Good morning, ladies and gentlemen:

Today I will talk about my article called “Principal component analysis of the influencing factors of phase cost”. We often meet a question which is how much does a project cost, of course, we have many methods, such as the linear regression method, neural-net algorithms, and so on. On this occasion I will use principal component analysis to forecast a project’s cost.

Today I will introduce my article using the three categories: the concept of Principal component analysis, case of Principal component analysis, and forecast and results analysis.

Let’s begin from the concept of Principal component analysis.

Generally, in the study of practical problems many related variables are always encountered, which can provide certain information, but their importance is different. In many cases, a certain correlation exists between the variables, which results to a certain extent an overlap but these overlaps do provide information. By processing these variables, with the smaller number of uncorrelated variables, it is possible to reflect most of the information provided by the original variables, and then by analyzing these new variables to be able to solve the overlapping information. This method of reducing the variable dimension overlap and to deal with the problem of high-dimensional data is the basic idea of the principal component analysis. Based on the description of n-dimensional data, a simple and objective evaluation will reflect the maximum difference between the data and put forward an improvement using the scientific and effective methodology from principal component analysis.

To establish a more accurate cost forecast model, the principal component analysis should be used to avoid any similar relationships between the cost implication factors, and that is basic idea of Principal component analysis.

Now let's look at the method of Principal component analysis.

From the supplied analysis, it can be known that each of the main components of x are equivalent to a respective feature value from the covariance matrix Σ and the corresponding orthogonal unitized feature vector. The linear combination of descending and corresponding orthogonal unit characterizations are respectively for the first, second, and so on until the p principal component, and the variance of the respective principal component is equal to the corresponding feature value.

Therefore, the principal component analysis is that dismantling the total variance of the original variables $x_1, x_2, \ldots x_p$ into the total variance of uncorrelated variables $z_1, z_2, \ldots z_p$, and this is the sum of characteristic values of the $x_1, x_2, \ldots x_p$ covariance matrix.

The contribution rate of the first principal component indicates the largest proportion of the comprehensive information contained within the original variables, while the remainder decreases in order of size from the original variables. In practical applications, it is usually selected $m < p$, so that the principal component of a cumulative contribution rate can achieve a higher proportion (e.g., 80% to 90%).

So m principal components instead of the original variables are used not only to reduce the number of variable dimensions, but also to retain the vast majority of the information in the original variables.

The processing of Principal component analysis has mainly two steps. Step one is to normalize data and make a correlation coefficient matrix, Step two is to determine principal component by characteristic value, from formula 1 to formula 5, we get the equation for Principal component analysis.

Now let's look at the case.

By using previously known facts, some factors that affect the cost amount are the research capacity factor, research target, research content, key technologies, technical approach, hard achievements, soft achievements, research cycle, personnel, test conditions, hardware coefficient, current research evaluation and so on.

Table 1 shows the quantized values of the influencing factors in the previous phase. “Total cost” in Table 1 is decided by the other columns, the project’s attributes are divided into three aspects, such as the project theoretical basis or method, the concrete implementation of the project and the results of the project.

This synthesis data affects the cost of the project from a multiple of angles comprehensively. Based on multidimensional information, the project’s cost influencing factors can be found, which reflects most of the differences, and the maximum effective improvements.

Table 2 shows correlation coefficient matrix. It gives the influencing factors and the correlation coefficient of the total cost. The first line shows the larger correlation coefficients of the five independent variables such as research targets, key technologies, test conditions, hardware coefficient and soft achievements.

A Negative number indicates that the variables and the dependent variables have a negative correlation. Taking into account the quantified research targets scoring, based on the project's technical maturity, the technical maturity is a comprehensive reflection of research contents, results, and technical difficulty; The quantified test conditions scoring, is based on the rate of test cost from the total cost, reflecting the difficulty of the implementation process of the project. Other lines reflect the hidden relationship of the other 12 corresponding variables.

Table 3 shows total variance explained which reflects the number of principal components，Table 4 shows eigenvectors of component factors. Principal component analysis is to reassemble each costing variable into several new variables, and departmentalize each component differently based on the proportions of the original variables to the new variables. Every principal component is quantified by the percentage of its eigenvalue to total eigenvalue, and arranged according to size with a value greater than 1.

To bring eigenvectors f_i in Table 4 into the formula 3, the five principal components and their cumulative contribution rate as shown in Table 5.

Then we get formula 6.

Now we introduce the last part of this article: forecast and results analysis.

In the same way as previously, regarding the issue affecting the current cost amount, a graph can be drawn showing the appropriate principal component variables. Table 6 shows the principal component variables of the current project.

Putting the values of table 6 Into equation 6, calculating the current cost of each programe as shown in Table 7 which is called forecasting values V.S. true values.

As found in the analysis of Table 4 and Table 5, the cumulative contribution rate of the first principal component is 32.64%, mainly reflecting research capacity coefficient (negative), research objectives, soft achievement, research cycle and personal (negative) to the target results; The first and second principal components accumulative contribution rate becomes 56.73%, reflecting the research content, hard achievement and issue of evaluation for target results; The accumulative contribution rate of 77.22% from the first three principal components, reflects key technologies (negative) and hardware coefficient of the target results; The accumulative contribution rate of 87.66% from the first four principal components, reflects the negative impact of the test

situation to target results; The accumulative contribution from the first five principal components becomes 96.56%, reflecting the technical approach impact on the target results. The first five cumulative contribution rates is over 96%, the results of the principal component analysis is to establish an orthonormal base in the five-dimensional space, and the direction of the first principal component is research capacity coefficient, research objectives, soft achievement, research cycle and personal. From the remaining principal component data, the loss of information does not exceed 4%. Therefore, a twelve-dimensional problem can be transformed into a five-dimensional problem to research by using the principal component analysis, and contains most of the information of the original data. The principal component analysis has reduced the dimension and overlapping problems and a prediction error of between 9.58% and 17.2% for the current cost forecasting but this is within the permitted tolerance range.

Summary: In this paper, principal component analysis method has been used to analyze the twelve influencing factors. These factors are constructed from five principal component factors which have an accumulative contribution rate of 96.56% on the original data, and fit into the cost regression equation. Secondly, the principal component analysis of the twelve influencing factors for the next phase has got another five principal component factors. Lastly, we brought all of those principal components into the upper regression equation, and calculated the cost amount of the next stage of each program. Using this method, the size of the influencing factors had been reduced effectively, and a scientific and effective regression equation had been established in which a smaller number of unrelated new variables were reflected from all of the information provided by the original variables.