RC24268 (W0705-125) May 24, 2007
Computer Science

IBM Research Report

Analytics for Audit and Business Controls in Corporate Travel and Entertainment

Vijay Iyengar, Ioana Boier
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598 USA

Karen Kelley, Raymond Curatolo
IBM Global Technology Services
150 Kettletown Road
Southbury, CT 06488 USA

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Analytics for Audit and Business Controls in Corporate Travel & Entertainment

Vijay Iyengar, Ioana Boier
IBM Thomas J. Watson Research Center
19 Skyline Drive, Hawthorne, NY 10532, USA
(vsi, ioana)@us.ibm.com

Karen Kelley, Raymond Curatolo
IBM Global Technology Services
150 Kettletown Road, Southbury, CT 06488, USA
(langan, rac)@us.ibm.com

Abstract

Travel and Entertainment (T&E) expenses are under increasing scrutiny as one of the largest controllable indirect expenses in a firm. This involves internal audits and analysis by business controls personnel to identify fraud and misuse and to take appropriate corrective actions. We have developed a set of statistical models to identify suspicious behavior for further investigation. Our Behavioral Shift Models (BSM) leverage domain knowledge in the form of simple, generic templates that represent classes of fraud and abuse. The emphasis is on robustly detecting repeated, out-of-the-norm behaviors as opposed to single instance occurrences. In this paper, we describe the application of these models and characterize their detection capabilities empirically. We also present validated results and insights generated by our approach when applied to production data from multiple firms for several T&E scenarios.

Keywords: Audits, business controls, fraud and abuse.

1 Introduction

Travel and Entertainment (T&E) expenses are considered one of the largest controllable indirect expenses in a firm. The recent emphasis on business integrity and compliance in conjunction with a tight business environment and constant attention to the bottom line have led to a renewed focus on the implementation of effective management and controls for T&E. This entails multiple dimensions, including improvement of internal controls and related business processes, expense monitoring and timely auditing, and improved vendor procurement and management.

The problem we address in this paper is that of analyzing transaction data logged through a T&E system for the purpose of effective audit and business controls. The data consists of expense and approval records, but completely lacks historical information on the outcome of any subsequent actions. Unfortunately, it is rather common practice in the audit & business controls domain to process candidates deemed worthy of attention without documenting the results of the investigation within the original T&E environment. This poses an interesting set of challenges for the analysis of the data and considerably reduces the number of options among the techniques developed to date (see Section 2). The outcome of the analysis may be directed towards audit or business controls and may be relevant at different granularity levels (in what follows, the elements of each such level, e.g., individuals, organizational subunits such as accounting centers and divisions are referred to as *entities*).

The audit and business control functions serve different purposes. *Audit* refers to checks that are performed to ascertain the validity and reliability of the T&E information. For example, an audit could examine a specific subset of travel expenses claimed by an employee. Such an audit could uncover fraud, error in the claims process (e.g., incorrect expense type used to categorize a claim), or misuse (e.g., bypassing the expense approval process by inappropriately splitting transactions into ones of smaller, less conspicuous amounts). *Business controls* encompass activities that examine and analyze data from expense claims and expense approval processes for excessive violations of relevant corporate policies and guidelines. For example, excessive approval of violations of business class travel policy by an organization within a firm could trigger an investigation and potential action to improve compliance with business travel policy. Currently, the internal audit and business controller roles are being emphasized in many organizations due to an increased focus on corporate business integrity. A decision to audit is not simply viewed in terms of balancing the cost of the audit against the costs due to the abuse. An audit is frequently pursued if there is adequate evidence and sometimes the investigation exposes the "tip-of-the-iceberg" where the same entities are involved in violations in other domains beyond T&E.

Detection of candidates for auditing and/or business control actions is a critical and challenging task. A typical approach relies on the deep knowledge of domain experts (auditors and business controls personnel) to aim at specific scenarios that reveal potential mechanisms for fraud, errors and misuse of policy. Clearly, such an approach makes for a highly non-uniform process of identifying candidates depending on the method and expertise of the individual domain expert. In the context of a T&E software system, this approach entails capturing and updating all possible scenarios describing mechanisms for fraud and misuse as they are discovered. At the opposite end of the spectrum, an ambitious approach is to try to detect entities for further investigation without explicit domain knowledge about

fraud and abuse mechanisms. This approach is relatively new and has been less explored in the literature or in the commercial space.

In our work we have adopted a middle path, by developing a set of Behavioral Shift Models (BSM), i.e., statistical models that identify suspicious behavior while relying only partially on domain knowledge. The latter is used solely to define a set of simple, generic templates that represent classes of fraud and abuse that may be of interest. The parameters of our statistical models for any given template are learned from the data. We contend that our models provide a balance between the amount of detailed domain knowledge required and the robustness of the insights generated (e.g., few false positives).

In this paper, we present two models that cover two classes of scenarios. The first model is applicable to cases that involve positive real-valued variables (e.g., categorized expense amounts, time durations such as payment delays). As an example, consider tip expenses of individual employees incurred during business travel. These expenses tend to be paid in cash and may not require receipts for reimbursement. A typical analysis scenario would seek to detect those employees with significantly high tip claims. Section 3 describes our first model, its empirical characterization, and results from scenarios in this class (referred to as *Expense Amount Scenarios*) using production T&E data from multiple firms. The second model is applicable to scenarios involving count data (referred to as *Event Count Scenarios*) for events like business rule exceptions. For example, organizations typically have well-defined business rules regarding the class of air travel allowed for business trips. They also have a business process for approving exceptions to this rule. From a business controls perspective, it is important to monitor and assess whether an organizational unit is lax in its business controls by excessively approving this type of exception. Section 4 describes the details of our second model and the corresponding results. The remainder of the paper is dedicated to the background for this work (Section 2), discussion (Section 5) and our conclusions (Section 6).

2 Background

Our work touches upon a multitude of aspects, some generic and some domain-specific. Broad topics like outlier identification, statistical inference, and hypothesis testing are examples in the former category. Analysis of transaction data in T&E systems for purposes such as reporting, monitoring, and compliance are representative of the latter. In this section we attempt to narrow down this rich field starting from the problem we are trying to solve and its desired (if not required) outcomes as described in the previous section.

Outlier detection pertains to the detection of anomalous observations (outliers) in data sets. The abnormality is typically defined with respect to other samples within the same data set. A broad spectrum of techniques has been developed for different applications on a varied theoretical backdrop that includes Statistics, Machine Learning, Neural Networks, etc. A comprehensive overview of outlier detection is provided by Hodge and Austin (2004). Using the taxonomy proposed in that work as our reference, we note that methods that fall in the Type 1 (i.e., unsupervised clustering) or Type 2 (i.e., supervised classification) categories do not offer suitable solutions to our problem: an a priori proximity metric to be used for clustering would be difficult to conjecture and labeled data is unavailable. The closest to our approach are the Type 3 methods (i.e., semi-supervised detection) that model normality and use it to pinpoint abnormal cases.

There has been extensive research done on outlier detection methods to identify observations (or points) in n-dimensional space that deviate from other observations. For example, statistical and data mining methods for this task are compared by Williams et al (2002). Such methods do not address the problem of analyzing repeat behavior that is of interest in our domain. We are interested in identifying entities with outlying behavior and the data contains varying numbers of behavioral observations for each entity.

A comprehensive review of statistical fraud detection is provided by Bolton and Hand (2002). Their exposition on unsupervised methods is clearly relevant to the problem addressed in this paper. The notion of using a statistical profile of the normal behavior has been used in earlier works. The computer intrusion detection work by Denning (1987) is an example that uses this approach. One of the statistical models used by Denning for representing the normal profile consists of the summarization using the mean and standard deviation. Any single observation is tested and scored for deviation from this normal characterization. The more recent work on peer group analysis by Bolton and Hand (2001) incorporates a key refinement by using local models in the form of peer groups that define normal behavior for any entity being analyzed for deviant behavior. However, in both examples the normal profile is used to score the deviation of a single observation. As mentioned earlier, our problem requires analyzing multiple observations for each entity to determine entities with repeated and significant outlying behavior.

Formulations developed in the area of Scan Statistics (Kulldorff 1997, Glaz et al 2001, Huang et al 2007) are well-suited to the problem at hand. The approach is to use hypothesis testing (Lehmann 1986) using the Likelihood Ratio Test (LRT) to scan for clusters of abnormality that stand out within the entire space of data considered. The expanding body of work in this area includes development of models suited for various underlying distributions and applications to various domains. Our work could be viewed as an adaptation of this approach to the problem of identifying suspicious behavior for audit and business controls purposes.

Our resulting solution is novel in the T&E domain for at least two essential reasons: (a) it uses a robust scoring mechanism that considers the magnitude of abnormality without requiring specification of boundaries between normal and abnormal; (b) it emphasizes the repetition of abnormal behavior as an important metric in

characterizing outliers. In T&E both aspects are crucial. For example, expenditure limits are set through policies and exceptions to these trigger alarms. However, there is considerable room for fraud under these limits which may not always be caught through additional thresholds (see Section 3). Capturing repetitiveness is also of essence: it corresponds to the amount of evidence to justify an audit and the cost of the corresponding follow-up investigation and it may reveal integrity gaps that may point to other problems. The importance of gathering sufficient audit evidence has been highlighted in other financial areas by Beasley et al (2001).

The importance of financial controls and policy adherence in the T&E domain is emphasized by the National Business Travel Association (NBTA). NBTA is a leading forum in the business travel domain and a source for information about the domain, including commercial service and product providers in this space. Commercial packages typically provide reporting functions that summarize and sort the data based on domain knowledge of the important metrics in this space. However, to the best of our knowledge our model and method represents the first analysis in this domain that evaluates each entity based on the magnitude of deviation from normal and the repetitiveness of the behavior after appropriate normalization.

3 Expense Amount Scenarios

In this section we describe our method as it applies to scenarios that involve positive, real-valued variables. Consider the tip expenses scenario introduced in Section 1. A typical analysis scenario would seek to detect those employees with significantly high tip claims. There are two important aspects we consider: (a) repetitiveness – we are interested in candidate entities (individual employees in this case) that exhibit a profile / pattern of repeated excessive tipping (in contrast with methods that focus on finding isolated outlier tip amounts); and (b) significance: to properly quantify excessiveness we incorporate domain knowledge that helps us normalize the range of our variables. For tips, the amount must be normalized by the location where the tip expense was incurred. The template for this scenario would specify the expenses to be analyzed (tip expenses), the covariate structure for normalization (location where expense was incurred), and the target entities (employees).

3.1 The Model

To analyse expense amount scenarios, we apply hypothesis testing using the LRT formulation. We compare the distribution of values for a given entity ξ with that of the baseline B of values computed from all the entities:

(H0: null hypothesis) $\mathbf{E}[\xi] = \mathbf{E}[B]$

(H1: alternate hypothesis) $\mathbf{E}[\xi] > \mathbf{E}[B]$

where $\mathbf{E}[]$ denotes the expectation (mean) operator. As previously explained, the values considered in the LRT must be normalized by taking into consideration all the relevant factors (determined through domain knowledge).

Through empirical analysis of many expense amount scenarios across datasets from multiple firms, we found that the exponential model developed in a recent work by Huang, Kulldorff and Gregario (2007) on a spatial scan statistic for survival data provides an excellent characterization for the majority of T&E baselines after proper normalization. In addition, we observed that in practice it has good power for a broader class of distributions (e.g., Gamma) which is in line with the observations of Huang et al (2007).

For each scenario we specify the entity space (e.g., employees, department, divisions, and business units) being targeted by the analysis. We also specify the covariates according to domain knowledge. Referring back to our tip example, the entities are individual employees and the location where the tip expense was incurred is the only covariate. The target variable is the amount of tip expense claimed. Categorical covariates are handled directly by determining a normalization factor F for each combination of covariate values. Typically, this factor F is the mean (or max) value for that combination of covariate values over all the entities. Normalization of an individual value simply becomes the ratio of the raw value and F. For example, each tip expense can be normalized by dividing it by the mean tip value for the location where the expense was incurred. Consider an entity ξ with M normalized values which sum up to S. Let the total number of normalized values over all the entities be N and their sum be T. The test statistic $\mathbf{Y}(\xi)$ for the exponential model is given by:

$$\mathbf{Y}(\xi) = M \times \mathbf{log}\left(\frac{M}{S}\right) + (N-M) \times \mathbf{log}\left(\frac{N-M}{T-S}\right) - N \times \mathbf{log}\left(\frac{N}{T}\right).$$

Following the methodology used with most scan statistics (Kulldorff 1997, Huang et al 2007), a p-value is computed by performing a number Z of Monte Carlo experiments. In each experiment, the entity values are determined by sampling from the baseline and the maximum test statistic achieved by any entity is computed. We use sampling with replacement instead of the permutation approach used by Huang et al (2007) since in our domain a small number of extreme values are deleted from the baseline. The p-value for the entity ξ is computed using the formula $(L+1)/Z$, where L is the number of Monte Carlo experiments with a maximum test statistic exceeding $\mathbf{Y}(\xi)$. Entities with p-values that reject the null hypothesis at the prescribed α level are considered candidates for further investigation and ranked in decreasing order of their test statistic values Y.

3.2 Empirical Characterization

We will analyse the power of our model empirically for a range of Gamma distributions that are based on our characterization of production T&E data. Our experimental procedure for this empirical characterization is sketched in Figure 1. Specifically, we report on one set of experiments each of which simulates 1000 entities. Each entity has varying number of data items representing the expense claims submitted. The number of data items for an entity is modeled by a Gamma

distribution (Γ_1) with shape parameter 1.0 and scale parameter 16 (i.e., mean = 16). The data values are characterized by various two-parameter Gamma distributions (Γ_2). In each case, without loss of generality, we choose the value distributions to have a mean 1.0 (the test statistic is invariant under multiplicative scaling). The following values were considered for the Gamma shape parameter: {0.25, 0.5, 1.0, 2.0, 3.0, 4.0, 5.0, and 6.0}. In each experiment, a single target entity is chosen to have its values increased by a percentage δ that is varied in the range [10, 400]. Note that the test statistic **Y** is not sensitive to the distribution of the increase across individual values for the target entity since it is based only on the sum of all its values.

This empirical analysis provides a characterization for the ability of our method to detect a target entity with inflated values at a given α level for p-values. Similarly, we also determine the number of non-target entities that are detected at the given α level and we use the two resulting characterizations to quantify the performance of our model in terms of false negatives and false positives in this idealized experimental setup.

Input: N_ξ number of entities in the population

N_b number of baselines

N_{exper} number of simulation experiments for each baseline

Γ_1{shape, scale}for the distribution of the number of expense items across entities

Γ_2{shape, scale}for the distribution of expense amounts across entities the population (mean = 1)

Output: statistics regarding successful identification of engineered increases

Algorithm:

Generate N_ξ random numbers according to Γ_1 distribution; these represent the number of expense items for each entity

for each baseline 1.. N_b **do**

  generate data for this baseline according to Γ_2 distribution; this data represents the amount of each expense for every entity

  for each δ amount of percentage increase **do**

    for each experiment 1..N_{exper} **do**

      select an entity ξ^* to be engineered

      apply a δ% increase to each expense amount in ξ^*

      run BSM model and compute p-values and detection statistics

      record entities with p-values below chosen α level & detection statistics

    end for

  end for

end for

Figure 1. Experimental procedure to evaluate the detection capabilities of BSM.

For illustration, we use $\alpha = 0.01$. First, we consider the experiments done with the Gamma shape parameter of 1.0 for the value distribution. We use a classification model (Duda, Hart and Stork 2001) to discriminate the class of experimental cases in which the target entity was detected by our method from those in which it was not. Figure 2 shows both these classes with the x-axis representing the number of data elements (ϕ) in the target entity and the y-axis representing the total excess added to the target's values (λ). The points marked with a "+" are instances of the class where the target entity was detected at the given α level. A linear classifier was generated using a training set composed of 25% of the data using an SVM formulation (Christianini and Shawe-Taylor 2000). The accuracy on the test set (remaining 75% of the data) was 95% in this case indicating that this linear discriminator is a reasonable characterization of the target detection achieved by our method. The equation for the linear discriminant is $\lambda = 0.513 \times \varphi + 13.26$ and it characterizes the amount of excess that is detectable by this model. For example, a target entity with 40 data entries is detected only if, on average, its values are increased by 84% of the mean value. In the limit, as the number of values in the target entity increases, the excess has to be 51% of the mean for it to be detected. The relatively high value for the excess needed in this case is due in part to the skewed nature of the exponential distribution (i.e., Gamma shape parameter of 1.0). The relatively long tail for values even under "normal" circumstances results in the need to have sizeable excess before it is deemed significant. Our experience with production T&E data summarized in the next subsection shows that even with this conservative performance our model detects many interesting candidates for further investigation.

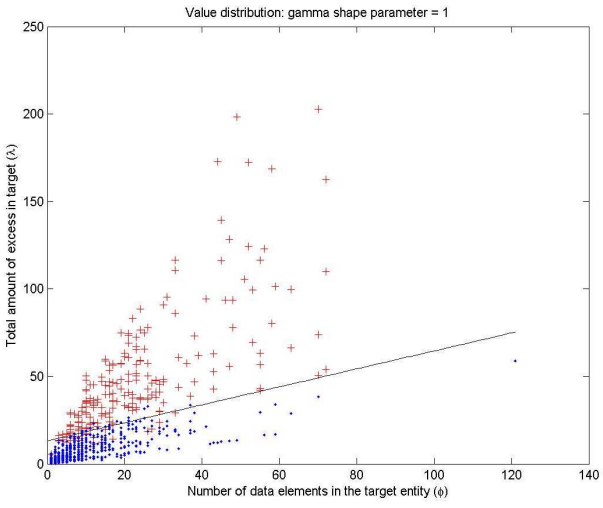

Figure 2. Detection of target entity for Gamma shape parameter of 1.0

Repeating the analysis by choosing other Gamma distributions for the baseline (i.e., choosing the Gamma shape parameter), we can characterize the detection ability of our model using the linear classifier learned for each Gamma distribution. The linear discriminants for the various Gamma distributions are shown in Figures 3 and 4. The SVM accuracy is higher than 93% in all the cases confirming the validity of these characterizations. The skewed value distributions when the shape parameter

is ≤ 1.0 result in larger excesses being required for detection (Figure 3). The tighter distributions that result as the Gamma shape parameter is increased lead to detection with much smaller excess (Figure 4). For example, when the shape parameter is 6.0, a target entity with 40 data entries is detected if, on average, its values are increased by as little as 25%.

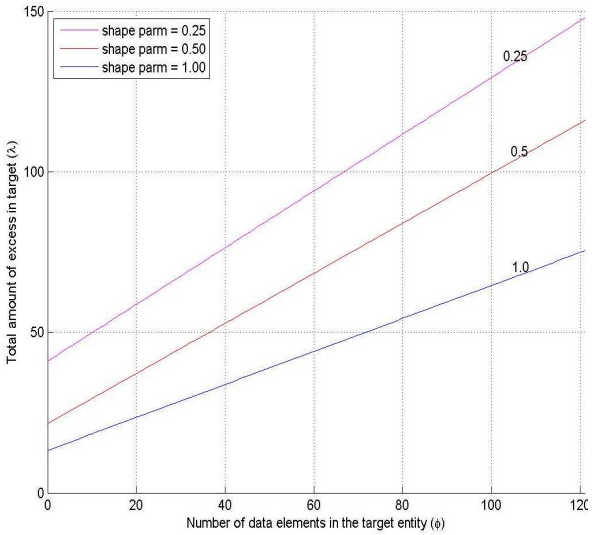

Figure 3. Linear classifier results for Gamma shape parameter values 0.25, 0.5, and 1.0.

The previous characterization indicates the ability of the model to detect the target for a wide range of Gamma distributions for the baseline values though the magnitude of excess needed is quite high when the baseline distribution itself has a long tail.

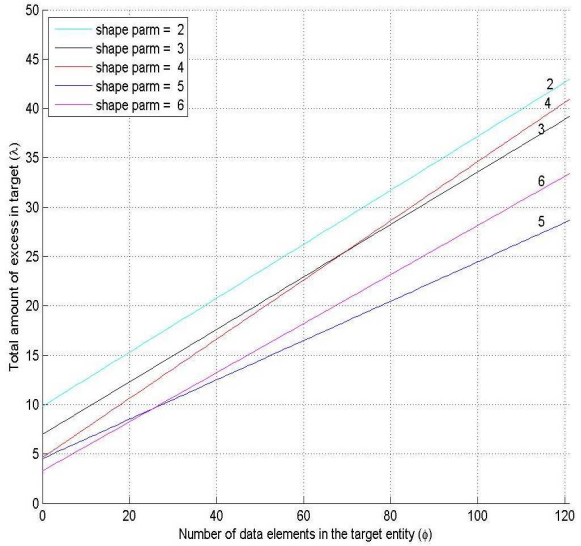

Figure 4. Linear classifier results for Gamma shape parameter values 2.0, 3.0, 4.0, 5.0, and 6.0.

Next, we consider the detection of non-target entities which gives us an indication of the false positive rate. At $\alpha = 0.01$, non-target entities were not detected in any of the experiments. The tradeoff between sensitivity and false positives can be illustrated if compare these results with those for $\alpha = 0.05$. The number of experiments (expressed as a percentage of the total number) in which non-target entities were detected is given for both α levels (0.01 and 0.05) in Table 1. At $\alpha = 0.05$ non-target entities are detected when the baseline Gamma distribution has shape parameter values of 2.0 and 3.0. In each instance, when a non-target entity was detected only one such entity was detected. The detection sensitivities for the two α levels can be compared by considering the corresponding equations for the linear discriminants. For the value Gamma shape parameter value of 1.0 considered earlier, the equation for the linear discriminant is $\lambda = 0.342 \times \varphi + 13.00$ at $\alpha = 0.05$ indicating the increase in sensitivity compared to the equation $\lambda = 0.513 \times \varphi + 13.26$ we had for $\alpha = 0.01$. The user can control this tradeoff between detection sensitivity and false positive rate by the choice of the α level.

The empirical characterization with the idealized Gamma distribution for the baselines indicates the magnitude of excess that is detectable by our method while keeping the false alarms rate in check. We show the utility of our method in the next subsection with results obtained by applying our method to production T&E data.

Gamma Shape Parameter	α -level =0.01		α -level =0.05	
	Target detected	Target not detected	Target detected	Target not detected
0.25, 0.5, 1.0	0	0	0	0
2.0	0	0	6.4%	6.7%
3.0	0	0	9.0%	10.3%
4.0, 5.0, 6.0	0	0	0	0

Table 1. Percentage of instances with non-target entity detection

3.3 Application to Production T&E Data

We applied our model to production T&E data from multiple firms in an enterprise expense reporting environment (GERS) and we reviewed the results of various scenarios with audit and business control professionals. The reviews were of a qualitative, not quantitative nature, i.e., they did not provide a quantitative assessment of false positive and false negative rates, but they did confirm the usefulness of our model. The top significant candidates detected by our technique in each scenario were found by the auditors to be interesting targets for further investigation. Interestingly, most of the candidates identified were not previously known to the domain experts as suspicious cases. In addition, we also did a few controlled experiments in which known cases were added to the data to confirm that BSM correctly detected them as

candidates for further investigation. In this section we present some of our analysis results. All these examples are based on data for one year. Other time periods of interest include calendar month and quarter. In all our analyses p-values were estimated using Z = 9999 Monte Carlo experiments.

3.3.1 Receipt limits scenario

This scenario focuses on employee behavior with respect to business rules that set limits for travel expenses. Specifically, we consider a rule that states that only the actual expenses should be claimed and that the limits should not be viewed as an entitlement. Under this rule, we explore expenses incurred that do not require receipts to be submitted since they are below the corresponding specified limits. We seek to detect individuals who are likely violating this business rule and, in particular, we are looking for those who are trying to exploit the receipt limits by claiming expenses just below them (i.e., "flying just under the radar" behavior). Specifically, we will consider expense types that require a receipt above $25. Note that the converted US$ value will be presented in this paper even when the expense was incurred in another currency. We will also focus the analysis on expenses paid with cash (not by corporate credit card) over a one year time period. No covariates are used in this analysis. We present the results from two different firms {A, B} for this scenario.

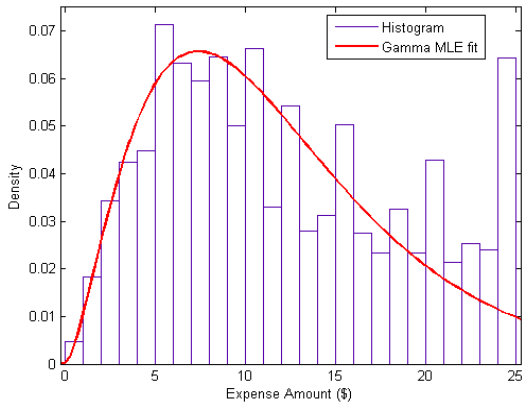

Figure 5. Histogram of expenses in firm A subject to $25 receipt limit and the corresponding Gamma fit

In the first firm A, the receipt limit of $25 is applied to expense categories like employee meals, business meals, ground transportation, parking, tips and tolls. Our analysis was performed on 660K expenses of these types that were below the $25 receipt limit. These expenses were claimed by 27K employees. The histogram of these expenses is shown in Figure 5. The maximum likelihood estimate for a Gamma distribution fit to these expenses has parameters {shape = 2.63, scale = 4.52} and the corresponding probability density function is also plotted in Figure 5. The histogram shows an increase in the counts near the maximum value of $25. It is important to note that the p-value computation described in Section 3.1 samples actual expense amounts and hence factors in the increased counts at the limit that occur across the firm. While this phenomenon of increased counts at the limit across the population is intuitive, a disproportionate increase in counts near the limit for any particular employee would be worthy of detection. Figure 6 shows the corresponding expenses for the top three employees in firm A identified by BSM in this scenario. The disproportionate concentration of expenses near the limit value of $25 is clear for these three employees.

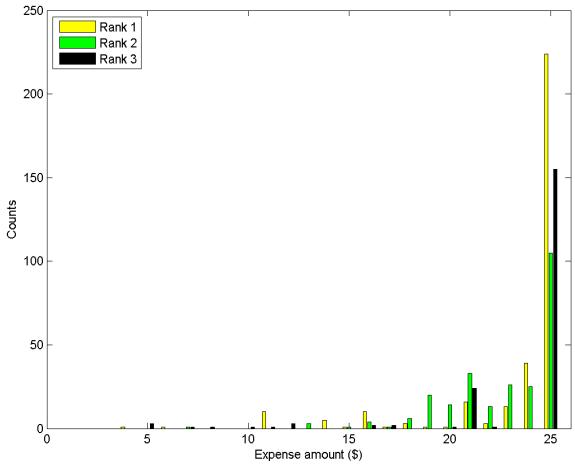

Figure 6. Histogram of expenses for the top three employees (in firm A) identified by BSM

Considering the second firm B, the receipt limit of $25 is applied to expense categories like employee meals, hotel, ground transportation, tolls/parking, tips and laundry. The data analyzed corresponds to a subset of the employees in the firm B. The analysis considered 110K expenses that were submitted by 3.6K employees. The histogram of these expenses is shown in Figure 7. The maximum likelihood estimate for a Gamma distribution fit to these expenses has parameters {shape = 1.76, scale = 4.98} and the corresponding probability density function is also plotted in Figure 7.

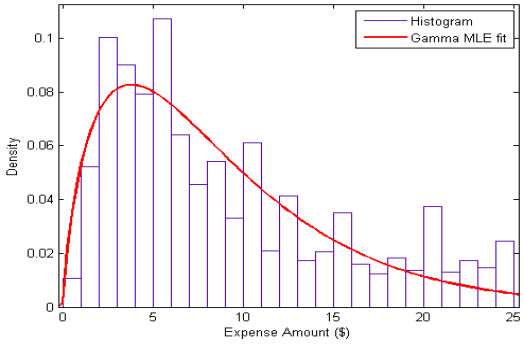

Figure 7. Histogram of expenses in firm B subject to $25 receipt limit and the corresponding Gamma fit

Figure 8 shows the corresponding expenses for the top three employees in firm B identified by BSM in this scenario. Again, the disproportionate excess concentration near the limit of $25 for these employees is clearly worthy of further investigation. Interestingly, the top employee in this case also has a concentration of $1 expenses (all corresponding to tips). We have observed a variety of expense amount patterns for the entities

identified by BSM. These would not be easily detected by simple filters considering disproportionate behavior in fixed expense amount windows below the limit.

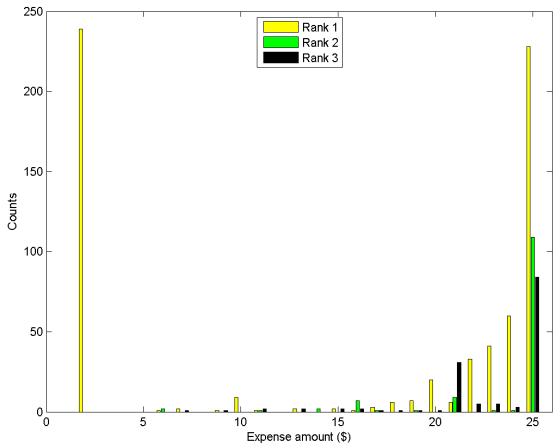

Figure 8. Histogram of expenses for the top three employees (in firm B) identified by BSM

3.3.2 Procurement analysis scenario

An important feature of our approach is the ability to do the analysis focusing on targets at different levels of granularity. Scenario 2 will utilize this capability by analyzing vendors, specifically hotel chains, to identify those that are significantly more expensive even after normalization for location is done. The analysis was done by considering the hotel room rates paid during business travel over a period of one year. The location (country, city) of the hotel was considered as a covariate for normalization. The average hotel room rate paid in each location was used as the normalization factor F (i.e., a normalized expense of 1 implies that the corresponding location's average room rate was paid). The analysis considered 523K expenses for hotel room nights that were paid to roughly 300 hotel vendors. The top ranked hotel vendor identified by BSM had significant usage (39K room nights) and a total excess charge of 9.3% after normalization by location. The second ranked hotel vendor identified by BSM had much less usage (around 2K) but a significantly higher percentage excess of 28.4% compared to the location based normalization factors.

The baseline of normalized room rates considering all the vendors and locations can be visualized using the cumulative distribution function (cdf) shown in Figure 9. The maximum likelihood estimate for a Gamma distribution fit for this baseline has parameters {shape = 7.78, scale = 0.129}. Figure 9 also shows the cumulative distribution function for the top two hotel vendors discussed above. Note that the ranking by BSM takes into account both the repetitiveness and the magnitude of excess compared to normal but the visualization in Figure 9 only depicts the latter.

In addition to the filtering of significant entities and their ranking, the BSM approach lends itself to providing diagnostic information that can help the user gain further insight on the identified entities. We have found it to be very useful to break down an identified entity's excessive deviation from the normal baseline by the covariate segments. For example, a single location is responsible for almost all of the excess exhibited by the second ranked entity. An excess of around 30% was charged by this entity at this location based on the data from all the relevant hotel vendors. This kind of diagnostic information can help focus the further investigation and corrective action.

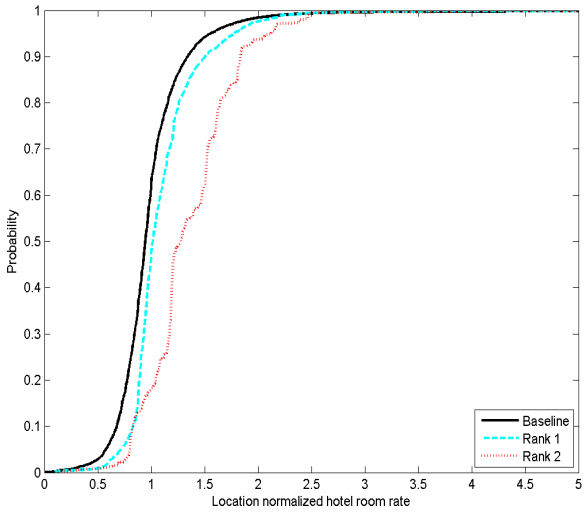

Figure 9. Cumulative distribution function for normalized room rate (baseline and BSM ranked top two hotel vendors)

3.3.3 Submission delay scenario

This scenario illustrates the application of BSM to other types of positive real valued quantities besides dollar amounts, for example, delays in submitting expense claims for approval. Organizations typically have a business rule specifying the maximum allowed time for submission after the expense was incurred. However, the guidelines typically suggest making an effort towards prompt submissions. Habitual delays in submission might indicate issues worthy of investigation even if the maximum delay limits are not always violated. This scenario identifies employees with repetitive excessive claim submission delays. The analysis was performed for firm B and focused on expenses charged to the corporate credit card.

The analysis considered 414K individual expense submissions from 4K employees over a period of a year. The average delay in claiming expenses was around 10 days for the baseline considering all 4K employees. The claim submission delay distribution was characterized by a Gamma distribution with maximum likelihood parameters {shape = 1.25, scale = 7.82}. The results for the top three employees ranked by BSM for having repetitive excessive delays are given in Table 2 and clearly show the repetitive deviation from the norm.

BSM Rank	Number of claims	Average submission delay
1	94	122
2	74	128
3	426	38

Table 2. Expense claim submission delays for the top three employees in firm B identified by BSM

4 Event Count Scenarios

In this section we consider scenarios involving discrete occurrences of events such as approvals of exceptions to specific business rules. A typical template would aim to determine if an entity has excessive (or insufficient) counts for a specified event type given the counts for the opportunities for the events. For example, consider the scenario from Section 1 to identify organizational units with excessive approval of exceptions to the prescribed class of air travel. Clearly, for this scenario we would need to consider, for each organizational unit, both the number of air travel expenses claimed and the number of air travel class exceptions that were approved. It might seem intuitive to consider some attribute of the air travel as a covariate, e.g., international versus domestic travel. However, our experience is that business controls professionals do not make accommodations for such attributes (beyond any use of such attributes in the corresponding business rule) when assessing if an organizational unit is being lax. There are other scenarios where the use of covariates is more appropriate. One such example is the approval of exceptions to the business rule that defines when receipts have to be submitted for T&E expense claims. There could be different reasons provided for why, on occasion, a receipt is missing (e.g., receipt lost, receipt not available). The rates of occurrence and approval of missing receipt exceptions clearly varies by expense type. For example, it is typically the case that missing receipt exception rates for hotel room expenses are low. On the other hand, missing receipt exception rates for ground transportation expenses like cab fares are much higher. Therefore, the expense type is an appropriate covariate when we are trying to detect organizational units with excessive approvals of missing receipts exceptions.

4.1 The Model

Our approach to detect entities with excessive (or insufficient) counts for specific events is similar to the one for expense amount scenarios in the use of the LRT. The LRT based on a Poisson model is well suited to model event counts that are proportional to known opportunities with possible categorical covariates. The LRT using the Poisson model has been used extensively in various surveillance applications (especially in public health) following the work on the spatial scan statistic by Kulldorff 1997. Indirect standardization was proposed in that work as one approach to handle categorical covariates. Let $O(\xi, F)$ and $V(\xi, F)$ represent the count of opportunities and the count of target event occurrences for entity ξ for the combination F of categorical values for the covariates, respectively. The expected number of target event occurrences $X(\xi)$ for an entity ξ is calculated using indirect standardization as:

$$X(\xi) = \sum_{F} \left\{ \left(\frac{\sum_{\xi'} V(\xi', F)}{\sum_{\xi''} O(\xi'', F)} \right) \times O(\xi, F) \right\}.$$

Following Kulldorff 1997, the test statistic $W(\xi)$ for Poisson model is given by

$$W(\xi) = Y(\xi) \times log\left(\frac{Y(\xi)}{X(\xi)} \right) + \left(U - Y(\xi) \right) \times log\left(\frac{U - Y(\xi)}{U - X(\xi)} \right),$$

where $Y(\xi)$ represents the aggregate number of target event occurrences for entity ξ over all combinations of categorical covariate values and U represents the total number of occurrences of target events over all the entities and the covariate value combinations.

The p-value is computed by performing a number Z of Monte Carlo experiments, where, in each experiment the target event counts for an entity ξ are determined by sampling from a Poisson distribution with mean equal to the expected count $X(\xi)$. As before, the p-value for the entity ξ is computed using the formula $(L+1)/Z$, where L is the number of Monte Carlo experiments with a maximum test statistic exceeding $W(\xi)$. Entities with p-values that reject the null hypothesis at the prescribed α level are candidates for further investigation and are ranked in decreasing order of their test statistic values W. The behavior of the LRT model using the Poisson model has been well-studied given its wide usage in domains like public health and epidemiology. In the next subsection we present results on production T&E data that demonstrate its applicability to this domain.

4.2 Application to Production T&E Data

As described earlier in Section 3.3, we applied our model for event count scenarios to production T&E data from multiple firms in an enterprise expense reporting environment (GERS) and reviewed the results of various scenarios with audit and business control professionals. In this section we will present results from two of these scenarios. The chosen scenarios will also illustrate the ability of our approach to do the analysis at different organizational levels. This is important feature since business controls are typically exercised by monitoring expenses for organizational units that are more suitable for expense management and policy guidance.

4.2.1 Hotel limit exceptions scenario

This scenario is related to the business rule that specifies upper limits by location on hotel room rates and requires management approval of exceptions to this rule. The goal of the analysis is to identify organizational units that are approving exceptions to this rule excessively. The analysis was done for firm B targeting 15 organizational units. In the time period of the year considered, there

were 4.6K exception approvals (events) for 43K underlying hotel expenses (opportunities) implying a baseline event rate of 10.7%. The top three organizational units identified by BSM as having significantly excessive ($\alpha = 0.01$) exception approvals are listed in Table 3. Clearly, the counts of approval events and opportunities indicate patterns of excessive approvals in these three organizational units that warrant further investigation.

BSM Rank	Number of hotel expenses	Number of hotel limit exception approvals (expected number)	Poisson test statistic W
1	777	235 (83.2)	99.75
2	609	144 (65.2)	35.96
3	1371	247 (146.8)	29.43

Table 3. Results for the top three organizational units identified by BSM as having excessive hotel limit exception approvals

4.2.2 Missing receipt exceptions scenario

This scenario addresses the business rule that requires submission of receipts based on the expense category and amount. The goal of the analysis is to identify approvers who are approving exceptions to this rule excessively. As discussed earlier, the rates of occurrence and approval of missing receipt exceptions across the firm clearly varies from one expense category to another. Therefore, the expense category is an appropriate covariate for this analysis.

BSM Rank	Number of exception opportunities	Number of exception approvals (expected number from indirect standardization)	Poisson test statistic W
1	403	245 (22.4)	363.5
2	1255	375 (72.5)	314.7
3	624	234 (25.7)	309.1

Table 4. Results for the top three approvers identified by BSM as having excessive missing receipt approvals

The analysis was done for firm A considering the exception approvals over a one year period. The analysis considered 18K exception approvals by 12K approvers that resulted from 159K opportunities for this exception. Table 4 shows the results for the top three approvers identified by BSM as having excessive approval rates after normalization by expense categories. Table 4 clearly indicates the repeated approvals by these approvers and its excessiveness when compared to expected numbers based on behavior across all approvers. Examining the diagnostic information for the top ranked approver in Table 4 led to the actionable insight that the dominant expense categories for the corresponding exceptions were employee lunch and dinner and also that one employee was the main contributor.

5 Discussion

The diversity of the application areas for fraud detection has been pointed out by Bolton and Hand (2002). Bolton and Hand also stress that operational and data characteristics of the application domain determine suitable fraud detection methods and tools. Analysis of expense claims for audit and business controls purposes is an application domain with specific characteristics and requirements. The models and methods presented in this paper address the following needs in this domain:

- Conservative analysis that identifies entities for further investigation when significant evidence is available.
- Entities analyzed at various levels of granularity in the firm based on the scenario and the corrective action that will follow.
- Analysis that can handle the data and operational characteristics like lack of labeled data, significant tails in the value distributions, impact of business rules (e.g., limits), and the need for normalization considering one or more covariates.
- Provide detailed evidence for the entities identified to help audit and business controls professionals determine if an investigation is warranted and to bootstrap it if the investigation is pursued.

Our simple and intuitive template structure has been used to create over 50 specific scenarios for the analysis of T&E data in an enterprise expense reporting environment. Our scenarios utilize only the structured data logged in the expense claim process. Unstructured data for the entities identified like explanations for triggering exceptions are presented as part of the evidence used for further investigation. Including the unstructured data in the automated analysis is unlikely to be useful due to its unreliable nature (inconsistent and possibly inaccurate or even misrepresented information).

Future work also includes utilizing the BSM scoring of entities based on their outlying behavior to impact the controls and management actions for selected entities within the travel expense management system.

The BSM model has also been applied to other domains like procurement (one such scenario was illustrated in Section 3.3.2). Our ongoing work in other domains suggests that BSM can be a valuable part of a toolkit for identifying entities with outlying behavior in various domains.

6 Conclusion

We have described a set of Behavioral Shift Models developed in the context of Travel and Entertainment (T&E) expense management for efficient auditing and business controls. Our models combine recent advances in unsupervised statistical analyses with T&E domain knowledge to profile and rank entities in a firm based on the deviation of their travel spending behavior from that of the general population. The focus is on repeated suspicious behavior as opposed to a one-time outlying case, in line with the domain practice of conservative filtering that takes into account the amount of evidence available. We have modeled two broad classes of data: one for continuous, real-valued variables and one for discrete, Poisson-type variables covering a large number of scenarios in the T&E domain. We characterized the discriminating power of our method using a systematic simulation approach that evaluates the detection capability of the BSM for different data distributions with different amounts of engineered deviations from the population norm. Lastly, we have presented several example scenarios with validated results of our analyses of T&E data from several firms.

7 Acknowledgements

We would like to thank the audit and business controls professionals who shared their deep knowledge of this domain with us and helped validate our results.

8 References

Beasley, M.S., Carcello, J.V., and Hermanson, D.R. (2001): Top 10 Audit Deficiencies, *Journal of Accountancy*, *American Institute of Certified Public Accountants*.

Bolton, R. J. and Hand, D.J. (2002): Statistical Fraud Detection: A Review (with discussion), *Statistical Science*, 17(3), 235-255.

Bolton, R. J. and Hand, D.J. (2001): Unsupervised Profiling Methods for Fraud Detection, *Credit Scoring and Credit Control VII*, Edinburgh, UK.

Christianini, N., and Shawe-Taylor, J. (2000): *Support Vector Machines*, Cambridge University Press, Cambridge.

Denning, D. E. (1987): An Intrusion-Detection Model, *IEEE Transactions on Software Engineering*, Vol. SE-13(2).

Duda, R.O., Hart P.E., and Stork, D.S. (2001): *Pattern Classification*, John Wiley & Sons.

GERS: IBM Global Expense Reporting Solutions, http://www-935.ibm.com/services/us/index.wss/offering/igs/a1009035

Glaz, J., Naus, J., and Wallenstein, S. (2001): *Scan Statistics*, Springer-Verlag, New York.

Hodge, V., J., and Austin, J. (2004): A Survey of Outlier Detection Methodologies. *AI Review*, 22 , 2004, pp. 85-126.

Huang, L., Kulldorff, M., and Gregorio, D. (2007): A Spatial Scan Statistic for Survival Data, *Biometrics* 63 (1), pp. 109-118.

Kulldorff, M (1997): A Spatial Scan Statistic, *Communications in Statistics: Theory and Methods*, 26:1481-1496.

Lehmann, L.E. (1986): *Testing Statistical Hypothesis*, Springer-Verlag, New York.

National Business Travel Association (NBTA): http://www.nbta.org/About/TheValueofManagedTravel.

Williams, G., Baxter, R., He, H., Hawkins, S., and Gu, L. (2002): A Comparative Study of RNN for Outlier Detection in Data Mining, International Conference of Data Mining & CSIRO Technical Report CMIS02/102.