Statistics Based Information Security Risk Management Methodology

Upasna Saluja, Dato Norbik Bashah Idris

Faculty of Computing, University of Technology (UTM), Malaysia

Summary
On the one hand organizations are confronted with increasing sophistication, severity and number of threats and on the other hand organizations are getting even more dependent on IT which is rapidly changing with introduction of new technologies such as outsourcing, cloud, mobility and social media. Traditional risk management methodologies are proving ineffective in addressing these risks and in keeping pace with the complexity and dynamically changing IT environment. In such a situation, there is a need for an effective Risk Management methodology that can address diverse kinds of risks and leverage data from within the organization to analyze risks scientifically rather than through primitive and subjective methods based on rudimentary calculations. This paper presents a methodology which addresses these issues. Adapting from Medical and Finance fields, this methodology has generated information security risk indicators for the IT environment. These Risk Indicators are observed over a period of time leading to data driven factual process that inspires greater confidence among stakeholders. Drawing inspiration once again from the fields of medicine and finance, this methodology has conducted risk analysis statistically using second generation statistical technique Structured Equation Modeling (SEM). The methodology provides a prediction model that predicts future risks scientifically. The Relative Risk Benchmark that this methodology has developed improves decision making when organizations need to prioritize risks, by providing a scientifically generated contribution of each risk towards the negative impact that organization faces. The path breaking information security risk management methodology cuts costs by enabling organizations to focus efforts and resources only on the risks that matter. This methodology inspires greater confidence in the results of the risk assessment since risks are assessed scientifically thus removing assessor bias while reducing the dependence of risk assessments on expert judgment.

Keywords:
Information Security Risk Assessment, Qualitative Risk Assessment, Quantitative, Statistical.

1. Introduction to Information Security Risk Management

Information security is an organisation's approach to maintain confidentiality, integrity and availability of the information it processes and manages. Organisations need to identify and manage the risks to information assets for accomplishment of the organisation's objectives. Management of risks entails number of activities including establishing of a framework [1].

In order to manage the risks organisations first need to understand the risks they face. This is accomplished through the overall process of risk assessment which encompasses the process of risk identification, risk analysis and risk evaluation.

2. Inadequacy of existing risk assessment methodologies

The key question that organisations face is "how to assess risks" [2]. Traditional Risk Assessments methodologies are quite qualitative in nature. They were effective in the past as threats were simple and IT infrastructure was not that complicated. In the recent times, threats have become quite complicated and infrastructure has become quite complex which has resulted in making existing Information Security Risk Assessments quite ineffective.

Information Security Risk management has been criticized for being shallow rather being based upon scientific approach. Even major US government programs like FISMA which require organisations to implement risk management when managing IT systems has also been criticized for the lack of scientific rigor. The subjectivity in assessing the value of assets, the likelihood of threats occurrence and the significance of the impact has also been criticized [3].

3. Necessity of the day

Information Security Risk management has been criticized for being shallow rather being based upon scientific approach. Even major US government programs like FISMA which require organisations to implement risk management when managing IT systems has also been criticized for the lack of scientific rigor. The subjectivity in assessing the value of assets, the likelihood of threats occurrence and the significance of the impact has also been criticized [3]. Necessity of the day is to innovate an

Manuscript received October 5, 2015
Manuscript revised October 20, 2015

Information Security Risk Management methodology which has scientific foundation.

4. Objective of the study

To develop an Information Security Risk Management methodology that is more scientific and less subjective in nature, which facilitates better risk-related decision making. This paper presents statistics based Information Security Risk Management Methodology 'SQRC: Statistical Quantitative Risk Calculator' [4]. SQRC Information Security Risk Management process consists of five key processes:

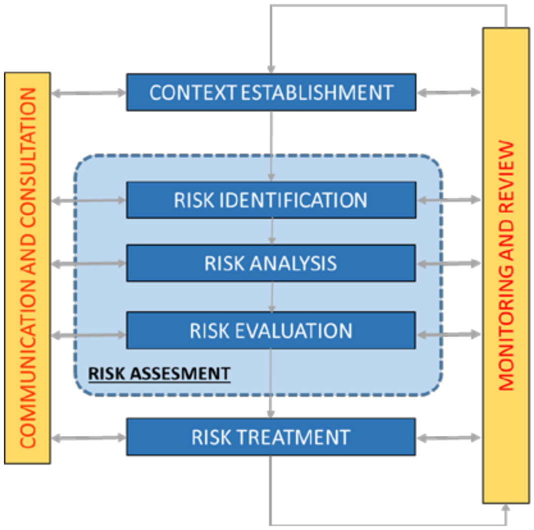


Figure 1 SQRC Risk Management

Context Establishment

- Information Security Risk Assessment
- Information Security Risk Treatment
- Information Security Risk Communication and Consultation
- Information security Risk Monitoring and Review

A description of SQRC Information Security Risk Management process is given in the following sections. This method is designed to give the risk assessor a pragmatic and repeatable process of risk assessment which is appropriate to organisations of any size or industry.

4.1. Context Establishment

SQRC Context Establishment Phase considers the external environment, internal organisational imperatives and the overall context in which risk management is undertaken. This includes recognizing key drivers of Risk Management, and finalizing the approach and scope of risk management.

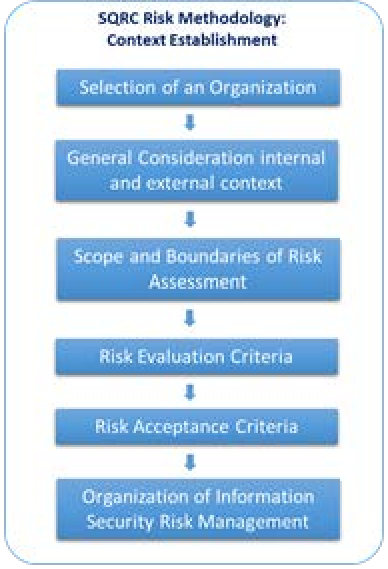

SQRC Risk Methodology:
Context Establishment
Selection of an Organization
General Consideration internal and external context
Scope and Boundaries of Risk Assessment
Risk Evaluation Criteria
Risk Acceptance Criteria
Organization of Information Security Risk Management

Figure 2: Context Establishment

4.2. Information Security Risk Assessment

The following section focuses on Risk Assessment process:

4.2.1. Information Security Risk Identification

This phase starts with the identification of Risk Indicators. Risk Indicators represent the different scenarios which have the potential of negative impact on the organization. ISO 27005 standard is one of the most established standard that is widely accepted. The list of threats provided by ISO 27005 is found closest to the concept of Risk Indicators and that is why this list is considered as the foundation for Risk Indicators, while relevant pointers from other methodologies are also considered. Besides interviews with relevant stakeholders involved in managing of information security risks are undertaken to further refine potential risks indicators.
Referring to finance and medical risk assessment approaches, SQRC conducts statistical analysis to identify Information Security risks. SQRC conducts regression analysis which demonstrates the relationship between predictor variables and response variables and is widely used for data modelling. Accordingly, Consequence Indicators represent the areas where an organisation is expected to observe impact and are considered as response variables. Since information security deals with confidentiality, integrity and availability of information, SQRC looks into impact in these three areas. Availability aspect is somewhat tangible while confidentiality and integrity are non-tangible aspects. Additionally,

organisational processes and infrastructure too bear negative impact.

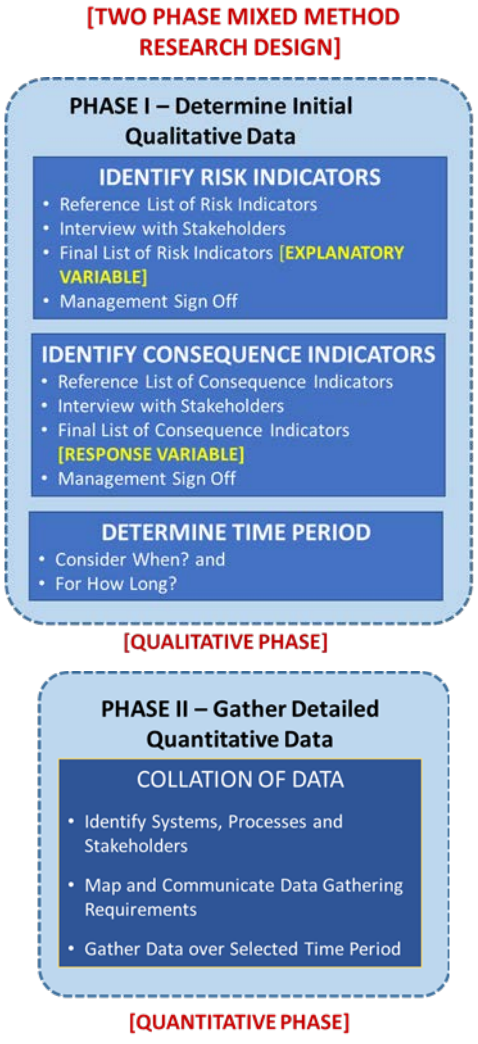

[TWO PHASE MIXED METHOD RESEARCH DESIGN]
PHASE I – Determine Initial Qualitative Data
IDENTIFY RISK INDICATORS
• Reference List of Risk Indicators
• Interview with Stakeholders
• Final List of Risk Indicators [EXPLANATORY VARIABLE]
• Management Sign Off
IDENTIFY CONSEQUENCE INDICATORS
• Reference List of Consequence Indicators
• Interview with Stakeholders
• Final List of Consequence Indicators [RESPONSE VARIABLE]
• Management Sign Off
DETERMINE TIME PERIOD
• Consider When? and
• For How Long?
[QUALITATIVE PHASE]
PHASE II – Gather Detailed Quantitative Data
COLLATION OF DATA
• Identify Systems, Processes and Stakeholders
• Map and Communicate Data Gathering Requirements
• Gather Data over Selected Time Period
[QUANTITATIVE PHASE]

Figure 3: Mixed Method Research Design

The key question is "How to measure impact?" Referring to 'Exploratory Sequential Mixed Method Design' of Two Phase Mixed Method Research design, observe identified Indicators over a period of time. Next, determine an appropriate unit of measurement for each risk indicator. SQRC considers three different units of measurement namely downtime which is appropriate for "downtime of systems, applications and networking equipment" for which unit of measurement is "Hours"; secondly "Frequency of Incidents" where each incident counts, and thirdly binary format (Yes / No) where the data represents presence or absence of a control.

4.2.2. Information Security Risk Analysis

Phase 1: Define Risk Indicators and create Data Set for statistical analysis - The objective of Risk Analysis is to bring out an understanding of risks faced by the organisation based upon analysis of the data collected in Risk Identification phase. Hence, in this phase conduct Risk Analysis on the data collected during the previous Risk Identification phase.

Phase 2: Review of Statistical Attributes of Data to understand statistical parameters including mean, stand deviation, frequency etc of data.

Phase 3: Statistical Analysis - SQRC statistical techniques known as Partial Least Square [5]. PLS is a regression technique under Structural Equation Modeling. PLS analysis combines Principal Components Analysis and Multiple Regression. Principal Component Regression [6] always captures maximum variance in X; and Multinomial Linear Regression (MLR) achieves maximum correlation between X and Y. Partial Least Square [7] captures both by maximizing covariance between X and Y. In PLS, firstly, latent factors are extracted which explain as much covariance as possible, between the explanatory and response variables. Subsequently, a regression analysis is used to predict values analysis is based on second generation statistical for the response variables on the basis of the decomposition of the explanatory variables.

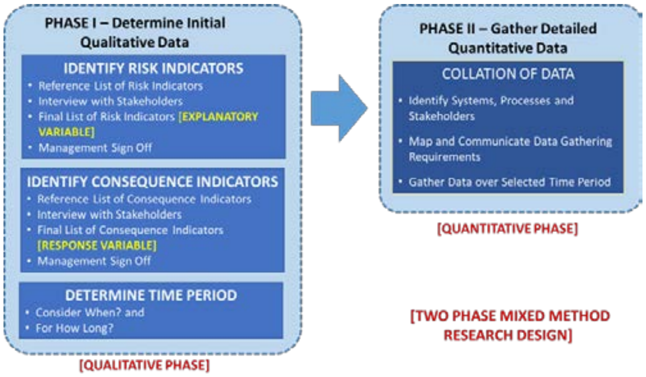
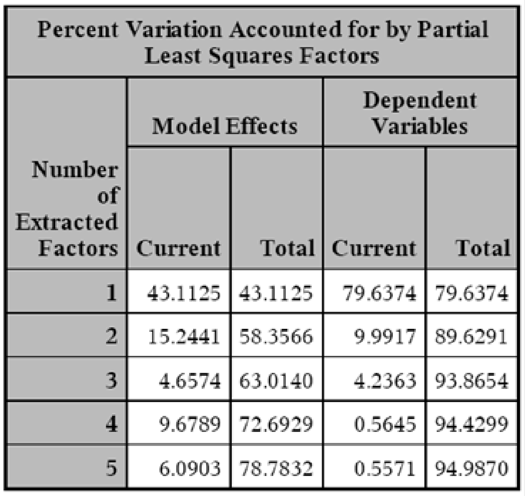

PHASE I – Determine Initial Qualitative Data
IDENTIFY RISK INDICATORS
• Reference List of Risk Indicators
• Interview with Stakeholders
• Final List of Risk Indicators [EXPLANATORY VARIABLE]
• Management Sign Off
IDENTIFY CONSEQUENCE INDICATORS
• Reference List of Consequence Indicators
• Interview with Stakeholders
• Final List of Consequence Indicators [RESPONSE VARIABLE]
• Management Sign Off
DETERMINE TIME PERIOD
• Consider When? and
• For How Long?
[QUALITATIVE PHASE]
PHASE II – Gather Detailed Quantitative Data
COLLATION OF DATA
• Identify Systems, Processes and Stakeholders
• Map and Communicate Data Gathering Requirements
• Gather Data over Selected Time Period
[QUANTITATIVE PHASE]
[TWO PHASE MIXED METHOD RESEARCH DESIGN]

Figure 4 Risk Analysis

Partial Least Squares [6] regression is a way to estimate parameters in a scientific model. It is essentially an extended version of the multiple linear regression. A linear model defines the relation between a dependent variable Y and a set of predictor variables, Xs, in such a way:

$$Y = b0 + b1X1 + b2X2 + \ldots + bpXp$$

Here the regression coefficient for the intercept is defined by b0 while the remaining bi define the regression coefficients for the variables 1 through p



Percent Variation Accounted for by Partial Least Squares Factors				
	Model Effects		Dependent Variables	
Number of Extracted Factors	Current	Total	Current	Total
1	43.1125	43.1125	79.6374	79.6374
2	15.2441	58.3566	9.9917	89.6291
3	4.6574	63.0140	4.2363	93.8654
4	9.6789	72.6929	0.5645	94.4299
5	6.0903	78.7832	0.5571	94.9870

Figure 5 Proportion of Variance Experienced

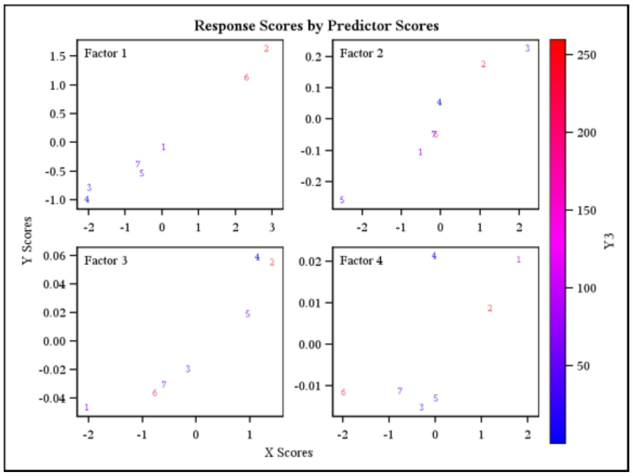
Partial least squares [8] regression is used to fit a model for one or more response variables based on one or more predictors. Overall the goal of PLS regression is to predict Y based on the values of X and to describe their common structure. The core objective of PLS regression analysis is to create a linear model,

Y=XB + E

Where Y is an n case by m variables response matrix, X is an n case by p variables. Explanatory matrix, B is a p by m regression coefficient matrix; and E is a noise term for the model which has the same dimensions as Y.

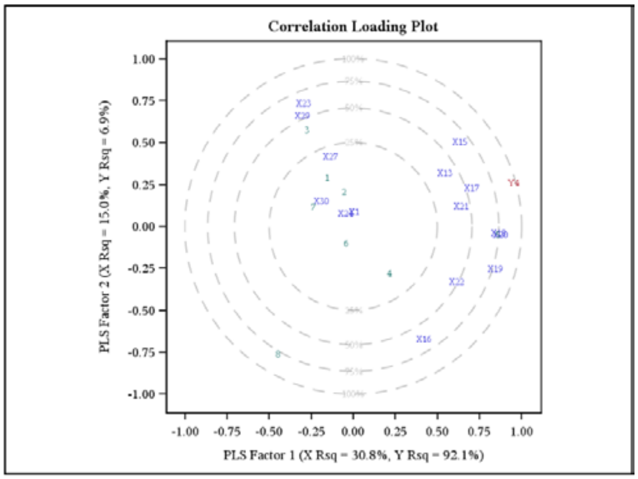
Model Building Strategies: While conducting statistical analysis, SQRC takes care of the following model building strategies:

- **Proportion of Variance Explained** (PVE) in X and Y: SQRC considers the proportion of variance explained (PVE) as a measure to see how good the regression line predicts obtained or actual scores. The value of PVE ranges from zero (stating that the regression line has no predictive value) to one (stating that the regression line predicts each obtained score in a perfect manner). The Proportion of Variation Accounted for by Partial Least Squares Factors explains the maximum variation in both explanatory variables (X) and response variable (Y). As per the best practices, a statistical model is considered to be good as long as it explains 75% of variation in both X and Y. A representative sample depiction is shown in Fig 4 which depicts Proportion of Variation Accounted for by Partial Least Squares Factors explaining maximum variation in both predictors (X) and responses (Y).
- **Response Scores by Predictor Scores**: SQRC PLS analysis, creates graph “Response Scores by Predictor Scores” for displaying XY-Score correlation which helps to understand the association between X scores and Y scores.



- Figure 6 Response Scores by Predictor Scores

- For a good PLS model, the first few factors should show a high correlation between X and Y scores. This graph provides a good visual impression of X score and Y score correlations for each of the components.



- Figure 7 Correlation Loading Plot

- **Correlation Loading Plot**: Component plots exhibit the component score of each component loading for each variable for a pair of principal components.
- The loading plot is a plot of the relationship between the original variables and the subspace dimension. It is used to interpret relationships between variables.
-

Performance of the Model: To assess performance of the model, SQRC considers MAPE metrics. Mean

Absolute Percent Error (MAPE) is a measure of how high or low is the difference between the predictions and the actual data. MAPE close to 0 indicates perfect prediction. A value of MAPE which is less than 15% is generally considered as good.

Model Diagnostics: To make sure that the model is following normality of the data, SQRC performs the following which further helps to identify or flag outlier, if existing.

- **Quantile - Quantile plot (Q-Q Plots)**: The quantile-quantile or q-q plot is an exploratory method used to check the normal distribution assumption for error term. This helps to detect outliers and non-normality of residuals. In a good situation, all points are expected to fall on or close to a diagonal straight line.
- **Actual by Predicted Scatter Plot:** This graph shows Actual vs. Predicted values for a dependent variable, which exhibits a fair idea of accuracy of the model. This graph provides a clear indication of how well the model fits the data. For a perfect fit, all the points would be on the diagonal.
- **Derivation of the Best Performance Model**: Variable selection process is vital for a model with high performance. To determine which explanatory variables are significantly related to the response variable, the results of initial model generated at the first instance should be examined. To work towards improving the quality of a model, “variable screening” is advised. It is recommended to exclude the variables which do not provide enough contribution towards model.Consider the following in order to derive the best model:
- **Variable Importance in the Projection:** The Variable Importance in the Projection (VIP) is based on the Canonical Powered Partial Least Square Regression (CPPLS). The CPPLS algorithm assumes that the column space of X has a subspace of dimension M containing all information relevant for predicting Y (known as the relevant subspace). The strategy for variable selection is usually based on a rotation of the solution by a manipulation of the PLS weight vector (i.e. w) or the regression coefficient vector (i.e. b). The VIP statistics is computed for each variable and latent factor. Variable Importance in Projection (VIP) values indicates the importance of each variable in the prediction model.
- **Parameter Estimates for Centred and Scaled Data:** These sets of parameters represent parameters computed based on the standardized data. The approach is to scale each variable to unit variance by dividing them by their Standard Deviation and then centre them by subtracting their averages. This helps to provide each variable with the same weight, meaning the same importance from the analysis point of view. Hence, these estimates help to provide relative importance of each variable in influencing dependent variable.

Each specific explanatory variable [9] is looked into, in the light of above mentioned points. For example, Risk Indicator (explanatory variable) having a relatively small coefficient (in absolute value) coupled with a small value of VIP makes a trivial contribution to the prediction model. Such variables are worthwhile to exclude from the model. After removing such a specific variable, an analysis needs to be conducted again without this variable. Again, if any variable is not significant enough then an analysis needs to be conducted again after removing that insignificant variable. SQRC method involves iteratively dropping of such an explanatory variable and keeps checking the impact of dropping specific variable on the model by going through the steps mentioned above.

This exercise needs to be done repeatedly until the quality of model is good and all variables are worth retaining. In this manner, the less significant variables get excluded and the new model gets generated based on new (which are reduced) set of variables in every cycle. This process gets iterated many times until further exclusion of any variable does not result in better quality of the model.

Phase 4: Model Validation - While generating prediction model, it is imperative to validate the prediction model. SQRC addresses it through Cross Validation.

Phase 5: Results of Risk Analysis

- Relative Risk Benchmark: SQRC identifies the contribution made by an explanatory variable towards the statistical model of response variable through Relative Risk Benchmark (RRB). SQRC takes into consideration the ‘variable in projection’ while creating RRB. Relative Risk Benchmark is presented as a “pie chart”, displaying all the risk indicators which have contributed towards statistical prediction model of the response variable.

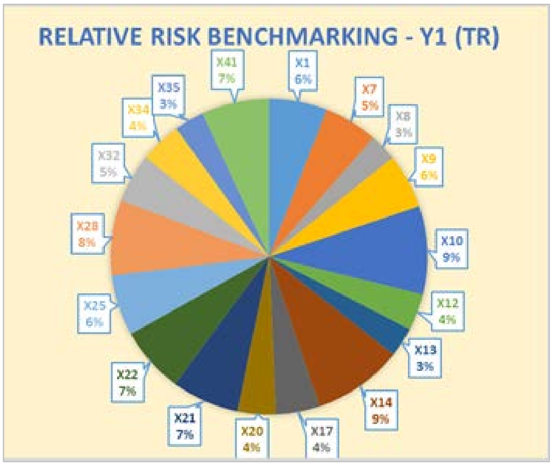


Figure 8 Relative Risk Benchmark

- Key Risk Influencers: SQRC derives the statistically significant risk influencers from the identified risks that apply to the organisation on the basis of Wolds's criterion. Rather than relying upon subjective estimates, SQRC presents the list of key risk influencers derived objectively through statistical techniques.

4.2.3. Information Security Risk Evaluation

Just as SQRC conducts statistical analysis during risk analysis phase, risk evaluation in SQRC is also based upon statistical parameters. The Risk Evaluation phase relies upon the two statistical parameters - Variable in Projection and Coefficient of Regression to determine the significance of the risk based on thresholds defined by the organisation. The following steps apply:
Step 1 Determine Regression Coefficient – Consider statistical model resulted after statistical analysis of Risk Analysis phase. Examine each risk indicator in the light of its relationship with the consequence variable and consider risk indicators which indicate high risk.
Step 2 – Determine value of Variable in Projection of each variable. Risk Indicator is statistically significant only if Variable in Projection > 0.8.
Step 3 – Determine Overall Risk Level of an organization considering results of steps 1 and step 2 for each risk indicators which is a part of statistical model.

4.3. Information Security Risk Treatment

For identified risks, organizations need to take action. In this risk treatment phase, when organizations choose to implement new or modified controls and measures to reduce risk to bring it within the risk acceptance limits, they need a mechanism to predict what reduction in risk will be accomplished on deploying the new or changed control. In order to conduct this exercise, refer to results of Risk Evaluation. Referring to Risk Evaluation, consider all those Key Risk Influencers which need to be addressed for Risk Treatment. SQRC Statistical Prediction Model further helps to evaluate the benefit of implementing any additional control in order to mitigate information security risk.

4.4. Information Security Risk Communication and Consultation

Risk Communication needs to start from the very inception of the risk management process, i.e. the context setting phase and must continue iteratively till the final phases of risk treatment. Having risk communication and consultation at each phase of the risk management process ensures that on the one hand, the stakeholders are actively involved in the process and on the other hand, this ensures that risk management process is aligned with the business sponsor's expectation.

4.5. Information Security Risk Monitoring and Review

Risk management process should be continuously monitored. It needs to be amended and updated according to changes in the environment.

5. Research Methodology

The research was conducted using mixed method approach specifically Exploratory Sequential method. This design consists of a recognized two-phase sequential design where a topic is qualitatively explored before building the quantitative phase [10].

6. Key Strengths of SQRC

SQRC Caters for Interdependence among Diverse Risks: In IT environment, different aspects which can pose risks to the environment are not independent. Conventional Risk Assessments do not factor this in and cater for association and collinearity among various risk areas while conducting the risk assessment. SQRC statistical analyzes the association among different aspects while looking into their effect on outcome observed by an organisation. SQRC thus provides a more realistic representation of risks.

Diverse Risk Areas Analysed Statistically: SQRC Risk Assessment handles the complicated situation of analyzing different risk areas which are measured in different units of measurements. SQRC captures details of incidents at granular level and is able to consider risk areas measured in different units in a holistic risk analysis.

SQRC Incorporates Statistical Analysis & Provides Prediction Model: Conventional Risk Assessment approaches calculate the probability of potential vulnerabilities and impact of any incident largely based on rough estimates. In the SQRC approach, however, there is no room for guess work. Analysis is based on the observed scenarios and the data collected from various resources pertaining to incidents.

Evaluation of Risks in Terms of Business Linked Outcomes: SQRC approach independently considers the real business impact as observed by the organisation, under context specific business outcomes, while also considering various incidents that have happened in the organisation.
Objectivity in Risk Assessment: In conventional and traditional approaches, the knowledge, experience and attitude of an assessor leave an impression on the results of Risk Assessment. As a result, when a new assessor conducts the same assessment, there could be a deviation in the results, influenced by individual preferences and assessor

biasness. However, SQRC RA approach is objective and is not subjective by nature. Since the analysis is data dependent, the results of the Risk Analysis stay consistent irrespective of the person who conducts the RA.

Promotes Security Metrics: Factual and Reliable: Risk Analysis is factual since it is based upon observed data, which has been collected by the organisation. There is no room for guesses made by different people whether system or process owners or Risk Assessors.

SQRC inspires greater confidence among stakeholders since it is based upon data and scientific analysis.

7. Future Research Potential

Future research and efforts need to be directed towards converting the methodology into easily deployable tools that can be used for Info Sec Risk Analysis.

References

[1] Radack, S., CONDUCTING INFORMATION SECURITY-RELATED RISK ASSESSMENTS. 2012, National Institute of Standards and Technology U.S. Department of Commerce].

[2] Ellen McDermott, B.B.a.D.G. Information Security is Information Risk Management. in New security paradigms. 2001.

[3] Katsicas, S.K., Computer and Information Security Handbook. 2009: Elsevier Inc.

[4] Idris, U.S.a.D.N.B., Statistical Quantitative Risk Calculator (SQRC). International Conference on Computer Technology and Science (ICCTS 2012), 2012.

[5] Haenlein, M.K., Andreas semPLS: Structural Equation Modeling Using Partial Least Squares, p. doi:10.1207/s15328031us0304_4, Editor. 2010. p. pages: 283–297.

[6] Jong, S.d., PLS fits closer than PCR. Journal of Chemometrics, 7(6), pages 551–557, 1993.

[7] Kaplan, M.H.a.A.M., A Beginner’s Guide to Partial Least Squares Analysis, in UNDERSTANDING STATISTICS Lawrence Erlbaum Associates, Inc. p. 283–297

[8] Svante Wold, M.S., Lennart Eriksson, PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems 58 2001 109–130 Ž, 2001.

[9] Hill, M. Credit Risk Indicators 2013.

[10] John W. Creswell, V.L.P.C., Designing and Conducting Mixed Methods Research. 2nd Edition ed. 2011, California: SAGE Publications