Philip Morris International: PMI IMPACT Project

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

Data Science Delivery Report 2:

Information Requirements and Initial Trends Evaluation

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

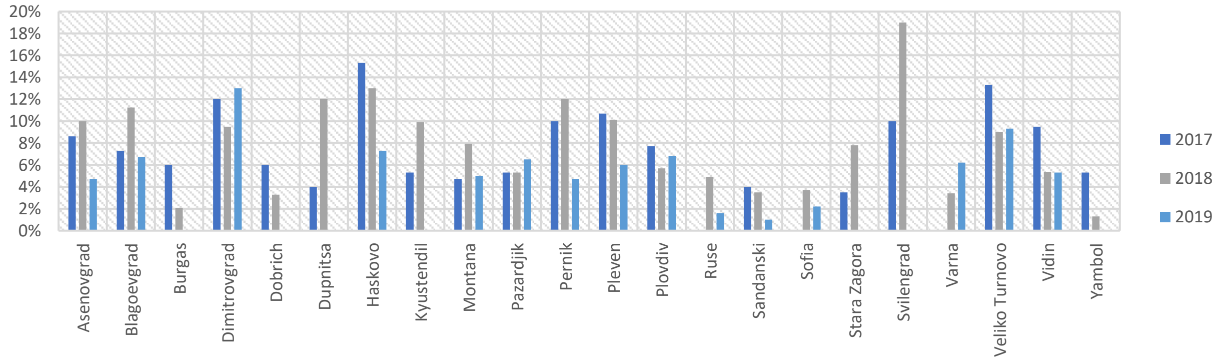
Delivery Date: 31. October 2019

Contents

List of Figures

Executive Summary

The current delivery report focuses on the second quarter of the project and presents an overview of data collection efforts, corrective measures taken as a result of previous report and an overview of key trends in the illicit cigarettes trade and incidence. These data for illicit cigarette incidence are mostly available for Bulgaria, but not available for other target countries and we cannot assume that other countries follow the same contraband dynamics as Bulgaria. Thus, we can neither impose the structure of data variability, nor the ultimate results on other target countries.



Incidence of illicit cigarettes, % of total consumption

Investigating data correlations, we observe the changing sign of the correlatoin coeffiecient, meaning that the increase of illicit cigarette occurrence in a given region is connected to an increase in some, but a decrease in certain others. This points to the importance of the regional dimension in cigarette contraband.

The review of the updated dataset and the illicit cigarette trends description lead to a few **key conclusions**:

- The database is now expanded with an indicator for incidence of illicit trade that can be used as a dependent variable in the modeling exercises.
- Illicit cigarette incidence in Bulgaria is relatively stable over the initial part of the period under investigation (2010-2013), but more dynamics over the later part of the period under investigation (2017-2019) when a clear downward trend is observed.
- The correlational structure of the data shows that the country as a whole may not exhibit a single trend, but regions may have different and opposite dynamics of incidence themselves.

The **main recommendations** for the next quarter are as follows:

1. Expand the database to include longer time series of the relevant dependent variables.
2. Find ways to compute or impute missing incidence data for target countries in order to meet information requirements of the forecasting and risk management models.
3. Make a more detailed overview of the overall correlational structure of all the data across all geographies.
4. Construct the two models initially on the country with most data – Bulgaria – and then expand to other focus countries.

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 second quarter of the project and presents an overview of data collection efforts, corrective measures taken as a result of previous report and an overview of key trends in the illicit cigarettes trade and incidence.

Information Needs

The current phase aims to complete the database for illicit cigarette trade by adding relevant data that can serve as dependent variables in both the forecasting and the risk management model. These specific pieces of data were not available at the curation of the first version of the database, but they are crucial for the successful modelling exercise (see Figure 1) and thus for the successful completion of the project at hand.

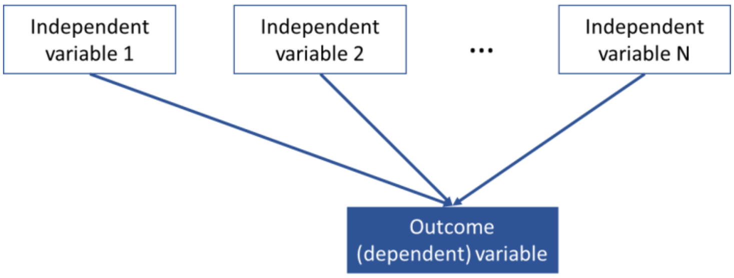


Figure 1: A Schematic Representation of a Supervised Learning Model

In the selected modelling approach 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. The second quarter of the project focuses on the collection of those and an overview of key trends in illicit cigarette consumption that will aid the modelling exercise.

Evaluation of Scope Fulfilment

The key indicators that were defined in the project proposal for the target counties (Bulgaria, Turkey, Serbia, Ukraine) are as follows:

- 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 current version of the database addresses Recommendations 1, 2, and 3 from the previous Delivery Report and thus includes the following:

- Non-domestic cigarette incidence for Bulgaria spanning the years 2010, 2012, and 2013 at a regional level on a bi-annual basis, sourced from PMI public press releases
- Non-domestic cigarette incidence for Bulgaria spanning the years 2017, 2018, 2019 at a regional level on a quarterly basis, sourced from PMI public press releases

Those indicators are indispensable for model training and come at sufficient level of granularity. The time series is of somewhat short length and may have to be supplemented with additional data. However, initial modelling can begin based on the sample provided.

These data is not available for other target countries and we cannot assume that other countries follow the same contraband dynamics as Bulgaria. Thus, we can neither impose the structure of data variability, nor the ultimate results on other target countries. Instead, we can reconstruct the time series for other countries by uncovering structural relationships between relevant indicators and so create an unbiased estimate of their likely non-domestic cigarette incidence and illicit cigarette trade.

Evaluation of Data Quality

Data collection and processing is outsourced to an external partner – Code Runners, and over the second quarter of the project they have delivered an updated data inventory across all countries of interest, together with the data in raw .csv format. The inventory is nearing completion, of high quality, and adequately identifies the relevant project data.

The main outstanding issues, as also reported in Delivery Report 1 remains the quality of data on crime, consumption of illicit cigarettes and intercepted imports that is collated from different sources, each with its own particularities. 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. At this point we recommend statistical adjustment to the data to improve their levels of comparability. This holds particularly true for cases where the periods under study diverge – i.e. one of the time series is on annual/bi-annual basis, whereas others are on a quarterly or even monthly. A neutral approach here would be to assume linear dynamics, thus smoothing the trends whenever missing data has to be imputed. While some unexpected shock and extreme observations may be underplayed this way, the expected value of the errors generated should approach zero, thus having relatively limited impact on the quality of the model produced.

Non-domestic Cigarette Incidence

The next step for modeling is to describe and understand the trends in illicit cigarette incidence across both the temporal as well as the spatial dimension. To achieve the former, we use the time series for contraband trade incidence for both the early period (2010 through 2013) and the late period (2017 through 2019). To achieve the latter, we conduct a deeper dive into the regional peculiarities of this phenomenon. As Bulgaria is the country with the largest scope of the data in this version of the database, we focus on it.

Trend Evaluation: Early Period 2010-2013

Over the early part of the investigated period, there is relatively little dynamics of total consumption of illicit cigarettes. From 2010 to 2013 it is prevalent in large cities and no clear trend of movement in the direction of decrease can be observed. The largest nominal consumption of contraband cigarettes is found in Sofia, followed at a significant distance by other large cities – Plovdiv, Varna, and Burgas. This dynamics shows the relatively more circumscribed efforts at fighting contraband that understandably produce only limited results.

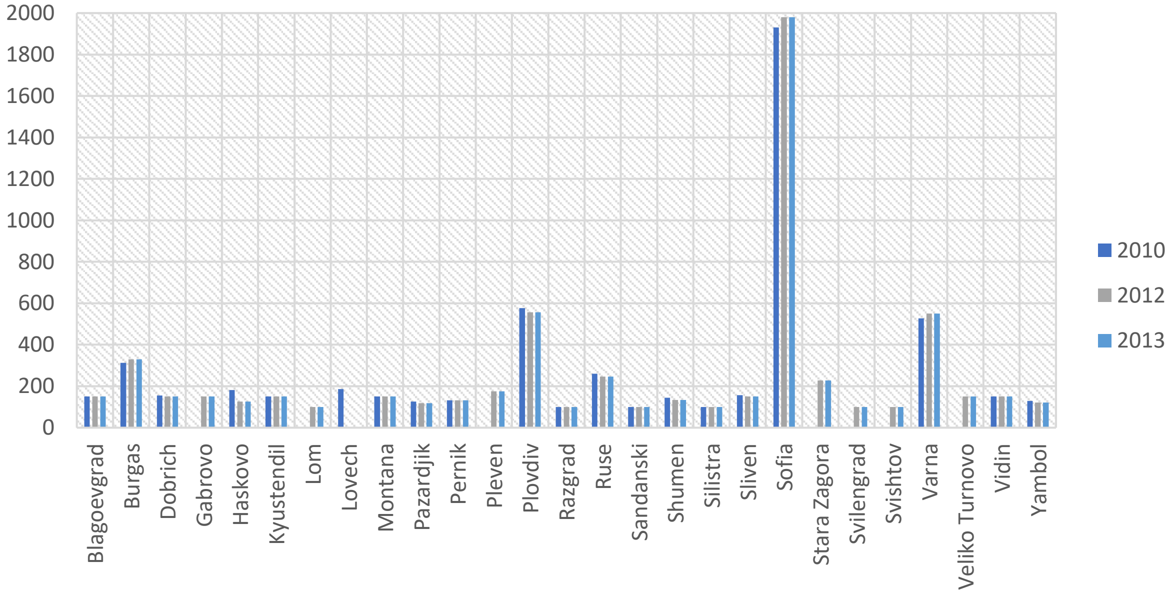


Figure 2: Incidence of illicit cigarettes, number of boxes

From the data perspective we need to note that the limited changes in the data, i.e. the small variance observed, may pose challenges as to the modeling. It is imperative that the early period data be combined with data exhibiting larger variability so that key pertinent trends may be captured by the statistical models used. The early part of the period is also indicative of what

Trend Evaluation: Late Period 2017-2019

The later part of the period under study shows significantly more variability. There are notable changes year on year in the average incidence. Most regions are on an downward average trend, with this being particularly pronounced in Asenovgrad, Blagoevgrad, Haskovo, Pernik, Pleven, Sandanski, Sofia, Vidin, and Yambol. This change in data clearly follows a policy shift in the real world and reflects a more conscious effort at combatting cigarette contraband.

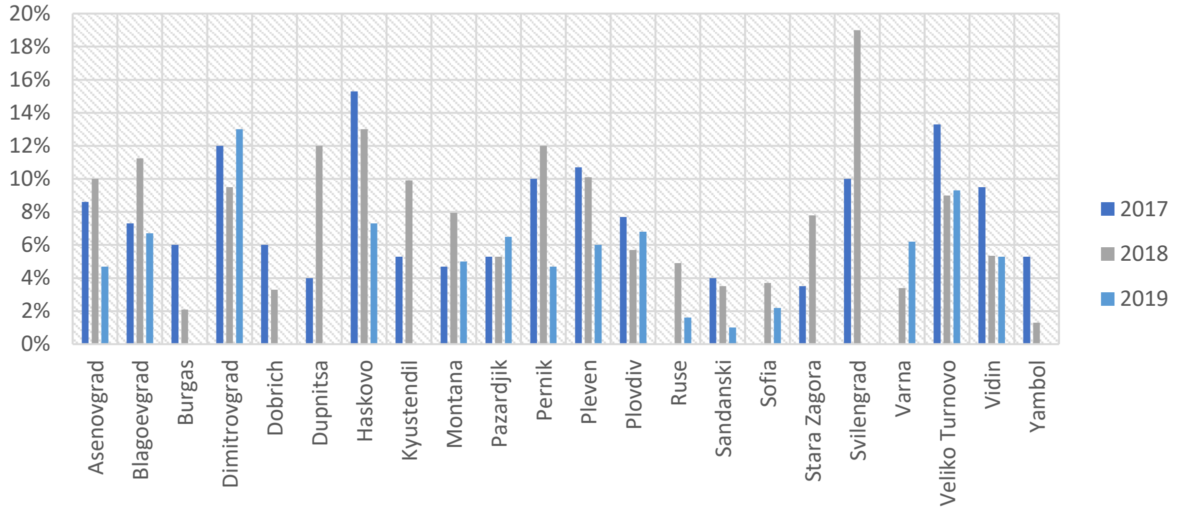


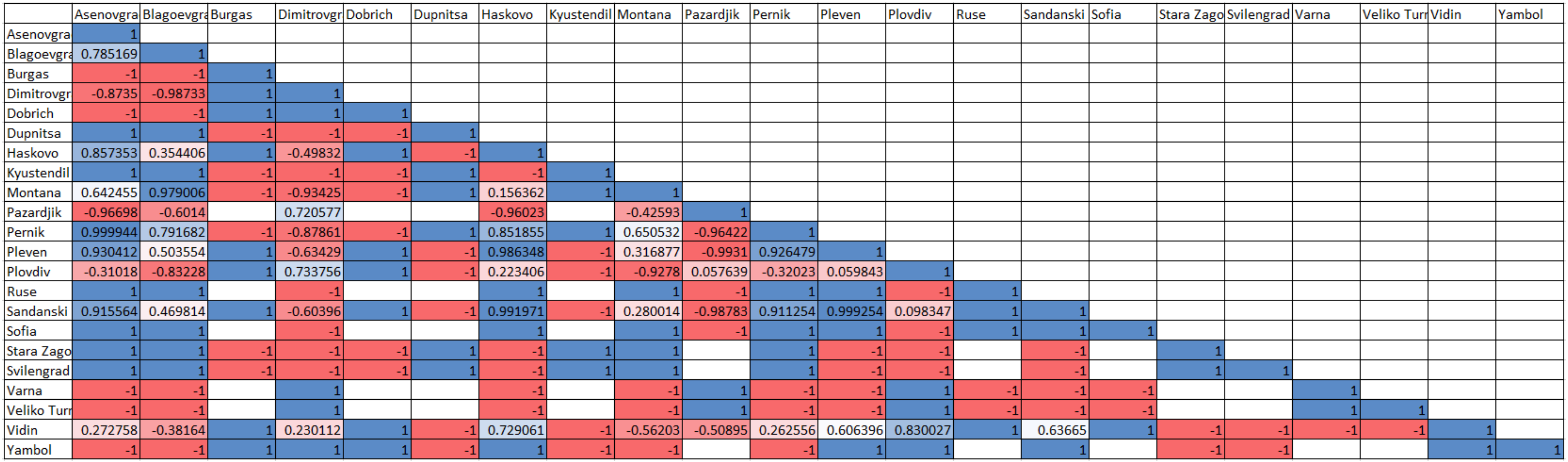
Figure 3: Incidence of illicit cigarettes, % of total consumption

From a data perspective these data show significant variance and are thus suitable for the modeling exercise. The span of the time series is relatively short and thus needs to be expanded but the regional coverage is satisfactory and can serve as a useful basis for both the risk management and the contranband forecasting model.

Here we note that the current database only contains such data for Bulgaria and other countries do not features data on this indicator. This may be problematic as illicit cigarette incidence will have to serve as a regressand in a possible forecasting equation. To resolve this problem we propose to investigate the correlatoinal structure of the dataset for other focus countries and to find a proxy indicator for incidence – i.e. an indicator that features a very high correlation with it and can be used interchangably in the statistical sense of the word. This will give an additional indication of whether the constructed times series for missing data are reliable and whether those imputations may be fruitfully used.

Regional Level Correlation Matrix

The trends between non-domestic cigarette incidence at the regional levels are investigated by means of a correlational analysis. We only use recent data for that spanning over the past three years on a quarterly basis, investigating occurrence (in percentage points) in the different regions of Bulgaria. Results are presented in the correlogram (Figure 4). The key trends that we observe is the changing sign of the correlatoin coeffiecient, meaning that the increase of illicit cigarette occurrence in a given region is connected to an increase in some, but a decrease in certain others. For example, an increase in Pernik is connected to an increase in Asenovrad, Blagoevgrad, Haskovo and Montana, but to a decrease in Dimitrovgrad and Pazardjik. We observe the same trend across the board, including Sofia. The major conclusion stemming from these results is that illicit cigarette trade dynamics do not move synchronously at the country level but have specific regional dynamics.



	Asenovgra	Blagoevgra	Burgas	Dimitrovgr	Dobrich	Dupnitsa	Haskovo	Kyustendil	Montana	Pazardjik	Pernik	Pleven	Plovdiv	Ruse	Sandanski	Sofia	Stara Zago	Svilengrad	Varna	Veliko Turr	Vidin	Yambol
Asenovgra	1																					
Blagoevgra	0.785169	1																				
Burgas	-1	-1	1																			
Dimitrovgr	-0.8735	-0.98733	1	1																		
Dobrich	-1	-1	1	1	1																	
Dupnitsa	1	1	-1	-1	-1	1																
Haskovo	0.857353	0.354406	1	-0.49832	1	-1	1															
Kyustendil	1	1	-1	-1	-1	1	-1	1														
Montana	0.642455	0.979006	-1	-0.93425	-1	1	0.156362	1	1													
Pazardjik	-0.96698	-0.6014		0.720577			-0.96023		-0.42593	1												
Pernik	0.999944	0.791682	-1	-0.87861	-1	1	0.851855	1	0.650532	-0.96422	1											
Pleven	0.930412	0.503554	1	-0.63429	1	-1	0.986348	-1	0.316877	-0.9931	0.926479	1										
Plovdiv	-0.31018	-0.83228	1	0.733756	1	-1	0.223406	-1	-0.9278	0.057639	-0.32023	0.059843	1									
Ruse	1	1		-1			1		1	-1	1	1	-1	1								
Sandanski	0.915564	0.469814	1	-0.60396	1	-1	0.991971	-1	0.280014	-0.98783	0.911254	0.999254	0.098347	1	1							
Sofia	1	1		-1			1		1	-1	1	1	-1	1	1	1						
Stara Zago	1	1	-1	-1	-1	1	-1	1	1		1	-1	-1		-1		1					
Svilengrad	1	1	-1	-1	-1	1	-1	1	1		1	-1	-1		-1		1	1				
Varna	-1	-1		1			-1		-1	1	-1	-1	1	-1	-1	-1			1			
Veliko Turr	-1	-1		1			-1		-1	1	-1	-1	1	-1	-1	-1			1	1		
Vidin	0.272758	-0.38164	1	0.230112	1	-1	0.729061	-1	-0.56203	-0.50895	0.262556	0.606396	0.830027	1	0.63665	1	-1	-1	-1	-1	1	
Yambol	-1	-1	1	1	1	-1	1	-1	-1		-1	1	1		1		-1	-1			1	1

Figure 4: Correlation Matrix of Illicit Cigarette Incidence across Regions in Bulgaria

Gap Analysis and Recommendations

The review of the updated dataset and the illicit cigarette trends description lead to a few **key conclusions**:

- The database is now expanded with an indicator for incidence of illicit trade that can be used as a dependent variable in the modeling exercises.
- The database is augmented with crime statistics from the target countries that can be used as independent variables in the modeling exercises. Given that criminal activities tend to cluster, it is likely that those variables will have relatively high explanatory power.
- Illicit cigarette incidence in Bulgaria is relatively stable over the initial part of the period under investigation (2010-2013) and this can be used as a natural experiment to see the variability of what factors does not produce change.
- Illicit trade incidence in Bulgaria is more dynamics over the later part of the period under investigation (2017-2019) and a clear downward trend is observed. This variance is useful for investigating factor influence and model coefficients estimation.
- The correlational structure of the data shows that the country as a whole may not exhibit a single trend, but regions may have different and opposite dynamics of incidence themselves.

The **main recommendations** for the next quarter are as follows:

1. Expand the database to include longer time series of the relevant dependent variables. This holds particularly true for the illicit cigarette incidence. Proprietary data from PMI will be useful to have at a minimum a full coverage of the period for Bulgaria, and preferably – for all target countries.
2. Find ways to compute or impute missing incidence data for target countries in order to meet information requirements of the forecasting and risk management models. Two possible approaches to achieve this are:
    a. Reconstruct the time series by leveraging their structural relationships with other relevant variables (income, unemployment, criminal activity, demographics, etc.). The coefficients that quantify these structural relationships can be estimated in a general linear model framework (e.g. multiple regression).
    b. Find a proxy variable that can be used instead of cigarette incidence in the modeling exercise for those countries that have no data on cigarette incidence. This proxy may be overall contraband or another variable that has sufficiently high correlation with incidence. The threshold is preferable a value of the Pearson correlation coefficient of above 0.80.
3. Make a more detailed overview of the overall correlational structure of all the data across all geographies. This is imperative as it will show whether the target countries have similar or different structural relationships and will thus inform the modeling part.
4. Construct the two models initially on the country with most data – Bulgaria – and then expand to other focus countries. This ensures that relevant insights from the information rich case can be applied to information-poorer ones.