Multi-variate flood loss estimation model for residential sector

Wahab, Roshan

2016

Wahab, R. (2016). Multi-variate flood loss estimation model for residential sector. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/69012>

<https://doi.org/10.32657/10356/69012>

NANYANG
TECHNOLOGICAL
UNIVERSITY

**MULTI-VARIATE FLOOD LOSS ESTIMATION
MODEL FOR RESIDENTIAL SECTOR**

ROSHAN WAHAB

SCHOOL OF CIVIL AND ENVIRONMENTAL ENGINEERING

2016

Multi-Variate Flood Loss Estimation Model for
Residential Sector

ROSHAN WAHAB

School of Civil and Environmental Engineering

A thesis submitted to the Nanyang Technological University in
partial fulfilment of the requirement for the degree of Doctor
of Philosophy

2016

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor, Associate Professor Robert Tiong, for his continuous support, warm encouragement, patient guidance and valuable advice during this research.

I wish to thank Emeritus Professor Chen Charng Ning, A/P Edmond Lo, A/P Shuy Eng Ban, and A/P Qin Xiaosheng for their continuous support and constructive advices. I also acknowledge the invaluable assistance and insightful questions from Dr. Weimin Dong and Dr. Pane Stojanovski.

I also wish to thank Dr. Ade Asmi, Head of Civil Engineering Department, Universitas Bakrie, Jakarta, and the students from civil engineering for helping me during the field survey in Jakarta.

I would like to express my thanks to Nanyang Technological University for the financial support provided during the course of this research works.

Special thanks are given to Mr Suresh Betapudi, Project Manager ICRM, Mrs. Li Juan, Mrs. Arsyanti Nurlailamari, Mr. Shankar Acharya K, Ms. Tan Juivy, Mrs. Chia Lan, Mr. Shao Zhe, Mr. Amiya Patnaik, Mr. F Rozy, Mr. Vivek B, Mr. Vipin S P, Mr. Varun M, Mr. Pramodh V and Mr. Xiao Zhe for their constant helps and constructive advices to this research work. Without them, the progress of this work would be arduous.

A special mention goes to all my friends who have made my life at NTU memorable and enjoyable.

Finally, but by no means least, thanks to my family for your love, support and the sacrifices you have made on my behalf. I would like to express my sincere thanks to my loving wife Sasna for encouraging and supporting me throughout this experience.

Table of Contents

Acknowledgements.....	i
Table of Contents.....	ii
Abstract.....	v
List of Figures.....	vii
List of Tables.....	x
Symbols and Abbreviations	xii
CHAPTER 1. INTRODUCTION	1
1.1 Background and Justification	1
1.2 Objectives.....	5
1.3 Organisation of the Research	6
CHAPTER 2. REVIEW OF DIRECT FLOOD LOSS ESTIMATION.....	8
2.1 Introduction	8
2.2 Definitions.....	8
2.2.1 Hazard.....	8
2.2.2 Vulnerability	8
2.2.3 Damage and loss	10
2.3 Urban Flooding	10
2.4 Assessment of Flood Damage	11
2.4.1 Flood damage typology.....	11
2.4.2 Levels of assessment.....	12
2.4.3 Depth – damage curves	13
2.4.4 Factors influencing flood damages or losses	15
2.4.5 Past studies.....	17
2.4.6 Overall analysis of the past studies	32
2.5 Conclusion.....	34
CHAPTER 3. METHODOLOGY	36
3.1 Introduction	36
3.2 Multiple Kernel Regression	36
3.2.1 Introduction to kernel methods	36
3.2.2 Support vector machines.....	39
3.2.3 Multiple kernel learning.....	41
3.3 Baseline Model Development	46
3.3.1 Principal component analysis	47

3.3.2	Mutual information	48
3.3.3	Cluster analysis	50
3.4	Performance Measures	51
CHAPTER 4. DEVELOPMENT OF A RESIDENTIAL FLOOD LOSS ESTIMATION MODEL FOR JAKARTA.....		52
4.1	Introduction	52
4.2	Study Region and Flood Loss Database.....	52
4.2.1	Study area.....	52
4.2.2	Data	55
4.3	Derivation of Indicators	58
4.3.1	Emergency measures	58
4.3.2	Precautionary measures	59
4.3.3	Flood experience index	59
4.4	Scales used for Representing Velocity and Contamination	60
4.4.1	Flow velocity	60
4.4.2	Contamination.....	61
4.5	Proposed Model Development	61
4.5.1	Principal component analysis	64
4.5.2	Mutual information	67
4.5.3	Cluster analysis	69
4.5.4	Development of residential loss functions	75
4.6	Model Performance Evaluation.....	86
4.7	Summary	90
CHAPTER 5. EVALUATION OF FACTORS INFLUENCING FLOOD LOSS.....		91
5.1	Introduction	91
5.2	Variable and Dataset	92
5.3	Proposed Input Feature Selection Method	95
5.3.1	Kernel function selection	96
5.3.2	Multiple kernel learning for regression (MKr)	99
5.3.3	Recursive feature elimination	100
5.4	Results	104
5.4.1	Factor/Feature identification.....	104
5.4.2	Model development	111
5.4.3	Model performance evaluation	114
5.5	Summary	118

CHAPTER 6. An INTEGRATED DATA DRIVEN AND STATISTICAL FRAMEWORK FOR DEVELOPING MULTI-VARIATE RESIDENTIAL FLOOD LOSS ESTIMATION MODEL	119
6.1 Introduction	119
6.2 Proposed Enhanced Sample Generator	120
6.2.1 Gibbs sampler	121
6.2.2 Hybrid Bayesian networks	124
6.2.3 TAN model for regression	127
6.3 Framework Process	131
6.4 Results and Discussions	134
6.4.1 Factors influencing flood loss	134
6.4.2 Model development	143
6.4.3 Model performance evaluation	143
6.5 Summary	148
CHAPTER 7. CONCLUSIONS AND RECOMMENDATIONS	149
7.1 Conclusions	149
7.1.1 Development of residential flood loss estimation model for Jakarta	149
7.1.2 Identification of prominent factors influencing building and content loss from floods.....	150
7.1.3 Integrated data driven and statistical framework for developing multi-variate flood loss estimation model in a data deficient environment.....	151
7.2 Limitations and Recommendations for Future Works	152
REFERENCES.....	154
APPENDIX.....	168
Appendix A	168
Appendix B	177

Abstract

Depth-damage curves or loss functions which are used to assess the direct loss from flooding are highly inadequate for accurate and reliable analysis. These models have a simplified structure which only takes into account the inundation depth and building characteristics while defining flood losses. Although a plethora of factors which influence the residential flood loss have been identified in the literature, relatively few models and studies have been attempted in assessing their interaction in predicting flood loss. This dissertation demonstrates that notable enhancements could be made in the approaches used for developing flood loss/damage functions. To supplement the knowledge of the multi-variate flood loss estimation this thesis attempts to address the issues pertained to (i) identifying the key factors which influence flood loss and (ii) development of comprehensive loss estimation model in a data deficient environment.

To develop a residential flood loss estimation model for Jakarta, the regions of Pluit and Kelapa Gading in North Jakarta and Cengkareng in West Jakarta were selected to conduct structured questionnaire survey. The household surveys determined the characteristics of flood and the loss influencing factors in the study area. Due to the lack of information on factors for segregating the residential sector in the study area for flood loss estimation, a standardized procedure for categorising the flood loss cases objectively into homogenous groups is presented. The first part of the approach, which is a combination of Principal component analysis (PCA) supplemented with correlation and Mutual information (MI), showed that monthly household income and people's response measures against flood along with flood water depth essentially determine the extent of flood loss in Jakarta. Using monthly income as similarity measures in Hierarchical cluster analysis (HCA), the second part yielded 3 income groups with homogenous flood loss data. Depth-damage curves were developed for these groups and scaling factors were derived from the surveyed data to represent the effect of people's response measures against floods.

To identify the main factors which influence flood damage and loss, a novel hybrid heterogeneous data feature selection algorithm based on kernel methods is proposed

herein. This algorithm captures the importance and relevance of the factors using Multiple kernel learning (MKL) and mutual information (MI) and capture the interaction among them using Multiple kernel support vector regression (MKr) in order to identify those factors whose single and joint interaction determine the flood loss . The outcome of the proposed approach when implemented in Jakarta showed that apart from water depth other variables belonging to the domain of flood response actions, socio-economic characteristics and flood warning play a significant role during flooding. Lastly this data driven multi-variate flood loss estimation model (MKr model) has been found to perform significantly better than the loss estimation models developed using existing methodologies.

Finally a new data driven and statistical framework for modelling residential flood loss in a data deficient environment was developed and presented in the dissertation. Key component of the framework is the probabilistic generation of synthetic data for underrepresented variables using an enhanced Gibbs sampler. The Enhanced MKr model developed by employing the proposed framework in Jakarta displayed comparable performance with the MKr model. However the factor selection stability of the proposed algorithm increased substantially under the new framework. Moreover, the following precautionary and emergency measures such as temporary water barriers, waterproofing of external walls, raised floors and using low value & easily movable objects in flood prone floor were identified by the En-MKr model which has the potential in reducing the damage from floods in Jakarta. The whole framework is beneficial in developing a comprehensive multi-variate flood loss estimation model especially with small datasets.

List of Figures

Figure 1-1: Natural catastrophes 2014- World map	1
Figure 2-1: Flood vulnerability indicators	9
Figure 2-2: Expected flood damage - slow rise flood – average	32
Figure 3-1: Kernel mapping.....	37
Figure 4-1: Rivers and main drains of Jakarta	53
Figure 4-2: Jakarta inundation map, 2013 Floods	57
Figure 4-3: Scatter plot depicting total loss for residential sector	62
Figure 4-4: Histogram of the total damage value	63
Figure 4-5: Histogram of damage to building structure.....	63
Figure 4-6: Histograms of damage to contents	64
Figure 4-7: Dendrogram for Wards clustering applied to building loss (Dash-dot line indicates the locations where dendrogram was sliced.)	72
Figure 4-8: Dendrogram for Wards clustering applied to content loss (Dash-dot line indicates the locations where dendrogram was sliced)	73
Figure 4-9: Flowchart showing the approach used for demarcating flood loss data into homogenous groups	74
Figure 4-10: Micro-scale residential flood loss estimation model for building structure. Bars represent 1 standard deviation	76
Figure 4-11: Micro-scale residential flood loss estimation model for contents. Bars represent 1 standard deviation	76
Figure 4-12: Flood experience before 2013 January	77
Figure 4-13: The influence of raising the floor level at the entry point on building and content loss (bars=means, points = median and whiskers = 25 and 75 percentile). Significant difference observed for both building and content loss.	79
Figure 4-14: The influence of using low cost/value and easily movable items on building and content loss (bars=means, points = median and whiskers = 25 and 75 percentile). Significant difference in loss is observed only for content.....	80
Figure 4-15: Building and content loss for household which has not undertaken waterproofing (bars=means, points = median and whiskers = 25 and 75 percentile). No significant difference found for both building and content loss.	81

Figure 4-16: Building and content loss for household which has installed and not installed watertight shields over openings such as doors, windows etc. (bars=means, points = median and whiskers = 25 and 75 percentile). No significant difference found for both Building and content loss.....	82
Figure 4-17: The influence of permanent or mobile flood barriers on the content and building loss (bars=means, points = median and whiskers = 25 and 75 percentile). Significant difference found for both building and content loss.....	83
Figure 4-18: Building and content loss for households who has and has not moved the objects including furniture and other valuables to safety (bars=means, points = median and whiskers = 25 and 75 percentile) Significant difference observed for content loss only.	84
Figure 4-19: Comparison of RMSE values for the flood loss estimation models ...	87
Figure 4-20: Box plot representing the correlation coefficient of flood loss estimation models	87
Figure 4-21: Comparison of flood loss estimation model in terms of their mean bias error (MBE)	89
Figure 5-1: Prediction performance for each size of selected factors for building loss	106
Figure 5-2: Prediction performance for each size of selected factors for content loss	106
Figure 5-3: Factor selection stability for building and contents	108
Figure 5-4: Composition of factors and their proportion in the top 10 and 14 ranked factors selected by 50 building (a) and content (b) loss prediction models using the Hy-MKr-RFE algorithm. Purple colour indicates the lesser significant factor among total factors selected for each signature sizes for building and content dataset.....	110
Figure 5-5: Feature importance for MKr flood loss estimation model for buildings (a) and contents (b)	112
Figure 5-6: Comparison of RMSE values of flood loss estimation models	115
Figure 5-7: Comparison of correlation coefficient for the flood loss estimation models (dd-f: - depth-damage functions, s-dd-f: - scaled depth-damage functions, MKr: - Multiple Kernel Regression Model)	115

Figure 5-8: Comparison of mean bias error (MBE) for the flood loss estimation models (dd-f: - depth-damage functions, s-dd-f: - scaled depth-damage functions, MKr: - Multiple Kernel Regression Model).	117
Figure 6-1: Sample markov chain.....	122
Figure 6-2: Illustration of Bayesian network with four binary variables.....	124
Figure 6-3:Naïve Bayes model - structure	127
Figure 6-4: TAN model - structure	127
Figure 6-5: Hybrid MTE-TAN regression model for content loss	129
Figure 6-6: Hybrid MTE-TAN regression model for building loss.....	130
Figure 6-7: Flowchart of the integrated data driven statistical framework.....	132
Figure 6-8: Prediction performance for each signature size of building and content loss dataset	136
Figure 6-9: Comparison of factor selection stability of the hybrid mixed data feature selection algorithm for the surveyed (original) and enhanced Building dataset....	138
Figure 6-10: Comparison of factor selection stability of the hybrid mixed data feature selection algorithm for the surveyed (original) and enhanced Content dataset	138
Figure 6-11: Composition of factors and its proportion among the top 7 and 10 selected factors by the 50 building (a) and content (b) loss prediction models using the propose Hy-MKr-RFE algorithm. Purple colour indicates the lesser significant factor among total factors selected for each signature sizes for building and content dataset.	140
Figure 6-12: Comparison of RMSE (a) and Pearson correlation coefficient (b) of flood loss estimation models (MKr: - Multiple Kernel regression model, En- MKr: - Enhanced Multiple Kernel regression model).	144
Figure 6-13: Effectiveness of mitigation measures in reducing content damage for the three income groups identified for Jakarta.....	147

List of Tables

Table 1-1: Costliest flood disasters	2
Table 2-1: Flood damage classification	11
Table 2-2: Comparison of relative and absolute damage functions.....	14
Table 2-3: Damage influencing factors.....	15
Table 2-4: Review of existing flood loss estimation models/methods	24
Table 2-5: Flood damage assessment scale	30
Table 4-1: Major floods in Jakarta.....	54
Table 4-2: List of Emergency measures and the number of household which has undertaken n(A)	58
Table 4-3: List of precautionary measures and the number of household which has undertaken n (A)	59
Table 4-4: Ranking scheme for the development of flood experience indicator.....	60
Table 4-5: Principal component loadings for variables that probably influence residential flood loss	66
Table 4-6: MI measure for building loss.....	67
Table 4-7: MI measure for content Loss.....	68
Table 4-8: Loss data counts in each group.....	70
Table 4-9: Coefficient of variation	71
Table 4-10: Ranks form Kruskal-Wallis Test.....	71
Table 4-11: Test statistic	71
Table 4-12: Factors considered in the residential flood loss estimation model for Jakarta	75
Table 4-13: Scaling factor for buildings and contents due to people's response measure to reduce damage	86
Table 5-1: Description of factors	93
Table 5-2: Sample case	97
Table 5-3: Variables influencing building loss	113
Table 5-4: Variables influencing content loss	114
Table 6-1: Minority factors and the domain to which they belong within the flood loss database.....	135

Table 6-2: Factors influencing building loss	142
Table 6-3: Factors influencing content loss	142

Symbols and Abbreviations

A	–	Absolute loss functions
BAPPENAS	–	Badan Perencanaan dan Pembangunan Nasional
BPBD	–	Badan Penanggulangan Bencana Daerah
BPS	–	Badan Pusat Statistik
C	–	Contamination
D	–	Water Depth (m)
Dur	–	Flood Duration
E	–	Empirical method
En-MK _r	–	Enhanced Multiple Kernel regression
FEMA	–	Federal Emergency Management Agency
FHRC	–	Flood Hazard Research Centre
FIA	–	Federal Insurance Administration
FIMA	–	Federal Insurance and Mitigation Administration
FLEMO _{cs}	–	Flood Loss Estimation Model for commercial sector
FLEMO _{ps}	–	Flood Loss Estimation Model for private sector
HAZUS- MH	–	Hazards U.S. Multi-Hazard
KI	–	Kuncheva Index
MCM	–	Multi Color Manual
MI	–	Mutual Information
MKL	–	Multiple Kernel Learning
MK _r	–	Multiple Kernel Regression
NFIP	–	National Flood Insurance Program
PCA	–	Principal Component Analysis
R	–	Relative loss functions

RFE	–	Recursive Feature Elimination
S	–	Synthetic method
SVC	–	Support Vector Classification
SVM	–	Support Vector Machines
SVR	–	Support Vector Regression
UNDRO	–	United Nations Disaster Relief Organisation
USACE	–	United States Army Corps of Engineers
X	–	Set of input vectors
Y	–	Output Vector
dd-f	–	Depth-damage functions/curves
n()	–	Number of elements
s-dd-f	–	Scaled Depth-damage functions/curves
$k(\cdot, \cdot)$	–	Kernel Function

CHAPTER 1. INTRODUCTION

1.1 Background and Justification

Natural and man-made disasters such as floods, earthquakes, terrorist attacks, industrial accidents etc. have impacted on human beings from time immemorial. They extort a heavy toll on society's vulnerability. The number of events and economic losses from the disaster are also on the rise. In the year 2014 alone 980 natural catastrophes have been reported worldwide with the total economic losses estimated to be around 110 billion USD (Munich Re 2015; Swiss Re 2015). Out of this, Asia alone accounted for 37% (Figure 1-1) of total events, 46% of global disaster damages and 75% of global victims.

Figure 1-1: Natural catastrophes 2014- World map

Among the natural hazards, flood being one of the most common hazards poses a serious threat to environment and humans. When the humans and environmental systems get influenced by the natural phenomenon such as flood, it becomes a risk. With a large percentage of the inhabited area located within the flood zones, makes flooding the biggest natural hazard in the world. Swiss Re in 2012 reported that the cost of floods has more than doubled in the last 10 years. The significant impact

from floods to society in terms of economic loss is evident in the world. Table 1-1 lists the costliest flood disaster for the period 1990 to 2013 (EM-DAT 2014). Still the risk from flooding is often underestimated.

Table 1-1: Costliest flood disasters (EM-DAT 2014)

Country	Date	Damage (000 US\$)
Thailand, Riverine flood	5-8-2011	40000000
China, Riverine flood	1-7-1998	30000000
China, Riverine flood	29-5-2010	18000000
Korea, Riverine flood	1-8-1995	15000000
Germany, Riverine flood	28-5-2013	12900000
China, Riverine flood	30-6-1996	12600000
USA, Riverine flood	24-6-1993	12000000
Germany	11-8-2002	11600000
USA, Riverine flood	9-6-2008	10000000
Pakistan, Flash flood	28-7-2010	9500000

However, with the advent of climate change and this immense impact from flood disasters, both the public and government have felt the need for appropriate flood management. Traditional flood management approaches commonly referred to as ‘Flood Control Approach’ refers to design standards and the structural flood defences measures such as dams, retention basins etc. as the dominant flood management approaches to control a predetermined flood. This approach of the older times was probed in depth leading to the development of the new methodology usually referred to as flood risk management. This new approach has had a paradigm shift from focusing on flood hazard to flood risk where the emphasis is given on the impacts from flooding (Merz et al. 2010). So it has

become necessary to take in to account the damage aspect whenever considering flood risk management.

Though the application of damage assessment within policy making is relatively new, it provides crucial information to aid governments and agencies in formulating better decision for flood risk management (Messer et al., 2007). The impacts from flooding could be minimized by taking appropriate actions prior to and after the events. Comprehensive assessment of flood damage/losses and the factors leading to it is the key for it. This reliable information could help communities in mitigating and reducing the impact of the events in future. Flood loss estimation and prediction has now become a key tool in formulating food control policies.

Hence there is a need to develop a comprehensive flood loss estimation model which not only helps in assessing the impacts from floods but also helps in the decision making process of flood risk management. Furthermore, they are useful for

- Assessment of vulnerability and risk mapping
- Carrying out Cost-benefit analysis to determine the optimal flood mitigation measures and
- Comparative risk analysis in regions prone to multiple risks.

In the past lot of importance and attention is given to the hazard assessment part while little or negligent effort is put in for the assessment of damage or losses. Thus the damage estimation methods are very crude and the information available on it is very less compared with the abundance of techniques and knowledge available on flood hazard (Merz et al. 2010). An insight in to the available damage estimation methodologies portrays the dependencies of many assumptions and pitfalls in economic evaluation. Most of the flood loss estimation models derive the loss function based on the damage data obtained for the specific building type or building use and the depth of water (Smith 1994; Wind et al. 1999) there by ignoring the other damage driving parameters. Omission of these parameters leads to model uncertainty in flood loss estimation resulting in inferior damage prediction/estimation especially when transferring in space. Thieken et al. (2008) illustrated this by applying FLEMOps, a model developed for Elbe and Danube

catchment, to the five municipalities in Baden-Württemberg in Germany. The mean relative error reported in this case was more than 1000%. Apart from this, another study which analysed the flood loss data from river Meuse and Cologne in 1995(Wind et al. 1999) showed reduction in flood loss compared to previous estimates. This has been credited to the improved capability and awareness of people to respond against flood. These observations show the importance of different factors in modelling flood loss. Therefore there is a need for an effective flood loss estimation model which takes into account the single and joint interaction of various damage influencing factors. It is also imperative to identify the prominent loss influencing factors so as to make the model usable in the physical world.

Indonesia's capital city Jakarta is long known for its risk from flooding (Budiyono et al. 2014; Marfai et al. 2015). The flooding problem in Jakarta has posted an existential threat to the city for many decades. The 13 rivers which congregate in the low lying urban area, rising sea water levels under climate change, land subsidence, rapid growth in population and the resultant urbanization has exacerbated the flooding situation in Jakarta from bad to worse (Nasir 2008; Julian et al. 2014; Marfai et al. 2015). As a result annual flooding during the wet season is a common sight in Jakarta. Flooding from both rivers and sea have resulted damages worth billions of dollars to property and infrastructure. As a result, they are currently the biggest climate change issue in Jakarta.

Jakarta experienced devastating floods in 2002, 2007 and 2013 (Julian et al. 2014). The 2007 flood event inundated an area of 454.8km² in Jakarta with over 80 people dead and caused approximately 5.2 trillion IDR (2007 values) in direct loss (BAPPENAS -Badan Perencanaan dan Pembangunan Nasional, 2007). Whereas the 2013 flood killed 20 people and caused 7.3 trillion IDR of direct damage (BPBD - Badan Penanggulangan Bencana Daerah, 2013). In spite of having a prolonged history of flooding, only two micro-scale flood loss estimation model has been developed for the low income communities living along the banks of river (Marschiavelli 2008; Wijayanti et al. 2015), As a result there is a need for developing a comprehensive flood loss estimation model for Jakarta. This drawback

has also been reported by Ward et al. (2011) In addition, there is also a need for developing and identifying efficient indicators for classifying the elements at risk in developing economies.

The current research work aims to address the broad range of concerns and issues of the previous approaches by developing a new method for multi-variate residential flood loss estimation; which will be described in later parts. The methodology developed will be used in identifying

- Indicators for classifying residential loss data in to homogenous groups for developing baseline flood loss functions
- the prominent factors influencing residential flood loss for Jakarta and

1.2 Objectives

The primary objective of this thesis, characterising a novel contribution to research, are

1. To demonstrate that notable enhancements should be made in the approaches which are used for developing flood loss/damage functions.
2. Develop a methodology based on a combination of statistical approaches for categorising the elements at risk into homogenous groups of flood damage and loss thereby eliminating subjectivity. This is accomplished using a combination of multivariate statistical analyses applied to multiple flood parameters, household characteristics, building characteristics and precautionary and emergency indicators.
3. Development of an enhanced mixed data feature selection algorithm for identifying the prominent factors influencing flood loss. The proposed algorithm is based on an enhanced feature weighing index and a modified search procedure incorporated inside the recursive feature elimination approach using multiple kernel regression. Our algorithm identified the prominent flood loss influencing factors by capturing its ‘relevance’, ‘importance’ and ‘interaction’.

4. Define a methodology for the development of multi-variate flood loss estimation model by capturing the single and joint interaction of the significant factors affecting flood loss in a data deficient environment and enhancing feature selection stability. This is achieved by combining the proposed feature selection algorithm with Gibbs sampler.

The methodology and framework developed as part of this research should provide support and direction to meet the following secondary objectives:

1. To illustrate the above objective for riverine residential flood loss estimation in Jakarta, Indonesia
2. To conduct extensive survey for the collection of damage and loss data which are vital for assessing the flood impacts
3. Investigate the existing approaches used for estimating flood loss and develop a riverine flood loss estimation model for residential sector in Jakarta using them.
4. Identify the significant factors which influence building and content loss from flooding in Jakarta
5. Develop a multi-variate residential flood loss estimation model by capturing the single and joint interaction of factors using kernel methods.
6. Identify a global representation of flood loss influencing factors in a data deficient environment.

The final loss estimation model thus developed will assist the decision makers, insurers and the public to facilitate appropriate mitigation measures, formulate flood control policies and plan and strategize according to the available resources.

1.3 Organisation of the Research

The structure of this dissertation is outlined as following:

Chapter 2 summarizes the review on existing methodologies and the relevant literature on a flood loss estimation model.

An overview of the multi-variate statistics and the background of the concept of kernel methods used for the development of multi-variate residential flood loss estimation model and performance evaluation metric used in the subsequent chapters are presented in **Chapter 3**.

Chapter 4 presents the standardized approach for demarcating flood loss data into homogenous groups. The proposed approach has been applied to the flood loss data from the residential sector of Jakarta and a baseline flood loss estimation model has been developed for these groups by the application of existing approaches. The database is obtained as result of extensive household survey conducted after the floods of 2013. This chapter also presents a basic multi-variate residential flood loss estimation model developed by employing typically used techniques.

Chapter 5 proposes a novel approach for identifying the prominent factors which influencing flood loss for both buildings and contents. The proposed approach combines the importance of each factors and interaction among factors to identify the significant factors among the available heterogeneous (Categorical and continuous) factors. Finally a comprehensive multi-variate residential flood loss estimation model is developed using multiple kernel regression (MKr) from the identified factors.

In **Chapter 6**, the complete framework for developing a multi-variate residential flood loss estimation model under a limited data environment is presented. The framework starts with (i) generating synthetic data for the underrepresented factors using Bayesian networks and Gibbs sampler, (ii) identification of a global representation of prominent flood loss influencing factors and (iii) development of multi-variate flood loss estimation model using MKr. The proposed framework is applied to the flood loss data from Jakarta.

Chapter 7 draws the conclusion of the research work. The chapter summarises the important findings, limitations and suggest possible directions for future research

CHAPTER 2. REVIEW OF DIRECT FLOOD LOSS ESTIMATION

2.1 Introduction

Flood actions which cause damage can be either physical or nonphysical. The mechanism which causes damages are hydrostatic actions, hydrodynamic actions, erosion actions, buoyancy actions, debris impact actions, chemical actions, biological actions and nuclear actions (Kelman et al. 2004). This chapter reviews the existing methodologies which have been used internationally to assess the direct impacts of flooding and also various flood actions and damage influencing factors considered in developing a flood loss model. The existing flood damage and loss assessment techniques are explored in details to develop a detailed multi-variate methodology in assessing the damages and losses caused by riverine floods.

2.2 Definitions

2.2.1 Hazard

According to Oxford English Dictionary (OED 2012) the noun hazard refers to a danger or risk or a potential source of danger. Relating it within the context of this report, the hazard here refer to the peril (flood) quantified by selecting parameters including water depth, velocity and a combination of both which have the potential to cause harm or damage to the fabric of society.

2.2.2 Vulnerability

It is important to define the scope of vulnerability before assessing existing methodologies for flood loss estimation.

United Nations Disaster Relief Organisation (UNDRO 1982) defines vulnerability as the extent of damages or loss to elements at risk which occur as a result of any hazard.

Capobianco et al. (1999) expressed vulnerability as a function of resistance, resilience and susceptibility.

For elements at risk, its vulnerability determines the extent of damage caused by any hazard. A highly vulnerable element is one if it is disproportionately exposed to any hazard i.e. it is highly susceptible to the flood actions. An interaction between these indicators needs to be established (Figure 2-1) to carry out any flood vulnerability analysis (Messner et al. 2005). In context to this report, the vulnerability of residential buildings and its components and contents is defined by taking into account

- The ability of the house to withstand failure (i.e. resistance)
- The material properties (i.e. resilience)
- Location and elevation of house (i.e. exposure) and
- Flood loading characteristics especially water depth, contamination, velocity and duration

The flood vulnerability of buildings and any of its components and contents is typically represented by vulnerability functions or loss functions or depth – damage curves. The vulnerability of objects is usually determined based on observations, modelling, experimentation or a combination. Detailed description of vulnerability functions appear in the following sections.

Figure 2-1: Flood vulnerability indicators

2.2.3 Damage and loss

Anything which suffered a reduction of utility or functions or purpose or integrity is considered as damaged. In almost all the events, the term damage has been used to reflect the degree of physical damage incurred without mentioning the economic loss. In the current study of flood damage and loss to residential buildings and its contents, the term damage refers to the direct consequences expressed as the reduction in the usefulness or the value of any component under study. Loss is defined as the cost incurred to restore the component to its previous state.

2.3 Urban Flooding

With the conversion of watershed with progressively impervious surfaces, exploitation of flood plains by communities due to population explosion, threats to coastal settlements due to continued rise in sea water level, increased migration to cities due to economic opportunities, unplanned development and climate change related problems have led to significant increase in the occurrences of urban flooding (Hsu et al. 2000; Dutta et al. 2003; Brown et al. 2007; Wilby 2007; Kawahara et al. 2008). The population density and economic value in urban areas also magnify the consequences of flooding. The flood frequency and the damages from urban floods are showing an upward trend in the last few decades (Ashley et al. 2005). The complex ways by which the urban landscape is being changed by the humans makes it highly vulnerable to floods. Thus there is a need for damage assessment to manage the risk of flooding in urban areas.

Generally urban floods can be fluvial, pluvial, coastal or combination of any of these. When there is a high intensity rainfall event the urban drainage becomes saturated and overwhelms the surface watercourses. As a result pluvial flooding occurs. The overflow of excess water from the stream banks as the capacity of the natural or constructed channel is exceeded when accommodating the surface runoff causes fluvial or riverine floods. Accidental situations such a levee breach, dam break can also cause riverine and flash floods. Unusual high tides, severe weather breaching sea defences, tsunamis and coastal surge can also lead to coastal flooding.

The focus of this report is on riverine/fluvial flooding on Asian mega cities (urban areas).

2.4 Assessment of Flood Damage

2.4.1 Flood damage typology

Basically the flood damages can be categorized into two: direct and indirect damages as shown in Table 2-1 (Grigg et al. 1975; Green et al. 1983; Smith 1994; Dutta et al. 2003).

- Direct damages: These include all damages caused by the physical contact of flood waters with humans, buildings, property or others.
- Indirect Damages: These refer to the economic impacts induced by direct damages and occur over time and space by means of disruption of social and economic activities.

They can further be categorised into tangible and intangible damages

- Tangible Damages: Damages to assets and resource flows which can be evaluated quantitatively in monetary terms
- Intangible Damages: Asset damages which are arduous to be expressed in financial terms

Table 2-1: Flood damage classification

Loss Typology	Measurement	
	Tangible	Intangible
Direct	Content and Building Damage	Loss of an archaeological site.
Indirect	Production loss of industries	Hindrances experienced during post-flood disaster recovery

2.4.2 Levels of assessment

Selection of appropriate temporal and spatial scales has a significant impact in the outcomes of the flood loss assessment. The different spatial scales on which the flood damage assessment performed are shown below.

- **Macro-scale analysis**

The flood damages are estimated at larger administrative units typically municipalities, regions and countries. The distribution of damages is assumed to be uniform throughout the area. Data which are generally obtained from the official statistics are used in calculating damages which has to be disaggregated or aggregated depending on the scale of study. Macro approach can be used to provide an overall picture of loss or damage from floods because of the low level of accuracy. The impact of sea water rise over a vast geographical area can be estimated using macro approach. Flood mitigation strategies for a national scale are framed here.

- **Meso-scale analysis**

In meso-scale analysis, aggregated data is used to obtain the damage potential. For better assessment of loss or damage, the data are more realistically localized according to land use categories such as residential, commercial, industrial etc. The distributions of damages are not uniform unlike the macro approach. Flood losses are calculated using the relative depth-damage functions. Flood mitigation strategies on a large scale are framed with this approach.

- **Micro-scale analysis**

The damage potential and the expected damages are evaluated based on the single elements at risk. This approach requires detailed information regarding the different building types, occupancy class, individual construction cost and other delineation parameters (Penning-Rowsell et al. 1977). The damages are calculated using relative depth-damage functions or absolute depth-damage functions. Cost benefit analysis of single flood protection measure at local level can be conducted using this approach

2.4.3 Depth – damage curves

Gilbert F. White in the year 1945 proposed the concept of depth – damage curves which is used to represent the relationship between the damage suffered by the element at risk and the depth of flooding. This is the fundamental building block upon which the direct flood damage assessments are based (Smith 1994). Since then damage functions have become the predominant idea in flood loss estimation. It represents the average damage at different inundation depth while considering building characteristics such as building type, building material and occupancy.

Depth – Damage curves can be developed using two main approaches, the empirical approach and the synthetic approach. Empirical approach conducts extensive survey to gather data in the after-math of a flood event and derives the relationship of flood damage or losses with flood depth. The data collected contains real damage information, thus there is greater accuracy in the loss function and better manifestation of inconsistencies within each category and the corresponding water depth (Gissing et al. 2004; Merz et al. 2004). However the scarcity of information regarding floods of different magnitude and damage details at higher depth requires extrapolation.

After the flooding, trauma faced by the families makes it difficult to employ extensive field surveys to get the damage and loss data. Delaying the loss survey may result in reduction of quality of data obtained as people will start forgetting their loss. It is in this situation, the synthetic methods come in handy. The synthetic method uses the damage data gather by what-if (hypothetical) queries. Compared with empirical approach, the exhaustive damage information can be retrieved easily (Penning-Rowsell et al. 1977). Loss functions can be developed easily in any areas with a higher degree of standardization as there is no requirement of any actual flood event (Smith 1994). However this process is cumbersome involving detailed and large surveys in order to attain adequate data for each category. The method also fails to incorporate the impact of mitigation actions and the disparity of buildings within a category.

The Australian Model Anuflood (Dutta et al. 2003), FLEMOps (Thieken et al. 2008) and FLEMOcs (Kreibich et al. 2010) from Germany are examples of first approach whereas UK's MCM model is the example for the latter. It is also possible to combine both the approaches as was done by USACE and USA's HAZUS-MH loss/damage estimation models.

The depth – damage curves can be either relative or absolute. The relative damage functions represent the loss in terms of a percentage of total replacement value. The value of the asset will be multiplied to the loss ratio to get the absolute loss. The absolute damage functions represent the losses directly in monetary units. The advantages and disadvantages of both the functions are shown in Table 2-2

Table 2-2: Comparison of relative and absolute damage functions

	Advantages	Disadvantages
Relative loss functions	<ul style="list-style-type: none">• Simple to derive as the asset values are available from public data sources• They are independent of market fluctuations	<ul style="list-style-type: none">• Asset value estimation can add in uncertainty
Absolute loss functions	<ul style="list-style-type: none">• Monetary Loss is computed directly from the value of the damaged object	<ul style="list-style-type: none">• Calibration of loss data has to be done at regular intervals

2.4.4 Factors influencing flood damages or losses

Direct tangible economic damage caused by floods depends on a lot of factors. These damage influencing parameters can be differentiated into two (Thieken et al. 2005).

- Impact or Loading or parameters - It reflects the explicit characteristics of the flood event for the elements at risk. The magnitude and nature of flood define the loading parameters.
- Resistance Parameters – The competence or incompetence of the element to oppose the onset of flood is being reflected by the resistance parameters. It depends on the attributes of the component at risk.

Table 2-3 compiles the factors influencing damage by floods.

Thus $Damage = f(Impact\ parameters, Resistance\ parameters)$

Table 2-3: Damage influencing factors (Merz et al. 2010)

Impact\Loading Parameters	Resistance Parameters
Water depth	Use of building or Line of Business
Velocity of flood water	Building Material
Rise rate of flooding	Building type
Duration of inundation	Precaution
Presence of Debris	Early Warning
Contamination	Emergency measures
Time of occurrence of flood	
Frequency of inundation	

These factors can further be classified according to their functional characteristics.

2.4.4.1 Flood characteristics

All the variables which belong to impact or loading parameters fall under flood characteristics. These variables include water depth, velocity, rising rate of flooding, flood duration, debris and sedimentation, contamination, flood frequency, time of flooding and frequency of inundation. Among these factors, only flood water depth has received significant amount of attention while estimating losses from flooding (Merz et al. 2004). Importance of the remaining factors in inducing flood damage and explaining uncertainly surrounding depth-damage curves were highlighted by many (Smith 1994). However, the effects of these factors are scantily studied and rarely used in developing flood loss estimation models (Vogel et al, 2012).

2.4.4.2 Building characteristics

The factors representing the characteristics of building which plays a pivotal role in determining flood damage are building type or use, construction type, building age, building materials, building value, content value, building and content quality, building area and location of building. These factors belong to resistance parameters as it opposes the onset of floods and reduces the extent of damage. Current standard flood loss estimation model uses, flood water depth and building type while developing depth-damage curves (Kelman et al. 2004, Merz et al. 2010). Apart from few studies conducted in Germany (Kreibich et al. 2011 Merz et al. 2013, Vogel et al. 2012), no other studies have quantified the effects and impacts of these factors while modelling flood loss.

2.4.4.3 Emergency and precautionary measures

Studies have provided qualitative hints on the effectiveness of the emergency and precautionary measures combined with early warning mechanisms in reducing flood damage. Apart from the study conducted by Vogel et al. (2012), there has been no other comprehensive approach developed so far in capturing the functioning of these measures while estimating flood losses. Emergency and precautionary measures vary regionally and have to be identified by literature reviews, questionnaire surveys and expert interviews (Thieken et al. 2005).

The various emergency measures identified from literature are ,documents and valuables stored more securely, furniture and movable objects brought to safety , using water barriers , electrical appliances turned off, switching off gas, sealing the openings in the buildings and vehicles driven to safety.

The various precautionary measures followed in different parts of the world includes, avoiding construction of buildings with basements in flood plains, construction of buildings with flood resistant materials, using water membranes to prevent infiltration, using low value materials and contents, water proofing of buildings, taking flood insurance and construction of dykes, dams or other barriers.

2.4.4.4 Socio-economic variables

The various socio-economic variables which has been used to model flood losses are age, size of household, ownership structure, household income and socio-economic status. Incorporation of socio-economic factors in modelling flood loss is relatively new. Except for the studies conducted by Merz et al. (2010), Vogel et al. (2012) and Chinh et al. (2015), no other studies have incorporated and evaluated the effect of these variables in defining the nature of damage.

Though various factors have been identified or proposed to affect flood loss, Kelman et al (2004) and Merz et al. (2013) still confirmed the need to comprehensively understand the mechanism leading to damage and loss from floods.

2.4.5 Past studies

2.4.5.1 United States

White in 1945 developed the path breaking concept of using depth – damage curves to assess the damages caused by floods. Due to inadequacies in the historic data on events of different magnitude, the study proposed the concept of using synthetic loss functions in developing damage curve (White 1964). Loss functions were developed for residential and commercial sector for buildings and their contents based on the repair and replacement cost by taking into accounts the effect of flood inundation. In the year 1965, Kates in his study of Lehigh valley made some

enhancements to the synthetic damage estimation process. The study undertook a detailed synthetic survey which included

- The assessment of critical damage level of inventory and equipment
- Height of the openings present in the buildings and
- Elevation of these openings at several points inside buildings.

Based on the enumerated details, inflection points were determined and depth – damage curves were developed. The framework was further expanded to include loss functions for inventory and equipment (Davis 1985). Similar approach is being used by Corp’s district and Federal Insurance and Mitigation Administration (FIMA) in developing the loss functions (Davis et al. 1992).

FIMA formerly known as Federal Insurance Administration (FIA) which is a part of Federal Emergency Management Agency (FEMA) is accountable for the flood coverage in the United States through the National Flood Insurance Program (NFIP) (Davis et al. 1992). Nationwide depth – damage relationships known as FIA depth – damage curves were computed by FIMA to determine the flood insurance premiums. They express flood damage cost to buildings and their contents separately. Using the information from Corps, the first FIA depth – damage curves were compiled in the year 1970 which were later adjusted based on actual claims information in the year 1973. These 1973 curves, according to FEMA (FEMA 2003b) were known as theoretical base tables. Based on the flood insurance rate review process and actual data, these curves are updated annually. The resulting curves are the modified FIA depth – damage curves which are used by flood plain managers, emergency managers etc. to address the flood risk in areas under their purview.

However due to the limitation of FIA insurance coverage, the items which are not covered under their policies especially basement flooring and finishes aren’t reflected in the damage curves (FEMA 2003b). On the contrary the generic depth – damage functions established by the U.S Army Corps of Engineers (USACE) represent a substantive improvement over the functions developed by the FIA (USACE 2003). Based on the damage data collected between 1961 and 2001,

USACE developed the depth – damage curves for building with (USACE 2003) and without basements (USACE 2000). By including basement flooding characteristics, it can be concluded that USACE depth-damage curves are more representative of the real time flood damages. The models were built by considering multiple factors such as lead time of warning and flood duration. The damage functions were then created using general regression analysis. USACE in the Engineering Manual No.1110-2-1619 outlines the methodology used for the development of flood loss curves for building structure and contents (USACE 1996).

In the aforementioned procedures, a major source of uncertainty in the development of loss functions can be attributable to the fact that water depth is the only damage influencing parameter considered. According to USACE (USACE 1988a), the critical damage influencing parameters described in section 2.4.4 were not accounted while developing the depth – damage functions. USACE (1996) also stated that the content and structure damage would worsen if the impact of flood water velocity and duration is taken in to account (USACE 1996). Several studies in have USA tried to address the relationship between velocity and damage. Richard Black (Black 1975) in his report ‘Project Agnes’ developed the Black curves which presented the critical water velocity and depth combination required to initiate the movement of a house based on buoyancy, self-weight, hydrostatic pressure and dynamic pressure. The composite effect of hydrodynamic and buoyancy forces on frictional resistance were studied by (Sangrey et al. 1975) for developing empirical building collapse potential functions for multiple occupancies i.e. structure destroyed or survived.

The most comprehensive flood damage assessment studies conducted by FEMA contributed to the development of loss estimation software called HAZUS-MH. It is used extensively to estimate potential losses from natural disasters. HAZUS-MH Flood Model produces loss estimates for riverine floods, coastal floods and both. It quantifies the financial, property and social impacts due to floods under existing scenarios or after any mitigation measures that were taken. HAZUS –MH uses the depth – damage curves developed by FIA and USACE and integrates the building collapse curve developed by the Portland Corps. But the model does not consider

the hydrodynamic forces from the flood water velocity in determining the damage. The velocity – depth – damage curves are used only to relate the velocity and depth to the collapse potential, which varies from very low to total collapse. If the critical combination of flood water velocity and depth is unattained, then the damage is estimated based only on inundation scenarios. Otherwise the building is assumed to have collapsed and the damage is estimated at 100%. The possible damage to buildings is estimated from the flood water depths and the computed depth – damage functions based on

- Building occupancy
- First floor elevation
- Number of stories
- Provision of basement
- Foundation type and
- Age of building

The direct damage to the buildings is expressed in proportion to the replacement cost at the census block. Finally the damage percent is multiplied with the replacement value based on the specific occupancy class to produce the total expected loss in monetary units as required.

2.4.5.2 England

In England the damage data developed by the Flood Hazard Research Centre (FHRC) at Middlesex University is currently used for estimating flood damages. The loss function calculates the pecuniary damage as a function of the flood depth. FHRC's Multi Coloured Manual (MCM) published in the year 2005 estimates the benefits from alleviating flood for both agricultural lands and urban areas (Penning-Rowsell et al. 2010). The first major publication from FHRC, the 'Blue Manual' assesses the direct damage to entire built environment of the UK using the synthesized data from floods (Penning-Rowsell et al. 1977). For residential flood damage assessment, the susceptibility of 203 household inventory items and 77 building fabric repair works at different depths of flooding were established. It presents the absolute loss functions by:

- Specific building types
5 Types which are detached, semi-detached, terrace, bungalow, and flat
Social class
- 4 types which are AB, C1, C2, DE
Year built (Age)
4 types which are pre-war (1918), interwar (1918-1938), post war 1939-1965, and modern (1965)
- Duration of flooding
Short duration: - < 12 hours and long duration: - > 12 hours
- Flood depths
10 to 17 water depths
- “Pre-fab” housing type.

Depth-damage curves were also developed by considering the impacts of flood warnings (warning scenario: 0.5 h, 2 h, and 4 h). For all the loss curves, interpolation between water depth values is assumed.

Further flood damage estimation studies in UK used this work as the prime guiding work. The depth – damage curves were later modified by considering the cleanup costs (Suleman et al. 1988). Land use classifications for calculating depth-damage curves were further revised by N’Jai et al (1990). This work was ultimately modified by Experian (2000), who developed the depth-damage curves based on the lifestyles rather than the established norm of building types. 52 lifestyle scenarios were framed using 86 housing and demographic variables for UK. These were clustered in 12 lifestyle groups and separate tables were provided for both the flood durations.

In conclusion, the studies by FHRC do not analyse the detailed effects of damage influencing parameters other than water depth. Penning-Rowsell (1981) stated that the extent of damage(s) produced by flood water velocities could be assumed miniscule except in the cases of building structure failure and depth of flooding can alone be used for measuring the minor effects. (Cited in Kelman et al. 2004). Later from the research conducted by Clausen (1989), Penning-Rowsell briefly indicated that the effect of velocity is negligible on flood damage (Penning-Rowsell 1992).

Clausen proposed a Dale Dyke dam failure study which identified water depth, flow velocity, rate of flood rise, bed shear stress, Probability of debris and dynamic forces as the most important hazard parameters for estimating flood damage. From the empirical data on the Dale Dyke dam failure in Sheffield, the norm proposed by Clausen in estimating flood damage for brick and masonry building is:

- Damage from Inundation occurs at: - Velocity < 2 m/s or
Depth $\times$ Velocity < 3 m²/s
- Partial damage occurs at: - Velocity > 2 m/s and
 3 m²/s $<$ Depth $\times$ Velocity < 7 m²/s
- Total Damage occurs at: - Velocity > 2 m/s and
Depth $\times$ Velocity > 7 m²/s

Clausen concluded that the velocity criterion proposed in the research study was intended as an initial guideline when more extensive results are lacking.

2.4.5.3 Jakarta

The flooding problem in Jakarta has posed an existential threat to the city for many decades. It has been the topic of many research studies which have acknowledged the magnitude of damage Jakarta faces. Floods of 2002, 2007 and 2013 have inflicted rampant damage. In spite of all these there exist only two flood damage assessment models for Jakarta and that too for assessing the damage to buildings located along the banks of the river. The first study was done by Marschiavelli (2008) who developed vulnerability functions for 5 different building types by conducting a household survey in Kampung Melayu, a village located in Jatinegara District, East Jakarta. The vulnerability functions assess the damages to building components such as walls, roofs and floors qualitatively in a scale ranging from 1 to 5. The second study developed depth-damage curves by conducting questionnaire survey in residential and business units along the Pessangrahan River. Depth-damage curves were developed for 3 different monthly income groups (low - < 226 USD, medium - $226 - 514$ USD and high > 514 USD) and two different business groups

(Wijayanti et al. 2015). However, this model has shown to be highly underestimating the flood loss when applied to 2013 floods. This has been attributed to the fact that the survey hasn't covered the entire spectrum of income and business groups and other significant flood damage parameters while developing the depth-damage curves. These studies thus highlight the need for the development of a comprehensive flood loss estimation model for Jakarta.

2.4.5.4 Other studies

Table 2-4 provides the various studies which analyse and estimate the flood damage and losses. The literature is reviewed providing detailed information, including hazard parameter (Depth, velocity, surge, contamination etc.), flood type (Fresh water, river, sea), type of damage functions (relative, absolute or collapse potential), development (empirical or synthetic), construction material type (masonry, wood, concrete, steel, brick or other) and scale (micro, meso or macro).

Table 2-4: Review of existing flood loss estimation models/methods

Study	Country	Scale	Hazard	Development	Functions	Construction Material	Remarks
(Dutta et al. 2003)	Japan	Micro	D	E	R	Wood RCC	Developed a mathematical model to assess the flood damages from published data. Grid based loss estimation model with a uniform network of square grids which are similar to the hydrologic model is employed. Depth-damage curves were developed for both structure and its contents.
Model of ICPR	Germany	Meso	D	E & S	R	N.A	Depth-damage curves for buildings, immobile and equipment were developed from the flood loss data base of Germany- HOWAS. Flood damage is calculated by using 5 loss functions.
Model of MURL	Germany	Meso	D	E	R	N.A	Loss functions were established for residential and commercial sectors for buildings, contents and inventory.
Model of Hydrotec	Germany	Meso	D	E	R	N.A	Damage data from the HOWAS 21 data base was used to develop the loss functions for residential properties. Losses were represented in a single figure including building structure loss and content loss.
FLEMOps (Thieken et al. 2008)	Germany	Meso – Micro	D C	E	R	N.A	Multi-factorial residential flood loss estimation model was developed from the large dataset of flooding in August 2002 in Elbe and Danube catchment. Depth-damage functions were developed for building and its contents. Apart from water depth, FLEMOps considered 3 building types and 2 quality classes while developing the loss functions. Scaling factors were the derived for the 3 classes of contamination and the 3 precaution

Review of direct flood loss estimation

							classes in FLEMOps ⁺ model
FLEMOcs (Kreibich et al. 2010)	Germany	Meso – Micro	D C P	E	R	N.A	The commercial loss assessment model sector estimated the damage in two stages to the buildings, products and stocks its goods and equipment of commercial establishments. In first stage, loss is estimated by considering the type of commercial sector the company belongs, water depth and size of company and in the second stage, the above obtained loss value will be multiplied with scaling factors by considering the effect of precautionary measures employed and contamination from flooding.
Kang et al (Kang et al. 2005)	Taiwan	Meso	D	S	A	N.A	Stage-damage curves were developed for single and multiple family dwelling based on the socio-economic data. Spatial characteristic of the damage were captured using spatial data which was then used to disaggregate the aggregated social data. Model households were built based on the socio-economic data and the appliances whose possession rate above 40% were included as contents in the model house. Synthetic assessments were based on repair cost data collected from interviewing contractors, carpenter and workers.
Su et al	Taiwan	Meso	D	E	A	N.A	Based on the damage data obtained from the tax reduction fillings after the flood event of the Nari typhoon, depth-damage curves for industrial and commercial sectors were developed. These sectors were further dived into manufacturing, retail trade, wholesale trade and services. They were also grouped into small scale and large scale. Loss functions were then developed for each category and subgroup.
(Dale K. et al. 2004)	Australia		D V		C P	Brick Fibre Wood	The critical combination of water depth and velocity at which the residential structure collapses were derived based on the methodology used by Richard Black (1975) in US. For the various combinations of water depth and velocity, collapse curves were developed by assessing hydrodynamic forces, buoyancy and self-weight for each structure.

Review of direct flood loss estimation

CH2M Hill (1974)	USA	N.A	D V		D	N.A	Depth-damage curves were developed and losses were expressed in monetary terms. The Collapse potential was established for the critical combination of velocity and depth.
JICA	Thailand	Micro	D Dur	E	A	N.A	Loss functions were developed based on the surveyed data. Flood losses for residential and industrial areas exhibited a linear relationship between flood depths and duration of flooding.
(McBean et al. 1988)	Canada	N.A	D	S	A	Masonry, Wood	Potential structural and content damages obtained from interview surveys were used to develop the depth-damage curves of each type of buildings. 28 different structural categories were surveyed in to 7 basic groups. Flood losses were then determined and found to vary a lot between structure types. Final flood losses were estimated using the mean damage curves from the 7 basic structure types.
(McBean et al. 1988)	Canada	N.A	D Dur V Ice	S	A	N.A	Adjustment factors were developed to reflect the effects of long duration of flooding, flood warning, high velocity and ice for the 7 depth-damage curves in Canada. The effects of household characteristics on flood damages were investigated by regression analysis. However the variability in flood damages across different housing categories were not established with reliability.
(Pistrika et al. 2010)	USA	Meso	D V	E	C P	N.A	The relationship between the economic damage and flood characteristics were analysed for the residential sector after hurricane Katrina in New Orleans. Statistical analysis in this study concluded that the direct relationship between velocity or flood depth or product of depth & velocity and percentage of damage were absent. Flood characteristics played a major role in the nature of damage in the studied polders. In order to ascertain the extent of flood damages due to levee failures, an alternate approach was proposed to distinguish the diverse zones of damage based on the product of depth & velocity which is similar to

Review of direct flood loss estimation

							Clausen's (1989) criterion. Damage function based on depth-velocity product was then derived
(Schwarz et al. 2008)	Germany	Micro, Meso, Macro	D V	E	R	Clay Pre-Fab Frame Masonry RCC	Using engineering evaluation of buildings which took into account the structural parameters (resistance parameter) along with water depth a new damage and loss prediction model was proposed. The study classified the damages into damage grades which links the flood impact and losses. Damage grades were then developed based on the procedure used for the developing of EMS 98. Using the data from August 2002 floods, 2005 flood in Bavaria and 2006 flood in Saxony and using the developed damage grade, vulnerability classes of structures were identified.
RiskScape Flood Model(Reese et al. 2010)	New Zealand	Macro – Meso	D V	E & S	R	Wood RCC Masonry	Fragility curves were developed for both structure and its contents. Structural damages were expressed in damage states with the corresponding damage ratios associated with each building type. Content damages were represented using a single fragility curve which can be applied to each storey separately. Impact of velocity was accounted by utilizing the approach detailed in other research studies (Black 1975; Dale K. et al. 2004). Depth-velocity product classified damage into damage caused by inundation, additional damage from velocity and finally total collapse.
Damage Scanner Model	Holland	Micro	D	S	R		Based on the flood water depth and information on land use, potential flood damages were calculated with seven depth – damage functions.

(D: Depth, V: Velocity, Dur: Duration, E: Empirical, S: Synthetic, R: Relative and A: Absolute)

2.4.5.5 Regression based flood loss estimation

Another approach which is widely used in modelling flood damage and loss is by making use of regression methods, more specifically ordinary least squares regression and geographic weighted regression. However, the use of regression analysis on depth-damage data has been cautioned by Davis et al. (1992) in his report “*Catalog of residential depth-damage functions used by the army corps of engineers in flood damage estimation.*” According to this report, the ability of regression equation in explaining the variance in damage caused by the multitude of variables is very low.

Regression based flood loss estimation model has been developed for Taiwan which incorporated factors such as water depth, building area, velocity and location (Chang et al. 2008). However the resulting model has been found to have a poor fit with R^2 equals to 0.15 and consistently under estimating the damage. Later on this model was improved by using geographic weighted regression technique, which incorporated geographical proximity to water bodies and spatial correlation of the loss data when developing flood loss functions. The new model's R^2 value increased from 0.15 to 0.21. However, the model continued to underestimate the losses (Chang et al. 2008).

Another model based on regression was developed by Zhai et al. (2005) for the private household in Japan. The model incorporated the following parameters such as inundation depth, income, building type and duration of stay. However, this model was neither verified nor compared with the existing flood loss models in Japan. Consequently its usability and performance is unknown.

Prettenhaler et al. (2010) using the recorded flood loss data from 2006 Austrian floods developed an absolute flood loss functions using regression methods. The form of the regression functions were chosen based on the previous findings from the literature for Austria. The developed flood loss function were compared with existing loss functions in Austria. The best model in terms of prediction accuracy was only found to be performing at par with the simple depth-damage curves (Prettenhaler et al. 2010).

The study conducted by Prettenhaler et al. (2010) highlights the necessity of selecting the correct functional form of the regression equation while developing a loss estimation model. The appropriate functional forms are identified based on the regional knowledge concerning the behaviour of factors influencing flood damage. The functional form has found to play an important role in improving the explanatory power of the model, its potential application, theoretical significance and also the prediction accuracy.

2.4.5.6 Alternate approach in damage assessment

Most of the studies mentioned in the above table use methodologies based on statistical analysis of claims data and aftermath surveys to estimate flood damages. These methodologies rarely analyse the various flood actions which could cause potential damage. Hydrostatics and hydrodynamic forces, energy transfers, impact from flood borne debris, wave actions (storm surge), buoyancy actions, erosion, water induced scouring, sediment transport and consequences of contamination which causes direct probable damage by directly imposing on a building is referred to as flood actions. Above mentioned methodologies also fail to make a distinction concerning the damage inflicted by contact of water or still water and other flood actions. In general the mechanism behind the damage caused from floods is rarely analysed.

The research conducted by Kelman on assessing the physical flood vulnerability in eastern and coastal England for residential buildings is among the more pertinent studies existing in the literature (Kelman 2003). The study made efforts to quantitatively and qualitatively portray risk from coastal flooding. The importance of flood velocity and flood depth differential on flood damage to residential buildings was examined by, therefore introducing relatively a novel and better approach for profiling vulnerability. The study evaluated the susceptibility of glass windows and masonry walls and developed a flood failure flowchart to aid develop the new form of vulnerability profile. Kelman's methodology evaluated the vulnerability and translated the results to damage scale as shown in Table 2-5. (Kelman 2003).

Table 2-5: Flood damage assessment scale (Kelman 2003).

Damage Scale (DS) Level	Interaction of water with building structure
DS0	No damage as no water touches with structure
DS1	External surface of the structure in contact with water
DS2	Minor damage as water infiltrates through small opening but no structural failure
DS3	Penetration of debris or water by damaging any covered or closed opening: for example, doors or windows
DS4	Structural failure occurs as debris or water enters by compromising the structural integrity
DS5	Total structural failure; irreparable; for example, undermining of foundation, collapse of roof or walls etc.

Another study which looked beyond the traditional approach of flood loss estimation based on depth – damage curves was performed in Netherlands by the Delft cluster (Roos et al. 2003). This model used two failure mechanisms namely the wall failure and scouring of foundation to quantify the damage to buildings. A load – resistance analysis with the strength of structure acting as the resistance and the four flood actions:

- Velocity of flood water
- Hydrostatic pressure resulting from the difference in water level between inside and outside of the building
- Loading from wave and
- Impact from debris,

as loads were used to assess the vulnerability of load bearing concrete and masonry buildings. The model which utilizes geographic information system incorporates the effect of location of buildings while assessing its vulnerability. Based on the two failure mechanisms, the model calculates the possibility of partial collapse. Buildings are assumed to collapse completely if its damages exceed 70%.

A stochastic model to determine the expected damage and direct impact of flood actions were proposed by Nadal in which the failure mechanism of individual building components were modelled to assess vulnerability (Nadal et al. 2010). Compared with the previous methods, this study analysed hydrodynamic and hydrostatics forces, wave actions, scouring of foundations and debris impact stochastically to evaluate the vulnerability of four components of buildings:

- RCC - frames
- Walls made from Concrete-Block
- Windows and Doors
- Utilities and Finishes

Detailed residential Building characteristics (geometry and structural) in Puerto Rico were collected from 28 buildings were synthetically expanded using Monte Carlo simulation to account for uncertainty and to perform load – resistance analysis. The expected flood damage computed per building unit is represented as a function of flood water velocity and depth in a three dimensional format as shown in Figure 2-2. Building vulnerability was further evaluated for two different loading cases for the following flooding events:

1. Riverine Floods
 - a. Slow rise floods
 - b. Flash floods
2. Storm Surges
 - a. Surging flood
 - b. Breaking waves
3. Tsunamis
 - a. Turbulent bores
 - b. Tsunami waves

Figure 2-2: Expected flood damage - slow rise flood – average

In general these methodologies analyse the failure mechanism of individual components by taking into account the effect of flood water velocity and flood water depths. However damage estimation by these models is only valid if the velocity is high enough to cause destruction to the built environment. Thus these approaches are mostly valid for dam break situation, extreme flash floods and coastal inundations.

2.4.6 Overall analysis of the past studies

In most of the existing models the damage is quantified based on water depth and the use or type of objects at risk (Wind et al. 1999; NRC 2000). However, the variation in damage influencing parameters brings in enormous variability in the experienced/observed damage for two similar houses when exposed to the same flood event. Due to the difficulty in predicting, modelling and attaining quantitative information on their effects, most of the damage influencing parameters as mentioned in Table 2-3 are generally ignored while modelling damage. Ignoring their role certainly induces significant reduction in the prediction accuracy of the loss estimation model especially when adapting the model in space. This is exemplified in the research by Thieken et al. (2008) where 1000% mean relative error was observed while applying the FLEMOps (A loss model developed for Elbe

and Danube catchment. refer Table 2-4) to the five municipalities in Baden-Wuerttemberg in Germany. Analysis of the flood damages in Germany from 1978 to 1994 by (Merz et al. in 2004) also confirm significantly poor performance of depth-damage functions when categorised according to the water depth and use of buildings. Exhaustive studies conducted by Merz et al. (2004); Pistrika et al. (2009); Thielen et al. (2009) also showed that the association between the flood water depth and the resulting damage is very feeble for smaller spatial scale. Substantial variation in damages is observed for limited sample size at higher level of detail.

To further strengthen the importance of various factors influencing flood damage and loss the following study conducted on loss data from the 1995 flood event in Cologne observed reduction in damages. Similar trends were also reported by (Wind et al. 1999) for the River Meuse catchment. Municipalities which experience frequent flooding suffer fewer damages per house when compared to the ones with less frequent flood events. This could be attributed to the improvement in the flood proofing measures, improved information, increased awareness and recuperative capabilities of the people. Thus it is necessary to account for the temporal changes in vulnerability as the system under study is never stationary. Despite the above mentioned facts the importance of non-depth flood characteristics and human measures in flood loss estimation has not been investigated thoroughly and systematically as the traditional depth – damage curves. This has introduced a gap in our knowledge in understanding and analysing flood risk.

Recently a flood loss estimation model developed for Germany using Bayesian Networks has incorporated a multitude of flood loss influencing factors. (Vogel et al. 2012). However the performance of this multi-variate model was just comparable with the existing flood loss estimation models for Germany. As a result it is imperative to investigate and identify the prominent factors that influence the flood loss thereby improving the prediction accuracy. This is also necessary to make the model suitable for real-world applications.

Though the approach of estimating damages based on traditional depth – damage curves is shown to be generally appropriate, a gap still exists in identifying and

quantifying the effect of the diverse damage influencing parameters. Omission of these, leads to model uncertainty in flood loss estimation resulting in inferior prediction when transferring in space. This unsuitability of transferring flood loss model from one region and time frame to another has been conclusively proven as demonstrated earlier. It is recommended that incorporation of critical damage influencing parameters and investigating local similarities in flood damages to residences by objective methods at progressively lower resolutions should be used in improving flood loss estimation models. Magnitude of damages normally varies from the interaction between various susceptibility indicators (building material, number of storeys, warning system, awareness, preparedness, emergency measures, etc.) and exposure indicators (flood frequency, duration, velocity, contamination, proximity to river, elevation, etc.). A thorough analysis on their effect on flood losses and damage to effectively determine the vulnerability of elements at risk has not yet been conducted. Among these factors, it is imperative to identify the critical factors which dominate especially in a developing economy, to map an understanding of this complex damage mechanism. Also there is a strong need for the development of a residential flood loss estimation model for Jakarta.

2.5 Conclusion

In general, riverine flood damage(s) and losses to residential sector need to be investigated in further detail. The prominent flood loss estimation models only take into account the inundation depth while defining flood losses. This simplified structure of depth-damage curves as discussed exhaustively in the sections above, is one of the most significant sources of uncertainty in flood loss estimation. In the recent years, approaches used in managing flood has seen a paradigm shift from flood control approaches to flood risk management, where emphasis is given on a holistic approach to reduce the impact of flood. Thus it is imperative to reflect the influence of these factors while developing a flood damage and loss estimation model. In order to make the model usable in real –life situations, it is imperative to identify the significant factors among the wealth of available variables whose single and joint interaction can effectively represent the nature and extent of flood loss in any area. To conduct a meaningful flood risk management, usage/development of

loss functions customized to the local situation is imperative. As no such functions have been developed for Jakarta, this study aims to fulfil the gap. This study also aims to infer more meaning to flood damage and losses and its multidimensional factors.

CHAPTER 3. METHODOLOGY

3.1 Introduction

This chapter elaborates the fundamental concepts of multiple kernel support vector regression involved in developing a multi-variate riverine flood loss estimation model. The basic concepts of principal component analysis, Mutual information and hierarchical clustering technique which are used for developing the baseline model for comparison are presented in Section 3.3.1, 3.3.2 and 3.3.3, respectively. Development of the baseline flood loss estimation model is elaborated in Chapter 4. Section 3.4 provides the performance measures used for model comparisons.

3.2 Multiple Kernel Regression

3.2.1 Introduction to kernel methods

Kernel methods (KM) are a family of algorithms which gives a principled and methodical approach in training learning machines to achieve a good generalization on data. They have shown promising results in extracting prominent features, prediction and classification on nonlinear and complicated data (Vapnik 1998). KM works on the concept of kernel functions which provides an implicit mapping of the original data to a high dimensional feature space where they are linearly separable, thus the name Kernel Methods. This process of mapping using kernel function is usually referred to as “kernel trick”. Kernel trick is a very powerful tool as it allows nonlinear formulation to be represented in terms of their inner product. So wherever a dot product is used, it can be replaced by Kernel function $k(\cdot, \cdot)$

$$k(x_a, x_b) = \phi(x_a) \cdot \phi(x_b), \quad \text{Equation 3-1}$$

thus avoiding explicit mapping into feature space.

Kernel function corresponds to a dot product on all pairs of inputs and measures some domain specific notion of similarity. Here ϕ is called the feature map as it maps the input features into some dot product high dimensional space called the feature space (Figure 3-1).

Figure 3-1: Kernel mapping

Kernel functions which are symmetric actually correspond to some feature space, i.e. $k(x_a, x_b) = \phi(x_a) \cdot \phi(x_b) = \phi(x_b) \cdot \phi(x_a) = k(x_b, x_a)$. Also kernel functions which satisfies Mercer's Theorem are said to be positive semi-definite.

Commonly used kernel functions which can be seen in literature are given below:

- Linear Kernel

This is the simplest kernel function obtained by adding an optional constant to the inner product of $\langle x_a, x_b \rangle$

$$k(x_a, x_b) = x^t x_b + c \quad \text{Equation 3-2}$$

- Polynomial Kernel

A non-stationary kernel widely used for non-linear mapping.

$$k(x_a, x_b) = (\alpha x_a^t x_b + c)^i \quad \text{Equation 3-3}$$

Where i is the polynomial degree, c is the constant and α is the slope.

- Gaussian Kernel

Gaussian Kernels are radial basis function (RBF) kernel represented by:

$$k(x_a, x_b) = \exp\left(\frac{||x_a - x_b||^2}{2\sigma^2}\right) \quad \text{Equation 3-4}$$

Alternate form of it is

$$k(x_a, x_b) = \exp(-\gamma ||x_a - x_b||^2) \quad \text{Equation 3-5}$$

- Exponential Kernel

This is another RBF which is similar to Gaussian kernel and can produce piecewise linear solution.

$$k(x_a, x_b) = \exp\left(\frac{||x_a - x_b||}{2\sigma^2}\right) \quad \text{Equation 3-6}$$

- Power Kernel

An un-rectified triangular kernel which is of scale-invariant

$$k(x_a, x_b) = -||x_a - x_b||^d \quad \text{Equation 3-7}$$

- Spline Kernel

A piece-wise cubic polynomial is used to represent this kernel

$$\begin{aligned} k(x_a, x_b) = & 1 + x_a \cdot x_b + xymin(x_a, x_b) \\ & - \frac{x + x_i}{2} \min(x_a, x_b)^2 \\ & + \frac{1}{3} \min(x_a, x_b)^3 \end{aligned} \quad \text{Equation 3-8}$$

- B-Spline Kernel

This kernel is defined in the closed interval [1,-1] and is given by

$$k(x_a, x_b) = B_{2p+1}(x_a - x_b), \quad \text{Equation 3-9}$$

where $p \in N$

- Additive Kernels

Summing kernels can give more complicated kernels as the summation of any number of positive definite functions are always positive definite.

$$k(x_a, x_b) = \sum_i K_i(x_a, x_b) \quad \text{Equation 3-10}$$

3.2.2 Support vector machines

Support vector machines (SVM) is one of the most well-known learning system based on kernel methods which embodies the principle of structural risk minimization (SRM) in obtaining a good generalization of learning patterns (Vapnik 1998). Though SVM has initially been designed in the domain of classification problem, it has been expanded to treat regression problems (Vapnik et al. 1997). The two categories of SVMs are support vector regression (SVR) and support vector classification (SVC).

Consider the following training dataset of N independent and identically distributed instances $\{(x_i, y_i)\}_{i=1}^N \subset \mathcal{X} \times \mathbb{R}$, where x_i denotes the D dimensional input vector. The output vector y_i can be either discrete or real valued depending classification and regression tasks respectively. In SVC, considering the binary classification case where each input point can take two labels $y = \{+1, -1\}$, it basically finds the hyper plane which maximises the margin or optimizes the empirical classification error. In SVR, it attains a good performance of learning patterns by finding a function $f(x)$ which has the most ε deviation from the actual targets y_i (Basak et al. 2007)). In this thesis, the problem at hand is the case of regression. So this section will focus on SVR and Multiple Kernel Learning for regression problems.

In SVR for the given training samples, the decision function which has to be learned is of the form

$$f(x) = w^T \phi(x) + b, \quad \text{Equation 3-11}$$

where $\phi(x)$ is the kernel feature map, w is the weight vector, and $b \in \mathbb{R}$, bias term.

The objective in SVR is to learn global optimum values of b and w from training data $\{(x_i, y_i)\}_{i=1}^N$. To ensure this, Euclidian norm has to be minimized. One way to achieve this is to write the problem as a convex optimization problem and solving it.

$$\min : \frac{1}{2} \|w\|^2$$

$$\text{subject to: } w^T \phi(x_i) + b - y_i \leq \varepsilon$$

$$y_i - w^T \phi(x_i) - b \leq \varepsilon$$

In addition, to cater for some errors and to handle the challenging constraints associated with the optimization problem, slack variables ξ_i^+ and ξ_i^- can be introduced. As a result the objective of SVR can be obtained by solving the quadratic optimization problem below.

$$\begin{aligned} & \text{minimize } \frac{1}{2} w^T w + C \sum_{i=1}^N \xi_i^+ + C \sum_{i=1}^N \xi_i^- && \text{Equation 3-12} \\ & \text{with respect to } w, \xi^+, \xi^-, b \\ & \text{s.t. } w^T \phi(x_i) + b - y_i \leq \varepsilon + \xi_i^- \\ & y_i - w^T \phi(x_i) - b \leq \varepsilon + \xi_i^+ \\ & \xi^+, \xi^- \geq 0, i = 1, \dots, N \end{aligned}$$

where C represents the positive trade off parameter between regression error and model simplicity and, ξ^+, ξ^- are vectors of slack variables and $\varepsilon > 0$ is the ε -insensitive loss functions which control the accuracy of model.

The optimization problem presented above can easily be solved indirectly by considering the Lagrangian multipliers α_i^+ and α_i^- on the constraints corresponding to ξ^+ and ξ^- . The dual for the above mentioned primal is thus represented as:

$$\begin{aligned} & \max_{\alpha_i^+, \alpha_i^-} \begin{cases} -\frac{1}{2} \sum_{i=1}^N (\alpha_i^+ - \alpha_i^-)^T (\alpha_i^+ - \alpha_i^-) k \\ -\varepsilon \sum_{i=1}^N (\alpha_i^+ - \alpha_i^-) + \sum_{i=1}^N y_i (\alpha_i^+ - \alpha_i^-) \end{cases} && \text{Equation 3-13} \\ & \text{s.t. } \sum_{i=1}^N (\alpha_i^+ - \alpha_i^-) = 0 \\ & \alpha_i^+, \alpha_i^- \in [0, C], i = 1, \dots, N \end{aligned}$$

where $\alpha_i^+, \alpha_i^- \in \mathbb{R}^N$ are vectors of dual variables and $k \in \mathbb{R}^{N \times N}$ is the kernel matrix calculated from the kernel function $k_{ij} = k(x_i, x_j)$.

Solving b, α_i^+ and α_i^- using the Karush-Kuhn-Tucker (KKT) conditions in above Equation 3-13, the resulting regression functions is:

$$f(x) = \sum_{i=1}^N (\alpha_i^+ - \alpha_i^-) k(x, x_i) + b \quad \text{Equation 3-14}$$

Here $k(x, x_i)$ can be any of the kernel functions explained above or custom made kernel functions such as those developed in bioinformatics (Schölkopf et al. 2004), natural language processing (Lodhi et al. 2002) etc. Selecting the most appropriate kernel functions and learning its parameters highly depends on the problem/data at hand.

3.2.3 Multiple kernel learning

In the above formulation of SVR, instead of using a single kernel function, one can use multiple kernels. In recent years, a tremendous increase in the use of multiple kernels for regression has been observed. Reason for this is the possibility of representing data from heterogeneous sources and different notions of similarity by using different kernels thus significantly improving the performance of learning machines (Gönen et al. 2011). The problem of learning a range of kernel parameters when different types of kernels are applied concurrently in one training process is referred to as Multiple kernel learning (MKL) (Qiu et al. 2005). MKL learns the kernel from the training data. Multiple kernels can be represented as:

$$k_\eta(x_i, x_j) = f_\eta(\{k_m(x_i^m, x_j^m)\}_{m=1}^p) \quad \text{Equation 3-15}$$

where $f_\eta: \mathbb{R}^p \rightarrow \mathbb{R}$, is the combination functions that can be linear or non linear, $k_m(\cdot, \cdot)$ is the kernel function that take P representation of feature data. $x_i = \{x_i^m\}$, represent the data instances, where $x_i^m \in \mathbb{R}^{D_m}$, and D_m is the feature dimensionality. η , parameterises the combination function.

The most common implementation of combination function is the case where the kernel functions and the corresponding parameters are known before training. So the predefined kernels are combined using the parameters. The other version is the one where kernel functions integrated with parameters are optimized during training.

3.2.3.1 Functional form

Functional form defines the numerous ways in which the combination of kernels kernel can be done. They can be grouped into three main categories

- Linear Combination

This is the most popular combination method where the kernels can be combined as a weighted or unweighted sum. In unweighted sum, the mean or sum of kernel is used as the combined kernels.

The weighted sum can be linearly parameterised as

$$\begin{aligned} k_{\eta}(x_i, x_j) &= f_{\eta}(\{k_m(x_i^m, x_j^m)\}_{m=1}^p) | \eta \\ &= \sum_{m=1}^P \eta_m k_m(x_i^m, x_j^m) \end{aligned} \quad \text{Equation 3-16}$$

where η denotes the kernel weights. η can be restricted as a linear sum (i.e. $\eta \in \mathbb{R}^p$), convex sum (i.e. $\eta \in \mathbb{R}_+^p$ and $\sum_{m=1}^p \eta_m = 1$) or the conic sum (i.e. $\eta \in \mathbb{R}_+^p$).

- Nonlinear combination.

The kernels are combined using multiplication, exponentiation and power functions which are nonlinear.

- Data-dependent combination.

Kernels functions are combined by assigning specific weights based on the data.

3.2.3.2 Target function

On the basis of target functions for optimizing the combination function parameters, MKL methods can be grouped into 3 groups

- Similarity based functions

They select the functions parameters that maximises the similarity between the optimum kernel calculated from training instances and the combined kernel matrix. Euclidean distance, Kullback-Leibler (KL) divergence, kernel alignment etc. are the methods to compute kernel similarity.

- Structural risk functions

These methods are based on the fundamental concept of structural risk minimization. They try to minimize the error associated with the system performance and the sum of a regularization term that corresponds to the model complexity. Model parameters can be learned using the ℓ_1 - norm, the ℓ_2 -norm, or a mixed-norm on feature spaces.

- Bayesian functions

They use Bayesian formulation in assessing the quality of combined kernel functions from the candidate kernel. Maximum a posterior (MAP) or maximum likelihood estimate (MLE) methods are used to estimate model parameters

3.2.3.3 Training method

MKL algorithms can be trained either as one step methods or two step methods.

- One-step methods

This method computes the parameters of combination function and combined kernel in a single pass either sequentially or simultaneously. In sequential approach, first the parameters of the combination functions are learned and then the combined kernel is trained using a kernel based learner. However both the parameters are learning together in simultaneous approach.

- Two-step methods.

They use an iterative approach in which by fixing the parameter of base learner, the combination functions parameters are updated and vice versa. This process is repeated until convergence.

3.2.3.4 MKL learning methods

Algorithms that are used in learning the combination of kernel functions in MKL can be grouped into 5 categories

- Fixed rules

These are straight forward methods where the kernel combinations are learned without any parameters. Methods such as summation (Equation 3-17) or multiplication (Equation 3-18) belong to this category.

$$k_{\eta}(x_i, x_j) = k_1(x_i^1, x_j^1) + k_2(x_i^2, x_j^2) = \sum_{m=1}^P k_m(x_i^m, x_j^m) \quad \text{Equation 3-17}$$

$$k_{\eta}(x_i, x_j) = k_1(x_i^1, x_j^1) \cdot k_2(x_i^2, x_j^2) = \prod_{m=1}^P k_m(x_i^m, x_j^m) \quad \text{Equation 3-18}$$

All notations have the same meaning as explained earlier.

- Heuristic approaches

This method learns the parameters of the combination function by obtaining some measures from each kernel functions separately. The measure can either be performance value of a single kernel based learner or value calculated from kernel matrices.

Qiu et al. (2009) proposed two simple heuristics to learn weights of a linear combination of kernel functions (Equation 3-16) for regression problem.

$$\eta_m = \frac{R_m}{\sum_{h=1}^P R_h} \quad \forall m \quad \text{Equation 3-19}$$

$$\eta_m = \frac{\sum_{h=1}^P M_h - M_m}{(P-1) \sum_{h=1}^P R_h} \quad \forall m \quad \text{Equation 3-20}$$

Where, R_m is the Pearson correlation coefficient estimated between the predicted data produced by the regressor learned using the kernel matrix k_m and the observed values. M_m is the mean squared error produced by the kernel matrix k_m . The combined kernel will then be a convex combination of the input kernels.

- Optimization approaches

This method learns the parameters of the parameterized combination function by solving an optimization problem. This optimization can be formulated as a different mathematical model or integrated to a kernel-based learner for obtaining only the combination parameters.

For example, Qiu et al. (2005) proposed a linear formulation of similarity based optimization function for MKL in SVR. Another multiple kernel regression formulation is proposed by Varma (Varma et al. 2009) called the generalized multiple kernel learning (GMKL) uses structural risk optimization for linear and nonlinear formulations of candidate kernels. They use projected gradient descent optimization for updating the kernel weights and standard SVM solvers to find the support vector coefficients. The primal formulation of GMKL can be written as:

$$\min_{w, b, \eta} \quad \frac{1}{2} w^T w + \sum_i l(y_i, f(x_i)) + r(\eta) \quad \text{Equation 3-21}$$

subject to $\eta \geq 0$

where l is the loss function for regression,

$$l = C \max(0, |y_i - f(x_i)| - \epsilon)$$

r is the regularizer with continuous derivative.

The algorithm has to learn the global optimum values of weight vector w , bias term b and kernel parameter η .

GMKL follows a two-step training method wherein the kernel is learned from the dataset by optimizing over η and then the SVR parameters are learned by keeping the kernel fixed.

To achieve this, the primal formulation of the target functions can be rewritten as

$$\min_{w, b} J(\eta) = \frac{1}{2} w^T w + \sum_i l(y_i, f(x_i)) + r(\eta) \quad \text{Equation 3-22}$$

subject to $\eta \geq 0$

For the gradient step to update weight, the dual formulation used for optimisation is given by

$$\text{Maximize } J(\eta) = -\frac{1}{2} (\alpha)^t k_\eta(\alpha) - \epsilon 1^t |\alpha| + 1^t Y \alpha + r(\eta) \quad \text{Equation 3-23}$$

with respect to $\alpha \in \mathbb{R}_+^P$

subject to $1^t \alpha = 0, 0 \leq \alpha \leq C$

where Y is the diagonal matrix of output.

Here the $k_\eta(\cdot, \cdot)$, the kernel matrix and $r(\cdot)$ can be any differentiable function of η with continuous derivative. So the gradient with respect to weights is:

$$\frac{\partial J(\eta)}{\partial \eta_k} = \frac{\partial r}{\partial \eta_k} - \frac{1}{2} \alpha^* \frac{\partial K}{\partial \eta_k} \alpha^* \quad \forall k \quad \text{Equation 3-24}$$

The kernel weights can then be updated using these gradients while considering all other constraints.

- Bayesian approaches

In Bayesian methods, the kernel combination parameters are treated as random variables and inference is performed for learning it and the base learner parameter by assigning priors.

In this thesis, multiple kernels having a weighted linear functional form will be learned using SVR as the base learner. A two-step iterative optimisation approach based on the concept of structural risk minimisation will be employed to learn the linear combination of multiple kernels.

3.3 Baseline Model Development

Loss estimation model for residential sector has depth-damage curves developed for specific sub categories (Merz et al. 2010). The commonly adopted approach for categorising the residential sector is either based on building type, occupancy type (Penning-Rowsell et al. 1992; Thieken et al. 2008) or based on the socio-economic status (SES) (Nascimento et al. 2007). These classification approaches are possible provided there is a clear demarcation between building/construction types and adequate amount of loss data. Conversely in most of the developing Asian cities neither such delineation nor the data is available.

So a major concern remains in identifying which factor(s) should be used in segregating the residential sector so as to reduce uncertainty and improve prediction in loss estimation. This issue/problem will be addressed in Chapter 4 and the

sections below will illustrate the concepts of various multivariate statistical techniques implemented there in.

3.3.1 Principal component analysis

Principal component analysis (PCA) is a non-parametric multi-variate approach used for extracting relevant information by expressing the dataset by means of Principal component (PCs). Principal components are the linear combination of uncorrelated variables obtained as a result of orthogonal transformation on the primary dataset. The components are selected successively so as to account for highest amount of total variance in the primary dataset. Thus the majority of the differences in the dataset can be explained by the initial few variables. Jolliffe (2002) projected this capability of PCA in identifying the minimum number of factors, which without much loss in information can be used for further analysis. PCA has found abundant applications in a various domains such as social sciences (Schyns 1998), delineation of hydrologically homogenous regions (Ilorme et al. 2013), spatial data analysis (Demšar et al. 2012) etc. PCA has also been employed to identify flood damage influencing variables (Thieken et al. 2008; Kreibich et al. 2010).

The PCA procedure is described as follows. Let $\mathbf{X}$ be a random vector of m variables $\mathbf{X} = [X_1, X_2, \dots, X_m]^T$ with the variance covariance matrix as

$$\Sigma_{\mathbf{x}} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1m} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1} & \sigma_{m2} & \dots & \sigma_m^2 \end{pmatrix}$$

Then the eigen value of the matrix $\mathbf{X}$, lets say λ , is defined as the root of

$$\det(\Sigma_{\mathbf{x}} - \lambda \mathbf{I}_m) = 0 \quad \text{Equation 3-25}$$

Where $\mathbf{I}_m$ is the identity matrix. Now there exists a vector $\mathbf{e}_i$ such that

$$[\Sigma_{\mathbf{x}}] \mathbf{e}_i = \lambda_i \mathbf{e}_i \quad \text{Equation 3-26}$$

This vector is the eigen vector of $\mathbf{X}$. The eigen values are ordered such that λ_1 has the largest eigen value and λ_m has the smallest. The principal components (PC) can now be defined from the eigen values and eigen vectors as

$$PC_i = [E^T]X \quad \text{Equation 3-27}$$

Where E is the matrix of eigen vectors $[\{e_1\}, \{e_2\}, \dots, \{e_m\}]$. If correlation matrix is used instead of covariance matrix, then the eigen vectors thus obtained are called componet loadings. Since the eigen vectors are normalized to unit magnitude, they represents the contribution of original variable (X_i) to that particular principal component (PC_i). So a larger eigen vector means higher the contribution of that original variable in that component. In addition, all the PC's are uncorreled with each other.

$$Cov(PC_i, PC_j) = e_i^T [[\Sigma_x] e_j = 0 \quad \text{For } i \neq j \quad \text{Equation 3-28}$$

In estimating the number of principal components, the general procedure is to produce a scree plot of cumulative variance against corresponding PC and search for a plateau. Or when using a correlation matrix, retain only those components whose eigenvalue is greater than 1. In this research, we select only those few components that explain most of the variability in the loss dataset and the specific variables from the PC's will be selected for subsequent cluster analysis.

3.3.2 Mutual information

Mutual information (MI) measures the dependency/relationship between the two variables X and Y such that the relationship is not limited to be linear. Probability density functions and joint density functions are used to define the MI between any two random variables.

$$I(X; Y) = \iint p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dx dy \quad \text{Equation 3-29}$$

where $p(x)$ and $p(y)$ represents the probability density functions of X and Y respectively and $p(x, y)$ denotes their joint density function. The variables X and Y are said to be independent if the MI between them is zero whereas higher values of MI corresponds to strong dependence.

Since none of the probability density functions are known, MI has to be estimated from the data set. Traditional estimators of MI which are either based on histogram or kernels are highly unreliable for high-dimensional data (Scott 2012) due to the curse of dimensionality. In the context of this research the author is dealing with 31

variables/features thereby making it necessary to address the above limitations to effectively estimating the mutual information. These problems were overcome with the estimators introduced by Kraskov (2004) which are based on the principle of nearest neighbours. So when dealing with large dimensions, Kraskov's estimators are found to be more robust as they do not estimate the entropy directly (Gómez-Verdejo et al., 2009). So in this study the nearest neighbours based estimators of MI proposed by Kraskov et al. (2004) are used. They are well suited to deal with mixed data types too.

Kraskov's proposed two different estimators of MI for regression problem. The most extensively used estimator is given below:

$$I_{con}(X; Y) = \psi(N) + \psi(K) - \frac{1}{k} - \frac{1}{N} \sum_{n=1}^N \psi(\tau_x(n)) + \psi(\tau_y(n)) \quad \text{Equation 3-30}$$

where ψ is the digamma function:

$$\psi(K) = \frac{\Gamma'(k)}{\Gamma(k)} = \frac{d}{dk} \ln \Gamma(k)$$

$$\Gamma(k) = \int_0^{\infty} x^{k-1} e^{-x} dx$$

In order to estimate the MI between discrete and continuous data types, a modified version of Kraskov's estimator proposed by Gómez-Verdejo et al. (2009) is used and is given below:

$$I_{cat}(X; Y) = \psi(N) + \frac{1}{N} n_l \psi(n_l) + \frac{d}{N} \left[\sum_{n=1}^N \log(\varepsilon_x(n, K)) - \sum_{l=1}^L \sum_{n \in y_l} \log(\varepsilon_l(n, K)) \right] \quad \text{Equation 3-31}$$

k , the number of nearest neighbours, X and Y are the random samples with d dimensions and N instances, $\tau_x(n)$ & $\tau_y(n)$ are the number of points from the n^{th} observation in X and Y space which are located no further than $\varepsilon_x(n, k)$ and $\varepsilon_y(n, k)$ respectively. $\varepsilon_l(n, K)$ is defined exactly as $\varepsilon_x(n, k)$ and $\varepsilon_y(n, k)$. In Equation 3-30 X and Y are continuous whereas in Equation 3-31 X is continuous and Y is categorical. Kraskov has suggested to use a mid-range value for parameter ‘ K ’, since larger values of ‘ K ’ brings in sizable bias and little variance and vice versa. So in both MI estimators the value of the parameter ‘ K ’ was set to 6 as prescribed by Kraskov et al. (2004)

3.3.3 Cluster analysis

Cluster analysis is statistical technique which can be used to group the flood loss data based on some distance measure reflecting the measure of association among the attributes. There are different classifications of clustering techniques available in literature such as model-based clustering, hierarchical clustering, k-means clustering etc. out of which agglomerated hierarchical clustering techniques have been used widely in grouping the level of damage from natural disasters (Han et al. 2012; Strupczewski 2015) and disaster risk management (Chen et al. 2015). Among the hierarchical cluster analysis, Ward’s methods is chosen to be used in this study as it tends to cluster groups which are extremely relevant in the context of evaluating damage and losses from urban floods (Strupczewski 2015).

Ward’s method is a clustering algorithm which begins with N clusters consisting of exactly one loss data and hierarchically agglomerates the clusters based on the principle of minimizing variance. The algorithm measures the similarity among the pair of clusters using analysis of variance to successively merge the clusters. The method computes the Error Sum of Squares (ESS) between clusters which is used to measure the similarity. Those clusters which minimises the total within group ESS are then merged. To illustrate the merger happening at successive level, a dendrogram is usually used with the value of ESS as the vertical axis.

Considering N loss data represented by n attributes ($X_1, X_2, \dots, X_n$), the ESS is given by

$$ESS = \sum_{p=1}^N (\mathbf{x}_p - \bar{\mathbf{x}})'(\mathbf{x}_p - \bar{\mathbf{x}}) \quad \text{Equation 3-32}$$

where $\mathbf{x}_p = (x_1, x_2, \dots, x_p)$ is the attribute vector of loss data and $\bar{\mathbf{x}} = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p)$ is the corresponding mean of the attributes.

PCA and CA were undertaken using the SPSS software package for Windows, Version 20.

3.4 Performance Measures

The performance of the models are evaluated using the subsequent error statistics

- Root mean square error (RMSE $\$/m^2$)

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (Loss_{obs,i} - Loss_{pred,i})^2}{N}} \quad \text{Equation 3-33}$$

- Pearson correlation coefficient

$$CC = \frac{\sum_{i=1}^N (Loss_{obs,i} - \overline{Loss_{obs,i}})(Loss_{pred,i} - \overline{Loss_{pred,i}})}{\sqrt{\sum_{i=1}^N (Loss_{obs,i} - \overline{Loss_{obs,i}})^2 \sum_{i=1}^N (Loss_{pred,i} - \overline{Loss_{pred,i}})^2}} \quad \text{Equation 3-34}$$

- Mean Bias Error

$$MBE = \sum_{i=1}^N \frac{Loss_{pred,i} - Loss_{obs,i}}{N} \quad \text{Equation 3-35}$$

where $Loss_{pred,i}$ = flood loss predicted by the model and $Loss_{obs,i}$ = observed or the surveyed flood loss for the i^{th} observation and N is the number of cases.

CHAPTER 4. DEVELOPMENT OF A RESIDENTIAL FLOOD LOSS ESTIMATION MODEL FOR JAKARTA

4.1 Introduction

In the recent risk oriented flood management paradigm, the emphasis is given on the holistic approach to reduce the impact of flood. Thus estimation of losses from flooding is an indispensable component for efficiently conducting the risk management. Although Jakarta has a history of (witnessed a series of) severe flooding with some places experiencing annual floods, a comprehensive tool for estimating losses specific to residential sector has not yet been developed (Budiyo et al. 2014). It can also be seen from the literature that the majority of studies on the generation of flood loss prediction/estimation models are for the developed countries and relatively fewer studies are done for developing countries. Thus there is a significant need to understand the nature and the behaviour of flood damages in developing region and that too in an Asian city. Furthermore, the central idea in estimating losses from floods is the notion of depth-damage curves or flood loss functions (Merz et al. 2010). They represent the average loss to the objects at risk at different inundation depth. The goal of this chapter is develop a micro scale comprehensive flood loss estimation model for the residential segment of Jakarta. Further, a novel objective approach for categorising the flood loss data into homogenous groups is also presented.

4.2 Study Region and Flood Loss Database

4.2.1 Study area

Jakarta, (DKI Jakarta) which was established in the year 1527 is the largest and the capital city of Indonesia. Covering a land area of 662.33 sq. km, with a population of 10,187,595 as of November 2011, Jakarta is also the biggest city in the South East Asia (BPS 2011). Housing and industrial sector accounts for the majority of the land use in Jakarta (BPS 2011). Approximately 40% of Jakarta is below sea level, particularly the northern parts; while the southern parts are comparatively hilly. 13 rivers (Figure 4-1) flow evenly through Jakarta with Ciliwung River being

the dominant source of urban drainage. It divides Jakarta into eastern and western halves.

Figure 4-1: Rivers and main drains of Jakarta

The flooding problem in Jakarta has posted an existential threat to the city for many decades. The first recorded flood in Jakarta dates back to 1600's. The 13 rivers which congregate in the low lying urban area, the rising sea water levels under climate change, rapid growth in population and the resultant urbanization (Verburg et al. 1999) has exacerbated the flooding situation in Jakarta from bad to worse (Nasir 2008). In addition to the above, clogging of waterways from household and industrial wastes deforestation, exploitation of resources, encroachment of river systems etc. have exacerbated the extent of flooding in Jakarta. As a consequence both the frequency and magnitude of annual floods in Jakarta have increased. Due to the increase in asset values and population, the impacts from future floods could be devastating (Ward et al. 2011a; Ward et al. 2011b). Table 4-1 shows the major floods which affected Jakarta (compiled from WHO 2007 and Joint Cooperation Program JCP 2013).

Table 4-1: Major floods in Jakarta

No	Year	Affect
1	1699	Flooding of Ciliwung river after the eruption of Mount Salak
2	1714	Deforestation in Puncak resulting in the overflow of Ciliwung river
3	1854	Ciliwung floods the new Batavia
4	1918	Widespread Flooding. Construction work of the Western Flood Canal begins by the colonial Dutch government
5	1942	Jakarta still under water even though the canal work is completed
6	1996	Jakarta floods resulting in a death of 10 people
7	2002	Largest reported flood in Jakarta's history as noted by Dartmouth Flood Observatory. 25 people died
8	2007	Greatest flood in Jakarta
9	2013	70m section of a dike along the West flood canal collapsed.

4.2.2 Data

Jakarta and Bogor experienced heavy rains from 15 to 18th January 2013 leading to severe inflow of water to Jakarta. Notwithstanding this high water level, about 76 meter long section of one levee along the Western flood canal collapsed, inundating Menteng and nearby regions. Simultaneously, heavy rains of over 150mm/day resulted in substantial flooding in the north and western parts of Jakarta and to a lesser extent in the north eastern regions. Figure 4-2 shows the 2013 flood map of Jakarta.

The data required for the development of residential flood loss model has been compiled after the 2013 January floods. The empirical data used were obtained by implementing a questionnaire survey to the household in the selected regions of the ravaged area. The regions were selected in such a way that the flood damage data can be obtained for the representative sampling of reported flood depth for different building/construction/occupancy type. In addition, these regions are also found to exhibit three typical forms of community observable in Jakarta: 1) the working-class with moderate income levels 2) the middle class family which exhibit strong family orientation and 3) the very exclusive residential communities (Ellisa 2014). So the records of all the flood affected regions were compiled with the help of information from NGO's, press releases, flood reports and government agencies. These regions were then superimposed over the flood maps obtained from the regional Disaster Mitigation Agency (BPBD) and point flood depth information from various online data sharing platforms. Then the sub districts which encompass a wide range of flood depth were selected. From these sub districts, villages which were able to demonstrate a broad range of living standards (via expert opinion) of the residents were selected for survey. As a result Pluit in Sub district Penjaringan, Cengkareng Barat and Timur in Sub district Cengkareng and Kelpa Gading Timur in Sub district Kelapa Gading are selected to conduct the survey.

In order to obtain random samples of household, a minimum of two household were randomly selected for each of the affected block/street for interview. Trained interviewers then conducted structured face to face interviews with the consent of the person who has the most comprehensive information regarding flood damage.

So the point of contact was either the head of the house or his wife. The survey resulted in 491 completed interviews. Questionnaire was developed to address the following topics: Flood impact parameters, emergency and precautionary measures, early warning systems, socio-economic factors, building characteristics and flood damage to building and contents. The flood impact parameters comprises of information regarding the flood depth, flood duration, flow velocity and extent of contamination. The questionnaire also gathered information on the various precautionary measures and emergency measures undertaken before and during flood respectively. For instance, interviewees were asked if they had undertaken any or a combination of precautionary measures such as raising the floor level of the entry point of water, flood adapted use of buildings, waterproofing etc. This section also collected information regarding the households flood experience, knowledge of flood hazard and details regarding the availability of flood early warning. The questionnaire was also designed to extract specific information regarding the type, extent and nature of damage inflicted to buildings and contents. This part of the questionnaire was developed after conducting a series of interviews with qualified loss adjusters from the insurance industry in Jakarta. Their knowledge and experience was used in tailoring specific questions to extract exclusive and exhaustive information on flood damages from the interviewees. For example, a list of all possible and plausible damages to buildings and contents from flooding was compiled from loss adjusters experience and from the sample claims report provided. This information was then incorporated in the questionnaire to assess the extent of damage for both buildings and contents. Under the section of building characteristics, information regarding materials used in constructing walls, roofs ceiling and flooring were collected along with building type, number of storey and building area. The type of construction material along with the nature and extent of damage observed to building structure were then used in estimating the repair and restoration costs. The questionnaire also collected information on socio-economic parameters such as age, household size, monthly income, ownership structure and period of stay. The survey questionnaire used for the study is attached in the Appendix A. In order to handle the detailed data, indicator variable was developed by aggregating the answers concerning to one particular topic. Section

4.3 describes in detail the procedure undertaken to aggregate the data set concerning a particular topic to one indicator variable. This information is then processed to develop the loss functions for Jakarta.

The survey thus generated the most accurate information regarding the extent of all possible damage. The damage information was then converted to loss estimates by using the repair and restoration cost for building structure and replacement cost data for contents. Normally the depreciated replacement cost is usually used in assessing the content loss from flooding. However in Jakarta, to popularize flood insurance among public, the insurance companies do provide the entire replacement cost to the insured, provided they replace their damaged items with new. So in this study, instead of employing depreciated replacement cost, full replacement cost was used in estimating the losses. Thus 491 raw loss data complemented by information of water depth, duration of flooding, flow velocity, flood experience, flood warning, precautionary and emergency measures, building characteristics and socio economic factors are used for the analysis.

Figure 4-2: Jakarta inundation map, 2013 Floods

4.3 Derivation of Indicators

4.3.1 Emergency measures

The approach used to develop the indicator of the emergency measure is same as that used in Germany (Thieken et al. 2005). From the check list consisting of 6 different emergency measures, interviewees were asked to select all the measures they have undertaken during the flood. These performed measures were given 7 points. They were also asked to rate the effectiveness of the undertaken action on a scale ranging from 1 (= highly effective) to 6 (= highly ineffective). The aggregate indicator is then developed by subtracting the efficiency score from performed measure score so that the very ineffective measure gets one point and effective measure gets 6 points. Table 4-2 lists out the different emergency measures taken by the people in Jakarta.

Table 4-2: List of Emergency measures and the number of household which has undertaken n(A)

Emergency Measure	n(A)
Documents and valuables stored more securely	317
Furniture and movable objects brought to safety (e.g. 2 nd floor, top of table)	300
Put water barriers (sand bags, flood boards etc.) to divert water	44
Electricity, gas and electrical appliances turned off	166
Openings in buildings (e.g. doors, windows etc.) were sealed against the ingress of water	42
Vehicles driven to safety	80

4.3.2 Precautionary measures

From a check list consisting of 8 different precautionary measures (Table 4-3), interviewees were asked to select all the measures they have undertaken before the flood. The indicator is developed by simply adding the number of measures considered. Since the two informational measures (last 2 in table) and acquiring flood insurance does not have significant effect in reducing the damage, these were not considered when building the indicator (Thieken et al. 2005).

Table 4-3: List of precautionary measures and the number of household which has undertaken n (A)

Precautionary Measure	n(A)
Raised floor at entry point for water	98
Exterior walls of the building coated waterproofed	102
Doors, windows and other openings covered with watertight shields	93
Permanent or mobile water barriers were placed to divert floods e.g. sand bags or local small walls	41
Flood prone floors were used with low value goods and easily movable furniture	81
Taken flood insurance	2
Find out more information about flood hazard of my house	24
Attending Community based capacity development programs	14

4.3.3 Flood experience index

The indicator for flood experience was developed by considering the following two questions: The time period since last experienced flood and the total number of floods experienced. The two variables were ranked in a scale of 0 to 5, where 0 corresponds to the lowest rating and 5 corresponds to the highest rating, and the

indicator was developed by adding the ranks. The range of the indicator lies between 0 and 10 with zero corresponding to least flood experience and 10 corresponding to maximum flood experience. The ranking scheme used is given in Table 4-4.

For example, If “No: of Floods experienced” is 1, from table 4-4, we get the rank as 1 and “Time since last flooded (years)” is 5, similarly rank is 3, then the flood experience index is calculated as $1+3 = 4$.

Table 4-4: Ranking scheme for the development of flood experience indicator

No. of Floods experienced	Time since last flood(years)	Rank
0		0
1	>10 - 25	1
2	>5 - 10	2
3	>2 - 5	3
4	>1 - 2	4
>4	<=1	5

4.4 Scales used for Representing Velocity and Contamination

4.4.1 Flow velocity

Depending on the danger for an adult personnel in relation to flood water depth, the Bureau of Reclamation (Bureau of Reclamation 1998) developed a three point scale for the categorising flood velocity. This scale ranges from ‘0’ for standing water to ‘3’ for high velocity flow. It was used to as the velocity indicator for Jakarta.

4.4.2 Contamination

The values used to represent the extent of contamination by flood water are as follows: 0 – No contamination, 1 – Contaminated with waste or sewage. 2 – Multiple sources of contamination (Heavy Contamination).

4.5 Proposed Model Development

All the available residential flood loss estimation models have depth-damage curves developed for specific classes. These classes are obtained by grouping the element at risk either by the type or use of the building (Thieken et al. 2008) or by the socio-economic classification of the region (Nascimento et al. 2007). The former approaches are used in the developed world whereas the latter was proposed for the developing economies. The rationale behind such classification being the paucity of information on the nature of damage incurred for each object and also the huge effort required in obtaining the same. All the elements under the same group will have the same loss functions.

The logic behind SES as an indicator for classification was purely subjective and more efforts are needed for classifying the elements using rational approaches and indicators (Nascimento et al. 2007). Since the housing section in developing economies exhibit large variability, standardized approaches for classification are a must. Thus it is necessary to group objects at risk of similar nature into same classes for Jakarta too.

To further understand the need to develop a grouping scheme, the variability in flood loss from the surveyed dataset is analysed by means of scatter plot and histogram. The scatter plot for total damage with water depth is shown in Figure 4-3. The plot shows substantial variability especially from water depth greater than 1m. For example at water depth 1.0m, total damage varies from 630 USD to 10153 USD. In addition to it Figure 4-4 to Figure 4-6 shows the histogram of total damage (a), building damage (b) and content damage (c). The samples are plotted by their logarithms as they are positively skewed. The Figure 4-4 also illustrates the large variability with coefficient of variation (CV) of the surveyed dataset at 111.23%. This large variability stems from the fact that the underlying database contains

different objects which were damaged under diverse situations. Slight reduction in variability was achieved when the dataset was differentiated into building structure and content with CV being 105.58% and 105.54% respectively. In order to address this variability and to develop better loss estimation models, it is thus necessary to classify the loss cases that have similar nature of damages or loss using suitable criteria.

Figure 4-3: Scatter plot depicting total loss for residential sector

From the survey conducted in Jakarta, all the buildings were single family dwellings which were mostly detached houses and houses which share a common wall. So classifying buildings based on their type is not possible. Also all the construction type employed in the study area was of brick masonry. Now the approach left with us was to classify the loss data on the basis of socio-economic status (SES). SES classification available for Jakarta used monthly household expenditure as the basis for categorising the residential sector in various SES classes (Nielson, 2010). Obtaining household expenditure was beyond the scope of this study, so such a classification was also not possible. So a new criterion has to be developed for classifying the cases that have similar type of damages into identical groups. The section below describes the procedure in finding out the criteria for grouping the flood loss cases in to homogeneous classes. The method

identifies critical variables that should be incorporated with depth-damage curves/loss functions to reduce uncertainty and improve accuracy of the loss estimation for Jakarta.

Figure 4-4: Histogram of the total damage value

Figure 4-5: Histogram of damage to building structure

Figure 4-6: Histograms of damage to contents

4.5.1 Principal component analysis

In order to comprehend the interaction between the variables that affect flood loss a principal component analysis (PCA) is performed. PCA identifies the independent linear combination of variables that can best describe the variability in the data set. PCA thus reduces the dimensionality of the data set to few underlying variables that can be used with little loss of information in subsequent analysis (Jolliffe 2002). Table 4-5 contains the 6 principal component (PC) loadings which accounts for 61.87% of the total variance, extracted for the Jakarta loss data set. Components which have their Eigen values greater than 1 are only extracted for the analysis. In order to have a better understanding on which of these components have a strong influence on structural and content damage, correlation analysis is performed between the factor score and damage variables. Last two rows in the Table 4-5 give the result of the correlation analysis. Bold values indicate the variables whose absolute loading are greater than 0.5 and are capable for providing better interpretation of the components. Since the loading explains the relative importance

of a variable in a given PC, a variable that has higher loading means it has more impact on that component.

First component (PC1) is characterised by the higher loadings to monthly income, total area and precautionary measures. They also have the highest negative correlation with the damage variables. So household with lower income, minimal floor area and those which have taken effective precautionary measures tend to have lower damages. First component also has the second highest positive correlation for flood experience index and flood warning. This hints that people who have previous flood experience or those received proper flood warning, tend to undertake effective response measures resulting in reduced flood loss. Whereas the PC₂ and PC₃ are mostly driven by parameters which are characteristics of households such as the size, period of stay, educational qualifications and age profile. A correlation, though not significant is observed for the second component (size) indicating that a larger household comprising more of elderly and children (0.326, a positive correlation) tend to have more flood loss. In addition, PC3 which describes the household characteristics is found to have no significant correlation with flood loss. Within PC₄, emergency measure and no. of storey forms the most important variable and has significant negative correlation with flood loss. So households which have undertaken significant emergency measures or having more than 1 storey tend to have lower damages. Flood impact parameters dominate the fifth component (PC5) and have strong correlation with loss variables. Last component (PC6) is marked by higher loadings for house ownership and flood experience index. Indicator for velocity though not prominent still has a negative correlation.

Apart from flood impact parameters, PCA supplemented with correlation analysis has identified household income, floor area and peoples response/reaction to floods (precautionary and emergency measures) as other prominent factors that probably influence flood damage in Jakarta. Though all these variables play some role in the mechanism behind flood damage, these results alone cannot assure whether these factors are most relevant in characterising the flood damage. The factors must also capture the non linear relationships and dependencies with flood loss.

Table 4-5: Principal component loadings for variables that probably influence residential flood loss

Variables	Component					
	1	2	3	4	5	6
Monthly Household Income	-.788	-.055	.050	.164	.062	-.158
Precautionary Measure index	.690	.140	.086	.281	-.220	.018
Floor Area (m ²)	-.650	.272	.147	-.086	-.180	-.155
Total no. of Household	.008	.882	-.156	.104	.141	.097
No. of Elderly	.056	.777	.213	.067	.032	.004
Age, years	.059	.084	.831	-.035	-.204	-.019
Employment Status	-.337	-.083	.741	.032	.232	.077
Years of Stay	.504	.222	.555	-.213	-.025	-.076
Emergency Measure Index	.102	-.034	-.051	.781	.091	-.021
No. of Storeys	-.063	.184	.009	.596	-.026	-.115
Flood Depth (m)	.091	-.105	.034	.400	.707	.031
Flood Duration (hours)	-.233	.322	-.087	-.148	.677	-.123
No. of Children	-.088	.326	-.402	-.286	.432	.090
House Ownership	-.064	-.027	-.082	-.207	.058	.713
Flood Exp. Index	.285	.042	.038	.015	.060	.557
Velocity Indicator	.189	.169	.131	-.349	.334	-.479
Flood Warning	.305	.245	.141	-.046	-.195	.403
<i>Pearson's correlation</i>						
Structural Damage	-0.509 ^{**}	0.192	0.003	-0.036 [*]	0.482 ^{**}	-0.116
Content damage	-0.518 ^{**}	0.221	-0.011	-0.105 [*]	0.487 ^{**}	-0.101

^{**}. Correlation is significant at the 0.01 level (2-tailed) Variance 61.147%

^{*}. Correlation is significant at the 0.05 level (2-tailed)

4.5.2 Mutual information

Mutual information (MI) measure is used to identify the factors which has the most amount of knowledge with the flood loss variables, and thus could be used in differentiating homogenous flood loss groups. Herein, MI is used to measure the dependence between the seventeen variables and flood loss variables (Building loss and Content loss). They can efficiently detect the non linear relationships between variables. In this study, Kraskov's estimator is used to estimate MI between continuous variable and the modified Kraskov's estimator proposed by Gomez is used to estimate MI between categorical and continuous variables. Table 4-6 and Table 4-7 report the MI measure between the variables and flood loss.

Table 4-6: MI measure for building loss

Variables	Mutual Information
Flood Depth (m)	1.7744
Monthly Household Income (USD)	0.8113
Precautionary Measure index	0.4255
Years of stay	0.0643
Emergency Measure Index	0.0628
Age (years)	0.0517
Employment Status	0.0487
Velocity Indicator	0.0450
Flood Experience Index	0.0328
Flood Duration (hours)	0.0314
Flood Warning	0.0243
No. of Children	0.0205
Floor Area (m ²)	0.0165
Total no. of Household	0.0123
No. of Elderly	0.0091
House Ownership	0.0013
No. of Storeys	0.0008

Table 4-7: MI measure for content Loss

Variables	Mutual Information
Flood Depth (m)	1.3692
Monthly Household Income (USD)	0.7869
Precautionary Measure	0.5103
Emergency Measure	0.1508
Years of Stay	0.0895
Velocity Indicator	0.0451
warning	0.0447
Employment	0.0356
Flood Duration (Hours)	0.0316
Total No. of Households	0.0289
Age (years)	0.0203
Flood Experience Index	0.0175
Floor Area (m2)	0.0173
No. of Children	0.0152
No. of Elderly	0.0026
No. of Storeys	0.0016
House Ownership	0.0006

Here also, the flood impact parameter - water depth has the maximum dependency for both building loss (1.77) and content loss (1.36) followed by monthly household income. The next variable that have the maximal information with flood loss are precautionary measures. Emergency measures, on the other hand exhibit a certain dependency with the content loss. These are thus the most relevant factors that exemplify the residential flood loss.

Apart from water depth, the results from PCA supplemented with correlation analysis and mutual information (MI) all concur in that the Monthly household income and people's response/loss reduction measures to flood as the most indicate variables that could be used for categorising the flood loss data of Jakarta. However

to make the flood loss functions usable in real world situation it may be better to have a simple classification scheme. For example, when a flood occurs or when the government or insurance agencies want to assess the risk, they may or may not be having the information on household's private flood impact reduction measures. In this context, it would be logical to borrow the concept adopted in Germany for developing the flood loss functions (Thieken et al. 2008; Kreibich et al. 2010). They represented the influence of damage reduction measures as scaling factors to the original depth-damage or loss functions. Thus for Jakarta, the absolute loss functions are developed with monthly household income and water depth as the grouping variable with the loss expressed in unit area. This study also represents people's response to floods with a scaling factor to the developed depth-damage curves/loss functions.

4.5.3 Cluster analysis

Cluster analysis is performed for building and content losses separately along with water depth and monthly household income so as to identify the income groups for which depth-damage curves need to be developed. The values of these variables are used as attributes in the Ward's algorithm to demarcate the income groups that have different scales of loss at different water depths. Figure 4-7 and Figure 4-8 presents the resulting dendrogram for building and content losses. Since the dendrogram was quite large, it has been sliced at appropriate locations to illustrate the formation of clusters from loss data. The dashed lines in Figure 4-7 and Figure 4-8 represent the corresponding locations. A dendrogram acts as a visual aid in understanding the distance at which clusters are combined. The clusters which are joined are shown by the vertical lines and the corresponding position on the scale (Y axis) indicates the distance. Clusters which are joined by the vertical lines near the left side of the dendrogram are very similar. The observed distances are rescaled to fall into the range of 1 to 25 in such a way that the ratio of the rescaled distances is the same as the ratio of the original distances. Consequently the Y axis is dimensionless.

For this research, a rescaled cluster distance of 3 units has been selected to classify the flood loss data into different income groups. The distance measure of 3 units is so selected such that there are ample numbers of loss data points that can be

grouped under different groups and also it is quite near to the left side of the dendrogram resulting in highly similar clusters. From the dendrogram it can be seen that the flood loss data from Jakarta can be clustered either into 3 or 4 different income groups for both building and content flood loss. This can be visualized by drawing a vertical line at 3 rescaled units. Then the numbers of clusters into which loss data fall into are obtained by counting the number of vertical lines intersecting the line at 3 distance. Table 4-8 shows the distribution of flood loss data when clustered into 3 and 4 income groups respectively. Since the quantity of loss data available to develop depth – damage curve is quite low (41 cases) for a particular income group in the 4 cluster model; it was decided to segregate the monthly household income into 3 groups. Further, significant reduction in the variability of loss data can be seen from Table 4-9 when the dataset has been stratified according to the 3 identified income groups. The variability has been reduced from 105% to 53%, 63% and 59% when classified into 3 groups respectively.

Table 4-8: Loss data counts in each group

Income Groups	Number of data		
	3 Clusters	4 Clusters	
5Mil. – < 9 Mil.	160	135	5Mil. ≤ 9 Mil.
9Mil. – <15Mil.	166	41	≥ 20 Mil.
≥ 15 Mil.	165	124	15Mil. ≤ 20Mil.
		191	9Mil. ≤ 15Mil.

The ranges of monthly household income to which the groups fall into are:

- Group 1 : 5Million to < 9 Million IDR (480 to < 900 USD)
- Group 2 : 9Million to <15Million IDR (900 to < 1440 USD)
- Group 3: ≥ 15 Million IDR (≥ 1440 USD)

Loss functions are developed for each of these income groups.

Table 4-9: Coefficient of variation

	Coefficient of Variation		
Structure Loss	0.53	0.63	0.59
Content Loss	0.52	0.61	0.55

Kruskal-Wallis-H-Test was applied to test whether the three independent groups of data are statistically significant. The resulting test statistics is provided in Table 4-10 and Table 4-11 indicates the groups are statistically significant with $p < 0.05$.

Table 4-10: Ranks form Kruskal-Wallis Test

Income Groups (IDR)	N	Mean Rank	
		Building Loss (USD/m ²)	Content Loss (USD/m ²)
5Mil. – < 9 Mil.	160	165.76	170.81
9Mil. – <15Mil.	166	229.55	229.45
≥ 15 Mil.	165	340.36	335.55

Table 4-11: Test statistic

	Building Loss (USD/m ²)	Content Loss (USD/m ²)
Chi-Square	126.380	113.004
Degree of freedom	2	2
Asymp. Sig.	0.000	0.000

Figure 4-7: Dendrogram for Wards clustering applied to building loss (Dash-dot line indicates the locations where dendrogram was sliced.)

Figure 4-8: Dendrogram for Wards clustering applied to content loss (Dash-dot line indicates the locations where dendrogram was sliced)

Figure 4-9: Flowchart showing the approach used for demarcating flood loss data into homogenous groups

4.5.4 Development of residential loss functions

The loss functions for Jakarta were developed for the 3 income groups for damages expressed in US dollars (US\$) per surface area of the first floor of the building (m^2). Since the loss information was not available for every water depth, average loss value was computed for selected bins of water depth for the three groups. The bins were so selected that there exists significant difference in the mean loss between the bins.

Table 4-12 enumerate the factors along with their classification considered for the development of Residential Loss Estimation Model for Jakarta. The core loss model was derived by using the first two variables and the third variable (optional) was used to quantify the consequence of response measures. The model developed is a micro-scale loss model that can be used to estimate losses on a building-by-building basis. Figure 4-10 and Figure 4-11 represents the micro-scale building and content loss functions/curves developed for the residential sector of Jakarta.

Table 4-12: Factors considered in the residential flood loss estimation model for Jakarta

Factor	Classification
Water level	$\leq 0.2\text{m}$, $0.2 - 0.5\text{m}$, $0.5 - 1.0\text{m}$, $1.0 - 1.5\text{m}$ and $1.5 - 2.0\text{m}$
Income Group	Group 1 (G1) : 5Million to < 9 Million IDR
	Group 2 (G2) : 9Million to <15Million IDR
	Group 3 (G3) : ≥ 15 Million IDR
People response measures	None, Good and Excellent measures.

Figure 4-10: Micro-scale residential flood loss estimation model for building structure. Bars represent 1 standard deviation

Figure 4-11: Micro-scale residential flood loss estimation model for contents. Bars represent 1 standard deviation

From the results of PCA and MI, it can be observed that precautionary measures have significant influence on both building and content loss whereas emergency measures effect content loss only. The impact of people's response to flood in terms of either precautionary or emergency measures was represented by scaling factors to the core loss model. The scaling factors are derived by comparing the average losses from those cases which has taken excellent measures, good measures and no measures with the average loss computed in the first stage of the model. Description of excellent measures and good measures are provided in the sections below.

For the people of Jakarta, the flood problem is nothing new. Among the surveyed households, 97% had pervious flood experience. Out of this 38% have experienced flood in past 2 years (Figure 4-12). Several studies have confirmed that the past flood experience and perceived & observed consequences from floods are some of the major drivers for implementing private household flood protection strategies (Grothmann et al. 2006; Reynaud et al. 2013). Thus assumptions were made that learning from past experience the households in Jakarta might have taken enough flood loss reduction measures to be better prepared for the floods. However, based on the results of the survey, only a few household were able to accomplish elaborative damage reduction measures.

Figure 4-12: Flood experience before 2013 January

To assess the efficiency of each precautionary measure, the reduction in damage was estimated by comparing the content and building loss of those cases where a particular measure was undertaken with other cases where that particular measure was not undertaken regardless of other measures. However, for emergency measures the comparison was done for all cases for which the specific measure was rated to have performed effectively with all cases where the measure was not performed. As a result, the second measure in Table 4-2 was solely considered from the list of emergency measures to estimate the loss reduction efficiency. Mann-Whitney-U-Test was used to check if there exists any significant difference between these two independent groups of data (Kreibich et al. 2005, Thielen et al. 2005). In addition, the number of cases was also considered in assessing the significance of damage reduction by each measure.

For the extreme flood event in 2013, many precautionary measures have helped in reducing the flood loss. Substantial reduction in both content and structural damage was observed through flood adapted construction by elevating the floor level at the point of entry of flood water. On an average the building loss was 39% lower for buildings that have their floor level raised in comparison with those who haven't (Figure 4-13). For contents, this difference was 30.3% (Figure 4-13). Similar observations were also made for household which has undertaken flood adapted building usage where in the flood prone floors were used with easily movable and low value objects. Content loss in this case got reduced by 24.9% (Figure 4-14) whereas no significant reduction was observed to building loss.

Waterproofing of buildings and installation of watertight shield over the openings had on an average trivial effect on both building and content damage (Figure 4-15 and Figure 4-16 respectively). This could be attributed to the intensity and magnitude of the 2013 floods where flood depths greater than 2m were reached for several days (Joint Cooperation Program (JCP) 2013). The efficiency of these structural measures under these extreme conditions would be negligible. However some studies from UK (Kelman et al. 2004) and Germany (Thielen et al. 2005) has projected significant damage reduction potential for these structural measures under less severe floods.

Figure 4-13: The influence of raising the floor level at the entry point on building and content loss (bars=means, points = median and whiskers = 25 and 75 percentile). Significant difference observed for both building and content loss.

Figure 4-14: The influence of using low cost/value and easily movable items on building and content loss (bars=means, points = median and whiskers = 25 and 75 percentile). Significant difference in loss is observed only for content.

Figure 4-15: Building and content loss for household which has not undertaken waterproofing (bars=means, points = median and whiskers = 25 and 75 percentile). No significant difference found for both building and content loss.

Figure 4-16: Building and content loss for household which has installed and not installed watertight shields over openings such as doors, windows etc. (bars=means, points = median and whiskers = 25 and 75 percentile). No significant difference found for both Building and content loss.

Figure 4-17: The influence of permanent or mobile flood barriers on the content and building loss (bars=means, points = median and whiskers = 25 and 75 percentile). Significant difference found for both building and content loss.

Figure 4-18: Building and content loss for households who has and has not moved the objects including furniture and other valuables to safety (bars=means, points = median and whiskers = 25 and 75 percentile) Significant difference observed for content loss only.

The private water barriers either permanent or mobile erected during the flood has demonstrated substantial reduction in damage to both contents and building structure. For buildings installed with water barriers, the damage was 41.8% lower than those households which have not installed the water barriers. For contents, this difference in loss was 38.3% (Figure 4-17). On the other hand the number of household which has undertaken this measure is quite low (Table 4-3). Since the sample size also determine the consequence of damage reduction, it is reasonable to provide a lower weight for this measure.

Figure 4-18 depicts the damage reduction potential of precautionary measures – moving the objects including furniture and other valuables to safety. Content loss was reduced by 19.25% whereas no significant reduction in damage to building structure was observed.

Based on the above results it can be concluded that, the measures such as raised/elevated floors and low value & easily movable furniture can be considered as the most significant measures in reducing flood loss. These ‘flood adapted use and construction’ measures have significant damage reduction potential to both buildings and contents. Thus the household which has undertaken any or both of these ‘flood adapted measures’ solely or in combination with any other flood loss reduction measures are categorised under excellent response measures. Good measures were characterised by the presence of private water barriers and its combination with any precautionary measures mainly moving furniture & valuable to safer location. Finally the households which have not employed any damage reduction measures were categorised into ‘No measure’ action group. The flood loss data set are then segmented according to the above mentioned 3 groups and scaling factors (Table 4-13) are developed separately for buildings and contents within these categories.

Table 4-13: Scaling factor for buildings and contents due to people's response measure to reduce damage

Action	Scaling factor for	
	Building Structure	Contents
No Response	1.03	1.13
Good Response	0.97	0.92
Excellent response	0.91	0.85

4.6 Model Performance Evaluation

To evaluate the performance of the two models namely; flood loss functions and scaled flood loss functions, a widespread approach followed in the field of hydrology called re-sampling procedure is employed. 100 bootstrap samples each with 100 affected household are drawn randomly with replacement from the dataset. The two loss prediction models are then applied to each household record and the predictions are quantified using following error statistics. The measures: Mean bias error (MBE), root mean square error (RMSE) and the Pearson correlation coefficient are provided in box plots.

An emphasis has to be made at this point that no separate test-sets are used to evaluate the performance of the flood loss prediction models. The bootstrap approach generates samples for evaluation by re-sampling from the training dataset. This approach, being not legitimate per se, may considerably influence the prediction result. However, since these two loss models contain relatively less number of free parameters they will not over fit. Consequently, the results yielding from evaluating the performance on training data is not overly optimistic.

Figure 4-19 to Figure 4-21 compares the performance of the flood loss estimation model (depth-damage curves) and scaled flood loss estimation model. It can be seen that the model developed from scaled flood loss functions shows a clear improvement especially in prediction content loss from flooding. The prediction

Figure 4-19: Comparison of RMSE values for the flood loss estimation models

Figure 4-20: Box plot representing the correlation coefficient of flood loss estimation models

accuracy of the content loss model has improved by 12.5%. The stronger value of the correlation coefficient also supports the same (Figure 4-20). However, only a marginal improvement is observed for the case with predicting building loss (i.e.5%). This can be explained by the fact that the scaled damage function is multi-variate model which incorporates the effect of people's response measure to flood, which improved prediction accuracy.

From Figure 4-19, the RMSE values of building loss and content loss for the depth-damage function model are 1.74 and 3.16 respectively. Similarly the RMSE values for the scaled damage function model are found out to be 1.65 and 2.766 respectively. Comparing the error deviations, it can be said that the scaled damage function model has better accuracy in predicting losses from flooding. The improvement in prediction accuracy of the latter model when compared to the former can be calculated as follows:

$$\text{Building loss} = \left(\frac{1.74 - 1.65}{1.74} \right) * 100 = 5\%$$

$$\text{Content loss} = \left(\frac{3.16 - 2.766}{3.16} \right) * 100 = 12.5\%$$

Figure 4-21: Comparison of flood loss estimation model in terms of their mean bias error (MBE)

In contrast to the results observed from RMSE and Pearson correlation coefficient, the MBE estimate of the scale damage model is relatively on the higher side compared to the simple depth-damage functions. This is because the scaled damage function model has a tendency to over-predict the building and content loss from floods for some cases. This can be attributed to the variance associated with the scaling factors which has been developed to quantify the impact of damage reduction measures. The scaling factors were derived as an average estimate representing the entire dataset within each action groups (Table 4-13). Due to the limited size of data per action groups, it is possible for these factors to be biased. Thus the mean error can get skewed to either sides depending on the data from which the scaling factors are derived. Apart from this measure, all other measures concur in that the multi-variate flood loss model: ‘scaled flood loss functions’ is better in predicting residential flood loss for Jakarta.

The extent and nature of flood damage depends on the single and joint interactions of all the factors. However in this study, the damage reduction potential of each measure was assessed individually. Still the scaled loss functions developed from such minimal assessment has shown considerable improvement in predicting flood loss. As a result, if these interactions can be captured and represented appropriately then the prediction performance of flood loss models can be enhanced significantly. This limitation will be addressed in the next chapter where new methods to capture the interaction and development of loss models are proposed.

4.7 Summary

The study presented the development of two residential flood loss estimation models for Jakarta. The first model, depth-damage functions is a uni-variate model which relates flood loss to water depth. This is the most widely used methodology in estimating and predicting flood loss. Whereas the second model which is a multi-variate model incorporates the effect of private precautionary and emergency measures by means of a scaling factor to the former model. This chapter also presented a standardized approach for demarcating flood loss data into homogenous groups and identifying prominent factors influencing flood loss. The outcome of the approach when applied to the flood loss data collected from Jakarta presented herein demonstrate that monthly household income, private precautionary measures and emergency measures as the appropriate variables for categorizing flood loss data along with water depth. Among these variables, monthly household income is used as attributes in Ward's clustering algorithm to yield homogenous flood damage groups. Influence of precautionary and emergency measures were incorporated as scaling factors. Flood loss estimation models are developed for each of these income groups.

CHAPTER 5. EVALUATION OF FACTORS INFLUENCING FLOOD LOSS

5.1 Introduction

In the classical flood risk assessment paradigm, depth-damage functions forms the central idea for estimating damage to flooded object. They relate absolute or relative damage to the corresponding water depth. However a variety of factors such as quality of response measures, duration, velocity, contamination, flood experience etc. which affect flood damages are rarely taken into account (Wind et al. 1999; Penning-Rowsell et al. 2000; Thielen et al. 2005). The extent of damage caused by the single and joint interaction of these parameters with flood impact parameters is largely unknown. The review in Chapter 2 has mentioned some studies where the loss model has been developed by incorporating different damage influencing parameters. These studies have focussed on the application of various machine learning methods in developing multi-variate flood loss estimation models. (Vogel et al. 2012; Merz et al. 2013). Though the studies of these kinds are important, further research is needed. This is important because it is completely feasible to develop a loss estimation model incorporating a plethora of factors, but the viability of using these models in real time situation is difficult. In addition one may not possess the required detailed information and the time, money to collect data. Instead of incorporating all the factors, it is necessary to identify those prominent factors which are considered as inputs for loss model development. In addition, studies investigating the interaction of the factors which influence damage have seldom or never been attempted for a developing Asian region.

The purpose of this chapter is to propose a standard procedure to eliminate subjectivity in identifying the key factors which influence the flood loss by capturing single and joint interaction among the variables. Key component of this procedure is an enhanced approach for mixed data feature selection. In addition, a multi-variate flood loss estimation model based on multiple kernel support vector regression (MKr) is also presented.

5.2 Variable and Dataset

The data used in the study were collected after the 2013 January floods in Jakarta. This data set consists of a total of 491 flood loss data with detailed information concerning all the plausible factors which influence flood loss. These details were collected via a structured questionnaire survey. Efforts were undertaken during the survey to obtain independent and identically distributed household samples (Further details can be found at section 4.2.2.) The previous chapter documents that the raw data were supplemented with indicators for emergency measures, precautionary measures and flood experience. The indicator variable was used to aggregate the answers relating to a particular topic into single measure, thereby making it comfortable to handle the large dataset. In order to develop the indicator for precautionary and emergency measures, households were posed with one critical question. They were asked to provide a rating for the efficiency of various measures they have undertaken with respect to its damage reduction potential. Households made use of their past and current flood experience in answering this question. As a result this measure has imparted an element of subjectivity in the development of respective indicators. This is because, even for similar households the same response measures may behave differently as they may be subjected to diverse flooding situations. Besides, these indicators may not fully capture the joint interaction between the resistance measures and the flood impact parameters. Thus for the current study, instead of using these indicators, the various measures employed by the households were incorporated as a dichotomous nominal variables. For example, if a particular house has installed water shields then that house is provided a ranking of 1 (Yes) otherwise 0 (No). Ordinal ranking scales were used to indicate the flow velocity and extent of contamination. Table 5-1 enumerates the list of all variables appropriated to seven domains. The last two variables, building loss and content loss form the response variables. The dataset consists of variables in all levels of measurement such as continuous variables (C) and categorical variables both nominal (N) and ordinal (O).

Table 5-1: Description of factors

X	Predictor Variables	Scale and range
Flood Impact Parameters		
Dep	Flood water depth	C :- 0 to 2m inside house
Dur.	Inundation duration	C : - 1 to 360 hours
Vel.	Flood Velocity indicator	O :- 0 - still to 3 - High flow velocity
Con.	Contamination Indicator	O :- 0 - no contamination to 3 - Heavy Contamination
Flood Warning		
Warning	Received any warning	N :- Yes/No
Warn_Time	Warning lead time	C :- 0 to 48 hours
Warn_src	Flood warning source	N :- 1- Government, 2- Mosque, 3- Media, 4- Others
Emergency Measures		
Doc_safety	Documents and valuables stored more securely	N :- Yes/No
Furn_safe	Furniture and movable objects brought to safety (e.g. 2nd floor, top of table)	N :- Yes/No
Temp. Barr:	Put water barriers (sand bags, flood boards etc.) to divert water	N:- Yes/No
Elec_Off	Electricity, gas and electrical appliances turned off	N :- Yes/No
Open_sealed	Openings in buildings (e.g. doors, windows etc.) were sealed against the ingress of water	N :- Yes/No
Veh_safe	Vehicles driven to safety	N :- Yes/No
Precautionary Measures		
Raised_Flr	Raised floor at entry point for water	N :- Yes/No
Water_proof	Apply waterproof coating/membrane to the exterior walls of the building	N :- Yes/No

Shields	Install watertight shields over doors, windows and other openings	N :- Yes/No
Water_Barr	Permanent or mobile water barriers were placed to divert floods e.g. sand bags or local small walls	N :- Yes/No
Low_Val_Furn	Flood prone floors were used with low value goods and easily movable furniture	N :- Yes/No
Fld_Insr	Taken flood insurance	N :- Yes/No
FH_Info	Find out more information about flood hazard of my house	N :- Yes/No
Comm_pgms	Attending Community based capacity development programs	N :- Yes/No
Flooding Information		
KFH	Knowledge about flood hazard	N :- Yes/No
Fld_Exp_Idx	Flood Experience Index	O :- 0 - No experience to 9 - Most recent experience
Characteristics of Building		
BT	Building Type	N :- 1- Single Family Dwelling, 2 - Semi-Detached, 3-Multi Family Apartments
TFA	Total floor area	C :- 36m ² to 350m ²
Storey	No: of Storeys	O :- 1- Single Storey , 2 - Multiple Storeys
Socio-economic factors		
Child	Number of Children	C :- 0 to 3
Adt	Number of Adults	C :- 1 to 5
Eld	Number of Elderly	C :- 1 to 3
Own	House Ownership	N :- 1 -Own, 2- Rented
Inc	Monthly Household Income	C : - 480 to 3000 USD
Str_Loss	Building loss of residential Buildings	C:- 0.5 to 30 USD/m ²
Cont_Loss	Content loss of residential Building	C :- 0 to 60 USD/m ²

5.3 Proposed Input Feature Selection Method

The fundamental idea for any flood loss estimation model is the concept of depth – damage curves which relates the damage suffered by the object and the depth of flooding. Accordingly the basic input for all flood loss estimation models is water depth. A recent study on the German flood loss database resulted in developing a comprehensive loss model which incorporated the entire surveyed features (Vogel et al. 2012). Though this model was able to capture the interactions and dependencies among the variables, its performance was just comparable with the existing flood loss estimation models for Germany. The lack of improvement could be attributed to the following reasons:

1. Most of the factors collected are either irrelevant or redundant. Inclusion of these futile factors can adversely affect the performance of the learning algorithm (Peng et al. 2005) and
2. The algorithm used for developing the model was capable of handling only discrete or categorical input. So they had to use standard discretization techniques such as binning approaches to convert the continuous factors to categorical ones. This approach has a strong drawback as it can lead to sufficient loss of information especially when dealing with small datasets (Doquire et al. 2011).

Feature selection not only improves the prediction capacity but also lead to better interpretability and understanding of the problem. In order to improve the performance and efficiency of a learning machine in predicting flood loss, key problem lies in the identification of appropriate input factors from the entire feature dataset. In the literature apart from expert opinion, no specific methods has been suggested for selecting inputs or factors for developing the loss model. In this study, a data mining based approach called kernel methods coupled with recursive feature elimination is proposed to select the prominent factors that influence flood damage. The method has been modified to cater to the mixed parameters for regression problem and also identified most prominent factors by capturing its ‘relevance’, ‘importance’ and ‘interaction’.

5.3.1 Kernel function selection

The data set consists of 31 candidate variables which belong to either categorical or continuous data type. Among the categorical variables, 4 variables belong to ordinal category and 19 belong to nominal category. The rest 8 variables belong to continuous data type. An ordinal variable is a type of categorical variable that comprises of two or more categories with an intrinsic ordering whereas nominal variables are those which lacks this ordering. A continuous variable is one which can take any value within certain limits defined by the lower and upper bound. For a meaningful and robust analysis, a clear distinction is required between the ordinal, nominal and continuous variables.

In kernel methods, kernel functions compute the inner product of one or more variables, providing a measure of similarity between the two cases. When computing the inner product for continuous and ordinal variables, the range of these variables plays an important role. For example, continuous variables such as water depth and age can range from 0 to 2 m and 20 to 70 years respectively. Since they are not comparable in range, the influence of each variable computed by their inner product will be dissimilar. The situation is similar for variables of ordinal data types. In addition to it, for ordinal variables there is a possibility of having a category equals zero. Thus the inner products for such case will always be zero irrespective of its dissimilarity with other cases. For nominal variables, the inner product should be greater than zero only if the cases are in the same category. The commonly used kernel function does not take into account the above mentioned problems associated with variable of different type when measuring the similarity. This could lead to significant loss in information and interpretability of the resulting learned model.

The kernel functions proposed by Daemen (2009) called the clinical kernel redefines the way how the inner product is normally computed. Their approach accounted for different types of variables, the variation in range and also ascertained the same level of influence for each variable. Clinical kernel functions proposed for each type of variables are provided below:

1. Ordinal and continuous variables

The following kernel function developed for continuous and ordinal variable types:

$$k_x(m, n) = \frac{(\max - \min) - |x_m - x_n|}{(\max - \min)} \quad \text{Equation 5-1}$$

where x can be any continuous or ordinal variable and x_m, x_n denotes the values of m^{th} and n^{th} case.

2. Nominal variables

The kernel function defined for nominal variables is as follows:

$$k_x(m, n) = \begin{cases} 1 & \text{if } x_m = x_n \\ 0 & \text{if } x_m \neq x_n \end{cases} \quad \text{Equation 5-2}$$

The following example illustrates the usage of clinical kernels for the flood loss dataset. Consider three flood loss cases a, b and c where we would like to calculate the kernel matrix (similarity matrix) for variables water depth, indicator of velocity and water proofing measure. Water depth is a continuous variable having its range between 0.05 m to 2 m, velocity indicator is an ordinal variable which takes values from 0 to 3 and water proofing measure is a nominal variable. Table 5-2 lists out the values for each variable for the sample cases considered.

Table 5-2: Sample case

Cases	Water Depth (m)	Velocity Indicator	Water Proofing
a	0.5	0	No
b	1.2	1	Yes
c	2	3	Yes

The similarity matrix or kernel matrix for age can be calculated as follows:

$$K_{depth}(a, b) = \frac{(2 - 0.05) - |0.5 - 1.2|}{2 - 0.05}$$

$$= 0.64$$

$$K_{depth}(a, c) = \frac{(2 - 0.05) - |0.5 - 2|}{2 - 0.05}$$

$$= 0.23$$

$$K_{depth}(b, c) = \frac{(2 - 0.05) - |1.2 - 2|}{2 - 0.05}$$

$$= 0.59$$

The kernel matrix for water depth can be represented as:

$$K_{depth} = \begin{bmatrix} 1 & 0.64 & 0.23 \\ 0.64 & 1 & 0.59 \\ 0.23 & 0.59 & 1 \end{bmatrix}$$

Similarly the kernel matrix for velocity indicator can be expressed as:

$$K_{vel} = \begin{bmatrix} 1 & 0.67 & 0 \\ 0.67 & 1 & 0.33 \\ 0 & 0.33 & 1 \end{bmatrix}$$

For the nominal variable water proofing, cases ‘b’ and ‘c’ has undertaken this measures whereas ‘a’ hasn’t. Applying the kernel function of nominal variables on variable waterproofing, the resulting kernel matrix becomes

$$K_{proof} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

The resulting kernel matrix for each variable has been normalized to the interval [0, 1]. These matrices can be combined using appropriate weights to represent a global heterogeneous kernel matrix. This matrix could then be used to describe the similarity between the cases in the flood loss database. Thus the clinical functions were employed in this study to capture the importance of each variable without any loss in information. The influence and importance of each variable was determined by the weights learned from the optimal linear and nonlinear combination of clinical base kernels using multiple kernel learning.

5.3.2 Multiple kernel learning for regression (MKr)

Multiple kernel learning (MKL) was used to learn optimal linear combination of 32 clinical base kernels where each kernel represents one of the 32 input variables. The MKL was integrated with Support Vector Regression (SVR) as the base learner to model and predict the flood loss. To optimally combine the kernel matrices for SVR, a nested two step iterative MKL optimization approach was employed. In the first stage, the weight vector η is kept constant and the heterogeneous kernel matrix is learned as a linear combination of base clinical kernels:

$$K_{\eta}(x_i, x_j) = \sum_{m=1}^P \eta_m k_m(x_i^m, x_j^m) \quad \text{Equation 5-3}$$

such that $\eta_m > 0$

Now using SVR as the base learner, the single kernel function $K_{\eta}(\cdot, \cdot)$ obtained from Equation 5-3 was solved to obtain the Lagrange multipliers(α) . Formulations of primal for the regression problems (Equation 3-22), the corresponding dual (Equation 3-23) and the approach to obtain α are explained in section 3.2.2 and 3.2.3.4. In the second step, the Lagrange multipliers are considered known, and the weights were updated using the projected gradient descent method (Varma et al. 2007). Equation 3-24 is used in calculating the gradient. The whole process was repeated until convergence. More details on this algorithm are provided in the optimization approaches used for MKL under section 3.2.3.4. Algorithm 5-1 portrays the two steps multiple kernel learning procedure employed in learning the linear combination of base kernels.

By keeping the weights of the base kernels positive in Equation 5-3 ($\eta_m > 0$), the combined variable representation is equivalent to concatenation of the respective feature maps. Also by restricting the kernel weights to be nonnegative relative importance of each kernel in the decision function is directly the value of the weights itself. Thus the weights can be seen as the importance of the feature in predicting flood loss.

Algorithm 5-1: Multiple Kernel Learning

1. Set n to 0
2. Initialize the weight vector $\eta^0 = \eta_1 \dots \dots \eta_n$ randomly, n = number of variables
3. Repeat till convergence
4. Heterogeneous kernel $K \leftarrow \sum_n \eta_n k_n = \eta_1 k_1 + \eta_2 k_2 + \dots \eta_n k_n$
5. Use the SVR as the solver, solve the single kernel K to obtain α
6. Use gradient descent update the weights by making sure $\eta \geq 0$
7. $n \leftarrow n + 1$

5.3.3 Recursive feature elimination

Recursive feature elimination (RFE) is one of the most commonly used methods in feature selection for small sample problems. In RFE features are ranked based on some measure of their importance in a recursive fashion. Basically RFE improves the generalization performance by iteratively removing the least important feature based on embedded backward elimination strategy (Guyon et al. 2002). This process is continued until all the variables are ranked. Then the first n features from this ranking are selected as part of feature selection process. For feature selection problems using RFE, a linear SVM is generally combined with RFE where the weights extracted from SVM classifier are used to rank features. Apart from SVM, any other classifier can also be used with the RFE method for feature selection.

For flood loss estimation the predictand variable is loss or damage value which is a continuous variable. Therefore the flood loss estimation problem shifts from classification to regression. So instead of a classification model, a regression model namely multiple kernel support vector regression (MKr) is used to replace SVM in the crux of the RFE method. Algorithm 5-2 provides the pseudo code for RFE for regression problem.

Algorithm 5-2: Recursive Feature Elimination

Inputs: $R = []$ empty ranking list and

$F = [f_1, f_2 \dots f_n]$ feature set

Step 1: While $F \neq []$ do

Step 2: Train a regression model RM (MKr) with features in F

Step 3: Compute the weight η and rank the features based on it.

Step 4: Remove the least important feature from F

Step 5: $R \leftarrow$ least important feature

Step 6: End while

Step 7: return R

It can be seen that the RFE algorithm discovers a subset of features which yields good prediction performance. This is a main drawback as the best feature subset identified may not represent the best feature combination. A weak and the redundant feature rejected by RFE can provide a significant improvement in performance when used with other features, i.e. by capturing the interaction among features the weak feature can become an important feature. Even though the optimization process of MKr can be interpreted as a ‘relevance maximising method’, the weights extracted from MKr alone is not a comprehensive measures in capturing feature’s ‘importance’ with respect to the response variable.. The weights only describe the contribution of each feature in the function $f(x)$ (Equation 3-11) so as to predict the target y_i with at most ε deviation plus some error. Also MKr only uses a subset of examples i.e. support vectors when determining the feature’s importance. Thus an appropriate metric representing the relevance of each feature along with the weights is required to comprehensively represent feature’s importance while ranking. So the existing RFE algorithm has to be improved to:

- Incorporate better representation of the feature importance and
- Capture the interaction among features for performance improvement

To measure the relevance of feature with the target variable, Peng (Peng et al. 2005) introduced a criteria using mutual information called maximum relevance. In maximum relevance criteria, the features are so selected that they individually

reflect a large dependency with the target variable by having largest mutual information $I(x_i: y_i)$.

The mutual information (MI) between two random variable X and Y defined in Equation 3-29 cannot be estimated analytically. Traditional estimators which are either kernel based or histogram based methods used to approximate the value of MI suffer from the curse of dimensionality (Verleysen 2003) and loss of information due to the discretization. The flood database in this study consists of multi-dimensional data points with continuous predictand variable and mixed (continuous and categorical) predictor variables. Since the traditional estimators become unreliable with increase in dimensions, nearest neighbours based estimator of MI proposed by Kraskov and Gómez (Kraskov et al. 2004; Gómez-Verdejo et al. 2009) was implemented. In addition these estimators can handle mixed data types which are very much relevant to this study. These estimators for MI have been found to give exemplary results for datasets with mixed variables (Doquire et al. 2011).

To estimate MI with continuous data types Equation 3-30 was used. Further to estimate MI between continuous and categorical data, Equation 3-31 was used. MI estimated by these equations for each feature represents the feature relevance. These were then incorporated with the weights learned from MKr to represent the comprehensive importance of each feature to the dependent variables i.e. flood loss.

In order to capture the interaction among the features for improved prediction performance, the weak feature based on their weights is not allowed to be removed as in RFE. Instead the prediction performance of the model estimated without the weak feature is compared with the performance of original model. If the prediction performance degrades, then the weak feature is retained and the model evaluates the performance with the second weak feature. This process is repeated until a feature is found whose exclusion does not result in the degradation of prediction performance.

To overcome the drawbacks of the existing RFE and to use it with mixed data, an enhanced RFE algorithm by incorporating the measures/methods described above is proposed. The new hybrid mixed data feature selection algorithm is developed by

coupling MKr and MI with RFE. The core of the new algorithm mainly consists of the following five steps. First, compute the feature importance and prediction accuracy using MKr and the feature relevance (MI) using Kraskov's estimator for all features. Second, normalize each feature using the maximum values of feature importance and MI ($\max[\eta_m]$ and $\max(MI_m)$). Third, sort the feature set S with respect to the sum of normalized values for each feature. Fourth, assess the prediction performance by removing the lower ranked and check if the prediction performance has degraded using MKr. Fifth, eliminate the feature whose exclusion does not degrade the prediction performance and place it on top of the ranked feature list R . Continue the iteration until all features are ranked.

Algorithm 5-3 provides the pseudo code for Hybrid mixed data recursive feature selection is given in

Algorithm 5-3: Hybrid Mixed Data Recursive Feature Selection

Inputs: $R = []$ empty ranking list and

$S = [f_1, f_2 \dots f_n]$ feature set

1. While $S \neq []$ do
2. Train a regression model (MKr) with features in S and obtain prediction performance PF_1
3. for all d features in S do
4. Extract the weight $\eta(k)$
5. Compute feature relevance $MI(k)$ using Eq.5-5 or Eq. 5-6 depending on continuous and categorical predictor
6. End for
7. Normalize η and MI by $Max(\eta(k))$ and $Max(MI(k))$
8. $S \leftarrow \text{Importance} = \eta(k)/Max(\eta(k)) + MI(k)/Max(MI(k))$
9. Sort S in ascending order
10. For $i = 1$ to $size(S)$
11. New feature list $\leftarrow [S - S_i]$
12. Evaluate the prediction performance PF_2 using MKr with the new feature set

13. If $PF_1 < PF_2$,
 $R \leftarrow$ Least important feature (S_i)
 Remove i^{th} feature from S
 go to step 15
 else $i = i + 1$
14. End for
15. Repeat steps 1 – 14 until all $S = \phi$

5.4 Results

5.4.1 Factor/Feature identification

Among the 31 candidate variables which have been surveyed, certain variables belonging to precautionary measures such as attending community programs, finding more information about flood hazard and taken flood insurance has no or insignificant influence in reducing both building and content loss (Thieken et al. 2005) . Similarly the variable: vehicle driven to safety belonging to the domain of emergency measure do no influence the flood loss. These variables are thus removed from further analysis. In addition to it, every surveyed household, which were single family dwellings, were aware about the flooding problem in their region and all of them were flooded with water contaminated by sewage. As a result the variables characterising the above properties such as knowledge about flood hazard, building type and contamination has only got responses to one of the many categories available. Since no variability is recorded among these 3 factors, they were eliminated from further study. Thus the candidate variables which could be used for identifying the factors which influence the flood loss in Jakarta have been reduced to 24. These 24 variables are then used as inputs to Algorithm 5-3 and its performance has been compared with the existing RFE approach (Algorithm 5-2).

Due to the limited size of the dataset, a resampling strategy is used to estimate the model performance. Specifically, from the flood loss dataset by using the bootstrapping method, 90% training samples and 10% test samples are randomly generated with replacement. This process is repeated 50 times to obtain 50 subgroups of training and test set. This approach is similar to a 10 fold cross

validation (CV) but benefits from large training size and larger number of tests. In addition, the performance of the model learned will be comparable to those models evaluated from the whole dataset (Kohavi 1995). For SVR, parameters C - the error penalty parameter and ϵ - the insensitive loss function are chosen by 10 fold CV (Fröhlich et al. 2005).

In each subgroup, the training set is used to rate the factors and learn predictive models using the selected number and sequence of factors. The model prediction performance is evaluated over the test set and the average performances from the 50 test groups are used for comparison.

Predictive performances of the model are evaluated in terms of root mean square error (RMSE). Figure 5-1 & Figure 5-2 shows the prediction performance of the two algorithms using clinical kernels for different number of features. It is clear that the Hybrid mixed data RFE algorithm (Algorithm 5-3) consistently outperform the traditional RFE method for both building and content loss prediction. The proposed approach even results in better prediction, especially with a smaller set of selected factors, i.e. when the number of factors selected is less than 10 the prediction performance of the proposed algorithm is 12.5% higher than that of traditional RFE for building flood loss.

This is because, in traditional RFE factors are ranked based on the weights learned from SVR which only represent their contribution in predicting flood loss. These weights represent only the ‘importance’ of the variables in predicting the flood loss. Whereas the proposed algorithm rank the factors based on its ‘importance’ and ‘relevance’ to the response variable. Furthermore, upon restricting the number of factors to be used in the model, RFE tends to eliminate weak factor which can consequently become significant when combined with other factors, thereby leading to a reduction in predictive accuracy. This interaction among factors and response variable has been captured in the proposed approach resulting better predictive accuracy. However as more factors are included in the analysis (i.e. greater than 20), both methods produce similar results. This is attributed to the fact that, under such situations both methods retain the essential factors required for loss prediction.

Figure 5-1: Prediction performance for each size of selected factors for building loss

Figure 5-2: Prediction performance for each size of selected factors for content loss

In practice when building a prediction model, it is always preferable to use small number of factors because:

1. To overcome the problems of model overfitting.
2. It may not be economically and technically feasible to obtain information on all factors.

Thus the proposed approach is quite suitable as it identifies fewer prominent factors with better prediction accuracy. To estimate flood loss to buildings, it can be seen from Figure 5-1 that the predictive model could have the first 10 ranked factors for optimal performance. This is because the building loss estimation model upon selecting 10 factors has the best prediction accuracy (lowest RMSE values) when compared with all other combination of variables. Similarly for content loss model, the factor size determined is 7 (Figure 5-2).

Since the testing environment consists of 50 sub-groups of data, we have a total of 50 loss prediction models. In each of these models, depending on the training dataset, factors may be ranked differently. So the factors which constitute the top 10 and 7 ranking sequence for building and content dataset could remain same for some models while varying elsewhere. Thus the selection stability of the algorithm has to be assessed in order to identify the factors that need to be included in the loss prediction model.

Here we follow the Kuncheva index (KI) in assessing the selection stability (Kuncheva 2007). It measures the extent to which the N sets of m selected features share common elements.

$$KI(f_1, \dots, f_N) = \frac{2}{N(N-1)} \left\{ \sum_{i=1}^{N-1} \sum_{j=i+1}^N \frac{|f_i \cap f_j| - \frac{m^2}{k}}{m - \frac{m^2}{k}} \right\} \quad \text{Equation 5-4}$$

where k is the total number of features and $\frac{m^2}{k}$ is the factor which corrects the bias due to random chance for two feature sets f_i and f_j sharing common features. The index can take values between (-1, 1] with larger KI indicating larger number of common elements within the feature sets. Random selection is indicated with value of KI equal to 0.

Figure 5-3: Factor selection stability for building and contents

Figure 5-3 shows the factor selection stability of the Hybrid mixed data RFE algorithm for both building and content loss dataset. It can be seen that the KI is neither 0 nor 1 but in-between for both building and content loss model for the signature size of 10 and 7 respectively. This means there are only some factors which are common to all the models. In order to select the most prominent factors among all the factors belonging to top 10 and 7 ranked sequences, it is essential to obtain further insight into the composition of all factors within those sequences. As a result the factors that have been represented at least once in any of the 50 models have been identified for both building and contents. Figure 5-4 (a) & (b) enumerates these factors and their proportion within the selected factors. It can be seen that a total of 16 and 14 factors are obtained for building and content dataset respectively.

In both cases, water depth, and monthly household income are selected by all the models (50 times); hence not surprisingly, they are the most important flood loss predictors. This result is in accordance with all other flood damage estimation models available in the literature (Penning-Rowsell et al. 1977; Thieken et al. 2008)

wherein depth-damage curves are used in for flood loss estimation. This result also confirms the outcomes from the previous study (Chapter 4) wherein the study used household income as the grouping variable to demarcate the flood loss dataset into homogenous groups. Other major flood loss influencing parameters are warning time, years of stay, duration, flood experience index, raised floors, number of adults, employment, total floor area, velocity indicator, and permanent or mobile water barriers. These are the factors which have been selected by at least 50% of the models except for the variable - number of adults (40%) in content dataset. As expected apart from flood depth, other flood impact parameters such as flood duration and velocity also influence flood loss. In addition, various flood resistance parameters such as raised floors, temporary water barriers, flood experience index also play a decisive role in influencing flood loss.

However, incorporating all the factors (16 and 14) in the loss model voids the whole purpose and objective of the study. In order to avoid model overfitting problems and increase the usability of the model in real situations, it is necessary to eliminate those factors which are underrepresented among the top ranked features. Thus the factors 1, 3, 4 and 5 from building dataset and 1, 3, 4, 7 and 11 from content dataset are removed. The factors 1, 3 and 4 in building dataset i.e. moving furniture to safety, gas and sealing openings do not significantly influence building loss. So it's reasonable to remove them from further analysis. Similarly, the factors 1, 3, 7 and 11 from content dataset were also removed from further analysis. For the factor - ownership, an argument can be raised as in the person who owns the house will naturally have a tendency to protect it from flooding. However, it depends on the income level and past flood experience of the household. Again, the efficiency of the measures undertaken will in turn depend on extent and nature of the flood. In addition, these weakly represented factors are located at the bottom of the ranked sequence for both content and building dataset, i.e. they have the least importance among the prominent factors within the model for loss prediction. Thus it is logical to remove these factors from the model development process. Accordingly, the number of factors selected for modeling building and content loss is 12 and 9 respectively.

Figure 5-4: Composition of factors and their proportion in the top 10 and 14 ranked factors selected by 50 building (a) and content (b) loss prediction models using the Hy-MKr-RFE algorithm. Purple colour indicates the lesser significant factor among total factors selected for each signature sizes for building and content dataset.

5.4.2 Model development

The factors identified from the above algorithm are used for developing the building and content flood loss estimation model for residential buildings in Jakarta. Model development is employed using MKr with MKL approach to learn the optimal linear combination of 12 and 9 factors with clinical kernels. The MKr flood loss model is developed by integrating the MKL approach with SVR to model and predict building and content loss. More details on model development are provided in section 5.3.2.

Figure 5-5 shows the importance of each factor in predicting building and content loss in terms of feature weights obtained using MKL on the flood loss dataset from Jakarta. Water depth and monthly household income, followed by years of stay and total flood area has the highest influence in predicting flood loss. It can also be seen that flood early warning in combination with flood resistant parameters such as precautionary and emergency measures and past flood experience contributes significantly in determining the extent of flood loss at different depths of water.

It is clear that apart from the flood impact parameters, various resistance and socio-economic parameters also play an important role in determining the extent of flood loss. However, there still remains the question on how these parameters influence the flood loss. In order to understand the underlying relationship or association between these variables and flood loss, a correlation analysis is performed. Point biserial correlation is performed when one of the variables is a dichotomous and Pearson correlation analysis is performed when both the variables are continuous.

Figure 5-5: Feature importance for MKr flood loss estimation model for buildings (a) and contents (b)

Table 5-3 and Table 5-4 lists out the critical variables which influence the building and content loss and their relationship with loss respectively. For both building and content loss flood impact parameters such as water depth, flood duration and velocity are positively correlated, i.e. larger these factors, higher the loss. Similarly for both the content and building loss, the monthly household income is positively correlated. As expected, the precautionary and emergency measures such as raised floors, installing water barriers are negatively correlated with the flood loss, i.e. the households which have undertaken these precautionary measures have lower building and content loss. This result is confirming the outcome from the previous study (Chapter 4) on quantifying the influence of individual flood loss reduction measures. Flood warning and flood experience index also shows an inverse relation to both building and content loss. This result confirms the importance of flood early warning systems in reducing the impact from flooding and is in accordance with the studies conducted in UK (Penning-Rowsell et al. 1977).

Table 5-3: Variables influencing building loss

Predictor Variables	Type of Correlation
Employment Status	Positive
Monthly Household Income	Positive
Years of Stay	Negative
Total Floor Area	Positive
Flood Depth	Positive
Flood Duration	Positive
Flood Experience Index	Negative
Velocity Indicator	Positive
Number of Adults	Negative
Flood Warning	Negative
Water Barriers	Negative
Raised Floor	Negative

Table 5-4: Variables influencing content loss

Predictor Variables	Type of Correlation
Monthly Household Income	Positive
Years of Stay	Negative
Flood Depth	Positive
Flood Duration	Positive
Flood Experience Index	Negative
Velocity Indicator	Positive
Number of Adults	Negative
Flood Warning	Negative
Raised Floor	Negative

5.4.3 Model performance evaluation

To evaluate the performance of the MKr flood loss estimation model, its flood loss prediction accuracy is compared with the damage models developed in chapter 4 namely, the flood loss functions (depth-damage curves) and the scaled flood loss functions. Performance comparison is based on the resampling procedure which is widely followed in the field of hydrology. 100 bootstrap samples each with 100 affected household are drawn randomly with replacement from the dataset. The flood loss prediction models are then applied to each household record to estimate the building and content loss. The model performance is evaluated in term of root mean square error (RMSE), mean bias error (MBE) and the Pearson correlation coefficient. As mentioned in section 4.6, separate test-sets to evaluate the performance of the models are not employed

Figure 5-6: Comparison of RMSE values of flood loss estimation models

Figure 5-7: Comparison of correlation coefficient for the flood loss estimation models (dd-f: - depth-damage functions, s-dd-f: - scaled depth-damage functions, MKr: - Multiple Kernel Regression Model)

Figure 5-6, Figure 5-7, and Figure 5-8 compares the performance measures of the bootstrap samples for the 3 models in boxplots. In terms of predicting both the building and content flood loss, the MKr model shows a clear improvement on comparing to the established approaches (i.e. depth-damage functions). The building loss prediction accuracy of MKr model improved by 44% and 41% when compared to models developed from flood loss functions and scaled flood loss functions. The corresponding values for the content loss prediction model are 27.8% and 17.9%. Among the 3 models, the model which demonstrated the worst performance is depth-damage functions (dd-f). This is can be attributed to the fact that, it is a single dimensional model with water depth as the lone predictor.

The multivariate model, scaled depth-damage functions (s-dd-f), outperforms the simple depth-damage model except for the mean bias. This means the s-dd-f model has a slight tendency to over-predict the building and content loss from floods for some cases. The comparatively higher bias error reported for s-dd-f model could be attributed to the usage of scaling factors in accounting peoples response measures. These scaling factors are applied to the entire dataset irrespective of any weightage accounting for the number of cases in each water depth category or income groups. This can be explained by the fact that only limited data was available while deriving the scaling factor per action group (Table 4-13). As a result the derived factor can be biased. So the mean scaling factor over the entire dataset was considered as a better representation rather than deriving imprecise estimate for every water depth for each income groups. As a result the mean error can get skewed to either side depending on the number of cases available to each water depth category and income groups. However s-dd-f model falls behind unfavourably when compare to the MKr model both in terms of RMSE and correlation coefficient. The correlation coefficient of the MKr model (0.94) indicates a strong relationship between the predicted loss and actual observed loss. It also reports an improvement of 41% and 18% over s-dd-f model in predicting content and building loss.

Though s-dd-f and MKr are multi-variate model, it is practically unfair to compare both these models directly as the latter model captures the single and joint interaction of factors while predicting the losses, accordingly better prediction

performance. The MKr model was able to project the important precautionary and emergency measures such as temporary and permanent water barriers, raising the floor level at the point of entry of flood waters and to a lesser extent - moving furniture to safety which influences the extent of flood loss. These private precautionary and emergency measures could be undertaken before, during and after the flood by the people of Jakarta to reduce the negative impacts. In addition, the MKr model was able to capture some non - trivial interactions between flood loss and warning time, flood experience, duration of stay and household size which are widely overlooked in flood risk assessment.

Figure 5-8: Comparison of mean bias error (MBE) for the flood loss estimation models (dd-f: - depth-damage functions, s-dd-f: - scaled depth-damage functions, MKr: - Multiple Kernel Regression Model).

However the main drawback in the proposed hybrid mixed data feature selection algorithm is its lesser efficiency in selecting stable features. Stability Index (KI) was measured at 0.5 and 0.57 for the identified features size from the building and content dataset (Figure 5-3). As a result, additional features were forced to be incorporated into the flood loss prediction model than the numbers of factor already identified by the proposed algorithm (which was 10 and 7). This problem could be explained in terms of the data imbalance which is already present in the flood loss database. To overcome this problem, a synthetic data generation approach is proposed. This limitation along with the approach to overcome it will be explained in detail in the next chapter.

5.5 Summary

This chapter presented a novel hybrid methodology using RFE for identifying the significant factors which influence the flood loss to both buildings and contents. The approach combines the ‘factor importance’ extracted from MI & MKL and ‘factor interaction’ from the model prediction performance with RFE to identify and rank the heterogeneous features. The methodology incorporates dedicated clinical kernels for combining the categorical and continuous factors. Application of this methodology (Hy-EnRFE) to the 24 candidate variables from Jakarta flood database results in identifying 10 and 7 significant factors for building and content loss data respectively. However, due to the instability present in the feature selection algorithm due to data imbalance, a total of 12 and 9 factors are finally selected to model flood loss to building and content loss. Finally a residential flood loss estimation model is developed using MKr based on the identified factors. The prediction performance of the MKr model is shown to provide significant improvement over the state of the art methods used in estimating flood loss. It is able to capture the interaction among the factors while predicting or estimating flood loss. This characteristic of the MKr model could be advantageous in providing necessary information regarding which flood reduction measures one should employ and in other flood risk related decision making processes.

CHAPTER 6. AN INTEGRATED DATA DRIVEN AND STATISTICAL FRAMEWORK FOR DEVELOPING MULTI-VARIATE RESIDENTIAL FLOOD LOSS ESTIMATION MODEL

6.1 Introduction

The Multiple kernel regression (MKr) and Mutual information (MI) based Recursive feature elimination (RFE) method has been applied to the Jakarta flood loss data to identify the prominent factors which influence flood loss. The selected factors were then used in building the residential flood loss prediction model. This method has demonstrated a significant improvement in accuracy in predicting flood loss and proven its suitability as an alternate approach in modelling flood loss. However a major limitation has been observed while using the hybrid mixed data feature selection algorithm in identifying the prominent factors. The efficiency of the proposed algorithm in selecting stable features is quite low. This limitation is due to the data imbalance present in the surveyed flood loss database from Jakarta. This is case for most of the real life datasets wherein every factor does not possess a perfectly uniform distribution within the database. In majority of the cases, some factors or at least the combination of critical factors may be highly under represented.

The limitation of data imbalance found in the dataset is not due to the natural imbalance problems (like rare diseases, credit card frauds etc) but those problems where it's very expensive and tedious to obtain adequate number of samples for the minority categories. These imbalanced dataset thus cannot represent the broad characteristics of the problem and the complexity of the concept. As a result, it will compromise the efficiency of the feature selection algorithms thereby making it unable to detect critical factors efficiently. Accordingly, the performance of learning algorithms as well as the selection stability depreciates, resulting in either suboptimal prediction accuracy or poor judgement in decision making.

To overcome the problem of data imbalances a Markov Chain Monte Carlo (MCMC) based sampling method is proposed in this study for generating synthetic data. In addition, a novel integrated data driven and statistical framework is developed for identifying the most prominent factors affecting flood damage. These factors are then used in developing a multivariate residential flood loss estimation model. The proposed framework mitigates the limitations of data imbalance, offers a cheap alternative in identifying the factors influencing flood loss and develops a robust and comprehensive residential flood loss estimation model.

6.2 Proposed Enhanced Sample Generator

When dealing with data samples that has high disparity in its distribution across the domain, it's quite probable for the learning machine to predict a rare event wrongly. Consider an example dataset with 98% of type 'A' instances and 2% of type 'B' instances. A predictor or classifier learned in the traditional manner in this dataset will predict or classify all the dataset as type A with an overall accuracy of 98%, since only the 2% of type 'B' instances are predicted incorrectly. The reason being, the standard prediction algorithm is overwhelmed with type 'A' instances thereby ignoring the less represented 'B'. In real life situations, this can lead to serious consequences compared to mispredicting a frequent event.

The flood loss dataset obtained as a result of direct household survey after the 2013 January floods in Jakarta is clearly disproportionate. Of the 491 household surveyed the proportion of household which has undertaken one or the other precautionary measures ranges from as low as 8.1% to 20.7% (Table 4-3). This proportion becomes even more insignificant when the loss cases are categorised into different water depths. The situation is the same for household which has undertaken one or the other emergency measures where the range varies from 8.5% to 64.5% (Table 4-2). The percentage of household which has undertaken more than one measures or a combination of precautionary and emergency measures is also quite low. Regarding the ownership of the building 92.6% household owns the house while the remaining are a rented properties. Of the total surveyed households, almost half of them are multi storied. In addition, the proportion of household which has received

any form of flood warning is only 11.2%. From this diverse and disproportionate dataset, it can be argued that the prediction performance of traditional algorithms will not be up to the mark. This has been proved by Barandela (2013) for classification problems.

Various methods are available in the literature to tackle this issue of data imbalance both at data level and algorithmic level. At the algorithmic level, the approach is to adapt the algorithms to perform better with the available data. This can be done by using ensemble methods such as bagging and boosting techniques or cost sensitive learning approaches etc. Whereas at the data level, the approach is to adjust minority and majority samples by data re-sampling methods such as under-sampling and over-sampling of data or synthetically creating the minority samples (Chawla et al. 2002). Though algorithmic level approaches yield better results, it exclusively creates a local predictive model as it learns from the existing dataset thereby lacking generality. Also all these methods tend to alleviate the problem by relying only on local information of the samples. So in this study, a data level probabilistic approach has been employed to handle data imbalance by synthetically generating a global representation of new samples.

6.2.1 Gibbs sampler

A Markov Chain Monte Carlo (MCMC) based sampling method used in this study for generating and strategically selecting new samples. The idea behind MCMC algorithm is to collect samples from a target distribution by constructing a Markov Chain (MC) with stationary distribution. A MC is mathematical model comprising of a finite number of states having a specific property that its future is conditionally independent of the past given its present (Geyer 2011). A sequence of random variables $\{x^{(1)}, \dots, x^{(k)}\}$ is defined as a first order MC if the Equation 6-1 (Markov property) holds true for all $k \in \{1, \dots, K - 1\}$.

$$P(x^{(k+1)} | x^{(1)}, \dots, x^{(k)}) = P(x^{(k+1)} | x^{(k)}) \quad \text{Equation 6-1}$$

The numbers $P(x^{(k+1)}|x^{(k)})$ are called the transition probabilities. So by knowing the probability distribution of first variable and the transition probabilities, a MC can be established.

Figure 6-1: Sample markov chain

For this study, using a Gibbs sampler, multivariate samples are synthetically generated by factoring the joint probability distribution of data attributes. Gibbs sampling has originated with the Metropolis algorithm (Metropolis et al. 1949) and can be used when direct sampling is not possible. In Gibbs sampling, rather than selecting the values for entire set of attributes simultaneously, the idea is to formulate a probabilistic choice separately for each of the ‘m’ dimensions based on the values of other m-1 dimensions and the previous value of the selected dimension. Thus the Gibbs sampler generates a MC whose sample converges to a target distribution. The Algorithm 6-1 below describes the standard Gibbs sampler. The new samples can be generated probabilistically by exploiting the conditional distribution of each dimension as explained in step 4.

Algorithm 6-1: Gibbs Sampler

- 1 $x^{(0)} := \langle x_1^0, x_2^0, \dots, x_k^0 \rangle$
- 2 *Do for every iteration till T*
- 3 *for i = 1 to k do*
- 4 $x_1^{(t+1)} \sim P(X_i | x_1^{t+1}, \dots, x_{i-1}^{t+1}, x_{i+1}^t, \dots, x_k^t)$
- 5 *end for*
- 6 *end for*

One of the major barriers in successful implementation of the Gibbs sampler is in estimating the underlying n -dimensional conditional probability distribution from samples (i.e. step 4). In addition to it, due to the imbalance in the dataset, the Gibbs sampler (sampling in step 4) is much less likely to explore the entire space of attributes and is therefore less likely to generate underrepresented data. To overcome this problem Das et al. (2015) proposed to use the Chow-Liu algorithm (Chow et al. 1968) to sample from the entire state space by factoring the joint distribution of the variables into a tree. Chow-Liu algorithm explicitly imposes dependencies between attributes using a tree structure. However the Chow-Liu approximation has the following drawbacks.

- The factorisation can only represent a first-order dependency tree. Since higher order interactions are not accounted, the resulting tree structure cannot capture multi-dimensional interactions among the original variables thereby missing the complete underlying structure
- It is incapable of handling mixed data attributes and is suitable for classification problems only.

Thus the use of Chow-Liu tree violates the main objective of the study, which is to identify the prominent flood loss influencing factors by considering their single and joint interaction. In addition, the flood loss estimation problem is a mixed data (both continuous and categorical predictors) regression problem in comparison to the discrete data classification problem dealt by Das et al (2015). So the above framework by Das et al (2013) has to be extended:

- To cater mixed data problems,
- To capture higher order interactions among variables and
- To address regression problems.

As a result, in this study a Hybrid Bayesian networks (BNs) instead of Chow-Liu tree is proposed. Gibbs sampler can then be built around the Hybrid-BNs for sample generation. Bayesian networks (BNs) or Directed acyclic graph (DAG), which impose explicit dependencies among features/attributes can be used to thoroughly explore the entire domain of the attributes and to generate data samples which are

consistent with underrepresented features. They acts as the MC (Step 4) on which the Gibbs sampling can be applied to synthetically generate samples.

6.2.2 Hybrid Bayesian networks

Direct acyclic graphs (DAGs) or Bayesian networks (BNs) belongs to the family of probabilistic graphical models (PGMs) which represents the knowledge concerning any domain under uncertainty using graphical structures (Jensen et al. 1990; Koller et al. 2009). The graphical interpretation of DAGs based on nodes and arcs explain the qualitative information i.e. independence relations (Figure 6-2 (a)) between the variables and the quantitative information concerning each variable is specified using the conditional distribution $p(X_i|pa(X_i))$ (Figure 6-2(b)). $pa(X_i)$ represent the parent set to each variable X_i . DAG decomposes the joint probability distribution over all variables into a product of the conditional distribution associated with each node:

$$p(X) = \prod_i p(X_i|pa(X_i)) \quad \text{Equation 6-2}$$

For the example network in Figure 6-2(a) the joint probability distribution can be specified as $p(V_a, V_b, V_c, V_d) = p(V_a) \cdot p(V_b|V_a) \cdot p(V_c|V_a) \cdot p(V_d|V_b, V_c)$

(a) Qualitative

$$\begin{aligned} P(V_a = 1) &= 0.80 \\ P(V_b = 1|V_a = 1) &= 0.80 \\ P(V_b = 1|V_a = 0) &= 0.20 \\ P(V_c = 1|V_a = 1) &= 0.80 \\ P(V_c = 1|V_a = 0) &= 0.20 \\ P(V_d = 1|V_b = 1, V_c = 1) &= 0.80 \\ P(V_d = 1|V_b = 0, V_c = 1) &= 0.60 \\ P(V_d = 1|V_b = 1, V_c = 0) &= 0.60 \\ P(V_d = 1|V_b = 0, V_c = 0) &= 0.20 \end{aligned}$$

(b) Quantitative

Figure 6-2: Illustration of Bayesian network with four binary variables

For any given scenario, efficient reasoning can be carried out by BNs under uncertainty. This property of the BNs is referred to as probabilistic inference or propagation. Traditional BNs were designed to handle discrete variables and much studies have been carried out for the probability propagation in such networks (Lauritzen et al. 1988). However in most of the real world datasets, it's natural to find discrete and continuous variables simultaneously. BNs which deal with mixed data types are referred to as Hybrid Networks. They handle the continuous variables either by discretization and then treating as if they were discrete or employing a Conditional Gaussian (CG) model to represent the probability distribution. However the direct discretization approach results in loss of information thereby reducing the precision of model (Kozlov et al. 1997). On the other hand, in CG models discrete variables are not allowed to have continuous parents (Lerner et al. 2001) consequently making it incompetent in applying to flood dataset.

These drawbacks were overcome with the use Mixture of Truncated Exponentials (MTE) as an alternative to discretization in representing the joint distribution of variable in the network (Moral et al. 2001; Cobb et al. 2004). The basic idea behind MTE model is in representing the domain of continuous variables using exponential functions. These functions, which are closed under combination, restriction and marginalisation, improve the accuracy of the model with better fitting properties (Romero et al. 2006; Rumí et al. 2006). The MTE model can be characterised by MTE potential and its corresponding density. For probability inference processes, standard propagation algorithms as well as approximate methods can be performed over MTE potentials (Cobb et al. 2006; Rumí et al. 2007). More details on MTE potential, density, conditional MTE density and probability propagation with MTE can be found in Rumí et al. (2007) and Romero et al. (2006).

The flood loss dataset obtained from the outcome of household survey for 2013 January flood consists of variables which are both discrete and continuous with continuous response variable i.e. flood loss (Content and Building loss) (Table 5-1). Since the variable of interest being continuous, we are here dealing with regressing problem (Hastie et al. 2009). Though classification and regression problem have been analysed by BNs, only a single study has considered BNs in flood loss

modelling (Vogel et al. 2012). The study employed discretization in handling continuous variables and the target variable i.e. flood loss has been considered as a class variable (classification problem). No attempts have been made till now in literature to study the relationship between the various flood loss influencing parameters using Hybrid BNs for regression problems. The advantages in using Hybrid BN based regression model are: (Roperio et al. 2014)

1. Without having full observation of variables, it is still possible to predict the response/predictant variable.
2. From a semantic perspective, the model is richer.

For the Hybrid MTE BNs for regression, the ideal way is to learn the structure and the parameters from data without any restrictions. However this approach is usually not feasible due to the limited availability of data. Instead the structures of the networks are fixed and the parameters and the relationships are learned from the data. This method reduces the computational cost involved in estimating the conditional density $f(\text{Response}|\text{variables})$ using the factorisation resolved by the BN's structure.

An extreme case of Hybrid MTE based BNs for regression is the Naïve Bayes (NB) structure where the predictand variable forms the single root node and all other variables have only one parent (the response variable). It assumes strong independence between the variables given the response variable, thus the name naïve (Figure 6-3). However if the dependence or the correlation between variables could be incorporated in the networks i.e. by permitting features to have more than one parent in the network, the error in the regression would diminish. More importantly, it helps in capturing multi-dimensional interaction among the variables. Several such structures are available for BNs for instances, Forest Augmented Naïve Bayes (FAN), kDB, Tree Augmented Naïve Bayes (TAN) etc. These models are richer with better representation of the underlying structure of the problem.

Figure 6-3: Naïve Bayes model - structure

Figure 6-4: TAN model - structure

In this study the objective is to develop a Markov Chain using BNs which can simultaneously handle continuous and discrete variables so as to synthetically generate samples. Thus it is proposed to use MTE-TAN regression model (Figure 6-4) as MC for both building and content loss. TAN model is so selected that the structure is complex enough to capture the single and joint interactions of variables better than NB and computationally simple when compared to kDB and FAN model.

6.2.3 TAN model for regression

The explanatory variables which are used to develop the Hybrid BNs for building and content loss are provided in Table 5-1. A 5 parameter MTE model as shown in Equation 6-3 is selected to fit the domain of the variables. This model can accurately fit most of the common distributions and at the same time the complexity of the model and amount of parameters required to evaluate is less (Cobb et al. 2006; Romero et al. 2006; Rumí et al. 2006).

$$f(y) = a_1 + a_2 e^{a_3 c} + a_4 e^{a_5 c} \quad A < y < B \quad \text{Equation 6-3}$$

where the density is estimated in the interval defined by A and B and a_i , $i=1$ to 5 are the coefficients.

By fixing the network structure as TAN, the task of learning the model from data has been simplified. The MTE-TAN regression model was learned from data using the Elvira software (Elvira Consortium 2002). Algorithm TAN-MTE in Appendix B describes the detailed steps involved in constructing the TAN-MTE model. Accordingly hybrid regression models were learned for building structure loss (Figure 6-6) and content loss separately (Figure 6-6).

Once the MTE model and the TAN network has been learned from the data, the predicted values for the variables ‘ x ’ can be estimated from the distribution using its expectation. The expected value for the random variable X whose density is defined as in Equation 6-3 is estimated as

$$\begin{aligned}
 E[Y] &= \int_{-\infty}^{-\infty} yf(y)dy = \int_A^B x(a_1 + a_2e^{a_3c} + a_4e^{a_5c})dy & \text{Equation 6-4} \\
 &= a_1 \left(\frac{B^2 - A^2}{2} \right) + \frac{a_2}{a_3^2} ((a_2B - 1)e^{a_2B}) - (a_2A - 1)e^{a_2A} \\
 &\quad + \frac{a_4}{a_5^2} ((a_5B - 1)e^{a_5B}) - (a_5A - 1)e^{a_5A}
 \end{aligned}$$

Thus the approach used in generating synthetic samples can be pictured using the following four steps

1. Create Hybrid MTE-TAN regression model for both building and Content loss as explained in section TAN model for regression (Section 6.2.3) – These act as the Markov Chains
2. Apply the MCMC algorithm called Gibbs Sampler (Section 6.2.1) to the MC (used in step 4 of MCMC algorithm) developed in step 1
3. In order to reduce time and the computational complexity, synthetic data are only generated for the underrepresented samples. To induce faster convergence of the algorithm, minority data points are provided as the initial samples as suggested by Das et al. (2015).
4. Select samples from the Gibbs Sampler and mix it with the original data set for further analysis.

The synthetically generated samples are then provided as input to the proposed novel Hybrid mixed data feature selection (Algorithm 5-3). Section below presents the complete framework in identifying the key flood loss influencing factors and developing a multi-variate residential flood loss estimation model.

Figure 6-5: Hybrid MTE-TAN regression model for content loss

6.3 Framework Process

An integrated data driven and statistical framework is proposed for identifying the prominent factors which influence both building and content loss from flooding. These significant factors are then used to develop a multivariate residential flood loss estimation model. The overall framework is shown in Figure 6-7. Each component of the framework is described below:

(1) Flood loss database

To implement the proposed framework, first and foremost a detailed flood loss database has to be created. Loss information can be obtained from any of the following sources such as: insurance companies, government agencies, field survey etc. These loss cases have to be complemented with information regarding damage influencing parameters (Table 5-1) which can be obtained via telephonic or household survey (Thieken et al. 2005; Vogel et al. 2012). If the database is developed via survey, it is important to take into account the sampling distribution in order to avoid bias. Once the database is compiled with the information regarding the attributes influencing flood loss, next steps can be performed.

(2) Construct Hybrid Bayesian networks

The second operation in the framework is to transform the flood loss database into BNs (for building and content separately) which represents the knowledge concerning any domain under uncertainty using graphical structures. In this phase, the dependencies among the attributes and with building/content loss are captured explicitly by using BNs. Various approaches for learning BNs from data are available in literature (Koller et al. 2009). In this framework, Hybrid MTE-TAN regression model (Fernández et al. 2008) is proposed for generating the network for building and content loss separately from the database. Often the attributes/variables dealing with flood loss are of mixed discrete and continuous types, thus this approach has been chosen.

Figure 6-7: Flowchart of the integrated data driven statistical framework

(3) Generate synthetic samples

The next step is to conduct a stochastic data generation process by applying appropriate sampling methods in order to obtain multivariate samples. In this framework, a MCMC based sampling method known as Gibbs sampler (Metropolis et al. 1949) is applied to generate multivariate samples synthetically by factoring the joint probability distribution of data attributes. The BNs learned from the previous step will act as the MC. This network decomposes the joint probability distribution over all variables into a product of the conditional distribution. Gibbs sampling will utilize the joint probability distribution determined by the BNs to generate new samples.

(4) Estimate factor importance

After the samples have been generated, in order to identify the prominent factors influencing flood loss, the importance of each variable corresponding to the building and content loss has to be estimated separately. In this framework, factor importance is computed from the factor weights and factor relevance. The factor weights are estimated based on the contribution of each factors in predicting the flood loss using MKr (Varma et al. 2009; Herrera et al. 2014) and the factor relevance are estimated using Kraskov (Kraskov et al. 2004) and Gomez (Gómez-Verdejo et al. 2009) MI estimator. Summation of the normalized values of each index is used to represent the total factor/feature importance index.

(5) Extraction of prominent factor

Synthetically generated samples and associated index for factor/variable importance obtained from the previous step are provided as inputs to the novel Hybrid mixed data feature selection (Algorithm 5-3). The proposed algorithm in this framework identifies the factors whose single and joint interaction can explain and better represent the flood loss in an iterative manner. For each iteration, Step 4 in the framework will be executed to

estimate the new factor importance before applying the proposed algorithm. More details on the algorithm can be found in Section 5.3.3 in CHAPTER 5

(6) Generate flood loss estimation model

Once the factors influencing the building and content loss has been estimated, the multi-variate residential flood loss estimation model is developed using MKr. The MKr model with the factors identified in Step 5 will be trained using the flood loss data. This trained model will be used to predict or estimate the flood loss under any scenarios.

6.4 Results and Discussions

6.4.1 Factors influencing flood loss

Among the 31 candidate variables, 24 variables as explained in section 5.4.1 are selected for identifying the significant factors that influence flood loss. Since the flood loss database is highly diverse and disproportionate, not all variables enjoy a perfectly uniform distribution within the dataset. For this study the minority variables are identified as those variables whose distribution within each water level groups in the loss dataset is less than 10%. 10% has been considered as the standard measure in representing minority class sample in classification problems (Das et al. 2015). So the same measure has been used for the regression problem too. Water level groups are the same 5 groups which have been used in developing the flood loss functions as in Chapter 4. Table 6-1 enumerates the list of the underrepresented factors along with the domain to which they belong in the flood loss database. The list is mostly made up of variables belonging to precautionary and emergency measures. This is quite evident from the fact that only 20.7% of surveyed household has undertaken at least one precautionary measure.

Table 6-1: Minority factors and the domain to which they belong within the flood loss database

Variables/Factors/Features	Domain
Temp. Water Barriers	Emergency Measures
Openings sealed	
.....	
Raised Floor	
Permanent Water Barriers	
Watertight Shields	Precautionary Measures
Low value furniture	
Valves	
.....	
Flood Warning	Warning
.....	
Velocity Indicator	Flood impact Parameter

To overcome the data imbalance within the flood loss database, synthetic samples corresponding to the identified minority factors are generated using the Gibbs sampler. Hybrid MTE-TAN regression BNs developed for building (Figure 6-6) and content (Figure 6-5) loss will act as the MC for sampling data by factoring the joint probability distribution of the variables. Once the data are generated, it is then mixed with the original flood loss dataset. This enhanced dataset is then fed as input to the hybrid mixed data feature selection (Algorithm 5-3) to identify the prominent flood loss influencing factors.

To identify the critical factors, a resampling procedure is employed. 50 set of bootstrap samples consisting of 90% training data and 10% test data are drawn randomly with replacement from the dataset. Training data is used for learning the loss prediction model and also to rank the variables/feature according to its importance. The performance of the prediction model is then evaluated over the test set. This process is repeated over the 50 subgroups of bootstrap samples for each size of the selected features and the average performances from the 50 test groups are used for comparison. Predictive performances of the model are assessed in terms of Root mean square error (RMSE). The factor signature size factors (i.e. the number of factors used in the prediction model) at which the RMSE values reaches its lowest or “plateau” off is selected as the minimum number of factors/variables to be used for flood loss modelling.

Figure 6-8: Prediction performance for each signature size of building and content loss dataset

Figure 6-8 shows the average prediction performance of the Building and Content flood loss model for different number of selected features. The signature size at which RMSE values have reached its minimum for building and content loss models is 7 and 10 respectively. Thus a minimum of 7 and 10 factors are necessary to model the building and content flood loss in Jakarta.

Since the study environment consists of 50 sets of bootstrap samples, each of the 50 data driven models learned from the training dataset can rank the importance of factors differently. In order to compare the variable/factor selection stability of the proposed algorithm with the enhanced dataset, Kuncheva Index (KI) is calculated for each signature size of features from all the 50 models. Figure 6-9 & Figure 6-10 compares the stability of the Hybrid mixed data feature selection algorithm (Algorithm 5-2) for the enhanced and the surveyed flood loss database for building and content respectively. From both the figures it is clear that there is substantial improvement in the factor selection stability of the proposed algorithm for the enhanced dataset especially in selecting the top 50% of the prominent factors influencing flood loss. Stability index for the building and content model is 0.89 and 0.90 at the signature size of 7 and 10 respectively. The high value for KI means that there exists only a slight variability among the factors selected by all 50 building and content loss models from the enhanced dataset. Figure 6-11 (a) and (b) portrays the significant factors and their composition within the factors selected by both the flood loss models. Since the representation of the factor - duration of flooding, is quite less within the building loss model, it is decided to eliminate this factor from further analysis. The effect of duration of flood damage and loss cannot be pronounced immediately, it seems logical to eliminate the variable duration of inundation from further analysis. Similarly, for the development of content loss model, the variable: velocity is similarly eliminated from further analysis. Thus a total of 7 and 10 significant factors are identified by the feature selection algorithm from the Jakarta residential flood loss database.

Figure 6-9: Comparison of factor selection stability of the hybrid mixed data feature selection algorithm for the surveyed (original) and enhanced Building dataset

Figure 6-10: Comparison of factor selection stability of the hybrid mixed data feature selection algorithm for the surveyed (original) and enhanced Content dataset

For building and content loss models, water depth and income has been identified as the most important flood loss predictors. These variables are being ranked 1st and 2nd respectively by all the 50 models. This result is in agreement with all other flood loss estimation models available in the literature (Penning-Rowsell et al. 1977; Thieken et al. 2008) wherein depth-damage curves are used in for flood loss estimation. These variables are then followed by precautionary and emergency measures. Among the precautionary measures, waterproofing of the buildings has been identified as the important factor by the building loss prediction model whereas raised floors and usage of low value or easily movable furniture in the first floor are selected by the content loss model. Within the domain of emergency measures both the models identified one factor each. The corresponding factor recognized by building and content loss models is the usage of temporary barriers and moving furniture & valuables to safety respectively. The importance of precautionary and emergency measures has already been illustrated in Chapter 4 by means of scaling factors whereas here, the specific measure which significantly influences the extent of flood loss has been identified. Flood experience index has been selected by both the models. Apart from the above mentioned factors, other important factors which influence building loss are velocity of flood water, duration of inundation and years of stay in the building. Similarly the factors which influence the content loss are warning time, number of storeys, duration of flooding and number of adults. As expected apart from the flood impact parameters, various factors belonging to the flood resistance, flood warning and socio-economic parameters have been identified in playing a decisive role in determining flood loss.

The proposed hybrid mixed data feature selection algorithm when implemented on the dataset infused with synthetically generated samples was able to capture a stable and global representation of factors which affects the flood loss. These factors are generally overlooked while estimating flood loss and in conducting flood risk assessment. The outcome of the proposed framework recommends the people of Jakarta to invest their time and money on the above identified private precautionary and emergency measures

Figure 6-11: Composition of factors and its proportion among the top 7 and 10 selected factors by the 50 building (a) and content (b) loss prediction models using the propose Hy-MKr-RFE algorithm. Purple colour indicates the lesser significant factor among total factors selected for each signature sizes for building and content dataset.

Apart from the flood impact parameters, it is apparent that various resistance, flood warning and socio-economic parameters play a crucial role in determining the degree of flood loss. However, the nature of interaction of these parameters with the flood loss still remains unknown. In order to understand the underlying relationship or association between these variables and the corresponding building and content loss, a correlation analysis is performed on the enhanced dataset. Point biserial correlation is performed when one of the variables is a dichotomous, while the other is continuous and Pearson correlation analysis is performed when both the variables are continuous.

Table 6-2 and Table 6-3 enumerate the significant variables which influence the building and content loss and their relationship with loss respectively. Monthly household income along with the flood impact parameters such as water depth and velocity shows a positive correlation with building loss. For content loss, instead of velocity, the factor flood duration along with income and flood depth is positively correlated. Thus higher the values of flood impact parameters and richer the households, more is the experienced flood loss. The strong correlation of flood impact parameters and flood loss is in accordance with the findings reported in chapter 4 and 5 and that of Thieken et al. (2005) for Germany and Penning-Rowsell and Green (2000) for UK. The importance of precautionary and emergency measures in reducing the building damage from floods has been demonstrated earlier by Kreibich et al. (2005). The outcome from this study for Jakarta confirms the same for building loss and extends the finding for content loss. The factors water barriers and waterproofing exhibits a negative correlation with building loss. In addition, the content loss is also negatively correlated to the following precautionary and emergency measures: raised floor, low value furniture and moving furniture to safety. This result also confirms the outcome from the previous study (Chapter 4) on quantifying the influence of individual flood loss reduction measures.

Flood warning, flood experience index, number of adults and number of storeys shows an inverse relation to content loss. This outcome hints that a larger household having multiple storeys with previous flood experience can have significantly

reduce content loss, confirming expert judgement and study by Marfai et al. (2015) The importance of flood early warning systems in reducing the impact from flood and is in accordance with the studies conducted in UK (Penning-Rowsell et al. 1977). The proposed framework was able to identify the specific private precautionary and emergency measures that would be useful in reducing flood damage for the residential sector of Jakarta.

Table 6-2: Factors influencing building loss

Predictor Variables	Type of Correlation
Monthly Household Income	Positive
Years of Stay	Negative
Flood Depth	Positive
Flood Experience Index	Negative
Velocity Indicator	Positive
Water Barriers	Negative
Waterproofing	Negative

Table 6-3: Factors influencing content loss

Predictor Variables	Type of Correlation
Monthly Household Income	Positive
Flood Depth	Positive
Flood Duration	Positive
Number of Storeys	Negative
Low value furniture	Negative
Flood Experience Index	Negative
Number of Adults	Negative
Flood Warning	Negative
Moving furniture and objects to safety	Negative
Raised Floor	Negative

6.4.2 Model development

The significant factors which influence the flood damage to buildings and contents as obtained from the above step are used in developing the corresponding residential flood loss estimation model for Jakarta. Model development is employed using MKr by integrating the MKL approach with SVR. MKL will learn the optimal linear combination of 7 and 10 factors with clinical kernels. The resulting heterogeneous kernel matrix is then integrated with SVR for prediction purposes. The flood loss prediction model is then learned from the enhanced dataset. This data driven Enhanced-MKr model (En-MKr) could then be used for prediction or estimating flood loss. More details on model development can be found in section 5.3.2.

6.4.3 Model performance evaluation

To evaluate the performance of the En-MKr model, its prediction accuracy is compared with the MKr model developed in chapter 5. Prediction performance is evaluated using the same resampling strategy as explained in section 4.6. 100 bootstrap samples each with 100 affected household are drawn randomly with replacement from the dataset. The En-MKr and MKr model are first learned from their respective datasets and the learned models are used to estimate the building and content loss for each record of the randomly drawn samples. This process is repeated 100 times and the predictions are quantified using the following error measures namely; Root Mean Square Error (RMSE), and the Pearson correlation coefficient.

Figure 6-12: Comparison of RMSE (a) and Pearson correlation coefficient (b) of flood loss estimation models (MKr: - Multiple Kernel regression model, En- MKr: - Enhanced Multiple Kernel regression model).

Figure 6-12 (a) & (b) compares the performance measures of the bootstrap samples for the 2 models in box plots. In general the spread in the prediction error for En-MKr model is quite low. This model also displays a slight improvement in predicting the content loss. Overall both the plots indicate that the two models have very similar prediction performance. This is quite an interesting observation as the En-MKr model, the one learned from the enhanced dataset, is able to predict the flood loss with accuracy very much similar or better than the MKr model that has been learned from the original dataset. This result thus confirms the hypothesis on the inclusion of the synthetic data generated using Gibbs sampler yields in better results. In addition, it also confirms that generation of synthetic samples using TAN-MTE BNs which incorporates the higher order interaction through joint probability distribution identifies the prominent factors which influence flood loss globally.

Thus for Jakarta, as evident from this study, the government or other agencies working on flood risk management should invest more efforts in flood warning and dissemination of specific information for reducing the impacts. The specific measures, as identified from this study are:

- Usage of temporary water barriers

The barriers can prevent the entry of floods of low magnitude into the buildings and can reduce the velocity of the incoming flow thereby lowering the impact on buildings and contents

- Waterproofing of buildings

Installation of waterproof membranes and applying waterproof paints along the external surface of the wall should be sought for the building located in regions experiencing frequent floods. However, due to the costs involved, this measure may not be possible for everyone.

- Construction or availability of additional storey

Provision of additional storey can reduce the extent of content damage provided adequate dissemination of flood warning information. Household can move their valuables to the higher storeys at the time of flood.

- Raising the floor level especially at the point of entry for flood water

Though this measure is quite expensive, it is one of the most effective methods for reducing flood loss. It requires elevating the lower levels of the building so as to forbid the entry of water. This measure has significantly reduced the loss to both buildings and contents in Jakarta. However, it can make the interior of the buildings unsightly.

- Flood prone floors used with easily movable or low value furniture/objects

Content damage can be reduced significantly by promoting this flood adapted usage of inventories within the households. By using low cost interior accessories and furniture the cost of flooding can be reduced significantly even when no flood warning is issued.

- Moving furniture and other valuables to safety

This emergency measure when provided with ample warning time and larger number of households can reduce the impacts from flood by moving the susceptible objects to higher levels or higher grounds.

In order to understand and quantify the impacts of mitigation measures in reducing flood damage in Jakarta, two scenarios were identified and evaluated. The first scenario is a 'do nothing scenario' which is characterised by the absence of the mitigation measures identified by our algorithm. The second scenarios describes the situation as characterised by the implementation of effective mitigation measures such as using low value furniture, presence of second storey, elevated floor level at the point of entry of water and moving furniture to safety. These scenarios have been evaluated for the three different income groups identified in Chapter 4 for Jakarta at different water depths.

The reduction in loss achieved for the three different income groups of Jakarta at different water levels while implementing and non implementing mitigation measures can be seen from Figure 6-13. Dashed lines represent the do nothing scenario and the solid lines represents the second scenario. It is evident from Figure 6-13 that significant reduction in damage can be achieved by implementing proper mitigation measures such as those identified by our algorithm. Application of mitigation measures have helped in reducing the damage on an average by 41% for low income group, 23.7% by medium income group and 23.1% by high income groups. Despite the increase in flood hazard characteristics, implementing adequate mitigation measures is always the key in reducing the extent and nature of flood damage and loss. Thus, this new scenarios can promote adaptation of mitigation measures for reducing flood loss.

Figure 6-13: Effectiveness of mitigation measures in reducing content damage for the three income groups identified for Jakarta

6.5 Summary

An integrated data driven and statistical framework for developing a multi-variate flood loss estimation model is presented in this chapter. Within this framework is an enhanced Gibbs sampler based probabilistic method for generating synthetic data for underrepresented factors. Key aspect of this approach is the incorporation of Hybrid MTE BNs as the Markov Chain in Gibbs sampler. This model captures interaction among heterogeneous factors while predicting/estimating flood loss. The synthetic data mixed with the original data is fed into the Hy-EnRFE algorithm to identify stable and significant damage influencing factors. They are then used to build the flood loss estimation/prediction model for both buildings and contents using MKr model.

The framework based on kernel methods and Bayesian networks identified the following variables which significantly influences the flood loss to buildings and contents: Flood depth, monthly household income, flood experience index and specific actions belonging to precautionary and emergency measures. However no characteristic relationship has been identified for velocity and duration of flooding with the flood loss. This can be attributed to the fact that:

- The surveyed dataset lacked the presence of cases belonging to high velocity class and
- The impact of duration on flood loss cannot be pronounced immediately after the flooding. It takes months or years to identify the impacts arising due to the duration of flooding.

Overall, the outcome of the framework when applied on the Jakarta flood loss database presented herein demonstrates significant improvement in the factor selection stability of the proposed algorithm. In addition, the resulting flood loss estimation model displayed similar prediction performance to the model learned from the original database. This outcome is highly beneficial for multi-variate modelling of flood loss in any regions especially when dealing with a database having very high dimensionality and small sample sizes. The proposed framework had the advantage in exploiting the global characteristics from the surveyed dataset.

CHAPTER 7. CONCLUSIONS AND RECOMMENDATIONS

Flood loss estimation models, such as depth-damage curves or loss functions solely explore the relationship between water depth and direct flood damage & loss while predicting and estimating the loss. However, various resistance and loading parameters have been identified in the literature which plays significant role in defining the extent and nature of damage and loss from flooding. In spite of this many flood loss estimation models do not take into account their role while assessing the loss. Unfortunately, the approaches followed by the available multi-variate models are extremely subjective and are dependent on expert's opinion in identifying the critical factors. In addition, no methodology currently exists in identifying a global representation of critical flood loss influencing factors from a data deficient environment. Apart from this no comprehensive flood loss estimation model has been developed for Jakarta, Indonesia. Thus there exists a need to address these basic concerns with regard to the development and use of multi-variate flood loss estimation models in flood risk management. The focus of this thesis was on these issues which are listed as objectives in Chapter 1.

7.1 Conclusions

The important conclusions obtained from this research are given below:

7.1.1 Development of residential flood loss estimation model for Jakarta

- a. The standardized procedure when applied on the flood loss database of Jakarta demonstrated monthly household income, precautionary and emergency measures are the appropriate variables for categorizing flood loss data into homogenous groups. Among these variables, using monthly household income as similarity measure yielded three different income groups. Loss functions were developed for these 3 income groups and scaling factors were derived to represent the influence of people's action to reduce the impact of flood.
- b. Compared to the mean loss functions, households which have undertaken excellent measures have seen a reduction in building and content damage by

9% and 15% respectively. In addition, those households which haven't responded to floods have their content loss magnified by 13%. The basic multi-variate model called the Scaled Flood Loss model developed in Chapter 4 was able to improve the prediction accuracy of content loss estimation model by 12.5%. Such an improvement, using a model developed on the basis of individual interaction of variables, has demonstrated the capability of imparting notable enhancement in the approaches used for developing flood loss or damage functions.

7.1.2 Identification of prominent factors influencing building and content loss from floods

- a. The outcome from this study on factor selection identified only 12 and 9 variables as the significant factors which effect building and content loss respectively for Jakarta, instead of using the entire plausible variables in the flood loss estimation model. Apart from water depth, parameters such as velocity and duration of flooding have been identified as the other significant flood impact parameters that effect flood loss. The resistance parameters which can significantly reduce the impact from flooding were determined to be: flood experience index, warning time, raised floors and temporary water barriers. In addition the algorithm also identified the interaction among the socio-economic characteristics of the household to flood loss. Variables such as household income and household size were identified in playing a significant role in determining flood loss.
- a. Among the two factor selection algorithms, the algorithm proposed by us called the Hybrid mixed data RFE (Hy-MKr-RFE) was found to be appropriate in selecting minimum number of significant factors with better loss prediction. This result was confirmed by comparing the prediction performance of the multi-variate model developed with the factors (identified via Hy-MKr-RFE) using Multiple Kernel Regression (MKr) with flood loss functions (dd-f) and scaled loss functions (s-dd-f). Overall an improvement of 41% and 17% in the prediction accuracy has been determined for building and content loss respectively.

7.1.3 Integrated data driven and statistical framework for developing multi-variate flood loss estimation model in a data deficient environment.

- b. The results from this study demonstrated the improvement in the factor selection stability of the propose Hy-MKr-RFE algorithm. By generating synthetic data using BNs and Gibbs sampler, the stability index of the proposed algorithms on buildings and content dataset improved from 0.51 to 0.89 and 0.58 to 0.90 respectively.
- c. Moreover, the flood loss estimation model (En-MKr) developed from factors identified by implementing the novel framework on the data from Jakarta has demonstrated comparable performance with the model (MKR) learned from the original dataset. This result implies that the enhanced dataset has been able to capture the global distribution of flood loss from Jakarta and the resulting prominent factors are representative for the whole of Jakarta.
- d. Apart from water depth and income the significant factors which influence the building loss in Jakarta were identified. They include:
 - a. Temporary water barriers,
 - b. Waterproofing,
 - c. Velocity and
 - d. Length of stay.
- e. Similarly for content loss the identified significant factors were
 - a. Flood adapted usage of buildings such as raised floors, low value and easily movable objects,
 - b. Household size (adults),
 - c. Moving objects to higher level,
 - d. Presence of a higher storey,
 - e. Past flood experience and
 - f. Flood warning time.

In general this research presents an improvement to the existing methodologies used for flood loss and damage estimation. The outcome could guide both the people and the government in evaluating their flood protection measures and policies which is indispensable due to climate change, urbanization, population explosion etc.

7.2 Limitations and Recommendations for Future Works

Based on the outcome of this research, the following limitations and recommendations are identified and proposed for future study:

- (1) The major limitation for this study is the data itself. Due to the limitation in time, this study confined only on the surveyed data obtained from Jakarta to test the validity of the framework and methodology proposed in this research. The proposed approach can be implemented to other geographically, climatically and economically similar regions in identifying the significant flood loss and damage influencing factors. Further analysis could be conducted in investigating the relationship between the factors and flood loss from these regions.
- (2) The loss data used in this study is the replacement cost data (current market prices). However, the losses to properties must be based on their pre-flood values so as to represent national economic losses. Thus the depreciated cost must be used instead of replacement cost of the items. Due to the difficulty in obtaining the age of the contents and property components, depreciated cost data was not factored in the study.
- (3) Nature and type of damage from flooding also depends on the quality of the affected buildings and its contents. Such data is only available for local building and construction authorities. In addition, due to difficulty in obtaining reliable data for validating the quality of the elements at risk especially from questionnaire surveys and interview, this variable was not considered while developing the loss estimation model.
- (4) The framework proposed in this dissertation could be expanded so as to include expert opinion while creating Bayesian Networks for synthetic data generation. This addition may add in the relationships between the variables that could be missed while learning the network from data. However due to the lack of availability of empirical studies describing the damage behaviour for the city of Jakarta, the proposed hypothesis has not been tested.

- (5) A further intensive flood risk assessment for Jakarta may be conducted by estimating the Expected Annual Loss (EAL) and Probable Maximum Flood (PML) index. EAL is computed by calculating the expected flood loss resulting from different flood return periods (10, 25, 50, 100, 500 year floods etc.) obtained through detailed hydraulic analysis. PML of a building is the value of loss experienced from the largest possible event. Both these indices can be used by government and insurance agencies in framing their flood management policies
- (6) The framework can be extended to include other hazards such as tsunami, landslide, typhoon etc. This could be achieved by replacing the flood loss database with the loss database of the corresponding hazard. Finally a software tool with a graphical user interface could be developed for multi hazard loss estimation based on the proposed framework developed in this thesis.

REFERENCES

- Ashley, R. M., D. J. Balmforth, A. J. Saul and J. D. Blanskby (2005). "Flooding in the future : predicting climate change, risks and responses in urban areas." Water Science and Technology **52**(5): 265-273.
- BAPPENAS (Badan Perencanaan dan Pembangunan Nasional),(2007). Laporan Perkiraan Kerusakan dan Kerugian Pasca Bencana Banjir Awal Februari 2007 di Wilayah Jabodetabek (Jakarta, Bogor, Depok,Tangerang, dan Bekasi). Jakarta, Badan Perencanaan Pembangunan Nasional.
- Barandela, R., J. S. Sánchez, V. Garcia and E. Rangel (2003). "Strategies for learning in class imbalance problems." Pattern Recognition **36**(3): 849-851.
- Basak, D., S. Pal and D. C. Patranabis (2007). "Support vector regression." Neural Information Processing-Letters and Reviews **11**(10): 203-224.
- Black, R. D.,(1975). Flood Proofing Rural Residences. A Project Agnes Report. Pennsylvania. New York Cornell University.
- BPBD (Badan Penanggulangan Bencana Daerah),(2013). Kajian Kerusakan, Kerugian Dan Kebutuhan Pemulihan Dampak Bencana Banjir Jakarta Januari 2013. Jakarta, BPBD DKI Jakarta.
- BPS,(2011). Jakarta Dalam Angka 2011. Jakarta, Indonesia.
- Brown, J. D., T. Spencer and I. Moeller (2007). "Modeling storm surge flooding of an urban area with particular reference to modeling uncertainties: A case study of Canvey Island, United Kingdom." Water Resources Research **43**(6).
- Budiyono, Y., J. Aerts, J. J. Brinkman, M. A. Marfai and P. Ward (2014). "Flood risk assessment for delta mega-cities: a case study of Jakarta." Natural Hazards **75**(1): 389-413.
- Bureau of Reclamation (1998). Downstream Hazard Classification Guidelines. Acer Technical Memorandum No. 11. U.S. Department of the Interior. Denver, Colo.

- Capobianco, M., H. J. DeVriend, R. J. Nicholls and M. J. F. Stive (1999). "Coastal area impact and vulnerability assessment: The point of view of a morphodynamic modeler." Journal of Coastal Research **15**(3): 701-716.
- Chawla, N. V., K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer (2002). "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research: 321-357.
- Chen, Q., J. Zhang and L. Zhang (2015). "Risk assessment, partition and economic loss estimation of rice production in China." Sustainability (Switzerland) **7**(1): 563-583.
- Chinh, D. T., Gain, A. K., Dung, N. V., Haase, D., & Kreibich, H. (2015). Multi-Variate Analyses of Flood Loss in Can Tho City, Mekong Delta. Water, **8**(1): 6.
- Chow, C. and C. Liu (1968). "Approximating discrete probability distributions with dependence trees." Information Theory, IEEE Transactions on **14**(3): 462-467.
- Clausen, L. K. (1989). Potential Dam Failure: Estimation of Consequences, and Implications for Planning, Unpublished M.Phil. thesis at the School of Geography and Planning at Middlesex Polytechnic collaborating with Binnie and Partners, Redhill.
- Cobb, B. R. and P. P. Shenoy (2006). "Inference in hybrid Bayesian networks with mixtures of truncated exponentials." International Journal of Approximate Reasoning **41**(3): 257-286.
- Cobb, B. R., P. P. Shenoy and R. Rumi (2004). Approximating probability density functions with mixtures of truncated exponentials. Proceedings of the Tenth Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU-04).
- Daemen, A. and B. De Moor (2009). Development of a kernel function for clinical data. Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE, IEEE.

- Dale K., Edwards M., Middelmann M and Zoppou C (2004). Structural flood vulnerability and the Australianisation of Black's curves. Risk 2004. Risk Engineering Society, Melbourne, 2004.
- Das, B., N. C. Krishnan and D. J. Cook (2015). "RACOG and wRACOG: Two Probabilistic Oversampling Techniques." Knowledge and Data Engineering, IEEE Transactions on **27**(1): 222-234.
- Davis, S. A. (1985). Business depth-damage analysis procedures. Ft. Belvoir, Va, U.S Army Corps of Engineers, Engineer Institute for Water Resources.
- Davis, S. A. and L. L. Skaggs (1992). Catalog of residential depth-damage functions used by the Army Corps of Engineers in flood damage estimations U.S. Army Corps of Engineers, Water Resources Support Center, Institute for Water Resources, Available from National Technical Information Service, Ft. Belvoir, Va. : Springfield, VA :.
- Demšar, U., P. Harris, C. Brunsdon, A. S. Fotheringham and S. McLoone (2012). "Principal Component Analysis on Spatial Data: An Overview." Annals of the Association of American Geographers **103**(1): 106-128.
- Doquire, G. and M. Verleysen (2011). An Hybrid Approach to Feature Selection for Mixed Categorical and Continuous Data. KDIR.
- Dutta, D., S. Herath and K. Musiake (2003). "A mathematical model for flood loss estimation." Journal of hydrology **277**(1–2): 24-49.
- Elvira Consortium (2002). Elvira: An environment for creating and using probabilistic graphical models. Proceedings of the first European workshop on probabilistic graphical models.
- Ellisa, E. (2014). The entrepreneurial city of Kelapa Gading, Jakarta. Journal of Urbanism: International Research on Placemaking and Urban Sustainability, 7(2), 130-151.

- EM-DAT (2014). The OFDA/CRED (Office of Foreign Disaster Assistance / Centre for Research on the Epidemiology of Disasters) International Disaster Database Université Catholique de Louvain, Brussels (Belgium).
- Experian Limited (2000). Great Britain MOSAIC Descriptions along with separate data tables for flood damage, Experian, Nottingham, U.K.
- FEMA (2003b). HAZUS-MH, Flood Model: Technical Manual, Federal Emergency Management Agency.
- Fernández, A. and A. Salmerón (2008). Extension of Bayesian network classifiers to regression problems. Advances in Artificial Intelligence–IBERAMIA 2008, Springer: 83-92.
- Fröhlich, H. and A. Zell (2005). Efficient parameter selection for support vector machines in classification and regression via model-based global optimization. Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on, IEEE.
- Geyer, C. (2011). "Introduction to Markov Chain Monte Carlo." Handbook of Markov Chain Monte Carlo: 3-48.
- Gissing, A. and R. Blong (2004). "Accounting for variability in commercial flood damage estimation." Australian geographer **35**(2): 209-222.
- Gómez-Verdejo, V., M. Verleysen and J. Fleury (2009). "Information-theoretic feature selection for functional data classification." Neurocomputing **72**(16–18): 3580-3589.
- Gönen, M. and E. Alpaydın (2011). "Multiple kernel learning algorithms." The Journal of Machine Learning Research **12**: 2211-2268.
- Green, C. H., D. J. Parker, P. M. Thompson and E. C. Penning-Roswell (1983). Indirect Losses from Urban Flooding: An Analytical Framework, Flood Hazard Research Centre, Middlesex Polytechnic, UK.

- Grigg, N. S. and O. J. Helweg (1975). "STATE-OF-THE-ART OF ESTIMATING FLOOD DAMAGE IN URBAN AREAS1." JAWRA Journal of the American Water Resources Association **11**(2): 379-390.
- Grothmann, T. and F. Reusswig (2006). "People at Risk of Flooding: Why Some Residents Take Precautionary Action While Others Do Not." Natural Hazards **38**(1-2): 101-120.
- Guyon, I., J. Weston, S. Barnhill and V. Vapnik (2002). "Gene selection for cancer classification using support vector machines." Machine learning **46**(1-3): 389-422.
- Han, Y., S. Wang, Y. Zhou, F. Wang, W. Liu and Y. Hou (2012). "Temporal and spatial variability of flood/drought in china during the last 530 years." Disaster Advances **5**(4): 1760-1765.
- Hastie, T., R. Tibshirani and J. Friedman (2009). Unsupervised learning, Springer.
- Herrera, M., A. Guglielmetti, M. Xiao and R. F. Coelho (2014). "Metamodel-assisted optimization based on multiple kernel regression for mixed variables." Structural and multidisciplinary optimization **49**(6): 979-991.
- Hsu, M. H., S. H. Chen and T. J. Chang (2000). "Inundation simulation for urban drainage basin with storm sewer system." Journal of hydrology **234**(1-2): 21-37.
- Ilorme, F. and V. W. Griffis (2013). "A novel procedure for delineation of hydrologically homogeneous regions and the classification of ungauged sites for design flood estimation." Journal of hydrology **492**: 151-162.
- Jensen, F. V., S. L. Lauritzen and K. G. Olesen (1990). "Bayesian updating in causal probabilistic networks by local computations." Computational statistics quarterly **4**: 269-282.
- Joint Cooperation Program (JCP),(2013). Jakarta Floods January 2013. Flood Early Warning System.
- Jolliffe, I. (2002). Principal component analysis, Wiley Online Library.

- Julian, M. M. and P. J. Ward (2014). "Assessment of the effects of climate and land cover changes on river discharge and sediment yield, and an adaptive spatial planning in the Jakarta region." Natural Hazards **73**(2): 507-530.
- Kang, J.-L., M.-D. su and L.-F. Chang (2005). "Loss functions and framework for regional flood damage estimation in residential area." Journal of Marine Science and Technology **13**(3): 193-199.
- Kates, R. W. (1965). Industrial flood losses: damage estimation in the Lehigh Valley, Dept. of Geography, University of Chicago.
- Kawahara, Y. and T. Uchida (2008). Integrated Modeling for Inundation Flows in Urban Areas, Honolulu, Hawaii, ASCE.
- Kelman, I. (2003) Physical flood vulnerability of residential properties in coastal, eastern England. PhD Thesis, University of Cambridge
- Kelman, I. and R. Spence (2004). "An overview of flood actions on buildings." Engineering geology **73**(3-4): 297.
- Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Ijcai.
- Koller, D. and N. Friedman (2009). Probabilistic graphical models: principles and techniques, MIT press.
- Kozlov, A. V. and D. Koller (1997). Nonuniform dynamic discretization in hybrid networks. Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc.
- Kraskov, A., H. Stögbauer and P. Grassberger (2004). "Estimating mutual information." Physical review E **69**(6): 066138.
- Kreibich, H., I. Seifert, B. Merz and A. H. Thieken (2010). "Development of FLEMOcs - a new model for the estimation of flood losses in the commercial sector." Développement de FLEMOcs - un nouveau modèle pour l'estimation des dommages dus aux inondations dans le secteur commercial **55**(8): 1302-1314.

- Kuncheva, L. I. (2007). A stability index for feature selection. Artificial intelligence and applications.
- Lauritzen, S. L. and D. J. Spiegelhalter (1988). "Local computations with probabilities on graphical structures and their application to expert systems." Journal of the Royal Statistical Society. Series B (Methodological): 157-224.
- Lerner, U., E. Segal and D. Koller (2001). Exact inference in networks with discrete children of continuous parents. Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc.
- Lodhi, H., C. Saunders, J. Shawe-Taylor, N. Cristianini and C. Watkins (2002). "Text classification using string kernels." J. Mach. Learn. Res. **2**: 419-444.
- Marfai, M. A., A. B. Sekaranom and P. Ward (2015). "Community responses and adaptation strategies toward flood hazard in Jakarta, Indonesia." Natural Hazards **75**(2): 1127-1144.
- Marschiavelli, M. I. C. (2008) Vulnerability assessment and coping mechanism related to floods in urban areas: a community-based case study in Kampung Melayu, Indonesia. Gadjah Mada University, Indonesia
- McBean, E. A., J. Gorrie, M. Fortin, J. Ding and R. Monlton (1988). "Adjustment Factors for Flood Damage Curves." Journal of Water Resources Planning and Management **114**(6): 635-646.
- McBean, E. A., J. Gorrie, M. Fortin, J. Ding and R. Moulton (1988). "Flood Depth--Damage Curves By Interview Survey." Journal of Water Resources Planning and Management **114**(6): 613-634.
- Merz, B., H. Kreibich and U. Lall (2013). "Multi-variate flood damage assessment: a tree-based data-mining approach." Natural Hazards and Earth System Science **13**(1): 53-64.
- Merz, B., H. Kreibich, R. Schwarze and A. Thieken (2010). "Review article "assessment of economic flood damage"." Natural Hazards and Earth System Science **10**(8): 1697-1724.
-

- Merz, B., H. Kreibich, A. Thielen and R. Schmidtke (2004). "Estimation uncertainty of direct monetary flood damage to buildings." Nat. Hazards Earth Syst. Sci **4**(1): 153-163.
- Messner, F. and V. Meyer (2005) Flood damage, vulnerability and risk perception - challenges for flood damage research. UFZ-Diskussionspapiere **No. 13/2005**,
- Metropolis, N. and S. Ulam (1949). "The monte carlo method." Journal of the American statistical association **44**(247): 335-341.
- Moral, S., R. Rumí and A. Salmerón (2001). Mixtures of truncated exponentials in hybrid Bayesian networks. Symbolic and Quantitative Approaches to Reasoning with Uncertainty, Springer: 156-167.
- Munich Re,(2015). Geo Natural Catastrophes 2014 Geo Risks Research, NatCatSERVICE. Munich, Germany, Munich Re.
- N'Jai, A., S. M. Tapsell, D. Taylor, P. M. Thompson and R. C. Witts (1990). FLAIR 1990 (Flood Loss Assessment Information Report), Middlesex Polytechnic Flood Hazard Research Centre, London, December 1990.
- Nadal, N. C., R. E. Zapata, I. Pagan, R. Lopez and J. Agudelo (2010). "Building Damage due to Riverine and Coastal Floods." Journal of Water Resources Planning and Management **136**(3): 327-336.
- Nascimento, N., M. L. Machado, M. Baptista and A. De Paula E Silva (2007). "The assessment of damage caused by floods in the Brazilian context." Urban Water Journal **4**(3): 195-210.
- Nasir, H.,(2008). Flood since Batavia until Jakarta (1614-2008) problems and solutions.
- NRC (2000). (National Research Council):Risk Analysis and Uncertainty in Flood Damage Reduction Studies. Washington DC, The National Academies Press.
- Peng, H., F. Long and C. Ding (2005). "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy."

Pattern Analysis and Machine Intelligence, IEEE Transactions on **27**(8): 1226-1238.

Penning-Rowsell, E. (1992). The Economics of coastal management: a manual of benefit assessment techniques, Belhaven Press.

Penning-Rowsell, E. and J. B. Chatterton (1977). The benefits of flood alleviation: a manual of assessment techniques, Saxon House.

Penning-Rowsell, E. and F. H. R. C. S. Middlesex University (2008). The Benefits of Flood and Coastal Risk Management: A Manual of Assessment Techniques, Middlesex University Press.

Penning-Rowsell, E., C. Viavattene, J. Pardoe, J. B. Chatterton, D. J. Parker and J. Morris (2010). The Benefits of Flood and Coastal Risk Management: A Handbook of Assessment Techniques 2010, Middlesex University, Centre for Higher Education Research.

Penning-Rowsell, E. C. and C. Green (2000). "New insights into the appraisal of flood-alleviation benefits: (1) flood damage and flood loss information." Journal of the Chartered Institution of Water and Environmental Management **14**(5): 347-353.

Penning-Rowsell, E. C., C. H. Green, P. M. Thompson, A. M. Coker, S. M. Tunstall, C. Richards and D. J. Parker (1992). The Economics of Coastal Management: A Manual of Benefit Assessment Techniques, Belhaven Press, London, U.K. and Florida, U.S.A.

Pistrika, A. K. and S. N. Jonkman (2009). "Damage to residential buildings due to flooding of New Orleans after hurricane Katrina." Natural Hazards and Earth System Science **2**(54): 413-434.

Pistrika, A. K. and S. N. Jonkman (2010). "Damage to residential buildings due to flooding of New Orleans after hurricane Katrina." Natural Hazards **54**(2): 413-434.

Prettenthaler, F., Amrusch, P., & Habsburg-Lothringen, C. (2010). "Estimation of an absolute flood damage curve based on an Austrian case study under a dam breach scenario". Natural Hazards and Earth System Science, **10**(4): 881-894.

- Qiu, S. and T. Lane (2005). "Multiple kernel learning for support vector regression." Computer Science Department, The University of New Mexico, Albuquerque, NM, USA, Tech. Rep: 1.
- Qiu, S. and T. Lane (2009). "A framework for multiple kernel support vector regression and its applications to siRNA efficacy prediction." Computational Biology and Bioinformatics, IEEE/ACM Transactions on **6**(2): 190-199.
- Reese, S. and D. Ramsay,(2010). RiskScape: Flood Fragility Methodology. Wellington, New Zealand, New Zealand Climate Change Research Institute, Victoria University of Wellington.
- Reynaud, A., C. Aubert and M.-H. Nguyen (2013). "Living with floods: protective behaviours and risk perception of Vietnamese households." The Geneva Papers on Risk and Insurance-Issues and Practice **38**(3): 547-579.
- Romero, V., R. Rumí and A. Salmerón (2006). "Learning hybrid Bayesian networks using mixtures of truncated exponentials." International Journal of Approximate Reasoning **42**(1): 54-68.
- Roos, W., P. Waarts and A. Vrouwenvelder,(2003). Damage to buildings. Delft Cluster.
- Ropero, R., P. Aguilera, A. Fernández and R. Rumí (2014). "Regression using hybrid Bayesian networks: Modelling landscape–socioeconomy relationships." Environmental Modelling & Software **57**: 127-137.
- Rumí, R. and A. Salmerón (2007). "Approximate probability propagation with mixtures of truncated exponentials." International Journal of Approximate Reasoning **45**(2): 191-210.
- Rumí, R., A. Salmerón and S. Moral (2006). "Estimating mixtures of truncated exponentials in hybrid Bayesian networks." Test **15**(2): 397-421.
- Sangrey, D. A., P. J. Murphy and J. L. Nieber,(1975). Evaluating the Impact of Structurally Interrupted Flood Plain Flows. Technical Report No. 98, Cornell University Water Resources and Marine Sciences Center, Ithaca, New York, USA.
-

- Schölkopf, B., K. Tsuda and J.-P. Vert (2004). Kernel methods in computational biology, MIT press.
- Schwarz, J. and H. Maiwald (2008). Damage and loss prediction model based on the vulnerability of building types. 4th International Symposium on Flood Defence: Managing Flood Risk, reliability and Vulnerability. Toronto, Ontario, Canada.
- Scott, D. W. (2012). Multivariate density estimation and visualization. Handbook of Computational Statistics, Springer: 549-569.
- Smith, D. I. (1994). "Flood Damage Estimation - A review of urban stage-damage curves and loss functions." Water SA(20): 231 - 238.
- STRUPCZEWSKI, G. (2015). "Cluster analysis of natural disaster losses in Polish agriculture." Scientific Papers Series-Management, Economic Engineering in Agriculture and Rural Development **15**(1): 513-526.
- Suleman, M. S., A. N'Jai, C. H. Green, E. C. Penning-Rowsell and (1988). Potential Flood Damage Data: A Major Update., Middlesex Polytechnic Flood Hazard Research Centre, London, U.K.
- Swiss Re,(2012). Flood – an underestimated risk : Inspect, inform, insure, Swiss Reinsurance Company, Zurich.
- Swiss Re,(2015). Natural Catastrophe and Man-made Disasters in 2014:Convective and winter storms generate most losses. Sigma. Zurich.
- Thieken, A. H., V. Ackermann, F. Elmer, H. Kreibich, B. Kuhlmann and U. Kunert (2009). Methods for the evaluation of direct and indirect flood losses. In RIMAX Contributions at the 4th International Symposium on Flood Defence (ISFD4), Potsdam: Deutsches GeoForschungsZentrum GFZ.
- Thieken, A. H., M. Müller, H. Kreibich and B. Merz (2005). "Flood damage and influencing factors: New insights from the August 2002 flood in Germany." Water Resour. Res. **41**(12): W12430.

- Thieken, A. H., A. Olschewski, H. Kreibich, S. Kobsch and B. Merz (2008). Development and evaluation of FLEMOps - A new Flood Loss Estimation MOdel for the private sector.
- UNDRO (United Nations Disaster Relief Organization),(1982). Natural Disasters and Vulnerability Analysis. Geneva, Office of the United Nations Disaster Relief Coordinator.
- USACE (1988a). National Economic Development Procedures Manual - Urban Flood Damage, U.S. Army Corps of Engineers,nstitute for Water Resources, IWR Report 88-R-2,Ft. Belvoir, VA.
- USACE (1996). Risk-based analysis for flood damage reduction studies. Washington, D.C, U.S. Army Corps of Engineers, CECW-EH-Y.
- USACE (2000). Economic Guidance Memorandum (EGM) 01-03, Generic Depth-Damage Relationships. Washington DC, U.S. Army Corps of Engineers Memorandum, CECW - PG 4.
- USACE (2003). Economic Guidance Memorandum (EGM) 04-01, Generic Depth-Damage Relationships for Residential Structures with Basements. Washington DC, U.S. Army Corps of Engineers Memorandum,CECW-PG 10.
- Vapnik, V. N. (1998). Statistical Learning Theory, Wiley, New York.
- Vapnik, V. N., S. E. Golowich and A. J. Smola (1997). Support vector method for function approximation,regression estimation, and signal processing. Advances in Neural Information Processing Systems9 --- Proceedings of the 1996 Neural Information Processing Systems Conference (NIPS 1996), Dever, CO, USA, MIT Press, Cambridge, MA, USA.
- Varma, M. and B. R. Babu (2009). More generality in efficient multiple kernel learning. Proceedings of the 26th International Conference On Machine Learning, ICML 2009.

- Varma, M. and B. R. Babu (2009). More generality in efficient multiple kernel learning. Proceedings of the 26th Annual International Conference on Machine Learning, ACM.
- Varma, M. and D. Ray (2007). Learning the discriminative power-invariance trade-off. Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, IEEE.
- Verburg, P. H., T. Veldkamp and J. Bouma (1999). "Land use change under conditions of high population pressure: the case of Java." Global Environmental Change **9**(4): 303-312.
- Verleysen, M. (2003). "Learning high-dimensional data." Nato Science Series Sub Series III Computer And Systems Sciences **186**: 141-162.
- Vogel, K., C. Riggelsen, B. Merz, H. Kreibich and F. Scherbaum (2012). Flood damage and influencing factors: a Bayesian network perspective. 6th European Workshop on Probabilistic Graphical Models (PGM 2012), University of Granada, Granada, Spain.
- Ward, P., H. d. Moel and J. Aerts (2011a). "How are flood risk estimates affected by the choice of return-periods?" Natural Hazards and Earth System Science **11**(12): 3181-3195.
- Ward, P. J., M. A. Marfai, F. Yulianto, D. R. Hizbaron and J. C. J. H. Aerts (2011b). "Coastal inundation and damage exposure estimation: a case study for Jakarta." Natural Hazards **56**(3): 899-916.
- White, G. F. (1945) Human adjustment to floods : a geographical approach to the flood problem in the United States. University of Chicago
- White, G. F. (1964). "Choice of adjustment to floods: Gilbert F. White: University of Chicago, Chicago Ill., 1964, 150 p. (Research paper no. 93)." Journal of hydrology **3**(3-4): 344.
- WHO,(2007). Floods in Jakarta, Banten, and West Java Province, Republic of Indonesia. Emergency Situation Report # 6.
-

Wijayanti, P., Tono, Hastuti and D. Pramudita,(2015). Estimation of River Flood Damage in Jakarta: The Case of Pessanggrahan River. EEPSEA Research Reports Series No. 2015-RR3. Philippines, IDRC.

Wilby, R. L. (2007). "A Review of Climate Change Impacts on the Built Environment." Built Environment **33**(1): 31-45.

Wind, H. G., T. M. Nierop, C. J. de Blois and J. L. de Kok (1999). "Analysis of flood damages from the 1993 and 1995 Meuse Floods." Water Resour. Res. **35**(11): 3459-3465.

Zhai, Guofang, Teruki Fukuzono, and Saburo Ikeda. "Modeling flood damage: case of tokai flood 20001." (2005): 77-92.

APPENDIX

Appendix A

Assessment of flood damage and losses to Residential Properties

Part I - Background Characteristics

Date: / /

1. Location : X and Y Coordinates* (Lat,Long)

Name and Address*:

Postal Code*:

Age:

2. Employment Status of the head of the family*

A. Full Time Employee	B. Part Time Employed
C. Self Employed	D. Unemployed
E. Retired	F. Others

3. How many persons live in there in total*?

Children (<18 Years):

Adults:

Elderly (>60Years):

4. Educational Experience: (Highest degree earned by any individual in house)*

A. No Education	B. Secondary School or below
C. High School	D. Diploma
E. Graduate (Bachelors)	F. Post Graduate and above

5. Average Household monthly income*:

Rupiah

6. Are you aware that you are living in a flood risk area*? Yes / No / Don't Know

Part 2 –Building Information

1. Which of the following options applies to your current living**

A. One Family House	B. Semi- Detached House
C. Apartments	

2. Select the number of story of this home**

A. Single Storey	B. Double Storey
C. 3 or more floors	

3. How long have you been staying in this house*? _____ Years

4. House Ownership*:

1. Own	2. Rent	3. Other
--------	---------	----------

5. Does this home contain basement*? Yes \No

6. First Floor area*: _____ square meters

Total area of the house*: _____ Square meters

7. The type of exterior wall material that best matches most of the exterior of this home**.

A. Mud	B. Brick Masonry
C. Stone Masonry	D. Concrete
E. Woods	F. Hollow Blocks
G. Precast	H. Others

8. Type of exterior wall covering that best matches most of the exterior of this home**.

A. Plaster/Stucco with paint	B. Concrete/Brick work
C. Tiles	D. Wood work
E. Other	

9. What is the brand of paint used*? _____

10. The type of floor best matches most of the floors of this home**.

A. Cement	B. Wood
C. Tile/Ceramics	D. Marble/Granite
E. Others	

11. Roof Covering*.

A. Concrete	B. Concrete tile
C. Ceramic Tiles	D. Metal Tiles
E. Asbestos tiles	

12. Roof Truss material *

F. Wood	G. Light weight steel
H. Steel	I. Others

13. Ceiling Cover*

J. Gypsum	K. Plywood
L. Asbestos	M. Others

14. Ceiling frame material *

N. Wood	O. Light weight steel
P. Iron	Q. Others

Part 3 – Flood Event Information

1. Please indicate the flood depth inside house and duration of flooding.*

Year	Flood Depth (m)	Duration (Hrs)
2013		
2007		

2. Have you had any previous flood experience*? Yes \ No.

If Yes how many times*?

3. When was the last time your house was flooded*?_____

4. How big were the losses in the previous event?_____ Rp

5. What do you think about the flood*? (Multiple answers allowed- Select all Applicable)

Flood Nature	Select
Back water from sewer which comes through toilets	
Rising ground water	
Slow rise flood from nearby river or stream (Low flow rates) from rain	
Dynamic flood from nearby river or stream (High flow rates) from rain	
Flooding due to heavy rain	
Coastal Flooding	
Failure of flood defence system (Dam/Dyke Breach)	
Don't Know	
Others (Specify)	

6. What do you think about the velocity of flood water*

Velocity	Select
0. Standing Water	
1. Low - Adult person can easily stand	
2. Medium - Adult person can stand with difficulty	
3. High - Adult person can get carried away	

7. What are the materials transported and deposited by the flood*?(Select all applicable)

A. Mud	B. Sand
C. Stones	D. Boulders
E. Tree Branches	Others please specify

8. Has the water been polluted by other materials*? (Select all applicable)

Wastewater or sewerage	
Paint or varnish	
Oils or fuels	
Chemical	
No Contamination	

Part 4– Cost and Damages

1. Height of first floor from the ground*? (Give your best estimation) _____meter
2. Did you have flood insurance?* Yes/ No

If Yes, how much was the insurance Payout?

Total Payout: _____Rupiah

Structure Damage payout: _____ Rupiah

Content Damage Payout: _____Rupiah

- List their names

Items Damaged	Brand & Size	Nos: Damaged
TV		
Plasma TV		
Komputer		
DVD player		
VCD Player		
Amplifier		
Mobile phone		
Tape Recorder		
AC		
Printer		
Fan		
Speaker		
Piano/Organ		
Mesin Cuci		
Kulkas		

- No: of days:

5. Please list out the nature and type of damages in the following categories*

Section	Rooms & Extent of damage Damage Types	Living Room				Dining Room				Master Bed Room				Childrens Room				Kitchen			
		1/4	1/2	3/4	1	1/4	1/2	3/4	1	1/4	1/2	3/4	1	1/4	1/2	3/4	1	1/4	1/2	3/4	1
Floors	Debonding of floor tiles																				
	Floor Tiles damaged/Broken																				
	Granite/marble lost it's shine																				
Walls	Peeling of paints from wall																				
	Peeling of plaster from wall																				
	Wall paper damaged																				
	Skirting damaged																				
Ceiling	Plywood Ceiling covering damaged																				
	Gypsum Ceiling covering damaged																				
	Walls Stained from floods	Yes/No				Yes/No				Yes/No				Yes/No				Yes/No			
Walls	Cracks on wall	Yes/No				Yes/No				Yes/No				Yes/No				Yes/No			
	Length of Crack																				

Collapse of Individual Components	Yes / No
If Yes	Nos: Damaged
Wall	
Column	
Beam	
Roof	

Foundation	Cracks in foundation	Y/N
	Settlement of foundation	Y/N

Item	Damage Type	No:
Doors & Windows	Swelling of doors	
	Warping and distortion of doors	
	Wooden doors cracked	
	Wooden doors completely damaged	
	Metal doors distorted	
	Paint gone from metal doors	
	Metal doors completely damaged	
	Door frame separated from wall	
	Window frame damaged	
	Window glasses broken	
Electrical	Switch boards, electrical panels inundated	
	Lighting Damaged	
Roof	Asbestos Roof Damaged	

Toilet	Clogging of toilet and sewers	Y/N
--------	-------------------------------	-----

Gates corroded / Paints peeled	Y/N
--------------------------------	-----

Trustworthiness Scale	
Very Low	1
Low	2
Moderate	3
High	4
Very High	5

Part 5 – Warning and Precautionary Measures

1. Did you receive an emergency warning before flood*?

A. Official Warning	B. Other Warning	C. No warning
---------------------	------------------	---------------

2. If yes, how many hours before the onset of flood has the warning been received*? _____
Hours

3. If yes, What are the contents of the warning*? *(Select all applicable)*

Select Contents of flood Warning

	Information about time to peak water level
	Information about maximum inundation level
	Evacuation Information
	Emergency measures for damage reduction measures
	Information about inundated areas

4. What are the sources of Flood Warning and please rate it's trustworthiness*? *(Select all applicable)*

Select Sources

	Warning by official sources (Govt. Agencies)	
	Warning by media (TV, Radio, Internet, mobile phones)	
	Warning by NGO's and mosques	
	Warning by friends, relatives, neighbours etc.	

5. Please rate the reliability of the flood warning source?

Sources

Warning by official sources (Govt. Agencies)	
Warning by media (TV, Radio, Internet, mobile phones)	
Warning by NGO's and mosques	
Warning by friends, relatives, neighbours etc.	

6. What emergency measures were taken to mitigate the damage due to flooding*?
(Select which ever measures you have undertaken)

Select Actions Actions which you may take in future

	No emergency measures were taken because of no time	
	No emergency measures were taken because I don't know what to do	
	Documents and valuables stored more securely	
	Furniture and movable objects brought to safety (e.g. 2 nd floor, top of table)	
	Put water barriers (sand bags, flood boards etc) to divert water	
	Electricity and gas turned off	
	Openings in buildings (e.g. doors, windows etc.) were sealed against the ingress of water	
	Electrical appliances unplugged	
	Vehicles driven to safety	
	Block Toilets	
	Use of pumps to pump the water out	

7. What information you would like to have in flood warning *? (Select all applicable)

Select Contents of flood Warning

	Information about time to peak water level
	Information about maximum inundation level
	Evacuation Information
	Emergency measures for damage reduction measures
	Information about inundated areas

8. Have you taken self protection measures before the flood*? Yes / No

9. If 'No', why haven't you?

A. No necessity	B. No means available	C. Not my responsibility	D. Not possible	E. Others
-----------------	-----------------------	--------------------------	-----------------	-----------

10. If Yes, What flood precautionary measures had been implemented prior to the flood event, it's effectiveness *?(Select which ever measures you have undertaken) Please

Select	Actions	Future Steps
	I have taken flood insurance	
	Adapting elevated configuration.	
	Waterproof the exterior walls of the building	
	Installation of watertight barriers/shields over doors, windows and other openings	
	Install backflow valves in sanitary and storm sewer lines	
	Permanent or mobile water barriers were placed to divert floods e.g. sand bags or local small walls	
	Flood prone floors were used with low value goods and easily movable furniture	
	Flood prone floors constructed with waterproof and flood resistant materials.	
	Moving to flood protected area	
	Find out more information about flood hazard of my house	
	Attending Community based capacity development programs	
Others (Please Specify Below)		

11. During the time of flood, how do you rate your level of private precautionary preparedness*?

A. Not at all prepared	B. Not very prepared	C. Don't Know	D. Slightly Prepared	E. Very well prepared
------------------------	----------------------	---------------	----------------------	-----------------------

12. Do you believe that private flood precautionary measures taken by you are effective in reducing the damages from flooding*? Yes/No

13. Are you worried about flooding in this locality*?

A. Not worried at all	B. Slightly worried	C. Moderately worried	D. Greatly worried	E. Extremely Worried
-----------------------	---------------------	-----------------------	--------------------	----------------------

Appendix B

Algorithm for constructing TAN-MTE network

1. Provide the variables $X_1 \dots \dots X_n$ and *Strcuture and Content loss* as Input
2. Using the variables $X_1 \dots \dots X_n$ as nodes construct the graph C
3. Compute the mutual information between the variables and the structure and content loss
4. Using the conditional mutual information between the variables, label each link (X_1, X_2)

$$I(X_1, X_2 | Loss) = \iiint f(x_1, x_2, loss) \log \frac{f(x_1, x_2 | loss)}{f(x_1 | loss) f(x_2 | loss)} dx_1 dx_2 dloss$$

5. From the graph C, obtain the maximum spanning tree T and direct the links in T
6. Construct a new Graph G by inserting the links from loss to variable $X_{i=1:n}$ to the tree T
7. For each variable $X_{i=1:n}$, estimate the conditional MTE density and estimate only the MTE density for the variable loss
8. The resulting network with the Graph G and the above computed densities constitutes the TAN.