Developing Scenario Segmentation and Anomaly Detection Models

How analytics can be used for BSA/AML compliance programs
## Contents

- **Introduction** ........................................................................................................... 1
- **Segmentation Is the Logical First Step** .............................................................. 1
- **The SAS Approach** .................................................................................................. 2
  - Development of the Segmentation Model .......................................................... 2
  - Developing the Initial Segmentation Model Framework ......................................... 3
  - Development of the Peer Group Model ............................................................... 4
- **Conclusion** ............................................................................................................... 5

## Contributors

- **Jim West**, Senior Data Scientist for Fraud and Compliance Solutions Development, SAS
- **Carl Suplee**, Senior Solutions Architect for Banking Security Intelligence Practice, SAS
Introduction

Since 2012 when the US Office of the Comptroller of the Currency (OCC) began applying the OCC’s Supervisory Guidance on Model Risk Management (OCC 2011-12) to Bank Secrecy Act/Anti-Money Laundering compliance practices, financial institutions have increased their adoption of more rigorous analytics use to improve their BSA/AML monitoring programs. Financial institutions have incentives to find new ways to optimize AML transaction monitoring processes, but it’s uncertain if their reasons for performing such improvements have changed over time.

Firms are quickly finding themselves caught within the oversight scrutiny. Enhanced regulatory pressure to continuously evaluate the firm’s risks, identify emerging trends, report risky activity and expeditiously make changes has caused firms to seek out new and aggressive approaches. To meet these demands, the BSA/AML industry has turned to analytical/statistical methodologies to improve their monitoring programs in order to reduce false-positive alerts, increase monitoring coverage and reduce the rapidly escalating financial cost of maintaining their AML programs.

The economic cost to perform BSA/AML compliance can no longer be ignored by firms, many of which are already struggling to make a profit in an ever-evolving competitive market. Not only are AML officers judged on their ability to react to these regulatory changes and quickly implement their solutions, they must also be accountable for BSA/AML program expenses. BSA officers are increasingly required to wear many different hats in an ever-changing environment, and it takes real leadership to balance and manage all of these expectations.

There are two real questions: Are firms learning how to effectively blend quantitative and qualitative transaction monitoring approaches in order to implement a risk-based program? Or, are they solely relying on either an all-qualitative or all-quantitative solution? These are important questions since banks within the industry often rely solely on one strategy or the other instead of an effective blend of both methods.

Segmentation Is the Logical First Step

So, where do banks start? They can begin with segmenting the customer base by analyzing customer activity and risk characteristics in order to monitor them more effectively. An effective AML transaction monitoring strategy begins with a sound foundation for monitoring the customer’s activities - a quality segmentation model provides just that foundation.

Segmentation is the process of grouping together customers and accounts that have similar characteristics and transactional behaviors with the objective of setting risk-based thresholds that are appropriate for each particular segment. Segments of customers and/or accounts may be grouped together based on one or more of their inherent characteristics, such as average transaction amount, average transactional volume, net worth, product usage, region, customer/account type, etc. Segmentation is the primary foundation for risk-based scenario threshold setting, and the quality of the segmentation model directly affects the transaction monitoring system’s ability to perform in an effective and efficient manner.

A quality segmentation model that segments homogeneous (similar) groups of customers and/or accounts together - and allows high-risk groups of customers and/or accounts to be monitored separately – provides the opportunity to set the appropriate threshold levels for the monitored segments. It also allows for enhanced monitoring of high-risk segments. In addition, having a quality segmentation model allows for scenario threshold values to be set in a way that provides
effective coverage throughout the customer and account populations. In fact, most banks that perform transaction monitoring without the use of a segmentation model (e.g., separating only personal and commercial customers) find that they have very poor alert coverage within their customer and account populations. As a result, the vast majority of alerts are generated for their largest and lowest-risk customer groups and few alerts are generated for smaller, high-risk customer groups.

The SAS Approach

The SAS approach to segmentation generally requires two primary projects to be undertaken:

1. Customer and/or account population segmentation.
2. Further refinement of individual segments into peer groups (only needed if anomaly detection will be performed).

In addition, SAS adheres to the guiding principles of OCC 2011-12 when developing, implementing and validating segmentation and peer group models, including the process of initial threshold setting.

Development of the Segmentation Model

The development of the segmentation model is further broken down into eight subtasks, where the first four tasks are often referred to as the business-driven components; the fifth task referred to as the data quality component; and the last three tasks referred to as the data-driven components. These eight subtasks are listed below:

- Decide on the model focus that needs to be developed: customer and/or account.
- Identify the key attributes and data items available for defining the segment groups.
- Ascertain the high-risk groups of customers and/or accounts that are believed by the bank to encompass the greatest level of risk.
- Develop a preliminary segmentation model based on the bank’s customer experience.
- Perform data quality/reasonability assessments.
- Analyze the preliminary business-driven segmentation model, and validate that it produces a fairly homogeneous population of customers and/or accounts.
- Refine the preliminary segmentation model and determine if the high-risk entity classifications should be segmented separately.
- Validate that the model produces homogeneous segments in terms of customer risk, attributes and activity that meet the bank’s key objectives.

Before the actual segmentation work can begin, make an initial decision on the focus of the segmentation models to be developed. The models can be at the customer level, account level or both depending on the needs of the bank. While SAS party-focused scenarios generally require a customer-level segmentation model, account-focused scenarios can be run using either a customer- or account-level model. In the case where an account-focused scenario is being run using a customer-level segmentation model, the segment containing the primary customer on the account is used for determining the account’s applicable segment for alert generation. Although developing separate models for account- and party-focused scenarios allows for more refined threshold setting, it also means that two models must be developed, maintained and periodically validated – all adding additional cost and complexity.

The first step in developing the segmentation model(s) is to identify potential risk attributes (data points) that should be considered when segmenting, including:

- Client type.
- Industry codes (e.g., NAICs, SICs).
- Income.
- Revenue.
- Account or product types.
- Geography/location.
- Transaction types.
- Channels (e.g., branch, ATM, online).
- Transaction volumes and/or amounts.
- Customer risk ratings.
- Other customer attributes.

The key to proper risk attribute selection is to conduct data discovery and clearly understand the data lineage, quality and sourcing. While numerous attributes may exist that potentially could be pertinent to the segmentation model development, poor data quality and inconsistent collection may prevent their use. As was articulated in the regulatory guidance, it is important to specifically call out any weaknesses or limitations in the data that will be used in analyzing the customer population and developing the segmentation model(s).

To begin the preliminary segmentation model development, start with the “business knowledge” of the customer population and historical alert experience. Then, further refine the model based on a combination of business- and data-driven analysis. This involves identifying the customer groups and associated characteristics that the bank feels add superfluous AML risk to the business. There are common entity types and geographical
regions that the FFIEC Manual and FinCEN have listed as potentially affecting AML risk. However, the true extent to which these risk factors affect a bank’s AML risk will depend on the customer base, key products and internal controls.

Developing the Initial Segmentation Model Framework

Once the model(s) focus has been determined, key data attributes selected and high-risk customer groups identified, it’s time to develop an initial segmentation model framework. The primary goal here is simply to form a high-level framework to start from when building out the various segments. During the Data-Driven phase a single segment containing large variability may be split into several smaller, more homogeneous groups; different segments found not to be overly heterogeneous may be combined; or additional segments may be added to accommodate customers containing similar previously unconsidered risk attributes. The main idea is that an initial starting point needs to be developed, however unrefined it may be, in order to focus the data-driven analysis that will follow.

Remember the old adage “garbage in, garbage out” before using the data collected to begin building the segmentation model(s). Explore the bank’s data and generate some basic metrics looking for: missing values, extreme values (outliers), inappropriate values, customers without accounts, and accounts without transactions or with excessive transactions. The results of this study will help the bank ensure that the data used to build the model is free from material defects. It is also a good practice to review the data summaries in order to apply a “smell test” as to whether or not they seem reasonable given the bank’s knowledge of the business. Also keep in mind that an assessment of the data quality used to develop a model is specifically mentioned in the regulatory guidance.

Once the bank has applied the business-driven components to begin their segmentation model development, the data-driven components come into play and data analytics becomes important in refining, testing and validating the segmentation model. At this stage, various analytical/statistical approaches are utilized to explore the data and determine the relationships between the variables, and identify key attributes that can be used to segment similar customers and accounts together.

Graphical approaches can be heavily utilized in order to dissect the population under review and understand the relationship between the various customer and account attributes. The advantage of using graphical techniques is that they allow the analyst to quickly get a general sense of the population distribution, correlation between variables and the variability (or spread) within the data. These graphical approaches include scatter plots, frequency plots, histograms, box plots, stacked bar charts, pie charts, heat maps, and so on. Graphical plots are also an effective way to identify outlying groups of customers and/or accounts. SAS offers several high-performance products for developing graphical visualizations, including SAS® Visual Analytics and SAS® Visual Statistics.

In addition to the use of graphical approaches, using common statistical methods and tests can further refine the segmentation model. Perform various clustering approaches to identify groups of entities with similar characteristics. While k-means is probably the most commonly used clustering technique within AML, there are certainly more advanced clustering algorithms that can be applied to provide banks with superior ways to identify separation within the population.

It is important to keep in mind the bank’s ultimate goal in building the segmentation model(s) – to allow for effective risk-based transaction monitoring by applying different threshold values to different groups of customers. Analysts, who commonly have a strong academic background, often get bogged down in trying to decide the “optimal” segmentation or clustering approach to use rather than focusing on the “true” end goal: to allow effective risk-based thresholds to be set and promote alert coverage within the customer/account population. Before spending lots of time implementing complex clustering algorithms or developing advanced quantitative segment assessments, analysts should ask themselves this question: “Will using this advanced approach allow for significantly greater risk-based transaction monitoring and will the regulators and management understand what I have done?” Keep in mind a fundamental rule in model building: All else being equal, a simpler model is preferred over a more complex model.
Also perform statistical tests in order to: (1) verify that segments of customers are in fact from different subpopulations; (2) identify outliers; and (3) validate normality of the population (or non-normality, which is more often the case.) SAS also recommends generating distributional metrics to be used in determining the similarity of the segments - including the skewness, kurtosis, coefficient of variation, mean, median and the empirical percentile distribution.

Now that the bank has done its analytical homework, it’s time to evaluate the data used to support the final segments within the model. There may still be a need to separate segments into smaller, more homogeneous segments. Here, there are three things to consider: (1) Does the segment have a wide range of total monthly transactional activity?; (2) Does the segment still comprise several independent statistical distributions?; and/or (3) Does the segment contain customers with widely varying degrees of inherent AML risk?

Once the bank has completed its analysis and developed the segmentation model that best fits the firm’s needs, there is no better time to put together comprehensive documentation (in accordance with the OCC’s 2011-12 Supervisory Guidance on Model Risk Management). The value of clearly documenting the entire segmentation process and providing a concise overview cannot be overstated. In the minds of many regulators, “If it wasn’t documented, it wasn’t performed.”

Do not overlook the effective challenge discussion in the OCC 2011-12 guidance. Take the time to ensure the bank has picked the best segments based on what is known (and has at least considered alternatives). But understand that this will not be the last time that the bank will evaluate its segmentation model. It needs to continuously monitor and conduct segmentation analytics to ensure that the model is still valid and that the underlying customer population or product offering hasn’t changed, causing a deterioration in the segmentation model over time. A segmentation model is a “model” as defined by the OCC and thus requires ongoing periodic validation just like any other model used by the bank.

Development of the Peer Group Model

The primary reason to develop a peer group model is to allow for anomaly detection at the peer group level. This allows anomalous activity to be monitored against both the individual customer’s historical transactional activity and against the customer’s peers (i.e., other customers that behave in a similar manner and/or contain a similar risk profile). The ultimate goal of monitoring activity at the peer group level is to identify customers who are expected to behave in a similar fashion as other customers (i.e., their peers) - but don’t.

Peer groups are essentially more refined segments, and thus the peer group model is developed after the segmentation model has been approved by the bank. Each of the segments has its customer population broken down into subsets based on the customer attribute(s) selected for differentiating between the various customers. (Note that different segments can use different attributes when defining peer groups). The attribute(s) can be used to separate the segment into peer groups either by rank order, classification or clustering techniques, depending on the data type(s) involved and the goal of the particular peer group model.

When developing peer groups it is important to balance the need for homogeneity within the groups with the number of customers contained in each group. Peer groups should never be comprised of less than 100 customers since this can result in having a statistically invalid population size from

![Figure 3. Illustrative Example of a Segmentation Model.](Image)
which to compare the individual customers against. The most common customer attributes considered for peer group development are:

- Transaction amount (average or total).
- Transaction count (average or total).
- Customer net worth (i.e., the sum of the account balance where the customer is listed as the primary).
- Industry type code (for commercial enterprises).
- Geography (in combination with another attribute).
- Transaction type (generally used in combination with either transaction amount or count).

However for all practical purposes, peer groups can be defined based on any set of customer attributes that allow similar customers to be grouped together.

For the rank-order allocation approach, the attribute for each customer is sorted from smallest to largest. Then the customers are separated into ‘N’ different equally sized groups based on the ordered attribute values. This approach is generally used when the attribute used to create the peer group involves either the transaction amount, transaction count or net worth amount.

With the classification approach, various distinct groups of attribute categories are created and customers are then assigned to peer groups that contain an attribute value within the group. This approach is generally used when the attribute used to create the peer group involves either the geography, industry code or a unique combination of attributes.

The cluster analysis allocation approach utilizes a clustering technique, such as k-means, to assign the customers to various peer groups based on one or more customer attributes. While it is possible to use this approach when considering only one attribute, it is more commonly used when multiple attributes are used. This method also makes creating relatively equally sized peer groups difficult to do and is heavily affected by outliers contained in the data.

### Conclusion

Enhanced regulatory pressure requires continuous evaluation of a bank’s risks. To meet these demands, the BSA/AML industry has turned to analytical/statistical methodologies to: improve monitoring programs in order to reduce false-positive alerts; increase monitoring coverage; and reduce the rapidly escalating financial cost of maintaining AML programs.

A well-designed segmentation model can significantly increase the AML monitoring coverage of the customer population at the bank while focusing the alert generation through risk-based thresholds on the customers that pose the most AML risk. This involves generating a greater proportion of alerts compared to their underlying population for some higher-risk segments, and a lower proportion of alerts as compared to their underlying population for less risky segments. For example, a bank that historically generated 80 percent of its personal banking alerts solely on the top 10 percent of customers based on average monthly transactional amount, may find that post-segmentation, only 5 percent of alerts are now being generated for that same personal banking customer population. The other 75 percent of the alerts that previously were being generated for those customers are now alerting on higher risk customers and customers well below the 90th percentile in terms of average monthly transactional amount.

Another benefit of a segmentation model integrated into transaction monitoring is that both the number of productive alerts detected and the number of productive alerts as a percent of total analyst reviewed alerts generally increases substantially. This is a direct result of enhancing the transaction monitoring coverage for the higher risk groups of customers and accounts.

By effectively blending both quantitative and qualitative methods, banks can monitor more effectively by segmenting the customer base and tuning the scenarios to identify the activity that poses the most risk to the bank.

For more information on high-performance products for developing graphical visualizations, visit sas.com/va and sas.com/visualstatistics.