Philip Morris International: PMI IMPACT Project

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

Data Science Delivery Report 5:

Risk Management through Rapid Contraband Forecasting at Regional Level

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

Delivery Date: 30. September 2020

Contents

List of Figures

Executive Summary

The current delivery report builds upon the insights gleaned from the previous modeling exercises and **attempts to undertake a forecasting exercise at a much lower level of granularity, i.e. from national to regional, in order to generate automated, rapid and useful forecasts that can then be utilized for feeding a risk-scoring model**. Forecasting contraband cigarettes at the regional level has a number of specifics, most notably the lack of either long, or detailed, time series of relevant data. This calls for an approach that is radically different from the one utilized with the national level forecasting model. This project installment outlines such an approach and presents all the forecasts generated by the optimal time series analysis models.

Therefore, a time series modeling approach within the autoregressive integrated moving average (ARIMA) is appropriate for this class of problems. It deals effectively with limited data by leveraging only incidence time series and can be efficiently estimated for a wide range of cases. This allows separate granular modeling of each specific region. To ensure an efficient regional model in every case**, for each region we estimate a large number of possible models, then select the best** and use it for forecasting.

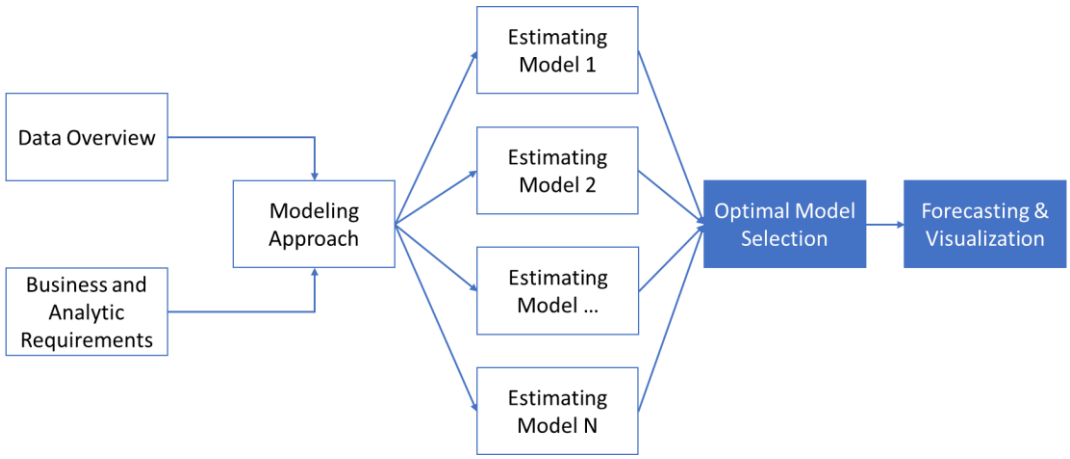

Figure: Overall Modeling Approach Logic

The **steps in this approach** are thus as follows:

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. Export forecast numbers and create visualizations

We estimated 1,944 models for each of the 55 regions under study (for a total of 106,920 alternatives) and using the AICc selected the optimal ones. Based on this exercise there are **a few key conclusions from the results** obtained:

- Given the length of the time series, the optimal models are unsurprisingly very parsimonious. Most of them include 1 or at most two lagged terms of either the regressor or the moving averaged error.
- Some of the regional incidence cannot be adequately forecasted by the ARIMA models and thus this modeling approach has reverted to the long-run average (mean) of the time series. In case of severe data limitations this is a reasonable approach as it generates both a forecast

that can be used, as well as constructs confidence intervals that give indication of the range of uncertainty.

- The dynamics of the forecasting is determined by the whole time series in most case and there is no undue effect of the last observations. This has resulted in regional forecasts that break from trend in the next period ahead which is useful for risk scoring.
- The quality of the forecast is naturally dependent upon both the quality and the quantity of the data that is fed into the model. In this sense it is unsurprising that the forecasting exercise seems most reliable for Turkey, and least reliable – for Serbia.

Detailed results of the selection process, as well as the **generated forecasts and relevant visualizations are to be found in the accompanying Technical Report** on Risk Management Model Selection.

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 current delivery period saw the completion of the risk management forecasting at the regional level. 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**:

1. Develop a risk scoring methodology based on the granular forecasting model that can be used to rank regions according to incidence risk
2. Visualize the risk scorecard in an intuitive way
3. Procure regional level data for Ukraine and include it in the analysis, estimating forecasts and risk scorecards
4. 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
5. Wrap-up and hand over the data the data science component of the project

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 attempts to undertake a forecasting exercise at a much lower level of granularity, i.e. from national to regional, in order to generate automated, rapid and useful forecasts that can then be utilized for feeding a risk-scoring model. Forecasting contraband cigarettes at the regional level has a number of specifics, most notably the lack of either long, or detailed, time series of relevant data. This calls for an approach that is radically different from the one utilized with the national level forecasting model. This project installment outlines such an approach and presents all the forecasts generated by the optimal time series analysis models.

More specifically, we cover the following tasks:

- **Trends Overview at the Regional Level** – a detailed granular regional-level analysis of illicit cigarette incidence at the regional level in Bulgaria, Turkey and Serbia
- **Modeling Strategy and Approach** – overall considerations for choosing a general modeling approach and brief presentation of model logic
- **Model Selection and Estimation** – procedures for estimating the parameters of a vast array of possible models for each single region and criteria for selecting the best performers
- **Regional Level Forecasts and Uncertainty Estimates** – estimation of regional-level forecasts, visualizations, and uncertainty markers
- **Next Steps** – final activities and deliverables needed over the next two quarters in order to complete project scope from a data science perspective

In short, this delivery report aims to present the major portion of work – from approach, through assumptions, estimation, and results for the risk management model. This is accompanied by a technical report with detailed model selection and complete results that can be referenced as needed. Furthermore, the results are exported in a .csv for easier visualization within order project parts and a visual plot with a four-period ahead forecast, including confidence intervals, is presented.

Trends Overview at the Regional Level

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.

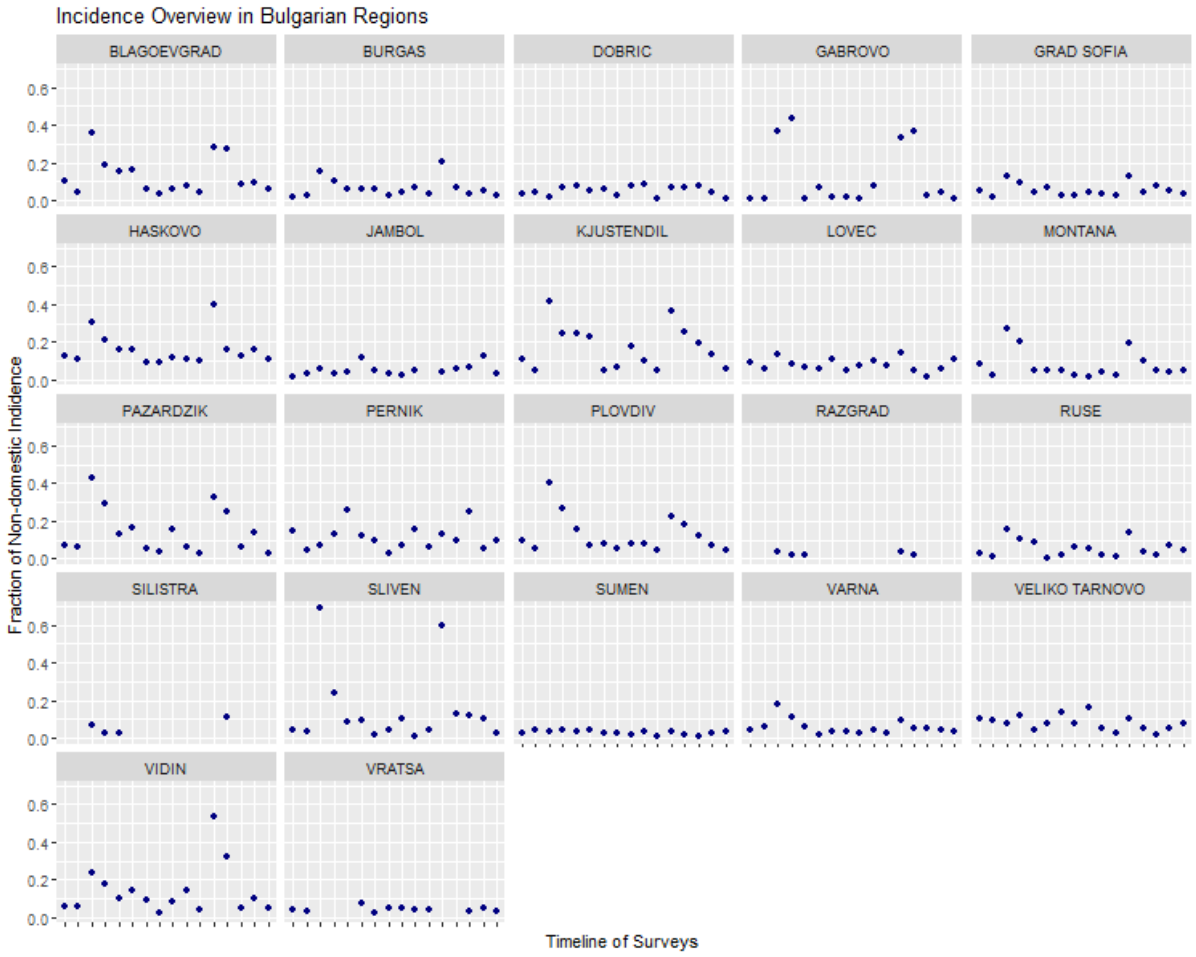

Figure 1: Illicit Cigarettes Proportion of Total Cigarettes Across Different Regions in Bulgaria

We begin the granular forecasting exercise by presenting a regional overview of illicit cigarette incidence. Figure 1 presents incidence dynamics in the Bulgarian regions. It is quite obvious that the capital has among the lowest proportions of illicit cigarette incidence, while other regions such as Blagoevgrad, Haskovo, Kjustendil, and Vidin register much higher proportion of contraband cigarette consumption. Generally, the incidence seems to be positively correlated with both the economic and social development of the regions, as well as with their proximity to borders. It is hardly surprising that Vidin – one of the poorest regions in the country also registers one the highest proportions of illegal cigarettes. On the other hand, data for Plovdiv – the second largest and one the most affluent cities in the country is somewhat surprising in this respect. It shows significant consumption of illicit cigarettes of up to 30-40% in some of the research waves.

All in all, Bulgarian regional data shows two striking conclusions:

- There **seems to be an overall downward trend in illicit cigarette incidence** across the board, and this is true for all regions. However, this conclusion needs to be qualified by the fact that some regions that had very low incidence in the first place, e.g. Dobrich, Sofia, Sumen, Varna, Veliko Tarnovo, Vratsa show no discernible trend in either direction;
- Some of **the regional data shows one or two striking outliers that are in much disagreement** with the rest of the time series. Usually, after them the time series tend to return to either its long-run mean or its long-run trend. This may need to be further investigated at the data

collection level to ensure whether this data is indeed correct and can reasonably be included in further analytics.

On the statistical and modeling side, such outliers break the usual trends and make the forecasting of the time series somewhat more difficult. Given the short length of the time series, this is an additional challenge for the exercise. In practical terms, this means that the reliability of regional-level forecasting needs to be carefully examined and **only the next one or two periods ahead of the forecast should be used**.

Data for Serbia at the regional (city) level is presented visually in Figure 2. It is significantly different from the data for Bulgaria – generally, there is much less contraband in Serbia in the start of the period, and there is limited trend in either direction in most of the regions. The only exception in this respect seems to be Subotica. The data for Novi Pazar is particularly notable as well – it has only four data points and these vary widely – from 5% to 25%. It is highly unlikely that the generating process of those is so volatile, and these results may also be an artefact of data collection. Either way, Novi Pazar has the highest proportion of incidence and will likely need to be prioritized in anti-contraband enforcement activities.

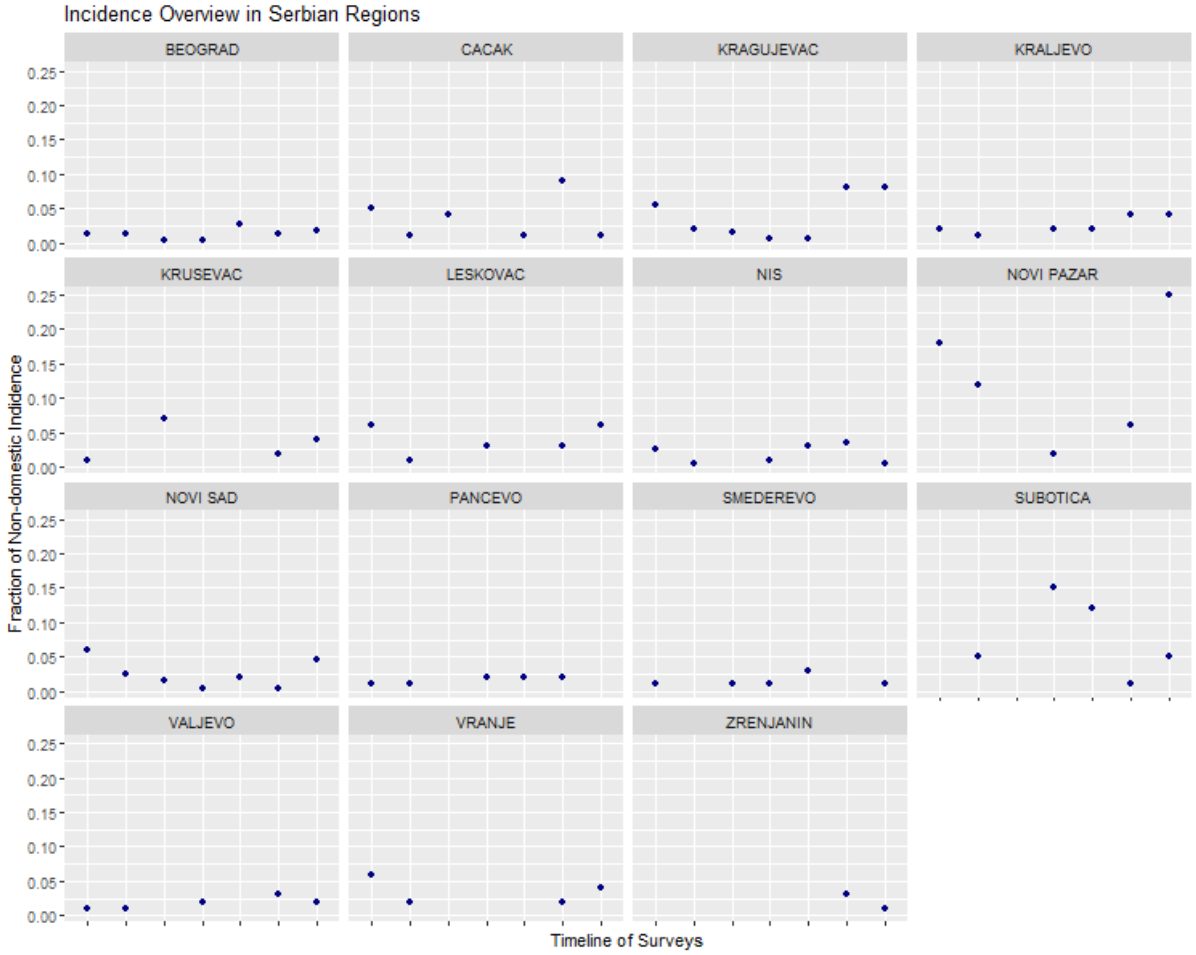

Figure 2: Illicit Cigarettes Proportion of Total Cigarettes Across Different Cities in Serbia

Two major conclusions stem from the Serbian data:

- The **overall contraband levels in Serbia are comparatively low** and show no specific regional trend. The only city and its adjacent region that registers higher levels and thus needs to be prioritized is Novi Pazar.

- **Data for Serbian regions is extremely limited** with some regions such as Zrenjanin having as few as two observations. On the one hand this underlines that forecasting results need to be interpreted with caution. On the other, it calls for expanding the dataset and including additional new data from the last year for the successful project finalization.

Finally, data for Turkish regions is presented in Figure 3. Of the total of eight regions for which data is collected, five register low and stable levels of illicit cigarette incidence. Those are the Aegean region, the Black Sea Region, Central Antalia, Istanbul, and Marmara. The incidence in the other three regions (Eastern Anatolia, Mediterranean, and South East Anatolia) are significantly higher. In the latter three there seems to be a downward trend in the last few quarters of data, showing an overall improvement, but they still do register values that are above the rest.

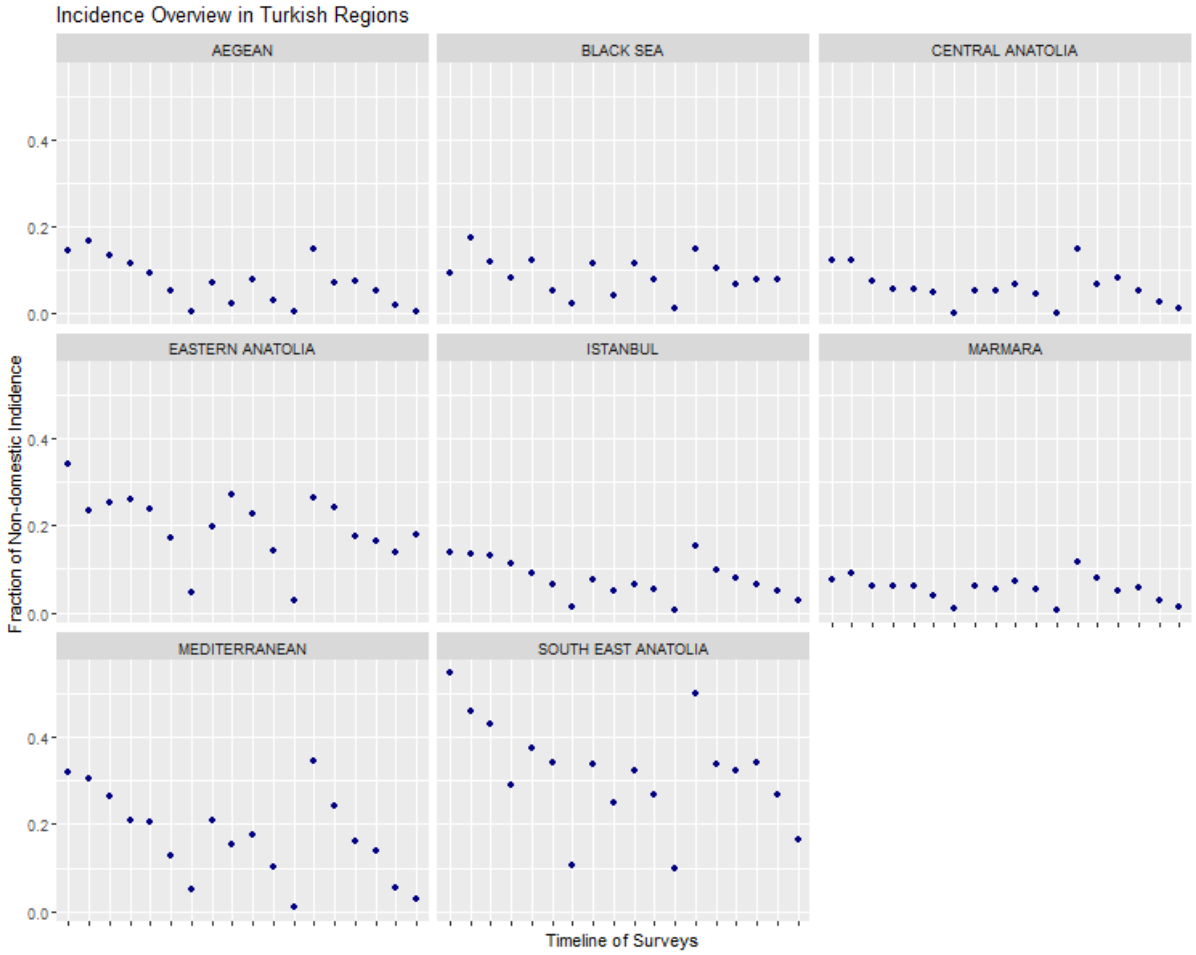

Figure 3: Illicit Cigarettes Proportion of Total Cigarettes Across Different Regions in Turkey

We can a couple of conclusions from the Turkish data under investigation:

- **Three Turkish regions show a notably higher illicit cigarette trade incidence** than the rest and thus law enforcement and prevention will be more effective if focused on those. The remaining five tend to have less than 10% consumption of illegal cigarettes, which however, remains stable.
- **Turkey has the most data** of all the countries under investigation and this will lead to the most accurate and reliable forecasting. In this sense, forecasts for Turkey may be used for neighboring countries such as Bulgaria as potential leading indicator for increases in contraband in border regions.

Data for Ukraine at the regional level is insufficient on this point to undertake such granular forecasting. However, as data becomes available the approaches and models may be seamlessly applied to Ukraine, as well. Based on the regional overview and data availability, we can proceed to creating granular forecasts for the sub-national level.

Modeling Strategy and Approach

The forecasting approach for granular level estimates of illicit cigarette consumption needs to take recourse to both objective analytic and business needs, as well as data availability and relevant assumptions. More specifically, the approach needs to be appropriate given the following considerations:

- Need to create rapid granular level forecasts with limited human interactions;
- Need to estimate a relatively large number of forecasts – for Bulgaria, Serbia, and Turkey those are 55 in total, with potential increase for Ukraine regions to a total of upwards of 60;
- Very limited time series for illicit cigarette incidence;
- Regional level economic and social data tend to be unavailable or published long after the relevant time period has ended;
- Large difference in regional level data dynamics making it necessary to estimate separate models for each region of each country.

Overall Approach

Therefore, a time series modeling approach within the autoregressive integrated moving average (ARIMA) is appropriate for this class of problems. It deals effectively with limited data by leveraging only incidence time series and can be efficiently estimated for a wide range of cases. This allows separate granular modeling of each specific region. To ensure an efficient regional model in every case, for each region we estimate a large number of possible models, then select the best and use for forecasting.

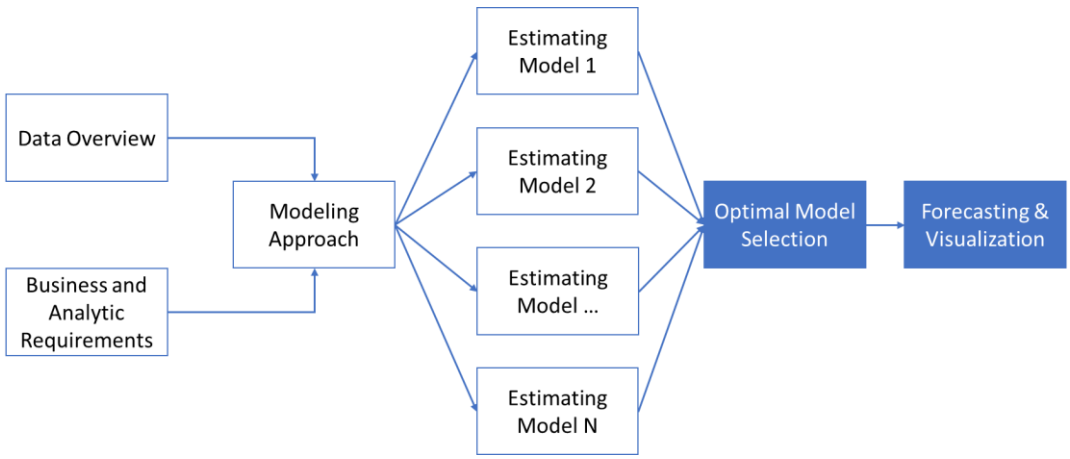

Figure 4: Overall Modeling Approach Logic

The steps in this approach are thus as follows (summarized in Figure 4):

6. Prepare granular level data for forecasting
7. Estimate a large number of ARIMA models
8. Select best (optimal) model by using a relevant information criterion
9. Use optimal model to generate forecasts and confidence intervals

10. Export forecast numbers and create visualizations

ARIMA Models

It is the task of the ARIMA models to try and capture the information, contained in the time series and model current variable realization as a function of past ones. In the simplest version of the model, the current realization of a given metric y_t is presented as a weighted function of p previous values y_{t-p} (is an error term). This AR(p) model is defined as follows:

$$y_t = \theta + \sum_{i=1}^{p} \beta_i y_i + \varepsilon_t$$

Additional information can be contained in the error structure of the time series. This can be modeled through a moving average of the error term. Should the analyst use q past values of the error terms to model current variable realization, then we reach a MA(q) of the following form:

$$y_t = \mu + \epsilon_t + \sum_{i=1}^{q} \alpha_i \varepsilon_i$$

Combining those two equations one gets a more fuller perspective on the time series, thus reaching the classical ARMA(p, q) model:

$$y_t = \beta_0 + \sum_{i=1}^{p} \beta_i y_i + \sum_{i=1}^{q} \alpha_i \varepsilon_i + \varepsilon_t$$

Should the time series be integrated of a certain order d, this can also be taken into account, finally reaching the ARIMA(p, d, q) model. A particular strength of this class of models is their ability to accommodate seasonality in the time series, and its mirrors the structure of the de-seasoned model. A model with seasonality is thus denoted as ARIMA(p, d, q)(P, D, Q) to account to the autoregressive, integrated, and moving average parts in the seasonal component of the data.

We use the framework of ARIMA modelling to address the forecasting problem at hands. Initially, we consider the maximum number of lags that would be useful for the series. Given the characteristics of the data, we use the following parametrization of the model:

- p = 5
- d = 2
- q = 5
- P = 2
- D = 1
- Q = 2

With those parameters we reach a pool of 1,944 possible forecasting models for each of the regions under study. In total we thus estimate 106,920 different models for the current granular forecasting exercise for the 55 regions under study.

Model Selection and Information Criteria

Initially we fit all specified alternative models to the data set. The problem is now straightforward – to select the 55 best models out of a total of 106,920 alternatives. For this purpose, we can use a number of information criteria. Three criteria have become particularly popular in practice – the Bayesian Information Criterion (BIC), the Akaike Information Criterion (AIC), and the corrected Akaike Information Criterion (AICc). To better understand model fit, we define the likelihood function L, equal to the probability p of observing the data x given a model M with a parameter set of ϑ, or:

$$L = p(x|\theta, M)$$

If we denote the maximized value of this likelihood function as L_{max}, then the BIC of a model with k parameters and a sample size of n is defined as follows:

$$BIC = -2 \ln L_{max} + k \ln n$$

Essentially, the information criterion is a measure of model quality, which represents informational loss as data is presented by a given model. Thus, it can serve to select the best model among a set of alternatives taking into account the tradeoff between fit and parsimony (or number of parameters). For a given dataset better models have lower values of their information criteria. The BIC is often criticized on the grounds of its difficulty of handling complex collections of model or feature selection, and it is only valid as $n >> k$. This, together with some derivation considerations and performance issues lead many authors to propose using the Akaike Information Criterion instead. It is defined as follows:

$$AIC = 2k - 2 \ln L_{max}$$

The AIC estimate is valid asymptotically, which means that some corrections needs to be made for finite sample sizes, leading to the corrected version of AIC, or AICc. The formula for univariate series with normally distributed residuals is as follows:

$$AICc = AIC + 2k(k+1)/(n-k-1)$$

The AICc penalizes more heavily models with more parameters than AIC and will thus lead to the selection of more parsimonious ones. In addition to that we should keep in mind that as the sample size grows AICc converges to AIC and this is why many authors recommend it as the primary criterion to use for model selection exercises. **We follow the best practices in the literature and use AICc to select the optimal models for our time series.** From an empirical perspective, the differences between the AIC and AICc for the series under study are very small and any of the criteria will lead to the selection of the same optimal model.

Model Selection and Estimation

We estimated 1,944 models for each of the 55 regions under study and using the AICc selected the optimal ones. Detailed results of the selection process, as well as the generated forecasts and relevant visualizations are to be found in the accompanying Technical Report on Risk Management Model Selection.

Based on this exercise there are **a few key conclusions from the results** obtained:

- Given the length of the time series, the optimal models are unsurprisingly very parsimonious. Most of them include 1 or at most two lagged terms of either the regressor or the moving averaged error.
- Some of the regional incidence cannot be adequately forecasted by the ARIMA models and thus this modeling approach has reverted to the long-run average (mean) of the time series. In case of severe data limitations this is a reasonable approach as it generates both a forecast that can be used, as well as constructs confidence intervals that give indication of the range of uncertainty.
- The dynamics of the forecasting is determined by the whole time series in most case and there is no undue effect of the last observations. This has resulted in regional forecasts that break from trend in the next period ahead which is useful for risk scoring.
- The quality of the forecast is naturally dependent upon both the quality and the quantity of the data that is fed into the model. In this sense it is unsurprising that the forecasting exercise seems most reliable for Turkey, and least reliable – for Serbia.

Regional Level Forecasts and Uncertainty Estimates

Using the approach proposed in the previous section we automatically (and rapidly) estimate forecasting models for all 55 regions, and then select the best performer. This section reviews the results obtained for both the region of the nation's capital as well as for the riskiest other region (whereby we define risk by observed data volatility).

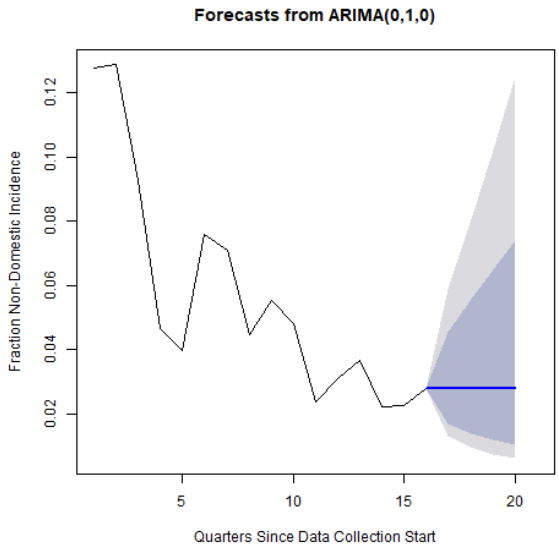

Figure 5: Regional Level Forecast for Sofia

In Bulgaria, the Sofia region has shown a dramatic decline in illicit cigarette incidence, dropping from over 12% in the beginning of the survey to less than 4% in 2019. The forecasting model (Figure 5) estimates that incidence will stabilize at levels of around 3% over the next four quarters. Since the time series is of somewhat adequate size, the 90%-confidence intervals can be reliably estimated.

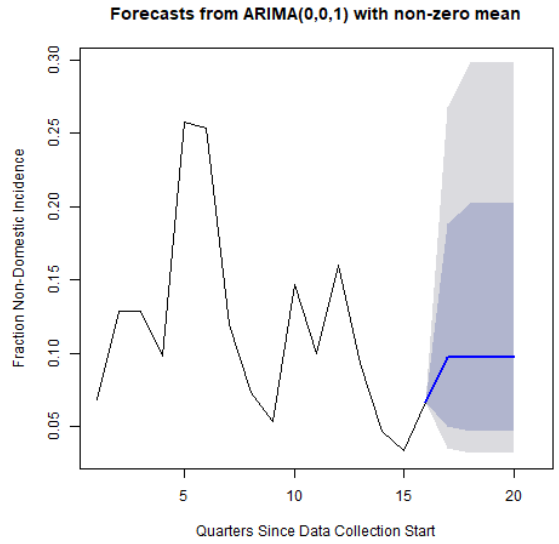

Figure 6: Regional Level Forecast for Pernik

The Sofia forecast ranges with 90% certainty between 1% and 7% illicit incidence. Those are acceptable values and thus the region is not particularly risky. On the other hand, the forecast for Pernik shows a possible uptick in illegal cigarette incidence. The historical values have ranged significantly from below 5% to more than 25%, until they fell to about 6-7% at the end of the available time series. The forecasting model projects an increase to almost 9% in the upcoming period. Given the large observed variance, the confidence intervals are also correspondingly large, varying from 5% to 20% with 90% confidence. The Pernik region is thus of higher risk than Sofia.

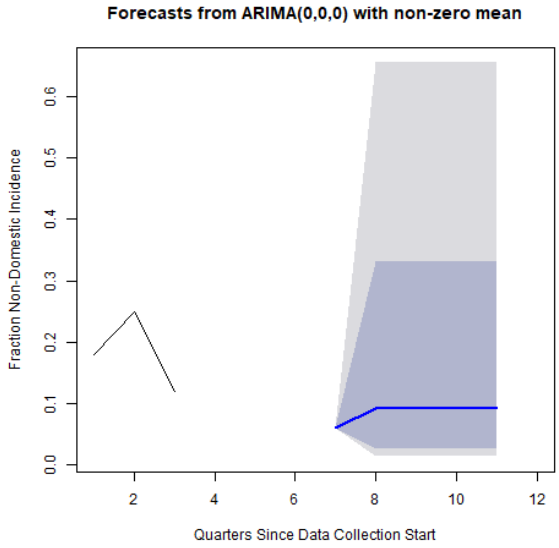

Figure 7: Regional Level Forecast for Novi Pazar

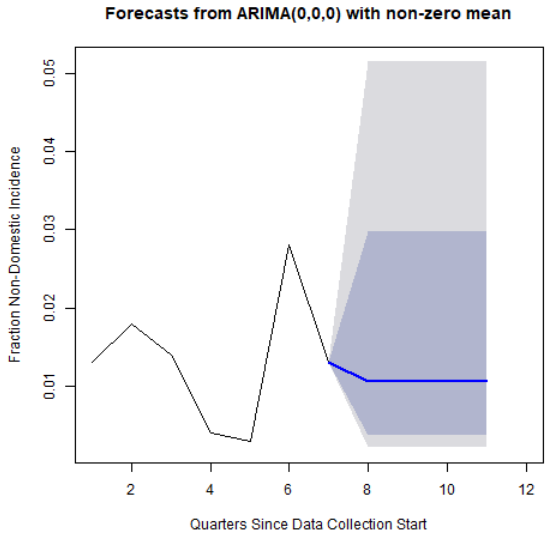

Figure 8: Regional Level Forecast for Belgrade

Data for two of Serbia's cities and, by assumption, their corresponding regions are presented in Figure 7 and Figure 8. Novi Pazar (Figure 7) is a region with significant volatility and somewhat more limited data exposure. It is a classic example of the approach in such cases. The automated algorithm finds that there is too limited structure to usefully model such short time series and thus find the best forecasting approach to be using the mean of time series, while indicating large confidence intervals to show the uncertainty of the forecast generated. All in all, the expectation is for the next quarter the

incidence in Novi Pazar to be a bit lower than 10%. However, this point estimate may vary significantly. Serbia's capital, Belgrade, show a completely different story. The time series here is somewhat longer and the forecast considers the unexpected spike in the end of the period of available data as somewhat of an outlier. Thus, the point forecast estimates a much lower level of incidence in the periods ahead that stabilizes at levels of 1-2%. Even the 90% confidence intervals are rather narrow, ranging from below 1% to up to 3% incidence. This data shows that the Serbian capital is not a particularly risky region.

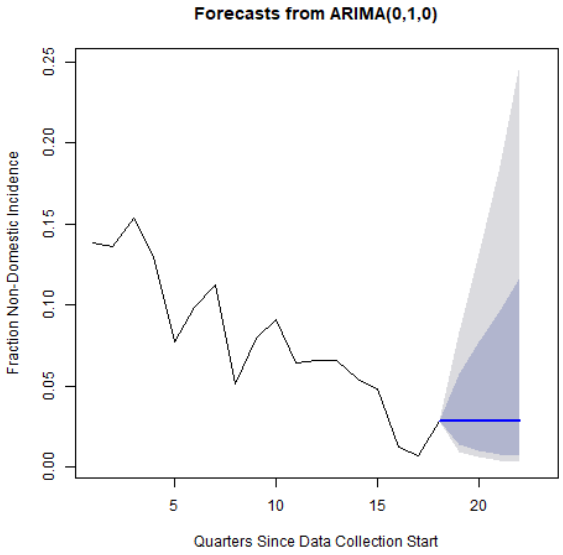

Figure 9: Regional Level Forecast for Istanbul

Finally, data for Turkey is presented in Figure 9 and Figure 10. Among the three countries under investigation, Turkey has the greatest data availability and is thus most prone to modeling. Forecasts here tend to be more reliable, and the overall estimates – more stable. However, results still need to be interpreted with appropriate caution and only take recourse to one of two periods ahead.

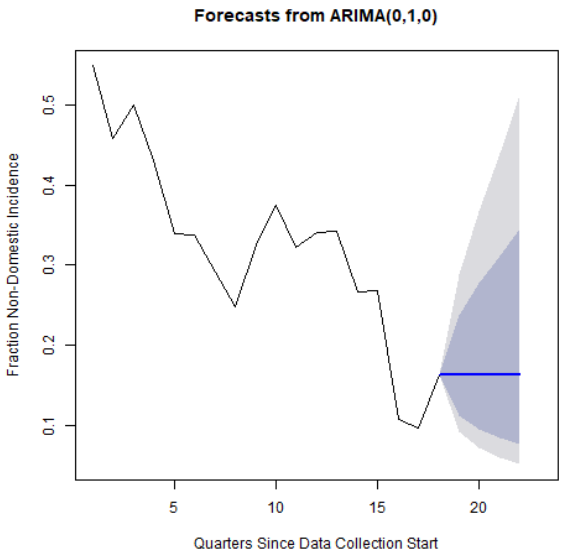

Figure 10: Regional Level Forecast for South East Anatolia

Istanbul (Figure 9) has seen a dramatic decline of illicit no-domestic incidence, dropping from about 15% to less than 5%. The forecasting model assumes that this trend will stabilize at levels below 5%. The 90% confidence intervals imply that the peak expected incidence should not be more than 10%. Thus, Turkey's capital is of lower risk than its other regions. South East Anatolia (Figure 10) lies at the other end of the spectrum. It has registered very high values for incidence – upwards of 50% and has registered a rather dramatic decline to about 10% over a period of 5 years. This decline, however, may be reversible. The forecasting model estimates that incidence levels may go up to 15%, with a significant upward risk – the 90% confidence high estimate stands at above 30%. Thus, the South Anatolia region poses notably more risk than other regions in Turkey.

Next Steps

The modeling exercise of the first six 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

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 current delivery period saw the completion of the risk management forecasting at the regional level. 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**:

- Develop a risk scoring methodology based on the granular forecasting model that can be used to rank regions according to incidence risk
- Visualize the risk scorecard in an intuitive way
- 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
- Wrap-up and hand over the data the data science component of the project