International Journal of Advanced Science and Technology
Vol.91 (2016), pp.39-48
http://dx.doi.org/10.14257/ijast.2016.91.04

Principle Component Analysis of Function Point Elements

Dr. Masood Uzzafer

Associate Professor
Amity University - Dubai
muzzafer@amityuniversity.ae

Abstract

Function point elements are metrics to measure the size of software projects. This article investigates the relationship among function point elements using principle component analysis. Principle component analysis reveals the relationship of different function point elements such that they may be measuring the same attribute of a software project. Therefore, principle component analysis brings out the influence of a function point element over each other. Principle component analysis can help to integrate the function point elements.

Keywords: *Software projects, Software size estimation, Function point element*

1. Introduction

Function Point elements have attracted much attention in the software research and development industry. Ever since function point were introduced by IBM in 70's their nature, behaviour, impact, distribution and correlation have been studied by the software researchers and the software practitioners. The idea behind function points is to standardize the measurement of the various software functions to estimate the software development effort which is independent of the computer language, development methodology, technology and the capability of the team developed the software. The International Function Point Users Group (IFPUG) was founded in the late 80s and is a membership governed, non-profit organization committed to promoting and supporting function point analysis and other software measurement techniques. There have been various releases of the Function Point by the International Function Point Users Group (IFPUG) which includes release 'Counting Practice Manual – 4.2' release.

It is critical to understand the relationship of the function point elements with other components of software projects; for example software quality, software reliability, software defects and testing requirements. Similarly, it is essential to understand the relationship of function point elements with each other. The degree of influence of a function point element over other elements shows that the function point element may be measuring the same attribute of a software project which other function point elements are measuring. Furthermore, based on this relationship, a function point element can be predicted from other function point elements. Previous studies [1-3] discussed the relationships among function point elements. This paper further extends this discussion and use principle component analysis to investigate the relationship of function point elements.

The paper is organized in the following way: Section 2 discusses the function point elements, Section 3 describes the dataset used for the principle component analysis, Section 4 presents the principle component analysis of the function point elements and Section 5 draws some conclusions.

ISSN: 2005-4238 IJAST

2. Function Point Elements

Function point estimates the size of a software project using five elements: Internal Logical Files (ILF), External Interface Files (EIF), External Inputs (EI), External Outputs (EO) and external Enquiries (EQ). Function point calculations begin with counting the five elements. Each function point element is assigned a complexity level (Low, Average, High) based on its associated file number. The associated file numbers are described as Data Element Type (DET), File Type Referenced (FTR) and Record Element Types (RET). The complexity metrics of function point elements is presented in Table 1.

Table 1. Function Point Element Complexity Metrics

ILF/EIF	DET			EI	DET			EO/EQ	DET		
RET	1-19	20-50	51+	FTR	1-4	5-15	16+	FTR	1-5	6-19	20
1	Low	Low	Avg	0-1	Low	Low	Avg	0-1	Low	Low	Avg
2-5	Low	Avg	High	2	Low	Avg	High	2-3	Low	Avg	High
6+	Avg	High	High	3+	Avg	High	High	4+	Avg	High	High

Each function component is then assigned a weight according to its complexity shown in Table 2.

Table 2. Function Point Complexity Weights

Component	Low	Average	High
External Inputs	3	4	6
External Outputs	4	5	7
External Inquiries	3	4	6
Internal Logical Files	7	10	15
External Interface Files	5	7	10

Unadjusted Function Point (UFP) is the total number function points counted together. The unadjusted function point is computed from the following equation.

$$UFP = \sum_{i=i}^{5} \sum_{j=1}^{3} w_{ij} x_{ij} \quad (1)$$

Where w_{ij} is the complexity weight and x_{ij} is the count of each function element. UFP is then multiplied by the Value Adjustment Factor (VAF) to get the function point (FP) count using equation 2. The VAF is calculated from 14 General System Characteristics (GSC). These characteristics are 1) Data Communication 2) Distributed Functions 3) Performance 4) heavily used configuration 5) transaction rate 6) on-line data entry 7) end user efficiency 8) on-line update 9) complex processing 10) reusability 11) installation ease 12) operational ease 13) multiple sites and 14) facilities change. The GSC values are summed to calculate the VAF.

$$VAF = 0.65 + 0.01 \sum_{i=1}^{14} c_i \quad (2)$$

Where c_i are the GSC values. Finally the UFP and VAF are multiplied to get the function point (FP) count.

$$FP = UFP \times VAF \quad (3)$$

3. Understanding the Dataset

The dataset for this research is taken from the International Software Benchmarking Standards (ISBSG) repository release 4.3 [4]. ISBSG performs the data validation of the contributed data and ensures data quality and consistency. The data repository release 4.3 contains data from 3024 different projects, where almost all the projects used IFPUG standard [5] for function points. Projects which used other methods then IFPUG were excluded from the study. Furthermore, datasets with the missing function point's values were also excluded.

In the selected projects largest projects were contributed from the financial industry (banking, financial services, and accounting) the rest of the projects were from engineering (software, hardware and telecommunication), insurance, public administration, government, manufacturing, consulting and education. The collected dataset is not homogenous which ensures linearity in statistical analysis. The variety in the dataset ensures that the data samples represent different scenarios and possibilities in the software development industry.

Unadjusted function points (UFP) represents the size of the software projects in the dataset. Figure 1 shows the histogram of unadjusted function points. The minimum project size is 13 UFP; the largest is 4943 UFP; mean is 579.33 UFP and standard deviation 715.46. Majority of the projects sizes are in the range of 13 UFP to 500 UFP, while there are few projects of size more than 2000 UFP.

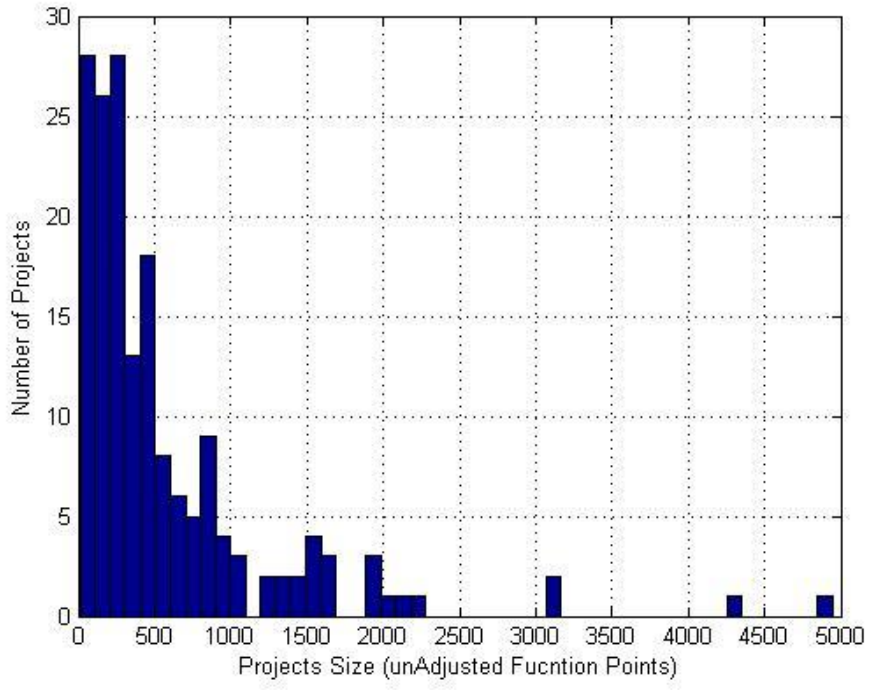

Figure 1. Projects Size (unadjusted FP)

Box plots helps to understand the measure of central tendency and dispersion. The box plot of function point elements included in the dataset is drawn in Figure 2.

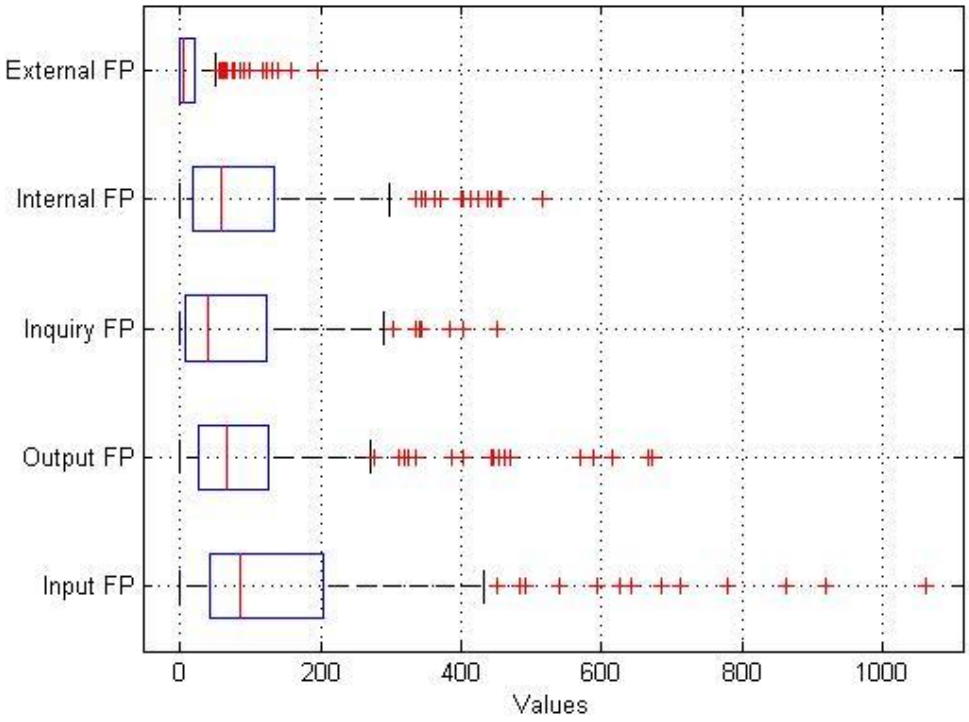

Figure 2. Box Plot of Function Point Elements

All the function point data elements which were 3 times the standard deviation away from the sample mean were classified as the outliers and were removed. The line in the middle of the box represents the median if the line is not in the center of the box that is an indication of the skewness. Skewness is a measure of asymmetry of the data around the sample mean/median. The lower and upper lines of the box are 25^{th} and 75^{th} percentiles respectively. The distance between the upper and lower lines is the interquartile range. Whiskers, lines extending above and below the box, show the rest of the data. The length of the whiskers is set to 1.5 times the interquartile range. Plus sign shows the data point which are 1.5 times away from the interquartile range. Table 3 gives the median and percentile values of the function point elements in the dataset.

Table 3. Median and Percentiles of the Function Point Elements

	Median	25^{th} percentile	75^{th} percentile	100^{th} percentile
External Input (EI)	86.5	41.5	204.5	1061
External Output (EO)	67	26	125	673
External Inquiry (EQ)	39	7.25	122.5	450
Internal Logical Files (ILF)	58	17.75	133	516
External Logical Files (ELF)	5	0	20	195

The median of EI is 86.5 UFP while the highest value is 1061 UFP, the EO has a median of 67 UFP with the highest value of 673 UFP, the median of EQ is 39 and highest is 450 UFP, the ILF has a median of 58 UFP while its highest is 133 UFP and ELF has a median of 5 UFP with the highest UPF is 195.

The box plot shows that the EI function point element has the longest tail above the upper whisker and the data values are more widely spread over the upper

whisker then in any other function point element. The median line is not in the middle of the box and has a large area after it in the box. This trend represents positive skewness of the dataset meaning that the data values are more spread out after the median. This phenomenon is also observed with other function point elements and found to be common in all function point elements. The ELF function point has the smallest set of values with fairly small upper and none existing lower whisker.

4. Principle Component Analysis

In multivariate analysis, visualizing multidimensional dataset stretches the imagination to visualize the relationship between different driving factors of a system. Various techniques have been devised to help with such visualizing; principle component analysis is one such technique. In datasets consisting of many variables, the groups of data may move together because many variables may be measuring the same driving factor of the system. Often the variables or groups of variables are correlated with each other and provide the same information about a system. Principle component analysis takes advantage of such redundancy and replace group of variables with a single new variable where each variable is a linear combination of the original variables. The variables are orthogonal to each other and hence do not measure any redundant information of the system under analysis.

First principle component is a single axis in space all original data observations are projected on to that axis to forms a new variable such that the variance of this variable is maximum among all the possible choices. Second principle component is another axis in space which is perpendicular to the first principle component, original data observations are projected onto that axis to form a new variable such that the variance of this variable is maximum among all possible choices. Full set of principle components could be as large as the original dataset but it is the first few principle components which accounts for more than 80% of the total variance in the original dataset. Interpretation of the principle components is subjective and requires knowledge of the original dataset.

Table 4. Principle Component Coefficients

	PC1	PC2	PC3	PC4	PC5
EI FP	0.5434	-0.1577	0.040	0.1241	0.8141
EO FP	0.4517	-0.0018	-0.8111	-0.3031	-0.2149
EQ FP	0.4770	0.1751	0.5551	-0.6215	-0.2177
ILF FP	0.5209	-0.0742	0.1591	0.6873	-0.4748
ELF FP	0.0430	0.9690	-0.0830	0.184	0.1351

Correlation is an issue in function point analysis and open opportunities to investigate the collection of redundant information. Principle component analysis of function point elements can provide an understanding of the relationships among function elements and it may help to reduce the redundancy involved in measuring the size of a software project using the function point elements.

Principle component variables are the linear combination of the original variables that account for the variance in the dataset and the maximum number of principle components is equal to the number of original variables. Principle component coefficients are the principle component values drawn from the original dataset on the basis that all the principle components should be orthogonal to each other. Coefficients indicate the relative weight of each variable in the component, larger the value of the coefficient, more important the corresponding variable is in constructing the component. Orthogonality can be tested by taking the transpose of

the principle component variables and multiplying with the principle component variables yielding the identity matrix. Principle component analysis is performed on the function point elements of the selected dataset. Table 5 shows the weight of coefficients called loading for each function point element in the principle component.

Principle component analysis of the function point elements provides interesting insights into the function point elements. A Pareto plot of the variances of all the function point principle components is shown in Figure 3, which explains that first three principle components cover more than 84 % of the variance. First principle component variance accounts for more than 51%, second principle component accounts for 20% and the third principle component accounts for 12% of the variance. The principle component variances are shown in Table 5.

Table 5. Principle Components Variances

	PC1	PC2	PC3	PC4	PC5
Variances	51.17	20.65	12.83	8.98	6.35

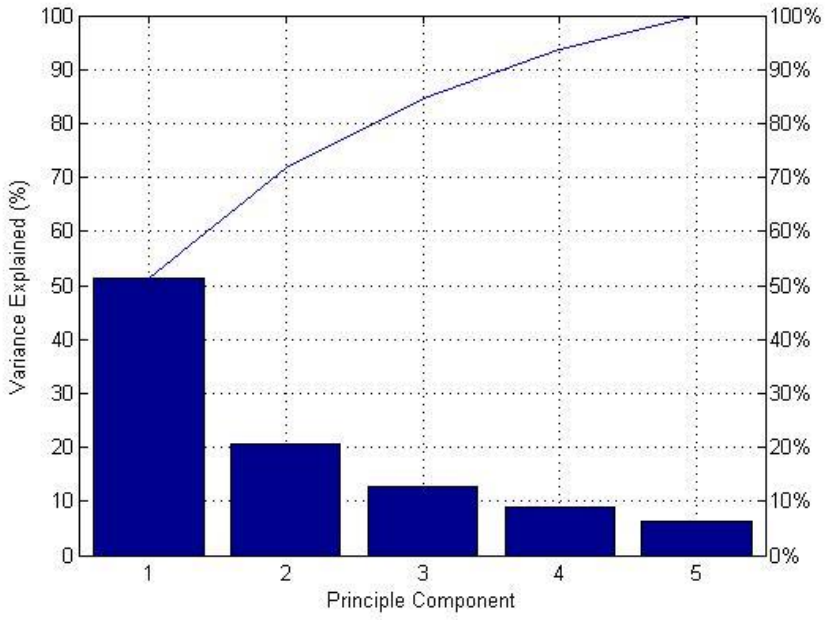

Figure 3. Function Point Principle Component Variance Pareto

From Table 4 following observations are drawn:

- EI FP (0.5434), EO FP (0.4517), EQ FP (0.4770) and ILF FP (0.5209) have large positive loading on first principle component (PC1).
- EIF FP (0.9690) have large positive loading on second principle component (PC2).
- EO FP has large negative loading on the third principle component (PC3) while EQ FP has lager positive loading on the PC3.

From these observations it can be deduced that EI, EO and ILF FP (having large positive loading on PC1) can be combined together to form a principle component, while EIF FP (having large positive loading on PC2) can be the second principle component and Inquiry FP can be the third principle component.

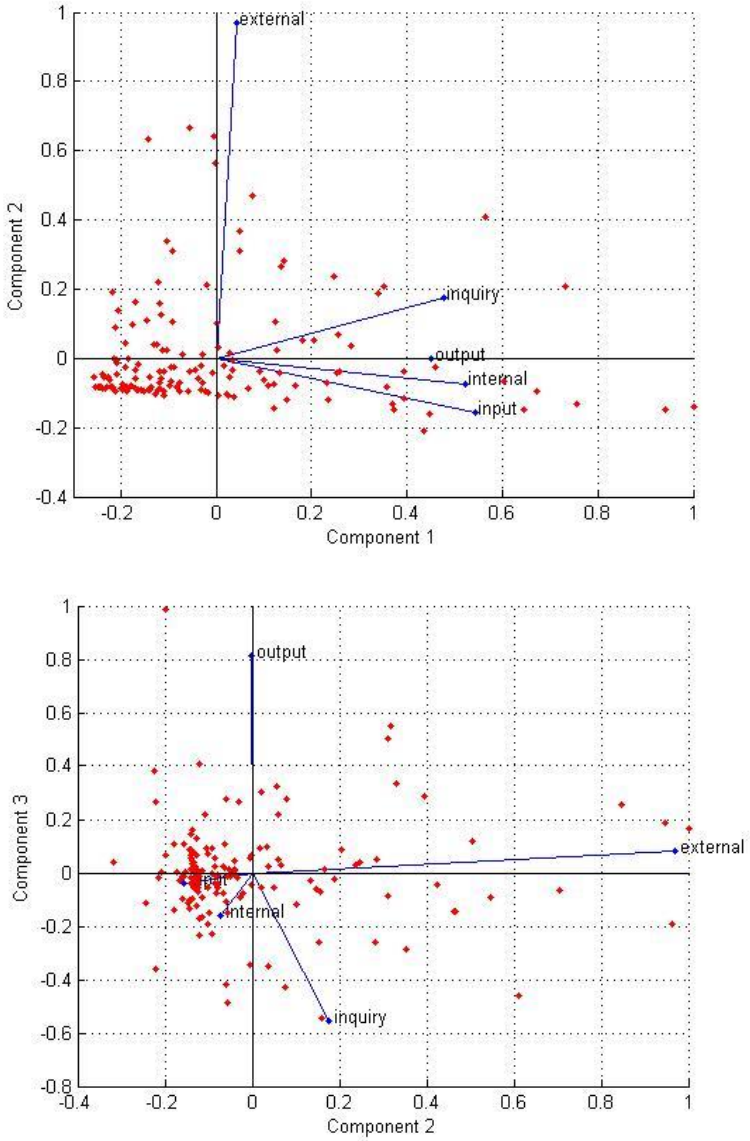

Figure 4. Biplot of Principle Component 1 vs. 2 and 2 vs. 3

It is interesting to note the logical explanation for EI, EO and ILF FP elements to be combined together is that they all deal with the data. Internal FP (ILF) contains logically related data while Input FP (EI) maintains the ILF and Output FP (EO) is the external data which passes through the application and also updates the ILF. These function points are related to data access and updates to Record Element Type (RET), Data Element Type (DET) and File Type Reference (FTR). Hence, they are related as revealed by the principle component analysis. We name this principle component Data Interface (DI). External FP (EIF) resides external to the application boundary and do not have access to the data. It is unique in way that it is outside the application and hence is not related to any other function point element and principle component analysis confirms it. The third principle component is EQ FP, it is unique because it deals with the data display and does not update the data.

Principle component analysis bring out the following three different groups of FP elements: EI, EO and ILF which are combined to form Data Interface; EO and EQ are unique in which forms the second and third principle components, respectively.

Therefore, principle component analysis suggests that five function point elements can be reduced to three elements: DI, EO and EQ.

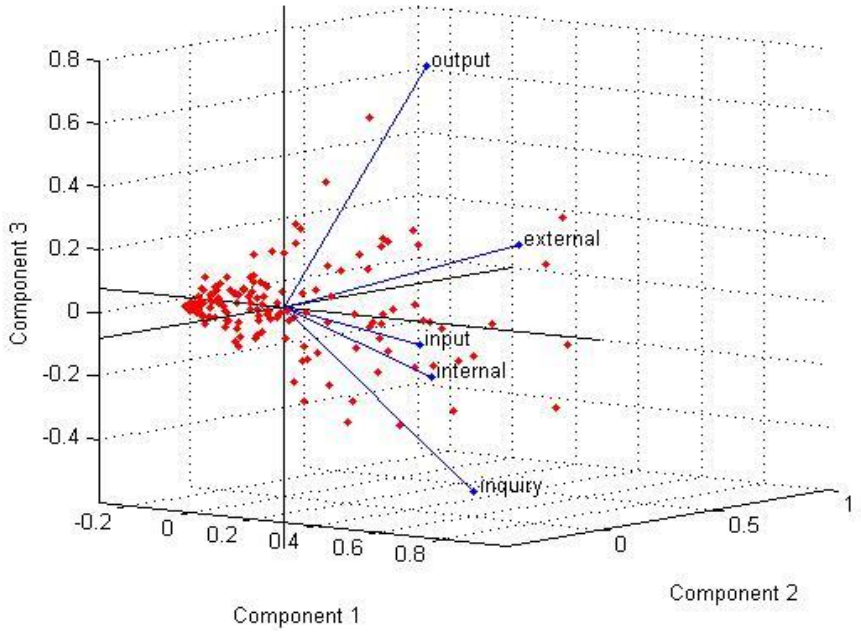

Figure 5. 3D Biplot of Three Principle Components

Biplots helps to visualize the contribution of variable to the principle component. A biplot represents each variable with a vector and the length and direction of the vector indicates the contribution of each variable to the principle components. Figure 4 shows the biplot of principle component 1 against 2 and 2 against 3.

All of the FP elements (EI, EO, EQ, ILF and ELF) for principle component 1 are positive (Table 4); therefore, all the vectors are in the right half of the first plot. For the principle component 2, EI, EO and EQ are negative making the vectors below the axis. The length of vectors on the principle component 1 is large except for ELF which has a strong contribution on principle component 2. The EO, EI and ILF have negative loading on principle component 21 meaning that they are inversely related. For principle component 2 values for EI, EO and ILF are negative keeping them in the right side of the plot.

A 3D plot shown in Figure 5 of three principle components has more interesting observations; principle component 3 has positive values for EI, EQ and ILF FP's and negative values for Output and External. When both the vectors from principle component 2 and 3 are combined it shows that EI, EO and ILF have negative loading on principle component 2 and EQ and ELF have positive loading. It is interesting to note that EI, EO and ILF FP's tends to be on the same side of the biplot, which confirms that our previous discussion that EI, EO and ILF can be combined together to form a new component.

5. Conclusions

Principle component analysis for function point elements is presented. The analysis suggests that the function point elements EI, EO and ILF have heavy loading on the first principle component and hence they can be combined together. It indicates that these function point elements capture the same attribute of a software project. Although, the EQ function point element has a larger loading on first principle component but it has even

larger loading on the third principle component; therefore, EQ function point element alone forms the third principle component. Therefore, five function elements can be combined to form three function point elements: i.e., Data Interface, ELF and EQ function point.

References

[1] C. J. Lokan, "An Empirical Study of the Correlation between Function Point Elements", Software Metrics Symposium, 1999 Sixth International Proceedings, pp. 200-206.
[2] D. R. Jeffery and J. Stathis, "Function Point Sizing: Structures, validity and applications", Journal of Empirical Software Engineering, vol. 11-30, **(1996)**.
[3] B. Kitchenham and K. Kansala, "Intr-item correlations among function points", Proceedings 15th International Conference on Software Engineering, IEEE, **(1993)** May, pp. 477-480.
[4] International Software Benchmarking Standards Group. Release 9.
[5] IFPUG. Function Point Counting Practices Manual release 4.2.

Author

Dr. Masood Uzzafer has 20 years of experience of which 10 years in the US high-tech industry and 10 years in the research and academics. Dr. Masood has worked for companies like Allied-Signal Aero-Space in Florida and Philips Semiconductors Silicon Valley California. His industrial experience encompasses designing and developing software for variety of products including Radar Systems, Digital TV, cable modem and Multi-Media processors.

His research has multiple streams including software engineering, project management and risk management. Currently, he is leading a research project to calibrate the Framingham Cardiovascular risk prediction model and Framingham Coronary heart disease risk prediction model for the population in Dubai.