Philip Morris International: PMI IMPACT Project

IT for Illicit Trade Risk Management (IT^{2}RM)

Data Science Delivery Report 1:

Information Gap Analysis and Data Quality Evaluation

Prepared by: Prof. Anton Gerunov, Ph.D., PMP

Delivery Date: 01. July 2019

Contents

List of Figures

Executive Summary

The modelling exercise for the risk management and forecasting models is designed to create a probabilistic model that can aid in the forecasting of illegal cigarette trafficking and aid state and non-state stakeholders. The model has to be valid, reliable, accurate, data-driven, parsimonious, auditable, recognizable. Since the project goal is to forecast and manage the risk of illegal cigarette trafficking, the **preferred approach is supervised learning**.

Across the four countries under study, data availability varies significantly. There is substantially more economic, social, and demographic data on EU member states and candidates in the Eurostat database, and there is relatively less data on non-member states. The World Bank database provides a broad overview on mostly economic matters, as does the OECD database.

The modelling approach together with the data inventory overview reveal a few **key conclusions**:

- Economic, social and demographic data is largely available.
- The data in the inventory is structured and suitable for the modelling exercise.
- Data on illicit cigarette import and consumption is mostly not available.
- Data on criminal activity is irregular and will have to be sourced from a number of alternative databases.
- The comprehensive database for the current project will have to additionally rely on specific national and regional databases.
- The quality of the currently covered data is good but additional data from other sources will have to undergo a quality audit.

The **main recommendations** to fill the gaps are as follows:

1. Expand data inventory to alternative sources.
2. Find and collate data on criminal activities for all countries in the studied sample.
3. Proactively add data on illicit cigarette import and consumption, using internal PMI sources.
4. Experiment with training models on individual and aggregated samples to outline differences.

Background

This PMI Impact project – IT for Illicit Trade Risk Management (IT^{2}RM) aims at utilizing publicly and privately available data, link them in a unified data warehouse and develop sophisticated analytic capabilities on top of it. Leveraging data on crime, socio-economic development, consumer sentiment, legitimate trade, consumer behavior, illicit cigarette and tobacco market and intercepted illegal imports the project will create a unified database that can be used to visualize and analyze key trends in illicit trade and outline the main drivers at a regional level. This will be used to gain insight into the connection between illicit trade in cigarettes and other criminal activities at a detailed level of granularity. Furthermore, a sophisticated forecasting and risk management system is to be built on top of that, dynamically showing increases in the risk of illicit cigarette trade in different regions that can guide both producers and law enforcement authorities.

The current delivery report focuses on the first quarter of the project and presents an overview of the modeling approach, data availability, critical evaluation of data, and presents further recommendation that ensure the successful completion of the project and the fulfilment of its goals.

Model Overview

The modelling exercise for the risk management and forecasting models is designed to create a probabilistic model that can aid in the forecasting of illegal cigarette trafficking and aid state and non-state stakeholders in analysing criminal activities around illicit cigarettes, improving the quality and targeting of field checks, and serve as a communication and analytic tool that raises awareness in both the general public and expert circles.

Model Properties

In order for this model to be useful, it has to have the following characteristics:

- **Valid** – the model needs to meaningfully capture the connections between real-world phenomena and be able to explain the links between different social, economic, and demographic determinants of illicit activity and the resulting amount of cigarettes contraband. From a statistical point of view, these connections can be seen as correlations between relevant variables.
- **Reliable** – the model results must be robust to new data, and alternative model specifications. Reliability essentially means producing similar results (within given confidence intervals) as the modelling exercise is repeated. This measure of robustness provides for a model that can be used continuously and can inspire trust.
- **Accurate** – the model's parameters need to be precisely calculated and its results have to show a high degree of accuracy, outlining its usefulness and practical applicability in a wide range of applications for both analytics and field use.
- **Data-driven** – a sufficient amount of data needs to be analysed in order to reach meaningful conclusions and serve as a sound basis for forecasting and risk management. From a statistical and econometric point of view, a minimum amount of data needs to be leveraged in order to provide for a sound estimation of model parameters.

- **Parsimonious** – this property provides for a relatively large explanatory power with a limited number of data inputs (or explanatory/independent variables). Parsimonious models are easier to interpret, use, and maintain over time.
- **Auditable** – the model should be amenable to human audit in order to understand process drivers. Additionally, auditability ensures that the results are more credible and thus easier to be accepted by a variety of stakeholders with different levels of statistical expertise.
- **Recognizable** – the model aims to combine a set of widely accepted statistical algorithms in a novel way in order to generate innovative insights.

Types of Modelling

There are two large groups of modelling approaches depending on the specifics of data and the task at hand:

- **Supervised learning** – in this case, there is clear target variable (also known as outcome or dependent variable) and statistical algorithms are used to delineate the relationships between this variable and alternative explanatory (independent) ones. Most statistical models also quantify the relationship between dependent and independent variables, and this can be used for forecasting and risk management purposes.
- **Unsupervised learning** – in this case there is no clearly labeled outcome or dependent variable and thus statistical algorithms look for anomalous observations that are markedly different from the rest and are able to point them out. This can be done via methods like clustering, analysis of time series or similar.

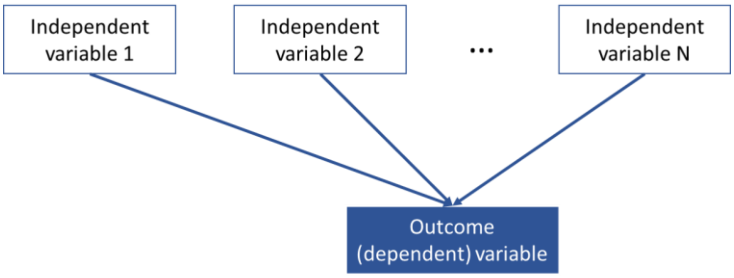


Figure 1: A Schematic Representation of a Supervised Learning Model

Since the project goal is to forecast and manage the risk of illegal cigarette trafficking, the **preferred approach is supervised learning**. To this end the data needs of the project are as follows:

- A set of dependent variables of interest (e.g. illegal cigarette consumption or intercepted cigarette contraband);
- A large set of independent explanatory variables that drive the process of cigarette contraband and can be used to forecast it (e.g. socio-economic conditions, other criminal activity, excise levels, etc.)

The overall logic of a supervised model is shown diagrammatically in Figure 1: A Schematic Representation of a Supervised Learning Model.

Data Evaluation

The fulfilment of the model requirements is crucially dependent upon leveraging a large amount of high quality data and feeding it into the analytic cycle of the project. This section reviews data requirements, the inventory, its quality, and makes a critical evaluation of current data assets. It further outlines the gaps between the available and the delivered data and comes up with concrete recommendations for expanding the scope of the data collection effort.

Data Requirements

The selected supervised learning approach presupposes the collection, curation, and processing of large amounts of relevant data that are to be summarized in a data warehouse, properly visualized, and used for the statistical estimation of the size of the link between different variables. More specifically, the project outline includes, as a minimum, the following data sets:

- Illicit cigarette and tobacco trade (per regions in Bulgaria and national level in the EU)
- Crime statistic per different groups of crime
- Intercepted illegal imports of cigarettes and tobacco products
- Gross domestic product (GDP), GDP per capita, GDP per capita in purchasing power parity, disposable incomes, economic cycle
- Population size
- Educational attainment
- Consumer behavior, including overall levels of cigarette and tobacco consumption
- Consumer sentiment, including sub-indicators such as household's financial conditions and expectations
- Tax environment, including effective tobacco excise levels
- Other relevant indicators

The key project countries are as follows:

1. Bulgaria
2. Turkey
3. Serbia
4. Ukraine

Data collection and processing is outsourced to an external partner – Code Runners, and over the first quarter of the project they have delivered a preliminary data inventory across all countries of interest, together with the data in raw .csv format.

Data Inventory

The delivered data inventory is a detailed analysis of three key databases for relevant data:

- The Eurostat database
- The OECD database
- The World Bank database

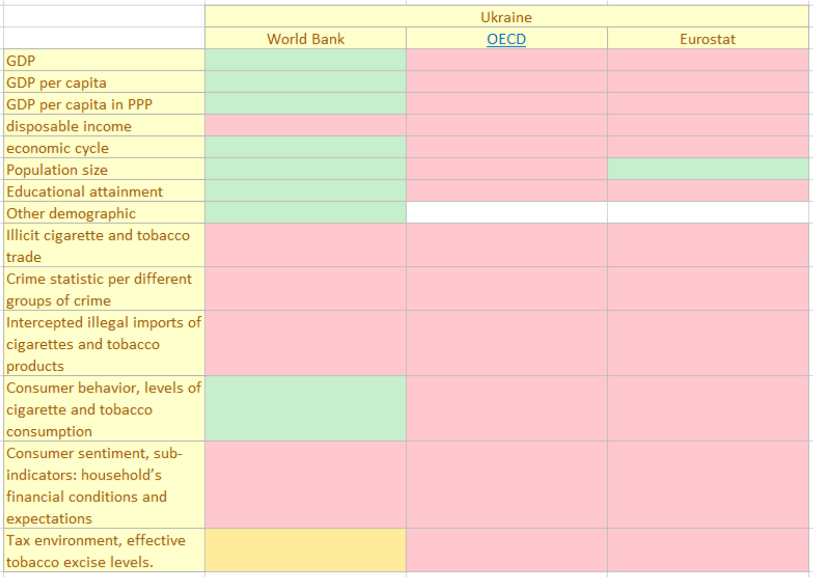
The inventory is complete and of high quality, and aptly identifies the relevant project data.

Data Availability

Across the four countries under study, data availability varies significantly. There is substantially more economic, social, and demographic data on EU member states and candidates in the Eurostat database, and there is relatively less data on non-member states. The World Bank database provides a broad overview on mostly economic matters, as does the OECD database.

Figure 2: Data Availability for Ukraine shows an overview of data about the Ukraine. The main conclusions are as follows:

- There is sufficient economic and demographic data
- Main coverage comes from the World Bank database
- There is relatively limited data on Eurostat and OECD
- No data on illicit cigarette and tobacco trade
- No data on intercepted illegal imports of cigarettes and tobacco products
- No data on crime statistics
- Data on excise taxes is obtainable
- Local Ukrainian databases are not covered at this point

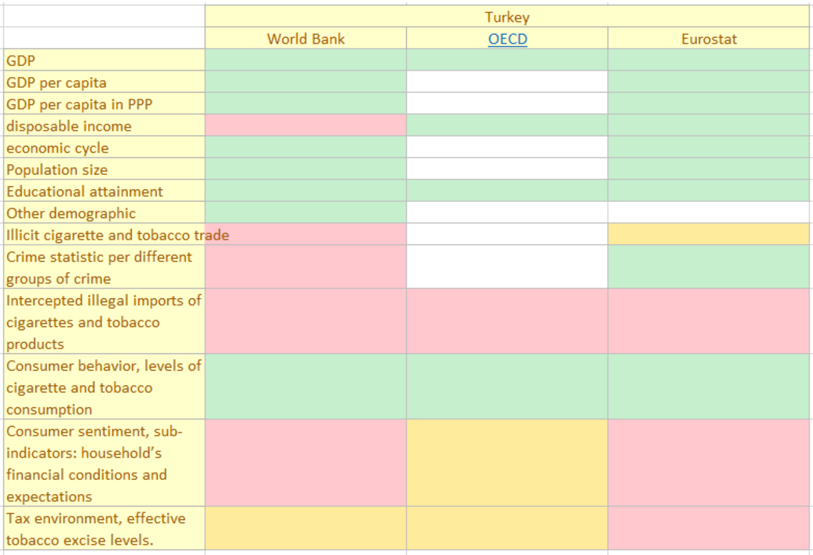


	Ukraine		
	World Bank	OECD	Eurostat
GDP	🟩	🟥	🟥
GDP per capita	🟩	🟥	🟥
GDP per capita in PPP	🟩	🟥	🟥
disposable income	🟥	🟥	🟥
economic cycle	🟩	🟥	🟥
Population size	🟩	🟥	🟩
Educational attainment	🟩	🟥	🟥
Other demographic	🟩		
Illicit cigarette and tobacco trade	🟥	🟥	🟥
Crime statistic per different groups of crime	🟥	🟥	🟥
Intercepted illegal imports of cigarettes and tobacco products	🟥	🟥	🟥
Consumer behavior, levels of cigarette and tobacco consumption	🟩	🟥	🟥
Consumer sentiment, sub-indicators: household's financial conditions and expectations	🟥	🟥	🟥
Tax environment, effective tobacco excise levels.	🟨	🟥	🟥

Figure 2: Data Availability for Ukraine

Figure 3: Data Availability for Turkey shows an overview of data about Turkey. The main conclusions are as follows:

- There is sufficient economic and demographic data
- Data on the country is abundant across all three databases
- No data on illicit cigarette and tobacco trade
- No data on intercepted illegal imports of cigarettes and tobacco products
- No data on crime statistics
- Data on excise taxes is obtainable
- Local Turkish databases are not covered at this point

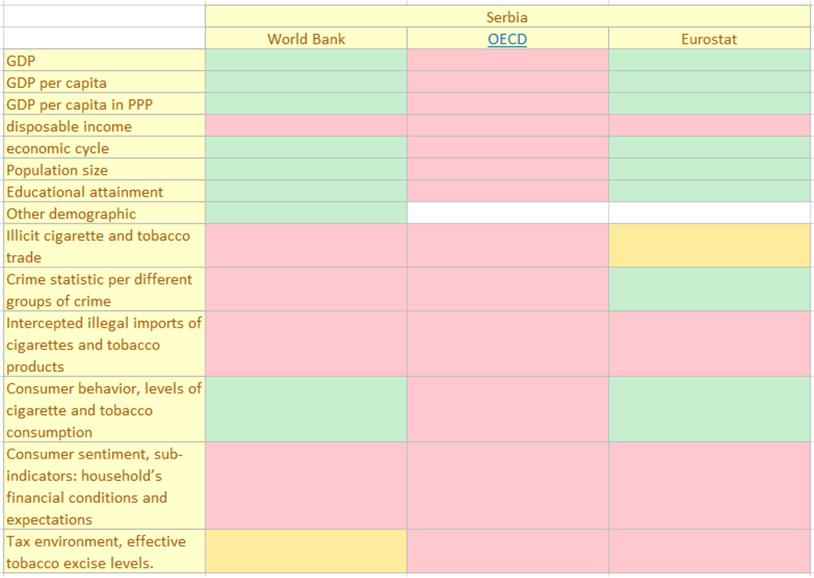


	Turkey		
	World Bank	OECD	Eurostat
GDP	🟩	🟩	🟩
GDP per capita	🟩		🟩
GDP per capita in PPP	🟩		🟩
disposable income	🟥	🟩	🟩
economic cycle	🟩		🟩
Population size	🟩		🟩
Educational attainment	🟩	🟩	🟩
Other demographic	🟩		
Illicit cigarette and tobacco trade	🟥		🟨
Crime statistic per different groups of crime	🟥		🟩
Intercepted illegal imports of cigarettes and tobacco products	🟥	🟥	🟥
Consumer behavior, levels of cigarette and tobacco consumption	🟩	🟩	🟩
Consumer sentiment, sub-indicators: household's financial conditions and expectations	🟥	🟨	🟥
Tax environment, effective tobacco excise levels.	🟨	🟨	🟥

Figure 3: Data Availability for Turkey

Figure 4: Data Availability for Serbia shows an overview of data about Serbia. The main conclusions are as follows:

- There is sufficient economic and demographic data
- Data on the country is abundant across the World Bank and the Eurostat database
- No data on intercepted illegal imports of cigarettes and tobacco products
- Data on crime statistics is present and potentially useful
- Data on excise taxes is obtainable
- Local Serbian databases are not covered at this point



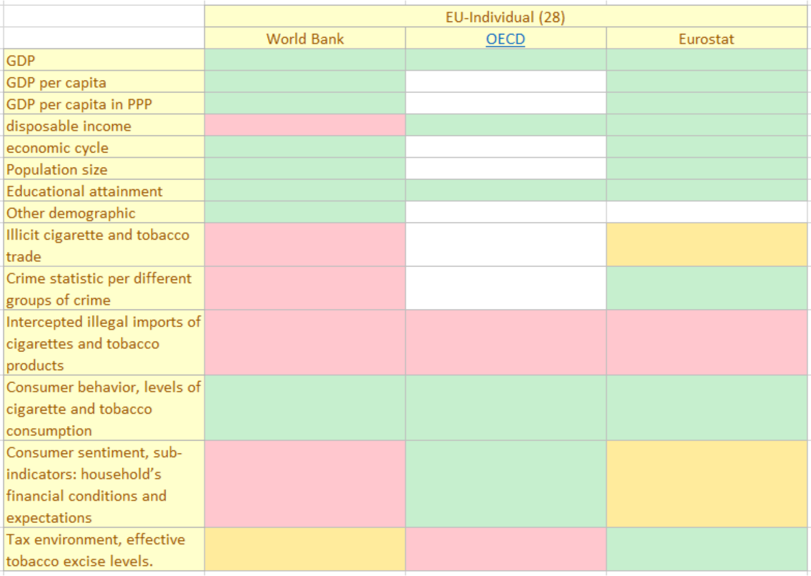
	Serbia		
	World Bank	OECD	Eurostat
GDP	🟩	🟥	🟩
GDP per capita	🟩	🟥	🟩
GDP per capita in PPP	🟩	🟥	🟩
disposable income	🟥	🟥	🟥
economic cycle	🟩	🟥	🟩
Population size	🟩	🟥	🟩
Educational attainment	🟩	🟥	🟩
Other demographic	🟩		
Illicit cigarette and tobacco trade	🟥	🟥	🟨
Crime statistic per different groups of crime	🟥	🟥	🟩
Intercepted illegal imports of cigarettes and tobacco products	🟥	🟥	🟥
Consumer behavior, levels of cigarette and tobacco consumption	🟩	🟥	🟩
Consumer sentiment, sub-indicators: household's financial conditions and expectations	🟥	🟥	🟥
Tax environment, effective tobacco excise levels.	🟨	🟥	🟥

Figure 4: Data Availability for Serbia

Figure 5: Data Availability for Individual EU Member States shows an overview of data about Bulgaria and other individual EU member states. The main conclusions are as follows:

- There is sufficient economic and demographic data

- Data on the country is abundant across all three databases
- No data on intercepted illegal imports of cigarettes and tobacco products
- Data on crime statistics is present and potentially useful
- Data on excise taxes is obtainable
- National databases are not covered at this point
- The EU Open data portal is not covered at this point



	EU-Individual (28)		
	World Bank	OECD	Eurostat
GDP	🟩	🟩	🟩
GDP per capita	🟩		🟩
GDP per capita in PPP	🟩		🟩
disposable income	🟥	🟩	🟩
economic cycle	🟩		🟩
Population size	🟩		🟩
Educational attainment	🟩	🟩	🟩
Other demographic	🟩		
Illicit cigarette and tobacco trade	🟥		🟨
Crime statistic per different groups of crime	🟥		🟩
Intercepted illegal imports of cigarettes and tobacco products	🟥	🟥	🟥
Consumer behavior, levels of cigarette and tobacco consumption	🟩	🟩	🟩
Consumer sentiment, sub-indicators: household's financial conditions and expectations	🟥	🟩	🟨
Tax environment, effective tobacco excise levels.	🟨	🟥	🟩

Figure 5: Data Availability for Individual EU Member States

Data Quality

The data overview includes data sets from recognized and standardized databases. This ensures that they are collected with a sufficient level of quality for the current modeling exercise. Moreover, data are mostly official government-provided data, thus giving more credibility to both the visualization and the modeling part of the project. The use of data from a single (or a few) sources ensures better comparability across the time series, which leads to a higher level of validity and quality of the analysis.

There may be a few **outstanding issues with quality**:

- Economic and demographic time series for Ukraine may have to be obtained from the World Bank, whereas the same data for other countries will be taken from Eurostat. The comparability of the data, and the consistency of the methodologies used have to be checked.
- Crime statistics are prone to misreporting and each country may have different measurement issues. This is due to both the collection methodology and the cultural embeddedness of measurement. This data has to be interpreted with great care.
- Consumer sentiment data is a soft measure that is largely driven by subjective, as well as objective factors, and it is collected through a survey methodology. As such this data is comparable across time for a given geography but harder to compare across different geographies.
- Collection of crime, consumption of illicit cigarettes and intercepted imports will have to be collected from different databases that are not among the currently surveyed. The collection methodologies will have to aligned and data may have to be adjusted to achieve comparability

and ensure content validity of the data. At the same time, those statistics are crucial outcome variables and extensive focus on them is warranted.

Gap Analysis and Recommendations

The modelling approach together with the data inventory overview reveal a few **key conclusions**:

- Economic, social and demographic data is largely available.
- The data in the inventory is structured and suitable for the modelling exercise.
- Data on illicit cigarette import and consumption is mostly not available.
- Data on criminal activity is irregular and will have to be sourced from a number of alternative databases, not currently included in the inventory exercise. This data is important as the illegal cigarette import (and hence consumption) is probably highly correlated with other criminal activities.
- Major databases cover a large proportion of the data needs but the complete and comprehensive database for the current project will have to additionally rely on specific national and regional databases.
- The quality of the currently covered data is good, or at least sufficient, but additional data added from other databases will have to be audited carefully.

The **main recommendations** to fill the gaps are as follows:

1. Expand data inventory to alternative sources such as national or EU open data portals and studies to identify additional relevant sources and types of data.
2. Find and collate data on criminal activities for all countries in the studied sample. Possible sources include national statistical portals, statistics published by the Ministries of Interior, or databases collected by national and international bodies (e.g. NGOs).
3. Proactively add data on illicit cigarette import and consumption, using internal sources. A major possible source is Philip Morris International internal survey and counting data. Access to the structured databases with results over time is needed so as to cover as large and granular a sample as possible over the maximum period of collection.
4. Experiment with training models on individual and aggregated samples to outline differences. Due to data availability it may be necessary to train the two models on an aggregated data set instead of training individual per country models. A panel regression model with fixed effects design is to be used to check for presence of differences between countries.