Comparing Segmentations Based on Clustering Algorithms to Classify Patients of Intensive Care Units

Carlos Hernández, Jaime Castillo
Departamento de Procesos Industriales
Universidad Católica de Temuco
Temuco, Chile
carlos.hernandez.zavala@uct.cl, jcastill@uct.cl

Abstract

It is common to classify patients that arrive at an intensive care unit (ICU) by means of a classification based on gender, age and health record. However, it is interesting and helpful to take advantage of additional data to propose alternative patient segmentation that might help allocate more efficiently the existing infrastructure, supplies, medical staff. In this investigation several ICU patients' segmentations were implemented and compared. Different from a supervised task like classification, where datasets have to be a *priori* labelled to train and to test prediction models, clustering algorithms requires no labelling. Instead, data are grouped according to their degree of similarity. The research was carried out following a 4-phase methodology: analysis, design, development, and validation. During the analysis, a large database with record of the medical care received by patients at the ICU of a public hospital located in the south of Chile was preprocessed and analyzed. During the design, several datasets were prepared to conduct experiments. At this point, the advantages and disadvantages of different clustering algorithms were analyzed and compared, selecting Simple K-Means Algorithm (SKMA) and Expectation-Maximization Clustering (EMC) to proceed with the investigation. Whereas SKMA creates clusters of equal variance, EMC assumes a Gaussian distribution of data. The phase of development was carried out using the data mining software WEKA 3.9.6. To complete the investigation, four datasets of five, ten, fifteen, and twenty thousand ICU records were used. Since no target class was defined, the clustering was the result of applying the selected algorithms: EMC and SKMA. For both cases, different number of clusters (k) were required to establish a comparison. Results revealed clear differences in the outputs generated by each clustering algorithm. For instance, with 5 clusters (k=5) EMC distributes data in the following proportions: 19%, 15%, 10%, 43%, and 28%. With SKMA, instead, the proportion were: 44%, 13%, 10%m 9%, and 24%. In conclusion, the investigation showed that popular clustering algorithms such as EMC and SKMA can be used for segmenting not only consumers but also ICU patients according to criteria that are not easy to visualize with classical tools and techniques. An adequate segmentation can provide valuable information to help estimate the requirement of medical staff, supplies and infrastructure, and also to define specific healthcare services.

Keywords

Segmentation, Clustering Algorithm, Machine Learning, Intensive Care Unit, Unsupervised Leaning.

1. Introduction

The recent sanitary crisis has revealed the difficulties that public health systems face to cope with an unusual growing number of incoming patients. In many cases, the situation was already complicated before the COVID-19 pandemic and the crisis only made things even worse.

The unexpected increase in the demand for medical care catalyzed the creativity and forced the improvement in the management of the limited existing resources such as medical personnel, supplies and hospital infrastructure.

Along with new challenges, innovative technologies have also emerged to facility the work in health centers and hospitals. Some good examples of these widely accepted new trends are: the so-called P4 medicine, where P4 stands for Predictive, Personalized, Preventive and Participatory (Ruiz and Velásquez, 2023), the advances in clinical medicine based on artificial intelligence (Pan et al., 2022), and the increasing use of machine learning algorithms to detect and to diagnose diseases using existing patient data (Kejriwal and Rajagopalan, 2023).

According to Kotler (2001), organizations must identify the user segments to identify specific user needs. Although classical criteria have been used to classify patients, this work investigates an approach that combines aspects taken from the consumer segmentation used in marketing with clustering algorithms based on machine learning to classify ICU patients. Previous comparison of clustering algorithms has been already proposed with promising results (Jung et al., 2014)

1.1 Objective

To implement and to compare segmentations based on machine learning and clustering algorithms to classify patients of an intensive care unit (ICU).

2. Literature Review

2.1 Market segmentation

Market segmentation is usually referred as the process of classifying or grouping customers with different characteristics and behavior to implement more efficient marketing strategies and tactics (Kotler and Armstrong, 2006). Some of the most studied segmentation types in the literature are:

Behavioral	: brand loyalty, buyer journey stage, price sensitivity, purchasing style, etc.
Benefit	: customer service, quality, etc.
Demographic	: age, education level, gender, income, family members, status, religion, etc.
Geographic	: country, city, district, etc.
Psychographic	: hobbies, interests, lifestyle, etc.

2.2 Machine learning

Machine learning is usually referred as the branch of artificial intelligence (AI) that uses algorithms to find patterns and to learn from datasets through experience. There several types of machine learning algorithms: supervised, unsupervised, and reinforcement algorithms. In supervised learning, the training is carried out using labelled datasets. This means that the class or the value to be predicted is included in the dataset so it can be used for training. In the case of unsupervised learning, instead, the desired class is not known. The machine learning algorithms used in this work have been implemented with WEKA 3.8.5. (Witten et al., 2017)

2.3 Clustering algorithms

Clustering algorithms are used to discover patterns and to group data points (Figure 1). They are a particular case of machine learning algorithms employed to analyzed unlabeled datasets. Some of the most popular clustering algorithms are:

- Agglomerative hierarchical clustering
- Density-based spatial clustering
- Expectation-maximization clustering (EMC)
- Simple K-Means algorithm (SKMA)
- Mean-shift clustering

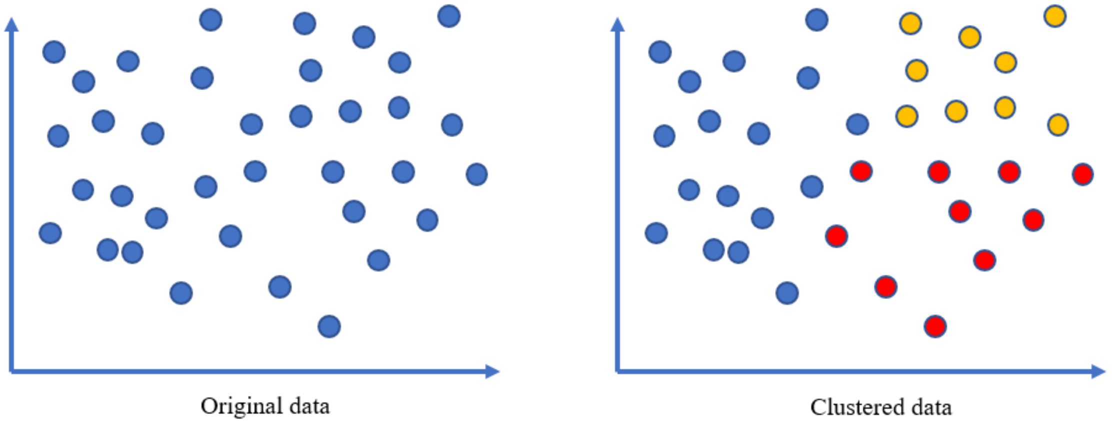

Figure 1. Clustering data

2.4 Simple k-means algorithm (SKMA)

K-means clustering is an unsupervised machine learning algorithm that is used to categorize unlabeled data. The algorithm works iteratively and assign every new instance to one of the existing k clusters. The classification criteria are based on the feature similarity of the instances.

2.4.1 Number of clusters

Finding the best clustering scheme might be useful when optimization is the goal. It can be found by means of varying k, distance measures, and clustering method. There are several methods to determine the optimal number of clusters. Some of the most common are: average silhouette method, elbow method, and gap statistic method.

2.5 Expectation-maximization clustering (EMC)

The expectation-maximization (EM) algorithm is an iterative procedure for the maximum likelihood estimate of a parametric distribution. A particular case of this algorithm is the parameter estimation of a Gaussian Mixture Model (GMM) when the generating Gaussian of each observation is unknown, commonly known as Expectation-Maximization Clustering (EMC) (Garriga et al., 2016; Jung et al., 2014).

3. Methods

This investigation is carried out following a 4-stage model: analysis, design, construction, and discussion (Figure 2).

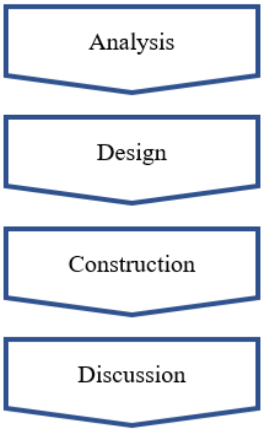

Figure 2. Four-phase model

3.1 Analysis

The investigation commenced with a complete review of data collected during 2020 by personnel of the intensive care unit (ICU) of one of the largest public hospitals located in the south of Chile. The databases contained approximately seventy thousand records corresponding the patients attended in a lapse of 12 months. Each record can be understood as a collection of fields with the data generated during the stay of a patients at ICU. Not all but a few of them were considered in this work (Table 1).

Table 1. Record selected fields

Field (attribute)	Description
Gender	The gender of a patient
City	The city, town or community where a patient resides.
Insurance type	Heath insurance policy.
Arrival mode	The principal means by which a patient arrives at ICU.
Source of admission	Place from where patients came.
Medical specialist	Doctor focused on a defined group of patients, diseases, skills, or philosophy.
Destination	The destination of the patient after leaving ICU.

Since clustering is an unsupervised learning task, no target class was ever defined. In other words, all selected fields were used to create the clusters. The type and the possible values of each filed is presented in Table 2.

Table 2. Field type and values

Field (attribute)	**Type**	**Number of values**
Gender	nominal	3
City	nominal	195
Insurance type	nominal	18
Arrival mode	nominal	11
Source of admission	nominal	23
Medical specialist	nominal	2
Destination	nominal	15

3.2 Design

From original database, which can be seen as a large matrix of 70,748 rows (records or instances) by 53 columns (fields or attributes), four subsets having five, ten, fifteen, and twenty thousand records each were prepared to. Only 7 out of the existing 53 were really used (Table 3).

Table 3. Dataset creation

Dataset	**Records (instances)**	**Fields (attributes)**
DS-05	5,000	7
DS-10	10,000	7
DS-15	15,000	7
DS-20	20,000	7

3.3 Construction

The objective is to implement and to compare patient segmentations based in clustering algorithms. As aforementioned, only the widely used SKMA and EMC algorithms were selected to proceed with the experimental wok.

Every dataset was split up to create two subsets in a proportion of 80% and 20% respectively. The first dataset contained records for training and test with 80% of data, whereas the second dataset contained records for validation only (Figure 3 and Table 4).

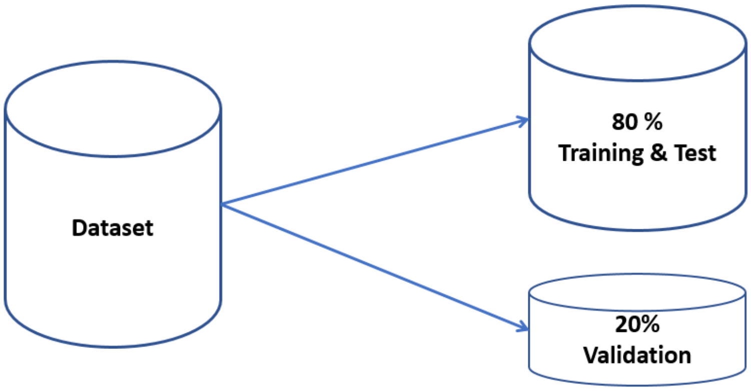

Figure 3. Dataset split up

Table 4. Datasets for training and test, and for validation

Dataset	**Training and test**	**Validation**
DS-05	4,000	1,000
DS-10	8,000	2,000
DS-15	12,000	3,000
DS-20	16,000	4,000

The validation dataset is used to confirm whether the proposed prediction model, prepared with the dataset for training and test, is able to generalized properly when predicting with unknown validation data.

3.4 Discussion

SKMA and EMC are two of the most popular clustering algorithms, both belong to a broader family usually called Gaussian mixture models. While SMKA requires the definition of k centroids and the iterations until certain degree of convergence to a local minimum is achieved, EMC is meant to solve some of the weaknesses of SKMA. Rather than focusing on the accuracy of the grouping, due to the nature of the patient segmentation, during the investigation the interest was set on the number of clusters and on their number of records.

4. Data Collection

Although some clustering algorithms are capable of finding the optimal number of clusters (k), the experimental work was carried out in such a way that it was possible to compare directly the output of EMC and SKMA when 2, 3, 4, and 5 cluster were required.

Clusters generated by using EMC with datasets DS-05, DS-10, DS-15, and DS-20 are presented in Table 5, Table 6, Table 7, and Table 8 respectively. Each of the following tables consolidates the output of the testing and test data, and that of the validation data.

Table 5. EMC with 4,000 training & test (TT) records and 1,000 validation (V) records

	k=2		**k=3**		**k=4**		**k=5**	
Cluster	TT	V	TT	V	TT	V	TT	V
1	3,355	817	744	196	951	247	944	247
2	645	183	2,686	651	1,862	436	1872	440
3			570	153	629	167	626	163
4					558	150	558	150
5							0	0

Table 6. EMC with 8,000 training & test (TT) records and 2,000 validation (V) records

	k=2		**k=3**		**k=4**		**k=5**	
Cluster	TT	V	TT	V	TT	V	TT	V
1	6,125	1,513	3,225	683	1,817	387	3,660	886
2	1,875	487	1,213	321	1,311	300	838	199
3			3,562	996	3,726	1,006	2,071	547
4					1,146	307	1,141	310
5							290	58

Table 7. EMC with 12,000 training & test (TT) records and 3,000 validation (V) records

	k=2		**k=3**		**k=4**		**k=5**	
Cluster	TT	V	TT	V	TT	V	TT	V
1	9,114	2,226	5,852	1,484	1,738	467	3,884	934
2	2,886	774	4,276	994	1,842	512	1,594	430

3			1,872	522	2,544	521	718	191
4					5,875	1,500	4,070	976
5							1,734	469

Table 8. EMC with 16,000 training & test (TT) records and 4,000 validation (V) records

	k=2		k=3		k=4		k=5	
Cluster	TT	V	TT	V	TT	V	TT	V
1	12,126	3,021	6,471	1,528	3,394	767	3,056	705
2	3,874	979	2,596	660	2,341	588	2,341	589
3			6,933	1,812	2,621	654	1,664	429
4					7,644	1,991	6,940	1,869
5							1,999	408

Clusters generated by using SKMA with datasets DS-05, DS-10, DS-15, and DS-20 are presented in Table 9, Table 10, Table 11, and Table 12 respectively. Each of the following tables consolidates the output of the testing and test data, and that of the validation data.

Table 9. SKMA with 4,000 training & test (TT) records and 1,000 validation (V) records

	k=2		k=3		k=4		k=5	
Cluster	TT	V	TT	V	TT	V	TT	V
1	3,328	819	2,039	419	1,501	386	1,116	290
2	672	181	673	186	689	189	755	199
3			1,288	323	1,281	316	1,224	312
4					529	109	572	121
5							333	78

Table 10. SKMA with 8,000 training & test (TT) records and 2,000 validation (V) records

	k=2		k=3		k=4		k=5	
Cluster	TT	V	TT	V	TT	V	TT	V
1	4,567	1,204	1,781	446	3,051	803	2,760	722
2	3,433	796	4,016	972	2,667	620	2,680	613
3			2,203	582	1,020	267	1,015	267
4					1,262	310	1,117	279
5							428	119

Table 11. SKMA with 12,000 training & test (TT) records and 3,000 validation (V) records

	k=2		k=3		k=4		k=5	
Cluster	TT	V	TT	V	TT	V	TT	V
1	10,007	2,599	8.026	2,054	4,744	1,218	4,592	1,181
2	1,993	401	1,884	369	1,884	369	1,848	361
3			2,090	577	2,198	617	2,188	616
4					3,174	796	3,153	792
5							219	50

Table 12. SKMA with 16,000 training & test (TT) records and 4,000 validation (V) records

	k=2		k=3		k=4		k=5	
Cluster	TT	V	TT	V	TT	V	TT	V
1	12,687	3,140	11,170	2,709	10,658	2,605	7,058	1,651
2	3,313	860	2,301	619	2,292	599	2,113	595
3			2,529	672	1,481	367	1,578	388

4					1,566	429	1,445	398
5							3,806	968

5. Results and Discussion

Finding the optimal number of clusters can be achieved with the help of optimization algorithms. Although interesting and challenging, in some cases it might be not applicable when the number of cluster is too large.

By means of applying an iterative method, SKMA can determine the number of clusters that minimized distance between each data point and its closest centroid. With EMC it is possible too. However, this approach can produce a larger number of clusters and some of home having just a few data points. In practice, having too many segments or cluster might not be helpful for decision making.

The optimization generated by EMC with the dataset DS-20 is presented in Table 13 where results are expressed in terms of the numbers of records and they corresponding percentage. In this case, 8 clusters were generated with EMC and to compare the performance of both algorithms, the same task was completed with SKMA.

The situation in both cases, EC and SKMA, is clearly different. While EMC generated one big cluster with almost 50% of data points and four with less than 5% of data points, SKMA generated more balanced clusters in terms of their size.

Table 13. Comparison EMC v/s SKMA with 16,000 training & test (TT) records and 4,000 validation (V) records

Cluster	**k=8**							
	EMC (records)		SKMA (records)		EMC (percentage)		SKMA (percentage)	
	TT	V	TT	V	TT	V	TT	V
1	7,252	1,945	5,042	1,197	45	49	32	30
2	190	23	1,867	524	1	1	12	13
3	1511	393	1,441	367	9	10	9	9
4	770	171	953	262	5	4	6	7
5	831	165	2,441	627	5	4	15	16
6	634	152	769	188	4	4	5	5
7	2,520	570	1,820	423	16	14	11	11
8	2,292	581	1,667	412	14	15	10	10

5.1 Numerical Results

Since the kind of segmentation generated by clustering algorithms is not obvious as other types such as demographics or demographic segmentation, understanding and characterization of the resulting clusters or patient segments demand additional labor since the interpretation is more complex.

A collection of complete side-by-side comparisons of the experimental results with dataset DS-05, DS-10, D-15, and DS-20 is presented in Table 14, Table 15, Table 16, and Table 17 respectively. All numbers are expressed as a percentage.

Table 14. Comparison (%) EMC v/s SKMA with 4,000 training & test (TT) records and 1,000 validation (V) records

Clu ster	**k=2**				**k=3**				**k=4**				**k=5**			
	EMC		SKMA		EMC		SKMA		EMC		SKMA		EMC		SKMA	
	TT	V	TT	V	TT	V	TT	V	TT	V	TT	V	TT	V	TT	V
1	84	82	83	82	19	20	51	49	24	25	38	39	24	25	28	29
2	16	18	17	18	67	65	17	19	47	44	17	19	47	44	19	20
3					14	15	32	32	16	16	32	32	14	16	31	31
4									14	14	13	11	14	15	14	12
5													0	0	8	8

Table 15. Comparison (%) EMC v/s SKMA with 8,000 training & test (TT) records and 2,000 validation (V) records

Cluster	**k=2**				**k=3**				**k=4**				**k=5**			
	EMC		SKMA		EMC		SKMA		EMC		SKMA		EMC		SKMA	
	TT	V	TT	V	TT	V	TT	V	TT	V	TT	V	TT	V	TT	V
1	77	76	57	60	40	34	22	22	23	19	38	40	46	44	35	36
2	23	24	43	40	15	15	50	49	16	15	33	31	10	10	34	31
3					45	45	28	29	47	50	13	13	26	27	13	13
4									14	15	16	16	14	16	14	14
5													4	3	5	6

Table 16. Comparison (%) EMC v/s SKMA with 12,000 training&test (TT) records and 3,000 validation (V) records

Cluster	**k=2**				**k=3**				**k=4**				**k=5**			
	EMC		SKMA		EMC		SKMA		EMC		SKMA		EMC		SKMA	
	TT	V	TT	V	TT	V	TT	V	TT	V	TT	V	TT	V	TT	V
1	76	74	83	87	49	49	67	68	14	16	40	41	32	31	38	39
2	24	26	17	13	36	33	16	12	15	17	16	12	13	14	15	12
3					16	17	17	19	21	17	18	21	6	6	18	21
4									49	50	26	27	34	33	26	26
5													14	16	2	2

Table 17. Comparison (%) EMC v/s SKMA with 16,000 training&test (TT) records and 4,000 validation (V) records

Cluster	**k=2**				**k=3**				**k=4**				**k=5**			
	EMC		SKMA		EMC		SKMA		EMC		SKMA		EMC		SKMA	
	TT	V	TT	V	TT	V	TT	V	TT	V	TT	V	TT	V	TT	V
1	76	76	79	78	40	38	70	68	21	19	67	65	19	18	44	41
2	24	24	21	22	16	17	14	15	15	15	14	15	15	15	13	15
3					43	45	16	17	16	16	9	9	10	11	10	10
4									48	50	10	11	43	47	9	10
5													12	10	24	24

6. Conclusion

Contrary to classical segmentation approaches based on geography or demography where clusters are rather evident, segmentations generated using clustering require a deeper analysis. This investigation presented the experimental results of applying two popular clustering algorithms to segment the patients received the ICU of n public hospital. A database containing twenty thousand records was used in this work.

Nowadays, clustering algorithms are widely used to identify patterns and classify unlabeled data by means of grouping similar data points in clusters that shares some degree of similarity. This application has been getting more attention in recent years because it can be of great help in decision making for different areas. Including that related to medical care and services.

Both SKMA and EMC are iterative optimization methods to cluster data points. Depending on the needs and the number of iteration is possible to determine the optimal number of clusters. Although interesting, it is not always practical. In some cases, when the number of cluster is too large, while few of them concentrate much of the data, other clusters have only a few data points. In this research the optimal number was found to be 8 clusters. With EMC the proportion of the resulting cluster showed that one cluster had 49% of data and four had less than 5% of data. SKMA, instead, generated a more balance set of clusters with proportion that fluctuated between 5% and 30% of data point.

An interesting fact is that the size of the larger datasets (DS-05, DS-10, DS-15, and DS-20) affected the proportions of the generated clusters. Is did not occur the same to the smallest dataset DS-05 (5,000 records), independently from the values of k.

A segmentation of ICU patients based on machine learning algorithms can lead to clusters that are not obvious. However, that unusual segmentations can help decision makers define better services and optimize the allocation of resources for each group of patients.
In conclusion, both clustering algorithms SKMA and EMC can be of great help when applied of the segmentation of patients. But, the interpretation of the clusters requires additional effort since is not evident.

References

Garriga J., Palmer J., Oltra A., and Bartumeus F., Expectation-Maximization Binary Clustering for Behavioural Annotation, PLoS ONE, vol. 11, no. 3, 2016.

Jung Y., Kang m., and Heo M., Clustering performance comparison using K-means and expectation maximization algorithms, Biotechnology & Biotechnological Equipment, vol. 28, pp. 44-48, 2014.

Kejriwal, S., and Rajagopalan, N., A technical review on machine learning-based prediction on COVID-19 diagnosis, Smart Innovation, Systems and Technologies, vol. 311, 2023.

Kotler, P., Dirección de Mercadotecnia. Análisis, Planeación, Implementación y Control, 8th edition, Pearson Education, 2001.

Kotler, P., Armstrong, G., Principles of marketing, Pearson Prentice Hall, 11th edition, 2006.

Pan, L.-C., Wu, X.-R., Lu, Y., Zhang, H.-Q., Zhou, Y.-L., Liu, X., Liu, S.-L., and Yan, Q.-Y., Artificial intelligence empowered digital health technologies in cancer survivorship care: A scoping review, Asia-Pacific Journal of Oncology Nursing, vol. 9, no. 12, 2022.

Ruiz, R. and Velásquez, J., Artificial intelligence for the future of medicine, Intelligent Systems Reference Library, vol. 229, 2023.

Witten, I., Frank, E., Hall, M., and Pal, C., Data Mining: Practical Machine Learning Tools and Techniques, 4th Edition, Morgan Kaufmann, Cambridge, 2017.

Biographies

Carlos Hernández is an industrial engineer, consultant, and university professor. He earned Master of Sciences in Engineering and Doctor of Engineering from Technische Universität Braunschweig, Brunswick, Germany. He is the author of several scientific and engineering articles. Through the years he has taught lectures in Discrete Event Simulation, Engineering Economics, Corporate Finances, Data Mining and Machine Learning for engineering students. He has developed a professional career working for multinational companies such as PricewaterhouseCoopers, BHP Billiton, and Merck Sharp & Dohme. He also worked as a scientific researcher in the Institut für Produktionsmesstechnick at TU Braunschweig, Germany. His research interests include manufacturing process simulation, supply chain design and simulation, and machine learning for finances. He is a member of IEOM.

Jaime Castillo is an industrial engineer, consultant, and university professor. He earned Licentiate Degree in Forest Engineering from Universidad de Concepción, Concepción, Chile, and Master in Industrial Engineering from Universidad del Desarrollo, Santiago, Chile. He has taught lectures in Project Planning & Management, Project Evaluation, Decision Theory, and Process Simulation for engineering students. He has extensive experience working in sustainable development projects for the local forest industry. During her academic tenure he has been appointed in different management positions and has mentored over a fifty students. His research interests include project management, logistic risk assessment, and decision theory.