Detecting and Dissuading Money Laundering Through Gambling

A Data Mining Approach

Master Thesis in Information Technology and Business Intelligence

CentraleSupélec

Specialization DSBI

Author: MIRONESCU Miruna Mihaela

Supervisor: BENNACER Nacéra

Advisor: QUERCINI Gianluca

WORSLEY Robert

CentraleSupélec

September 2015

Content

Introduction

Context

The research is being carried out at France International Business Machines also known as IBM France in the Business Analytics Department. The client is a company whose business is part of the betting industry.

Gambling industry in France has a very long history and some of the oldest and most popular gambling establishments located on the territory of this country date back to the 1800s. These establishments still heavily influence the gambling market today, but betting itself has been around since ancient times. France also contributed to the development of popular casino games and it has recently opened its doors to internet betting (French Casinos, 2015). The Law No 2010-476 of 12 May 2010 on the introduction of competition and sector regulation of gambling and online gambling is often referred to as the French Gambling Act. Betting is defined by the Gambling Act as the making or accepting of a bet on the outcome of a race, competition or other event or process; the likelihood of anything occurring or not occurring; or whether anything is or is not true.

Motivation

Revenues of companies whose target domain is gambling have soared in the past years. Over a time span of 18 years, these companies' gains have increased by over 265% from 16.7 million euros in 1995 to nearly 44.3 million euros at the end of 2013 (Influence du TRJ sur le blanchiment, 2012). The wide range of bets has given room to numerous fraudful transactions. In the current thesis, a transaction is represented by any bet related operation (e.g. the act of betting, the payment of a winning bet, the cancellation of a bet etc.).

With almost 200 casinos in France grouped in approximately 15 chains, the largest chain is comprised of 44 casinos (Casino en France). Other major gambling organizations include 17 online legal betting companies and numerous traditional horse-racing betting events. More than half of all European racetracks are built in France, and they are also some of the largest in the world. With over 239 racecourses spread all over France, horse races occur daily with an average of 8 race track meetings on the card every day of the week, car racing like "The 24 Hours of Le Mans" (in French: *24 Heures du Man)* is the world's oldest active sports car race in endurance racing, held annually since 1923 near the town of Le Mans, France and it is considered to be one of the most prestigious automobile races in the world (24 Hours le Le Mans). With other betting sports, both seasonal and regular, France is now one major betting market (BetMinded).

This large diversity of betting opportunities, including chance games, skill games, online games and many others (Arjel), has led to the invention of ingenious fraud methods, especially money laundering. Evolving from basic techniques such as claiming to have won money from a poker game to using stolen credit cards on the online gaming websites, thieves have become extremely clever in finding the loopholes in the gaming industry.

Scope

The scope of anti-fraud data mining is to protect businesses such as credit card, commerce and e-commerce, insurances, telecommunication industries and retails. In this thesis, the ultimate scope is detecting and dissuading money laundering in the commerce domain. Regulators have dramatically stepped up enforcement

of anti-money laundering whose consequences can be seen in the national economy (which concerns the economical balance) justice system (with respect to corruption) and especially level of terrorism (since money laundering is actually one of the key financial sources of terrorism).

In France, all banks and many other institutions have the obligation to report suspicious transactions to TRACFIN (Les Cles de la Banque), also known as "Traitement du renseignement et action contre les circuits financiers clandestins", which is a service provided by the Ministry of Economics, whose aim is to fight against money laundering and terrorism funding. Its scope is to gather, analyze and if needed further investigate all concern declarations regarding transactions. This institution then decides whether to further investigate or not the suspicious transaction. In the current case, this research project was commissioned by TRACFIN after numerous anomalistic transactions reported in the gambling industry.

Objectives

- Defining existing challenges in this domain (e.g. Big Data, Money Laundering Indicators and Data Quality Issues). This will be done with respect to different types of large datasets and streams
- Making use of existing anti-fraud methods (from both academia and industry). It is believed that experience and knowledge from the previous studies can prove to be reusable and thus it will help prevent the repetition of common mistakes and the "reinvention of the wheel".
- Defining an approach to detect potential money laundering in case of gambling venues but also gamblers, since money laundering is connected to both gambling venues and gamblers.

Other secondary objectives include:

- preparing data for investigation
- studying different types of fraud i.e. money laundering
- analyzing anomalous historical transaction data
- validating its quality with respect to necessary standards
- creating statistical indicators using the data mentioned above
- developing a scoring methodology to measure the level of risk of money laundering
- taking care of the development and follow up
- automating the process as much as possible
- identifying further research

Structure of the document

In section 1, the scientific background is presented, more precisely basic concepts, related work with characterization of current approaches and detected opportunities and challenges. Section 2 describes how the data was prepared and presents the system architecture. Section 3 is the technical part which contains information about the applied approach. Section 4 describes an additional analysis made to stabilize the results. Section 5 concludes the presentation with a brief discussion on future research.

1 Scientific background

The scope of data mining is to discover unknown insights from the data. In order to do that, the first step is to have relevant and clean data. It must also be mentioned that data mining problems must be well-defined and cannot be solved through traditional means such as SQL queries or reporting tools. That is because these traditional tools are limited as it will further be explained.

Data mining is the computational process of discovering patterns in large data sets. It has emerged as a result of recent tremendous technical advances in storage capacity, processing power and inter-connectivity of computer technology which created and it is still creating an unprecedented quantity of digital data. Data Mining, which is also known as Knowledge Discovery in Data (abbreviated KDD) is hence a young and interdisciplinary field in computer science.

Another important step is to define the term “money laundering”. It can be defined as the process of transforming the proceeds of crime into ostensibly legitimate money or other assets.

The research will be conducted on the client’s data, which is mainly composed of bet transaction history (e.g. large and small gains, beneficiaries, timestamp etc.), gambling venues (e.g. inspection report, profile of the unit, staff etc.) as well other additional information. In the context of the research, INSEE data will also be used. INSEE data comprises statistical information from a variety of fields w.r.t. IRIS. The IRIS can be defined as a territorial map “cut-out” in homogenous sizes which targets about 2 000 inhabitants per cut-out. It has to respect geographic criteria but also demographic ones and it also must have identifiable (non-ambiguous) edges which are stable in time. The IRIS concept will be used later on in the research to determine whether there is a correlation between the environment of a gambling venue and the predicted money laundering risk.

In terms of technical background, clustering methods are used (among others) in the development of the anti-money laundering solution. **Two Step**, for example, is one of the clustering methods used in the approach .It takes place in two steps. The first one consists in a data exploration in which the brute input data is compressed into sub-clusters in order to make them easier to manipulate. In the second step, the utilization of a hierarchical (agglomerative) classification method allows the progressive merging of sub-clusters into bigger and bigger clusters without the need of a new data exploration (IBM Internal Documentation). Another clustering algorithm which we used is **K-means**. It starts by defining a set of cluster centers which are taken from the data. It assigns each record to a cluster to which the record is most related to. Once all the records have been assigned, the centers of the clusters are updated to reflect better the set of records within the clusters. The records are re-evaluated in order to determine whether they need to be assigned to other clusters. The process is repeated until the maximal number of iterations is reached or the induced change is below a certain threshold (Jain Anil K., Dubes Richard C., 1988).

Other useful algorithm which is utilized is **CHAID**. It creates a decision tree with the help of the chi-squared measure used to identify the most optimal division. The first step is to create categorical predictors (i.e. independent variables on which the prediction is made) out of any continuous predictors by dividing the respective continuous distributions into a number of categories with an approximately equal number of observations. The next step is to cycle through the predictors to determine for each predictor the pair of (predictor) categories that is least significantly different with respect to the dependent variable (i.e. the

prediction). Afterwards, the algorithm will chose to split the predictor which yields the most significant split. The algorithm continues until no more splits can be made due to the presence of a threshold (Basic Tree-Building Algorithm: CHAID and Exhaustive CHAID, 2015). This method will be used at a certain point to determine the most influential variables in the money laundering process.

The approach also makes use of the **Selection Algorithm Function**. One of the exploratory data problems is found in the fact that hundreds, maybe thousands fields can serve as entry fields. The selection algorithm function (which allows the identification of the most important fields) can help reduce the number of choices. The algorithm follows 3 steps: Filter which reduces the number of fields, Classification which classifies the remaining entries and Selection which identifies the sub category of functions to be used in the following models and it preserves only the most significant entries (IBM). At a certain point, this algorithm will prove useful when selecting the entry fields to be used in the algorithm.

Related work

This section's scope is to compare, categorize and summarize automated fraud detection considering almost all the related published articles and reviews. According to a research article (Clifton Phua, 2010), the common data mining approaches when dealing with fraud are: **single supervised algorithms**, **supervised hybrid algorithms**, **supervised/unsupervised hybrids**, **semi-supervised algorithm**. A paragraph will be devoted to each of these approaches.

In case of **single supervised algorithms,** the scope is to examine all labeled data (e.g. transactions) in order to assign a risk score. This is of course a mathematical approach. A way of doing this is to use hard-coded rules so that each transaction would meet specific criteria such as matching addresses, phone numbers, price, amount limits etc. (Sherman, 2002). A different approach consists in using a three-layer Neural Network with only two training passes to produce a fraud score every two hours (Ghosh, 1994). Specific attributes are not revealed but they should comprise a timestamp, transaction amount, geographical location, merchant industry and other account information. The datasets contain more than 1 million credit card transactions. The results are reflected in increased precision.

For the **supervised hybrid approaches** (e.g. **multiple supervised algorithms)**, some popular supervised methods (e.g. Bayesian networks, decision trees) were combined in sequential fashion for the sole purpose of improving the results. For example there is an approach that uses naïve Bayes, C4.5, CART and RIPPER as base classifiers and stacks them to combine them (Chan, 1999). The results prove to be very good: high cost savings and improved efficiency on credit card transactions. Another approach (Phua, 2004) uses neural networks, naïve Bayes and C4.5 as base classifiers on data partitions derived from minority oversampling. In the end, it results in cost saving in automobile insurance claims.

Supervised/unsupervised hybrids (i.e. a **combination of supervised and unsupervised algorithms)** have been used a lot in the telecommunication domain. Attributes used in this case are usually timestamp, call duration, type of call, geographical origin, source number etc. An example of supervised/unsupervised approach (Cortes, 2003) consists in the use of signatures (i.e. telecommunication account summary) which update daily (i.e. time-driven). Fraudulent signatures are added to the training data set then they are processed by supervised algorithms. The results are that the authors remark extensive late night activity (and other characteristics) in

case of the fraudulent phone calls. By using association rules, the authors are also able to determine interesting country combinations. In another experiment (Cahill, 2002), average suspicion scores are assigned to each call based on their similarity to known fraudulent examples / dissimilarity of it to known legal (i.e. non fraud) examples (in terms of average time between calls, call cost etc.). The result is an increase in fraud precision.

For the **Semi-Supervised Approaches** with only Legal Data, one approach (Kim, 2003) managed to implement a novel fraud detection method that is meant to discover rules which detect anomalies and replicate them by adding tiny random mutations. In the end, the method provides a comprehensive representation of customer behavior and it is able to deal with the dynamic nature of telecommunication fraud. Another methodology (Murad, 1999) uses daily profiling from each telecommunication account which is extracted using a clustering algorithm. An alert is raised if the daily profile's call duration, destination and quantity exceed the threshold and standard deviation of the overall profile. The results show a superiority of this method over rule-based method.

The analyst can also benefit from using SNA (**Social Network Analysis**) in order to prevent money laundering. In the context of banking, the gain of analyzing fraudster network is indisputably relevant (CGI, 2011). The idea is to identify the fraudulent relationship between people as clusters. Density measures were applied to these clusters in order to see which clusters were at most risks. The identified clusters were subject to centrality measure to identify key actors within each cluster. The result is that future clients that fall into the patterns described by the clusters are rejected prior to becoming a client of the bank.

Opportunities and challenges detected

It is impossible to be 100% certain about the legitimacy of a transaction, regardless of its intention. A feasible solution is to use possible fraud evidence from the available ready data using mathematical algorithms.

The first most important challenge concerns the creation, selection and grouping of **Money Laundering Indicators.** These are general indicators intended to help financial and non-financial professionals identify money laundering transactions. Criminals use deposits in cash or in the form of checks to launder the proceeds of crime. The size of the deposits, the fact that cash is deposited at different branches of the same bank, the lack of credibility of the economic explanation for the transactions are only some of the factors that financial institutions need to pay attention to. Gambling venues need to pay a particular attention when a client purchases for an amount disproportionate to his known financial status or when a client has a playing behavior that does not correspond to that of the normal gambler and the aim to win is apparently absent or secondary. Hence a series of indicators were created in order discover these suspicious transactions. These indicators are presented in this thesis, classified by axis: gambling venue axes or player axes.

The second most important challenge is dealing with **Big Data**. Taking into consideration the complex nature of the data and its constant growing size, the real challenge here is, of course, extracting value from the data. The main idea is to use Predictive Analytics to extract patterns from the data. In this case, the scope is to be able to distinguish a normal player behavior and the usual gambling venue gambling activity from anomalous (potentially fraudful) actions. After using predictive models to exploit the patterns in the historical

transactional data, the client will be better able to identify risks and opportunities, hence an improved decision making process. That is because the defining functional effect of these technical approaches is to provide a predictive score (i.e. probability for which each individual and/or venue is assigned a degree of risk). This part is represented by the last phase of the project which is called “simple scoring”.

A third challenge consists in dealing with **Data Quality Issues**. Data is impacted by numerous processes, most of which affect its quality to a certain degree. Data is deemed of high quality if it correctly represents the real world construct to which it refers. The problem is that when data tends to be rapidly growing in size, the question of reliability arises. This is usually measured in completeness (all relevant data e.g. accounts, addresses and relationships for a given customer are linked) accuracy (common data problems e.g. misspellings, typos, random abbreviations etc.) availability **(**the required data must be accessible on demand) and timely **(**up-to-date information is readily available to support decisions).

A forth challenge is deciding how to use the client’s **historical data**. A close investigation must be conducted to decide which attributes to be used. This is mainly because data structure was not consistent over the years (e.g. many database fields, basic units of data entry in a record bearing a name and representing a type of data, which are now very important for the research, were added in recent years, hence scarce data). The time frame that will be used must also be established with respect to the client’s overall evolution (amount of revenues, number of clients etc.)

Among other challenges in this project there is the geographical analysis of the anomalous transactions by means of **INSEE data** as well as the determination of the correlation report between predicted risk score and real risk score.

2 Description of the Data

Tools

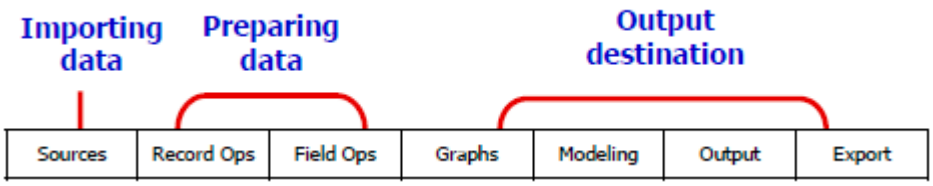
The tools that will be used in this approach are mainly IBM SPSS, IBM I2 and DB2.

IBM SPSS modeler is a data mining and text analytics software application built by IBM. The software was released in its first version in 1968 as the Statistical Package for the Social Sciences (SPSS) after being developed by Norman H. Nie, Dale H. Bent, and C. Hadlai Hull. It is used to build predictive models and conduct other analytic tasks. It has a visual interface which allows users to leverage statistical and data mining algorithms without programming. A snapshot of the interface can be seen in Figure 1.

Figure 1 IBM SPSS Interface

The functions that are used are reflected by nodes which are grouped in palettes. Each palette contains a related group of nodes. The palettes follow the order of the stages in analysis. In the Sources palette, the analyst can choose the nodes to import data from various sources. Two data preparation palettes organize record operations (such as selecting records) and fields operations (such as deriving new fields). Four output destination palettes complete the Sources*, Records Ops* and Field *Ops palettes*. The output destination palettes contain nodes for various analysis tasks, such as graphical displays (the Graphs palette), data-mining models (the Modeling palette), reports (the Output palette) and nodes to export your data (the Export palette). Nodes from the output destination palettes appear at the end of a stream and thus are called terminal nodes. No other node can have its input from a terminal node.

Figure 2 IBM SPSS palettes



IBM I2 is a software tool meant to explore data e.g. see people network, investigate historical data, connect different types of information etc. It provides an extensible, service-oriented environment designed to integrate into the existing enterprise infrastructure. The platform helps facilitate and support operational analysis, improving situational awareness and providing faster, more informed decision making across and inside organizations. IBM i2 solutions help law enforcement, national security, defense and commercial organizations detect, investigate and combat criminal and terrorist activity. In the current project, its scope is to explore data in order to see people network, investigate historical data, connect different types of information etc. all for fraud detection purpose. IBM I2 interface can be seen in Appendix C.

The infrastructure is supported by DB2, which is an IBM Database Solution. IBM DB2 is a family of database server products developed by IBM. These products all support the relational model, but in recent years some products have been extended to support object-relational features and non-relational structures, in particular XML.

Description of Data

In order to start working with the data, we need to gather all the necessary data, to verify the quality of it with respect to necessary imposed standards, derive the necessary attributes which will later be used in the experiments and integrate the data. The steps that need to be followed in this setting are therefore Data Selection, Data Cleaning, Data Construction and Data Formatting and Dataset Combination. Additionally, we can identify the unit of analysis. These tasks are critical but nevertheless critical for the success of a data-mining project, especially in the experimentation phase.

The data that was used consists of several data sources, which are detailed in Table 1. For instance, the data source Bets contains the bet transaction history with its associated information (e.g. stakes, opening bet date, closing end date, match cancelled, odds at opening date, odds at closing date, type of sport bet etc.). Other important datasets are Large Gains and Small Gains datasets which contain highly relevant information regarding the gain evolution of a beneficiary. The dataset describing the profile of gambling venues will prove useful in the gambling venue analysis. To have a complete view of the database, refer to Appendix A.

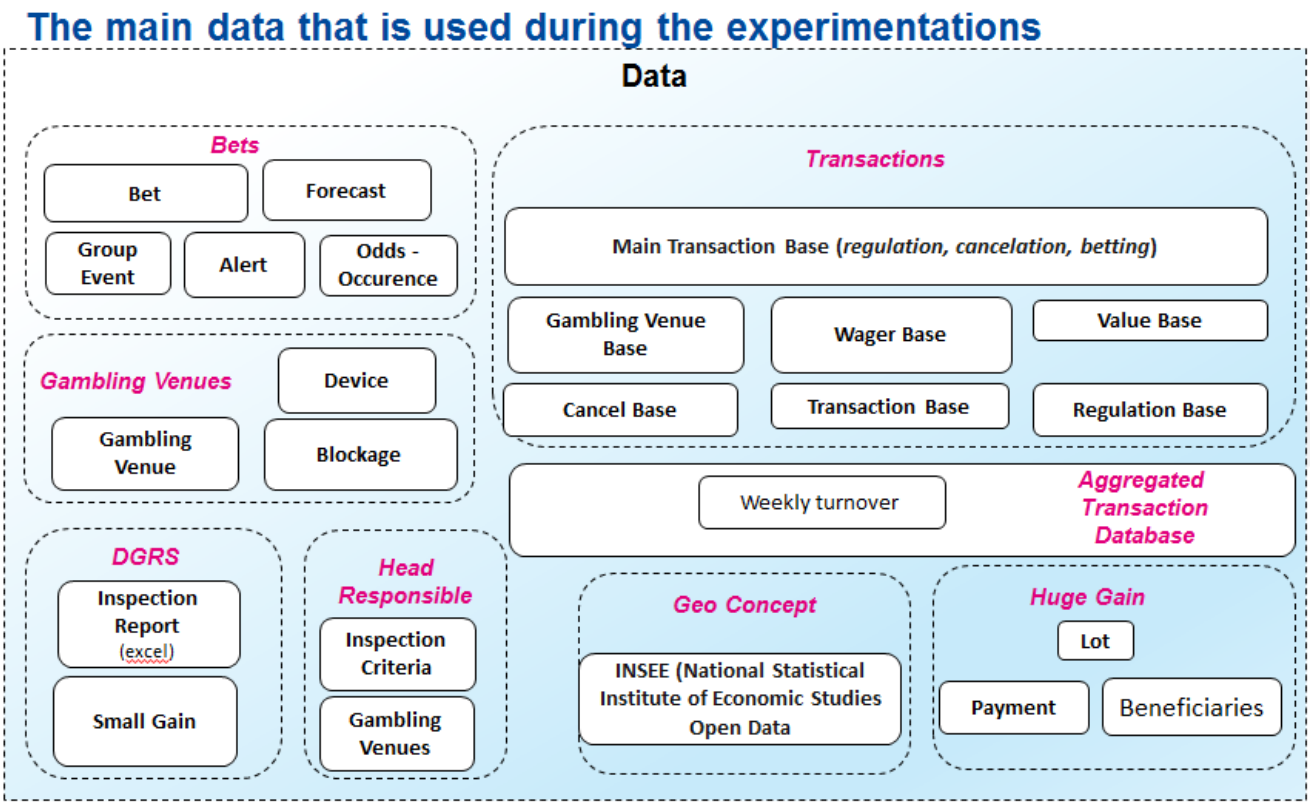
Table 1 Database

Data Sources	Description	Number of lines per year	Use
Bets	Administrative tool for different types of bets (e.g. stakes, opening bet date, odds etc.) and alerts	23 million lines/year	Yes
Large Gains	Information on beneficiaries of a check or bank transaction whose earnings are bigger or equal to 300 euros	2 million lines/year	Yes
Alarm Database	Alarm manipulation (available information on alarms according to gambling venue, bet, forecast, dates,	1 million	Yes

	number of occurrences, global amount of invested money per player etc.)	lines/year	
Gambling Venue	Information about gambling venues (type of contract, rights, opening date, data about manager, annual taxes and revenues etc.)	2 million lines	Yes
Transactions	Transaction History (transaction date, transaction series, amount, beneficiary, payment date etc.)	4 million lines /year	Yes
Inspection Report	Inspection history of the different gambling venues (information about venues where suspicious transaction have taken place, but also regular normal transactions)	1 million lines/year	No
Small Gains	Information on beneficiaries of a check, bank transaction or direct payment whose earnings are smaller than 300 euros	2.5 million lines/year	Yes
Head Responsible	Information about the person responsible of the gambling venue (e.g. retailer, manager, supervisor etc.)	1.5 million lines	No
Geo Concept	INSEE open data which reveal geographical information about the environment of the gambling venue	1 million lines /year	No
Aggregated Transaction Database	Data aggregation made by the company	1 million lines /year	Yes
Risk	Data file about the gambling venues (responsible) at risk	1 million lines /year	Yes

The data that we mostly use in the context of this thesis can be succinctly seen in Figure 3:

Figure 3 Chosen solution and used data



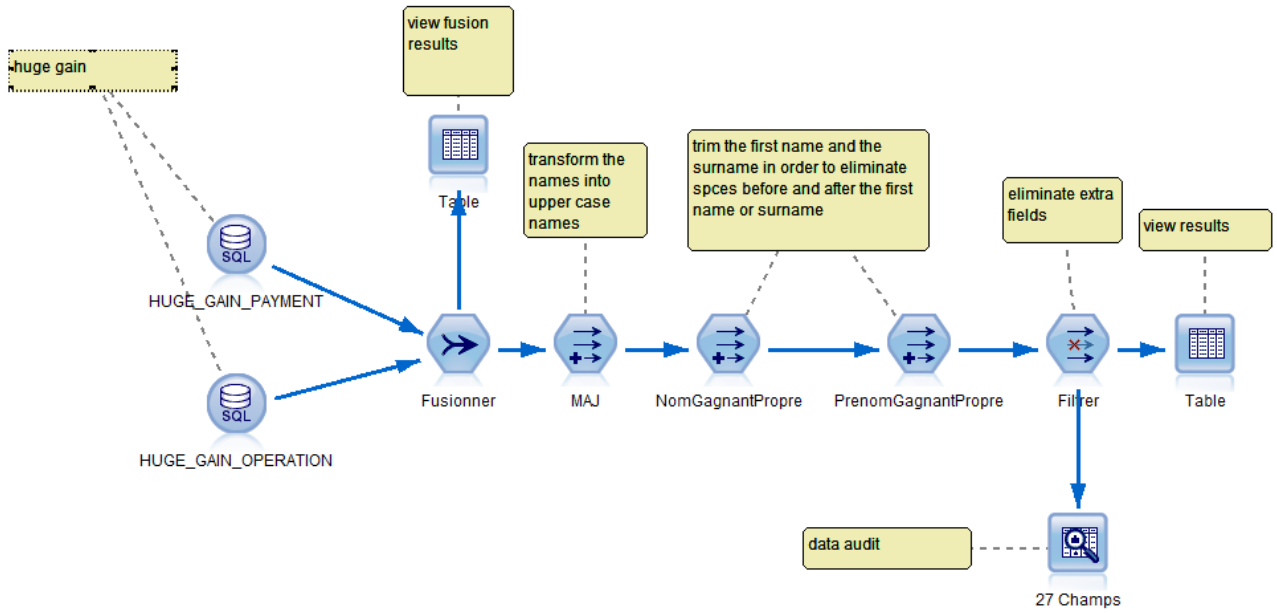
Data Preparation

The most common data quality issues include duplicates, inconsistencies, null values, wrong format, misspellings and unnecessary data fields. The tool that we used to deal with these issues is IBM SPSS.

In the construction of the flows, different nodes are used to deal with these problems, including deduplication nodes, aggregation nodes, renaming nodes, reclassification nodes etc.

In this case, the unit of analysis is bets, since the main target of the project is to find customers that do money laundering.

Figure 4 Huge Gain



In order to give a better sense of how we prepared the data, we show an example in Figure 4. In order to use the data contained in the data source Large Gains, we combine the tables Huge_Gain_Payment and Huge_Gain_Operation. The first step is to join the two tables using an inner join on the key fields Name and Surname. Secondly, for future aggregation purposes, all surnames and names are transformed to upper case strings. The two newly created fields are called “PrenomGagnantPropre” and “NomGagnantPropre”. Thirdly, the trim function is used in order to eliminate spaces before and after the fields “PrenomGagnantPropre” and “NomGagnantPropre”. Finally, the results are outputted into a table. The data audit of these results is displayed in Figure 5.

Figure 5 Data Audit Huge Gains before Cleaning

Field	Measure	Min	Max	Average	St Deviation	Unique	Valid
Amount	Continue	-1	371056	159,63	48451,63		1127014
Last Name	Categorial						1126965
First Name	Categorial						1126947
City	Categorial						89012
Department	Categorial					103	125458
Birth Date	Continue	01/01/1900	31/12/1999				146773
Telephone Number	Categorial						684667
Payment Date	Continue	19/03/2010	22/03/2015				3126305

In Figure 5, the problem of **missing data** can be seen, judging by the column “Valid” which stands for the number of values in the database. It can be deduced that, if all the data was present in the table, the column Valid would have the same (maximal) value for all the columns (e.g. First Name, Last Name which have different count values). Also, looking at “Department” and “City”, the user can see that there are more departments than cities. Considering the fact that a department contains multiple cities hence there could be

at most one city per department, the user can easily deduce that the number of cities should always be greater or at least equal to the number of departments. Also, **erroneous data** can be spotted. More precisely, the number of unique departments is 103, as shown by Figure 5. Considering the fact that France has 96 departments in the metropolitan area and 5 overseas departments, that means the maximum number of departments that can appear in the database is actually 101 (96 metropolitan departments plus 5 overseas departments) and not 103.

Figure 6 Cleaning Data - In Depth Analysis Excel File

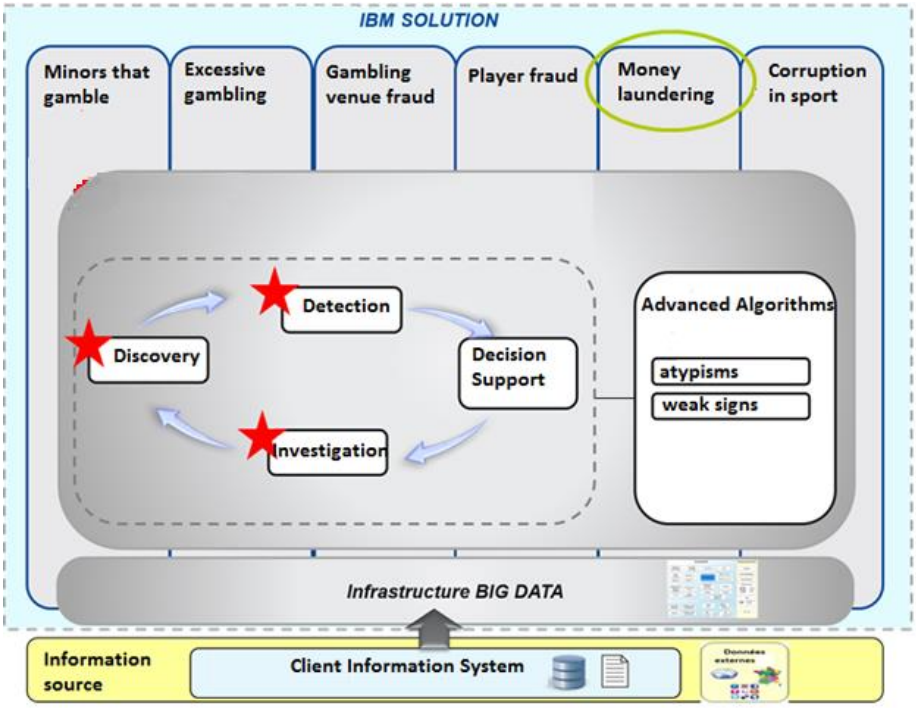
Field	Outliers	*Percentage of non null, non blanks, non empty string data	Number of of non null, non blanks or non empty string data	Null Values	Empty Strings	Blank Values	Non Loaded Values
Amount	756	78%	912268	10230	679	23430	0
Last Name	0	72%	822684	15673	830	243	0
First Name	0	68%	777593	13654	765	230	0
City	0	82%	74011	967	304	101	0
Department	0	57%	72458	1090	251	151	0

Figure 6 presents more detailed statistics about Figure 5; there are 756 outlier values for the field Amount. This measure (i.e. outliers) only makes sense for continuous data; hence they are not defined for categorical data. Furthermore, out of all available data, only 77% is "useful" data meaning data without nulls (i.e. no value), blanks values or empty strings (i.e. ""). Although sometimes null values, blanks and empty strings can also be good values (null values for example may render non applicability of the field). One important remark is the fact that City appears to contain more "useful" data than department in terms of null and blanks (but not for empty string). Hence, the analyst can take into consideration the possibility of deriving the department from the city. As a general observation, all values were correctly loaded, as can be seen in the last column named "Non Loaded Values". Another observation is the fact that null values represent the biggest problem in the data base, since in most cases, it represent the majority of value problems. It must be noted that only the data of 2013 was present and no data from 2014 was introduced. Another example of data cleaning can be seen in Appendix D.

System Architecture

Given the nature of the problem and the available data, the following architecture in Figure 7 was decided upon:

Figure 7 General Architecture



As Figure 7 shows, the procedure to analyze money laundering relies on an iterative approach. The approach is divided into phases comprised of 4 steps. The steps in each phase are the following:

- *Discovery*, discovering new patterns or rules (in this case, indicators which will be seen later on) to detect money laudering
- *Detection*, applying these rules
- *Decision Support* , deciding, based on the evidence, whether the money laudering cases will be further investigated or not
- *Investigation*, the concrete investigation of the selected money laudering cases

The scope is to discover new money laundering patterns at the end of each phase. This will lead to more and more complex money laundering profiles each time for both gambling venues but also gamblers. The phases which incorporate these steps are *Simple Profiling*, *Stake Analysis*, *Series Analysis*, *Complex Segmentation Profiling* and *Scoring Alerts* as can be seen from Figure 8.

Figure 8 Organization of phases - Detection/Discovery steps

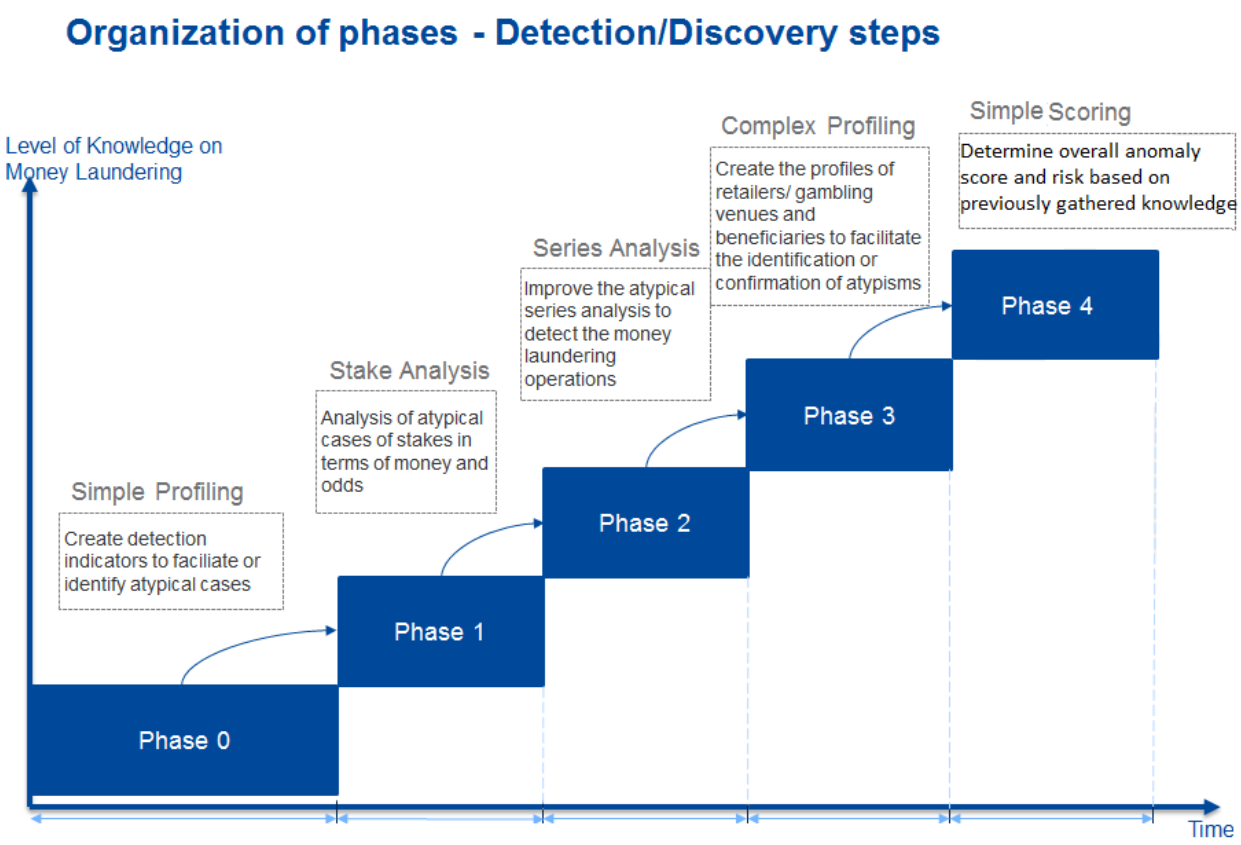
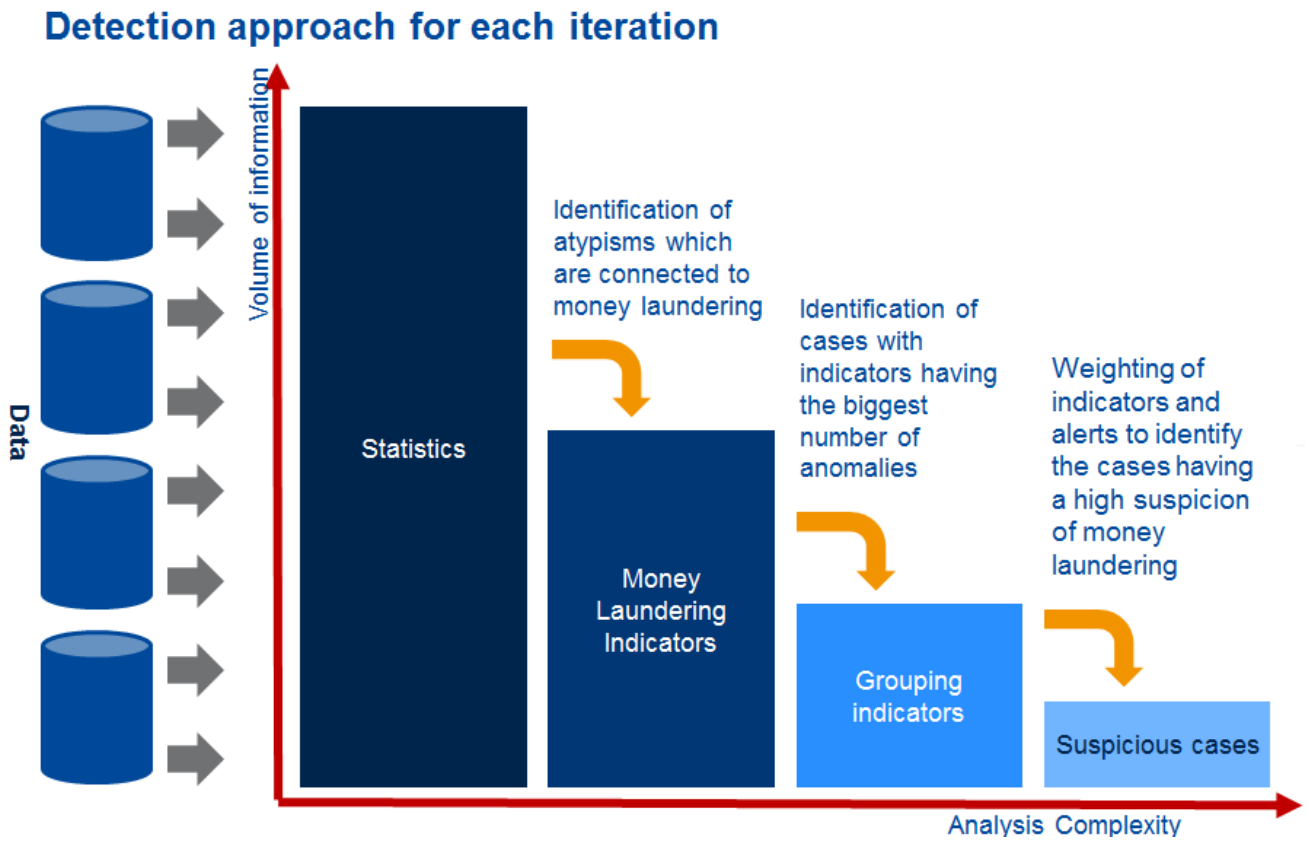


Figure 8 shows the sequence of phases as well as a brief description of them. The main idea is to better understand the concept of anomaly and use it to determine new ways/rules/patterns to detect more money laundering anomalies at each phase. In this case, the rules will take the shape of indicators, which are meant to “capture” these anomalies. The analysis is targeted at gambling venues and players. In the last phase, the scope is to create an overall anomaly score for both gambling venues and players based on all the previously gathered knowledge. For a more in depth description, refer to Appendix B.

Figure 9 Detection Approach for each step



In Figure 9, the detection approach is explained in detail. The idea behind it is to first use statistics in order to identify atypical cases (connected to money laundering). Secondly, derive money laundering indicators from the relevant statistics. The grouping of these indicators will give a fairly good idea about the target atypical cases which will be identified by their high anomalous scores. By weighting of indicators, the cases with high suspicion will obviously attract the highest interest.

3 Data Analysis

Phase 0: Simple Profiling

In this subsection, a list of indicators was created in order to create simple profiles, as described in System Architecture. These indicators must take into consideration both the beneficiary and gambling venue. Moreover, the indicators must be constructed with respect to time and location which forms the two main axes. Alternatively, the type of bet can also be seen as an axis. The granularity of the axes range from specific to general as can be seen from Table 2.

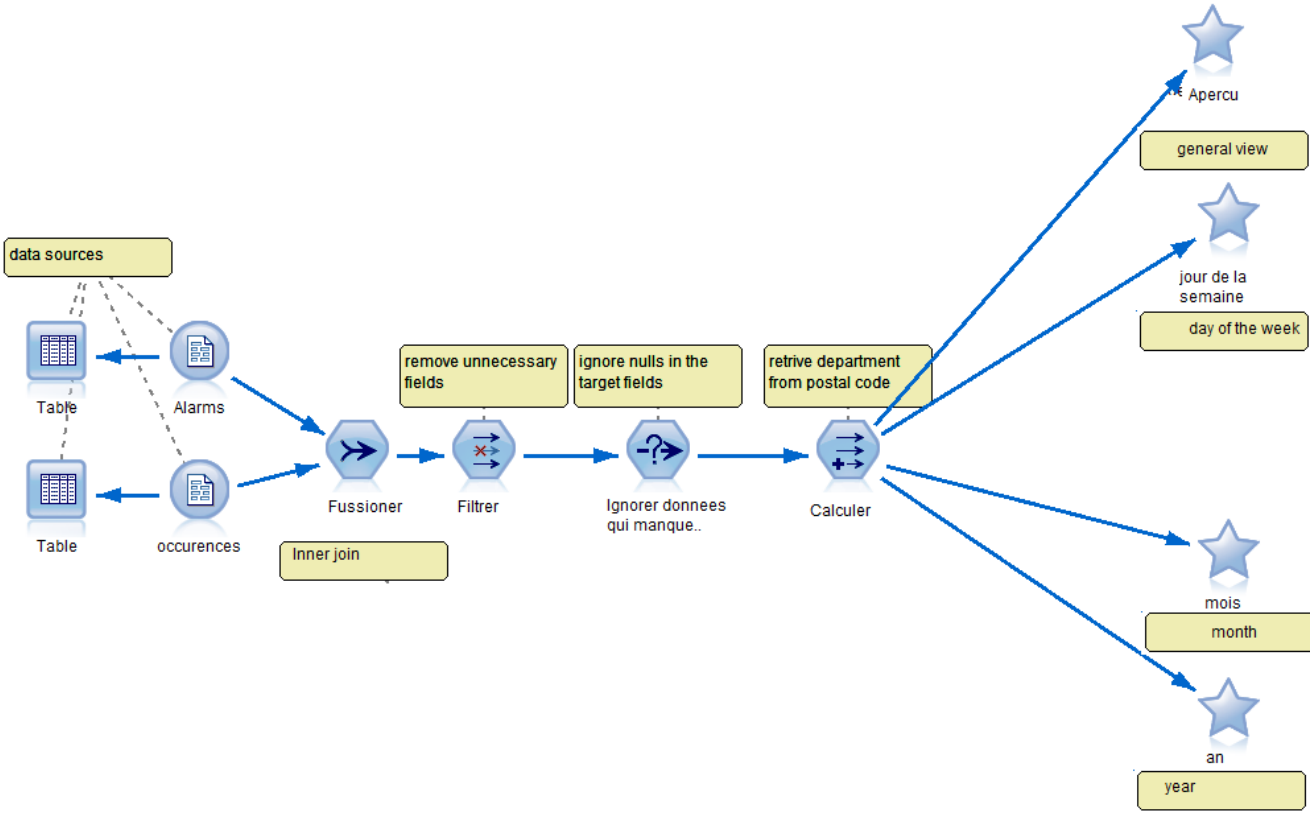
Table 2 Indicators: Phase 0

Indicator Index	Analysis Domain	Indicator and Axes
001b	Beneficiary	Nb of huge gain beneficiaries - by day of the week/month/year - by gambling venue /postal code/ department
013	Beneficiary	Total amount of small gains (i.e. gains smaller than 300 euros) - by day of the week/month/year - by gambling venue/postal code/department
027	Beneficiary	Number of small gains - by day of the week/month/year - by gambling venue/postal code/department
030	Gambling venue	Number of alarms - by day of the week/month/year - by gambling venue/postal code/department
060	Beneficiary	Number of units (i.e. in a small gain) - by day of the week/month/year - by gambling venue/postal code/department
062	Gambling venue	Number of occurrences - by day of the week/month/year - by gambling venue/postal

		code/department

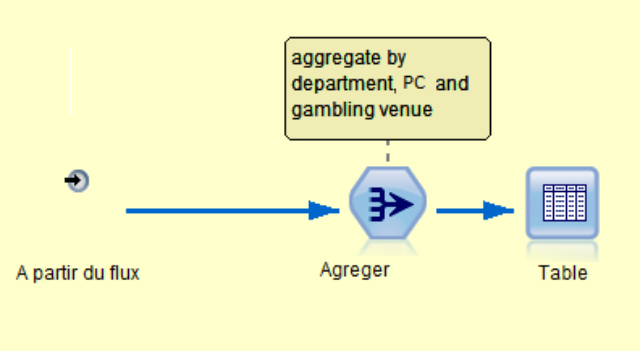
The time axes granularity ranges from day of the week to month and year. The day of the week must be taken into consideration in order to see the distribution of bets in a normal week. Also, it would be interesting to investigate which games players prefer for money laundering. The geographical aspect must also be taken into consideration (i.e. gambling venue, postal code, and department). An example of indicator implementation can be seen in the Figure 10 .

Figure 10 Overlook I030 Time Scale



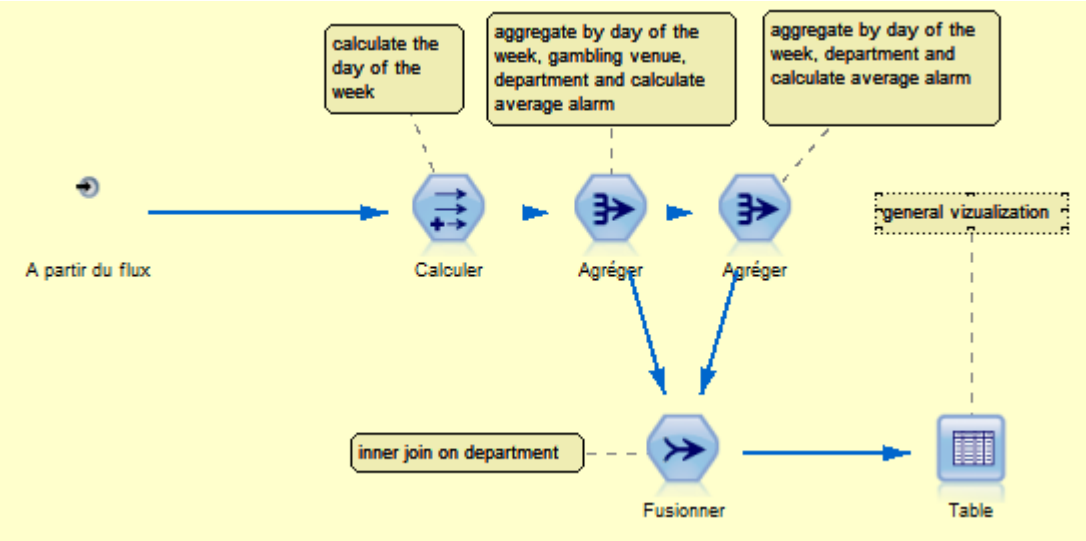
After loading the necessary data, the first step is to do mandatory data preparations (e.g. ignore nulls, eliminate extra fields, deal with blanks, outliers etc.). In the second step, the analyst calculates required fields (i.e. department from postal code) and the third step is the actual analysis with the general view, day of week, month and year analysis, which can be seen in Figure 11, Figure 12 and Figure 15.

Figure 11 I030 Time Scale - General View



For the General View, the sole requirement is to have the data aggregated by department, postal code and gambling venue.

Figure 12 I030 Time Scale - Day of Week Analysis



In the Day of the week analysis as well as the Month Analysis, the analyst first calculates the day of week and month from the timestamp. The idea is to calculate the number of alarms of the gambling venue and later compare it with the norm number of alarms of gambling venues in the same department, in order to check the existance of an anomaly as can be seen in Figure 13.

Figure 13 represents the general anomaly detection condition applied to all indicators (The right hand side of the formula is the anomaly threshold and the left hand side is the sought measure). In this case, the indicator is the number of alarms and the measure is the exact number of alarms.

The result is later outputted into a table.

Figure 13 Anomaly Detection Condition

*If Calculated Measure > Average of Calculated Measure + 2*Standard Deviation of Calculated Measure*

Then anomaly = true

In the case of Year Analysis, a different approach to construct the anomaly score is used, as can be seen in Figure 14.

Figure 14 Year Anomaly Score

If year = 2014

*If Calculated Measure > Average of Calculated Measure + 2*Standard Deviation of Calculated Measure*

Then anomaly = 2

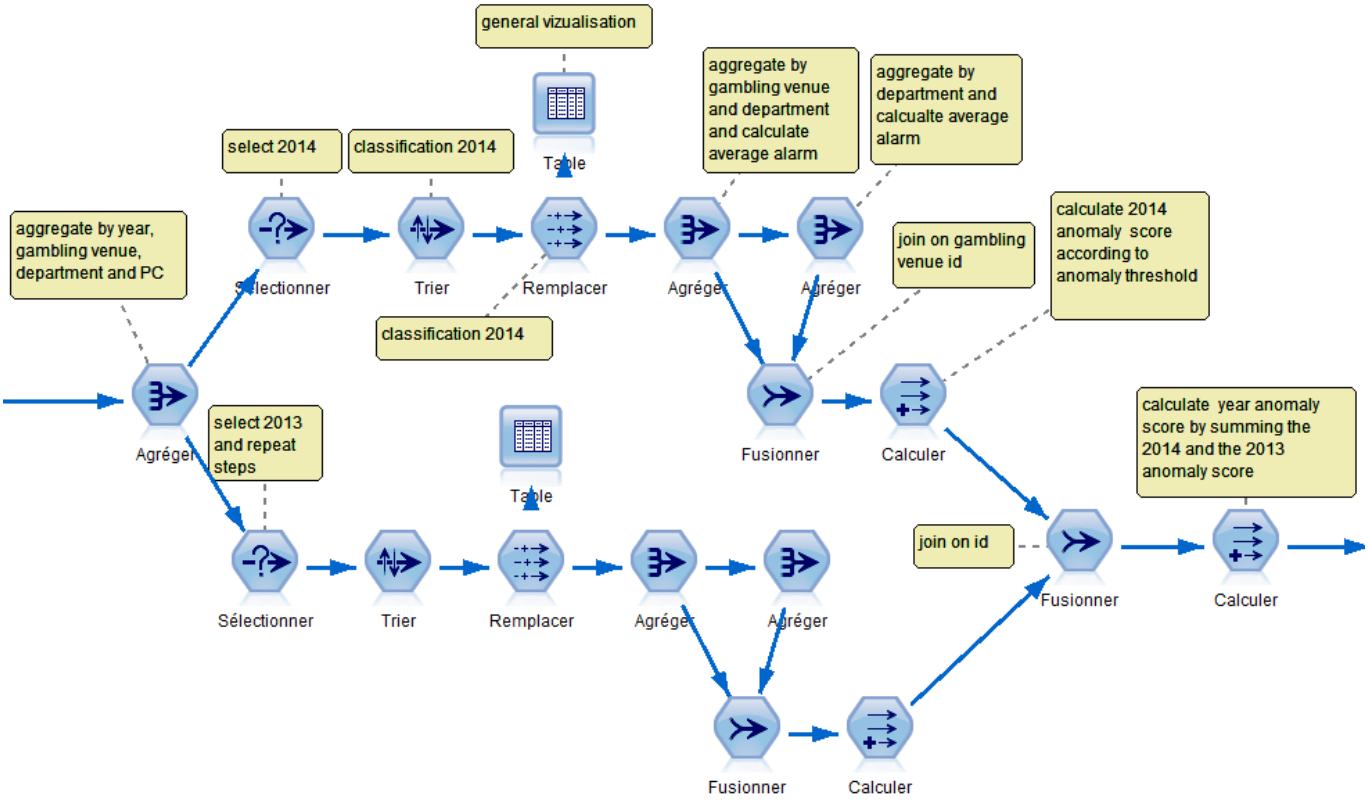
If year = 2013

*If Calculated Measure > Average of Calculated Measure + 2*Standard Deviation of Calculated Measure*

Then anomaly = 1

Final Score = 2014 Score + 2013 Score

Figure 15 I030 Time Scale - Year Analysis



The Year Analysis is based on multiple year analysis whose range is predefined (i.e. 2013 and 2014). After aggregating on the required fields (e.g. gambling venue, department, year), the year is selected from the year range (i.e. 2014, 2013), the data is ordered and indexed in order to obtain partial results (i.e. number of alarms classification 2014). The results are then joined by id and the anomaly score, as can be seen in Figure 15, is assigned.

In the case of year anomaly score, the scope is to use a weighted score sum to give more importance to recent years rather than distributing the weights equally among the target years, as can be seen in Figure 14. More weight is given to recent year (2 for 2014) and less weight for distant years (1 for 2013).

The general pseudo code algorithm for an indicator with respect to the time axes is:

0. Pseudo Code Algorithm

1. *load data sources*
2. *prepare data*
    a. *ignore nulls and/or blanks on key fields*
    b. *detect and manage out-of-range values*
    c. *eliminate extra fields*
3. *join data sources on id field(s)*
4. *calculate necessary (missing) fields*
5. *split analysis on*
    a. *general view*
        i. *aggregate on required fields*
        ii. *sort data*
        iii. *view partial results*
    b. *day of week analysis*
        i. *position analysis on day of week time scale*
        ii. *calculate measure w.r.t. the required aggregation fields at gambling venue granularity level*
        iii. *calculate demanded measure w.r.t. required aggregation at department granularity level*
        iv. *derive anomaly score based on the previously determined measures*

c. *month analysis*

   i. *position analysis on month time scale*

   ii. *calculate measure w.r.t. the required aggregation fields at gambling venue granularity level*

   iii. *calculate demanded measure w.r.t. required aggregation at department granularity level*

   iv. *derive anomaly score based on the previously determined measures*

d. *year analysis*

   i. *position analysis on year time scale*

   ii. *for each year in the year range*

      1. *presort results w.r.t. demanded measure*

      2. *create ranking*

      3. *output preliminary results*

      4. *calculate measure w.r.t. the required aggregation fields at gambling venue granularity level*

      5. *calculate demanded measure w.r.t. required aggregation at department granularity level*

      6. *derive anomaly score based on the previously determined measures*

   iii. *join results on id*

   iv. *determine overall anomaly score through weighted score method*

6. *consolidate results*

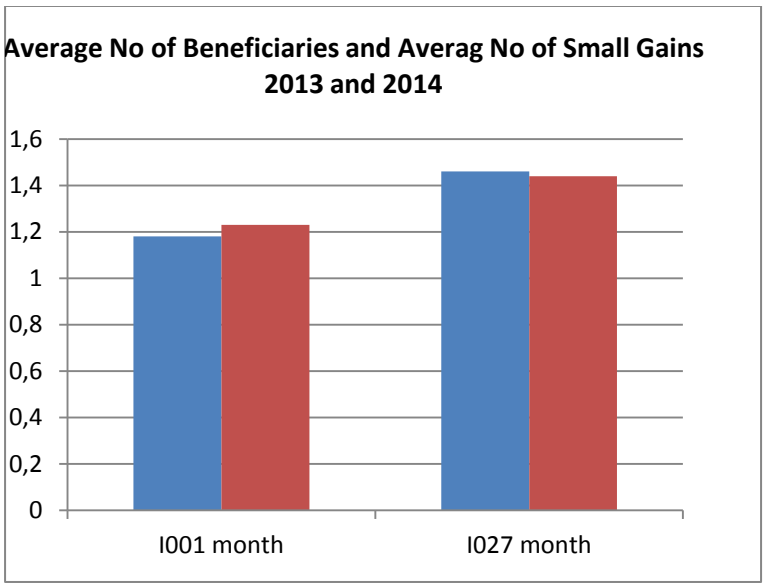
The algorithm needed to construct indicators for location shares the same principle and can be seen in Appendix E, Appendix F, Appendix G and Appendix H.

Conclusions Phase 0 Simple Profiling

After creating these indicators, the results were gathered in order to further analyze them. The values that were considered for each indictor were the minimal value, the maximal value, the average value, the standard deviation value and anomaly score w.r.t. the time scale, but also the location scale.

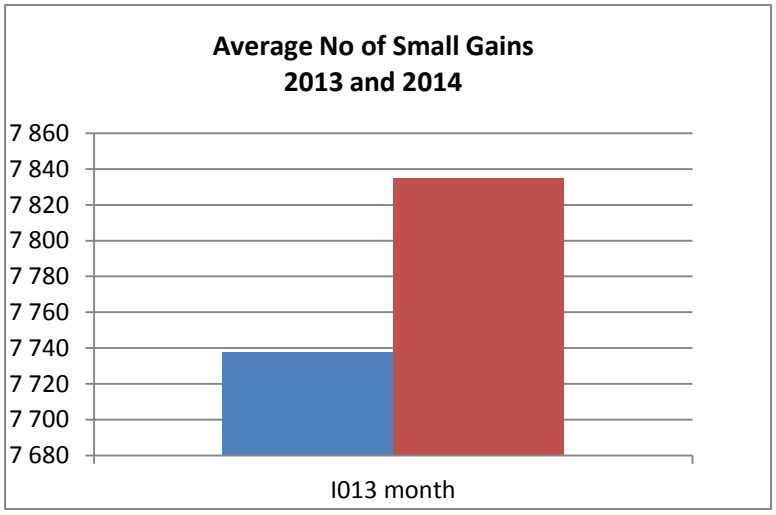
Note that indicators I001, I027, I013 and I060 refer to Number of beneficiaries, Number of small gains, Total amount of small gains, and Number of units (i.e. winning tickets in a small gain). Note that blue represents the values for 2013 and red renders the ones in 2014.

Figure 16 Average Values Indicator I001, I027 per months in 2013 and 2014



The analyst can see that the average number of small gains is well higher than the average number of beneficiaries. The opposite might have meant an error in the data.

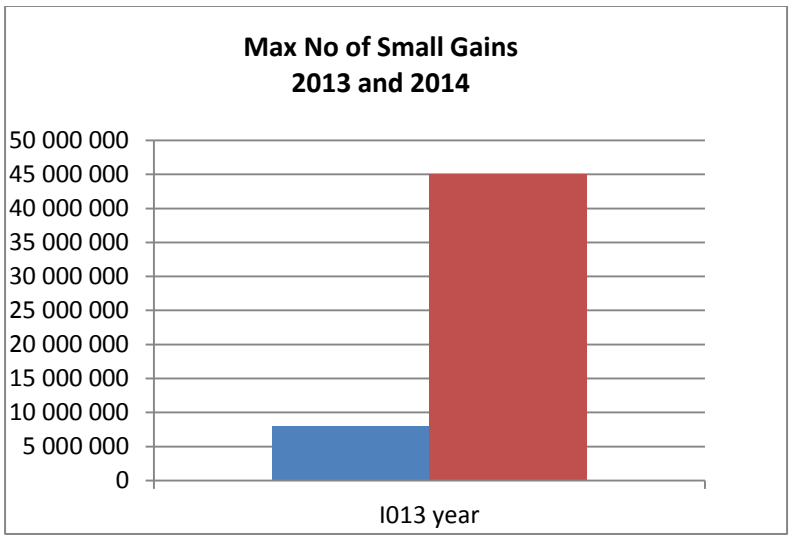
Figure 17 Average Values Indicator I013 per months in 2013 and 2014



It is interesting to see that the average amount of small gains of a beneficiary have increased 3 times in 2014 when compared to 2013. The analyst must consider the data distribution (min, max, standard deviation, mean etc.) to determine the threshold above which a beneficiary's small gains become anomalous.

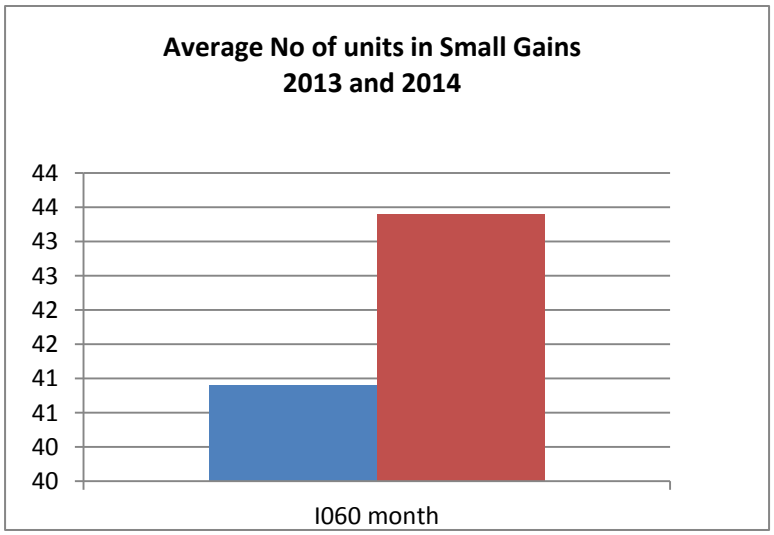
It is also worth mentioning that this increase may also be due to the 2014 World Cup.

Figure 18 Max Values Indicator I013



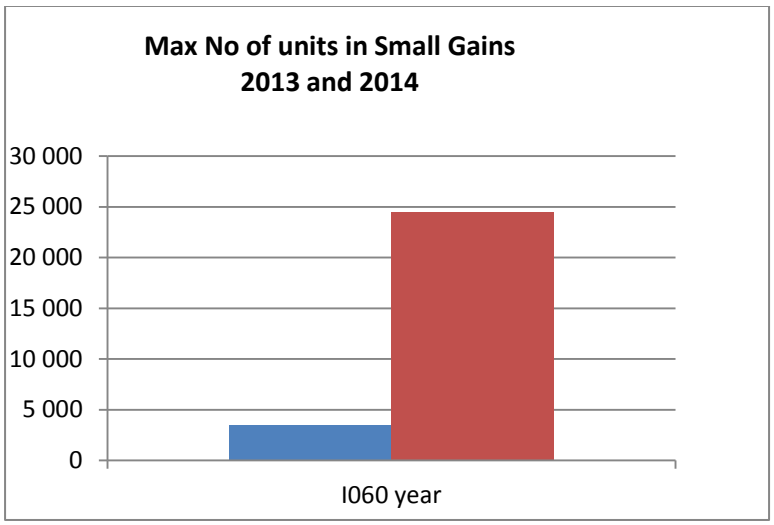
According to Figure 18, the proportion between total amounts of small gains in 2014 versus 2013 seems to be around 5, 08. Now the analyst knows the maximum value for the threshold.

Figure 19 Average Values Indicator I060 per months in 2013 and 2014



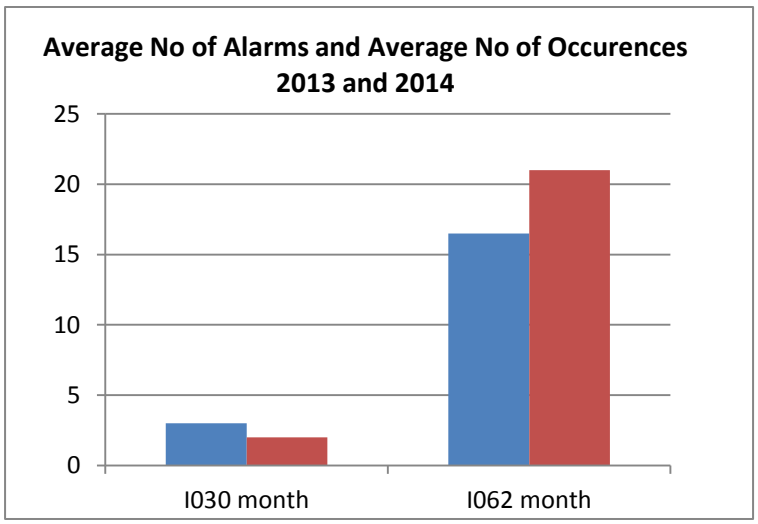
According to Figure 19, the number of units in a small gain has also increased approximately the same proportion as the amount of small gains from Figure 17, hence it is a general increase, but again the maximum values must be considered, as well as other measures (average, min etc.) before determining a threshold.

Figure 20 Max Value Indicator I060



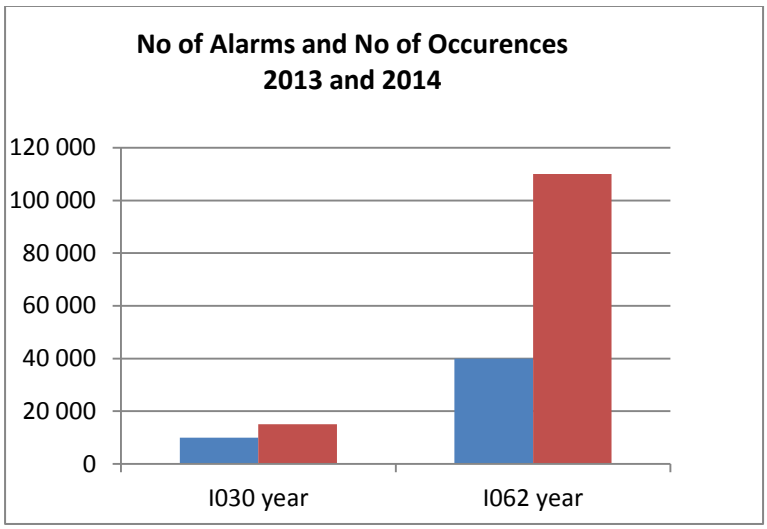
According to Figure 20, the proportion between 2013 and 2014 in terms of number of units seems to be 1:7 (more precisely 0, 213).

Figure 21 Average Values Indicator I030, I062 per month in 2013 and 2014



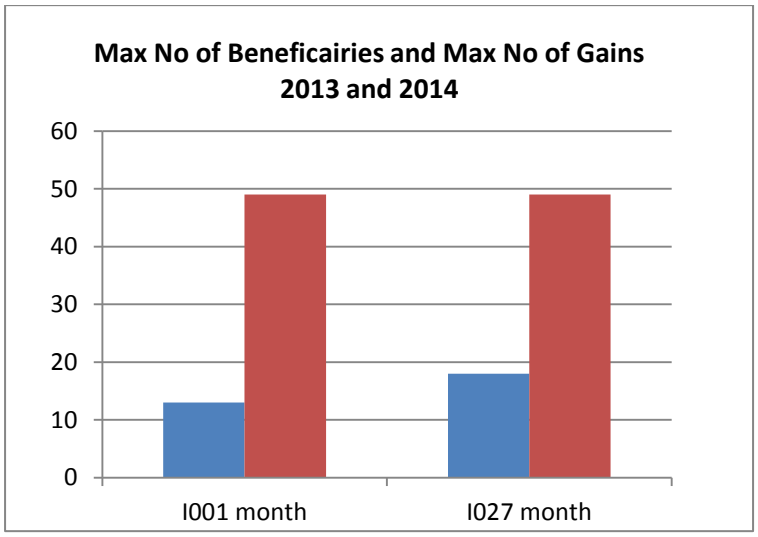
In Figure 21, the user can clearly see that the ratio between alerts (the company's security system meant to detect anomalous transactions) and alerts occurrences (i.e. the same alert repeated several times within the same day) is well above 1:6 which was constant across the 2 years (I030 are the alerts and I062 are the occurrences).

Figure 22 Sum of Indicator I030, I062 per year in 2013 and 2014



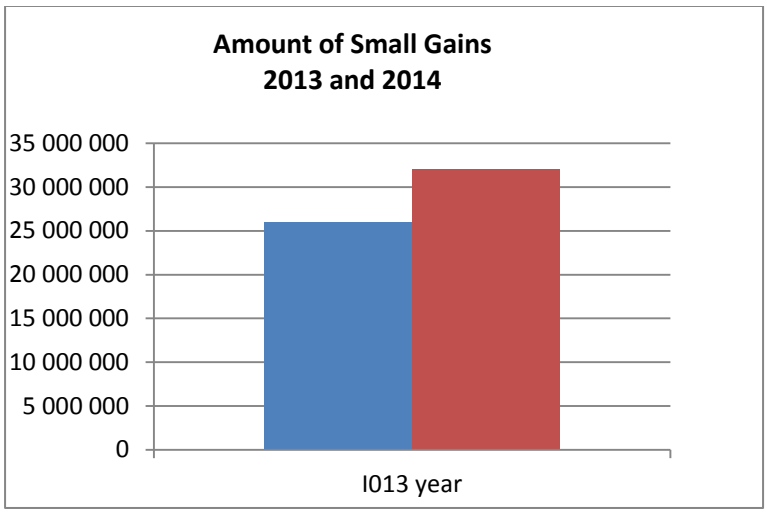
From Figure 22, the reader can see that while the number of alerts has only slightly increased when compared to the previous year (i.e. I030 in 2013 and 2014) the number of occurrences has drastically grown (i.e. I062 in 2013 compared 2014 has increased 3 times).

Figure 23 Max Values Indicator I001, I027 per month in 2013 and 2014



The ratio between the max number of beneficiaries (I001) and the max number of gains (I027) is equal to 1:3 when considering 2013 and 2014. Note that this is the maximum value for the 2 years.

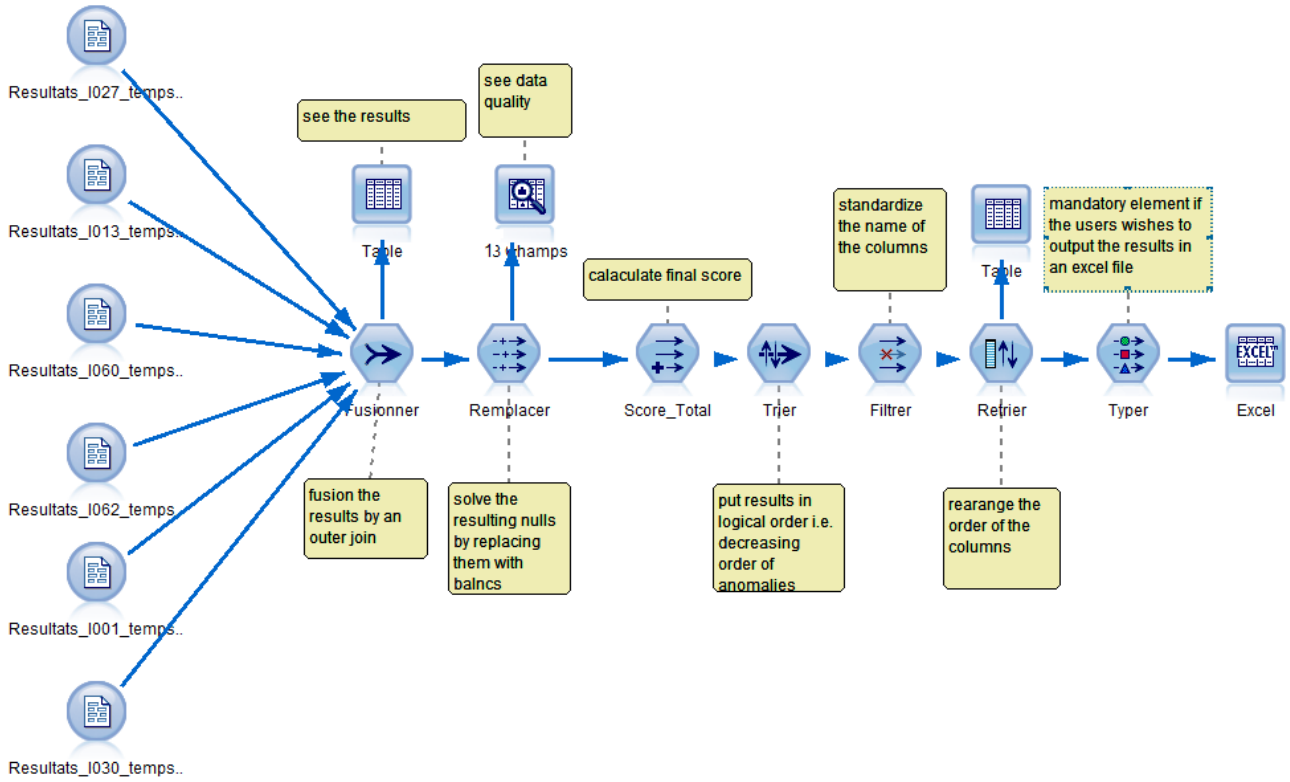
Figure 24 Sum of Indicator I013 per year in 2013 and 2014



According to Figure 24, the general increase in small gains from 2013 to 2014 was more than 5 million euros. From Figure 17, it was noted that the increase in proportion of small gains for a beneficiary was about 1:3 from 2013 to 2014.

The results were consolidated as can be seen in Figure 25. Its scope is to calculate the final anomaly score for all the indicators. In order to do so, it must first join all the indicators on gambling venue id by an outer join, replace all possible resulting nulls by zero, and then calculate the final score. For clarity reasons, the results are ordered; the names of the fields are standardized, the columns are rearranged in a more logical order using the node “Retrier” and then the results are finally outputted into an Excel file.

Figure 25 Results Consolidation Phase 0



Hence, an example of the top 10 most suspicious gambling venues was possible to be listed:

Figure 26 Top 10 Most Suspicious Cases

Gambling Venue	Sum of Anomaly Score	Sum of Score I027	Sum of Score I062	Sum of Score I030	Sum of Score I001	Sum of Score I060	Sum of Score I013
30258	14	3	2	3	0	3	3
247330	13	2	2	1	2	3	3
291785	13	1	2	1	3	3	3
303834	13	2	2	3	2	2	3
307783	13	3	2	0	2	3	3
319238	13	2	2	0	3	3	3
336133	13	2	2	3	2	2	2
349290	14	2	2	1	3	3	3
350220	14	3	2	1	2	3	3
358670	15	2	2	2	3	3	3
Grand Total	**135**	**22**	**20**	**15**	**22**	**28**	**29**

After investigation, the first observation was that some gambling venues were deemed as suspicious (i.e. anomaly score was higher than zero) even though their average gain was not significant (e.g. gambling venues whose gain per month was 5 000 euros). This is because, even though this was detected as an anomaly, the value itself was still too small to cause concern. In other words, at this phase, these gambling venues appear as false positive and the cost of investigating a false positive is something that must be taken into account. In order to improve the final precision of the model (i.e. prevent these false positive to appear in the final top suspicious gambling venues, which will provided in the last phase of the thesis) an additional global threshold will be added for some indicators. Hence, below this global threshold, the anomaly is not worth investigating

(e.g. if the number of small gains for a gambling venue is below 4 000, the gambling venue is not worth investigating).

The algorithm (seen in ***Pseudo Code Algorithm***) was updated:

……

d. year analysis

……

1. check whether the measured unit is greater than the global threshold

b. if true: continue to step 7

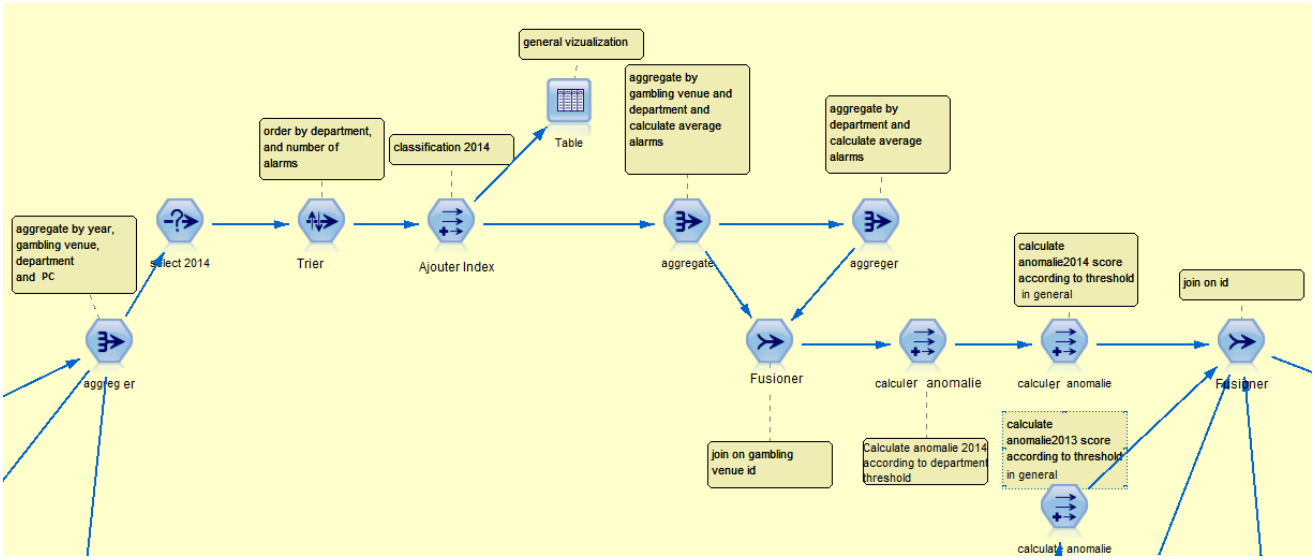
c. if false: anomaly score = 0, continue to step iii

7 derive anomaly score based on the previously determined measures

iii. join results on id

……

Figure 27 Year Analysis with global threshold



The global threshold should be made based on the data tendencies of each year (there are significant differences from one year to another). The next idea was to potentially consider different thresholds for months. The value distribution according to days may also be taken into account. Hence, a finer granularity level was added i.e. day.

One more important pattern was discovered when analyzing the data set w.r.t the number of alerts and number of occurrences in the potentially fradful gambling venues. An interesting average ratio of 1 to 8 was observed for the top most anomalous gambling venues.

The most important discovery from the ranking was the fact that, although many results were interesting, one result with high values for anomalies was discovered to be false. An extreme case of a player who was earning more than 1 million euros per year was suggested to be an exception from the typical pattern of a money launderer, hence an extreme outlier. The conclusions is that the atypical set of extremely successful players should be kept in mind since they influence the average value, the maximal value and the standard deviation value that were used to detect outliers.

It is interesting to see the gambling venues or players that appear all the time in the results. In order to improve the understanding of the player behaviors, the analyst needs to know the differences between an expert gambler, a money launderer and an addict. By observing the data set and the existing player records, it was discovered that an expert gambler has constant high gains, not necessarily equally distributed in time. A money launderer tends to have constant evolution in time, balancing winning and losing (obviously prefers winning but he may also lose sometimes). The time distribution is again not even. An addict's frequency is constant and has random earnings or money loss.

This can be explained as follows: an expert is good at gambling (i.e. knows how to assess risk). The money laundered on the other hand does not want to risk his money. He prefers betting on the most probable winner (i.e. lowest risks) but he does not mind loosing small amounts of money (e.g. Betting on all possible game results to assure some gain). An addict constantly plays because of his addiction.

Another important point is that some results were not "solid" (e.g. the results from Indicator No 27). This is because there is not a sufficient amount of data to form a Gaussian Curve in order to apply the condition in Figure 13. For Gaussian Curve, refer to Appendix I. It may be the case that there are few gambling venues in the department and the one found as being anomalous is actually just successful. In this case, a more practical solution was implemented using a histogram to determine the threshold.

There are also two main aspects to be considered. The first one is whether the money launderer is using multiple gambling venues to obtain money or, on the contrary, he is frequently going to the same gambling venue, in which case, there are high chances that he has an accomplice. In the latter case, it would be a good idea to keep in mind these gambling venues for further studies. And furthermore, the use of statistics is limited and cannot itself construct a solution, as demonstrated.

Phase 1: Stake Analysis

In this subsection, the scope is to analyze in detail the betting trends in order to be able to identify suspicious / atypical transactions that can be related to money laundering activities.

Indicator Index	Analysis Domain	Indicator and Axes
I029	Gambling venue	% of amount of low stakes bets / high stake bets - by month/year - by gambling venue / postal

		code/department
I073	Gambling venue	Number of winners having the same family name but different first names - by month/year - by gambling venue / postal code/department
016	Beneficiary	Number of gambling venues - postal code/department (statistics only)
019	Beneficiary	Staff age - by city / postal code/department (statistics only)
031	Beneficiary	Age of gambling venue - by city / postal code/department (statistics only)
I006	Beneficiary	Amount of gains (payments) - by beneficiary - by month/year
I007	Beneficiary	No of gains (payments) - by beneficiary - by month/year

I012	Beneficiary	Number of small gains - by beneficiary - by month/year
I014	Beneficiary	Proportion of small gains compared to huge gains - by beneficiary
I038	Beneficiary	No of bank accounts - by beneficiary - by year
I039	Beneficiary	No of beneficiaries by payment - by type of operation (check or transaction)
I074	Beneficiary	No of beneficiaries by payment - by account - by year
I075	Beneficiary	No of times when the beneficiary played in multiple gambling venues and earned money in multiple gambling venues - by gambling venue by month/year - by beneficiary
I076	Beneficiary	Amount of money of a beneficiary when the beneficiary played in multiple gambling venues and earned money in multiple gambling venues

		- by month/year - by beneficiary
I079	Beneficiary	Proportion of amount of small gains compared to large gains - by beneficiary

Conclusions Phase 1

The same steps as in the Conclusions Phase 0 Simple Profiling were followed (i.e. data audit, anomaly detection, result consolidation etc.), before rendering graphs and conclusions. For simplicity reasons, they will not be repeated.

For the figures (as in Conclusions Phase 0 Simple Profiling), refer to Appendix J, Appendix K, Appendix L, Appendix M, Appendix N and Appendix O.

In conclusion, the gamblers prefer bank transfer payments. Suspicious cases are reflected by the beneficiaries that chose to distribute their gains over 8 different bank accounts. Indicator I005 (which shows the average values of stakes people bet on) is only for statistical purposes, since money launderers are known to use high stakes in order to ensure their gains (e.g. betting 500 euros on a low odds bet, hence high chances of winning). It is again suspicious the cases where people decide to use multiple gambling venues to get paid (except once again for professional players). There is an increase in frequency of play but also in amount of money gained. The ratio of small to large gains has proven to be significantly important. The average proportion in 2014 is 6.25 but the maximum can go as high as 4.000. In this case, it is an astonishing 266% proportion increase from 2013 to 2014 in terms of maximal values.

Phase 2: Series Analysis

In this subsection, the scope is to analyze series of bets in order to detect money-laundering operations. First, a few definitions will be made clear. For example, an **ISA alert** (which stand for "Internal Security Alert" and it is described in Appendix A) is defined only within the time limit of a day. In order to create an alert, 2 parameters are taken into account: time interval (T) and amount of money (A). In the current state, A is equal to 900 euros while T is represented by a 180 second time frame. If a series of bets has an accumulated amount of money superior to 900 euros within the given time frame, it triggers an alert (e.g. a series of 10 bets of 100 euros each within 180 seconds). If the same series is repeated in the same day, it does not create a new alert but a second occurrence. But if the same series is repeated the next day, that creates a new alert. Hence, if the same series is repeated n times in one day, one alert and n occurrences are created. If the same series reappears the following day, that causes a second alert.

The types of bets taken into consideration for the analysis are **simple bets** (the player wins if his forecast is correct) **multiple bets** (the player wins if all his forecasts are correct) and **combined bets** (the player wins if a subset of the forecasts is correct e.g. 2 bets out of 4 are correct).

Among the hypothesis, which will be used, there is the one that states that there are about 12 million transactions per month. This means that analyzing all the available data would be very heavy; hence a period of 3 months will be considered (October, November and December 2014).

What are interesting to detect are the elements that interrupt the series of bets. These **dissimulators** can be recognized as the insertion of:

- A different type of game (e.g. Bingo bet within a series of 10 sport bets).
- A different odd bet. It also can be seen as the alternation of high and weak odd bets (e.g. 10 bets on low odds, followed by a bet on high odds)
- Bets on different stakes (e.g. 10 bets of 100 euros and 1 bet on 2 euros)

The following study was conducted:

1) An analysis consisted in detecting alerts given a pair of parameters such as amount of money (A) equal to 900 euros and time frame (T) set to 200 seconds. This analysis will be called "**Parameter analysis**".
2) The next study was based on bets having different stakes (hence non homogenous amounts of money). Hence, the amount will not be taken into consideration. This analysis will be called "**No amount analysis**".
3) The analysis was conducted considering the different series of bets whose cumulated value was 700, 900 or 1100 for different values for T. Note that only homogenous series were considered. The analysis will be called "**Amount analysis**".
4) An analysis on the odds of the series was also undertaken in order to reveal possible alerts. This is the "**Odds Analysis**"

Note the scope of each step is to create **new alerts** (**w.r.t. the analysis described in each step**) and compare them with the **ISA Alerts**.

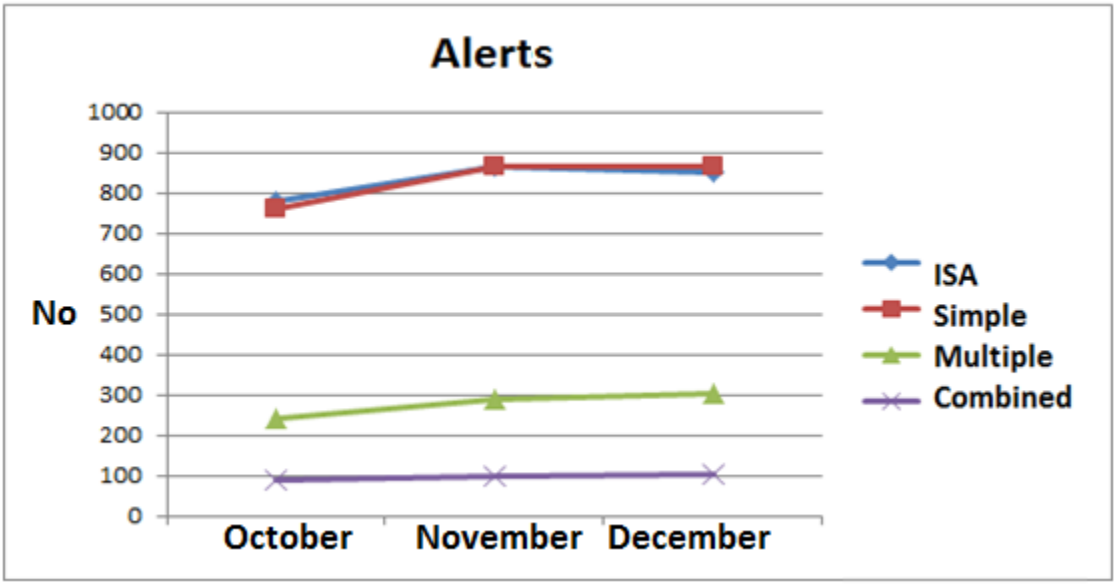
Parameter analysis

The **new alerts** (which are created in this analysis) are triggered by series having a combination of bets of the same odds and stakes (i.e. a set of repeated bets). Hence, the amount is homogenous.

For simplicity reasons, simple bets will be noted as S, multiple bets will be M and combined bets will be C.

After calculation, the results show that the alerts on M series and the ones on C series represent 40% of the alerts on S series in the month of October (and slightly more for November and December).

Figure 28 Alerts - October November December 2014



As can be seen from Figure 28, the number of ISA alerts (i.e. a series of bet detected within 180 seconds of 900 euros worth) for the 3-month time frame has approximately the same values as for the alerts on S series. In fact, these values actually superpose for November (i.e. 850 alerts). Since the trend is persistent, the analysis can clearly state that it is a constant phenomenon. The more complicated bets (C or M) significantly show a lesser number of alerts. They also are fewer in number.

It must be noted that the time breaks are not taken into consideration here (e.g. time breaks between bets). They surely have an impact on the parameter T (i.e. they take a few seconds from the 200 second frame). The algorithm must be improved by making it see these time breaks (and even disregard them in future calculations).

The analyst can also clearly state that there are a higher number of alerts in 2014 when comparing the data with 2013. Refer to Appendix Q.

Note that from observations, some series of bets were made on all possible outcomes of a match (i.e. sure gain for a money launderer).

No amount analysis

In this analysis, the **new alerts** (created in this analysis) were calculated over series whose amount of money (i.e. stakes) was not necessarily homogenous. Hence, the analyst is looking at all the series. The following phenomenon is observed: the series take the shape of blocs composed of different number of bets separated by time gaps. For example, the analyst can see a series of 3 blocs where each bloc is composed of 10 bets of 100 euros each. Each bloc can be separated from the previous one by an approximately 30 minute time gap, as seen in Figure 29.

Figure 29 Series

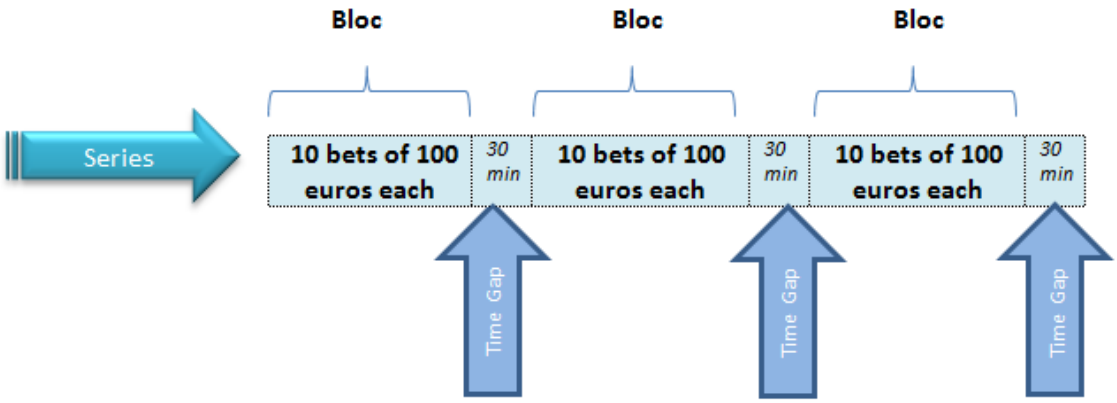
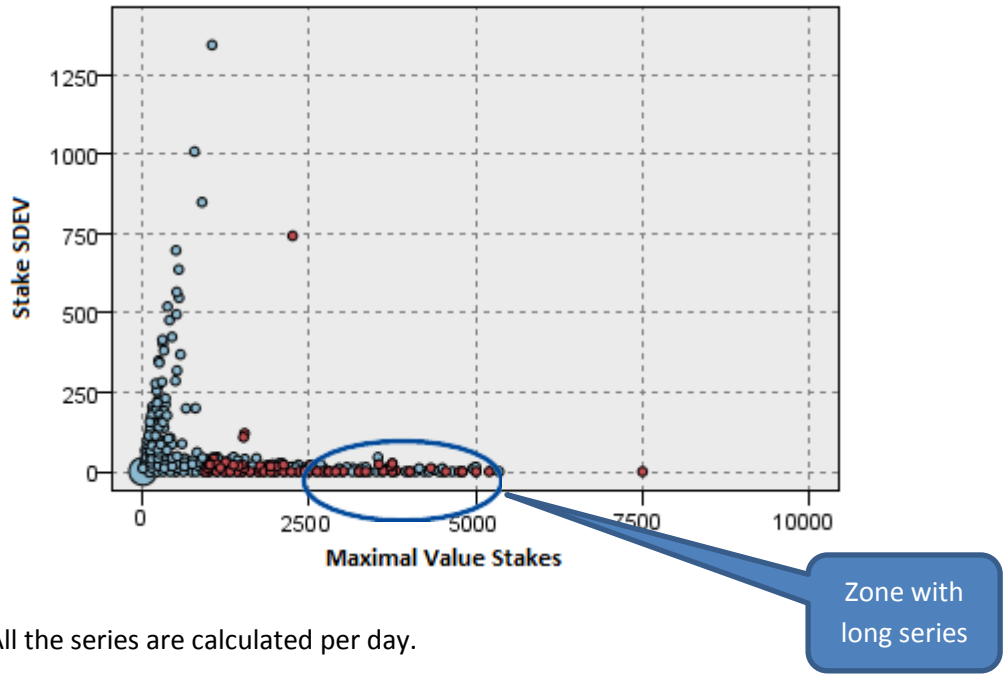


Figure 30 Series Analysis with no control over the stakes



All the series are calculated per day.

From the Figure 30, the analysis can see on the vertical axis as being the Stake Standard Deviation and the horizontal axis the Maximal Value Stake (Note that standard deviation is a measure meant to quantify the amount of variation in the data set).

Analyzing the first axis, the analyst can see that most series are homogenous, since they tend to be closer to the bottom of the figure (i.e. closer to zero). Very few are non-homogenous.

Looking at the second axis, the distribution of values is much higher, ranging from 2 euros to 7 500 euros. There are some peculiar series that were rendered in blue in the graphic. These are the series that are composed of blocs. They accumulated amount reaches 5 000 euros. Due to the length criteria, it was possible

to detect these series. They are very useful in order to detect people that do money laundering since they render some sort of repetition.

In the cases of long series (e.g. 80 bets), there are no ISA alerts. They range from 2 500 to 5 000 euros (hence significant amount of money). Their blocs tend to be repetitive yet no ISA alert is triggered. For example, there is the series defined in Figure 29. ISA alerts take into consideration 900 euros in 180 seconds; hence these 1 000 euros series is never detected.

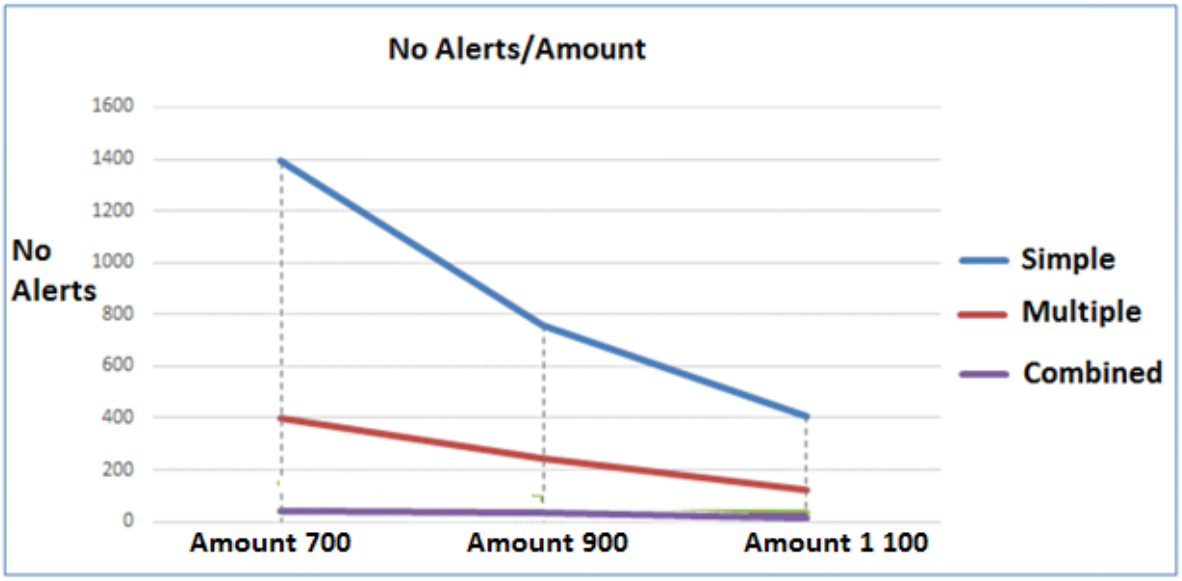
Therefore, for money laundering, volume is considered. The repetitive nature of the bet is something that causes suspicion i.e. bloc, time gap, bloc, time gap etc. It is worth mentioning that also dissimulators can indirectly form time gaps between blocs.

It must be noted that some small chains can also appear at the beginning of each series which can "delay" the start of the series (i.e. a few 2 euro bets followed by 20 bets of 50 euros). ISA takes into consideration series whose amount of money passes 900 euros in 180 seconds. Because of the "attached" 2 euro header, ISA may not see that the actual value of the chain is more than 1 000 euros (since it calculates the sum of the series within 180 seconds. It does not have visibility after 180 seconds.). It should be designed to "skip" these small chains to better generate alerts. The algorithm needs to be updated.

Amount analysis

The same analysis is done but with an additional constraint. This time, the stakes need to be homogenous. The amount of money subsequently takes the values 700, 900 and 1 100 euros and the process is repeated.

Figure 31 Analysis with control over the stakes



As can be seen from Figure 31, the evolution of the **new alerts** considering the amount is linear. Each time the amount is augmented by 200 euros, the number of **new alerts** is decreased two times each time. The advantage of using a 700 euro threshold is the fact that the blocs are more detectable. The S series is rapidly decreasing while the M series is less rapid, still the same effect can be seen. This is a "volume effect". In decreasing order, the series that are most likely to "cause" fraud are S series, followed by M series and C series.

Odds Analysis

The analysis is made according to the odds. The classification of odds can be seen as:

Indicator	Odds
High	Superior to 2,5
Medium	Between 2,5 and 1,2
Low	Inferior to 1,2

The highest number of alerts can be observed on low odds bets. But it must be mentioned that although money launderers use low odd bets to "clean" money, average people may also prefer them just because they are simple to us. Multiple bets and combined bets have a more complicated way to determine the final odds.

Phase 3: Complex Profiling

In this stage, the scope is to create complex profiles for the beneficiaries as well as for gambling venues by means of clustering algorithms with the aid of IBM SPSS.

Steps for Clustering:

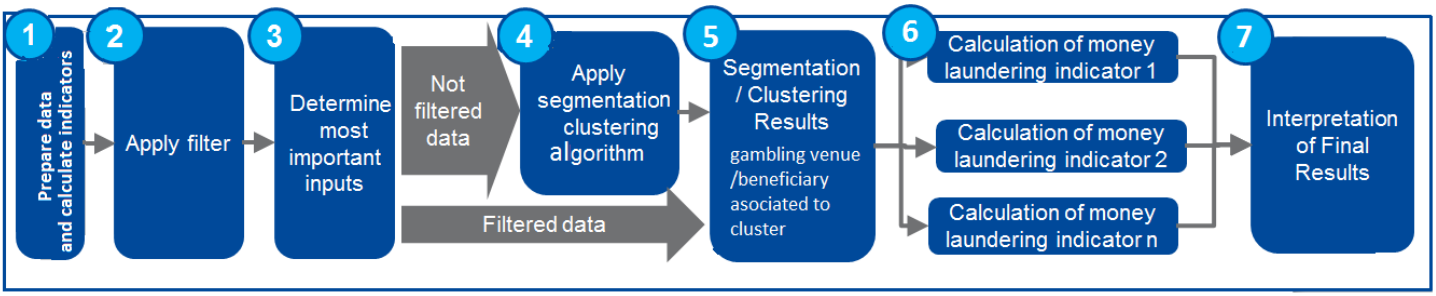
1. Prepare data and calculate new statistics to use in the clustering.
2. Use filters to segment the data before applying statistical clustering algorithms by means of the most relevant factors.
3. Determine the most important attributes for the clustering (e.g. type of gambling venues, competition present, etc.).
4. Apply a clustering algorithm for each group (i.e. gambling venue and beneficiary) in order to obtain 4-5 clusters per group (and selecting about 4-5 attributes for the clustering algorithm from the entire set of attributes).
5. Consolidate results for gambling venues and for beneficiaries.
6. Following the identification of coherent clusters, each money laundering indicator is calculated w.r.t. each cluster allowing the identification of clusters that are at high risk.
7. Interpretation of the results

Two remarks regarding the algorithm must be made. The first is (step 3 in the algorithm) the fact that, by segmenting the data prior to the cluster method, the analyst already sets aside groups that are known to have distinctly different characteristics without the need to apply clustering (e.g. payment centers[1]). The second (step 5) is related to why this number of inputs and clusters are a good recommendation. The reason is that it makes the clusters easier to interpret and therefore more relevant to the business. Also the greater the number of clusters, the more difficult it will become to distinguish between them.

Numerous iterations were made, modifying each time the attributes, in order to obtain good clustering results in terms of data mining (i.e. silhouette measure which is a good measure to evaluate the quality of the clusters.) but also meaningful for the business.

Note that the silhouette measure (which can be seen in Figure 33) combines two concepts. First, there is the concept of cluster cohesion (which favors clusters that contain highly similar values). Secondly, there is cluster separation (which favors clusters that contain highly separated clusters).

Figure 32 Steps for Segmentation



Factors to consider for the interpretation of results:

- The results of money laundering indicators that are already calculated. For each indicator, it can be stated whether the gambling venue /beneficiary is anomalous or not.
- The indicators that are purely statistical for each gambling venue / beneficiary (e.g. the number of transactions, the total gained amount etc.)
- Indicators to categorize the gambling venues / beneficiaries which are not money laundering indicators but they are still interesting to use in order to understand the clusters (e.g. number of terminals in the gambling venue)

Points of attention:

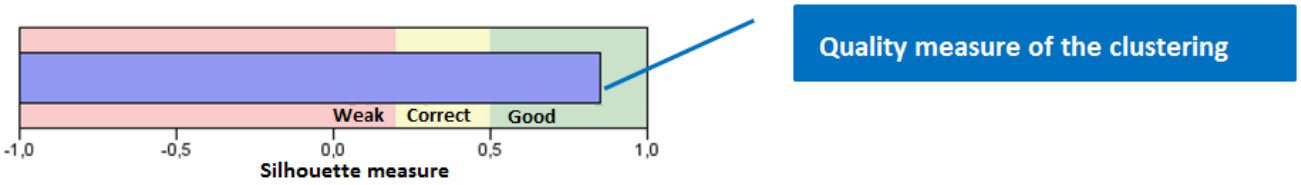
- Coherence of the clustering to be confirmed / tested at all points. This is done by using the silhouette coefficient.

[1] Payment Centers are special centers whose sole purpose is to ensure payment services. They are also the only ones authorized to make payments larger than 300 euros. Gambling venues can only pay up to 300 euros.

- Special care with the possible use of geographical characteristics as attributes in the clustering algorithms (e.g. communes, regions etc.). Maybe consider possible regrouping of the departments.

In this case, the quality measure is about 0, 9, as can be seen in Figure 33. Of course, the quality measure is important, but also the ease in which we can interpret the clusters into something meaningful for the business. There is no point in having excellent quality measure if the clusters are difficult to understand.

Figure 33 Quality Measure Gambling Venue clustering



The results in case of gambling venues are:

Figure 34 Gambling Venue Clusters – General View

	PC cluster	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5	cluster 6
Proportion of gambling venues in the cluster	0%	19%	10%	30%	6%	22%	14%
Type of Gambling Venue	PC(100%)	bar(34,4%)	bar,tobacconist(23,6%)	bar,tobacconist(43,7%)	newsagent(100%)	tobacconist,newsagent(100%)	bar,tobacconist,newsagent(100%)
Competition present?	no(94,3%)	yes(100%)	no(65%)	no(100%)	no(100%)	no(100%)	no(100%)
Manager is a winning player?	no(91,8%)	no(100%)	yes(100%)	no(100%)	no(100%)	no(100%)	no(100%)
Average number of bets by week	1878,27	1963,125	4330,53	1496,115	1263,48	1446,975	1126,125
Average amount of bets by week	156,00 €	177,00 €	279,00 €	126,00 €	114,00 €	136,50 €	109,50 €
Average amount of large gains by gambling venue	195,45 €	11,10 €	18,00 €	11,40 €	8,25 €	9,45 €	8,10 €
Win to bet ratio	NA	138	141,45	154,65	138,45	146,25	138,9
Average duration since the last manager replacement	62,4	61,5	64,2	63,3	65,1	66,75	62,85
Average number of managers	2,39	3,66	3,41	2,90	2,84	3,11	3,26

Note that these figures are illustrative.

12 attributes was the initial set of attributes to be considered in order to create the clusters. The attributes that were finally selected were: type of gambling venue activity (bar, tobacconist, newsagent etc.), competition present, if the manager is a winning player, amount of bets per week. These four generate the best silhouette but they are also the most meaningful in terms of business.

Note that the indicator called “Competition present” was used to reveal the % of gambling venues that have competing gambling services in the same location. It is interesting to see that 100% of the gambling venues contained within cluster 2 have a competitor gambling service available in the same location. With the

exception of cluster PC and cluster 2, the remaining clusters have no competing gambling services within the same gambling venue.

The indicator "average duration since the last manager replacement" is used rather to understand the segments, not to create them. If the duration is long, it means that the gambling venue tends to be stable, hence a lower chance of money laundering (at least in theory).

The interpretation of cluster results:

- **Cluster PC** (i.e. payment centers) is a cluster that was manually created due to the filter.
- **Cluster 1 is** the cluster that has 100% competition present, hence its name „Gambling venues with competing gambling services in the same location".
- **Cluster 2** has 100% of its managers as active winning players. It also has the highest average amount of bets per week. Perhaps active managers increase revenues a lot.
- **Cluster 3** is characterized by the fact that it has the largest win to bet ration. It also has bar, tobacconist in 43% of the cases, but the most significant feature remains the win to bet ration. Perhaps these gambling venues are visited by good players that gamble on high stakes.
- **Cluster 4, 5** and **6** are differentiated in terms of type of gambling venue. Cluster 4 has 100% newsagent, cluster 5 has 100% tobacconist, newsagent and cluster 6 has bar, tobacconist, newsagent 100%.

Hence, the name as well as the profile of the clusters is shown in Figure 35:

Figure 35 Gambling Venue Clusters - Names and Profiles

	PC cluster	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5	cluster 6
Names	PC cluster	Gambling venues with competing gambling services in the same location	Gambling Venues with Manager Player and Winner	Gambling Venue with the most significant amount of gains	Typical newsagent Gambling Venues	Typical tobacconist, newsagent Gambling Venues	Typical bar, tobacconist, newsagent Gambling Venues
Description	cluster where the gambling venues are only payment centers	cluster where the competition is 100% present	cluster where manager is a winning player(1 or more times between 2013 and 2014). It is also the cluster where the average mount of gains per week is the highest	cluster where the gambling venues paid the most gains but with lesser bets received. It is also the cluster with the highest number of gambling venues.	cluster with newsagent, no competition and no winning manager. This represent 81% of newsagent.	cluster with tobacconist, newsagent but no competition and no winning manager. This represents 77% of all gambling venues tobacconist, newsagent	cluster with bar, tobacconist and newsagent and no winning manager. This represent 67% of gambling venues bar, tobacconist, newsagent

As general observation, the largest cluster is the 3rd one. In terms of win to bet ration, the most important, as specified before, is cluster 3 followed by cluster 5 and 2.

Figure 36 Anomaly by Gambling Venue cluster

	PC cluster	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5	cluster 6
Name of the cluster	PC cluster	Gambling venues with competing gambling services in the same	Gambling Venues with Manager Player and Winner	Gambling Venue with the most significant amount of gains	Typical newsagent Gambling Venues	Typical tobacconist, newsagent Gambling	Typical bar, tobacconist, newsagent Gambling
No of large gain beneficiaries	8	28	58	78	6	18	2
Amount of low odd bets/amount of bets by validating gambling venue	0	396	348	572	230	510	202
Amount of bets paid in gambling venue/amount of bets in the same gambling venue	0	166	74	348	74	134	104
No of checks of small amount of money	2	10	38	20	6	6	0
No of managers that do not gamble in their own gambling venue	2	0	278	0	0	0	0
Total	**12**	**600**	**796**	**1018**	**316**	**668**	**308**

The anomaly distribution within each cluster was calculated using the money laundering indicators: number of large gain beneficiaries, amount of low odd bet/amount of bets, amount of bets paid in a gambling venues/amount of bets in the same gambling venue, number of checks of small amount of money, number of managers that do not gamble in their own gambling venue. Note that the anomalies were calculated w.r.t. a given threshold, as described in Phase 0: Simple Profiling and Phase 1: Stake Analysis. The idea is to see which cluster is the one at most at risk since the threshold for the indicators are the same across all clusters (except for payment centers). The next step (future research) is to vary these thresholds w.r.t each cluster in order to adjust them according to the normal behavior of the gambling venue population. From Figure 36, the analyst can see that the most suspicious cluster is cluster number 3. It must be noted that cluster 3 is the one with the most significant amount of gains. In the gambling venues of Cluster 3, players also play a lot on low stake odds, and we already mentioned that money launderers prefer low stake odds.

In the case of beneficiary clustering, the quality measure is 0, 8.

But before presenting the results for beneficiaries, it must be noted that, depending on the frequency of wins, winners were divided into 3 groups: unique winners (which have won only once), multiple winners (that have won 2 – 5 times) and recurrent winners (more than 10 wins) w.r.t. 2013 and 2014.

The results in case of beneficiaries are shown in Figure 37:

Figure 37 Beneficiary Clusters – General View

	cluster 1	cluster 2	cluster 3	cluster 4	cluster 6	cluster 5	cluster 7
Proportion of beneficiaries	52,70%	1,60%	8,44%	0,40%	7,40%	12,36%	17,10%
Type of beneficiary (frequence of gains)	Unique(100%)	Unique(42,2%)	Unique(80,9%)	Recurrent(99,9%)	Recurrent(100%)	Multiple(100%)	Multiple(100%)
Beneficiary is a manager?	No(100%)	Yes(100%)	No(100%)	No(75%)	No(100%)	No(100%)	No(100%)
No of gambling venues used by the beneficiary	Unique(100%)	Unique(82,3%)	Unique(100%)	Multiple(80,9%)	Multiple(66,4%)	Multiple(100%)	Unique(100%)
Average Gained Amount	1 184,78 €	5 044,62 €	4 077,78 €	144 167,94 €	15 993,34 €	4 553,24 €	3 186,14 €
Average No of Gains	1,5	5,25	1,95	87,75	16,05	4,5	3,9
Average amount gained by win	789,85 €	960,88 €	2 091,17 €	1 642,94 €	996,47 €	1 011,83 €	816,96 €
Average No of Large Gains	1,5	5,1	1,65	82,35	15,6	4,2	3,9
Average No of small gains	0	0,15	0,3	5,4	0,45	0,3	0

As attributes for the clusters, the indicators that were used were: average gained amount, no of gains, type of beneficiary (frequency of gains), no of gambling venues used by the beneficiary and beneficiary is manager.

In order to understand the clusters, the analyst needs to understand that, for example for cluster 3, 80.9% of beneficiaries have won only once during the analysis period (2013 - 2014). The average amount gained per win is calculated by dividing the average gained amount by the average number of gains. The average number of gains is of course the sum of the average number of large gains and average number of small gains.

The interpretation of cluster results:

- **Cluster 4**, which is the large gain winners, has obviously the largest gains (9 times higher than the next cluster, which is cluster 6). They have the highest number of small and large gains. They represent less than 1% of the population. They may be professional players.
- **Cluster 6** is characterized by the fact that its players have won more than 10 times, but the amount gained is significantly less than the large gain winners. They won less in terms of money but also in terms of frequency. Consequently, the average amount gained per gain is less.
- **Cluster 5 and 7** are both multiple winners (between 2 and 5 gains), but there are two differences between the clusters. The first difference consists in the number of gambling venues visited by the winner. The second difference is about the average amount of gains. In cluster 5 for example, all the winners tend to visit multiple gambling venues, while in cluster 7, they only use one gambling venue. Cluster 5 has a higher amount of gain, and also amount of gains per win, when compared to cluster 7.
- **Cluster 1** can be called the simple players as they have won only once. It is also the largest cluster. They only use one gambling venue and they have the lowest amount of gains per gain. Hence, they are not regular winners and therefore are likely to be occasional gamblers that frequent their local gambling venue.

- **Cluster 2**'s main feature is the fact that the beneficiaries are also managers of a gambling venue. It is smaller than regular players in terms of average gain, but bigger than the multiple winners.
- **Cluster 3** is the cluster with the good players, which means that when they play and win, they win more than other clusters.

Hence, the name as well as the profile of the clusters is:

	cluster 1	cluster 2	cluster 3	cluster 4	cluster 6	cluster 5	cluster 7
Name of the cluster	The simple winners	The manager winners	The good players	The large gain winners	The regular players	The multiple winners that cash their earning in multiple gambling venues	The mutiple winners
Description	The biggest cluster where 100% of the beneficiaries have won only once between 2013 and 2014 and where the gained amount is the lowest.	The cluster where the beneficiaries are the managers.	The cluster where the beneficiaries have not won many times, but they have the best average gains.	The smallest cluster(only 0,4% of winners) but with a gained amount 9 times bigger than other clusters. These beneficiaries have also won many times - potential professional players.	The cluster with players that have regularly won (more than 10 times in 2 years), but with a gained amount relatively low when compared to cluster 4.	The cluster with beneficiaries that have won 2 - 5 times and that visited multiple gambling venues to cash their gains.	The cluster with beneficiries that have won 2 - 5 times and that frequently go to the same gambling venue to cash their gains.

Figure 38 Anomaly by Beneficiary cluster

	cluster 1	cluster 2	cluster 3	cluster 4	cluster 6	cluster 5	cluster 7
Name of the cluster	The simple winners	The manager winners	The good players	The large gain winners	The regular players	The multiple winners that cash their earning in multiple gambling venues	The mutiple winners
No of large gain payments	0	0	0	998	56	0	0
No of cumulated small gains	0	2	0	190	72	0	0
For the same beneficiary, proportion of small gains/large gains	0	4	2	10	32	6	0
No of bank accounts	138	522	0	58	144	154	66
Amount of gains per beneficairy having received more than 5 checks in one day	4	6	12	170	70	10	0
Total	**142**	**534**	**14**	**1426**	**374**	**170**	**66**

As can be seen from Figure 38, cluster 4 presents the highest risk (large gain players) because the indicators that were calculated are connected to gains (number of gains, amount of gains, proportion of gains etc.). Considering the first indicator "no of large gain payments" of course the large gain clusters will have a high score on this criterion. But on the indicator "no of banks accounts", the cluster that is most suspicious is cluster number 2 (the cluster where managers are winning players). Apparently, they tend to distribute their

gains over multiple bank accounts. The last indicator reveals information about the amount of money that the beneficiary obtained having received at least 5 different checks in one day. Again, the large gain cluster is the most at risk.

As future improvement, different thresholds might be settled for large gain winners in order not to have too many of these players as suspicious players. Maybe what they are doing is a normal activity for them. In other words, perhaps for a simple player it makes sense for this threshold to be 2 (i.e. maximum 2 bank accounts used) because they do not play that often anyway. But for the professional players (or at least for the more experienced players) and also for manager players, it would be more realistic to have a threshold higher than 2 (business bank account, private bank account, wife's bank account. Hence it can be more than 3 bank accounts). Therefore, each threshold must be adapted to the profile of the player. The same can be said for gambling venues.

As a conclusion, the objective was to create a profile of the gambling venues and of the beneficiaries in order to better identify anomalies. Clusters that make sense for the business have been identified. Also the clusters that rendered the highest risks have been revealed.

Phase 4: Simple Scoring

The scope of this phase is to regroup all the anomaly indicators according to a target domain (gambling venue or beneficiary) and calculate the overall anomaly score according to the weight assigned to each anomaly indicator. This weighting will indicate the importance of the indicator in relation to the detection of money laundering. The assessment of importance is done by experienced gambling industry professionals. Therefore, the first step is to take the list of calculated indicators and assign a weight to each of the money laundering indicators. The second step is to make the sum of weighted anomaly indicators that will provide an overall score at gambling venue level and at beneficiary level.

An example of a list of weights can be seen in Table 3.

Table 3 Weight of the indicators for Gambling Venues

Analysis Domain	Indicators	Importance	Weight
Gambling Venue	No of large gain beneficiaries	High	3
Gambling Venue	Amount of low odd bets/amount of bets	Medium	2
Gambling Venue	Amount of bets paid in a gambling venue/amount of bets in the same gambling venue	Low	1

Gambling Venue	No of checks of small amount	Low	1

Moreover, it was deemed by gambling fraud experts that it is possible to group some of the indicators in order to create rules that give a clearer indication of money laundering. In Table 4, a demonstration of grouping can be seen. Note that many such groupings were made, for both beneficiary but also for gambling venues.

Table 4 Grouping of Money Laundering Indicators

Analysis Domain	Indicators	Importance	Weight	Grouped Importance	Grouped Weighting
Beneficiary	% of small odd gains/total amount of gains - by gambling venue - by year	Low	1	If these 4 indicators are in alert for one beneficiary (i.e. all of associated measures pass the anomaly threshold for at least one beneficiary), then the importance of these 4 indicators becomes high.	3
Beneficiary	No of large gains	High	3		
Beneficiary	No of small gains	High	3		
Beneficiary	Proportion of small gains/large gains	Low	1		

Hence, a result sample (in the case of gambling venues) can be seen in Table 5. Note that not all indicators are present.

Table 5 Sample of Weighted Anomaly Score for Beneficiary

Beneficiary ID	I007: No of gains	I039: No of beneficiaries per payment	I014: No of small gains/large gains	I038: No of bank account	I086: No of telephone numbers of the gambler	I097: No of gambling venues visited by the gambler	Total Anomaly Score
1002357	75	21	34	1	5	5	115
1002324	32	8	12,8	2	1	8	32

1002393	229	23	5,2	2	1	10	32
1002370	53	32	2,9	5	5	9	27
1002378	52	6	67,9	1	1	4	27
1002385	51	7	102,8	2	4	6	26
1002374	50	3	176,2	9	2	4	26
1002389	54	9	89	2	2	5	25
1002344	63	8	50	12	3	11	16
1002349	47	12	20	5	3	9	8
1002354	49	15	15,4	2	1	8	7
1002359	34	4	2,3	1	1	5	6
1002368	21	3	4,8	5	1	2	1
1002373	5	3	3,7	1	1	2	0
1002380	8	2	2,8	1	1	1	0

Code color for lines:

Anomaly
Anomaly in 2013
Anomaly in 2014
Anomaly in 2013 and 2014

In the header of the table:

Weight
Weight 3
Weight 2
Weight 1

Note that the figures are illustrative.

As a small example, considering the first beneficiary (ID 1002357), he shows proof of anomaly in 2014 and 2013 for the indicators I007 and I039 because both values (i.e. 75 and 21) exceed the thresholds (hence both cells are in red). He is not anomalous in 2013 (no cell in yellow). His values for I014 and I038 are non-anomalous, but he is anomalous for I09 7in 2014 (cell in orange).

The final scope is to calculate a weighted score for each line, where each line represents a beneficiary or a gambling venue. More precisely:

For each gambling venue:

$$Final\ Anomaly\ Score = \sum_{anomaly\ indicator} anomaly\ indicator\ score * weight\ of\ the\ anomaly\ indicator$$

The same will be applied for beneficiary. The results can be seen in Table 5. Hence, a file called "Money_Laundering_Scoring" is created containing the final anomaly score for beneficiary and gambling venue, the anomaly score for each indicator, as well as their associated value. An extra column was added including information whether the record was a proven fraud case or not (therefore the name of the column is "FRAUD" having a Yes/No value).

4. Additional Analysis

The scope of this additional phase is to see whether by using the INSEE data we are able to make a correlation between them and the anomaly score (question is whether the environment of the gambling venue influences money laundering fraud).

The INSEE 2013 Data are a group of files that describe:

- Social action services (temporary housing center, housing center for disabled children etc.)
- Commerce (supermarket, clothing shop etc.)
- First-degree education for both public and private sector (private and public elementary school, public and private kindergarten etc.)
- Second-degree education for both public and private sector (public and private university, public and private boarding school etc.)
- Medical and paramedical functions (cardiologist, gynecologist etc.)
- Private services (driving school, plumber etc.)
- Health units (pharmacy, ambulance etc.)
- Sports, leisure and culture (cinema, theater etc.)
- Tourism and transport (taxi, airport etc.)
- Residential activities (active population, unemployed etc.)
- Couples, families and household (e.g., widowed, divorced etc.)
- Housing (no. of rooms, house equipment etc.)
- Population (gender distribution, immigrants etc.)
- Formation (send to school or not, higher education etc.)

The process has the following steps:

1. Prepare data sources (i.e. the scored gambling venues from the previous phases of the experiment and the INSEE files. Note that the possibility of regrouping the INSEE files must be considered e.g. “First degree education” and “Second degree education” INSEE files can be merged)
2. Calculate real anomaly score per IRIS. This is done by aggregating gambling venues by IRIS and calculating their total anomaly score.
3. Calculate the predicted anomaly score per IRIS (i.e. geographical statistical unit) based on INSEE variables. In other words, the analyst will use the CHAID model to predict the target anomaly score having as attributes the INSEE variables.
4. Analysis at IRIS level to see the correlation between the predicted anomaly score per IRIS and INSEE variables. This correlation is shown by the CHAID model.

For the data sources preparations, refer to Appendix R. Note that some files were merged (i.e. first degree education and second degree education) and some fields were aggregated (e.g. number of commerce) prior to the data preparation.

Figure 39 Flow INSEE Data and Anomaly Score

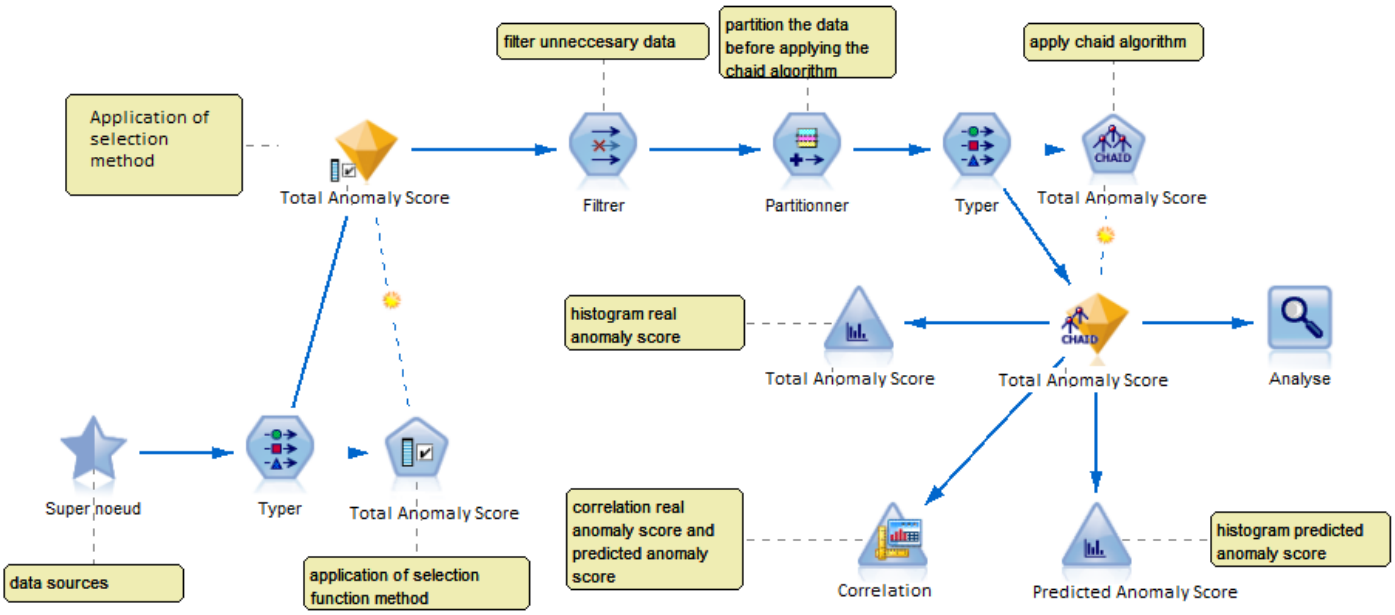
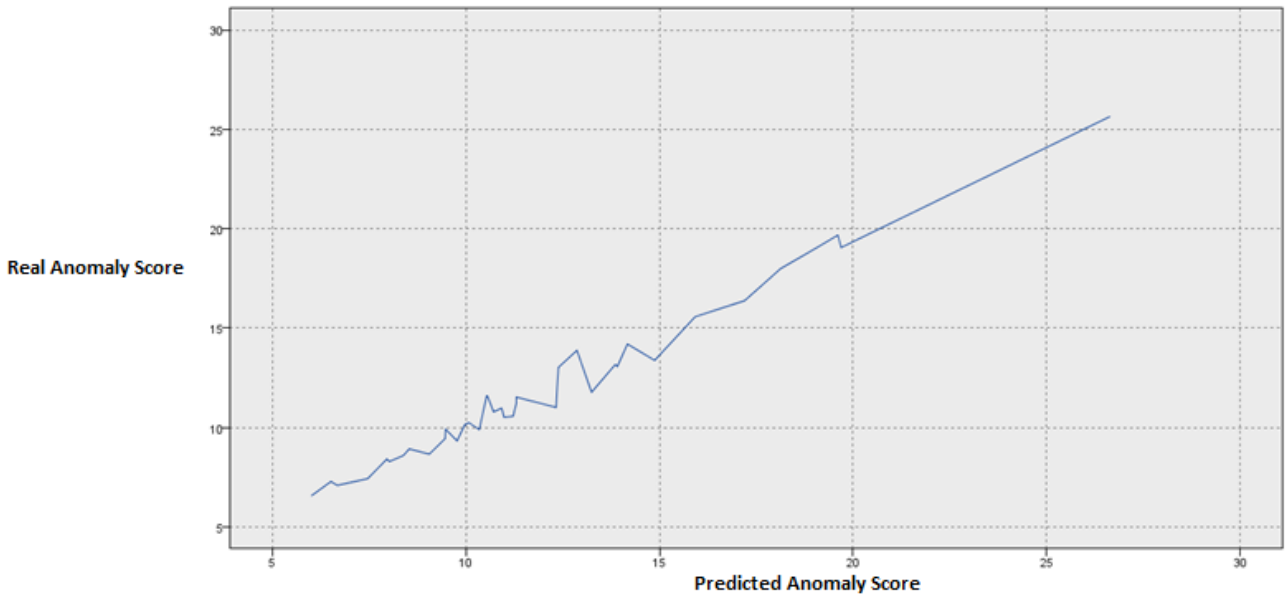


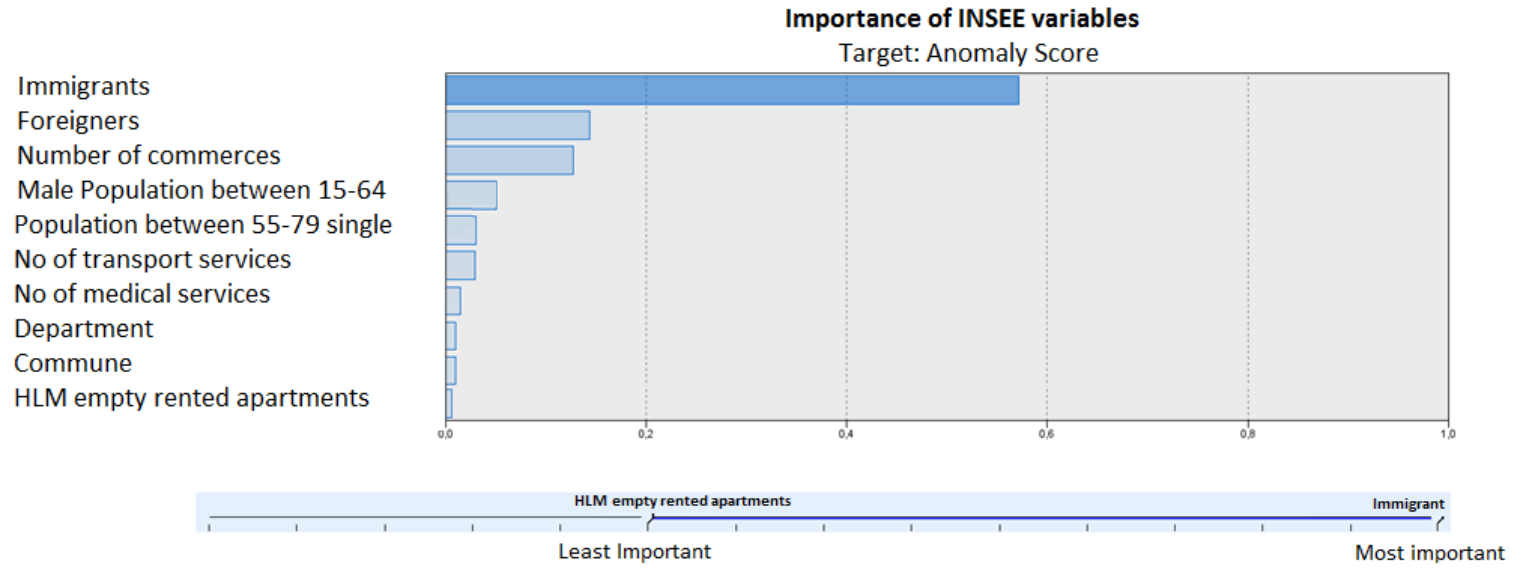
Figure 39 shows the stream that was implemented to analyze the correlation. After loading the data sources, the selected method, which is an algorithm that identifies the factors that have the most influence on the target variable, is used. It enables the analyst to limit the number of input fields in a model. Thus, the analyst selects those INSEE fields which are most connected to the target field (i.e. total anomaly score). The next step is to use a partitioning node prior to the CHAID modeling implementation (random 70% training set and 30% test set). After applying the CHAID algorithm, the analyst is able to see the correlation of the real average anomaly score per IRIS (calculated in Phase 4: Simple Scoring) but also of the predicted anomaly score per IRIS based on the INSEE data (i.e. those INSEE previously selected by the select method). From Figure 40, a clear correlation between the two values can be seen.

Figure 40 Correlation



Since CHAID is a decision tree algorithm, the analyst can see the most important INSEE variables that influence the anomaly score per IRIS.

Figure 41 Importance of INSEE variables



In order of the most important variables to the least important, there are:

- Immigrants (i.e. non EU). A possible explanation may be the facts that, since immigrants tend to have low income, money launderers tend to take advantage of their situation and use them to commit fraud.

- Foreigners (EU), which are closely related to immigrants.
- Number of commerces, which denotes the prosperity of a region but also the density.
- Male population between 15 – 64 and population between 55 – 79 single.
- Number of transport services which indicates the facility of transport (i.e. from one gambling venue to another)
- Number of medical services. This value is usually reported at the population density.
- Department and commune. People have preferences in terms of department and commune.
- HLM rented empty apartments. HLM indicates subsidized housing.

5. Evaluation

In order to evaluate the results, a further classification was made based on "Money_Laundering_Scoring". This file ranks the alerts by their risk score and then segments them in terms of their level of risk of being money laundering cases: top 5% is high risk, 6-15% medium risk and >15% low risk as can be seen from Figure 42.

Figure 42 Money_Laundering_Scoring

Data Set	Anomaly	Size	Risk	Level	Percentage	Description
Alerts	Anomalies	7 000	High	Level 1	5%	Suspicious Fraud Cases
			Medium	Level 2	15%	Anomalies Requiring Investigation
			Low	Level 3	80%	Normal Anomalies
Not Alerts	No anomaly	1 400 000	None			

*Note that the figures are illustrative

For gambling venues, as can be seen in Figure 43, in the first branch, all the proven fraud cases were selected (i.e. those cases already identified by the business through their standard detection methods). In the second branch, non-fraud cases were chosen. In total, 350 records were selected. Type Node is inserted to determine which are the attributes and target of the model. In this case, the attributes were all the indicator values (e.g. number of small gains) and the target was FRAUD. Then, a Partition node was inserted to create training set

and test set. 70% of the data was used for the training set, while the remaining 30% was used in the test set. Then, the model was created after applying the CHAID algorithm.

The accuracy, precision and recall for the test set were calculated, as can be seen in Table 6:

Table 6 Test Set

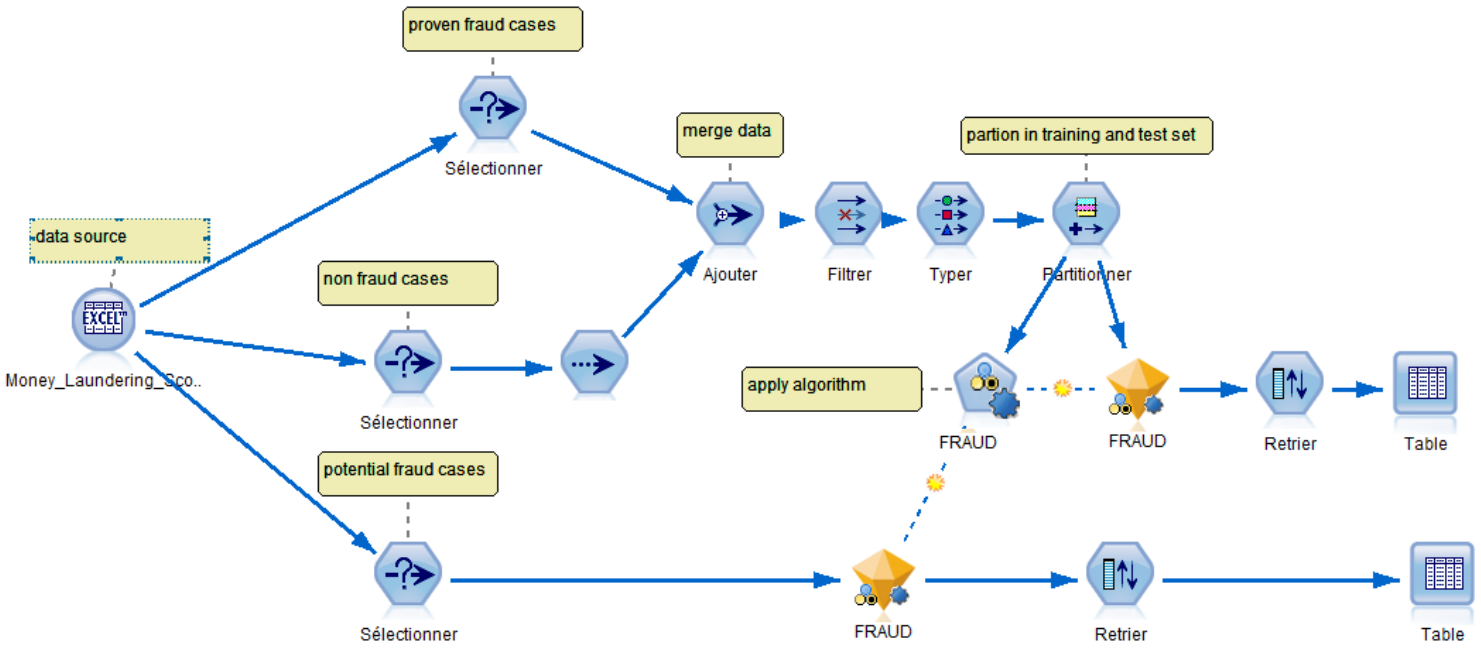
Measures	Values
Precision	0,89
Recall	0,41
Accuracy	0,93

AUC	0,97
Gini	0.95

According to Table 6, the precision is good (i.e. 89% of positive fraud predictions were correct) as well as the accuracy (i.e. 93% of fraud predictions were correct). The recall needs to be improved (i.e. only 41% of positive fraud cases were caught).

The Gain Chart related to testing set can be seen in Appendix S.

Figure 43 Evaluation

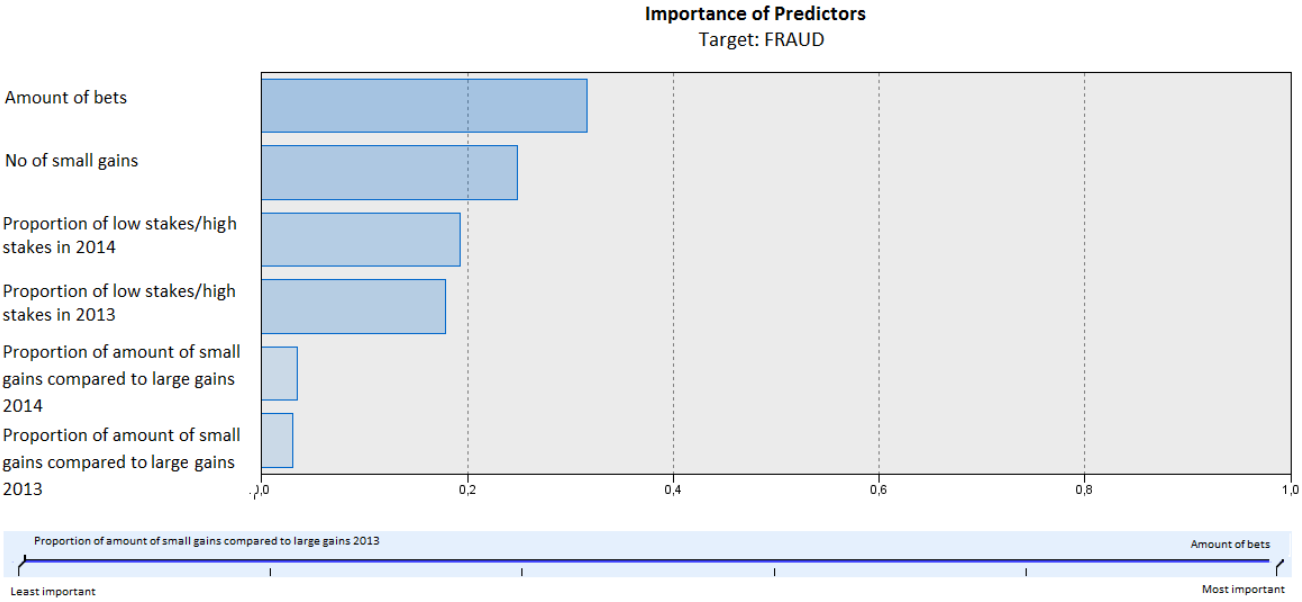


After creating the model, the scope is to be able to test the potential fraud cases which were selected in the third branch. These records are run through the model. The precision of the model is supposed to be 89% (taken from the test set).

Those cases that were deemed as anomalous by the model and having an anomaly score >=90% are most likely fraud. The efficiency of the detection methodology is deemed to be (number of fraud cases + number of predicted fraud cases with a high probability) / total number of cases.

The most important predictors of the CHAID model are amount of bets, number of small gains, proportion of low stakes/high stakes in 2014 and 2013, and proportion of amount of small gains / large gains in 2014 and 2013 as can be seen from Figure 44. As an observation, it can be stated the fact that the proportions of low stakes/high stakes for the two years (2014 and 1023) have relatively close values. The same can be said for the proportions of small gains/large gains for 2014 and 2013. But more recent values (i.e. 2014) always have higher impact on the final prediction than later values (i.e. 2013).

Figure 44 Importance of Anomaly Indicators



6. Conclusions and Future Work

The following conclusions can be mentioned:

- Outlier detection is highly suitable for distinguishing anomalous data from "good" data. Hence, it deserves more investigation.
- Visualization techniques have the power to render data anomalies. This makes the identification and also the quantification of fraud schemas much easier.
- The most pressing matter is bridging the gap between researchers and practitioners. This can be done by increasing the shared amount of sensitive financial data. This will have the benefit of increasing investigations of data mining techniques that could be applied to sensitive data (i.e. privacy-preserving data). Consequentially, the problem becomes a question of anonymizing the confidential/sensitive data.
- Indicators like frequency of gain, amount of small gains, number of gambling venues visited by the player, number of gains are highly crucial. Long bet series are also extremely important because they render a high risk of money laundering, mostly because they tend to be homogenous.

As future research, the following can be mentioned:

- Cases of bet series on all possible outcomes of an event (i.e. team 1 wins, draw or team 2 wins) are interesting to analyze.

- A possible different approach would have consisted in taking advantage of the known fraud cases in the detection phase, rather than using them in the evaluation part. But there are few such cases.

- A review of the weighting of the indicators should be considered. By analyzing the most frequent indicators that appear in the top 5% anomalous cases, a distribution of frequency may be constructed, in order to assign new weights to these indicators and observe the impact on the anomaly scores.

- Consider applying the anomaly detection algorithm on the beneficiary and the gambling venue clusters, as well as the entire dataset for insights on the data. This outlier detection on the clusters would prove beneficial to tailor the thresholds in order to better identify the anomalies per cluster.
- A more in depth analysis and selection of the INSEE variables before using them (e.g. the field "male population between 15 -64" can be further split into multiple age categories).
- ISA alert analysis w.r.t. gambling venue and beneficiary.
- Geographical analysis of the stakes (i.e. see the distribution of stakes per different geographical units like department, city, district etc. Specific community behavior may be identified.).

- The creation of indicators related to gambling venue turnover (e.g. proportion of 2014 turnover compared to 2013 turnover by type of bet and gambling venue, the same proportion but only considering gambling turnover, the turnover 6 months before and after the departure of the manager etc.)

- An analysis over the bet series validation history. It appears that money launderers validate the series then divide it into multiple sub series to be paid in different gambling venues. In this way, money becomes harder to track. Multiple validations can also reveal networks of people that buy winning tickets for money laundering purposes (e.g. they buy multiple winning tickets and cash their tickets by means of bank transfer. In this way, they could justify their revenues.)
- Further analysis of atypical or illogical behavior (i.e. bets made outside the working hours of a gambling venue).

Bibliography

13. (s.d.). Retrived from L'Institut national de la statistique et des études économiques: http://www.insee.fr/fr/

Influence du TRJ sur le blanchiment. (2012, 11 22). Consulted on March 14, 2015, sur Enonomie Gouvernement Francais: http://www.economie.gouv.fr/observatoire-des-jeux/influence-trj-sur-blanchiment

Basic Tree-Building Algorithm: CHAID and Exhaustive CHAID. (2015). Consulted on June 2015, 26, sur STATISTICA Help: http://documentation.statsoft.com/STATISTICAHelp.aspx?path=GXX/GCM/Overviews/BasicTreeBuildingAlgorithmCHAIDandExhaustiveCHAID

French Casinos. (2015). Consulted on March 15, 2015, sur Casinos Europe: http://www.casinoseurope.com/france/

24 Hours le Le Mans. (s.d.). Consulted on March 21, 2015, sur Wikipedia: http://en.wikipedia.org/wiki/24_Hours_of_Le_Mans

Arjel. (s.d.). Consulted on February 28, 2015, sur http://www.arjel.fr/-Liste-des-operateurs-agrees-.html

BetMinded. (s.d.). Consulted on March 20, 2015, sur http://www.betminded.com/bet-french-horse-racing-10955.html

Cahill, M. C. (2002). Detecting Fraud in the Real World. *Handbook of Massive Datasets*, 911-930.

Casino en France. (s.d.). Consulted on March 21, 2015, sur Wikipedia: http://fr.wikipedia.org/wiki/Casino_en_France

CGI. (2011). Implementing social network.

Chan, P. F. (1999). Distributed Data Mining in Credit Card Fraud Detection. *IEEE*, 14: 67-74.

Clifton Phua, V. L. (2010). A Comprehensive Survey of Data Mining-based Fraud Detection Research. 5.

Cortes, C. P. (2003). Computational Methods for Dynamic Graphs. *Journal of Computational and Graphical Statistics*, 950-970.

Ghosh, S. &. (1994). Credit Card Fraud Detection with a Neural Network. *Proc. of 27th Hawaii International Conference on Systems Science 3*, 621-630.

IBM Internal Documentation. (s.d.). *Noeud Classification TwoStep*. Consulted on June 2015, 2015, sur Aide de IBM SPSS Modeler: http://127.0.0.1:62857/help/index.jsp?topic=/com.ibm.spss.modeler.help/clementine/clusternode_general.htm

IBM. (s.d.). *Noeud Sélection de fonction*. Consulted on June 2015, 26, sur AIDE DE IBM SPSS MODELER: http://127.0.0.1:62857/help/index.jsp?topic=/com.ibm.spss.modeler.help/clementine/featureselectionnode_general.htm

Jain Anil K., Dubes Richard C. (1988). *Algorithms for clustering data.* New Jersey: Prentice Hall.

Kim, H. P. (2003). Constructing Support Vector Machine Ensemble. *Pattern Recognition*, 2757-2767.

Les Cles de la Banque. (s.d.). Consulted on March 18, 2015, sur http://www.lesclesdelabanque.com/Web/Cdb/Particuliers/Content.nsf/DocumentsByIDWeb/6WECHD?OpenDocument

Murad, U. &. (1999). Unsupervised Profiling for Identifying Superimposed Fraud. *Proc. of PKDD99*.

Phua, C. A. (2004). Minority Report in Fraud Detection: Classification of Skewed Data. *Explorations SIGKDD*, 50-59.

Appendix

Appendix A

Source	Name of the table	Description	Size	Usage
Bets	Bet	Details of the bets in the gambling venues	145 Mo	Yes
	Forecast	Details of the possible forecasts of bets	291 Mo	Yes
	Odds - Occurrence	History of odds for a given forecast of a bet	248 Mo	Yes
	Event	Description of the game event (e.g. Formula One)	57 Mo	Yes
	Group Event	Information on the group of the event (if there is one)	0,1 Mo	Yes
	Game type	Information on game type (e.g. horse racing)	0,1 Mo	Yes
	Transaction	Information on Transactions	135 Mo	Yes
	Occurrence Transaction	Link Table between Occurrences and Transaction	23 Mo	Yes
	Alert occurrence	Information on the number of times an Alert was risen	5 Mo	Yes
	Alert	Information on Alerts	2 Mo	Yes
	Alert Transaction	Link Table between Alerts and Transaction	19 Mo	Yes

	Extra	Supplementary Information on Bets	5 Mo	Yes
Huge Gain	Pile	Winning Bets	25 Mo	Yes
	Payment	General Information on the payments which were carried out (who, what, where, why)	10 Mo	Yes
	Operation	Detail of the Financial Operation that supported the Payment (bank transaction, check or cash)	10 Mo	Yes
Calog	Alarm type	Reference Table of the type of Alarm (60 types of Alarm)	20 Mo	Yes
	Alarm Group Criteria	Definition table of Regrouping Criteria of an alarm occurrence (central criteria : type of Alarm, day, group of attributes)	10 Mo	Yes
	Attribute type	Table which references the different types of Attributes which define the occurrence of an alarm	5 Mo	Yes
	Alarm Attribute	Attribute definition of an Alarm	10 Mo	Yes
	Alarm	Main table of Alarms (level, status, date of last occurrence, ID gambling venue, game ID)	5 Mo	Yes
	Alarm Level	Reference table of level of Alarms	5 Mo	Yes
	Alarm Status	Reference table of status of Alarms	5 Mo	Yes
	Alarm Occurrence	Main table	15 Mo	Yes
	Alarm Occurrence Attribute	Definitions of the attributes of an occurrence	20 Mo	Yes
Transaction	Main Transaction	General Characteristics of a Transaction (regulation, cancelation,	2,3 Go	Yes

	Base	betting)		
	Linkage	Link Table between Transactions (a record exists if there is a link between 2 transactions: betting and regulation, betting and cancelation …)	278 Mo	Yes
	Wager Base	General Characteristics of Betting	1,1 Go	Yes
	Wager Board Base	Information of Types of Games	1,8 Go	Yes
	Gambling Venue Base	Detail of a Gain of a Bet	146 Mo	Yes
	Value Base	Payment Transaction or "In Query" Transaction	428 Mo	Yes
	Value Gambling Venue Base	Details of the Payment Transaction	91 Mo	Yes
	Transaction Base	Payment Transaction	0,1 Mo	Yes
	Cancel Base	Cancelation of a bet	6,4 Mo	Yes
	Regulation Base	Regulation of a bet	20 Ko	Yes
	Main Transaction Base	Non sport bets	4,4 Go	Yes
Gambling Venues	Gambling Venue	General Information about the Gambling Venue (type of contract, rights, duration) at a given moment	32 Mo	No
	Address	Physical Address of the Gambling Venue	33 Mo	Yes
	Blockage	Table used for the information of different blockages imposed on the Gambling Venue (e.g. roulette blockage, contract suspension)	1,6 Mo	Yes

	Agreement Type	Table referencing possible Agreements	0,1 Mo	No
	Geo	Geographical Characteristics of Gambling Venues	0,1 Mo	Yes
	Gambling Venue Geo	Link Table between Geo and Gambling Venue	6 Mo	Yes
	Teller	Optional Information associated to the technician	45 Mo	Yes
	Device	Optional Information associated to the machine	31 Mo	Yes
DGRS	Small Gain	Information on the small gains from gambling centers	40 Mo	Yes
RDSA	Weekly Sales	Data Aggregations made by the Company	8,5 Mo	Yes
Geo Concept		INSEE (13) Open Data revealing information on the gambling venue	780 Mo	No
Head Responsible	Inspection Criteria	Details on the Inspection Criteria	940 Mo	No
	Inspection Results	Results of the last Inspection of a Gambling Venue. If a Gambling Venue is inspected more than twice per year, this will result in contract suspension	16 Mo	No
	Gambling venues	List of Gambling Venues present here	1,2 Go	No
DGRS	Inspection Report	List of all inspection criteria	5 Mo	No
Risk		Information File about gambling venues, followed by the gambling venues which are considered to be at risk	0,1 Mo	Yes

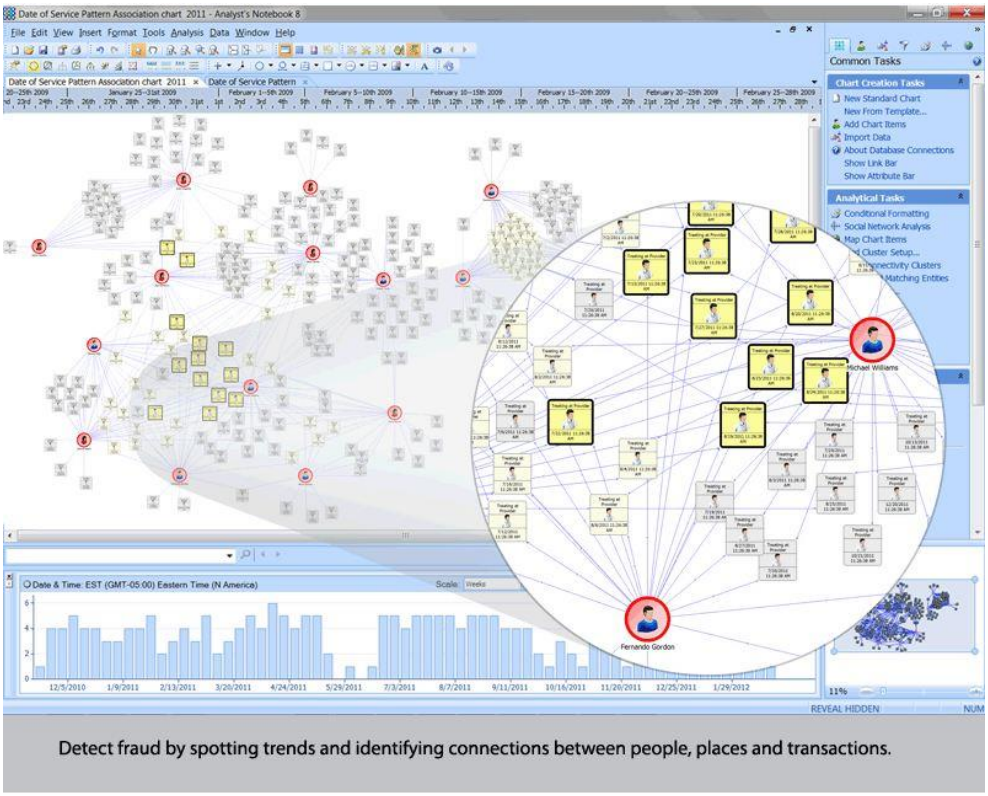
Appendix B

Phases	Comprehensive View	
Phase 0	Name	Simple Profiling
	Scope	This phase consists of creating detection indicators to facilitate the identification or the confirmation of atypisms based on atypical transaction(s).
	Description	The first half of the phase will regroup the existing information connected to *gambling venues* and *retailers.* The following indicators will be considered: o Basic statistics, for example: • The number of gambling venues/retailers by department • The average amount and frequency of bets by gambling venue/retailer • The standard deviation of the previous indicators in order to detect atypisms. o indicators which are specific to a certain domain, for example: • the average number of bets by gambling venues/retailers • the average number of bets by day, month, year by gambling venues/retailers • idem for the amount of bets The second half will regroup the existing information connected to beneficiaries. The following indicators will be constructed: o Basic statistics, for example: • The number/average number of beneficiaries by gambling venue • The average number of gains/frequency by players/gambling venue • The standard deviation of the previous indicators in order to detect atypisms o indicators which are specific to a certain domain, for example: • The average, maximum of beneficiaries by gambling venue/department by day/month/year • The average number and amount of money of bets by day/month/year by gambling venue/department
Phase 1	Name	Stake Analysis
	Scope	This phase consists of analyzing the atypical bets in terms of amount of money and odds
	Description	The scope is to analyze the bets in order to detect atypisms in terms of amount of money and stakes. The focus will be on the bets whose odds are low and their evolution in time. The following analysis is proposed: 1. Analyze the low odds bets, construct the distribution of bets w.r.t. their stakes and odds. 2. Calculate the % of low odds compared to the overall % of odds by gambling venue. 3. Analyze the atypical/illogical behavior (e.g. players betting against their local teams) 4. Analyze the location of bets (few gambling venues register the same bet more than 80% of the time)

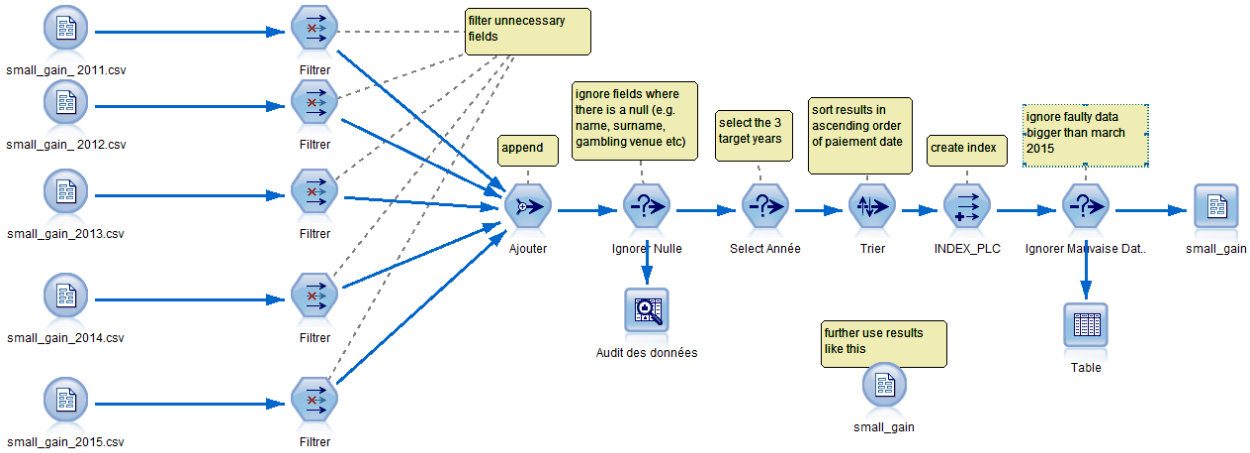
		Options: 5. Analyze the temporal atypisms which are repeated in the same gambling venue in order to detect the time intervals of atypical bets, in terms of amount and frequency. These analyses will be conducted w.r.t. gambling venues/retailer. The atypisms that will thus be discovered will permit the construction of money laundering indicators.
Phase 2	Name	Series Analysis
	Scope	This phase aims at improving the atypical series analysis in real time and in batches to detect the money laundering operations
	Description	In the present moment, the series (i.e. series of bets) that are detected consist of series of bets whose value is greater or equal to 900 euros in a given amount of time (e.g. 180 seconds for example) by gambling venue and by bet. The moment such a series is record, called ISA alert, is triggered. The following analysis will be proposed: 1. Analyze the atypical series according to the odds of the bet 2. Analyze the distribution of (a) typical series w.r.t. amount of money, type of transaction and number of transactions within the series, duration of a series 3. Differentiate the statistical parameters (duration and amount) according to the geography and/or context (important events drawing significant number of bets) 4. Detecting elements that interrupt series (i.e. elements that are inserted within a series of bets to make them harder to detect by the company's alert system) These analyses will be conducted w.r.t. gambling venues/retailer.
Phase 3	Name	Complex Profiling
	Scope	This phase consists of creating profiles of gambling venues/beneficiaries in order to identify or confirm atypisms
	Description	The main idea is to cluster each analysis domain. 1. Gambling venue / retailer 2. Beneficiary The approach is defined by using data mining solutions as well as the client's knowledge to create clusters. For each cluster, a profile will be created which will include basic statistics, specific indicators but also other indicators concerning the series and the stakes (from previous phases). These profiles will allow the understanding of the different types of behavior (clusters) which exist in each domain and evaluate the threshold alert relevance w.r.t. each cluster. By using a threshold, the analyst will be able to draw a line between normal behavior and suspicious (potentially fraudful) behavior. In addition to this, these profiles will provide recommendations to improve the thresholds to detect atypisms. These thresholds will be different for each profile/cluster.
Phase 4	Name	Simple Scoring
	Scope	This phases consists of a creation of an overall anomaly score and a correlation with money laundering cases

	Description	The main idea is to create an overall anomaly score. The set of indicators will be used to produce a mixed score: each indicator will be weighted by a number of points given by the experts in money laundering domain. This number of points will indicate the importance of the indicator w.r.t. the money laundering phenomena. The sum of the weighted indicator scores per gambling venue/beneficiary will provide a score at gambling venue/beneficiary level. This analysis will allow the verification of indicator relevance. If it is the case, the score will allow the detection and in the same time automatic prevention of money laundering cases.

Appendix C

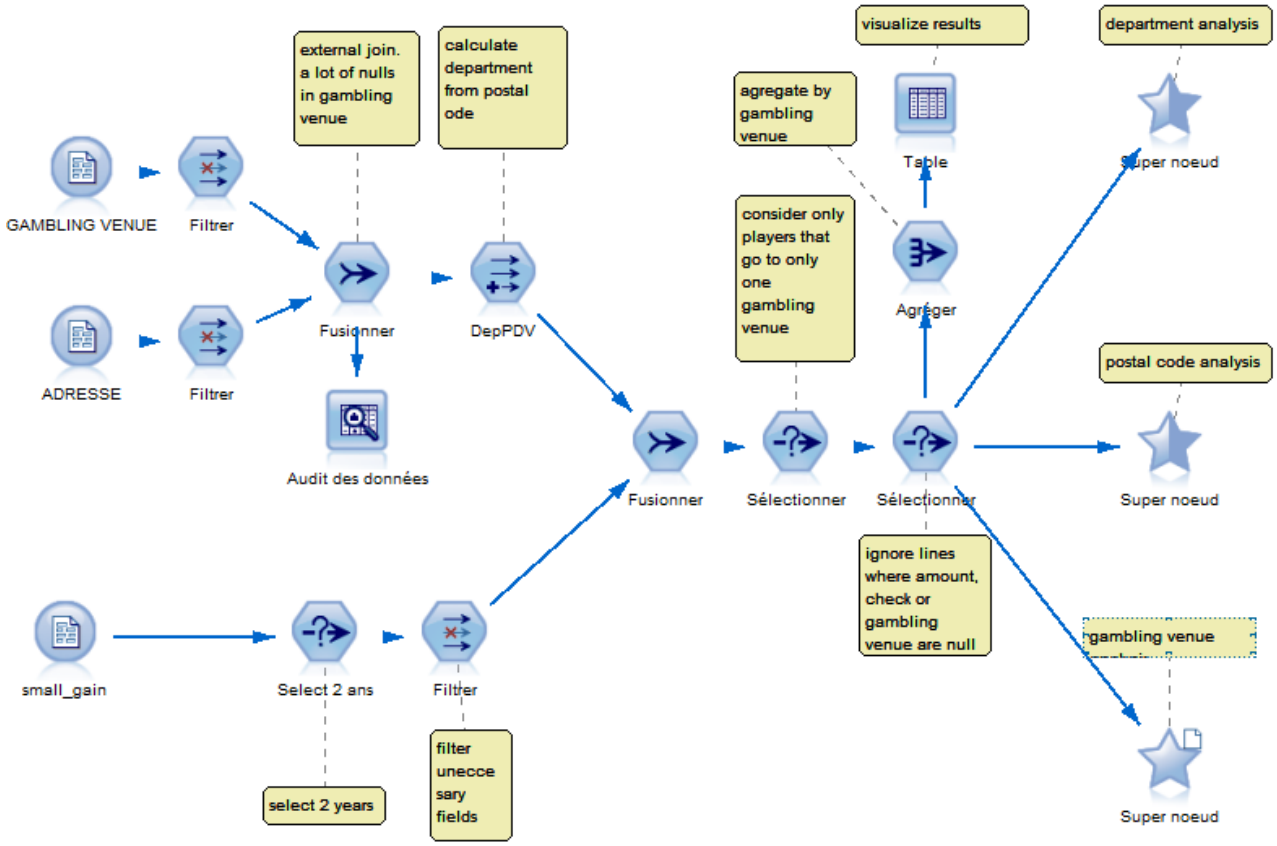


Appendix D

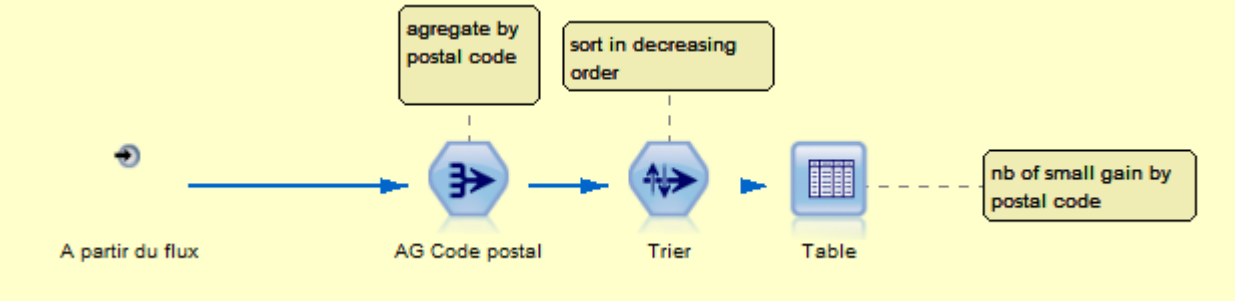


The inputs in this case are two .csv files.

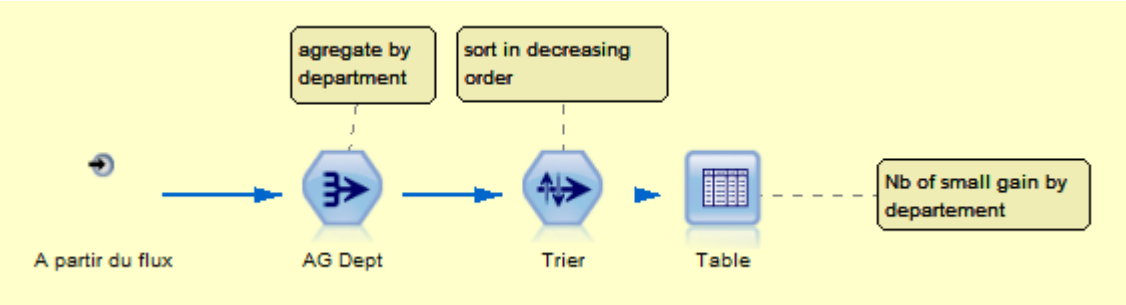
Appendix E



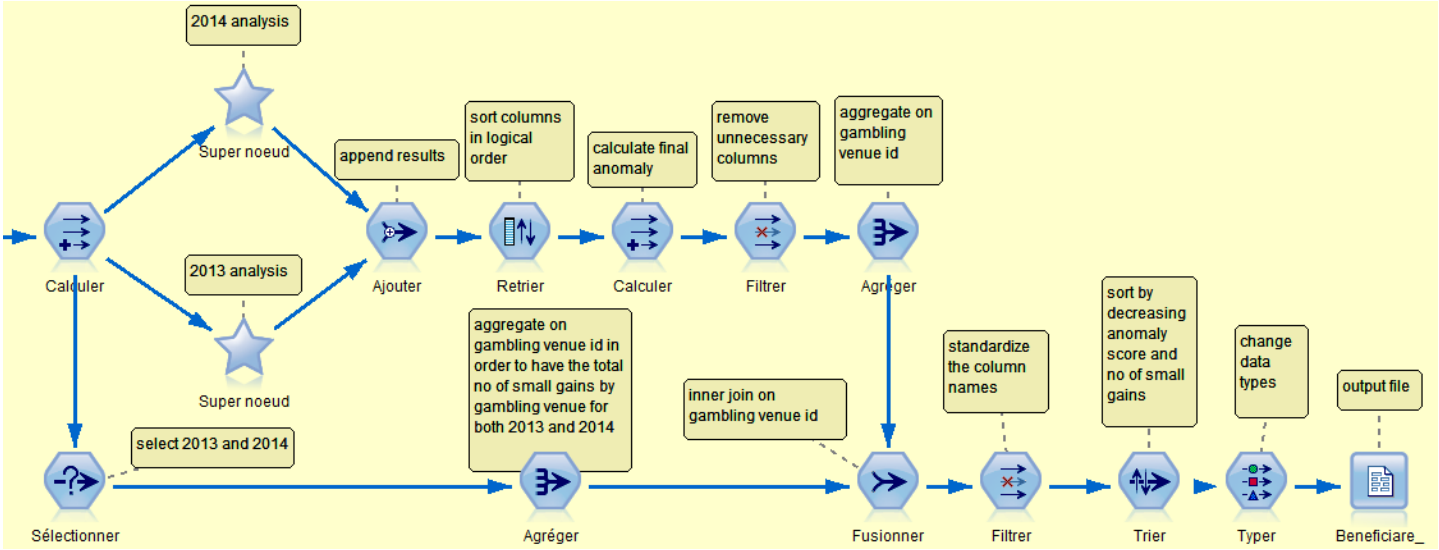
Appendix F



Appendix G

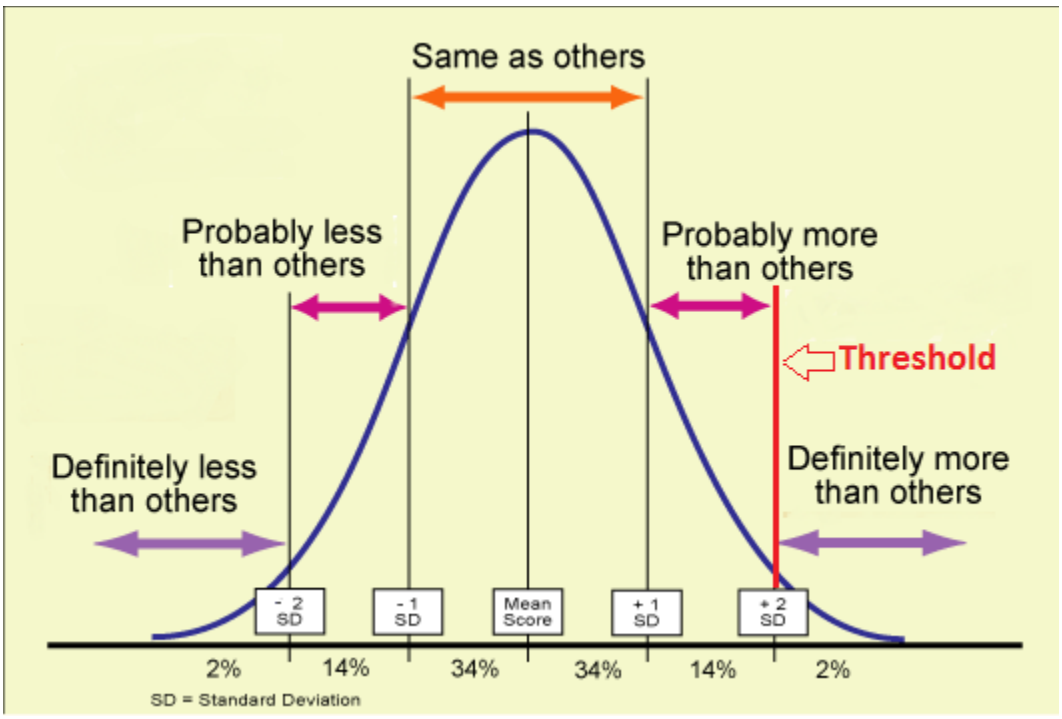


Appendix H

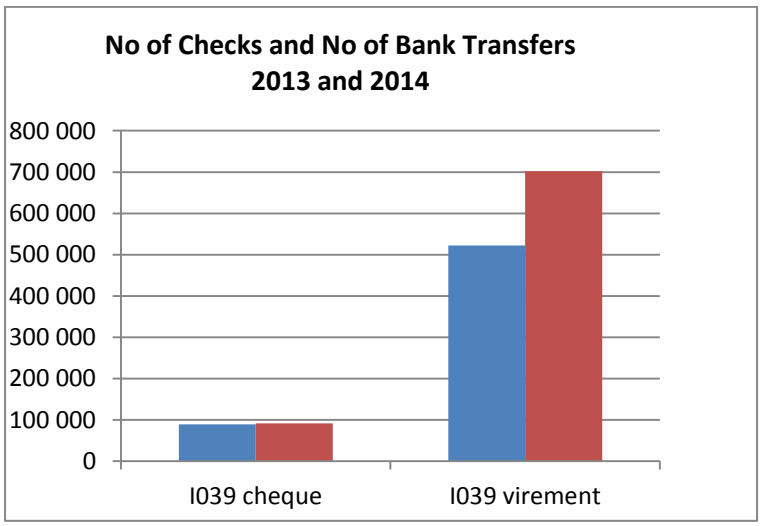


For the aggregation on gambling venue, the total anomaly score for each gambling venue was also constructed with respect to the 2 years in the data range (i.e. 2014 and 2013) in a similar way as the previous case. The analysis of the 2 years are encapsulated in 2 Super Nodes called ”Analyses 2014” and “Analyses 2013” (which appear as two stars) as can be seen in Appendix H.

Appendix I

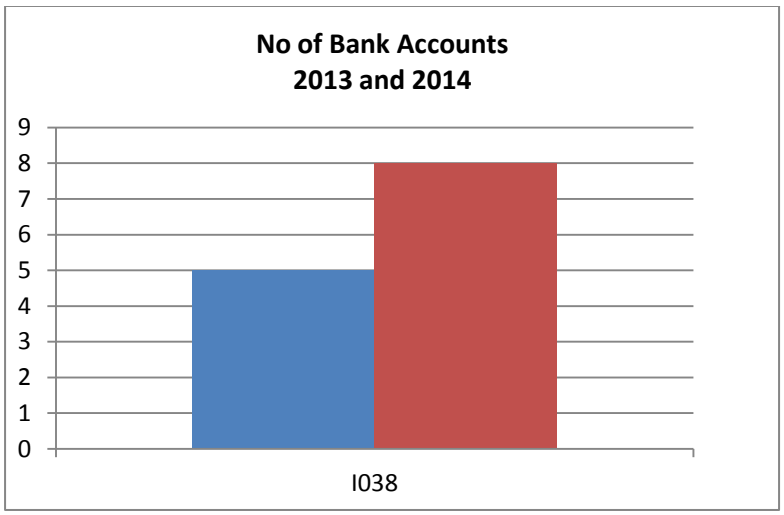


Appendix J



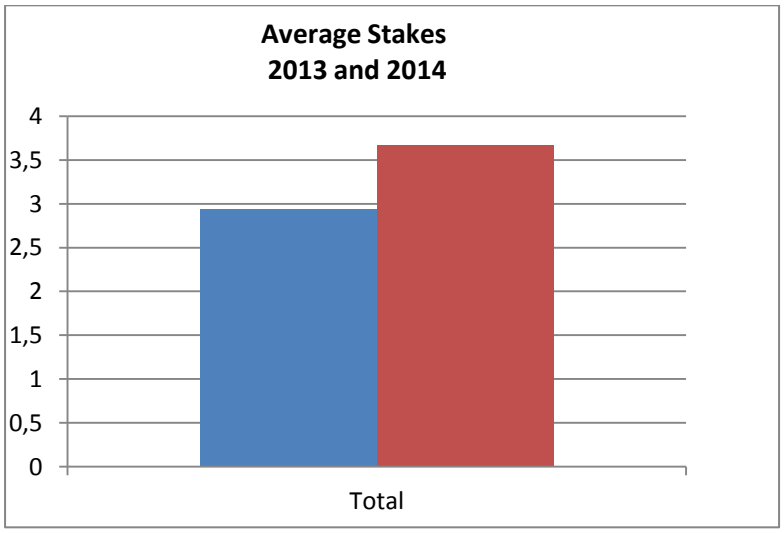
From Appendix J, there is a clear tendency for beneficiaries to be paid by bank transfer.

Appendix K



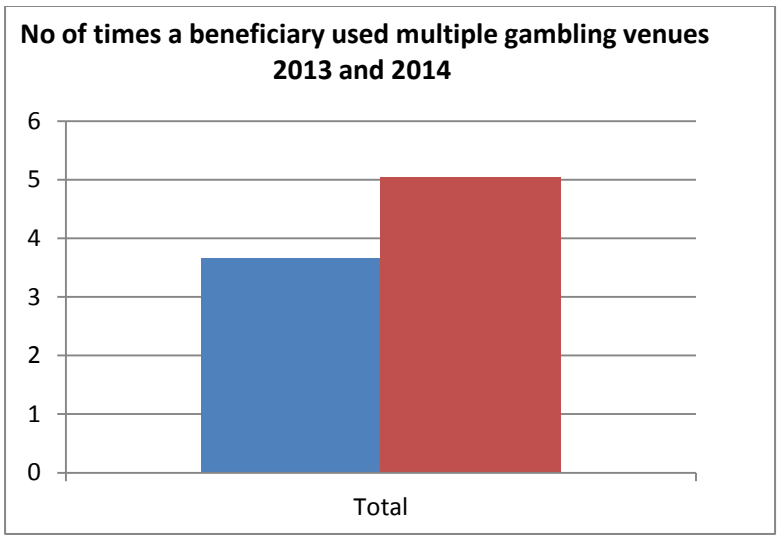
As can be seen from Appendix K, in 2014 some beneficiaries distribute their gain over 8 different bank accounts, a clear increase when compared to 2013 when the maximal number of bank accounts for a beneficiary was 5. Unless he wanted to take advantage of multiple banking offers, it is strange why a beneficiary would have so many different accounts.

Appendix L



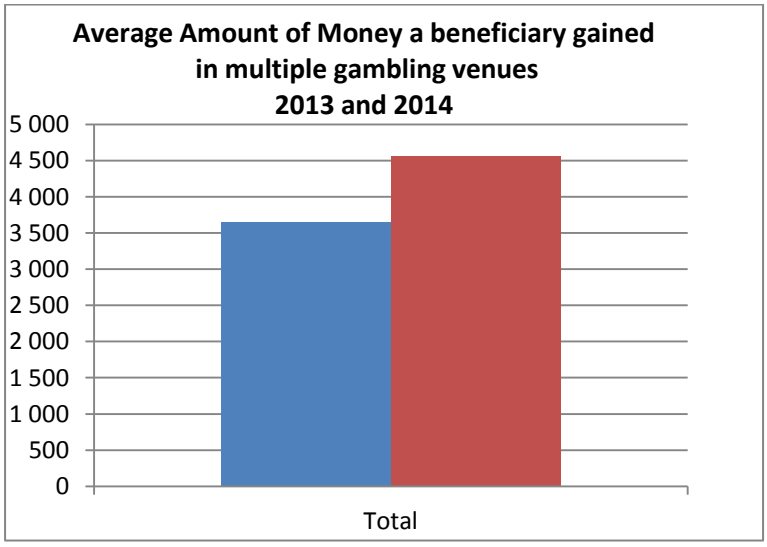
In Appendix L, the average stakes in 2013 are almost 3 while in 2014 it is about 3, 6. Note that money launderers tend to make bets on low stakes; hence the ones that are targeted are those whose bets are generally below the average value of stakes. It does not make sense however to construct a chart showing the lowest stakes that potential money launderers bet on, because by default, the value is 1, 1.

Appendix M



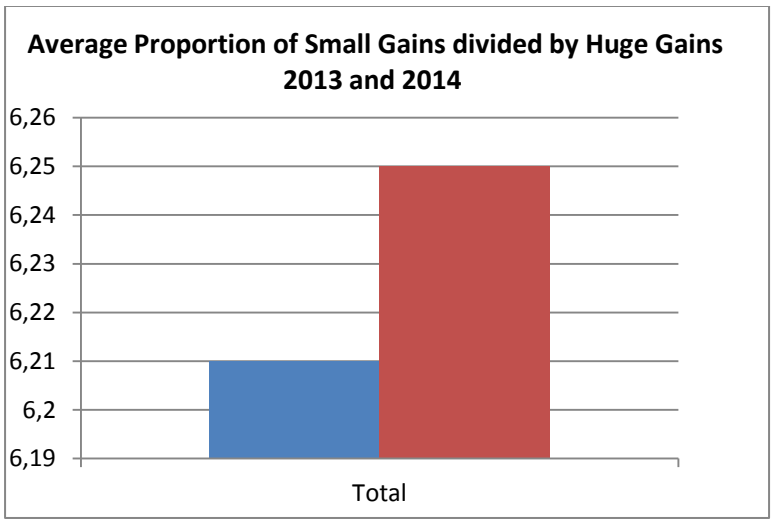
Appendix M demonstrates the fact that the number of times a winner decided to cash in his revenues in a different gambling venue then the one he purchased his tickets, is 25% larger in 2014 than in 2013. While it is not strange for a passionate player to visit multiple gambling venues, some money launderers use many gambling venues in the same time to distribute their gains. It is a way for them to hide their large gains (i.e. "getting lost in the crowd")

Appendix N



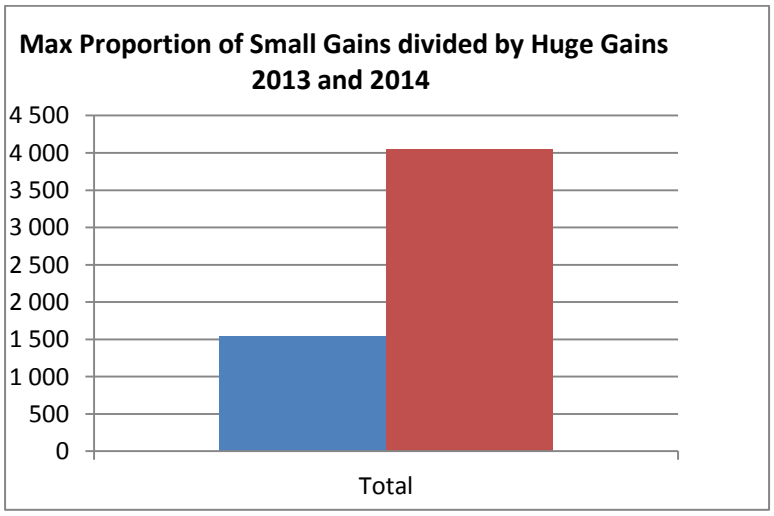
This increase can also be seen in the amount of money people cash in. As can be seen from Appendix N, the increase between 2014 and 2013 is about 20%.

Appendix O



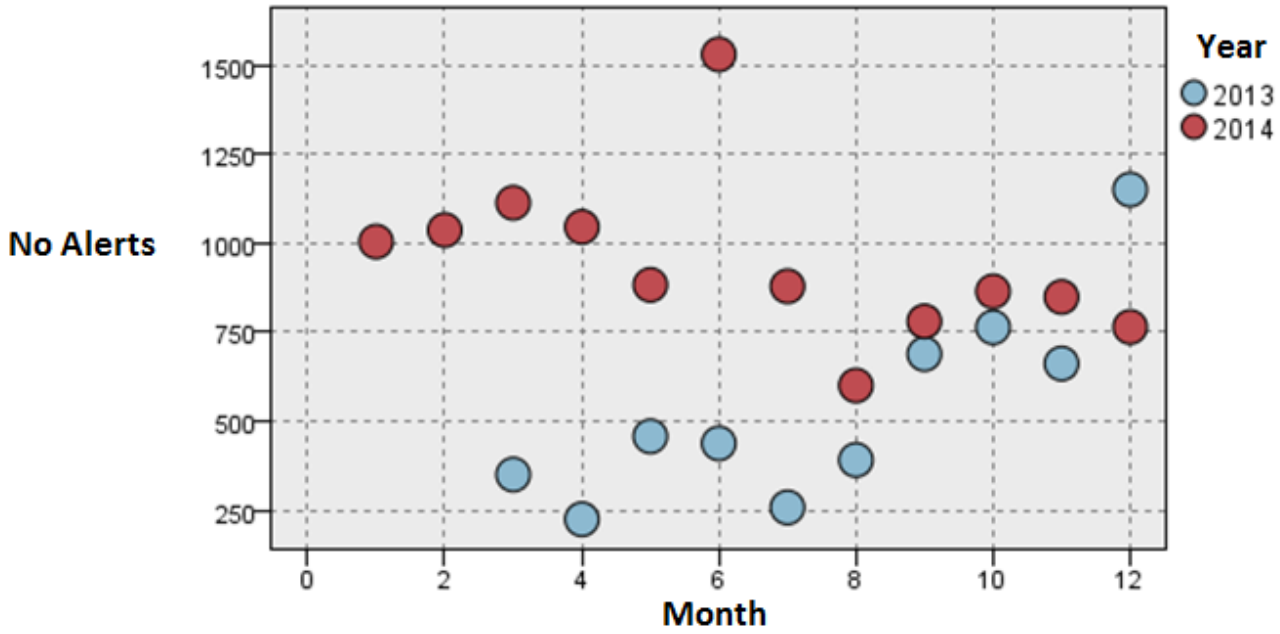
In Appendix O, the proportion in 2014 is around 6, 25 while in 2013 it is about 6, 20.

Appendix P

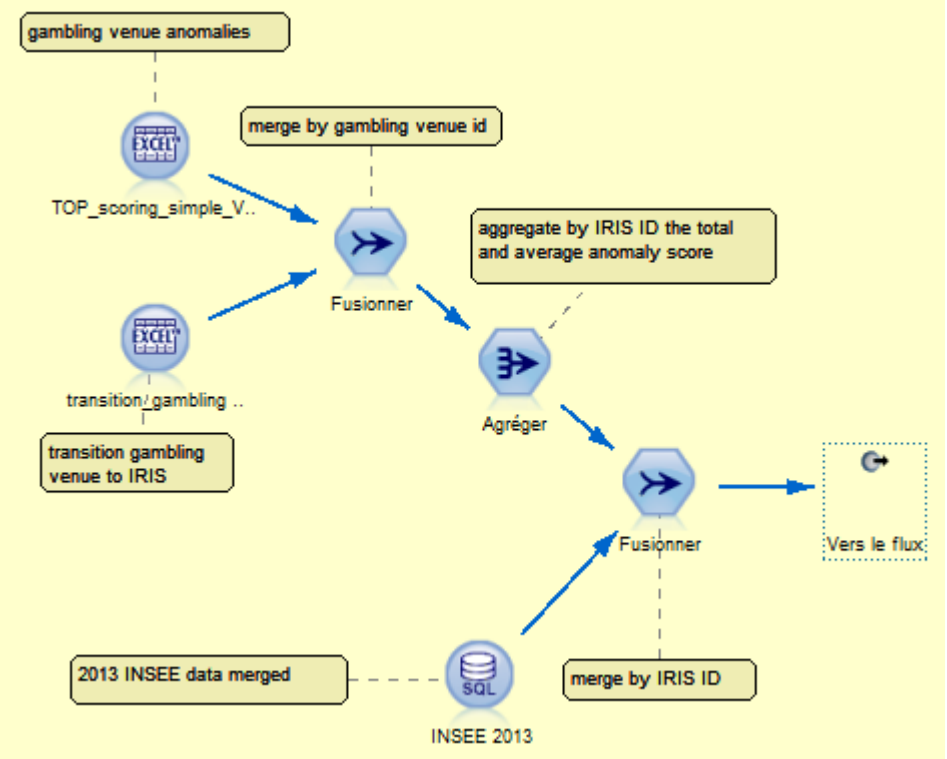


In the case of maximal values, the proportion in 2013 is 1500 and in 2014 it soars to 4000.

Appendix Q



Appendix R



As can be seen from Appendix R, which renders the data sources for the 4. Additional Analysis, the analyst makes use of the file “Top_scoring_simple” and merges it with “transition_gambling venue_iris” by means of gambling venue id. The set is later merged with the INSEE 2013 Data by IRIS ID.

Appendix S