Philip Morris International: PMI IMPACT Project

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

Data Science Delivery Report 6:

Risk Scoring Algorithm: Approaches and Utility

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

Delivery Date: 31. March 2021

Contents

List of Figures

Executive Summary

The current delivery report builds upon the insights gleaned from the previous modeling exercises and aims to use the granular-level forecasting model to create an intuitive and easy-to-use risk scoring and visualization that can guide stakeholders in understanding the dynamics of illicit cigarette contraband as well as to take action.

We propose the following **risk scoring dimensions**:

- Forecasted illicit cigarette incidence
- Current incidence of illicit cigarettes

Those two dimensions can be used to construct a Risk Prioritization Matrix. For simplicity, each of those dimensions can be ranked on a given scale, and the resulting overall expected risk is simply the product of the two ranks.

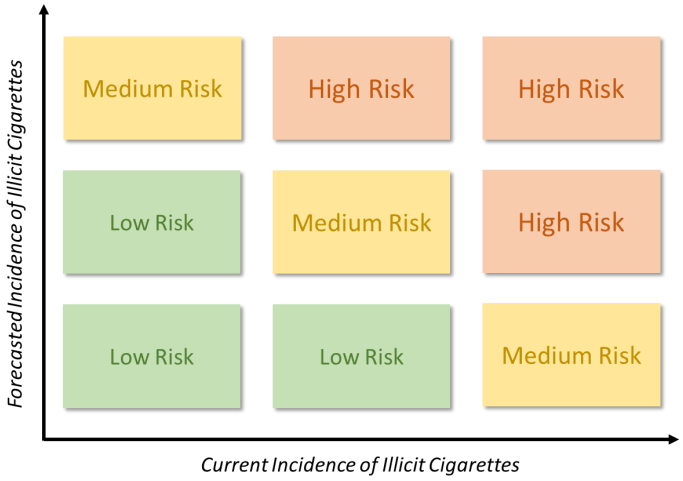

Figure: Topline Graph of Illicit Cigarette Incidence Risk Prioritization Matrix

While this definition is rooted in risk science, its adapted version has a very intuitive interpretation in the context of cigarette contraband. It broadly states that a region which has had a significant about of illicit incidence and is likely to continue to exhibit high contraband levels, is very high risk. Conversely, a region with no or limited incidence that is not expected to develop it is market as low risk. Finally, intermediate regions in terms of current and forecasted incidence occupy a middle ground in terms of their risk score.

The **Turkish** Risk Map at the regional level shows a clear-cut display of what regions pose the highest level of illicit incidence risk. Those are:

- Eastern Anatolia
- South East Anatolia

Both of these regions have a much higher contraband incidence that the rest of the country, and this is a trend which is highly likely to carry over into the future.

The two regions in **Serbia** with highest values on both dimensions, and thus the highest-risk ones are the following:

- Subotica
- Novi Pazar

The riskiest regions in **Bulgaria** in terms of cigarette contraband are the following:

- Haskovo
- Pernik
- Plovdiv
- Kjustendil

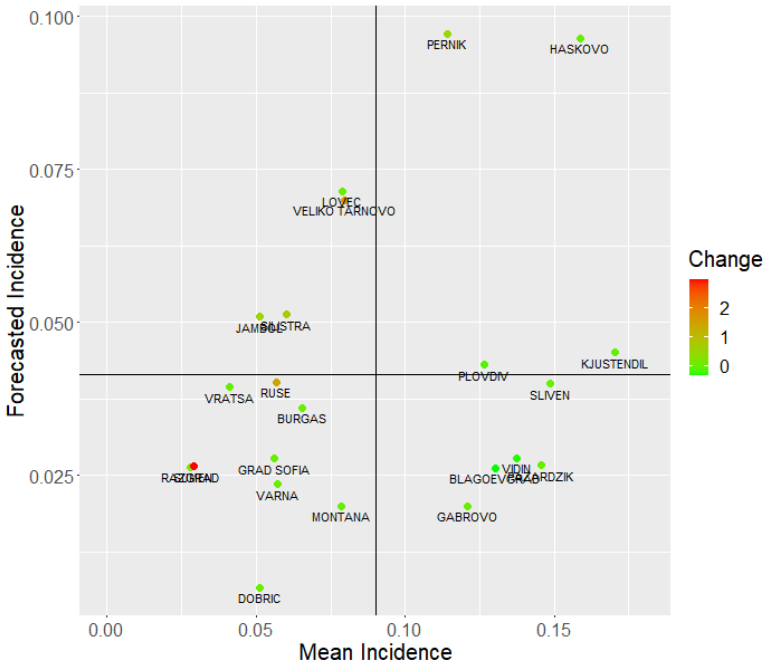

Figure: Visualization of Risk Map with Automated Identification of High-Risk Regions

The proposed risk scoring, prioritization and visualization approach stems from key results in the science of risk but is also adapted to the concrete realities of illicit cigarette incidence. Its major benefit is that despite it is data-driven and informed by sophisticated modeling, it is still intuitive and easy to understand and derive actionable insight from. The traffic light system of color-coding risk prioritization should be familiar to a wide range of different stakeholders and thus the whole approach will likely be easier to adopt. The main beneficiaries of this risk management exercise include local authorities, national and local law enforcement agencies, as well as NGOs working in the field. In addition to this, PMI can leverage results for both its internal operations as well as external communication.

Overall, **the data science component is proceeding according to the timeline**. Data curation is complete and needs only updating as new data becomes available. The structural forecasting model is estimated and first conclusion for policy and law enforcement are outlined. The previous delivery period saw the completion of the risk management forecasting at the regional level. Results and insights from the risk management forecasting model are used to construct risk scores and show which ones of a country regions present the greatest risk for illicit cigarette incidence.

Components from the data science part of the IT^{2}RM project are made available to other project participants on time so that overall delivery is assured.

Background

This PMI Impact project – IT for Illicit Trade Risk Management (IT^2RM) 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 builds upon the insights gleaned from the previous modeling exercises and aims to use the granular-level forecasting model to create an intuitive and easy-to-use risk scoring and visualization that can guide stakeholders in understanding the dynamics of illicit cigarette contraband as well as to take action.

More specifically, the risk scoring must answer to the following requirements:

- **Produce granular level risk scoring** – a detailed granular regional-level analysis of illicit cigarette incidence at the regional level in Bulgaria, Turkey and Serbia is obtained in the previous delivery period, and it is now to be used for scoring which regions have a higher or lower level of risk.
- **Require limited data** – since regional level data for all target countries is extremely sparse, the risk scoring needs to take correspondingly limited input in order to generate adequate scores. The best route to follow would be to highly leverage forecasts generated from the risk management model.
- **Ensure quick time to value** – the risk scoring procedure aims to ensure that scores are produced rapidly and unambiguously so that stakeholders may use them in their activities and be able to generate significant value from them.
- **Lends itself to intuitive visualization** – working with a full-scale forecasting model may be undesirable for some stakeholders due to matters of time and convenience. The risk scoring approach therefore must be able to generate results that are easy to visualize and understand by a wide range of users with widely varying backgrounds and interests
- **Next Steps** – the risk scoring exercise should take into account next project steps and easily incorporate activities included in them. Most notably, next steps of the project include data updates that need to be integrated seamlessly.

In short, this delivery report outlines the approach to using forecasts from the risk management model in order to generate risk scores for different regions under study, allowing project stakeholders to effectively prioritize time and resources to the regions that need them most.

Overall Risk Scoring Approach

Illegal cigarette incidence can be measured on the national level, and this has critical repercussion for forming a coherent policy and response by the government and central authorities against overall contraband activities in the country. For this purpose, a more encompassing structural forecasting model with both accuracy and high explanatory power was designed throughout the previous project periods. However, there is often a necessity for a more granular understanding of illicit cigarette incidence at the regional level that can guide local law enforcement and preventive activities. A more

granular regional understanding is a natural first step towards building a risk-scoring model that can more effectively guide enforcement and also inform local economic and social policy that address the root causes of contraband. The granular risk management forecasting model was delivered during the work described in Delivery Report 5, and the current task revolves around the need to create a simple and intuitive risk score for each region.

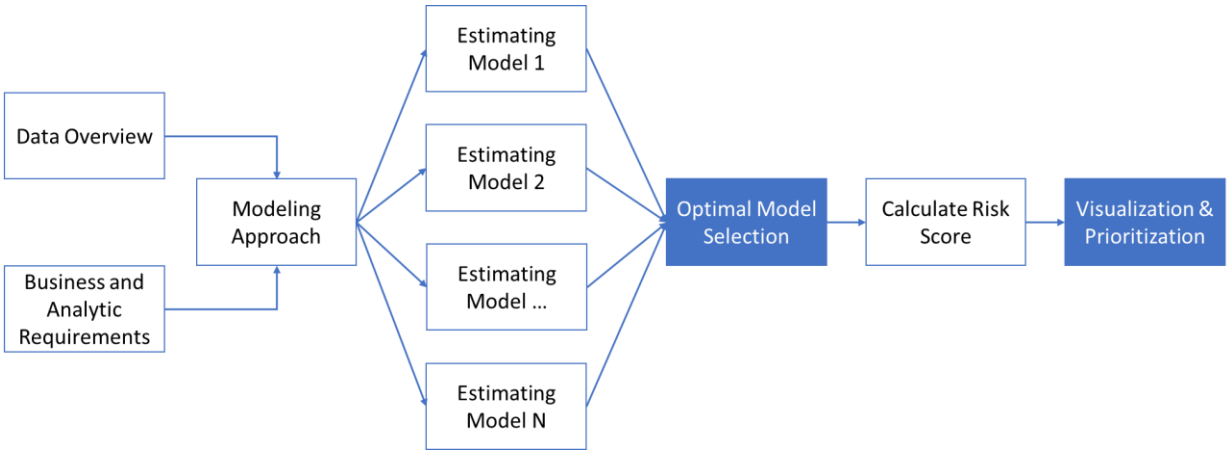

Figure 1: Overall risk management analytic process

This **enriches the analytic pipeline** so that it currently has the following steps:

1. Prepare granular level data for forecasting
2. Estimate a large number of ARIMA models
3. Select best (optimal) model by using a relevant information criterion
4. Use optimal model to generate forecasts and confidence intervals
5. Calculate the percentage of expected deviations of future state from current level
6. Rank expected deviations in descending order
7. Apply color coding as follows:
    a. Red – high risk
    b. Yellow – medium risk
    c. Green – low risk
8. Export and visualize risk scores at a regional level

A schematic representation of the process in presented in Figure 1. The key question for creating a risk scorecard would be what constitute a high-risk region. Risk management traditionally leverages two types of concepts:

- Probability of occurrence of a given event – how likely it is that this event will materialize given historical data.
- Impact of the said event – how large is the resulting damage from this event's occurrence.

Current data in the project does not fully support those definitions and they need to be further adapted to both the available information and the practical need to apply risk management concepts in a widely divergent context. Nevertheless, the resulting risk scoring methodology still needs to be able to a produce a risk prioritization matrix that empowers different stakeholders to glean actionable insight from the analytic product and be able to create value based upon it. Therefore we propose the following risk scoring dimensions:

- **Forecasted illicit cigarette incidence** – this dimension proxies the probability of illicit cigarette incidence occurring in a given region in the subsequent time period. This forecast in generated by the risk management forecasting model, developed in the previous delivery period and leverages quick and simple to use ARIMA time series models. Forecast generation is

automated via means of an extensive codebase and once the data is fed, any stakeholder can run the modeling, and receive the necessary forecasts.

- **Current incidence of illicit cigarettes** – a proxy for the impact of risk realization in the contraband context is the overall average incidence of illicit cigarettes. It is quite clear that large incidence is related to a large impact, and a small one – to a correspondingly smaller impact.

Those two dimensions can be used to construct a Risk Prioritization Matrix. For simplicity, each of those dimensions can be ranked on a given scale, and the resulting overall expected risk is simply the product of the two ranks. Ranking can proceed in one of two ways. First, it can be automated and data-driven by dividing the total results according to some statistical measure. The simplest one would be the average, thus receiving rankings of each region as being below or above average in terms of current incidence, and an additional ranking of it being below or above average in terms of forecasted incidence. This leaves the analyst with a Risk Matrix with four quadrants with the topmost one (above average current and above average forecasted incidence) containing the highest risk regions. One can divide the results in even more groups as desired. If for example three groups are needed (for a 3x3 matrix), then one can easily divide the sample into three groups with cutoff points at the 33rd and 67th percentile.

Alternatively, rule-based scoring may be applied to the results. In this way both the forecasted and the current incidence may be ranked according to some rules which denotes them as low, medium or high incidence level. By combining the rankings, one obtains the Risk Prioritization Matrix which shows how those two dimensions stack against each other (see Figure 2). The highest risk regions are characterized by one dimension ranked as high and the other as at least medium risk. Those tend to be the highest priority for intervention. An important note is in order here – one may argue that collapsing continuous numeric variables into ranks diminishes their informational value. While this is technically true, one should note that there might be some measurement errors due to the data collection and thus both actual and forecasted incidence may only be approximations of the true values. Thus, turning them into ranks addresses possible errors and focuses on what matters most – prioritizing regions by comparing them among each other.

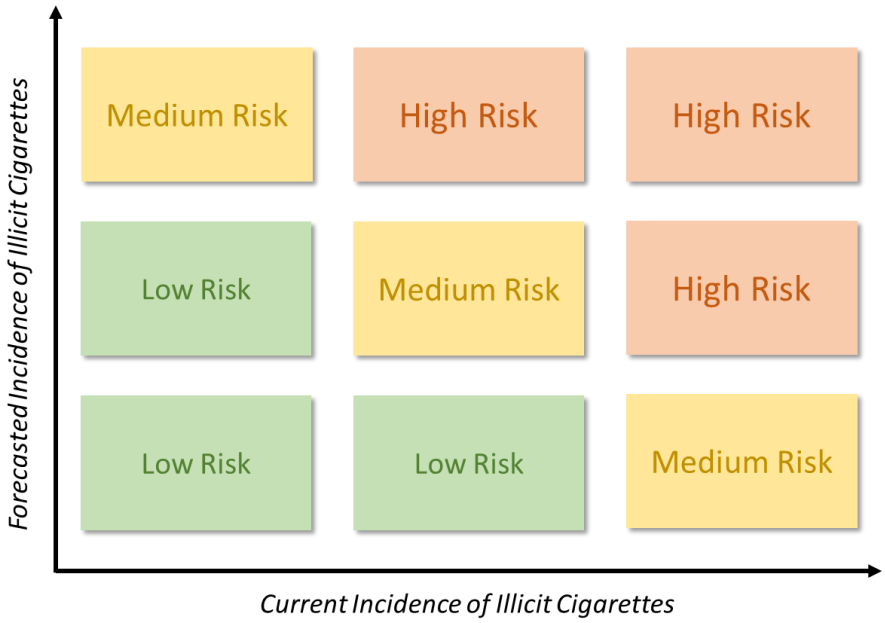

Figure 2: Topline Graph of Illicit Cigarette Incidence Risk Prioritization Matrix

While this definition is rooted in risk science, its adapted version has a very intuitive interpretation in the context of cigarette contraband. It broadly states that a region which has had a significant about of illicit incidence and is likely to continue to exhibit high contraband levels, is very high risk. Conversely, a region with no or limited incidence that is not expected to develop it s market as low risk. Finally, intermediate regions in terms of current and forecasted incidence occupy a middle ground in terms of their risk score.

Once the Risk Prioritization Matrix is prepared, concrete risk management and resource prioritization actions need to be taken. While the specific actions depend on the stakeholder who leverage the analysis, including their motivation, resources, and goals, there are still some generic strategies that can be applied to regions at different risk levels:

- **High Risk Regions** – those are regions that exhibit risks which need to be actively managed. Resources need to be allocated at both tackling the underlying drivers of contraband (i.e. those outlined in the forecasting and policy model) as well as taking tactical action such as more focus on law enforcement.

- **Medium Risk Regions** – these are regions characterized by risks that can be passively managed. While they need to be considered as part of an overall strategy for the decrease of illicit cigarette incidence, they need not be the main focus of policy-making, lobbying and enforcement efforts.

- **Low Risk Regions** – those regions are mostly calm – they exhibit relatively lower levels of contraband incidence and it is not expected to spike dramatically. The risk here can only be monitored in case of change but does not need to be actively managed.

This analysis can either be performed at the level of an individual country, or within a group of countries, depending on the level of decision-making and resource allocation.

Risk Maps and Relative Risk Scores

Leveraging the risk management model forecasts, as well as long-run historical data on illicit cigarette incidence we can create a Risk Map for the countries under study in the current project. We briefly review the risk maps for:

- Turkey;
- Serbia;
- Bulgaria.

Data for Ukraine is still not available at a granular regional level and the expectation is that in the next delivery period this will be obtained and appropriately included in the overall analysis.

The Case of Turkey

The Turkish Risk Map at the regional level is displayed in Figure 3. It shows a clear-cut display of what regions pose the **highest level of illicit incidence risk**. Those are:

- Eastern Anatolia
- South East Anatolia

Both of these regions have a much higher contraband incidence that the rest of the country, and this is a trend which is highly likely to carry over into the future. Law enforcement efforts need to be particularly focused on those two regions. The Mediterranean region is also of interest due to its relatively high incidence.

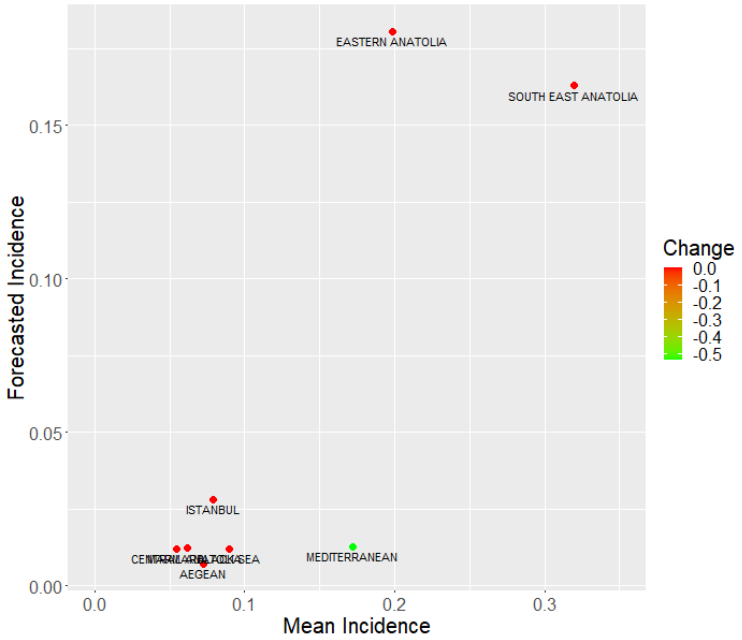

Figure 3: Illicit Cigarette Incidence Risk Map for Turkey

A data-driven risk prioritization is shown in Figure 4. The overall risk map is divided into four quadrants by the mean values of realized and forecasted incidence. Quadrant I shows regions that have above average values on both dimensions and are therefore particularly high risk. Eastern Anatolia and South East Anatolia neatly cluster in this quadrant. Most regions are in the least risky quadrant (Quadrant III), displaying below average value on both dimensions, and thus only need to be monitored.

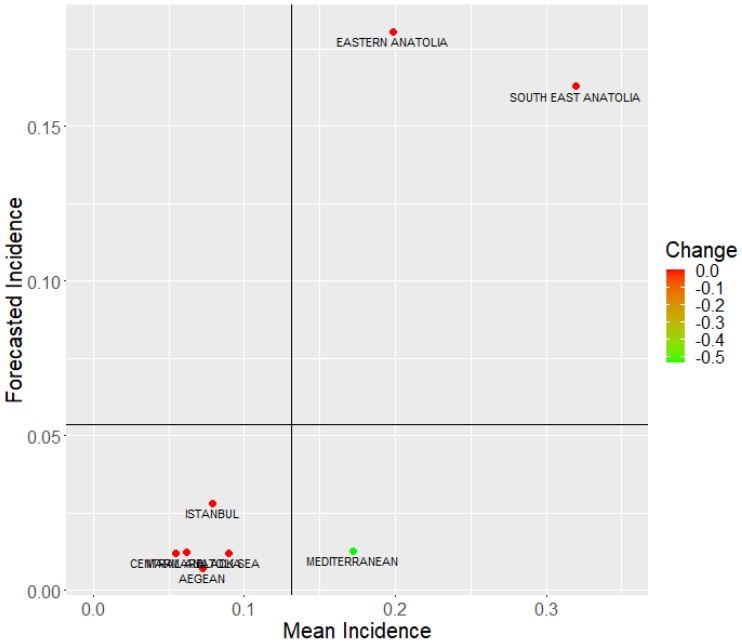

Figure 4: Illicit Cigarette Incidence Risk Map with Automated Scoring for Turkey

The Case of Serbia

A similar Risk Map for the Serbian regions is shown in Figure 5. Again, it displays the mean and expected illicit cigarette incidence with coloring that reflects the forecast dynamics.

The two regions with highest values on both dimensions, and thus the **highest-risk ones are the following**:

- Subotica
- Novi Pazar

This is even clearer when the Risk Map is divided into four quadrants by the average values (see Figure 6). Again, Subotica and Novi Pazar neatly cluster in the high-risk Quadrant I. We should note that the automated data-driven prioritization (e.g. by dividing the results according to the average, or into percentiles) has numerous benefits. It automatically focuses the stakeholder's attention to the key part of the risk map and can also be done and fed into an automated decision-making system. This serves to formalize expert intuition as can be seen in the data for Serbia.

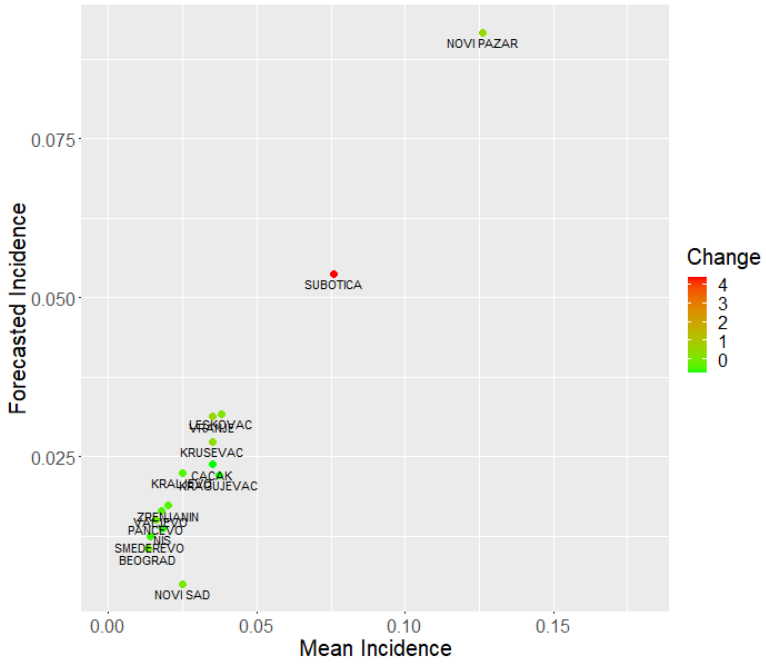

Figure 5: Illicit Cigarette Incidence Risk Map for Serbia

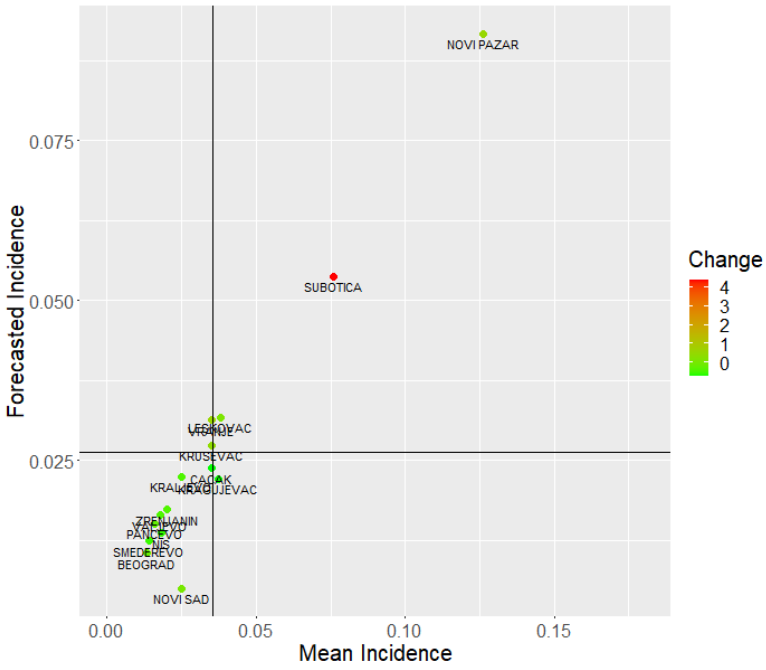

Figure 6: Illicit Cigarette Incidence Risk Map with Automated Scoring for Serbia

The Case of Bulgaria

Using data for Bulgaria's regions we also prepare a complete Risk Map for the country (see Figure 7), following the exact conventions as we did with Turkey and Serbia. Since Bulgaria has data for more regions that the other two countries under study, and since there is much more variability in them, it is not as easy to pinpoint highest priority regions from simply looking at the Risk Map visualization. This holds particularly true for edge cases that are not clearly high or medium risk such as Kjustendil or Veliko Tarnovo.

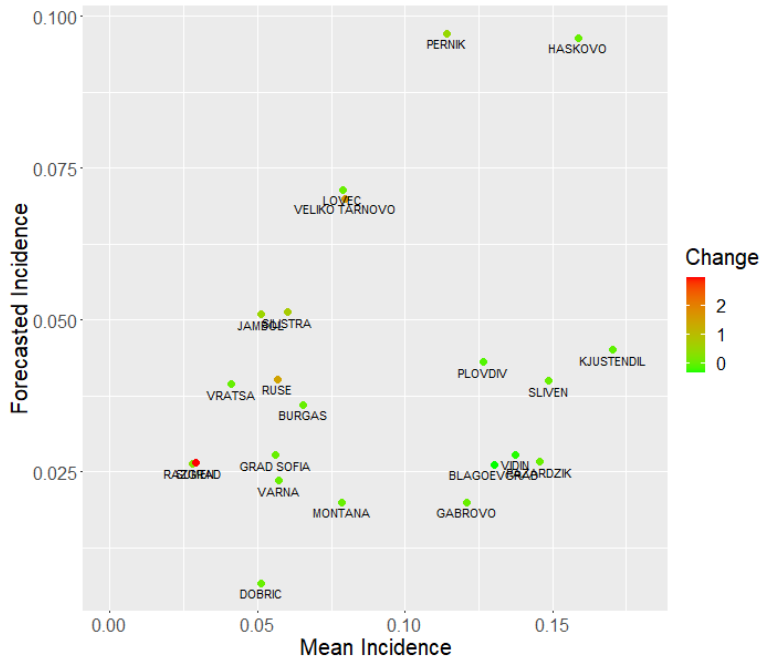

Figure 7: Illicit Cigarette Incidence Risk Map for Bulgaria

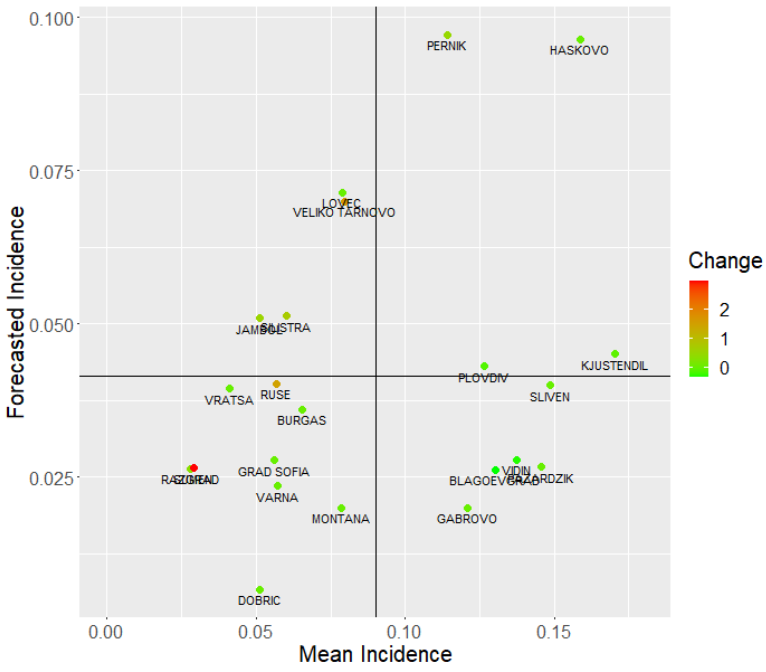

Figure 8: Illicit Cigarette Incidence Risk Map with Automated Scoring for Bulgaria

To aid the risk priority identification process, we divide the Risk Map into four quadrants, whereby the cutoff lines present the average realized incidence on the one hand, and the average forecasted incidence on the other. Results are shown in Figure 8. High-risk regions are positioned in the first quadrant where their incidence values are above average across both dimensions. The **riskiest regions in Bulgaria in terms of cigarette contraband are the following**:

- Haskovo
- Pernik
- Plovdiv
- Kjustendil

In addition to that the map displays many regions as medium-risk (Quadrants II and IV) and quite a few with a low level of risk (Quadrant III).

Risk Ranks and Risk Matrix

An alternative approach to designating risk levels is by applying rules derived from both business logic and the data at hand. This can be leveraged to design simple ranking of both dimensions ranging from 1 (Low) through 2 (Medium) to 3 (High), essentially achieving a Risk Matrix with nine possible quadrants where different regions may fall. The risk prioritization is done in accordance with the Prioritization Matrix, shown in Figure 2.

Proposed Risk Ranking Rules for Mean Incidence are as follows:

- **Low (1)** – this corresponds to mean incidence per region below 5%
- **Medium (2)** – this corresponds to mean incidence per region between 5% and 10%
- **High (3)** – this corresponds to mean incidence per region above 10%

Proposed Risk Ranking Rules for Forecast Incidence are as follows:

- **Low (1)** – this corresponds to forecast incidence per region below 5%
- **Medium (2)** – this corresponds to forecast incidence per region between 5% and 7.5%
- **High (3)** – this corresponds to forecast incidence per region above 7.5%

The resulting regional risk levels then are as follows:

- **High-risk regions** (6-9) – those should be actively managed
- **Medium-risk regions** (3-5) – those can be passively managed
- **Low-risk regions** (1-2) – those can be monitored in case any change occurs

Using those simple heuristics, one can divide the whole data space into nine quadrants, with each of them corresponding to a specific combination of ranks for the mean and forecasted incidence. This can be done either visually or in a table, with both representations providing an easy-to-grasp and intuitive prioritized map of risks. We further propose color coding, whereby the conventional traffic light approach is used: high-risk regions are colored red, medium-risk ones are colored yellow, and finally – low-risk ones are green. The application of this approach to the three countries under study follows.

The Case of Turkey

Figure 9 displays the nine quadrants following from this approach and maps data on Turkish region onto this space. Again, the regions of Eastern Anatolia and South East Anatolia fall neatly into the highest risk quadrant, showing that these are to ones to which sufficient attention needs to be given in terms of both resources and efforts.

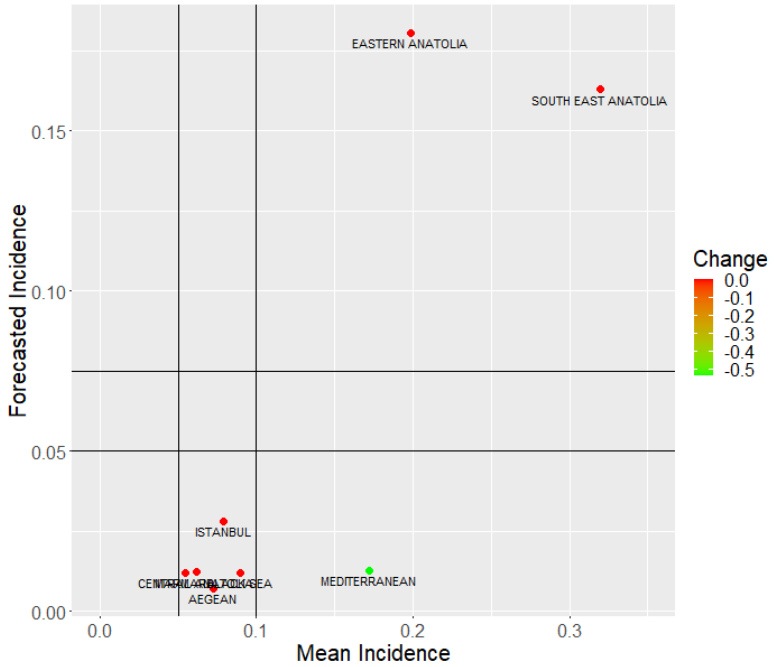

Figure 9: Illicit Cigarette Incidence Risk Map with Rule-Based Scoring for Turkey

A similar visualization is presented in Table 1, whereby each of the regions is ranked 1 to 3 in terms of long-run average (mean) and forecasted incidence. The resulting risk score is the product of those and from it stems the overall color-coded risk level. The table allows the easy identification of the two high-risk regions, and also of one which is medium risk – the Mediterranean region.

Table 1: Risk Prioritization Matrix for Turkey

Country	Region	Mean Incidence Rank	Forecast Incidence Rank	Risk Score	Risk Level
Turkey	MEDITERRANEAN	3	1	3	Medium Risk
Turkey	SOUTH EAST ANATOLIA	3	3	9	High Risk
Turkey	CENTRAL ANATOLIA	2	1	2	Low Risk
Turkey	MARMARA	2	1	2	Low Risk
Turkey	BLACK SEA	2	1	2	Low Risk
Turkey	AEGEAN	2	1	2	Low Risk
Turkey	EASTERN ANATOLIA	3	3	9	High Risk
Turkey	ISTANBUL	2	1	2	Low Risk

The Case of Serbia

The Serbian Illicit Cigarette Incidence Risk Map is also divided in nine quadrants according to the rules outlined above (see Figure 10). The highest risk region is Novi Pazar, which is followed by a medium risk region – Subotica. All the other regions in Serbia are classified as low risk. Overall, one should keep in mind that risk is a fundamental characteristic of assets, activities, and geographies. Different metrics and approaches merely serve to approximate and explicate it. It is thus hardly surprising that both the data-driven and rules-driven approaches show consistent and robust results: they are measuring the same thing.

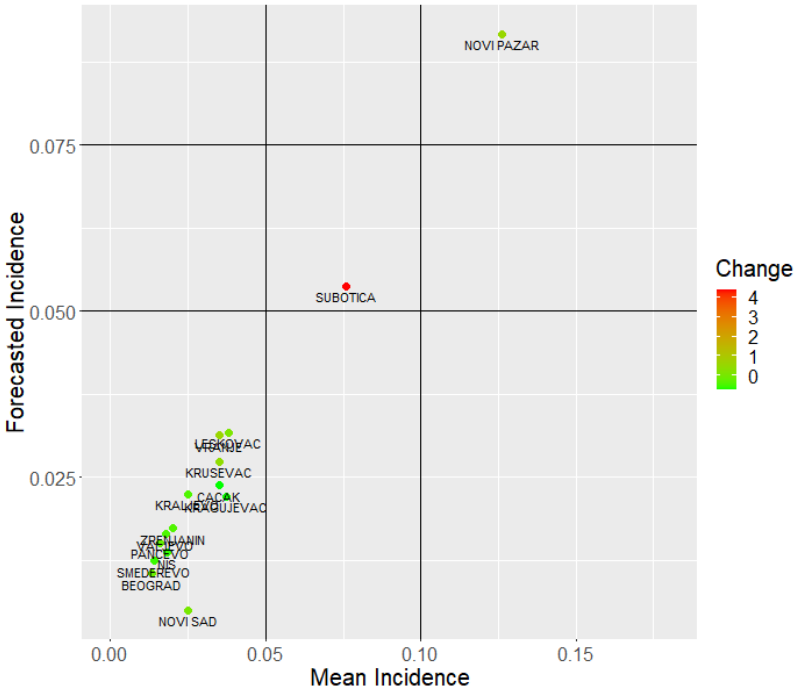

Figure 10: Illicit Cigarette Incidence Risk Map with Rule-Based Scoring for Serbia

The results explicated in the Illicit Cigarette Incidence Risk Map can also be observed in the Risk Prioritization Matrix in Table 2. Which type of visualization is to be used depends on the needs and preferences of the user, but it seems appropriate to introduce the Risk Map first, and only after that to resort to the more detailed Risk Prioritization Matrix, if needed.

Table 2: Risk Prioritization Matrix for Serbia

Country	Region	Mean Incidence Rank	Forecast Incidence Rank	Risk Score	Risk Level
Serbia	BEOGRAD	1	1	1	Low Risk
Serbia	CACAK	1	1	1	Low Risk
Serbia	KRAGUJEVAC	1	1	1	Low Risk
Serbia	KRALJEVO	1	1	1	Low Risk
Serbia	KRUSEVAC	1	1	1	Low Risk
Serbia	LESKOVAC	1	1	1	Low Risk
Serbia	NIS	1	1	1	Low Risk
Serbia	NOVI PAZAR	3	3	9	High Risk
Serbia	NOVI SAD	1	1	1	Low Risk
Serbia	PANCEVO	1	1	1	Low Risk
Serbia	SMEDEREVO	1	1	1	Low Risk
Serbia	SUBOTICA	2	2	4	Medium Risk
Serbia	VALJEVO	1	1	1	Low Risk
Serbia	VRANJE	1	1	1	Low Risk
Serbia	ZRENJANIN	1	1	1	Low Risk

The Case of Bulgaria

The Bulgarian data is characterized by much more variability and a somewhat higher number of regions under study. It is particularly for this type of more complex data that the risk scoring and visualization is best used to make sense of the data.

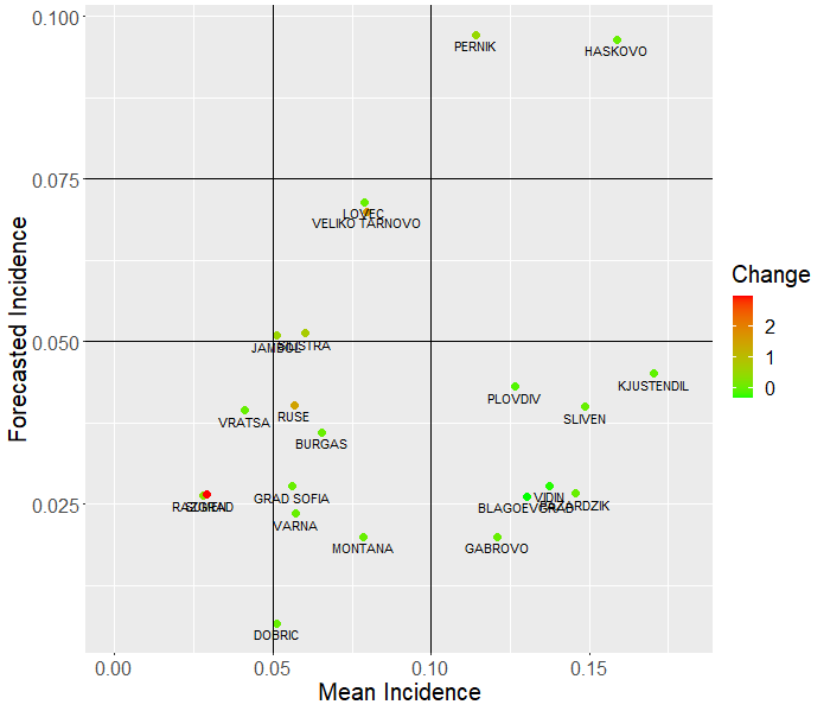

Figure 11: Illicit Cigarette Incidence Risk Map with Rule-Based Scoring for Bulgaria

Again, Figure 11 displays the Risk Map with nine quadrants, but here we observe that most of the quadrants are filled by at least two regions. The highest risk quadrant (topmost on the right) shows two high-risk regions – Pernik and Haskovo that both have a large average incidence over the study and are forecasted to remain the regions with the highest incidence over the next period.

Table 3: Risk Prioritization Matrix for Bulgaria

Country	Region	Mean Incidence Rank	Forecast Incidence Rank	Risk Score	Risk Level
Bulgaria	BLAGOEVGRAD	3	1	3	Medium Risk
Bulgaria	BURGAS	2	1	2	Low Risk
Bulgaria	DOBRIC	2	1	2	Low Risk
Bulgaria	GABROVO	3	1	3	Medium Risk
Bulgaria	HASKOVO	3	3	9	High Risk
Bulgaria	JAMBOL	2	2	4	Medium Risk
Bulgaria	KJUSTENDIL	3	1	3	Medium Risk
Bulgaria	MONTANA	2	1	2	Low Risk
Bulgaria	PAZARDZIK	3	1	3	Medium Risk
Bulgaria	PERNIK	3	3	9	High Risk
Bulgaria	LOVEC	2	2	4	Medium Risk
Bulgaria	PLOVDIV	3	1	3	Medium Risk
Bulgaria	RAZGRAD	1	1	1	Low Risk
Bulgaria	RUSE	2	1	2	Low Risk
Bulgaria	SILISTRA	2	2	4	Medium Risk
Bulgaria	SLIVEN	3	1	3	Medium Risk
Bulgaria	GRAD SOFIA	2	1	2	Low Risk
Bulgaria	SUMEN	1	1	1	Low Risk
Bulgaria	VELIKO TARNOVO	2	2	4	Medium Risk
Bulgaria	VARNA	2	1	2	Low Risk
Bulgaria	VIDIN	3	1	3	Medium Risk
Bulgaria	VRATSA	1	1	1	Low Risk

The figure also shows that there are a number of regions with medium risk that may merit some attention – those are the following: Blagoevgrad, Gabrovo, Jambol, Kjustendil, Pazardzik, Lovec, Plovdiv, Silistra, Sliven, Veliko Tarnovo. The other ten regions of the country are classified as low risk. This is shown in more details in Table 3, which presents the complete Risk Prioritization Matrix for the country, giving a granular overview of all the regions under study.

Use and Applicability

The proposed risk scoring, prioritization and visualization approach stems from key results in the science of risk but is also adapted to the concrete realities of illicit cigarette incidence. Its major benefit is that despite it is data-driven and informed by sophisticated modeling, it is still intuitive and easy to understand and derive actionable insight from. The traffic light system of color-coding risk prioritization should be familiar to a wide range of different stakeholders and thus the whole approach will likely be easier to adopt. The main beneficiaries of this risk management exercise include local authorities, national and local law enforcement agencies, as well as NGOs working in the field. In addition to this, PMI can leverage results for both its internal operations as well as external communication.

Next Steps

The modeling exercise of the first seven quarters of the **project has achieved the following**:

- Collated, procured and organized a large database of socio-economic, demographic, business, and crime statistics for the purposes of the project.
- Performed initial data analysis and visualization of the data.
- Outlined possible drivers of non-domestic cigarette incidence and interpreted their associations within the framework of correlational analysis.
- Selected best predictors and constructed an optimal structural forecasting model and estimated it for the national level using data at hand.
- Investigated a wide variety of approaches for making granular regional-level forecasts that can be used for risk management.
- Selected an ARIMA-based automatic forecasting and optimal model determination approach.
- Estimated forecasts for 55 regions in Bulgaria, Serbia, and Turkey.
- Generated visualizations and exported forecasts in machine-readable form for inclusion in the other parts of the project.
- Used risk management forecasts to generate risk scores and prioritize regions.
- Proposed appropriate color coding and intuitive visualization that can be used by stakeholders with varied backgrounds and interests.

Overall, **the data science component is proceeding according to the timeline**. Data curation is complete and needs only updating as new data becomes available. The structural forecasting model is estimated and first conclusion for policy and law enforcement are outlined. The previous delivery period saw the completion of the risk management forecasting at the regional level. Results and insights from the risk management forecasting model are used to construct risk scores and show which ones of a country regions present the greatest risk for illicit cigarette incidence.

Components from the data science part of the IT^{2}RM project are made available to other project participants on time so that overall delivery is assured.

Some additional work needs to be done to finalize the data science component of the project over the next two quarters. More specifically**, the upcoming tasks are as follows**:

- Procure regional level data for Ukraine and include it in the analysis, estimating forecasts and risk scorecards
- Update data for all countries with 2020 values and re-estimate both the national-level structural forecasting model as well as regional level ARIMA-models and the risk scores
- Further test the visual dashboard in terms of both interface, visualization and user experience (UX) to ensure wide adoption
- Disseminate results
- Wrap-up and hand over the data the data science component of the project