Complex Systems Informatics and Modeling Quarterly
CSIMQ, Issue 9, December 2016\January 2017, Pages 44–66
Published online by RTU Press, https://csimq-journals.rtu.lv
https://doi.org/10.7250/csimq.2016-9.03
ISSN: 2255-9922 online

Using Data-Driven and Process Mining Techniques for Identifying and Characterizing Problem Gamblers in New Zealand

Suriadi Suriadi[1⋆], Teo Susnjak[2⋆], Agate M. Ponder-Sutton[2], Paul A. Watters[2] and Christoph Schumacher[2]

[1]Queensland University of Technology, Information Systems School,
2 George St, Brisbane, QLD 4000, Australia
[2]Institute of Natural and Mathematical Sciences, Massey University, Albany Expressway (SH17), Albany, Auckland 0632, New Zealand

s.suriadi@qut.edu.au (orcid.org/0000-0002-6311-5927), t.susnjak@massey.ac.nz,
a.m.ponder-sutton@massey.ac.nz (orcid.org/0000-0002-9683-8718),
p.a.watters@massey.ac.nz, c.schumacher@massey.ac.nz

Abstract. This article uses data-driven techniques combined with established theory in order to analyse gambling behavioural patterns of 91 thousand individuals on a real-world fixed-odds gambling dataset in New Zealand. This research uniquely integrates a mixture of process mining, data mining and confirmatory statistical techniques in order to categorise different sub-groups of gamblers, with the explicit motivation of identifying problem gambling behaviours and reporting on the challenges and lessons learned from our case study.

We demonstrate how techniques from various disciplines can be combined in order to gain insight into the behavioural patterns exhibited by different types of gamblers, as well as provide assurances of the correctness of our approach and findings. A highlight of this case study is both the methodology which demonstrates how such a combination of techniques provides a rich set of effective tools to undertake an exploratory and open-ended data analysis project that is guided by the process cube concept, as well as the findings themselves which indicate that the contribution that problem gamblers make to the total volume, expenditure, and revenue is higher than previous studies have maintained.

Keywords: Data mining, process mining, confirmatory statistics, problem gambling.

1 Introduction

Pervasive use of digital technologies has left rich digital traces about not only the physical objects of our world (e.g. purchase orders, sales figures) but also about our activities and our *behaviours*

⋆ Corresponding author

Reference: S. Suriadi, T. Susnjak, A.M. Ponder-Sutton, P.A. Watters and C. Schumacher, "Using Data-Driven and Process Mining Techniques for Identifying and Characterizing Problem Gamblers in New Zealand," Complex Systems Informatics and Modeling Quarterly, CSIMQ, no. 9, pp. 44–66, 2016. [Online]. Available: https://doi.org/10.7250/csimq.2016-9.03

(e.g. emotions, sentiments, relationships, and interactions) [1]. The encoding of our activities and behaviours as digital traces enables new evidence-based studies that would have been difficult just a few decades ago, e.g. the prediction of students' success through the use of learning management systems data [2], and the discovery of working patterns of human resources in organisations [3].

Techniques from the domain of data mining [4], machine learning [5], and other statistical analysis have been frequently and successfully applied to extract useful insights from data. The increased availability of detailed event logs (due to the widespread use of process-aware information systems) coupled with maturing process mining techniques [6] have recently enabled wider applications of *process mining* in organisations around the world, such as in a large Australian insurance organisation [7] and others [8], [9], [10], [11].

This article[1] reports on the techniques applied, challenges and lessons learned from our case study where a mixture of process mining, data mining, and confirmatory statistical techniques are applied to analyse a data set containing information about all fixed-odds bets (FOBs) recorded by a gambling service provider in New Zealand during a timeframe between 2013-2014. This study attempts to identify and characterize various groups of gamblers (with a focus on problem gamblers) directly from the data. The nature of this case study is exploratory: we do not "label" our data with various classes of gamblers; rather, we attempt to learn how many groups of gamblers can be discerned from the data.

To address these questions, we had to analyse the data both at an "aggregated level" (that is, treating each gambler as an individual entity), as well as at an "event level" (that is, diving into the details of the behaviours of each group of gamblers). The need to go into the details of the behaviours of gamblers also posed a practical challenge that arose from computational resources limitation, relative to the size and the complexity of the data to be analysed.

The exploratory nature of this study suggested the use of unsupervised learning techniques, such as k-means clustering [13] as the starting point of analysis. While the use of clustering to analyse problem gamblers is not new, we found that cluster analysis alone is not sufficient as it tends to focus on aggregated, "static" characteristics of gamblers (such as gambling frequency and socio-demographic status) and ignores the behavioural aspects of a gambler, such as *how does a gambler behave when he/she loses a bet*. More importantly, it does not offer any insights into the behaviour of gamblers seen in each cluster. We contend that this could be a contributing factor to the experience found in other similar studies, e.g. [14] which found the use of k-means clustering in their analysis to be rather ineffective.

In our case study, *we demonstrate how both process mining analysis*, with its proven ability to extract detailed behavioural insights from fine-grained and chronologically-arranged data, as evidenced by the insights obtained from previous process mining case studies [7], [8], [9], [15], *and confirmatory statistics, can be weaved together with clustering analysis to not only understand the variety of behaviours exhibited by gamblers, but also to evaluate the results*.

The integration of data mining techniques in process mining case studies is not new, e.g. [16], [17]. It is mostly performed to provide an explanation of behaviours extracted from process mining analysis. However, the integration is often lightweight: often performed towards the end of the analysis stage and is mostly limited to simple *classification/regression* analysis [16], [17]. Jain [18], on reviewing k-means clustering [13] also reported that cluster analysis may have to be supplemented with other analysis techniques as necessary. The inclusion of confirmatory statistical analysis techniques in process mining case studies is, however, rarer, e.g. hypothesis testing can be effectively used to assert, with high assurance, the significance of differences observed in various intra- or inter-organisation processes. However, existing cross-organisational process mining case studies (which could benefit most from such hypothesis testing), e.g. [10], [15], have not done so.

The case study highlights how process mining, clustering analysis, classification analysis, and confirmatory statistics can contribute equally, as an expression of the process cube concept [19], to

[1] This article is an extension of the work published in [12] by the same authors.

extract *sound* insights about problem gamblers' behaviours. With a wide range of tools at our disposal, we were able to divide our data into multiple clusters and reframed our case study question as a problem of extracting and understanding the similarities and differences in the behaviours exhibited by gamblers within each, and across various, clusters. We could, thus, build a picture starting from complex clusters and "funnel down" (drill-down) our attention to those few clusters that exhibited interesting behaviours for more refined analysis[2]. By focusing our attention on behaviours, we saw this as the first step towards eventually building a robust predictive model for problem gamblers.

The main contribution of this article is on the reporting of challenges and lessons learned in the application of these three classes of analysis techniques. Most importantly, this article details how these three techniques can be strategically interwoven to extract sound insights about problem gamblers' behaviours. Furthermore, we also highlight practical challenges that arise when one analyses a relatively large size of data with limited computational resources.

Section 2 describes the classes of analysis techniques employed and brief psychological theory about problem gamblers that informs our analysis. Section 3 summarizes the approach taken in this case study. Section 4 details the analyses performed at each stage of the case study (including evaluation of analysis results), along with the challenges and lessons learned.

A discussion about the related work is provided in Section 5, followed by the conclusion.

2 Background

To understand the context that drives the way in which we conducted our data analysis, this section provides a succinct introduction into psychological theory underpinning problem gamblers and a brief description of process mining, clustering and classification analysis, and confirmatory statistics.

Problem gamblers. From a psychological perspective, the basic mechanics of how addictive behaviours, such as problem gambling, are learned are well understood: they involve a complex set of interactions between actions and rewards. Reinforcement learning (or operant conditioning) works on the principle that a behaviour will increase when it leads to a pleasurable outcome. Depending on the salience (or importance) of the reward to an individual, they will seek to maximise the pleasure experienced by increasing the frequency of the behaviour. At the same time, studies have shown that brains tend to develop a tolerance level as the frequency of such pleasurable experience increases, leading to the need to increase the intensity of the pleasure [22].

This tolerance reveals a very powerful feedback loop between stimuli, responses, and rewards, which can be characterised as a process, hence the focus in this article is on using process mining as a tool to understand problem gambling, since the sequence of events is critical for learning to occur. For instance, if a stimulus is presented after (or at the same time of) a behaviour occurring, then no association is learned between the two, and without the reinforcement of that behaviour by the presentation of a reward, learning will fail. This also reveals the potential of the process analytics approach, since it may be possible to suggest changes or interventions in the gaming process which might reduce the harm from problem gambling, perhaps by providing additional parallel feedback about losses accumulated and the likelihood of success.

Process mining. Process mining [6] is a specialized class of techniques used to analyse data related to the processes of organisations. A process consists of multiple activities, executed in a particular order, to achieve a particular goal (e.g. an insurance claim process). A process is

[2] As elaborated later in this article, the "drill-down" approach, manifested through a form of recursive clustering, is an appropriate approach as recent studies shows that problem gamblers only affect 0.7% of the population in New Zealand [20], [21].

typically expressed as a process model using graphical notations, e.g. BPMN.[3] The term *case* or *process instance* represents *one execution* of a particular process (e.g. the insurance claim process for Customer A). Data used in process mining is known as an *event log* which consists of *events*. At minimum, an *event* is described by an activity (i.e. the task related to the event, such as placing a bet), a timestamp (i.e. the moment the event occurred), and a *case identifier* (i.e. the *case* to which the event belongs).

Typical process mining analysis includes *process discovery* (where process models are extracted from event logs), *conformance checking* (where the actual behaviour captured in the log is compared against the expected behaviour of the process), process performance analysis (e.g. average case duration), and resource analysis (e.g. collaboration among resources, resources working pattern). A process mining case study typically results in *detailed evidence-based insights* about the *behaviours and characteristics* of the process being studied.

Cluster and classification analysis. Clustering and classification algorithms have been extensively studied and are common in data mining and machine learning text books [4], [5].

Cluster analysis attempts to find groups whose individual members share sufficient commonality within a given cluster, and distinctiveness from samples in remaining clusters. One of the simplest and most widely-used clustering algorithms is k-means clustering [13] due to its ease of use and empirical success. As suggested by Jain [18], for multidimensional data, best practices recommend the use of other complementing analytical techniques to assert the distinct characteristics of each cluster. In this article we use process mining analysis to do so: we extract distinct behavioural patterns from each cluster and use them to support the uniqueness of each cluster (see Section 4.5). Cluster analysis is primarily exploratory and is categorised as unsupervised machine learning since objects' labels or class memberships are not known *a priori*.

Unlike clustering, *classification analysis* which is a form of supervised machine learning, relies on each data point having a unique attribute called "Label" which assigns the data item to a particular class to which it belongs (e.g. a gold customer vs. a platinum customer). Other variables in each data item (e.g. income, postcode, profession), also known as predictor variables, are used by the classification algorithm as sources of information to learn the strength of the correlation between one or more of those predictor variables and the label. Classification analysis aims to explain the response variable using predictor variables.

Confirmatory statistics. Confirmatory statistics can measure the strength and importance of the patterns and results found by the exploratory analysis. The data used in this article had highly non-parametric, long tailed distributions. Kruskal-Wallis [23] is one of the most commonly used non-parametric tests; as such it makes no assumptions about the distribution of the data and was, thus, deemed suitable. To determine significant differences between two groups, pair-wise Dunn's test [24] is frequently applied in literature in post hoc analysis since it retains the rankings used by the Kruskal-Wallis test. The larger the number of pair-wise Dunn's tests performed (especially true for a large number of groups), the greater the likelihood of finding false positive statistical differences. To account for this increased probability, the Bonferoni correction was applied. In this article, we used the Kruskal-Wallis test [23] with Dunn's test [24] and the Bonferoni correction to find statistical differences across various clusters of gamblers.

[3] www.bpmn.org

3 Approach

While we applied a range of techniques from multiple disciplines, our case study employed the PM^2 approach [25] because: (1) the starting point of our analysis was an event log (the log used in our study consisted of betting *events* per account holder), of which process mining is designed for, and (2) the PM^2 methodology is rather detailed in its guidance on each stage, and flexible enough to allow the inclusion of other types of classical data mining techniques (e.g. clustering).

The PM^2 methodology consists of 6 stages: planning (identification of process to analyse and research questions, project team formation), data extraction (determining scope of data to be extracted, data interpretation), data processing (refinement of extracted data to build the optimal "views" for analysis), mining and analysis (application of various analysis techniques to answer research questions), evaluation (diagnosis, verification, and validation of results), and process improvement (application of results to improve processes).

We regarded process cube as a fine-grained concept that guided us in the way in which we pre-processed and analysed our data, while the PM^2 provided the overall structure for our case study.

4 Case Study

This case study entailed the first five stages of PM^2, with a focus on the data processing, data mining, clustering and analysis, and evaluation stages. Key activities that were undertaken at each stage are explained, followed by the challenges encountered and lessons learned.

4.1 Planning

The stakeholder (an economist) was consulted about the open questions from the data set. In our case study, our stakeholder was interested in understanding the strategy and behaviours of various types of gamblers. The questions that are of interest include: (RQ1) "How many groups of gamblers can be discerned from the data set?" and (RQ2) "Which users were most likely problem gamblers from the data?". The team formed for this case study consisted of experts in the domain of data mining, process mining, statistics, and the stakeholder.

4.2 Data Extraction

The data set was extracted from a New Zealand gambling service provider[4]. It contained information about all FOBs placed in New Zealand over a 9-month period (from August 2013 to May 2014) which were associated with anonymised customer accounts. SOGS-R and SOGS-3M are some of the most cited screening instruments used in psychology for diagnosing pathological and problem gambling [26]. These survey-based instruments examine gambling related responses from gamblers over the previous six and three months of their gambling patterns respectively [27], [28], which is in line with the timeframe covered by our data. Furthermore, practical limitation of our computing resources (three 8-16GB/i5 Core workstations and one 32GB/i5 Core virtual machine) also limited the amount of data that we could analyse.

The following figures depict descriptive statistics about the dataset. Figure 1 shows the distribution of the monetary volume of betting across the nine months that the dataset covers which includes some of the most active betting months in New Zealand (which typically occur during southern hemisphere summer).

The data set extracted was in an event log format where each line of data represents a betting event. The majority of the bets in our data set were placed through the Internet channel, followed by mobile betting (see Figure 2). Consultation with the stakeholder was needed to interpret the data

[4] New Zealand Racing Board

and to decide the scope of data used for analysis. Important attributes from the data set included: “Account ID” (the unique identifier for each gambler), “Timestamp” (the time when the bet was lodged to the system), “Bet Outcome” (whether the gambler won or lost the bet), and “Bet Value” (the bet amount). Other attributes from the dataset were not used in this study.

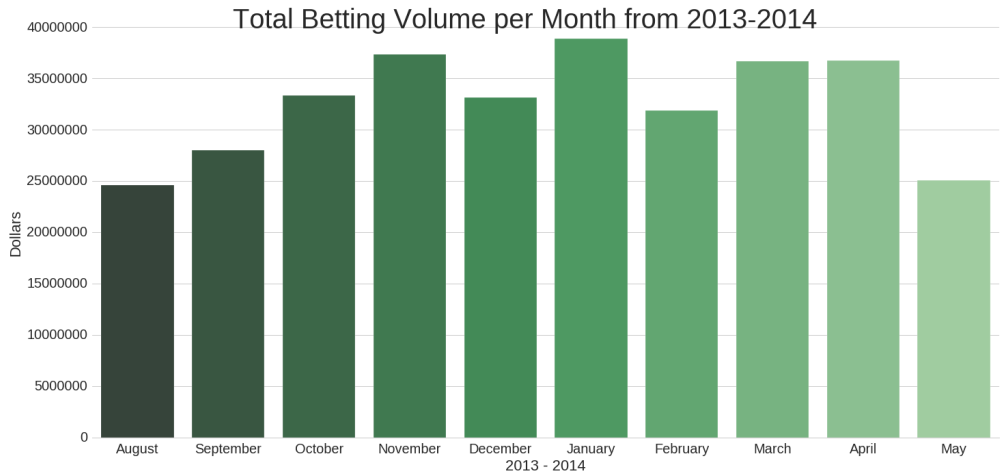

Figure 1. Total number of bets per month

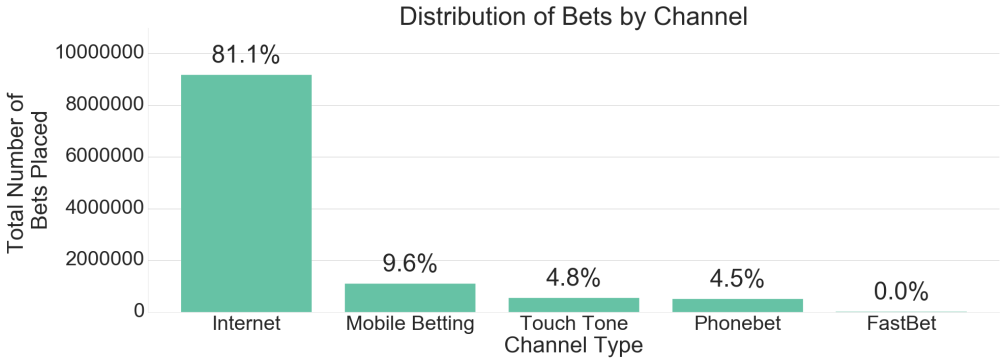

Figure 2. Distribution of bets by betting channel

Lastly, Figure 3 shows the distribution of the frequency of bets placed per product[5]. Various forms of racing generated the bulk of the betting activity, followed by regionally popular sports such as Rugby, Football and Cricket.

4.3 Data Processing

The dataset consisted of 11,311,892 betting events executed by 91,405 account-holders. The number of betting events per case (or per gambler) varies quite widely, from only a few gambling events to millions of gambling events (for the latter, we suspect them to be professional gamblers). The data was pre-processed: (1) because our experiences show existing implementations of many process mining techniques do not scale well; many process mining case studies used event logs that are significantly smaller (between a few thousand events to just over 1 million events, e.g. [7], [29]); and (2) because the unsupervised and supervised learning algorithms used in our case study were run using aggregated “case log” granularity, not using detailed event-level granularity.

[5] For readability, the figure excludes less popular products with under 0.5% frequency such as: Darts, Boxing, Motorsport, Yachting, Hockey, Snooker, Athletics, Speedway, FOB Racing, Bowls, Surfing, Motorcycling, Cycling, Shearing, Softball, Triathlon.

Applying the concept of the process cube [19] would call for slicing and dicing event logs from multiple dimensions. However, the exploratory nature of this case study meant that there were no clear filtering dimensions that could be used to slice the data. Therefore, an unsupervised learning technique was a suitable tool as it allowed us to find the "natural separation" of gamblers in the data set. We used the k-means algorithm [13] to cluster three problem gambling features from the data: bet frequency (absolute frequency), the ratio (in dollar amount) of winnings over amount lost (referred to as *win/loss dollar ratio*, and ratio of the number (count) of bets placed that result in winning as compared to bets that were lost (referred to as *win/loss ratio*). Seven clusters were chosen due to: the distribution of the data, the doubling plus one of the number of groups the data owner used; and the insights from the *Within Sum of Squares* (WSS) analysis which showed that the cohesiveness of the clusters converged to an optimal in the range of k=7 and k=8 (k referring to the number of clusters).

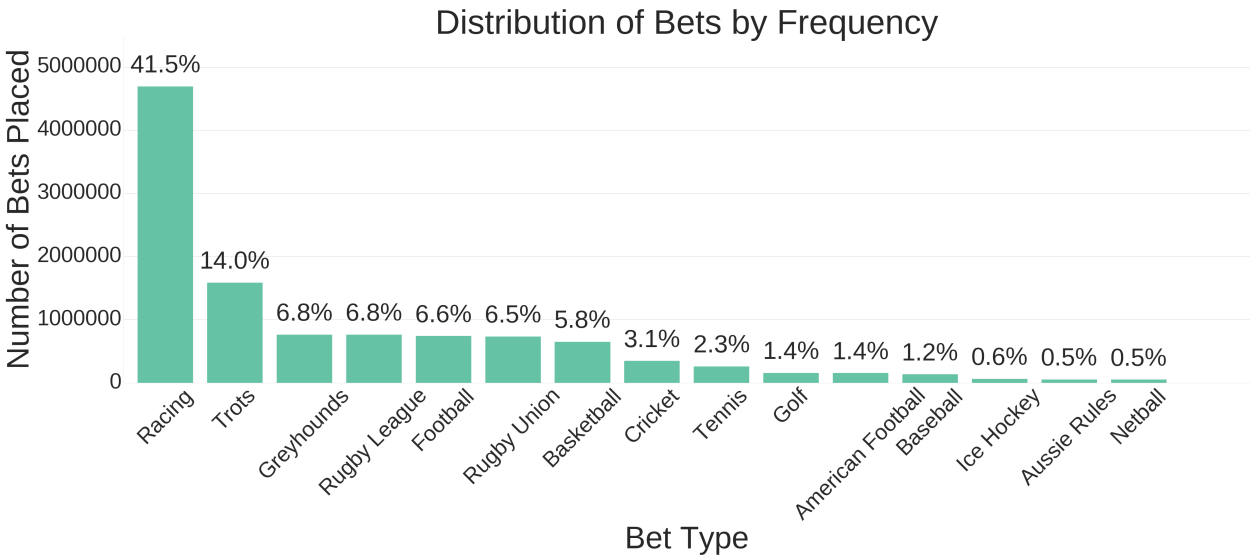

Figure 3. Distribution of bets by betting product for the top 15 products

Next, an event log was derived for each cluster. We used a gambler's account identifier as the case ID in the event log to trace the behaviour of each gambler. The timestamp values in the original data were interpreted as the times respective bets were made. For the activity field, the data was preprocessed to approximate a gambler's reward feedback loop mechanism (see Section 2) in terms of their subsequent action upon win or loss of a previous bet – referred to as a gambler's *clawback* behaviour.

An activity name was created as the concatenation of (1) the outcome of the previous bet (`win` or `loss`) and (2) the amount of money placed in the next immediate bet as a proportion of the amount of money placed in the previous bet: less than or equal to the previous bet amount ($<= 1$), up to double the previous amount (`1to2`), or more than double (>2). For instance, the activity name `loss_<= 1` would mean that a user having lost the previous bet, had subsequently placed another bet where the bet amount was less than or equal to the previous bet. Table 1 shows a sample of the event log obtained for each cluster. The name of the activity for the very first bet is `Initial Bet`.

Outcome. We obtained 7 smaller event logs for each cluster that are of manageable size for deeper analysis.

Challenges. The large dataset posed analysis challenges due to the lack of scalability of the software tools used. Attempts to generate an XES file (i.e. the log format expected by most process mining tools) were a problem: we were limited by the number of events that

Table 1. Event log derived from the raw gambling data set

Case ID	Timestamp	Activity	Bet Outcome	Bet Value
12345	03/03/2012 08:56:12	Initial Bet	Win	25.00
12345	03/04/2012 07:34:22	Win_1to2	Loss	28.00
6789	03/03/2012 22:30:01	Initial Bet	Loss	2.00
6789	03/05/2012 23:11:34	Loss_>2	Loss	8.00
...	..	...	...	...

can be imported in Disco (www.fluxicon.com/disco) tool. We could have used the ProM Tool (www.processmining.org) to convert the original CSV-formatted log into XES; however, it would not have been practical: the generated XES file (in XML format) would have been substantially larger than the original CSV data. It would not have scaled well as is proven later in our analysis.

Slicing the original data into smaller sets was crucial. Trace clustering algorithms, e.g. [30], could not have been applied as they would have required the entire event log to be analysed (not feasible with our resources). This was tackled by using k-means clustering [13] based on aggregated features at a coarser case-level granularity. This required deciding the optimal number of clusters (k) to be used. We combined the advice from the stakeholder with the WSS analysis which showed that the cohesiveness of the clusters converged optimally in the range of k=7 and k=8. We, therefore, decided to use 7 clusters.

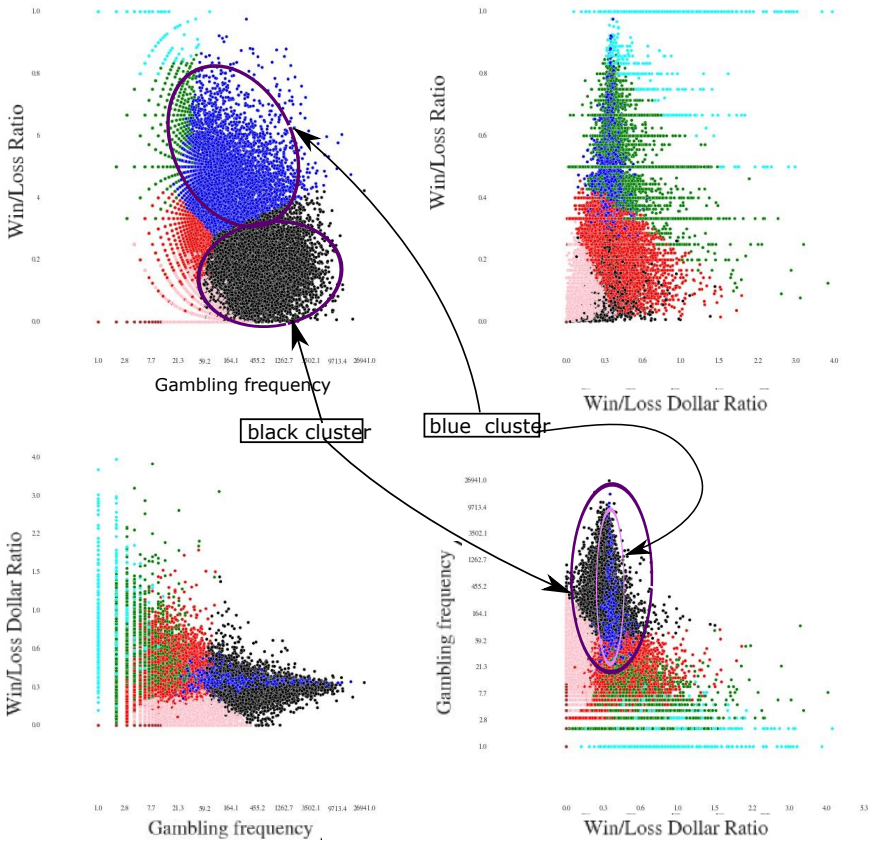

Figure 4. Two-dimensional projections of 3D visualisation of the 7 clusters (colour-coded) based on gambling frequency, win/loss ratio, and win/loss dollar ratio

Another challenge that we faced was related to the long-tailed distribution and large gaps in data points (sparse data) – a phenomenon that is commonly seen in process mining case

studies, e.g. [7], [9]. Such a data set renders k-means clustering less effective and makes data visualisation challenging (as high-ranged values would dominate the graphs). This was, unfortunately, the state of our data. The aggregated attributes used for clustering (i.e. bet frequency, win/loss ratio, and win/loss dollar ratio) were sparse. For instance, the absolute frequency of bet ranges from just one bet per gambler, to over one millions bets.

We overcame this issue by applying two data normalisation procedures: logarithmic normalisation of the values (which unfortunately did not manage to "squeeze" the data range to an acceptable scale) and min-max normalisation to force a particular range of values to be respected. Through these normalisation procedures, we managed to ensure that the range of values used in our k-means clustering were within an acceptable range, and that the results of the clustering could be properly visualized and interpreted. Figure 4 visualizes the population of each cluster (colour-coded) against the three attributes used.

Lessons learned. *The effectiveness of combining unsupervised k-means clustering and aggregated case-level attributes* to slice a substantially large data set was highlighted in this analysis. We "deflated" the data size from over 11 million events to just over 94 thousand lines of data (each line represented one gambler). By "deflating", we mean that instead of slicing the data at event-level, we used case-level perspective (by aggregating event-level attributes to case-level attributes). Therefore, instead of having to deal with 11 million events, we considered only 94 thousand lines of data (each representing a case with its aggregated attributes) and cluster them into 7 clusters using k-means clustering. Once we knew the assignment of each gambler to a particular cluster, we could build a smaller event log for each cluster by retrieving the relevant events for each gambler from the original data set using the case ID as the link.

We also developed a strategy to find a meaningful number of clusters. By *doubling the number of initially-suspected clusters and cross-referencing it with WSS analysis*, an estimate of an optimal number of clusters could be made which was conservative enough to reduce the error probability of not finding enough clusters and fine-grained enough to account for the complexity of the data. This insight was supported by the subsequent evaluation stage (Section 4.5).

4.4 Mining and Analysis

To answer RQ1 and RQ2, we had to determine if the key markers for problem gamblers differed across clusters. To this end, we applied process mining techniques to extract behavioural features which allowed further analysis.

Analysis of the initial 7 clusters. The clawback (chasing losses) behaviour is refered to as a key problem gambling psychological feature (see Section 2). We studied the clawback behaviours in each cluster by generating *process models* for each cluster using the Disco tool. Then, through visual comparisons (similar to Suriadi et al. [10] and Partington et al. [15]), we extracted 9 distinct flow patterns that are dominant in some clusters but not in others. These patterns capture a sequence of activities which may include one loop. For instance, consider the pattern `loss_<=1` followed by `loss_<=1`. In the log, if we see three consecutive occurrence of the activity `loss_<=1`, we counted that pattern as having occurred twice. Interestingly, *6 out of the 9* patterns represented lost-related clawback behaviours, e.g. continuous iterations of the activity `loss_<=1` and the loop between `loss_<=1` and `win_<=1` activities (Figure 5).

Visual analysis only informed whether certain patterns were more "prominent" in one cluster over another. It did not imply any statistical significance. We, therefore, extracted the distribution (in terms of frequency) of the incidence of the 9 patterns from all clusters. For instance, the distribution of the frequency of Pattern 3 and Pattern 4 is shown in Figure 6. Based on the frequency distribution of these patterns, we then performed Kruskal-Wallis tests in conjunction with Dunn's tests. The results of these tests confirmed that *the distribution of these patterns across the seven*

clusters were indeed different at the p-value of 0.05. In addition, from box-and-whisker plots, we also observed that the distribution of these 9 patterns were quite distinct in the blue and black clusters (see Figure 6, for instance).

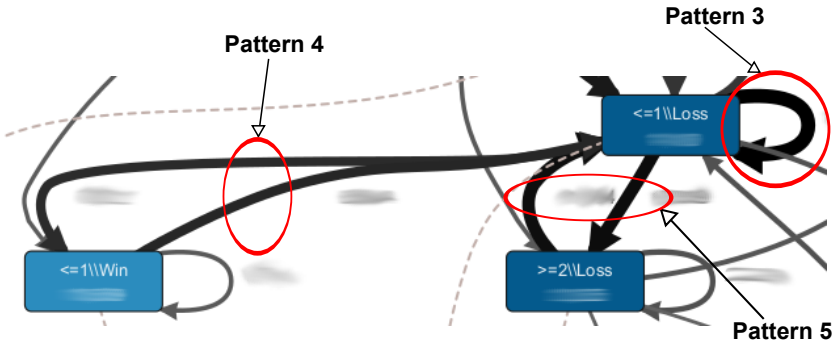

Figure 5. A visualisation of clawback behaviours identified from the process models of each cluster. Pattern 3 depicts a gambler who repeatedly lost bets and kept on betting to recoup the loss; Pattern 4 captures a continuous feedback loop winning and then losing a bet; Pattern 5 captures a clawback behaviour where the bet was double the amount of previous bet following a lost bet. The frequency of each event is blurred due to data confidentiality reasons.

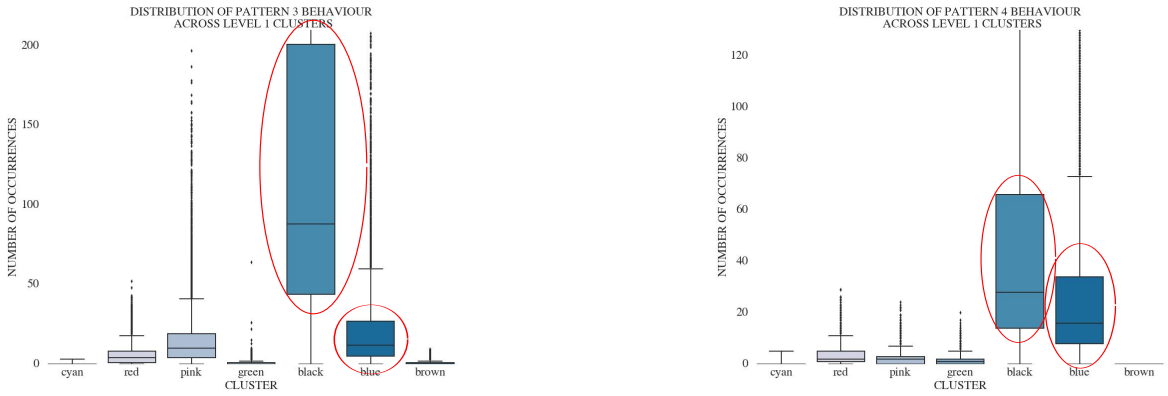

Figure 6. A comparison of the distribution of Pattern 3 and Pattern 4. Similar distribution for other patterns is also observed (not shown above).

The time elapsed between bets (bet interval) was another marker used to approximate betting intensity (another marker of problem gambling). We applied the *event interval analysis* plug-in [3] used with the ProM Tool to extract the bet interval for *each bet* placed by *every gambler* in *each cluster*. We aggregated the bet interval information *per gambler*, using the median value, to obtain the distribution of the bet interval values per cluster (as shown in Figure 7).

Next, Kruskal-Wallis and Dunn's tests were applied, the results of which asserted that the distributions of the bet intervals across the 7 clusters were indeed different at p-value 0.05. This was supported by the box-and-whisker graph (Figure 6 – right) which shows that the *blue* and *black* clusters had significantly lower bet interval values compared to other clusters. The exceptions being *cyan* and *brown* clusters. However, from Figure 4, we discounted these two clusters since their actual bet numbers were very low (e.g. the gamblers in the both clusters had mostly only bet once).

Interestingly, the pair-wise Dunn's tests for several behavioural patterns between the blue and black clusters resulted in a p-value higher than 0.05, which indicated insignificant behavioural differences across these two clusters (in the context of the overall 7 clusters). Thus, we decided to

re-cluster ("roll-up") just the instances belonging to the *blue* and *black* clusters for second-level clustering analysis. Our expectation was that the second-level clustering would further refine the clusters in from these two original groups such that additional behavioural differences could be extracted.

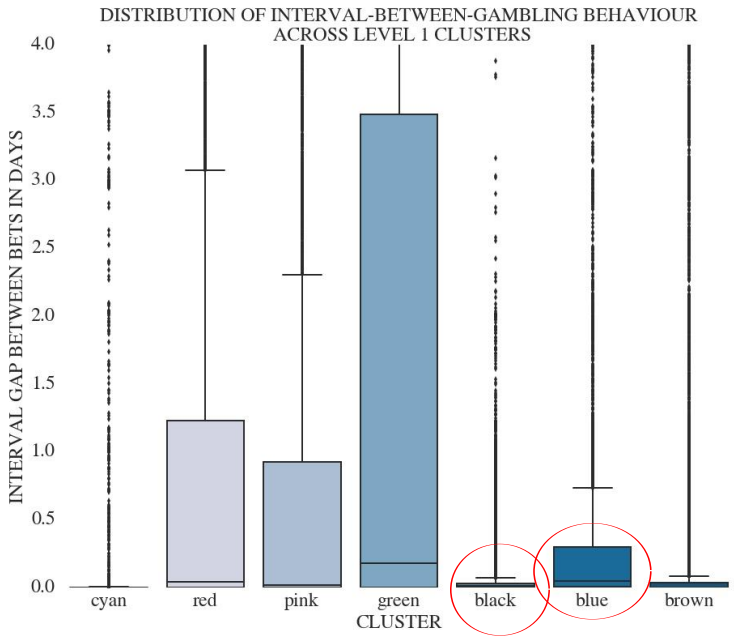

Figure 7. The distribution of the bet interval across the 7 clusters

Outcome. Through clawback-behaviour pattern and bet interval analysis, as well as Kruskal-Wallis and Dunn's statistical tests, we were able to group and rank the severity of gambling patters across the 7 clusters. We addressed RQ1 by having found mainly 4 clusters that are statistically different. Our ranking exercise also suggested that the *blue* and *black* clusters were more likely to contain problem gamblers as these groups exhibit significantly more intense bet intervals and gambling patterns, suggestive of gambling addiction. Results from this stage warranted a "drilling-down" analysis into these two clusters and by further re-clustering.

Challenges. A key challenge faced was process comparison. Existing literature on comparative analysis of processes [7], [10], [11], [15] seemed to go no further than visual analysis of process models derived from process mining tools (e.g. Disco tool), or through some forms of multi-perspective visualisation techniques (e.g. [11]). However, such approaches could be theoretically unsound. For instance, the *frequency perspective* visualisation in the Disco tool showed the total number of times a particular path was traversed by *all* of the process instances of a cluster. A process instance within a cluster that had an abnormally high number of such path traversals will skew the overall picture. This issue was solved by *extracting the distribution of potentially-distinct flow patterns identified through visual analysis*, and by having applied *confirmatory statistics testing* of the significance of pattern differences across multiple process groups.

From a practical perspective, we faced computational limitation (as severe as a machine re-boot due to overloading) during use of the ProM Tool plug-ins: event interval analysis [3] and BPMN Miner [31]. This was handled by *further division* of the event log which belonged to the largest cluster into three sub-logs. However, division of large XML-formatted event log was also a challenge as it has more than 105 million XML elements/lines. We used basic, powerful,

Linux-based text processing utility, such as `less`, `vim`, and `sed` to be able to cope with large XML files.

Lessons learned. The extraction by clustering was key to our work, as was the application of confirmatory statistical testing to the distributions of potentially-distinct process flows identified through visual comparison of process models. We were reasonably confident that the differences observed across multiple processes within, or across various, clusters were significant, and not by part of the same distribution. The statistical significance amongst clusters allowed use of other forms of graphical analysis tools, such as box-and-whisker, to identify how the clusters differ. The non-parametric assumptions of Kruskal-Wallis tests, paired with Dunn's tests, made them flexible and useful. Our experiences with process mining analysis suggest that we rarely see data that is normally distributed in practice. Therefore, such a combination of techniques could be recommended to be applied to many process comparison exercises frequently seen in cross-organisational process mining [10], [11], [15].

The Inductive Miner [32] implementation in the ProM Tool was shown to be scalable to even the largest event log in our clusters. However, the process model was less so: it produced models that exhibited "flower model" characteristics. We also used the BPMN Miner [31]; however, our computational resources were too limited for it to work properly. So far, the Disco tool was still the most scalable tool in terms of extracting process models even on an event log of over 4GB in size. The event interval analysis plug-in [3] was unable to handle the event log of the largest cluster; however, it worked though slowly (overnight for the largest cluster) after the log was divided.

Analysis of second-level clusters. The population within the *blue* and *black* clusters still exhibited a wide range of behaviours. It was unlikely that the majority of the population in those clusters were problem gamblers: there were 21,642 gamblers (23.67% of the total) in these two clusters. The data and diversity of these two clusters supported *further clustering* of the original population in the *blue* and *black* clusters into another 7 clusters. Each of the second-level cluster is also identified by a colour.[6]

This second level clustering was performed using k-means clustering using similar features as the initial clustering. For these second-level clusters, our focus was to establish statistical differences in terms of gambling clawback behaviour. To do so, we used the frequency distribution of those activities signifying loss-related clawback behaviours. `loss_<=1` signified the gambling pattern of following a losing bet with another that is of equal or smaller value, while `loss_1to2` encoded following a loss with another bet that was up to twice the size of the previous and `loss_>=2` indicated chasing a loss by more than doubling the size of the immediately preceding losing bet. The Kruskal-Wallis and Dunn's tests show an interesting insights: while the second level clusters do show mixed results for the statistical differences of the distribution for the `loss_<=1` and `loss_>=2` activities, there are no statistical similarities for the `loss_1to2` activity across all 7 clusters.

We also extracted the distribution of the Pattern 3 and Pattern 4 behaviours and the bet interval behaviour across these second-level clusters. As shown in Figure 8 and Figure 9, we can see that the `pink_2` and `brown_2` clusters contained disproportionately higher number of Pattern 3 and Pattern 4 compared to the other clusters. Furthermore, the median bet interval values for these two clusters were also noticeably lower than other clusters.

[6] For clarity, these 7 clusters obtained in the second-level clustering can be seen as more refined clustering of the original two clusters (i.e. the *blue* and *black* clusters from the intial clusters detailed in Section 4.4). In combination, we now have 12 clusters in total: 5 from the initial clusters, and 7 from the second-level clusters.

Outcome. We used the Kruskal-Wallis and Dunn's tests results with our second-level clustering 3D visualisation (Figure 10), and cross-referenced with box-and-whisker plots (Figure 8) to provide a more refined severity ranking of problem gamblers within the second-level clusters. These results indicated that the `pink_2` and `brown_2` clusters were likely to be the location where the problem gamblers were situated. We, thus, addressed RQ2: we managed to narrow down the most likely problem gamblers to two clusters (from the second-level clusters) with a combined population size of 6,389 accounts.

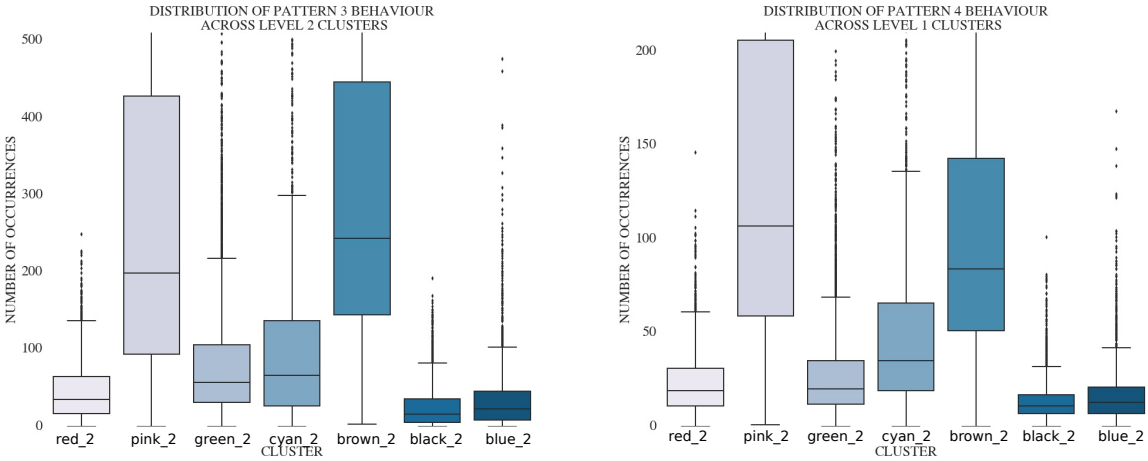

Figure 8. Distribution of Pattern 3 and Pattern 4 in the second-level cluster

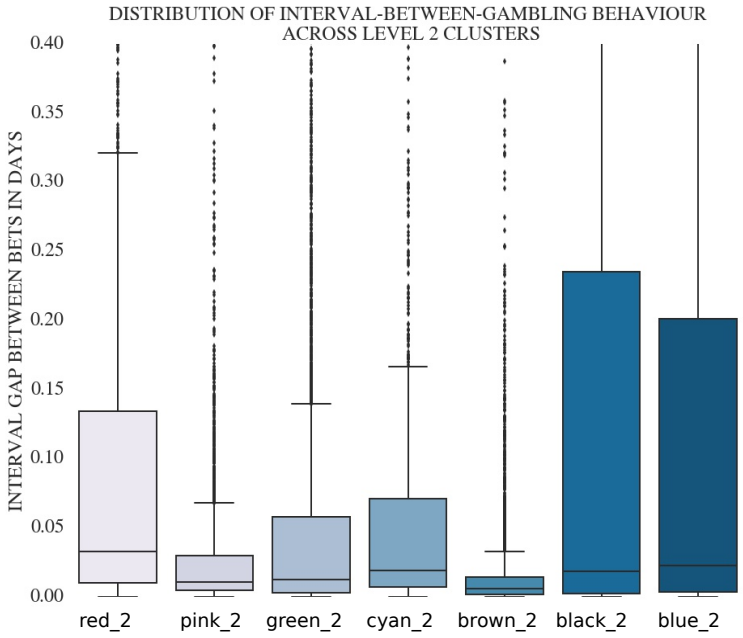

Figure 9. Distribution of bet interval across second-level clusters

Challenges and Lessons Learned. The main challenge was to narrow down the problem gambling patterns from total gambling population. We overcame this by using a recursive clustering approach as described above.

More importantly, we found that *recursive clustering approaches with confirmatory statistics were a suitable expression of the process cube concept* [19]. Recursive clustering based on

aggregated case-level attributes allowed us to "slice and dice" event logs in an unsupervised manner, while confirmatory statistics informed us as to which groups of processes that should be "rolled-up" (that is, those two groups whose p-value from the Dunn test indicates statistical similarities) and which to 'drill down' into. We found a *unique manifestation of the process cube concept* which hitherto has been dominated by slicing processes based on control-flow perspectives, e.g. Ekanayake et al. [33], or simple log filtering based on some known case-level attributes, e.g. Suriadi et al. [7].

4.5 Evaluation

This case study relied heavily on clustering to identify various classes of gamblers and to identify problem gambling (see Section 4.3). For these clusters to have been meaningful, we had to assert that (1) the population within each cluster shared some unique characteristics and/or behaviours that were distinct from other clusters, (2) the existence of those clusters were within reasonable expectation of the stakeholders and could be supported by other studies in the domain of problem gambling.

For the former, we applied classification analysis on the combined first- and second-level clusters (12 clusters in total: 5 from the initial clusters and 7 clusters from the second-level clusters). We used features that *had not previously been used to generate the k-means clustering*. These features included the median bet interval values (Section 4.4), the frequency of the 9 gambling patterns (Section 4.4), the frequency of clawback behaviour activities (Section 4.4), the total sum of money won, and the number of times a gambler won a bet. These features when used with the random forest algorithm (with 300 trees and 5-fold cross validation) generated a classification accuracy of 84.9% with a standard deviation of +/- 0.8%. These classification results, having used independent features from the clustering process, provide support that the clusters that were produced were relevant and meaningful.

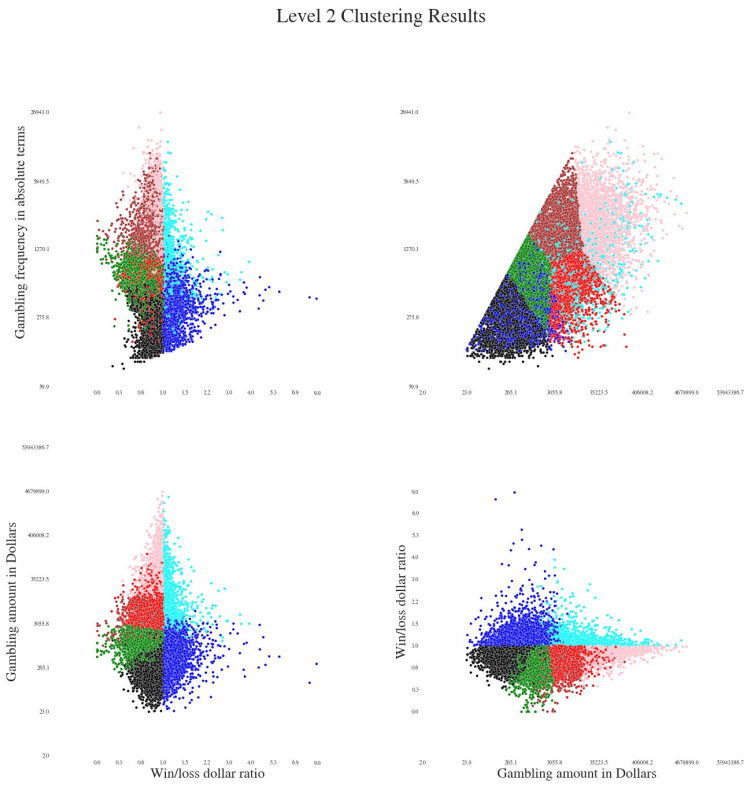

Figure 10. Visualisation of the Level 2 Clustering Result

We also presented our results to the stakeholder who found the results to be reasonable. Some insights will require further study, e.g. the dominance of Pattern 3 and Pattern 4 (which suggested that gamblers had bet equal or less than their previous amount after they had won or lost the previous bet). This contradicted currently-understood risk-taking behaviour of problem gamblers where the expectation was for them to increase the amount of bet money following a loss. While further analysis may be required, there was an explanation for this: the dominance of Pattern 3 and Pattern 4 was evident when we took all 12 clusters into account (which could mean that such behaviour is discriminatory for deciding if a gambler is a problem gambler or not). In the second level clustering, however, it was the frequency of high-risk clawback behaviour, i.e. the `loss_1to2` activity that was statistically different across all second-level clusters (thus, in line with the expected behaviour).

To further evaluate the results, we also compared our results with the results gathered from other studies about problem gambling, including the Problem Gambling Foundation of New Zealand[7], the Ministry of Health New Zealand [21], and the study by Orford et al. [34].

In particular, the validity of the clusters obtained from our study can be further supported by the distribution of gamblers who placed bets of $500 or more as this is one of the key markers of a problem gambler according to Problem Gambling Foundation of New Zealand. As shown in Figure 11, two second level clusters (identified as `pink_2` and `cyan_2`) contained disproportionately higher number of gamblers who placed bets over $500 dollars. This observation confirmed our earlier observation that the `pink_2` cluster likely contained problem gamblers; however, the fact that the `brown_2` cluster contained very low number of gamblers who placed bets over $500 dollars weakened the support for our earlier observation that this cluster may also contain problem gamblers. This phenomenon will require further investigation.

Figure 11. Distribution of gamblers who placed bets greater than $500

Another typical result obtained from other studies include the (dis)proportionate contribution, in terms of revenue and volume of bets (that is, the total number of bets placed), by problem gamblers. A study in New Zealand showed a consistent phenomenon (that has lasted for more than a decade) whereby the majority of gambling revenue comes from a relatively few people [21]. A similar study in Britain also showed a similar phenomenon whereby, in certain types of gambling, around 30% of both gambling volume and revenues came from problem gamblers [34].

These are interesting observations that correspond to our results, though to a varying degree. If we were being very conservative and assume that 1% of accounts are potentially problem gamblers, our data analysis shows that the top 1% of the accounts by frequency account for 25% of the total betting volume. Likewise, the top 20% of the accounts by frequency account for 88% of the total

[7] www.pgf.nz

betting volume. This phenomenon is supported by the cumulative distribution function shown in Figure 12 which shows that the top 1% of gamblers placed disproportionately higher number of bets compared to the other 99% of gamblers.

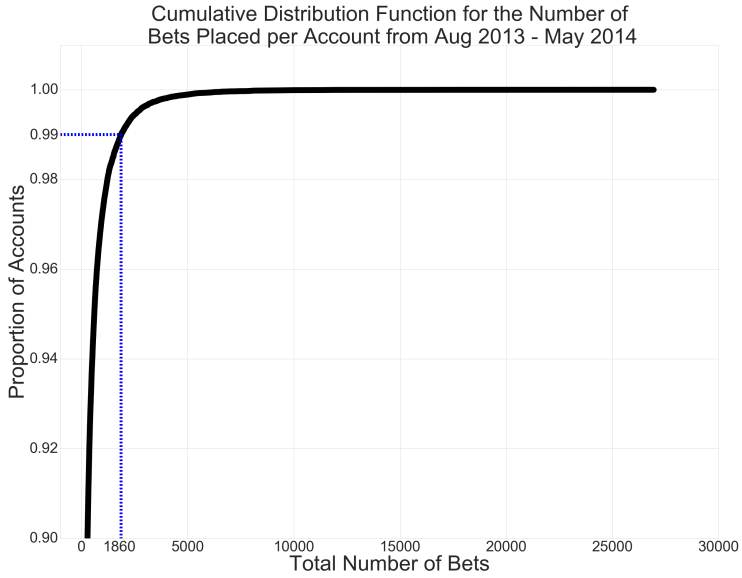

Figure 12. Cumulative distribution function graph showing total number of bets placed per gambler

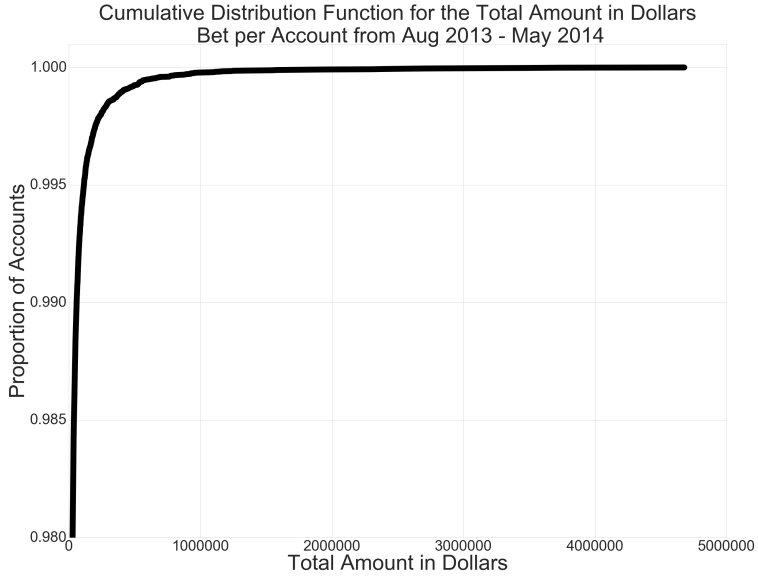

Figure 13. Cumulative distribution function graph showing total amount of money in dollars placed per gambler

When we viewed the gambling volume through total expenditure (i.e. total dollar amount from all bets, Figure 13), more interestingly, we saw that the top 1% (914) of gamblers (by betting expenditure) contributed to about 60% of the total betting expenditure. Meanwhile, the top 20% (18281) of gamblers (by betting expenditure) account for 96% of the total betting expenditure for FOBs. Our analysis also calculated the total balance (sum of all wins minus the sum of all bets placed) for each account and sorted the 91,405 accounts from the highest losing to most profitable. Strikingly, the analysis showed that the top 1% (739) of account holders with the highest negative balance account for 43% of revenue (having a median loss of $14,782), while the top 5% (3697) of account holders with the highest negative balance account for 71% of revenue (having a median loss of $3,900). The analysis shows that the contribution by those most likely to be problem

gamblers to the gambling volume, turnover and revenue is larger than the British based study [34] reported in 2013.

Of immediate interest to our analysis, is to link and understand the distribution of these top 1% and 5% gamblers across the 12 clusters that we have obtained from our analysis. As shown in Figure 14, the both 1% gamblers (by frequency/volume) and top 1–5% of gamblers (i.e. excluding the top 1%) reside mainly within the `pink_2` and the `brown_2` clusters.

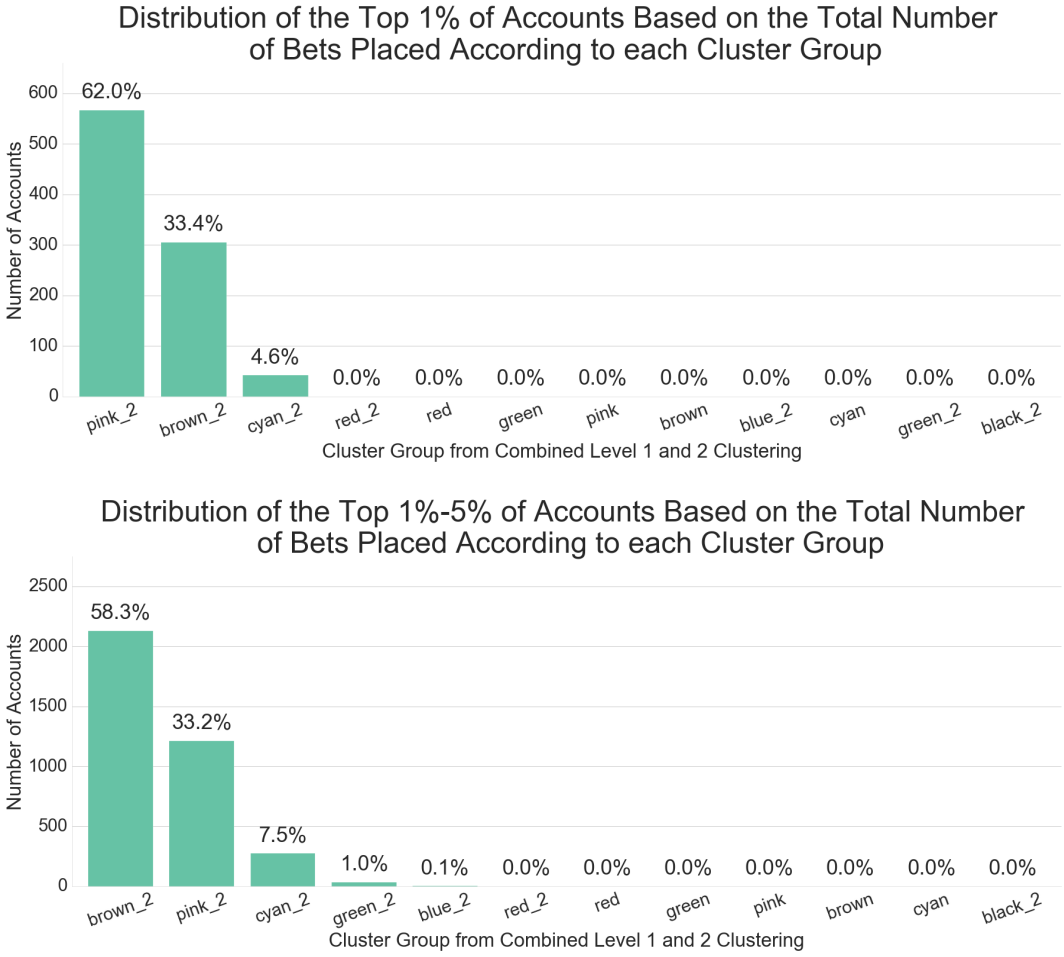

Figure 14. The distribution of the top 1% and 1–5% gamblers (by frequency) across the 12 clusters

Figure 15 (top) shows an interesting observation: while the top 1% of gamblers by total amount of dollars bet, reside mainly within the `pink_2` cluster, they do not reside within the `brown_2` cluster; instead, they were distributed within the `cyan_2` and `red_2` cluster. This may be explained by our earlier observation whereby the gamblers within the `brown_2` cluster hardly spent more than $500 per bet. As expected, Figure 16 also shows that the top 1% of gamblers (in terms of total loss) also resides mainly within the `pink_2` cluster, but none existed within the `brown_2` cluster.

However, when we look at the top 1–5% of gamblers in Figures 15 and 16, we can see that they are more or less well-represented within the `brown_2` cluster.

From these observations, we can see that when we extract the top 5% of gamblers as a whole, our results do correspond to the British study [34], thus, further supporting the validity of our approach. However, our analysis approach also allows us to reveal a rather distinct behaviour between the `pink_2` and `brown_2` clusters: both exhibit typical problem gambler behaviours, in terms of frequency of bet, clawback behaviour, and betting intensity; however, this is no longer the case in terms of the dollar amount spent and total lost.

Challenges and Lessons Learned. The main challenge was establishing that the clusters that were obtained were meaningful. A detractor of k-means clustering is that it *will* produce as many clusters

as parameterised. We performed classification analysis using features that were not used as input to the k-means clustering. This is important since previous work which used k-means clustering in a process mining case study [10] also applied classification analysis to provide meaning to each cluster. However, it was done using *the same features* as those used for clustering, which resulted in a very high accuracy rate but added nothing more to our understanding of each cluster than the threshold values of each variable used for generating the cluster groupings. By using features that are independent from those used in clustering, we have confidently established that each cluster has been "defined", and exhibited unique behaviours and characteristics.

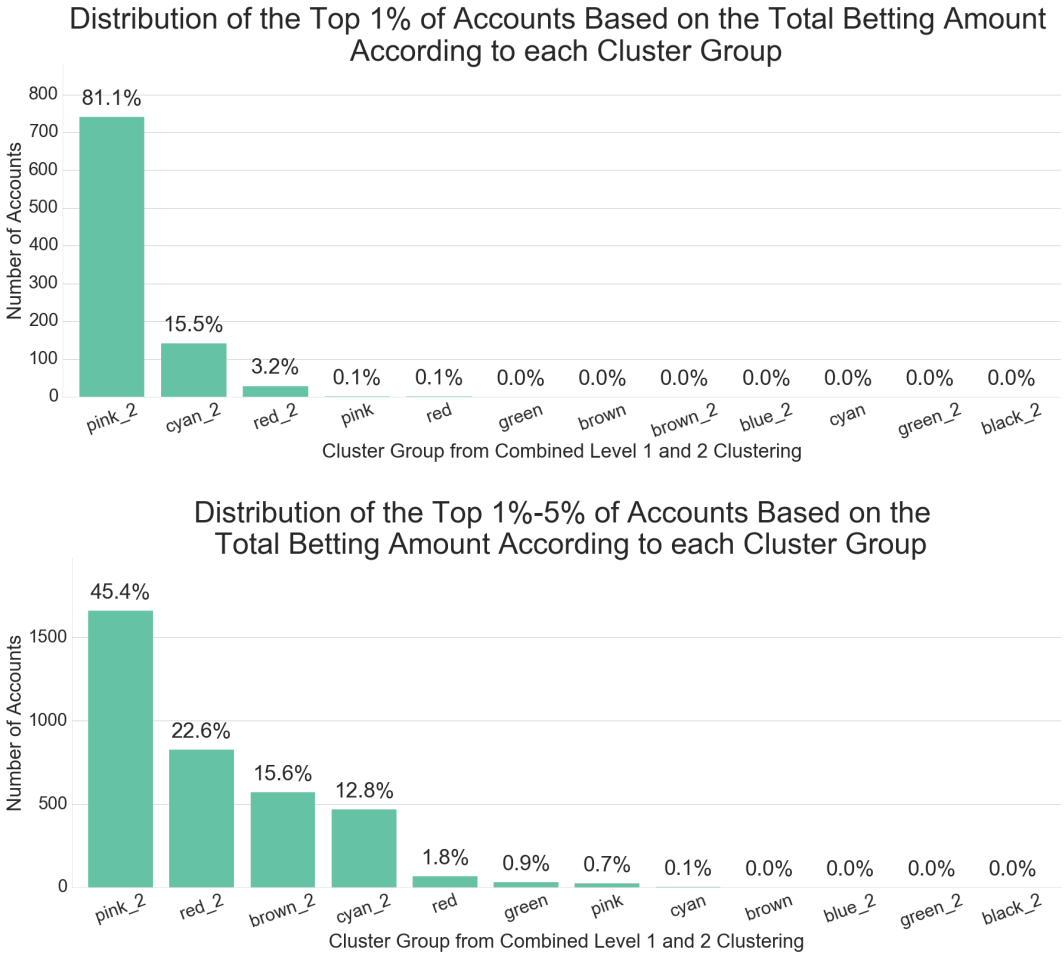

Figure 15. The distribution of the top 1% and 1–5% gamblers (by amount of dollars) across the 12 clusters

Other Observations and Key Recommendation. In addition to the lessons learned in Section 4, we summarize here a few observations and recommendations that may benefit other practitioners.

We recommend the integration of unsupervised clustering with confirmatory statistics when undertaking a process-cube approach to process mining projects. This allows an efficient and theoretically-sound way to slice, dice, roll-up, and drill-down groups of processes. This is especially true in process mining projects involving comparative analyses of various processes.

We question the need for XML-formatted data in process mining. Raw data often come in a table-like format with fields whose meaning can be understood through discussions with stakeholders. The self-defining property of XML is, thus, less necessary in practice. Processing XML files requires a significantly more powerful computational capacity, not only to parse every single XML element but also to handle the expanded file sizes: our original CSV-formatted data (of less than 1.5 GB) was unnecessarily "blown up" to over 5.8 GB after converting it to the XES format (with the majority of attributes that exist in the original data set *excluded*). We do not see any practical benefit in using XML-formatted log vs. CSV-formatted log. In fact, extracting and

manipulating CSV log is often faster and less heavyweight than dealing with XML documents as the former does not entail the parsing and semantic interpretation of each XML element.

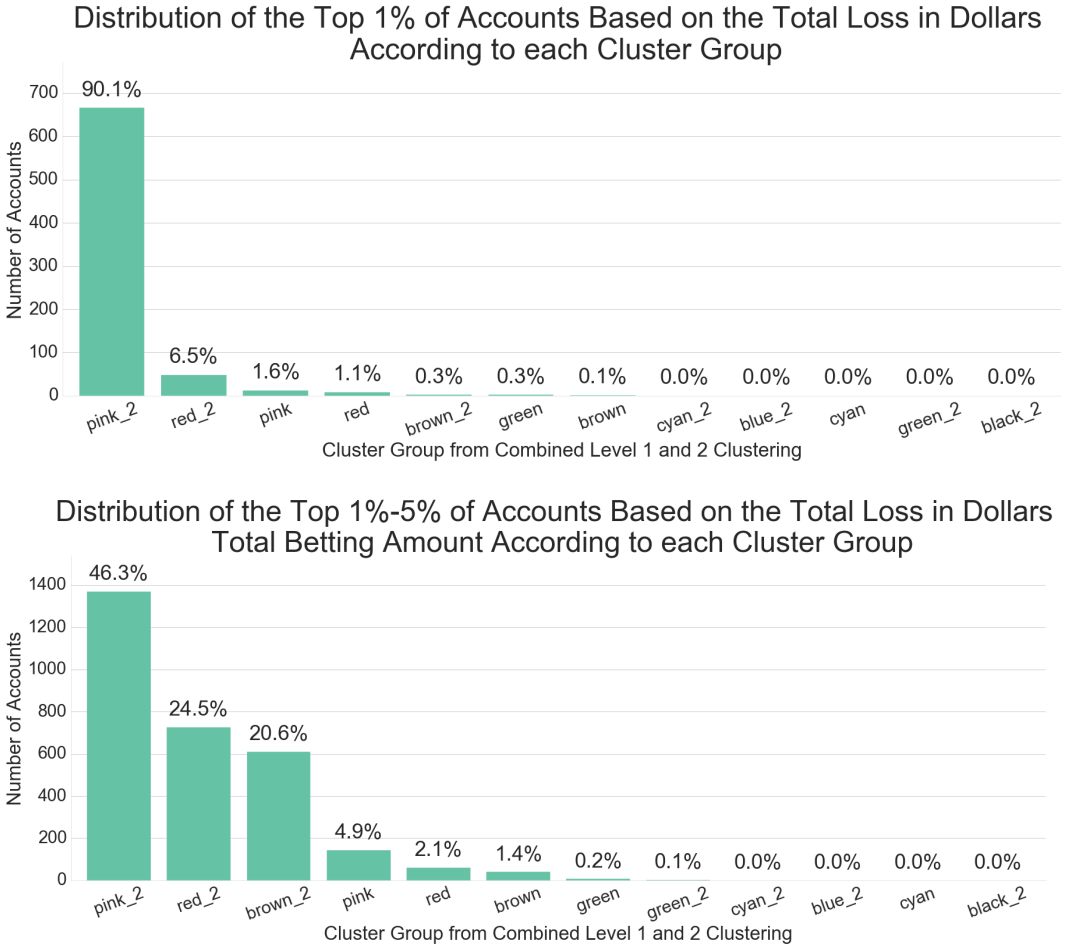

Figure 16. The distribution of the top 1% and 1–5% gamblers (by total loss) across the 12 clusters

5 Related Work

This section briefly discusses the approaches taken in other process mining case studies and compare them with the approach taken in this article.

Early process mining case studies [9], [35], [36], focused mainly on the application of standard process mining techniques, such as process discovery, performance analysis, and trace clustering. Later process mining case studies [15], [16], [17] applied a combination of process and data mining techniques. Our case study fits closer to the latter: we applied a balanced amalgamation of process and data mining techniques. However, we ventured further by also using a multi faceted approach involving process mining and well-established confirmatory statistics to assert, with high confidence, the significance of insights/observations gained from our case study. Nevertheless, the use of confirmatory statistics to assert the similarity/differences between various groups of "cases" as demonstrated in this article is quite similar to a very recent work by Bolt et al. [37].

This case study confirms, to a certain extent, some of the observations and experiences reported in other process mining studies. For instance, the use of clustering techniques to split original event logs into smaller pieces was applied in our case study. However, instead of using trace/sequence clustering as in [29], [36], we used k-means clustering [13] based on aggregated case-level features. We did this since in our case, we were more concerned with obtaining logs that are properly split based on gamblers' characteristics than with reducing trace variants within each cluster (a feature that is the focus of most trace clustering techniques, e.g. [30]). This is because two classes of

gamblers may behave very differently, but can still be grouped within the same cluster, had trace clustering been used. For instance, the (non-)existence of loop behaviours may not necessarily add visual complexity to a process model, but they are a key differentiator between recreational and problem gamblers.

We confirmed the scalability, and the ease-of-use of the Disco Tool (as reported in [7]) in terms of its ability to quickly generate abstracted and easy-to-understand process models from a relatively large event log. However, we encountered mixed experiences with respect to an earlier observation made in [35] about the feasibility of conducting process mining analysis using the ProM Tool: the quality and the robustness of the plug-ins in the ProM Tool proved to be highly variable with some being robust and scalable enough to handle large data sets (e.g. the Inductive Miner plug-in), while others tended to perform rather poorly (e.g. [3], [31]).

While our approach to process comparison was similar to other case studies, e.g. patient flows comparisons [11], [15], our case study went further by studying the distribution of observed differences, and asserted their statistical significance using well-established confirmatory statistics. This is in contrast to the simple visual observations about behavioural differences conducted in other case studies [7], [11], [15].

In relation to the concept of *proces cube*, our work has demonstrated how we can apply various data mining and confirmatory statistical techniques as a way to slice, dice, roll-up, and drill-down original data sets into appropriate chunks. This is in contrast to existing works (e.g. Ekanayake et al. [33] and Suriadi et al. [7]) which focus on either using simple filtering techniques based on some attribute values to split a data set into its appropriate "cubes", or proposing the necessary tool support to facilitate process cube-based analysis, such as the work by Vogelgesang et al. [38].

Most importantly, this article reported new lessons learned and recommendations that may be novel and helpful to other process mining practitioners.

6 Conclusion

We demonstrated how to apply a diverse and complementary set of techniques from the domains of process mining, data mining, and confirmatory statistics, as a unique expression of the process cube concept, to characterize and assert differences in gambling behaviours exhibited by more than 94 thousand gamblers in New Zealand.

Most importantly, this case study reported a number of challenges and lessons learned that are novel and would likely be beneficial to not only other process mining practitioners in general, but also to researchers within the domain of problem gambling. Particularly, our approach has demonstrated the viability of using a variety of data analysis techniques to perform evidence-based identification of problem gamblers. These techniques can indeed be automated, thus, opening up the possibility to perform continuous evidence-based analysis of gamblers in various community with the goal of facilitating early-intervention for problem gamblers.

While this work has focused on establishing a foundation for a data-driven approach to understanding different sub-groups of gamblers and problem gamblers, our future work will focus on developing automated classification methods for detecting the early onset of problem gambling behaviour in individuals for the purpose of intervention and prevention.

References

[1] C. Strong, "Humanizing Big Data: Marketing at the Meeting of Data, Social Science and Consumer Insight," Kogan Page, 1. edition, pp. 224, 2015.

[2] K.E. Arnold and M.D. Pistilli, "Course Signals at Purdue: Using Learning Analytics to Increase Student Success," in Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, ser. LAK'12, New York, NY, USA, ACM, pp. 267–270, 2012. [Online]. Available: https://doi.org/10.1145/2330601.2330666

[3] S. Suriadi, C. Ouyang, W. van der Aalst and A. ter Hofstede, "Event Interval Analysis: Why Do Processes Take Time?" Decision Support Systems, vol. 79, pp. 77–98, 2015. [Online]. Available: https://doi.org/10.1016/j.dss.2015.07.007

[4] J. Han and M. Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufmann, pp. 744, 2000.

[5] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006. [Online]. Available: https://doi.org/10.1016/c2009-0-22409-3

[6] W. van der Aalst, "Process Mining – Discovery, Conformance and Enhancement of Business Processes," in Discovery, Conformance and Enhancement of Business Processes, Springer, 2011. Available: https://doi.org/10.1007/978-3-642-19345-3

[7] S. Suriadi, M.T. Wynn, C. Ouyang, A.H.M. ter Hofstede and N. van Dijk, "Understanding Process Behaviours in a Large Insurance Company in Australia: A Case Study," in CAiSE, ser. LNCS, Springer, vol. 7908, pp. 449–464, 2013. [Online]. Available: https://doi.org/10.1007/978-3-642-38709-8_29

[8] R. Mans, M. Schonenberg, M. Song, W. van der Aalst and P. Bakker, "Application of Process Mining in Healthcare: A Case Study in a Dutch Hospital," in ICBES, ser. CCIS, vol. 25, Springer, pp. 425–438, 2009. [Online]. Available: https://doi.org/10.1007/978-3-540-92219-3_32

[9] A. Rozinat, I. de Jong, C. Gunther and W. van der Aalst, "Process Mining Applied to the Test Process of Wafer Scanners in ASML," in IEEE Trans. on System., Man, and Cybernetics, Part C, vol. 39, no. 4, pp. 474–479, July 2009. [Online]. Available: https://doi.org/10.1109/tsmcc.2009.2014169

[10] S. Suriadi, R.S. Mans, M.T. Wynn, A. Partington and J. Karnon, "Measuring Patient Flow Variations: A Cross-Organisational Process Mining Approach," in AP-BPM, ser. LNBIP, Springer, vol. 181, pp. 43–58, 2014. [Online]. Available: https://doi.org/10.1007/978-3-319-08222-6_4

[11] A. Pini, R. Brown and M. Wynn, "Process Visualization Techniques for Multi-Perspective Process Comparisons," in AP-BPM, ser. LNBIP, vol. 219, 2015. [Online]. Available: https://doi.org/10.1007/978-3-319-19509-4_14

[12] S. Suriadi, T. Susnjak, A. Ponder-Sutton, P. Watters and C. Schumacher, "Characterizing Problem Gamblers in New Zealand: A Novel Expression of Process Cubes," in Proceedings of the CAiSE'16 Forum. CEUR, vol. 1612, pp. 185–192, 2016. [Online]. Available: http://ceur-ws.org/Vol-1612/paper24.pdf

[13] J.A. Hartigan and M.A. Wong, "Algorithm as 136: A K-Means Clustering Algorithm," Journal of the Royal Statistical Society, Series C (Applied Statistics), vol. 28, no. 1, 1979. [Online]. Available: https://doi.org/10.2307/2346830

[14] S. Dragicevic, G. Tsogas and A. Kudic, "Analysis of Casino Online Gambling Data in Relation to Bebehavior Risk Markers for High-Risk Gambling and Player Protection," International Gambling Studies, vol. 11, 2011. [Online]. Available: https://doi.org/10.1080/14459795.2011.629204

[15] A. Partington, M. Wynn, S. Suriadi, C. Ouyang and J. Karnon, "Process Mining for Clinical Processes: A Comparative Analysis of Four Australian Hospitals," ACM Trans. on Management Information Systems, vol. 5, no. 4, 2015. [Online]. Available: https://doi.org/10.1145/2629446

[16] J. Nakatumba, "Resource-Aware Business Process Management: Analysis and Support," Ph.D. dissertation, Eindhoven University of Technology, 2013. [Online]. Available: http://doi.org/10.6100/IR760115

[17] H. Nguyen, M. Dumas, M. La Rosa, F. Maggi and S. Suriadi, "Mining Business Process Deviance: A Quest for Accuracy," in OTM 2014, ser. LNCS. Springer, vol. 8841, pp. 436–445, 2014. [Online]. Available: https://doi.org/10.1007/978-3-662-45563-0_25

[18] A.K. Jain, "Data Clustering: 50 Years Beyond K-Means," Pattern Recognition Letters, vol. 31, no. 8, pp. 651–666, 2010. [Online]. Available: https://doi.org/10.1016/j.patrec.2009.09.011

[19] W. van der Aalst, "Process Cubes: Slicing, Dicing, Rolling Up and Drilling Down Event Data for Process Mining," in Asia Pacific Conference on Business Process Management, ser. LNBIP, vol. 159, pp. 1–22, 2013. [Online]. Available: https://doi.org/10.1007/978-3-319-02922-1_1

[20] M. Abbott, M. Bellringer, N. Garrett and S. Mundy-McPherson, "New Zealand 2012 National Gambling Study: Gambling Harm and Problem Gambling," Report no. 2, Ministry of Health, Wellington, July 2014.

[21] Ministry of Health, "Strategy to Prevent and Minimise Gambling Harm 2016/17 to 2018/19," Ministry of Health, Wellington, May 2016. [Online]. Available: http://www.health.govt.nz/publication/strategy-prevent-and-minimise-gambling-harm-2016-17-2018-19

[22] P.A. Watters and M. Patel, "Competition, Inhibition, and Semantic Judgment Errors in Parkinson's Disease," Brain and Language, vol. 80, no. 3, pp. 328–339, 2002. [Online]. Available: https://doi.org/10.1006/brln.2001.2592

[23] K. Wallis, "Use of Ranks in One-Criterion Variance Analysis," Journal of the American Statistical Association, vol. 47, no. 260, pp. 583–621, 1952. [Online]. Available: https://doi.org/10.2307/2280779

[24] O.J. Dunn, "Multiple Comparisons Among Means," Journal of the American Statistical Association, vol. 56, no. 293, pp. 52–64, 1961. [Online]. Available: https://doi.org/10.2307/2282330

[25] M.L. van Eck, X. Lu, S. Leemans and W.M. van der Aalst, "PM^2: A Process Mining Project Methodologya," in Advanced Information Systems Engineering, ser. Lecture Notes in Computer Science, J. Zdravkovic, M. Kirikova and P. Johannesson, Eds., Springer International Publishing, vol. 9097, pp. 297–313, 2015. [Online]. Available: https://doi.org/10.1007/978-3-319-19069-3_19

[26] G. Meyer, T. Hayer and M. Griffiths, "Problem Gambling in Europe: Challenges, Prevention, and Interventions," Springer Science & Business Media, vol. 3, Springer, 2009. [Online]. Available: https://doi.org/10.1007/978-0-387-09486-1

[27] M.E. Devlin and D. Walton, "The Prevalence of Problem Gambling in New Zealand as Measured by the PGSI: Adjusting Prevalence Estimates Using Meta-Analysis," International Gambling Studies, vol. 12, no. 2, pp. 177–197, 2012. [Online]. Available: https://doi.org/10.1080/14459795.2011.653384

[28] H.R. Lesieur and S.B. Blume, "The South Oaks Gambling Screen (SOGS): A New Instrument for the Identification of Pathological Gamblers," The American Journal of Psychiatry, vol. 144, no. 9, pp. 1184–1188, 1987. [Online]. Available: https://doi.org/10.1176/ajp.144.9.1184

[29] J. De Weerdt, S. vanden Brouckev, J. Vanthienen and B. Baesens, "Leveraging Process Discovery with Trace Clustering and Text Mining for Intelligent Analysis of Incident Management Processes," in IEEE CEC, 2012. [Online]. Available: https://doi.org/10.2139/ssrn.2165170

[30] J. D. Weerdt, S. vanden Broucke, J. Vanthienen and B. Baesens, "Active Trace Clustering for Improved Process Discovery," Knowledge and Data Engineering, vol. 25, no. 12, pp. 2708–2720, 2013. [Online]. Available: https://doi.org/10.1109/tkde.2013.64

[31] R. Conforti, M. Dumas, L. Garcia-Banuelos and M. La Rosa, "Beyond Tasks and Gateways: Discovering BPMN Models With Subprocesses, Boundary Events and Activity Markers," in BPM, ser. LNCS, Springer, vol. 8659, pp. 101–117, 2014. [Online]. Available: https://doi.org/10.1007/978-3-319-10172-9_7

[32] S. Leemans, D. Fahland and W. van der Aalst, "Discovering Block-Structured Process Models From Incomplete Event Logs," in Petri Nets, ser. LNCS, Springer, vol. 8489, pp. 91–110, 2014. [Online]. Available: https://doi.org/10.1007/978-3-319-07734-5_6

[33] C.C. Ekanayake, M. Dumas, L. Garcia-Banuelos and M.L. Rosa, "Slice, Mine and Dice: Complexity-Aware Automated Discovery of Business Process Models," in BPM, ser. LNCS, vol. 8094, pp. 49–64, 2013. [Online]. Available: https://doi.org/10.1007/978-3-642-40176-3_6

[34] J. Orford, H. Wardle and M. Griffiths, "What Proportion of Gambling is Problem Gambling? Estimates From the 2010 British Gambling Prevalence Survey," International Gambling Studies, vol. 13, no. 1, pp. 4–18, 2013. [Online]. Available: https://doi.org/10.1080/14459795.2012.689001

[35] W. van der Aalst, H. Reijers, A.M.M. Weijters, B.F. van Dongen, A. Medeiros, M. Song and H. Verbeek, "Business Process Mining: An Industrial Application," Information Systems, vol. 32, no. 5, pp. 713–732, 2007. [Online]. Available: https://doi.org/10.1016/j.is.2006.05.003

[36] A. Rebuge and D. Ferreira, "Business Process Analysis in Healthcare Environments: A Methodology Based on Process Mining," Inf. Syst., vol. 37, no. 2, pp. 99–116, Apr. 2012. [Online]. Available: https://doi.org/10.1016/j.is.2011.01.003

[37] A. Bolt, M. de Leoni and W.M.P. van der Aalst, "A Visual Approach to Spot Statistical-Significant Differences in Event Logs Based on Cofigurable Metrics," in CAiSE 2016, ser. LNCS, Springer, pp. 151–166, 2016. [Online]. Available: https://doi.org/10.1007/978-3-319-39696-5_10

[38] T. Vogelgesang, G. Kaes, S. Rinderle-Ma and H.J. Appelrath, "Multidimensional Process Mining: Quesquest, Requirements, and Limitations," CAiSE Forum, vol. 1612, CEUR, pp. 169–176, 2016.