

STATISTICAL ANALYSIS OF MARITIME ACTIVITIES AND ACCIDENTS IN THE
STRAIT OF ISTANBUL

by

Işıl Baş

B.S., Mathematics Engineering, Istanbul Technical University, 2006

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Industrial Engineering
Boğaziçi University
2010

ACKNOWLEDGEMENTS

I am happy to state my sincere gratitude to my thesis supervisor Prof. İlhan Or, for his wisdom and helpfulness throughout this thesis. He was always understanding and motivating, and it was a pleasure working with him.

I would also like to thank Birnur Özbaş for being very helpful in the improvement of this study. Without her help, it would be hard to proceed in this thesis. I am also grateful to Onur Özgün for his technical support in this project.

Finally, I gratefully recognize the support of TÜBİTAK – Turkish Technological and Scientific Research Institute, which provided a scholarship during my graduate education.

ABSTRACT

STATISTICAL ANALYSIS OF MARITIME ACTIVITIES AND ACCIDENTS IN THE STRAIT OF ISTANBUL

The Strait of Istanbul has a great strategic importance in the world since it is the key component of the unique waterway system between the Black Sea and the Mediterranean. Around 160 transit vessel passages are daily accomplished, many of them carrying hazardous cargo. Due to its narrow and winding geography, adverse meteorological conditions (current, fog, wind and rain which affect the visibility conditions in the Channel) and very high local traffic density, maritime traffic becomes very complex and risky. This study, which is based on statistical data analyses of several datasets, consists of two core sections.

In the first part, in order to support and provide input to simulation-based risk studies of the Strait, transit vessel arrival process and visibility conditions (which are the key factors affecting the maritime transit traffic of the Strait and hence, the accident potential of the vessels) are investigated and analyzed. In this regard, distribution fitting and parameter estimations are performed via detailed analyses of historic transit and visibility data. For the distribution fitting study of the arrival process, Input Analyzer function of the ARENA software is used. For visibility conditions, the modeling of fog start times and their durations are done with the aid of empirical and “Mixtures of Generalized Erlang” distributions.

The second part of the study is based on maritime accidents occurred in the Strait of Istanbul in recent years. The first sub-section contains descriptive analyses of the accidents that happened in the Strait: these analyses can be cited as frequency histograms, chi-square tests (in order to identify the statistical relationships between different accident factors), discriminant and cluster analyses to classify the accident causes. In the second sub-section, the likelihood of maritime accidents involving transit vessels is investigated through

logistic regression modeling. The related logistic regression models are generated from several data sources covering the years 2005-2007 and are aimed at identifying the most significant factors influencing the probability of a vessel to play a role in an accident. Six different models are generated based on various assumptions of human error occurrence in the vessel during its transit. As a result, significant factors triggering an accident are found to be the pilot status, daytime/nighttime status, season and existence of bad visibility conditions, failure and human error. Moreover, the accidents occurring in the anchoring areas are also discussed in this section. Due to the lack of information about the anchoring areas, the weaknesses of the model on these areas are discussed.

ÖZET

İSTANBUL BOĞAZI DENİZ OLAYLARI VE KAZALARININ İSTATİSTİKSEL İNCELEMESİ

Karadeniz ve Akdeniz'i bağlayan su yolu sisteminin önemli bir bileşeni olması nedeniyle İstanbul Boğazı dünyada çok büyük stratejik öneme sahiptir. Boğazdan günlük olarak ortalama 160 gemi geçişi sağlanmaktadır ve bunların önemli bir kısmı tehlikeli madde taşımaktadır. Boğazın dar ve kavisli yapısı, elverişsiz meteorolojik koşullar (boğazdaki görüş koşullarını etkileyen akıntı, sis, rüzgar ve yağışlar), yoğun yerel trafik, boğazdaki gemi trafiğini karmaşık ve riskli kılmaktadır. Pek çok veri kümesinin istatistiksel analizine dayalı olan bu çalışma, iki ana bölümden oluşmaktadır.

İlk bölümde, benzetim tabanlı risk çalışmalarını desteklemek ve çalışmalara girdi sağlamak amacıyla, deniz trafiğini ve böylelikle kaza potansiyelini etkileyen önemli iki faktör, transit gemi geliş süreci ve görüş koşulları araştırılmış ve incelenmiştir. Bu bağlamda, geçmiş transit geçiş ve görüş verileri yardımıyla dağılım uydurma ve parametre tahmini çalışmaları yürütülmüştür. Bir diğer yanda, görüş şartları için, sis başlangıç zamanları ve süreleri, deneysel dağılımlar ve Genel Erlang dağılımları yardımıyla modellenmiştir.

Çalışmanın ikinci kısmı ise geçmiş yıllarda meydana gelmiş kazalara dayanmaktadır. İlk alt bölümde boğazda meydana gelmiş olan geçmiş kazaların birçok istatistiksel analizi yapılmıştır: bunlar, frekans histogramları, kazayla ilgili faktörler arasındaki istatistiksel ilişkiyi gösteren ki-kare testleri ve kazaların nedenlerine göre sınıflandırılmasında kullanılan diskriminant ve kümeleme analizleri olarak sayılabilir. İkinci alt bölümde ise transit gemilerin kazaya karışma ihtimallerini bulmak için oluşturulan lojistik regresyon modelleri yer almaktadır. Bu modeller, 2005-2007 yıllarını içeren birçok veri kaynağından elde edilmiş olup, geminin bir transit kazada rol alması olasılığını arttıran faktörleri bulmak amacını taşımaktadır. Transit geçiş yapan gemide insan hatası olması durumu ile

ilgili, farklı varsayımlara dayanan, altı adet model yaratılmıştır. Sonuç olarak, pilot alma durumu, gündüz/gece durumu, mevsim, kötü görüş koşulları, arıza ve insan hatası faktörleri kazayı tetikleyen en önemli faktörler olarak elde edilmiştir. Aynı zamanda bu bölümde, demir yeri lojistik regresyon modellerinde yaşanan problemler de ele alınmıştır. Bu bölgelerle ilgili bilgi eksikliğinden ötürü, modelin zayıf kaldığı yönler tartışılmıştır.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS.....	iii
ABSTRACT	iv
ÖZET	vi
LIST OF FIGURES	xi
LIST OF TABLES	xvi
LIST OF SYMBOLS/ABBREVIATIONS	xxvi
1. INTRODUCTION AND OBJECTIVES OF THE STUDY	1
2. LITERATURE SURVEY	7
3. MODELING OF THE ARRIVAL PROCESS	12
3.1. Introduction and Description of the Data	12
3.2. Methodology and Interpretations.....	15
3.3. Trend Tests	19
4. MODELING OF THE VISIBILITY CONDITIONS	23
4.1. Data Sources and Graphical Interpretations	23
4.1.1. Turkish State Meteorological Service Visibility Data	24
4.1.2. Weather Underground Visibility Data	25
4.1.3. The Channel Closure Data of the Directorate General of Coastal Safety	29
4.2. The Visibility Model: On / Off Process.....	32
4.2.1. General overview	32
4.2.2. Modeling of Fog Initiations in the Summer and Transition Seasons	35
4.2.3. Comparison of WU and Channel Closure Data	36
4.2.4. Modeling of Fog Initiations in the Winter Season – On Process	38
4.2.5. Modeling of Fog Durations – Off process	41
5. STATISTICAL ANALYSIS OF MARITIME ACCIDENTS	44
5.1. Data Description	44
5.1.1. Vessel Data	45
5.1.2. Accident Data.....	45
5.1.3. Data Completion: Tonnage – Length Regression	45
5.2. Descriptive Statistics	47
5.2.1. Vessel Descriptions	47

5.2.2. Accident Descriptions	52
5.3. Chi-Square Tests for Nominal Scale Variables and Interpretations	56
5.3.1. Definitions	56
5.3.2. Chi- Square Testing For Vessel Attributes	57
5.3.3. Chi Square Testing For Accident Attributes	69
5.4. Near-Miss Analyses	74
5.4.1. Data Description	74
5.4.2. Near-Miss Descriptions.....	75
5.4.3. Chi-Square Tests and Results	82
5.5. Discriminant Analysis	89
5.6. Cluster Analysis	96
6. LOGISTIC REGRESSION MODEL OF ACCIDENT LIKELIHOOD	103
6.1. Logistic Regression Definition	103
6.2. Logistic Regression Methodology and Description of the Data	105
6.2.1. Cardinality Value Determination	106
6.2.1.1. Type Cardinality	106
6.2.1.2. Age & Flag Cardinality.....	107
6.2.2. Correlation Analysis	108
6.3. The Logistic Regression Model for Transit Accidents.....	109
6.3.1. Model Inputs.....	109
6.3.2. Introducing Human Error into the Model	111
6.3.3. Model Interpretations	113
6.3.4. Verification of the Models	119
6.3.5. Validation of the Models	120
6.3.6. Weakness of the Models	124
6.4. Anchorage Areas Logistic Regression Model	125
7. CONCLUSION.....	126
8. FURTHER STUDIES	129
APPENDIX A: ARRIVAL PROCESS PROBABILITY DISTRIBUTIONS	131
APPENDIX B: TREND ANALYSIS RESULTS	135
APPENDIX C: TONNAGE-LENGTH REGRESSION.....	137
APPENDIX D: LOGISTIC REGRESSION MODEL OUTPUTS.....	139

APPENDIX E: ACCIDENT FREQUENCIES CALCULATED FOR THE REAL DATA.....	141
APPENDIX F: MODEL VERIFICATION OUTPUTS	145
APPENDIX G: MODEL VALIDATION OUTPUTS – ACCIDENT DATABASE	152
APPENDIX H: MODEL VALIDATION OUTPUTS – YEAR 2009.....	156
REFERENCES	160

LIST OF FIGURES

Figure 3.1.	An example of the transit vessel dataset from year 2007	13
Figure 3.2.	Priority classes used for scheduling	14
Figure 3.3.	Histogram of interarrival times (in minutes) of southbound Class D vessels passing directly at summer nights	17
Figure 4.1.	31 days moving average of frequency distribution of daily fog occurrences over 32 years in Meteorological Service data	25
Figure 4.2.	Frequency distributions of fog start times in hours of day based on the WU data	27
Figure 4.3.	Probability distribution of fog start times in daytime and nighttime	27
Figure 4.4.	2.5 hours moving average of frequency distributions of fog occurrences in hours of day in the WU data	28
Figure 4.5.	Scatter diagram of fog start times versus fog durations in the WU data ..	29
Figure 4.6.	Frequency distribution of closure start times in hours	30
Figure 4.7.	Probability distribution of closure start times in day and night in the closure data	30
Figure 4.8.	2.5 hours moving average of frequency distributions of fog occurrences in hours of day in the closure data	31
Figure 4.9.	Scatter diagram of fog start times versus fog durations in the closure data	32

Figure 4.10.	On/off process layout	33
Figure 4.11.	Graphical representation of the generalized Erlang distribution	34
Figure 4.12.	MGE-k distribution with $Cv_x^2 > 1$	35
Figure 4.13.	Graphical representation of MGE-3 distribution	40
Figure 4.14.	Graphical representation of MGE-2 distribution	42
Figure 5.1.	Frequency histogram of vessels involved in accidents according to their types	47
Figure 5.2.	Frequency histogram of vessels involved in accidents according to their tonnages	48
Figure 5.3.	Frequency histogram of vessels involved in accidents according to their lengths	49
Figure 5.4.	Frequency histogram of vessels involved in accidents according to their flag types	50
Figure 5.5.	Frequency histogram of vessels involved in accidents according to their ages	51
Figure 5.6.	Frequency histogram of vessels involved in accidents according to their pilot status	51
Figure 5.7.	Frequency histogram of accidents according to year	52
Figure 5.8.	Frequency histogram of accidents according to season	53
Figure 5.9.	Frequency histogram of accidents according to daytime/nighttime status	54

Figure 5.10.	Frequency histogram of accidents according to accident type	54
Figure 5.11.	Frequency histogram of accidents according to region	55
Figure 5.12.	Frequency histogram of accidents according to accident causes	56
Figure 5.13.	Accident regions map	59
Figure 5.14.	Frequency histogram of failures according to vessel type	76
Figure 5.15.	Frequency histogram of failures according to vessel length	76
Figure 5.16.	Frequency histogram of failures according to pilot captain status	77
Figure 5.17.	Frequency histogram of failures according to vessel flag type	78
Figure 5.18.	Frequency histogram of failures according to vessel age	78
Figure 5.19.	Frequency histogram of failures according to years	79
Figure 5.20.	Frequency histogram of failures according to season	80
Figure 5.21.	Frequency histogram of failures according to time status	80
Figure 5.22.	Frequency histogram of failures according to region	81
Figure 5.23.	Frequency histogram of failures according to failure type	82
Figure 5.24.	Categorical variablewise importance of Cluster 1	100
Figure 5.25.	Categorical variablewise importance of Cluster 2	101
Figure 5.26.	Categorical variablewise importance of Cluster 3	101

Figure 6.1.	Local traffic density calculation	111
Figure 6.2.	An example from the validation dataset	122
Figure E.1.	Accident frequencies with respect to the vessel type.....	141
Figure E.2.	Accident frequencies with respect to the vessel GRT	141
Figure E.3.	Accident frequencies with respect to the vessel length	142
Figure E.4.	Accident frequencies with respect to the pilot status	142
Figure E.5.	Accident frequencies with respect to the accident season	143
Figure E.6.	Accident frequencies with respect to the accident time	143
Figure E.7.	Accident frequencies with respect to the vessel flag type	144
Figure E.8.	Accident frequencies with respect to the vessel age	144
Figure F.1.	Model probability histogram for non-accidents	145
Figure F.2.	Model A probability histogram for non-accidents	146
Figure F.3.	Model B probability histogram for non-accidents	146
Figure F.4.	Model C probability histogram for non-accidents	147
Figure F.5.	Model D probability histogram for non-accidents	147
Figure F.6.	Model E probability histogram for non-accidents	148
Figure F.7.	Model probability histogram for the accidents	148

Figure F.8.	Model A probability histogram for the accidents	149
Figure F.9.	Model B probability histogram for the accidents	149
Figure F.10.	Model C probability histogram for the accidents	150
Figure F.11.	Model D probability histogram for the accidents	150
Figure F.12.	Model E probability histogram for the accidents	151
Figure G.1.	Model probability histogram	152
Figure G.2.	Model A probability histogram	153
Figure G.3.	Model B probability histogram	153
Figure G.4.	Model C probability histogram	154
Figure G.5.	Model D probability histogram	154
Figure G.6.	Model E probability histogram	155
Figure H.1.	Model probability histogram	156
Figure H.2.	Model A probability histogram	157
Figure H.3.	Model B probability histogram	157
Figure H.4.	Model C probability histogram	158
Figure H.5.	Model D probability histogram	158
Figure H.6.	Model E probability histogram	159

LIST OF TABLES

Table 3.1.	Number of vessels transited in 2005, 2006 and 2007 according to the general categories	14
Table 3.2.	Frequency distribution of vessels transited in 2005, 2006 and 2007 according to the general categories	14
Table 3.3.	Number of vessels transited in 2005, 2006 and 2007 according to the priority classes	15
Table 3.4.	Frequency distribution of vessels transited in 2005, 2006 and 2007 according to the priority classes	15
Table 3.5.	Fitted distribution summary for interarrival times (in minutes) of southbound Class D vessels passing directly at summer nights	18
Table 3.6.	The number and the frequencies of the sub-classes where $p\text{-value} < 0.05$ for each priority class	21
Table 3.7.	Class D frequencies based on the direction and direct/indirect passing status	22
Table 4.1.	Number of fog observations for 32 years in Meteorological Service data	24
Table 4.2.	Total number of foggy days for 15-day time intervals in 32 years	25
Table 4.3.	Visibility Seasons	26
Table 4.4.	Probability distribution of fog occurrences in summer	35
Table 4.5.	Probability distribution of fog occurrences in transition season	36

Table 4.6.	Mean, variance and squared coefficient of variation of good visibility conditions in the WU data	36
Table 4.7.	Mean, variance and squared coefficient of variation of bad visibility conditions in the WU data	36
Table 4.8.	Mean, variance and squared coefficient of variation of open state of the Channel	36
Table 4.9.	Mean, variance and squared coefficient of variation of closed state of the Channel	37
Table 4.10.	Test results of the comparison between the original WU and Channel closure data	37
Table 4.11.	Test results of the comparison between adjusted WU and Channel closure data	38
Table 4.12.	Fog frequency comparison between adjusted WU and Channel closure data	38
Table 4.13.	Mean, variance and squared coefficient of variation of on-period data ...	38
Table 4.14.	The first three moments of on-period data	39
Table 4.15.	Model parameters of on-period data	41
Table 4.16.	Mean, variance and squared coefficient of variation of off-period data ..	41
Table 4.17.	The first three moments of off-period data	41
Table 4.18.	Model parameters of off-period data	43

Table 5.1.	Tonnage-Length Regression Equations	46
Table 5.2.	Frequency of vessels according to their types	47
Table 5.3.	Frequency of vessels according to their tonnages	48
Table 5.4.	Frequency of vessels according to their lengths	49
Table 5.5.	Frequency of vessels according to their flag types	49
Table 5.6.	Frequency of vessels according to their ages	50
Table 5.7.	Frequency of vessels according to their pilot status	51
Table 5.8.	Frequency of accidents according to year	52
Table 5.9.	Frequency of accidents according to season	53
Table 5.10.	Frequency of accidents according to daytime/nighttime status	53
Table 5.11.	Frequency of accidents according to the accident type	54
Table 5.12.	Frequency of accidents according to region	55
Table 5.13.	Frequency of accidents according to accident causes	56
Table 5.14.	Chi-square test results for the vessel database	60
Table 5.15.	Length * Vessel type crosstabulation	61
Table 5.16.	Flag type * Vessel type crosstabulation	61
Table 5.17.	Age * Vessel type crosstabulation	61

Table 5.18.	Pilot * Vessel type crosstabulation	62
Table 5.19.	Accident year * Vessel type crosstabulation	62
Table 5.20.	Season * Vessel type crosstabulation	63
Table 5.21.	Region * Vessel type crosstabulation	63
Table 5.22.	Accident type * Vessel type crosstabulation	64
Table 5.23.	Flag type * Length crosstabulation	64
Table 5.24..	Pilot * Length crosstabulation	64
Table 5.25.	Season * Length crosstabulation	65
Table 5.26.	Day/Night * Length crosstabulation	65
Table 5.27.	Accident type * Length crosstabulation	66
Table 5.28.	Flag type * Pilot crosstabulation	66
Table 5.29.	Year * Age crosstabulation	67
Table 5.30.	Region * Age crosstabulation	67
Table 5.31.	Year * Pilot crosstabulation	68
Table 5.32.	Accident cause * Pilot crosstabulation	68
Table 5.33.	Chi-square test results for the accident-based attributes	69
Table 5.34.	Day/Night * Year crosstabulation	69

Table 5.35.	Year * Region crosstabulation	70
Table 5.36.	Year * Accident type crosstabulation	70
Table 5.37.	Year * Accident cause crosstabulation	71
Table 5.38.	Season * Day/Night crosstabulation	71
Table 5.39.	Region * Season crosstabulation	72
Table 5.40.	Accident cause * Season crosstabulation	72
Table 5.41.	Region * Accident type crosstabulation	73
Table 5.42.	Region number * Accident cause crosstabulation	73
Table 5.43.	Accident cause * Accident type crosstabulation	74
Table 5.44.	Frequency of failures according to vessel type	75
Table 5.45.	Frequency of failures according to vessel length	76
Table 5.46.	Frequency of failures according to pilot captain status	77
Table 5.47.	Frequency of failures according to vessel flag type	77
Table 5.48.	Frequency histogram of failures according to vessel age	78
Table 5.49.	Frequency of failures according to years	79
Table 5.50.	Frequency of failures according to season	79
Table 5.51.	Frequency histogram of failures according to time status	80

Table 5.52.	Frequency of failures according to region	81
Table 5.53.	Frequency of failures according to failure type	81
Table 5.54.	Chi-square test results regarding possible attributes triggering failures ..	82
Table 5.55.	Length * Vessel type crosstabulation	83
Table 5.56.	Flag type * Vessel type crosstabulation	83
Table 5.57.	Age * Vessel type crosstabulation	84
Table 5.58.	Flag type * Length crosstabulation	84
Table 5.59.	Length * Age crosstabulation	84
Table 5.60.	Day/Night * Length crosstabulation	85
Table 5.61.	Region * Length crosstabulation	85
Table 5.62.	Length * Failure type crosstabulation	86
Table 5.63.	Pilot * Failure type crosstabulation	86
Table 5.64.	Flag type * Age crosstabulation	87
Table 5.65.	Season * Age crosstabulation	87
Table 5.66.	Failure type * Age crosstabulation	87
Table 5.67.	Year * Region crosstabulation	88
Table 5.68.	Season * Failure time crosstabulation	88

Table 5.69.	Region * Failure type crosstabulation	89
Table 5.70.	Tests of equality of group means for the full model	91
Table 5.71.	Tests of equality of group means for the reduced model	91
Table 5.72.	Pooled within-group matrices	91
Table 5.73.	Box's test of equality of covariance matrices: Log determinants	92
Table 5.74.	Box's test of equality of covariance matrices: Test results	92
Table 5.75.	Summary of canonical discriminant functions: Eigenvalues	93
Table 5.76.	Summary of canonical discriminant functions: Wilks' lambda	93
Table 5.77.	Standardized canonical discriminant function coefficients	93
Table 5.78.	Structure matrix	94
Table 5.79.	Functions at group centroids	94
Table 5.80.	Classification function coefficients	95
Table 5.81.	Classification Results	95
Table 5.82.	Auto-clustering table	97
Table 5.83.	Cluster distribution	98
Table 5.84.	Cluster profile regarding vessel type	98
Table 5.85.	Cluster profile regarding the pilot status of vessels	98

Table 5.86.	Cluster profile regarding accident type	99
Table 5.87.	Cluster profile regarding accident region	99
Table 5.88.	Cluster profile for the accident cause	100
Table 6.1.	Type Cardinality Values	107
Table 6.2.	Age & Flag Cardinality Values	107
Table 6.3.	Correlation analysis regarding all considered logistic regression independent variables	108
Table 6.4.	Presentation of the logistic regression models	113
Table 6.5.	Case Processing Summary	113
Table 6.6.	Dependent Variable Encoding	114
Table 6.7.	Iteration History of Step 0	114
Table 6.8.	Classification Table of Step 0	114
Table 6.9.	Variables in the Equation for Step 0	115
Table 6.10.	Variables Not in the Equation	115
Table 6.11.	Iteration History for Block 1	116
Table 6.12.	Omnibus Tests of Model Coefficients	116
Table 6.13.	Model Summary	116

Table 6.14.	Hosmer and Lemeshow Test	117
Table 6.15.	Contingency Table for Hosmer and Lemeshow Test	117
Table 6.16.	Classification Table for Block 1	117
Table 6.17.	Variables in the Equation	118
Table 6.18.	Sum of expected probabilities for six logistic regression models (verification)	120
Table 6.19.	Sum of expected probabilities for six logistic regression models (validation set 1)	122
Table 6.20.	Sum of expected probabilities for six logistic regression models (validation set 2)	124
Table A.1.	Fitted distributions to vessel clusters in arrival processes: Class A.....	131
Table A.2.	Fitted distributions to vessel clusters in arrival processes: Class B.....	131
Table A.3.	Fitted distributions to vessel clusters in arrival processes: Class C.....	132
Table A.4.	Fitted distributions to vessel clusters in arrival processes: Class E.....	132
Table A.5.	Fitted distributions to vessel clusters in arrival processes: Class D.....	133
Table A.6.	Fitted distributions to vessel clusters in arrival processes: Class P.....	134
Table B.1.	t, F and Mann-Whitney tests p-values for trend analysis: Class A	135
Table B.2.	t, F and Mann-Whitney tests p-values for trend analysis: Class B	135

Table B.3.	t, F and Mann-Whitney tests p-values for trend analysis: Class C	135
Table B.4.	t, F and Mann-Whitney tests p-values for trend analysis: Class D	135
Table B.5.	t, F and Mann-Whitney tests p-values for trend analysis: Class E.....	136
Table B.6.	t, F and Mann-Whitney tests p-values for trend analysis: Class P	136
Table C.1.	Passenger vessels regression model	137
Table C.2.	HazMat/LNG/LPG vessels regression model	137
Table C.3.	Tanker vessels regression model	137
Table C.4.	General cargo vessels regression model	138
Table D.1.	Logistic regression output: Model	139
Table D.2.	Logistic regression output: Model A.....	139
Table D.3.	Logistic regression output: Model B.....	139
Table D.4.	Logistic regression output: Model C.....	140
Table D.5.	Logistic regression output: Model D.....	140
Table D.6.	Logistic regression output: Model E	140
Table F.1.	Estimated accident probability frequencies for the original data	145
Table G.1.	Estimated accident probability frequencies for the accident data	152
Table H.1.	Estimated accident probability frequencies for the data of year 2009	156

LIST OF SYMBOLS/ABBREVIATIONS

Cv_x^2	Squared Coefficient of Variability
χ_0^2	Chi-Square
AHP	Analytical Hierarchy Process
AIC	Akaike's Information Criterion
BIC	Bayesian Information Criterion
BMTS	Bureau of Turkish Strait's Maritime Traffic Services
GLM	Generalized Linear Models
GRT	Gross Tonnage
HazMat	Hazardous Material
KOERI	Kandilli Observatory and Earthquake Research Institute
K-S	Kolmogorov-Smirnov
LNG	Liquefied Natural Gas
LPG	Liquefied Petroleum Gas
MGE	Mixtures of Generalized Erlang
R&R	Channel Traffic Rules and Regulations
VTs	Vessel Traffic Services
WU	Weather Underground

1. INTRODUCTION AND OBJECTIVE OF THE STUDY

The Strait of Istanbul has a great strategic importance since it is the only waterway connection of the Black Sea with the Sea of Marmara and open seas beyond. Because of this, the Strait has very heavy maritime traffic. Compared to other straits, the Strait of Istanbul has three times more traffic than the Suez Canal, and four times more traffic than the Panama Canal (Ulusçu *et al.*, 2009a). In fact, after the Malacca Strait, the Strait of Istanbul has the world's second most dense sea traffic. One other property of the Strait is that, it is one of the most difficult waterways to navigate in the world, due to its narrow width, many sharp turns, strong and frequently changing current conditions, in addition to the heavy local traffic crisscrossing between the shores (Birpınar *et al.*, 2009).

There are several reasons making the Strait a dangerous and a risky one. The Strait of Istanbul is a narrow and curly strait. The average width of the Strait is 1.5 kilometers and its length is 31 kilometers, the narrowest point being just 660 meters. Because of its geographical properties, it forces the vessels to change their course many times. At some sharp turns, the vessels coming from opposite directions cannot see each other (Ulusçu *et al.*, 2008). The two bridges on the Strait and the energy transfer lines are other factors affecting recognition in low visibility conditions. Weather conditions, especially frequent fogs reducing visibility and strong winds also contribute to the Strait maritime risks.

One other geographical property of the Strait of Istanbul contributing to its treacherous nature is the currents. There are two currents in the Strait, one in the surface, from the Black Sea to the Marmara Sea and the other one 15 meters below the surface, in the reverse direction. These currents can reach up to 12 km/hour, creating significant hazards for the maritime traffic. They may create turbulence in certain points in the Strait. The strong winds from the south not only affect the traffic by itself, but also cause alterations in the currents that affect the traffic. The highly populated shores of the Strait increase the risk factors (Ulusçu *et al.*, 2008).

Beside these geographical and natural factors, there are some other factors affecting maritime traffic in the Strait of Istanbul. One of these factors is the highly congested local

maritime traffic. Regarding the May 2007 report of the Undersecretariat for Maritime Affairs about the local traffic on the Strait of Istanbul, there are 2,025 planned passes daily (Ulusçu *et al.*, 2009b). When the unscheduled activities such as tugboats, fishing boats, pleasure boats, etc. are added, this number increases up to 2,500 per day. One other factor is the frequency of the transit vessels in the Strait and their cargos. The properties of the transit passing vehicles create an important risk factor on the navigation. In a year, there are about 55,000 transit vessels passing through the Strait on average, while about 20% of these carrying hazardous materials (Ulusçu *et al.*, 2009b).

The maritime transit traffic through the Strait of Istanbul has been regulated by the Montreux Convention. Through this treaty, while the Strait is considered International Waters, its management is carried by the Turkish state. At the time of the convention, only an average of just 5000 vessels transited the Strait annually whereas, at the moment, this number is 10 times larger (Ulusçu *et al.*, 2009a).

Over time, the size and the weight of the international maritime vessels have considerably increased. A reflection of this can be seen on the Strait of Istanbul, where the size of tankers carrying oil, LPG, LNG and other hazardous cargo increased considerably in time. This is the result of the Black Sea countries trying to integrate with the developed countries and using the Strait of Istanbul as a passage for energy products transportation (Birpınar *et al.*, 2009). This not only increases the strategic importance of the Strait, but also increases the accident risks on it.

There are some rules and regulations set by the Turkish state in the period 1994-1998 for the Turkish Straits and the Marmara Region. Basically, the primary purpose of these actions has been to ensure safety and protect the environment in the Strait. The rules and regulations include the Traffic Separation Schemes, which set up and defined the traffic lanes in the Strait regulating the traffic for the transit vessels (Ulusçu *et al.*, 2009a). Additionally, in 2003, an integrated radar, camera, sensors and communication system, called Vessel Traffic Services (VTS) is introduced by the Turkish state. This service is intended to increase the safety of the navigation in the Strait.

Even though these precautions are implemented, the Strait of Istanbul is still a dangerous place for maritime traffic. After the preventive actions, the number of hazards decreased. But there are still accidents happening there. Below is a set of examples on maritime accidents that occurred in the Strait:

- On 14.12.1960, Yugoslavian tanker Peter Zoranic collided with Greek tanker World Harmony at Kanlıca, leading to the death of 20 crew members. Moreover, 18,000 tons of oil spilled into the sea.
- On 15.09.1964, a Norwegian flagged vessel, Norborn crashed into Peter Zoranic's wreck which caused fire and pollution.
- On 01.03.1966, Soviet flagged tankers Lutsik and Cransky Oktiabr collided at Kızkulesi. 1,850 tons of oil spilled into the sea. Karaköy ferry terminal and a ferry burned because of this accident.
- In July 1966, the ferry Yeni Galatasaray collided with the Turkish coaster Aksaray. 13 people died in the ensuing fire.
- Collision between the Greek vessel Evriali and the Romanian tanker Independenta on 15.11.1979 at the southern entrance was the largest maritime accident ever happened in the Strait region. 43 crew members died and about 95,000 tons of crude oil from Independenta spilled into the sea and approximately 30,000 tons of this oil burned. The incident was ranked as the 10th worst tanker accident in the world.
- On 29.03.1990, Iraqi tanker Jambur and Chinese bulk carrier Da Tong Shan collided in the Strait. About 2,600 tons of oil from Jambur spilled into the sea.
- On 14.11.1991, Phillippine-flagged bulk carrier Madonna Lily and Lebanese live stock carrier Rabunion 18 collided in November. Rabunion 18 sank with its cargo of 21,000 sheep and three crew members died.

- On 13.03.1994, the Greek Cypriot tanker Nassia and bulk carrier Shipbroker collided in the Strait of Istanbul. 29 people died in this accident; approximately 20,000 tons of crude oil spilled over the sea and caused fire and severe pollution.
- In December 1999, the Russia-flagged tanker Volganef-248, carrying 4,365 tons of oil, split in two and grounded in the Strait. Approximately, 1,500 tons of oil spilled into the sea.
- On 06.10.2002, Malta-flagged bulk carrier Gotia rammed into the Emirgan Pier and damaged its fuel tank. 18 tons of oil spilled into the sea.
- On 10.11.2003, Georgian-flagged cargo carrier Svyatov Panteleymon grounded and broke apart. Its fuel spilled into the sea.
- On 13.02.2004, Cambodian bulk carrier Hera sank at the northern entrance of the Strait due to bad weather conditions. 19 members of the crew died.

Hence the accidents occurring in the Strait show that all transit vessels and local traffic elements constitute risk factors threatening the Strait and its entourage, thereby bringing about the need for a detailed analysis of accident causing factors and of accidents themselves. Within this context, this study incorporates two main sets of statistical analyses presented in Chapters 3, 4, 5 and 6. Chapters 3 and 4 constitute the distribution fitting analysis for vessel arrival process and bad visibility conditions, which are two key factors affecting transit traffic and hence, increasing the accident risks. Chapters 5 and 6 are mainly composed of several statistical analyses of historical maritime accidents, such as chi-square analysis and logistic regression model, whose details are presented below.

The distribution fitting part of the study consists of two primary aspects, the arrival process of the transit vessels and visibility conditions in the Strait. The distribution fitting for the arrival process (Chapter 3) is accomplished through selecting appropriate distributions and related parameter estimations regarding the vessels approaching the north and south entrances of the Strait for their intended transit. These “arrival distributions” are expected to provide important information and clues regarding the maritime traffic

congestion (and its possible negative effects on accidents, as well as providing support and input to related simulation based risk studies (such as (Özbaş, 2005), (Almaz, 2006) and (Ulusçu *et al.*, 2009b)). The arrival process investigation concentrates on forming vessel clusters based on their priority classes (which are used for scheduling purposes). The vessels clustered by their priorities are then split into several sub-categories with respect to their arrival direction, time of the day, season and direct/indirect passage choice for the years 2005, 2006 and 2007. Then, a trend analysis is conducted to see whether the samples obtained for all sub-classes differ or not according to the years.

In the second distribution fitting study (Chapter 4), visibility analysis consists on modeling fog occurrences and durations according to a detailed analysis of several data sources. While the methodology is very similar to the study conducted by (Almaz, 2006), more and detailed historic data are analyzed in this study. Moreover, the model generated based on the WU dataset is cross-checked with actual Channel closure data to see whether they have the same pattern or not. The model is then modified in order to reflect the actual Strait closures due to bad visibility conditions and the results are supported with several statistical tests as in the arrival process.

Apart from the distribution fitting analyses explained previously, two basic accident-based analyses are conducted in this study. Chapter 5 begins with various descriptive statistical analyses of the Strait maritime accidents that occurred between years 1982 and 2007. According to the datasets covering both vessel and accident attributes, frequency distributions and chi-square tests are performed. Chi-square test enables us to observe whether there is a statistical relationship between two categorical attributes of these datasets. The same analyses are conducted for near-miss cases, namely the dataset of engine/propeller, rudder and mechanical/electrical failures occurred between the years 2005-2007 with or without causing an accident to investigate statistically-supported results for the failures. The statistical analyses concerning the vessel dataset ends up with discriminant and cluster analyses with respect to accident cause. This part of the study intends to find a formulation for grouping all vessels involved in accidents according to the accident cause.

In Chapter 6, a logistic regression model is constituted in order to find out the most significant factors influencing the accident propensity of a vessel transiting the Strait. The model dataset is composed of transit vessel data, vessel accidents in transit and the near-miss cases composed of three types of failures covering the years 2005-2007. Accordingly, some additional model inputs are generated either randomly or by simulation-based risk studies of the department. Accordingly, since there is no exact information about the human error probability of a vessel which is not subject to an accident, six different cases are considered. The first model does not contain a human error variable and the remaining five models are generated based on the assumption that the probabilities of occurrence of human error in a vessel are 0.25%, 0.5%, 1%, 3% and 5% respectively. As the outcome of the run, significant factors triggering an accident are found to be the pilot status, daytime/nighttime status, season and existence of bad visibility conditions, failure or human error. The results obtained from the models are then compared in the verification and validation sub-sections of the study. Moreover, this section also contains the problems met in the anchorage area logistic regression model and the weak points of the model are discussed.

2. LITERATURE SURVEY

Safe navigation in waterway systems is an important issue for many narrow waterways and inland waters since the maritime risks do not simply affect the vessels themselves: a maritime accident can really have great negative impact on the environment, the population or the biodiversity. Furthermore, hazardous material carrying vessels increase the risks in waterways and the risk analysis of maritime accidents becomes necessary. In the literature, it is possible to find various studies on the likelihood and risk assessment of maritime accidents. Within this context, various techniques are used such as probabilistic risk assessment, simulation modeling and statistical analysis of historical data. Since this study concentrates on data analysis, the past studies accomplished with this technique are of great importance. In this section, literature on this topic (maritime accident risks in narrow waterways in general and the Strait of Istanbul in particular) and on the statistical analysis methodologies are introduced.

(Le Blanc and Rucks, 1996) study the accident data containing 936 vessel accidents between 1979 and 1987 on lower Mississippi River. The authors use cluster analysis to create four groups and then use discriminant analysis to create three discriminant functions. The results reveal that the accident clusters are defined as danger zone, bad conditions for good navigators, probability preventable and accidents that should not have happened.

In a study about the oil spill size in accidents in U.S. waters, (Anderson and Talley, 1995) examine a series of variables including vessel-related factors from a database containing detailed tanker and tank barge accidents. (Talley, 1996), in another study, investigates the risk of cargo damage in containership accidents in U.S. waters. The author makes Probit and Tobit analysis in order to draw conclusions from the containership accidents data. The results of this study show that the cargo damage risk is less if the vessel is manned by a licensed operator when compared to an unlicensed operator and for large vessels when compared to smaller-sized vessels, and it is greater for collision, fire & explosions when compared to grounding accidents.

In a study for establishing a safety level in Dutch inland waterway transport system, (Roeleven *et al.*, 1995) state that the visibility and wind are important factors triggering the occurrence of maritime accidents. The authors use Generalized Linear Models (GLM) to come to this conclusion.

In a study for predicting risks during transits into and out of port, (Kite-Powell *et al.*, 1998) collect and analyze data from five U.S. ports, which is then used to construct a port-level model having predictive ability. The results of the study reveal that operator skills, the traffic and the vessel properties and environmental attributes are the key factors affecting the risk.

(Merrick *et al.*, 2000) use a systems simulation model in order to determine risks in Prince William Sound, Alaska. The authors use the data to estimate the vessel interarrival time distributions and develop a Markov Chain model to predict the weather conditions. In another study for modeling and analyzing maritime transportation risks in the Washington State Ferries system, (Merrick *et al.*, 2001) analyze the maritime traffic data in order to build a model to represent vessel arrival patterns, which in turn, is deployed in a maritime simulation model. (Merrick *et al.*, 2003) analyze a proposed ferry service extension in San Francisco Bay. The authors also use a simulation model in their analysis and the maritime traffic activity and the visibility are two important factors in the study. While assessing visibility, the authors develop a statistical model of the visibility conditions which is fed into the simulation model. In this study, the authors also use the Analytical Hierarchy Process (AHP) methodology to analyze the expert opinions.

An analysis on Hong Kong port reveals that there is a certain risk pattern, and the accident types and vessel types are critical variables in this study (Yip, 2008). The author uses a negative binomial regression model in order to analyze the local marine accidents data.

There also exist many recent studies on the Strait of Istanbul. These studies include the construction of a simulation model enabling the maritime risk analysis. Environmental conditions, vessel arrival processes, transit scheduling of arrived vessels, navigation in the Strait are covered in the simulation model.

(Özbaş, 2005) creates a basic simulation model which involves basically the maritime transit traffic in the Strait of Istanbul, imitating the vessel entrance procedures and taking into account the overall Channel Traffic Rules and Regulations (R&R), vessel types, their cargo characteristics, the meteorological/geographical conditions and the pilot/tugboat services. Vessel arrival distributions are assumed to be exponential and their parameters are estimated from historical statistics. The visibility conditions are provided by external datasets. At the end of the study, an extensive scenario analysis is performed and the effects of the factors stated above are analyzed. It is concluded that the most significant factor affecting transit traffic is the pilot/tugboat availability and the second important factor is the arrival rate of the vessels.

(Almaz, 2006) extends the previous simulation model by integrating several new and updated modules. In this regard, a detailed arrival process, visibility and current module with current pilot/tugboat services, updated R&R, traffic lanes and overtaking conditions are integrated into the model. For the first three factors, a detailed historical data analysis is conducted: the arrival process distributions and parameters are estimated from transit vessel data of the year 2005. Visibility conditions are modeled with the help of the datasets provided from the Kandilli Observatory and Earthquake Research Institute (KOERI), the International Weather Information website and Directorate General of Coastal Safety. The fog occurrences and durations, which show strong seasonal patterns, are approximated by empirical and “Mixtures of Generalized Erlang” distributions. Hence, this new integrated model consists of three classes of parameters. Internal parameters (can be controlled by Channel Traffic Control Authority) are cited as pilot and tugboat availabilities, vessel priorities and overtaking rules. External parameters are the factors that cannot be controlled by the Control Authority, such as meteorological conditions, arriving vessel profile, arrival process, surface currents and pilot captain/tugboat requests. The third group is the R&R which are assumed unchangeable. The study continues with scenario analyses of two models: the basic model (which does not contain the current and visibility sub-models, overtaking rules and pursuit distances) and the integrated model (Almaz *et al.*, 2006). The scenario analysis performed on the basic model shows that pilot/tugboat availability, vessel arrival process and vessel properties are the significant factors of the model. On the other hand, the scenario analysis on the integrated model indicates that the new variables have an impact on the simulation outputs. For instance, total number of

passing vessels is influenced by the visibility conditions, the average transit time by the current, the average waiting time by the Kandilli rule. Moreover, pursuit distance is a powerful factor on all outputs.

(Or *et al.*, 2007) discuss the updating process of the integrated model of (Almaz *et al.*, 2006) with the new traffic regime that the Bureau of Turkish Strait's Maritime Traffic Services (BMTS) has applied in the Strait since the middle of 2005 (where the previous bi-directional regime is allowed only at night, while a new uni-directional traffic flow regime is implemented in the day). Hence, a new scheduling insight is created, which is based on the determination of time-window sizes for both unidirectional northbound and southbound vessel flows every morning, taking into account basically the vessel waiting times according their priority classes.

(Ulusçu *et al.*, 2009a) incorporate to the simulation model the scheduling process, which is applied currently by VTS as a compact and well-defined algorithm. The scheduling algorithm is based on parameters such as vessel priorities, minimum time displacement between vessels, waiting times, maximum operational duration per day, and number of vessels waiting at each entrance of the Strait. The basic idea of the formulation is scheduling the vessels with highest waiting time first, while giving priority to larger vessels carrying hazardous material. Due to vessel prioritization policies, the analysis is divided into two parts such as daytime and nighttime scheduling.

In (Ulusçu *et al.*, 2009b), safety risks of the maritime traffic in the Strait of Istanbul are analyzed through the risk simulation model developed in earlier studies and discussed above. The authors then build a dynamic risk measurement and assessment model onto the developed simulation model. The probabilistic risk model concerns the representation of maritime accident risks by a mathematical formulation based on conditional probabilities (regarding factors such as accidents cause, accident type and the damage it causes). These probabilities are derived based on the answers given by the experts of different background to the prepared questionnaires. The results are then checked for consistency and calibrated. Within this context, it is mentioned that the risk profile is highly related to vessel arrival rates, pilotage services, overtaking rules and local traffic density. Since the decrease in the local traffic in the Strait causes a significant decrease in overall risk in the Strait, it is

recommended that the passage of hazardous material carrying vessels can be scheduled mostly at night, when the local traffic is considerably low. Also, the vessels are encouraged to benefit more from pilotage services.

Apart from the simulation-based risk studies, maritime accidents in the Strait of Istanbul are analyzed statistically in (Ece, 2007). Firstly, the maritime accidents and the corresponding details are collected from several private and public data sources. After a detailed preliminary about the Strait, the frequencies corresponding to the vessel and accident attributes are presented. Then, chi-square and correlation analyses are conducted to find out the statistical relationships between the candidate attributes affecting accidents and near-miss cases. Moreover, discriminant and cluster analyses are applied in order to classify the accidents according to their causes. The study ends by several remedies to decrease accident risks in the Strait.

3. MODELING OF THE ARRIVAL PROCESS

3.1. Introduction and Description of the Data

Vessels transiting the Strait of Istanbul are the key components of maritime accidents in the Strait: if there were not any transit passages, there would be no transit accident. Accordingly, it is very much worthwhile to study the arrival behavior of vessels intending to make a transit. Arrival process can be defined as the random generation of interarrival times of the vessels arriving to the Strait according various attributes. In the study of (Almaz, 2006), vessel interarrival times, lengths, speeds, anchoring durations, pilot/tugboat demands and direct/indirect passing status are generated mainly according to vessel type, length and entrance direction. This study focuses on vessel interarrivals of several vessel groups, which are obtained according to the priority classes (the classes actually deployed to use in the scheduling process), entrance direction, daytime, season and direct/indirect passing status; while using the most recent transit data (which covers the year 2007). Interarrival time inputs of each group are analyzed through Input Analyzer of ARENA, which selects the best distribution with corresponding parameters for the input data (Banks *et al.*, 2000). After the analyses are done separately for the years 2005, 2006 and 2007, for each sub-class; t, F and Mann-Whitney tests are performed to the groups 2005-2006, 2005-2007 and 2006-2007. The intention of these tests is to see whether the sample means, variances and distributions differ significantly as the time changes.

The transit vessel dataset is obtained from the Strait of Istanbul VTS Center and contains detailed information of all vessels passing from the Strait throughout a year (Figure 3.1). The elements recorded for all vessels are as follows: entrance direction, direct/indirect passing status, vessel type, vessel gross tonnage (GRT), vessel length, vessel speed, whether the vessel benefits from pilot and/or tugboat services, whether the vessel carries hazardous material or not, first contact time, ready time (time at which the vessel is ready to enter the Strait), entrance time (time at which the vessel is allowed to enter to the Strait), vessel flag type and age. The data contains 54,790 vessels for the year 2005, 54,880 for the year 2006 and 56,606 for the year 2007.

YONU	UGRAKSIZ	TIPİ	GROS TONAJ	BOYU	HAZ	PILOT ALMIS	ROMORK REFAKATLI	IMDG1,2,7 TASIYOR	ILK KONTAKT	GIRISE HAZIR	GIRISE ZAMANI	BAYRAK DEGERI	YASI	YAS GRUBU
K-G	E	GENERAL CARGO	2.205	82,2	10,0			E	03.01.2007 07:27	03.01.2007 07:27	03.01.2007 15:15	IYI	30	YASLI
K-G	E	GENERAL CARGO	2.345	76,0	11,0				03.01.2007 08:16	03.01.2007 08:16	03.01.2007 15:16	IYI	16	YASLI
K-G	E	TANKER	18.023	167,0	12,5				03.01.2007 02:21	03.01.2007 02:21	03.01.2007 15:27	ORTA	20	YASLI
K-G	E	GENERAL CARGO	7.641	130,0	11,0				03.01.2007 10:23	03.01.2007 10:23	03.01.2007 15:50	ORTA	24	YASLI
K-G	E	GENERAL CARGO	2.352	95,4	12,0				03.01.2007 10:47	03.01.2007 10:47	03.01.2007 15:55	ORTA	37	YASLI
K-G	E	RORO CARGO	10.931	161,4	12,0				03.01.2007 04:00	03.01.2007 04:00	03.01.2007 16:01	IYI	29	YASLI
K-G	E	GENERAL CARGO	2.360	108,0	8,5				03.01.2007 10:59	03.01.2007 10:59	03.01.2007 16:08	IYI	27	YASLI
K-G	H	GENERAL CARGO	5.157	126,5	10,0				03.01.2007 11:29	03.01.2007 11:29	03.01.2007 16:24	ORTA	32	YASLI
K-G	E	GENERAL CARGO	3.990	109,1	11,0				03.01.2007 11:39	03.01.2007 11:39	03.01.2007 16:28	IYI	23	YASLI
K-G	E	GENERAL CARGO	22.046	186,4	10,5	E			02.01.2007 14:23	02.01.2007 14:23	03.01.2007 16:37	IYI	25	YASLI
K-G	E	BULK CARRIER	17.770	188,2	12,1	E			02.01.2007 17:42	02.01.2007 17:42	03.01.2007 17:09	IYI	24	YASLI
K-G	E	BULK CARRIER	24.987	181,5	10,0	E			03.01.2007 11:09	03.01.2007 11:09	03.01.2007 17:37	IYI	10	ORTA
K-G	E	LPG BARGE	26.207	197,0	13,0	E	E		02.01.2007 18:05	02.01.2007 18:05	03.01.2007 18:05	IYI	27	YASLI
K-G	E	BULK CARRIER	38.995	225,0	12,7	E			02.01.2007 02:39	02.01.2007 02:39	03.01.2007 18:16	IYI	9	ORTA
K-G	H	BULK CARRIER	27.011	189,9	11,7	E			03.01.2007 05:02	03.01.2007 05:02	03.01.2007 18:40	IYI	10	ORTA
K-G	E	TANKER	17.277	170,5	11,0	E			02.01.2007 21:28	02.01.2007 21:28	03.01.2007 18:56	IYI	21	YASLI
K-G	E	GENERAL CARGO	2.516	114,0	9,5				03.01.2007 17:16	03.01.2007 17:16	03.01.2007 19:08	IYI	30	YASLI
K-G	E	BULK CARRIER	12.765	158,8	10,0	E			03.01.2007 07:09	03.01.2007 07:09	03.01.2007 19:16	IYI	28	YASLI
K-G	E	GENERAL CARGO	2.466	114,0	8,5				03.01.2007 16:31	03.01.2007 16:31	03.01.2007 19:24	IYI	22	YASLI
K-G	E	BULK CARRIER	35.879	225,0	11,1	E			03.01.2007 12:00	03.01.2007 12:00	03.01.2007 19:39	IYI	12	ORTA
K-G	E	GENERAL CARGO	2.466	114,0	9,0				03.01.2007 16:38	03.01.2007 16:38	03.01.2007 19:46	ORTA	27	YASLI
K-G	H	TANKER	24.248	189,0	11,0	E			02.01.2007 23:35	02.01.2007 23:35	03.01.2007 20:01	IYI	20	YASLI
K-G	H	BULK CARRIER	745	64,8	7,0				03.01.2007 17:47	03.01.2007 17:47	03.01.2007 20:05	ORTA	53	YASLI
K-G	E	GENERAL CARGO	2.516	114,0	9,0				03.01.2007 17:59	03.01.2007 17:59	03.01.2007 20:18	IYI	22	YASLI
G-K	H	CONTAINER SHIP	8.583	134,4	17,5	E	E		03.01.2007 13:24	03.01.2007 13:24	03.01.2007 22:19	IYI	2	GENC

Figure 3.1. An example of the transit vessel dataset from year 2007

There are 53, 55 and 51 types of vessels in transit vessels data in the years 2005, 2006 and 2007, respectively. Naval vessels and submarines are extracted from the analysis since they are free from the rules and regulations. The remaining vessels are firstly classified into five general categories taking into account the cargo type they carry as follows:

- Passenger Vessels
- LNG-LPG Carrying Vessels
- Hazardous Material (HazMat) Carrying Vessels
- Tankers
- General Cargo Vessels

The number of vessels for each category and their frequency distributions concerning the years 2005, 2006 and 2007 are shown in Tables 3.1 and 3.2 respectively.

Table 3.1. Number of vessels transited in 2005, 2006 and 2007 according to the general categories

	2005	2006	2007
PASSENGER	1,331	1,681	1,747
LNG-LPG	332	360	376
HAZMAT	845	1,133	1,552
TANKER	9,161	9,166	9,151
GENERAL CARGO	42,954	42,372	43,614
OTHER	167	168	166
TOTAL	54,790	54,880	56,606

Table 3.2. Frequency distribution of vessels transited in 2005, 2006 and 2007 according to the general categories

	2005	2006	2007
PASSENGER	0.0243	0.0306	0.0309
LNG-LPG	0.0061	0.0066	0.0066
HAZMAT	0.0154	0.0206	0.0274
TANKER	0.1672	0.1670	0.1617
GENERAL CARGO	0.7840	0.7721	0.7705
OTHER	0.0030	0.0031	0.0029
TOTAL	1.0000	1.0000	1.0000

According to the scheduling rules applied in the Strait and based on these five vessel categories and their lengths, the vessels are then divided into seven priority classes, which can be observed in Figure 3.2.

		Type				
Length (m.)	Draft (m.)	Tanker	LNG-LPG	Carrying Dangerous Cargo	Dry Cargo	Passenger Vessels
< 50	< 15	Class E	Class C	Class E	Class D	Class P
50 - 100	< 15					
100 - 150	< 15					
150 - 200	< 15	Class B			Class C	
200 - 250	< 15	Class A				
250 - 300	> 15					
> 300	> 15	Class T6				

Figure 3.2. Priority classes used for scheduling

The number of vessels for each priority class and their frequency distributions concerning the years 2005, 2006 and 2007 are displayed in Table 3.3 and 3.4.

Table 3.3. Number of vessels transited in 2005, 2006 and 2007 according to the priority classes

	2005	2006	2007
CLASS A	2,193	2,320	2,282
CLASS B	2,660	2,576	2,762
CLASS C	8,645	8,973	8,393
CLASS D	36,875	36,145	38,428
CLASS E	2,919	2,997	2,737
CLASS P	1,331	1,681	1,747
CLASS T6	0	20	89
OTHER	167	168	166
TOTAL	54,790	54,880	56,606

Table 3.4. Frequency distribution of vessels transited in 2005, 2006 and 2007 according to the priority classes

	2005	2006	2007
CLASS A	0.0400	0.0423	0.0403
CLASS B	0.0485	0.0469	0.0488
CLASS C	0.1578	0.1635	0.1483
CLASS D	0.6730	0.6586	0.6789
CLASS E	0.0533	0.0546	0.0484
CLASS P	0.0243	0.0306	0.0309
CLASS T6	0.0000	0.0004	0.0016
OTHER	0.0030	0.0031	0.0029
TOTAL	1.0000	1.0000	1.0000

3.2. Methodology and Interpretations

First, the dataset is split into the given seven priority classes and then these groups are divided into several sub-classes, so that the behavior of class members are more coherent and the number of data points becomes appropriate for a distribution fitting study. The attributes used to construct the sub-classes are the entrance direction, daytime versus nighttime, season and the direct/indirect passing status. For the season attribute, three seasons are taken into account: the winter season covers December, January and February; the spring/autumn seasons cover March, April and November and the summer season covers the rest. The time of day attribute is divided into two categories: daylight is considered between 07.00 am-19.00 pm and nighttime between 19.00 pm-07:00 am. Since

for our purposes the key time factor is the time a vessel is “ready” to enter the Strait (rather than, for example, contact time or time at refueling or loading), interarrival times of the vessels in the dataset are computed based on the ready time column and calculated in minutes. The resulting interarrival times are analyzed through “fit all” command of the Input Analyzer. “Fit all” command estimates appropriate parameters for all applicable distributions defined in Input Analyzer, ranks them according to the respective square errors, from best to worst (namely, smallest to largest square error). It then returns the distribution having minimum square error as the best fit with the corresponding parameters.

After all appropriate sub-classes are analyzed through Input Analyzer; the next step is to decide whether the fit is good or bad, with the help of the goodness-of fit tests. Goodness-of-fit tests are helpful guides for the suitability of a potential probability distribution, while providing the discrepancy between the observed and expected values under the model in question. The two most well-known goodness-of-fit tests are chi-square and Kolmogorov-Smirnov (K-S) tests, which are reported by the Input Analyzer summary report. The chi-square test is based on comparing the histogram of the given data with the shape of the candidate density (or mass) function. The chi-square test can be applied to a continuous sample as well as a discrete one and it is even powerful for large-sized samples. On the other hand, K-S test is based on examining the quantile-quantile plots, it is used when no parameters are estimated from the sample and it is more functional for small sample sizes (Banks *et al.*, 2000). Since the samples displayed in the arrival process are large, the results of chi-square test are taken into consideration.

The chi-square test statistic is given by

$$\chi_0^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \quad (3.1)$$

where O_i is observed and E_i is expected frequencies in i^{th} class interval, k being total number of intervals. χ_0^2 approximately follows the chi-square distribution with $n-s-1$ degrees of freedom, where s represents the number of estimated parameters and n the sample size. The null and alternative hypotheses are:

- H_0 : the random variable, X , conforms to the distributional assumption with the parameter(s) given by the parameter estimate(s).
- H_1 : the random variable X does not conform.

H_0 is rejected when $\chi_0^2 > \chi_{\alpha, n-s-1}^2$.

The results of the goodness-of-fit tests are interpreted with the aid of p-values, namely, the largest value of the type-I error probability that allows the distribution to fit the data. A p-value larger than the significance level α (the probability of falsely rejecting H_0) indicates a good fit since the null hypothesis would not be rejected, whereas a smaller p-value means poor fit (Banks *et al.*, 2000). In this study, the significance level is chosen as 0.05.

As an example, histogram and distribution summary of “interarrival times (in minutes) of southbound Class D vessels passing directly at summer nights”, are displayed in Figure 3.3 and Table 3.5 respectively. According to Input Analyzer, the distribution with the minimum square error is found to be the exponential distribution with parameter λ equal to 51.5. Moreover, the p-value of the chi-square goodness-of-fit is 0.687, which is quite high. Since the p-value is greater than $\alpha=0.05$, the null hypothesis cannot be rejected: accordingly, it can safely be assumed that the sample consisting of interarrival times drawn from this sub-class conforms to the $Exp(51.5)$ distribution.

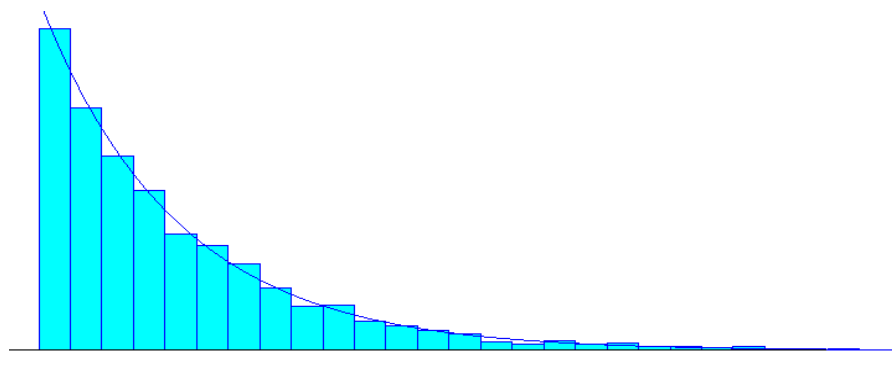


Figure 3.3. Histogram of interarrival times (in minutes) of southbound Class D vessels passing directly at summer nights

Table 3.5. Fitted distribution summary for interarrival times (in minutes) of southbound Class D vessels passing directly at summer nights

Distribution Summary	
Distribution:	Exponential
Expression:	EXPO(51.5)
Square Error:	0.000177
Chi Square Test	
Number of intervals	= 21
Degrees of freedom	= 19
Test Statistic	= 15.5
Corresponding p-value	= 0.687
Kolmogorov-Smirnov Test	
Test Statistic	= 0.0157
Corresponding p-value	> 0.15
Data Summary	
Number of Data Points	= 2581
Min Data Value	= 0.117
Max Data Value	= 473
Sample Mean	= 51.5
Sample Std Dev	= 51.9
Histogram Summary	
Histogram Range	= 0 to 473
Number of Intervals	= 40

One of the most important factors that affect the goodness-of-fit test results is the sample size. “If very little data are available, then a goodness-of-fit test is unlikely to reject any candidate distribution; but if a lot of data are available, then a goodness-of-fit test will likely reject all candidate distributions” (Banks *et al.*, 2000). Therefore, the sample sizes for all the sub-classes, besides the aim of having a detailed analysis on vessel arrivals, are determined so that the tests do not definitely reject all distributions. For instance, from Table 3.3, it can be checked that Class D contains at least 36,000 vessels. Hence, since the sample size is quite high, it can be split into several sub-groups: the criteria used to divide Class D into sub-classes are direction, nighttime/daytime status, season and direct/indirect passing status, which form 24 classes in total. Therefore, it becomes possible to obtain many samples with more accurate sample sizes, which do not always reject the null hypothesis.

As indicated in the beginning of the chapter, “fit all” command provides parameter estimations for all candidate distributions. So, for some cases, more than one distribution may have square errors close to each other. In these cases, it is hard to select the best fit because the function with the smallest square error is not necessarily the best. When this

situation is encountered, the p-values of the fitted distributions are also compared for a better selection. For example, when the subset of southbound Class C vessels at winter nighttime is considered, resulting square errors of lognormal, exponential, Erlang and Weibull distributions are 0.000784, 0.00164, 0.00164 and 0.0017 respectively. Among these, lognormal distribution is accepted as the best fit by Input Analyzer, since it has the smallest square error. However, its p-value is <0.005 , which implies that the distribution does not really conform to the sample. On the other hand, the p-value associated with the exponential distribution is 0.0349. Hence, since its square error is not too high, it can be selected as the best fit. However, for some sub-classes, all p-values of the fitted distributions are <0.005 . In such cases, as Input Analyzer does, the distribution with the smallest square error is chosen as the best fit.

After all the probability distributions with their associated parameters are determined for all sub-classes, the results are passed to the risk-simulation team to be deployed in the risk-simulation model. The resulting distributions and estimated parameters are presented in Appendix A.

3.3. Trend Tests

In this part of the study, vessel interarrival times realized in the years 2005-2007 are compared. This analysis is about whether interarrival time behavior of vessel classes differs with respect to different years. For this purpose, for each sub-class, three interarrival data groups representing each year are created and pairwise analyses are conducted. Namely, three tests are applied for each sub-class to check the relations for the years 2005-2006, 2005-2007 and 2006-2007: the first test is the t-test, which compares the means of the groups in the trend analysis; the second one is the F-test, which compares the variances and finally the non-parametric Mann-Whitney test, which tests whether the samples come from the same distribution.

Some details about these three tests are as follows. The parametric t-test is used for comparing the means of two samples to see whether they come from the same population. The null and the alternative hypotheses in this test are as follows:

- $H_0: \mu_1 = \mu_2$
- $H_1: \mu_1 \neq \mu_2$

Next, the parametric F-test is used for comparing the variation of two samples. The null and alternative hypotheses in the F-test are as follows:

- $H_0: \sigma^2_1 = \sigma^2_2$
- $H_1: \sigma^2_1 \neq \sigma^2_2$

When the assumptions of parametric tests are not met, there is always a chance to conduct non-parametric tests which are generally less efficient, but in return, less restrictive in the assumptions (Field, 2005). The K-S and Mann-Whitney tests are the most widely used nonparametric tests to test the independence of two samples. The two-sample K-S test gives information about whether two independent samples come from the same distribution. Hence, the null and alternative hypotheses are:

- H_0 : Data variable X has equal cumulative probability distributions for the two samples, $F(x) = G(x)$.
- H_1 : $F(x) \neq G(x)$.

The K-S test has a power efficiency of 95% when compared to parametric t-test and is especially efficient in testing small samples. Alternatively, the Mann-Whitney test has the same aim as the previous test and the elements of the samples must be at least ordinal. It operates under the philosophy of ranking the data and then analyzes the differences in the ranked positions of scores in different groups instead of dealing with the original data itself (Field, 2005). The power-efficiency of this test is 95.5%, when compared to the t-test and it is widely used for moderate and large sample-sized datasets.

Since the sub-classes introduced in Appendix A satisfy the normality assumptions, applying only the parametric t and F tests may suffice for trend analysis. However, as noted previously, the power of non-parametric Mann-Whitney test cannot be underestimated. Hence, this test is also included in the analysis, to check whether the samples in question follow the same distribution. The obtained results in terms of p-values

are presented in Appendix B: when p-value of one of these three tests is smaller than the 0.05 significance level, the null hypothesis is rejected, meaning that the two datasets differ in mean, variance or distribution.

Table 3.6 summarizes, for each priority class, the number and the frequencies of the sub-classes, where p-value <0.05 with respect to the pairwise year groups and the test types.

Table 3.6. The number and the frequencies of the sub-classes where p-value <0.05 for each priority class

		<i>T-TEST</i>			<i>F-TEST</i>			<i>M-W TEST</i>		
		2005-2006	2005-2007	2006-2007	2005-2006	2005-2007	2006-2007	2005-2006	2005-2007	2006-2007
CLASS A	<i>freq (p-val<0,05)</i>	0	0	0	2	1	4	0	1	0
	<i>prob (p-val<0,05)</i>	0.000	0.000	0.000	0.167	0.083	0.333	0.000	0.083	0.000
CLASS B	<i>freq (p-val<0,05)</i>	1	0	1	0	1	1	1	1	1
	<i>prob (p-val<0,05)</i>	0.250	0.000	0.250	0.000	0.250	0.250	0.250	0.250	0.250
CLASS C	<i>freq (p-val<0,05)</i>	1	0	2	1	1	2	0	0	3
	<i>prob (p-val<0,05)</i>	0.250	0.000	0.500	0.250	0.250	0.500	0.000	0.000	0.750
CLASS D	<i>freq (p-val<0,05)</i>	17	17	6	16	18	8	18	20	4
	<i>prob (p-val<0,05)</i>	0.708	0.708	0.250	0.667	0.750	0.333	0.750	0.833	0.167
CLASS E	<i>freq (p-val<0,05)</i>	0	1	2	2	1	4	0	2	2
	<i>prob (p-val<0,05)</i>	0.000	0.083	0.167	0.167	0.083	0.333	0.000	0.167	0.167
CLASS P	<i>freq (p-val<0,05)</i>	2	3	0	2	2	0	1	3	2
	<i>prob (p-val<0,05)</i>	0.333	0.500	0.000	0.333	0.333	0.000	0.167	0.500	0.333

When the sum of the frequencies for the cases where p-value <0.05 are taken into account (meaning that the sum of all columns in Table 3.6), it can be checked that the dissimilarity is maximum between years 2005-2007 and minimum between years 2006-2007 for all three test results. This means that, even if there are small changes in consecutive years, the change in the vessel arrivals grow up as the time passes and the inputs of this study require an update so that the most recent situation is achieved. When the frequencies are analyzed class by class, it can be inferred that the probabilities of having p-value <0.05 for smaller-sized classes (Class A, B, E and P) are smaller than that of large classes (Class D and C). This is due to the fact that the difference in mean, variance and distribution is slight for smaller classes, whereas the changes in larger samples are more significant. For instance, as displayed in Table 3.7, there is a significant change regarding the number of northbound direct and indirect passing Class D vessels occurring between 2005 and 2006, which directly affect the interarrival means and variances between 2005-2006 and 2005-2007. This change also increases the number of

cases where $p\text{-value} < 0.05$ for all three tests for the years in question, which can be checked from Table 3.6.

Table 3.7. Class D frequencies based on the direction and direct/indirect passing status

<i>Direction</i>	<i>Passing Status</i>	<i>2005</i>	<i>2006</i>	<i>2007</i>
<i>S-N</i>	<i>Indirect</i>	7009	9047	9645
<i>S-N</i>	<i>Direct</i>	11326	8995	9508
<i>N-S</i>	<i>Indirect</i>	7004	7888	8551
<i>N-S</i>	<i>Direct</i>	11532	10211	10719

4. MODELING OF THE VISIBILITY CONDITIONS

Adverse visibility conditions in the Strait of Istanbul (primarily caused by fog occurrences) not only contribute to the difficulty of navigation in the Strait, but also may lead to its closure as outlined below, leading to dangerous and unwanted vessel pile-ups. Accordingly, analysis of the visibility condition in the Strait of Istanbul is a worthwhile undertaking. Additionally, the results (in the form of arrival and duration distributions of fog related low visibility conditions) form an important input for the risk-simulation model.

4.1. Data Sources and Graphical Interpretations

According to the 36th article of the Strait of Istanbul and Marmara Sea Traffic Regulations, there exist two rules regarding visibility:

- (i) If the visibility in any part of the Strait is below 1 knot, the maritime traffic is kept open on one direction and closed on the reverse direction. In these cases, vessels carrying dangerous cargo, larger vessels, and deep draft vessels cannot enter into the Strait.
- (ii) If the visibility in any part of the Strait is below 0.5 knots, the traffic is closed on both directions.

Fog occurrences in the Strait and implied visibility conditions are analyzed through three different data sources:

- (i) Historical visibility conditions data covering the years 1975-2006, obtained from the Turkish State Meteorological Service.
- (ii) Historical meteorological data obtained from the international weather website “www.wunderground.com” (WU), covering the years 2000-2007.

- (iii) The data of the Channel closures between years 2004-2007, obtained from the Directorate General of Coastal Safety.

Among these three datasets, the information taken from the Turkish State Meteorological Service is the most extensive dataset in terms of overall duration, but includes only three measurements per day. This dataset is used to analyze the seasonality effect and to construct the visibility model, for the summer and the transition seasons (which are spring and fall). The data taken from the WU website covers a shorter period, but includes 30 minutes measurements and is used to construct the model of fog occurrences in winter and the duration of fogs in all seasons. The closure data obtained from the Coastal Safety and Salvage Administration is then used to check whether the model, based on the fog occurrence data, obtained from the WU website is compatible with the closure decisions. This data is also used to refine and validate the decision mechanism for closures of the Strait depending on the visibility conditions.

4.1.1. Turkish State Meteorological Service Visibility Data

The Turkish State Meteorological Service data includes three observations for every day taken at 07:00, 14:00 and 21:00 hours. The values showing the visibility distance in terms of kilometers are transformed into the binary system: 1 refers to fog conditions necessitating closure of the Strait and 0 refers to non-closure conditions.

Table 4.1. Number of fog observations for 32 years in Meteorological Service data

Month/Year	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	AVG
January	3	1	1	1	0	0	1	1	4	2	4	1	0	2	4	6	4	6	6	7	1	3	8	4	3	0	0	6	0	2	6	6	2.91
February	1	0	1	2	1	2	1	0	0	2	6	1	0	2	4	4	4	8	2	10	7	7	4	6	1	0	1	8	3	6	7	7	3.38
March	2	6	0	5	0	0	3	6	1	6	6	10	2	0	11	3	7	4	4	4	3	6	2	1	3	1	1	1	2	3	3	0	3.31
April	1	8	1	0	2	4	1	3	0	5	2	7	3	5	4	2	4	7	1	7	4	12	3	4	3	1	2	5	2	1	1	2	3.34
May	4	3	5	1	0	3	1	2	0	0	3	4	1	1	0	1	1	4	4	0	3	3	1	1	0	1	0	1	1	0	4	0	1.66
June	1	0	1	0	0	0	1	0	0	0	1	1	1	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.28
July	1	0	2	0	0	0	1	1	0	0	0	2	1	2	1	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0.44
August	0	2	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	2	0	2	0	0	0	1	0	0	0	0	0	0	0	0.31
September	0	1	2	2	0	1	0	0	0	1	3	0	0	1	0	0	0	0	0	2	0	0	0	0	0	0	0	3	0	1	1	0	0.56
October	4	2	3	3	1	0	0	0	2	1	4	0	0	0	2	2	2	1	2	2	1	2	1	2	1	3	0	1	1	0	0	0	1.34
November	1	3	2	4	0	3	1	1	1	5	1	6	3	2	4	4	1	1	2	5	6	10	0	0	1	5	1	6	7	1	0	2	2.78
December	2	1	0	3	2	0	1	0	1	3	7	4	0	3	4	2	1	1	6	8	1	1	1	1	1	2	4	0	2	1	1	9	2.28
TOTAL	20	27	18	22	6	13	11	14	9	26	37	36	12	20	35	24	24	33	31	45	28	44	20	19	14	13	9	31	18	15	23	26	22.59

Monthly and yearly frequencies of fog occurrences are displayed in Table 4.1, each cell in the table indicates the number of foggy days in a month. In this table, the seasonality effect can easily be detected. In order to better represent the seasonality effect, 31 days moving average of the daily fog occurrence frequencies over the indicated 32 years are calculated and shown in Figure 4.1. Moreover, sum of foggy days for 15-day periods are presented in Table 4.2.

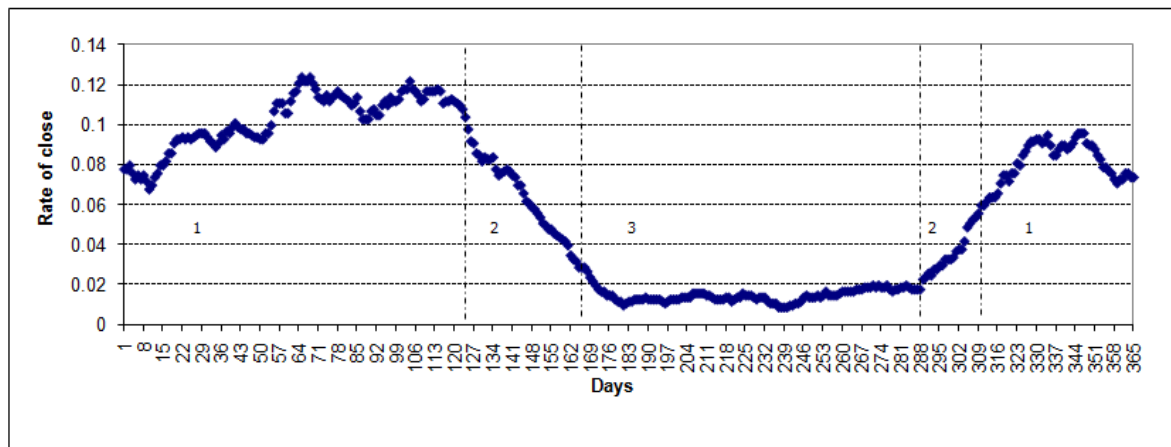


Figure 4.1. 31 days moving average of frequency distribution of daily fog occurrences over 32 years in Meteorological Service data

Table 4.2. Total number of foggy days for 15-day time intervals in 32 years

<i>Dates</i>	<i># of Fog Occurrences</i>	<i>Dates</i>	<i># of Fog Occurrences</i>
1-15 January	49	1-15 July	7
15-31 January	44	15-31 July	7
1-15 February	52	1-15 August	7
15-29 February	57	15-31 August	3
1-15 March	53	1-15 September	11
15-31 March	53	15-30 September	7
1-15 April	64	1-15 October	10
15-30 April	43	15-31 October	31
1-15 May	31	1-15 November	43
15-31 May	22	15-30 November	46
1-15 June	6	1-15 December	43
15-30 June	3	15-31 December	30

Based on the statistics presented in Figure 4.1 and Table 4.2, a year is divided into three rough seasons (1.Winter, 2.Fall/Spring, 3.Summer) for the purpose of further analysis

of fog arrival and duration behavior. The corresponding months for these seasons are displayed in Table 4.3.

Table 4.3. Visibility Seasons

<i>(1) Winter</i>	<i>(2) Fall / Spring</i>	<i>(3) Summer</i>
November	May	16-30 June
December	1-15 June	July
January	16-31 October	August
February		September
March		1-15 October
April		

4.1.2. Weather Underground Visibility Data

The WU data consist of information about visibility conditions in terms of visible distance taken in half-hour intervals for the years between 2000 and 2007 from Yeşilköy. These visibility data are also converted to the binary form (where 1 refers to fog induced closure conditions and 0 refers to clear visibility). Hence, the distances below 1 knot (~1.85 km) are converted as 0 and above 1 knot are flagged as 1. However, since the measurement is done in Yeşilköy (which is about 10 km away from the Strait and enjoys milder conditions) that does not totally reflect the Channel visibility conditions. So, an adjustment is performed on the original data. The details of the adjustment process are presented in Section 4.2.3.

The frequencies of fog initiation times show that they have a fluctuating behavior throughout the day, being far more frequent during nighttime (Figure 4.2). Because of this fluctuation, accompanied by the scarcity of the data, it is hard to precisely fit a distribution modeling fog start times. However, it can be represented by the empirical distribution displayed in Figure 4.3. From the graph, it can be inferred that the fogs generally starts in nighttime rather than the daytime.

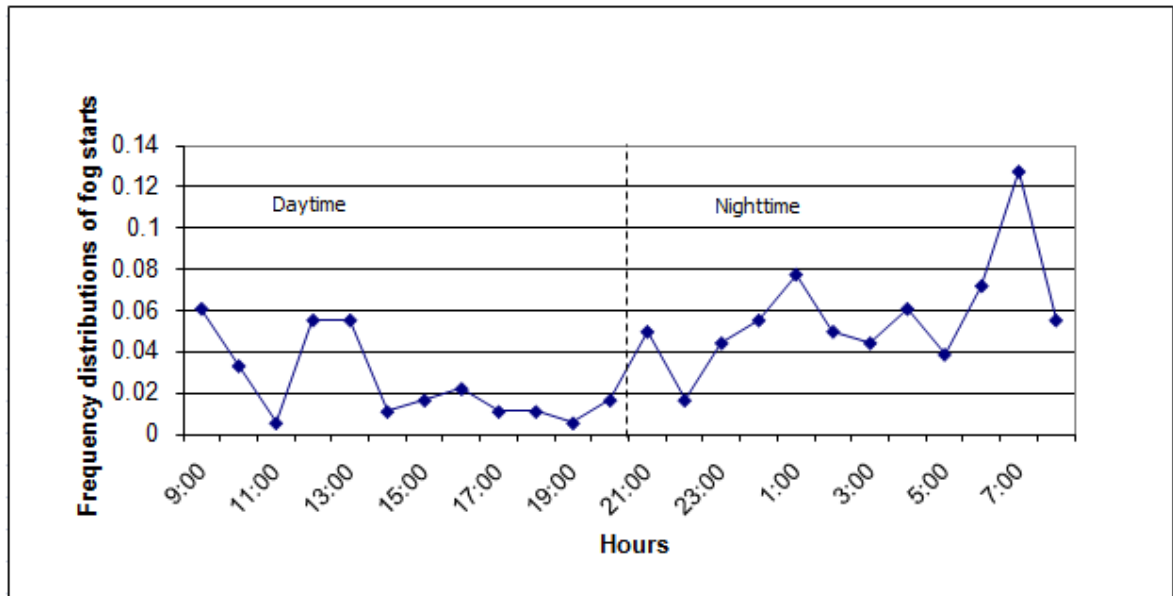


Figure 4.2. Frequency distributions of fog start times in hours of day based on the WU data

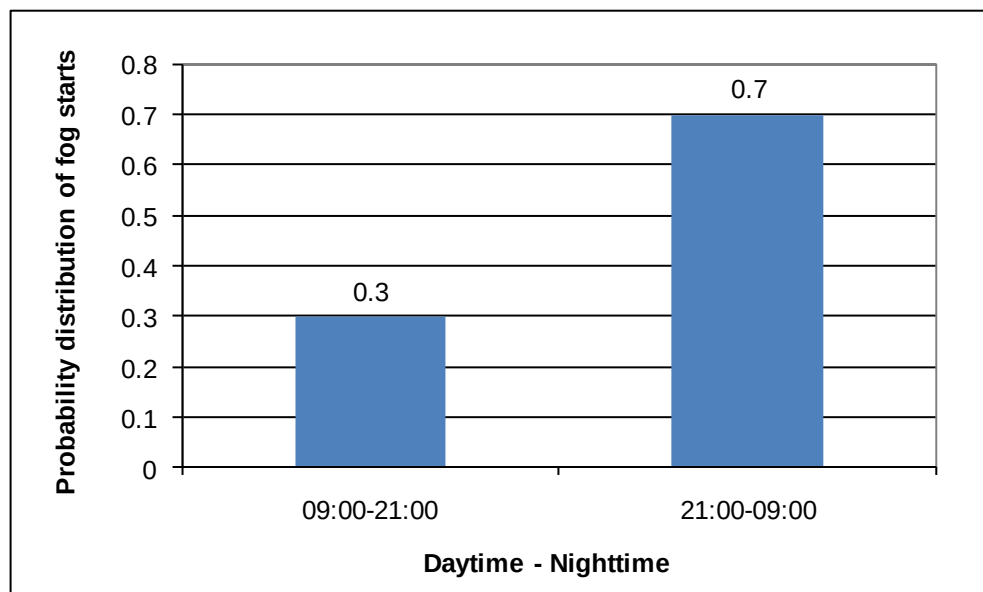


Figure 4.3. Probability distribution of fog start times in daytime and nighttime

It is important to analyze not only fog start times but also fog existence (or duration). Figure 4.4 displays the frequencies of “fog sightings” at different times throughout a day, where the numbers presented are averages obtained over the full duration of the WU data. Additionally, five periods moving average (namely, 2.5 hours) is taken into account when

constructing the figure, in order to decrease the fluctuation caused by the scarcity of the data. It can be observed from the figure that fog frequency begins to increase in the very early hours of the day and decreases all the way to the midnight. This behavior is compatible with the previous figure because the fogs mostly begin at night and can continue until morning.

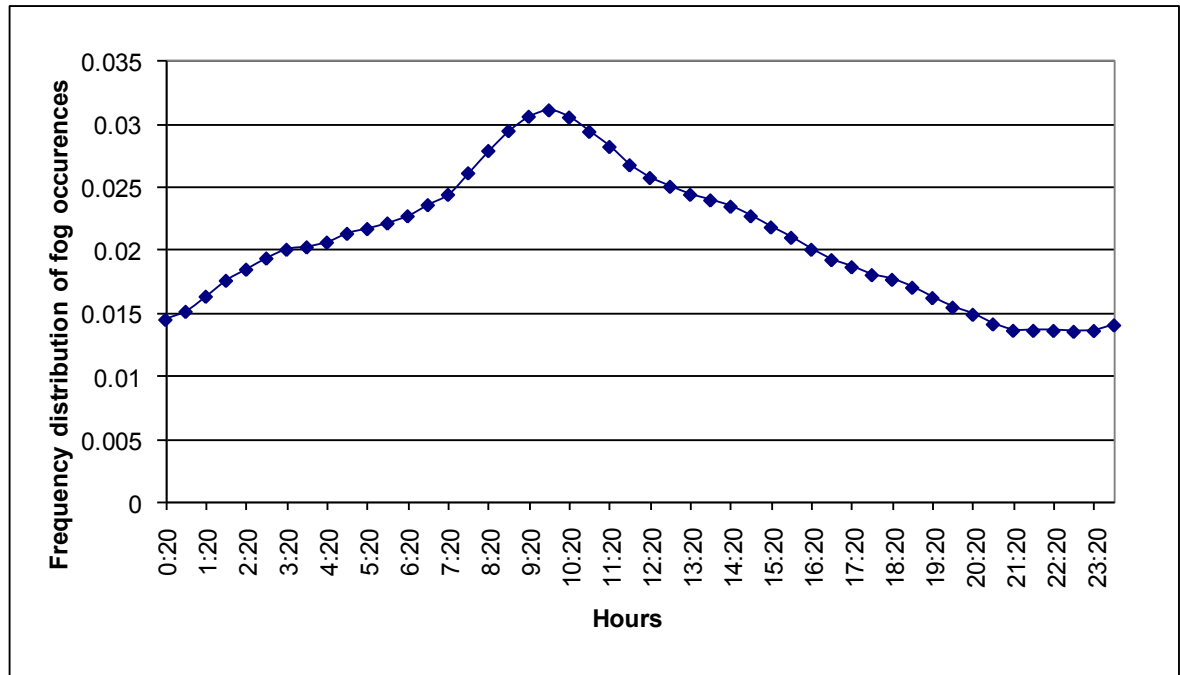


Figure 4.4. 2.5 hours moving average of frequency distributions of fog occurrences in hours of day in the WU data

The relationship between the fog durations and the fog start times can be observed in Figure 4.5. When the trend line is considered, fogs starting at night tend to last a little longer than fogs starting throughout the day. But this relation is very weak and in this study, it is assumed that the fog start times and fogs durations are independent.

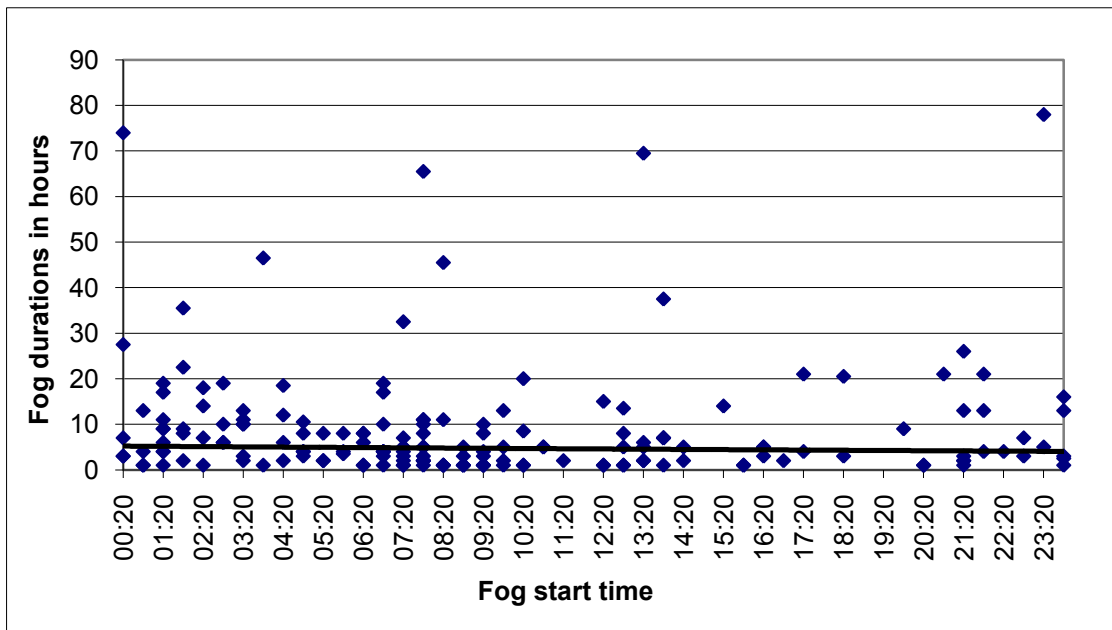


Figure 4.5. Scatter diagram of fog start times versus fog durations in the WU data

4.1.3. The Channel Closure Data of the Directorate General of Coastal Safety

The Channel closure data obtained from the Directorate General of Coastal Safety includes the start time, end time and direction (being one or two-way) information regarding fog (and other causes) induced closures in the Strait of Istanbul, spanning the period between April 2004 and December 2007. It may also be noted that in 2007, only two-way directional closures occur due to Marmaray project. Since the interval is not too large, precise distributions cannot be obtained regarding fog start times or durations, or the seasonality effect. But, this dataset is used to validate the models obtained from the other datasets.

To support the findings from the WU data, same graphical analyses are conducted with this dataset. Frequency distributions of the fog induced closure start times can be seen in Figure 4.6. It can be concluded that the start times of the Channel closures due to fog are more likely to fall into nighttimes than into daytimes. Again it is hard to fit a probability distribution to this fluctuating and scarce data. But, when the frequencies are calculated for only by day and night, as seen in Figure 4.7, it is quite consistent with the WU data.

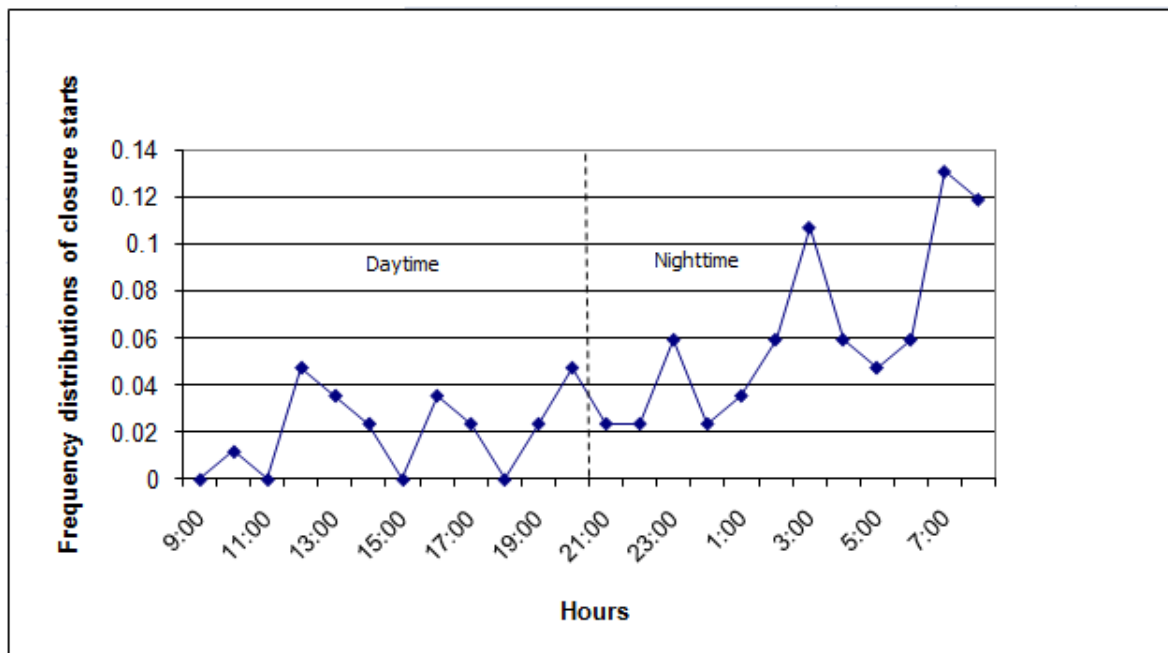


Figure 4.6. Frequency distribution of closure start times in hours

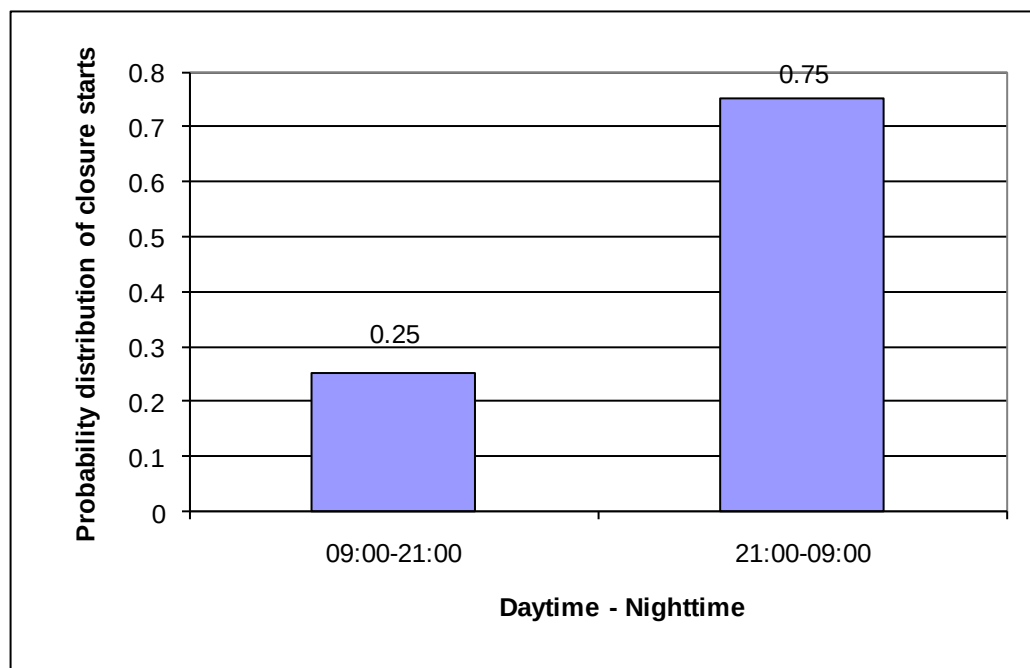


Figure 4.7. Probability distribution of closure start times in day and night in the closure data

Figure 4.8 shows the frequency of the closure times in the Strait. To distinguish the similarities between WU and Channel closure data, the closure data is adjusted as it indicates good and bad visibility conditions being a 0 or 1 taken in half hour intervals. As in the WU data, the fluctuations caused by the scarcity of the data are reduced by taking five periods (2.5 hours) moving average. It can be seen from the figure that it is in accordance with WU data.

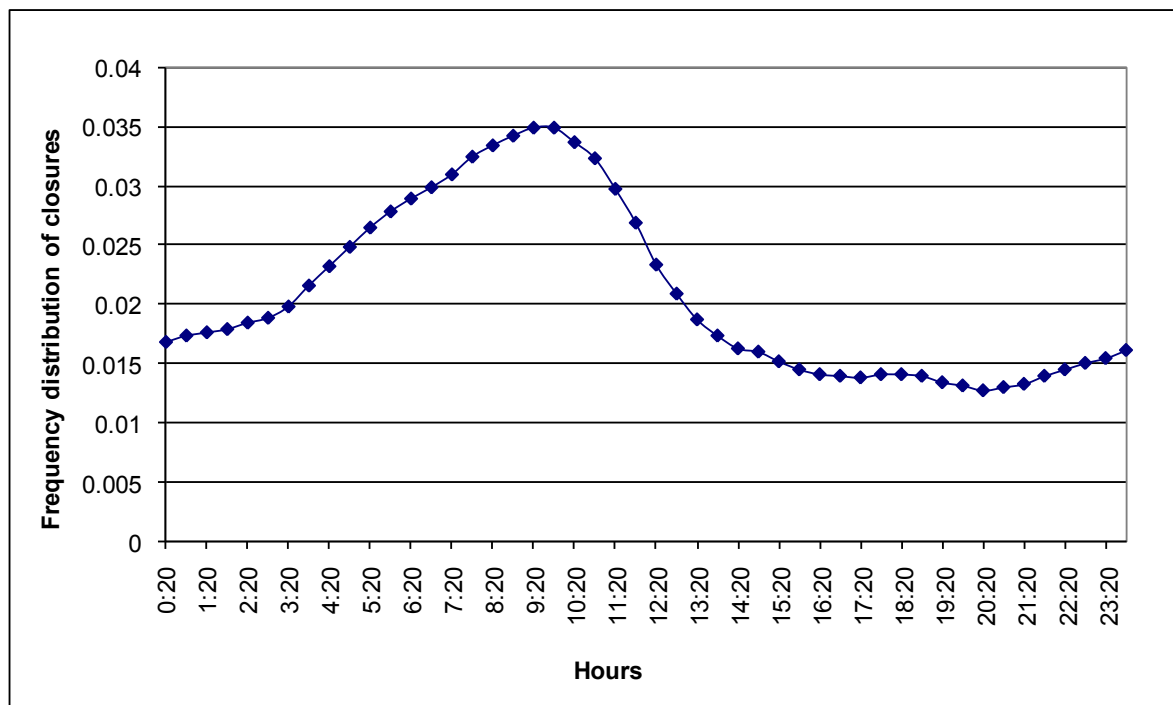


Figure 4.8. 2.5 hours moving average of frequency distributions of fog occurrences in hours of day in the closure data

In addition, the relationship between the duration of the closures and the start times of the closures are investigated. The trend line in Figure 4.9 shows that the duration of the closures which start at night is a little bit greater than the duration of the closures which start in the daytime. But, this is not a strong relationship and it can be inferred that, similar to the case of fog start times and durations, the start times and the durations of the closures are independent.

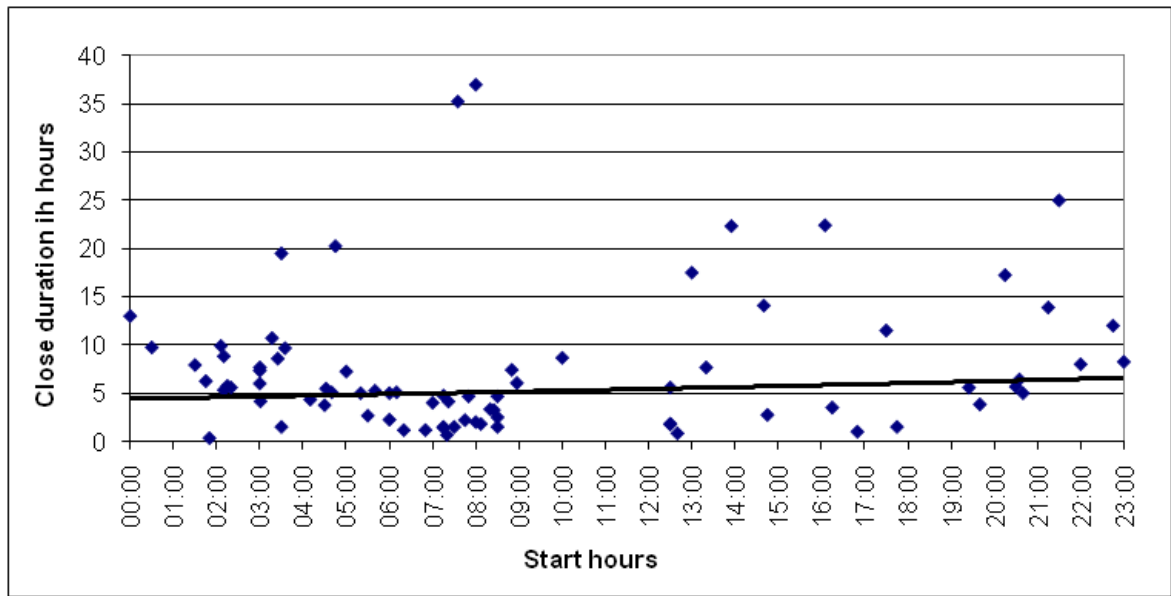


Figure 4.9. Scatter diagram of fog start times versus fog durations in the closure data

4.2. The Visibility Model: On/Off Process

4.2.1. General overview

As previously noted, fog occurrence in the Strait is one of the key factors affecting safe navigation and increasing the maritime accident probability. Hence, a much more detailed analysis than graphical interpretations is conducted to model the fog occurrence and durations. Within this context, the data interpreted above are reused to fit appropriate fog distributions and to estimate the model parameters, in a way to best reflect the real conditions occurring in the Strait. Good and bad visibility, namely, open and closed Channel conditions occur successively in the system, one following another. So, this can be thought as an on/off process, where on process indicates clear sight, whereas off process indicates fog conditions that require Channel closure. The model layout is presented in Figure 4.10.

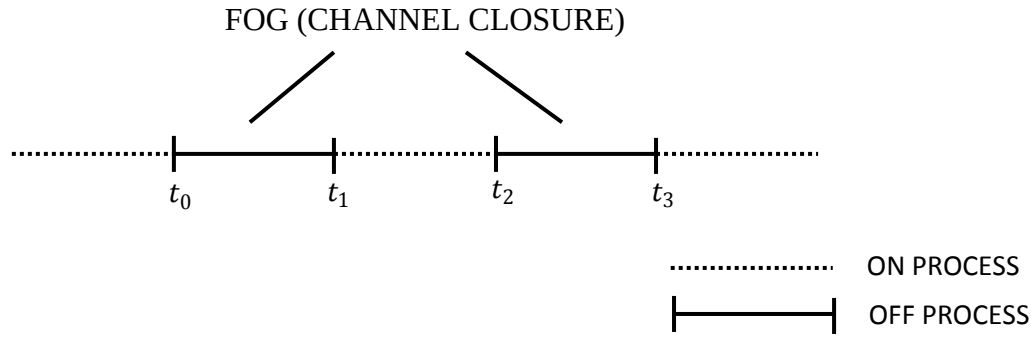


Figure 4.10. On/off process layout

As mentioned in section 4.1.1, the analyses of Turkish State Meteorological Service data showed that the fog occurrences are highly seasonal: low visibility conditions are more frequent in winter than in summer and transition seasons. Hence, three distinct fog initiation models are developed for each season. However, since the bad visibility conditions occur rarely in summer and transition seasons, exact distributions with appropriate parameters cannot be obtained for these seasons. But the fog initiations in winter can be modeled based on the WU data. In addition, for the fog duration model, it is inferred from the Figures 4.5 and 4.9, that there is no evidence of correlation between fog start times and durations. Hence, fog durations are uniquely modeled for all seasons. In order to model fog initiations in winter and fog durations in all seasons, mixtures of generalized Erlang (MGE) distributions are used and appropriate parameters are estimated from the WU data.

One of the key parameters giving general information about a random variable X is squared coefficient of variability, denoted by Cv_x^2 . Cv_x^2 is a measure of variability between elements of X . It is formulized as,

$$Cv_x^2 = \frac{V(X)}{E^2[X]} \quad (4.1)$$

Therefore, if the difference between the elements of a random sample X decreases, it means that Cv_x^2 of X gets closer to zero (Altıok, 1996). For instance, the mean and variance of the exponential distribution are $1/\lambda$ and $1/\lambda^2$ respectively, so, $Cv_x^2=1$. Since time points corresponding to fog and no fog conditions can be regarded as interarrival times, at

the first look, their modeling via the exponential distribution seems like a good idea. However, it is found from the data that $Cv_x^2 > 1$ for both cases. Hence, a more flexible distribution ought to be chosen: phase-type distribution, which is frequently used in stochastic modeling, is a good alternative in this case. “A phase-type distribution can be considered as the distribution of the time until absorption in a Markov Chain with a single absorbing state” (Altıok, 1996). One of the sub-classes of phase-type distribution is the MGE distribution. Its squared coefficient of variability range is $(0, \infty)$ and it is commonly used in modeling manufacturing, computer and communication systems. It is a convolution of k successive exponential distributions and its structure is displayed in Figure 4.11.

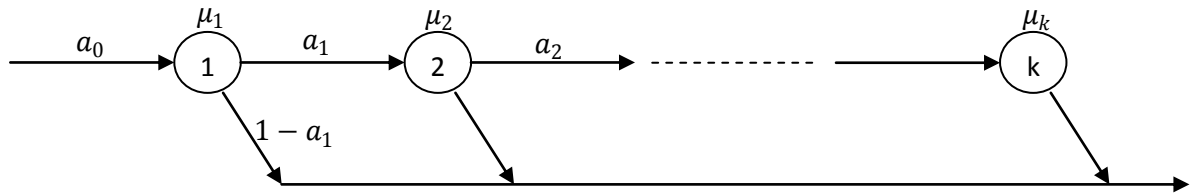


Figure 4.11. Graphical representation of the generalized Erlang distribution

According to this distribution, an entity enters the system with a probability of a_0 and spends there a certain amount of time, which is exponentially distributed with a rate of μ_1 . Hence; the average spent time is equal to $1/\mu_1$. After completion of this stage, the entity enters a new phase with a probability of a_1 or leaves the system with a probability of $(1 - a_1)$. If it enters to the next phase, the time it spends there is $1/\mu_2$ on the average. The last process is repeated until k^{th} stage is completed. However, the calculation of all these parameters is cumbersome. In other words, there is a need for simplifying the model with appropriate mathematical formulations. This simplification is performed via the moment-matching approximations.

Moment-matching approximation method is a good way of estimating the MGE distribution parameters (Altıok, 1996). There are two distinct approaches in calculation depending on Cv_x^2 ((i): $Cv_x^2 > 1$ and (ii): $Cv_x^2 < 1$). Since both on and off processes of the WU data have $Cv_x^2 > 1$, the first case is considered: the model presented in Figure 4.11 is simplified as in Figure 4.12. In this case, the entity certainly enters the first phase (a_0 is equal to 1 in this case), spends there an average time of $1/\mu_1$ and continues to enter

following $(k-1)$ phases of rate μ_2 successively with a probability of a or leaves the system with a probability of $(1 - a)$. When the visibility issues are taken into consideration within this context, successive MGE distributions will indicate the good (on process) and bad visibility (off process) conditions successively. Hence, two distinct MGE distributions representing these two processes are derived from the historical data, which occur one after another. This means that, if clear visibility condition is modeled with MGE- k distribution, the average time that it continues without fog occurrences is $1/\mu_1$ with a probability of $(1 - a)$ and $1/\mu_1 + (k-1) * 1/\mu_2$ with a probability of a . After the process is completed, the MGE- p distribution of the off process starts and the average time that it continues with fog occurrences is $1/\mu_1$ with a probability of $(1 - a)$ and $1/\mu_1 + (p-1) * 1/\mu_2$ with a probability of a . Then, it again returns to the MGE- k model of the on process and the same processes are repeated.

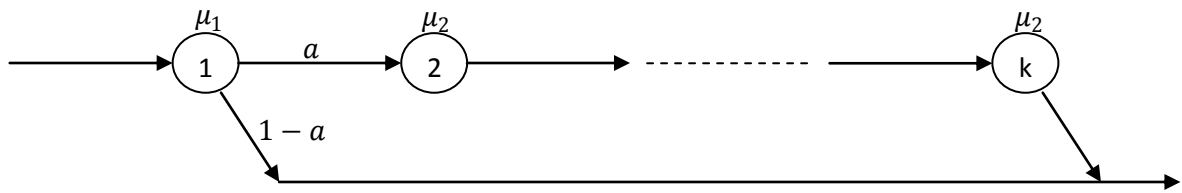


Figure 4.12. MGE- k distribution with $Cv_x^2 > 1$

4.2.2. Modeling of Fog Initiations in the Summer and Transition Seasons

Fog initiations in summer and transition seasons are modeled via empirical probability distributions. To this end, the dataset of Turkish State Meteorological Service is used, because it contains more precise information about these two seasons and covers 32 years. So, it is suggested that in these time periods, fog occurrences for each day should be randomly generated, based on the probability distributions presented in Tables 4.4 and 4.5.

Table 4.4. Probability distribution of fog occurrences in summer

<i>Status</i>	<i>Frequency</i>	<i>Probability</i>
<i>Fog</i>	55	0.0141
<i>Clear</i>	3849	0.9859
<i>Total</i>	3904	

Table 4.5. Probability distribution of fog occurrences in transition season

<i>Status</i>	<i>Frequency</i>	<i>Probability</i>
<i>Fog</i>	90	0.0454
<i>Clear</i>	1894	0.9546
<i>Total</i>	1984	

4.2.3. Comparison of WU and Channel Closure Data

As noted in previous sections, the original WU data contain visibility distances whose values are taken from the Yeşilköy station. Hence, a comparison with a dataset better representing the characteristics of the Strait can give indication about whether this dataset is a good choice to estimate parameters or not. For this purpose, general characteristics of the Channel closure data are taken as the reference. The mean, variance and squared coefficient of variation of good/bad visibility conditions based on the original WU data and of open/close status in the Channel closure data are displayed in Figures 4.6, 4.7, 4.8 and 4.9.

Table 4.6. Mean, variance and squared coefficient of variation of good visibility conditions in the WU data

<i>Mean (days)</i>	6.650407
<i>Variance</i>	160.8147
<i>Cv²</i>	3.636046

Table 4.7. Mean, variance and squared coefficient of variation of bad visibility conditions in the WU data

<i>Mean (hours)</i>	2.969613
<i>Variance</i>	18.03657
<i>Cv²</i>	2.045287

Table 4.8. Mean, variance and squared coefficient of variation of open state of the Channel

<i>Mean (days)</i>	6.966645
<i>Variance</i>	125.589
<i>Cv²</i>	2.587642

Table 4.9. Mean, variance and squared coefficient of variation of closed state of the

Channel	
<i>Mean (hours)</i>	8.710913
<i>Variance</i>	160.8502
<i>Cv²</i>	2.119797

Since open Channel conditions last very much longer than that of closed, for ease of interpretation, the values in on process are calculated in days, whereas closed conditions are in hours. One side effect of this time unit change is that the difference between the WU and the Channel closure data is much more distinctive in the “closed” case because of the smaller time unit selected. As it in the case of the arrival process, t, F and Mann-Whitney tests are performed to see whether these two datasets may come from the same distribution. p-values of these tests obtained for the comparison of the original WU and Channel closure datasets are tabulated in Table 4.10.

Table 4.10. Test results of the comparison between the original WU and Channel closure data

	<i>T-TEST</i>	<i>F-TEST</i>	<i>M-W TEST</i>
<i>WU Original (ON) / Channel (OPEN)</i>	0.491	0.737	0.001
<i>WU Original (OFF) / Channel (CLOSED)</i>	0.000	0.000	0.000

It can be observed that p-values for t, F, Mann-Whitney tests for the off process are smaller than the significance level 0.05, which means that the data obtained from the WU and Channel closure data do not follow the same distribution. Hence, an adjustment is administered to the WU dataset, such that it better mimics the Channel closures. At this stage, it is assumed that fogs occurring in the Strait are more intense and longer than those occurring in Yeşilköy because of the geographical conditions. So, in order to have the WU data better reflecting the Channel closures, the lower limit of “clear sight” is redefined as 2.5 km (~1.35 knots) in terms of the WU visibility distances and then the fog durations are doubled.

This “adjustment” can be explained as follows: A 2.5 km visibility at Yeşilköy, along with the doubling of fog duration, is an indicator of a much denser fog at the Strait, due to the harsher meteorological conditions there. The new test results are displayed in Table

4.11. It can be seen that all p-values are greater than 0.05, which means that the means and variances of the datasets can be considered equal and the assumption that they follow the same distribution may be accepted. As a result, this adjusted dataset is used to model the on/off processes of the visibility conditions explained in Sections 4.2.4 and 4.2.5.

Table 4.11. Test results of the comparison between adjusted WU and Channel closure data

	<i>T-TEST</i>	<i>F-TEST</i>	<i>M-W TEST</i>
<i>WU Adjusted (ON) / Channel (OPEN)</i>	0.954	0.970	0.760
<i>WU Adjusted (OFF) / Channel (CLOSED)</i>	0.827	0.813	0.353

Moreover, comparison of the fog frequencies for adjusted WU and Channel closure datasets is made in Table 4.12: 43 fogs out of 85 occur at the same time with the Channel closures, which means that these two datasets are 50% compatible. Moreover, total number of fogs occurred in the WU data is compatible with the total number of closures.

Table 4.12. Fog frequency comparison between adjusted WU and Channel closure data

	<i>2004</i>	<i>2005</i>	<i>2006</i>	<i>2007</i>	<i>TOTAL</i>
<i>Common</i>	4	18	16	5	43
<i>Only in WU data</i>	2	11	19	10	42
<i>Only in Closure data</i>	5	16	15	15	51

4.2.4. Modeling of Fog Initiations in the Winter Season – On Process

According to the adjusted WU data, on process is the case where there do not exist any low visibility condition requiring Channel closure. The duration constituting on process is calculated by subtracting the finish time of a fog from the start time of the next fog. The mean, variance and squared coefficient of variation of the WU on-period data (in days) is displayed in the Table 4.13.

Table 4.13. Mean, variance and squared coefficient of variation of on-period data

<i>Mean (days)</i>	6.878787879
<i>Variance</i>	94.40958864
<i>Cv²</i>	1.995226805

Since moment matching approximation is chosen as the methodology, the first step is to calculate the first three moments of the on-period sample, which are tabulated in Table 4.14.

Table 4.14. The first three moments of on-period data

m_1	6.878787879
m_2	141.114262
m_3	4213.287874

In order to test whether this data fits to a MGE-2 distribution, the necessary condition is given by the formula ($Cv_x^2=c$ for ease of notation),

$$\frac{m_3}{m_1^3} > \frac{3}{2}(c+1)^2 \quad (4.2)$$

In this case, $12.9445 < 13.4571$, and the condition is not satisfied. So, it is required to add more phases to the MGE-distribution and the number of phases k is calculated as the smallest integer satisfying,

$$k \geq \left(\frac{m_3 \cdot m_1}{m_2^2} - 1 \right)^{-1} \quad (4.3)$$

The right-hand side of the equation is 2.1957, so, $k=3$ and the on process is modeled by MGE-3 distribution displayed in Figure 4.13. This means that, when clear visibility condition begins, the average time T that it continues without fog occurrences is,

$$T = \begin{cases} \frac{1}{\mu_1}, & \text{with a probability of } (1 - \alpha) \\ \frac{1}{\mu_1} + 2 * \frac{1}{\mu_2}, & \text{with a probability of } \alpha \end{cases} \quad (4.4)$$

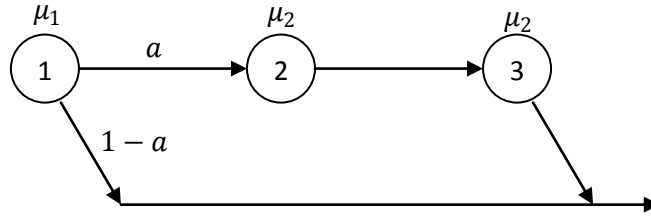


Figure 4.13. Graphical representation of MGE-3 distribution

The next step is to calculate α , which is an auxiliary variable provided for the calculation of model parameters. α is the smallest root satisfying,

$$\alpha^2[4m_1^2(k+1) - 3m_2k] + \alpha[km_3 - (4+k)m_1m_2] + (k+1)m_2^2 - km_1m_3 = 0 \quad (4.5)$$

The numerical version of this equation is,

$$-512.9448\alpha^2 + 5844.9981\alpha - 7294.0009 = 0 \quad (4.6)$$

Hence,

$$\alpha = 1.4265 \quad (4.7)$$

Then, model parameters are given by,

$$\mu_1 = \frac{1}{\alpha} \quad (4.8)$$

$$\mu_2 = \frac{k(m_1 - \alpha)}{m_2 - 2m_1\alpha} \quad (4.9)$$

$$a = \frac{k(m_1 - \alpha)^2}{(k-1)(m_2 - 2m_1\alpha)} \quad (4.10)$$

The results are presented in Table 4.15. According to these results, when clear visibility condition begins, the average time that it continues without fog occurrences is

1/0.701 days with a probability of 0.63296 and $1/0.701+2*1/0.134673$ with a probability of 0.36704.

Table 4.15. Model parameters of on-period data

μ_1	0.701026
μ_2	0.134637
a	0.36704

4.2.5. Modeling of Fog Durations – Off process

In previous sections, it is noted that the duration of fog occurrences is assumed to be independent from the fog start times and seasons. So, the fog duration will be modeled in a unique way for all seasons. Fog durations are calculated, in hours, by subtracting the start time from the finish time of a fog. The mean, variance and squared coefficient of variation values of off-periods in the WU data are presented in Table 4.16.

Table 4.16. Mean, variance and squared coefficient of variation of off-period data

<i>Mean (hours)</i>	8.836111
<i>Variance</i>	153.8012
<i>Cv²</i>	1.969869

As in the on-process, the first step is to calculate the first three moments which are displayed in Table 4.17.

Table 4.17. The first three moments of off-period data

m_1	8.836111
m_2	231.0236
m_3	11258.49

Then, the necessary condition for the MGE-2 distributions given by (4.2) is applied. Since $16.3191 > 13.2302$, MGE-2 distribution is selected for the off process and it is represented in Figure 4.14. MGE-2 means that, when bad visibility condition begins, the average time T that it continues with fog occurrences is,

$$T = \begin{cases} \frac{1}{\mu_1}, & \text{with a probability of } (1 - \alpha) \\ \frac{1}{\mu_1} + 1 * \frac{1}{\mu_2}, & \text{with a probability of } \alpha \end{cases} \quad (4.11)$$

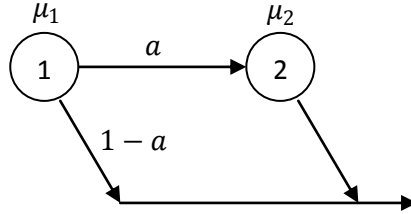


Figure 4.14. Graphical representation of the MGE-2 distribution

The appropriate formulae concerning model parameters are as follows:

$$a = \frac{\mu_2}{\mu_1} (m_1 \mu_1 - 1) \quad (4.12)$$

$$2m_1(\mu_1 + \mu_2) - m_2\mu_1\mu_2 = 2 \quad (4.13)$$

$$6m_1(\mu_1 + \mu_2)^2 - 6m_1\mu_1\mu_2 - 6(\mu_1 + \mu_2) - m_3\mu_1^2\mu_2^2 = 0 \quad (4.14)$$

These formulae can be simplified by the substitution of two new variables as follows:

$$X = \mu_1 + \mu_2 \quad (4.15)$$

$$Y = \mu_1 \cdot \mu_2 \quad (4.16)$$

These two variables in terms of three moments are given in (4.17) and (4.18).

$$Y = \frac{6m_1 - 3(m_2/m_1)}{\left(6m_2^2/4m_1\right) - m_3} \quad (4.17)$$

$$X = \frac{1}{m_1} + \frac{m_2 Y}{2m_1} \quad (4.18)$$

Based on the WU data, Y and X values are calculated as follows:

$$Y=0.01156 \quad (4.19)$$

$$X=0.26434 \quad (4.20)$$

Then, μ_1 and μ_2 can be calculated from equations,

$$\mu_1 = \frac{X + \sqrt{X^2 - 4Y}}{2} \quad (4.21)$$

$$\mu_2 = X - \mu_1 \quad (4.22)$$

and a from (4.10). According to the results displayed in Table 4.18, when a bad visibility condition starts, the average time that it stays in the system is 1/0.209018 hours with a probability of 0.775834 and 1/0.209018 + 1/0.55325 with a probability of 0.224166.

Table 4.18. Model parameters of off-period data

μ_1	0.209018
μ_2	0.055325
a	0.224166

5. STATISTICAL ANALYSIS OF MARITIME ACCIDENTS

This section contains various statistical analyses of past accidents and near accidents that have occurred in the Strait of Istanbul. First, frequency histograms are presented in order to give a general insight about the past accidents. Then, chi-square analysis is conducted to find out whether the vessel and accident attributes are statistically related or not. Next, discriminant analysis is conducted to classify the vessels according to the corresponding accident cause. Cluster analysis is used to group the vessels according to the selected attributes. Moreover, the frequency and chi-square analyses are also performed for the near-miss cases.

5.1. Data Description

All analyses of this section are based on historical data. The data taken into consideration comprise all vessel accidents that have occurred in the Strait of Istanbul between the years 1982 and 2007 and it is a compilation of various data sources, such as the Undersecretariat for Maritime Affairs database, the data included in (Ece, 2007), the VTS records and other sources. Moreover, some missing vessel attributes -mainly tonnage, length and age of the vessel- are supplied from international vessel databases through vessel names and/or known attributes. The consistency of the information compiled from all these sources is checked by pairwise comparisons. All the sources used in this process are stated in the reference section.

As a result, an accident dataset containing two groups of attributes is obtained:

- Attributes related to vessels that were involved in an accident: name, type, gross tonnage (GRT), length, age, flag and pilot request of the vessel.
- Attributes related to accidents associated with each vessel: date, place, cause of the accident, accident type, death toll, number of injured and missing people.

5.1.1. Vessel Data

The vessel data comprises 1128 cases, each containing the name, type, GRT, length, age, flag and pilot request of the vessel. Each case (namely, each row) introduces a vessel involved in an accident, with its associated accident attributes. Out of these 1128 vessels, 247 are local traffic vessels and 68 are of unknown type. Since the focus in this study is the analysis of accidents involving transit vessels, to render the data more reliable, these unknown-type vessels are only kept in frequency histograms and excluded from all other following analyses (chi-square, discriminant and cluster analyses).

5.1.2. Accident Data

Although the vessel data contain 1128 rows (in other words, 1128 vessels that were involved in accidents), there exist 729 distinct accidents in the data, since many accidents such as collision or ramming involve more than one vessel. As a consequence, the accident attributes of the vessel data are repeated in the cases where there exists a second or third vessel participating in the accident and this can lead to erroneous conclusions for accident based part of the analyses. To correct this situation, besides the original dataset, an accident dataset is formed containing only distinct accidents with the corresponding attributes (date, place, accident cause, accident type, death toll, number of injured and missing people). Moreover, in 167 accidents out of 729, only local traffic vessels were involved, whereas in the remaining 562 accidents, at least one transit vessel played a role. Hence, while 729 accidents are considered in frequency analyses, transit-based analyses are applied to the related 562 accidents.

5.1.3. Data Completion: Tonnage – Length Regression

Although the accident database is compiled with the aid of several data sources, it still has many missing data. The aim of this part of the study is to investigate whether the data can be rendered more complete and coherent through prediction and/or projection methods. Hence, simple linear regression is displayed as a tool to associate and predict the missing entries according to two key factors: length and GRT. As in the arrival process, the tonnage-length regression analysis is applied to the 2005-2007 vessel arrival data, since it

gives precise information about all vessel attributes. From this data, Pearson correlation coefficient between GRT and length is calculated as 0.902. This is a quite high value; so, one of these variables can give a good prediction for the other. Since the length of the vessel is thought to be more significant regarding the likelihood of being involved in an accident, vessel length is chosen as the dependent variable and GRT as the independent variable.

In the vessel data, in 472 cases out of 813, vessel tonnage values are known and length values are missing. Therefore, with the aid of the data on all transit vessels in the 2005-2007 period, it became possible to predict the missing length values through simple linear regression. For this purpose, all transit vessels dataset is divided to four categories according vessel types and four different regression models, where tonnage is explanatory and length is output variable, are developed. During this analysis, LNG/LPG and HazMat classes are grouped in one class. The general formula of the models is as follows:

$$\text{Length} = \beta * \text{GRT} + \text{Intercept} \quad (5.1)$$

The regression models obtained for four types are summarized in Table 5.1. The details of the models are given in Appendix C. From the appendix, it can be seen that the t-test p-values are equal to 0.000 for all models, which means that the corresponding β coefficients are significant. Moreover, F-test p-values are also equal to 0.000 which indicates that the models are significant.

Table 5.1. Tonnage-Length Regression Equations

<i>Vessel Type</i>	<i>Regression Equation</i>
<i>Passenger</i>	LENGTH= 84.969+0.003*GRT
<i>HazMat/Lng/Lpg</i>	LENGTH=95.227+0.004*GRT
<i>Tanker</i>	LENGTH=107.653+0.002*GRT
<i>General Cargo</i>	LENGTH=96.466+0.003*GRT

5.2. Descriptive Statistics

5.2.1. Vessel Descriptions

In this section, frequency distributions and histograms for vessel attributes associated to the vessel database are presented.

- Vessel type: As displayed in Table 5.2 and Figure 5.1, more than half of the vessels involved in accidents are of general cargo type. Moreover, when the ratio of accident frequency versus transit frequency (accident/transit ratio) is taken into consideration for each category, the maximum value is attained by the passenger vessels.

Table 5.2. Frequency of vessels according to their types

	<i>Accident Frequency</i>	<i>Accident Percent</i>	<i>Transit Frequency</i>	<i>Transit Percent</i>
GENERAL CARGO	642	56.90	128,940	77.82
LNG/LPG/HAZMAT	4	0.40	4,598	2.78
LOCAL	247	21.90	0	0.00
PASSENGER	60	5.30	4,667	2.82
TANKER	107	9.50	27,475	16.58
UNKNOWN	68	6.00	0	0.00
<i>Total</i>	1,128	100	165,680	100

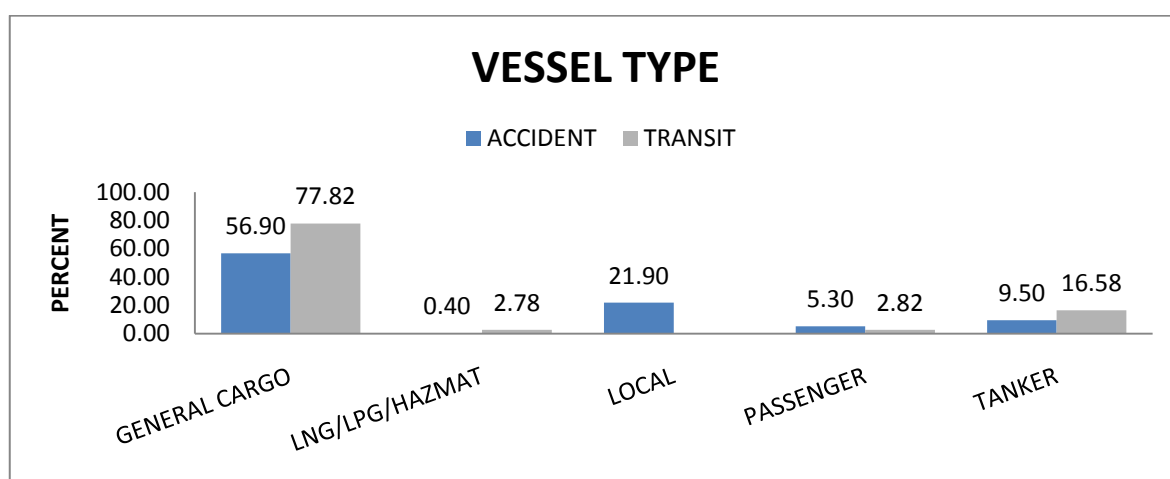


Figure 5.1. Frequency histogram of vessels involved in accidents according to their types

- Vessel GRT: As displayed in Table 5.3 and Figure 5.2, nearly 70% of the vessels involved in accidents are mostly in the (0-10,000 GRT) class. Moreover, accident/transit ratio is also maximized in the same cluster.

Table 5.3. Frequency of vessels according to their tonnages

	<i>Accident Frequency</i>	<i>Accident Percent</i>	<i>Transit Frequency</i>	<i>Transit Percent</i>
<i>0-10000</i>	788	69.90	131,296	79.25
<i>10000-25000</i>	101	9.00	19,671	11.87
<i>25000 and ABOVE</i>	52	4.60	14,713	8.88
<i>UNKNOWN</i>	187	16.60	0	0.00
<i>Total</i>	1,128	100	165,680	100

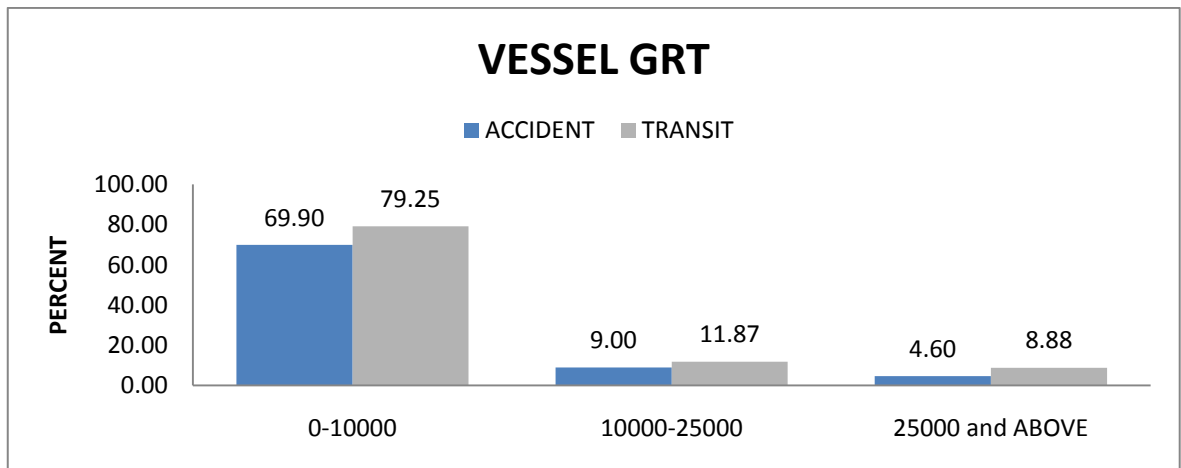


Figure 5.2. Frequency histogram of vessels involved in accidents according to their tonnages

- Vessel Length: As displayed in Table 5.4 and Figure 5.3, the vessels involved in accidents are mostly between 100 and 200 meters (with a percentage of 46.2%) while 3.2% of these vessels are of 200 meters or more in length. This statistics includes vessel length values predicted through the tonnage-length regression model mentioned in subsection 5.1.3. There is not a significant accident potential change according accident/transit ratio.

Table 5.4. Frequency of vessels according to their lengths

	<i>Accident Frequency</i>	<i>Accident Percent</i>	<i>Transit Frequency</i>	<i>Transit Percent</i>
0-100	296	26.24	64,984	39.22
100-200	521	46.19	89,887	54.25
200 and ABOVE	36	3.19	10,809	6.52
UNKNOWN	275	24.38	0	0.00
<i>Total</i>	1,128	100	165,680	100

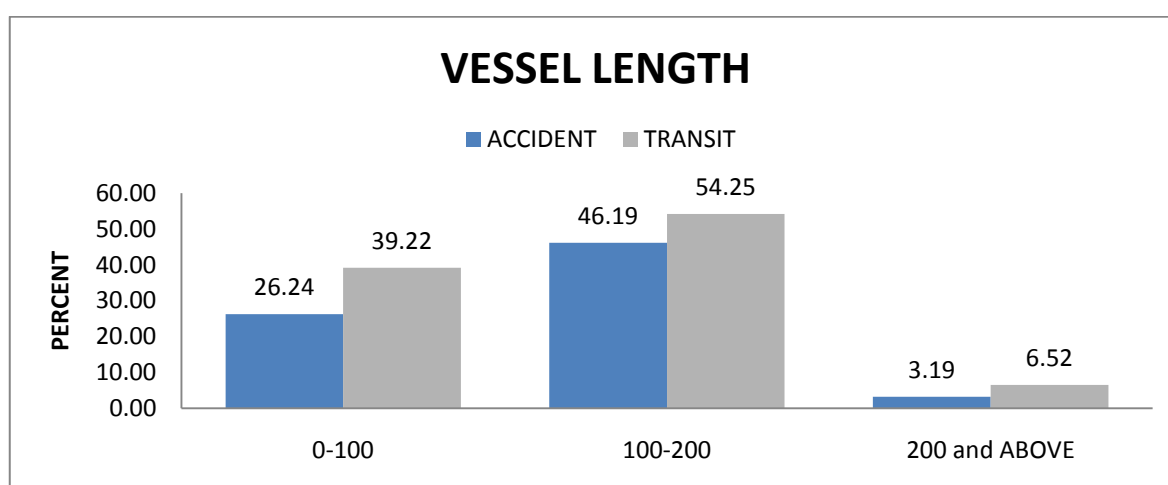


Figure 5.3. Frequency histogram of vessels involved in accidents according to their lengths

- Vessel Flag Type: As displayed in Table 5.5 and Figure 5.4, the highest percentage of flag type observed in vessels involved in accidents is of medium-risk (64.3%) vessels and the lowest percentage is of the high-risk vessels (2.7%). The accident/transit ratio attains its maximum at medium-risk vessel group.

Table 5.5. Frequency of vessels according to their flag types

	<i>Accident Frequency</i>	<i>Accident Percent</i>	<i>Transit Frequency</i>	<i>Transit Percent</i>
LOW-RISK	360	31.90	78,370	47.30
MEDIUM-RISK	725	64.30	80,186	48.40
HIGH-RISK	30	2.70	4,966	3.00
UNKNOWN	13	1.20	2,158	1.30
<i>Total</i>	1,128	100	165,680	100

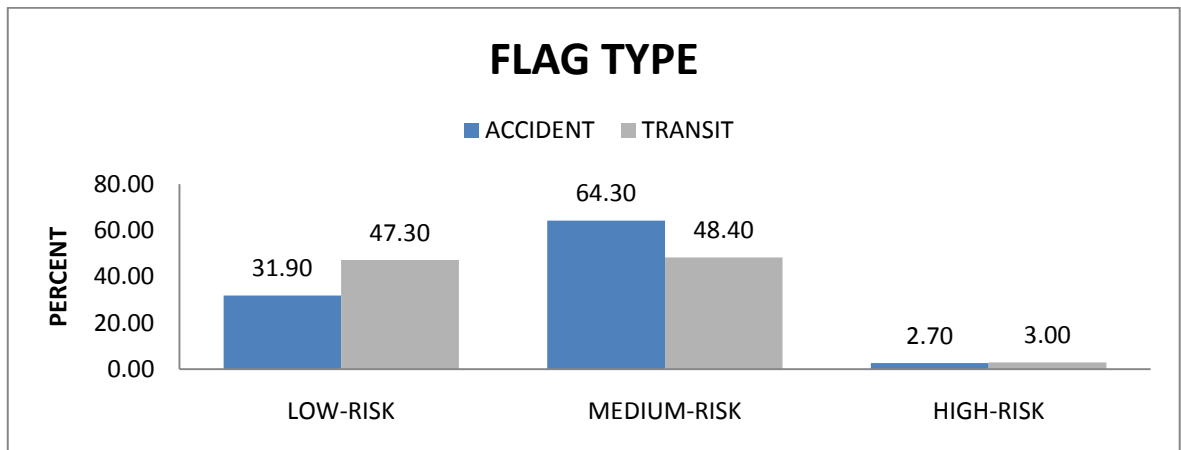


Figure 5.4. Frequency histogram of vessels involved in accidents according to their flag types

- **Vessel Age:** As displayed in Table 5.6 and Figure 5.5, the age of nearly half of the vessels that were involved in accidents are unknown. The lowest percentage belongs to the young group (6.3%). Moreover, the accident/transit ratio is maximized for middle-aged vessels.

Table 5.6. Frequency of vessels according to their ages

	<i>Accident Frequency</i>	<i>Accident Percent</i>	<i>Transit Frequency</i>	<i>Transit Percent</i>
<i>YOUNG</i>	71	6.30	19,619	11.84
<i>MIDDLE AGED</i>	117	10.40	13,798	8.33
<i>OLD</i>	416	36.90	130,915	79.02
<i>UNKNOWN</i>	524	46.50	1,348	0.81
<i>Total</i>	1,128	100	165,680	100

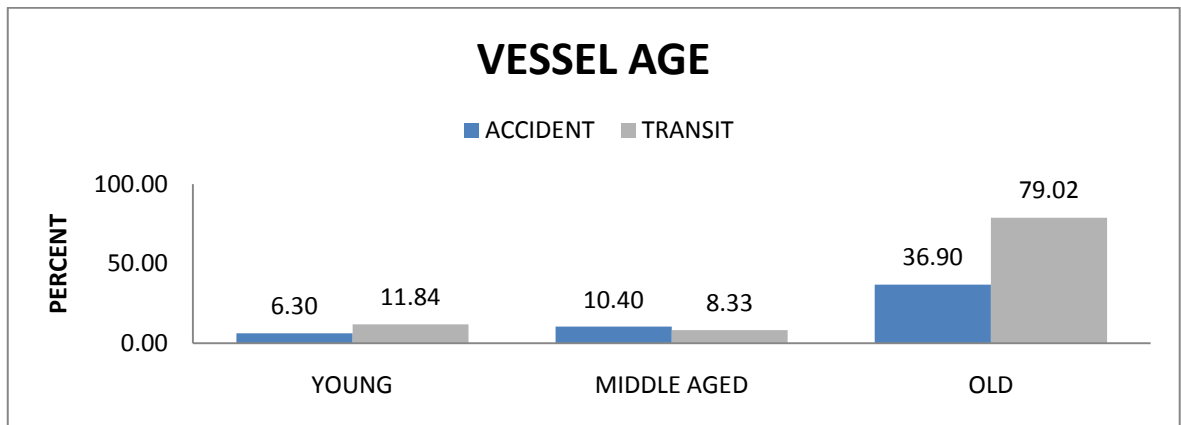


Figure 5.5. Frequency histogram of vessels involved in accidents according to their ages

- **Vessel Pilot Status:** As displayed in Table 5.7 and Figure 5.6, the majority of the vessels involved in accidents did not have a pilot on board. There is a significant difference between accident/transit ratios of these categories; which signifies that the accident potential of the vessels not having a pilot on board is much greater than those of having a pilot.

Table 5.7. Frequency of vessels according to their pilot status

	<i>Accident Frequency</i>	<i>Accident Percent</i>	<i>Transit Frequency</i>	<i>Transit Percent</i>
<i>0</i>	1,068	94.70	87,965	53.09
<i>1</i>	60	5.30	77,715	46.91
<i>Total</i>	1,128	100	165,680	100

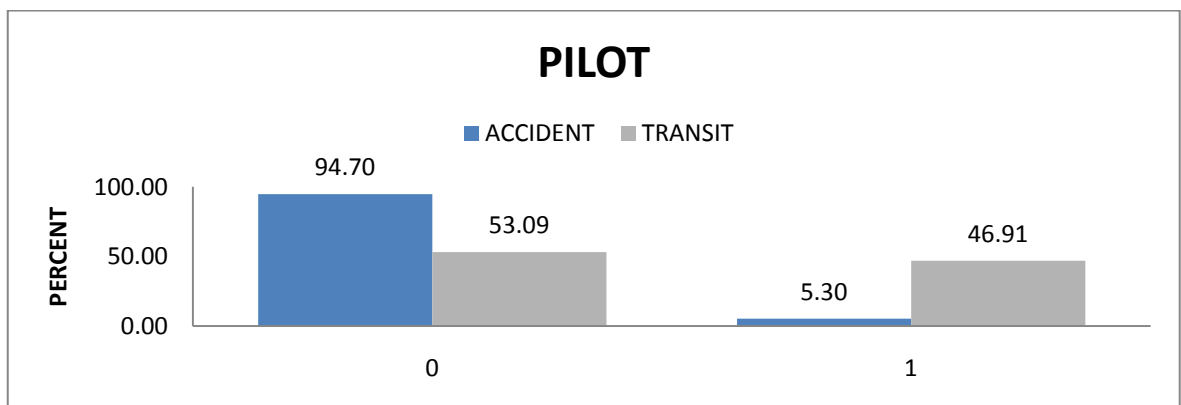


Figure 5.6. Frequency histogram of vessels involved in accidents according to their pilot status

5.2.2. Accident Descriptions

In this part, frequency distributions and histograms for the accident attributes associated with the accident database are presented.

- Accident Year: As displayed in Table 5.8 and Figure 5.7, most accidents occurred between the years 2003 and 2007 (24.3%) but the difference in years is slight.

Table 5.8. Frequency of accidents according to year

	<i>Frequency</i>	<i>Percent</i>
1982-1987	122	16.8
1988-1992	133	18.3
1993-1997	157	21.6
1998-2002	140	19.2
2003-2007	177	24.3
<i>Total</i>	729	100

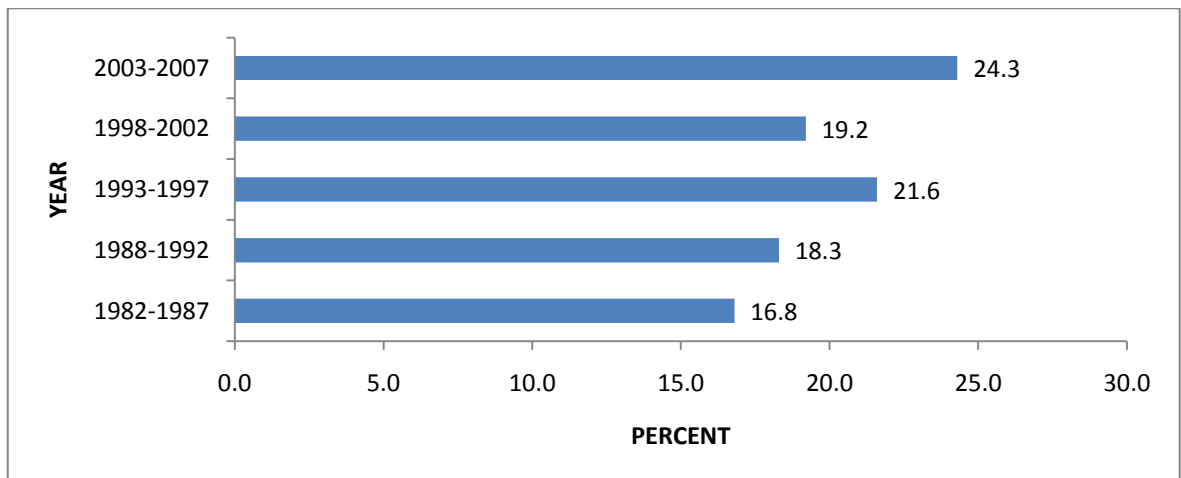


Figure 5.7. Frequency histogram of accidents according to year

- Accident Season: As displayed in Table 5.9 and Figure 5.8, accidents occurred more frequently in the winter season, and there are fewer accidents in spring (transition season) and summer. Moreover, accident/transit ratio is maximized for the vessels in winter season.

Table 5.9. Frequency of accidents according to season

	<i>Accident Frequency</i>	<i>Accident Percent</i>	<i>Transit Frequency</i>	<i>Transit Percent</i>
<i>SPRING</i>	89	12.20	29,717	17.94
<i>SUMMER</i>	190	26.10	56,627	34.18
<i>WINTER</i>	450	61.70	79,336	47.89
<i>Total</i>	729	100	165,680	100

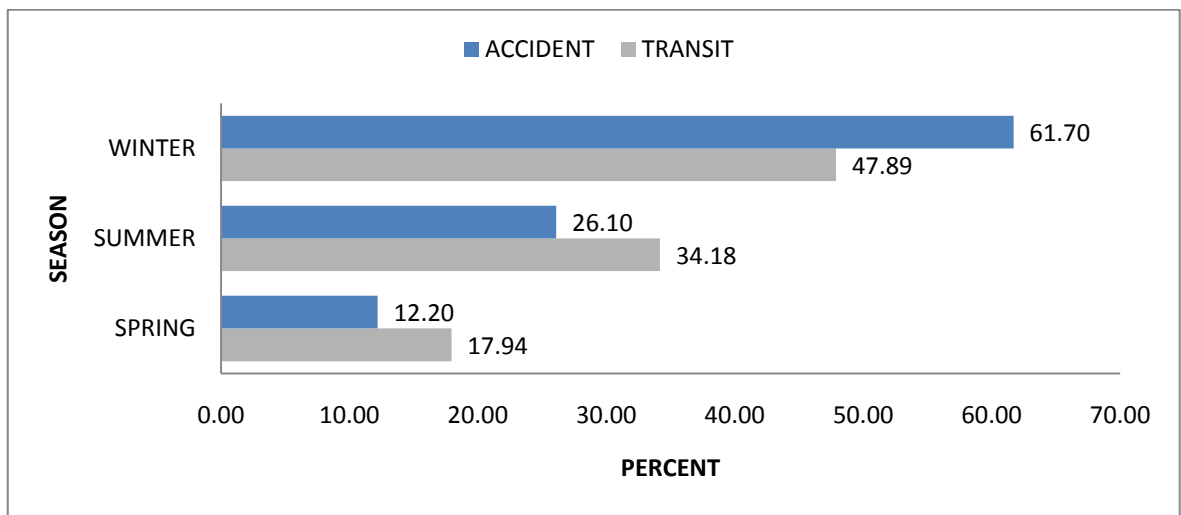


Figure 5.8. Frequency histogram of accidents according to season

- **Accident Time (Daytime/Nighttime):** As displayed in Table 5.10 and Figure 5.9, the accidents happen mostly at night. But there is not a significant difference in accident/transit ratio for both daytime and nighttime passages.

Table 5.10. Frequency of accidents according to daytime/nighttime status

	<i>Accident Frequency</i>	<i>Accident Percent</i>	<i>Transit Frequency</i>	<i>Transit Percent</i>
<i>DAY</i>	260	35.70	82,683	49.91
<i>NIGHT</i>	286	39.20	82,997	50.09
<i>UNKNOWN</i>	183	25.10	0	0.00
<i>Total</i>	729	100	165,680	100

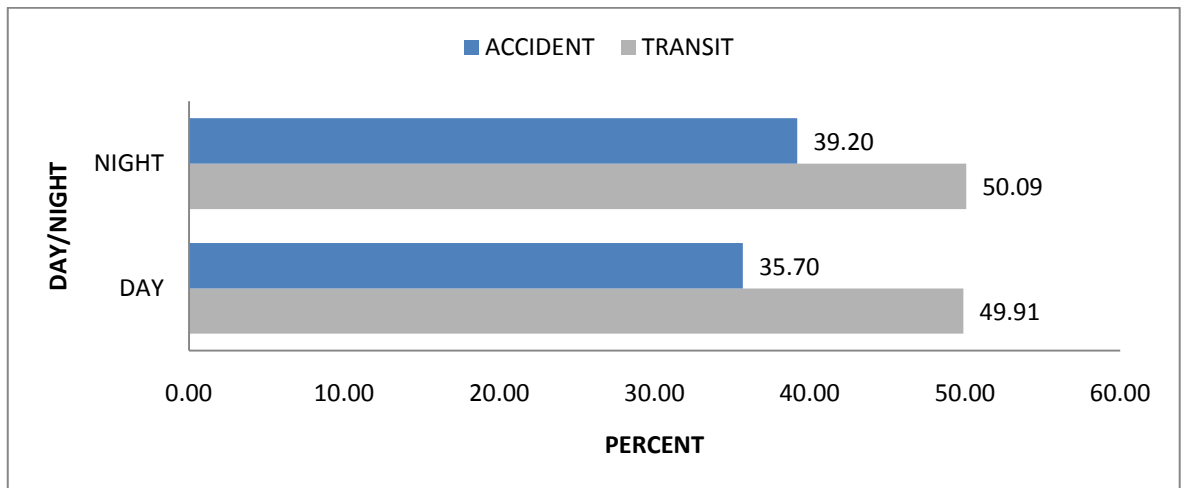


Figure 5.9. Frequency histogram of accidents according to daytime/nighttime status

- Accident type: As displayed in Table 5.11 and Figure 5.10, collision accidents are the major type of accidents in the Strait of Istanbul, (about 39.8%) and the least frequent type of accidents is in “Sinking” category.

Table 5.11. Frequency of accidents according to the accident type

	Frequency	Percent
<i>COLLISION</i>	290	39.8
<i>FIRE & EXPLOSION</i>	73	10.0
<i>GROUNDING</i>	183	25.1
<i>RAMMING</i>	155	21.3
<i>SINKING</i>	27	3.7
<i>UNKNOWN</i>	1	0.1
<i>Total</i>	729	100

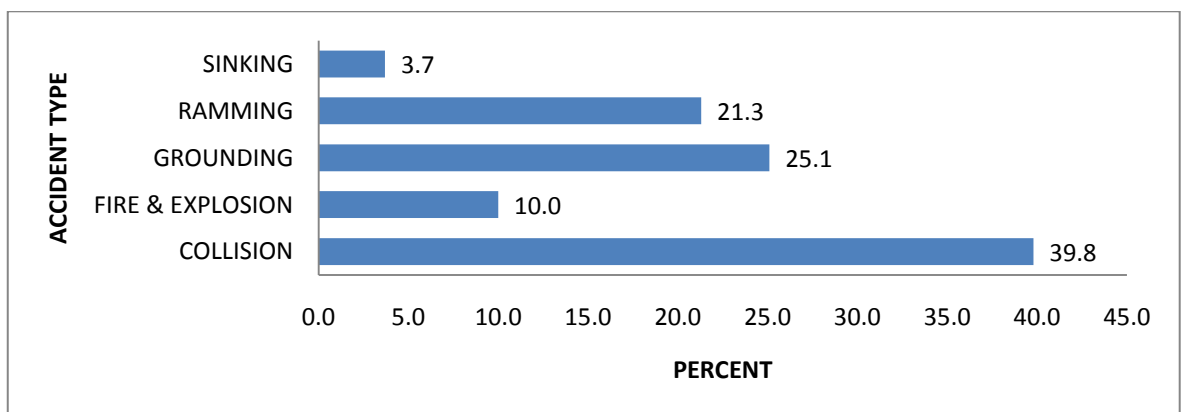


Figure 5.10. Frequency histogram of accidents according to accident type

- Accident region: As displayed in Table 5.12 and Figure 5.11, the south anchoring area (Region 7) is the place where most accidents took place (25.1%), while the north anchoring area (Region 8) hosted the smallest number of accidents.

Table 5.12. Frequency of accidents according to region

	<i>Frequency</i>	<i>Percent</i>	<i>Cumulative Percent</i>
1	47	6.4	6.4
2	80	11.0	17.4
3	60	8.2	25.7
4	74	10.2	35.8
5	45	6.2	42.0
6	149	20.4	62.4
7	183	25.1	87.5
8	33	4.5	92.0
UNKNOWN	58	8.0	100.0
<i>Total</i>	729	100	

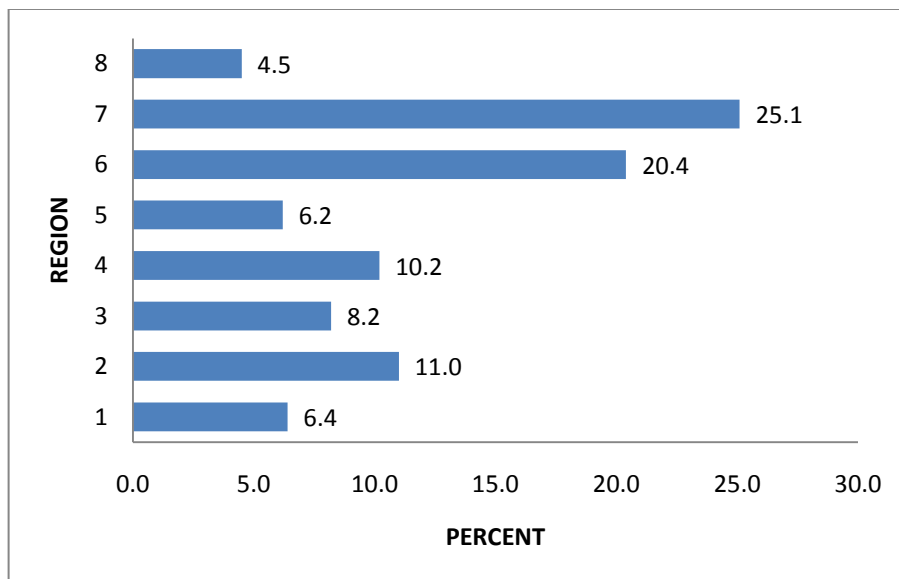


Figure 5.11. Frequency histogram of accidents according to region

- Accident cause: As displayed in Table 5.13 and Figure 5.12, apart from the “cause unknown” class, most of the accidents occurred because of human error.

Table 5.13. Frequency of accidents according to accident causes

	<i>Frequency</i>	<i>Percent</i>
<i>BAD WEATHER CONDITIONS</i>	106	14.5
<i>FAILURE</i>	106	14.5
<i>HUMAN ERROR</i>	202	27.7
<i>SABOTAGE</i>	7	1.0
<i>STRONG CURRENT</i>	37	5.1
<i>UNKNOWN</i>	271	37.2
<i>Total</i>	729	100

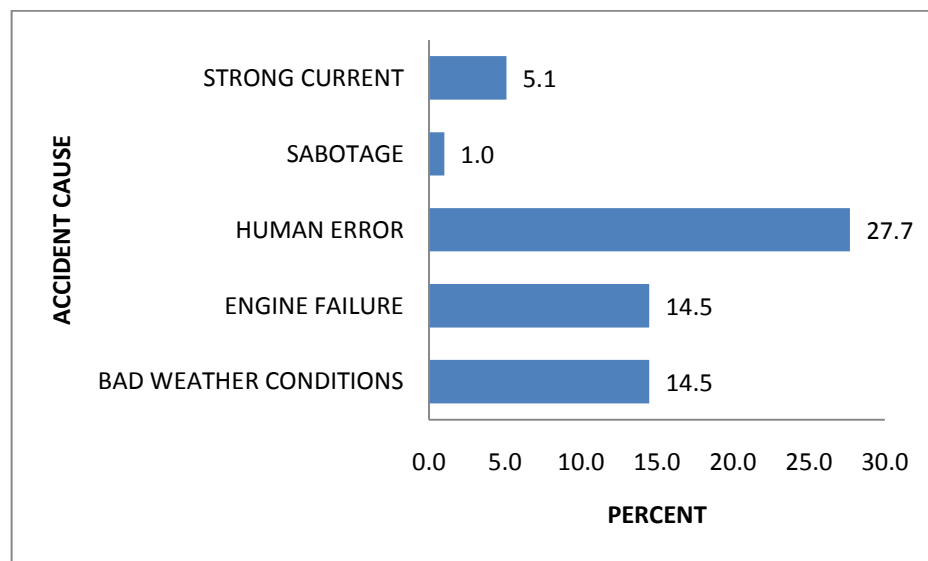


Figure 5.12. Frequency histogram of accidents according to accident causes

5.3. Chi-Square Tests for Nominal Scale Variables and Interpretations

5.3.1. Definitions

The aim of the chi-square test is to detect the existence of a relationship between two cross-tabulated variables. It gives information about the degree to which the dependent variable distribution across independent variable values differ from expected frequencies when statistical independence is assumed. It is more commonly used when the variables are categorical. The null and alternative hypotheses are:

- H_0 : There is no statistical relationship between the cross-tabulated variables.

- H_1 : There is a statistical relationship between the cross-tabulated variables.

Hence, if p-value of the test result is greater than the significance level (in this study, $\alpha=0.05$), H_0 cannot be rejected, which means that the differences between observed and expected counts is due to random sampling error. On the other hand, if $p\text{-value} < \alpha$, this indicates a statistical relationship between these two variables (Crowder and Glynn, 2000).

The chi-square tests are based on two assumptions which render the corresponding test statistics reliable and stronger. These are:

- If the cross-table is of size 2×2 , expected count for each cell must be greater than 5.
- When the cross-table size is greater than 2×2 , expected count for each cell must be greater than 1 and at least 80% of the cells must be greater than 5.

When these conditions are not satisfied, the test results become unreliable. There can be two remedies to prevent this condition: collecting more data or merging some categories, in order to have the resulting cross-table satisfying the above stated conditions.

Moreover, the chi-square test result only tells whether there is a relationship or not between the numeric variables: a large chi-square statistics does not mean a strong relationship because it is totally affected by the sample size. For this purpose, Phi and Cramer's V scores, which can be thought as the correlation of the numeric variables, can be investigated. When two variables are statistically independent, the scores are close to 0 whereas if there is a statistical relationship between two variables, the scores approach to 1. But the interpretation of these two results would sometimes be hard since the maximum possible value of the scores would not reach to 1 in all cases (Morgan *et al.*, 2004).

5.3.2. Chi- Square Testing For Vessel Attributes

Chi-square test is performed to the vessel accident database to specify whether there is a statistical relationship between two vessel attributes or between a vessel attribute and the corresponding accident attribute. Since SPSS crosstab option and chi-square test is

performed to categorical (nominal) variables, firstly, all the attributes are categorized. Since logistic regression analysis presented in Chapter 6 is performed to transit vessels, the vessel dataset composed of 813 vessels, namely, the dataset obtained by extracting local and unknown type vessels from the main database is used for the chi-square analysis.

The attributes with their categories are as follows:

- Vessel Type: Classified as general cargo, LNG/LPG/HazMat, passenger and tanker.
- Length: Comprises both original lengths and the length values obtained by GRT-length regression. The corresponding categories are 100 meters and below, between 100 and 200 meters, 200 meters and above.
- Flag Type: Classified as low-risk, medium-risk and high-risk flag.
- Age: Categorized as young, middle-aged and old.
- Pilot: “0” represents vessels without pilot request, “1” represents vessels with pilot request.
- Year: The data set comprises the accidents between 1982 and 2007. To simplify the interpretations of the output, the database has been divided into 5 groups (1: 1982-1987, 2: 1988-1992, 3: 1993-1997, 4: 1998-2002, 5: 2003-2007).
- Season: Three seasons (summer, spring/transition and winter) are deployed based on the results obtained from the fog analysis cited in Chapter 4.
- Day/Night Status: Day/night determination is accomplished via a dataset submitted by Bogazici University Kandilli Observatory and Earthquake Research Institute (KOERI). The dataset contains the information of sunrise and sunset times in Istanbul. It also includes the time intervals at which the European Summer time is applied since 1990. All these inputs are projected to the accident and logistic regression datasets to determine the daytime and the nighttime status.

- Region: Regions 1 to 6 correspond to various sections of the Strait (ordered from north to south), while region 7 is the south anchoring area and 8 is the north anchoring area. The regions are displayed in Figure 5.13.
- Accident Type: It is split into five categories as collision, fire&explosion, grounding, ramming and sinking.
- Accident Cause: The accident causes are classified as bad weather conditions, equipment failure, human error and strong current.

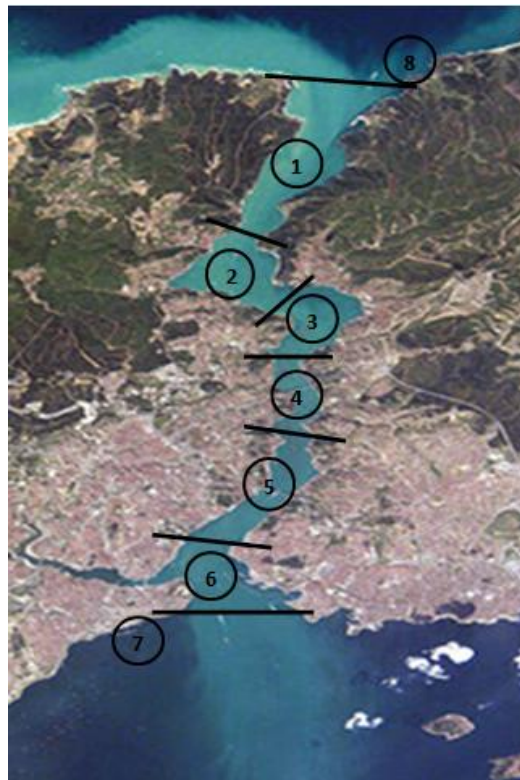


Figure 5.13. Accident regions map

Apart from these categories, since a good number of data records have missing values, an additional category named “Unknown” is defined for most of the attributes (which can also be seen in frequency analysis). Since chi-square analysis is based on pairwise frequencies (namely, crosstabs), the category “Unknown” behaves as an artificial category and modifies the test statistic: for example, some unrelated variables may become

related due to these unknown classes. Since the majority of the rows (representing vessels) of the database contain at least one unknown component, eliminating all such rows would cause a huge data loss. As a consequence, instead of the elimination of all the rows containing unknown/undefined pieces of data, a pairwise elimination is accomplished for each test statistics calculation. For example, when the relation between vessel age and accident time is taken into consideration, records containing unknown values for only these two attributes are deleted. After the database is prepared and chi-square test is applied for all variables pairwise, the next step is to identify the relationships via p-values.

As stated in Sub-section 5.3.1, when p-value < 0.05 , the null hypothesis is rejected, which means that there exist a relationship between the two variables. The obtained p-values are presented in Table 5.14.

Table 5.14. Chi-square test results for the vessel database

	VESSEL TYPE	LENGTH	FLAG TYPE	AGE	PILOT
VESSEL TYPE	.	0.000	0.005	0.001	0.003
LENGTH	0.000	.	0.000	0.230	0.007
FLAG TYPE	0.005	0.000	.	0.050	0.000
AGE	0.001	0.230	0.050	.	0.636
PILOT	0.003	0.007	0.000	0.636	.
YEAR	0.000	0.331	0.067	0.000	0.000
SEASON	0.013	0.000	0.702	0.070	0.207
DAY/NIGHT	0.423	0.001	0.419	0.168	0.179
REGION NUMBER	0.000	0.045	0.043	0.008	0.000
ACCIDENT TYPE	0.006	0.001	0.064	0.035	0.165
ACCIDENT CAUSE	0.062	0.586	0.875	0.051	0.000

The interpretations for the cases where the variables are found to be related are given as follows.

As displayed in Table 5.15, there is a relationship between vessel length and type. General cargo, LNG/LPG/HazMat and tanker vessels involved in accidents are mostly of length 100 meters and above. In contrast, passenger vessels involved in accidents are of length below 200 meters.

Table 5.15. Length * Vessel type crosstabulation

			VESSEL TYPE				Total
			GENERAL CARGO	LNG/LPG/HAZMAT	PASSENGER	TANKER	
LENGTH	100 AND BELOW	Count	158	1	24	17	200
		% within VESSEL TYPE	26.1%	25.0%	55.8%	16.8%	26.6%
	100-200	Count	426	3	19	70	518
		% within VESSEL TYPE	70.4%	75.0%	44.2%	69.3%	68.8%
	200 AND ABOVE	Count	21	0	0	14	35
		% within VESSEL TYPE	3.5%	0.0%	0.0%	13.9%	4.6%
Total	Count	605	4	43	101	753	
	% within VESSEL TYPE	100.0%	100.0%	100.0%	100.0%	100.0%	

As displayed in Table 5.16, there is a relationship between flag type and vessel type. LNG/LPG/HazMat and tanker vessels involved in accidents are mostly of low-risk flag. In contrast, general cargo and passenger vessels involved in accidents are mostly of medium-risk.

Table 5.16. Flag type * Vessel type crosstabulation

			VESSEL TYPE				Total
			GENERAL CARGO	LNG/LPG/HAZMAT	PASSENGER	TANKER	
FLAG TYPE	HIGH-RISK	Count	246	4	25	55	330
		% within VESSEL TYPE	38.6%	100.0%	41.7%	52.4%	40.9%
	MEDIUM-RISK	Count	365	0	35	50	450
		% within VESSEL TYPE	57.2%	0.0%	58.3%	47.6%	55.8%
	LOW-RISK	Count	27	0	0	0	27
		% within VESSEL TYPE	4.2%	0.0%	0.0%	0.0%	3.3%
Total	Count	638	4	60	105	807	
	% within VESSEL TYPE	100.0%	100.0%	100.0%	100.0%	100.0%	

As displayed in Table 5.17, there is a relationship between vessel age and type. LNG/LPG/HazMat and passenger vessels involved in accidents are distributed homogeneously in terms of age categories. In contrast, general cargo and tanker vessels involved in accidents are mostly old.

Table 5.17. Age * Vessel type crosstabulation

		VESSEL TYPE				Total
		GENERAL CARGO	LNG/LPG/HAZMAT	PASSENGER	TANKER	
AGE YOUNG	Count	30	1	5	8	44
	% within VESSEL TYPE	7.6%	33.3%	27.8%	12.7%	9.2%
MIDDLE-AGED	Count	62	1	6	16	85
	% within VESSEL TYPE	15.7%	33.3%	33.3%	25.4%	17.8%
OLD	Count	302	1	7	39	349
	% within VESSEL TYPE	76.6%	33.3%	38.9%	61.9%	73.0%
Total	Count	394	3	18	63	478
	% within VESSEL TYPE	100.0%	100.0%	100.0%	100.0%	100.0%

As displayed in Table 5.18, there is a relationship between pilot request and vessel type. General cargo and passenger vessels involved in accidents do not generally request pilot. The same holds true for LNG/LPG/HazMat and tanker vessels, but these two types have a little more tendency to take a pilot on board.

Table 5.18. Pilot * Vessel type crosstabulation

		VESSEL TYPE				Total
		GENERAL CARGO	LNG/LPG/HAZMAT	PASSENGER	TANKER	
PILOT	0 Count	605	3	56	91	755
	% within VESSEL TYPE	94.2%	75.0%	93.3%	85.0%	92.9%
	1 Count	37	1	4	16	58
	% within VESSEL TYPE	5.8%	25.0%	6.7%	15.0%	7.1%
Total	Count	642	4	60	107	813
	% within VESSEL TYPE	100.0%	100.0%	100.0%	100.0%	100.0%

As displayed in Table 5.19, there is a relationship between accident year and vessel type. Tankers involved in an accident mostly show a homogenous distribution throughout time, but general cargo vessel accidents have an increasing trend and passenger vessel accidents have a decreasing trend with respect to time.

Table 5.19. Accident year * Vessel type crosstabulation

			VESSEL TYPE				Total
			GENERAL CARGO	LNG/LPG/HAZMAT	PASSENGER	TANKER	
YEAR	1982-1987	Count	64	0	7	21	92
		% within VESSEL TYPE	10.0%	0.0%	11.7%	19.6%	11.3%
	1988-1992	Count	106	3	14	21	144
		% within VESSEL TYPE	16.5%	75.0%	23.3%	19.6%	17.7%
	1993-1997	Count	154	0	16	21	191
		% within VESSEL TYPE	24.0%	0.0%	26.7%	19.6%	23.5%
	1998-2002	Count	124	0	20	21	165
		% within VESSEL TYPE	19.3%	0.0%	33.3%	19.6%	20.3%
	2003-2007	Count	194	1	3	23	221
		% within VESSEL TYPE	30.2%	25.0%	5.0%	21.5%	27.2%
	Total	Count	642	4	60	107	813
		% within VESSEL TYPE	100.0%	100.0%	100.0%	100.0%	100.0%

As displayed in Table 5.20, there is a relationship between accident season and vessel type. General cargo, passenger and tanker type vessels involved in accidents mostly have accidents in winter. In contrast, the LNG/LPG/HazMat vessels do not have a specific season for accidents, but it is because the number of vessels of this type is low. Besides, there is a statistical relationship between season and vessel type at 0.05 confidence level but not at 0.01.

Table 5.20. Season * Vessel type crosstabulation

			VESSEL TYPE				Total
			GENERAL CARGO	LNG/LPG/HAZMAT	PASSENGER	TANKER	
SEASON	SUMMER	Count	139	2	19	22	182
		% within VESSEL TYPE	21.7%	50.0%	31.7%	20.6%	22.4%
	SPRING	Count	59	1	12	14	86
		% within VESSEL TYPE	9.2%	25.0%	20.0%	13.1%	10.6%
	WINTER	Count	444	1	29	71	545
		% within VESSEL TYPE	69.2%	25.0%	48.3%	66.4%	67.0%
	Total	Count	642	4	60	107	813
		% within VESSEL TYPE	100.0%	100.0%	100.0%	100.0%	100.0%

As displayed in Table 5.21, there is a relationship between accident region and vessel type. General cargo and tanker vessels are mostly involved in accidents at the south anchoring area (Region 7). In contrast, passenger vessels are mostly involved in accidents in the southern transit area (Region 6).

Table 5.21. Region * Vessel type crosstabulation

		VESSEL TYPE				Total
		GENERAL CARGO	LNG/LPG/HAZMAT	PASSENGER	TANKER	
REGION	1 Count	41	0	1	9	51
	% within VESSEL TYPE	6.6%	0.0%	1.8%	8.7%	6.5%
	2 Count	58	0	8	16	82
	% within VESSEL TYPE	9.3%	0.0%	14.0%	15.4%	10.4%
	3 Count	45	0	3	18	66
	% within VESSEL TYPE	7.2%	0.0%	5.3%	17.3%	8.4%
	4 Count	64	0	4	9	77
	% within VESSEL TYPE	10.3%	0.0%	7.0%	8.7%	9.8%
	5 Count	27	0	9	4	40
	% within VESSEL TYPE	4.3%	0.0%	15.8%	3.8%	5.1%
	6 Count	74	2	22	15	113
	% within VESSEL TYPE	11.9%	50.0%	38.6%	14.4%	14.3%
	7 Count	285	2	9	27	323
	% within VESSEL TYPE	45.7%	50.0%	15.8%	26.0%	41.0%
	8 Count	29	0	1	6	36
	% within VESSEL TYPE	4.7%	0.0%	1.8%	5.8%	4.6%
Total	Count	623	4	57	104	788
	% within VESSEL TYPE	100.0%	100.0%	100.0%	100.0%	100.0%

As displayed in Table 5.22, there is a relationship between accident type and vessel type. General cargo, LNG/LPG/HazMat and passenger vessels are mostly involved in collision accidents and for each type; other accident rates do not exceed 25%. But for tankers, not only collision, but also grounding accidents take important place.

Table 5.22. Accident type * Vessel type crosstabulation

			VESSEL TYPE				Total
			GENERAL CARGO	LNG/LPG/HAZMAT	PASSENGER	TANKER	
ACCIDENT TYPE	COLLISION	Count	330	2	31	45	408
		% within VESSEL TYPE	51.4%	50.0%	51.7%	42.1%	50.2%
	FIRE & EXPLOSION	Count	28	0	8	6	42
		% within VESSEL TYPE	4.4%	0.0%	13.3%	5.6%	5.2%
	GROUNDING	Count	117	1	5	35	158
		% within VESSEL TYPE	18.2%	25.0%	8.3%	32.7%	19.4%
	RAMMING	Count	158	1	14	19	192
		% within VESSEL TYPE	24.6%	25.0%	23.3%	17.8%	23.6%
	SINKING	Count	9	0	2	2	13
		% within VESSEL TYPE	1.4%	0.0%	3.3%	1.9%	1.6%
	Total	Count	642	4	60	107	813
		% within VESSEL TYPE	100.0%	100.0%	100.0%	100.0%	100.0%

As displayed in Table 5.23, there is a relationship between vessel flag type and vessel length. Vessels having a length smaller than 100 meters are mostly of medium-risk flag, whereas vessels having a length larger than 200 meters are generally of low-risk flag. Furthermore, vessels of length between 100 and 200 m. are equally distributed between low and medium-risk. It can be concluded that most of large vessels are of low-risk.

Table 5.23. Flag type * Length crosstabulation

			LENGTH			Total
			100 AND BELOW	100-200	200 AND ABOVE	
FLAG TYPE	HIGH-RISK	Count	28	252	27	307
		% within LENGTH	14.0%	48.9%	77.1%	40.9%
	MEDIUM-RISK	Count	164	246	7	417
		% within LENGTH	82.0%	47.8%	20.0%	55.6%
	LOW-RISK	Count	8	17	1	26
		% within LENGTH	4.0%	3.3%	2.9%	3.5%
Total	Count	Count	200	515	35	750
		% within LENGTH	100.0%	100.0%	100.0%	100.0%

As displayed in Table 5.24, there is a relationship between pilot request and vessel length. Vessels larger than 200 meters have more tendencies to benefit from pilotage services than smaller vessels.

Table 5.24. Pilot * Length crosstabulation

		LENGTH			Total
		100 AND BELOW	100-200	200 AND ABOVE	
PILOT	0 Count	193	473	29	695
	% within LENGTH	96.5%	91.3%	82.9%	92.3%
	1 Count	7	45	6	58
	% within LENGTH	3.5%	8.7%	17.1%	7.7%
Total	Count	200	518	35	753
	% within LENGTH	100.0%	100.0%	100.0%	100.0%

As displayed in Table 5.25, there is a relationship between accident season and vessel length. The vessels involved in accidents mostly have their accidents in winter. But for the vessels of length smaller than 100 meters, there are also a significant number of summer accidents, so it can be inferred that the smaller vessels are more insensitive to season.

Table 5.25. Season * Length crosstabulation

			LENGTH			Total
			100 AND BELOW	100-200	200 AND ABOVE	
SEASON	SUMMER	Count	66	94	7	167
		% within LENGTH	33.0%	18.1%	20.0%	22.2%
	SPRING	Count	24	55	2	81
		% within LENGTH	12.0%	10.6%	5.7%	10.8%
	WINTER	Count	110	369	26	505
		% within LENGTH	55.0%	71.2%	74.3%	67.1%
Total	Count	200	518	35	753	
	% within LENGTH	100.0%	100.0%	100.0%	100.0%	

As displayed in Table 5.26, there is a relationship between accident time (i.e daytime or nighttime) and vessel length. The majority of the vessels of length larger than 200 meters are involved in accidents mostly in the day, whereas smaller vessels have approximately equal frequencies between day and night.

Table 5.26. Day/Night * Length crosstabulation

			LENGTH			Total
			100 AND BELOW	100-200	200 AND ABOVE	
DAY/NIGHT	DAY	Count	82	195	21	298
		% within LENGTH	50.0%	43.3%	77.8%	46.5%
	NIGHT	Count	82	255	6	343
		% within LENGTH	50.0%	56.7%	22.2%	53.5%
Total	Count	164	450	27	641	
	% within LENGTH	100.0%	100.0%	100.0%	100.0%	

As displayed in Table 5.27, there is a relationship between accident type and vessel length. The majority of the vessels of length larger than 200 meters are involved mostly in grounding accidents, whereas smaller vessels are mostly involved in collision accidents.

Table 5.27. Accident type * Length crosstabulation

			LENGTH			Total
			100 AND BELOW	100-200	200 AND ABOVE	
ACCIDENT TYPE	COLLISION	Count	108	257	10	375
		% within LENGTH	54.0%	49.6%	28.6%	49.8%
	FIRE & EXPLOSION	Count	9	25	4	38
		% within LENGTH	4.5%	4.8%	11.4%	5.0%
	GROUNDING	Count	31	98	14	143
		% within LENGTH	15.5%	18.9%	40.0%	19.0%
	RAMMING	Count	45	134	7	186
		% within LENGTH	22.5%	25.9%	20.0%	24.7%
	SINKING	Count	7	4	0	11
		% within LENGTH	3.5%	0.8%	0.0%	1.5%
	Total	Count	200	518	35	753
		% within LENGTH	100.0%	100.0%	100.0%	100.0%

As displayed in Table 5.28, there is a relationship between flag type and pilot request. The majority of the vessels requesting pilot are of low-risk class, whereas vessels without pilot are of medium-risk. Low-risk vessels have a higher tendency to benefit from pilotage services than the others.

Table 5.28. Flag type * Pilot crosstabulation

			PILOT		Total
			0	1	
FLAG TYPE	HIGH-RISK	Count	284	46	330
		% within PILOT	37.9%	79.3%	40.9%
	MEDIUM-RISK	Count	439	11	450
		% within PILOT	58.6%	19.0%	55.8%
	LOW-RISK	Count	26	1	27
		% within PILOT	3.5%	1.7%	3.3%
Total	Count	749	58	807	
	% within PILOT	100.0%	100.0%	100.0%	

As displayed in Table 5.29, there is a relationship between accident year and vessel age. There is a significant increase in the frequency of old vessels involved in an accident in the near past.

Table 5.29. Year * Age crosstabulation

		AGE			Total
		YOUNG	MIDDLE-AGED	OLD	
YEAR	1982-1987 Count	8	18	34	60
	% within AGE	18.2%	21.2%	9.7%	12.6%
	1988-1992 Count	8	23	51	82
	% within AGE	18.2%	27.1%	14.6%	17.2%
	1993-1997 Count	13	19	54	86
	% within AGE	29.5%	22.4%	15.5%	18.0%
	1998-2002 Count	4	14	40	58
	% within AGE	9.1%	16.5%	11.5%	12.1%
	2003-2007 Count	11	11	170	192
	% within AGE	25.0%	12.9%	48.7%	40.2%
	Total Count	44	85	349	478
	% within AGE	100.0%	100.0%	100.0%	100.0%

As displayed in Table 5.30, there is a relationship between accident region and vessel age. Middle-aged and young vessels are mostly involved in accidents at the regions 6 and 7 (southern transit and anchorage regions). In contrast, the majority of the accidents that old vessels involved occurred at region 7 (south anchorage area).

Table 5.30. Region * Age crosstabulation

		AGE			Total
		YOUNG	MIDDLE-AGED	OLD	
REGION	1 Count	1	8	23	32
	% within AGE	2.4%	9.6%	6.8%	6.9%
	2 Count	4	9	32	45
	% within AGE	9.5%	10.8%	9.4%	9.7%
	3 Count	5	9	27	41
	% within AGE	11.9%	10.8%	8.0%	8.8%
	4 Count	6	10	27	43
	% within AGE	14.3%	12.0%	8.0%	9.3%
	5 Count	2	5	5	12
	% within AGE	4.8%	6.0%	1.5%	2.6%
	6 Count	9	19	42	70
	% within AGE	21.4%	22.9%	12.4%	15.1%
	7 Count	13	22	165	200
	% within AGE	31.0%	26.5%	48.7%	43.1%
	8 Count	2	1	18	21
	% within AGE	4.8%	1.2%	5.3%	4.5%
	Total Count	42	83	339	464
	% within AGE	100.0%	100.0%	100.0%	100.0%

As displayed in Table 5.31, there is a relationship between accident year and pilot request. The number of accidents that the vessels not taking pilots increases throughout the years, but the vessels taking pilots have less accidents as the years passed. One reason for the vessels taking pilots having less accidents through the years may be that the pilots may

have become more experienced or that rules/navigational aids, which the pilot captains are far more knowledgeable about, have become effective in reducing the likelihood of accidents.

Table 5.31. Year * Pilot crosstabulation

			PILOT		Total
			0	1	
YEAR	1982-1987	Count	73	19	92
		% within PILOT	9.7%	32.8%	11.3%
	1988-1992	Count	127	17	144
		% within PILOT	16.8%	29.3%	17.7%
	1993-1997	Count	189	2	191
		% within PILOT	25.0%	3.4%	23.5%
	1998-2002	Count	154	11	165
		% within PILOT	20.4%	19.0%	20.3%
	2003-2007	Count	212	9	221
		% within PILOT	28.1%	15.5%	27.2%
	Total	Count	755	58	813
		% within PILOT	100.0%	100.0%	100.0%

As displayed in Table 5.32, there is a relationship between accident cause and pilot request. The majority of accidents in the vessels without pilots are caused by human errors, but the vessels with pilots mostly had equipment failure triggered accidents. This is an interesting observation indicating that having pilot decreases the likelihood of human error based accidents, but does not have the same effect on failure based accidents.

Table 5.32. Accident cause * Pilot crosstabulation

			PILOT		Total
			0	1	
ACCIDENT CAUSE	BAD WEATHER CONDITIONS	Count	132	14	146
		% within PILOT	24.9%	28.0%	25.1%
	FAILURE	Count	83	19	102
		% within PILOT	15.6%	38.0%	17.6%
	HUMAN ERROR	Count	276	14	290
		% within PILOT	52.0%	28.0%	49.9%
	STRONG CURRENT	Count	40	3	43
		% within PILOT	7.5%	6.0%	7.4%
	Total	Count	531	50	581
		% within PILOT	100.0%	100.0%	100.0%

5.3.3. Chi Square Testing For Accident Attributes

In order to define pairwise relationships between year, season, date, region, accident type and cause, the accident dataset constructed in subsection 5.1.2 is used. Since chi-square analysis is conducted for records corresponding to transit cases, this dataset only covers the accidents where at least one transit vessel is involved. Moreover, as in subsection 5.3.2, only the records containing unknown cells associated with the attribute pair under consideration are excluded from the analysis. The obtained p-values are presented in Table 5.33.

Table 5.33. Chi-square test results for the accident-based attributes

	YEAR	SEASON	DAY/NIGHT	REGION	ACCIDENT TYPE	ACCIDENT CAUSE
YEAR	.	0.055	0.002	0.000	0.000	0.000
SEASON	0.055	.	0.000	0.001	0.326	0.001
DAY/NIGHT	0.002	0.000	.	0.593	0.544	0.219
REGION	0.000	0.001	0.593	.	0.000	0.000
ACCIDENT TYPE	0.000	0.326	0.544	0.000	.	0.000
ACCIDENT CAUSE	0.000	0.001	0.219	0.000	0.000	.

The interpretations for the cases which are found to be related are as follows.

As displayed in Table 5.34, there is a relationship between day/night status and year of the accident. Until early 1990's, accidents are generally occurring in the day, while in later years, they are generally occurring at night.

Table 5.34. Day/Night * Year crosstabulation

			YEAR					
			1982-1987	1988-1992	1993-1997	1998-2002	2003-2007	
DAY/NIGHT	DAY	Count	26	42	48	42	52	210
		% within YEAR	65.0%	56.8%	49.0%	38.5%	36.6%	45.4%
	NIGHT	Count	14	32	50	67	90	253
		% within YEAR	35.0%	43.2%	51.0%	61.5%	63.4%	54.6%
Total	Count	40	74	98	109	142	463	
	% within YEAR	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	

As displayed in Table 5.35, there is a relationship between accident year and accident region. The frequency of the anchoring area accidents increases and that of transit area

accidents decreases as time passes. The reason for this increase may be the better record keeping for anchoring areas at the present time.

Table 5.35. Year * Region crosstabulation

			REGION								Total
			1	2	3	4	5	6	7	8	
YEAR	1982-1987	Count	6	13	11	14	2	15	4	0	65
		% within REGION	15.0%	18.8%	19.6%	23.0%	6.5%	17.0%	2.3%	0.0%	11.8%
	1988-1992	Count	11	17	9	14	10	14	24	0	99
		% within REGION	27.5%	24.6%	16.1%	23.0%	32.3%	15.9%	13.8%	0.0%	18.0%
	1993-1997	Count	7	13	14	11	4	26	46	3	124
		% within REGION	17.5%	18.8%	25.0%	18.0%	12.9%	29.5%	26.4%	10.0%	22.6%
	1998-2002	Count	6	18	12	11	12	20	32	8	119
		% within REGION	15.0%	26.1%	21.4%	18.0%	38.7%	22.7%	18.4%	26.7%	21.7%
	2003-2007	Count	10	8	10	11	3	13	68	19	142
		% within REGION	25.0%	11.6%	17.9%	18.0%	9.7%	14.8%	39.1%	63.3%	25.9%
	Total	Count	40	69	56	61	31	88	174	30	549
		% within REGION	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

As displayed in Table 5.36, there is a relationship between accident year and accident type. Collision, fire & explosion and grounding accidents are homogeneously distributed regarding years, but ramming and sinking accidents increase in the near past.

Table 5.36. Year * Accident type crosstabulation

			ACCIDENT TYPE					Total
			COLLISION	FIRE & EXPLOSION	GROUNDING	RAMMING	SINKING	
YEAR	1982-1987	Count	25	5	24	18	0	72
		% within ACCIDENT TYPE	10.5%	12.5%	15.6%	14.5%	0.0%	12.7%
	1988-1992	Count	48	12	32	10	2	104
		% within ACCIDENT TYPE	20.3%	30.0%	20.8%	8.1%	15.4%	18.3%
	1993-1997	Count	64	6	27	27	1	125
		% within ACCIDENT TYPE	27.0%	15.0%	17.5%	21.8%	7.7%	22.0%
	1998-2002	Count	62	7	34	16	3	122
		% within ACCIDENT TYPE	26.2%	17.5%	22.1%	12.9%	23.1%	21.5%
	2003-2007	Count	38	10	37	53	7	145
		% within ACCIDENT TYPE	16.0%	25.0%	24.0%	42.7%	53.8%	25.5%
	Total	Count	237	40	154	124	13	568
		% within ACCIDENT TYPE	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

As displayed in Table 5.37, there is a relationship between year and accident cause. Bad weather conditions, equipment failure and human error triggered accidents have increased in recent accidents, whereas the number of accidents caused by strong current has decreased in time.

Table 5.37. Year * Accident cause crosstabulation

		ACCIDENT CAUSE				Total
		BAD WEATHER CONDITIONS	FAILURE	HUMAN ERROR	STRONG CURRENT	
YEAR 1982-1987	Count	22	5	8	2	37
	% within ACCIDENT CAUSE	23.7%	5.7%	4.4%	6.1%	9.3%
1988-1992	Count	16	17	22	13	68
	% within ACCIDENT CAUSE	17.2%	19.5%	12.0%	39.4%	17.2%
1993-1997	Count	13	12	42	14	81
	% within ACCIDENT CAUSE	14.0%	13.8%	23.0%	42.4%	20.5%
1998-2002	Count	18	22	47	3	90
	% within ACCIDENT CAUSE	19.4%	25.3%	25.7%	9.1%	22.7%
2003-2007	Count	24	31	64	1	120
	% within ACCIDENT CAUSE	25.8%	35.6%	35.0%	3.0%	30.3%
Total	Count	93	87	183	33	396
	% within ACCIDENT CAUSE	100.0%	100.0%	100.0%	100.0%	100.0%

As displayed in Table 5.38, there is a relationship between season and day/night status. Nighttime accidents occur more frequently than daytime accidents in winter, while the reverse is true in summer (this may be because of longer winter nights). Moreover, in spring, nighttime accidents occur slightly more than daytime accidents.

Table 5.38. Season * Day/Night crosstabulation

		DAY/NIGHT		Total
		DAY	NIGHT	
SEASON SUMMER	Count	71	40	111
	% within DAY/NIGHT	33.8%	15.8%	24.0%
SPRING	Count	22	26	48
	% within DAY/NIGHT	10.5%	10.3%	10.4%
WINTER	Count	117	187	304
	% within DAY/NIGHT	55.7%	73.9%	65.7%
Total	Count	210	253	463
	% within DAY/NIGHT	100.0%	100.0%	100.0%

As displayed in Table 5.39, there is a relationship between accident region and season. In summer, most accidents happen at the southern transit area (region 6), whereas in spring and winter, most accidents happen at the south anchoring area (region 7). This observation may indicate that the south anchoring area becomes more treacherous in the harsher climate conditions of winters.

Table 5.39. Region * Season crosstabulation

		SEASON			Total
		SUMMER	SPRING	WINTER	
REGION	1 Count	13	5	22	40
	% within SEASON	9.8%	7.8%	6.2%	7.3%
	2 Count	22	10	37	69
	% within SEASON	16.7%	15.6%	10.5%	12.6%
	3 Count	14	4	38	56
	% within SEASON	10.6%	6.3%	10.8%	10.2%
	4 Count	10	8	43	61
	% within SEASON	7.6%	12.5%	12.2%	11.1%
	5 Count	13	4	14	31
	% within SEASON	9.8%	6.3%	4.0%	5.6%
	6 Count	29	13	46	88
	% within SEASON	22.0%	20.3%	13.0%	16.0%
	7 Count	22	17	135	174
	% within SEASON	16.7%	26.6%	38.2%	31.7%
	8 Count	9	3	18	30
	% within SEASON	6.8%	4.7%	5.1%	5.5%
Total		132	64	353	549
		100.0%	100.0%	100.0%	100.0%

As displayed in Table 5.40, there is a relationship between accident cause and season of the accident. In spring and summer, the majority of the accidents are caused by failure and human error, whereas in winter, the majority of the accidents are caused by bad weather conditions and human error.

Table 5.40. Accident cause * Season crosstabulation

			SEASON			Total
			SUMMER	SPRING	WINTER	
ACCIDENT CAUSE	BAD WEATHER CONDITIONS	Count	9	6	78	93
		% within SEASON	10.2%	13.3%	29.7%	23.5%
	FAILURE	Count	27	13	47	87
		% within SEASON	30.7%	28.9%	17.9%	22.0%
	HUMAN ERROR	Count	47	24	112	183
		% within SEASON	53.4%	53.3%	42.6%	46.2%
	STRONG CURRENT	Count	5	2	26	33
		% within SEASON	5.7%	4.4%	9.9%	8.3%
	Total		88	45	263	396
			100.0%	100.0%	100.0%	100.0%

As displayed in Table 5.41, there is a relationship between region and accident type. Collision and ramming cases occur mostly in region 7 (south anchoring area), fire & explosion cases in region 6 and 7 (southern transit and anchorage regions), grounding cases in region 2 (transit region) and sinking cases in region 8 (north anchoring area).

Table 5.41. Region * Accident type crosstabulation

		ACCIDENT TYPE					Total
		COLLISION	FIRE & EXPLOSION	GROUNDING	RAMMING	SINKING	
REGION	1 Count	18	4	10	6	2	40
	% within ACCIDENT TYPE	7.8%	10.8%	6.7%	5.1%	15.4%	7.3%
	2 Count	14	1	50	4	0	69
	% within ACCIDENT TYPE	6.0%	2.7%	33.3%	3.4%	0.0%	12.6%
	3 Count	10	3	25	18	0	56
	% within ACCIDENT TYPE	4.3%	8.1%	16.7%	15.4%	0.0%	10.2%
	4 Count	17	3	15	26	0	61
	% within ACCIDENT TYPE	7.3%	8.1%	10.0%	22.2%	0.0%	11.1%
	5 Count	18	4	6	3	0	31
	% within ACCIDENT TYPE	7.8%	10.8%	4.0%	2.6%	0.0%	5.6%
	6 Count	49	11	13	14	1	88
	% within ACCIDENT TYPE	21.1%	29.7%	8.7%	12.0%	7.7%	16.0%
	7 Count	98	9	19	45	3	174
	% within ACCIDENT TYPE	42.2%	24.3%	12.7%	38.5%	23.1%	31.7%
	8 Count	8	2	12	1	7	30
	% within ACCIDENT TYPE	3.4%	5.4%	8.0%	0.9%	53.8%	5.5%
Total		232	37	150	117	13	549
		100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

As displayed in Table 5.42, there is a relationship between region and accident cause. Accidents caused by bad weather conditions and human error mostly occur in region 7 (south anchoring area), failure mostly occur in regions 2, 3 and 4, which are the difficult-to-navigate transit regions in the middle of the Strait, strong current mostly occur in region 4. In region 7, bad weather condition-based accidents occur from the fact that the vessels drag anchor because of strong wind.

Table 5.42. Region number * Accident cause crosstabulation

		ACCIDENT CAUSE				Total
		BAD WEATHER CONDITIONS	FAILURE	HUMAN ERROR	STRONG CURRENT	
REGION NUMBER	1 Count	12	4	8	1	25
	% within ACCIDENT CAUSE	13.0%	4.8%	4.4%	3.0%	6.4%
	2 Count	11	16	25	0	52
	% within ACCIDENT CAUSE	12.0%	19.3%	13.7%	0.0%	13.3%
	3 Count	5	19	11	8	43
	% within ACCIDENT CAUSE	5.4%	22.9%	6.0%	24.2%	11.0%
	4 Count	13	17	11	12	53
	% within ACCIDENT CAUSE	14.1%	20.5%	6.0%	36.4%	13.6%
	5 Count	2	10	6	2	20
	% within ACCIDENT CAUSE	2.2%	12.0%	3.3%	6.1%	5.1%
	6 Count	6	3	34	6	49
	% within ACCIDENT CAUSE	6.5%	3.6%	18.7%	18.2%	12.6%
	7 Count	31	9	81	4	125
	% within ACCIDENT CAUSE	33.7%	10.8%	44.5%	12.1%	32.1%
	8 Count	12	5	6	0	23
	% within ACCIDENT CAUSE	13.0%	6.0%	3.3%	0.0%	5.9%
Total		92	83	182	33	390
		100.0%	100.0%	100.0%	100.0%	100.0%

As displayed in Table 5.43, there is a relationship between accident cause and accident type. Collision, grounding, ramming and sinking accidents mostly occur because

of human error, whereas fire & explosion accidents mostly occur because of equipment failures.

Table 5.43. Accident cause * Accident type crosstabulation

			ACCIDENT TYPE					Total
			COLLISION	FIRE & EXPLOSION	GROUNDING	RAMMING	SINKING	
ACCIDENT CAUSE	BAD WEATHER	Count	46	0	27	17	3	93
	CONDITIONS	% within ACCIDENT TYPE	29.5%	0.0%	23.1%	17.9%	30.0%	23.5%
	FAILURE	Count	8	15	36	26	2	87
		% within ACCIDENT TYPE	5.1%	83.3%	30.8%	27.4%	20.0%	22.0%
	HUMAN ERROR	Count	93	2	44	39	5	183
		% within ACCIDENT TYPE	59.6%	11.1%	37.6%	41.1%	50.0%	46.2%
	STRONG CURRENT	Count	9	1	10	13	0	33
		% within ACCIDENT TYPE	5.8%	5.6%	8.5%	13.7%	0.0%	8.3%
Total		Count	156	18	117	95	10	396
		% within ACCIDENT TYPE	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

5.4. Near-Miss Analyses

5.4.1. Data Description

Besides the accident data, the data provided from Prime Ministry Undersecretariat for Maritime Affairs also contain information about near miss events covering various failures. Hence, frequency and chi-square analyses are conducted to investigate the attributes triggering various types of failure/near-miss events. Since the number of failures in 2004 and before is remarkably smaller than that of the 2005 and later data, it is assumed that failures before 2005 have not been properly recorded. Hence, the analysis covered just 503 failures that occurred between the years 2005 and 2007. The original near-miss data comprises information about date, hour, type, description of the failure, a binary indicator showing whether the failure has occurred in or out of the Strait, another binary indicator showing whether a pilot captain was deployed in the vessel when the failure occurred, vessel length, type and age.

First, the near-miss information is cross-checked with the transit vessels data and the vessels which hosted the failures are identified in order to prepare the logistic regression model data. This process served to identify most of the flag types of the failed vessels from the transit vessels data and this flag type information is imported to the failure data. On the other hand, the information related to failure region is derived from the failure description. For the cases where the region is not properly cited, the distinction between the entrance

and anchorage areas in both directions (north and south) are made via the keywords of the details. For instance, if it is noted that a failure has occurred when the vessel is approaching to the Strait, it is assumed that the failure has occurred at the anchoring region. Moreover, if it is noted that the failure occurred when entering the Strait, it is assumed that it has occurred at the entrance. The north/south identification of these regions is decided via the comparison of the transit vessels and the failure datasets. The cases where the transit regions are not properly distinguished are noted as unknown in the near miss analyses. The failure types are grouped in three main categories:

- Mechanical and Electrical Failures (consisting of the failures such as radar failure, gyro failure, generator failure, navigation light failure, VHF failure, fuel pump failure, rope breakage etc...)
- Engine/Propeller Failures
- Rudder Failures

5.4.2. Near-Miss Descriptions

As in the accident frequency and chi-square analyses, all the attributes in near-miss dataset are transformed into categorical variables. The related frequency results are as follows:

- Vessel Type: As displayed in Table 5.44 and Figure 5.14, the general cargo vessels host most of the near-miss events, while LNG/LPG vessels feature the fewest near-miss events. The failure frequency versus the transit frequency (failure/transit ratio) does not differ significantly over vessel types.

Table 5.44. Frequency of failures according to vessel type

	<i>Failure Frequency</i>	<i>Failure Percent</i>	<i>Transit Frequency</i>	<i>Transit Percent</i>
<i>GENERAL CARGO</i>	403	80.10	128,940	77.82
<i>HAZMAT</i>	11	2.20	3,530	2.13
<i>LNG/LPG</i>	3	0.60	1,068	0.64
<i>PASSENGER</i>	10	2.00	4,667	2.82
<i>TANKER</i>	76	15.10	27,475	16.58
<i>Total</i>	503	100	165,680	100

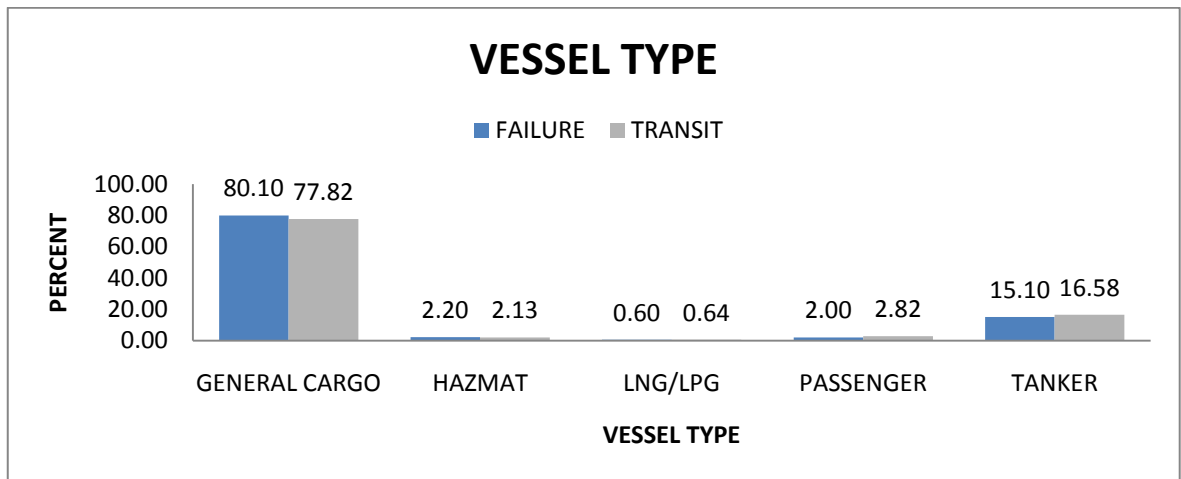


Figure 5.14. Frequency histogram of failures according to vessel type

- **Vessel Length:** As displayed in Table 5.45 and Figure 5.15, the vessels involved in near-miss cases are mostly between 100 and 200 meters. The failure/transit ratio does not differ significantly over vessel types, but is a little higher for vessels of length 200 meters and above.

Table 5.45. Frequency of failures according to vessel length

	<i>Failure Frequency</i>	<i>Failure Percent</i>	<i>Transit Frequency</i>	<i>Transit Percent</i>
<i>100 AND BELOW</i>	201	40.00	64,984	39.22
<i>100-200</i>	280	55.70	89,887	54.25
<i>200 AND ABOVE</i>	22	4.40	10,809	6.52
<i>Total</i>	503	100	165,680	100

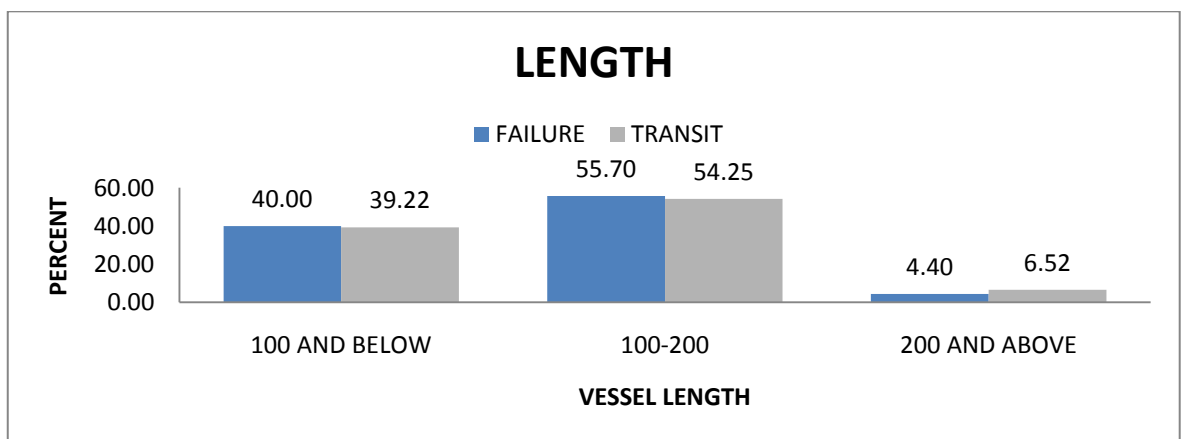


Figure 5.15. Frequency histogram of failures according to vessel length

- **Pilot Captain Status:** As displayed in Table 5.46 and Figure 5.16, about 2/3 of the vessels involved in near-misses did not have a pilot on board. The failure/transit ratio is higher for the vessels without pilot captains.

Table 5.46. Frequency of failures according to pilot captain status

	<i>Failure Frequency</i>	<i>Failure Percent</i>	<i>Transit Frequency</i>	<i>Transit Percent</i>
0	329	65.40	87,965	53.09
1	174	34.60	77,715	46.91
<i>Total</i>	503	100	165,680	100

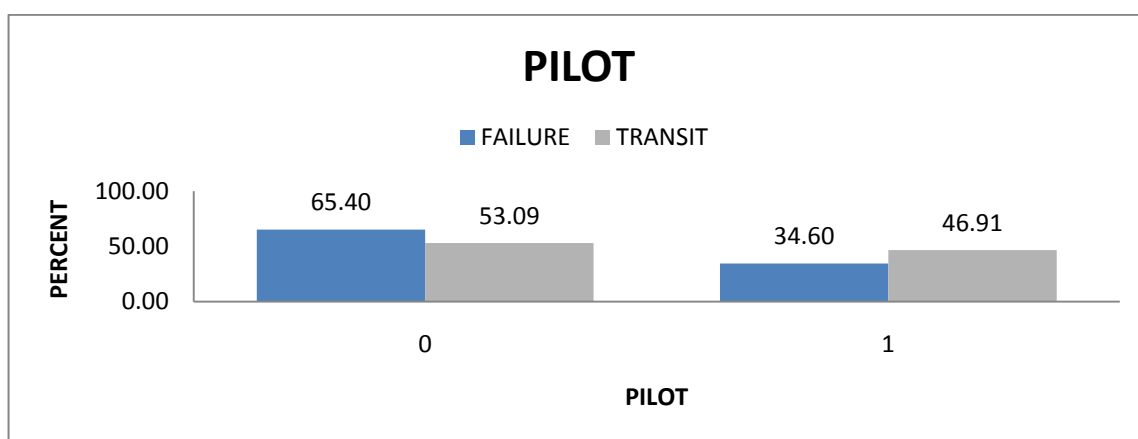


Figure 5.16. Frequency histogram of failures according to pilot captain status

- **Vessel Flag Type:** As displayed in Table 5.47 and Figure 5.17, most of the vessels involved in near-miss cases are of low-risk or medium-risk flag. However, the failure/transit ratio is the highest for the high-risk class of flags.

Table 5.47. Frequency of failures according to vessel flag type

	<i>Failure Frequency</i>	<i>Failure Percent</i>	<i>Transit Frequency</i>	<i>Transit Percent</i>
LOW-RISK	260	51.70	78,370	47.30
MEDIUM-RISK	206	41.00	80,186	48.40
HIGH-RISK	28	5.60	4,966	3.00
UNKNOWN	9	1.80	2,158	1.30
<i>Total</i>	503	100	165,680	100

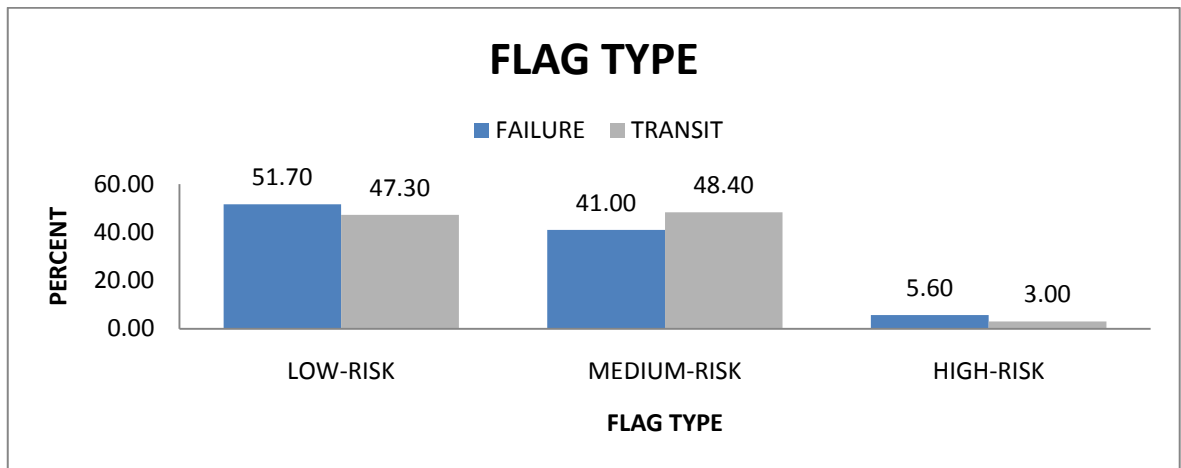


Figure 5.17. Frequency histogram of failures according to vessel flag type

- **Vessel Age:** As displayed in Table 5.48 and Figure 5.18, a significant portion of the vessels in near-miss events are old and there are very few middle-aged vessels involved in near-miss events. The failure/transit ratio is also higher for the old-aged vessels.

Table 5.48. Frequency histogram of failures according to vessel age

	<i>Failure Frequency</i>	<i>Failure Percent</i>	<i>Transit Frequency</i>	<i>Transit Percent</i>
YOUNG	49	9.70	19,619	11.84
MIDDLE-AGED	32	6.40	13,798	8.33
OLD	422	83.90	130,915	79.02
UNKNOWN	0	0.00	1,348	0.81
Total	503	100	165,680	100.00

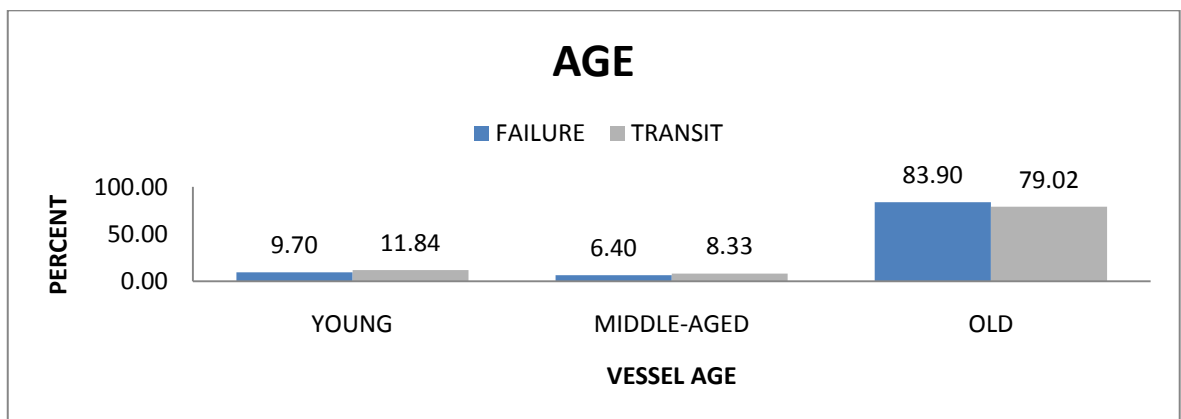


Figure 5.18. Frequency histogram of failures according to vessel age

- **Failure Year:** As displayed in Table 5.49 and Figure 5.19, most near-miss cases occurred in 2007, while the least occurred in 2005, when the 2005-2007 time frame is considered.

Table 5.49. Frequency of failures according to years

	<i>Frequency</i>	<i>Percent</i>
2005	133	26.4
2006	156	31.0
2007	214	42.5
<i>Total</i>	503	100

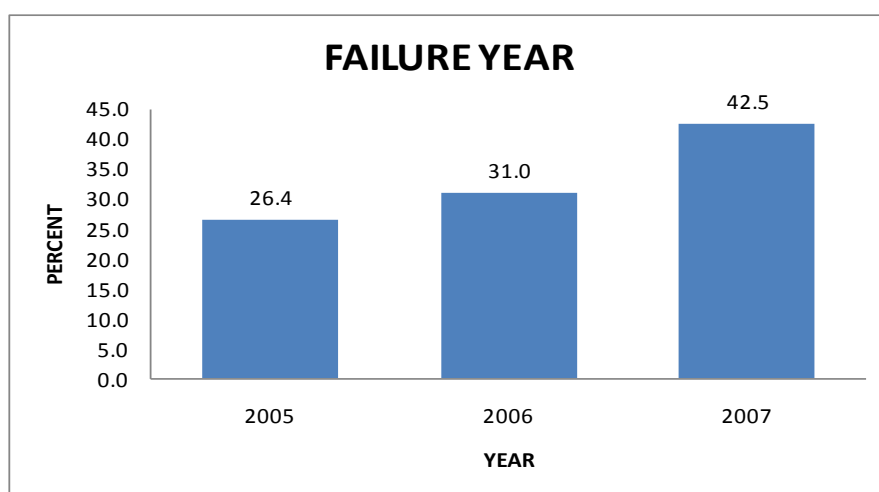


Figure 5.19. Frequency histogram of failures according to years

- **Failure Season:** As displayed in Table 5.50 and Figure 5.20, most of the near-miss events occurred in the winter season. However, the failure/transit ratio shows no significant variation over seasons.

Table 5.50. Frequency of failures according to season

	<i>Failure Frequency</i>	<i>Failure Percent</i>	<i>Transit Frequency</i>	<i>Transit Percent</i>
WINTER	244	48.50	79,336	47.89
SPRING	85	16.90	29,717	17.94
SUMMER	174	34.60	56,627	34.18
<i>Total</i>	503	100	165,680	100

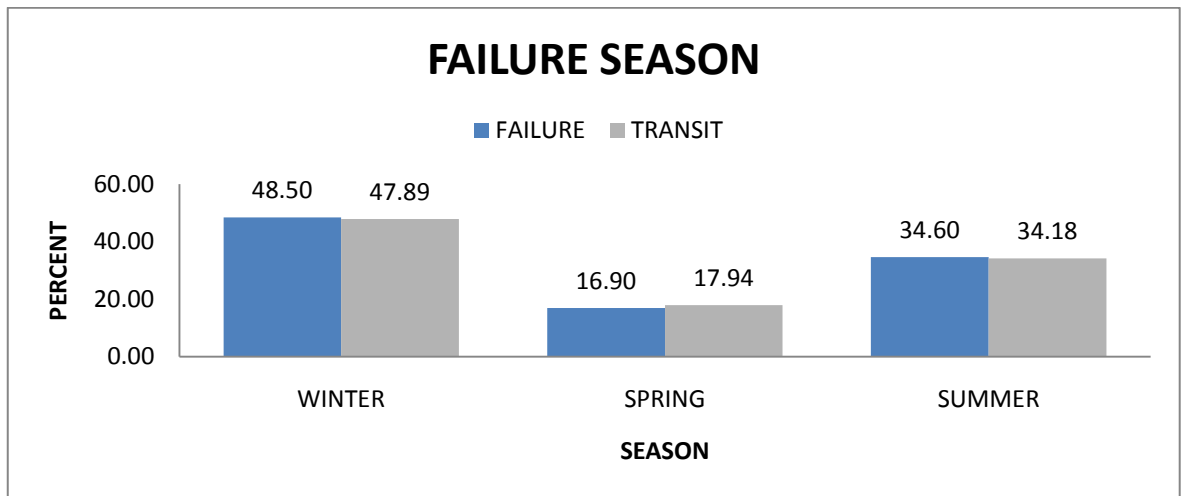


Figure 5.20. Frequency histogram of failures according to season

- **Failure Time Status:** As displayed in Table 5.51 and Figure 5.21, near-miss cases do not differ significantly regarding daytime. But when compared to the transit frequencies, nighttime is seen to be more likely for having a failure.

Table 5.51. Frequency histogram of failures according to time status

	<i>Failure Frequency</i>	<i>Failure Percent</i>	<i>Transit Frequency</i>	<i>Transit Percent</i>
<i>DAY</i>	243	48.30	82,683	49.91
<i>NIGHT</i>	260	51.70	82,997	50.09
<i>Total</i>	503	100	165,680	100

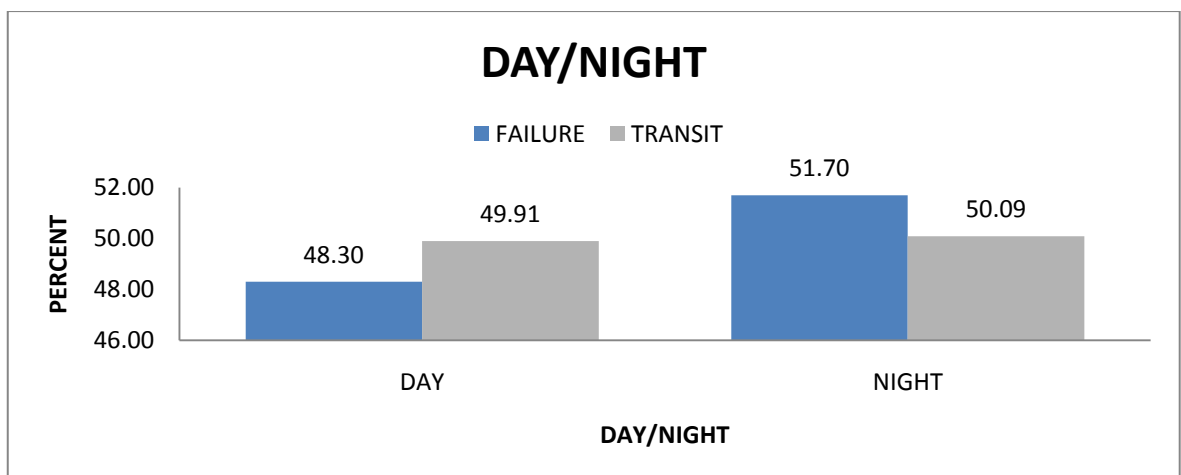


Figure 5.21. Frequency histogram of failures according to time status

- Failure Region: As displayed in Table 5.52 and Figure 5.22, “Region 7” (south anchoring area) has most of the near-miss cases.

Table 5.52. Frequency of failures according to region

	<i>Frequency</i>	<i>Percent</i>
<i>UNKNOWN</i>	42	8.3
<i>1</i>	49	9.7
<i>2</i>	29	5.8
<i>3</i>	27	5.4
<i>4</i>	25	5.0
<i>5</i>	16	3.2
<i>6</i>	80	15.9
<i>7</i>	133	26.4
<i>8</i>	102	20.3
<i>Total</i>	503	100

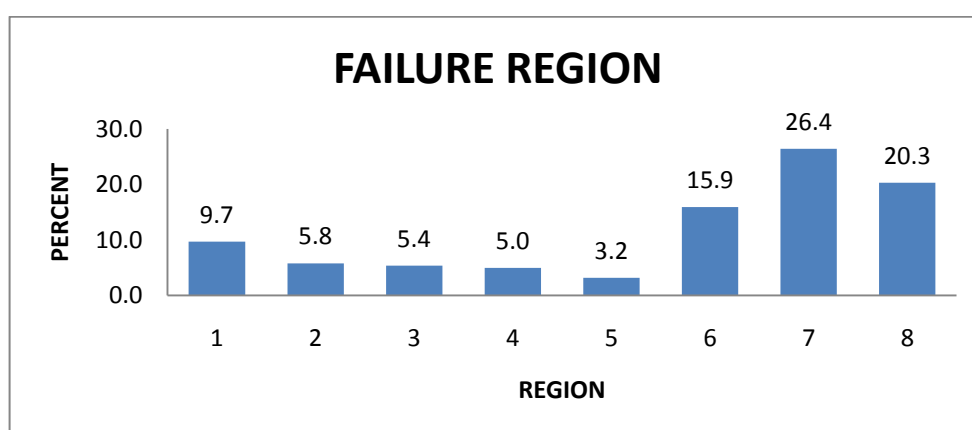


Figure 5.22. Frequency histogram of failures according to region

- Failure Type: As displayed in Table 5.53 and Figure 5.23, the vessels involving in near-miss cases had mostly engine/propeller failures, while rudder failures occurred the least.

Table 5.53. Frequency of failures according to failure type

	<i>Frequency</i>	<i>Percent</i>
<i>ENGINE / PROPELLER FAILURE</i>	323	64.2
<i>MECHANIC AND ELECTRONIC FAILURE</i>	120	23.9
<i>RUDDER FAILURE</i>	60	11.9
<i>Total</i>	503	100

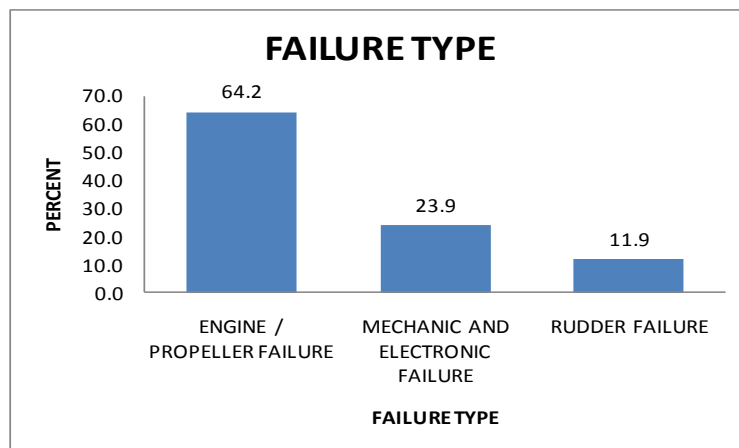


Figure 5.23. Frequency histogram of failures according to failure type

5.4.3. Chi-Square tests and results

As explained in section 5.4.1, the missing cases for the failure dataset are only in flag type and region attributes. Besides, for this analysis, the same missing case elimination methods are performed as in the accident data. The obtained p-values are presented in Table 5.54.

Table 5.54. Chi-square test results regarding possible attributes triggering failures

	TYPE	LENGTH	PILOT	FLAG TYPE	AGE	YEAR	SEASON	DAYTIME	REGION	FAILURE TYPE
TYPE	-	0.000	0.207	0.000	0.000	0.223	0.731	0.245	0.679	0.544
LENGTH	0.000	-	0.057	0.000	0.000	0.391	0.784	0.001	0.003	0.001
PILOT	0.207	0.057	-	0.423	0.385	0.671	0.149	0.446	0.000	0.000
FLAG TYPE	0.000	0.000	0.423	-	0.000	0.030	0.488	0.797	0.297	0.136
AGE	0.000	0.000	0.385	0.000	-	0.627	0.011	0.640	0.568	0.003
YEAR	0.223	0.391	0.671	0.030	0.627	-	0.886	0.171	0.001	0.070
SEASON	0.731	0.784	0.149	0.488	0.011	0.886	-	0.000	0.336	0.177
DAYTIME	0.245	0.001	0.446	0.797	0.640	0.171	0.000	-	0.164	0.915
REGION	0.679	0.003	0.000	0.297	0.568	0.001	0.336	0.164	-	0.000
FAILURE TYPE	0.544	0.001	0.000	0.136	0.003	0.070	0.177	0.915	0.000	-

The interpretations for the cases where the variables are found to be related are given in the following tables.

As displayed in Table 5.55, there is a relationship between vessel length and vessel type regarding near-misses. The majority of passenger vessels that experienced failure between the years 2005 and 2007 are of length 100 meters and below, whereas the other types are mostly of length between 100 and 200 meters.

Table 5.55. Length * Vessel type crosstabulation

			VESSEL TYPE				Total
			GENERAL CARGO	LNG/LPG/HAZMAT	PASSENGER	TANKER	
LENGTH	100 AND BELOW	Count	165	3	8	25	201
		% within TYPE	40.9%	21.4%	80.0%	32.9%	40.0%
	100-200	Count	226	11	2	41	280
		% within TYPE	56.1%	78.6%	20.0%	53.9%	55.7%
	200 AND ABOVE	Count	12	0	0	10	22
		% within TYPE	3.0%	0.0%	0.0%	13.2%	4.4%
Total		Count	403	14	10	76	503
		% within TYPE	100.0%	100.0%	100.0%	100.0%	100.0%

As displayed in Table 5.56, there is a relationship between vessel flag type and vessel type regarding near-misses. Among the vessels which experienced failure between years 2005 and 2007, general cargo vessels are mostly of low and medium-risk flag; LNG/LPG/HazMat and tanker vessels of low-risk flag and passenger vessels are of medium-risk flag.

Table 5.56. Flag type * Vessel type crosstabulation

			VESSEL TYPE				Total
			GENERAL CARGO	LNG/LPG/HAZMAT	PASSENGER	TANKER	
FLAG TYPE	LOW-RISK	Count	181	11	1	67	260
		% within TYPE	45.8%	78.6%	10.0%	89.3%	52.6%
	MEDIUM-RISK	Count	188	3	8	7	206
		% within TYPE	47.6%	21.4%	80.0%	9.3%	41.7%
	HIGH-RISK	Count	26	0	1	1	28
		% within TYPE	6.6%	0.0%	10.0%	1.3%	5.7%
Total		Count	395	14	10	75	494
		% within TYPE	100.0%	100.0%	100.0%	100.0%	100.0%

As displayed in Table 5.57, there is a relationship between vessel age and vessel type regarding near-misses. Among the vessels which experienced failure between years 2005 and 2007, the rates of young vessels of tanker and LNG/LPG/HazMat type are not negligible as of young vessels of general cargo and passenger type.

Table 5.57. Age * Vessel type crosstabulation

			VESSEL TYPE				Total
			GENERAL CARGO	LNG/LPG/HAZMAT	PASSENGER	TANKER	
AGE	YOUNG	Count	28	2	0	19	49
		% within TYPE	6.9%	14.3%	0.0%	25.0%	9.7%
	MIDDLE-AGED	Count	23	2	0	7	32
		% within TYPE	5.7%	14.3%	0.0%	9.2%	6.4%
	OLD	Count	352	10	10	50	422
		% within TYPE	87.3%	71.4%	100.0%	65.8%	83.9%
Total	Count	403	14	10	76	503	
	% within TYPE	100.0%	100.0%	100.0%	100.0%	100.0%	

As displayed in Table 5.58, there is a relationship between vessel flag type and vessel length. Near-miss experienced vessels of length below 100 meters are mostly of medium-risk flag, whereas vessels of length above 100 meters are mostly of low-risk flag.

Table 5.58. Flag type * Length crosstabulation

			LENGTH			Total
			100 AND BELOW	100-200	200 AND ABOVE	
FLAG TYPE	HIGH-RISK	Count	15	13	0	28
		% within LENGTH	7.7%	4.7%	0.0%	5.7%
	LOW-RISK	Count	72	166	22	260
		% within LENGTH	36.9%	59.9%	100.0%	52.6%
	MEDIUM-RISK	Count	108	98	0	206
		% within LENGTH	55.4%	35.4%	0.0%	41.7%
Total	Count	195	277	22	494	
	% within LENGTH	100.0%	100.0%	100.0%	100.0%	

As displayed in Table 5.59, there is a relationship between the length and age of near-miss experienced vessels. The likelihood of middle-aged and old such vessels to be smaller than 100 meters is higher than that of young vessels.

Table 5.59. Length * Age crosstabulation

			AGE			Total
			YOUNG	MIDDLE-AGED	OLD	
LENGTH	100 AND BELOW	Count	7	10	184	201
		% within AGE	14.3%	31.3%	43.6%	40.0%
	100-200	Count	38	18	224	280
		% within AGE	77.6%	56.3%	53.1%	55.7%
	200 AND ABOVE	Count	4	4	14	22
		% within AGE	8.2%	12.5%	3.3%	4.4%
Total	Count	49	32	422	503	
	% within AGE	100.0%	100.0%	100.0%	100.0%	

As displayed in Table 5.60, there is a relationship between failure time and vessel length regarding near-misses. The vessels of length below 200 meters experienced failure mostly at night, whereas vessels of length above 200 meters experienced failure mostly in the day.

Table 5.60. Day/Night * Length crosstabulation

			LENGTH			Total
			100 AND BELOW	100-200	200 AND ABOVE	
DAY/NIGHT	DAY	Count	89	135	19	243
		% within LENGTH	44.3%	48.2%	86.4%	48.3%
	NIGHT	Count	112	145	3	260
		% within LENGTH	55.7%	51.8%	13.6%	51.7%
Total		Count	201	280	22	503
		% within LENGTH	100.0%	100.0%	100.0%	100.0%

As displayed in Table 5.61, there is a relationship between failure region and vessel length regarding near-misses. Vessels of length below 100 meters experienced failure mostly in southern transit region (region 6) and south anchoring area (region 7), whereas vessels of length above 100 meters experienced failure mostly in south anchoring area.

Table 5.61. Region * Length crosstabulation

		LENGTH			Total	
		100 AND BELOW	100-200	200 AND ABOVE		
REGION	1	Count	23	24	2	49
		% within LENGTH	12.8%	9.2%	10.0%	10.6%
	2	Count	19	10	0	29
		% within LENGTH	10.6%	3.8%	0.0%	6.3%
	3	Count	11	16	0	27
		% within LENGTH	6.1%	6.1%	0.0%	5.9%
	4	Count	12	12	1	25
		% within LENGTH	6.7%	4.6%	5.0%	5.4%
	5	Count	9	6	1	16
		% within LENGTH	5.0%	2.3%	5.0%	3.5%
	6	Count	39	40	1	80
		% within LENGTH	21.8%	15.3%	5.0%	17.4%
	7	Count	37	86	10	133
		% within LENGTH	20.7%	32.8%	50.0%	28.9%
	8	Count	29	68	5	102
		% within LENGTH	16.2%	26.0%	25.0%	22.1%
Total	Count	179	262	20	461	
	% within LENGTH	100.0%	100.0%	100.0%	100.0%	

As displayed in Table 5.62, there is a relationship between vessel length and failure type regarding near-misses. Vessels of length below 100 meters experienced failure mostly

in southern transit region (region 6) and south anchoring area (region 7), whereas vessels of length above 100 meters experienced failure mostly in region 7.

Table 5.62. Length * Failure type crosstabulation

			FAILURE TYPE			Total
			ENGINE / PROPELLER FAILURE	MECHANIC AND ELECTRONIC FAILURE	RUDDER FAILURE	
LENGTH	100 AND BELOW	Count	110	54	37	201
		% within FAILURE TYPE	34.1%	45.0%	61.7%	40.0%
	100-200	Count	196	62	22	280
		% within FAILURE TYPE	60.7%	51.7%	36.7%	55.7%
	200 AND ABOVE	Count	17	4	1	22
		% within FAILURE TYPE	5.3%	3.3%	1.7%	4.4%
Total	Count	323	120	60	503	
	% within FAILURE TYPE	100.0%	100.0%	100.0%	100.0%	

As displayed in Table 5.63, there is a relationship between vessel pilot status and failure type regarding near-misses. The majority of vessels experiencing engine/propeller failure does not benefit from pilotage services, whereas the other cases approximately have the same distribution in taking and not taking a pilot.

Table 5.63. Pilot * Failure type crosstabulation

			FAILURE TYPE			Total
			ENGINE / PROPELLER FAILURE	MECHANIC AND ELECTRONIC FAILURE	RUDDER FAILURE	
PILOT	0	Count	244	53	32	329
		% within FAILURE TYPE	75.5%	44.2%	53.3%	65.4%
	1	Count	79	67	28	174
		% within FAILURE TYPE	24.5%	55.8%	46.7%	34.6%
Total		Count	323	120	60	503
		% within FAILURE TYPE	100.0%	100.0%	100.0%	100.0%

As displayed in Table 5.64, there is a relationship between flag type and age regarding near-misses. The majority of young and middle-aged vessels which experienced failure are of low-risk flag, whereas old vessels are approximately equally distributed in low-risk and medium-risk. It can be derived that higher-risk countries have a higher tendency to utilize older vessels.

Table 5.64. Flag type * Age crosstabulation

			AGE			Total
			YOUNG	MIDDLE-AGED	OLD	
FLAG TYPE	LOW-RISK	Count	38	24	198	260
		% within AGE	79.2%	82.8%	47.5%	52.6%
	MEDIUM-RISK	Count	9	5	192	200
		% within AGE	18.8%	17.2%	46.0%	41.7%
	HIGH-RISK	Count	1	0	27	28
		% within AGE	2.1%	0.0%	6.5%	5.7%
Total	Count	48	29	417	494	
	% within AGE	100.0%	100.0%	100.0%	100.0%	

As displayed in Table 5.65, there is a relationship between failure season and vessel age regarding near-misses. The majority of the middle-aged and young vessels experienced failures in winter, whereas the old vessels experienced failures in spring and summer rather than in winter.

Table 5.65. Season * Age crosstabulation

			AGE			Total
			YOUNG	MIDDLE-AGED	OLD	
SEASON	SUMMER	Count	12	6	156	174
		% within AGE	24.5%	18.8%	37.0%	34.6%
	SPRING	Count	5	4	76	85
		% within AGE	10.2%	12.5%	18.0%	16.9%
	WINTER	Count	32	22	190	244
		% within AGE	65.3%	68.8%	45.0%	48.5%
Total	Count	49	32	422	503	
	% within AGE	100.0%	100.0%	100.0%	100.0%	

As displayed in Table 5.66, there is a relationship between failure type and vessel age regarding near-misses. For middle-aged and young vessels, the rate of the engine/propeller failure is significantly high when compared to other failure types. For the old vessels, again the engine propeller failure rate is high but the mechanic/electronic failure is also significant by 26.8%.

Table 5.66. Failure type * Age crosstabulation

			AGE			Total
			YOUNG	MIDDLE-AGED	OLD	
FAILURE TYPE	ENGINE / PROPELLER FAILURE	Count	41	26	256	323
		% within AGE	83.7%	81.3%	60.7%	64.2%
	MECHANIC AND ELECTRONIC FAILURE	Count	5	2	113	120
		% within AGE	10.2%	6.3%	26.8%	23.9%
	RUDDER FAILURE	Count	3	4	53	60
		% within AGE	6.1%	12.5%	12.6%	11.9%
Total	Count	49	32	422	503	
	% within AGE	100.0%	100.0%	100.0%	100.0%	

As displayed in Table 5.67, there is a relationship between failure year and failure region. It is seen that the failures at the region 7 and 8 (namely, south and north anchoring areas) increase year by year.

Table 5.67. Year * Region crosstabulation

			REGION								Total
			1	2	3	4	5	6	7	8	
YEAR	2005	Count	21	11	10	9	7	22	29	17	126
		% within REGION	42.9%	37.9%	37.0%	36.0%	43.8%	27.5%	21.8%	16.7%	27.3%
	2006	Count	15	12	10	10	7	21	36	30	141
		% within REGION	30.6%	41.4%	37.0%	40.0%	43.8%	26.3%	27.1%	29.4%	30.6%
	2007	Count	13	6	7	6	2	37	68	55	194
		% within REGION	26.5%	20.7%	25.9%	24.0%	12.5%	46.3%	51.1%	53.9%	42.1%
Total	Count	49	29	27	25	16	80	133	102	461	
	% within REGION	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	

As displayed in Table 5.68, there is a relationship between season and failure time regarding near-misses. The majority of the day failures occur in spring and summer, whereas the majority of the night failures occur in winter. This is straightforward since the night lasts longer in winter.

Table 5.68. Season * Failure time crosstabulation

			DAY/NIGHT		Total	
			DAY	NIGHT		
SEASON	SUMMER	Count	98	76	174	
		% within DAYTIME	40.3%	29.2%	34.6%	
	SPRING	Count	50	35	85	
		% within DAYTIME	20.6%	13.5%	16.9%	
	WINTER	Count	95	149	244	
		% within DAYTIME	39.1%	57.3%	48.5%	
	Total		Count	243	260	503
			% within DAYTIME	100.0%	100.0%	100.0%

As displayed in Table 5.69, there is a relationship between failure region and failure type regarding near-misses. Even though the engine/propeller failures mostly occur in the south anchoring area (region 7), the majority of mechanic/electronic and rudder failures occur in the southern transit region (region 6).

Table 5.69. Region * Failure type crosstabulation

			FAILURE TYPE			Total
			ENGINE / PROPELLER FAILURE	MECHANIC AND ELECTRONIC FAILURE	RUDDER FAILURE	
REGION	1	Count	31	13	5	49
		% within FAILURE TYPE	10.4%	12.0%	8.9%	10.6%
	2	Count	17	8	4	29
		% within FAILURE TYPE	5.7%	7.4%	7.1%	6.3%
	3	Count	13	4	10	27
		% within FAILURE TYPE	4.4%	3.7%	17.9%	5.9%
	4	Count	15	3	7	25
		% within FAILURE TYPE	5.1%	2.8%	12.5%	5.4%
	5	Count	8	0	8	16
		% within FAILURE TYPE	2.7%	0.0%	14.3%	3.5%
	6	Count	38	31	11	80
		% within FAILURE TYPE	12.8%	28.7%	19.6%	17.4%
	7	Count	100	25	8	133
		% within FAILURE TYPE	33.7%	23.1%	14.3%	28.9%
	8	Count	75	24	3	102
		% within FAILURE TYPE	25.3%	22.2%	5.4%	22.1%
Total		Count	297	108	56	461
		% within FAILURE TYPE	100.0%	100.0%	100.0%	100.0%

5.5. Discriminant Analysis

Discriminant analysis is a tool for modeling the value of a dependent categorical variable according to its relationship to one or more predictors (denoted also as independent variables). It attempts to find linear combinations of the independent variables that best represent the categories of the dependent variable. The concentration of discriminant analysis is profiling, differentiation and classification; whereas the regression analysis deals with description, inference and prediction. “Profiling” is about showing the dissimilarities between groups regarding the independent variables. “Differentiation” relates to the significance about the differences in group means and finally the “classification” associates the observations to the groups (Garson, 2008). The assumptions of the discriminant analysis are:

- The correlation between the independent variables is not high.
- The values of independent variables follow the normal distribution.
- The mean and the variance of an independent variable are not correlated.
- Correlation between any two independent variables is constant.

Discriminant analysis is composed of two concerns. The first one is the F-test (Wilks’ lambda), which helps to assess the significance of the “whole” model. The second

portion is on the individual significances of the independent variables regarding the differences in means (provided that the first portion is significant).

Logistic regression is generally preferred over the discriminant analysis, because logistic regression has less stringent assumptions, is a more robust procedure and accepts both continuous and categorical variables. If the dataset is not normally distributed or the group sizes are unequal, then it is better to use logistic regression method. However, when the assumptions of the discriminant analysis are met, it is better to use the discriminant analysis rather than the logistic regression, since the discriminant analysis is more powerful, meaning that it is less likely to make type-II errors (accepting a false null hypothesis) compared to logistic regression.

In this study, accident causes for each vessel involved in an accident are classified. This classification is derived from some vessel and accident properties. At first, vessel type, GRT, length, flag type, age, pilot captain status as well as season, daytime, accident type and region are chosen as independent variables, which are expressed as categorical variables. The missing values in the accident database are defined as the Category 0. However, the records having missing vessel types or accident causes are excluded in this analysis, since this information is regarded crucial.

The accident cause, which is the dependent variable of the analysis, has four different categories:

- (i) Bad weather conditions;
- (ii) Equipment failure;
- (iii) Geographical structure, human error, traffic density and sabotage grouped as human error;
- (iv) Strong current.

When the independent attributes stated above are incorporated into the analysis, the significance of Wilks' Lambda of the attributes GRT, length and flag type turned out to be insignificant, as displayed in Table 5.70. After eliminating these insignificant variables, the

analysis is repeated and the reduced model is obtained. The significance of the attributes of the reduced model is given in Table 5.71.

Table 5.70. Tests of equality of group means for the full model

	Wilks' Lambda	F	df1	df2	Sig.
TYPE	0.984	3.740	3	693	0.011
GRT	1.000	0.090	3	693	0.965
LENGTH	0.995	1.269	3	693	0.284
FLAG TYPE	0.998	0.460	3	693	0.710
AGE	0.985	3.427	3	693	0.017
PILOT	0.975	5.852	3	693	0.001
SEASON	0.936	15.836	3	693	0.000
DAYTIME	0.982	4.245	3	693	0.006
ACCIDENT TYPE	0.935	16.073	3	693	0.000
REGION	0.942	14.314	3	693	0.000

Table 5.71. Tests of equality of group means for the reduced model

	Wilks' Lambda	F	df1	df2	Sig.
VESSEL TYPE	0.984	3.740	3	693	0.011
AGE	0.985	3.427	3	693	0.017
PILOT	0.975	5.852	3	693	0.001
SEASON	0.936	15.836	3	693	0.000
DAYTIME	0.982	4.245	3	693	0.006
ACCIDENT TYPE	0.935	16.073	3	693	0.000
REGION	0.942	14.314	3	693	0.000

Table 5.72 shows the correlations between the predictors; it can be seen from this table that there are no significantly high correlations in this case.

Table 5.72. Pooled within-group matrices

	VESSEL TYPE	AGE	PILOT	SEASON	DAYTIME	ACCIDENT TYPE	REGION
Correlation VESSEL TYPE	1.000	-0.123	0.099	0.086	-0.079	-0.086	-0.095
AGE	-0.123	1.000	0.038	0.010	0.108	0.147	0.084
PILOT	0.099	0.038	1.000	0.062	-0.052	0.015	-0.124
SEASON	0.086	0.010	0.062	1.000	-0.184	-0.015	-0.087
DAYTIME	-0.079	0.108	-0.052	-0.184	1.000	0.034	-0.016
ACCIDENT TYPE	-0.086	0.147	0.015	-0.015	0.034	1.000	0.011
REGION	-0.095	0.084	-0.124	-0.087	-0.016	0.011	1.000

Log determinants are measures of the variability of each category of the dependent attribute. Larger log determinants correspond to more variable groups and the rank is the number of the independent attributes. According to Table 5.73, the accidents caused by strong current show much more variability inside the group when compared to the remaining groups.

Table 5.73. Box's test of equality of covariance matrices: Log determinants

ACCIDENT CAUSE	Rank	Log Determinant
BAD WEATHER CONDITIONS	7	-0.869
EQUIPMENT FAILURE	7	-0.342
HUMAN ERROR	7	-1.733
STRONG CURRENT	7	-3.327
Pooled within-groups	7	-1.003

As given in Table 5.74, Box's M test tests the assumption of homogeneity of covariance matrices and also gives information about the multivariate normality assumption. The p-value of 0.000 shows that the data differ significantly from multivariate normality, thus, violates an assumption of the analysis. However, discriminant function analysis is robust even when the homogeneity of variances assumption is not met.

Table 5.74. Box's test of equality of covariance matrices: Test results

Box's M	264.462
F	3.055
Approx.	
df1	84
df2	127117.5
Sig.	0.000

Discriminant analysis gives the model in forms of the discriminant functions and the number of the functions is defined as $\min(k-1, p)$, where k is the number of dependent groups and p the number of independent attributes. Since in this study, the independent variable includes four groups and the number of independent attributes is seven, the number of discriminant functions is selected as three. The eigenvalues, which are given in Table 5.75, show how much of the variance in the accident cause, is accounted for by each of the functions. From the table, it can be derived that the function best explaining the variability of the accident cause categories is the "Function 1".

Table 5.75. Summary of canonical discriminant functions: Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	0.169	59.7	59.7	0.380
2	0.108	38	97.7	0.312
3	0.007	2.3	100	0.081

Wilks' lambda shows the significance of each function, which is displayed in Table 5.76. Hence, it can be inferred that the third function does not have a big contribution in the discriminant model since its significance is greater than 0.05.

Table 5.76. Summary of canonical discriminant functions: Wilks' lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 3	0.767	182.780	21	0.000
2 through 3	0.897	75.045	12	0.000
3	0.993	4.510	5	0.479

The standardized discriminant function coefficients in Table 5.77 serve the same purpose as beta weights in the multiple regression analysis: these standardized coefficients indicate the relative importance of the independent attributes in predicting the accident cause for each discriminant function.

Table 5.77. Standardized canonical discriminant function coefficients

	Function		
	1	2	3
VESSEL TYPE	-0.211	0.07	0.696
AGE	0.286	-0.028	0.543
PILOT	-0.289	0.02	0.356
SEASON	0.185	0.891	0.055
DAYTIME	0.194	0.494	-0.216
ACCIDENT TYPE	-0.642	0.355	-0.255
REGION	0.552	0.123	0.104

The structure matrix in Table 5.78 shows the simple correlation between each independent variable and each function. The difference of this table from Table 5.77 is that standardized canonical coefficient interprets the partial contribution of the independent attributes to the functions. According to Table 5.78, "Function 1" is more correlated with

region and accident type; “Function 2” with season and daytime and “Function 3” with vessel type, pilot captain status and vessel age.

Table 5.78. Structure matrix

	Function		
	1	2	3
REGION	0.605(*)	0.030	0.036
ACCIDENT TYPE	-0.576(*)	0.350	-0.237
SEASON	0.078	0.791(*)	0.178
DAYTIME	0.192	0.331(*)	-0.252
VESSEL TYPE	-0.272	0.071	0.699(*)
PILOT	-0.376	0.045	0.443(*)
AGE	0.276	0.090	0.419(*)

Table 5.79 is used to establish the “cutting points” for classifying the cases. In order to determine the cutting points, discriminating scores for each function are processed automatically. For example, when the dependent attributes of a vessel are inserted into “Function 1”, if its score is closer to 0.351, it means that this vessel most probably involved in a human error based accident. When it is closer to -0.100, the accident is most probably due to the bad visibility conditions. Accordingly, the optimal cutting point is the weighted average of the scores taken pairwise. For instance, the average for these two causes is 0.1255, meaning that if the vessel has a score below 0.1255, it enters the group “accidents caused by the bad visibility conditions” and if it is above, it enters the group “accidents caused by human error”.

Table 5.79. Functions at group centroids

ACCIDENT CAUSE	Function		
	1	2	3
BAD WEATHER CONDITIONS	-0.100	-0.513	0.061
EQUIPMENT FAILURE	-0.741	0.348	0.024
HUMAN ERROR	0.351	0.167	0.001
STRONG CURRENT	-0.225	-0.301	-0.274

Table 5.80 shows the classification function coefficients output, when the Fisher’s function coefficients are selected under the “Statistics” options. In this table, unstandardized linear discriminant coefficients are displayed, and then the coefficients are used to calculate the scores and to classify the vessels.

Table 5.80. Classification function coefficients

	ACCIDENT CAUSE			
	BAD WEATHER CONDITIONS	EQUIPMENT FAILURE	HUMAN ERROR	STRONG CURRENT
VESSEL TYPE	1.327	1.450	1.263	1.189
AGE	0.379	0.213	0.435	0.216
PILOT	1.786	2.524	1.245	1.476
SEASON	2.561	3.352	3.401	2.741
DAYTIME	4.259	4.762	4.969	4.508
ACCIDENT TYPE	1.198	1.755	1.173	1.383
REGION	1.386	1.264	1.545	1.348
(Constant)	-12.136	-14.942	-15.118	-12.428

Table 5.81 is used to assess how well the discriminant function works, and whether it works equally well for each accident cause. All the accident cause categories are correctly predicted in different percentages. For example, it can be seen that the correct prediction probability of equipment failure is 52.0%, when cross-validated group is taken into consideration. As a result, the model in question has 44.6% efficiency in grouping the vessels according to the correct accident cause. This is not a very strong result, which may mainly be caused by the problem that the data covers the vessels which are involved in an accident instead of the vessels causing an accident.

Table 5.81. Classification Results

		ACCIDENT CAUSE	Predicted Group Membership				Total
			BAD WEATHER CONDITIONS	EQUIPMENT FAILURE	HUMAN ERROR	STRONG CURRENT	
Original	Count	BAD WEATHER CONDITIONS	71	25	50	24	170
		EQUIPMENT FAILURE	23	67	18	19	127
		HUMAN ERROR	71	59	175	44	349
		STRONG CURRENT	14	17	7	13	51
	%	BAD WEATHER CONDITIONS	41.8	14.7	29.4	14.1	100.0
		EQUIPMENT FAILURE	18.1	52.8	14.2	15.0	100.0
		HUMAN ERROR	20.3	16.9	50.1	12.6	100.0
		STRONG CURRENT	27.5	33.3	13.7	25.5	100.0
Cross-validated	Count	BAD WEATHER CONDITIONS	63	25	50	32	170
		EQUIPMENT FAILURE	24	66	18	19	127
		HUMAN ERROR	71	60	174	44	349
		STRONG CURRENT	15	17	11	8	51
	%	BAD WEATHER CONDITIONS	37.1	14.7	29.4	18.8	100.0
		EQUIPMENT FAILURE	18.9	52.0	14.2	15.0	100.0
		HUMAN ERROR	20.3	17.2	49.9	12.6	100.0
		STRONG CURRENT	29.4	33.3	21.6	15.7	100.0

- Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.
- 46.8% of original grouped cases correctly classified.
- 44.6% of cross-validated grouped cases correctly classified.

5.6. Cluster Analysis

Cluster analysis is a statistical data analysis technique used for assigning of a set of observations into coherent subsets called clusters. The goal in cluster analysis is to identify a set of groups which minimizes within-group variation and maximizes between-group variation.

There are different types of classifications of clustering methods. In SPSS, clustering analysis is divided into three sections. The first one is hierarchical clustering, where successive clusters are sought based on previously established clusters. It can be either agglomerative or divisive. Agglomerative algorithms start with many clusters and merge them, whereas divisive algorithms start with one and divide it into pieces. The second type of clustering in SPSS is *K*-means clustering, where the user enters the number of desired clusters and SPSS determines how to assign observations to *K* clusters. Thirdly, there is a two-step clustering procedure, which is a good choice when the data is large and categorical. It has the largest array of output options, including variable importance plots. This procedure gives the user the choice of entering the desired number of clusters or it can determine the optimum number of clusters itself. In the first step of the algorithm, the dataset is split into pre-clusters. Then each sub-class is separately divided into clusters using the hierarchical methods.

The two-step cluster analysis also works if the variables are continuous. In two-step cluster analysis, there are two types of distance measures. One of them is the Log-likelihood distance and the other is the Euclidean distance. If all the attributes of the observations are continuous, the Euclidean distance must be used, whereas if at least one of the attributes is categorical, the Log-likelihood distance is applied. The assumptions of these distance measures are that the attributes in the model are independent. Further, each continuous variable is assumed to have a normal (Gaussian) distribution and each categorical variable is assumed to have a multinomial distribution. It should also be noted that even when these assumptions are not fully met, the procedure is still fairly a robust one.

In this study, the vessels of the accident data are attempted to be clustered in order to reveal similarities and groupings for all the data. As in the chi-square testing and discriminant analysis, the vessels that have some missing values are removed. By trial and error; combinations of vessel type, pilot, accident type, region and accident cause reveal useful information and clusters. The outputs of the procedure are displayed in the following tables.

Table 5.82 shows the advance of the process and the iterations in terms of the Bayesian Information Criterion (BIC). In general, the smallest BIC value represents the best number of clusters. But in some cases, the BIC values tend to decrease as the number of clusters increases. Additionally, the ratio of the BIC changes is also tracked: decreasing BIC change means that BIC values start to decrease more slowly. So, even though the BIC still continues to decrease, the model gets more and more complex. In these cases, it is better to consider the “Ratio of Distance Measures”, which is based on the current number of clusters against the previous number of clusters, along with “Ratio of BIC Changes”, since the philosophy of all the clustering algorithms is to minimize within-group variances and maximize between-group variances. In this case, the largest combination of distance measure is observed as 1.628. Hence, three is chosen as the number of clusters.

Table 5.82. Auto-clustering table

Number of Clusters	Schwarz's Bayesian Criterion (BIC)	BIC Change	Ratio of BIC Changes	Ratio of Distance Measures
1	7,602.236			
2	6,753.315	-848.921	1.000	1.518
3	6,234.034	-519.282	0.612	1.628
4	5,960.478	-273.556	0.322	1.120
5	5,728.767	-231.710	0.273	1.139
6	5,539.564	-189.203	0.223	1.318
7	5,424.387	-115.176	0.136	1.045
8	5,319.191	-105.196	0.124	1.087
9	5,231.837	-87.355	0.103	1.172
10	5,174.567	-57.270	0.067	1.067
11	5,128.261	-46.305	0.055	1.112
12	5,098.407	-29.854	0.035	1.022
13	5,071.785	-26.623	0.031	1.023
14	5,048.446	-23.339	0.027	1.163
15	5,044.911	-3.535	0.004	1.004

Table 5.83 summarizes the determined number of clusters along with the number and the percentage of the cases that fall into the related groups. Tables 5.84, 5.85, 5.86, 5.87 and 5.88 show the detailed cluster profiles for each attribute.

Table 5.83. Cluster distribution

	N	% of Combined	% of Total
Cluster 1	260	37.7%	37.7%
2	200	29.0%	29.0%
3	229	33.2%	33.2%
Combined	689	100.0%	100.0%
Total	689		100.0%

As can be observed from Table 5.84, a third of the general cargo vessels and two thirds of tanker vessels fall into Clusters 3, whereas the majority of local and passenger vessels fall into Cluster 2. Cluster 1 is absolutely covered by general cargo vessels.

Table 5.84. Cluster profile regarding vessel type

	Cluster							
	1		2		3		Combined	
	Frequency	Percent	Frequency	Percent	Frequency	Percent	Frequency	Percent
GENERAL CARGO	260	55.2%	44	9.3%	167	35.5%	471	100.0%
LOCAL	0	0.0%	106	92.2%	9	7.8%	115	100.0%
PASSENGER	0	0.0%	20	71.4%	8	28.6%	28	100.0%
TANKER	0	0.0%	30	40.0%	45	60.0%	75	100.0%

As displayed in Table 5.85, the vessels having pilots on board fall into Cluster 3, indicating that mostly the tanker and general cargo vessels benefit from this service rather than local and passenger vessels.

Table 5.85. Cluster profile regarding the pilot status of vessels

	Cluster							
	1		2		3		Combined	
	Frequency	Percent	Frequency	Percent	Frequency	Percent	Frequency	Percent
0	260	40.7%	198	31.0%	181	28.3%	639	100.0%
1	0	0.0%	2	4.0%	48	96.0%	50	100.0%

As can be observed from Table 5.86, collision accidents mostly fall into Clusters 1 and 2, while fire & explosion accidents fall into Clusters 2 and 3, the majority of the grounding accidents fall into Cluster 3, ramming accidents into Cluster 1 and sinking accidents into Cluster 2.

Table 5.86. Cluster profile regarding accident type

	Cluster							
	1		2		3		Combined	
	Frequency	Percent	Frequency	Percent	Frequency	Percent	Frequency	Percent
COLLISION	168	48.3%	148	42.5%	32	9.2%	348	100.0%
FIRE & EXPLOSION	0	0.0%	16	55.2%	13	44.8%	29	100.0%
GROUNDING	0	0.0%	3	2.3%	129	97.7%	132	100.0%
RAMMING	86	53.1%	21	13.0%	55	34.0%	162	100.0%
SINKING	6	33.3%	12	66.7%	0	0.0%	18	100.0%

As can be observed from Table 5.87, accidents that occurred in transit regions mostly fall into Clusters 2 and 3, whereas south anchoring area accidents fall into Cluster 1.

Table 5.87. Cluster profile regarding accident region

	Cluster							
	1		2		3		Combined	
	Frequency	Percent	Frequency	Percent	Frequency	Percent	Frequency	Percent
1	18	47.4%	8	21.1%	12	31.6%	38	100.0%
2	15	21.7%	4	5.8%	50	72.5%	69	100.0%
3	13	22.8%	5	8.8%	39	68.4%	57	100.0%
4	17	22.4%	13	17.1%	46	60.5%	76	100.0%
5	0	0.0%	30	68.2%	14	31.8%	44	100.0%
6	0	0.0%	103	83.7%	20	16.3%	123	100.0%
7	186	73.2%	34	13.4%	34	13.4%	254	100.0%
8	11	39.3%	3	10.7%	14	50.0%	28	100.0%

As can be observed from Table 5.88, the accidents caused by failures and strong current are mostly grouped in Cluster 3, whereas the accidents caused by bad weather conditions and human errors fall into Cluster 1. Noting that Cluster 3 includes the vessels with pilot on board, it can be derived that the pilot has a contributive effect on preventing human error related accidents.

Table 5.88. Cluster profile for the accident cause

	Cluster							
	1		2		3		Combined	
	Frequency	Percent	Frequency	Percent	Frequency	Percent	Frequency	Percent
BAD WEATHER CONDITIONS	84	49.7%	41	24.3%	44	26.0%	169	100.0%
FAILURE	0	0.0%	27	22.1%	95	77.9%	122	100.0%
HUMAN ERROR	176	50.7%	117	33.7%	54	15.6%	347	100.0%
STRONG CURRENT	0	0.0%	15	29.4%	36	70.6%	51	100.0%

Figures 5.24, 5.25 and 5.26 display the “Categorical Variablewise Importance” tables which show the importance ranking of each attribute in the model regarding all three clusters. According to these tables, the first cluster introduces the region, the second cluster introduces the vessel type and the third cluster introduces the accident type as the most important variables. When the critical values are taken into account, only in Cluster 2, accident cause is below the critical value and therefore less important.

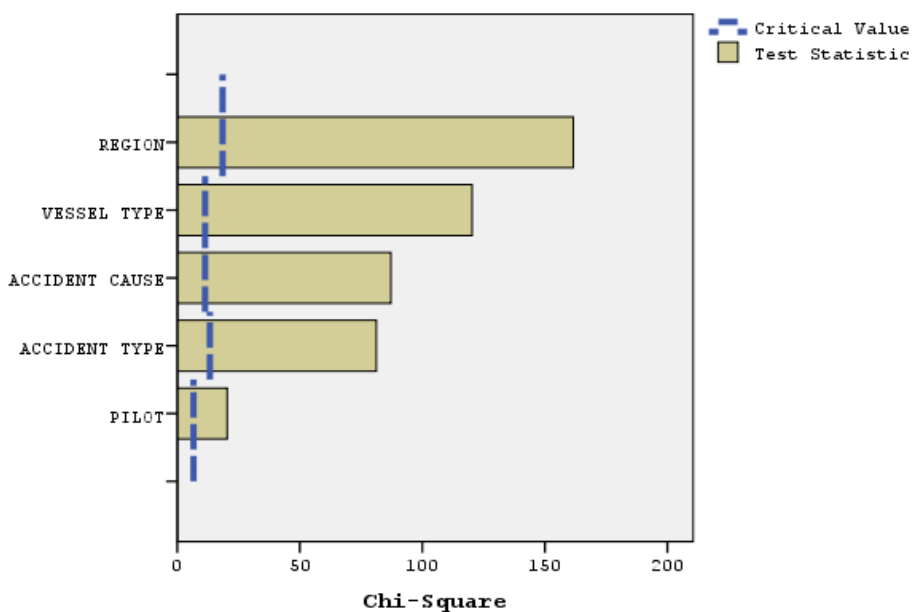


Figure 5.24. Categorical variablewise importance of Cluster 1

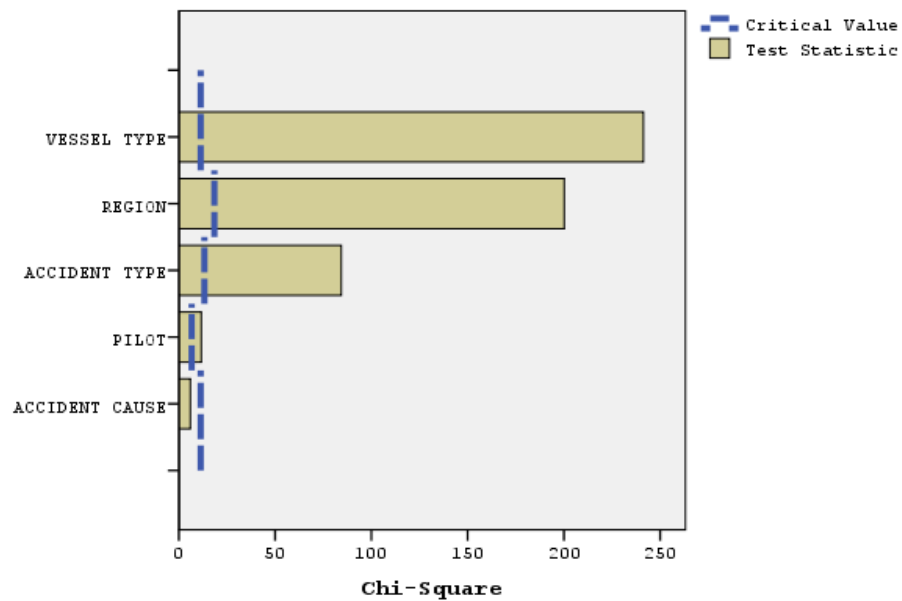


Figure 5.25. Categorical variablewise importance of Cluster 2

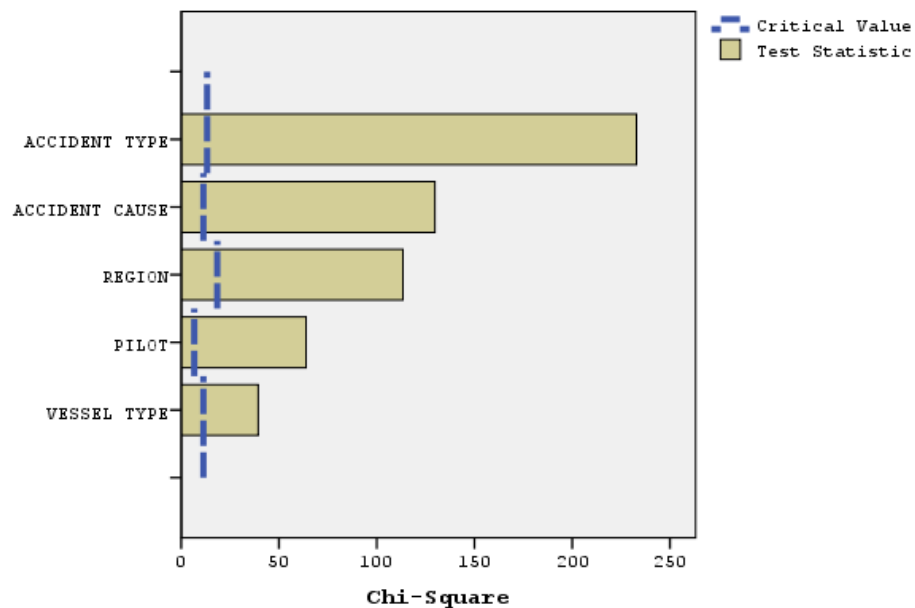


Figure 5.26. Categorical variablewise importance of Cluster 3

As a result, three vessel clusters are generated for the vessels that are involved in an accident according to the vessel type, pilot status, accident type, region and accident cause. Cluster 1 is composed of general cargo vessels not benefiting from pilotage services, which are subject to the collision or ramming accidents mostly occurring in the south anchoring area and caused by human error and bad weather conditions. Cluster 2 is the group of passenger and local vessels, which do not have a pilot on board and which mostly have

collision accidents in southern transit region. Cluster 3 is composed of general cargo and tanker vessels including all the vessels with pilot on board with a frequency of 20%. Their accident types are mostly grounding and ramming. Moreover, the majority of the vessels with accidents caused by the engine failures are grouped in this cluster.

6. LOGISTIC REGRESSION MODEL OF ACCIDENT LIKELIHOOD

In this section, six different logistic regression models are obtained in order to find out the most significant factors contributing to a vessel's involvement in an accident and in order to find the probability of accident occurrence for each vessel. For this purpose, the explaining variables are proposed as vessel type cardinality, gross tonnage, vessel length, age & flag cardinality, pilot request, season, day/night status, visibility condition, local traffic density, engine/propeller failure, rudder failure, mechanic and electronic failure and human error. After the determination of models' coefficients, the verification and validation of the models are accomplished. The verification study is based on the utilized dataset itself (which is based on the years 2005-2007). In the validation process two distinct datasets are used: the first one is the transit, accident and near-miss dataset of year 2009 and the second one is composed of the earlier accidents. Moreover, the weaknesses of the models are discussed at the end of the chapter.

6.1. Logistic Regression Definition

Logistic regression is a special case of multiple regression, where the outcome (dependent) variable of the model is dichotomous (separated into two groups) and the independent variables are either continuous or categorical. It is especially applied in biomedical literature and social sciences.

In simple linear and multiple regressions, the value of the dependent variable Y is calculated in terms of the independent variables X s. In contrast, in logistic regression, the X s are used to predict the probability of occurrence of Y , rather than its exact value. Logistic regression equation is defined by the formula

$$\text{logit } p = \log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \quad (6.1)$$

which can be transformed into

$$P(Y) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)}} \quad (6.2)$$

where $P(Y) = p$ is the probability of occurrence of Y ; X_i is the value of the i^{th} predictor and β_i is the regression coefficient associated with X_i .

Even though linear and logistic regressions may look similar, they should be used in different cases. If the outcome variable is a dichotomy, then simple and multiple regressions cannot be applied. The reason is that, in most of the cases, linearity assumption is not satisfied. To overcome the problem of violating model assumptions, the data is logarithmically transformed, which gives rise to logistic regression.

When (6.2) is taken into account, since left-hand side of the equation is a probabilistic value, it varies between 0 and 1 and the more $P(Y)$ is closer to 0, the less likely that the event Y occurs. In addition to $P(Y)$, the model also calculates appropriate β_i s which is estimated with the help of maximum likelihood principle.

To analyze how good the model is fitted, a measure can be built using observed and predicted values: the log-likelihood measure formula is,

$$\ell\ell = \sum_{i=1}^N \left(Y_i \ln(P(Y_i)) + (1 - Y_i) \ln(1 - P(Y_i)) \right) \quad (6.3)$$

where, Y_i indicates the observed output value and $P(Y_i)$ the probability computed by the model. Log likelihood measure can be approximated by the chi-square distribution.

Apart from the significance of the model, the significance of each predictor is also important for the interpretation of the model. Therefore, to test whether β_i coefficients for each corresponding X_i are significantly different from zero, Wald statistic is calculated by,

$$z = \frac{\hat{\beta}}{SE_{\beta}} \quad (6.4)$$

where, $\hat{\beta}$ is the estimated β value and $SE_{\hat{\beta}}$ is the standard error of the estimate. Wald statistic follows the chi-square distribution. So, if the coefficient β_i is found to be significant, that means that its associated predictor X_i has a considerable contribution to the output variable Y (Field, 2005).

In this study, logistic regression model is used to determine the significant factors contributing to the probability of the occurrence of an accident in the Strait of Istanbul. In this context, for each vessel, the probability of being involved in an accident is predicted based on the level of a set of proposed “accident triggering” factors, whose weights “ β ” are estimated by the model.

6.2. Logistic Regression Methodology and Description of the Data

In order to construct the logistic regression model, a database constituting of past vessel transits (with no incidents), past maritime accidents and near accidents (in the context of failures) covering the years 2005 to 2007 is compiled. The integration of these three distinct datasets is done via identifying common standardized vessel and environment information regarding the vessels which were involved in an accident or exposed to a failure or accomplished a routine transit. The primary information used in this process is vessel type, length, age, flag type, accident/failure time (in case of accident/failure events) with transit time (in case of routine transits).

Additionally, there are also some missing cases in transit vessels data. For 3,524 cases out of 165,680, at least one value regarding gross tonnage (GRT), length, age or flag information is missing. These missing cases are completed with the aid of the dataset itself in order to prevent excluding all these cases from the analysis. Since a good number of the transit vessels pass through the Strait of Istanbul several times in a year, or have very similar characteristics to other vessels that have done the transit, the vessels with same type and length (GRT) are taken into account to complete the missing cases for GRT (length). In addition, to complete the missing flag and age information, the frequencies for each category (low, medium, high risk for flag; young, middle-aged, old for age) are calculated for the non-missing cases, and then these frequencies are used in the random regeneration of the missing data.

As explained in Chapter 5, accidents are categorized by their nature as collision, ramming, fire & explosion, grounding and sinking. Especially in the majority of the collision and ramming accidents, more than one vessel play a role but the main responsible party is not cited properly. For this reason, the analysis in this section is related to the act of being involved in an accident, rather than causing an accident. Hence, the response variable is defined as;

$$Y = \begin{cases} 1 & \text{if the vessel is involved in an accident} \\ 0 & \text{if the vessel is not involved} \end{cases} \quad (6.5)$$

The accident database contains not only the accidents that have occurred in actual transit, but also the accidents that occurred in the north and south anchoring areas. But these accidents are different in nature, since they do not have the same characteristics. For instance, transit accidents occur while the vessels are moving along the Strait, whereas anchoring area accidents mostly occur because they drag anchor due to strong current and wind. Moreover, for a transit accident, whether the vessel benefits from the pilotage services or not may be an important factor, whereas this is not an issue for the accidents that occurred in the anchoring areas. Therefore, the logistic regression analysis is accomplished only for the actual transit cases. On the other hand, it is attempted to construct a model concerning the south anchorage area accidents, but since reliable results could not be inferred from the output, it is excluded from the analysis.

6.2.1. Cardinality Value Determination

Although some of the independent variables deployed in a logistic regression model can also be categorical, in this study, it is attempted to transform categorical variables into ordinary variables taking cardinal values as much as possible (by taking into account the potential of these attributes to play a role in a maritime accident). The aim of this attempt is to facilitate the interpretation of the model coefficients.

6.2.1.1. Type Cardinality. In the accident dataset, vessel type is split into four categories: passenger, LNG/LPG/HazMat, tanker and general cargo. To introduce this factor to the logistic regression model as a cardinal variable, type cardinality values associated with

each vessel type are estimated. These estimations are done based on the relative frequencies of these vessel types both in accidents and in routine transits in the years 2005-2007. For this purpose, the total number of transit passages and accidents are found for each vessel type. Then, the frequency of accidents P_i s are calculated through the formula,

$$P_i = \frac{\text{total \# of vessels of type } i \text{ involved in an accident}}{\text{total \# of vessels of type } i} \quad (6.6)$$

Next, these frequencies are normalized with respect to the maximum P_i , which are then deployed as “type cardinality” values. The results are displayed in Table 6.1.

Table 6.1. Type Cardinality Values

VESSEL TYPE	NUMBER OF TOTAL ACCIDENTS	NUMBER OF TRANSITS	ACCIDENT FREQUENCIES	NORMALIZED ACCIDENT FREQUENCIES
PASSENGER	3	4667	0.000643	0.596288
HAZMAT-LPG	2	4598	0.000435	0.403491
TANKER	12	27474	0.000437	0.405165
GENERAL CARGO	139	128940	0.001078	1.000000

6.2.1.2. Age & Flag Cardinality. In the accident data, there exist three categories both for vessel age (young, medium, old) and for vessel flag (low risk, medium risk, high risk) attributes. In order to include both factors in the logistic regression model, a combined age & flag cardinality value is computed. For such age-flag cardinality estimations, a previous survey that the captains participated is used. On this survey, the experts were asked to grade the potential of having an accident on a 0 to 100 scale, for every age-flag combination. The answers are firstly normalized with respect to the maximum score of each captain. Then, for each age-flag category, the average of all the normalized values is obtained. The resulting values, which are deployed as age/flag cardinality values, are displayed in Table 6.2.

Table 6.2. Age & Flag Cardinality Values

FLAG/AGE	YOUNG	MIDDLE-AGED	OLD
LOW-RISK	0.032273	0.157812	0.380215
MEDIUM-RISK	0.092234	0.266905	0.616893
HIGH-RISK	0.298724	0.572080	1.000000

6.2.2. Correlation Analysis

Since the independent variables deployed in the logistic regression model must, in theory, be independent of one another, a correlation analysis is performed for the variables, which are at least ordinal. (These variables are also included in the analysis based on the chi-square analysis of the previous chapter.) The correlation analysis is then deployed as a guide in order to decide the variables that will enter the logistic regression analysis (highly correlated variables being excluded from the model). In this context, transit vessel dataset covering the years 2005-2007, which is furnished by the type and age & flag cardinality values, is used.

The initial set of variables considered are type cardinality (TYCRD), gross tonnage (GRT), vessel length (LEN), age & flag cardinality (AFL), pilot request (PIL), season (SEA) and daytime (DAY), respectively. The results are presented in Table 6.3.

Table 6.3. Correlation analysis regarding all considered logistic regression independent variables

		TYCRD	GRT	LEN	PIL	AFL	SEA	DAY
TYCRD	Pearson Correlation	1	-.374	-.297	-.256	.368	.005	.023
	Sig. (2-tailed)		.000	.000	.000	.000	.066	.000
	N	165680	165680	165680	165680	165680	165680	165680
GRT	Pearson Correlation	-.374	1	.884	.368	-.469	-.003	-.140
	Sig. (2-tailed)	.000		.000	.000	.000	.193	.000
	N	165680	165680	165680	165680	165680	165680	165680
LEN	Pearson Correlation	-.297	.884	1	.411	-.492	-.009	-.095
	Sig. (2-tailed)	.000	.000		.000	.000	.000	.000
	N	165680	165680	165680	165680	165680	165680	165680
PIL	Pearson Correlation	-.256	.368	.411	1	-.254	-.012	.011
	Sig. (2-tailed)	.000	.000	.000		.000	.000	.000
	N	165680	165680	165680	165680	165680	165680	165680
AFL	Pearson Correlation	.368	-.469	-.492	-.254	1	.007	.041
	Sig. (2-tailed)	.000	.000	.000	.000		.003	.000
	N	165680	165680	165680	165680	165680	165680	165680
SEA	Pearson Correlation	.005	-.003	-.009	-.012	.007	1	-.127
	Sig. (2-tailed)	.066	.193	.000	.000	.003		.000
	N	165680	165680	165680	165680	165680	165680	165680
DAY	Pearson Correlation	.023	-.140	-.095	.011	.041	-.127	1
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	
	N	165680	165680	165680	165680	165680	165680	165680

It can be seen that the factors related to vessel (TYCRD, GRT, LEN, AFL and PIL) are all correlated and the absolute values of correlation coefficients vary between 0.256 and 0.884. On the other hand, the correlations related to season and daytime are not too

strong (the lower limit for the correlation significance is assumed to be 0.2). Hence, it is decided not to include all the vessel attributes in the logistic regression model at the same time or to use combined interaction variables.

6.3. The Logistic Regression Model for Transit Accidents

6.3.1. Model Inputs

The candidate predictor variables considered are as follows:

- **Type Cardinality (TYCRD):** This attribute takes one of four different values, between 0 and 1, associated with the four different vessel types, indicating the (normalized) potential of that vessel type to be involved in an accident, as calculated in Section 6.2.1.1.
- **Gross Tonnage (GRT):** This attribute is the actual GRT of the vessel.
- **Length (LEN):** The attribute LEN is the actual vessel length value. Since LEN and GRT turned out to be highly correlated (as expected), one needed to be given preference (regarding inclusion in the model). The attribute LEN is given preference, since it is considered to be a more important characteristic regarding accident causality.
- **Age & Flag Cardinality (AFL):** This attribute is provided from the cardinality value determination of Section 6.2.1.2. It takes nine numeric values (between 0 and 1) for each combination of age and flag categories defined in Chapter 5.
- **Pilot Request (PIL):** This attribute has a binary nature. “1” indicates that there is a pilot in the vessel under consideration and “0” indicates that there is no pilot.
- **Season (SEA):** The seasons are based on the seasonality study of the visibility conditions, which is explained in Section 4.1.1. In this context, “1” indicates the winter, “2” indicates the transition and “3” indicates the summer season.

- Daytime (DAY): This attribute is based on the data obtained from Kandilli Observatory and Earthquake Research Institute, which is also used in Chapter 5. According to this dataset, “0” indicates daytime and “1” indicates nighttime.
- Visibility Condition (VIS): This is also a binary variable, where “0” indicates clear sight and “1” indicates bad visibility conditions. The accident database includes visibility condition information on the vessels involved in an accident. For the vessels which are not involved in an accident, adjusted fog model generated from the Weather Underground data is used. In adapting this data to the logistic regression dataset, vessel entrance and exit times are taken into consideration. The exit time of the vessels is calculated by adding three hours to their entrance time, (which is the average duration of completing a passage throughout the Strait). Accordingly, if there is at least one “1” value in the fog data between the entrance and exit times of a vessel, bad weather condition is assumed to exist during the transit of this vessel. Otherwise, if there is no fog during the assumed transit window of a vessel, bad weather condition is set to “0” value.
- Local Traffic Density (LOC): This attribute indicates the number of local traffic vessels that a vessel meets during its transit. It is included in the model in order to see whether the local traffic density has any effect on accident likelihood. Since the accident, traffic density and near miss databases do not contain any information of that kind; suitable values for this attribute are obtained from the simulation model. Moreover, although local traffic density is primarily related to region, a time-based approach is used to describe the density since the logistic regression model does not contain any information about the region. The output file of the simulation model contains number of local traffic elements for 21 slices (from north to south) in the Strait, at ten-minute intervals, during 24 hours. The outline of the output is presented in Figure 6.1. The values of the matrix show the number of vessels at a specific time interval, specific slice. Since the average time of the passage from the Strait is assumed to be three hours, a vessel transits each slice approximately in ten minutes. If the diagonal values from left to right are summed up, it gives the total number of vessels that a vessel will encounter during passing the Strait from north to south and it is presented as A_i for each time interval i in the figure. Similarly, diagonal sum

taken from right to left gives the number of local traffic elements for the vessels moving from south to north (B_i).

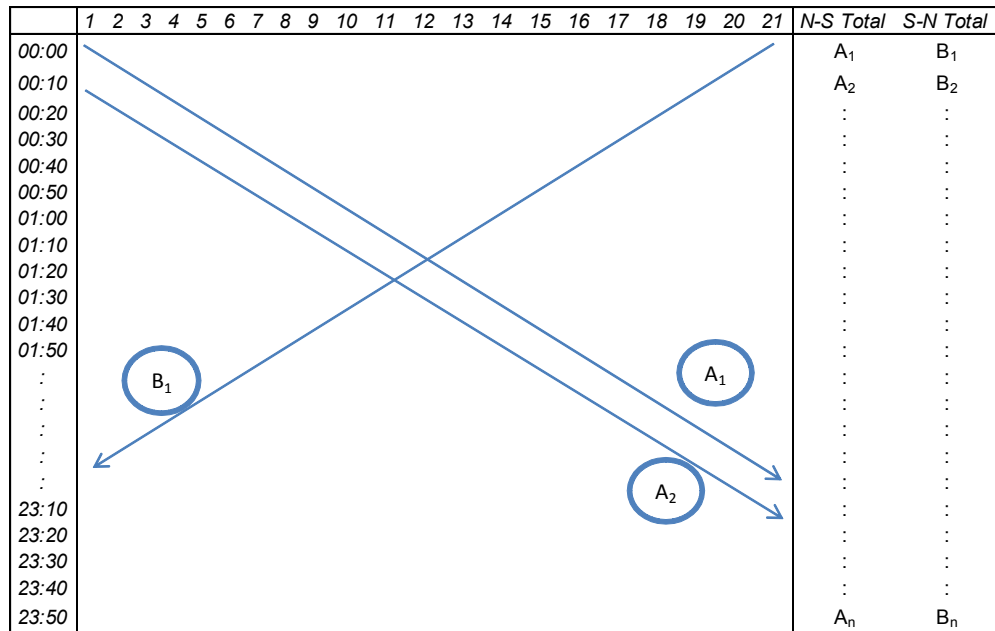


Figure 6.1. Local traffic density calculation

- Engine/Propeller Failure (EPF): This is a binary variable based on the near-misses presented in Chapter 5. Value “1” indicates failure while “0” indicates no failure.
- Rudder Failure (RUF): This is also a binary variable based on the near-miss cases presented in Chapter 5. Value “1” indicates failure while “0” indicates no failure.
- Mechanic and Electronic Failure (MEF): This is also a binary variable based on the near-miss cases presented in Chapter 5. Value “1” indicates failure while “0” indicates no failure.

6.3.2. Introducing Human Error into the Model

Values of all the explanatory variables stated above are either obtained from the transit, near-miss and accident datasets or through a generation mechanism based on specific probability distribution driven random generations or a simulation model.

However, the human error factor cannot be generated in such a way: vessels which are subject to a human error (such as maneuver fault, navigation fault, etc.) during their passage from the Strait are identified only if they are involved in maritime accidents (since this is recorded in the accident database as the accident cause). However, when there are no accidents, possible human error realizations are not recognized. Moreover, this information cannot be obtained from the simulation-based risk study, since in that study, only conditional human error probabilities are considered (human error probabilities in the mathematical risk model stated in Chapter 2 are associated with specific environmental and vessel related conditions). Since neither exact nor predicted information can be obtained for possible human errors in cases of no accident, one alternative is to exclude this variable from the analysis. However, the accident database and the statistics mentioned in Chapter 5 indicate that human error is actually a substantial accident causing factor between the years 2005-2007: 18 transit vessels out of 37 that were involved in an accident suffered some human error. Since there is no exact information about the human error probabilities, five potential percentage values are considered in order to randomly generate the human error column of the database, which are presented in Table 6.4. Within this context, the results of the original model without human error and five models including different human error levels are generated and compared regarding their power of estimation. Actually, the impact of human error on accidents is also an issue taken into consideration in the literature. (Guedes Soares and Teixeira, 2001) mention that 80% of the accidents are caused by human error. Additionally, a parametric study accomplished on the logistic regression model show that this factor is significant even if more than 70% of the vessels are subjected to a human error. When preparing the dataset, the vessels, which are involved in an accident due to human error cases, are directly inserted into the models with the value “1” and the human error cases without any accidents are randomly generated based on the parameters presented in Table 6.4. The models are run in the statistical package programs SPSS and R and the significant factors and coefficients of the models are presented in Appendix D.

Table 6.4. Presentation of the logistic regression models

<i>Model Name</i>	<i>Human Error Prob.</i>	<i># of Human Error Cases</i>	<i>Variable Name</i>
MODEL	-	-	-
MODEL A	0.25%	415	HEA
MODEL B	0.50%	830	HEB
MODEL C	1%	1660	HEC
MODEL D	3%	4970	HED
MODEL E	5%	8285	HEE

6.3.3. Model Interpretations

The models covering the inputs presented above are run both in SPSS and R package programs. In this sub-section, resulting outputs of the models are presented and discussed. Model A with 0.25% human error probability is taken as the reference for the interpretation. All the candidate variables stated in Sections 6.3.1 and 6.3.2 are entered into the models. Then, the insignificant variables are eliminated from the model, so that the model is left with the significant attributes.

When all the attributes enter the model at the same time (in this section, the model with only the significant variables are presented), which is defined as the “Method Enter” in the SPSS, the outputs generated consist of successive tables presented below. Table 6.5 shows how many cases are selected and whether there are missing cases. The model has 165,680 rows, each representing a vessel with corresponding attributes.

Table 6.5. Case Processing Summary

Unweighted Cases		N	Percent
Selected Cases	Included in Analysis	165,680	100.0
	Missing Cases	0	0.0
	Total	165,680	100.0
Unselected Cases		0	0.0
Total		165,680	100.0

Table 6.6 reflects the encoded values of the categorical dependent variable in the Internal Value column. 0 indicates that the vessel is not involved in an accident and 1 indicates that the vessel is involved in an accident.

Table 6.6. Dependent Variable Encoding

Original Value	Internal Value
0	0
1	1

After the presentation of the data, the model consisting of the constant term only is presented. This step is called “Beginning Block (Step 0)”. Table 6.7 shows all the iterations in order to achieve non-changing (stable) parameter estimates for the Step 0. Estimation terminates at iteration 11 because parameter estimates changed by less than .001. The “-2 Log Likelihood” and “Constant” columns show the change in the estimates and the constant value as the iterations proceed.

Table 6.7. Iteration History of Step 0

Iteration		-2 Log likelihood	Coefficients
			Constant
Step 0	1	42,242.098	-1.999
	2	14,376.621	-3.132
	3	5,387.341	-4.170
	4	2,259.074	-5.171
	5	1,169.622	-6.137
	6	812.162	-7.035
	7	714.072	-7.782
	8	697.102	-8.247
	9	696.107	-8.395
	10	696.102	-8.407
	11	696.102	-8.407

Table 6.8 shows the number of original and predicted cases under the Beginning Block. In this model, since the model predicts the value of the independent variable by only using the constant term, all the cases are predicted as 0, which means that none of the vessels is involved in a maritime accident.

Table 6.8. Classification Table of Step 0

Observed		Predicted		
		ACC		Percentage Correct
		0	1	
Step 0	ACC 0	165,643	0	100.0
	1	37	0	0.0
	Overall Percentage			100.0

Tables 6.9 and 6.10 show which variables are or are not in the equation for the Beginning Block. Since Step 0 is evaluated, the only variable in the equation is the constant term and all the other variables are not in the equation. The significance values in Table 6.10 show whether the variables are worth adding to the equation. If the p-value is smaller than 0.05; adding that variable into the equation will significantly improve the predictive power of the model. Although the significance terms of the season and the visibility status are greater than 0.05 in this model, they are kept in the following steps since they become significant for the resulting model (which is presented in Table 6.17).

Table 6.9. Variables in the Equation for Step 0

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-8.407	0.164	2,614.285	1	0.000	0.000

Table 6.10. Variables Not in the Equation

			Score	df	Sig.
Step 0	Variables	PIL	5.873	1	0.015
		SEA	2.196	1	0.138
		VIS	1.900	1	0.168
		EPF	483.288	1	0.000
		RUF	5,104.406	1	0.000
		MEF	62.952	1	0.000
		HEA	3,469.507	1	0.000
	Overall Statistics		8,955.335	7	0.000

The “Block 1 (Step 1)” signifies the next step in the SPSS. Since “Method Enter” is chosen, Block 1 takes all the variables at once and creates the model. According to Table 6.11, “-2 Log likelihood” column shows the improvement in the estimate as the model iterates. The iterations are terminated when the improvements to the estimate is lower than 0.001. The next columns indicate the estimated coefficients of the model in each iteration.

Table 6.11. Iteration History for Block 1

Iteration		-2 Log likelihood	Coefficients							
			PIL	SEA	VIS	EPF	RUF	MEF	HEA	Constant
Step 1	1	42,229.823	-2.000	-0.001	0.000	0.002	0.099	0.562	0.058	0.170
	2	14,342.429	-3.136	-0.003	0.001	0.007	0.300	1.352	0.182	0.500
	3	5,317.581	-4.179	-0.008	0.003	0.020	0.713	2.299	0.481	1.137
	4	2,143.647	-5.196	-0.022	0.009	0.054	1.336	3.201	1.086	2.024
	5	1,004.651	-6.204	-0.058	0.024	0.144	2.086	4.092	1.985	2.981
	6	598.254	-7.212	-0.152	0.062	0.362	2.896	5.011	2.975	3.941
	7	457.029	-8.216	-0.363	0.147	0.803	3.676	5.924	3.948	4.863
	8	410.948	-9.173	-0.721	0.283	1.422	4.389	6.757	4.838	5.677
	9	399.265	-9.916	-1.092	0.410	1.917	4.921	7.356	5.491	6.253
	10	397.901	-10.253	-1.284	0.462	2.120	5.162	7.617	5.788	6.503
	11	397.872	-10.307	-1.318	0.469	2.150	5.202	7.660	5.838	6.544
	12	397.872	-10.308	-1.319	0.469	2.150	5.203	7.661	5.839	6.545
	13	397.872	-10.308	-1.319	0.469	2.150	5.203	7.661	5.839	6.545

Table 6.12 shows the statistics regarding the model created by the “Method Enter”. As seen in the significance column, the generated model is significant with its associated chi-square value.

Table 6.12. Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	298.230	7	0.000
	Block	298.230	7	0.000
	Model	298.230	7	0.000

Table 6.13 lists the log-likelihood value multiplied by -2. When the log-likelihood statistic has smaller values, this shows a sign of improvement.

Table 6.13. Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	397.872	0.002	0.429

The Hosmer and Lemeshow test divides subjects into ten pieces and then computes a chi-square from observed and expected frequencies. If the significance value is greater than 0.05, null hypothesis, which says that the observed and predicted values have no difference, is rejected. This means that the model’s estimates fit the data at an acceptable level. The resulting outputs are given in Tables 6.14 and 6.15.

Table 6.14. Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	1.873	4	0.759

Table 6.15. Contingency Table for Hosmer and Lemeshow Test

		ACC = 0		ACC = 1		Total
		Observed	Expected	Observed	Expected	
Step 1	1	35,903	35,903.488	1	0.512	35,904
	2	13,574	13,573.690	0	0.310	13,574
	3	26,048	26,047.049	0	0.951	26,048
	4	39,941	39,940.869	2	2.131	39,943
	5	17,378	17,378.456	2	1.544	17,380
	6	32,799	32,799.448	32	31.552	32,831

Table 6.16 shows how well the model predicts the group membership. The overall percentage shows how well the model correctly makes the classification. It is noted that the accident prediction power of the model is very low (because the number of accidents are very low when compared to the total number of transits), since the accident occurrence probabilities are very low and cannot reach the cut value, which is 0.5.

Table 6.16. Classification Table for Block 1

Observed		Predicted		
		ACC		Percentage Correct
		0	1	
Step 1	ACC 0	165,642	1	100.0
	1	34	3	8.1
Overall Percentage				100.0

Table 6.17 shows the coefficients of the model and related statistics. “B” column shows the coefficients; “Wald” column shows whether the coefficients are significantly different from zero. All the variables stated in Table 6.17 are significant since their significance levels are smaller than 0.05.

Table 6.17. Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	PIL	-1.319	0.426	9.565	1	0.002	0.267
	SEA	0.469	0.203	5.343	1	0.021	1.599
	VIS	2.150	0.800	7.224	1	0.007	8.586
	EPF	5.203	0.785	43.916	1	0.000	181.765
	RUF	7.661	0.564	184.435	1	0.000	2,122.978
	MEF	5.839	1.093	28.515	1	0.000	343.461
	HEA	6.545	0.404	262.757	1	0.000	695.538
	Constant	-10.308	0.568	329.824	1	0.000	0.000

The same model is obtained by the statistical package program R, whose outputs are displayed in Appendix D. It is concluded that the obtained model coefficients both in SPSS and R are the same since they use the same algorithm but the comparison criteria to select the best model has a slight difference: SPSS uses -2 Log-likelihood, whereas R is based on Akaike's Information criterion given in (6.7): it simply performs penalty for each parameter entered in the model (presented by k) and the best model is selected based on minimizing this value.

$$AIC = -2 * \ell\ell + 2 * k \quad (6.7)$$

Although the significant factors of the models presented in Appendix D vary, it can be observed that pilot status, all three types of failures and human error are the common factors. Moreover, the explanatory variables season, daytime/nighttime status and visibility condition significance differ from model to model. But these differences do not apply to the coefficient signs and, since the input variables are numeric and ranked in order, the signs give an insight about the impact level of the explanatory variables increasing the accident probability. More specifically, it can be observed that,

- The model coefficient associated with the pilot/no pilot status is negative, indicating that the existence of a pilot in the vessel while transiting the Strait decreases accident occurrence probability.
- The season coefficient is positive, hence, it can be inferred that the accidents are more likely to occur in summer rather than in winter (which is an unexpected result, since meteorological conditions are harsher in winter).

- The daytime/nighttime status coefficient is positive indicating that the accidents are more probable at nighttime.
- The coefficients for the binary variables visibility, three types of failures and human error are positive, which indicate that the occurrence of these events increases the accident potential of the considered vessel.

6.3.4. Verification of the Models

The verification study of the models generated is accomplished with the help of the dataset itself. The consistency of the resulting models is firstly checked by the descriptive statistics of the dataset. The transit accident frequencies calculated from the original explanatory attributes (not generated by simulation or randomly) are presented in Appendix E. From the graphs, it is concluded that the accidents between the years 2005-2007 occurred mostly when the vessels do not benefit from the pilotage services and when it is nighttime. Moreover, the probability of accidents increases when it is summer, which is an unexpected result for the reality but which is compatible with the model outputs. This inconsistency is explained due to the scarcity of the data.

As a second analysis, using the same dataset, the predicted accident probabilities of each vessel can be calculated. Each column i of the dataset gives the i^{th} independent variable input denoted by X_i . To calculate the predicted accident probabilities, firstly, the X_i s are multiplied by their corresponding β_i coefficients of each model and their sum including the intercept is calculated (such as in (6.8)).

$$\text{logit } p = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \quad (6.8)$$

Secondly, the probabilities are calculated with the transformation stated in (6.9)

$$P(Y) = \frac{1}{1 + e^{-(\text{logit } p)}} \quad (6.9)$$

As a result, the predicted value frequencies for each model are tabulated in Appendix F. Appendix F also covers graphical versions of these frequencies separately for accidents and non accident cases. Since these probabilities are very low except some outliers, instead of predicting the accidents directly, sum of expected probabilities are taken into consideration. These values are presented in Table 6.18.

Table 6.18. Sum of expected probabilities for six logistic regression models (verification)

<i>Model Name</i>	<i>Sum of expected probabilities</i>	<i>Sum of expected probabilities/Total number of transits</i>
MODEL	38.60572577	0.000233014
MODEL A	36.99678623	0.000223303
MODEL B	36.99910234	0.000223317
MODEL C	37.00216385	0.000223335
MODEL D	37.00001368	0.000223322
MODEL E	37.0019413	0.000223334

6.3.5. Validation of the Models

The validation study consists of two different analyses regarding two distinct datasets, which are the accident database covering the years 1982-2004 and the combination of the transit, accident and near miss data of the year 2009.

For this purpose, firstly, the full accident database is taken into account and a filtered dataset is obtained in order to calculate the predicted accident probabilities using the corresponding coefficients. With the aim of obtaining an accurate dataset, some eliminations and restrictions are applied to the original data. The important criteria in the formation of this dataset are as follows:

- The accident database includes all vessels playing a role in an accident between the years 1982 and 2007. Hence, the local traffic and unknown type vessels are eliminated from the model.
- The accident data cover the accidents occurred both in anchorage and transit areas. Moreover, there exist cases where the accident places are not known. These unknown cases and anchorage area vessels are eliminated from the data in order to include only the transit accidents.

- Since for some models the daytime/nighttime status is significant, in order to be able to calculate the accident probabilities, the cases, in which the accident times are unknown, are extracted from the validation dataset.
- The GRT, length, age and flag attributes of the vessels are insignificant in all of the models. So, the records with missing values for these attributes are retained.
- In order to have coherence in the frequency and other statistical analyses conducted in Chapter 5, it was assumed that each accident could be represented by at most one reason. But, for some cases, the database introduces the combination of two accident causes (which is especially related to the fact that the accident data is derived from several data sources).
- The cases where the triggering factors of the accident are unknown are excluded from the analysis.
- The 2005-2007 data is deleted from the dataset since they are already used in the logistic model database itself.

After all these updates are applied on the data, a restricted dataset containing 214 vessels is obtained. A portion of this validation data is displayed in Figure 6.2.

As performed in verification section, the sums of the expected accident probabilities for each model are calculated and the results are displayed in Table 6.19. According to the table, the sums of expected accident probabilities for this restricted sample set of vessels already involved in accidents are much smaller than total number of accidents. But, when the ratios “Sum of expected probabilities/Total number of accidents” are taken into consideration, it can be observed these ratios of the sample set are at least 100 times higher than those of the ordinary vessels, as displayed in Table 6.18. In other words, the model is behaving as expected: the collective accident prediction power of the explanatory variables can be regarded as very much significant. Moreover, since the “Sum of expected probabilities/Total number of accidents” ratios in Table 6.19 give the average accident probabilities for the vessels that were already involved in accidents, these values can be

regarded as the cut-off values of an accident. The detailed results of this validation part are given in Appendix G.

SHIP NO	YEAR	PIL	SEA	DAY	VIS	EPF	RUF	MEF	HE
20	1982	1	3	1	0	0	0	0	1
21	1982	0	3	1	0	0	0	0	1
28	1982	0	1	1	1	0	0	0	0
30	1983	1	1	0	1	0	0	0	0
38	1983	0	1	0	0	0	0	0	1
53	1984	0	1	1	1	0	0	0	0
69	1985	0	1	1	1	0	0	0	0
78	1985	0	1	0	1	0	0	0	0
79	1985	0	1	0	1	0	0	0	0
83	1985	0	1	0	0	0	0	0	1
104	1985	1	1	0	1	0	0	0	0
107	1986	0	1	0	1	0	0	0	0
108	1986	1	1	0	1	0	0	0	0
113	1986	0	1	0	1	0	0	0	0
116	1986	1	1	0	1	0	0	0	0

Figure 6.2. An example from the validation dataset

Table 6.19. Sum of expected probabilities for six logistic regression models (validation set 1)

<i>Model Name</i>	<i>Sum of expected probabilities</i>	<i>Sum of expected probabilities/Total number of accidents</i>
MODEL	5.371833839	0.025102027
MODEL A	17.36652051	0.081151965
MODEL B	13.45017909	0.062851304
MODEL C	8.992248787	0.042019854
MODEL D	6.650930437	0.031079114
MODEL E	5.66467478	0.026470443

The second part of the validation analysis is based on the year 2009. For this purpose, as in the logistic regression model formation, three main databases are taken into consideration:

- The transit vessels data which covers all the vessels transiting the Strait with the corresponding attributes.
- The near miss cases categorized as the engine/propeller, rudder and mechanical & electronic failures.
- The accident database of the same year.

The important characteristics of this new set of vessels are as follows:

- A total of 51,421 vessels transit the Strait in 2009.
- There exist seven vessels involved in transit accidents in 2009: five of them are caused by human error and two of them are due to rudder failure.
- 24,951 of the vessels considered have a pilot on board when transiting the Strait.
- Associated with these 51,421 transits, there exist 50 engine/propeller failures, 17 rudder failures and 39 mechanical and electronic failures which are recorded in the transit region.

The six logistic regression models are run for this dataset and the resulting sum of the expected accident probabilities are calculated and displayed in Table 6.20. The detailed results are given in Appendix H. Table 6.20 shows that the numbers of accidents predicted by the model are somewhat overestimated but in general in the same range as the actual number of accidents observed in 2009 (10-12 versus 7). The overestimation is greatest in Model B and least in Model E. This is probably due to the fact that, no accident caused by bad visibility conditions, engine/propeller failure and mechanic and electronic failures is observed in 2009, while these are the significant factors in the models. For instance, when the estimated coefficients corresponding to visibility attribute are taken into account (Appendix D), the greatest coefficient is in the Model A, which gives one of the largest accident probability estimation in Table 6.20. On the other hand, the visibility is insignificant in Models D and E, which is compatible with the accident probability estimations because they are smaller when compared to the remaining models. Moreover, when the ratios “Sum of expected probabilities/Total number of accidents” are taken into consideration, it can be observed that it is compatible with those of the original data displayed in Table 6.18: there is a slight difference in the average accident probability ratios of the years 2005-2007 and those of the year 2009.

Table 6.20. Sum of expected probabilities for six logistic regression models (validation set 2)

<i>Model Name</i>	<i>Sum of expected probabilities</i>	<i>Sum of expected probabilities/Total number of accidents</i>
MODEL	10.57289129	0.000205614
MODEL A	12.03767526	0.0002341
MODEL B	12.48903521	0.000242878
MODEL C	11.15049377	0.000216847
MODEL D	10.44857248	0.000203197
MODEL E	9.909225958	0.000192708

6.3.6. Weakness of the Models

The transit logistic regression models presented in previous sections generally give us logical expected results regarding the significant factors. However, as it can be checked from SPSS output Table 6.16, the cutoff 0.5 is caught only for some outliers. For instance, two of the vessels having high predicted probabilities are those which have two failures at the same time. Hence, the model is very weak in predicting the accident potential of the vessels already involved in accidents. The most important reason for this is the rare occurrence of the transit accidents between the years 2005 and 2007. Since only 37 vessels out of 165,680 are subject to an accident, the accident probabilities calculated by the model are very low. In fact, no exact probabilities can be computed from the data since the visibility conditions and human error cases are not based on the reality. Although adjusted fog occurrences are obtained from the WU data, it is not 100% real visibility values. One other factor that makes the model weaker is that the outcome variable Y is defined by “facing an accident” rather than “causing an accident” because of the missing explanations in the accident dataset. For instance, when a collision or ramming occurs between two vessels because of the human error, both sides are accepted as faulty and flagged as “1” in the human error column. Because of these issues, namely, the scarcity of the accident occurrences, the variables generated from non real-time sources and the missing details about the accident occurrences; the reliability of the model gets lower.

6.4. Anchorage Areas Logistic Regression Model

The accident descriptives show that the anchorage areas accidents are as important as the transit area accidents (due to the large number of accidents occurred in the anchorage areas). This makes the logistic regression for the anchorage areas a worthy endeavor.

However, since the candidate factors triggering the accidents in transit and anchorage areas differ in nature, it is attempted to model and analyze the anchorage areas separately. Although a logistic regression model is constructed for the south anchoring area, the obtained results are so weak that it is removed from the study. The possible reasons for this are discussed below.

One of the anchoring area accident reasons is “adrifting” which is especially caused by strong wind and strong current conditions. Strong wind can be added to the model from the WU data since its measurement station is in Yeşilköy, which is quite near the south anchoring area. Accordingly, visibility status can be derived from the same source. But, a criterion must be defined in order to insert these variables into the model. For the wind variable, this criterion is taken as the “maximum wind speed that the vessel faces during its anchorage”. In parallel, for the visibility status attribute, the criterion is “whether the vessel faces any fog occurrence during its anchorage or not”. However, the waiting times can be calculated for only the northbound vessels from the transit vessels data. Therefore, the model is constructed taking only these vessels into account. But in the phase of inserting the accident vessels into this dataset, it is noticed that many of the vessels facing anchoring area accidents are not included in the data. This leads to the fact that even some of the southbound vessels utilize the anchoring area, after completing their passages. Besides, the anchorage of all southbound vessels is not guaranteed and their anchoring durations are unknown. Therefore, including all southbound vessels into the model is not logical. But when only northbound vessels are considered, there is a significant decrease in accidents, which cause the unreliability of the model.

For the north anchorage area, there are not enough records and inputs to constitute the model. Moreover, some of the records are for the Black Sea, which are not totally related to the Strait of Istanbul. So, this case is also excluded from the study.

7. CONCLUSION

The focus of this study is the statistical analyses of accidents and several factors affecting maritime traffic in the Strait of Istanbul. The common part of all these analyses is that they are based on historical data, which are transit vessels data covering the years 2005-2007, the visibility data obtained from three different sources, accident data between the years 1982-2007 and near miss cases of the years 2005-2007.

The first part of the study is on determining the appropriate distributions and the corresponding parameters of the vessel arrivals, which are mainly based on the priority classes which have been constructed for scheduling purposes. These priority classes are split into several sub-classes so that the sample size for each sub-class becomes appropriate for the goodness-of-fit testing. The distributions are obtained separately for the years 2005, 2006 and 2007 and a trend analysis is conducted to see whether the vessel arrivals change over time. It is concluded that the largest classes (Classes C and D) are more sensitive to the changes in years.

The second part of the study is about the modeling of bad visibility conditions in the Strait, which are key factors increasing the maritime accident risks. For this purpose, the visibility data obtained from the Meteorological Services covering the years 1975-2006, the WU website data between the years 2000-2007 and the Channel closure data of the years 2004-2007 are used. Meteorological Services data capture the seasonality effect and are used to model the fog arrivals in the summer and the transition seasons. Accordingly, the data retrieved from the WU website serve to model the fog start times in winter and the fog durations during the year. Fog start times and durations are determined as an on-off process and approximated by the “Mixtures of Generalized Erlang” distributions, where the on-process is defined as having no fog and the off-process shows a foggy weather. Even though the Channel closure data provide the net effect of visibility conditions on Strait’s traffic efficiency, it is not enough (in terms of the number of cases) to create a model. But it is valuable to compare it with the WU data, so that the WU data can be adjusted for a better reflection of the weather conditions and the fog pattern of the Strait.

The third part of the study is about the statistical analysis of maritime accidents that happened in the Strait between the years 1982 and 2007. This part starts with the descriptive statistics of the attributes of vessels that were involved in accidents in the specified time frame and the attributes of the corresponding accidents. In order to check the statistical relationships between the attributes in question, chi-square test is applied and the possible relationships are interpreted. The same analyses are also performed for the near-miss cases that occurred in the Strait between the years 2005 and 2007. Accordingly, two more analyses are also applied to the accident data. The first one is the discriminant analysis, which is conducted to see whether the vessels can be correctly classified according to the accident causes. The predicting variables used in this analysis are vessel type, age, pilot, accident season, daytime/nighttime status, accident nature and region. As a result, it is derived that 44.6% of the cases are correctly classified. The second analysis is the two-step cluster procedure that is used to group the vessels with respect to five selected variables, which are type, pilot request, nature of the case, region and accident cause. As a result, three different clusters are formed.

In the fourth part of the study, a logistic regression model is implemented to reveal the most significant factors affecting the accident potential of a vessel transiting the Strait of Istanbul. For this purpose, three distinct databases are integrated in order to obtain the logistic regression model. These datasets are transit vessels data, previous transit accidents and near-miss cases covering the years 2005-2007. For the sake of better interpreting the logistic regression model output, the categorical variables of the data are transformed into numerical values. For this purpose, vessel type is changed to a type cardinality variable. Moreover, a combined age & flag cardinality variable is obtained for the flag type and age of the vessel. Hence, the model inputs obtained from the logistic regression data are vessel type cardinality, vessel length, vessel age & flag cardinality, pilot request, whether the vessel fails or not (obtained from the near-miss data) during its passage. Moreover, two more variables, which cannot be retrieved from the real system, are added with the help of different sources: the visibility condition is implemented into the model from the adjusted WU data mentioned in Chapter 4 and the local traffic density is obtained from the recent simulation studies. Moreover, since human error is one of the key factors affecting accidents risks, it is also implemented into the model. However, there is no source that records the human error cases when there is no accident; so, a parametric approach is used

to generate various related scenarios: 0.25%, 0.5%, 1%, 3% and 5% are the rates that represent the human error probability of these vessels. The model without human error variable is also taken into account. As a result, six different models are found and compared in the verification and validation processes. The significant variables are found to be the pilot request, daytime/nighttime status, season, visibility, failure and human error status. The obtained results are also supported by the verification and validation studies. Particularly, the validation process is composed of two parts. In the first case, the accident probability estimations that six models generate are calculated for the selected past accidents derived from the accident dataset of Chapter 5. In the second case, the vessel passages of year 2009 are used and the same procedure is applied. These two cases indicate that the models are quite successful in predicting the accidents.

7. FURTHER STUDIES

This study is an attempt to quantify and statistically analyze several factors influencing the maritime traffic and accident risks associated with the transit vessels in the Strait of Istanbul. Even though this study covers a wide range of statistical analyses, there are areas that need to be improved in the future.

For the arrival process, a simpler model can be obtained by joining some sub-classes. Within this context, the tests applied for the trend analysis can again be performed for the sub-classes of each priority class and the similarities and dissimilarities can be checked by the test results.

As explained in weakness of the logistic regression models, in Chapter 5 and 6, the analyses take into account the vessels being involved in accidents rather than the vessels causing the accidents. Since the database does not clearly contain this information, it would be a good idea to obtain more detailed explanation about the occurrence of the accidents. Hereby, more reliable results can be derived about maritime accident risks based on the real data. Moreover, the analysis can be furnished with the most recent data (i.e. year 2008 and 2009) to increase the number of cases and the consistency of the data. On the other hand, a detailed human error study can be accomplished. From the logistic regression study and according to the accident statistics, it is inferred that human error is one of the key factors threatening safe navigation in the Strait. Since it is hard to collect data about human error cases (because the vessels will not report all the errors by themselves), an expert opinion-based study can be conducted.

From the statistics, it can also be observed that the accidents in the anchoring areas, especially the south anchoring area, are numerous and diverse. Hence, an analysis of this region may be a worthwhile study. However, there is not enough information regarding this region. For instance, as noted in the anchoring area logistic regression study, the vessel profile waiting in the south anchorage area is unknown: not only the northbound vessels but also other vessels wait at this region and increase the accidents risks. In addition, their waiting times are unknown. Hence, to generate the model of the anchorage area, missing

information can be obtained by the experts and again a better record keeping can be proposed.

APPENDIX A: ARRIVAL PROCESS PROBABILITY DISTRIBUTIONS

Table A.1. Fitted distributions to vessel clusters in arrival processes: Class A

CLASS A	Direction	Daytime	Season	Expression	Square Error	Chi-Square P-Value	Number of Data Points
2005	S-N	Night	Spring	$3 + \text{LOGN}(470, 1.65\text{e}+003)$	0.008287	< 0.005	180
	S-N	Night	Winter	$1 + \text{WEIB}(252, 0.686)$	0.001798	< 0.005	140
	S-N	Night	Summer	$1 + \text{LOGN}(585, 2.09\text{e}+003)$	0.008181	< 0.005	456
	S-N	Day	Spring	$6 + 2.93\text{e}+003 * \text{BETA}(0.454, 1.37)$	0.019802	< 0.005	97
	S-N	Day	Winter	$2 + \text{EXPO}(729)$	0.002873	0.612	100
	S-N	Day	Summer	$1 + 2.39\text{e}+003 * \text{BETA}(0.625, 2.01)$	0.036765	< 0.005	133
	N-S	Night	Spring	$5 + \text{EXPO}(453)$	0.001464	0.621	139
	N-S	Night	Winter	$\text{EXPO}(502)$	0.008024	0.206	117
	N-S	Night	Summer	$1 + \text{GAMM}(470, 0.953)$	0.001757	0.661	314
	N-S	Day	Spring	$2 + 2.02\text{e}+003 * \text{BETA}(0.84, 2.64)$	0.005231	0.191	140
	N-S	Day	Winter	$\text{EXPO}(556)$	0.004965	0.407	118
	N-S	Day	Summer	$1 + \text{WEIB}(497, 1.14)$	0.003191	0.3	257
2006	S-N	Night	Spring	$2 + 2.73\text{e}+003 * \text{BETA}(0.362, 2.42)$	0.004923	0.0161	181
	S-N	Night	Winter	$1 + 5.87\text{e}+003 * \text{BETA}(0.139, 2.28)$	0.007012	< 0.005	146
	S-N	Night	Summer	$1 + \text{LOGN}(513, 1.7\text{e}+003)$	0.005184	< 0.005	445
	S-N	Day	Spring	$2.89\text{e}+003 * \text{BETA}(0.494, 1.29)$	0.020096	0.00751	100
	S-N	Day	Winter	$3 + \text{WEIB}(701, 0.725)$	0.004010	< 0.005	90
	S-N	Day	Summer	$1 + 2.24\text{e}+003 * \text{BETA}(0.682, 2.3)$	0.009020	0.00544	194
	N-S	Night	Spring	$5 + 1.94\text{e}+003 * \text{BETA}(0.704, 2.47)$	0.001918	0.727	150
	N-S	Night	Winter	$1 + \text{WEIB}(499, 0.907)$	0.003335	0.0281	135
	N-S	Night	Summer	$\text{EXPO}(429)$	0.003313	0.326	349
	N-S	Day	Spring	$4 + \text{EXPO}(474)$	0.004687	0.0844	138
	N-S	Day	Winter	$1 + \text{WEIB}(534, 1.03)$	0.005403	0.151	120
	N-S	Day	Summer	$\text{WEIB}(446, 1.19)$	0.004164	0.131	270
2007	S-N	Night	Spring	$1 + 2.66\text{e}+003 * \text{BETA}(0.217, 1.38)$	0.008383	< 0.005	169
	S-N	Night	Winter	$3.79\text{e}+003 * \text{BETA}(0.133, 2.11)$	0.001165	0.0216	173
	S-N	Night	Summer	$\text{LOGN}(540, 1.89\text{e}+003)$	0.012964	< 0.005	458
	S-N	Day	Spring	$3.96\text{e}+003 * \text{BETA}(0.446, 1.7)$	0.011776	< 0.005	107
	S-N	Day	Winter	$5 + \text{WEIB}(579, 0.653)$	0.005992	0.0299	120
	S-N	Day	Summer	$1 + 2.41\text{e}+003 * \text{BETA}(0.561, 1.43)$	0.023363	< 0.005	127
	N-S	Night	Spring	$1 + \text{WEIB}(428, 1.06)$	0.005199	0.0997	153
	N-S	Night	Winter	$5 + \text{LOGN}(703, 1.8\text{e}+003)$	0.006896	0.0308	142
	N-S	Night	Summer	$\text{EXPO}(453)$	0.001117	0.408	346
	N-S	Day	Spring	$2 + \text{GAMM}(423, 1.22)$	0.001590	0.553	129
	N-S	Day	Winter	$1 + 2.06\text{e}+003 * \text{BETA}(0.676, 2.21)$	0.002810	> 0.75	132
	N-S	Day	Summer	$2 + \text{WEIB}(501, 1.12)$	0.002220	0.336	224

Table A.2. Fitted distributions to vessel clusters in arrival processes: Class B

CLASS B	Direction	Daytime	Expression	Square Error	Chi-Square P-Value	Number of Data Points
2005	S-N	Night	$1 + 4.44\text{e}+003 * \text{BETA}(0.642, 5.17)$	0.009256	< 0.005	599
	S-N	Day	$\text{LOGN}(389, 1.06\text{e}+003)$	0.000984	< 0.005	748
	N-S	Night	$1 + \text{EXPO}(421)$	0.001282	0.307	637
	N-S	Day	$\text{WEIB}(342, 0.91)$	0.000449	0.692	710
2006	S-N	Night	$1 + 2.81\text{e}+003 * \text{BETA}(0.643, 2.92)$	0.001609	< 0.005	571
	S-N	Day	$1 + \text{WEIB}(264, 0.75)$	0.001395	< 0.005	732
	N-S	Night	$1 + \text{WEIB}(421, 1.01)$	0.00107	0.222	603
	N-S	Day	$\text{EXPO}(410)$	0.000859	> 0.75	665
2007	S-N	Night	$2 + 2.34\text{e}+003 * \text{BETA}(0.72, 3.14)$	0.002966	< 0.005	602
	S-N	Day	$1 + \text{WEIB}(305, 0.838)$	0.001756	0.0407	798
	N-S	Night	$1 + \text{EXPO}(421)$	0.000958	0.623	644
	N-S	Day	$1 + \text{WEIB}(264, 0.75)$	0.001234	0.386	716

Table A.3. Fitted distributions to vessel clusters in arrival processes: Class C

CLASS C	Direction	Daytime	Expression	Square Error	Chi-Square P-Value	Number of Data Points
2005	S-N	Night	EXPO(124)	0.000437	0.557	2139
	S-N	Day	-0.001 + GAMM(120, 0.954)	0.000422	0.0653	2270
	N-S	Night	WEIB(128, 0.944)	0.000354	0.0999	2103
	N-S	Day	-0.001 + GAMM(130, 0.942)	0.000273	0.357	2128
2006	S-N	Night	GAMM(127, 0.968)	0.000399	0.268	2160
	S-N	Day	GAMM(116, 0.951)	0.000286	0.633	2367
	N-S	Night	EXPO(117)	0.000379	0.219	2191
	N-S	Day	EXPO(119)	0.000291	0.282	2253
2007	S-N	Night	EXPO(128)	0.000461	0.314	2046
	S-N	Day	GAMM(123, 0.981)	0.000288	0.698	2183
	N-S	Night	WEIB(131, 0.979)	0.000208	>0.75	1996
	N-S	Day	GAMM(124, 0.99)	0.00037	0.653	2167

Table A.4. Fitted distributions to vessel clusters in arrival processes: Class E

CLASS E	Direction	Daytime	Season	Expression	Square Error	Chi-Square P-Value	Number of Data Points
2005	S-N	Night	Spring	1 + WEIB(320, 0.93)	0.002811	0.127	189
	S-N	Night	Winter	EXPO(438)	0.002184	0.25	146
	S-N	Night	Summer	1 + WEIB(343, 1.02)	0.000869	> 0.75	372
	S-N	Day	Spring	GAMM(488, 0.753)	0.004959	0.115	189
	S-N	Day	Winter	1 + WEIB(351, 0.816)	0.000903	0.107	162
	S-N	Day	Summer	1 + EXPO(348)	0.001136	0.158	392
	N-S	Night	Spring	2 + 1.87e+003 * BETA(0.738, 3.64)	0.002750	0.063	176
	N-S	Night	Winter	WEIB(419, 0.955)	0.001564	0.0908	146
	N-S	Night	Summer	WEIB(376, 1.08)	0.002127	0.0109	369
	N-S	Day	Spring	1 + GAMM(414, 0.879)	0.004854	0.0752	206
	N-S	Day	Winter	1 + EXPO(398)	0.001889	0.578	166
	N-S	Day	Summer	WEIB(317, 0.972)	0.000444	0.499	404
2006	S-N	Night	Spring	1 + EXPO(270)	0.001678	0.321	189
	S-N	Night	Winter	3 + WEIB(409, 1)	0.001618	0.453	139
	S-N	Night	Summer	WEIB(369, 1)	0.000538	0.675	348
	S-N	Day	Spring	WEIB(346, 0.983)	0.002330	0.343	230
	S-N	Day	Winter	2 + WEIB(379, 0.813)	0.004401	0.156	174
	S-N	Day	Summer	2 + WEIB(319, 0.936)	0.001617	0.36	412
	N-S	Night	Spring	1 + WEIB(351, 0.933)	0.000901	> 0.75	193
	N-S	Night	Winter	2 + EXPO(413)	0.005121	0.0658	157
	N-S	Night	Summer	1.67e+003 * BETA(0.788, 2.85)	0.001070	0.538	386
	N-S	Day	Spring	1 + 1.97e+003 * BETA(0.577, 3.19)	0.002012	0.19	202
	N-S	Day	Winter	1 + WEIB(386, 0.86)	0.005009	0.0657	155
	N-S	Day	Summer	2 + GAMM(327, 0.996)	0.001254	0.692	410
2007	S-N	Night	Spring	2 + GAMM(374, 1.15)	0.004026	0.0873	138
	S-N	Night	Winter	2 + EXPO(455)	0.000819	0.585	140
	S-N	Night	Summer	1 + GAMM(364, 0.99)	0.001694	0.144	334
	S-N	Day	Spring	2.36e+003 * BETA(0.638, 3.2)	0.002116	0.234	183
	S-N	Day	Winter	1 + WEIB(358, 0.918)	0.001676	0.259	169
	S-N	Day	Summer	1 + WEIB(353, 0.977)	0.000868	0.712	402
	N-S	Night	Spring	2.64e+003 * BETA(0.662, 2.92)	0.004949	0.00661	145
	N-S	Night	Winter	1 + GAMM(390, 1.16)	0.010285	0.0462	132
	N-S	Night	Summer	GAMM(398, 0.968)	0.001770	0.156	363
	N-S	Day	Spring	1 + WEIB(353, 0.863)	0.002207	0.249	183
	N-S	Day	Winter	1 + WEIB(348, 0.902)	0.006410	0.013	188
	N-S	Day	Summer	GAMM(315, 1.11)	0.001724	0.352	358

Table A.5. Fitted distributions to vessel clusters in arrival processes: Class D

CLASS D	Direct / Indirect				Expression	Square Error	Chi-Square		Number of Data Points
	Direction	Status	Daytime	Season			P-Value		
2005	S-N	Indirect	Night	Spring	-0.001 + WEIB(75.2, 0.871)	0.000760	0.248		790
	S-N	Indirect	Night	Winter	WEIB(88.7, 0.887)	0.002011	0.0389		663
	S-N	Indirect	Night	Summer	EXPO(76.7)	0.000657	0.319		1609
	S-N	Indirect	Day	Spring	WEIB(61.5, 0.939)	0.000708	< 0.005		986
	S-N	Indirect	Day	Winter	WEIB(67, 0.771)	0.001111	0.0162		856
	S-N	Indirect	Day	Summer	-0.001 + GAMM(77.2, 0.872)	0.001067	< 0.005		2105
	S-N	Direct	Night	Spring	GAMM(49.8, 0.885)	0.002174	0.00752		1483
	S-N	Direct	Night	Winter	-0.001 + WEIB(47.6, 0.803)	0.000421	0.0869		1224
	S-N	Direct	Night	Summer	-0.001 + GAMM(45.1, 0.972)	0.000385	0.0944		2955
	S-N	Direct	Day	Spring	GAMM(51.1, 0.886)	0.000582	0.042		1455
	S-N	Direct	Day	Winter	WEIB(42.7, 0.81)	0.000564	0.0414		1363
	S-N	Direct	Day	Summer	GAMM(51.6, 0.921)	0.000433	0.106		2846
	N-S	Indirect	Night	Spring	WEIB(77.2, 0.939)	0.001070	0.164		833
	N-S	Indirect	Night	Winter	WEIB(80.4, 0.865)	0.000326	0.472		704
	N-S	Indirect	Night	Summer	GAMM(85.7, 0.939)	0.000504	0.519		1671
	N-S	Indirect	Day	Spring	GAMM(75, 0.906)	0.001033	0.141		945
	N-S	Indirect	Day	Winter	WEIB(71.9, 0.82)	0.000992	0.000992		885
	N-S	Indirect	Day	Summer	-0.001 + WEIB(63.2, 0.965)	0.000243	0.555		1966
	N-S	Direct	Night	Spring	-0.001 + GAMM(50.7, 0.882)	0.000993	0.275		1475
	N-S	Direct	Night	Winter	GAMM(54.8, 0.868)	0.000684	0.18		1357
	N-S	Direct	Night	Summer	-0.001 + WEIB(42.1, 0.983)	0.000164	0.578		2996
	N-S	Direct	Day	Spring	-0.001 + GAMM(50.5, 0.864)	0.000462	0.184		1492
	N-S	Direct	Day	Winter	-0.001 + WEIB(47.7, 0.925)	0.000319	< 0.005		1268
	N-S	Direct	Day	Summer	GAMM(47.2, 0.955)	0.000209	0.338		2944
2006	S-N	Indirect	Night	Spring	GAMM(58.7, 0.964)	0.000757	0.0111		1120
	S-N	Indirect	Night	Winter	LOGN(87.7, 211)	0.001018	< 0.005		873
	S-N	Indirect	Night	Summer	GAMM(58, 0.938)	0.000397	0.00641		2374
	S-N	Indirect	Day	Spring	GAMM(64.3, 0.859)	0.000810	0.222		1232
	S-N	Indirect	Day	Winter	-0.001 + WEIB(62.5, 0.75)	0.001580	0.00539		1027
	S-N	Indirect	Day	Summer	-0.001 + GAMM(64.2, 0.872)	0.000470	0.0379		2421
	S-N	Direct	Night	Spring	WEIB(63.8, 0.956)	0.001116	0.0247		981
	S-N	Direct	Night	Winter	WEIB(77.8, 0.827)	0.003542	< 0.005		846
	S-N	Direct	Night	Summer	GAMM(64.2, 0.952)	0.000337	0.226		2164
	S-N	Direct	Day	Spring	WEIB(52.8, 0.928)	0.000845	0.024		1273
	S-N	Direct	Day	Winter	-0.001 + WEIB(52.1, 0.793)	0.001050	0.0201		1177
	S-N	Direct	Day	Summer	GAMM(58, 0.893)	0.000414	0.182		2554
	N-S	Indirect	Night	Spring	EXPO(67.4)	0.000578	0.524		953
	N-S	Indirect	Night	Winter	WEIB(86, 0.904)	0.002258	0.00612		790
	N-S	Indirect	Night	Summer	GAMM(64.2, 0.938)	0.000327	0.451		2170
	N-S	Indirect	Day	Spring	WEIB(64.9, 0.953)	0.000923	0.0733		1042
	N-S	Indirect	Day	Winter	EXPO(79.5)	0.002644	0.00851		845
	N-S	Indirect	Day	Summer	GAMM(67.1, 0.957)	0.000438	0.399		2088
	N-S	Direct	Night	Spring	GAMM(56.3, 0.945)	0.000387	0.275		1230
	N-S	Direct	Night	Winter	GAMM(67.4, 0.928)	0.000727	0.243		1011
	N-S	Direct	Night	Summer	EXPO(52.7)	0.000283	> 0.75		2513
	N-S	Direct	Day	Spring	-0.001 + GAMM(52.7, 0.909)	0.000739	0.295		1378
	N-S	Direct	Day	Winter	-0.001 + GAMM(59.7, 0.887)	0.001575	0.0276		1264
	N-S	Direct	Day	Summer	EXPO(47)	0.000565	0.174		2815
2007	S-N	Indirect	Night	Spring	GAMM(60, 0.978)	0.000727	0.102		1077
	S-N	Indirect	Night	Winter	EXPO(64.2)	0.000009	< 0.005		1078
	S-N	Indirect	Night	Summer	-0.001 + GAMM(54.6, 0.929)	0.000433	0.27		2532
	S-N	Indirect	Day	Spring	-0.001 + WEIB(55, 0.942)	0.000482	0.354		1193
	S-N	Indirect	Day	Winter	WEIB(56.1, 0.881)	0.002246	0.0448		1237
	S-N	Indirect	Day	Summer	-0.001 + GAMM(58.8, 0.92)	0.000812	0.209		2528
	S-N	Direct	Night	Spring	GAMM(63.1, 0.987)	0.000433	0.0706		1055
	S-N	Direct	Night	Winter	WEIB(64.6, 0.861)	0.000400	0.177		958
	S-N	Direct	Night	Summer	EXPO(57.7)	0.000197	0.358		2266
	S-N	Direct	Day	Spring	WEIB(47.5, 0.966)	0.000526	0.419		1328
	S-N	Direct	Day	Winter	GAMM(58.8, 0.919)	0.000750	0.183		1217
	S-N	Direct	Day	Summer	WEIB(48.8, 0.968)	0.000110	> 0.75		2684
	N-S	Indirect	Night	Spring	GAMM(70.5, 0.926)	0.000917	0.315		985
	N-S	Indirect	Night	Winter	WEIB(67.6, 0.928)	0.001255	0.102		980
	N-S	Indirect	Night	Summer	GAMM(66.1, 0.956)	0.000589	0.202		2108
	N-S	Indirect	Day	Spring	-0.001 + EXPO(62)	0.000898	0.039		1080
	N-S	Indirect	Day	Winter	-0.001 + WEIB(59.5, 0.903)	0.000599	0.148		1063
	N-S	Indirect	Day	Summer	-0.001 + GAMM(58.7, 0.96)	0.000264	> 0.75		2335
	N-S	Direct	Night	Spring	GAMM(52.3, 0.961)	0.000718	0.22		1300
	N-S	Direct	Night	Winter	WEIB(50, 0.916)	0.000475	0.567		1211
	N-S	Direct	Night	Summer	EXPO(51.5)	0.000177	0.687		2581
	N-S	Direct	Day	Spring	-0.001 + WEIB(44.3, 0.994)	0.000285	0.281		1413
	N-S	Direct	Day	Winter	GAMM(54.7, 0.921)	0.000175	> 0.75		1300
	N-S	Direct	Day	Summer	EXPO(45.3)	0.000234	> 0.75		2914

Table A.6. Fitted distributions to vessel clusters in arrival processes: Class P

CLASS P	Direction	Season	Expression	Square Error	Chi-Square P-Value	Number of Data Points
2005	S-N	Spring	$8.8e+003 * \text{BETA}(0.383, 2.59)$	0.004157	0.0411	158
	S-N	Winter	$\text{EXPO}(1.1e+003)$	0.004671	0.17	109
	S-N	Summer	$4.33e+003 * \text{BETA}(0.312, 1.72)$	0.002971	0.0713	398
	N-S	Spring	$\text{EXPO}(818)$	0.003518	0.218	161
	N-S	Winter	$\text{LOGN}(1.67e+003, 4.73e+003)$	0.003255	< 0.005	100
	N-S	Summer	$4.57e+003 * \text{BETA}(0.626, 3.45)$	0.005059	< 0.005	403
2006	S-N	Spring	$4.29e+003 * \text{BETA}(0.2, 1.15)$	0.010298	0.178	203
	S-N	Winter	$\text{WEIB}(507, 0.539)$	0.000178	< 0.005	139
	S-N	Summer	$\text{LOGN}(825, 3.8e+003)$	0.001276	0.0345	496
	N-S	Spring	$1 + \text{GAMM}(979, 0.668)$	0.005250	0.148	199
	N-S	Winter	$\text{EXPO}(961)$	0.002644	0.283	132
	N-S	Summer	$\text{WEIB}(457, 0.794)$	0.001646	0.0968	510
2007	S-N	Spring	$\text{LOGN}(1.23e+003, 8.44e+003)$	0.001544	0.0134	208
	S-N	Winter	$\text{EXPO}(786)$	0.001488	< 0.005	157
	S-N	Summer	$5.23e+003 * \text{BETA}(0.387, 3.48)$	0.002017	0.108	509
	N-S	Spring	$1 + \text{WEIB}(543, 0.742)$	0.002161	0.242	201
	N-S	Winter	$5.59e+003 * \text{BETA}(0.483, 2.42)$	0.004748	0.00665	160
	N-S	Summer	$\text{WEIB}(433, 0.732)$	0.003701	< 0.005	510

APPENDIX B: TREND ANALYSIS RESULTS

Table B.1. t, F and Mann-Whitney tests p-values for trend analysis: Class A

CLASS A			T-TEST			F-TEST			M-W TEST		
Direction	Daytime	Season	2005-2006	2005-2007	2006-2007	2005-2006	2005-2007	2006-2007	2005-2006	2005-2007	2006-2007
S-N	Night	Spring	0.580	0.673	0.925	0.049	0.631	0.122	0.201	0.638	0.092
S-N	Night	Winter	0.897	0.099	0.118	0.962	0.019	0.012	0.441	0.009	0.061
S-N	Night	Summer	0.183	0.633	0.411	0.010	0.916	0.010	0.567	0.534	0.906
S-N	Day	Spring	0.633	0.918	0.716	0.611	0.531	0.887	0.481	0.982	0.416
S-N	Day	Winter	0.362	0.706	0.568	0.072	0.101	0.643	0.723	0.664	0.392
S-N	Day	Summer	0.307	0.272	0.041	0.054	0.157	0.002	0.355	0.584	0.129
N-S	Night	Spring	0.664	0.438	0.706	0.622	0.290	0.487	0.849	0.627	0.858
N-S	Night	Winter	0.757	0.816	0.556	0.346	0.717	0.436	0.860	0.999	0.923
N-S	Night	Summer	0.538	0.896	0.446	0.685	0.619	0.170	0.710	0.967	0.707
N-S	Day	Spring	0.838	0.615	0.508	0.264	0.330	0.998	0.437	0.929	0.377
N-S	Day	Winter	0.687	0.248	0.447	0.333	0.155	0.735	0.817	0.365	0.450
N-S	Day	Summer	0.096	0.838	0.073	0.041	0.633	0.013	0.169	0.926	0.168

Table B.2. t, F and Mann-Whitney tests p-values for trend analysis: Class B

CLASS B			T-TEST			F-TEST			M-W TEST		
Direction	Daytime		2005-2006	2005-2007	2006-2007	2005-2006	2005-2007	2006-2007	2005-2006	2005-2007	2006-2007
S-N	Night		0.326	0.126	0.010	0.553	0.000	0.000	0.168	0.807	0.186
S-N	Day		0.852	0.556	0.696	0.800	0.203	0.312	0.473	0.003	0.035
N-S	Night		0.817	0.368	0.273	0.896	0.574	0.514	0.949	0.278	0.245
N-S	Night		0.016	0.426	0.085	0.252	0.518	0.064	0.001	0.076	0.102

Table B.3. t, F and Mann-Whitney tests p-values for trend analysis: Class C

CLASS C			T-TEST			F-TEST			M-W TEST		
Direction	Daytime		2005-2006	2005-2007	2006-2007	2005-2006	2005-2007	2006-2007	2005-2006	2005-2007	2006-2007
S-N	Night		0.935	0.228	0.204	0.511	0.482	0.960	0.571	0.156	0.046
S-N	Day		0.211	0.073	0.003	0.447	0.017	0.002	0.117	0.196	0.004
N-S	Night		0.036	0.130	0.000	0.004	0.264	0.000	0.116	0.117	0.002
N-S	Night		0.446	0.988	0.452	0.176	0.684	0.350	0.839	0.765	0.619

Table B.4. t, F and Mann-Whitney tests p-values for trend analysis: Class D

CLASS D Direct / Indirect				T-TEST			F-TEST			M-W TEST		
Direction	Status	Daytime	Season	2005-2006	2005-2007	2006-2007	2005-2006	2005-2007	2006-2007	2005-2006	2005-2007	2006-2007
S-N	Indirect	Night	Spring	0.000	0.000	0.430	0.000	0.000	0.407	0.000	0.000	0.283
S-N	Indirect	Night	Winter	0.000	0.000	0.486	0.000	0.001	0.535	0.000	0.000	0.130
S-N	Indirect	Night	Summer	0.000	0.000	0.033	0.000	0.000	0.022	0.000	0.000	0.095
S-N	Indirect	Day	Spring	0.000	0.000	0.578	0.000	0.000	0.866	0.000	0.002	0.223
S-N	Indirect	Day	Winter	0.137	0.000	0.001	0.286	0.000	0.000	0.006	0.000	0.130
S-N	Indirect	Day	Summer	0.000	0.000	0.298	0.000	0.000	0.071	0.000	0.000	0.920
S-N	Direct	Night	Spring	0.000	0.000	0.217	0.000	0.000	0.107	0.000	0.000	0.373
S-N	Direct	Night	Winter	0.000	0.000	0.169	0.000	0.000	0.060	0.000	0.000	0.401
S-N	Direct	Night	Summer	0.000	0.000	0.077	0.000	0.000	0.003	0.000	0.000	0.664
S-N	Direct	Day	Spring	0.001	0.057	0.146	0.013	0.158	0.233	0.000	0.011	0.129
S-N	Direct	Day	Winter	0.037	0.107	0.498	0.078	0.757	0.083	0.004	0.000	0.462
S-N	Direct	Day	Summer	0.006	0.077	0.265	0.000	0.071	0.001	0.040	0.004	0.485
N-S	Indirect	Night	Spring	0.001	0.000	0.532	0.000	0.000	0.733	0.051	0.100	0.512
N-S	Indirect	Night	Winter	0.060	0.000	0.001	0.007	0.000	0.000	0.275	0.000	0.006
N-S	Indirect	Night	Summer	0.000	0.000	0.141	0.000	0.000	0.175	0.000	0.000	0.079
N-S	Indirect	Day	Spring	0.286	0.086	0.425	0.079	0.028	0.434	0.901	0.154	0.179
N-S	Indirect	Day	Winter	0.340	0.003	0.001	0.452	0.000	0.000	0.226	0.018	0.000
N-S	Indirect	Day	Summer	0.379	0.000	0.000	0.250	0.000	0.000	0.622	0.000	0.002
N-S	Direct	Night	Spring	0.000	0.005	0.181	0.000	0.022	0.096	0.000	0.002	0.304
N-S	Direct	Night	Winter	0.000	0.007	0.001	0.000	0.017	0.001	0.000	0.006	0.001
N-S	Direct	Night	Summer	0.000	0.000	0.431	0.000	0.000	0.136	0.000	0.000	0.909
N-S	Direct	Day	Spring	0.027	0.132	0.489	0.192	0.976	0.198	0.003	0.003	0.899
N-S	Direct	Day	Winter	0.405	0.854	0.305	0.370	0.348	0.078	0.311	0.499	0.718
N-S	Direct	Day	Summer	0.132	0.854	0.171	0.832	0.572	0.412	0.008	0.294	0.104

Table B.5. t, F and Mann-Whitney tests p-values for trend analysis: Class E

CLASS E			T-TEST			F-TEST			M-W TEST		
Direction	Daytime	Season	2005-2006	2005-2007	2006-2007	2005-2006	2005-2007	2006-2007	2005-2006	2005-2007	2006-2007
S-N	Night	Spring	0.052	0.013	0.000	0.004	0.126	0.000	0.267	0.003	0.000
S-N	Night	Winter	0.578	0.734	0.383	0.273	0.824	0.229	0.908	0.725	0.664
S-N	Night	Summer	0.312	0.471	0.813	0.156	0.154	0.939	0.514	0.995	0.537
S-N	Day	Spring	0.610	0.538	0.247	0.262	0.255	0.021	0.210	0.566	0.644
S-N	Day	Winter	0.585	0.662	0.277	0.414	0.274	0.024	0.676	0.653	0.994
S-N	Day	Summer	0.452	0.751	0.276	0.622	0.973	0.588	0.343	0.612	0.145
N-S	Night	Spring	0.191	0.004	0.106	0.026	0.005	0.528	0.619	0.008	0.030
N-S	Night	Winter	0.805	0.590	0.438	0.485	0.930	0.478	0.934	0.454	0.459
N-S	Night	Summer	0.244	0.874	0.386	0.662	0.226	0.106	0.178	0.197	0.911
N-S	Day	Spring	0.065	0.467	0.015	0.057	0.119	0.001	0.080	0.816	0.068
N-S	Day	Winter	0.740	0.581	0.367	0.511	0.632	0.211	0.863	0.710	0.629
N-S	Day	Summer	0.772	0.130	0.225	0.882	0.130	0.189	0.699	0.149	0.268

Table B.6. t, F and Mann-Whitney tests p-values for trend analysis: Class P

CLASS P		T-TEST			F-TEST			M-W TEST		
Direction	Season	2005-2006	2005-2007	2006-2007	2005-2006	2005-2007	2006-2007	2005-2006	2005-2007	2006-2007
S-N	Spring	0.078	0.066	0.963	0.012	0.002	0.565	0.055	0.133	0.669
S-N	Winter	0.432	0.114	0.556	0.675	0.079	0.284	0.134	0.156	0.789
S-N	Summer	0.025	0.010	0.803	0.001	0.000	0.248	0.051	0.148	0.547
N-S	Spring	0.092	0.067	0.784	0.231	0.805	0.099	0.052	0.000	0.017
N-S	Winter	0.221	0.038	0.325	0.163	0.146	0.904	0.114	0.000	0.027
N-S	Summer	0.003	0.005	0.950	0.048	0.832	0.064	0.000	0.000	0.097

APPENDIX C: TONNAGE-LENGTH REGRESSION

Table C.1. Passenger vessels regression model

<i>Model Summary</i>	<i>R</i>	<i>R Square</i>	<i>Adjusted R Square</i>	<i>Std. Error of the Estimate</i>		
	0.906	0.821	0.819	25.8271		
<i>ANOVA</i>	<i>Sum of Squares</i>		<i>df</i>	<i>Mean Square</i>	<i>F</i>	<i>Sig.</i>
	Regression	430363.049	1	430363.049	645.183	0.000
	Residual	94052.684	141	667.040		
	Total	524415.733	142			
<i>Coefficients</i>	<i>Unstandardized Coefficients</i>				<i>Sig.</i>	
	<i>B</i>	<i>Std. Error</i>	<i>t</i>	<i>Std. Error</i>		
	(Constant)	84.969	2.815	30.187	0.000	
	grt	0.003	0.000	25.400	0.000	

Table C.2. HazMat/LNG/LPG vessels regression model

<i>Model Summary</i>	<i>R</i>	<i>R Square</i>	<i>Adjusted R Square</i>	<i>Std. Error of the Estimate</i>		
	0.943	0.889	0.889	17.94664		
<i>ANOVA</i>	<i>Sum of Squares</i>		<i>df</i>	<i>Mean Square</i>	<i>F</i>	<i>Sig.</i>
	Regression	1658327.703	1	1658327.703	5148.779	0.000
	Residual	206454.378	641	322.082		
	Total	1864782.081	642			
<i>Coefficients</i>	<i>Unstandardized Coefficients</i>				<i>Sig.</i>	
	<i>B</i>	<i>Std. Error</i>	<i>t</i>	<i>Std. Error</i>		
	(Constant)	95.227	0.985	96.653	0.000	
	grt	0.004	0.000	71.755	0.000	

Table C.3. Tanker vessels regression model

<i>Model Summary</i>	<i>R</i>	<i>R Square</i>	<i>Adjusted R Square</i>	<i>Std. Error of the Estimate</i>		
	0.941	0.885	0.885	20.48991		
<i>ANOVA</i>	<i>Sum of Squares</i>		<i>df</i>	<i>Mean Square</i>	<i>F</i>	<i>Sig.</i>
	Regression	5096554.261	1	5096554.261	12139.380	0.000
	Residual	660822.567	1574	419.836		
	Total	5757376.828	1575			
<i>Coefficients</i>	<i>Unstandardized Coefficients</i>				<i>Sig.</i>	
	<i>B</i>	<i>Std. Error</i>	<i>t</i>	<i>Std. Error</i>		
	(Constant)	107.653	0.734	146.759	0.000	
	grt	0.002	0.000	110.179	0.000	

Table C.4. General cargo vessels regression model

<i>Model Summary</i>	<i>R</i>	<i>R Square</i>	<i>Adjusted R Square</i>	<i>Std. Error of the Estimate</i>		
	0.9	0.81	0.81	25.5654		
<i>ANOVA</i>	<i>Sum of Squares</i>		<i>df</i>	<i>Mean Square</i>	<i>F</i>	<i>Sig.</i>
	Regression	20739887.17	1	20739887.168	31732.213	0.000
	Residual	4859448.655	7435	653.591		
	Total	25599335.82	7436			
<i>Coefficients</i>	<i>Unstandardized Coefficients</i>			<i>Sig.</i>		
	<i>B</i>	<i>Std. Error</i>	<i>t</i>	<i>Std. Error</i>		
	(Constant)	96.466	0.389	248.065	0.000	
	grt	0.003	0.000	178.135	0.000	

APPENDIX D: LOGISTIC REGRESSION MODEL OUTPUTS

Table D.1. Logistic regression output: Model

Model coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-9.6841	0.5582	-17.350	< 2e-16
PIL	-1.1135	0.3960	-2.812	0.00492
SEA	0.3906	0.1954	1.999	0.04557
DAY	0.7788	0.3661	2.128	0.03337
VIS	1.5747	0.7590	2.075	0.03801
EPF	4.6434	0.6638	6.995	2.66e-12
RUF	6.8756	0.4766	14.426	< 2e-16
MEF	4.8641	1.0579	4.598	4.27e-06

AIC: 573.5

Table D.2. Logistic regression output: Model A

Model A Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-10.3079	0.5676	-18.162	< 2e-16
PIL	-1.3188	0.4264	-3.093	0.00198
SEA	0.4694	0.2031	2.312	0.02080
VIS	2.1502	0.7999	2.688	0.00719
EPF	5.2027	0.7851	6.627	3.43e-11
RUF	7.6606	0.5641	13.581	< 2e-16
MEF	5.8391	1.0934	5.340	9.29e-08
HEA	6.5447	0.4037	16.210	< 2e-16

AIC: 413.87

Table D.3. Logistic regression output: Model B

Model B Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-9.4220	0.3346	-28.156	< 2e-16
PIL	-1.3898	0.4275	-3.251	0.00115
VIS	1.7616	0.7629	2.309	0.02094
EPF	5.3201	0.8018	6.635	3.24e-11
RUF	7.8131	0.5730	13.635	< 2e-16
MEF	6.1152	1.0888	5.616	1.95e-08
HEB	5.9270	0.4048	14.643	< 2e-16

AIC: 427.24

Table D.4. Logistic regression output: Model C

Model C Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-10.1000	0.5575	-18.118	< 2e-16
PIL	-1.3132	0.4245	-3.094	0.00198
SEA	0.4151	0.2030	2.044	0.04091
VIS	1.6318	0.7719	2.114	0.03451
EPF	4.9999	0.7451	6.710	1.95e-11
RUF	7.6013	0.5468	13.901	< 2e-16
MEF	5.4338	1.1268	4.822	1.42e-06
HEC	4.9111	0.3809	12.893	< 2e-16

AIC: 463.5

Table D.5. Logistic regression output: Model D

Model D Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-9.2232	0.3056	-30.181	< 2e-16
PIL	-1.2674	0.4137	-3.064	0.00219
EPF	4.5307	0.7410	6.115	9.68e-10
RUF	7.6363	0.5349	14.276	< 2e-16
MEF	5.1825	1.1137	4.653	3.26e-06
HED	3.8710	0.3762	10.289	< 2e-16

AIC: 494.22

Table D.6. Logistic regression output: Model E

Model E Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-9.6047	0.3944	-24.352	< 2e-16
PIL	-1.3675	0.4251	-3.217	0.00130
DAY	0.7340	0.3697	1.985	0.04713
EPF	5.1757	0.7150	7.239	4.52e-13
RUF	7.3086	0.5303	13.782	< 2e-16
MEF	5.6201	1.0917	5.148	2.63e-07
HEE	3.2906	0.3672	8.962	< 2e-16

AIC: 510.86

APPENDIX E: ACCIDENT FREQUENCIES CALCULATED FOR THE REAL DATA

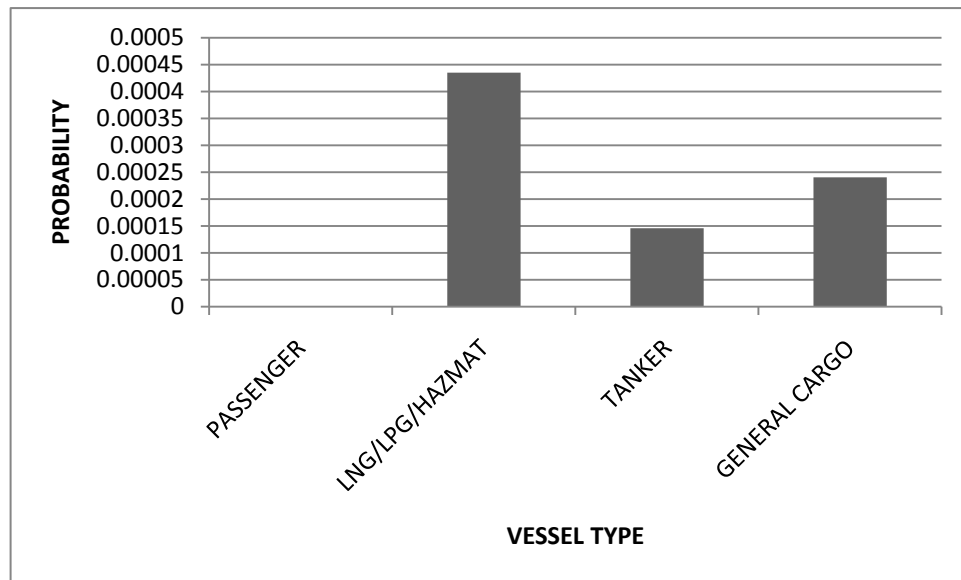


Figure E.1. Accident frequencies with respect to the vessel type

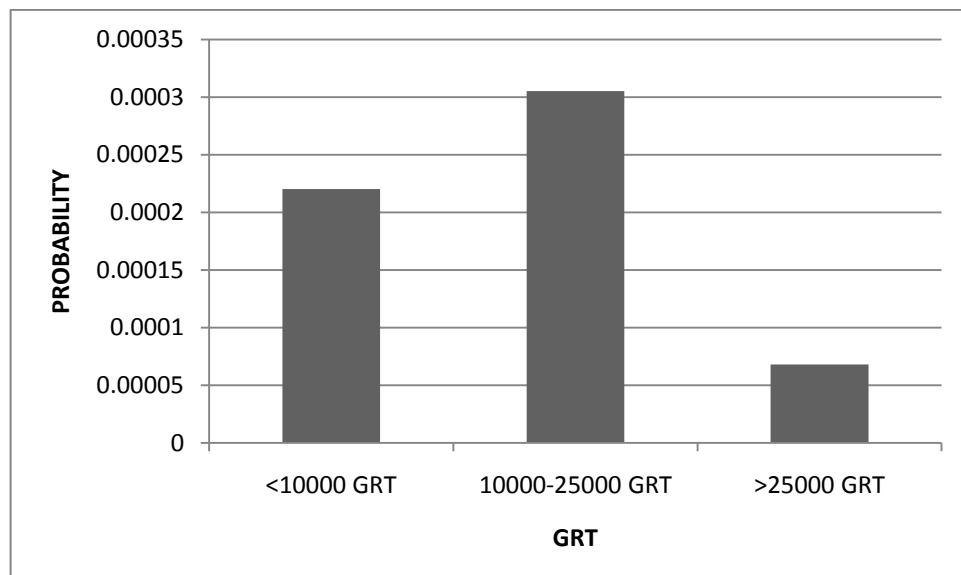


Figure E.2. Accident frequencies with respect to the vessel GRT

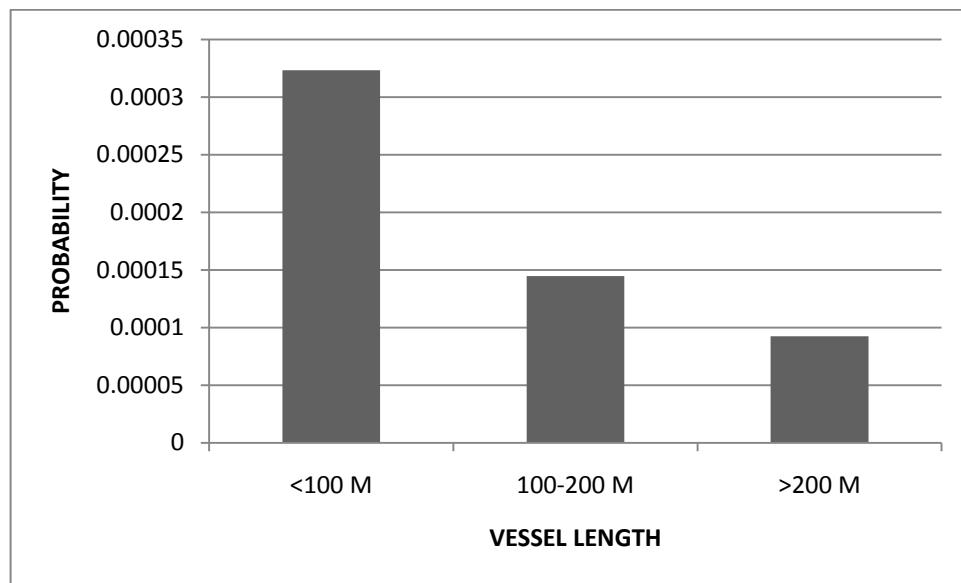


Figure E.3. Accident frequencies with respect to the vessel length

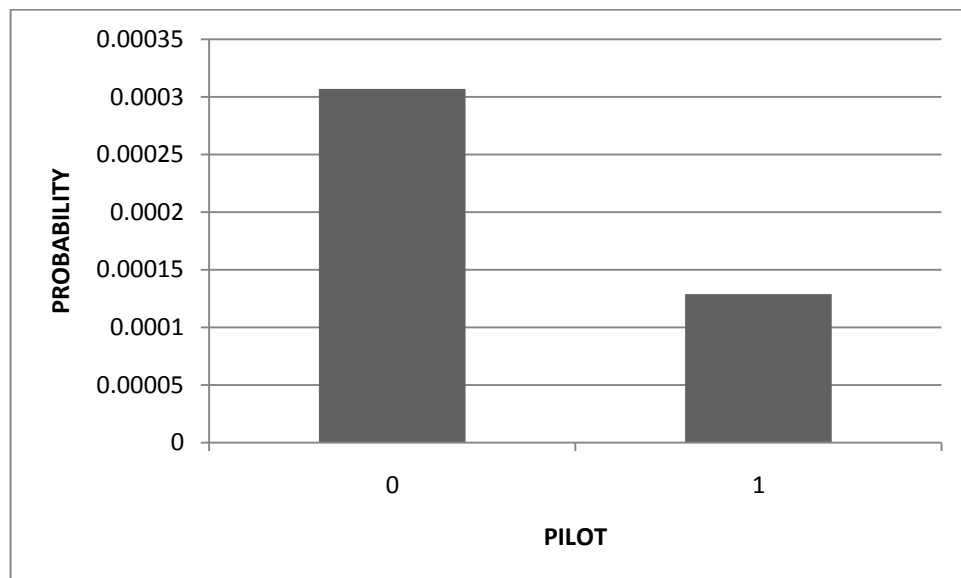


Figure E.4. Accident frequencies with respect to the pilot status

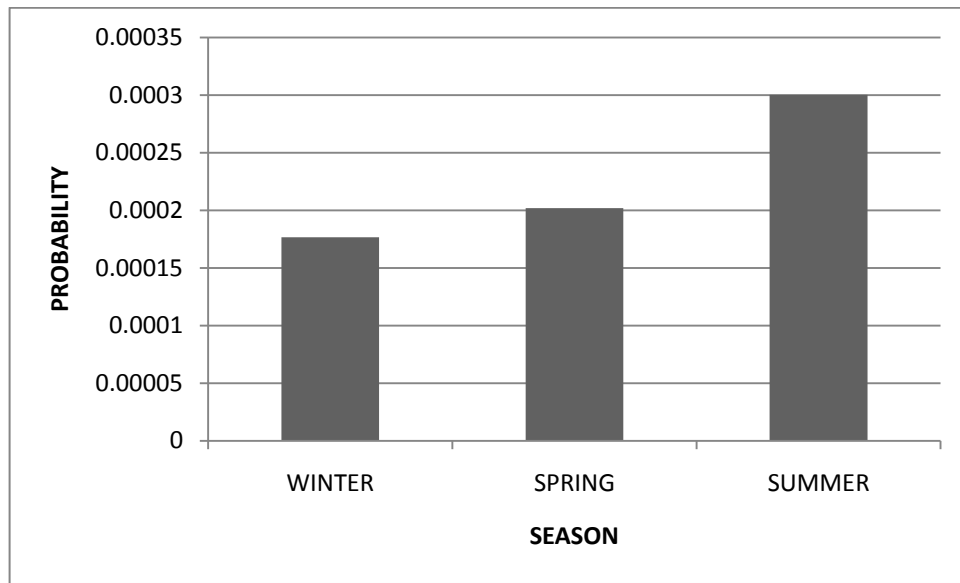


Figure E.5. Accident frequencies with respect to the accident season

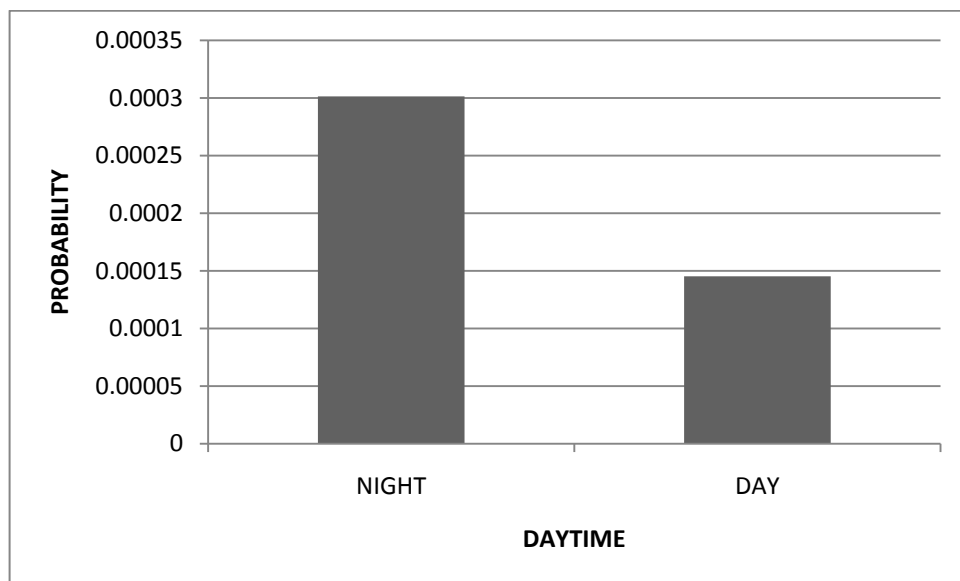


Figure E.6. Accident frequencies with respect to the accident time

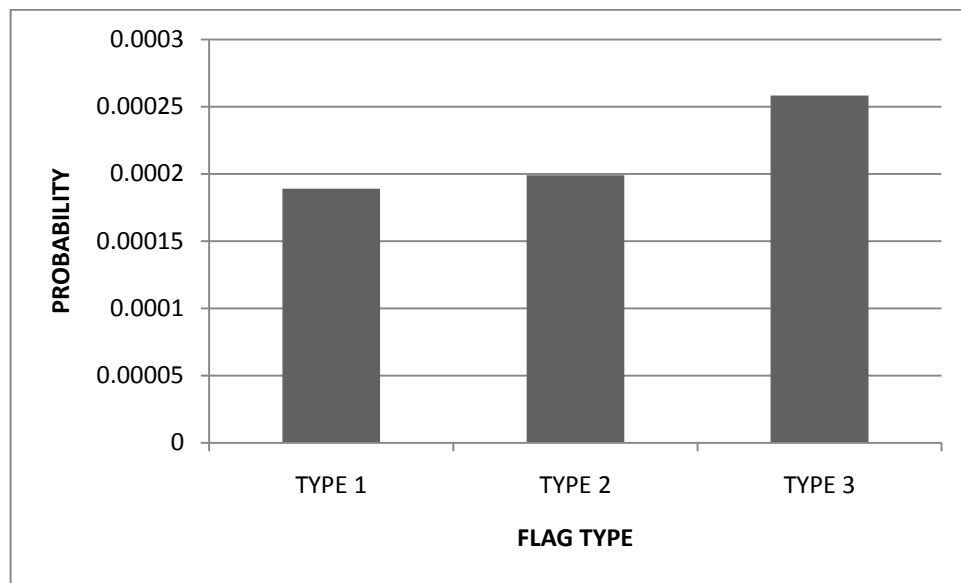


Figure E.7. Accident frequencies with respect to the vessel flag type

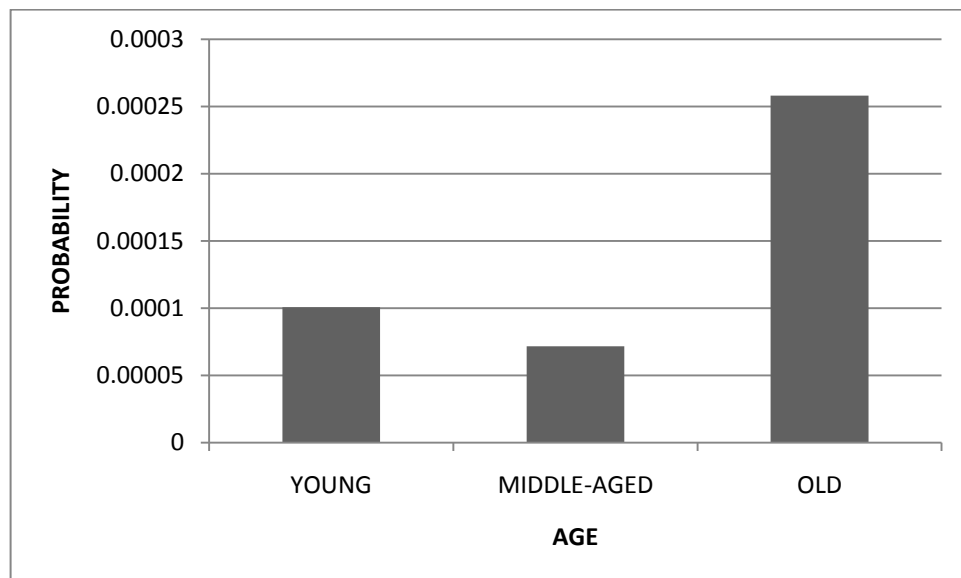


Figure E.8. Accident frequencies with respect to the vessel age

APPENDIX F: MODEL VERIFICATION OUTPUTS

Table F.1. Estimated accident probability frequencies for the original data

<i>PROB</i>	<i>MODEL FREQ</i>	<i>MODEL A FREQ</i>	<i>MODEL B FREQ</i>	<i>MODEL C FREQ</i>	<i>MODEL D FREQ</i>	<i>MODEL E FREQ</i>
0.00001-0.000055	23053	115469	75337	74982	75171	73705
0.000055-0.0001	57967	15733	85732	56881	85282	42141
0.0001-0.00055	83338	33607	3515	31862	0	43130
0.00055-0.001	967	158	0	32	0	2004
0.001-0.0055	100	94	70	792	5084	4484
0.0055-0.01	50	177	435	509	69	76
0.01-0.055	150	302	563	585	39	107
0.055-0.1	24	112	0	8	0	10
0.1-0.55	29	24	24	26	32	19
0.55-1	2	4	4	3	3	4

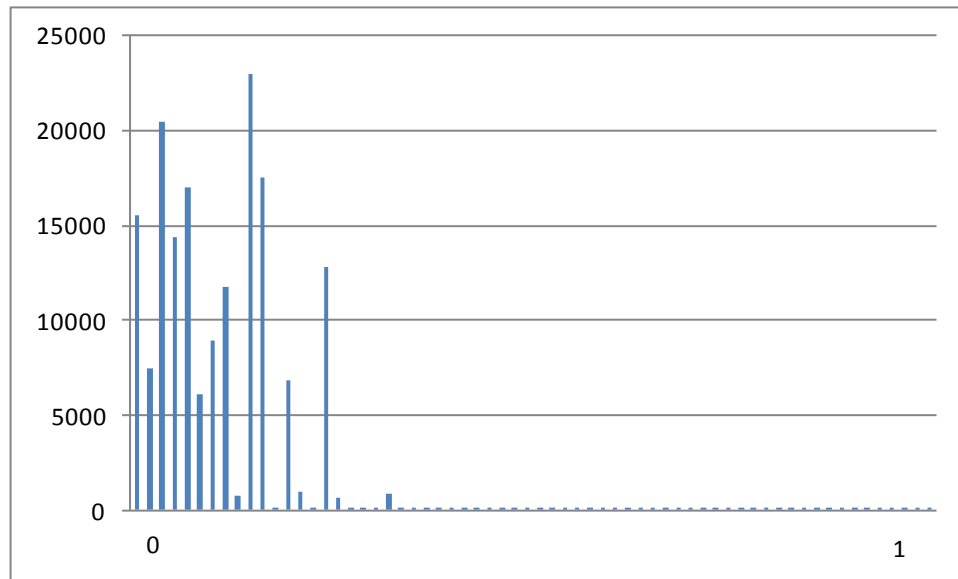


Figure F.1. Model probability histogram for non-accidents

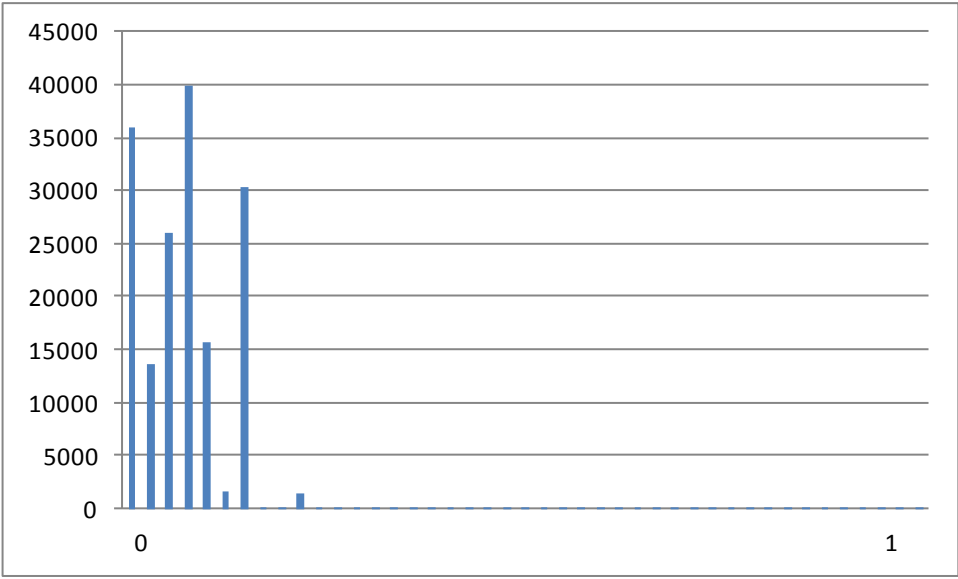


Figure F.2. Model A probability histogram for non-accidents

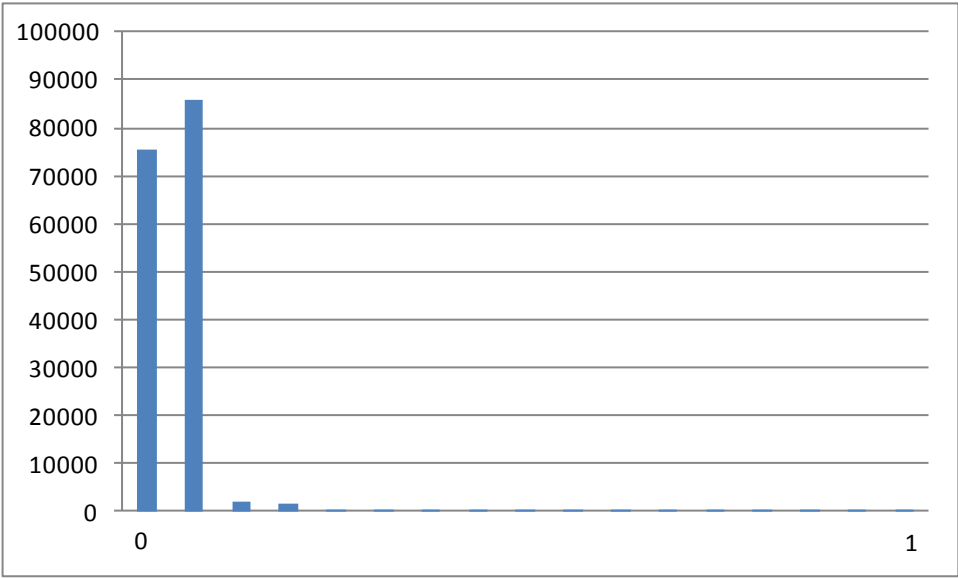


Figure F.3. Model B probability histogram for non-accidents

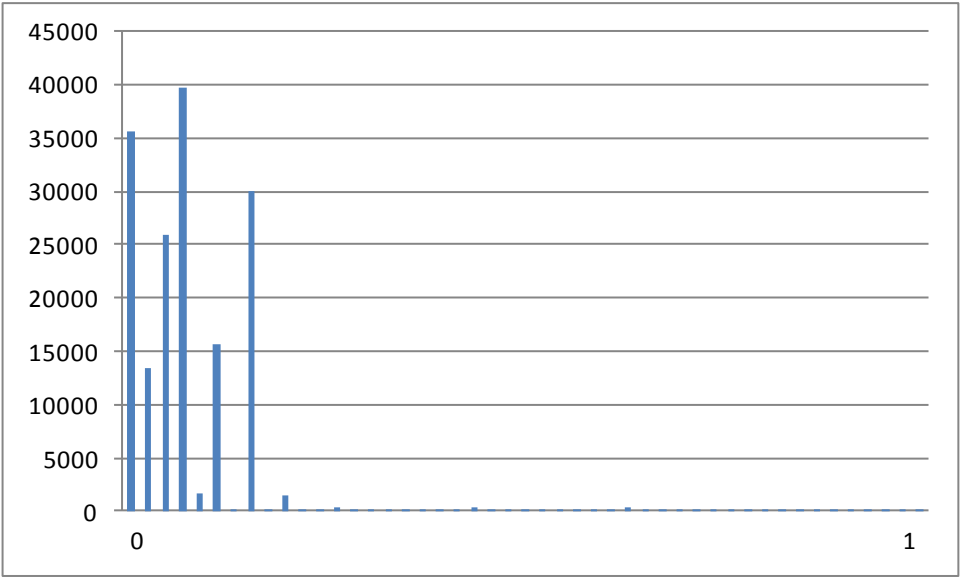


Figure F.4. Model C probability histogram for non-accidents

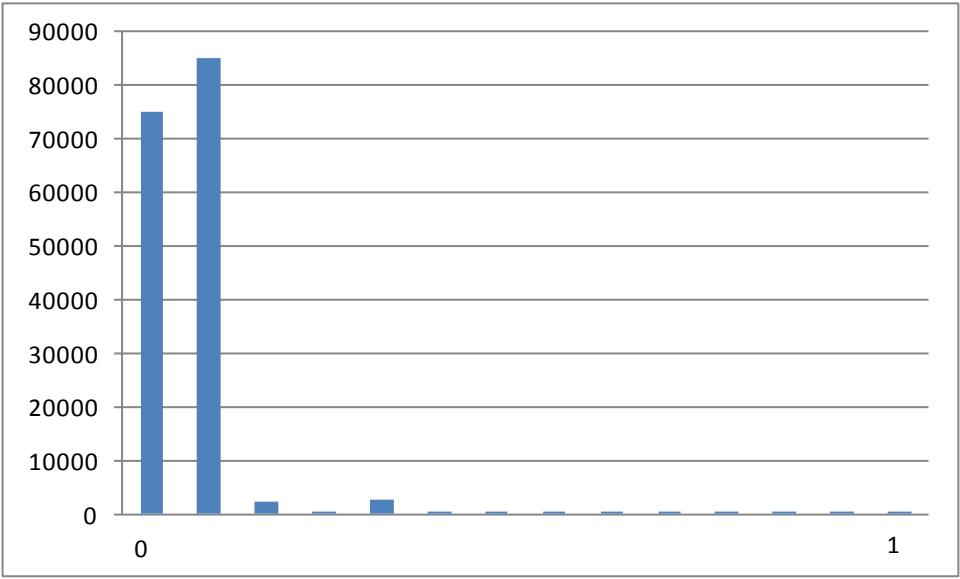


Figure F.5. Model D probability histogram for non-accidents

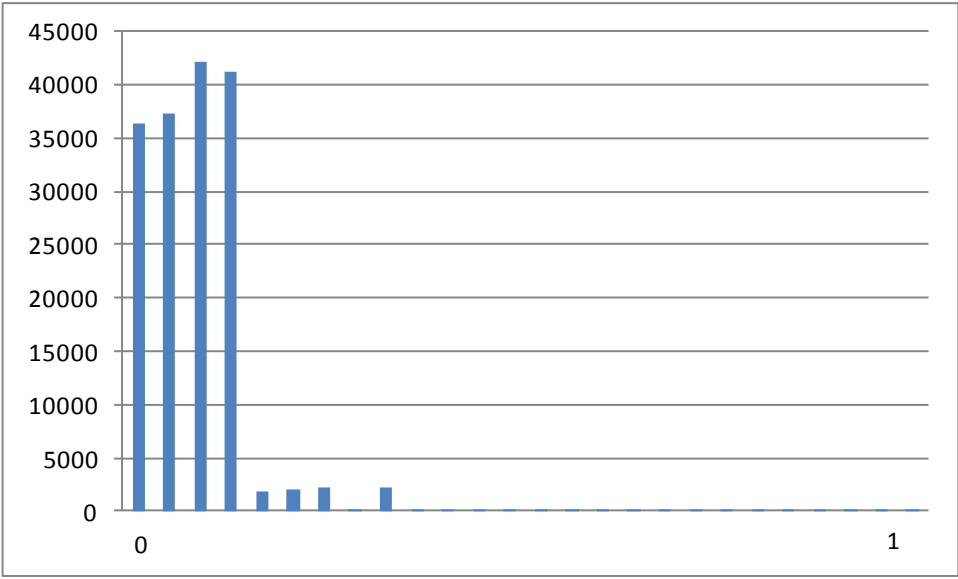


Figure F.6. Model E probability histogram for non-accidents

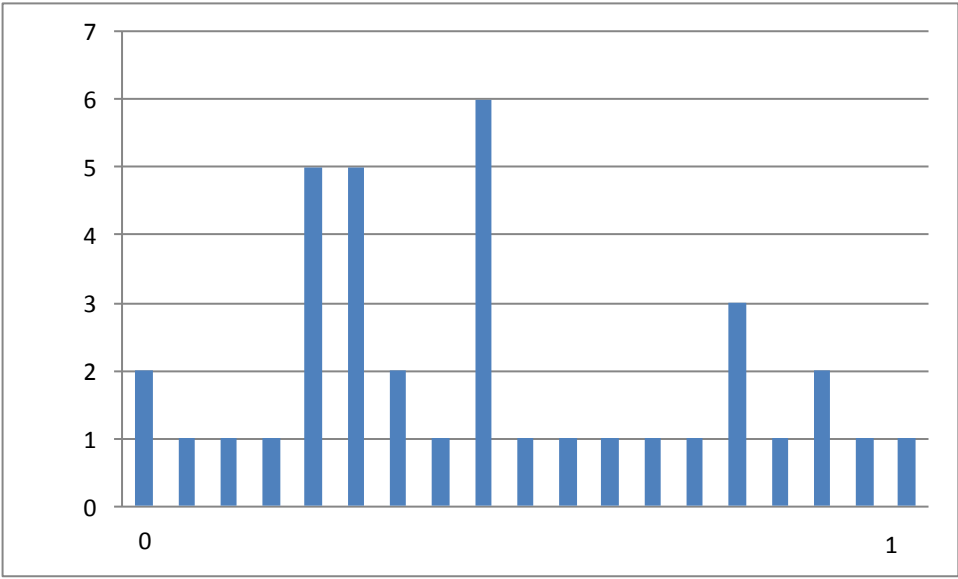


Figure F.7. Model probability histogram for the accidents

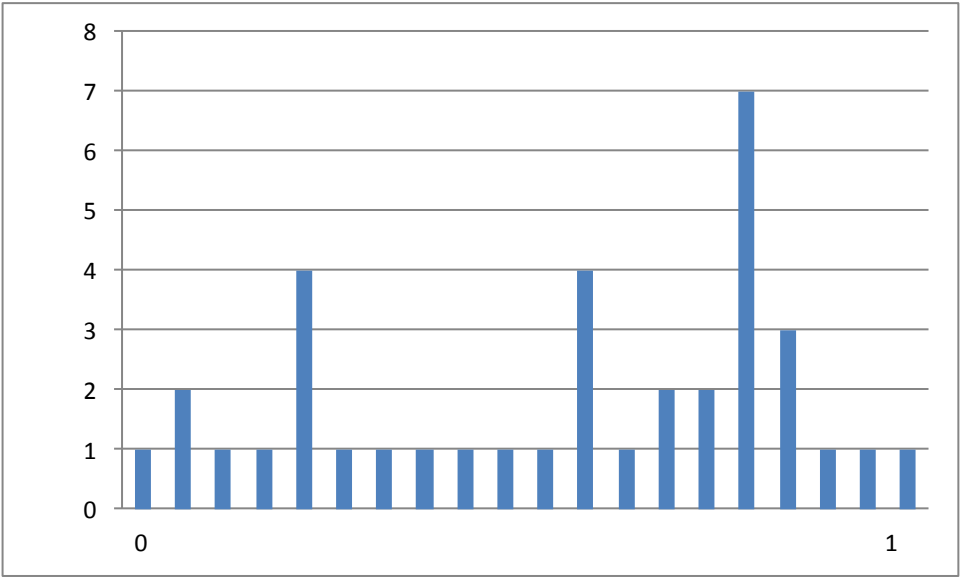


Figure F.8. Model A probability histogram for the accidents

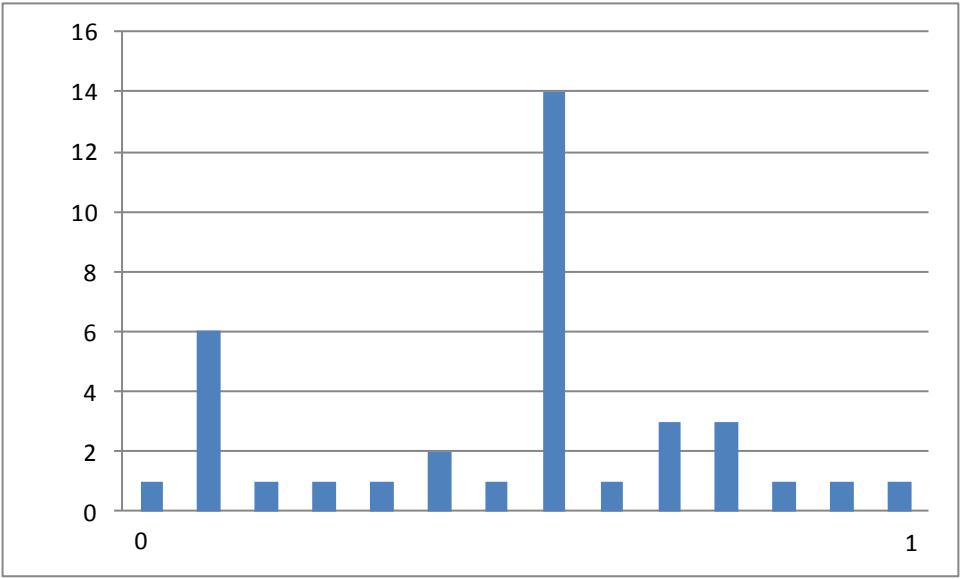


Figure F.9. Model B probability histogram for the accidents

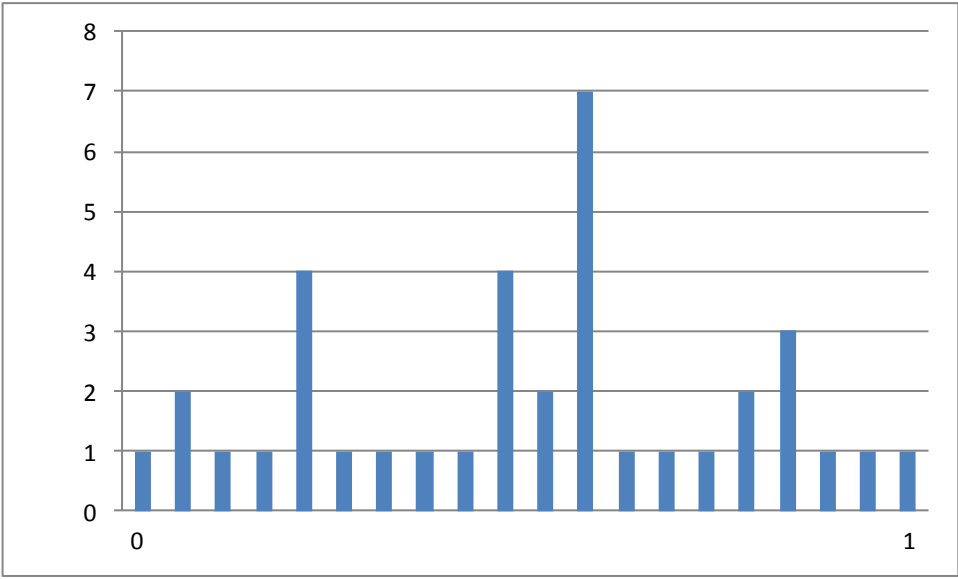


Figure F.10. Model C probability histogram for the accidents

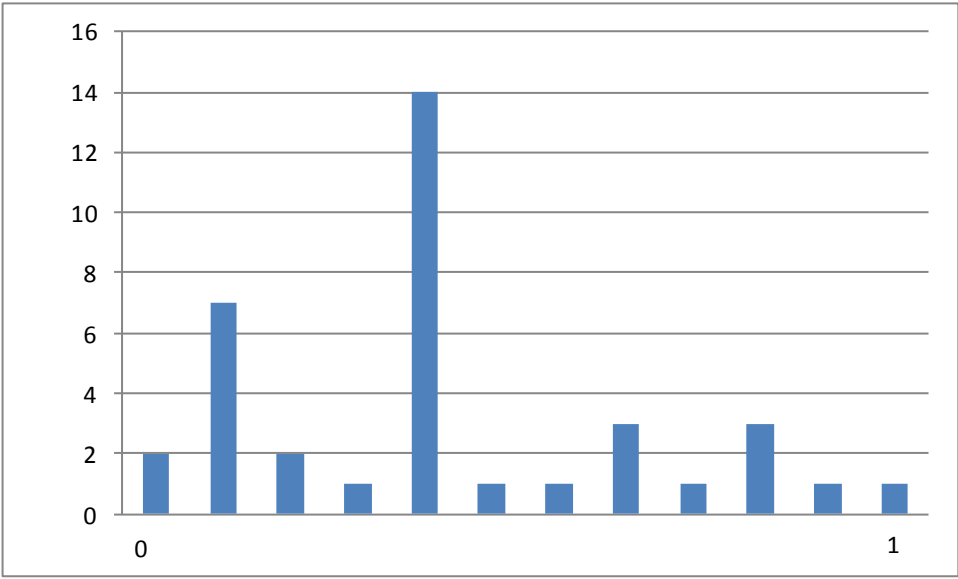


Figure F.11. Model D probability histogram for the accidents

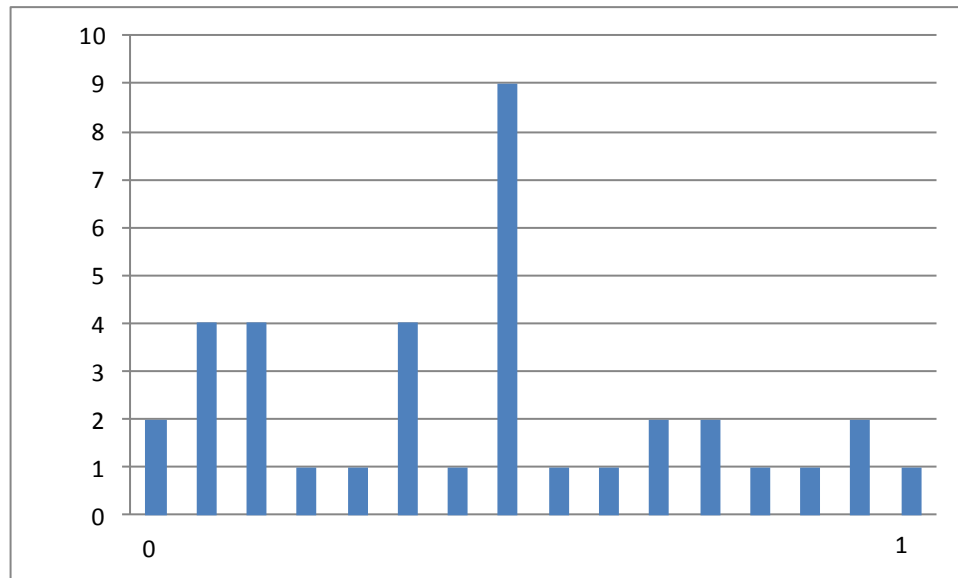


Figure F.12. Model E probability histogram for the accidents

APPENDIX G: MODEL VALIDATION OUTPUTS – ACCIDENT DATABASE

Table G.1. Estimated accident probability frequencies for the accident data

<i>PROB</i>	<i>MODEL FREQ</i>	<i>MODEL A FREQ</i>	<i>MODEL B FREQ</i>	<i>MODEL C FREQ</i>	<i>MODEL D FREQ</i>	<i>MODEL E FREQ</i>
0.00001-0.000055	3	0	0	0	7	7
0.000055-0.0001	20	0	0	5	19	11
0.0001-0.00055	113	23	26	19	0	13
0.00055-0.001	16	1	0	2	0	5
0.001-0.0055	2	8	2	14	135	123
0.0055-0.01	4	11	14	70	13	2
0.01-0.055	21	86	133	75	14	27
0.055-0.1	13	46	0	3	0	10
0.1-0.55	21	34	34	22	24	14
0.55-1	0	4	4	3	1	1

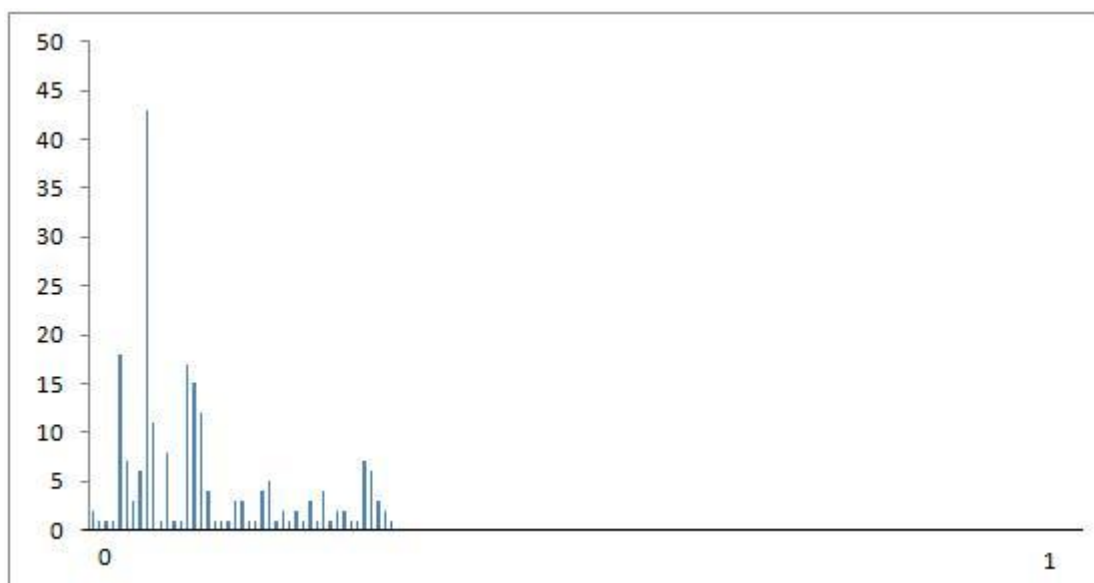


Figure G.1. Model probability histogram

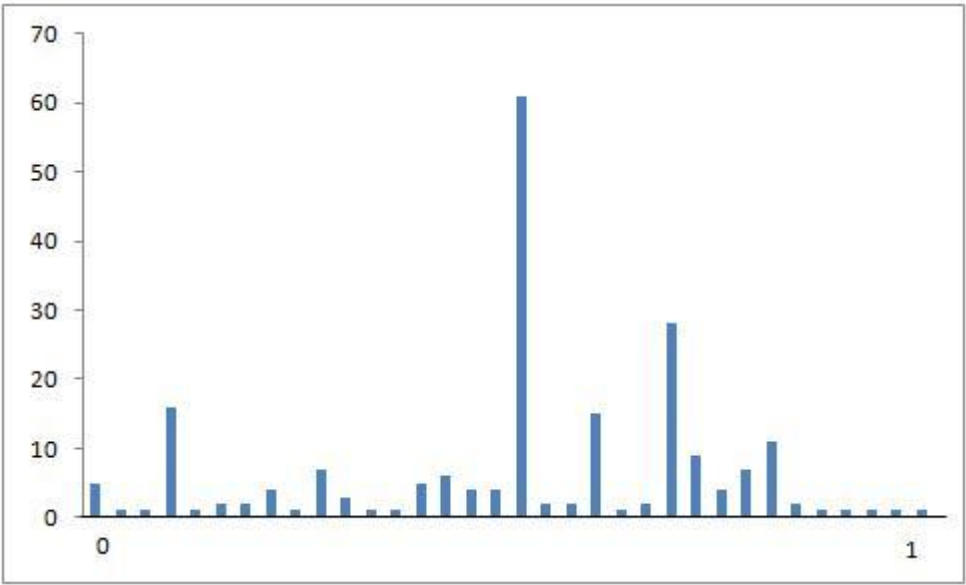


Figure G.2. Model A probability histogram

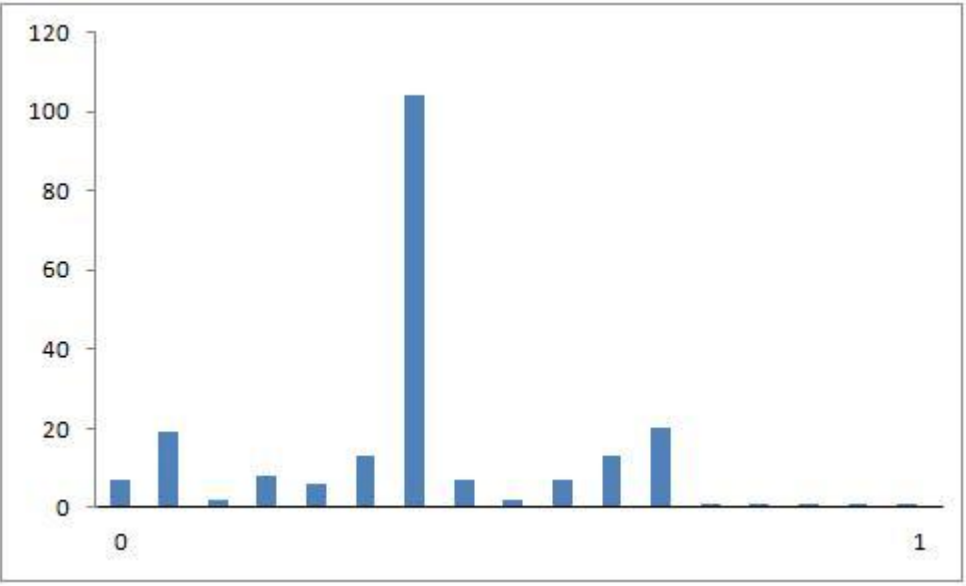


Figure G.3. Model B probability histogram

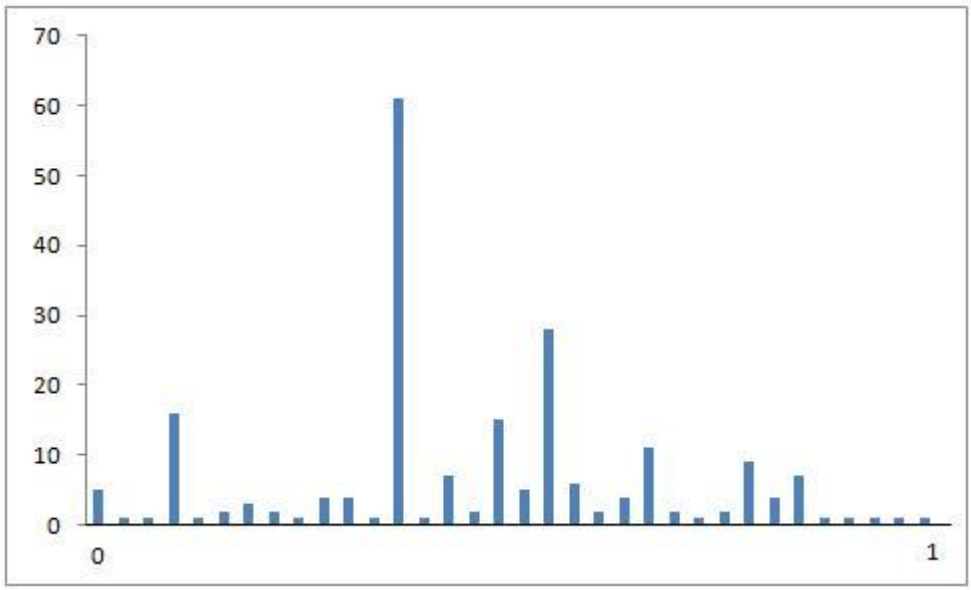


Figure G.4. Model C probability histogram

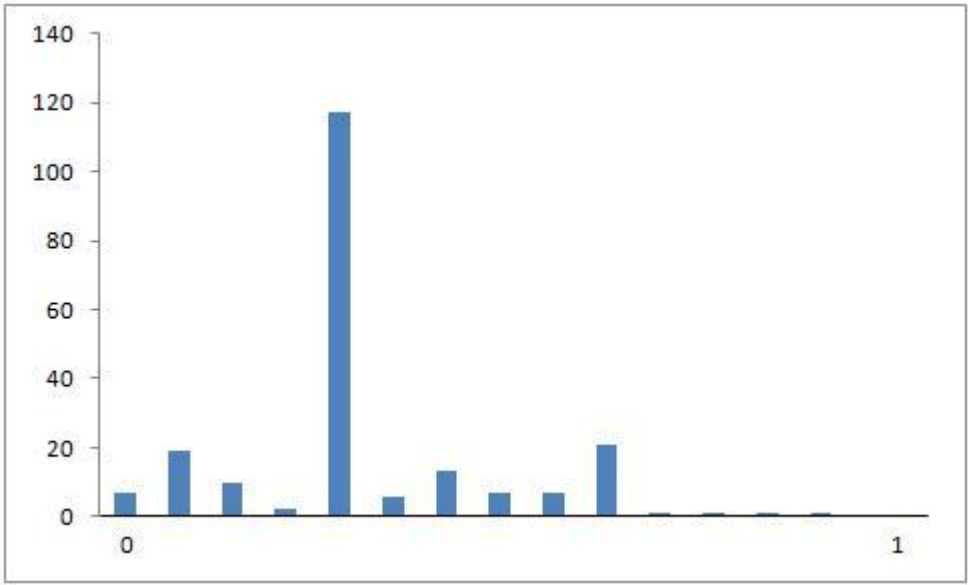


Figure G.5. Model D probability histogram

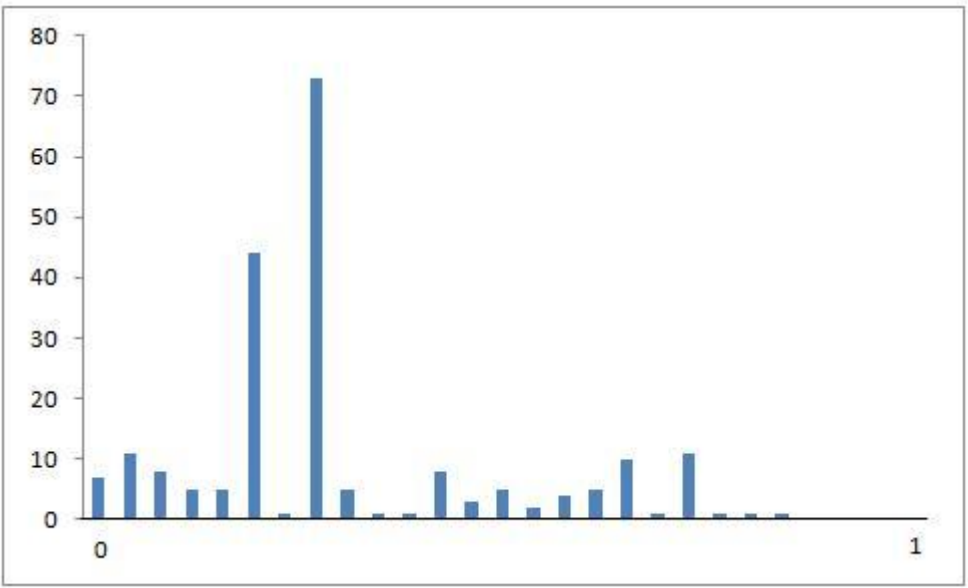


Figure G.6. Model E probability histogram

APPENDIX H: MODEL VALIDATION OUTPUTS – YEAR 2009

Table H.1. Estimated accident probability frequencies for the data of year 2009

<i>PROB</i>	<i>MODEL FREQ</i>	<i>MODEL A FREQ</i>	<i>MODEL B FREQ</i>	<i>MODEL C FREQ</i>	<i>MODEL D FREQ</i>	<i>MODEL E FREQ</i>
0.00001-0.000055	7724	37019	24531	24384	24072	23604
0.000055-0.0001	18816	4512	26091	17013	25698	13444
0.0001-0.00055	24591	9607	434	9383	0	12338
0.00055-0.001	156	27	0	19	0	625
0.001-0.0055	31	49	29	288	1609	1334
0.0055-0.01	25	48	146	161	18	33
0.01-0.055	63	113	183	164	17	36
0.055-0.1	13	39	0	3	0	3
0.1-0.55	2	5	4	4	7	4
0.55-1	0	2	3	2	0	0

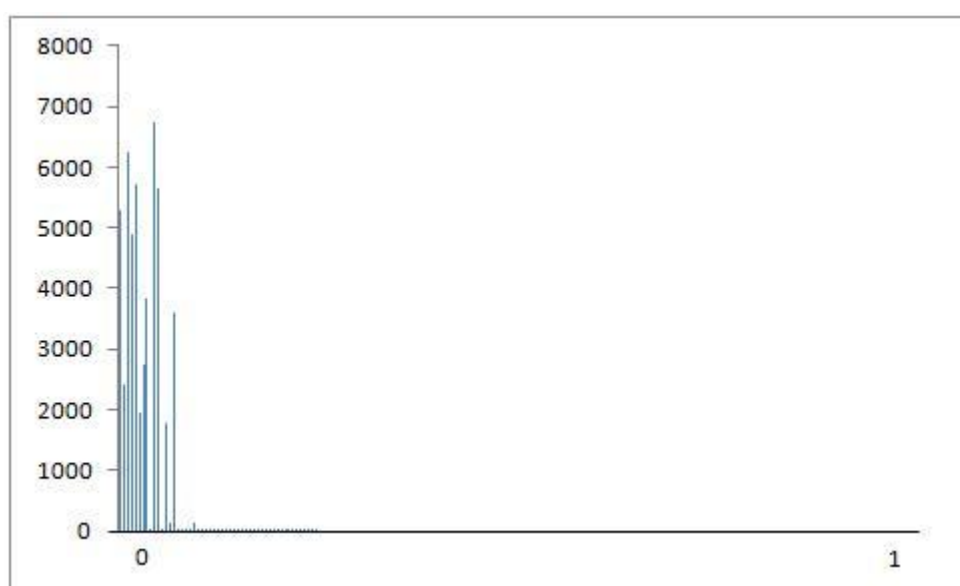


Figure H.1. Model probability histogram

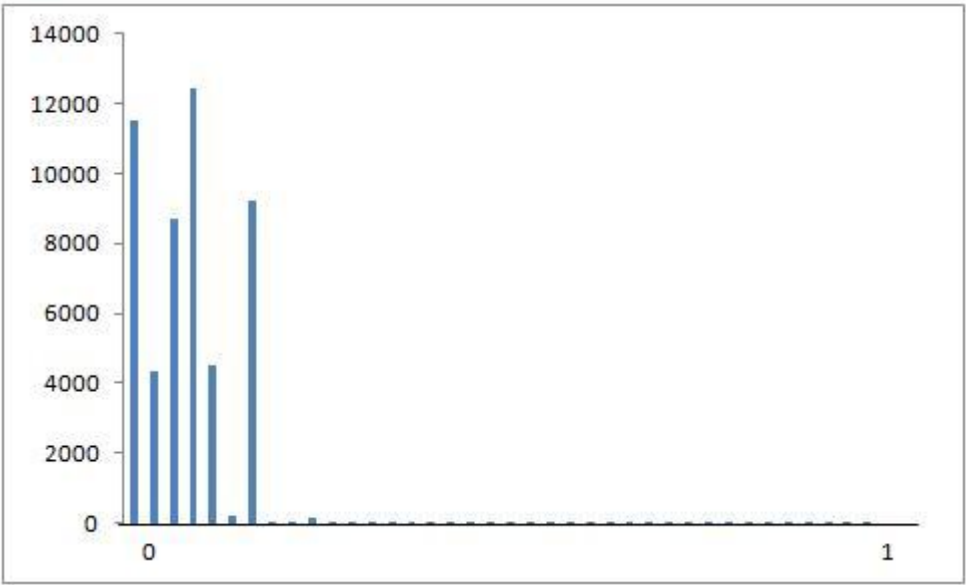


Figure H.2. Model A probability histogram

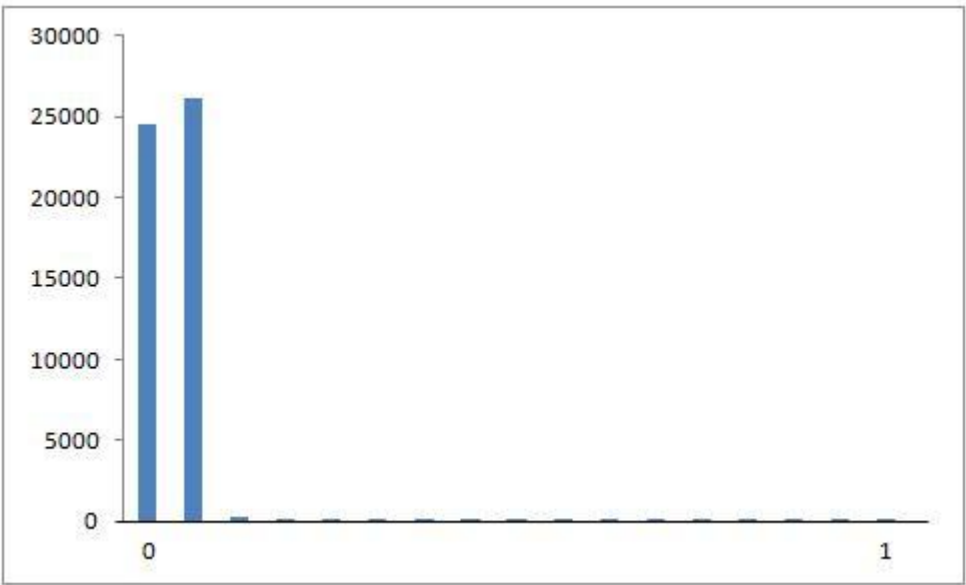


Figure H.3. Model B probability histogram

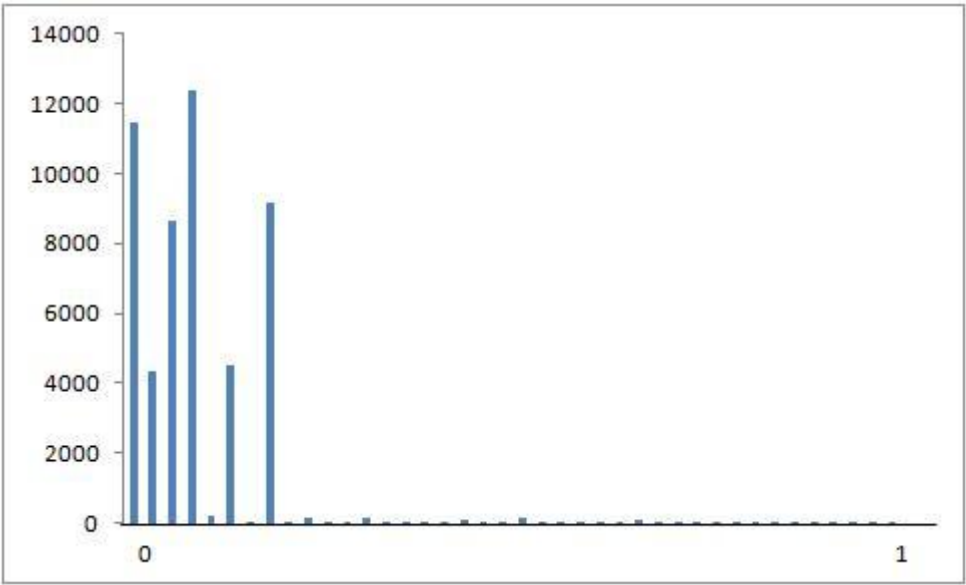


Figure H.4. Model C probability histogram

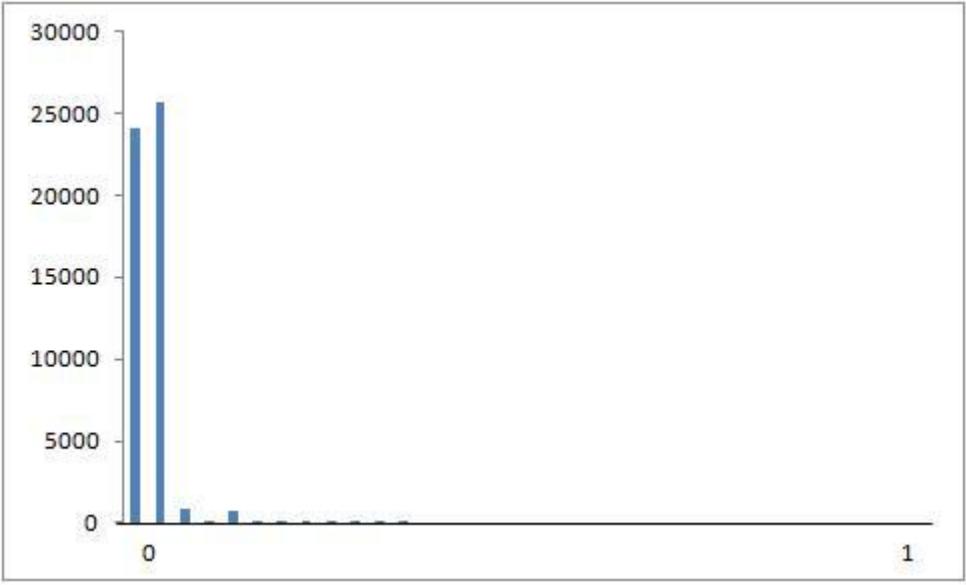


Figure H.5. Model D probability histogram

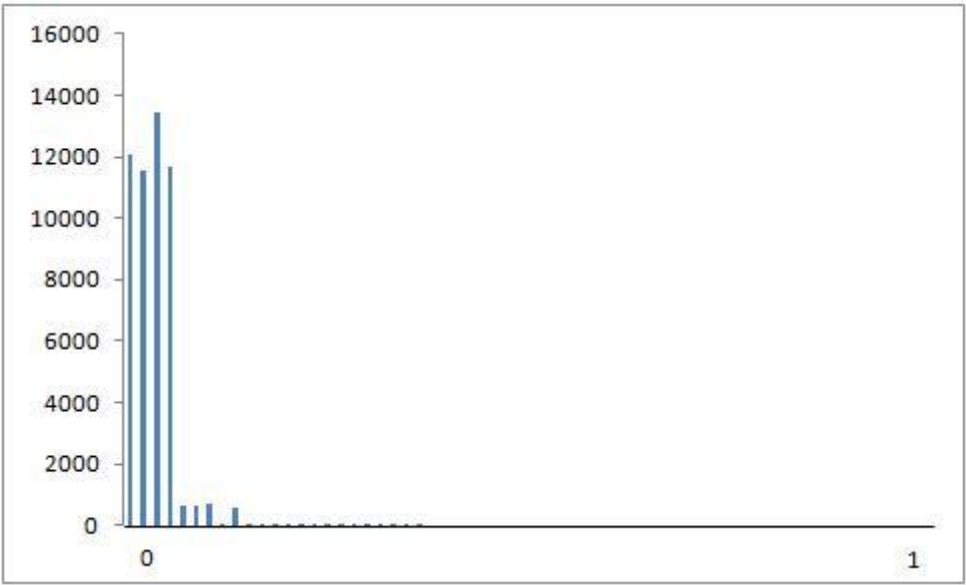


Figure H.6. Model E probability histogram

REFERENCES

- Almaz, A. Ö., 2006, *Investigation of the Maritime Transit Traffic in the Istanbul Channel through Simulation Modeling and Scenario Analysis*, M.S. Thesis, Department of Industrial Engineering, Boğaziçi University, Istanbul.
- Almaz, A. Ö., İ. Or and B. Özbaş, 2006, “Investigation of Transit Maritime Traffic in the Strait of Istanbul through Simulation Modeling and Scenario Analysis”, *Journal of Simulation*, Vol. 7, No. 7, pp. 1-9.
- Altıok, T., 1996, “Tools of Probability”, in *Performance Analysis of Manufacturing Systems*, pp. 15-65, Springer, New York.
- Anderson, E. and W. K. Talley, 1995, "The Oil Spill Size of Tanker and Barge Accidents: Determinants and Policy Implications", *Land Economy*, Vol. 71, pp. 216-228.
- Banks J., J. S. Carson II, B. L. Nelson and D. M. Nicol, 2000, “*Discrete-Event Systems Simulation*”, Prentice Hall, New Jersey.
- Bewick, V., L. Cheek and J. Ball, “Statistics Review 14: Logistic Regression”, *Critical Care*, Vol. 9, No.1, pp. 112-118, 2005.
- Birpınar, M. E., G. F. Talu and B. Gönençgil, 2009, “Environmental Effects of Maritime Traffic on the Istanbul Strait”, *Environmental Monitoring and Assessment*, Vol. 152, pp. 13–23.
- Crowder K. and P. Glynn, 2000, *SPSS PC Version 10: Crosstabs with Chi-Square Test*, <http://staff.washington.edu/glynn/chissps.pdf>
- Digital Seas Website*, <http://www.digital-seas.com>, 2010.

- Ece, J. N., *İstanbul Boğazı'ndaki Deniz Kazalarının Seyir ve Çevre Güvenliği Açısından Analizi ve Zararsız Geçiş Koşullarında Değerlendirilmesi*, Phd. Thesis, Institute of Science and Technology, Gazi University, Ankara, 2005.
- Ece, J. N., 2007, *İstanbul Boğazı: Deniz Kazaları ve Analizi*, 1st Edition, Deniz Kılavuzluk A.Ş., İstanbul.
- Field, A., 2005, *Discovering Statistics Using SPSS*, 2nd Edition, Sage Publications, London.
- Garson D., 2008, *Discriminant Function Analysis*, <http://faculty.chass.ncsu.edu/garson/PA765/discrim.htm>
- Guedes Soares, C. and A. P. Teixeira, 2001, "Risk Assessment in Maritime Transportation", *Reliability Engineering and System Safety*, Vol. 74, No. 3, pp. 299-309.
- Istanbul Seabuses and Fast Ferries Inc. Website*, <http://www.ido.com.tr>, 2010.
- Kite-Powell, H. L., N. M. Patrikalakis, D. Jin, S. L. Abrams, J. Jebsen, V. Papakonstantinou and S. C. Lin, 1998, *Formulation of a Model for Ship Transit Risk*, Final Project Report, Massachusetts Institute of Technology, Massachusetts.
- Boğaziçi University Kandilli Observatory and Earthquake Research Institute Website*, <http://www.koeri.boun.edu.tr>, n.d.
- Le Blanc, L. A. and C. T. Rucks, 1996, "Multiple Discriminant Analysis of Vessel Accidents", *Accident Analysis and Prevention*, Vol. 28, No. 4, pp. 501-510.
- Marques de Sá, J. P., *Applied Statistics Using SPSS, STATISTICA, MATLAB and R*, 2nd Edition, Springer, Heidelberg, 2007.

- Merrick, J. R. W., J. R. Van Dorp, J. Harrauld, T. Mazzuchi, J. E. Spahn, and M. Grabowski, 2000, "A Systems Approach to Managing Oil Transportation Risk in Prince William Sound", *Systems Engineering*, Vol. 3, No. 3, pp. 128-142.
- Merrick, J. R. W., J. R. Van Dorp, T. A. Mazzuchi and J. Harrauld, 2001, "Modeling Risk in the Dynamic Environment of Maritime Transportation", *Proceedings of the 2001 Winter Simulation Conference*, Vol. 2, pp. 1090-1098, Arlington, Virginia.
- Merrick, J. R. W., V. Dinesh, A. Singh, J. René van Dorp and T. A. Mazzuchi, 2003, "Propagation of Uncertainty in a Simulation-Based Maritime Risk Assessment Model Utilizing Bayesian Simulation Techniques", *Proceedings of the 2003 Winter Simulation Conference*, pp. 449-455, New Orleans, Louisiana.
- Morgan, G. A., N. L. Leech, G. W. Gloeckner and K. C. Barrett, 2004, *SPSS for Introductory Statistics*, 2nd Edition, Lawrence Erlbaum Associates, Mahwah, New Jersey.
- Or, İ., B. Özbaş and T. Yılmaz, 2007, "Simulation of Maritime Transit Traffic in the Istanbul Strait-II: Incorporating the Traffic Regime, Arrival Processes, Meteorological Conditions", *Proceedings of the 21st European Conference on Modelling and Simulation*, Prague.
- Özbaş, B., 2005, *Simulation of Maritime Transit Traffic in the Istanbul Channel*, M.S. Thesis, Department of Industrial Engineering, Boğaziçi University, Istanbul.
- Özbaş, B. and İ. Or, "Analysis and Control of Maritime Transit Traffic through the Istanbul Channel: a Simulation Approach", *CEJOR*, Vol. 15, pp. 235-252, 2007.
- Roeleven, D., M. Kok, H. L. Stipdonk and W. A. De Vries, 1995, "Inland Waterway Transport: Modelling the Probability of Accidents", *Safety Science*, Vol. 19, No. 2-3, pp. 191-202.
- Shipping Database Website*, <http://www.shippingdatabase.com>, n.d.

Ship Spotting Website, <http://www.shipspotting.com>, n.d.

Talley, W. K., 1996, "Determinants of Cargo Damage Risk and Severity: The Case of Containership Accidents", *Logistics and Transportation Review*, Vol. 32, pp. 377-388.

Ulusçu, S. Ö., B. Özbaş, T. Altıok and İ. Or, 2008, "*Risk Analysis of the Transit Vessel Traffic in the Strait of Istanbul*", Executive Summary, http://ise.rutgers.edu/research/working_paper/paper_08-018.pdf

Ulusçu, S. Ö., B. Özbaş, T. Altıok, İ. Or and T. Yılmaz, 2009, "Transit Vessel Scheduling in the Strait of Istanbul", *Journal of Navigation*, Vol. 62, No. 1, pp. 59-77.

Ulusçu, S. Ö., B. Özbaş, T. Altıok and İ. Or, 2009, "Risk Analysis of the Vessel Traffic in the Strait of Istanbul", *Risk Analysis*, Vol. 20, No. 10, pp. 1454-1472.

Walpole, R. E., R. H. Myers, S. L. Myers and K. Ye, *Probability and Statistics for Engineers and Scientists*, 8th Edition, Pearson Prentice Hall, New Jersey, 2007.

World Shipping Register Website, <http://e-ships.net/ships.htm>, n.d.

Weather Underground Website, <http://www.wunderground.com>, 2010.

Yip, T. L., 2008, "Port Traffic Risks – A Study of Accidents in Hong Kong Waters", *Transportation Research*, Vol. 44, Part E, pp. 921–931.