POLITECNICO
MILANO 1863

Nonparametric Statistics

Valentin Lacombe and Flavia Petruso

AA: 2022-2023

Analysis of Smoking trends in OECD countries

Abstract

In this work, we aim to examine how smoking prevalence has changed in Organization for Economic Co-Operation and Development (OECD) member states across the period 2008-2020. Additionally, we want to measure the level of deployment of anti-smoking policies and the extent to which they may impact cigarette smoking. To this aim, we start by analyzing the changes in cigarette smoking prevalence in OECD countries with a focus on socio-economic and gender-related differences. Then, we describe and quantify the deployment of anti-smoking measures in those countries. Finally, we construct a generalized additive model to predict smoking prevalence based on the affordability of cigarettes while controlling for country-specific and socioeconomic factors.

Keywords: *Smoking, tobacco, MPOWER, OECD, development, nonparametric statistics.*

The source code of the entire project,
including this report and the presentations, is available at
https://github.com/fl-hi1/Nonparametric-Statistics-project

Contents

1 Introduction **3**

1.1 Problem Statement . . . 3

1.2 Dataset Presentation . . . 4

1.3 Stakeholder Analysis . . . 5

1.4 Research Questions . . . 5

2 Data Preprocessing **7**

2.1 MPOWER Preprocessing . . . 7

2.2 Education Preprocessing . . . 7

2.3 Smoking Prevalence Preprocessing . . . 7

3 Analysis **9**

3.1 Human Development Index Functional Analysis . . . 9

3.2 Monitoring Changes in Prevalence . . . 10

3.2.1 Bagplot to Visualize Prevalence Changes . . . 10

3.2.2 Visualization of Prevalence Changes Using the HDI . . . 11

3.2.3 Univariate Permutational Tests on Prevalence Differences . . . 13

3.2.4 Permutational Test on Spearman Functional Correlation Coefficient . . . 13

3.2.5 Bootstrap Intervals on the Smoking Prevalence Differences . . . 13

3.3 Monitoring Changes in Anti-Smoking Policies . . . 14

3.3.1 Permutational Test on MPOWER Scores . . . 14

3.3.2 Bootstrap Reverse Percentile Intervals on Median MPOWER Scores . . . 15

3.3.3 Clustering Analysis on MPOWER . . . 15

3.3.4 Investigating the Impact of MPOWER Clusters on Smoking Decrease with Permutational Anova . . . 18

3.4 Prediction of Prevalence . . . 19

3.4.1 Conformal Prediction Intervals for Smoking Prevalence with 2020 Data . . . 19

3.4.2 Generalized Additive Model on Prevalence . . . 20

3.4.3 Bootstrap Intervals on GAM Predictions . . . 22

4 Conclusions **23**

1. Introduction

1.1. Problem Statement

The World Health Organization (WHO) estimates that tobacco smoking is responsible for the loss of approximately 8 million lives worldwide each year[1]. Alongside deaths, smoking has a detrimental impact on quality of life, as evidenced by the staggering figure of 200 million disability-adjusted life-years attributed to tobacco smoking globally in 2019[2].

To address this tobacco epidemic, WHO member states adopted the WHO Framework Convention on Tobacco Control (WHO FCTC) in 2003. This regulatory framework mandates that adhering countries implement and monitor the deployment of anti-smoking measures, such as increased taxation, stringent regulations on advertising and packaging, as well as comprehensive smoking cessation programs. This set of measures is known under the acronym of MPOWER, standing for

- M: **Monitor** tobacco use and prevention policies
- P: **Protect** people from tobacco smoke
- O: **Offer help** to quit tobacco use
- W: **Warn** about the dangers of tobacco
- E: **Enforce bans** on tobacco advertising, promotion, and sponsorship
- R: **Raise taxes** on tobacco products

The implementation and monitoring of these combined actions has been shown to save lives and reduce costs from averted healthcare expenditure[3,4,5].

The social, economic and healthcare costs of the smoking epidemic are measurable not only on a global scale but also specifically within **Organisation for Economic Co-operation and Development** (OECD) member states. These advanced economies, known for their high living standards, are not immune to the adverse consequences of smoking tobacco. Indeed, smoking-related diseases pose a significant burden on healthcare systems and public health within OECD nations[2]. Recognizing the severity of the issue, considerable efforts within these countries have been focused on implementing and reinforcing the above-mentioned MPOWER measures. By employing these comprehensive, multi-layered strategies, OECD member states aim to reduce the burden of smoking-related diseases and promote better health outcomes for the citizens. However, despite the tangible improvements made during the last decades[2], several countries are still far from the target outcomes set by WHO, and the effect of policy measures on smoking prevalence, although undoubted, has been difficult to quantify, also due to countries economic, social, and cultural heterogeneity, alongside the heterogeneity in the quantity and quality of data provided by different countries.

The aim of this work is to assess the changes in smoking prevalence in OECD countries across the period 2008-2020, while also measuring the level of deployment of anti-smoking policies and their potential impact on cigarette smoking.

1.2. Dataset Presentation

We gathered data on smoking prevalence, smoking-related policies, and socioeconomic factors for the 38 OECD member states from various publicly accessible databases. More specifically, we relied on three main resources:

- **WHO** database[6] for smoking prevalence, MPOWER, and cigarette-associated variables,
- **OECD** database[7] for GDP and Education,
- **Hunited Nations Development Program** database[8] for Human Development Index (HDI).

Table 1: Variables used for the current analysis with description and unit of measures

Variable name	Description	Unit of measure
	WHO database	
Prevalence_both	Smoking prevalence	Percentage of population above 15 years old (%)
Prevalence_males	Smoking prevalence across males	Percentage of male population above 15 years old (%)
Prevalence_females	Smoking prevalence across females	Percentage of female population above 15 years old (%)
Monitor	Initiatives to monitor tobacco smoking	Ordinal data: 1=no data, 2=lowest level, 5=highest level (4 for Monitor)
Protect	Initiatives to protect from tobacco smoking	
Help	Initiatives to help quitting smoking	
Warn	Initiatives to warn about the dangers of Tobacco	
Bans	Implementations of tobacco bans (television, social media, etc)	
Taxes	Raising of taxes on cigarettes	
Campaigns	Organizing anti-smoking campaigns	
Affordability	Price of cigarettes compared to GDP	Percentage of GDP to buy 2000 cigarettes of most famous brand
Cig_taxes	Proportion of taxes in final cigarette price	Percentage of final cigarette price which goes into taxes (%)
	OECD database	
Education_both	Tertiary education	Percentage of 35-64 y.o. population with tertiary education(%)
Education_females	Tertiary education across females	Percentage of 35-64 y.o. females with tertiary education (%)
Education_males	Tertiary education across males	Percentage of 35-64 y.o. males with tertiary education (%)
GDP	Gross domestic product per capita, corrected by purchasing power parity	US dollars ($)
	UNHD database	
HDI	Human Development Index (Population life expectancy, expected schooling level, GDP)	Adimensional index (0-1)

For each of the above variables, data from OECD member countries have been selected from 2008 and 2020, with a two-year interval. This selection allows for consistent time granularity matching the dataset for MPOWER. Indeed, the WHO Framework for Tobacco control has launched the MPOWER initiative in 2007-2008 and gathered data every two years[3]. In some of the analysis of the current report, different years have been used, based on data availability and the specific objectives of each analysis. For instance, since the majority of MPOWER data was available from 2007, we included 2007 in some analyses. On the other hand, socioeconomic data were available for all years, therefore for some analyses, we used the full yearly dataset, treating them as functional data. Additional information can be found in the Preprocessing section, which provides further details.

1.3. Stakeholder Analysis

This project aims to assess and quantify changes in smoking prevalence in OECD countries, with a focus on socioeconomic and gender-related differences. It will also try to evaluate how smoke-cessation policies have been implemented by OECD member states and to understand to which extent those policies have been effective.

The output of the present analysis is intended to be used within two documents with different objectives and levels of detail:

- A **dossier** to be distributed among policy-makers of OECD member states governments, aimed at informing political decisions and possibly influencing the healthcare and research sectors accordingly;
- A **chapter** to be included in the *Health at a glance 2023* report, a public document produced annually and shared through the OECD website to the general public and other actors, with the goal of summarizing the most important and up-to-date statistics related to smoking prevalence and measures.

The result of this analysis might therefore be of interest to several different stakeholders, ranging from governmental bodies of OECD countries to the general public, as detailed in **Table 2**.

1.4. Research Questions

As anticipated in previous sections, this analysis will try to answer to several questions, which can be divided into three main conceptual areas:

1. **Smoking prevalence changes**: have the OECD member countries decreased their smoking prevalence during the period considered and, if yes, to which extent? Do female and male smoking prevalence follow the same trends? Are the trends similar among countries with different socioeconomic conditions?
2. **Anti-smoking measures deployment**: how much has the MPOWER package been implemented by OECD countries overall?
3. **Prediction of prevalence, also based on anti-smoking measures**: can we build a model for smoking prevalence including the effect of anti-smoking polices? Based on this model, will the countries reach the SDG target of a 30% relative decrease of prevalence from 2010 to 2025?

Table 2: Stakeholder Analysis for the current work

	Stakeholder	**Role/Concerns**	**Power**	**Interest**
Government and Regulatory Bodies	OECD Member States	Implementing tobacco control policies and regulatory measures, fund initiatives	High	High
International Organizations	Public Health Organizations (OECD, WHO)	Advocacy, research, program implementation, public awareness	High	High
	Nonprofit Organizations	Advocacy, awareness campaigns, community support	High/Low	High
Healthcare Sector	Healthcare Providers	Patient care, smoking cessation support, policy implementation	High	High
Academic Sector	Research Institutions	Conducting studies, data collection, evidence-based recommendations	High/Low	High
	Academia	Research, education, policy recommendations	Low	High
General Public	Individuals and Communities	Public awareness, behavior change, support for policy interventions	Low	High
Tobacco industry	Tobacco Manufacturers and Distributors	Marketing, product regulation, profitability	High	Low
Others	Media	Disseminating information, shaping public opinion	Low	High

2. Data Preprocessing

2.1. MPOWER Preprocessing

First, we addressed missing values in the MPOWER dataset by replacing them with an average of the corresponding measurements from the same country in the immediately preceeding and subsequent year. In the case the first or the last year was missing, we approximated them with the value of the closest year available. In case this approximation yielded noninteger values, we opted for a conservative approach and truncated them to their integer value. The same approach was used on MPOWER data with level 1 (*i.e.,* not available data).

We also created a continuous MPOWER score which was built as follows:

$$\text{MPOWER}_{score} = \frac{(\text{Monitor} + \text{Protect} + \text{Help} + \text{Warn} + \text{Bans} + \text{Taxes}) - min}{max - min} \tag{1}$$

where *min* indicates the minimum score achievable, which was set to 12 (6x2, as level 1 was not allowed by our preprocessing), and *max* corresponds to 29, as 5 is the highest level of 5 MPOWER component and 4 is the highest level of *Monitor*, for a total of 5*5+4=29. This way, the score varies from 0 to 1, one being the highest policy achievable. The choice to build this score has been motivated by the need to create a unified measure of MPOWER policies and is also useful as the individual measures are difficult to process due to their high collinearity. Furthermore, it is a common way of assessing overall anti-smoking policies in the literature[5].

2.2. Education Preprocessing

2.3. Smoking Prevalence Preprocessing

The original data on prevalence provided by WHO (which was, to our knowledge, the only comprehensive and publicly available smoking prevalence source), were only collected for the years 2000, 2010, 2015, 2018, 2019, and 2020. However, as the other smoking-associated variables were present from 2007, 2008, 2010, 2012, 2014, 2016, 2018 and 2020, we decided to use the available smoking prevalence data to estimate the ones from the missing years. Based also on the visualization of our data, we made the crucial assumption that smoking prevalence from each country across different years can be seen as a **functional datum**, as it can be seen as a realization of an underlying regular phenomenon occurring through time.

We then decided to use two different approaches to estimate the missing data and confront the results. First, we used cubic b-splines (with knots at every original year) to interpolate our data and use the obtained coefficients to estimate the values for the missing years. Then, we used penalised smoothing splines with cubic b-spline basis and the smoothing parameter λ set to 100. We did the same work for the prevalence of both sexes, males and females. In **Figure 1**, we report one plot from the female prevalence.

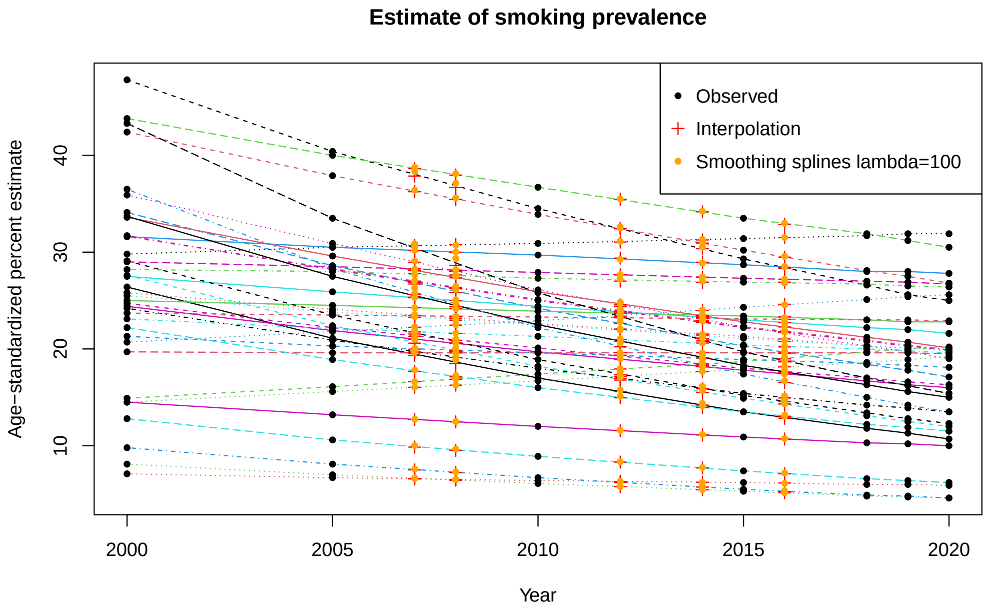


Figure 1: Estimates of female smoking prevalence on years 2007, 2008, 2012, 2014, 2016 and 2020 with interpolating splines (red crosses) and penalised splines (orange dots). Observed data are marked with black dots.

From the plot, it is clear that the two estimates are extremely similar. We also computed the Frobenious square norm of the difference between the two predictions and obtained extremely low values compared with the magnitude of the prevalences (percentages spanning from 0 to nearly 60 in 38x5 matrix). For the female prevalence, corresponding to **Figure 1**, the value of the squared Frobenious norm of the difference was 7.01.

Finally, we opted for the data estimates based on interpolation, considering the two predictions almost equivalent.

3. Analysis

3.1. Human Development Index Functional Analysis

Before moving to the exploratory data analysis and tests related to smoking prevalence, we wanted to introduce some ordering and a clustering structure based on socioeconomic factors to characterize the socioeconomic profile of OECD countries during the period under consideration to possibly use in further analysis. To this aim, we extracted the yearly HDI (2007-2020) for the OECD countries and we treated them as a functional datum. **Figure 2** shows the plot of the HDI with the median computed using modified band depth.

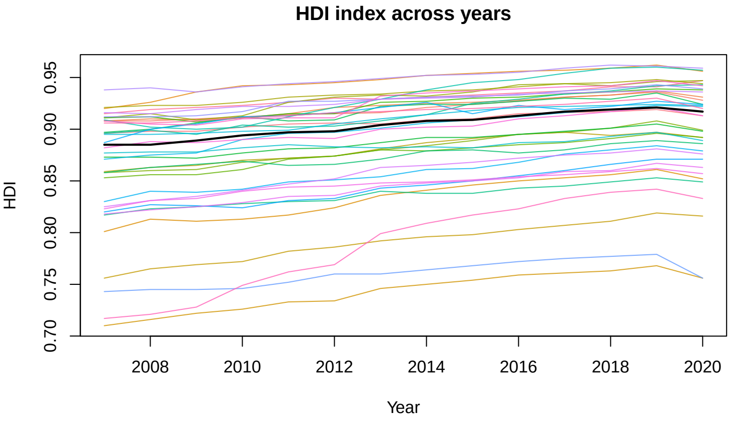
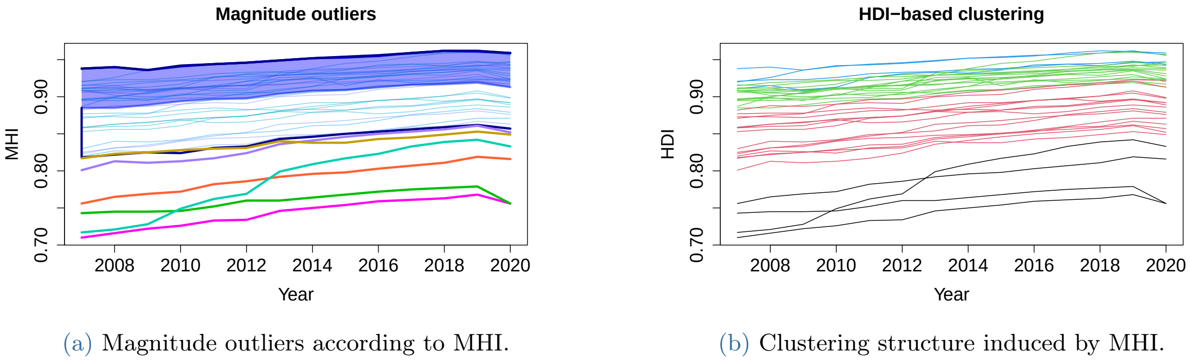


Figure 2: Plot of yearly Human Development Index in 2007-2020 with modified band depth median.

As curves seem to show a similar trend, we computed magnitude outliers based on the Modified Hypograhp Index (MHI). The plot with magnitude outliers based on MHI is shown in **Figure 3a**.



(a) Magnitude outliers according to MHI.

(b) Clustering structure induced by MHI.

Figure 3: Results on Human Development Index (HDI) based on Modified Hypograph Index (MHI).

We then used this functional depth to generate four clusters determined by the corresponding 10th, 50th, and 90th percentiles of MHI-induced depth scores. Results are shown in **Figure 3b**. With this choice of percentiles, the first cluster corresponds to the magnitude outliers of **Figure 3a**.

The countries corresponding to the clusters found are reported in **Table 3**.

Table 3: Clusters based on MHI computed on Human Development Index.

Cluster 1	Cluster 2	Cluster 3	Cluster 4
Colombia	Chile	Australia	Switzerland
Costa Rica	Czeck Republic	Austria	Germany
Mexico	Spain	Belgium	Denmark
Turkey	Estonia	Canada	Norway
	France	Finland	
	Greece	United Kingdom	
	Hungary	Ireland	
	Italy	Iceland	
	Korea	Israel	
	Lithuania	Japan	
	Latvia	Luxembourg	
	Poland	Netherlands	
	Portugal	New Zealand	
	Slovak Republic	Sweden	
	Slovenia	United states	

The HDI-based clustering is coherent with the health and wealth of the countries. In particular, the first cluster contains less socio-economically advanced countries from Latin America and Turkey; the second cluster contains mostly European countries with middle or middle-low economies and heterogeneous healthcare systems, plus some extra-European countries with similar features. On the other hand, the third and fourth clusters contain countries which are in general, richer and healthier. In particular, the fourth cluster contains northern European countries and Switzerland, renowned for their extremely efficient healthcare system and their rich economies, alongside Germany.

It is worth noticing that the choice of quantiles was driven by the goal of assessing the main groups of countries and the more extreme ones. Nevertheless, due to the absence of a net separation between clusters 2, 3, and 4, the merging of two or more of these clusters could also be performed, depending on the specific goal of the analysis and constraints on cluster numerosity.

We added the MHI index and the MHI-based clustering to the data which will be used in the next steps.

3.2. Monitoring Changes in Prevalence

3.2.1 Bagplot to Visualize Prevalence Changes

First, we aimed to assess the changes in smoking prevalence visually. To do so, we used three bagplots for overall prevalence, male prevalence and female prevalence in 2008 and 2020. As shown in **Figure 4**, all the countries seem to experience a decrease in prevalence, as shown by the bagplots lying almost entirely below the red dashed line which indicates the absence of change. From the bagplots, where the scale factor is set to 1.5, some countries appearing as outliers are Greece, France, Chile and Austria, with higher-general prevalence than the majority of the countries, whereas Latin American countries such as Colombia, Costa Rica, Chile and Mexico seems to experience lower smoking prevalence on both years, for both males and females. Interestingle, these countries are also among the top 20 countries for best MPOWER practices[3].

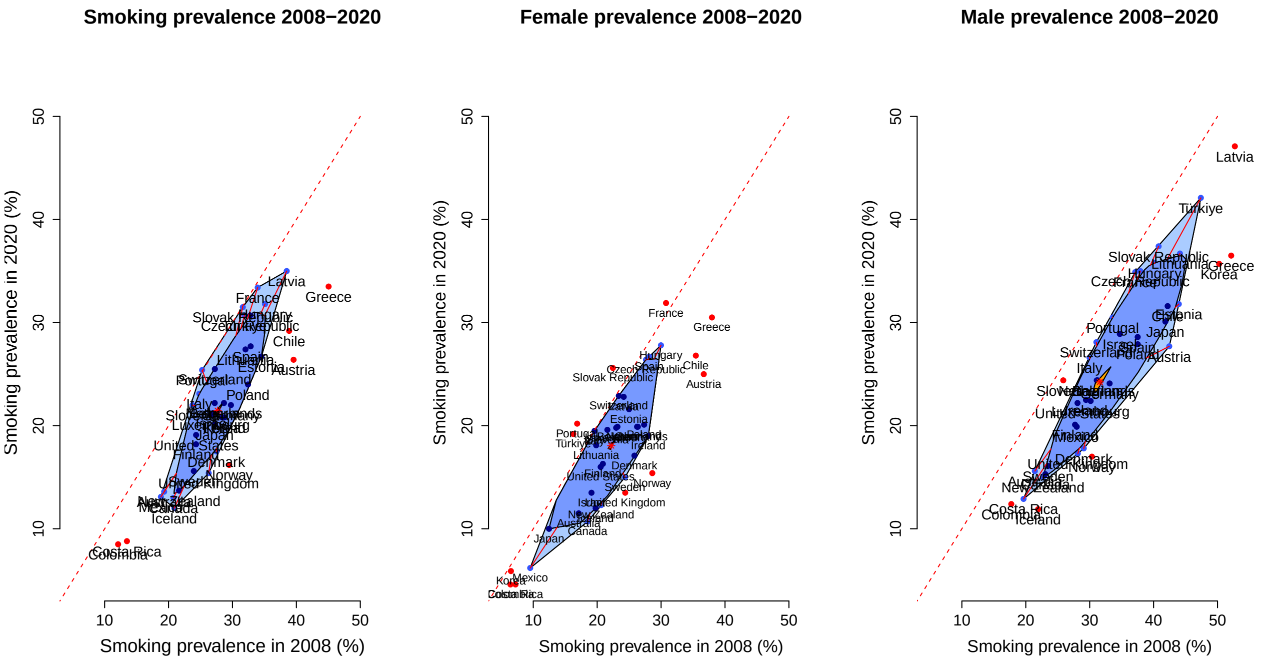


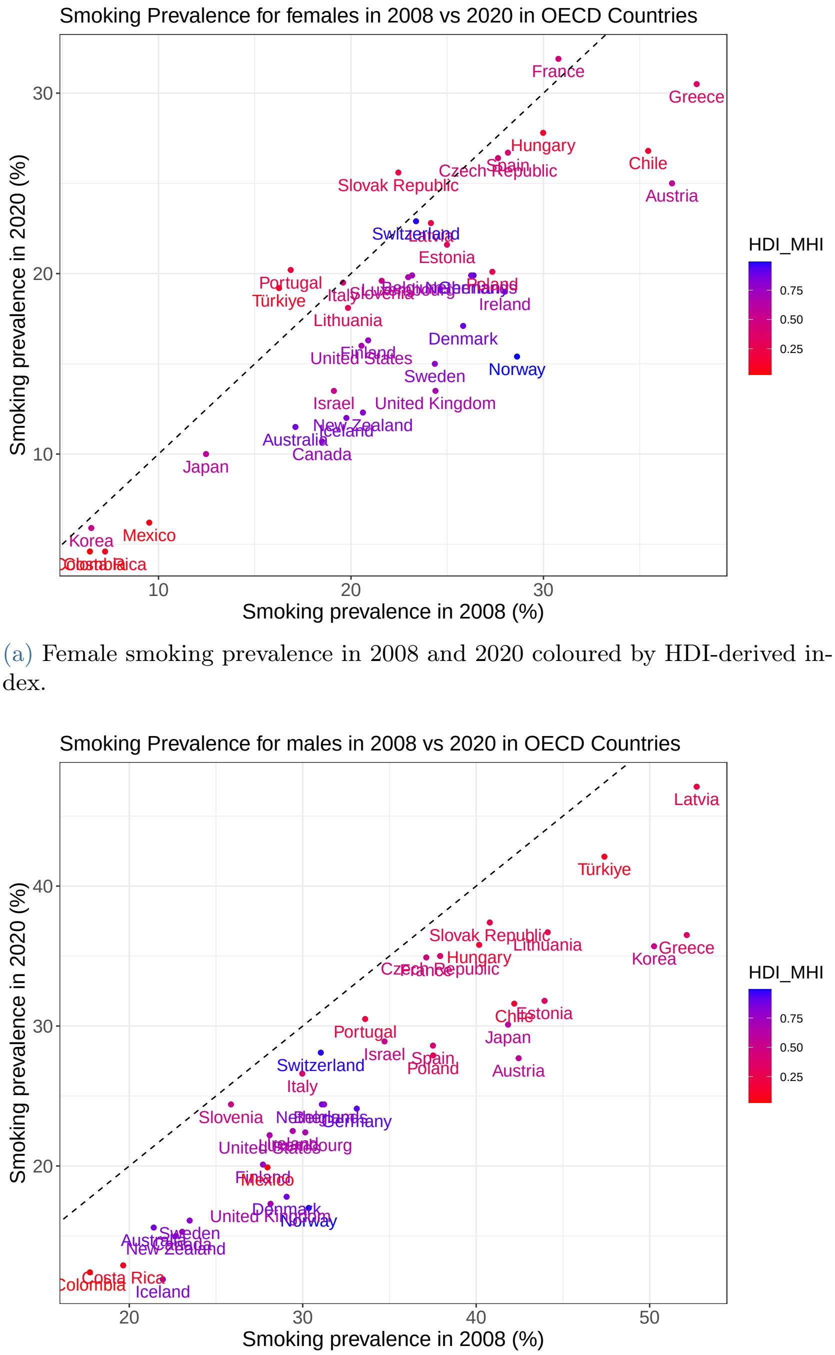
Figure 4: Bagplot of smoking prevalence in 2008 and 2020, factor=1.5. The red dashed line indicates no changes.

3.2.2 Visualization of Prevalence Changes Using the HDI

To obtain an alternative visualization which also took into account socioeconomic factors, we created a scatterplot for the prevalence in 2008 *vs* the one of 2020, where the color of each country quantifies the HDI-based index previously built (see **Section 3.1**). The plot is shown in **Figure 5**. From this visualization, two different types of considerations can be done:

- **Relation between prevalence changes and HDI**. The relation between prevalence, prevalence changes, and HDI seems complex and nonlinear: overall, OECD countries with higher scores position themselves in the middle of the prevalence scale in 2008 and 2020. Especially for the males, there seems to be a trend of high-HDI countries displaying lower smoking prevalence overall, but also a higher decrease from the baseline, as testified by a somewhat less steep slope in the countries within the prevalence range 20-30% in 2008 (including Austria, Sweden, New Zealand, Finland, United Kingdom), which can be framed as a "virtuous countries", and all belong to the HDI clusters 3 and 4. A particular case is represented by the Latin American countries Mexico, Colombia and Costa Rica which, despite the low HDI score (they all belong to cluster 1), have extremely low smoking prevalences, below 20% in 2008 and 10% in 2020.
- **Difference between male and female changes in smoking prevalence**. Although male prevalence has diminished in all OECD countries across the whole period, female trends are less homogeneous, with countries such as Portugal, Turkey, and the Slovak Republic experiencing a percentage increase. An interesting case is provided by Japan and especially South Korea, Asian countries where the difference in smoking prevalence among sexes is extreme. In 2020, in Korea 35% of over-15 year-old males are smokers, whereas less than 5% of women are. This phenomenon seems to a great extent attributed to the misperception of smoking women still

present in those cultures, resulting in lower prevalence. Furthermore, this may affect the result of self-reported surveys used to measure smoking estimates, which may produce results underestimating the number of smokers, due to the social desirability response bias[9]. This result claims for more reliable measures based on more objective evaluation criteria better to assess the entity of this phenomenon in those countries.



(a) Female smoking prevalence in 2008 and 2020 coloured by HDI-derived index.

(b) Male smoking prevalence in 2008 and 2020 coloured by HDI-derived index.

Figure 5: Male smoking prevalence in 2008 and 2020 coloured by MHI-based HDI index.

3.2.3 Univariate Permutational Tests on Prevalence Differences

Another aspect emerging clearly from this and other visualizations is the difference in smoking prevalence between sexes, with males displaying an average higher prevalence than females. We aimed to test this result not for all the years of the period under consideration. To do so, we ran seven univariate permutational tests, one for each year, for the difference between female and male smoking prevalence. Each test can be reformulated as a test for the center of symmetry of a univariate population. Here we report the test formulation:

$$\begin{aligned}
&H_0 : \mu(\text{male}_i - \text{female}_i) = 0 \quad vs \quad H_1 : \mu(\text{male}_i - \text{female}_i) \neq 0,\\
&\text{Test statistic}_i : |mean(\text{male}_i - \text{female}_i)|,\\
&\text{Permutation scheme}_i\text{: permutation of the intra-country differences}_i,\\
&\text{for } i \text{ in } \{2008, 2010, 2012, 2014, 2016, 2018, 2020\}.
\end{aligned} \tag{2}$$

For each test, we set the significance level α to 0.05 and the number of repetitions to 1000. For every year, the test yielded a p-value of 0, which in this case can be read as a confirmation that for the difference in prevalence between males and females is always significant. With such a low p-value, any correction for multiple comparisons would not change this conclusion.

3.2.4 Permutational Test on Spearman Functional Correlation Coefficient

Given the result of the previous permutational test, we proceeded to examine the presence of correlation between the smoking prevalence trends among females and males throughout the entire period under study. To account for the temporal dependence among the observations, we treated the smoking prevalence data from each country over the 7 years as a functional datum, and we computed the functional Spearman correlation coefficient based on the MHI. We obtained an estimate of ρ equal to 0.54, suggesting a moderate positive correlation. This means that the changes registered across the years are relatively similar between males and females, although not completely superimposable (a potential interpretation of this result will be given in the next paragraph). To assess the statistical significance of this value, we computed a functional permutational test with the following formulation:

$$\begin{aligned}
&H_0 : \rho_{MHI}(\text{male}, \text{female}) = 0 \quad vs \quad H_1 : \rho_{MHI}(\text{male}, \text{female}) \neq 0,\\
&\text{Test statistic : } |\text{rho}_{MHI}(\text{male}, \text{female})|,\\
&\text{Permutation scheme: permutation of functional observations}
\end{aligned} \tag{3}$$

we set the significance level α to 0.05 and the number of repetitions to 10000. The test yielded a p-value of 0.0005. Hence, we reject the null hypothesis of zero functional correlation.

3.2.5 Bootstrap Intervals on the Smoking Prevalence Differences

Motivated by the results of the previous test, ad by the presence of only a moderate positive coefficient, we hypothesized that this result could be due to the reduction of gender difference over the years.

To assess if the gender-related prevalence difference are changing across the considered period, we computed bootstrap reverse percentile confidence intervals on the difference between male and female prevalence for each year. To obtain an estimate which could take into account the majority of the

OECD countries, which could be of greater interest to our main stakeholder than the mean, we decided to compute bootstrap intervals on the 75th percentile of the distribution of the differences. We set the significance level α of each interval equal to 0.05, and the number of replications to 1000. Results are in **Table** 4.

Table 4: Bootstrap reverse percentile confidence intervals on the third quantile (corresponding to the 75^{th} percentile) of the distribution of smoking prevalence differences.

Year	Lower	75^{th} percentile	Upper
2008	15.60	17.29	24.39
2010	14.65	16.08	22.38
2012	8.42	14.55	19.85
2014	10.49	13.20	17.48
2016	7.68	11.87	15.52
2018	3.23	10.63	13.75
2020	4.05	9.73	12.35

Based on Table, it appears that the difference in prevalence is narrowing by approximately 1-1.5 points each two years. Interestingly, the bootstrap interval for 2008 is completely disjoint from the one for 2020 and contains strictly higher values, suggesting an important shift in the differences towards lower values.

In light of this and previous results, we can affirm that the overall trend of smoking prevalence for males and females is decreasing and that the gender gap is gradually closing.

3.3. Monitoring Changes in Anti-Smoking Policies

3.3.1 Permutational Test on MPOWER Scores

A second goal of the present analysis is to assess if and to which extent the MPOWER policies have been implemented by OECD countries. Boxplots of the MPOWER overall score (introduced in **Section 2**) distribution for each year is shown in **Figure 6**.

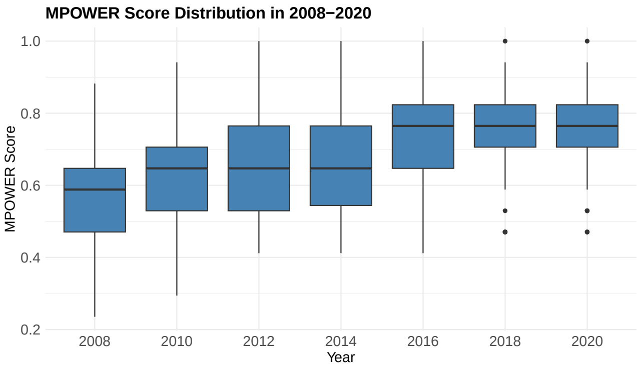


Figure 6: Distribution of overall MPOWER scores on each year.

We chose to investigate whether there was a notable alteration in the MPOWER score between 2008 and 2020. To accomplish this, we conducted a permutational test similar to the one employed for prevalence, but this time focusing on the median score. We opted for the median as a test statistic to obtain more robust estimates. Here, we report the test formulation:

$$\begin{aligned}
&H_0 : med(\text{mpower}_{2008} - \text{mpower}_{2020}) = 0 \quad vs \quad H_1 : med(\text{mpower}_{2008} - \text{mpower}_{2020}) \neq 0,\\
&\text{Test statistic} : |median(\text{mpower}_{2008} - \text{mpower}_{2020})|, \qquad (4)\\
&\text{Permutation scheme: permutation of the intra-country differences}
\end{aligned}$$

We set the significance level α to 0.05 and the number of repetitions to 1000. The test yielded a p-value of 0. We therefore reject the null hypothesis of zero difference in MPOWER score.

3.3.2 Bootstrap Reverse Percentile Intervals on Median MPOWER Scores

To have a quantification of the MPOWER changes between 2008 and 2020, we built bootstrap reverse percentile intervals on the median value of the MPOWER score in 2008, 2020 and on the median of the country-wise differences between MPOWER score in 2020 and 2008. We set the significance level α to 0.05 and the number of samples to 1000.

Table 5: Bootstrap reverse percentile intervals on median MPOWER scores and their difference.

Variable	Lower	Prediction	Upper
med(mpower$_{2008}$)	0.588	0.588	0.647
med(mpower$_{2020}$)	0.706	0.765	0.794
med(mpower$_{2020}$-mpower$_{2008}$)	0.147	0.176	0.235

From **Table 5**, it emerges that the bootstrap interval for the MPOWER score in 2008 is lower and completely disjoint from the one of the MPOWER score in 2020. Furthermore, when we look at the paired difference, we see that the bootstrap interval for the median does not contain 0 and is always greater than 0, suggesting that the MPOWER score median difference is greater than zero and indicates a median increase of between 0.147 and 0.235. This result testifies that the OECD countries have improved their anti-smoking policies over the years, in line with the global trends[3].

3.3.3 Clustering Analysis on MPOWER

This section aims to assess the similarities in the 38 OECD countries' MPOWER strategies by using all the MPOWER components in a less compact way than the overall score. Those strategies are quite challenging to cluster; indeed, they are composed of an evaluation of the six MPOWER components and are evolving through time. Moreover, only for this analysis, in addition to the MPOWER variables, we added the *Campaigns* feature, which is also provided by WHO and concerns the deployment of anti-tobacco media campaigns. To illustrate better what information is used from each country, see **Table 6** with an example based on Italian data.

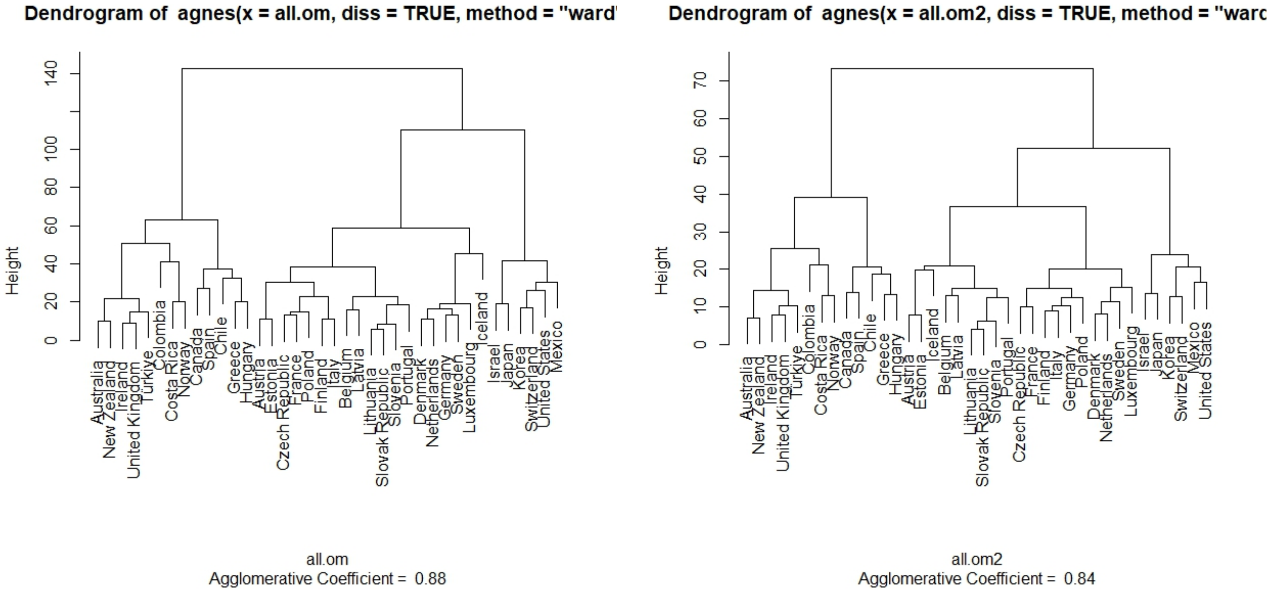
More precisely, out of those data, we would build objects named sequences, where we can define over a dissimilarity metric that would be used through a hierarchical clustering approach. To perform the different analyses, we will mostly rely on the *TraMineR*[10] package. Technically, a *sequence* is a succession of states through time. At first, we considered only one feature for the analysis of the strategies, with the set of states S being equal to $\{2, 3, 4, 5\}$. Then, we needed to set "dissimilarities" between

Table 6: Necessary information in order to build the sequence for the current analysis.

Country	Year	Campaign	Help	Warn	Bans	Protect	Monitor	Taxes
Italy	2008	4	4	3	4	2	4	5
Italy	2010	4	4	3	4	2	4	5
Italy	2012	3	4	3	4	2	4	5
Italy	2014	4	4	3	4	2	4	5
Italy	2016	5	4	5	4	2	4	5
Italy	2018	5	4	5	4	2	4	5
Italy	2020	2	4	5	4	2	4	5

those states i and j, which we decided by simplicity and common sense as $d(i, j) = |i - j|$, which will be associated as a *substitution cost* for what follows. Indeed, we need to define dissimilarity between sequences in the end, for that we will rely upon the notion of *optimal matching*. This dissimilarity between two sequences is defined as the minimum cost to go from one to another using only elementary operations, which are deletion/insertion (*i.e.*, *indel*) and substitution, where indel cost and substitution cost must be set. From this point, the implementation of the hierarchical algorithm is straightforward but would only account for one aspect of the policy of the countries.

To take into account all the aspects of the policy, we proceed as such. In the general wrap-up case, the number of all possible states would be equal to the number of 6-tuple made of basic states $\{2, 3, 4, 5\}$ which is of cardinality $4^6 = 4096$, which is huge for the standards of sequence analysis. Nevertheless, we have used a similar approach as before, and we generalized the distance between two states for substitution using two different approaches. The dissimilarity D_1 between two states S_1 and S_2 defined as $D_1(S_1, S_2) = \sum_{p_i \in policies} |p_i(S_1) - p_i(S_2)|$, and conversely D_2 such as $D_2(S_1, S_2) = \sqrt{\sum_{p_i \in policies} (p_i(S_1) - p_i(S_2))^2}$.

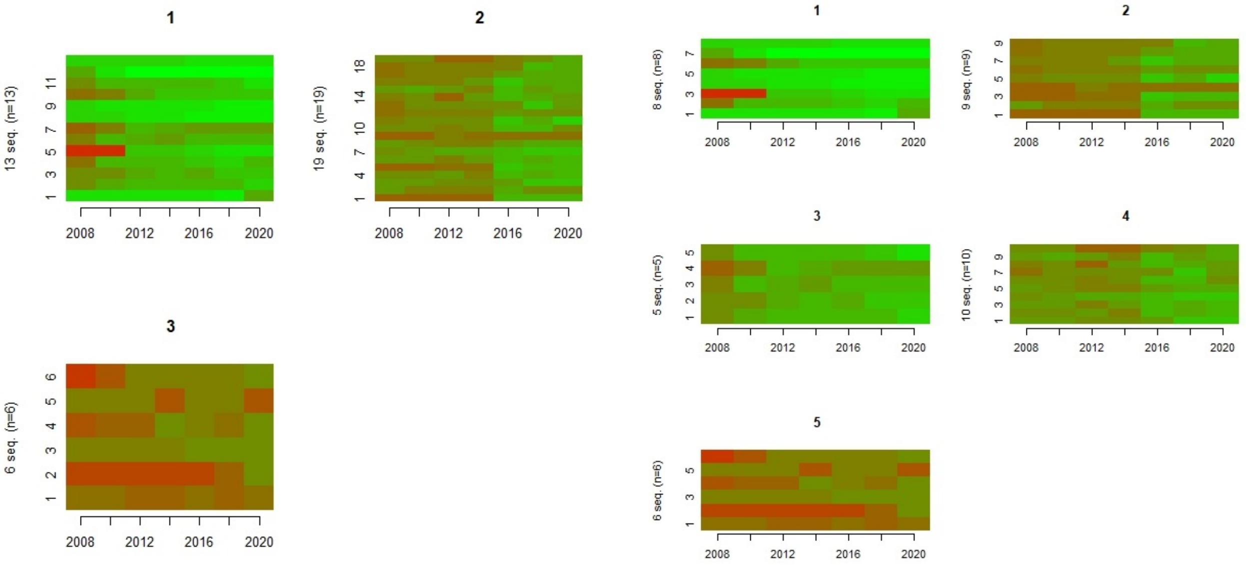


(a) Dendrogram using D_1.

(b) Dendrogram using D_2.

Figure 7: Dendrogram obtained using different metrics, both using Ward linkage.

After properly setting the indel cost, in both cases, as half of the maximum substitution cost, we computed the dendrogram using different linkages. By looking at its shape, the best linkage in both cases was the Ward one, intuition followed by the maximization of the *Agglomerative Coefficient*. By graphical analysis, see **Figure 7**, we observe that the possible cuts for D_1 make 2 or 3 clusters and for D_2 are made of 2,3 or 5 of them. As a further step, we are looking at the clustering structure which enables us to retain as much information as possible, so the most cluster as possible in both cases.



(a) 3-cluster structure using D_1

(b) 5-cluster structure using D_2

Figure 8: Sequences of retained clusters.

In **Figure 8**, we report a plot of the sequences for two different cluster structures considered. We have used a color grid from red to green as a scale from the worth possible state to the best possible.

In the 3-cluster structure, we observe very neatly three distinct behaviors.

1. Cluster one, with **massive involvement in tobacco policy throw time**. It represent the left one of the dendrogram of **Figure 7a** and also includes Latin American countries, Turkey (*i.e.*, the one with lower HDI already commented in previous sections).
2. Cluster two, with at first a poor involvement and then a **switch of behaviour for better policy around 2014**. It corresponds to the central one of the dendrogram and mostly includes countries with middle or middle-to-upper economies.
3. Cluster three, with **fairly poor involvement through time**. It is difficult to interpret in terms of socioeconomic factors, as it contains extremely heterogeneous countries such as United States, Korea, Mexico and Switzerland.

Concerning the other clustering version, it seems that clusters 1 and 2 have been separated. The second approach of clustering, therefore gave birth to a more refined version of the initial one.

3.3.4 Investigating the Impact of MPOWER Clusters on Smoking Decrease with Permutational Anova

Does the clustering structure really affect tobacco consumption, this is indeed a natural question that comes to mind, especially in the 3-cluster cases where different trends have been identified. More precisely, we would like to assess if the smoking variation of both sexes (in %) between 2008 and 2020, is significatively different according to the MPOWER strategy performed during these years.

To do so, we would perform a *One-way ANOVA permutational test* using as a T-statistics the F one, we set the number of permutations to 10000 and the significance level α at 0.05. Finally, we obtain a p-value equal to $p_{val} = 0,128$ which does not allow us to reject the fact that the different treatments (*i.e.*, tobacco policy) have no influence over the tobacco reduction in this scale of time. The result is somehow frustrating in the sense that this policy has been implemented for the purpose of reducing consumption, and there is no clear evidence of it considering those data. the coefficients associated with the different clusters are reported in table 7.

Cluster 1	Cluster 2	Cluster 3
-29.7%	-20%	-23.7%

Table 7: coefficient of the fitted ANOVA on the evolution of tobacco consumption 2008/2020 for both sexes.

As further steps, enlarging the group of countries to the analysis might give a better insight into the influence of the policies on tobacco consumption. Moreover, an approach aiming to consider mixed effects with other features might highlight more significant behavior.

As a further visualization including smoking prevalence, socioeconomic factors, and MPOWER score together, in **Figure 9** we plot the difference in smoking prevalence and MPOWER between 2008 and 2020, alongside the baseline smoking prevalence in 2008 (x-axis). The bubble size indicates the baseline MPOWER for 2008, whereas the colorbar is given by the HDI-based index, described in previous sections. From this plot, it seems that lower developed countries (in red), despite starting from a low MPOWER score and higher prevalence, had a greater increase in MPOWER measures, although these results have not emerged clearly from the permutational test.

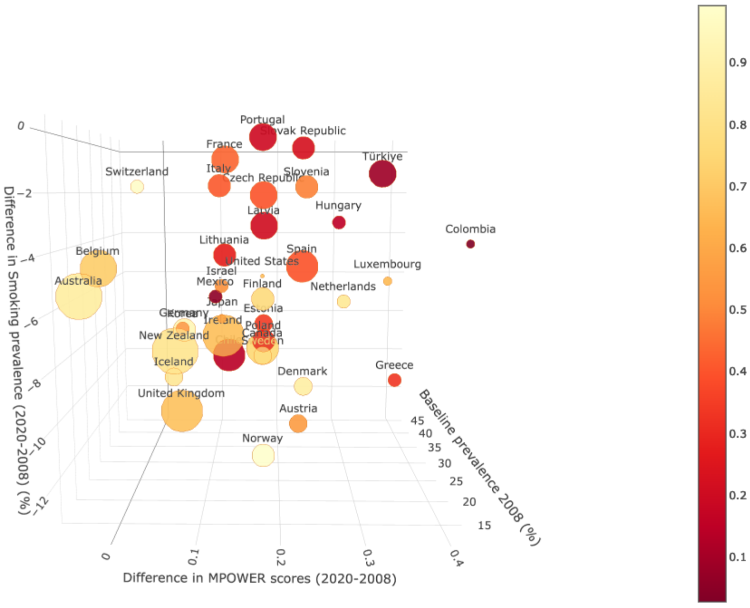
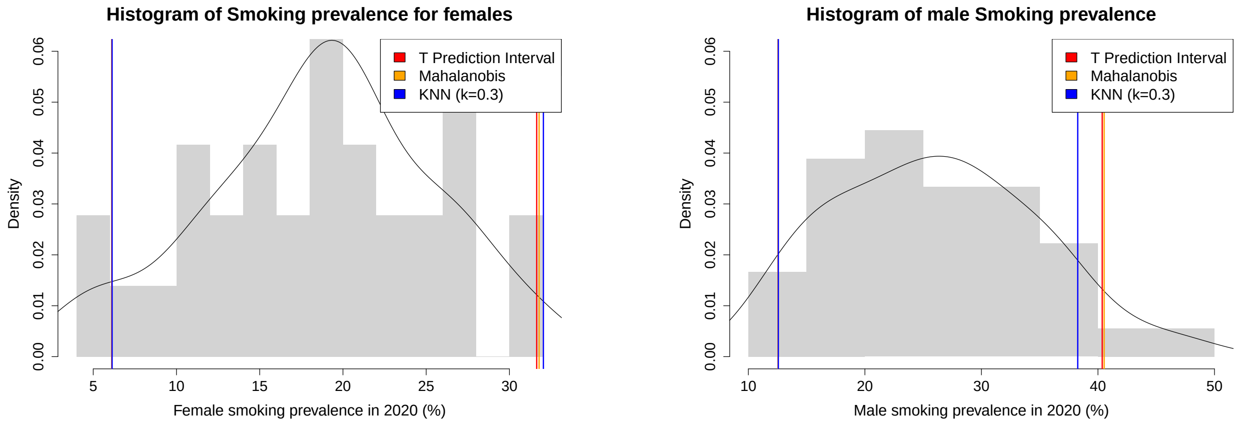


Figure 9: 5 dimensions components plot of OECD

3.4. Prediction of Prevalence

3.4.1 Conformal Prediction Intervals for Smoking Prevalence with 2020 Data

We first tried to predict possible future smoking prevalence in 2025 for an OECD country by using conformal prediction intervals based on the most recent prevalence data from 2020. We set the significance level α to 0.1 and a grid factor of 1.5. We confronted the results obtained from T-prediction intervals, Mahalanobis distance, and K-nearest neighbours (KNN) to compute the nonconformity score. For the KNN prediction interval, we set the fraction of the data used to build the interval to 0.3. The results are shown in **Figure 10** and **Table 8**.



(a) Conformal prediction intervals for female smoking prevalence using the data from 2020.

(b) Conformal prediction intervals for male smoking prevalence using the data from 2020.

Figure 10: Conformal prediction intervals for smoking prevalence using the data from 2020.

Table 8: Conformal prediction intervals for smoking prevalence using the data from 2020.

Variable	Lower	Upper
T Prediction Intervals		
Overall prevalence	10.65789	33.42105
Female prevalence	6.116165	31.660150
Male prevalence	12.57180	40.37143
Mahalanobis Prediction Intervals		
Overall prevalence	10.65789	33.55263
Female prevalence	6.116165	31.780075
Male prevalence	12.5718	40.5485
KNN Prediction Intervals		
Overall prevalence	10.78947	33.68421
Female prevalence	6.116165	32.019925
Male prevalence	12.57180	38.24662

From the plots and the table, we can see that the intervals obtained with the three approaches are extremely similar. However, the prevalence ranges are extremely wide, and therefore they do not seem suitable to draw any meaningful consideration beyond what has already been discussed.

3.4.2 Generalized Additive Model on Prevalence

To quantify the potential impact of anti-smoking policies on smoking prevalence in OECD countries, we decided to use a generalized additive model. Considering the data and the results from previous analyses, we decided to run two separate regression models for males and females.

In the original model, each observation was represented by the pair {Country, Year}, we wanted to account for potential sources of dependence between observations. For every model, we initially used the prevalence as a target and the following quantities as covariates:

- **Country**: factor variable added as fixed effect on the intercept.
- **Year**: continuous variable, added as a linear predictor.
- **Socioeconomic variable**s: HDI, GDP and Education (male or female) added as cubic b-spline terms to allow for more flexibility. We introduced also GDP and Education hoping that we could simplify some of the three variables using permutational tests. We made this choice since we wanted to identify for each model the socioeconomic aspect which was more appropriate to explain variability in the target.
- **Policy component**: for each model, we either included the MPOWER continuous score or the six MPOWER measures separated or the affordability variable. As we were interested in an effect which was easy to explain, interpret and communicate, we decided to add them as linear predictors.

From all the models we tried (including ones with lagged predictors), none of them showed significant coefficients for the MPOWER score. This may possibly be related to the fact that the overall MPOWER index we built is not fully adequate to explain prevalence changes when we control for all the other variables. Indeed, the index weights equally all components and does not allow to consider potential differences and lags between each measure implementation and its effect on the prevalence. Furthermore, we also tried some models considering the different MPOWER variables separately, but it was challenging to interpret the results due to the high collinearity among the measurements, and also the relatively few possible values (2 to 5). While we also considered techniques such as Multiple Correspondence Analysis to extract meaningful features from them, we decided to discard this approach due to the fact that it does not consider the ordinal nature of the variables.

An important methodological consideration concerns the link function. Indeed, as we are dealing with percentages, a more correct approach would require using a link function accounting for this. Indeed, we also tried several attempts using the *betar* function family with the *logit* link to build the GAM. However, as the prediction with both methods yielded similar results, and considering that the interpretation of the coefficients would be less straightforward to our potential stakeholders, we decided to simplify the approach and not change the link function.

Finally, we were able to obtain a model with a significant impact on policy when we considered the **female prevalence** and the **affordability** measure. After reducing the model to its significant components using permutational tests, the final semiparametric regression model is the following:

$$\begin{aligned}\text{Prevalence\_females}_i &= \beta_0 + \beta_{\text{Year}} \cdot \text{Year}_i \\ &\quad + \sum_{j=1}^{J} \beta_{\text{Country}_j} \cdot \mathbb{1}_{(\text{Country}_i = \text{Country}_j)} \\ &\quad + f(\text{HDI}_i) + \beta_{\text{Affordability}} \cdot \text{Affordability}_i, + \epsilon_i \\ \textit{with} \quad & \epsilon_i \sim N(0, \sigma^2), \\ & i \in \{1, ..., 222\}, \\ & J = 38. \\ & ,\end{aligned} \tag{5}$$

Here, the function f indicates cubic b-splines. In **Table 5**, we report the resulting coefficients for the linear terms (intercepts from the countries are in the "GAM" pdf file in the GitHub).

Table 9: GAM. coefficients

Variable	Coefficient	p-value
Intercept	601.49673	7.77e-09
Year	-0.29076	2.33e-08
Affordability	-0.69530	0.028110

As it is shown, the β coefficient for the year is negative and indicates that each year there is on average a 0.29% decrease in female smoking prevalence. As concerns affordability, a 1 unit increase (corresponding to 1% increase) in affordability on average causes a 0.69 % decrease in prevalence. As the affordability measure indicates the percentage of GDP per capita required to buy 2000 cigarettes, this result may suggest increasing the price of cigarettes, for example via taxation, results in a decrease of smoking prevalence.

We also provide the plot of the smooth term for HDI:

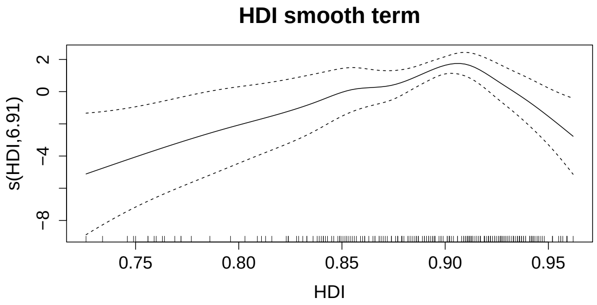


Figure 11: Smooth term for HDI.

It seems that for lower HDI values there is a positive effect of HDI on the prevalence (*i.e.*, increasing HDI results in increased prevalence), whereas for very high index values the prevalence decreases with decreasing HDI. Interestingly, this finding has the potential to validate and reinforce what has already been observed in **Figure 5a**. Over the years, in OECD countries with less advanced economies, the progressive female emancipation and societal role transformation might have contributed to a rise in female smoking rates. This phenomenon could explain the upward trend observed until approximately an HDI value of 0.90. In fact, throughout the period examined, all OECD countries experienced considerable economic growth, leading to higher HDI scores. Concurrently, in countries with initially low HDI scores (indicating less developed countries), the growth of this index closely coincided with advancements in women's emancipation. This connection may have resulted in a situation where increased HDI scores in developing nations were associated with higher smoking prevalence among females. On the other hand, this phenomenon was less pronounced in countries starting from a higher index, where the HDI growth has hence been paralleled by decreased female smoking prevalence, as we could expect as a general trend (and was indeed seen in regression on overall and male prevalence, although not shown in this report). However, despite the tempting speculations, further analysis should be done to investigate this hypothesis.

3.4.3 Bootstrap Intervals on GAM Predictions

Given that one of the objectives outlined by the WHO in Sustainable Development Goal 3 (SDG3) is for each country to achieve a 30% reduction in smoking prevalence between 2010 and 2025, our aim was to forecast female smoking prevalence for the year 2025 by using the GAM model obtained in the previous subsection. Subsequently, we calculated bootstrap intervals to estimate the uncertainty associated with these pointwise predictions.

To illustrate examples from countries belonging to HDI clusters, we report the case of Norway (a country with high HDI considered "virtuous"), Turkey (a low HDI country), and France (an middle-HDI country experiencing an increase in female smoking prevalence from 2010 to 2020). To determine the desired target prevalence for 2025 according to SDG3, we calculate it as prevalence2008 minus 0.3 times prevalence2008. For the prediction, we estimate HDI and affordability based on past values and the average rate of change observed in previous years. We refer to this approach as "business as usual". For the bootstrap confidence intervals, we set the level of significance α to 0.1. We report the results in **Table 10**.

Table 10: Bootstrap confidence interval on prediction for female prevalence using the GAM model with affordability. Red color indicates that the target has not been met.

Country	Est. HDI	Est. Affordability	Lower	Pred	Upper	Target
France	0.902	3.20	27.67214	28.32308	29.32093	21.63
Turkey	0.850	3.80	15.60299	16.41899	17.38266	11.69
Norway	0.966	2.22	14.57067	15.50573	16.28126	18.06

Based on the results provided in **Table 10**, under business as usual, neither France nor Turkey will be reaching their target by 2025, whereas Norway may even overcome it (the bootstrap interval is entirely below the target).

To conclude, this instrument, once refined, might prove extremely useful to identify countries needing more actions to reduce smoking prevalence, or to forecast the burden on the healthcare system.

4. Conclusions

This work aimed at providing information useful to policymakers of the OECD countries and figures suitable for the general public.

First, we conducted an analysis of **smoking prevalence** trends in OECD countries. The results revealed that member states are experiencing an overall decreasing trend with respect to 2008. Countries with more advanced economies seem to have lower prevalences than less advanced ones, while still experiencing a significant reduction. One interesting exception is provided by the Latin American countries, where, despite the low socioeconomic index, they have an extremely low prevalence. It is worth noticing that these countries are also the ones with the highest levels of MPOWER policies[3]. When looking at gender-based differences, the situation gets more nuanced. During the period under consideration, male and female prevalence underwent a significant reduction, and the difference in prevalence between males and females decreased. However, when we focus on female prevalence, some countries including Portugal and the Slovak Republic experienced an increase in smoking prevalence. One hypothesis -beyond cultural factors- is that in countries with a lower development at baseline, the increase in HDI was paralleled by a relative increase in female smoking prevalence (after controlling for other variables), whereas in countries with more advanced economies from the beginning the increase in HDI was paralleled by a decrease in female smoking. These considerations are also supported by the results of the HDI coefficient of the generalized additive model but deserve further investigation, as it might be necessary to devote more attention to interventions to avoid this "female emancipation" effect on prevalence in developing countries. Last, the extremely different results for males and females in Asian Countries such as Korea and Japan might lead to reconsidering the data collection strategies.

We then investigated the **deployment of anti-smoking policies** by member countries. Overall, we registered an increase in MPOWER score, in line with global trends. With our clustering analysis based on MPOWER components, we identified clusters of countries based on their level of compliance with the WHO policy targets, identifying the ones who may need more actions to reach better scores, including the United States and Mexico. However, when we tried to assess whether these clusters were differing in terms of prevalence change, we did not obtain any significant result, suggesting that possibly other factors should be taken into consideration.

Finally, we built a semiparametric model to **predict** female smoking prevalence, which suggested that an increase in cigarette final cost may indeed impact negatively on prevalence. This confirms the results reported by WHO, which indicate taxes (and hence their effects on cigarette prices) as one of the most effective measures for smoking control[3]. We used this model to check whether 3 OECD countries with different conditions would **reach SDG3** goal of reducing smoking prevalence by a relative 30% between 2010 and 2025. This approach could be a good starting point to identify countries where more action is necessary and ones which are more likely to reach the target.

Other works will need to characterize better the direct effect of MPOWER measures on smoking prevalence in OECD country states. Nevertheless, the results of our analysis pave the way for future investigations, possibly facilitated by the presence of a higher number of measurements and a further extension to a wider range of countries.

References

1. Global Burden of Disease [database].Washington, DC: Institute of Health Metrics; 2019. IHME. https://www.healthdata.org/gbd

2. OECD (2021), Health at a Glance 2021: OECD Indicators, OECD Publishing, Paris.https://doi.org/10.1787/ae3016b9-en.

3. World Health Organization. (2021). WHO report on the global tobacco epidemic, 2021: addressing new and emerging products. World Health Organization.

4. Flor, L. S., Reitsma, M. B., Gupta, V., Ng, M., & Gakidou, E. (2021). The effects of tobacco control policies on global smoking prevalence. Nature Medicine, 27(2), 239-243.https://doi.org/10.1038/s41591-020-01210-8

5. Ngo, A., Cheng, K. W., Chaloupka, F. J., & Shang, C. (2017). The effect of MPOWER scores on cigarette smoking prevalence and consumption. Preventive medicine, 105, S10-S14. https://doi.org/10.1016/j.ypmed.2017.05.006

6. WHO database on smoking https://apps.who.int/gho/data/node.main.TOBMPOWER?lang=en

7. OECD database https://stats.oecd.org/

8. United Nations Developent Program database for HDI data https://hdr.undp.org/data-center/human-development-index#/indicies/HDI

9. Park, M. B., Kim, C. B., Nam, E. W., & Hong, K. S. (2014). Does South Korea have hidden female smokers: discrepancies in smoking rates between self-reports and urinary cotinine level. BMC women's health, 14(1), 1-8. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4319222/

10. Gabadinho, A., Ritschard, G., Müller, N. S. and Studer, M. (2011) "Analyzing and Visualizing State Sequences in R with TraMineR", Journal of Statistical Software, 40(4), pp. 1–37. doi: 10.18637/jss.v040.i04. https://www.jstatsoft.org/article/view/v040i04