A Three-Phase Approach for R&D Project Scheduling

C. Çapa[1], G. Ulusoy[1], K. Kiliç[1]
[1]Manufacturing Systems Engineering Program, Faculty of Engineering and Natural Sciences, Sabanci University, İstanbul, TURKEY
(canancapa@sabanciuniv.edu, gunduz@sabanciuniv.edu, kkilic@sabanciuniv.edu)

Abstract - **During project execution unforeseen events which disrupt plans and budgets arise and yield higher costs due to missed due dates and deadlines, resource idleness, higher work-in-process inventory and increased system nervousness due to frequent rescheduling. In this study, we consider the resource constrained multi-project scheduling problem with multi-skilled resources in a stochastic and dynamic environment for modeling the scheduling of the R&D projects of a leading home appliances company in Turkey. For this purpose a three-phase model is developed. Phase I, which is referred to as the risk and deviation analysis phase, aims at predicting the resource usage deviation level of projects and the resource usage deviation level of the activities of the projects. Phase II and Phase III are the proactive and reactive scheduling modules, respectively.**

Keywords - **Data Mining, R&D, Project Scheduling**

I. INTRODUCTION

In all sectors of the economy, an appreciable amount of work is accomplished through managing projects. A group of organizations, called project based organizations, such as consulting firms and R&D organizations perform almost all their work through projects and operate in general on more than one project simultaneously. These projects are interrelated since the same pool of resources is employed to execute them. During project execution, unforeseen events which disrupt plans and budgets arise and yield higher costs due to missed due dates and deadlines, resource idleness, higher work-in-process inventory, and increased system nervousness due to frequent rescheduling. Hence, there is a significant requirement for risk integrated robust project scheduling techniques making risk analysis step an essential step for project scheduling. The goal of risk analysis is to generate insights into the risk profile of a project and to use these insights in order to drive the risk response process [1]. In literature, risk analysis process is divided into four main subprocesses, namely, risk identification, risk prioritization, quantitative risk assessment and quantitative risk evaluation. Hubbard[2] states that good risk management requires a risk analysis process that is scientifically sound and that is supported by quantitative techniques. A wide body of knowledge on quantitative techniques has been accumulated over the last two decades. Monte Carlo Simulation is the predominant quantitative risk evaluation technique both in practice and literature. With the risk information on hand, proactive scheduling aims at the construction of a protected initial schedule (baseline or predictive schedule) that anticipates possible future disruptions by exploiting statistical knowledge of uncertainties that have been detected and analyzed in the project planning phase. A change in the starting times of such activities could lead to infeasibilities at the organizational level or penalties in the form of higher costs. A possible measure for the deviation between the baseline schedule and the realized schedule is the *weighted instability cost*. It can be calculated by taking the sum of the expected weighted absolute deviations between the planned and the actually realized activity starting times. The weight w_i assigned to each activity i, reflects the activity's importance of starting it at its planned starting time in the baseline schedule. Minimizing instability then means looking for a schedule, which is able to accommodate disruptions without too much change in the activity starting times.

The problem on hand is the resource constrained multi-project scheduling problem (RCMPSP) with multi-skilled resources in a stochastic and dynamic environment present in the R&D department of a leading home appliances company in Turkey for scheduling the R&D projects. A three-phase model will be developed incorporating data mining and project scheduling techniques to schedule these R&D projects.

The literature on risk integrated proactive scheduling is scarce. Most of the research approaches on project scheduling involving risk do not model risks explicitly, but try to evaluate the risk of schedule and/or budget overruns using stochastic models for activity durations and/or costs. Jaafari[3], Shatteman et. al.[4], and Herroelen[5] are notable examples of the risk integrated proactive project scheduling methodologies. Still, there is no study on risk integrated multi-objective proactive project scheduling. The most common objectives in robust project scheduling are quality robustness and solution robustness[6]. Note that the quality robustness refers to the stability of the makespan over all projects whereas the solution robustness refers to the stability of the activity starting times.

In this paper, our focus will be limited to the Phase I of the three-phase approach proposed as a solution to the RCMPSP problem. For that purpose, risk tables of randomly selected 40 R&D projects in the firm are constructed and analyzed. The lack of some required components of the risk data precludes the implementation

of risk-driven approach proposed in the literature for robust project scheduling. As an alternative to consider the risk-based deviations in project scheduling, we propose making use of data mining tools. With the help of the feature selection, clustering, and classification -the most known data mining techniques- the important factors on the risk and deviation level of projects are identified. This information will be used later during the proactive project scheduling phase and whenever needed will trigger the reactive project scheduling phase on a disrupted project plan.

In the following, the only resource considered is the various types of human resource. This is due to the relatively high importance of human resource as well as the relatively unrestricted availability of other resources such as laboratory equipment in the problem that is dealt. In order to consider the human resource usage deviations of the projects as a risk measure, in the proposed model, the projects are classified into four groups making use of the feature selection, clustering and classification analysis. We kindly suggest the interested readers to refer to Tan et. al.[7], and Du[8] for detailed information on the data mining tools that are used. After project deviation level prediction, activities are classified into five groups and percentage human resource usage deviation assignment procedure is developed to predict the deviation levels of the activities.

II. RISK AND DEVIATION ANALYSIS PHASE

Phase I, which will be referred to as the risk and deviation analysis phase, is comprised of two steps: (i) Risk and Deviation Analysis of Projects and (ii) Activity Deviation Assignment Step. It aims at predicting the deviation level of the projects and the deviation level of the activities of the projects.

A. Step I: Risk and Deviation Analysis of Projects

The objective of the first step of Phase I, is establishing a classification model based on real data collected from a leading home appliances company in Turkey, in order to classify the R&D projects with respect to their percentage human resource usage deviation from mean. Thus, by using the classification model, in the planning phase that is to say before the project actually starts, predicting its human resource deviation level can be possible, and the needed precautions can be taken. Furthermore, this information will be used in the second step of the proposed methodology in order to obtain the percentage human resource deviation distributions of *activities*. The resulting human resource deviation distributions of the activities later will be used in Phase II when assigning start and finish times to projects and their activities.

For this purpose, each R&D project in the data set is labeled as NHD (*negative high deviation*), NLD (*negative low deviation*), PLD (*positive low deviation*) and PHD (*positive high deviation*) based on threshold levels which are determined by consulting the experts and the projects' percentage human resource deviation realized. Next a feature selection process is applied to the data in order to determine the relevant features. The resulting data is used to construct the classification model. Note that in the analysis an open source data mining tool, namely WEKA developed by Hall et. al. [9] is utilized.

Data Set

After several interviews with the project managers of the firm, the factors that might affect project risk levels and cause time overruns are determined. The input features determined after these interviews with the data types and ranges are presented in Table 1.

TABLE 1
Input Features

FEATURE_ID	FEATURE_NAME	TYPE	MINIMUM	MAXIMUM
FA1	Existence of the technology family "Liquid Dynamics"	(binary)	0	1
FA2	Existence of the technology family "Material Science"	(binary)	0	1
FA3	Existence of the technology family "Thermodynamics"	(binary)	0	1
FA3	Existence of the technology family "Cleaning"	(binary)	0	1
FA5	Existence of the technology family "Vibration and Acoustics"	(binary)	0	1
FA6	Existence of the technology family "Structural Design"	(binary)	0	1
FA7	Existence of the technology family "Power Electronics"	(binary)	0	1
FA8	Existence of the technology family "Electronic Assessment"	(binary)	0	1
FA9	Number of collobrative internal plants	(integer)	0	5
F1	Number of Technology families involved in the project	(integer)	2	9
F2	Required size of project team in numbers	(integer)	5	27
F3	Number of required equipment and machine type	(integer)	0	5
F4	Number of collobrations	(integer)	0	3
F5	First Usage of infrastructure	(binary)	0	1
F6	Existence of similar projects worked on before	(binary)	0	1
F7	Planned man-months needed	(double)	6.1	88.69
F8	Planned equipment-months needed	(double)	0	119,97
F9	Expected cost of the project	(integer)	32064	506825
F10	Technology maturity of the Project	(integer)	1	25
F11	Position of the project in the r&D-R&d spectrum	(integer)	1	3

In the analysis two types of output are considered, i.e., Numeric Output and Nominal Output. The numeric output is basically the percentage human resource usage deviations. On the other hand the nominal output is determined by the application of a simple K-Means clustering algorithm developed by MacQueen [10] to the numeric output. Based on the resulting clusters, four deviation levels (i.e., NHD, NLD, PLD and PHD) are determined and each project is labeled accordingly. As a result a data set with 20 input features and two output features is obtained.

Data Preprocessing: Feature Selection Analysis

Not only missing some of the significant input features but also existence of abundant number of irrelevant features makes it difficult (if not impossible) to establish the relation between the inputs and the output. Therefore, feature subset selection analysis is an essential step in data mining process and directly influences the classification performance.

In the analysis, 20 input features and the numeric output, i.e., the percentage human resource deviation of the projects is utilized. Various different filtering and wrapper algorithms with *n-fold* cross validation is utilized. Note that, different folds (i.e., different training

and test combinations) yield different subsets of significant inputs hence a threshold value of 70% is set in order to make a final decision for inclusion of a feature for the further analysis in the case of wrappers. On the other hand, for the filtering techniques 0.007 +-0.004 are assumed as threshold values for the merits in the final decision.

As a result of the analysis four different feature subsets are determined as significant, namely, {F1, F4, F5, F6, F10}, {F1, F2, F4, F5, F6}, {FA1, FA6, F4, F5, F6} and {FA1, FA4, FA8, F5, F6}. In order to evaluate the influence of the feature subset selection stage to the classification performance two extra feature sets are also included in the further analysis, i.e., one set with all of the features proposed by the managers, and the second set which consists of 11 features namely {F1, …, F11}.

Classification Analysis

For each one of the six feature sets that was determined as the result of the feature subset selection analysis two different classification analysis were conducted; one with the numerical output and one with the nominal output. Note that for the numerical output case various well known classification algorithms such as J48 Decision Tree or Naïve Bayes were not applicable and limited to only regression like algorithms. Therefore the nominal class labels were determined by utilizing K-Means Clustering algorithm.

The resulting thresholds that were used to label the projects with the four class labels (NHD, NLD, PLD and NLD) were determined as -0.20, 0.00 and 0.20. That is to say, the projects having percentage human resource deviation level less than or equal to -0.20 were labeled as NHD, the projects having percentage human resource deviation between -0.20 and 0.00 were labeled as NLD, the projects having percentage human resource deviation between 0.00 and 0.20 were labeled as PLD and the rest were labeled as PHD.

Classification Analysis with Numeric Output

As stated earlier, in the classification analysis with numeric output, only regression based classification algorithms were applied, namely, Linear Regression, Least Median Squared Linear Regression, Pace Regression and M5P Algorithm.

Table 2 tabulates the predictive performance of these algorithms based on various metrics, namely, Count of Exact Class Matches (True Count), Accuracy Rate and the Mean Squared Error (MSE), for each of the six input feature sets determined as the result of the Data Preprocessing Stage. Note that, for the numerical output analysis the *True Counts* are calculated based on the intervals determined as the labels of the numeric output. In order to calculate the MSE of classification methods, the labels of the projects are converted into numbers. The numbers 1, 2, 3, and 4 are used for the labels "NHD", "NLD", "PLD", and "PHD", respectively. In this manner, the error is simply the difference between the corresponding number of prediction and corresponding number of actual label.

In addition to the performance metrics, Table 2 also presents the used features in the class label assignment procedure of the corresponding classification method for each feature subset used in the analysis.

TABLE 2
Classification Results for the Numeric Output

RESULTS FOR HR DEVIATION							
	INPUT FEATURES						
PERFORMANCE	11 Feature	20 Feature	F1,F4,F5,F6,F10	F1,F2,F4,F5,F6	FA1,FA6,F4,F5,F6	FA1,FA4,FA8,F5,F6	CLASSIFICATION METHODS
True Count	21	9	17	16	17	12	Linear Regression
Accuracy Rate	0,488372093	0,209302326	0,395348837	0,372093023	0,395348837	0,279069767	
MSE	54	94	57	63	64	66	
Selected Features	F2,F4,F5,F10	FA1,FA4,FA6,F2,F4,F7,F9,F10	F1,F5,F10	F2,F4	FA1,FA6,F4	FA1,FA4	
True Count	18	10	18	17	21	24	Least Median Squared LR
Accuracy Rate	0,395348837	0,23255814	0,395348837	0,372093023	0,465116279	0,558139535	
MSE	34	93	37	47	42	34	
Selected Features	ALL	ALL	ALL	ALL	ALL	ALL	
True Count	17	21	22	16	18	22	Pace Regression
Accuracy Rate	0,395348837	0,488372093	0,511627907	0,372093023	0,418604651	0,511627907	
MSE	44	37	36	45	42	39	
Selected Features	F1,F2,F4,F5,F10,F11	FA1,FA4,FA6,F2,F4,F10	F1,F4,F5,F10	ALL	ALL	ALL	
True Count	20	24	19	16	17	12	M5P
Accuracy Rate	0,465116279	0,558139535	0,441860465	0,372093023	0,395348837	0,279069767	
MSE	35	49	36	45	64	55	
Selected Features	F2,F4,F7,F10,F11	FA6,F2,F4,F10	F1,F4,F5,F10	F2,F4	FA1,FA6,F4	FA1,FA4	

Table 2 shows that the best true count values, accuracy rates and MSE values are obtained with the Pace Regression classification method. Besides being good, the true count values, accuracy rates and MSE values are more robust among the input feature subsets.

Classification Analysis with Nominal Output

The classification algorithms applied to the data set with nominal output were J48 Decision Tree classification method and Naive Bayes classification method. Again the same predictive performance metrics are used. The results for the data set with nominal output are presented in Table 3.

TABLE 3
Classification Results for the Nominal Output Obtained from Simple K-Means Algorithm

RESULTS FOR K-MEANS (4)							
	INPUT FEATURES						CLASSIFICATION
PERFORMANCE	11 Feature	20 Feature	F1,F4,F5,F6,F10	F1,F2,F4,F5,F6	FA1,FA6,F4,F5,F6	FA1,FA4,FA8,F5,F6	METHODS
True Count	37	37	28	29	27	22	J48 DECIDION TREE
Accuracy Rate	0,837209302	0,837209302	0,651162791	0,674418605	0,627906977	0,488372093	
MSE	12	20	30	20	41	45	
Selected Features	F2,F3,F4,F5,F9,F10	FA1,FA3,FA4,FA5,FA6,F3, F4,F5,F6,F8,F11	F1,F4,F5,F10	F1,F2,F4,F5	FA1,F4,F5	FA1,FA4,FA8,F5	
True Count	26	30	23	23	25	22	NAIVE BAYES
Accuracy Rate	0,604651163	0,674418605	0,534883721	0,534883721	0,558139535	0,488372093	
MSE	29	22	44	38	33	45	

Table 3 demonstrates that the best true count values, accuracy rates and MSE values are obtained with J48 Decision Tree classification method. Besides being good, the true count values, accuracy rates and MSE values are more robust among the input feature subsets.

Comparisons of Classification Approaches

One way of comparing the classification approaches other than comparing accuracy performances is using average variability of each classification approach among the other approaches. This variability attribute is specific for each feature subset and classification method combination and can be calculated using the label numbers associated with the projects and in the same manner that was adopted while calculating MSE for the nominal analysis. The variability of a project for a feature subset and classification method is simply the sum of the squared difference between the corresponding label number of the result obtained from the combination in question and corresponding number labels of the results obtained from the other feature subsets and classification methods. The average variability is obtained summing these variability values of the projects among 43 projects and simply taking the average. Since the number of combinations for each output type is different (due to number of algorithms used in the analysis for the corresponding output type) in order to make the comparisons consistent we have divided the average variability values to the number of combinations. In this way, we were able to compare the feature subset and classification method combinations. The average variability values of the feature subset and classification method combinations for the prediction of percentage human resource deviation levels of projects as NHD, NLD, PLD and PHD are demonstrated in Table 4.

TABLE 4
Average Variability Results of the Classification Approaches

Output:	PERCENTAGE HUMAN RESOURCE DEVIATON		LABELS OBTAINED BY APPLYING SIMPLE K-MEANS CLUSTERING ALGORITHM TO THE PERCENTAGE HUMAN RESOURCE DEVIATON	
	Classification Method	Average Variability	Classification Method	Average Variability
CLASSIFICATION USING 11 FEATURE	Linear Regression	0,94		
	Least Median Squared LR	0,94	J48 Decision Tree Method	0,64
	Pace Regression	0,80		
	MP5	0,73	Naive Bayes Method	0,86
CLASSIFICATION USING 20 FEATURE	Linear Regression	1,76		
	Least Median Squared LR	1,89	J48 Decision Tree Method	0,71
	Pace Regression	0,73		
	MP5	1,00	Naive Bayes Method	0,86
CLASSIFICATION USING F1, F4,F5,F6,F10	Linear Regression	1,14		
	Least Median Squared LR	0,78	J48 Decision Tree Method	0,53
	Pace Regression	1,02		
	MP5	0,98	Naive Bayes Method	0,65
CLASSIFICATION USING F1, F2 F4,F5,F6	Linear Regression	1,01		
	Least Median Squared LR	0,78	J48 Decision Tree Method	0,66
	Pace Regression	0,90		
	MP5	1,01	Naive Bayes Method	0,59
CLASSIFICATION USING FA1, FA6 F4,F5,F6	Linear Regression	1,80		
	Least Median Squared LR	0,99	J48 Decision Tree Method	1,02
	Pace Regression	0,97		
	MP5	1,80	Naive Bayes Method	0,52
CLASSIFICATION USING FA1, FA4,FA8,F5,F6	Linear Regression	1,26		
	Least Median Squared LR	0,76	J48 Decision Tree Method	0,69
	Pace Regression	0,88		
	MP5	1,48	Naive Bayes Method	0,70

Table 4 reveals that among the classification approaches the feature subset of F1,F4,F5,F6,F10 and the classification method of J48 Decision Tree Method and the feature subset of FA1,FA6,F4,F5,F6 and the Naive Bayes classification method combinations give the lowest average variability results. In parallel with the accuracy results, using the labels obtained by applying Simple K-Means clustering algorithm to the percentage human resource deviations of projects yields better results than the results using the percentage human resource deviation of projects.

Another consideration we need to take into account while comparing classification approaches is the interpretability of the results. Since the Naive Bayes classification method is a black box only giving the classes of the given projects, it is hard to convince the decision-maker about the reliability of the method. Decision tree based algorithms are better for interpretability since they also give a tree as a rule of classification to the decision maker for the classification of the newly added data point (new project in our case). When selecting a classification approach the other consideration is the number of features used in the classification and their ease of obtainment.

B. Step II: Activity Deviation Assignment Procedure

In Step I, we have developed a model to predict the percentage human resource deviation level of a newly arrived project based on its various input features. Using this information, in Step II, we also developed a model to predict the percentage human resource deviation of the activities of this newly arrived project. Since we are dealing with R&D projects and the activities of R&D projects are unique and the work content is characteristic among all the activities, in order to obtain sufficiently large amount of data for a valid percentage human resource activity deviation distribution we have grouped the activities of projects in six activity classes. The classification of the activities was based on the work contents and the density of required resource types of the activities. The list of activity classes are as follows:

- Meeting and Reporting Activity Class
- Design Modeling and Visualizing Activity Class
- Test, Measurement and Analysis Activity Class
- Prototyping/Production Activity Class
- Literature and Patent Search Activity Class
- Other Activity Class

The aim of Step II of Phase I is to obtain percentage human resource deviation distributions for each project deviation class - activity class combination. Using the model developed in Step I, for a newly arrived project we predict its percentage human resource deviation class and for each activity class in the corresponding project; using the percentage human resource deviations of already completed activities in the associated activity class belonging the predicted project deviation class we form the human resource deviation distribution of that project deviation class - activity class combinations. Table 5 shows the frequency information used to obtain the deviation distribution for NHD Project Class - Test, Measurement and Analysis Activity Class combination and Figure 1 depicts the corresponding deviation distribution.

TABLE 5
Frequency and Probability Information for the NHD-Test, Measurement and Analysis Class Combination

Activity Class	Count of Activities in NHD Project Deviation Class	Percentage Deviation Range	Count of Activities	Probability of Being in that Range
Test Measurement and Analysis	102	(-1)-(-0,67)	23	22,55%
		(-0,67)-(-0,33)	30	29,41%
		(-0,33)-(0)	23	22,55%
		0-0,33	14	13,73%
		0,33-0,67	6	5,88%
		0,67-1	1	0,98%
		1-1,33	3	2,94%
		1,33-1,66	0	0,00%
		1,66-2	2	1,96%
		SUM	102	

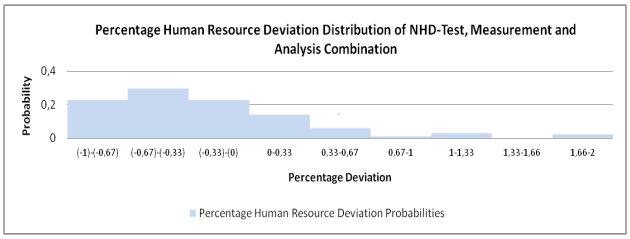

Fig. 1. Distribution the NHD- Test, Measurement and Analysis Combination

The percentage human resource usage deviation distributions of the activities belonging each activity class - project deviation class combinations are obtained following the same procedure and percentage human resource deviations are assigned all the activities belonging to the existing project set of sized-43 in order to compare the actual percentage human resource deviations with the percentage deviations assigned using the procedure we have just suggested. The results of comparisons are shown in Table 6.

TABLE 6
Performance of Proposed Model in Step II

Number of Activities	1008	Number of Activities Having Negative Deviation	628	Number of Activities Having Positive Deviation	380	
	ASSIGNMENT 1	ASSIGNMENT 2	ASSIGNMENT 3	ASSIGNMENT 4	ASSIGNMENT 5	AVERAGE
Total Negative Match Count	369	387	392	360	369	375,4
Total Positive Match Count	147	143	160	160	140	150
Total Match Count	516	530	552	520	509	525,4
Negative Match Probability	58,76%	61,62%	62,42%	57,32%	58,76%	59,78%
Positive Match Probability	38,68%	37,63%	42,11%	42,11%	36,84%	39,47%
Match Probability	51,19%	52,58%	54,76%	51,59%	50,50%	52,12%

Table 6 shows that using the procedure that we suggested, on the average with the probability of 52 % we are able to make correct predictions on the percentage deviations of activities. Our predictions are much better to predict the negative percentage deviations of activities than the positive percentage activity deviations of activities. This correct prediction rates cannot be underrated since the correct prediction rates of activity deviation levels even if the deviation level of the projects are exactly known in advance very similar with the results presented above. Table 7 shows the results of the activity deviation assignment procedure when project deviation labels are exactly given.

TABLE 7
Activity Deviation Assignment Results for the Actual Project Deviation Classes

ACTUAL STATISTICS	Number of Activities	1008	Number of Activities Having Negative Deviation	628	Number of Activities Having Positive Deviation	380	
		ASSIGNMENT 1	ASSIGNMENT 2	ASSIGNMENT 3	ASSIGNMENT 4	ASSIGNMENT 5	AVERAGE
ACTUAL PROJECT LABELS	Total Negative Match Count	377	355	373	365	353	364,6
	Total Positive Match Count	168	161	183	153	145	162
	Total Match Count	545	516	556	518	498	526,6
	Negative Match Probability	60,03%	56,53%	59,39%	58,12%	56,21%	58,06%
	Positive Match Probability	44,21%	42,37%	48,16%	40,26%	38,16%	42,63%
	Match Probability	54,07%	51,19%	55,16%	51,39%	49,40%	52,24%

III. CONCLUSION

In this study we have presented Phase I of the proposed three-phase approach for robust multi-objective R&D project scheduling. As a future direction it is planned to provide probabilistic results in Phase I for the prediction of newly arrived projects and a new activity deviation assignment procedure using this probabilistic results since it is expected that probabilistic results will yield better predictions for the percentage human resource usage deviations for each activity class. In this way we would not ignore the possibility of the newly arrived project's belonging to another project deviation class from predicted.

REFERENCES

[1] The Project Management Institute "*A Guide to the Project Management Body of Knowledge (PMBOK Guide)*" Project Management Institute, 2008.

[2] D. W. Hubbard, "*The Failure of Risk Management: Why It's Broken and How to Fix It.* " Wiley, 2009.

[3] A. Jaafari, "Management of risks, uncertainties and opportunities on projects: time for a fundamental shift." *International Journal of Project Management* 19, no. 2 ,pp.89-101, 2001.

[4] Schatteman, Damien, W. Herroelen, S. Van de Vonder, and A. Boone. "Methodology for integrated risk management and proactive scheduling of construction projects." *Journal of Construction Engineering and Management*134, no. 11, pp.885-893, 2008.

[5] W. Herroelen. "A risk integrated methodology for project planning under uncertainty", in: Sarin, S., Pulat, S., Uzsoy, R. (Ed.) *Festschrift for Salah Elmaghraby*, Springer Verlag, Berlin (to appear in 2013).

[6] W. Herroelen, and R. Leus. "Project scheduling under uncertainty: Survey and research potentials." *European Journal of Operational Research*, *165*(2), 289-306, 2005.

[7] P. N. Tan, M. Steinbach, and V. Kumar. "Introduction to Data Mining, Addison", Addison Wesley, 2006.

[8] H. Du. "*Data Mining Techniques and Applications: An Introduction*." Course Technology Cengage Learning, 2010.

[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. "The WEKA data mining software: an update." *ACM SIGKDD Explorations Newsletter*, *11*(1), pp. 10-18, 2010.

[10] J. MacQueen."Some methods for classification and analysis of multivariate observations". In *Proceedings of The Fifth Berkeley Symposium on Mathematical Statistics and Probability* Vol. 1, No. 281-297, p. 14, June, 1967.