Big Data and Information Analytics
©American Institute of Mathematical Sciences
Volume **2**, Number **2**, April **2017**
doi:10.3934/bdia.2017015
pp. **141–175**

IDENTIFYING ELECTRONIC GAMING MACHINE GAMBLING PERSONAE THROUGH UNSUPERVISED SESSION CLASSIFICATION

Maria Gabriella Mosquera and Vlado Keselj

Faculty of Computer Science, Dalhousie University
6050 University Avenue
Halifax, NS B3H 4R2, Canada

(Communicated by Aijun An)

Abstract. The rising accessibility in gambling products, such as Electronic Gaming Machines (EGM), has increased interest in the effects of gambling; in particular, the potential for impulse control disorders, such as problem gambling. Nevertheless, empirical research of EGM gambling behaviour is scarce. In this exploratory study, we apply data mining techniques on 46,416 gambling sessions, collected in situ from 288 EGMs. Our research focused on identifying the at-risk behavioural markers of sessions to help distinguish gambling personae. Our data included measures of gambling involvement, out-of pocket expense of sessions, amount won, and cost of gambling. This research, discusses the methodology used to collect and analyze the required gambling measures, explains the criteria used for identifying valid sessions, and combines outlier mining methods to identify instances of heavily involved gambling (i.e., outliers). Our results suggest that sessions were classified as potential non-problem, potential low-risk, potential moderate risk, and potential problem gambling sessions. Further, outlier sessions were more heavily involved in terms of gambling intensity and amount redeemed, despite having low duration times. Finally, our methods suggest that the lack of player identification does not prevent one from identifying the potential incidence of problem gambling behaviour.

1. **Introduction.** Due to the increasing accessibility of gambling products and the rising popularity of Internet gambling [13], [26], [32], [45], interest in the effects of gambling has grown over the past decade. Within the gambling industry, Video Lottery Terminals (VLT) and Electronic Gaming Machines (EGM) are considered the dominant segment [3]. EGMs, in particular, can be found in different types of venues, such as bars, restaurants and hotels, they can hold several types of games within a single machine, and can attract a high number of gamblers due to their structural characteristics (e.g., flashing lights, music, fast paced type of play) [41]. For example, in 2009, Australian gamblers spent $19 billion on gambling products [3], 55.3% of that revenue was generated through EGMs in clubs and hotels [3]. In 2010, EGMs generated 32.9% [16] of the Canadian gambling industry's $13.74 billion revenue [33]. By 2012, EGMs generated 65.32% of the Canadian gambling industry's $13.87 billion revenue, and EGMs located in bars and lounges

2010 *Mathematics Subject Classification.* Primary: 91C20, 62H30; Secondary: 03C45.
Key words and phrases. Behaviour analysis, unsupervised session classification, clustering, gambling personae, problem gambling.

were responsible for 47.8% of that total revenue [42]. Clearly, there has been an increase in the acceptance and accessibility of gambling products, as evidenced by the increase in revenue. Nevertheless, this growth has also increased the potential for impulse control disorders, such as problem gambling.

Problem gamblers, in particular, have difficulty limiting themselves from gambling excessively, regardless of the detrimental consequences that their actions can cause to themselves or others (e.g., family, friends, colleagues) [9], [32]. However, there is little consensus in regards to the prevalence rate of this disorder. Standardized prevalence rates of problem gambling, according to Williams et al. [52], range from 0.5% to 7.6%, with the lowest prevalence rates occurring in European countries and higher rates in Asian countries. According to Williams et al. [52], Sweden, Switzerland, Canada, Australia, Italy, United States, Estonia, and Finland share a prevalence rate on par with the worldwide average of 2.3%. While the root cause of the disorder is unknown, research suggests problem gambling is likely to occur parallel to other behavioural problems, such as substance abuse, eating disorders, or compulsive shopping [35]. In regards to its incidence, while problem gambling can occur in every demographic group [3], [13], the disorder does seem to be more predominant in men than women [45]; people of lesser means are at a greater risk of problem gambling than individuals in a higher socio-economic status, as they see a greater potential for financial gain and stability in gambling; and due to their increased accessibility to gambling products, casino workers are also at a higher risk for developing this disorder. Problem pathological gamblers, on the other hand, are considered to suffer from a severe type of problem gambling [9].

The American Psychiatric Association (APA) considers problem pathological gambling as an impulse control disorder due to the pleasure that the gambler obtains from the act of gambling [2]. The APA has defined ten criteria to guide in the process of diagnosing problem gambling (i.e., preoccupation, tolerance, withdrawal, escape, chasing, lying, loss of control, illegal acts, risked significant relationship, bailout) [35], some of which are similar to characteristics found in other impulse disorders such as substance abuse (e.g., tolerance, withdrawal, loss of control, preoccupation, mood alteration). However, not all gambling behaviour results in problem gambling, as the disorder can range from at-risk, problem, sub-clinical, pathological, probable pathological, extremely pathological, in-transition, and compulsive gambling [45]. The Canadian Problem Gambling Index (CPGI) [15], which includes questions centered on the factors contributing to gambling addiction [10], [13], is often used to study the social context and predominance of gambling and problem gambling. Though there is a growing need for studies from which to gather a greater understanding of the expression of the disorder [36], research focusing on actual gambling behaviour is limited [46], and empirical research of EGM gambling behaviour is particularly scarce. Thus, in this initial exploratory study we look into the application of data mining techniques on EGM gambling data, with the goal of identifying the at-risk behavioural markers of EGM gambling sessions and distinguishing types of gambling (i.e., gambling personae) based on the behavioural characteristics of gambling sessions.

To identify the gambling involvement measures [5], [6], [26], [55], that can serve as behavioural markers of EGM gambling sessions, our research uses EGM gambling measures collected in situ, over a one-month period (i.e., July 2010), from a single EGM manufacturer. Gambling measures, generated during a session, can provide more information about gambling events, such as game titles, wagered amounts, bet

outcomes, bonus rounds activity, out-of-pocket cost of a session, and amount cashed-out at the end of a gambling session. Gambling involvement measures, in particular, have been associated with excessive gambling behaviour [5] and can be measured in regards to financial involvement (e.g., amount wagered), time involvement (e.g., duration), gambling intensity (e.g., bets, bets per minute), and gambling cost (e.g., net loss) [26]. By conducting an analysis of gambling measures, our study aims to provide a better insight into the expression of EGM gambling behaviour and identify potential instances of problem gambling, as individuals suffering from this impulse control disorder may, at times, feel the need to gamble excessively (e.g., larger bets, bigger risks) in order to make up for previous losses [2].

However, the use of EGM data does pose some limitations on this research. For example, as players did not use logins or loyalty cards on the EGMs from which the data was collected, it was difficult to assess if a single subject generated a single gambling session or whether a single subject produced multiple sessions, thus this research did not assume an independence of sessions nor did it attempt to identify individual gamblers; instead, this research focused on identifying gambling personae. Furthermore, similar to LaBrie et al's [26] study of Internet sports gambling behaviour, in this initial exploratory study, we used aggregated measures of EGM gambling. While this approach allowed us to identify measures of gambling involvement in EGMs, the simplicity of this data limited our efforts to perform further analysis that could provide more detailed information in regards to each session (e.g., wager variability, trajectory of wagers, bonus round activity). However, the use of anonymous EGM gambling data has the potential to provide a better understanding of EGM gambling activity than what can be gathered through surveys, as it minimizes the risk of inaccurate results due to evaluation apprehension, as well as self-presentation and recall bias [26], [29].

The remaining sections of this article are organized as follows: In Section 2, a discussion of the background and related literature is provided with the aim of presenting the reader with a view of the research available in the field of gambling studies; Section 3 presents a methodology for unsupervised EGM gambling session detection, and describes our process for preprocessing the identified EGM sessions. Section 4 provides the clustering methodology for identifying the at-risk behavioural markers of EGM gambling sessions from which to recognize cases with similar behavioural characteristics. Section 5 provides a discussion of the identified types of EGM gambling. Finally, Section 6 describes the differences between normal sessions and outliers (i.e., heavily involved gambling sessions), in order to determine the likelihood of a session being assigned to a particular cluster based on its behavioural characteristics.

2. **Background and related work.**

The growing accessibility of gambling products [13], [32], [45], and rising popularity of Internet gambling [26], has increased interest in the affects of gambling over the past decade [13]; particularly, due to the potential for health risk factors for impulse control disorders such as problem gambling. And though the root cause of the disorder is unknown [35], the likelihood of the disorder to occur parallel to other behavioural problems (e.g., substance abuse, eating disorders, shopping addictions) is quite high [3], [13]. However, despite the growing need for studies that help understand excessive gambling [36], there is little research focusing on the analysis of actual gambling behaviour [46]. For example, the CPGI [15] survey, includes questions centred on the factors contributing to

gambling addiction such as the nature of gambling products, gambling experience, accessibility, anonymity, affordability, interactivity, and convenience [10], [13], and is used to study the social context and predominance of gambling, and problem gambling.

Individual's, based their CPGI score, can be classified into one of five categories: non-gambling, non-problem gambling (score=0), low risk gambling (score=1-2), moderate risk gambling (score=3-7), and problem gambling (score=8-27) [3]. Other standards, such as the South Oaks Gambling Screen (SOGS) [3], the National Opinion Research Center Diagnostic and Statistical Manual of Mental Disorders Screen for Gambling Problems (DSM-IV) [2], [32], and the Gambler's Anonymous Scale (GA20) [2], [32], also tend to be used in conjunction with the CPGI survey. However, the use of survey tools as the only means for assessing problem gambling can sometimes produce inaccurate results due to threats to construct validity [26], [29]. Controlled studies have also been done in order to assess the impact of external factors on gambling behaviour such as alcohol consumption and music tempo [12], [14], [31], [48], with results showing positive correlations between alcohol consumption and gambling time, and music tempos and gambling intensity. However, the use of laboratory studies for analyzing gambling behaviour, may impose limitations on the generalizability of the obtained results, as these studies lack the realism of in situ behaviour.

Observational studies by Harrigan and Dixon [20], Dixon et al. [11], and Harrigan [21], [22], have focused on analyzing the impact of the structural characteristics of slot machines on gambling activity (e.g., illusion of control, entrapment, frustration, near misses), with results indicating that certain structural characteristics, such as stop buttons, bonus modes, hand-pays, and 'near-misses', can lead to a player's increase in gambling involvement, and the development of inaccurate beliefs in regard to personal skill and win probability. Though these observational studies provide an insight into the impact of external factors on gambling, they do not focus solely on the gambling patterns generated from the observed gambling activity.

While conducting this literature review, it was evident that studies focusing on EGM gambling behaviour are scarce in the field of data mining; perhaps due to the lack of player identification (i.e., player ID) in EGM data, as these machines often require cash rather than logins or loyalty cards, which can be an obstacle when attempting to identify EGM gambling sessions. Nonetheless, the use of data mining techniques for behaviour analysis [4], [43], has, in the past, been successfully applied for facilitating behaviour classification and identification, such as customer classification (e.g., loyal, discount, opportunistic, wandering, need-based, impulse) [43], identification of at-risk behaviours (e.g., at-risk academic performance, credit risk evaluation) [38], [53], and the recognition of negative risk-taking behaviour such as dangerous driving or substance abuse [34].

This exploratory research focuses on the application of data mining techniques on aggregated measures of EGM gambling, with the goal of identifying the measures of gambling involvement that can serve as the behavioural markers of EGM gambling sessions. These gambling involvement measures allow for the behavioural characteristics of sessions to be explored. The use of real EGM gambling data may increase the likelihood of generalizing our results to the general EGM gambling population. However, unlike Internet live-sport gambling data, EGM gambling data does not contain player identification information, which makes it difficult to identify whether

a single individual generated a single session or whether they generated multiple sessions. As such, the lack of player IDs, as EGM gamblers do not tend to use logins or loyalty cards, limits our ability to identify individual gamblers as well as assume an independence of sessions, and suggests the need for unsupervised session detection in EGM data. Therefore, in this research, it was of particular importance to define what constitutes an EGM gambling session. While unsupervised session detection methods have been used for web session detection [30], the structure of EGM messages and the communication protocol used in these machines [17], [18] indicated that specifying time thresholds would not be suitable for the purposes of this research. Instead, specific gambling events could help determine a criterion for defining an EGM gambling session. Once sessions are identified, suitable data preprocessing and transformation techniques (e.g., smoothing, normalization, aggregation) can be applied to increase the quality of our results [19] by providing other variables (e.g., bets per session, cost of gambling, ratio of losses) to assist in our analysis.

Measures of gambling involvement have been associated with excessive gambling behaviour [5], i.e., increasing gambling involvement expressed during a session [36], and can be measured in regards to financial involvement (e.g., amount wagered), time involvement (e.g., duration), gambling intensity (e.g., bets, bets per minute), and gambling cost (e.g., net loss) [26]. In this regard, by conducting an analysis of gambling measures, our study could provide a better insight into the expression of gambling behaviour and identify potential instances of problem gambling, as sessions generated by subjects suffering from this impulse control disorder may incur larger bets or take bigger risks in order to make up for previous losses [2].

In the first of a series of longitudinal studies of Internet sport gambling data, and the only study found to be closely related to the research at hand, LaPlante et al. [29] found certain measures of gambling involvement, such as intensity and frequency, could significantly contribute to the incidence of problem gambling behaviour. Later on, LaBrie et al. [45] found gamblers who imposed limits on their gambling activity, incurred longer duration times than the rest of the sample, despite decreasing their total amount wagered. These findings suggested the importance of session duration (i.e., game time) as another measure of gambling involvement. Subsequently, LaBrie et al. [27] found Internet gamblers who played casino-style games (e.g., slots) incurred larger gambling costs (i.e., net loss) despite playing less than sports bettors, and suggested net loss and total amount wagered as important measures for gambling involvement.

Further studies of Internet live sport gambling [7], [55], focusing on analyzing the betting patterns of gamblers, have also identified gambling intensity, gambling frequency, variability of bet sizes, and the trajectory of gambling activity, to be important variables for analyzing problem gambling, as they take into consideration the general gambling behaviour of at-risk players, such as overconfidence from early large winnings and increasing bet sizes to achieve the same excitement experienced after their first large win [13], and are consistent with the personality traits of problem pathological gamblers, such as negative urgency and sensation seeking [32], [34]. Thus, the aggregated data used in this exploratory research contained information regarding the length of gambling sessions (i.e., duration), total number of bets (i.e., bets), gambling intensity (i.e., bets per minute), total amount wagered (i.e., redeemed), cost of gambling sessions (i.e., net loss), and the ratio of losses (i.e., %loss). While the lack of payer identification limited our ability to explore gender

differences in regards to gambling behaviour, the results shown in LaBrie et al. [27] suggested no gender differences in live-sport and casino-type gamblers.

In regards to identifying types of gambling behaviour in EGM gambling sessions, data mining techniques, such as clustering [19], can be used to group together sessions with similar behavioural characteristics (e.g., duration, intensity, frequency of bets). Among the various techniques for clustering data (e.g., partitioning, hierarchical, grid-based, model based, and constraint-based methods) [19], k-means clustering is one of the most widely used partitioning methods [4]. However, there are certain issues that emerge when using k-means clustering in large data sets, such as case order effect [37], selection of suitable evaluation variables [4], [26], [51], data comparability [7], [19], and k-means instability [7], [23], [40]. The results shown in Braverman and Shaffer [7] provide an example of k-means clustering for analyzing the betting patterns of Internet live-sport gamblers.

In their research, Braverman and Shaffer [7] classified gamblers into four clusters. Gamblers within the high-risk sub-group showed intensive and frequent betting, high wager variability, positive gambling trajectory, and were at a much higher-risk for closing their account due to gambling-related issues than the rest of the sample. Furthermore, these results were consistent with those of Xuan and Shaffer [55], who analyzed the gambling patterns of Internet live-sport bettors during their last month of gambling activity; both studies analyzed the same data set used in LaBrie et al. [27]. The results from Braverman and Shaffer [7] highlighted the significance of selecting a stable solution for k, as this process can be quite subjective.

There are numerous methods for selecting k [19], [39], for example, researchers can visually estimate the proper number of clusters by mapping data points to points in space, hierarchical clustering methods [19] can also be used to visually identify meaningful splits in the data. In cases where large data sets are used, efforts are often focused on identifying and creating more efficient and effective methods for cluster analysis [4], [23], [39], [40], [49], [54]. To identify a stable and high-quality yielding solution for k, this research applied a k-means stability test.

To assess the stability of a k-means solution, researchers can define a clustering criteria (e.g., $3 \leq k \leq 10$) and compare the movement between the initial and final cluster centers for each solution [19]; clusters that show minimal or no movement can be considered to be more stable. Moreover, researchers can also apply a split test where the full sample can be randomly split into two halves, which are then re-clustered, the movement of cluster centers is then compared on both halves. A Kappa degree of concordance test [37], [50] can then be applied to asses the level of agreement between cluster memberships of the resulting sub-samples and the full sample. Once a stable and high-quality yielding solution (i.e., clusters with high intra-class and low inter-class similarity) for k is found, differences between the resulting clusters can be identified through a one-way Analysis of Variance (ANOVA) [37]. A similar methodology for assessing the stability of a k-means solution was also used in Braverman and Shaffer [7]. Additionally, by clustering EGM gambling sessions it may be possible to identify those cases that deviate from the general model in a data set. Most data mining applications often remove outliers, as they may be caused by measurement errors [19]. However, outlier mining has been the focus of fraud detection, customized marketing, medical analysis, and network security [19]. Within the context of this research, outliers may represent sessions of a riskier gambling nature; thus an analysis of outliers has the potential

to provide results that may give a better insight into heavily involved gambling behaviour.

Outlier detection methods can be classified into statistical, proximity-based, density based, and clustering-based methods depending on the assumptions they make [19]. Statistical methods assume the data is normal and use a discordancy test [19] to find outliers; however, most statistical methods are only suitable for univariate data, and can at times miss outliers [19]. Density-based approaches assess the degree to which a data object can be an outlier (i.e., Local Outlier Factor) [19], though these methods fail to provide the level of detail that can be obtained by combining proximity-based and clustering-based methods [1], [19]. Proximity-based methods use a distance measure (e.g., standard deviation, median rule, Tukey's outlier labeling method) as a way of assessing the similarity between data points, and avoid excessive efforts associated with discordancy tests. Finally, clustering-based methods focus on exploring the relationship between data objects and their clusters to identify single outliers or a cluster of outliers [19]. Clustering-based and proximity-based outlier detection methods were of particular interest for this research, as the outlier mining methodology consisted of clustering the data points before using a distance measure to identify contextual outliers [19].

Tukey's Outlier Labeling Method (OLM) [44], a commonly used outlier detection method, makes no assumptions of normal distribution, and looks at the bottom (i.e., 25th percentile) and top (i.e., 75th percentile) quartiles of a sample to determine the upper and lower limits (i.e., hinges) of a distribution [24], [25] with data objects beyond these limits labeled as 'outliers'. However, Tukey's OLM is not appropriate for asymmetric data, as the number of outliers tends to increase in skewed data [44]. On the other hand, the standard deviation (SD) method, allows for researchers to examine the presence of data objects at x standard deviations from the mean value. The non-normality of the data used in this analysis, suggested the SD method as the most appropriate for exploring the existence of contextual outliers.

While the SD method is only appropriate for univariate data, the findings in LaBrie et al. [26] suggested this outlier detection method was well suited for this analysis. In their research, LaBrie et al. [26] showed heavily involved gamblers were discouraged by losses, as an increase in %loss often resulted in other variables decreasing (e.g., frequency, intensity, wagered amount); these findings suggested that heavily involved gamblers tended to assess the risk of a wager and self-moderate their behaviour (e.g., reducing intensity while increasing gambling duration), the latter similar to the controlled behaviour seen in substance abuse [26]. Similarly, Xuan and Shaffer [55] found heavily involved gamblers tended to have an involvement-seeking and risk-averse gambling behaviour. The results from these studies [26], [55] suggested problem gamblers were likely to show heavily involved gambling behaviour on one aspect of gambling rather than across variables.

Despite the growing need for studies to help understand gambling behaviour, as well as problem gambling behaviour, research of actual gambling data is scarce [46]; particularly, research focused on EGM gambling data. While analysis done on Internet sport gambling sheds a light on the behaviour of on-line problem gamblers, it is not possible to generalize their results to EGM gamblers, a deficiency also found in controlled gambling studies [12], [14], [31], [48], due to the impact that EGM gambler proximity may have on EGM gambling behaviour. Another difference is the lack of anonymity for each gambler, which is an appealing characteristic of online gambling [27], as EGM gambling cannot be done remotely. Nevertheless, the

results from longitudinal studies [6], [7], [26], [27], [28], [29], [46], [55], highlighted the importance of analyzing actual gambling data in order to more accurately identify unusual changes in gambling patterns. And while there are a number of factors that can facilitate problem gambling, such as gambling accessibility and availability of help services [10], [13], these aspects and their impact on EGM gambling behaviour are beyond the scope of this research.

Behavioral analysis of EGM gambling data can certainly increase understanding of the expression of the disorder during a gambling session. This initial exploratory study expands on the current literature by defining what constitutes an EGM gambling session, identifying the gambling involvement measures, distinguishing gambling personae (e.g., clusters of sessions) based on the behavioural markers of EGM gambling sessions, and recognizing differences between these clusters as well as between normal sessions and outliers (i.e., heavily involved gambling sessions).

3. Methodology.

The EGM data used in this exploratory study was collected during the month of July 2010 from a single EGM manufacturer. The purpose of this research was to conduct an analysis of gambling measures in order to identify the gambling involvement measures [5], [6], [26], [55], that could serve as the behavioural markers of EGM gambling sessions and, based on the behavioural characteristics of sessions, distinguish types of gambling. The machines from which the data was collected required cash rather than logins or loyalty cards. While the lack of player IDs can be an obstacle when attempting to identify EGM gambling sessions, as it makes it difficult to assume an independence of sessions, the use of anonymous gambling data allowed us to limit threats to construct validity [8] such as the good subject tendency and evaluation apprehension. As such, this research did not attempt to identify individual gamblers but rather focused on identifying types of gambling (i.e., gambling personae).

Thus, in the following sub-sections we discuss the methodology used to collect and analyze the gambling measures required for this study. First, in Section 3.1, we explain the criteria used for identifying EGM gambling sessions during the data selection process. Second, in Section 3.2, we define the necessary data preprocessing steps to help specify what constitutes a valid EGM gambling session. Third, in Section 3.3, we specify the data transformation tasks needed to increase the overall quality of the mined results.

3.1. EGM gambling session definition criteria.

The EGM gambling data used in this research consisted of a sequence of messages containing information related to gambling events, using the Game to System (G2S) protocol. Among other things, this XML-based protocol, developed by the Gaming Standards Association (GSA) [17], [18], supports real-time calculation of wins, remote EGM configurations, and player tracking [17], [18]. A major benefit of this XML-based standard is its extensibility, as EGM manufacturers are able to develop proprietary extensions in order to customize the implementation of this protocol. Our initial approach to identifying sessions consisted of applying a methodology similar to that used in Liu and Keselj [30] for unsupervised web session detection. The implementation of such methodology involved applying a time-lapsed between events approach; in this regard, our research would have used gambling events rather than web pages, and a time threshold of fifteen (15) seconds between events to determine which session an event belonged to. In other words, our assumption was that gambling events taking place within fifteen seconds of each other, and on the same EGM, would belong to

the same session; messages taking place after the 15-second mark would be assigned to a new session. However, after exploring the dataset, it was clear that a more precise approach could be considered.

G2S messages operate in one of two levels, the message level and the application level [17], [18]. Messages operating within the message level are responsible for acknowledging requests, and though they may be useful for EGM fault detection, they were not relevant to this research. On the other hand, messages operating in the application level are in charge of handling and processing requests [17], [18]. Within the application level, there are two types of messages, multicast messages and g2sBody messages. Multicast messages were not found to be relevant as they are used for remote configurations [17], [18]. Finally, g2sBody messages, used for communications between a single host and a single EGM, are responsible for processing EGM requests and can contain information related to game-play events such as game title, amount wagered, bet results, and bonus round activity.

In general, the context of a G2S message can be specified through the use of classes, which serve as containers of physical and/or logical devices (e.g., note acceptor device). Thus, in this research, in order to define a gambling session, we set out to identify the specific classes in charge of handling gambling requests (e.g., bets), and reporting the results of a gambling event. Identifying these classes allowed for the researchers to note the commands used to report a gambling event (e.g., a bill is redeemed). These commands make use of attributes, which provide further information on the events taking place during a gambling session (e.g., $5 redeemed). Using the information provided within these relevant messages, we were able to define parameters for identifying gambling sessions. In other words, rather than using a time-lapsed between events approach, we aimed at specifying actions that could serve as markers for the start and end of session.

For the purposes of this research, in order for gambling sessions and play-personae to be identified, messages must first be grouped together according to EGM. Second, sessions can only contain game-play related classes (i.e., g2sBody messages providing game-related information). Third, gambling sessions must start by indicating a session reported no money in the bank and some sort of currency (e.g., bills, coins, or vouchers) was entered into an EGM for the first time, this criteria was necessary as EGM players are able to enter bills throughout a gambling session; the specific type and method of currency is dependent on the EGM manufacturer's configuration. Fourth, if a session has winnings at cash-out, the session would end with a voucher being issued; in the event a session ends without any winnings (i.e., no credits remaining), the gambling session would end with a message showing the result of the last wager was a loss and no credits were left in that session. Once sessions were identified, relevant game-play information was extracted and gambling measures were aggregated. The measures collected included the EGMs IDs, session duration, intensity, amount redeemed throughout a session, and amount received in voucher form.

3.2. **Data preprocessing.** A total of 288 EGMs were identified as machines involved in actual game-play activity, these EGMs produced a total of 46,514 gambling sessions. The aggregated data within these sessions included the duration of a session measured in seconds, the average intensity of a session, the total amount of money redeemed by a player throughout a session, the amount of money obtained in voucher form per session, and the EGM's ID which was converted to a random number in order to ensure player and EGM anonymity.

In terms of gambling activity, there were five games played in these sessions, four slot-machine-type games and one poker game. The 46,514 sessions amounted to a approximately 35,095.44 hours of gambling activity, the dataset was then binned based on session duration. By binning the data, the researchers aimed to define what constitutes a valid EGM gambling session. As shown on Figure 1, sessions with a duration time within 24 hours, were binned into eight (8) hour bins (e.g., $0 < \ x \ \leq \ 8hrs$; $8hrs < x \leq 16hrs$; $16hrs < x \leq 24hrs$), sessions with a duration time longer than 24 hours and less than 168 hours (i.e., one week) were binned into 24 hour bins (e.g., $24 < x \leq 48hrs$; $48hrs < x \leq 72hrs$), sessions with a duration time longer than 168 hours were binned into weekly bins (e.g., $168 < \ x \ \leq \ 336hrs$). As illustrated on Figure 1, a total of 45,637 sessions had a duration time within the eight hour mark, of the remaining 877 sessions, seventeen (17) gambling sessions had a duration time greater than 24 hours, one of these cases had a duration time of over a week. The results shown on Figure 1 made it clear that additional preprocessing tasks were needed in order to increase the quality of our results.

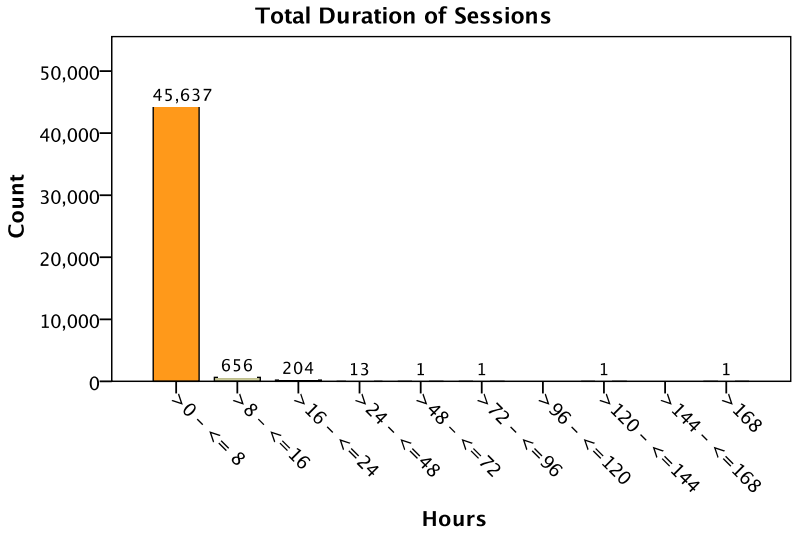


FIGURE 1. Binned EGM Gambling Sessions Based on Hours Played.

Thus, in order for sessions to be considered to be valid, certain conditions must be met. First, a session cannot be considered valid if no gambling activity occurred, thus, valid gambling sessions must contain at least one bet placed. Second, the amount redeemed in a session must be greater than 0. Third, the total duration of a session cannot exceed 18.5 hours (i.e., 1,110 minutes); the specified time threshold was based on the maximum number of hours a non-casino venue would likely be open during the Summer months (i.e., 7:30am to 2am). Though removing sessions with a duration time shorter than five (5) minutes was considered, the researchers noticed that these sessions had a minimum duration time of three minutes, a high gambling intensity and an amount redeemed greater than €5, with €200 being the maximum amount redeemed; as such, these short sessions were not removed from the research sample.

The original research sample consisted of 46,514 sessions, after applying the aforementioned conditions for identifying valid EGM gambling sessions, our final research sample consisted of 46,416 sessions. A total of 98 cases were removed from the original research sample, five of these cases reported no gambling activity (i.e., no bets placed), and 93 cases were removed for having a total duration time greater than the specified total duration threshold (i.e., 18.5 hours or 1,110 minutes). As shown on Figure 2, 98.3% of valid sessions fell within the 8 hour mark (i.e., 480 minutes) and only 0.3% of sessions had a duration time between the 16 and 18.5 hour mark.

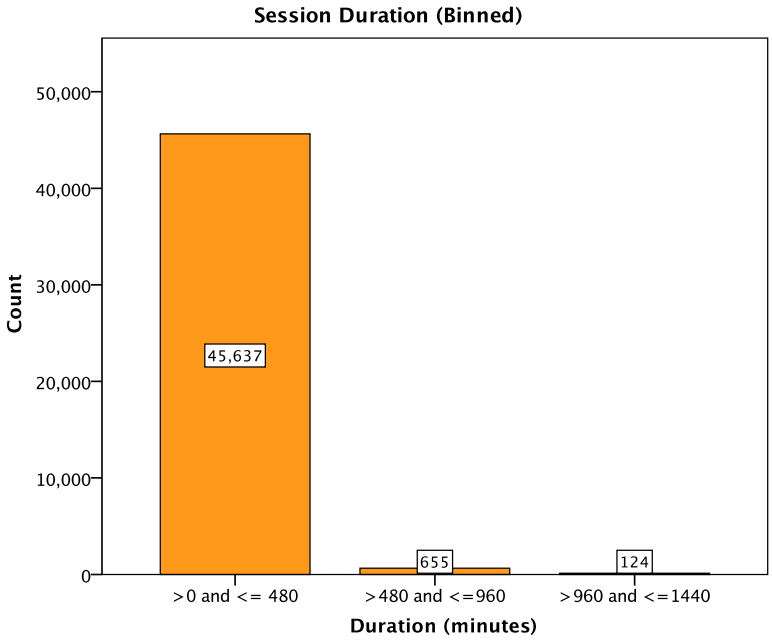


FIGURE 2. Valid EGM Gambling Sessions.

3.3. **Data transformation.** In our final dataset, session duration values were transformed from seconds to minutes to increase comparability with gambling intensity. In regards to gambling intensity, this value was calculated by dividing the total number of bets recorded in a session by the total duration (in minutes) of that session, the values for gambling intensity were smoothed by removing decimals. Calculating the intensity of sessions was of particular importance as it was one of the four variables used to assess problem gambling (i.e., intensity, frequency, variability, trajectory) [7], [13]. The total number of bets in a session was tallied, the cost of a session (i.e., net loss) was calculated by deducting the out-of-pocket cost of a session (i.e., redeemed amount) from the winnings reported in a session (i.e., amount issued in voucher form at cash-out), and the ratio for losses (i.e., %loss) was calculated based on the total amount redeemed during a session.

The 46,416 sessions amounted to approximately 32,516.75 hours of gambling activity (i.e., 1,951,005.05 minutes), with a total of 17,329,709 bets placed; the maximum duration of a session was 18.3 hours (i.e., 1099.72 minutes). For the

purposes of this research, Euros (€) were used as the unit of measure for the amount of money wagered within a session (i.e., Redeemed), and for the amount obtained in voucher form (i.e., Vouchers) at the end of a session. In these sessions, a total of €3,272,065.00 were redeemed (i.e., amount of money entered into an EGM), and a total of €2,341,148.58 were issued in voucher form (i.e., player winnings at the end of a session). These sessions had a total net loss of €2,417,693.39. In regards to gambling intensity (i.e., bets per minute), as shown on Table 1, the maximum intensity, reported in these sessions, was 68 bets per minute (bpm). Interestingly, only 8,981 sessions reported any winnings, all of these sessions had a duration time within the eight (8) hour mark. One case in particular reported a voucher amount of €40,833.05; in fact, there were sixteen (16) cases in which a significantly large voucher amount was issued, likely the result of a jackpot.

TABLE 1. Descriptive Statistics Measures of EGM Sessions

Variables	**Mean**	**SD**	**Median**	**Mode**	**Max.**	**Min.**
Duration[a]	42.03	109.09	15.67	3.87	1099.72	0.17
Bets	373	606	170	3	13,282	2
Intensity	16	10	19	20	68	0
Redeemed[b]	70.49	150.75	30.00	20.00	6,425.00	5.00
Vouchers[b]	50.44	384.18	0.05	0	40,833.05	0
Net Loss[b]	52.09	133.39	19.90	0	6,424.69	0
%Loss	75.46	40.97	99.97	100.00	100.00	0

a. Measured in minutes.

b. Measured in Euros.

While the mean values for all variables, shown on Table 1, do not seem high, the relationship between the mean, median, and standard deviation values suggested a non-normal sample distribution. As such, we explored the use of z-score and min-max normalization prior to conducting a correlation analysis. While both normalization methods preserve relationships among the data [45], z-score normalization has some limitations due to the skewness of the sample, as evidenced by the mean and standard deviation values shown on Table 1. Though there were no clear differences found between both normalization methods on our sample, the min-max normalization method was chosen, as this method has the potential to make outliers more noticeable.

Once the data was normalized, a boxplot analysis was done as part of a normality test. The results from this boxplot analysis, shown on Figure 3, illustrate the non-normality of the sample distribution. For example, Figure 3c shows numerous outliers present in terms of Intensity, with the median (i.e., 19 bpm) closer to the upper quartile of the distribution. Figure 3a, 3b, 3d, 3e, and 3f, show numerous extreme outliers within the Duration, Bets, Redeemed, Voucher, and Net Loss variables, respectively. The results from a normal Q-Q plot analysis, shown on Figure 4, also illustrate the data's clear deviation from the expected normal value. Furthermore, the results of a Skewness test confirmed the non-normal distribution of the sample, as the skewness coefficients were found to be more than twice the value of their respective standard error values. To explore the relationship among the

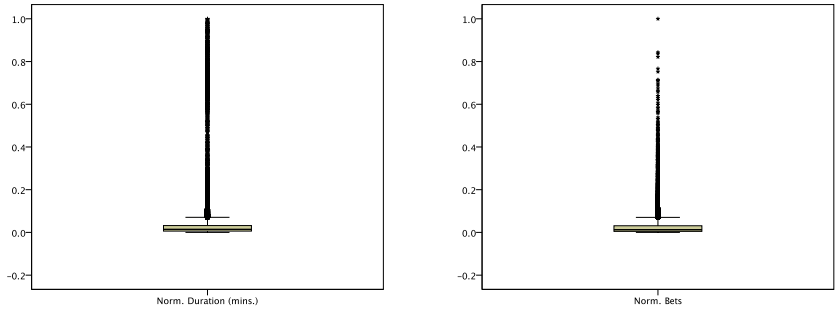
sample variables, the results of the boxplot, Q-Q plot analysis, and Skewness test, suggested the suitability for Spearman's Rank-Order Correlation analysis, as these results showed a monotonic relationship between the aforementioned variables. As explained in Section 4, for the purposes of this research, the results of a correlation analysis can help in identifying suitable evaluation attributes for recognizing gambling personae (i.e., clusters) based on the gambling behaviour expressed in these sessions.

4. **Clustering methodology.** In the following sections we provide an explanation of the methodology used for identifying types of gambling (i.e., gambling personae), based on the behaviour expressed in these sessions. In Section 4.1, we present our methodology for identifying the measures of EGM gambling involvement that can serve as behavioural markers of EGM sessions, which can then be used as evaluation variables for classifying sessions. In Section 4.2, we discuss the clustering techniques applied on the research sample. In particular, we discuss methods for selecting a stable and high-quality yielding solution for k. The results of our clustering analysis are discussed in Section 5.

4.1. **Selection of evaluation variables.** To identify suitable evaluation variables for classifying EGM sessions, the researchers conducted a correlation analysis using Spearman's Rank-Order Correlation coefficient. Also known as Spearman's rho, this non-parametric statistical measure is used for exploring the strength of monotonic relationships among variables of a non-normally distributed data set. One of the benefits of this statistical measure is its lack of sensitivity towards outliers and its assumption of variable independence [9], [35], [46]. Spearman's rho assigns values between -1 and +1 (i.e., $-1 \leq r_s \leq 1$) to variables, where positive values show a positive monotonic correlation and negative values show a negative monotonic correlation. The strength of the relationships can be described through the absolute values of r_s (i.e., Spearman's rho). Values for Spearman's rho between .00 and .19 describe very weak relationships; values between .20 and .39 describe weak relationships; values between .40 and .59 moderate relationships; values between .60 and .79 strong relationships; and values between .80 and 1.0 very strong relationships [9], [35].

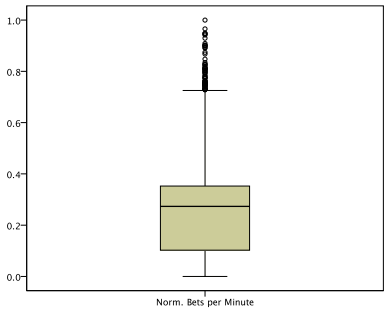
The results for Spearman's rho, shown on Table 2, indicated a very strong negative monotonic correlation between Vouchers and %Loss ($rho = -.980, n = 46,416$, $p < .001$). These results implied that as the total amount of money issued in voucher form increases, the %Loss in a session decreases. However, these perceived wins could still have represented a loss (i.e., the wins produced throughout a session were less than the out-of-pocket cost of a session). There was a strong positive monotonic correlation between the total number of bets in a session and amount redeemed ($rho = .690, n = 46,416, p < .001$). Another strong positive relationship was found between intensity (i.e., bets per minute) and the amount of issued in voucher form ($rho = .640, n = 46,416, p < .001$), an indication that as the intensity in a session increases, so does the potential for that session to produce winnings.

The results from this correlation analysis also showed a moderately positive correlation between intensity and amount redeemed ($rho = .433$), an indication that as the average number of plays per minute increases, the total amount of money entered into the EGM is likely to increase. As well, there was a strong negative relationship between intensity and %Loss ($rho = -.608, n = 46,416, p < .001$), which corroborates the relationship between intensity and vouchers, as sessions with a

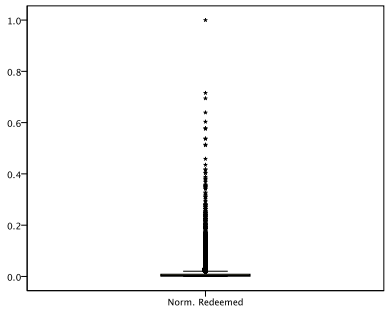


(A) Duration

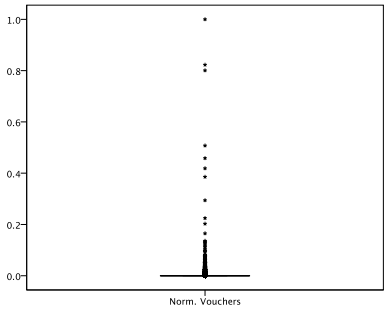
(B) Bets



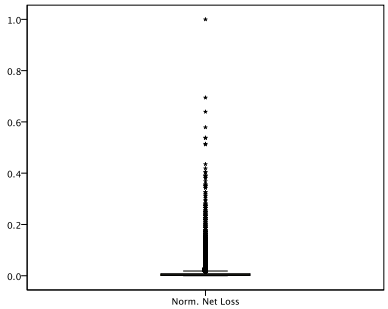
(C) Intensity



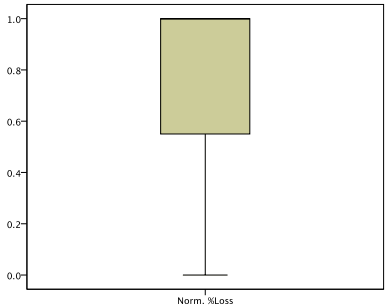
(D) Redeemed



(E) Voucher

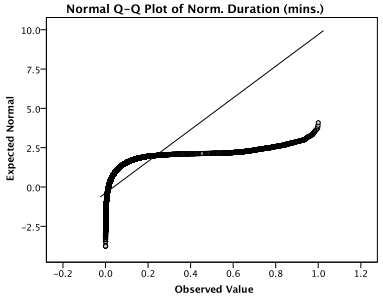


(F) Net Loss

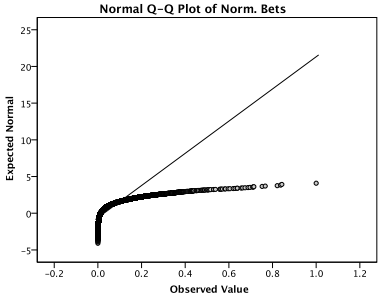


(G) %Loss

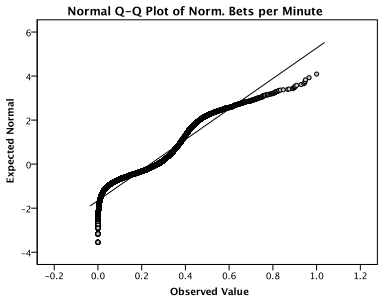
FIGURE 3. Test of Normality: Boxplot Analysis.



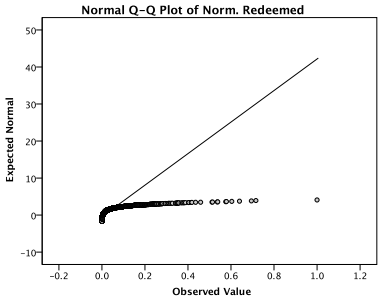
(A) Duration



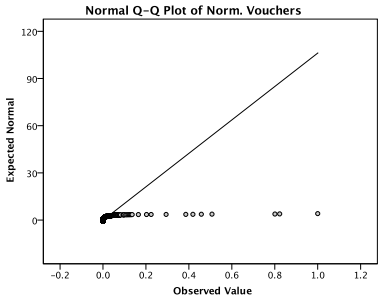
(B) Bets



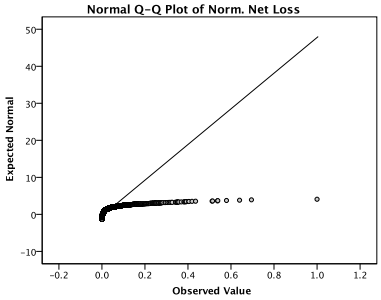
(C) Intensity



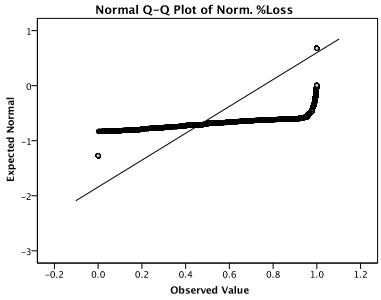
(D) Redeemed



(E) Voucher



(F) Net Loss



(G) %Loss

FIGURE 4. Test of Normality: Normal Q-Q Plot Analysis.

TABLE 2. Spearman's Rank-Order Correlation Analysis

Spearman's Rho								
		Duration	**Bets**	**Intensity**	**Redeemed**	**Vouchers**	**Net Loss**	**%Loss**
Duration	r_s	.	.515**	-.208**	.341**	-.334**	.374**	.406**
	Sig.	.	.000	.000	.000	.000	.000	.000
Bets	r_s	.515**	.	.606**	.690**	.224**	.356**	-.150**
	Sig.	.000	.	.000	.000	.000	.000	.000
Intensity	r_s	-.208**	.606**	.	.433**	.640**	-.003	-.608**
	Sig.	.000	.000	.000	.000	.511	.000	0.000
Redeemed	r_s	.341**	.690**	.433**	.	.177**	.584**	-.068**
	Sig.	.000	.000	.000	.	.000	.000	.000
Vouchers	r_s	-.334**	.224**	.640**	.177**	.	-.512**	-.980**
	Sig.	.000	.000	.000	.000	.	.000	.000
Net Loss	r_s	.374**	.356**	-.003	.584**	-.512**	.	.592**
	Sig.	.000	.000	.511	.000	.000	.	.000
%Loss	r_s	.406**	-.150**	-.608**	-.068**	-.980**	.592**	.
	Sig.	.000	.000	.000	.000	.000	.000	.

** Correlation is significant at the 0.01 level (2-tailed).

more intense style of play can have a smaller percent, of the amount redeemed in a session, lost during that session.

In regards to identifying suitable evaluation variables for our clustering analysis, the results shown on Table 2, allowed us to quickly rule out variables with high correlations, such as vouchers and %loss. Other variables such as bets and net loss were excluded due to their redundancy, as these variables can be derived from other variables (e.g., intensity, duration, redeemed, vouchers). Since our primary research question was to identify types of gambling, based on the behaviour expressed throughout these sessions, we used the attributes for duration, intensity, and amount redeemed. The methodology for identifying a suitable clustering criteria is explained in Section 4.2.

4.2. **Finding a suitable solution for K.** In general, clustering algorithms partition data into k groups or clusters by analyzing cases in a data set, cases that appear similar to others are grouped into the same partition [45]. These clusters are defined based on a dissimilarity function. There are various methods for clustering data, among the most widely used partitioning methods is k-means clustering [16]. However, there are a number of limitations to implementing k-means clustering in large datasets such as case order effect and instability of clustering solution.

For example, the k-means clustering algorithm is known to be vulnerable to the learning effect, particularly when dealing with large data sets [15], as this algorithm can learn to cluster cases based on their order within the dataset. To minimize the clustering algorithm's vulnerability to the learning effect, which could produce biased results, the researchers randomized the order of the cases within the data set, as the case order was originally based on EGM ID. Furthermore, the researchers did not use running means for this analysis.

In regards, to the lack of stability of the k-means clustering solution, the process of selecting k is a highly subjective one. While the k-means clustering algorithm allows data miners to partition the data into a fixed k number of clusters, it is often up to data scientists to select the number of clusters in which to partition the data. Ideally, a suitable solution for k is one that would produce high quality clusters with high intra-class and low inter-class similarity. In this research, the dissimilarity between data objects was calculated based on the distance between pairs of data objects using the Euclidean distance [4] on the normalized dataset.

To identify a stable and suitable solution for k, the dataset was clustered into $3 \leq k \leq 10$ using duration, intensity, and redeemed amount as evaluation variables. The initial and final cluster centers of each solution (i.e., $3 \leq k \leq 8$) were then compared to identify the solution with the least amount of movement in cluster centers. The final cluster centers were then aggregated and used to re-cluster the dataset for a second comparison. When examining the movement of cluster centers, $k = 3$, $k = 4$, $k = 5$, and $k = 7$ were identified as possible solutions for k. A split test was then applied on the dataset, which resulted in two sub-samples of approximately equal size, the k-means clustering procedure (i.e., $k = 3, k = 4, k = 5, k = 7$) was then repeated on each half.

To assess the measure of agreement between the cluster memberships for these two sub-samples and the full sample, a Kappa degree of concordance test was done; $k = 4$ was found to have the highest degree of agreement. The researchers were particularly interested on how well each sub-sample agreed with the full sample rather than the statistical significance of the results, as any Kappa value greater than 0 could be considered to be of statistical significance [50], especially if a large research sample is used. The measure of agreement between sub-sample 1 and the full cohort for $k = 4$ was .996, the measure of agreement between sub-sample 2 and the full cohort was .995, both measures suggested an almost perfect degree of agreement [47].

TABLE 3. Crosstabulation: Sample 1 v. Full Sample ($k = 4$)

Cluster Membership (Sample 1) v. Cluster Membership (Full Sample) Crosstabulation)							
			Cluster Mship. (Full Sample)				
			1	2	3	4	**Total**
Cluster Mship. (Sample 1)	**1**	Count	6944	15	0	0	6,959
		Expected Count	2,071.5	2,171.1	123.2	2,593.2	6,959
		% within Cluster (sample 1)	99.8%	0.2%	0%	0%	100%
		% within Cluster (full sample)	100%	0.2%	0%	0%	29.8%
		% of Total	29.8%	0.1%	0%	0%	29.8%
	2	Count	0	7,263	0	44	7,307
		Expected Count	2,175.1	2,279.7	129.4	2,722.9	7,307
		% within Cluster (sample 1)	0%	99.4%	0%	0.6%	100%
		% within Cluster (full sample)	0%	99.8%	0%	0.5%	31.3%
		% of Total	0%	31.1%	0%	0.2%	31.3%
	3	Count	0	0	413	0	413
		Expected Count	122.9	128.9	7.3	153.9	413
		% within Cluster (sample 1)	0%	0%	100%	0%	100%
		% within Cluster (full sample)	0%	0%	100%	0%	1.8%
		% of Total	0%	0%	1.8%	0%	1.8%
	4	Count	0	0	0	8,649	8,649
		Expected Count	2,574.5	2,698.4	153.1	3,223	8,649
		% within Cluster (sample 1)	0%	0%	0%	100%	100%
		% within Cluster (full sample)	0%	0%	0%	99.5%	37.1%
		% of Total	0%	0%	0%	37.1%	37.1%
Total		Count	6,944	7,278	413	8,693	23,328
		Expected Count	6,944	7,278	413	8,693	23,328
		% within Cluster (sample 1)	29.80%	31.20%	1.8%	37.3%	100%
		% within Cluster (full sample)	100%	100%	100%	100%	100%
		% of Total	29.80%	31.20%	1.8%	37.3%	100%

The results shown on Table 3 and Table 4 present the resulting cross tabulation tables for sample 1 with the full cohort, and sample 2 with the full cohort, respectively. These tables are useful for obtaining the sensitivity (i.e., *Recall*) and specificity of a measure. When calculating these values for sample 1, compared

TABLE 4. Crosstabulation: Sample 2 v. Full Sample ($k = 4$)

Cluster Membership (Sample 2) v. Cluster Membership (Full Sample) Crosstabulation)							
			Cluster Mship. (Full Sample)				
			1	2	3	4	**Total**
Cluster Mship. (Sample 2)	**1**	Count	6,871	0	0	0	6,871
		Expected Count	2,053.1	2,148.1	122.3	2,547.5	6,871
		% within Cluster (sample 2)	100%	0%	0%	0%	100%
		% within Cluster (full sample)	99.6%	0%	0%	0%	29.8%
		% of Total	29.8%	0%	0%	0%	29.8%
	2	Count	28	7,162	0	0	7,190
		Expected Count	2,148.5	2247.8	128	2,665.7	7,190
		% within Cluster (sample 2)	0.4%	9,9.6%	0%	0%	100%
		% within Cluster (full sample)	0.4%	99.2%	0%	0%	31.10%
		% of Total	0.1%	31.0%	0%	0%	31.10%
	3	Count	0	0	411	0	411
		Expected Count	122.8	128.5	7.3	152.4	411
		% within Cluster (sample 2)	0%	0%	100%	0%	100%
		% within Cluster (full sample)	0%	0%	100%	0%	1.8%
		% of Total	0%	0%	1.8%	0%	1.8%
	4	Count	0	56	0	8,560	8,616
		Expected Count	2,574.6	2,693.6	153.4	3,194.4	8,616
		% within Cluster (sample 2)	0%	0.6%	0%	99.4%	100%
		% within Cluster (full sample)	0%	0.8%	0%	100%	37.30%
		% of Total	0%	0.2%	0%	37.1%	37.3%
Total		Count	6,899	7,218	411	8,560	23,088
		Expected Count	6,899	7,218	411	8,560	23,088
		% within Cluster (sample 2)	29.9%	31.3%	1.8%	37.1%	100%
		% within Cluster (full sample)	100%	100%	100%	100%	100%
		% of Total	29.9%	31.3%	1.8%	37.1%	100%

with how sessions in this sub-sample were clustered in the full cohort, the results indicated that our test performed quite well when picking sessions as not belonging to a cluster when they did in fact not belong to that cluster (i.e., specificity), as well as when identifying sessions as part of a cluster when they did in fact belong to that cluster (i.e., sensitivity).

For example, as shown on Table 3, of the 23,328 sessions in sample 1, a total of 413 were classified into Cluster 3 when being analyzed as part of the full cohort; all of these sessions were correctly classified into Cluster 3 in sample 1, representing a sensitivity and specificity value of 100% ($C3R = .10$). In fact, the lowest sensitivity was obtained by Cluster 4; of the 8,693 sessions classified into Cluster 4, when analyzed as part of the full sample, a total of 8,649 sessions were correctly classified into Cluster 4 in sample 1, which suggests that only 44 sessions were missed ($C4R = .9949$).

In regards to specificity, Cluster 2 had the lowest specificity obtained in sample 1. Of 16,050 sessions correctly not classified into Cluster 2 as part of the full sample, 16,006 sessions were correctly not classified into this cluster in sample 1; in other words, in sample 1, 99.7% of sessions not belonging to Cluster 2 were in fact identified as not part of Cluster 2.

When calculating the values for sensitivity and specificity for the 23,088 sessions clustered in sample 2, shown on Table 4, the results seemed to be on par with those shown on Table 3. For example, Cluster 2 had the lowest sensitivity value as 99.2% ($C2R = .992$) of sessions belonging to this cluster, as part of the full sample, were correctly classified into Cluster 2 in sample 2. In regards to specificity, the lowest value was obtained by Cluster 4, where 99.6% of sessions not classified into Cluster 4, in the full research sample, were correctly identified as not belonging to Cluster 4 in sample 2.

Furthermore, our macro averaged precision values, as well as our macro averaged recall values, showed that our classification solution performed well on both samples ($P[s1] = 99.8\%$; $P[s2] = 99.7\%$; $R[s1] = 99.8\%$; $R[s2] = 99.7\%$). In regards to the overall performance of our clustering solution, our test seemed to perform slightly better on sample 1 ($F_{macro} = 99.8\%$; $F_{micro} = 99.8\%$) than sample 2 ($F_{macro} = 99.7\%$; $F_{micro} = 99.6\%$). The results from this cross tabulation analysis suggested that, when comparing the results between each sub-sample with the full research sample, there was a slightly higher degree of agreement between sessions in Sample 2 and the full research cohort. This slight difference in results could be due to the skewness of the data, as outliers were not removed from the sample since these cases could represent sessions of a riskier gambling nature.

However, it must be noted that while there was no significant change in the sample when using min-max normalization compared to z-score normalization, the application of a time threshold on sessions (i.e., 18.5 hours) did drastically improve the results of our split test. For example, without the removal of the 98 sessions with duration times exceeding the 18.5 hour mark, the Kappa value for sub-sample 1 was $-.081$ which suggested the agreement between the two samples (i.e., sample 1 and cohort) was less than would be expected by chance; while the Kappa value for sample 2 was .29. Nevertheless, these results were also consistent in suggesting $k = 4$ as the most stable solution for k. A more detailed explanation of the cluster profile results (i.e., gambling personae) is provided in Section 5.

5. **Identified types of EGM gambling.** When comparing the size of the four resulting clusters, Cluster 4 was the largest with 37.2% of sessions ($n = 17,253$) allocated to this cluster. Cluster 2 was the second largest with 31.2% of sessions ($n = 14,496$) classified into this group. Cluster 1, third in size, had 29.8% of sessions ($n = 13,843$) assigned to this cluster. Finally, Cluster 3 had the remaining 1.8% of sessions ($n = 824$) classified in this group, making it the smallest cluster. These cases were clustered based on the type of behaviour expressed throughout a gambling session using the duration of sessions, play intensity, and the out-of-pocket cost of a session (i.e., amount redeemed) as evaluation variables.

Cases in Cluster 1, shown on Table 5, seemed to have relatively low intensity ($n = 17,728$; $mean = 4.09bpm$), the average amount redeemed in a session further corroborates this notion as the mean redeemed amount was €30.02. Overall, sessions in this cluster seemed to have a somewhat passive style of gambling activity, though these sessions had a longer duration time than cases in Cluster 2 and Cluster 4 ($mean = 47.14mins.$). Despite an average %loss of 97.36%, the second highest among all clusters, some sessions in this cluster still reported wins. The maximum amount redeemed in these sessions was €1,215 which may have indicated the presence of more heavily involved sessions in this cluster.

On the other hand, cases in Cluster 2, as shown on Table 6, seemed to be characterized by a higher intensity than cases in Cluster 1 ($n = 14,496$; $mean = 17.26bpm$). In fact the mode value for intensity (i.e., most repeated value) was 20 bpm. The maximum intensity reported in Cluster 2 was 22 bpm, the second highest across clusters. When compared with cases in Cluster 1, sessions in Cluster 2 tended to have a slightly higher financial involvement ($mean =$ €54.98), though the standard deviation ($SD =$ €94.35) for the amount of money redeemed implied the presence of outliers. In fact the highest amount redeemed in these sessions was €2,450.00. Interestingly, cases in Cluster 2 lost a smaller percentage of their amount

TABLE 5. Descriptive Statistics: Cluster 1 Sessions

Cluster 1 Sessions [a]						
Variables	**Mean**	**SD**	**Median**	**Max.**	**25th**	**75th**
Duration[b]	47.14	56.17	27.65	412.77	12.33	59.20
Bets	122	168	68	3,648	34	142
Intensity	4.09	3.10	3.35	11.27	1.32	6.58
Redeemed[c]	30.02	45.04	20.00	1,215.00	10.00	40.00
Vouchers[c]	1.60	23.41	.00	1,962.93	.00	.00
Net Loss[c]	29.48	45.13	20.00	1,215.00	10.00	40.00
%Loss	97.36	15.70	100.00	100.00	100.00	100.00

a. $n = 13,843$

b. Measured in minutes.

c. Measured in Euros.

redeemed ($mean = 71.77\%$; $SD = 43.03\%$; $median = 99.25\%$) in a smaller amount of time ($mean = 14.34mins.$; $SD = 17.26mins.$; $median = 9.07mins.$) than sessions in Cluster 1. Further, the higher intensity and short duration of sessions in Cluster 2 seemed to imply a lack of strategy in their gambling style. However, the low net loss and %loss reported in some of these sessions suggested that perhaps some gamblers in this group did have a strategy, which may have been characterized by quick decisions that assessed the degree of risk in a wager [12], [26], [31].

TABLE 6. Descriptive Statistics: Cluster 2 Sessions

Cluster 2 Sessions [a]						
Variables	**Mean**	**SD**	**Median**	**Max.**	**25th**	**75th**
Duration[b]	14.34	17.26	9.07	326.85	4.57	17.32
Bets	247	299	153	5,344	75	302
Intensity	17.26	3.17	17.81	21.91	14.77	20.03
Redeemed[c]	54.98	94.35	30.00	2,450.00	10.00	50.00
Vouchers[c]	44.06	412.79	.20	40,833.05	.00	20.00
Net Loss[c]	42.74	90.52	19.75	2,450.00	4.60	50.00
%Loss	71.77	43.03	99.25	100.00	19.05	100.00

a. $n = 14,496$

b. Measured in minutes.

c. Measured in Euros.

The results shown on Table 7 indicate that sessions in Cluster 3 have a much longer duration time than those in Cluster 1 and Cluster 2 ($mean = 782.53mins.$; $SD = 160.19mins.$; $median = 780.66mins.$) with the lowest intensity across all clusters ($mean = .32bpm$; $SD = .66bpm$; $median = .12$); in fact, the highest intensity reported in this cluster was 7.25 bpm. In regards to financial involvement, sessions in Cluster 3 seemed to show a somewhat conservative type

of gambling behaviour ($mean$ = €52.34; SD = €127.88; $median$ = €20.00) despite the presence of more heavily involved sessions ($max.redeemed$ = €1,530.00; $max.duration$ = $1,099.72 mins.$). Nevertheless, sessions in this cluster reported the highest ratio of losses when compared with other clusters.

TABLE 7. Descriptive Statistics: Cluster 3 Sessions

Cluster 3 Sessions [a]						
Variables	**Mean**	**SD**	**Median**	**Max.**	**25th**	**75th**
Duration[b]	782.53	160.19	780.66	1,099.72	686.27	899.45
Bets	246	524	86	6,715	39	227
Intensity	.32	.66	.12	7.25	.05	.31
Redeemed[c]	52.34	127.88	20.00	1,530.00	10.00	50.00
Vouchers[c]	.00	.00	.00	.00	.00	.00
Net Loss[c]	52.34	127.88	20.00	1,530.00	10.00	50.00
%Loss	100.00	.00	100.00	100.00	100.00	100.00

a. $n = 824$

b. Measured in minutes.

c. Measured in Euros.

TABLE 8. Descriptive Statistics: Cluster 4 Sessions

Cluster 4 Sessions [a]						
Variables	**Mean**	**SD**	**Median**	**Max.**	**25th**	**75th**
Duration[b]	25.84	29.94	16.07	394.65	7.57	32.75
Bets	688	844	413	13,282	193	855
Intensity	26.12	4.00	25.35	68.32	23.59	27.43
Redeemed[c]	116.87	218.11	50.00	6,425.00	20.00	120.00
Vouchers[c]	97.40	499.21	.45	33,602.76	.20	100.00
Net Loss[c]	78.06	193.43	19.90	6,424.69	.00	70.00
%Loss	58.59	45.74	97.32	100.00	.00	99.73

a. $n = 17,253$

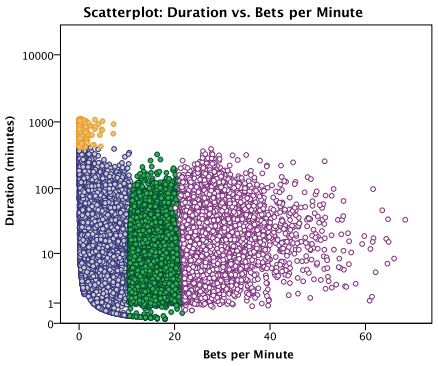
b. Measured in minutes.

c. Measured in Euros.

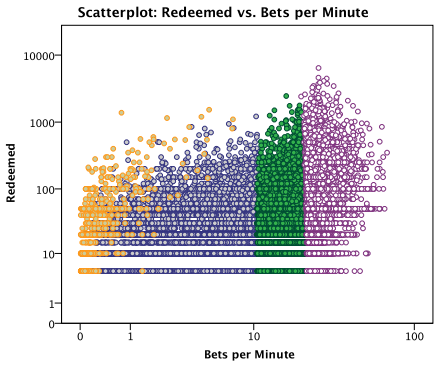
Sessions in Cluster 4, despite having a short duration time ($mean = 25.84 mins.$; $SD = 29.94 mins.$; $median = 16.07 mins.$), reported the highest intensity among all clusters ($mean = 26.12 bpm$; $SD = 4 bpm$; $median = 25.35 bpm$). The average total amount redeemed throughout a session was also higher than in other clusters ($mean$ = €116.87; $median$ = €218.11; $max.$ = €6,425.00). However, despite their much larger redeemed amount, and high intensity, the average %loss reported in sessions of this cluster was relatively low ($mean$ = 58.59%; SD = 45.74%; $median = 97.32\%$), which corroborates the findings discussed in Tables 2, 6, and 7, as more intense sessions are likely to have a shorter duration time and may produce more winnings. When comparing the results shown on Table 5 through Table 8, it

appears that sessions in Cluster 2 and Cluster 4 seemed to exhibit a riskier type of gambling behaviour. However, the longer duration of sessions in Cluster 4, as well as the higher intensity and amount redeemed, seemed to indicate a more involved and riskier type of gambling than sessions in Cluster 2.

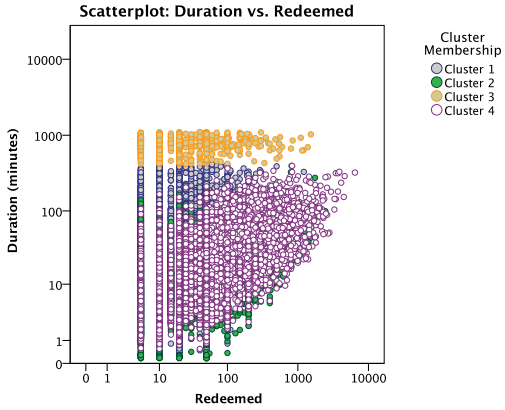
As shown on Tables 5 through 8, with the exception of amount redeemed in Cluster 3, the mean values for all evaluation variables (i.e., duration, intensity, redeemed) were slightly over their respective medians. Figure 5 illustrates the relationships between the evaluation variables. For example, Figure 5a shows the relationship between the duration of sessions and the gambling intensity expressed in these sessions, the relationship between the total amount redeemed throughout a session and the gambling intensity of sessions is displayed in Figure 5b, and the relationship between the total amount redeemed and the duration of sessions is illustrated in Figure 5c.



(A) Duration vs. Intensity



(B) Redeemed vs. Intensity



(C) Duration vs. Redeemed

FIGURE 5. Scatterplot: Clustered sessions.

Interestingly, the application of a time threshold on the duration of sessions had a noticeable impact on the skewness of the clustered results, as this threshold seemed to reduce the distance between data points within their respective clusters.

Nevertheless, the results shown on Tables 5 through 8 did indicate a non-normal distribution within our clusters, potentially caused by heavily involved sessions (i.e., outliers); the results of our Outlier Mining analysis are discussed in Section 6. To examine the structure and strength of the relationships between the resulting clusters and their means, and identify any differences between these clusters, a one-way analysis of variance (ANOVA) test, along with the post-hoc comparisons, were carried out on our clustered results. The results of our ANOVA test are discussed in Section 5.1.

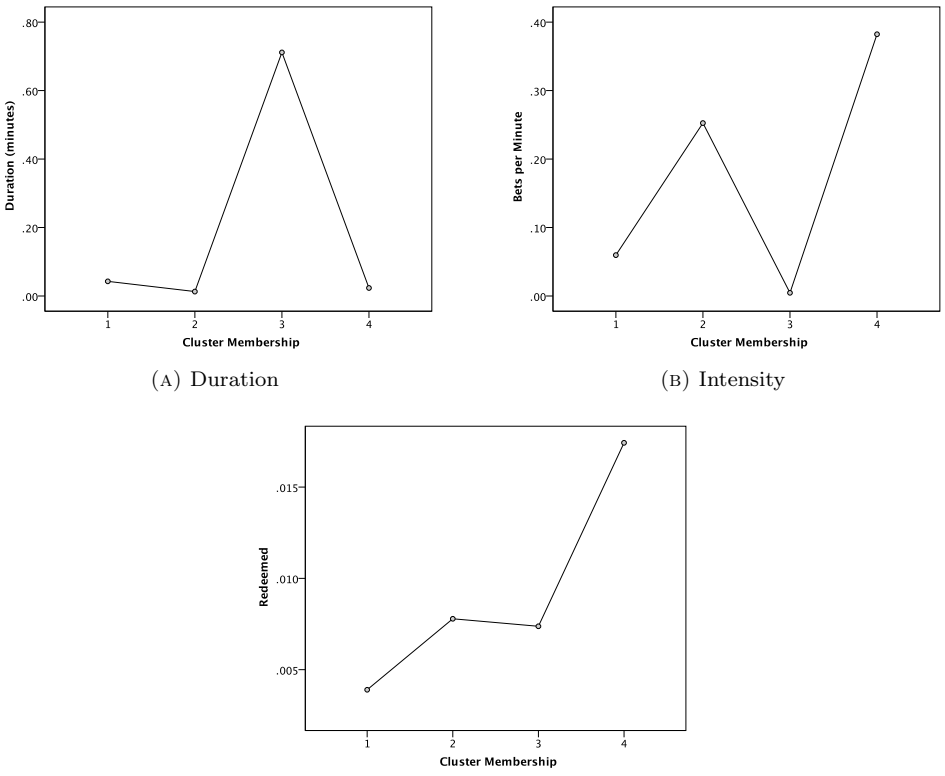
5.1. **Cluster differences.** In order to examine clusters more closely, the researchers conducted a comparison of means between these clusters to identify any meaningful differences. As the resulting clusters followed an independent groups design [36], a between-groups ANOVA test was chosen instead of a repeated measures ANOVA. In this regard, the researchers were particularly interested on the relationships between duration, intensity, and amount redeemed, with cluster membership. As shown on Table 9, there was a significant difference among the mean scores of the dependent variables across all four clusters.

TABLE 9. Between-Groups One-Way ANOVA

ANOVA						
		Sum of Squares	df	Mean Square	F	Sig.
Duration	Between Groups	386.957	3	128.986	85580.335	.000
	Within Groups	69.952	46412	.002		
	Total	456.909	46415			
Intensity	Between Groups	847.409	3	282.470	110089.371	.000
	Within Groups	119.085	46412	.003		
	Total	966.493	46415			
Redeemed	Between Groups	1.542*	3	.514	991.576	.000
	Within Groups	24.052	46412	.001		
	Total	25.593	46415			

As previously mentioned, the relationship between the mean, median, and standard deviation values shown on Table 1, and the boxplot analysis done as part of a normality test shown on Figure 3, suggested a non-normal distribution of our sample. Furthermore, the standard deviations of the dependent variables, shown on Tables 5 through 8, indicated a violation of the homogeneity of variances assumption, which implied the need for a post-hoc analysis [36]. As illustrated in Figure 6, the results of a Tukey's Honestly Significant Difference (HSD) test revealed significant differences ($p < 0.05$) between all clusters within each of the evaluation variables, with the exception of Clusters 2 and 3 in regards to amount redeemed.

However, due to the large size of this sample ($n = 46,416$), these small differences can become statistically significant even in cases where the difference is quite small [36]. In addition to plotting the means for each of the three dependent variables in each of the clusters, to explore the degree that cluster memberships were affected by our sample size, a test of between-subjects effects was done as part of our ANOVA test. In regards to session duration, the results from the between-subjects effects test showed a Partial Eta Squared (η_p^2) of 0.847. In other words, 84.7% of the variability in session duration was accounted for by which cluster a case belonged to. In regards to intensity, 87.7% ($\eta_p^2 = 0.877$) of the variability in gambling intensity was accounted for by cluster membership. Finally, 6% ($\eta_p^2 = 0.060$) of the variability in amount Redeemed was accounted for by cluster membership.



(A) Duration

(B) Intensity

(C) Redeemed

FIGURE 6. Between Groups One-Way ANOVA: Means Plots.

In summary, the results from the between groups one-way ANOVA test indicated that there was a statistically significant difference ($p < 0.05$) for all four clusters, with the exception the amount redeemed in sessions in Clusters 2 and 3. Furthermore, the Partial Eta Squared values for the three dependent variables (i.e., duration, intensity, redeemed) confirmed that the differences in duration ($\eta_p^2 = 0.847$) and intensity ($\eta_p^2 = 0.877$), for the four resulting clusters, was in fact small and may not necessarily be statistically significant with this particular research sample. Though, it is likely that a longitudinal analysis, with greater data granularity, may provide more insight into the gambling behaviour expressed in EGM sessions; particularly, in regards to gambling intensity, frequency of active gambling sessions, wager variability, and trajectory of wagers.

While some of the dependent variables (i.e., duration and redeemed) were approximately distributed, the values for the intensity expressed in gambling sessions in Cluster 1 and Cluster 2 were not. As a result, due to its assumption of non-normality of the data [36], the non-parametric Kruskal-Wallis test was used to confirm the differences identified through our ANOVA test. The Kruskal-Wallis test allows for continuous variables, from more than two clusters, to be compared by converting the values for each of the variables into ranks, the mean rank for each of the clusters can then compared [36] to find differences between the groups.

The results of the Kruskal-Wallis test confirmed a statistical difference in the dependent variables across clusters, and provided an insight into how each of the clusters were ranked based on each of the dependent variables. For example, as shown on Table 10, in regards to duration, Cluster 3 had the highest rank (i.e., longest mean session duration when compared to other clusters), followed by Cluster 1. In regards to intensity and amount redeemed, Cluster 4 had the highest mean intensity and redeemed amount when compared to other clusters, closely followed by Cluster 2.

TABLE 10. Kruskal-Wallis Results

Ranks			
	Cluster Membership	N	Mean Rank
Duration	1	13,843	28907.31
	2	14,496	16544.68
	3	824	46004.50
	4	17,253	23146.26
	Total	46,416	
Intensity	1	13,843	7706.21
	2	14,496	21919.71
	3	824	1090.35
	4	17,253	37786.01
	Total	46,416	
Redeemed	1	13,843	16455.75
	2	14,496	22349.74
	3	824	17893.58
	4	17,253	29601.97
	Total	46,416	

The results from both tests, ANOVA and Kruskal-Wallis, agree that there are significant differences across all clusters. However, the power of the one-way ANOVA test lies in its ability to provide more detailed information into where these differences may occur. For example, while all clusters were found to be different from each other in regards to amount redeemed, the results of our ANOVA test showed no significant difference between Clusters 2 and 3. Furthermore, the high Partial Eta Squared values obtained for duration ($\eta_p^2 = 0.847$) and intensity ($\eta_p^2 = 0.877$) implied that these differences may not necessarily be statistically significant.

6. **Outlier mining methodology.** The non-normality of our clustered results, as shown on Tables 5 through 8 in Section 5, indicated the possible presence of outliers. Though outliers may be caused by measurement errors [45], analyzing cases that do not follow the general model of the data set has been the focus of previous research in fraud detection, customized marketing, medical analysis, and network security [43], [45]. Similarly, in this research, outliers may represent sessions with a passive gambling activity or sessions with riskier behaviour (i.e., heavily involved gambling sessions), as such outliers were not removed from our research sample.

In general, outliers may be classified into global, contextual or collective outliers, though any one outlier may belong to more than one type [4]. Global outliers are the most common as these data points noticeably drift from the rest of the data set. Contextual outliers, on the other hand, drift from the rest of the data objects within a specific context (e.g., cluster of data points). Finally, collective outliers consist of an entire subset of data points that deviate from the rest of the data set [4]. In this case, our focus was to identify contextual outliers within our resulting clusters, and understand why these outliers were placed in these clusters. Thus the outlier identification method chosen must provide some justification of the detection [4], [45].

Detection methods can be classified into supervised, unsupervised, and semi-supervised, depending on whether data objects have been labeled as 'normal' or 'outlier' [4]. In our research, we focused on unsupervised detection methods, since cases had been clustered but not labeled. Outlier detection methods can also be classified into statistical, proximity-based, density based, and clustering-based methods, depending on the assumptions they make [4], [45]. For the purposes of this analysis, we combined two outlier detection methods, proximity-based and clustering-based methods since our main goal was to explore the relationship between data objects and the clusters they belonged to.

Proximity-based methods use a distance measure such as the standard deviation, median rule, or Tukey's OLM, as a way of assessing the similarity between data points. Clustering-based methods focus on exploring the relationship between data objects and their clusters to identify single outliers or a cluster of outliers [19]. Tukey's OLM [44] is one the most commonly used outlier detection methods, it makes no assumptions of normal distribution, and looks at the bottom (i.e., 25th percentile) and top (i.e., 75th percentile) quartiles of a sample to determine the upper and lower limits (i.e., hinges) of a distribution [24], [25]; data objects beyond these limits are labeled as 'outliers'.

However, Tukey's OLM is not always appropriate for asymmetric data, as the number of outliers tends to increase in skewed data [44]. On the other hand, the SD method, allows for researchers to examine the presence of data objects at x standard deviations from the mean value. The non-normality of the data used in this analysis, suggested the SD method as the most appropriate for exploring the existence of contextual outliers. While the SD method is only appropriate for univariate data, the findings in LaBrie et al. [26] suggested this outlier detection method was well suited for this analysis. In their research, LaBrie et al. [26] showed heavily involved gamblers were discouraged by losses, as an increase in %loss often resulted in other variables decreasing (e.g., frequency, intensity, wager amount); these findings suggested that heavily involved gamblers tend to assess the risk of a wager and self-moderate their behaviour (e.g., reducing intensity while increasing gambling duration), the latter similar to the controlled behaviour seen in substance abuse subjects [26]. Similarly, Xuan and Shaffer [55] found heavily involved gamblers tend to have an involvement-seeking and risk-averse gambling behaviour. The results from these studies [26], [55] suggested problem gamblers are likely to show heavily involved gambling behaviour on one aspect of gambling rather than across variables.

The scatterplots shown on Figure 5a through 5c, and the boxplots shown on Figure 7, indicated the existence of heavily involved gambling sessions within Clusters 1, 2, and 4, particularly within the session duration attribute. Thus our focus

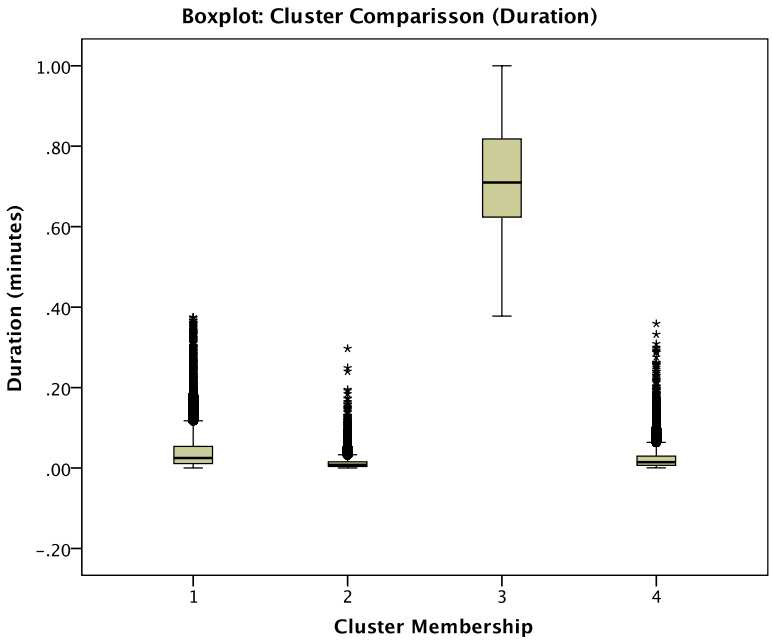


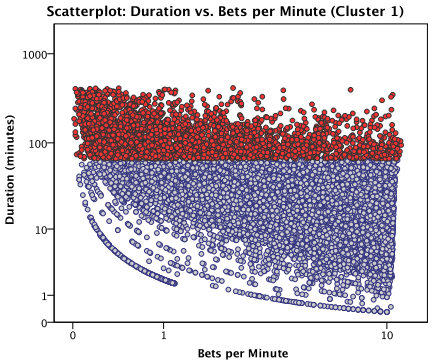
FIGURE 7. Boxplot Analysis on Clustered Data: Duration.

was to identify outliers, within these three clusters, with respect to amount of time spent gambling (i.e., duration). Specifically, we explored the application of the SD outlier detection method to explore the presence of data objects at three standard deviations from the mean duration (i.e., $Mean \pm 3SD$). The results of our outlier analysis are discussed in Section 6.1.

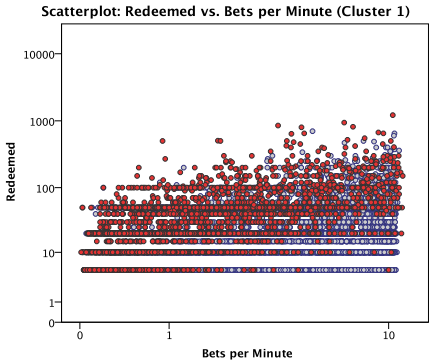
6.1. **Results: Heavily involved gambling sessions.**

6.1.1. *Cluster 1 outliers.* Initially, as shown on Table 5, the 13,843 sessions in Cluster 1 were characterized by having a passive type of gambling activity, with low intensity ($mean = 4.09bpm$), low redeemed amount for most sessions($mean = €30.02$), and an average duration time of 47.14 minutes. The results of this cluster seemed to show that sessions in Cluster 1 could represent non-problem or low-risk problem gambling sessions. When looking at data objects at three (3) standard deviation marks from the average duration time ($mean = 47.14$), we found 3,028 outliers.

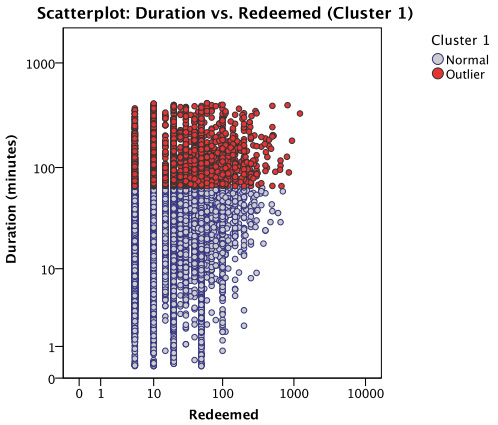
When comparing outliers to normal sessions, as shown on Table 11, Cluster 1 outliers seemed to be more heavily involved in their gambling activity than normal sessions, despite their relatively low involvement. For example, as shown on Figure 8, outliers in this cluster had higher average duration and amount redeemed than normal sessions; the mean duration time for outliers was over five (5) times the average for normal sessions. Though, interestingly, outliers experienced a higher ratio for losses than normal sessions. It seemed that outliers in this cluster were classified into Cluster 1 due to their low gambling intensity and financial involvement (i.e., amount redeemed).



(A) Duration vs. Intensity



(B) Redeemed vs. Intensity



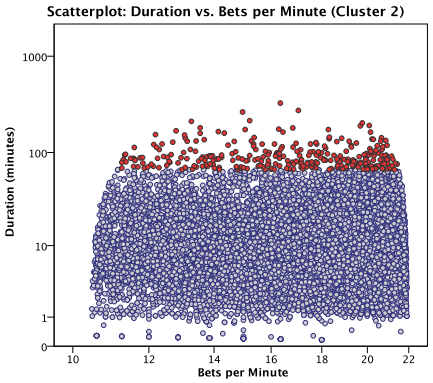
(C) Duration vs. Redeemed

FIGURE 8. Scatterdot: Cluster 1 Outlier Analysis.

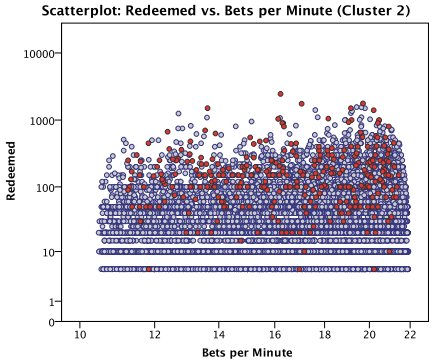
TABLE 11. Cluster 1 Sessions (Normal v. Outliers)

Descriptive Statistics: Cluster 1 Sessions (Normal v. Outliers)								
	Normal				Outlier			
	Mean	SD	Median	Count	Mean	SD	Median	Count
Duration	23.89	16.98		10815	130.20	67.56		3028
Bets	93	93	62		224	292	116	
Intensity	4.67	3.02	4.22		1.99	2.37	.98	
Redeemed	25.74	33.03	20.00		45.31	71.28	20.00	
Voucher	2.02	26.37	.00		.08	4.37	.00	
Net Loss	25.05	33.09	15.00		45.31	71.28	20.00	
%Loss	96.63	17.67	100.00		99.97	1.82	100.00	

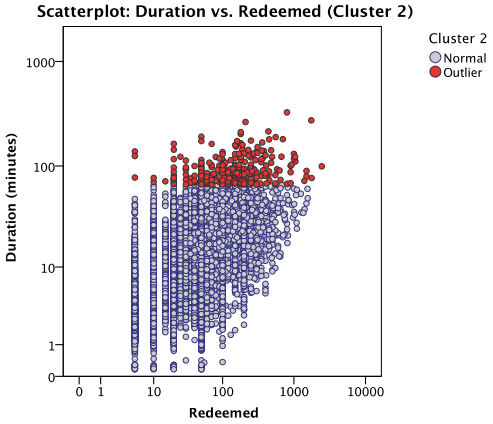
6.1.2. *Cluster 2 outliers.* The 14,496 sessions originally classified into Cluster 2, as shown on Table 6, seemed to be characterized by a more intense type of play than sessions in Cluster 1 and 3. However, their short duration time, and low ratio for losses, implied the gambling activity in these cases may have been influenced by the degree of risk in a wager. When comparing outliers to normal sessions within this cluster, as shown on Table 12, the 298 identified outliers had a much higher gambling involvement than Cluster 1 outliers. For example, the average duration time for Cluster 1 outliers was 7.5 times greater than that of normal sessions. The average amount redeemed in these outliers was over four (4) times greater than normal sessions, and the %Loss experienced by outliers was also higher than normal sessions.



(A) Duration vs. Intensity



(B) Redeemed vs. Intensity



(C) Duration vs. Redeemed

FIGURE 9. Scatterdot: Cluster 2 Outlier Analysis.

Cluster 2 outliers may have been classified into Cluster 2 due to their higher gambling involvement in regards to intensity and amount redeemed. These results suggested the gambling behaviour expressed Cluster 2 outliers could be at a higher risk of problem gambling than normal sessions. While there were clear differences

TABLE 12. Cluster 2 Sessions (Normal v. Outliers)

Descriptive Statistics: Cluster 2 Sessions (Normal v. Outliers)								
	Normal				Outlier			
	Mean	SD	Median	Count	Mean	SD	Median	Count
Duration	12.64	11.73		14198	95.54	34.68		298
Bets	218	209	149		1599	622	1461	
Intensity	17.27	3.18	17.84		16.80	2.95	16.99	
Redeemed	51.40	81.72	25.00		225.45	292.33	150.00	
Voucher	44.46	415.10	.20		24.98	281.31	.00	
Net Loss	39.02	76.65	19.65		219.91	294.73	147.50	
%Loss	71.28	43.23	99.10		94.85	21.61	100.00	

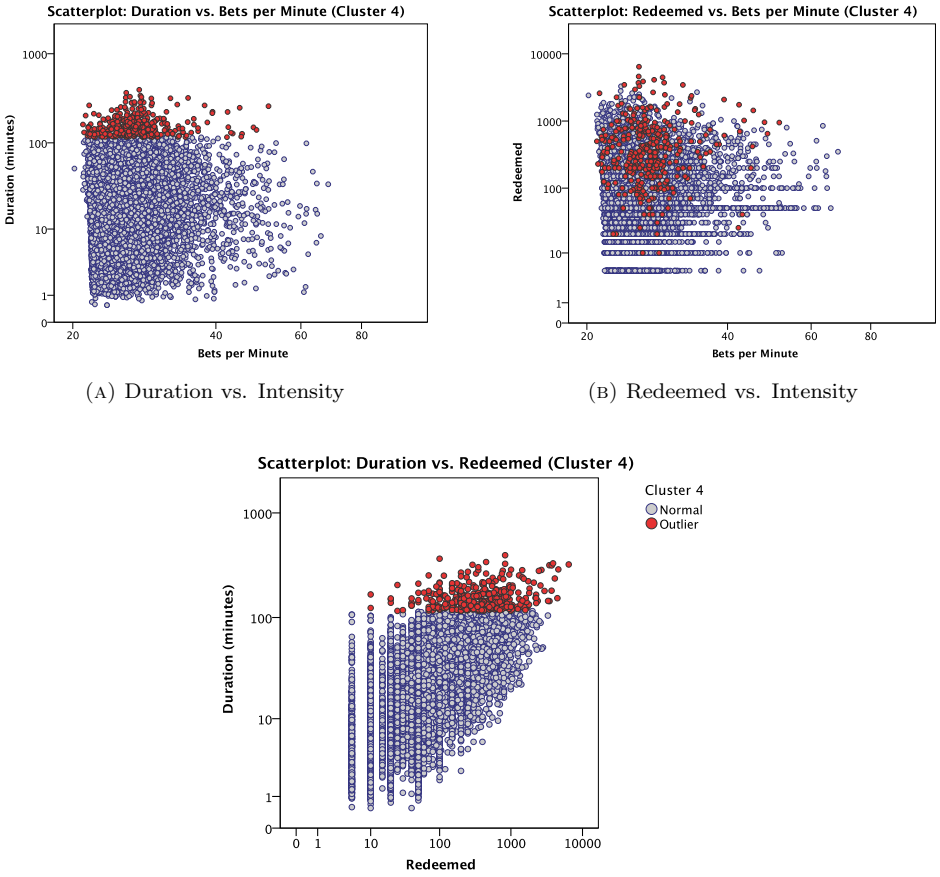
between these two types of sessions in Cluster 2, as illustrated in Figure 9, the results shown in Figure 8b and Figure 9b also suggested the importance of amount redeemed as a measure for EGM gambling involvement.

6.1.3. *Cluster 4 outliers.* A total of 17,253 sessions were classified into Cluster 4, this amount represented 37.2% of the total research sample. Sessions in this cluster, as shown on Table 8, were characterized by high gambling intensity, short duration times, and higher redeemed amounts, which suggested the expressed gambling behaviour in these sessions was, potentially, that of a riskier type of gambling. However, the smaller amount of losses produced by these sessions implied that shorter and more intense sessions may produce smaller losses. There were 346 sessions identified as outliers in this cluster.

TABLE 13. Cluster 4 Sessions (Normal v. Outliers)

Descriptive Statistics: Cluster 4 Sessions (Normal v. Outliers)								
	Normal				Outlier			
	Mean	SD	Median	Count	Mean	SD	Median	Count
Duration	23.04	21.79		16907	162.67	49.02		346
Bets	609	603	401		4519	1656	4071	
Bets per Minute	26.09	3.99	25.32		27.65	4.58	26.88	
Redeemed	106.51	173.11	50.00		623.16	805.07	350.00	
Voucher	94.80	496.65	.45		224.20	598.56	.20	
Net Loss	69.34	151.70	19.85		504.17	746.68	259.93	
%Loss	59.55	45.81	97.00		72.47	40.45	99.97	

When comparing outliers to normal cases within this cluster, as shown on Table 13, outliers were more heavily involved in certain aspects of their gambling activity; see Figure 10. For example, the average duration time for outliers was just over seven (7) times greater than that of normal sessions, and just under two (2) times greater than that of Cluster 2 outliers. Despite little differences in regards to intensity, Cluster 4 outliers had a much higher amount of total bets ($mean = 4,519$),



(A) Duration vs. Intensity

(B) Redeemed vs. Intensity

(C) Duration vs. Redeemed

FIGURE 10. Scatterdot: Cluster 4 Outlier Analysis.

and a higher amount of financial involvement compared to normal sessions. For example, Cluster 4 outliers tended to redeem just over six (6) times more than normal sessions, and over two (2) times the average amount redeemed in Cluster 2 outliers. Nevertheless, Cluster 4 outliers had an average ratio for losses just over the mean amount for normal sessions in Cluster 2. Still, the longer duration of sessions, higher intensity in regards to gambling activity, and high out-of-pocket cost of sessions, indicated a riskier and more involved type of gambling in Cluster 4 outliers, when compared to other clusters; particularly, Cluster 2.

7. **Conclusion.** The overall purpose of this research was to identify gambling personae (e.g., non-problem, low risk, moderate risk, high-risk problem gambler), based on the attributes found within these gambling sessions. Thus allowing the researchers to explore data mining techniques to not only analyze problem gambling through EGMs, but also explore ways to predict the incidence of this condition based on the type of gambling currently taking place. Therefore, a very important part in our research was to identify messages, and attributes related to gambling

activity from which we could extract gambling related data, and determine what constitutes a valid EGM gambling session.

After exploring the G2S protocol, and understanding the structure of G2S messages, it was clear that specific events could mark the start (e.g., player enters a bill) and end of a session (e.g., player cashes out) if particular criteria were met, such as explained in Section 3. The gambling sessions used in this research were collected in situ over a period of one-month. Once sessions were identified, details about their duration, intensity, amount redeemed, final amount won (i.e., vouchers), were extracted. Additional variables such as total number of bets, net loss reported within a session, and percent loss of a session, were later added. The sessions were then clustered into four groups, using k-means, based on the sessions' gambling intensity, duration, and amount redeemed.

A between groups one-way ANOVA test suggested significant differences among the variables' mean score, across all clusters. However, in regards to the intensity and duration of gambling activity in clusters, this difference was not necessarily significant. The results of the cluster analysis suggested that cases in Cluster 1 ($n = 13,843$, 29.8% of sessions), due to their relatively low intensity, the low average amount redeemed, and medium duration time, seemed to consist of non-problem or low-risk gambling sessions. However, the 3,028 outliers identified in this cluster seemed to be more heavily involved in their gambling activity than normal sessions, despite their relatively low intensity, and therefore at a higher risk than normal sessions. Cases in Cluster 2 ($n = 14,496$, 31.2% of sessions) were more involved in their gambling activity, as evidenced by their higher intensity ($mean = 17.26bpm$) and greater financial involvement; interestingly, despite their short duration time, Cluster 2 cases lost a smaller percentage of their amount redeemed, which suggested that cases in this cluster may have had a strategy characterized by quick decisions that assessed the degree of risk in a wager [12], [26], [31]. Cases in Cluster 2 seemed to be at a moderate risk of expressing problem gambling behaviour. The 298 outliers identified in this cluster had a higher level of involvement than normal cases.

Meanwhile, Cluster 3 ($n = 824$, 1.8% of sessions) consisted of sessions with a longer duration time than those cases in Cluster 1 and Cluster 2, despite having the lowest intensity across all clusters and low financial involvement. Cases in this cluster were considered to exhibit non-problem gambling behaviour. Finally, cases in Cluster 4 ($n = 17,253$, 37.2% of sessions) had a higher risk of problem gambling that cases in other clusters due to their high intensity ($mean = 26.12bpm$; $median = 25.35bpm$), and high redeemed amount, despite having a short duration. Sessions in this cluster also had a relatively low average %loss. Thus implying that more intense sessions are likely to have a shorter duration time and may produce more winnings; further evidenced by the results of a correlation analysis. The 346 sessions identified as outliers in this cluster were more heavily involved in regards to duration, intensity, and amount redeemed when compared to sessions in other clusters.

While the use of aggregated data was useful for identifying clusters, a longitudinal analysis, where each gambling event can be analyzed, would allow researchers to analyze the type of gambling strategy used in sessions. Furthermore, this type of study would provide a better understanding of the type of decisions taking place throughout an EGM gambling session based on other measures of gambling involvement, such as wager variability, frequency, and trajectory. Greater data granularity could give more insight into the how the outcome of a bet or bonus round may

affect gambling strategies, as the ability to increase wagers, upon entering bonus rounds, would clearly facilitate a change in strategy. Furthermore, a longitudinal study could help identify any differences based on the game being played. Breaks between gambling events (i.e., bets) could also be analyzed based on the distribution of bets per minute during the length of a session. Finally, a longitudinal analysis of EGM gambling measures could help describe cluster profiles, and identify play-personae, more accurately.

REFERENCES

[1] C. C. Aggarwal, *Outlier Analysis*, Springer, New York, 2013.

[2] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders*, 4^{th} edition, American Psychiatric Association, Washington, DC, 1994.

[3] G. Banks, R. Fitzgerald and L. Sylvan, *Gambling: Productivity Commission Inquiry Report*, Technical Report 50, 2010, http://www.pc.gov.au/inquiries/completed/gambling-2009/report/gambling-report-volume1.pdf(visited on: 09/12/2012).

[4] M. Berry and G. Linoff, *Data Mining Techniques for Marketing, Sales, and Customer Relationship Management*, 2^{nd} edition, Wiley Publishing Inc., Indianapolis, 2004.

[5] J. Braverman, R. A. LaBrie and H. J. Shaffer, A taxometric analysis of actual Internet sport gambling behavior, *Psychological Assessment*, **23** (2011), 234–244.

[6] J. Braverman, D. A. LaPlante, S. E. Nelson and H. J. Shaffer, Using cross-game behavioral markers for early identification of high-risk Internet gamblers, *Psychology of Addictive Behaviors*, **27** (2013), 868–877.

[7] J. Braverman and H. J. Shaffer, How do gamblers start gambling: Identifying behavioral markers for high-risk Internet gambling, *European Journal of Public Health*, **22** (2012), 273–278.

[8] S. Carpendale, Evaluating information visualizations, in *Information Visualization, Lecture Notes in Computer Science*, A simple univariate outlier identification procedure, **4950** (2008), 19–45.

[9] National Research Council, *Pathological Gambling: A Critical Review*, National Academies Press, Washington, DC, 1999.

[10] P. Delfabbro, A. Osborn, M. Nevile, L. Skelt and J. MacMillen, *Identifying Problem Gamblers in Gambling Venues*, Technical report, 2007.

[11] M. J. Dixon, K. A. Harrigan, M. Jarrick, V. MacLaren, J. A. Fugelsang and E. Sheepy, Psychophysiological arousal signatures of near-misses in slot machine play, *International Gambling Studies*, **11** (2011), 393–407.

[12] L. Dixon, R. Trigg and M. Griffiths, An empirical investigation of music and gambling behaviour, *International Gambling Studies*, **7** (2007), 315–326.

[13] S. Dragicevic, G. Tsogas and A. Kudic, Analysis of casino online gambling data in relation to behavioural risk markers for high-risk gambling and player protection, *International Gambling Studies*, **11** (2011), 377–391.

[14] M. Ellery, S. H. Stewart and P. Loba, Alcohol's effects on video lottery terminal (vlt) play among probable pathological and non-pathological gamblers, *Journal of Gambling Studies*, **21** (2005), 299–324.

[15] J. Ferris and H. Wynne, *The Canadian Problem Gambling Index: Final Report*, Technical Report, 2001, http://www.ccgr.ca/en/projects/resources/CPGI-Final-Report-English.pdf(visited on: 06/28/2013).

[16] G. Data, *Canadian Gaming Market Report*, Technical report, 2011, http://www.gamblingdata.com/files/Gambling%20Data%20Canadian%20Gaming%20Market%20Report%20Final_0.pdf (visited on: 04/10/2013).

[17] GSA, *G2S Message Protocol v1.1 Game-to-system*, Technical Report GSA-P0075.024.00-2011, GSA, 2011.

[18] GSA, *G2S Message Protocol v2.0 Game-to-system*, Technical Report GSA-P0075.0800.00-2006, GSA, 2006.

[19] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, 3^{rd} edition, Morgan Kaufmann, Waltham, 2012.

[20] K. A. Harrigan and M. Dixon, Par sheets, probabilities, and slot machine play: Implications of problem and non-problem gambling, *Journal of Gambling Issues*, **23** (2009), 81–110.

[21] K. A. Harrigan, Slot machine structural characteristics: Distorted player views of payback percentages, *Journal of Gambling Issues*, **20** (2007), 215–234.
[22] K. A. Harrigan, Slot machines: Pursuing responsible gaming practices for virtual reels and near misses, *International Journal of Mental Health Addiction*, **7** (2009), 68–83.
[23] C. Hennig, Cluster-wise assessment of cluster stability, *Computational Statistics & Data Analysis*, **52** (2007), 258–271.
[24] D. C. Hoaglin, John W. Tukey and data analysis, *Statistical Science*, **18** (2003), 311–318.
[25] B. Iglewicz and S. Banerjee, *A Simple Univariate Outlier Identification Procedure,* Proceedings of Annual Meeting of the American Statistical Association, 2001.
[26] R. A. LaBrie, D. A. LaPlante, S. E. Nelson, A. Schumann and H. J. Shaffer, Assessing the playing field: A prospective longitudinal study of Internet sports gambling behavior, *Journal of Gambling Studies*, **23** (2007), 347–362.
[27] R. A. LaBrie, S. A. Kaplan, D. A. LaPlante, S. E. Nelson and H. J. Shaffer, Inside the virtual casino: A prospective longitudinal study of actual Internet casino gambling, *European Journal of Public Health*, **18** (2008), 410–416.
[28] D. A. LaPlante, S. E. Nelson, R. A. LaBrie and H. J. Shaffer, Stability and progression of disordered gambling: Lessons from longitudinal studies, *Canadian Journal of Psychiatry*, **53** (2008), 52–60.
[29] D. A. LaPlante, S. E. Nelson, R. A. LaBrie and H. J. Shaffer, Disordered gambling, type of gambling and gambling involvement in the British gambling prevalence survey 2007, *European Journal of Public Health*, **21** (2011), 532–537.
[30] H. Liu and V. Keselj, Combined mining of web server logs and web contents for classifying user navigation patterns and predicting users' future requests, *Data & Knowledge Engineering*, **61** (2007), 304–330.
[31] P. Loba, S. H. Stewart, R. M. Klein and J. R. Blackburn, Manipulations of the features of standard video lottery terminal (VLT) games: Effects in pathological and non-pathological gamblers, *Journal of Gambling Studies*, **17** (2001), 94–98.
[32] V. V. MacLaren, J. A. Fugelsang, K. Harrigan and M. Dixon, The personality of pathological gamblers: A meta-analysis, *Clinical Psychology Review*, **31** (2011), 1057–1067.
[33] K. Marshall, *Gambling 2011*, Technical Report 4, 2011, http://www.statcan.gc.ca/pub/75-001-x/2011004/article/11551-eng.pdf(visited on: 04/10/2013).
[34] S. Mishra, M. L. Lumiére and R. J. Williams, Gambling as a form of risk-taking: Individual differences in personality, risk-accepting attitudes, and behavioral preferences for risk, *Personality and Individual Differences*, **49** (2010), 616–621.
[35] National Research Council, *Pathological Gambling: A Critical Review*, The National Academies Press, Washington D.C., 1999.
[36] S. R. Nelson, D. A. LaPlante, A. J. Peller, A. Schumann, R. A. LaBrie and H. J. Shaffer, Real limits in the virtual world: Self-limiting behavior of Internet gamblers, *Journal of Gambling Studies*, **24** (2008), 463–477.
[37] J. Pallant, *SPSS Survival Manual: A Step By Step Guide to Data Analysis Using SPSS*, 4^{th} edition, Allen & Unwin, Sydney, 2011.
[38] Y. Peng, K. Gang and Y. Shi (eds.), Knowledge–rich data mining in financial risk detection, in *Computational Science – ICCS 2009* (eds. G. Allen, J. Nabrzyski, E. Seidel, G. D. van Albada, J. Dongarra and P. M. A. Sloot), Springer Berlin Heidelberg, **5545** (2009), 534–542.
[39] D. T. Pham, S. S. Dimov and C. D. Nguyen, Selection of k in k-means clustering, *Journal of Mechanical Engineering Science*, **219** (2005), 103–119.
[40] A. Rakhlin and A. Caponnetto (eds.), Stability of k-means clustering, in *Advances in Neural Information Processing Systems 19* (eds. B. Schölkopf, J. Platt and T. Hoffman), MIT Press, (2006), 1121–1128. http://papers.nips.cc/paper/3116-stability-of-k-means-clustering (visited on: 12/10/2014)
[41] Responsible Gambling Council, *Electronic Gaming Machines and Problem Gambling*, Saskachewan Liquour and Gaming Authority, 2006, http://www.responsiblegambling.org/docs/research-reports/electronic-gaming-machines-and-problem-gambling.pdf?sfvrsn=10 (visited on: 06/28/2013).
[42] Responsible Gambling Council, *Canadian Gambling Digest 2011-2012*, Technical report, 2013, http://www.responsiblegambling.org/docs/default-document-library/20130605_canadian_gambling_digest_2011-12.pdf?sfvrsn=2 (visited on: 05/04/2015).
[43] G. Schwartz, *The Impulse Economy*, Atria Books, New York, 2011.

[44] S. Seo, *A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets*, M.S thesis, University of Pittsburg in Pensylvania, 2006.
[45] H. J. Shaffer and D. A. Korn, Gambling and related mental disorders: A public health analysis, *Annual Review of Public Health*, **23** (2002), 171–212.
[46] H. J. Shaffer, A. J. Peller, D. A. LaPlante, S. E. Nelson and R. A. LaBrie, Toward a paradigm shift in Internet gambling research: From opinion and self-report to actual behavior, *Addiction Research and Theory*, **18** (2010), 270–283.
[47] J. Sim and C. C. Wright, Understanding interobserver agreement: The Kappa statistic, *Family Medicine*, **37** (2005), 360–363.
[48] S. H. Stewart, P. Collins, J. R. Blackburn, M. Ellery and R. M. Klein, Heart rate increase to alcohol administration and video lottery terminal (VLT) play among regular VLT players, *Psychology of Addictive Behaviors*, **19** (2005), 94–98.
[49] S. Tufféry, *Data Mining and Statistics for Decision Making*, John Wiley & Sons, Ltd., Chichester, 2011.
[50] A. J. Viera and J. M. Garrett, The Kappa statistic in reliability studies: Use, interpretation, and sample size requirements, *Journal of the American Physical Therapy Association*, **85** (2005), 257–268.
[51] C. Wheelan, *Naked Statistics: Stripping the Dread from the Data*, W.W. Norton and Company, New York, 2013.
[52] R. J. Williams, R. A. Volberg and R. M. G. Stevens, *The Population Prevalence of Problem Gambling: Methodological Influences, Standardized Rates, Jurisdictional Differences, and Worldwide Trends*, Technical report, 2012, https://www.uleth.ca/dspace/bitstream/handle/10133/3068/2012-PREVALENCE-OPGRC%20(2).pdf?sequence=3 (visited on: 08/12/2013).
[53] D. S. Wilson, R. A. Kauffman and M. S. Purdy, A program for at-risk high school students informed by evolutionary science, *PLoS ONE*, **6** (2011), e27826.
[54] I. H. Witten and E. Frank, Data mining: Practical machine learning tools and techniques, *Newsletter: ACM SIGMOD Record Homepage archive,* **31** (2002), 76–77.
[55] Z. Xuan and H. Shaffer, How do gamblers end gambling: Longitudinal analysis of Internet gambling behaviors prior to account closure due to gambling related problems, *Journal of Gambling Studies*, **25** (2009), 239–252.

Received May 2017; revised October 2017.

E-mail address: mosquera@cs.dal.ca
E-mail address: vlado@cs.dal.ca