TheData Open

Presented by Citadel and Citadel Securities
In Partnership with CorrelationOne

East Coast Regional Datathon Fall 2021

Team 10

Murtaza Nomani[1], Nange Li[2], Pengyun Li[3], Enming Zhang[4]

[1] Georgia Institute of Technology

[2] Brown University

[3] Columbia University

[4] Cornell University

November 14, 2021

A Cost-Benefit Analysis on the Global Tobacco Policy with Kernel Density Estimation (KDE)

Part I: Non-Technical Executive Summary

Introduction

Tobacco consumption has been one of the biggest public health threats in the world. Recent studies conducted by the World Health Organization (WHO) and the Institute for Health Metrics and Evaluation have shown that around 7.2 million people die prematurely every year from smoking. However, despite the severe harms of tobacco usage, it is impractical to simply ban tobacco production as tobacco generates significant economic benefits such as agricultural employment, tax revenue, foreign exchange earnings, etc. Therefore, policymakers in different countries need to find a balance point between the harms of tobacco consumption on public health and the economic benefits generated from the tobacco industry. Our team aims to find this balance point with more detailed descriptions shown below.

Main Objective

Our main objective is to develop a win-win strategy to find the optimal policy balance between public health and the economic profits of the tobacco industry. To achieve this goal, we derived an empirical equilibrium curve using cost-benefit analysis on the global tobacco industry with kernel density estimations. We hope that our findings can help countries to adjust their tobacco policies.

Key Findings

In this part, we will summarize the key findings of our cost-benefit analysis on the tobacco industry and policy in each country. More details on data manipulations and methodology are included in the technical part of our report. Through the initial stage of exploratory data analysis, we have identified the benefits of the tobacco industry as well as the net ‘costs’ based on visualizing relevant variables in the datasets. Some interesting findings from the data exploration include that for some countries, especially in Africa, the enforcement levels of the “HelpToQuit” policy are inconsistent over the policy years, which may be attributed to factors like the GDP per capita and death rates. Also, as derived from our data, developed countries tend to have higher death rates compared to developing countries. This counterintuitive phenomenon may be explained by the fact that though healthcare in developed countries is usually more advanced the reporting bias is likely to be higher in developing countries; countries with little healthcare systems may face difficulties in determining whether the death is caused by smoking. These interesting findings are potential topics to dig further into, but we are only using these as rationales to define our cost and benefit functions to ensure an in-depth quality analysis of our report.

For the cost function, we have included the smoking deaths rate and the "NewQuit" policy indicator into our cost function. The death rate directly reflects the damages that tobacco brings to societies. As for the "NewQuit" policy indicator, we believe that it is an indirect cost as society realizes the damage of tobacco and provides more tobacco aids. For the benefit function, we have included Tax-to-GDP per capita ratio and the "EnforceBansTobaccoAd" indicator as our variables. The tax-to-GDP ratio indicates the importance of tobacco in generating economic benefits as a source of tax revenue in a country. The larger the ratio, the greater the benefits generated from tobacco production. For the "EnforceBansTobaccoAd" indicator, we think that a higher enforcement level implies that tobacco sales play a less important role in the economy because the government wants to diminish tobacco sales.

After defining our cost and benefit functions respectively, we formulated the variability function to conduct the kernel density estimation. The index of "Variability" is to show how unstable one country is in terms of restrictions of tobacco use and economic development policies, in the years 2007, 2010, 2012, and 2014. KDE is a non-parametric method for joint distribution estimation for cost and benefit, trying to draw a relationship between these two indices, especially in countries with low variability values. Since we have no information about the true distribution of cost and benefit, we applied cross-validation, leaving one data point out each time, to evaluate each model's performance on the left-out data. Also, we used EconValue and HarmValue to represent the benefit function and cost function.

Our findings show that the correlation coefficient of EconValue and HarmValue is -0.26 (p-value < 0.0001), and no countries show high EconValue and HarmValue simultaneously. As seen from our results, Nepal is the country with the highest EconValue and modest HarmValue. Countries with the highest HarmValue and modest EconValue include Bulgaria, Hungary, Serbia, and Montenegro. These four countries are close in the same continental region, which validates the consistency of our definition of indices.

We implemented KDE on countries whose policies and economic conditions are more stable in the past years, as we believed those countries were in an empirical "equilibrium". So the regression curve of Nadaraya-Watson Estimator for the conditional expectation gives suggestions for countries - if the effect of policies isn't satisfactory, how to adjust to an optimal state. And it shows that governments are able to improve the tobacco-related economic benefits, not at the cost of people's health.

In conclusion, our project has provided in-depth details on how to conduct a cost-benefit analysis based on government policies and economic metrics using a non-parametric model (KDE) - we just let the data speak for itself. We hope this report can provide a guideline to help countries to make better policy decisions to balance the harms of tobacco consumption and the economic benefits with the methodology to find the equilibrium state.

Part II: Technical Exposition

1 Data Descriptions and Preprocessing

1.1 Descriptions

We thoroughly explored our datasets to preprocess and merge our datasets into a single csv file that contains all the variables that we need for our cost-benefit analysis and kernel density estimations (KDEs). To begin with, we used several datasets provided to us, which included *death_rates_smoking_age.csv*, *tobacco_production.csv, and stop_smoking.csv*. To indicate how important the tobacco tax revenue is on the country's economic development, we extracted the external dataset of GDP per capita for all the countries in the world from World Bank DataBank. We named the file *GDP_Per_Capita.csv*. Since our main targeting groups are policymakers from different countries, we used the stop_smoking.csv to set the timeline, i.e., 2007, 2010, 2012, and 2014.

The *stop_smoking.csv* file is the core dataset. This dataset contains severity of tobacco regulation policies, which are "AvgCigarettePriceDollars", "AvgTaxesAsPctCigarettePrice", "EnforceBansTobaccoAd", and "HelpToQuit". There were many missing values in "AvgCigarettePriceDollars" and "AvgTaxesAsPctCigarettePrice". The iterative imputation machine learning algorithm, which applies the random forests regression estimator to fill in the missing values, works best in our scenario. We merged the other relevant datasets with the *stop_smoking.csv* file together to provide more information for iterative imputation. Columns we used for the imputation process included 'AvgCigarettePriceDollars', 'AvgTaxesAsPctCigarette Price', 'EnforceBansTobaccoAd', 'HelpToQuit', 'ProductionMT', 'SalePerDay', 'GDP_USD', and 'ProductionUSD'.

1.2 Missing Values

To better use the dataset, we have redefined the feature "HelpToQuit" by rescaling level 1-2 as 0 (since there is no level 1 and 2 in the entire dataset), level 3 as "1", level 4 as "2", and level 5 as "3". We have created the new column as "NewQuit", which facilitates us to calculate the variability of KDEs which we will talk about in more detail in the Methodology section.

The *death_rates_smoking_age.csv* file contained the number of early deaths due to smoking per 100,000 individuals by country-level from 1990 to 2017. We first cleaned the data by removing columns of the age group for “Under_5” and “5_14” since all the values are missing or equal to 0. After cleaning the data, we divided the total numbers of deaths per country by 100,000 to derive a percentage number and merged this csv file with *stop_smoking.csv* so that the years were limited to the four policy years we have mentioned above. The new column name is “Percentage of Death”.

The *tobacco_production.csv* file contains tobacco production in metric tons and monetary value. We first divided the data into two groups, i.e., units in metric tons and units in monetary value, and merged with the *stop_smoking.csv* with the two new columns named “ProductionMT” and “ProductionUSD” respectively.

The *GDP_Per_Capita.csv* file we extracted from the World Bank DataBank contained the GDP per capita in different countries by year. We merged this dataset with *stop_smoking.csv* with the new column name as “GDP_USD”.

Iterative imputation, or multivariable imputation, refers to a process where each feature is modeled as a function of the other features. Each feature is imputed sequentially, one after the other, allowing prior imputed values to be used as part of a model in predicting subsequent features. This process is repeated to form accurate estimations and ensure the quality of our dataset. After cleaning and preprocessing all the relevant datasets, we implemented the iterative imputation using the random forest regressor to fill in all the missing values in *stop_smoking.csv* as expected. Other model-related feature engineering and transformations are explained in the exploratory data analysis section below.

2 Exploratory Data Analysis

Given our objective of helping policymakers to find a balance point, we built a cost-benefit analysis by identifying the benefits of the tobacco industry as well as the net ‘costs’. It is a technique often used in the business world to evaluate capital projects’ viability. To figure out our cost components and benefit components from the datasets, we plot intuitively relevant indicators to see if there are any patterns. We first investigated the two policy indicators “EnforceBansTobaccoAd”, and “HelpToQuit”, as they relate to our main goal the most.

As we can see from Figure 1 below, for both policies, the enforcement levels over the years are generally increasing as the intensity of colors increases and color chunks become more widely spread by year. Another interesting pattern is that for some countries, especially in Africa, the enforcement levels of the “HelpToQuit” policy are inconsistent. The GDP per capita and death rates could potentially cause these inconsistencies. A decrease in GDP per capita may limit the government to provide services and help people quit smoking. Similarly, the difficulties to control the death rate may discourage the government to enforce the policy.

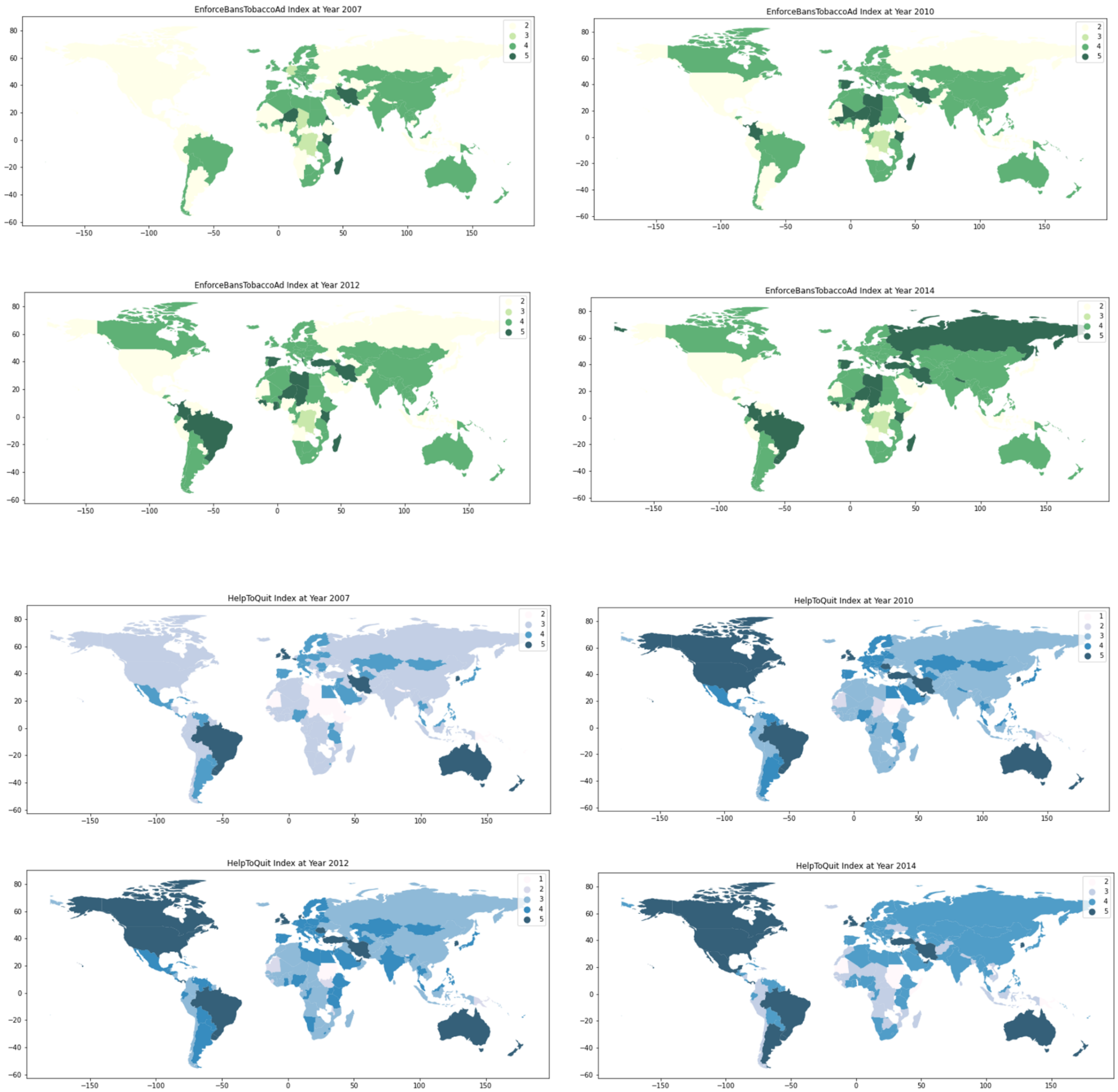
EnforceBansTobaccoAd Index at Year 2007
EnforceBansTobaccoAd Index at Year 2010
EnforceBansTobaccoAd Index at Year 2012
EnforceBansTobaccoAd Index at Year 2014
2
3
4
5
HelpToQuit Index at Year 2007
HelpToQuit Index at Year 2010
HelpToQuit Index at Year 2012
HelpToQuit Index at Year 2014
1
2
3
4
5
80
60
40
20
0
−20
−40
−60
−150
−100
−50
0
50
100
150

Figure 1: Index Visualization for “EnforceBansTobaccoAd” (green) and “HelpToQuit” (blue)

With the above considerations, we explored the death rates and GDP per capita datasets. As seen from Figure 2 shown below, it is surprising to find that developed countries tend to have higher death rates compared to developing countries. This is counter-intuitive in the way that healthcare in developed countries is usually more advanced, which should have led to lower death rates instead. However, the reporting bias is likely to be higher in developing countries; countries with little healthcare systems may face difficulties in determining whether the death is caused by smoking.

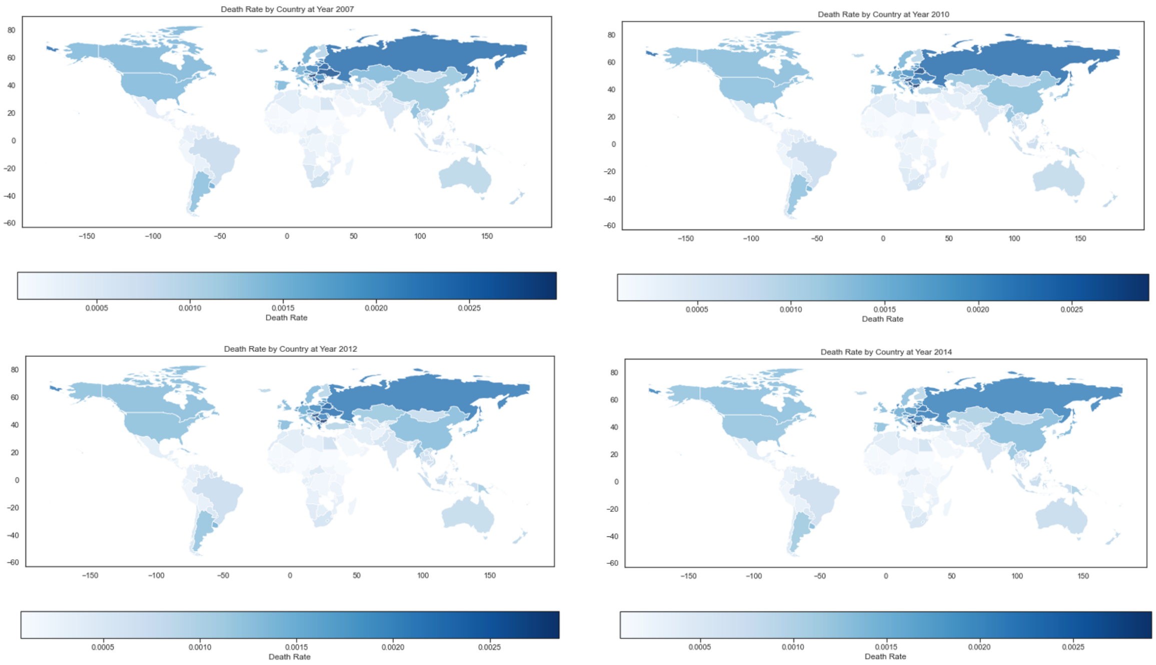
Death Rate by Country at Year 2007
Death Rate by Country at Year 2010
Death Rate by Country at Year 2012
Death Rate by Country at Year 2014
80
60
40
20
0
−20
−40
−60
−150
−100
−50
0
50
100
150
0.0005
0.0010
0.0015
0.0020
0.0025
Death Rate

Figure 2: Index Visualization for percentage of death rates

As the next step in our exploration, we break down all the relevant variables into two categories, i.e., the ones related to harms of tobacco to public health, and the ones that can bring economic benefits to the country. We want to formulate our cost and benefit functions from these two groups of variables. The correlation matrix for the relevant variables is shown in Figure 3 below. The SalesPerDay, AverageCigPrice, and the AvgTaxesAsPctCigarettePrice have high pair-wise correlation coefficients, which means we should only use one variable from these three when building our models.

For the cost function, we have included the smoking deaths rate and the “NewQuit” policy indicator into our cost function. The death rate directly reflects the damages that tobacco brings to societies. As for the “NewQuit” policy indicator, we believe that it is an indirect cost as society realizes the damage of tobacco and provides more tobacco aids. As described in the Data Schema, this original indicator “HelpToQuit” has 5 levels; The highest level indicates the highest availability of tobacco cessation aids provided in the country, which required the greatest amount of government expenditure. As mentioned above under the data preprocessing part, we have rescaled it into a 3-level “NewQuit”. We assume countries that find tobacco more harmful will have higher levels. Besides, the reason to put “NewQuit” into formulating our cost function instead of benefit function is that policymakers can lower damages by providing more tobacco aids. We did not include the production amount or sales value because there is too much bias after iterative imputation.

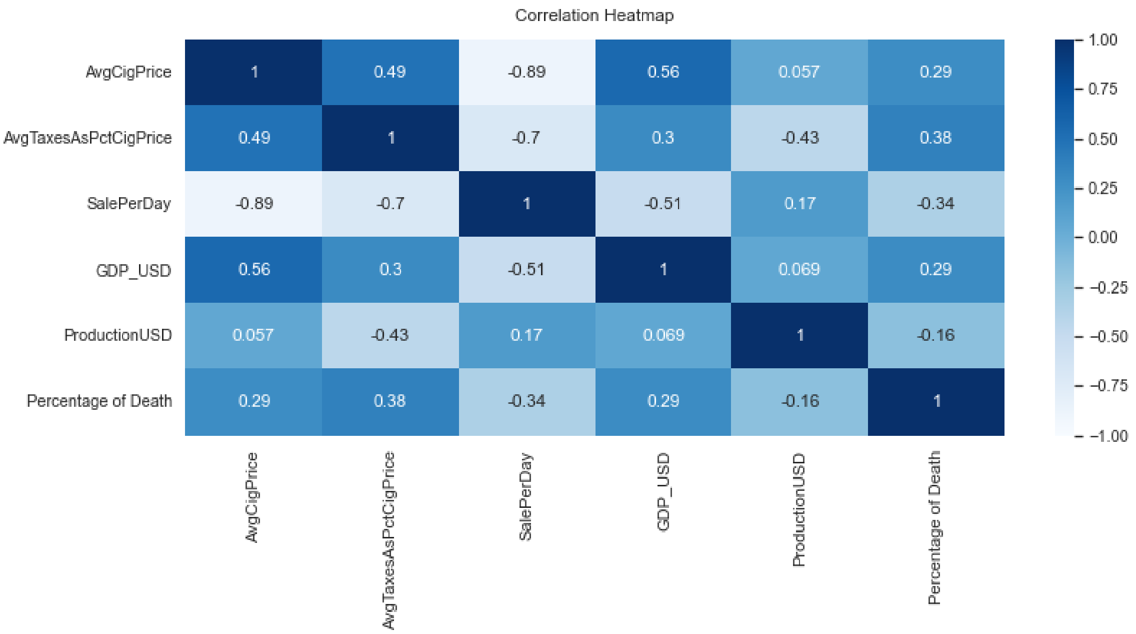

Figure 3: Correlation Heatmap of all relevant variables

For the benefit function, we used the Tax to GDP-per-capita ratio and the "EnforceBansTobaccoAd" indicator as our variables. The tax to GDP-per-capita ratio indicates the importance of tobacco in generating economic benefits as a source of tax revenue in a country. The larger the ratio, the greater the benefits generated from tobacco production. For the "EnforceBansTobaccoAd" indicator, we think that a higher enforcement level implies that tobacco sales play a less important role in the economy because the government wants to diminish tobacco sales.

3 Methodology

3.1 Index Definition

3.1.1 Cost

After exploring the datasets and deciding on what variables to include in our cost and benefit functions, we built models for our cost function and benefit function. For this part, we will talk in detail about our modelling process and kernel density estimation. As what we have mentioned in the previous section, for the cost function, we included the weighted number of deaths due to smoking and the "HelpToQuit" policy indicator into our formula, which we modeled as:

$$HarmValue\ =\ (Scale\ Factor\ \times\ Percentage\ of\ Death)^{Power\ Factor} - \frac{\max\{0, HelpToQuit-2\}}{2}$$

[Scale Factor = 1700 Power Factor = 1.3]

We model in this way as we assume that as the death rate increases, the harm value it generates will increase polynomially. For example, a 80% death rate from smoking will certainly raise much more concern compared to a 20% death rate. We set the scale factor to 1700 and the power factor to 1.3 for scaling purposes to make our kernel density estimations graph smoother. We have assigned a 0.5 weight to NewQuit as it is an indirect cost factor.

3.1.2 Benefit

For the benefit function, we modelled it in this way:

$$EconValue\ = \log(\frac{AvgTaxesAsPctCigarettePrice}{GDP_{per\ capita}} + 1)\ + \frac{1}{EnforceBansTobaccoAd}$$

The first part of the benefit function $\log(\frac{AvgTaxesAsPctCigarettePrice}{GDP_{per\ capita}} + 1)$ is built in this way because the log transformation provides us with an increasing function of diminishing returns. This follows economical intuitions as when the tax-GDP ratio increases, it will bring more economic benefits to the country, but the marginal benefit will decrease as the ratio keeps increasing. We add one in the bracket as some countries do not have tax rate data and log (0) will lead to a negative infinity value. The second part of the benefit function is the reciprocal of EnforceBansTobaccoAd as we want to reveal the inverse relationship between a country's enforcement level and benefit generation. By building this model, we update our dataset with two new columns named "InverseBan" and "EconValue" accordingly.

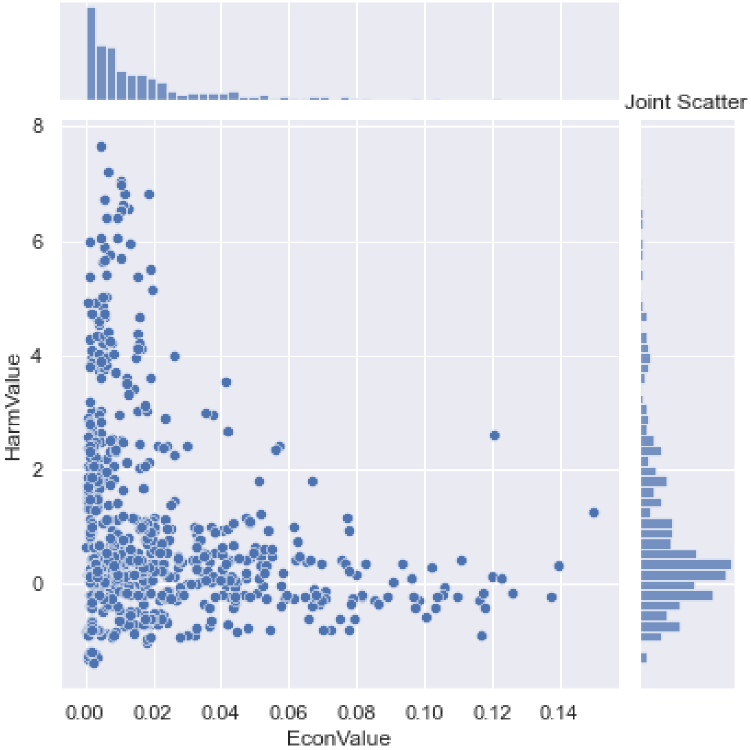

Figure 4: Scatter plot of EconValue and HarmValue

3.1.3 Variability

The index of “Variability” is to measure how unstable one country is in terms of policies on tobacco use and economic metrics, in the years of 2007, 2010, 2012, and 2014. As the time intervals between the four certain years are not uniform, we generated the data points of 2008 using the linear interpolation of points from 2007 and 2010.

For the input features of variability, we employed five columns (missing values have been dealt in preprocessing)- “AvgCigarettePriceDollars”, “AvgTaxesAsPctCigarettePrice”, “EnforceBansTobaccoAd”, “HelpToQuit”, and “GDP_USD”. We standardized these five features by using the StandardScaler. For a series of data, standard deviation is a metric for the uncertainty of this series. Therefore, for each country, the sum of the standard deviation of each feature computed on the four-year series was the variability index for that country.

$$Variability \ = \sum_{k \in features} std(\widehat{col}_k[08],\ col_k[10],\ col_k[12],\ col_k[14])$$

3.2 KDE and Nadaraya-Watson Regression

3.2.1 Kernel Density Estimation

We implemented Kernel Density Estimation (KDE), a non-parametric method for joint distribution estimation for cost and benefit, intending to draw a relationship between these two indices, especially in countries with low variability values. Since we have no information about the true distribution of cost and benefit, we applied cross-validation, leaving one data point out each time, to evaluate each model's performance on the left-out data. As the kernel we employed here is a bi-variate tricube function with a parameter λ, and the value of λ should be obtained with optimization on the cross-validation loss estimator.

The tricube function here is defined within unit value

$$D(u) = \begin{cases} \frac{70}{81}(1-|u|^3)^3 & \text{if } |u| < 1 \\ 0 & \text{if } |u| \geq 1. \end{cases}$$

And add a parameter positive λ,

$$D_\lambda(u) \ = \frac{1}{\lambda} D(\frac{u}{\lambda}) \ .$$

So we are able to adjust the shape of the tricube kernel by tuning the value of λ. The kernel function is then defined on the tricube function:

$$K_\lambda(x, y) \ = \ D_\lambda(x) D_\lambda(y)$$

The mass around one point is spreaded as a symmetric bell-shaped curve in the 3-dimensional space

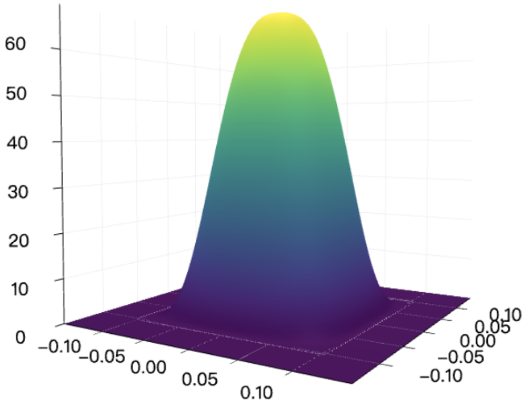

Figure 5: Kernel Function

With all the functions defined above, to estimate the value of joint probability density function at (x, y), KDE uses the sum of all observation points' mass spreaded to this certain point. Thus, the approximated joint distribution (probability mass function) for $f(x, y)$ is

$$\hat{f}_{\lambda}(x, y) = \sum_{i \in observations} K_{\lambda}(x - X_i, y - Y_i) .$$

By minimizing empirical cross validation loss

$$\hat{J}(x, y) = \int (\hat{f}_{\lambda})^2 - \frac{2}{n}\sum_{i} (\hat{f}_{\lambda})^{(-i)}(X_i, Y_i) ,$$

Where $(\hat{f}_{\lambda})^{(-i)}$ denotes the approximation with leaving the i-th point out strategy from the observations set for validation purpose. We obtained the optimal value $\lambda = 0.104$.

The observations in this problem are made up with country points with variability index smaller than the median value of variability, with the "EconValue" and "HarmValue" selected from 2014 only. To show the results of KDE in the 3D space and heatmap (Figure 6),

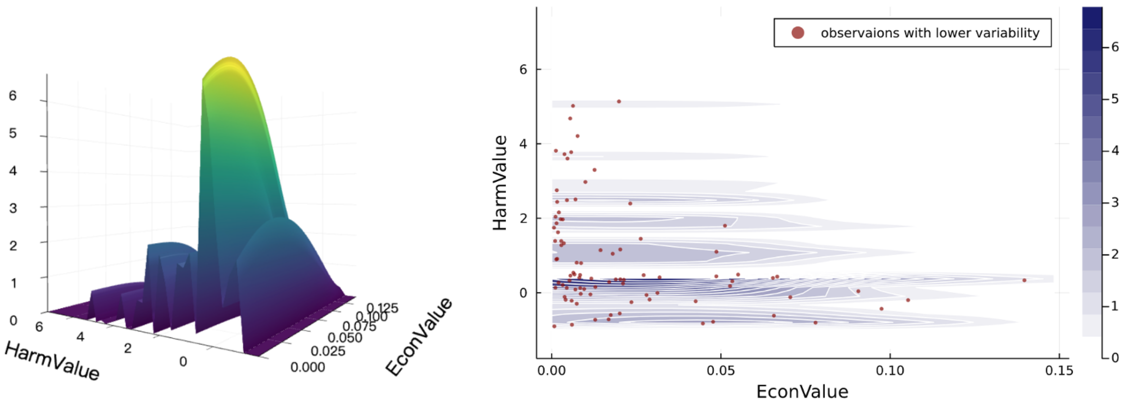

Figure 6: Approximated Joint Distribution by KDE with $\lambda = 0.104$

3.2.2 Nadaraya-Watson Regression

On the KDE results, we could implement Nadaraya-Watson estimator to compute the regression function, which is in essence a conditional expectation of Y given the value of X. The Nadaraya-Watson regression function is defined as:

$$\hat{r}(x) = \frac{\sum_{i} D_{\lambda}(x - X_l)Y_i}{\sum_{i} D_{\lambda}(x - X_l)}$$

A regression curve can be drawn from this estimator, and include all the observation points together in the Figure 7:

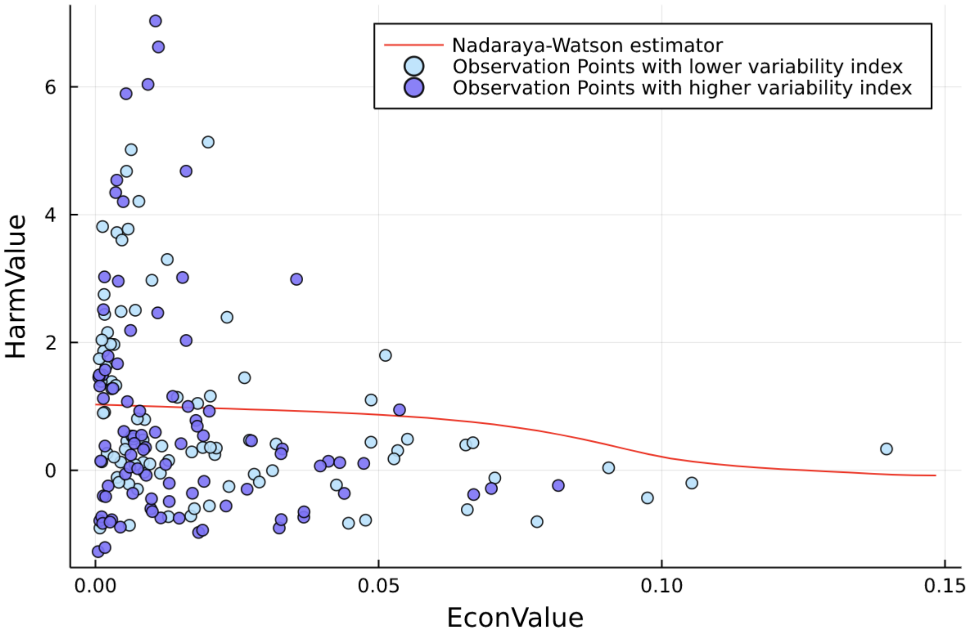

Figure 7: Nadaraya-Watson regression curve

3.3 Results

The correlation coefficient of EconValue and HarmValue is -0.26 (p-value < 10-4), and no countries show high EconValue and HarmValue simultaneously. In Figure 4 and Figure 6, the point to the rightmost corresponds to Nepal, with the highest EconValue and modest HarmValue. Countries with the highest HarmValue and modest EconValue include Bulgaria, Hungary, Serbia, and Montenegro. These four countries are close in the same continental region, which validates the consistency of our definition of indices.

We employed KDE in countries whose policies and economic conditions are more stable (low variability) in the past years, as we believed those countries were in a kind of “equilibrium”. The concept of “equilibrium” is, there’s no initiative for any variable to change given other variables are kept the same. It’s almost impractical to find general optimality for all countries, due to the difference in social conditions and limited access to data. That’s why we tried to draw some patterns regarding the equilibrium between economic benefits and harm to public health. In Figure 7, stable country data points are dyed to light blue and unstable ones are dyed to purple, and a large proportion of points do assemble in the bottom-left part of the graph.

So the regression curve estimating the conditional expectation gives us a handle to make suggestions for countries - if the EconValue or HarmValue isn't satisfactory, how to adjust the policies to a better state, such as tax rate on tobacco, restriction on advertisements, and help center setup. And the Nadaraya Watson estimator here gives us the faith that governments are able to improve the tobacco-related economic benefits, not at the cost of people's health.

Part III: Discussions

1 Strengths

In our analysis, we integrated the stability of a country's policies to help stop smoking to more accurately represent the countries with desirable and stable tobacco prevention policies. When performing statistical analysis, we didn't rely on any parametric family of models when estimating the joint distribution, since we had no prior information on the relationship between EconValue and HarmValue. So we were trying to make the most use of the empirical data, instead of introducing more assumptions into the model.

Another strength of our analysis is our calculation formula for the EconValue and HarmValue. Due to the fact that the harms of tobacco increase as the death rate increases, we used a polynomial rather than a linear model to estimate the cost values. To calculate the EconValue, we used a logarithmic function since the additional economic benefit from a higher tax/GDP ratio decreases as the tax/GDP ratio increases. These considerations make our model accurate to our intuition about how the economic costs and benefits change with changes in death rates and tax/GDP ratios.

2 Weaknesses

A weakness in our economic cost-benefit analysis was not considering the economic harms from serious disease and addiction due to tobacco, and only using the death rate to calculate the harm value. So the data and information are limited, and that might partly be a reason that our observations of countries with high variability and low variability don't separate well from each other.

Another weakness to consider is whether the data is biased. As discussed before, Figure 2 of the global map shows country-wise death rate is higher in developed countries, and this might be due to cases that developing countries have weaker abilities to detect and diagnose compared to developed countries. So we're skeptical that this kind of bias might still exist in the other part of our datasets. But without further information from reliable sources, we can actually do little things with those biases.

Part IV: References

1. The World Bank DataBank-https://databank.worldbank.org/source/world-development-indicators
2. Taxing Tobacco: A win-win for public health outcomes and mobilizing domestic resources-https://www.worldbank.org/en/topic/tobacco/brief/taxing-tobacco-a-win-win-for-public-health-outcomes-mobilizing-domestic-resources
3. Anderson, T. K. (2009). Kernel density estimation and K-means clustering to profile road accident hotspots. *Accident Analysis & Prevention*, *41*(3), 359–364. https://doi.org/10.1016/j.aap.2008.12.014
4. Minoiu, C., & Reddy, S. G. (2012). Kernel Density Estimation Based on Grouped Data: The Case of Poverty Assessment. *SSRN Electronic Journal*. Published. https://doi.org/10.2139/ssrn.1097182