reveal the exact values achieved. However, it was ensured that the sign of the impact as well as the relative magnitude among coefficients within related models were preserved.

Since not all countries are included in all models, it was chosen to present only the coefficients for the remaining significant input variables.

In terms of overall incidence in the country, represented in Table 5.2, it was found that the increase in the unemployment rate tends to decrease the number of smokers in the country given. On the other hand, it can be seen that the RRP affordability also has a significant impact in the smokers choice to continue smoking or give up. A conventional smoker when faced with the choice to stop smoking once and for all or switch for a healthier tobacco product, the decision is affected by the ability of the smoker to buy these alternative products.

Table 5.2: Market total incidence' significant variables and respective coefficients

Market total incidence	
Unemployment rate	-0,02
RRP affordability	-0,01

Regarding the balances between conventional products and RRP (in Table 5.3), the reason for migration between the two categories is mainly given by price, as well as the maturity and available products of RRP, meaning that the smaller the price gap, the more stability RRP products are in the market, as well as increasing the quantity of products available from RRP, the greater the tendency to migrate to RRP.

Table 5.3: Category balances' significant variables and respective coefficients

Conventional balance	
Conventional affordability	-0,98
Conventional regulation	0,01
Number of RRP categories	-0,04
RRP maturity	-0,02

RRP balance	
RRP affordability	-1,55
Conv. - RRP price gap	-1,40
Conventional regulation	-0,02
Number of RRP categories	0,02
RRP maturity	0,01

Regarding the balances between conventional products, in Table 5.4, it can be seen that in the case of increased income, smokers tend to smoke FCT tobacco. Price is also an important factor, and the greater the price gap between RMC and FCT the greater the tendency to migrate from RMC to FCT. In addition, the consumption of OTP, a category usually dominated by dualistic consumers, is especially negatively affected in case of unemployment growth.

Table 5.4: Conventional subcategory balances' significant variables and respective coefficients

RMC balance		**FCT balance**		**OTP balance**	
PDI per capita	-0,01	PDI per capita	0,02	PDI per capita	-0,01
OTP existence	-0,01	RMC-FCT price gap	0,02	Unemployment rate	-0,05

In Table 5.5, for the balances associated with RRP products, it can be seen that economic factors also have a major influence, namely the unemployment rate and PDI per capita, as well as the price itself.

Table 5.5: RRP subcategory balances' significant variables and respective coefficients

E-vapor balance		**Heated tobacco balance**		**Oral balance**	
PDI per capita	0,05	Unemployment rate	-2,42	PDI per capita	-0,04
Unemployment rate	1,81	Heated tobacco affordability	-3,88	Unemployment rate	-0,01
HT - E-vapor price gap	1,06	E-vapor existence	-0,28		

Regarding the average daily consumption of the different subcategories, whose values are presented in Tables 5.6 and 5.7, a common factor that stands out is solos. The higher the proportion of smokers in the subcategory who smoke exclusively that product, the higher is the average consumption. In addition, another relevant factor is the unemployment rate. An increase in the unemployment rate tends to reduce the consumption of RRP products and increase conventional consumption.

The regulations affects mainly the consumption of RMC and Heated tobacco products and price is specially relevant for RMC and Oral products.

Table 5.6: Conventional ADC' significant variables and respective coefficients

RMC ADC		**FCT ADC**		**OTP ADC**	
RMC affordability	-0,26	Unemployment rate	0,03	Unemployment rate	0,31
Unemployment rate	0,02	FCT solos	0,02	OTP solos	0,01
Conv. consumption regulation	0,01				

Table 5.7: RRP ADC' significant variables and respective coefficients

E-vapor ADC		**Heated Tobacco ADC**		**Oral ADC**	
Unemployment rate	-0,41	Unemployment rate	-0,63	Oral affordability	-0,73
		RRP regulation	-0,24	Oral solos	0,01
		RRP consumption regulation	0,14		
		RRP Heated tobacco solos	0,02		

5.2 Analysis of Covid-19's impact

As described in section 3.2, the goal of this analysis is to explore, at the selected country level, the impact of Covid on tobacco consumption according to a optimistic Covid scenario and a pessimistic one, taking into account the following effects: macroeconomic changes, downtrading, travel restrictions, labour migrations and behaviour shifts.

Although the analysis was developed at the country level and by subcategory, for the purpose of synthesizing the information and to preserve the confidentiality, it was decided, in Figure 5.2, to reflect the overall figures for the volume of tobacco expected to be sold for all the 57 countries in analysis. It reflects although very well the common pattern of consumption among the various countries.

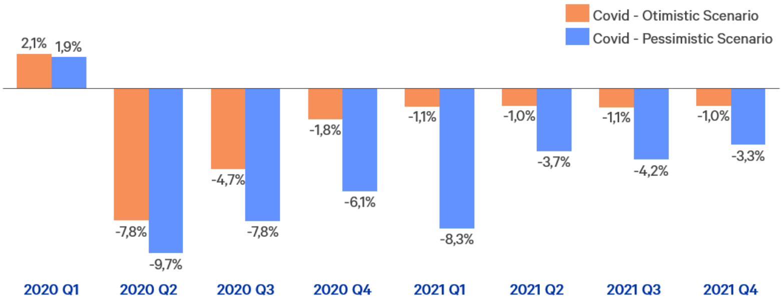


Figure 5.2: Percentual impact on market volume (vs no-Covid scenario)

Figure 5.2 illustrates, for each quarter and for the two scenarios under analysis, the percentage increase/decrease in volume comparatively to a non-covid scenario. There is an increase in volume in the first quarter of 2020 in both scenarios. This is mainly due to an increase in tobacco consumption during the first weeks of lockdown as people are at a high level of stress and still getting used to the new routine. On the other hand, the second quarter of 2020 is characterised by a more abrupt decline. In this period, people have become used to the new lockdown routine, the economic downturn has been more pronounced and, after disconfinement, restrictions imposed in public places encourage a reduction in consumption.

In the optimistic Covid scenario, after this wave of Covid-19 spread, in the following quarters a gradual return to normal consumption is expected. In the pessimistic scenario, since a second wave of Covid-19 proliferation is expected at the end of the year, in this scenario there is again an abrupt fall in consumption in the first quarter of 2021 (during the second phase of disconfinement).

In order to analyse more closely the impact of the effects described, Figure 5.3 details, for the most critical quarter (2020 Q2), how, in the optimistic Covid scenario, each effect contributes to the percentage reduction in volume compared to the scenario with no Covid.

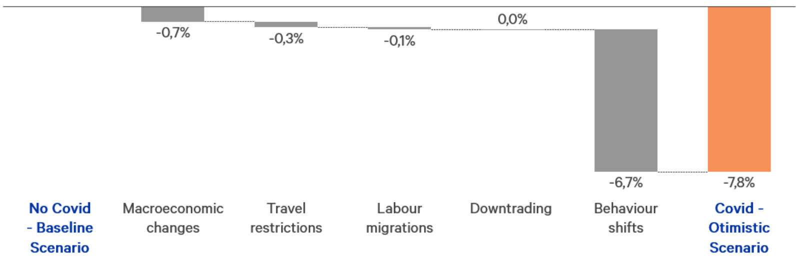


Figure 5.3: Split of percentual impact on market volume by the several effects, 2020 Q2

Consumer behaviour is mainly impacted by the new routine and restrictions to which it is subject due to lockdown and disconfinement, followed by macroeconomic changes that reduces the affordability of consumers to buy tobacco. The impact of these two effects are similar across the countries even thought the economic impact is more brutal in some countries than others.

Regarding travel restrictions and labour migrations, although the overall impact on the group of 57 countries is negative, it must be taken into account that these figures vary considerably from country to country. France is an example of a country that benefits from the closing borders, as residents start to consume more domestic tobacco. On the other hand, the Czech Republic, from where a lot of tobacco is exported, sees a bigger drop in tobacco sales due to travel restrictions.

On the other hand, in relation to labour migration, Switzerland is a country where the proportion of immigrants is substantially higher than emigrants and therefore, with workers returning to their origin countries during lockdown, the tobacco consumption in Switzerland is expected to be even lower. Unlike Lithuania, where the opposite is true.

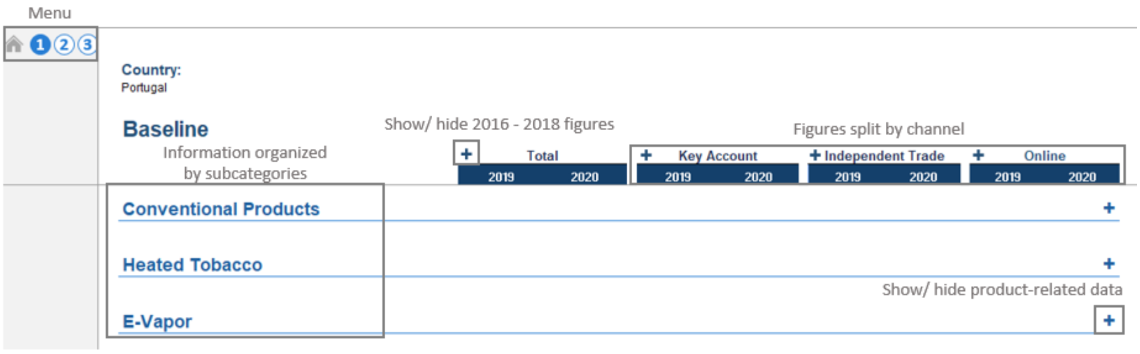
In this particular case, as the dimension represented in the Figures 5.2 and 5.3 is volume and not the market value generated, the effect of downtrading is ultimately not represented. However, due to the migration from more expensive to cheaper products, this effect enhances the percentage reduction of value generated from the no-Covid scenario to the Covid scenarios.

5.3 Operational monitoring

Taking into account the needs required regarding the interface for the workstream related to the route to consumer process, the MS Excel tool was designed as described next.

The first page is used by the user to select the country. Additionally, the user is presented with the index.

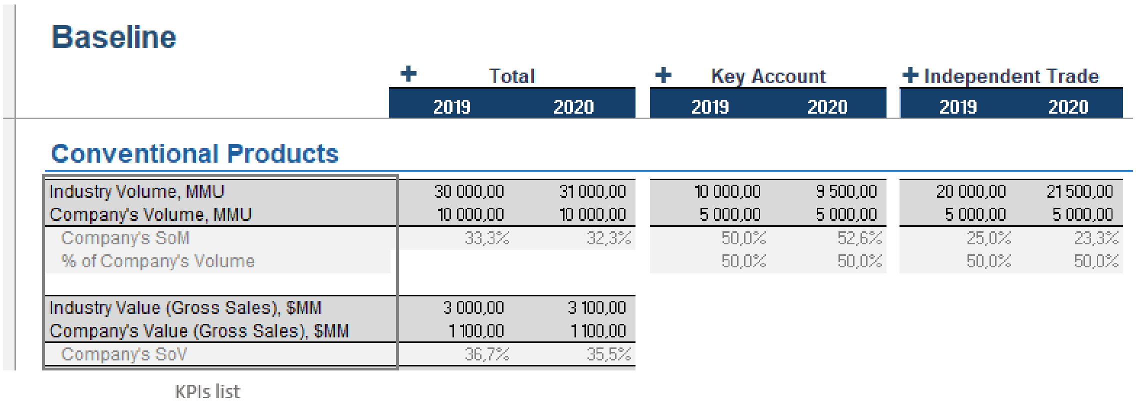
The design is similar between pages. As presented in Figure 5.4, there is a menu for page switch on the top left corner. On the vertical axis there are several tabs to subdivide the categories and, on the horizontal axis, the data is displayed in total and by channel.



Country:
Portugal

Figure 5.4: MS Excel tool standard layout

Each category can be expanded and a list of indicators is shown on the left, organized by classes of indicators (volume, value, costs and outlets), as partly exhibited in Figure 5.5[2].



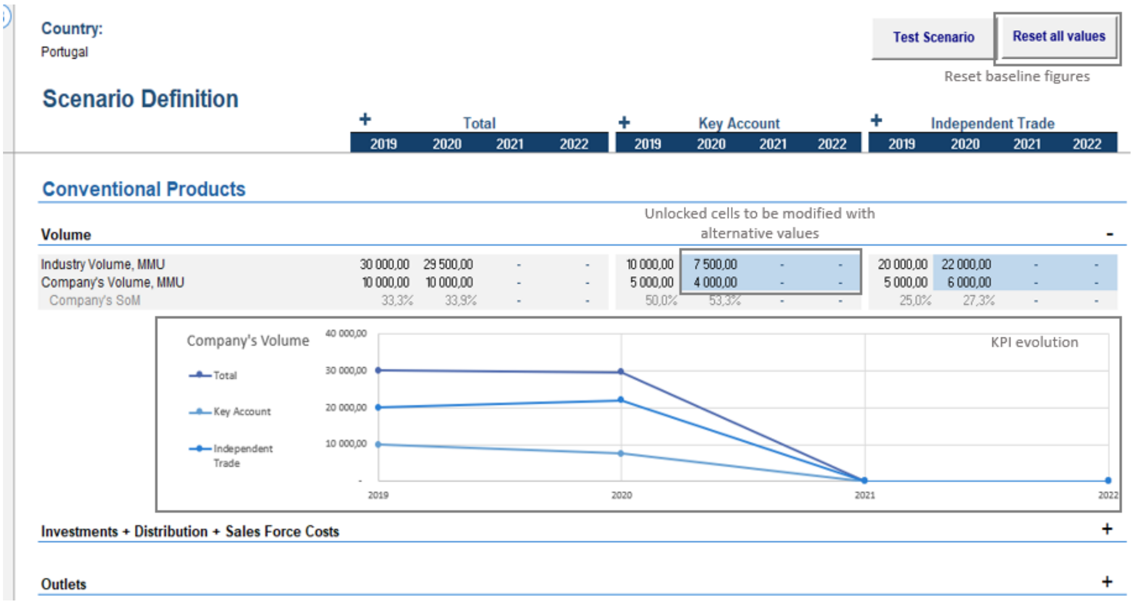
Baseline

	+ Total		+ Key Account		+ Independent Trade	
	2019	2020	2019	2020	2019	2020
Conventional Products						
Industry Volume, MMU	30 000,00	31 000,00	10 000,00	9 500,00	20 000,00	21 500,00
Company's Volume, MMU	10 000,00	10 000,00	5 000,00	5 000,00	5 000,00	5 000,00
Company's SoM	33,3%	32,3%	50,0%	52,6%	25,0%	23,3%
% of Company's Volume			50,0%	50,0%	50,0%	50,0%
Industry Value (Gross Sales), $MM	3 000,00	3 100,00				
Company's Value (Gross Sales), $MM	1 100,00	1 100,00				
Company's SoV	36,7%	35,5%				

KPIs list

Figure 5.5: Baseline page layout

Following, there is a page exclusively dedicated to alternative scenario definition where volume, value and costs values can be modified, as illustrated in Figure 5.6. For each set of indicators, by the analysis of the table and the graph presented, the user has the opportunity to check the historical values of the indicator at the channel-specific level and at the aggregated level.



Country:
Portugal

Reset baseline figures

Scenario Definition

	+ Total				+ Key Account				+ Independent Trade			
	2019	2020	2021	2022	2019	2020	2021	2022	2019	2020	2021	2022
Conventional Products												
Volume												-
Industry Volume, MMU	30 000,00	29 500,00	-	-	10 000,00	7 500,00	-	-	20 000,00	22 000,00	-	-
Company's Volume, MMU	10 000,00	10 000,00	-	-	5 000,00	4 000,00	-	-	5 000,00	6 000,00	-	-
Company's SoM	33,3%	33,9%	-	-	50,0%	53,3%	-	-	25,0%	27,3%	-	-

Unlocked cells to be modified with alternative values

KPI evolution

Investments + Distribution + Sales Force Costs +

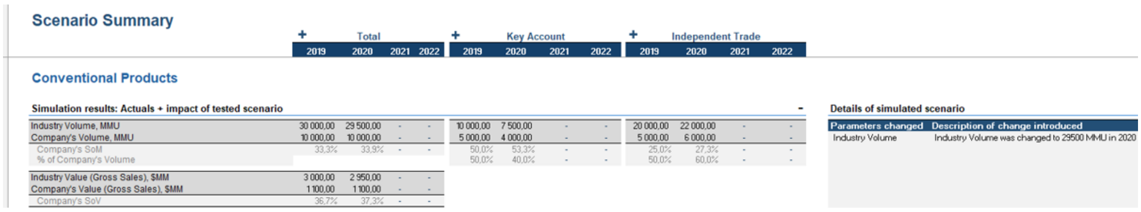
Outlets +

Figure 5.6: Scenario definition page layout

The blue cells represent figures that the user has the possibility to edit. For 2021 and 2022 years, the user, if desired, has the ability to indicate the volume of industry expected and the remaining KPIs are pre-filled. At any time, through the use of the button highlighted, the user has the possibility to reset the scenario.

[2] Due to the required confidentiality, in the interface highlights illustrated in section 5.3, the numbers presented are fictitious and the KPIs list is not fully displayed.

A third page, illustrated in Figure 5.7, is dedicated to the presentation of the alternative scenario summary. For each category there are two tables available for visualization. One, as illustrated in the Figure 5.7, reflects the total values of the new scenario, while the second table presents the differences between the alternative scenario and the baseline scenario.



Scenario Summary

Conventional Products

Simulation results: Actuals + impact of tested scenario

	+ Total				+ Key Account				+ Independent Trade			
	2019	2020	2021	2022	2019	2020	2021	2022	2019	2020	2021	2022
Industry Volume, MMU	30 000,00	29 500,00	-	-	10 000,00	7 500,00	-	-	20 000,00	22 000,00	-	-
Company's Volume, MMU	10 000,00	10 000,00	-	-	5 000,00	4 000,00	-	-	5 000,00	6 000,00	-	-
Company's SoM	33,3%	33,9%	-	-	50,0%	53,3%	-	-	25,0%	27,3%	-	-
% of Company's Volume					50,0%	40,0%	-	-	50,0%	60,0%	-	-
Industry Value (Gross Sales), $MM	3 000,00	2 950,00	-	-								
Company's Value (Gross Sales), $MM	1 100,00	1 100,00	-	-								
Company's SoV	36,7%	37,3%	-	-								

Details of simulated scenario

Parameters changed	Description of change introduced
Industry Volume	Industry Volume was changed to 29500 MMU in 2020

Figure 5.7: Scenario summary page layout

On the right side, the user is able to check the listed modified variables and the year for which the variable was changed.

Chapter 6

Conclusion

Given the features that shape the world nowadays, it is not an option for a company to do nothing to improve their operations and to innovate. If the company does not do anything, its competitors will certainly do it, so the company will lose for them its market share.

Recently, with the proliferation of Covid-19 all around the world, this need has became even more clear because companies had to quickly adjust their strategies to a new reality. In an uncertain world, the most agile companies are the ones that are most successful.

Therefore, more and more companies are choosing to leverage data and analytical methods to anticipate trends (e.g. demand) and monitor company's performance in order to provide a solid basis for key strategic decisions.

All the workstreams, developed and described in this dissertation, reflect the need the client company has perceived to leverage available data to create useful tools to improve strategic decision making.

Strategic cross-market insights

Since the client company is present in a large number of countries, totaling more than 100, each market currently has a considerably level of autonomy and it is responsible to design its own forecast. However, in some countries this forecast is not developed and, in other cases, there is a lack of information to analyse the impact of some events on tobacco consumption, for example, the implementation of a new regulation in the market. In addition, the assumptions used for analysis differ from country to country and the figures reported to the headquarters are sometimes misleading.

The first workstream of this project has the purpose of fighting all the gaps mentioned above by forecast the tobacco consumption at the national level for all the countries in which the client company is present and integrate it in a single interface for all markets.

Since eventual changes are expected to happen in the market (e.g. price changes and application of new regulations) as well as changes in the economy, the baseline forecast can be adjusted based on those variations[1].

[1] This stream was identified throughout the dissertation as 'Cross-market insights'

This tool provides the client company with a powerful decision support resource for strategic decisions.

Concerning the scenarios analysis segment, the aim was (i) to identify the variables with potential impact on consumption, (ii) to collect and prepare the information and (iii) to develop causal models to extract knowledge.

Initially, three categories of potential determinants of tobacco consumption were identified: macroeconomic, demographic and market variables. The information was collected through various sources of information. The market data was obtained through internal sources while the remaining data was obtained by public sources of information.

After collecting all the data, there was a detailed process of preparing the information, with identification of outliers and correlated variables.

The third step was to develop causal models in order to extract relevant insights. It was chosen to use causal models since those models have the particularity of allowing to play with relevant variables that influence tobacco consumption and, therefore, create a tool for sensitivity analysis based on those variables. However, for the design of the baseline forecast, the most precise methods are the time series forecasting models, since they better reflect the volume trend of conventional products (market stability or slight decrease), and the adoption curves used in cases of introduction of new products in the market. Both reflect time components that are more difficult to capture in causal models.

This difficulty was most evident in the development of models related to subcategories recently introduced in the market by tobacco companies (e-vapor, heated tobacco and oral), where the uncertainty of the model was higher compared to other models, as observed in Figure 5.1.

However, since slight changes in the market are expected to happen over the years, the forecast baseline readjustment must be done using causal models where the expected impact of variables can be measured.

Thus, four causal models were developed for each segment of analysis: GLM delta and absolute, panel regression and random forest. Due to reduced number of observations, more complex algorithms were not studied since the data was not enough to properly train those models.

For this project, linear models proved to be easier to interpret, give the developer the possibility to check if all effects were well captured and in the direction expected from a business point of view, as well as it was easier to implement.

From the results obtained, the panel regression model proved to be the best among the others, for all segments of analysis. The selection of variables, including macroeconomic and demographic variables, was carefully carried out, in order to test the impact of these variables on tobacco consumption, and also to illustrate the characteristics of the countries under study. However, it was found that there is an intrinsic effect to the country - the culture - that was only captured in the panel regression and turned out to be significant.

Besides the culture of each country, several variables were considered relevant across all countries in the analysis. In general, consumers have a greater propensity to give up smoking when the unemployment rate rises and/or the RRP affordability increases. The affordability tends to rise in cases of price increase or when the Economy is negatively affected and, consequently, people's disposable income decreases.

The migration between conventional to RRP products tends to increase in cases where the RRP price is relatively lower comparing to conventional prices or the supply of RRP products in the market increases.

Between subcategories of the same category, changes in the market's economy, translated by unemployment rate and PDI variables, as well as changes in price are the main reasons behind subcategory movement and changes in the daily consumption. The ADC is also affected by the market's regulations.

Although the impact of most variables turned out to be in the direction expected according to the business perspective, there is uncertainty associated with the coefficients due to data limitation (which makes the results very sensitive and unstable), misleading market information (due to template filling errors) as well as relevant time variables not considered in the models.

Also related with the absence of data, there are regulation variables that due to the low sample of countries where these restrictions, in the historical period, were implemented (e.g. menthol ban), the effect on the models is not significant. However, in this particular case of menthol ban, this restriction was implemented for all European Union countries in May 2020 (Gretler, 2020), which means that in the future these data can be used to capture the impact of this restriction on tobacco consumption.

In order to fight the uncertainty associated with the coefficients, two solutions arise. The first one, aligned with the topic mention above, is the inclusion of more countries into the analysis as well as more data from other years, in order to enrich the models with more information.

The second solution is the development of Ridge and Lasso regressions. The basic concept is the same as a simple linear regression, however, the way the coefficients are determined is different. These regularization techniques are a good alternative to a simple regression when there is a large number of input variables comparing to the number of observations to train - which is this case - because the model is likely to overfit the data (Jain, 2016).

These regressions, not only work towards minimizing the error between forecasts and actual values, but also penalize the magnitude of the coefficients. In the case of Ridge regression, the penalty is given by the square of the magnitude, while Lasso regression penalizes in terms of absolute value. Thus, the advantages, respectively, are the reduction of the magnitude of the coefficients and reduction of the complexity of the model while, in the case of Lasso regression, besides reducing the coefficients, also acts as a variable selector (Jain, 2016).

Covid-19's impact on tobacco industry

Sometimes companies are required to adapt quickly to the uncertain world in which we live, and in this case the emergence of Covid has meant a redefinition of priorities and an urgent focus on predicting its expected impact.

The main purpose of this analysis was a readjustment of the forecasts expected for the year 2020 and 2021 in order to empower the client company to better formulate decisions to mitigate the impact on tobacco consumption.

Five main effects of Covid-19, with impact on tobacco consumption, were identified: macroeconomic changes, downtrading, travel restrictions, labour migrations and behaviour shifts.

The synergies with the first workstream allowed to translate macroeconomic variations into changes in consumption. The approaches to analyse the impact of the remaining effects were based on the best proxy of the reality using the available data e.g., since the wave of Covid proliferation in China was earlier than in all other countries, it was possible to estimate the impact on other countries by extrapolating the information on the consumption pattern in China.

In general terms, the behaviour shifts that, due to lockdown period as well as disconfinement, led smokers to consume more tobacco in some periods and less in others, are expected to be the effect with most impact on the tobacco industry. Followed by macroeconomic changes. With the decline of the economy, more people became unemployed and family income decreased. Purchasing power decreases and some smokers are forced to consume less tobacco.

Operational monitoring

While the two previous workstreams focused on a broader perspective of tobacco consumption, this workstream emerged to focus on a process that is peculiarly important to any industry: the intermediate chain between company and end customer.

Each market is ruled by a common set of KPIs. The indicators are used to monitor the efficiency and effectiveness of the route to consumer process in each market over the years.

The main objective of this workstream was a consolidation of the company's KPIs in a single interface with the possibility to modify some figures and analyze the impact on the KPIs. This last feature is used over figures to which there are still uncertainty associated (predictions to 2020 and further years). For instance, with the appearance of Covid-19, it is expected that the initial values predicted for the year 2020 will turn out to be inaccurate. This tool allows markets to quickly and easily understand the impact in terms of performance metrics.

As a final conclusion to this dissertation, all the described workstreams empower the client company with agile tools for decision support, not only consolidating knowledge into a single interface but allowing the user to tailor the analysis to market uncertainty, and, therefore, have the ability to quickly and effectively adjust the analysis to new realities and take conscious strategic decisions to always be a better company.

Bibliography

Alam, M. (2020). Panel data regression: a powerful time series modeling technique. Available at: https://towardsdatascience.com/panel-data-regression-a-powerful-time-series-modeling-technique-7509ce043fa8 [Accessed on: 15-03-2020].

Awad, M. and Khanna, R. (2015). *Efficient learning machines: Theories, concepts, and applications for engineers and system designers*, pages 67–80. Apress Media LLC.

Bialous, S. and Glantz, S. (2018). Heated tobacco products: Another tobacco industry global strategy to slow progress in tobacco control. *Tobacco Control*, 27:s111–s117.

Borland, R. (2003). A strategy for controlling the marketing of tobacco products: A regulated market model. *Tobacco Control*, 12(4):374–382.

Brillio (2018). Demand Forecasting - Choosing The Right Forecasting Technique. Available at: https://www.brillio.com/insights/choosing-the-right-forecasting-technique/ [Accessed on: 20-03-2020].

Brownlee, J. (2018a). A Gentle Introduction to k-fold Cross-Validation. Available at: https://machinelearningmastery.com/k-fold-cross-validation/ [Accessed on: 11-04-2020].

Brownlee, J. (2018b). How to Remove Outliers for Machine Learning. Available at: https://machinelearningmastery.com/how-to-use-statistics-to-identify-outliers-in-data/ [Accessed on: 20-03-2020].

Brownlee, J. (2019). How to Choose a Feature Selection Method For Machine Learning. Available at: Available at: https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/.

DEI, M. (2019). Catalog of Variable Transformations To Make Your Model Work Better. Available at: https://towardsdatascience.com/catalog-of-variable-transformations-to-make-your-model-works-better-7b506bf80b97 [Accessed on: 21-03-2020].

Elite (2019). Overfitting in Machine Learning: What It Is and How to Prevent It. Available at: https://elitedatascience.com/overfitting-in-machine-learning#overfitting-vs-underfitting [Accessed on: 03-05-2020].

Frost, J. (2019). 5 Ways to Find Outliers in Your Data - Statistics By Jim. Available at: https://statisticsbyjim.com/basics/outliers/ [Accessed on: 05-03-2020].

Galarnyk, M. (2017). Understanding Boxplots - Towards Data Science. Available at: https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51 [Accessed on: 15-03-2020].

Garbade, M. J. (2018). Regression Versus Classification Machine Learning: What's the Difference? Available at: https://medium.com/quick-code/regression-versus-classification-machine-learning-whats-the-difference-345c56dd15f7 [Accessed on: 20-03-2020].

Garrett, B., Dube, S., Babb, S., and McAfee, T. (2015). Addressing the social determinants of health to reduce tobacco-related disparities. *Nicotine and Tobacco Research*, 17(8):892–897.

Gretler, C. (2020). EU Menthol Ban: Tobacco Firms Offer Alternatives to Cigarettes - Bloomberg. Available at: https://www.bloomberg.com/news/articles/2020-02-05/eu-menthol-ban-tobacco-firms-offer-alternatives-to-cigarettes [Accessed on: 20-05-2020].

Hawkins, B., Holden, C., Eckhardt, J., and Lee, K. (2018). Reassessing policy paradigms: A comparison of the global tobacco and alcohol industries. *Global Public Health*, 13(1):1–19. Available at: http://dx.doi.org/10.1080/17441692.2016.1161815.

Herman, C. (1974). External and internal cues as determinants of the smoking behavior of light and heavy smokers. *Journal of Personality and Social Psychology*, 30(5):664–672.

Jain, A. (2016). A Complete Tutorial on Ridge and Lasso Regression in Python. Available at: https://www.analyticsvidhya.com/blog/2016/01/ridge-lasso-regression-python-complete-tutorial/#three [Accessed on: 22-05-2020].

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *Introduction to Statistical Learning*, pages 24–26. Available at: http://www.springer.com/series/417.

Koehrsen, W. (2018a). An Implementation and Explanation of the Random Forest in Python. Available at: https://towardsdatascience.com/an-implementation-and-explanation-of-the-random-forest-in-python-77bf308a9b76 [Accessed on: 07-03-2020].

Koehrsen, W. (2018b). Hyperparameter Tuning the Random Forest in Python. Available at: https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74 [Accessed on: 06-04-2020].

Kuhn, M. and Johnson, K. (2013). *Applied predictive modeling*, page 488. Springer.

Mindrila, D. and Balentyne, P. (2017). Scatterplots and Correlation. Technical report.

Moffatt, M. (2018). What You Should Know About Econometrics. Available at: https://www.thoughtco.com/definition-of-econometrics-1146346 [Accessed on: 13-03-2020].

Murphy, C. (2019). Tobacco decline: Cigarette sales and advertisement drop in the U.S. Available at: https://eu.usatoday.com/story/money/2019/12/31/tobacco-decline-cigarette-sales-and-advertisement-drop-u-s/2777335001/ [Accessed on: 16-04-2020].

Palipudi, K., Gupta, P., Sinha, D., Andes, L., Asma, S., and McAfee, T. (2012). Social determinants of health and Tobacco use in thirteen low and middle income countries: Evidence from Global Adult Tobacco Survey. *PLoS ONE*, 7(3).

Pupale, R. (2018). Support Vector Machines(SVM) — An Overview. Available at: https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989 [Accessed on: 12-04-2020].

Ray, S. (2015). Regression Techniques in Machine Learning. Available at: https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/ [Accessed on: 17-03-2020].

Roka, A. (2019). Logarithmic Transformation in Linear Regression Models: Why & When. Available at: https://dev.to/rokaandy/logarithmic-transformation-in-linear-regression-models-why-when-3a7c [Accessed on: 20-03-2020].

Scollo, M., Bayly, M., White, S., Lindorff, K., and Wakefield, M. (2018). Tobacco product developments in the Australian market in the 4 years following plain packaging. *Tobacco Control*, 27(5):580–584.

Smolyakov, V. (2017). Ensemble Learning to Improve Machine Learning Results. Available at: https://blog.statsbot.co/ensemble-learning-d1dcd548e936 [Accessed on: 07-03-2020].

Sozańska, B., Pearce, N., Błaszczyk, M., Boznański, A., and Cullinan, P. (2016). Changes in the prevalence of cigarette smoking and quitting smoking determinants in adult inhabitants of rural areas in Poland between 2003 and 2012. *Public Health*, 141:178–184.

Subanti, S., Hakim, A., Sriwiyanto, H., and Hakim, I. (2019). The determinant of individual smoking consumption in Central Java province. 1321(2).

Walther, B. A., Moore Walther, J. L., Walther, B. A., and Moore, J. L. (2005). The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance Denmark (present address of B. Technical report.

Whiteside, E. (2019). Smokeless tobacco: 5 common questions about 'heat not burn' products answered. Available at: https://scienceblog.cancerresearchuk.org/2019/02/01/smokeless-tobacco-5-common-questions-about-heat-not-burn-products-answered/ [Accessed on: 23-04-2020].

World Bank (1999). Curbing the Epidemic. Technical report.

Wyckham, R. G. (1999). Smokeless tobacco in Canada: Deterring market development. *Tobacco Control*, 8(4):411–420.

Yadav, A. (2018). Support Vector Machines (SVM).
Available at: https://towardsdatascience.com/support-vector-machines-svm-c9ef22815589 [Accessed on: 12-04-2020].

Yiu, T. (2019). Understanding Random Forest.
Available at: https://towardsdatascience.com/understanding-random-forest-58381e0602d2 [Accessed on: 07-03-2020].

Appendix A

Selected regulations

The tobacco market is highly regulated, so the list of regulations is considerably extensive. However, there are some regulations whose impact is substantially higher than the others.

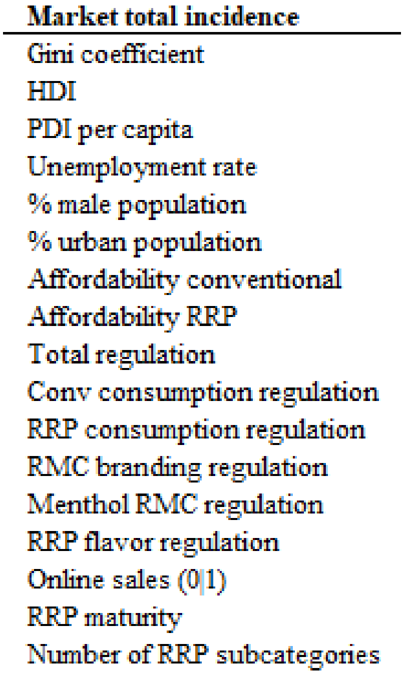
The following list comprises the most relevant topics subjected to regulation:

1. Product components:
   - Menthol flavour in RMC products;
   - RRP flavours;
2. Places to smoke:
   - Smoking in public places;
   - Smoking in Hotels, Restaurantes & Catering (HORECA) establishments;
3. Advertising:
   - Trade press advertising;
   - Media advertising;
4. Packaging branding;
5. RRP online sales.

Appendix B

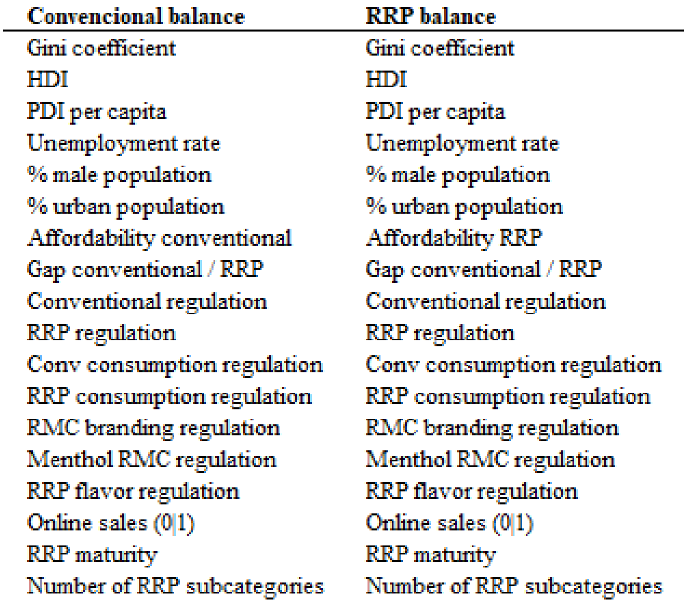
Variables selection

Since a detailed explanation of each analysis segment would imply a disproportionate extension of the document, in the main body of the document it was decided to define in general the variables that potentially affect the number of consumers and consumption of each, knowing, however, that the variables selected for each analysis segment should be critically selected according to the business perspective.



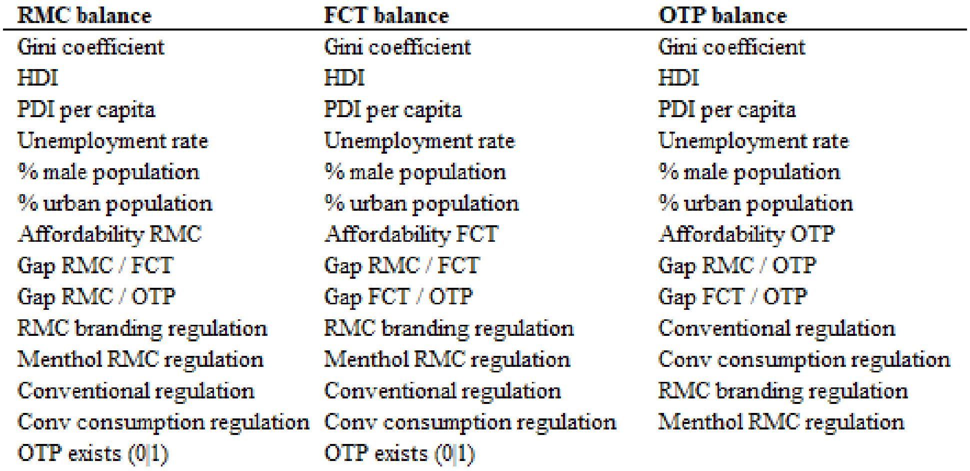
Market total incidence
Gini coefficient
HDI
PDI per capita
Unemployment rate
% male population
% urban population
Affordability conventional
Affordability RRP
Total regulation
Conv consumption regulation
RRP consumption regulation
RMC branding regulation
Menthol RMC regulation
RRP flavor regulation
Online sales (0|1)
RRP maturity
Number of RRP subcategories

Figure B.1: Explanatory variables of market total incidence



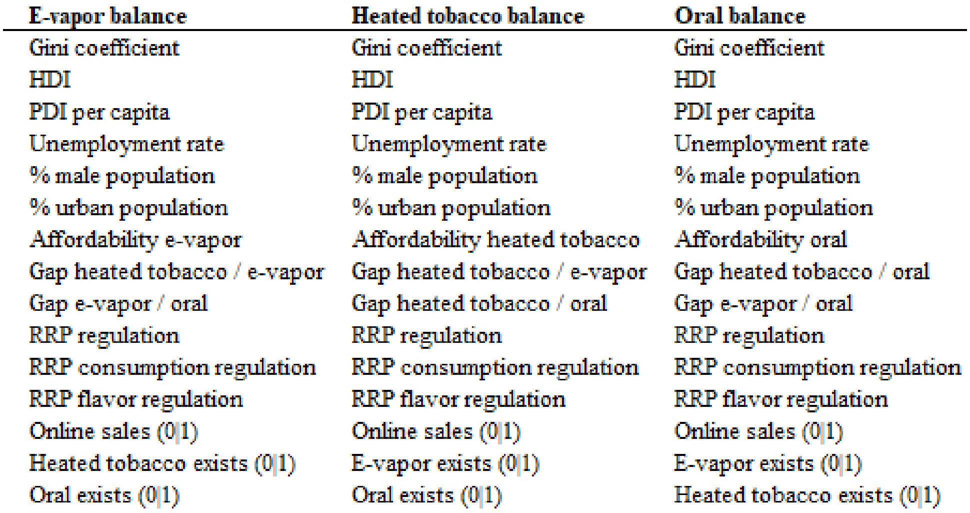
Convencional balance	RRP balance
Gini coefficient	Gini coefficient
HDI	HDI
PDI per capita	PDI per capita
Unemployment rate	Unemployment rate
% male population	% male population
% urban population	% urban population
Affordability conventional	Affordability RRP
Gap conventional / RRP	Gap conventional / RRP
Conventional regulation	Conventional regulation
RRP regulation	RRP regulation
Conv consumption regulation	Conv consumption regulation
RRP consumption regulation	RRP consumption regulation
RMC branding regulation	RMC branding regulation
Menthol RMC regulation	Menthol RMC regulation
RRP flavor regulation	RRP flavor regulation
Online sales (0|1)	Online sales (0|1)
RRP maturity	RRP maturity
Number of RRP subcategories	Number of RRP subcategories

Figure B.2: Explanatory variables used in category balances



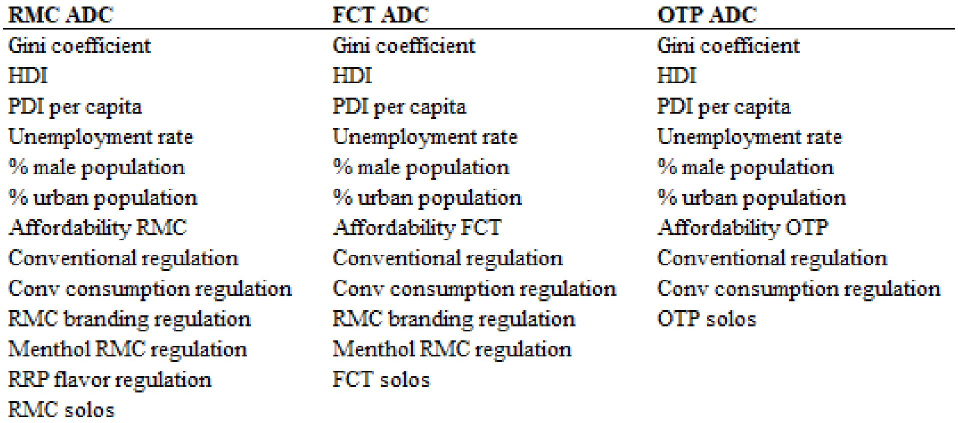
RMC balance	FCT balance	OTP balance
Gini coefficient	Gini coefficient	Gini coefficient
HDI	HDI	HDI
PDI per capita	PDI per capita	PDI per capita
Unemployment rate	Unemployment rate	Unemployment rate
% male population	% male population	% male population
% urban population	% urban population	% urban population
Affordability RMC	Affordability FCT	Affordability OTP
Gap RMC / FCT	Gap RMC / FCT	Gap RMC / OTP
Gap RMC / OTP	Gap FCT / OTP	Gap FCT / OTP
RMC branding regulation	RMC branding regulation	Conventional regulation
Menthol RMC regulation	Menthol RMC regulation	Conv consumption regulation
Conventional regulation	Conventional regulation	RMC branding regulation
Conv consumption regulation	Conv consumption regulation	Menthol RMC regulation
OTP exists (0|1)	OTP exists (0|1)	

Figure B.3: Explanatory variables selected to conventional' subcategories balances



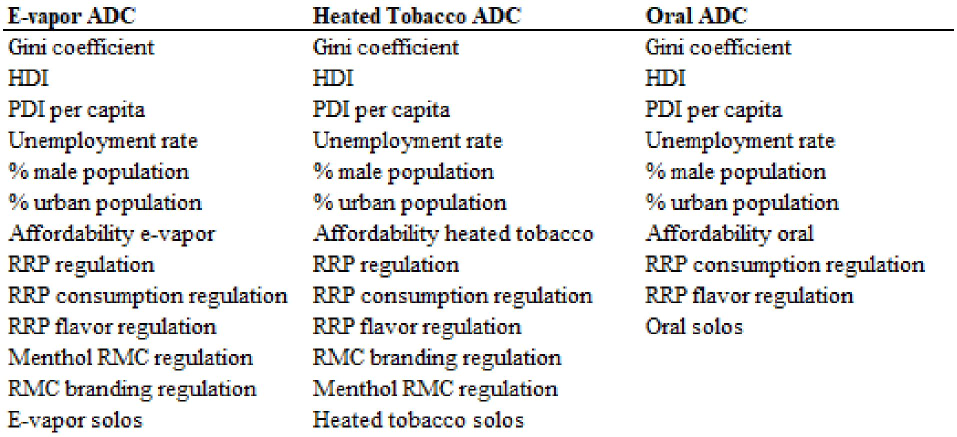
E-vapor balance	Heated tobacco balance	Oral balance
Gini coefficient	Gini coefficient	Gini coefficient
HDI	HDI	HDI
PDI per capita	PDI per capita	PDI per capita
Unemployment rate	Unemployment rate	Unemployment rate
% male population	% male population	% male population
% urban population	% urban population	% urban population
Affordability e-vapor	Affordability heated tobacco	Affordability oral
Gap heated tobacco / e-vapor	Gap heated tobacco / e-vapor	Gap heated tobacco / oral
Gap e-vapor / oral	Gap heated tobacco / oral	Gap e-vapor / oral
RRP regulation	RRP regulation	RRP regulation
RRP consumption regulation	RRP consumption regulation	RRP consumption regulation
RRP flavor regulation	RRP flavor regulation	RRP flavor regulation
Online sales (0|1)	Online sales (0|1)	Online sales (0|1)
Heated tobacco exists (0|1)	E-vapor exists (0|1)	E-vapor exists (0|1)
Oral exists (0|1)	Oral exists (0|1)	Heated tobacco exists (0|1)

Figure B.4: Explanatory variables selected to RRP' subcategories balances



RMC ADC	FCT ADC	OTP ADC
Gini coefficient	Gini coefficient	Gini coefficient
HDI	HDI	HDI
PDI per capita	PDI per capita	PDI per capita
Unemployment rate	Unemployment rate	Unemployment rate
% male population	% male population	% male population
% urban population	% urban population	% urban population
Affordability RMC	Affordability FCT	Affordability OTP
Conventional regulation	Conventional regulation	Conventional regulation
Conv consumption regulation	Conv consumption regulation	Conv consumption regulation
RMC branding regulation	RMC branding regulation	OTP solos
Menthol RMC regulation	Menthol RMC regulation	
RRP flavor regulation	FCT solos	
RMC solos		

Figure B.5: Explanatory variables selected to conventional' subcategories ADC



E-vapor ADC	Heated Tobacco ADC	Oral ADC
Gini coefficient	Gini coefficient	Gini coefficient
HDI	HDI	HDI
PDI per capita	PDI per capita	PDI per capita
Unemployment rate	Unemployment rate	Unemployment rate
% male population	% male population	% male population
% urban population	% urban population	% urban population
Affordability e-vapor	Affordability heated tobacco	Affordability oral
RRP regulation	RRP regulation	RRP consumption regulation
RRP consumption regulation	RRP consumption regulation	RRP flavor regulation
RRP flavor regulation	RRP flavor regulation	Oral solos
Menthol RMC regulation	RMC branding regulation	
RMC branding regulation	Menthol RMC regulation	
E-vapor solos	Heated tobacco solos	

Figure B.6: Explanatory variables selected to RRP' subcategories ADC