
Predicting Seismic Damage and Loss for Residential Buildings using Data Science

Samuel Roeslin

A thesis submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy in Civil Engineering,
The University of Auckland, 2021.

"All models are wrong but some are useful" (Box, 1979)

Abstract

This thesis presents the application of data science techniques, especially machine learning, for the development of seismic damage and loss prediction models for residential buildings. Current post-earthquake building damage evaluation forms are developed for a particular country in mind. The lack of consistency hinders the comparison of building damage between different regions. A new paper form has been developed to address the need for a global universal methodology for post-earthquake building damage assessment. The form was successfully trialled in the street 'La Morena' in Mexico City following the 2017 Puebla earthquake.

Aside from developing a framework for better input data for performance based earthquake engineering, this project also extended current techniques to derive insights from post-earthquake observations. Machine learning (ML) was applied to seismic damage data of residential buildings in Mexico City following the 2017 Puebla earthquake and in Christchurch following the 2010-2011 Canterbury earthquake sequence (CES). The experience showcased that it is readily possible to develop empirical data only driven models that can successfully identify key damage drivers and hidden underlying correlations without prior engineering knowledge. With adequate maintenance, such models have the potential to be rapidly and easily updated to allow improved damage and loss prediction accuracy and greater ability for models to be generalised.

For ML models developed for the key events of the CES, the model trained using data from the 22 February 2011 event generalised the best for loss prediction. This is thought to be because of the large number of instances available for this event and the relatively limited class imbalance between the categories of the target attribute. For the CES, ML highlighted the importance of peak ground acceleration (PGA), building age,

building size, liquefaction occurrence, and soil conditions as main factors which affected the losses in residential buildings in Christchurch. ML also highlighted the influence of liquefaction on the buildings losses related to the 22 February 2011 event.

Further to the ML model development, the application of post-hoc methodologies was shown to be an effective way to derive insights for ML algorithms that are not intrinsically interpretable. Overall, these provide a basis for the development of 'greybox' ML models.

Acknowledgements

I would like to express my deepest gratitude to the Earthquake Commission (EQC), especially Geoffrey Spurr. I extend my recognition to RiskScape and Ryan Paulik. This research project would not have been possible without their technical support.

Countless people supported me along my PhD journey, within the good times as well as the challenging one. I would like to acknowledge some people in particular.

I would like to express my deepest gratitude to the Earthquake Commission (EQC), especially Geoffrey Spurr. I extend my recognition to RiskScape and Ryan Paulik. This research project would not have been possible without their technical support.

Countless people supported me along my PhD journey, during the good times as well as the challenging ones. I would like to acknowledge some people in particular.

First and foremost, I would like to express my deepest gratitude to my supervisor Dr Quincy Ma. Your guidance, constructive feedback and thoughtful insights facilitated my transition from a wide-eyed graduate student to an independent researcher. Thank you for your patience, endless encouragement, and for granting me the freedom to be a curious researcher and devote time to explore new areas.

I would like to express special thanks to Dr Joerg Wicker for generously providing me with technical advice related to the computer science aspects of my research. Your advice and insights gave me the confidence to explore, learn, and publish in the computer science domain. Your trust and assistance will always be remembered.

I am indebted to Associate Professor Liam Wotherspoon. Your support and help made my journey a lot easier. Your invaluable assistance enabled me to meet professionals that contributed to my research. I extend my gratitude to Dr Sjoerd Van Ballegooy who made time available to provide me with insightful insights.

I am sincerely grateful to my advisor Professor Ken Elwood. Your support before the start of my PhD and assistance during my provisional year were key to the success of my PhD journey. Your expert knowledge, enthusiasm, and passion combined with sagacity, wisdom, and humility made me self-reflect on what it is to be an outstanding researcher. I was honored to be at your side during the reconnaissance mission following the 2017 Puebla earthquake. Your advice and insightful comments all participated in my growth as a young researcher.

I would like to thank Professors Rajesh Dhakal, Hugon Juarez-Garcia, and Alonso Gomez-Bernal for allowing me to collaborate with you and your students working on damage assessment in Mexico. I also would like to recognise the invaluable technical assistance of Professor Amador Terán Gilmore.

I would like to pay my special regards to Pavan Chigullapally, Sunil Nataraj, and Eyitayo Ademola Opabola. Besides being great colleagues, you have been exceptional friends and mentors. Your selfless help, technical, and mental support surely made my research seem a lot easier. I will always appreciate the numerous discussion around PhD and life as young researchers. I hope that we will always be available for each other and be willing to discuss science.

I am thankful to Amelia Lin. It was a pleasure to be a student rep with you. In addition, your help and advice enabled me to overcome and move forward again whenever my research hit GIS-related hurdles. I would like to acknowledge the help of Diego Ivan Hernandez Hernandez. Your expertise and insights in Mexico's seismicity were of invaluable help. Moreover the numerous events and activities you invited me to, kept me sane during the past four years. I am thankful for the support and discussions with Matt Cutfield. Your selfless assistance at the beginning of my PhD enabled me to get a smoother start in my research journey. I wish to extend my thanks to Kai Marder, Tongyue Zhang, Frank Bueker, Mehdi Sarrafzadeh, Rijalul Fikri, Harish Shivaramu, Rahul Kadam, Haozhi Tan, Ronald Gultom, Mehrdad Bisadi.

Special thanks to the Professors and friends from the UoA Machine Learning Group. The weekly meetings were a great opportunity to be exposed to current research in the field of machine learning. Additionally, your help and insights contributed to the successful application of machine learning for my research project. Thank you to

Professor Amador Terán-Gilmore and Professor Andreas Maurial. Your help extended beyond my master studies. I am grateful for your unwavering support.

Thank you to my flatmates Long Qi Yu, James Opie, and Timothy Christopher. You have been present for me, extremely understanding and supportive within the good times as well as the most challenging one. Thank you to my friend in New Zealand Arisarawan Tanasinsiri and Ondřej Balkánský as well as my friends in Europe: Maike Reuther, Maud Streissel, Oliver Scholz, Thomas Hardieck, Eric Hauss, Francois Bauer. You have always made yourself available for a call, notwithstanding the time difference.

Finally, I would like to thank my family, especially to my uncle, grand-parents, and parents for their relentless support despite the distance.

Contents

ABSTRACT	iii
ACKNOWLEDGEMENTS	v
LIST OF FIGURES	xv
LIST OF TABLES	xxi
1 INTRODUCTION	1
1.1 Motivation	1
1.2 Objectives	4
1.3 Organisation	5
2 BACKGROUND	7
2.1 Environmental hazards, natural risk, seismic risk	7
2.1.1 What are environmental hazards?	7
2.1.2 Hazard (peril) classification	8
2.1.3 What is "Natural Risk"?	8
2.1.4 Evolution and consequences of geophysical and climate-related events	10
2.1.5 Seismic risk	12
2.2 Post-earthquake damage collection	14
2.2.1 Review of existing post-earthquake building damage assessment data collection	14
2.2.2 GEM Inventory Data Capture Tools (IDCT)	15
2.2.3 GEM building Taxonomy v2.0	16
2.2.4 Non-structural components	17
2.3 The principle of "vital few and the useful many"	17
2.4 Seismic damage and loss assessment	18
2.4.1 PEER PBEE framework	18
2.4.2 Current practice in seismic damage and loss modelling	21
2.4.3 Current limitations	23
2.5 The 2010-2011 Canterbury earthquake sequence	24
2.5.1 Generalities	24
2.5.2 Seismic insurance following the Canterbury earthquake sequence	25
2.5.3 The Earthquake Commission (EQC)	27
2.5.4 EQC's catastrophe loss models	28
2.5.5 Earthquake Commission Amendment Act	29
2.6 A primer on machine learning	30
2.6.1 Machine learning, data science, and artificial intelligence	30
2.6.2 Machine learning compared to rule-based systems	31

2.6.3	<i>Types of learning</i>	34
2.6.4	<i>Examples of machine learning application</i>	34
2.6.5	<i>General framework of a machine learning model</i>	35
2.7	Data pre-processing/ Feature engineering	37
2.7.1	<i>Feature extraction and feature engineering</i>	38
2.7.2	<i>Feature selection</i>	40
2.7.3	<i>Feature transformation</i>	40
2.7.4	<i>Training, validation, and test set</i>	41
2.7.5	<i>Class imbalance</i>	41
2.8	Model training	44
2.8.1	<i>Algorithm selection</i>	44
2.8.2	<i>Linear regression</i>	47
2.8.3	<i>Logistic regression</i>	47
2.8.4	<i>Support Vector Machine (SVM)</i>	48
2.8.5	<i>Decision trees</i>	48
2.8.6	<i>Random forest</i>	49
2.9	Model evaluation	49
2.9.1	<i>Performance evaluation of classification models</i>	51
2.9.2	<i>Good fit, overfitting, underfitting</i>	52
2.10	Interpretability of machine learning models	54
2.10.1	<i>Background</i>	54
2.10.2	<i>SHapley Additive exPlanations (SHAP)</i>	55
3	DEVELOPMENT OF A UNIVERSAL DAMAGE DATA COLLECTION FRAMEWORK	57
3.1	Introduction	58
3.2	Improved paper form based on the GEM Building Taxonomy v2.0	59
3.2.1	<i>Field trial of the improved paper form</i>	60
3.2.2	<i>Future improvements</i>	61
3.3	Case study: Calle La Morena	62
3.3.1	<i>Building assessment</i>	62
3.3.2	<i>Statistical findings</i>	62
3.3.3	<i>Examples of observed damage</i>	66
3.4	Conclusion	70
4	DEVELOPMENT OF A MACHINE LEARNING MODEL FOR BUILDING DAMAGE PREDICTION IN THE ROMA AND CONDESA NEIGHBOURHOODS - MEXICO CITY, MEXICO	71
4.1	Introduction	72
4.2	Soil characteristics and seismic recording in Mexico	73
4.3	Building damage and damage distribution	74
4.4	Machine learning model development	76
4.4.1	<i>Problem framing</i>	76
4.4.2	<i>Data collection</i>	77
4.4.3	<i>Data exploration</i>	81
4.4.4	<i>Data preparation</i>	81
4.4.5	<i>Model selection and training</i>	88
4.5	Model prediction	89
4.5.1	<i>Prediction performance</i>	89
4.6	Feature importance Random Forest algorithm	89
4.7	Conclusion	92

5	DATA INTEGRATION FOR THE DEVELOPMENT OF A SEISMIC LOSS PREDICTION MODEL USING EQC'S RESIDENTIAL CLAIM DATABASE	93
5.1	Introduction	94
5.2	Residential building loss data: EQC claims data set	94
5.3	Data collection from additional databases	95
5.3.1	<i>Sourcing building characteristics</i>	95
5.3.2	<i>Seismic demand</i>	97
5.3.3	<i>Liquefaction occurrence</i>	98
5.3.4	<i>MBIE Technical categories</i>	98
5.3.5	<i>Soil conditions</i>	98
5.4	Feature extraction/selection	100
5.4.1	<i>Extract EQC residential building claims related to the CES</i>	100
5.4.2	<i>Select claim status</i>	104
5.5	Overview of the data merging	104
5.6	Merging building characteristics from RiskScape with EQC residential claims	104
5.6.1	<i>Initial merging attempts</i>	107
5.6.2	<i>Alternate approach: constraining the merging using property boundaries</i>	108
5.6.3	<i>Merge LINZ NZ Street Address with LINZ NZ Property Titles</i>	108
5.6.4	<i>Merge RiskScape with LINZ for instances with unique street address per property</i>	110
5.6.5	<i>Filtering the LINZ RiskScape data set for primary dwelling data</i>	110
5.6.6	<i>Properties with two street addresses and one or multiple RiskScape instances</i>	112
5.6.7	<i>LINZ and RiskScape merging for instances with unique and double street address(es) per property</i>	117
5.6.8	<i>Merge EQC claims data with street addresses</i>	117
5.6.9	<i>Multiple EQC instances</i>	117
5.6.10	<i>Merge EQC claims with RiskScape</i>	119
5.7	Add the seismic demand, liquefaction, and soil conditions information to EQC claims database	119
5.8	The number of usable data points through the data merging process	120
5.9	Conclusion	123
6	DATA PRE-PROCESSING AND MODEL DEVELOPMENT OF A SEISMIC LOSS PREDICTION MODEL FOR RESIDENTIAL BUILDINGS - CHRISTCHURCH, NEW ZEALAND	125
6.1	Introduction	126
6.2	Feature filtering	126
6.2.1	<i>EQC attributes</i>	127
6.2.2	<i>RiskScape attributes</i>	129
6.2.3	<i>Filtering of the target attribute: Building Paid</i>	135
6.2.4	<i>Evolution of the number of points during the feature filtering</i>	135
6.3	Processing of the target attribute	136
6.3.1	<i>Building loss ratio or Building Paid</i>	136
6.3.2	<i>Cap</i>	138
6.3.3	<i>Transform BuildingPaid to a categorical attribute</i>	138
6.4	Attribute selection	139
6.4.1	<i>Target attribute: Building Paid - Categorical</i>	140

6.4.2	<i>Liquefaction</i>	140
6.4.3	<i>PGA</i>	140
6.4.4	<i>Construction Type</i>	141
6.4.5	<i>Building Floor Area</i>	141
6.4.6	<i>Floor Type</i>	144
6.4.7	<i>Wall Cladding</i>	144
6.4.8	<i>Deprivation Index</i>	144
6.4.9	<i>Construction year</i>	145
6.4.10	<i>Soil type</i>	145
6.4.11	<i>Latitude and Longitude</i>	149
6.5	Attribute preparation	149
6.5.1	<i>Training, validation, and test set</i>	149
6.5.2	<i>Handling categorical features</i>	150
6.5.3	<i>Handling numerical features</i>	151
6.5.4	<i>Addressing class imbalance</i>	151
6.6	Algorithm selection and training	152
6.7	Model evaluation	153
6.8	Conclusion	156
7	MODEL TESTING AND KNOWLEDGE EXTRACTION FROM THE SEISMIC LOSS PREDICTION MODEL FOR CHRISTCHURCH RESIDENTIAL BUILDINGS	157
7.1	Introduction	157
7.2	Model testing on another event in the CES	158
7.3	Relationship between numerical variables	165
7.4	Feature importance from the random forest model	170
7.4.1	<i>SHAP feature importance</i>	170
7.4.2	<i>Discussion of the results</i>	173
7.5	Conclusion	174
8	CONCLUSIONS	175
8.1	Post-earthquake damage data collection	175
8.2	Machine learning for the seismic damage prediction for residential buildings in the Roma and Condesa neighbourhoods, Mexico City	176
8.3	Machine learning for the seismic loss prediction for residential buildings in Christchurch, New Zealand	177
8.4	Current challenges in the application of machine learning for the prediction of seismic damage and loss	178
8.5	Future work and opportunities	180
A	LOSS DATABASES	183
B	FORMS FOR THE EVALUATION OF SEISMIC BUILDING DAMAGE	185
B.1	ATC-20 Detailed Evaluation Safety Assessment Form	187
B.2	GEM paper based assessment tool	191
B.3	New paper form for the seismic assessment of building based on GEM Building Taxonomy v2.0	195
C	GEM BUILDING TAXONOMY v2.0	203
C.1	Overview of the GEM building taxonomy v2.0	204
C.2	Comparison of the GEM assessment methodology	206

<i>Contents</i>	xiii
D SOIL CODE	207
E NZ DEPRIVATION INDEX 2013 CHRISTCHURCH	211
REFERENCES	215

List of Figures

2.1	Peril classification for natural hazards (Integrated Research on Disaster Risk, 2014)	9
2.2	Natural hazards, risk and consequences (United Nations Office for Disaster Risk Reduction (UNDRR), 2019)	10
2.3	Annual number of geophysical and climate-related events between 1980 and 2019 (Source: EM-DAT, CRED (UCLouvain & Guha-Sapir, 2020))	11
2.4	Earthquakes which caused 10 or more deaths, and/or affected 100 or more persons, and/or triggered the declaration of a state of emergency between 1980 and 2019 (Source: EM-DAT, CRED (UCLouvain & Guha-Sapir, 2020))	13
2.5	Total losses and insurance contribution for earthquake events with absolute losses greater than USD 5B (Source: EM-DAT, CRED (UCLouvain & Guha-Sapir, 2020))	14
2.6	Pareto diagram of the number of customer queries, adapted from (Juran & De Feo, 2010)	19
2.7	Overview of the four steps of PEER PBEE analysis methodology, adapted from (Porter, 2003)	20
2.8	Components of an integrated seismic risk model	21
2.9	Damage assessment in the earthquake risk reduction	22
2.10	Uncertainties in the analytical fragility assessment methodology (Maio and Tsionis, 2015)	23
2.11	Maps of the Canterbury Region (A) New Zealand map with main cities labelled. (B) Canterbury Region, with districts labelled as 1) Kaikoura; 2) Hurunui; 3) Waimakariri; 4) Christchurch City; 5) Selwyn; 6) Ashburton; 7) Timaru; 8) Mackenzie; 9) Waimate; 10) Waitaki. (C) Map of the Christchurch city area and nearby towns (Potter et al., 2015).	26
2.12	Location of the main events in the 2010-2011 Canterbury earthquake sequence (O'Rourke et al., 2014).	26
2.13	Cumulative number of aftershocks (with magnitude $M_w \geq 3.0$) in the CES, adapted from (Reyners et al., 2014)	27
2.14	(a) Overall Minerva system architecture, (b) Schematic diagram of the Earthquake Loss sub system used in Minerva (Shephard et al., 2002)	29
2.15	Core concepts constituting the field of data science (Barber, 2014)	31
2.16	Venn diagram of artificial intelligence, machine learning, and data science (Kotu & Deshpande, 2019)	32
2.17	Schematic overview of rule-based systems, machine learning, and deep learning systems. Grayed box highlights elements that can learn from data (Goodfellow et al., 2016)	33
2.18	Machine learning can help humans understand a problem better (Géron, 2019)	33
2.19	Machine learning project life cycle (Burkov, 2020)	37
2.20	Main steps of a machine learning model, adapted from (Géron, 2019)	38

2.21	Tabular tidy data. The columns represents attributes and the rows examples (Burkov, 2020)	39
2.22	Overview of the train/test splitting when there is no validation set	42
2.23	Overview of techniques for over- and under-sampling implemented in the imbalanced-learn Python toolbox (Lemaitre et al., 2017a)	43
2.24	How to choose a machine learning algorithm? (Pedregosa et al., 2019)	45
2.25	Overview of main machine learning algorithms (Kuhn & Johnson, 2013)	46
2.26	Overview of the Random Forest algorithm, adapted from (Hastie et al., 2009) and (Efron & Tibshirani, 1986)	50
2.27	Details of a confusion matrix	51
2.28	Bias-variance trade-off	53
2.29	Feature importance computed using SHAP (Lundberg, 2020)	56
3.1	Screenshots of the GEM IDCT software. (a) Depicting building boundaries as available from shapefiles. (b) Location of the 25 buildings assessed.	63
3.2	Distribution of damage categories for the 25 buildings studied in Calle La Morena. Damage categories as per EMS-98.	64
3.3	(a) Building categorized by occupancy and number of storey. (b) Detail of the building occupancy.	65
3.4	Damage grade distribution categorised by adjoining buildings and plan shape. (a) Configuration of adjoining buildings. (b) Plan shape of each building. (c) Combination of plan shape and building position.	66
3.5	Damage grade distribution categorised by adjoining buildings and plan shapes. Material of structural system in (a) longitudinal direction and (b) transversal direction. Lateral Load Resisting system in (c) longitudinal direction and (d) transversal direction.	67
3.6	Distribution of damage for regular and irregular structures	67
3.7	Distribution of damage for irregular structures. Principal (a) vertical and (b) horizontal irregularities. Secondary (c) vertical and (d) horizontal irregularities.	68
3.8	Representative damage in La Morena. (a) Corner building with torsion eccentricity. (b) Pounding failure. (c) Soft storey in the ground floor.	69
3.9	(a) Failure of the column in shear. (b) Buckling of the longitudinal reinforcement bars.	69
4.1	Localisation of historical seismic events in Mexico (Servicio Sismológico Nacional (SSN), 2017)	72
4.2	Location of the CIRES recording stations (Centro de Instrumentación y Registro Sísmico (CIRES), 2017) over a map of Mexico City	73
4.3	Preliminary classification of inspected structures according to the level of damage, after (Colegio de Ingenieros Civiles de México (CICM), 2017a)	74
4.4	Damage severity of the surveyed buildings located in the Roma and Condesa neighborhoods	75
4.5	Statistics on the UAM building damage database: categorised by (a) number of storeys, (b) construction year, (c) material type, (d) lateral load resisting system	76
4.6	Locations of the buildings assessed by the UAM team superimposed over a map of the geotechnical zones of Mexico city	77
4.7	Buildings assessed by the UAM team in the Roma and Condesa neighbourhoods	78
4.8	Inverse distance weighted (IDW) interpolation of the PGA values between the CIRES recording stations	80

4.9	Steps to derive the seismic demand for each building	82
4.10	Detailed description of the damage grade (Grünthal, 1998)	84
4.11	(a) Number of training data point available as five damage classes; (b) Data points distribution after transforming target feature as a binary damage class	85
4.12	Pair plots showing the relationship between the variables: number of stories, natural period, PGA, and S_A . The hue represents the damage grade	86
4.13	Pearson correlation coefficient before pre-processing of the database	87
4.14	Graphical representation of missing values (on the raw database)	87
4.15	Overview of categorical features after one-hot encoding	88
4.16	Performance of the Random Forest (RF) algorithm	91
4.17	Feature importance based on Shapley value	91
5.1	Graphical overview of the raw data in the EQC claims database for the Canterbury earthquake sequence. The columns represent attributes and the rows examples. White areas represent missing values. Column 4 represents the PortfolioID, columns 5 and 6 the longitude and latitude respectively.	95
5.2	Interpolation of PGA 4Sep2010	97
5.3	Maps showing the central part of Christchurch with the Peak Ground Acceleration and liquefaction occurrence for the (a) 4 September 2010, (b) 22 February 2011, (c) 13 June 2011 and (d) 23 December 2011 earthquakes (J. Russell & van Ballegooy, 2015)	99
5.4	Liquefaction occurrence for 22 February 2011, data from (Earthquake Commission (EQC) et al., 2012)	100
5.5	Map of CERA “Red Zone” and MBIE Residential Technical Category (Ministry of Business Innovation & Employment (MBIE), 2012)	101
5.6	Map showing the NZSC soil order classification in Christchurch (layer obtained from (Land Resource Information Systems (LRIS), 2010)). Information of the soil codes can be found in Table D.1 in Appendix D.	102
5.7	Number of claims and property for events in the CES with more than 1,000 instances	103
5.8	Steps to extract EQC data for a unique event in the Canterbury earthquake sequence (CES)	103
5.9	Number of instances for each category in the attribute ClaimStatus	105
5.10	Number of claims and property for events in the CES after filtering for ClaimStatus. Only events with more than 1,000 instances prior to cleansing are shown.	106
5.11	Schematic overview of the merging of information on top of EQC’s claims data	106
5.12	Comparison of the spatial location of the EQC claims data points and the RiskScape buildings	108
5.13	Merging of LINZ NZ street address with LINZ NZ property tiles	109
5.14	Map of an urban block in Christchurch overlaid with the LINZ NZ Street Address and LINZ NZ Property Titles layers. This highlights some Property Titles do not have a matching LINZ NZ Street Address.	110
5.15	Satellite image of urban blocks in Christchurch overlaid with the LINZ NZ Street Address and LINZ NZ Property Titles layers. The polygons with a bold red border represent LINZ NZ property titles having only one street address.	111
5.16	Map view of selected LINZ NZ property titles	112
5.17	Satellite view of an urban block in Christchurch with RiskScape points and selected LINZ NZ Property Titles	113
5.18	Steps to filter RiskScape data including secondary buildings to RiskScape data with residential buildings only	113

5.19	Property with one LINZ street address but multiple RiskScape points	114
5.20	Property with two LINZ NZ street address points and two RiskScape points	115
5.21	Property with two LINZ NZ street address points and three RiskScape points	115
5.22	Neighbouring properties having two LINZ NZ street addresses and two RiskScape points each leading to issues with the “spatial join - closest”	116
5.23	Steps to merge EQC and RiskScape using the LINZ NZ street address	119
5.24	Steps to add the seismic demand, the liquefaction occurrence, the location of MBIE Technical categories, and the soil conditions on top of EQC claims	120
5.25	The number of data points after each processing step for event on 4 September 2010 and 22 February 2011	122
6.1	Overview of the steps for the filtering of the EQC features	127
6.2	Number of instances for each value in Number of Dwelling insured	128
6.3	Number of instances for EQC Building Sum Insured (categorised)	129
6.4	Number of instances for which Building Paid equals Building Net Incurred	130
6.5	Number of instances per ‘Use Category’	130
6.6	Distribution of Building Paid against Building Floor area	131
6.7	Distribution of Building Floor area (selected below 1,000 sqm)	131
6.8	Number of instances for each Construction Type category	132
6.9	Number of instances for each category of Floor Type	132
6.10	Number of instances for each category of Deprivation index (DepInd01 = least deprived to DepInd10=Most deprived)	133
6.11	Number of instances for each category of Wall Cladding	134
6.12	Number of instances for each category of Roof Cladding	134
6.13	Number of instances for each soil category	135
6.14	Distribution of BuildingPaid after selection of the instances between NZ\$0 and NZ\$115,000	136
6.15	Overview of the filtering steps for the RiskScape attributes	136
6.16	Evolution of the number of instances after each feature filtering step	137
6.17	Schematic overview of the thresholds for the transformation of Building Paid from a categorical to a numerical attribute	139
6.18	Selected attributes for the model using data from 4Sep2010	139
6.19	Number of instances in Building Paid categorical in the filtered data set	140
6.20	Number of instances in the filtered data set which experienced liquefaction	141
6.21	Distribution of PGA in the filtered data set	142
6.22	Number of instances per Construction Type in the filtered data set	143
6.23	Number of instances by Building Floor Area in the filtered data set	143
6.24	Number of instances by Building Floor Area in the filtered data set	144
6.25	Number of instances by Building Floor Area in the filtered data set	145
6.26	Number of instances by Deprivation Index in the filtered data set	146
6.27	Number of instances by construction period in the filtered data set	147
6.28	Number of instances per soil type in the filtered data set	148
6.29	Overview of the training, validation, and test data sets and their usage in the development of a machine learning seismic loss model for Christchurch	150
6.30	Correlation matrix between the numerical features	151
7.1	Confusion matrices for the random forest algorithm	159
7.2	Pairplots for the numerical attributes - 4 September 2010	166
7.3	Pairplots for the numerical attributes - 22 February 2011	167
7.4	Pairplots for the numerical attributes - 13 June 2011	168

7.5	Pairplots for the numerical attributes - 23 December 2011	169
7.6	SHAP feature importance for the random forest model (4 September 2010) . .	171
7.7	SHAP feature importance for the random forest model (22 February 2011) . .	171
7.8	SHAP feature importance for the random forest model (13 June 2011)	172
7.9	SHAP feature importance for the random forest model (23 Dec 2011)	172
E.1	NZDep2013 Index of Deprivation Christchurch City Area Units (Christchurch City Council, 2015)	213

List of Tables

2.1	Data science tasks (non exhaustive) and examples, adapted from (Kotu & Deshpande, 2019)	36
2.2	List of some intrinsic interpretable machine learning algorithms, after (Molnar, 2020)	55
4.1	PGA and max spectral acceleration for the CIRES stations	79
4.3	Coefficient a for building period T_1 in Mexico City, after (Muria-Vila & Gonzalez-Alcorta, 1995)	81
4.5	Features present in the final dataset (after addition of supplementary information)	83
4.6	Number of values available for each feature	88
4.7	Damage prediction accuracy for logistic regression, support vector machine, decision trees, and random forest models	90
5.1	Overview of selected features in the RiskScape data set	96
5.2	Future land performance and foundation criteria for MBIE Technical Categories (TC), adapted from (Ministry of Business Innovation & Employment (MBIE), 2012) and (J. Russell & van Ballegooy, 2015)	99
5.3	Overview of the action taken depending on the number of LINZ NZ street address and RiskScape point present per LINZ NZ property title	118
6.1	Model evaluation for logistic regression, SVM, decision trees, and Random Forest for the 4 September 2010 data	154
6.2	Model evaluation for Random Forest model for 4Sep2010, 22Feb2011, 13June2011, and 23Dec2011	155
7.1	Random forest model for the 4 September 2010 tested on the 22 February 2011, 13 June 2011, and 23 December 2011	161
7.2	Random forest model for the 22 February 2011 tested on the 4 September 2010, 13 June 2011, and 23 December 2011	162
7.3	Random Forest model for the 13 June 2011 tested on the 4 September 2010, 22 February 2011, and 23 December 2011	163
7.4	Random forest model for the 23 December 2011 tested on the 4 September 2010, 22 February 2011, and 13 June 2011	164
7.5	Five most important features according to the SHAP values for the random forest model	170
A.1	Overview of the main loss databases, adapted from (Integrated Research on Disaster Risk, 2014)	184
C.1	Overview of the GEM building taxonomy v2.0 categories	204

C.2 Comparison of the GEM assessment methodology vs. the local Mexican procedure 206

D.1 Soil code (Land Resource Information Systems (LRIS), 2010) 208

Co-Authorship Forms

Roeslin, S., Ma, Q. T. M., & García, H. J. (2018). Damage Assessment on Buildings Following the 19th September 2017 Puebla, Mexico Earthquake. *Frontiers in Built Environment*, 4, 18. <https://doi.org/10.3389/fbuil.2018.00072>

Roeslin, S., Ma, Q., Juárez-García, H., Gómez-Bernal, A., Wicker, J., & Wotherspoon, L. (2020). A machine learning damage prediction model for the 2017 Puebla-Morelos , Mexico, earthquake. *Earthquake Spectra*, 36(S2), 1–26. <https://doi.org/10.1177/8755293020936714>

Roeslin, S., Ma, Q., Wicker, J., & Wotherspoon, L. (2020). Data Integration for the Development of a Seismic Loss Prediction Model for Residential Buildings in New Zealand. In P. Cellier & K. Driessens (Eds.), *Machine Learning and Knowledge Discovery in Databases* (Vol. 1168, pp. 88–100). Springer International Publishing. https://doi.org/10.1007/978-3-030-43887-6_8

Roeslin, S., Ma, Q., Chigullapally, P., Wicker, J., & Wotherspoon, L. (2020). Feature Engineering for a Seismic Loss Prediction Model Using Machine Learning, Christchurch Experience. *Proceeding of the 17th World Conference on Earthquake Engineering, 17WCEE*.

Co-Authorship Form

This form is to accompany the submission of any PhD that contains published or unpublished co-authored work. **Please include one copy of this form for each co-authored work.** Completed forms should be included in all copies of your thesis submitted for examination and library deposit (including digital deposit), following your thesis Acknowledgements. Co-authored works may be included in a thesis if the candidate has written all or the majority of the text and had their contribution confirmed by all co-authors as not less than 65%.

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.

Chapter 3

Roeslin, S., Ma, Q. T. M., & García, H. J. (2018). Damage Assessment on Buildings Following the 19th September 2017 Puebla, Mexico Earthquake. *Frontiers in Built Environment*, 4, 18. <https://doi.org/10.3389/fbuil.2018.00072>

Nature of contribution by PhD candidate	All aspects of the paper
---	--------------------------

Extent of contribution by PhD candidate (%)	95
---	----



CO-AUTHORS

Name	Nature of Contribution
Quincy T. M. Ma	Manuscript review and edits
Prof. Hugón Juárez García	Help with data collection in Mexico City

Certification by Co-Authors

The undersigned hereby certify that:

- ❖ the above statement correctly reflects the nature and extent of the PhD candidate's contribution to this work, and the nature of the contribution of each of the co-authors; and
- ❖ that the candidate wrote all or the majority of the text.

Name	Signature	Date
Quincy T. M. Ma		14/01/2021
Hugón Juárez García		Jan, 13, 2021

Co-Authorship Form

This form is to accompany the submission of any PhD that contains published or unpublished co-authored work. **Please include one copy of this form for each co-authored work.** Completed forms should be included in all copies of your thesis submitted for examination and library deposit (including digital deposit), following your thesis Acknowledgements. Co-authored works may be included in a thesis if the candidate has written all or the majority of the text and had their contribution confirmed by all co-authors as not less than 65%.

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.

Chapter 4

Roeslin, S., Ma, Q., Juárez-García, H., Gómez-Bernal, A., Wicker, J., & Wotherspoon, L. (2020). A machine learning damage prediction model for the 2017 Puebla-Morelos, Mexico, earthquake. *Earthquake Spectra*, 36(S2), 1-26.
<https://doi.org/10.1177/8755293020936714>

Nature of contribution by PhD candidate	All aspects of the paper
---	--------------------------

Extent of contribution by PhD candidate (%)	90
---	----

CO-AUTHORS

Name	Nature of Contribution
Dr Quincy Ma	Methodology input, manuscript review and edits
Prof. Hugón Juárez García	Manuscript review and edits
Prof. Alonso Gómez-Bernal	Manuscript review and edits
Dr Joerg Wicker	Methodology input, manuscript review and edits
Assoc. Prof. Liam Wotherspoon	Manuscript review

Certification by Co-Authors

The undersigned hereby certify that:

- ❖ the above statement correctly reflects the nature and extent of the PhD candidate's contribution to this work, and the nature of the contribution of each of the co-authors; and
- ❖ that the candidate wrote all or the majority of the text.

Name	Signature	Date
Quincy Ma		14/01/2021
Hugón Juárez García		Dec / 3 / 2020
Alonso Gómez-Bernal		2/12/20
Joerg Wicker		26/11/20
Liam Wotherspoon		27/11/20

Co-Authorship Form

This form is to accompany the submission of any PhD that contains published or unpublished co-authored work. **Please include one copy of this form for each co-authored work.** Completed forms should be included in all copies of your thesis submitted for examination and library deposit (including digital deposit), following your thesis Acknowledgements. Co-authored works may be included in a thesis if the candidate has written all or the majority of the text and had their contribution confirmed by all co-authors as not less than 65%.

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.

Chapter 5

Roeslin S., Ma Q., Wicker J., Wotherspoon L. (2020) Data Integration for the Development of a Seismic Loss Prediction Model for Residential Buildings in New Zealand. In: Cellier P., Driessens K. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2019. Communications in Computer and Information Science, vol 1168. Springer, Cham. https://doi.org/10.1007/978-3-030-43887-6_8

Nature of contribution by PhD candidate	All aspects of the paper
Extent of contribution by PhD candidate (%)	95

CO-AUTHORS

Name	Nature of Contribution
Dr Quincy Ma	Manuscript review and edits
Dr Joerg Wicker	Manuscript review
Assoc. Prof. Liam Wotherspoon	Manuscript review

Certification by Co-Authors

The undersigned hereby certify that:

- ❖ the above statement correctly reflects the nature and extent of the PhD candidate's contribution to this work, and the nature of the contribution of each of the co-authors; and
- ❖ that the candidate wrote all or the majority of the text.

Name	Signature	Date
Quincy Ma		14/01/2021
Joerg Wicker		26/11/20
Liam Wotherspoon		27/11/20

Co-Authorship Form

This form is to accompany the submission of any PhD that contains published or unpublished co-authored work. **Please include one copy of this form for each co-authored work.** Completed forms should be included in all copies of your thesis submitted for examination and library deposit (including digital deposit), following your thesis Acknowledgements. Co-authored works may be included in a thesis if the candidate has written all or the majority of the text and had their contribution confirmed by all co-authors as not less than 65%.

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.

Chapter 5

Roeslin, S., Ma, Q., Chigullapally, P., Wicker, J., & Wotherspoon, L. (2020). Feature Engineering for a Seismic Loss Prediction Model Using Machine Learning, Christchurch Experience. Proceeding of the 17th World Conference on Earthquake Engineering, 17WCEE.

Nature of contribution by PhD candidate	All aspects of the paper
---	--------------------------

Extent of contribution by PhD candidate (%)	85
---	----

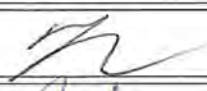

CO-AUTHORS

Name	Nature of Contribution
Dr Quincy Ma	Manuscript review and edits
Pavan Chigullapally	Code review and edits, Manuscript review
Dr Joerg Wicker	Manuscript review
Assoc. Prof. Liam Wotherspoon	Manuscript review

Certification by Co-Authors

The undersigned hereby certify that:

- ❖ the above statement correctly reflects the nature and extent of the PhD candidate's contribution to this work, and the nature of the contribution of each of the co-authors; and
- ❖ that the candidate wrote all or the majority of the text.

Name	Signature	Date
Dr Quincy Ma		14/01/2021
Pavan Chigullapally		26/11/2020
Dr Joerg Wicker		26/11/20
Assoc. Prof. Liam Wotherspoon		27/11/20

Introduction

1.1 Motivation

In 2010-2011, New Zealand experienced the costliest earthquakes in its history. The Canterbury earthquake sequence (CES) began on 4 September 2010 with the M_w 7.1 Darfield earthquake and it continued with more than 3,500 aftershocks, including major shaking events on 22 February 2011, 13 June 2011 and 23 December 2011 (Potter et al., 2015). The CES induced unprecedented and widespread damage in the Christchurch built environment, devastating many commercial buildings in the central business district (CBD) and residential houses in wider Christchurch. The CES led to a direct economic loss evaluated at 20% of New Zealand's GDP in 2011 (King et al., 2014). The losses far exceeded any previous model prediction (Feltham, 2011), prompting a rethink on seismic loss prediction models for New Zealand.

On the 19th September 2017, on the exact 28th anniversary of the 1985 Michoacán earthquake, a M_w 7.1 earthquake occurred 120 km southeast of Mexico City. The 2017 Puebla earthquake induced significant damage to structures located in the Mexico City urban area, led to the collapse of 46 buildings and 370 fatalities. These seismic damage and losses to the built environment emphasised the need for improved disaster risk reduction. Accurate seismic damage and loss prediction can improve efficient resources allocation to increase the seismic resilience of cities and communities.

Since 2000's, research efforts focused on performance based seismic engineering of buildings (Federal Emergency Management Agency (FEMA), 1997; Poland et al.,

1995). The Pacific Earthquake Engineering Research (PEER) Center further framed the Performance-based earthquake engineering (PBEE) methodology (Porter, 2003). First implementations of the PBEE methodology defined the building performance in terms of discrete performance levels (i.e. fully operational, operational, life-safe, and near collapse) linked to specific earthquake intensity levels. The definition of the building performance then evolved to be expressed via quantitative metrics (e.g. casualties, repair cost, repair time, environmental impacts, and unsafe placarding), which are usable by structural engineers as well as other stakeholders and decision-makers (Federal Emergency Management Agency (FEMA), 2018).

The PBEE takes into account uncertainties and thus treats performance probabilistically. This includes the use of performance functions, fragility and vulnerability functions, that communicate the probability of non-exceedance for a given impact quantity. Such functions are developed upon analytical data, expert opinion, and in some cases empirical data (Porter, 2020). Empirical observation data from actual events is highly valued as it comes from the direct inspection of buildings subjected to earthquakes and thus reflects actual achieved performance. In recent years, reconnaissance missions following earthquake events provided valuable perishable information related to building damage. The analysis of the damage data provides a better understanding of building performance pointing out causes of building failure and highlighting deficiencies in previous and current construction practice. The knowledge gained through the analysis of empirical damage data analysis enables the development of damage prediction tools that improve future building damage assessment and seismic risk mitigation. Perhaps owing to the historic lack of post-earthquake building performance data, current loss estimation tools primarily follow a probabilistic causation-based approach rather than an actuarial approach as it is common for non-natural perils. A central concept of this probabilistic-causation approach is represented by the PEER triple integral for calculating expected loss (Porter, 2003), as shown in Equation 1.1 below.

$$P(DV) = \int \int \int G(DV|DM)dG(DM|EDP)dG(EDP|IM)|d\lambda(IM)| \quad (1.1)$$

It is often a complex and time-consuming process to adopt new event observations to update fragility and vulnerability models in the probabilistic causation framework. Significant effort is required to prepare and interpret the empirical post-earthquake building assessment data. There are often sampling bias on damaged structures, inconsistencies and individual subjectivity due to the collection technique and the assessment form employed. There are also further problems capturing sufficiently detailed data to account for regional nuance and local building practices. Localised assessment form capture more detailed results but hamper a comparison between damage data collected from different regions of the world.

The analysis of empirical damage data also requires significant time, leading to a gap of several years between the data collection and new insights being implemented in updated damage and loss models. For example, the new damage probability matrices for Greece were published nine years following the 1999 Athens earthquake (Eleftheriadou & Karabinis, 2008), empirical fragility curves for reinforced concrete (RC) buildings were published seven years after the 2009 L'Aquila earthquake (Del Gaudio et al., 2016).

Lastly, damage and loss prediction models are important tools for governments, insurers, and engineers. They facilitate the planning and prioritisation of earthquake risk mitigation projects and the selection of risk financing options. Due to the many engineering, social and economical input to these models, it is often difficult to identify the key building parameters attributing to the most damage. A technique to understand the influence of each model parameter on the final outputs would enable more informed decisions.

It is the aforementioned challenges combined with the unrealised potential of empirical data collected following earthquake events that motivated this study, to apply data science techniques to the analysis of empirical post-earthquake damage and loss data. Kovačević et al. (2018) highlighted the potential for machine learning (ML) to allow for more flexible and more rapid earthquake loss assessment of residential buildings.

1.2 Objectives

Main objectives

- To improve the framework for post-earthquake damage data collection
- To develop a machine learning model for the seismic damage and loss prediction for residential buildings
- To develop a process to rapidly and accurately identify key building parameters contributing to seismic damage and loss

The first objective of this study is to improve the framework for collecting building damage data following earthquake events. The aim is to develop a building evaluation form that is general enough to compare building damage data across the world while providing sufficient flexibility to account for local construction nuances.

The second objective is to use empirical data to develop models for the seismic damage and loss prediction in residential buildings. Specifically, the following sub-objectives are to be addressed: collect post-event building evaluations (damage and loss data), obtain information on the seismic demand for the buildings observed, aggregate damage and loss evaluation with additional information related to the buildings and its surroundings (e.g. seismic demand, soil information), pre-process the data employing engineering judgement, apply machine learning to the curated data set, evaluate the performance of several algorithms taking into account the trade-off between accuracy and model interpretation.

The third and final objective is to identify key parameters contributing to damage and loss in residential buildings in Christchurch, New Zealand. This task is achieved through interrogating the previously developed machine learning models on issues such as feature importance. The study will focus on generating statistics and insights that are useful for engineers, the insurance sector, and risk managers.

1.3 Organisation

The outputs of this research are presented over Chapters 3 to 7. A review of the published literature is presented in Chapter 2, and a summary of the findings is provided in Chapter 8. Chapter 3 addresses the first objective related to the improvement of the methodology for seismic damage data collection. Chapters 4 to 6 cover the second objective on developing a damage and a loss prediction model using machine learning. Chapters 4 and 7 examine what parameters drive building damage and loss in earthquakes.

Chapter 2 introduces the background information related to seismic damage and covers the necessary material and concepts related to machine learning. First, the chapter covers generalities and the scientific treatment of natural events, their consequences and seismic risk. Secondly, it provides a review of current post-earthquake damage collection frameworks with a focus on the tools developed by the Global Earthquake Model (GEM). Thirdly, it introduces the concept of “vital few and useful many”. Fourthly, the chapter reviews and discusses current practices in seismic damage and loss modelling. Fifthly, it provides a background on the Canterbury earthquake sequence (CES). Finally, it introduces key concepts and current developments on data science and specifically machine learning.

Chapters 3 and 4 outline the experience and research following the author’s participation as a New Zealand team member in the 2017 Puebla earthquake reconnaissance mission.

Chapter 3 presents a new paper form for post-earthquake building damage data collection. It also report on the experience when the new form was trialled in November 2017 in Mexico City following the 2017 Puebla earthquake. The chapter presents a damage analysis case study for Calle La Morena based on the empirical data collected.

Chapter 4 presents the development of a machine learning model for seismic damage prediction in the Roma and Condesa neighbourhoods in Mexico City. The chapter describes in detail the data preparation, the addition of supplementary information, data pre-processing, training of the machine learning model, and the selection criteria for different algorithms.

Chapters 5 to 7 present the development of a loss prediction model for residential buildings using data collected in Christchurch, New Zealand, following the 2010-2011 Canterbury earthquake sequence.

Chapter 5 describes the process of aggregating information from multiple databases, or data merging. This process introduces useful information from databases belonging to public and private organisations related to the seismic demand, building inventory or soil conditions. This improved the machine learning model but is also necessary to overcome missing data in the EQC claim database.

Chapter 6 presents the development of a machine learning model for the seismic loss prediction of residential buildings in Christchurch, New Zealand. It explains the data filtering, pre-processing of the target, attribute selection and the preparation prior to the application of machine learning algorithms. Then, it describes the algorithm selection, training and the model evaluation.

Chapter 7 deals with the model testing and knowledge extraction from the merged data set and developed model. The first part of the chapter is concerned with the prediction performance of each model developed in the previous chapter on unseen data. The model generalisation and prediction performance are tested with data from other main events of the Canterbury earthquake sequence. The second part of the chapter deals with the presentation of findings and derivation of insights from the merged data set and machine learning models. This section studies the relationship between the numerical model attributes and the predicted and actual building losses from a statistical and machine learning model standpoint.

Finally, **Chapter 8** summarizes the main conclusions of this study. It highlights the challenges and limitations, discusses the importance and usefulness interpretable machine learning model for derivation of actionable insights for insurers, engineers and emergency planners, and provides recommendations for future extensions of this research.

Background

This chapter presents a review of the literature with the objective of introducing key concepts related to seismic risk and modelling. The review begins with defining seismic risk and damage and loss in the context of large recent earthquakes. The review then examines available post-earthquake damage collection methodologies and current practices in seismic damage and loss modelling. The review also presents the principle of “vital few and trivial many” also known as the Juran or Pareto principle and how it can be applied to seismic risk and impact studies. The chapter then describes the 2010-2011 Canterbury earthquake sequence (CES) and gives an introduction to New Zealand’s unique insurance setting. Finally, the review provides a primer on data science, especially machine learning (ML). It explains the concept of machine learning, describes the main elements in the development of ML models, explores the operation and details of key ML algorithms, discusses limitations of ML, and lists examples of several previous civil and earthquake engineering ML studies.

2.1 Environmental hazards, natural risk, seismic risk

2.1.1 What are environmental hazards?

The United Nations (2016) defines a hazard as “a process, phenomenon or human activity that may cause loss of life, injury or other health impacts, property damage, social and economic disruption or environmental degradation”. Hazards may have different

origins: natural, anthropogenic (human-induced), socionatural (combination of natural and anthropogenic factors). Hazards can happen as a single event but can also occur as a sequence and be combined in their source and outcomes. Hazards are usually distinguished by their intensity, frequency, probability and location of occurrence.

2.1.2 Hazard (peril) classification

In the insurance sector, the word 'peril' is sometimes used exchangeably for the concept of 'hazard' as defined previously (United Nations Office for Disaster Risk Reduction (UNDRR), 2015). In order to compare the consequences of various hazards, stakeholders of the Integrated Research on Disaster Risk (IRDR) research programme established a structure to classify perils (Integrated Research on Disaster Risk, 2014). Figure 2.1 shows a list of common perils and their respective main event and hazard family. A peril can be related to multiple main events. The IRDR classification distinguishes six general families of hazards: geophysical, hydrological, climatological, biological, extraterrestrial. Many loss databases follow the IRDR peril classification scheme. A list of the major loss databases is provided in Appendix A for information.

2.1.3 What is "Natural Risk"?

In a non-technical context, "risk" is often understood as a situation involving exposure to danger and set of circumstances that might lead to an undesirable outcome and adverse consequences (Oxford English Dictionary (OED) Online, 2010). In a technical context, Hansson (2018) differentiates the definition of "risk" from a qualitative and quantitative sense. From a qualitative definition, "risk" refers to the unwanted event itself or the cause of the unwanted event. In a quantitative sense, "risk" expresses the probability of the occurrence of an unwanted event. It sometimes describes the process of decision making in cognizance of the possible adverse consequences. The most common definition of risk in technical context is the expected value of an undesirable outcome expressed as the product of risk probabilities and its severity (Hansson, 2018).

In the context of natural hazards, risk is defined as the combination of the hazard, exposure and the vulnerability of the asset in question (Porter, 2020). Hazard defines the frequency and intensity of phenomena. Exposure defines the degree of presence of

Family	Main Event	Peril
Geophysical	Earthquake Mass Movement Volcanic Activity	Ash Fall Fire following EQ Ground Movement Landslide following EQ Lahar Lava Flow Liquefaction Pyroclastic Flow Tsunami
Hydrological	Flood Landslide Wave Action	Avalanche: Snow, Debris Coastal Flood Coastal Erosion Debris/Mud Flow/Rockfall Expansive Soil Flash Flood Ice Jam Flood Riverine Flood Rogue Wave Seiche Sinkhole
Meteorological	Convective Storm Extratropical Storm Extreme Temperature Fog Tropical Cyclone	Cold Wave Derecho Frost/Freeze Hail Heat Wave Lightning Rain Sandstorm/Dust storm Snow/Ice Storm Surge Tornado Wind Winter Storm/Blizzard
Climatological	Drought Glacial Lake Outburst Wildfire	Forest Fire Land fire: Brush, Bush, Pasture Subsidence
Biological	Animal Incident Disease Insect Infestation	Bacterial Disease Fungal Disease Parasitic Disease Prion Disease Viral Disease
Extraterrestrial	Impact Space Weather	Airburst Collision Energetic Particles Geomagnetic Storm Radio Disturbance Shockwave

Figure 2.1: Peril classification for natural hazards (Integrated Research on Disaster Risk, 2014)



Figure 2.2: Natural hazards, risk and consequences (United Nations Office for Disaster Risk Reduction (UNDRR), 2019)

people, buildings, infrastructure in the hazard prone areas. Vulnerability is a measure of the damageability or fragility of the asset which can be physical, social-economical, or environmental. Natural risk only arises if all three components are present together. The United Nations Office for Disaster Risk Reduction (UNDRR) expresses the consequences of natural risks in terms of death, damage and losses (Figure 2.2). Alternatively, recent engineering practice also uses the death, damage and downtime metrics (Dhakal, 2011).

2.1.4 Evolution and consequences of geophysical and climate-related events

Between 1980 and 2019, the Centre for Research on the Epidemiology of Disasters (CRED) reported more than 21,000 disaster events that led to 10 or more deaths, and/or 100 or more people affected, and/or the declaration of a state of emergency (EM-DAT, CRED (UCLouvain & Guha-Sapir, 2020)). Among them, 61.4% originated from natural hazards. While being natural hazards, the biological and extraterrestrial families as shown in Figure 2.1 are not considered in this study. Instead, the following observations focus on geophysical and climate-related events encompassing the geophysical, hydrological, meteorological, and climatological families of hazards.

Figure 2.3a shows the annual number of geophysical and climate-related events for the period 1980-2019 for each disaster type. Figure 2.3b presents this same data as a stacked column graph highlighting the relative occurrence frequency. Floods and storms are the most occurring geophysical and climate-related events contributing to 40% and 30% of the total respectively. Earthquakes, the next most frequent event type, only

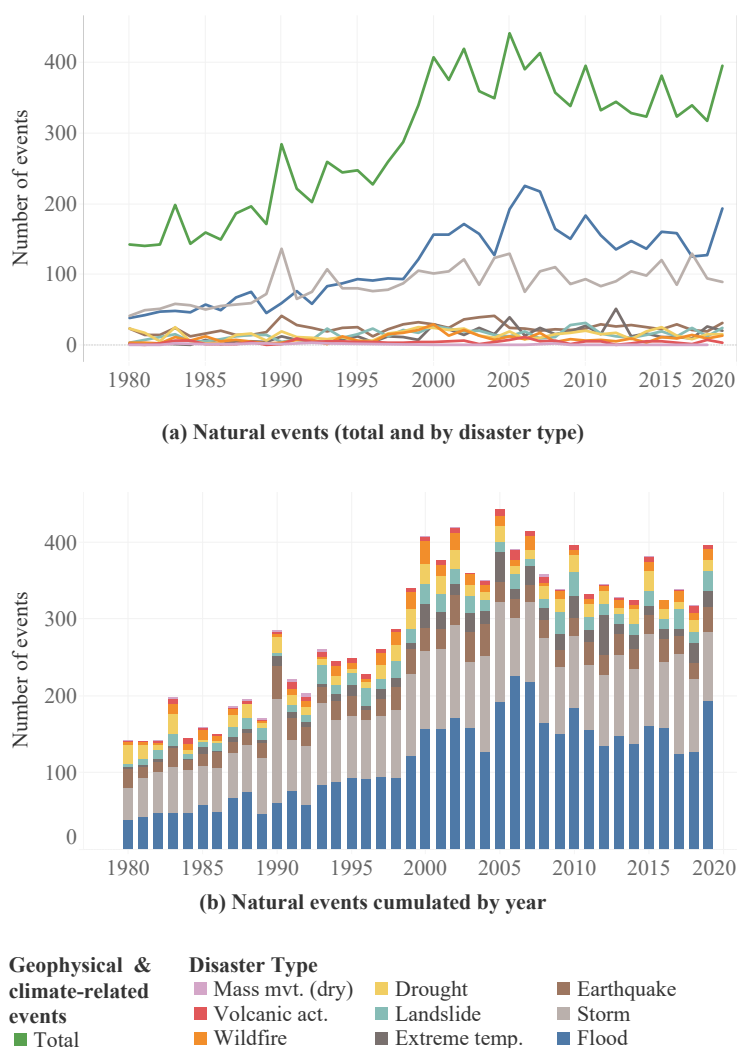


Figure 2.3: Annual number of geophysical and climate-related events between 1980 and 2019 (Source: EM-DAT, CRED (UCLouvain & Guha-Sapir, 2020))

account for 9% of the total. Over the 1980-2019 period, China experienced the highest number of geophysical and climate-related events (861 events), closely followed by the USA (849 events). China also recorded the largest number of earthquakes with over 140 seismic events which led to 100 or more people affected, and/or 10 or more deaths.

Geophysical and climate-related events led to more than 2.4 million casualties between 1980 and 2019. Earthquakes and droughts are the most critical and alone accounted for almost 1.5 million deaths. Ethiopia, Haiti, and Indonesia heavily suffered from droughts and earthquakes with 730,000 casualties over the 1980-2019 period.

Storm is the geophysical and climate-related hazard that led to most of the absolute losses (US\$1,472B) for the 1980-2019 period. Flood induced US\$813B absolute losses and

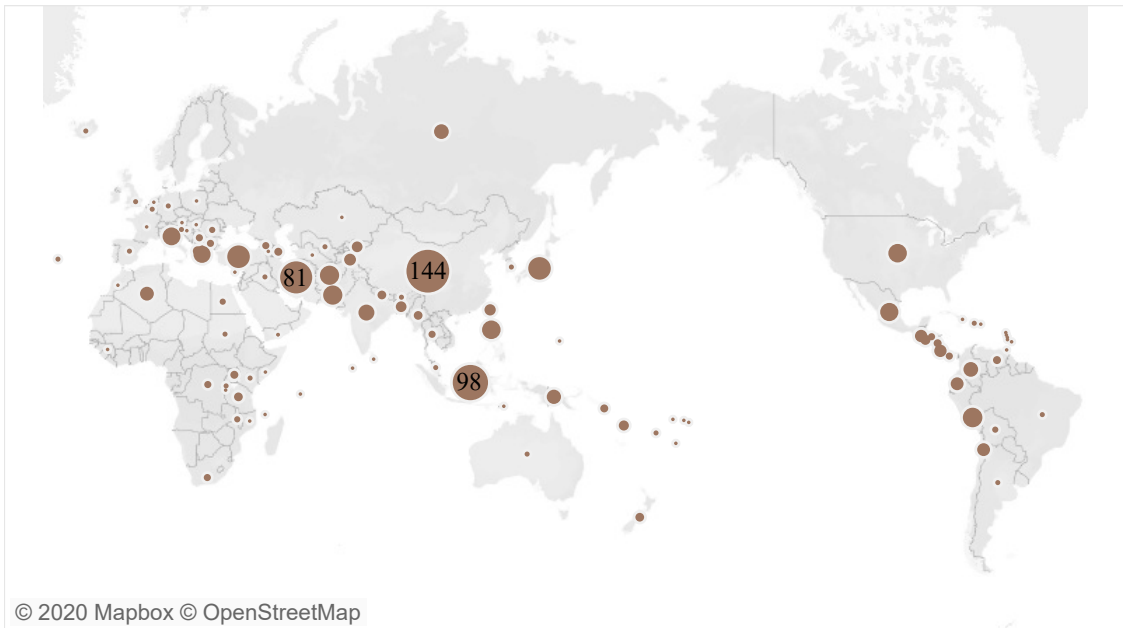
earthquakes US\$803B. The USA is the most affected country with direct economic losses of US\$1,103B. China and Japan follow with US\$555B and US\$511B respectively.

2.1.5 Seismic risk

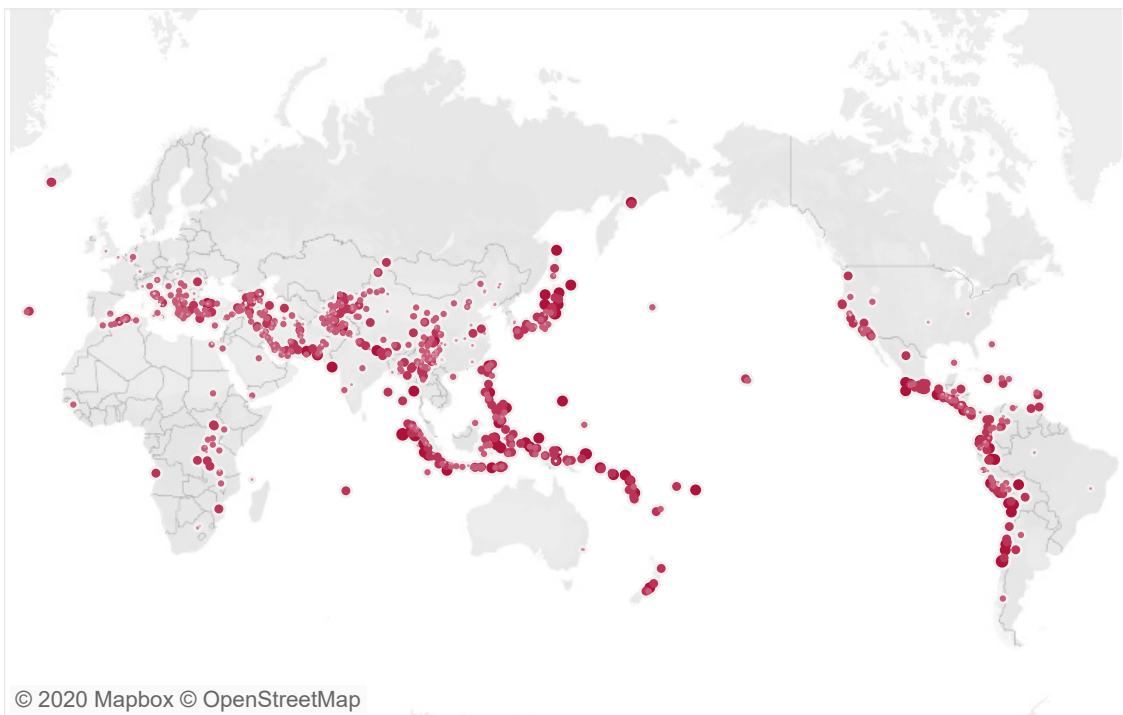
Figure 2.4a shows earthquakes which led to 100 or more people affected, and/or 10 or more deaths and Figure 2.4b shows the location of the corresponding epicentre and Richter magnitude of the earthquake events for the period 1980-2019. Most of the larger seismic events were located along plate boundaries, particularly in the Pacific Ring of Fire (United States Geological Survey (USGS), 1999).

While earthquakes only represented 9% of the geophysical and climate-related events, seismic events led to the largest number of casualties with over 884,700 deaths between 1980-2019. Earthquakes accounted for more than 36% of the total deaths for geophysical and climate-related events. Haiti was the most affected country with more than 200,000 lives lost during the 2010 Haiti earthquake.

Figure 2.5a shows the absolute losses, insured losses, and the insurance contribution for earthquake events between 1980 and 2019. Losses due to seismic events were the largest in 2011 due to the 2011 Tōhoku earthquake in Japan and 2011 Canterbury earthquake in New Zealand. Figure 2.5b presents the total losses and insurance contribution for largest earthquakes. The 2011 Tōhoku earthquake in Japan led to US\$210B absolute losses and was the costliest seismic event, of which US\$37.5B (18%) was paid by insurance. This was the individual event with the largest insurance payout. The 2010 Darfield earthquake and the 2011 Canterbury earthquake in New Zealand stood out with an insurance penetration above 75%. This was significantly higher than the next highest insurance participated earthquake event, the 1994 Northridge earthquake in the USA at 35%, confirming New Zealand's unique seismic insurance setting.

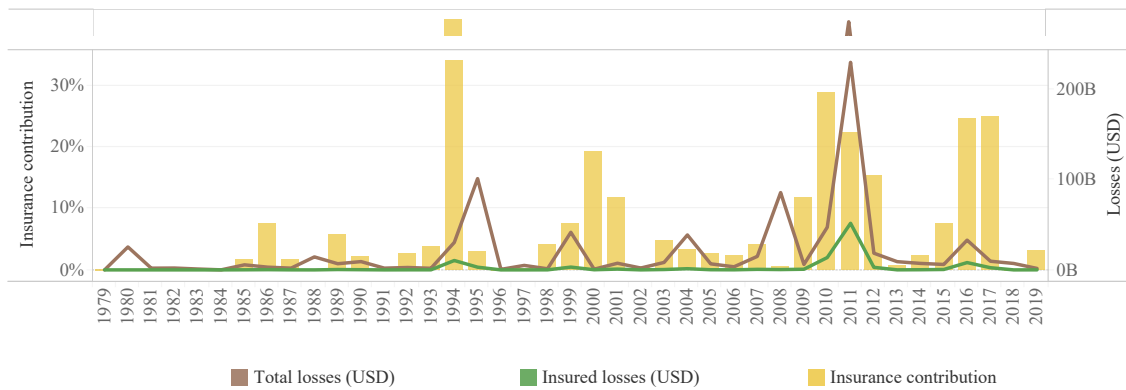


(a) Number of earthquakes by country for the period 1980-2019



(b) Epicentre and magnitude of earthquakes for the period 1980-2019

Figure 2.4: Earthquakes which caused 10 or more deaths, and/or affected 100 or more persons, and/or triggered the declaration of a state of emergency between 1980 and 2019 (Source: EM-DAT, CRED (UCLouvain & Guha-Sapir, 2020))



(a) Total losses, insured losses, and insurance contribution from earthquake events between 1980 and 2019



(b) Total losses and insurance contribution for largest earthquake events between 1980 and 2019

Figure 2.5: Total losses and insurance contribution for earthquake events with absolute losses greater than USD 5B (Source: EM-DAT, CRED (UCLouvain & Guha-Sapir, 2020))

2.2 Post-earthquake damage collection

2.2.1 Review of existing post-earthquake building damage assessment data collection

The first seismic damage assessment methodology for the US was developed by the Applied Technology Council (ATC) in the form of a report “ATC-13 Earthquake Damage Evaluation Data for California” (Rojahn et al., 1985) in 1985. The report presented damage and loss estimates as well as a framework to compute these estimates. ATC subsequently published “ATC-20 Procedures for Postearthquake Safety Evaluation of Buildings” (Applied Technology Council, 1989) and “ATC-20-2 Addendum to the ATC-20 Postearthquake Building Safety Procedures” (Applied Technology Council (ATC), 1995). These reports introduced a standardised approach for the seismic damage evaluation focused exclusively on Californian building stock. These supported the

information need during the response phase of an earthquake event, while developing a data frame that enabled further research connecting earthquake damage and losses. A copy of the ATC-20 Detailed Evaluation Safety Assessment Form is provided for reference in Appendix B.1.

In Europe, the European Seismological Commission developed the European Macroseismic Scale 1998 (EMS-98) in 1998 to collect information and classify seismic damage for masonry and concrete buildings (Grünthal, 1998). In Italy, the “Field Manual for post-earthquake damage and safety assessment and short-term countermeasures (AeDES)” provided tools for the evaluation of seismic damage data post-earthquake (Baggio et al., 2007).

In 1998, the New Zealand Society for Earthquake Engineering (NZSEE) published the guidelines “Post-earthquake building safety procedures”. The assessment forms and building evaluation methodology is identical to that of ATC-20. Nevertheless, the New Zealand document differentiated itself in the placard use. An update of the guidelines followed in 2009 (New Zealand Society for Earthquake Engineering (NZSEE), 2009) which expanded the evaluation process to include two levels of rapid assessment followed by a detailed engineering evaluation if necessary.

Today, the ATC-20 Building Safety Evaluation forms remain the blueprints for post-earthquake building damage data collection. However, it had been designed specifically for American building stock and characteristics. There is thus a need for a single form for capturing building features from different countries, to allow universal data collection anywhere in the world yet detailed enough to capture important region-specific features and local construction practice. In 2009, private and public entities collaboratively founded the Global Earthquake Model (GEM) aimed to develop a single model for evaluating earthquake risk in any location worldwide. One of its workstreams was the development of tools to capture damage data following earthquakes (Foulser-Piggott et al., 2014).

2.2.2 GEM Inventory Data Capture Tools (IDCT)

As part of GEM, scientists and engineers developed the opensource GEM Inventory Data Capture Tools (IDCT) for collecting building exposure data on site. The tool is

available on Windows and for Android devices at the time of writing (Global Earthquake Model (GEM), 2013; Jordan et al., 2014; Rosser et al., 2014). These tools streamlined the data processing process. However, due to the challenges of an emergency situation and the potential limit of devices, power and cellular communication availability, and inflexibility of input for onsite use, paper forms are still developed and used as a backup or as the main tool itself (Jordan et al., 2014). The GEM project developed paper forms are provided in Appendix B.2.

The tool development process included field tests in learning from earthquake exercises in L'Aquila, Italy, in Athens, Greece and Bishek, Kyrgyzstan (Foulser-Piggott et al., 2013). These experiences further improved aspects of the form construction such as colour coding for different categories, presentation order of attributes. Field tests pointed to the need to refine lateral-load resisting system options and data options (Foulser-Piggott et al., 2013).

2.2.3 GEM building Taxonomy v2.0

Building features classification and arrangements are described in building taxonomies. Early recorded use of building taxonomies can be found from the end of the 18th century. Insurance companies needed to accurately define and document the building characteristics to provide adequate fire insurance. The concept of recording building characteristics evolved during the 19th century. Prudent insurance companies recognised the need for accurately cataloguing of building inventory, and classification of buildings and their components in order to apply a technical actuarial approach to underwriting. This categorisation of material, elements and components into several groups is called "building taxonomy".

The use of building taxonomy is standard practice for the calculation of risk related to fire. Nevertheless, it is only in the mid-1900s that earthquake insurance used taxonomy systems for rating purposes (Brzev et al., 2013). The 1970s saw the emergence of Performance-Based Earthquake Engineering (PBEE) procedures (Porter, 2005). The building industry first started to adopt taxonomy classification systems in 1985 with the ATC-13 Earthquake Damage Evaluation Data for California (Rojahn et al., 1985). ATC-13 evolved to form FEMA P-154. This methodology first developed in 1985, is still in use

today in an updated version, the FEMA P-154 Rapid Visual Screening of Buildings for Potential Seismic Hazards: A Handbook (Rojahn et al., 2015).

The beginning of the 21st century saw the development of building taxonomies for use outside the U.S. Examples include The World Housing Encyclopedia (Earthquake Engineering Research Institute (EERI) & International Association for Earthquake Engineering (IAEE), 2000), the PAGER-STR taxonomy (Jaiswal & Wald, 2008), and the GEM Building Taxonomy v2.0 which is designed to be applicable worldwide (Brzev et al., 2013). The GEM taxonomy describes and uniformly classifies buildings according to thirteen attributes. These are presented in detail in Appendix C.1. Additional information on the GEM Building Taxonomy Version 2.0 is available in the technical report by Brzev et al. (2013).

2.2.4 Non-structural components

In PBEE, non-structural components significantly influence the damage and loss analyses (Porter, 2005). Taghavi and Miranda (2003) showed that non-structural components commonly make up 60% to 80% of the total value of a building. It is thus important to adequately identify and categorise these non-structural components with sufficient details so that components with different damageability are addressed to different categories (Porter, 2005).

2.3 The principle of “vital few and the useful many”

The principle of “vital few and the useful many” is also commonly referred to as the Pareto Principle, the 80/20 rule or the Juran Principle. In the late 1800’s, economist Vilfredo Pareto found that wealth is unequally distributed among the population. He observed that 80% of the land in Italy was in fact only owned by 20% of the individuals (Pareto, 1906). In the early 1950s, Joseph M. Juran, extended the application of the Pareto Principle to quality control (Juran, 1951). Juran affirmed that the Pareto principle applies to several situations in industry and in daily life. Juran demonstrated that when several elements contribute to a common effect, often a small part of them contribute significantly to the outcome. Juran used the expression the “vital few and the

useful many” defining the critical elements as the “vital few” and the components that contribute to a lesser extent as the “useful many”.

Juran applied this rule for quality control in industrial process in business. He emphasized the need to identify the vital few, to streamline and improve the entire production process as these critical, key parameters contribute the most to the total effects. A helpful property of the rule is that the Pareto principle applies to several levels of detail. In the fifth edition of his *Quality Control Handbook* (Juran & Godfrey, 1999), Juran presented the example of a paper mill. His objective was to find the key parameters to minimise wastage due to quality issues. The first broad level of analysis identified the key headline problem areas. Once the vital few were identified, the emphasis was then put on the optimization of processes within each individual area causing the highest percentage of total quality loss. In Juran’s case study, the broke (defect paper that should be reprocessed) was the critical accounting category. Once aware of this fact, the 80/20 rule was applied again examining the wastage broke category, which was now subdivided into product types. Here again, the vital few leading to the most of the annual broke loss were identified.

Figure 2.6 shows a graphical representation of a sample Pareto analysis. It lists the number of customer queries received and their corresponding query category. The Pareto diagram highlights that focusing efforts to address customer queries categories A, B and C will have the most significant contribution addressing more than 70% of the customers’ queries. Independently on the representation solution chosen, the methodology allows the user to portray the vital few ranked by the magnitude of their contribution.

2.4 Seismic damage and loss assessment

2.4.1 PEER PBEE framework

A milestone in the seismic performance and loss assessment is the development of the PBEE methodology developed by the Pacific Earthquake Engineering Research Center (PEER) in the late 1990s (Cornell & Krawinkler, 2000; Poland et al., 1995). In recent years, the PEER methodology has been extensively described (Broccardo et al., 2016; Günay & Mosalam, 2013; Yang et al., 2009).

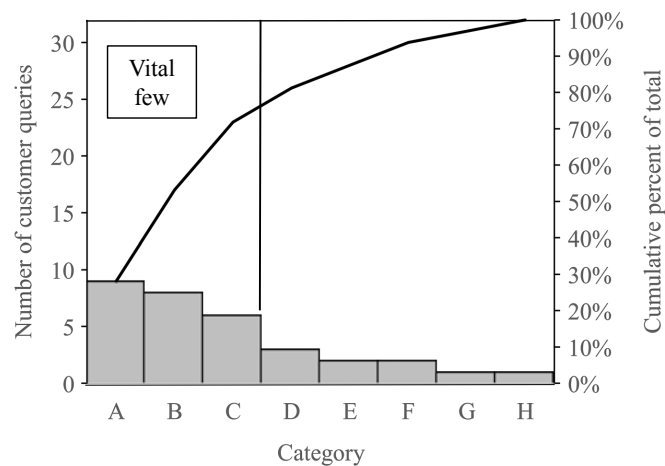


Figure 2.6: Pareto diagram of the number of customer queries, adapted from (Juran & De Feo, 2010)

The PEER PBEE methodology is an evolution of the “Vision 2000” framework which emerged in the aftermath of the 1994 Northridge earthquake in an effort to improve seismic codes. The framework developed new criteria based on field observations and quantitative evaluations. It established concepts of key engineering response parameters, defined acceptance limits for building performance objectives in various level of ground shaking (Poland et al., 1995). The “Vision 2000” methodology set out a relationship between the performance objective, the risk profile of a facility, the probability of an earthquake, and the response parameters related to each performance objective. The framework introduced new definitions for the building performance: fully operational, operational, life safe, and near collapse (Poland et al., 1995). Work such as FEMA 273, ATC-32, ATC-40, and FEMA 356 followed (Applied Technology Council (ATC), 1996a, 1996b; Federal Emergency Management Agency (FEMA), 1997, 2000).

Through these documents, PEER introduced a probabilistic framework to account for uncertainties which is now commonly referred to as the PEER PBEE methodology. The PEER PBEE methodology eased decision-making by providing a consistent and clear framework regarding the seismic performance of assets given location and design (Cornell & Krawinkler, 2000). The PBEE probabilistic framework allowed accounting for variability and inherent uncertainties in earthquake performance assessment (Moehle & Deierlein, 2004). The decision variable (DV) is expressed through performance metrics such as the casualties, the repair costs, and loss-of-use duration (3 D’s: death, dollars, and

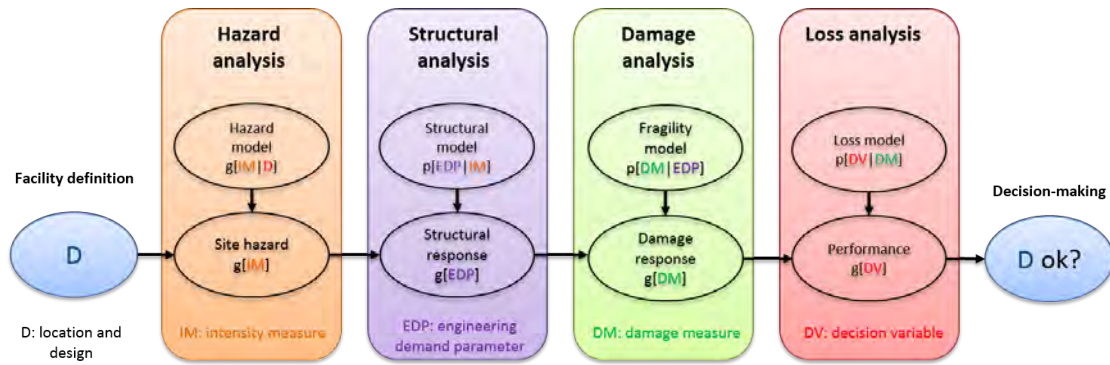


Figure 2.7: Overview of the four steps of PEER PBEE analysis methodology, adapted from (Porter, 2003)

downtime), which are meaningful to technical and non-technical stakeholders (Porter, 2003).

The first version of PBEE expressed DV as a function of the seismic hazard, in terms of an intensity measure (IM), and damage, characterised as damage measure (DM) (Cornell & Krawinkler, 2000). Porter (2003) improved the PEER methodology and introduced a fourth stage.

Figure 2.7 shows a graphical overview of the four-stage approach which combines the results of hazard analysis, structural analysis, damage analysis and loss analysis. The mathematical expression of the mean annual rate of a decision variable is shown in equation 1.1.

Equation 1.1 expresses the mean annual rate of the outcome or decision variable (DV) as an integral of a chain of conditional probabilities depending on

- the damage measure (DM), a measure of physical damage associated with a given engineering demand parameter (e.g. local failure, degree of collapse, loss of load capacity),
- the engineering demand parameter (EDP), a measure used to characterise structural response (e.g. floor acceleration, interstory drift, roof displacement), and
- the intensity measure (IM) of the earthquake, a measure used to characterise the intensity of ground shaking (e.g. ground acceleration, ground velocity, spectral displacement).

The United Nations Office for Disaster Risk Reduction (UNISDR) defines disaster risk¹ as 'the potential loss of life, injury, or destroyed or damaged assets which could occur to a system, society or a community in a specific period of time, determined probabilistically as a function of hazard, exposure, vulnerability and capacity.' Seismic hazard, exposure, and physical vulnerability comprise the components of a typical physical seismic risk model, as illustrated in Figure 2.



Figure 2. Schematic diagram illustrating the different components of a typical integrated seismic risk model

The UNISDR definitions of the terms hazard, exposure, and vulnerability are listed below, along with brief descriptions of these terms as they are used in the report.

The derivation of the conditional probabilities, $G(DM | EDP)$ and $G(DV | DM)$, require a damage analysis and loss analysis based on fragility and loss models respectively (Yang, 2013). The concept of the PEER PBEE methodology² is thoroughly explained in the seminal work of Porter (2003). In recent years, the earthquake and structural engineering community extensively studied and applied the PEER PBEE methodology.

The interested reader is directed to Kiureghian (2005), Krawinkler (2005), Mirani-Reiser (2007), Dhakal (2011), Gray and Mosman (2012), Cutfield (2015), Burton et al. (2016).

¹ 2017 UNISDR terminology on disaster risk reduction
 Disaster risk: <http://preventionweb.net/go/7818>
² Hazard: <http://preventionweb.net/go/488>
³ Hazardous event: <http://preventionweb.net/go/51759>
⁴ 2017 UNISDR terminology on disaster risk reduction.
 Exposure: <http://preventionweb.net/go/7822>

2.4.2 Current practice in seismic damage and loss modelling

2

Loss models for seismic risk rely on three main components (see Figure 2.8): a seismic hazard component often obtained using a probabilistic seismic hazard analysis (PSHA) for ground shaking levels for a region, an exposure component encapsulating information related to the building stock, and a damageability component (Silva et al., 2019). Damageability manifests as fragility or loss. Fragility is the probability of an undesirable outcome conditional to an environmental excitation (Porter, 2020). Loss is often broken down as the 3Ds (dollars, deaths, and downtime) (Dhakal, 2011).

To estimate the probability of damage and loss given an intensity measure, loss modelers employ fragility and vulnerability functions. Fragility and vulnerability are related but are not to be mixed. Fragility express the probability between a measure of the environmental excitation and the undesirable outcome. Vulnerability captures losses (Porter, 2020). Vulnerability functions express the relationship between a intensity measure and the repair or replacement cost (Silva, 2019). Vulnerability functions are

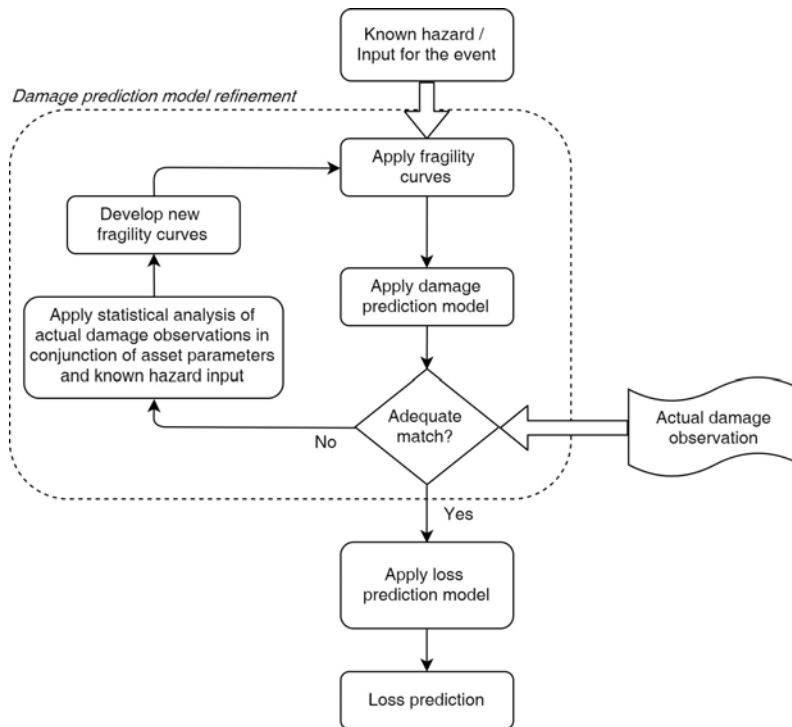


Figure 2.9: Damage assessment in the earthquake risk reduction

often defined for a specific asset class (Porter, 2020). Fragility functions, however, can present the fragility of an entire structure (Ellingwood et al., 2007), or for individual components such as structural components (Aslani & Miranda, 2005; Brown & Lowes, 2007), non-structural components (Badillo-Almaraz et al., 2007; Porter et al., 2007), and contents (Hutchinson & Ray Chaudhuri, 2006; Porter & Kiremidjian, 2001).

Figure 2.9 presents a flowchart showing the connections between damage data collection to the implementation of the loss prediction model. The damage state of a building following an earthquake is the key validation in the damage prediction process.

Fragility functions can be developed according to multiple procedures based on the data type available (Porter et al., 2007). They can be developed 1) empirically through laboratory experiments and/or real world observations, 2) analytically via simulation, 3) expert solicitation, or 4) a combination of the above (Porter, 2020). Each approach offers its own advantages; however, empirical data is often regarded as the most desirable and reliable as it comes from real buildings subjected to real earthquakes and thus reflects actual performance. Post-earthquake seismic damage assessment play a crucial role in the understanding of our buildings deficiencies and strengths. A better understanding of building failures can lead to design code improvements.

- Damage limitation states (DLS), associated with damage beyond which specified service requirements are no longer met.

The fragility analysis phase includes the choice of the fitting and sampling methods, the selection of models to express the fragility curves and the construction process itself, taking into account the uncertainties considered and measured in the previous structural and damage analysis phases (FEMA 2003a; Wen *et al.* 2004; Pagnini *et al.* 2011; ATC 2011).

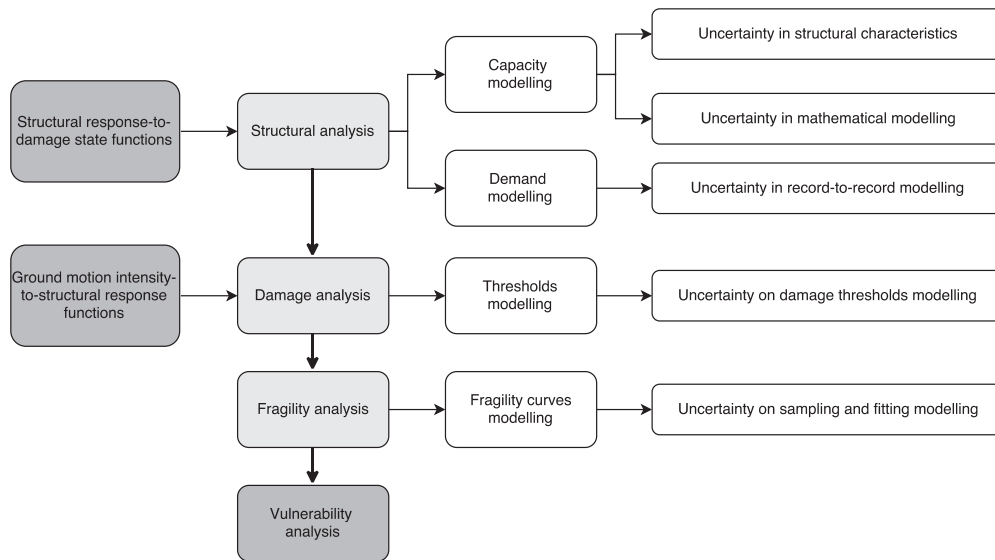


Figure 3. Main components and phases considered in analytical fragility assessment methodologies and associated uncertainties. Figure 2.10. Uncertainties in the analytical fragility assessment methodology (Maio and Tsionis, 2015)

2.4.3 Current limitations

Real-world damage observations are scarce but are extremely valuable as they capture the aleatoric variability and epistemic uncertainties that are difficult to simulate. The level of detail captured by assessment forms derived for particular purposes, often does not match the required level of detail necessary to find the main building parameters causing damage in buildings. There is an¹³ evident lack of damage data concerning nonstructural elements, which could account for up to 80% of a total building value (Taghavi & Miranda, 2003).

Current fragility models are most often developed using a probabilistic approach (Villar-Vega & Silva, 2017). So, whilst it is capable of accounting for uncertainties, it also highlights that the uncertainties can be large depending on understanding of the system (Bradley, 2010; Lallemand *et al.*, 2015). Figure 2.10 presents an overview of possible uncertainties related to analytical fragility assessment. Fragility and vulnerability functions are affected by three principal causes of variability: record-to-record variability, building-to-building variability, and uncertainty in the damage criterion (Silva, 2019).

The application of seismic risk assessment is also limited by the tools and data available. Seismic loss models are often region-specific and are restricted to developed countries (Silva *et al.*, 2019). Damage and loss data are often scarce and thus statistically

insufficient to update current damage models (Silva, 2019). Even when sufficient data is available, considerable efforts in analysing and interpreting data are usually required to update current damage models. In the case of the 2010-2011 Canterbury earthquake sequence, the high level of insurance involvement complicated the process of data collection. It is difficult to know what information was captured by various private insurance companies. When damage information or claims data are available, non-disclosure and confidentiality agreements often limit the possibilities to use the data openly.

2.5 The 2010-2011 Canterbury earthquake sequence

2.5.1 Generalities

In 2010-2011, New Zealand suffered the costliest natural disaster of its history with a series of earthquakes known as the Canterbury earthquake sequence (CES). Figure 2.11 shows an overview of the location of the Canterbury region within New Zealand as well as the location of Christchurch's CBD. The CES began on 4 September 2010 with the M_w 7.1 Darfield earthquake. The Darfield earthquake was centered approximately 40 km west of Christchurch Central Business District (CBD) (GeoNet, 2010), as shown on Figure 2.12. Christchurch was the second largest city by population in New Zealand. The Darfield earthquake collapsed many unreinforced masonry buildings in the CBD, affected residential houses in wider Christchurch, induced liquefaction in eastern suburbs, and led to one fatality due to sudden cardiac arrest. In the next 15 months, the Canterbury region experienced over 3,500 aftershocks with a magnitude above M_w 3 (see Figure 2.13) and around 60 earthquakes above M_w 5 (Christophersen et al., 2013). On 22 February 2011 12:51pm local time, a M_w 6.2 shallow aftershock occurred directly under Christchurch CBD at a depth of 5 km (GeoNet, 2011). This was the most significant event in the CES.

The event commonly referred to as the 2011 Christchurch earthquake occurred near lunch time when office and street pedestrian occupancies were at their peaks. It collapsed unreinforced masonry buildings that were not already removed from earlier aftershocks, irrecoverably damaged many mid-rise and high-rise buildings, and collapsed two

concrete buildings that led to 135 of the total 182 fatalities in the event (Kam et al., 2011). It also prompted liquefaction in Christchurch CBD and eastern residential areas which exacerbated building damage due to foundation displacement. Following this, there were a number of other aftershocks that led to further building damage.

The CES led to extensive building damage across the region, with over NZ\$50 billion of economic losses, the equivalent of 20% of New Zealand's GDP (Bevere & Balz, 2012; Munich RE, 2019). The CES also highlighted a number of civil and earthquake engineering challenges, such as building on liquefiable land, short-term heightened seismicity, rock slope stability, all of which impacted the reconstruction and recovery (Elwood et al., 2014). It has been estimated that 70% of Christchurch CBD was demolished or partly reconstructed. Significant parts of the CBD were cordoned off from public access for over 2 years from February 2011 until June 2013 (Kim et al., 2017). The CES, being the fourth most costliest insured global natural disaster in history at the time (Insurance Council of New Zealand (ICNZ), 2019), also extensively affected the local and global insurance sector regarding seismic building damage (King et al., 2014).

2.5.2 Seismic insurance following the Canterbury earthquake sequence

Many countries located near tectonic plate boundaries are exposed to frequent earthquakes. However, insurance uptake for geophysical events remains low (2% in Italy, 5% in Turkey, 9% to 11% in Japan, 10% in Mexico, 26% in Chile, 38% in US, and 80% in New Zealand (Bevere & Balz, 2012)). New Zealand is an exception with an insurance penetration of 80% (Bevere & Balz, 2012; King et al., 2014). Over the two years of the CES, major earthquake events and multiple aftershocks led to more than 650,000 insurance claims have been lodged (Insurance Council of New Zealand (ICNZ), 2019). 59% were residential claims and 41% were for commercial claims (Deloitte Access Economics, 2015). Most of the claims for residential buildings were lodged for the main events of the 4 September 2010 and 22 February 2011. However, it was difficult to assess the exact impact of each earthquake and aftershocks on buildings, as the time between the event was too short to permit detailed building assessments following each event, especially for such a large number of affected buildings. This led to significant legal challenges between claimants, insurers and reinsurers about the damage apportionment between

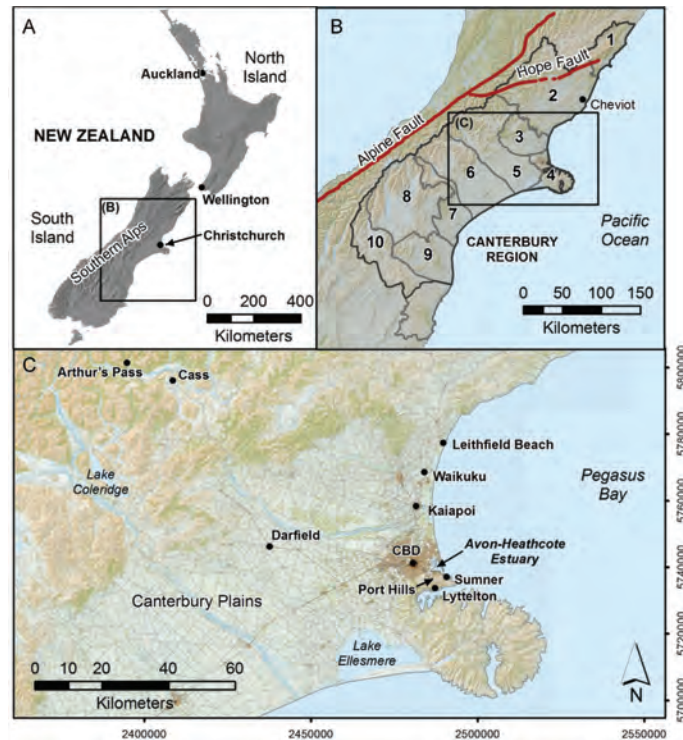


Figure 2.11: Maps of the Canterbury Region (A) New Zealand map with main cities labelled. (B) Canterbury Region, with districts labelled as 1) Kaikoura; 2) Hurunui; 3) Waimakariri; 4) Christchurch City; 5) Selwyn; 6) Ashburton; 7) Timaru; 8) Mackenzie; 9) Waimate; 10) Waitaki. (C) Map of the Christchurch city area and nearby towns (Potter et al., 2015).

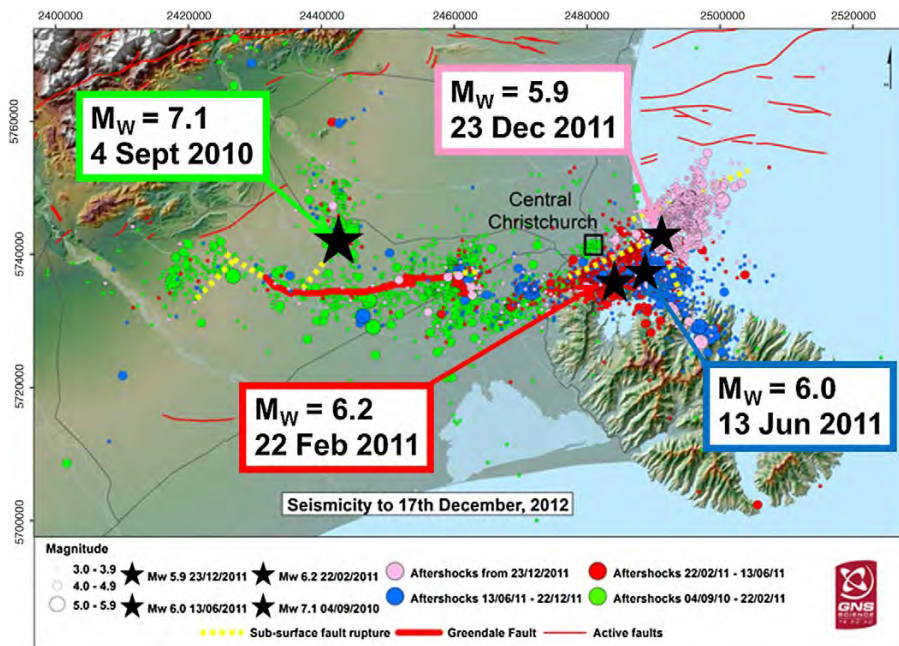


Figure 2.12: Location of the main events in the 2010-2011 Canterbury earthquake sequence (O'Rourke et al., 2014).

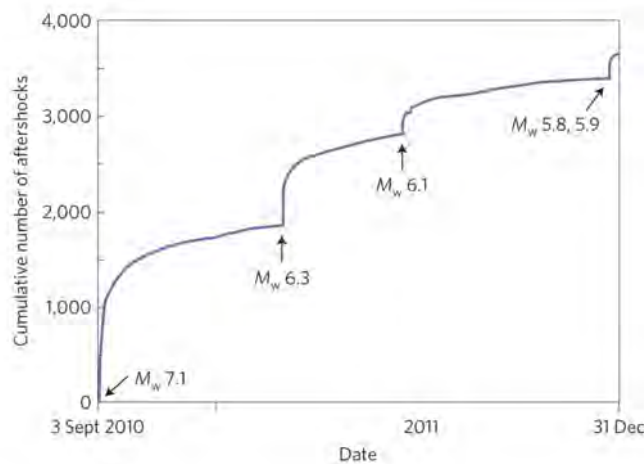


Figure 2.13: Cumulative number of aftershocks (with magnitude $M_w \geq 3.0$) in the CES, adapted from (Reyners et al., 2014)

events. Reports shows that 61% of the residential insurance claims were settled by the Earthquake Commission (EQC) and 39% by private insurers (Deloitte Access Economics, 2015). This distribution points the significant participation of EQC.

2.5.3 The Earthquake Commission (EQC)

The Earthquake Commission (EQC) is a Crown entity which has for its mission to provide natural disaster insurance for residential property. EQC also manages the Natural Disaster Fund (NDF) and promotes research and education on solutions for reducing the impact of natural disasters. EQC involvement is particularly visible with the EQC insurance, EQCover (Earthquake Commission (EQC), 2019b). EQCover provides home and land insurance for natural disaster for every home that is covered by private fire insurance. At the time of the CES, EQC provided coverage for the first NZ\$100,000 + 15% Goods and Service Tax (GST) of the building damage, NZ\$20,000 + GST for contents and land damage up to the value of the damaged land (since 1 July 2019 the cap for residential building cover was increased to NZ\$150,000 but no longer include coverage for contents). EQC accessed the NDF and its reinsurance cover to settle claims. Before the CES, the NDF had a value of NZ\$6.1 billion (more than US\$4 billion) though this has now been significantly depleted to less than NZ\$180 million following the CES and a smaller Kaikoura earthquake in 2016 (Earthquake Commission (EQC), 2019e; Feltham, 2011).

The CES brought major changes for New Zealand, especially for the insurance industry (Greater Christchurch Group - Department of the Prime Minister and Cabinet, 2017). EQC increased the annual levy in order to replenish the NDF (Earthquake Commission (EQC), 2017). Owing to the largely unexpected losses for the private insurers since the CES, there had been a trend of increased scrutiny of the risk profile of any insurance cover. Private insurers now applies risk-based premium pricing for earthquake covers. This had led to increased premiums and at times unavailability of earthquake insurance for some regions in New Zealand.

2.5.4 EQC's catastrophe loss models

Loss models are important for the insurance and reinsurance sector for quantifying probable losses to ensure adequate provisions in case of a catastrophe. EQC similarly relies on hazard and loss models for adjusting base cover, investment and reinsurance strategies and general planning for response to natural catastrophe (Middleton, 2002).

In early attempts to quantify the risk for New Zealand, EQC actuaries estimated possible annual claims from historical data, and probable earthquake intensities. With the evolution of individual computers in the 1980s, new modelling opportunities arose. EQC first employed a computer-based modelling software for loss simulation in 1993. In the past, EQC relied on two models that work in tandem: a system dynamics model (SDM) called 'Logjam' for the management of the claims and a hazard and financial risk management system called 'Minerva' (Middleton, 2002). EQC employed Minerva for estimating claims numbers and losses following a major disaster, as well as for the predicting earthquake loss risk over 10 years in the future to design EQC levy structures and deductibles and to maintain the reserves in the NDF. Minerva relied on an internal database as well as external sources such as the EQC Building Costs or Aon Soils database (2.14a). An earthquake loss subsystem, which entails an attenuation and a vulnerability model combined, was used to simulate the losses for any earthquake event (2.14b). Additionally, it has source models for New Zealand as well as 10-year portfolio models for predicting loss frequency. Outputs from these possible scenarios are stored in the Minerva database which can then be accessed by the financial management sub-system (Shephard et al., 2002). Nowadays, EQC works closely with reinsurance

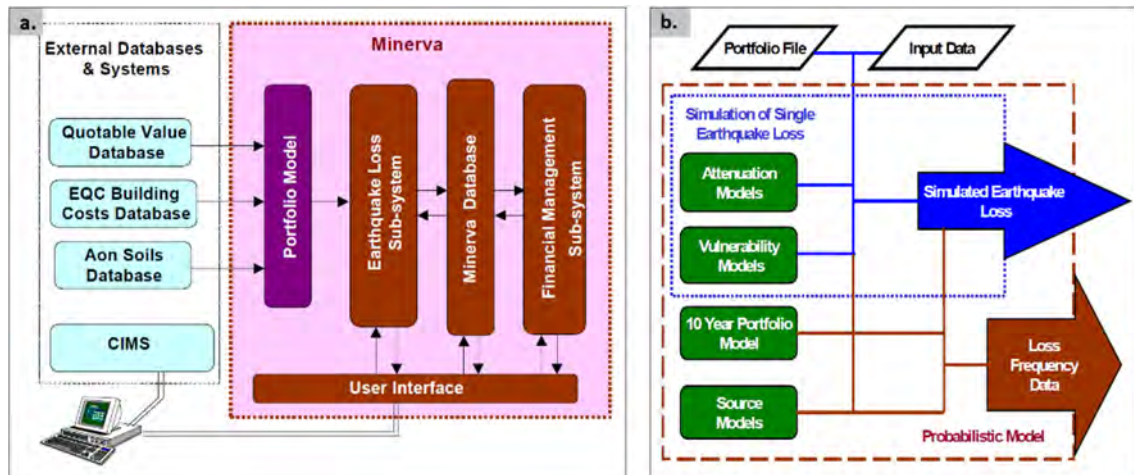


Figure 2.14: (a) Overall Minerva system architecture, (b) Schematic diagram of the Earthquake Loss sub system used in Minerva (Shephard et al., 2002)

companies to ensure that New Zealand retains the necessary international support in case of a disaster (Earthquake Commission (EQC), 2019d). EQC still uses Minerva as an impact estimation tool to predict likely losses for single events and one-year probabilistic analyses.

Without minimizing the great improvement that these tools offered to the New Zealand insurance sector, limitations are still present. Since EQC offers natural disaster insurance for residential building on top of existing private insurance, EQC does not retain a database of its policy holders. It thus uses New Zealand records of real estate property as a base of its calculation (Middleton, 2002). This led to limitations regarding the accuracy of the exact loss prediction per asset. Moreover, the CES highlighted that the existing loss models did not accurately capture liquefaction. Additionally, the models usually took the building stock as undamaged at the time of the earthquake. But in the CES, the time between the events was too short such that the structures could not have been repaired or rebuilt. Cumulative damage occurred in reality but was not taken into account by the loss models (Drayton & Verdon, 2013).

2.5.5 Earthquake Commission Amendment Act

The Earthquake Commission Amendment Act 2019 (2019/1) received royal assent on 18 February 2019 (New Zealand Parliament, 2019). The Earthquake Commission Amendment Act 2019 introduced changes including an increase in the time limit to

lodge a claim following an earthquake event from three months to two years, the removal of the insurance cover for content, but an increase in maximum building cover from NZ\$100,000 to NZ\$150,000+GST. At the same time, the Act brought revisions to the information sharing provision. EQC is now allowed to share information about the residential property claims, which have been lodged with EQC. Homeowners and prospective buyers can now ask EQC to provide them with information on residential property damage due to a natural disaster (Earthquake Commission (EQC), 2019c). The Act also enables EQC to share information for public good purposes which greatly assisted this study. Before March 2019, building data in EQC's property database were rounded to approximately 70 m to protect privacy. This made it impossible to merge EQC's claim information with additional databases. The change in legislation permitted the data to be used to its full potential and enabled new opportunities for this research. The more accurate building location in the data enabled spatial joining and merging with new information on liquefaction, soil conditions, and building characteristics.

2.6 A primer on machine learning

2.6.1 Machine learning, data science, and artificial intelligence

To build a model from the observation of data and make prediction through experience, machine learning employs computers and algorithms. Thus it is often illustrated as a field of study at the intersection of computer science and mathematics (see Figure 2.15).

Machine learning can also be regarded as a subpart of the greater data science discipline. Data science employs concepts from multiple fields such as Bayesian methods, computational complexity theory, control theory, information theory, and even neurobiology and philosophy (e.g. Occam's razor principle) (Mitchell, 1997).

Machine learning can also be seen as a subset of artificial intelligence (AI) (Goodfellow et al., 2016). Tasks such as computer vision, robotic, facial recognition, or automated driving are commonly associated or achieved through AI. However, a single definition of AI does not exist. Depending on the approach followed, historical definitions of AI can be grouped into four categories: thinking humanly, thinking rationally, acting humanly, and acting rationally. The first two groups relate to the

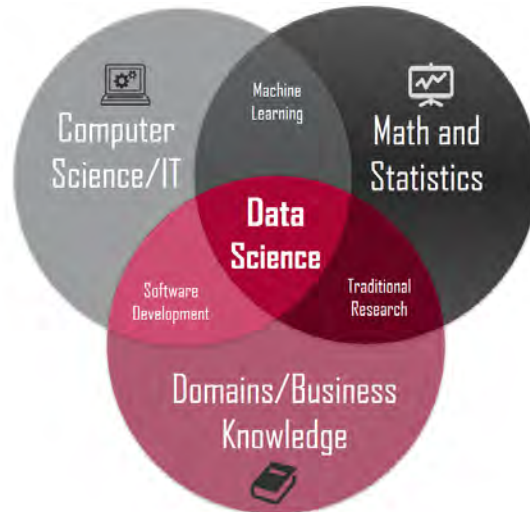


Figure 2.15: Core concepts constituting the field of data science (Barber, 2014)

thought process and reasoning, while the two other connects to behaviour. Another differentiation is the performance measure. Some compare the system performance against human performance, while others assess performance against rationality (S. Russell & Norvig, 2020).

Figure 2.16 presents a schematic overview of the overlap between machine learning and data science and also shows the position of machine learning within the AI field. The following review focuses on machine learning, data science and its practical applications. The reader interested in general technologies and concepts related to AI is directed to a textbook by Russell and Norvig (2020).

2.6.2 Machine learning compared to rule-based systems

Machine learning differentiates itself from rule-based programming as it is not procedurally predefined by human, instead, the computer analyses features of the input data and builds a representative model developed from the data. The model can then be used to extrapolate patterns and predict new possible outcomes (S. Russell & Norvig, 2020). Figure 2.17 shows a schematic overview of rule-based system (sometimes referred as “traditional approach”), “classic” machine learning where the feature engineering is done by a human, and representation learning where the system automatically selects the features.

The term machine learning first appeared in research literature in 1959 (Samuel, 1959).

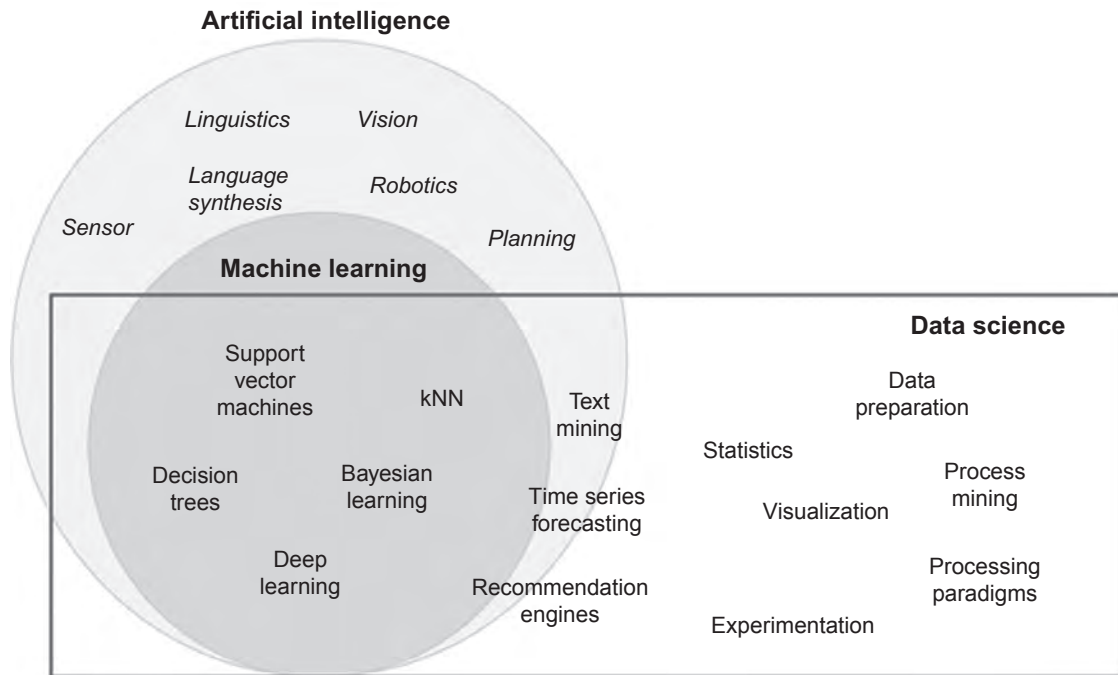


FIGURE 1.1 Figure 2.16: Venn diagram of artificial intelligence, machine learning, and data science (Kotlu & Deshpande, 2019)

Using the game of checkers, Samuel (1959) introduced the notion that computers can be programmed in a way to learn from experience. Mitchell (1997) refined the definition stating that “a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured

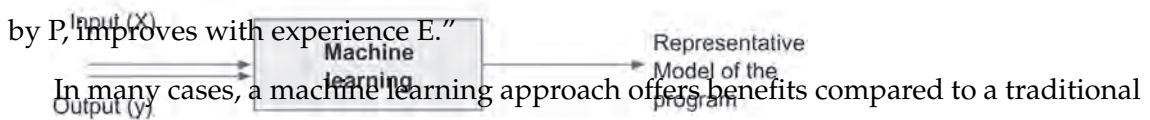


FIGURE 1.2 Traditional program and machine learning. In many cases, a machine learning approach offers benefits compared to a traditional

based system. Examples are models which require constant updating over time (e.g. spam filter), complex problems which cannot be easily implemented using rule-based systems, and problems for which the best performing algorithm is unknown (e.g. speech from experience). Experience for machines comes in the form of data. Data that is used for training (Géron, 2019). Although the benefit of machine learning approach is the ability to indiscriminately reveal correlation to discover insights through ML models using predetermined rules and relationships. Machine learning algorithms for a system (see Figure 2.18). Machine learning algorithms can process a significant

amount of input data, “study” the data during the model training and develop a solution. As humans, it is possible to observe the solution to get a better understanding of the problem, and depending on the algorithm chosen, it is even possible to inspect the relationships and quantify the influence between the model features.

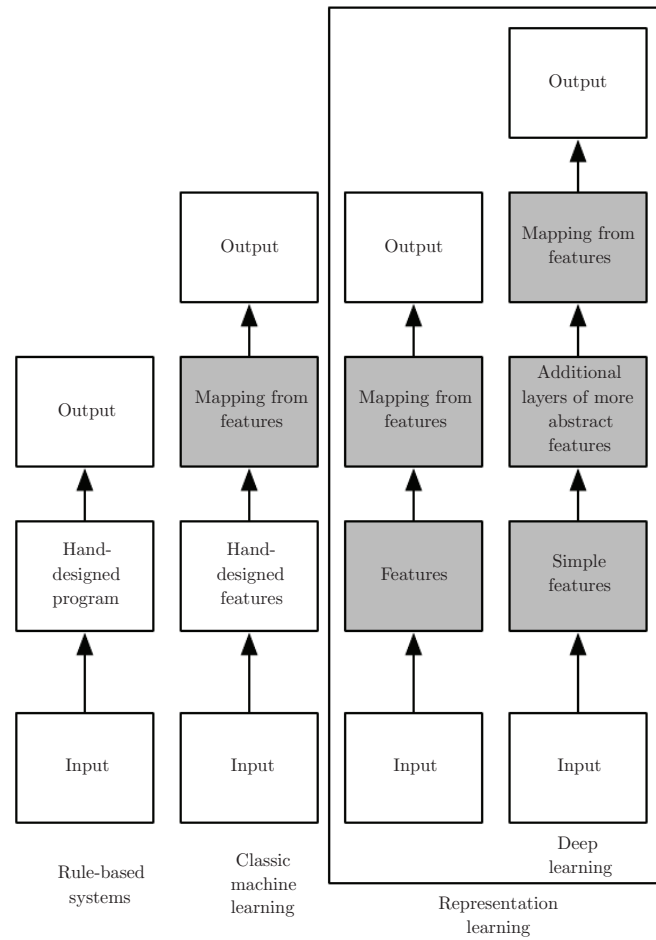


Figure 2.17: Schematic overview of rule-based systems, machine learning and deep learning systems. Grayed box highlights elements that can learn from data (Goodfellow et al., 2016)

correlations or new trends, and thereby lead to a better understanding of the problem. Readers should feel free to skip parts that are not relevant given their interests or background. Applying ML techniques to dig into large amounts of data can help discover patterns that were not immediately apparent. This is called *data mining*. machine learning concepts can skip part I, for example, while those who just want

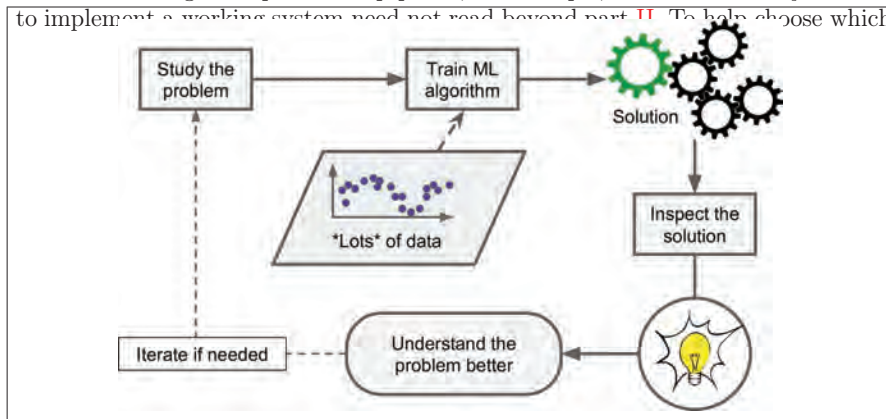


Figure 2.18: Machine Learning can help humans understand a problem better (Géron, 2019)

To summarize, Machine Learning is great for:

- Problems for which existing solutions require a lot of fine-tuning or long lists of rules: one Machine Learning algorithm can often simplify code and perform better than the traditional approach.
- Complex problems for which using a traditional approach yields no good solution: the best Machine Learning techniques can perhaps find a solution.
- Fluctuating environments: a Machine Learning system can adapt to new data.
- Getting insights about complex problems and large amounts of data.

2.6.3 Types of learning

Machine learning systems can be categorised according to the type of feedback that a ML system receives during the learning process. Three main types of learning are distinguished: supervised learning, unsupervised learning, and reinforcement learning. Semi-supervised learning, which combines supervised and unsupervised learning, is sometimes seen as a fourth category.

- In supervised learning, the ML system is provided with input-output pairs and is asked to find a function that best captures the input-output relationship. Depending on the type of the output, the learning problem is named “classification” if the objective is to predict categories, or “regression” if the target output is a number.
- In unsupervised learning, the ML system is provided with input data without specific outputs. It is asked to learn by itself. Common examples include clustering (ML should find groups sharing similar properties), anomaly detection (ML should highlight instances different from the standard ones), association rule learning (ML should find relationships between the features), and dimensionality reduction (ML should combine multiple correlated features to simplify the data with limited information loss).
- In reinforcement learning, the ML system learns over iteration, performing actions in an environment and getting rewards or punishments (negative rewards) depending on the action.

2.6.4 Examples of machine learning application

Nowadays, machine learning is applied in many fields for diverse applications. Examples include web related activities (e.g. engine search, spam filter, online shopping), assistance (e.g. personal virtual assistants, chatbots), gaming (e.g. bot for a game), payment and banking (e.g. credit card fraud detection, voice verification), real estate (e.g. house price prediction), business (e.g. forecasting a company’s revenue), transportation (e.g. route optimisation, predictive fleet maintenance), medicine (e.g. detection of

malignant tumours on CT scans). Machine learning applications are often grouped by the types of task performed. Table 2.1 lists some machine learning tasks and provides a description and examples for each category.

The application of machine learning in civil engineering is increasing in popularity. Below is a few notable relevant ML studies,

- the evaluation of post-earthquake structural safety (Zhang et al., 2018),
- the rapid loss assessment (Kovačević et al., 2018),
- the derivation of fragility curves (Kiani et al., 2019),
- the quality classification of ground motion records (Bellagamba et al., 2019),
- the classification of earthquake damage to buildings (Mangalathu & Burton, 2019; Mangalathu et al., 2020) and bridges (Mangalathu et al., 2019; Mangalathu & Jeon, 2019).

Xie et al. (2020) present an extensive review of the application of machine learning in earthquake engineering. Sun et al. (2020) give a review of machine learning applications for building structural design and performance assessment.

2.6.5 General framework of a machine learning model

Before starting a machine learning project, it is essential to carefully define the problem and objective. It should be noted that the machine learning model objectives can be different from the overall business goal (Burkov, 2020). If the machine learning is to be applied in business or industry, it is necessary for the data science team to know the specific purpose that the machine learning model should achieve but also to get an understanding of the business problem.

Figure 2.19 shows the main steps of a machine learning project. As soon as the objective is clearly defined, the data engineer starts to collect and prepare relevant data. The collected data is transferred to the data scientist (referred to as data analyst in Figure 2.19) for data pre-processing, model training and model evaluation. The time for the data preparation and data cleaning should not be underestimated. Experience showed that usually, 80% of the time and efforts of a data scientist lies

Table 2.1: Data science tasks (non exhaustive) and examples, adapted from (Kotu & Deshpande, 2019)

Type of ML learning	Tasks	Description	Examples
Supervised Learning	Classification	Predict if a data point belongs to one of the predefined classes. The prediction will be based on learning from a known data set.	<ul style="list-style-type: none"> • Assigning voters into known buckets by political parties • Bucketing new customers into known customer groups
Supervised Learning	Regression	Predict the numeric target label of a data point. The prediction will be based on learning from a known data set.	<ul style="list-style-type: none"> • Predicting the unemployment rate for the next year • Estimating insurance premium
Supervised Learning	Recommender system	Predict the preference of an item for a user.	<ul style="list-style-type: none"> • Finding the top recommended movies for a user
Supervised Learning	Time series forecasting	Predict the value of the target variable for a future timeframe based on historical values.	<ul style="list-style-type: none"> • Sales forecasting • Production forecasting • Any growth phenomenon that needs to be extrapolated
Unsupervised Learning	Clustering	Identify natural clusters within the data set based on inherent properties within the data set.	<ul style="list-style-type: none"> • Finding customer segments in a company based on transaction, web, and customer call data
Unsupervised Learning	Anomaly detection	Predict if a data point is an outlier compared to other data points in the data set.	<ul style="list-style-type: none"> • Detecting fraudulent credit card transactions • Detecting network intrusion
Unsupervised Learning	Association analysis	Identify relationships within an item set based on transaction data.	<ul style="list-style-type: none"> • Finding cross-selling opportunities for a retailer based on transaction purchase history

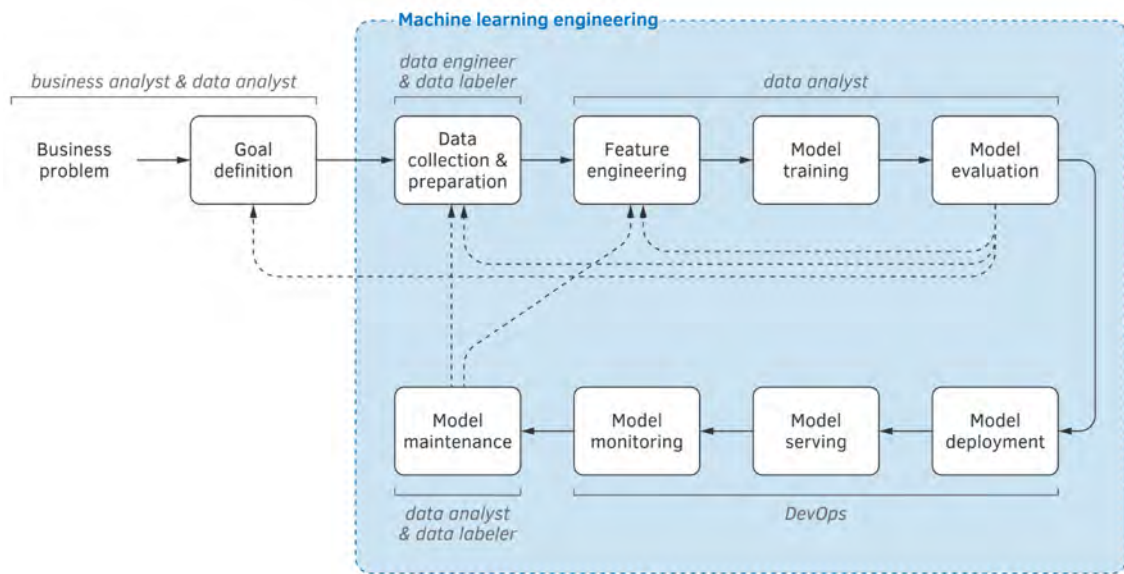


Figure 2.19: Machine learning project life cycle (Burkov, 2020)

in data pre-processing while modelling accounts only for the remaining 20% (Kotu & Deshpande, 2019). Furthermore, a machine learning project is often iterative, and it may be necessary to return to the feature engineering and data collection step, or even to the general goal definition in order to improve the model. Once a satisfactory model accuracy is achieved, the machine learning model can be deployed. After successful deployment, it should not be forgotten that a machine learning model must be monitored and maintained over time.

The main steps of machine learning are shown in an alternate format in Figure 2.20. While Figure 2.19 presented the key steps and highlighted the person responsible for each step, Figure 2.20 describes the tasks related to each step. It is reiterated that data pre-processing is a critical step and often require significant time and efforts.

2.7 Data pre-processing/ Feature engineering

The performance and the capacity of machine learning algorithms to learn from data is linked to the quality of the input data (Raschka & Mirjalili, 2019). Before training a machine learning model, it is necessary to carefully prepare the data by handling missing data, select useful features to train on (i.e. feature selection), remove outliers, remove skewness in the data, transform categorical data in a form that is usable by the machine learning algorithm (Kuhn & Johnson, 2013). To achieve this purpose, it might

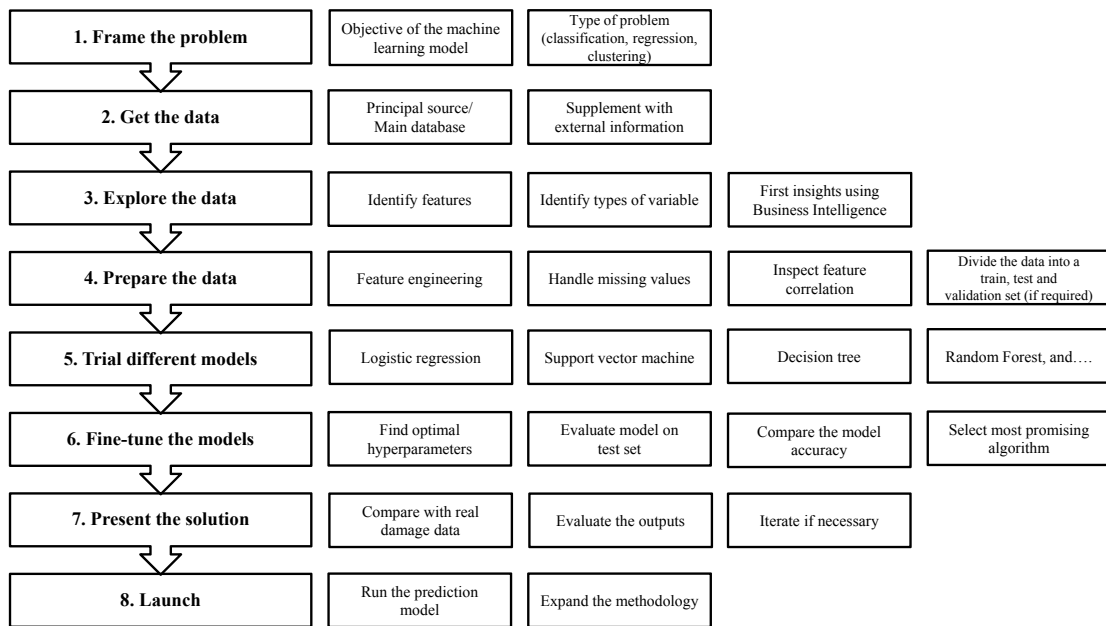


Figure 2.20: Main steps of a machine learning model, adapted from (Géron, 2019)

be necessary to combine existing features to generate new features more appropriate for the objective (i.e. feature extraction) or create new features by adding information from additional data (Géron, 2019).

The requirement for data pre-processing depends on the type of input data, which can for instance be numerical, categorial, image, text, or speech. This research project deals only with spatial, numerical, and categorial data. This review thus focus on the preparation of numerical and categorial input data.

2.7.1 Feature extraction and feature engineering

Feature extraction and feature engineering define the conceptual as well as programmatic process of mining, extracting, and transforming raw data into tidy data suitable for machine learning. For supervised learning, the input data should be labelled or tagged data. In other words, raw data that has been interpreted/labelled/tagged some informative target sets. For example, an unlabelled data x-ray image can be labelled as diseased or healthy. The objective of feature engineering is to transform raw data into feature vectors where feature vectors are one-dimensional arrays. Each feature vector is a sequence of values which describes the example and has a dimensionality (vector's length) related to the number of values in the sequence. The transformation from raw

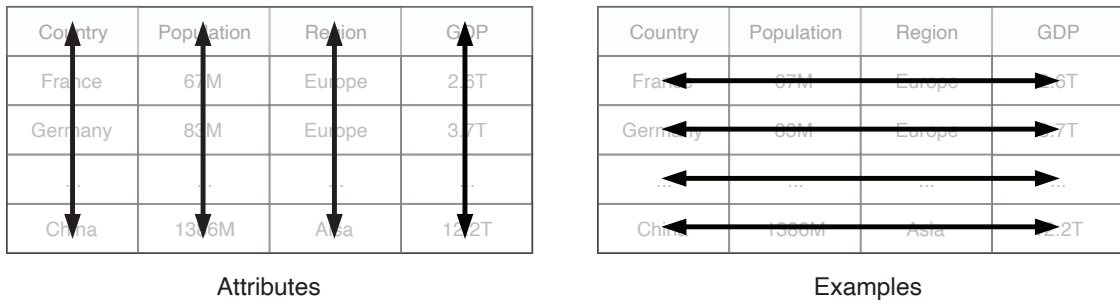


Figure 2.21: Tabular tidy data. The columns represents attributes and the rows examples (Burkov, 2020)

1.3.2 Raw and Tidy Data

data to tidy data in the form of feature vectors should be consistent for all the examples in the data set such that each feature vector entails the same type of information (attribute) at the same position in the sequence. Figure 2.21 shows an example of tabular tidy data.

Examples are represented by rows and columns represent attributes. For all the data set, the information pertaining to a same attribute are located at the same position (i.e. the same column).

In the literature, the terms *feature* and *attribute* are usually used interchangeably. Nevertheless, strictly speaking, an attribute is a data type, a specific property in a data set. A feature is an attribute with a value attached to it. Similarly, the terms “instance” or “sample” are employed in place of “example” (Burkov, 2020; Géron, 2019).

Unfortunately, there is no one-size-fits-all feature engineering process. Some transformations are optimal for some algorithms but are suboptimal for others. Some algorithms have the ability to select useful features and predictors that benefit the model accuracy, other algorithms do not. Consequently, feature engineering is machine learning project specific.

Handling missing data

Research has highlighted that the majority of the ML algorithms do not work, or perform very poorly when there is missing feature in any entry of the input data. Thus, a precondition critical to the success of a machine learning model is the requirement of “clean data” (Géron, 2019; Raschka & Mirjalili, 2019). Depending on the number of missing values compared to the number of examples in the data set, different approaches can be followed. It is possible to simply discard the instances or attributes entailing missing values. In some cases, it is also possible to fill in the missing values using for

example

- the mean, median or mode,
- the imputation of the missing value (e.g. using nearest neighbours or specific libraries),
- the prediction of the missing values (i.e. using correlation in the existing data).

2.7.2 Feature selection

Feature selection is the process of selecting relevant features for the construction of a machine learning model. Only the features relevant to the task are to be retained. The goal of feature selection is 1) to improve the prediction performance, 2) to provide a faster and less memory-intensive learning time, and 3) to enable a better understanding (Guyon & Elisseeff, 2003).

Guyon and Elisseeff (2003) showed that variables that are perfectly correlated are redundant. It is thus advised to remove correlated features as some algorithms perform poorly if input variables are highly correlated (Tang et al., 2014).

2.7.3 Feature transformation

Handling numerical features

Some algorithms might have specifications related to the form of the input. One of the standard requirements is the need for a common scale between attributes. Techniques such as normalisation and feature scaling should be applied before the model training. Logarithm, square root, inverse or statistical methods such as the Box-Cox transformation can be used to address data skewness. Outliers might also affect the model performance. While the removal of outliers can benefit the model performance, it should always follow critical judgement. Outliers are not always recording errors or mistakes. In some cases, outliers represent valid values.

Handling categorical features

Categorical features require processing before algorithms can make use of them. A typical process to transform categorical features into input variables suitable for ML is

one-hot encoding. One-hot encoding creates one binary attribute for each category of a categorical feature (Géron, 2019). For example if a data set includes an attribute “Color” with four possible values (red, yellow, green, and blue). It is possible to express each value with a binary vector, as listed below.

Red = [1, 0, 0, 0]

Yellow = [0, 1, 0, 0]

Green = [0, 0, 1, 0]

Blue = [0, 0, 0, 1]

The reader interested in more information related to transformation techniques is directed to Kuhn and Johnson (2013).

2.7.4 Training, validation, and test set

Before developing a machine learning model, it is necessary to divide the data into different sets: the training, validation and test sets. Such practice is required to ensure the trained model can make predictions not only memorised from the training examples and that the trained model can generalise for other samples (without over- or underfitting). Only the training set is used for training the machine learning model. The validation (sometimes called development set) and test set should be left untouched during the training process. These are the holdout sets. The validation set is used to select the model and tune the hyperparameters. Hyperparameters are variables of the ML algorithm that affect the performance of the ML model but are not related to the training data (e.g. regularisation hyperparameter). The test set is employed to evaluate the final model performance (Burkov, 2020).

Figure 2.22 shows a schematical overview of a data set that has been split into a training test data set. The situation presented in Figure 2.22 does not include a validation set. In some cases, for example when the instances in the training data set are scarce, the model parameters can be tuned using cross-validation.

2.7.5 Class imbalance

In real-world applications, some data sets exhibit a significant imbalance between the classes (e.g. credit card fraud, disease). However, most standard machine learning

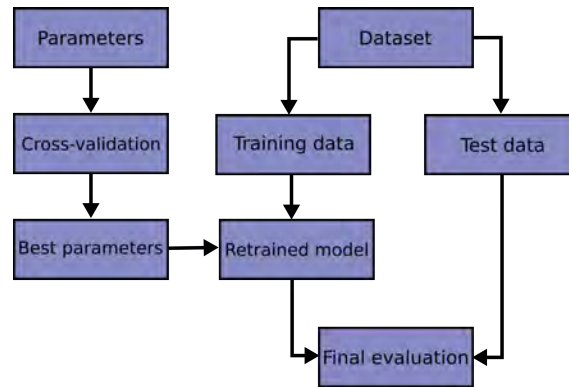


Figure 2.22: Overview of the train/test splitting when there is no validation set

algorithms for classification assume a balanced distribution in the classes. An imbalanced input data set can hamper the model performance as the algorithm will focus on the class with the largest number of instances (Haibo He & Garcia, 2009). This will lead to poor prediction performance, especially related to the prediction of the minority class.

In the last twenty year, many studies tried to overcome the issue of class imbalance (Prati et al., 2009). The proposed solutions can be grouped into two main categories depending on the approach: either at an algorithmic level or a data level. Algorithmic level solutions encompass approaches such as cost-sensitive learning, one-class classifiers, and ensembles of classifiers. The reader interested in more information related to these techniques is directed to Prati et al. (2009). The alternate data level solutions try to balance the data between the classes. Sampling methods are often used to transform the data and reduce the gap between the categories. It is either possible to remove instances in the majority classes to have a number of instances closer to the minority class (random under-sampling) or to replicate instances from the minority class (random over-sampling).

There are caveats from random under-sampling and random over-sampling. Random under-sampling might discard data that might be important for the model. Random over-sampling can increase the possibility of overfitting. The use of heuristics to remove only training examples that have a lesser importance for the model can be used to limit issues applying under-sampling methods. To avoid overfitting for over-sampling, interpolation can be applied for the under-represented class that are located close together (Prati et al., 2009).

Method	Over-sampling		Under-sampling	
	Binary	Mutli-class	Binary	Multiclass
ADASYN (He et al., 2008)	✓	✗	✗	✗
SMOTE (Chawla et al., 2002; Han et al., 2005; Nguyen et al., 2011)	✓	✗	✗	✗
ROS	✓	✓	✗	✗
CC	✗	✗	✓	✓
CNN (Hart, 1968)	✗	✗	✓	✓
ENN (Wilson, 1972)	✗	✗	✓	✓
RENN	✗	✗	✓	✓
AKNN	✗	✗	✓	✓
NM (Mani and Zhang, 2003)	✗	✗	✓	✓
NCL (Laurikkala, 2001)	✗	✗	✓	✓
OSS (Kubat et al., 1997)	✗	✗	✓	✓
RUS	✗	✗	✓	✓
IHT (Smith et al., 2014)	✗	✗	✓	✗
TL (Tomek, 1976)	✗	✗	✓	✗
BC (Liu et al., 2009)	✗	✗	✓	✗
EE (Liu et al., 2009)	✗	✗	✓	✓
SMOTE + ENN (Batista et al., 2003)	✓	✗	✓	✗
SMOTE + TL (Batista et al., 2003)	✓	✗	✓	✗

Figure 2.23: Overview of techniques for over- and under-sampling implemented in the `sample` performs the sampling and returns the data with the desired balancing ratio; and `imbalanced-learn` Python toolbox (Lemaitre et al., 2017a)

(iii) `fit_sample` is equivalent to calling the method `fit` followed by the method `sample`. A class `Pipeline` is inherited from the `scikit-learn` toolbox to automatically combine samplers, transformers, and estimators. Additionally, we provide some specific state-of-the-art metrics to evaluate classification performance.

Applying the solutions mentioned above to imbalanced data sets often improves the model predictions performance, especially for the under-represented class. However, it is often not possible to directly know which solution will work better. Depending on

The imbalanced-learn toolbox provides four different strategies to tackle the problem of the raw imbalanced data set: one of the other over-sampling (i) or under-sampling method and (iv) ensemble learning. The following subsections give an overview of the techniques implemented.

4.1 Notation and background

Let χ be an imbalanced dataset with χ_{min} and χ_{maj} being the subset of samples belonging to the minority and majority class, respectively. The balancing ratio of the dataset χ is defined as:

techniques in Python. Figure 2.23 presents an overview of the method that are embedded within the ‘imbalanced-learn’ toolbox. It includes random over-sampling (ROS) and

random under-sampling (RUS) as well as many heuristic under-sampling methods (e.g.

NearMiss (NM), Condensed Nearest Neighbor Rule (CNN), Tomek links (TL), One-sided Under-sampling (OSS), Neighborhood Cleaning Rule (NCL) and heuristic over-sampling

selection (AKNN) and cleaning under-sampling. *Fixed under-sampling* refer to the methods which perform under-sampling to obtain the appropriate balancing ratio $r_{\chi_{res}}$. Contrary

to the previous methods, *cleaning under-sampling* do not allow to reach specifically the nearest neighbour (ENN). The reader interested in more details on the background

behind each of the methods included in the imbalanced-learn package is directed to Lemaitre et al. (2017b).

2.8 Model training

2.8.1 Algorithm selection

Machine Learning can be classified based on their type of supervision: supervised, unsupervised, semi-supervised, and reinforcement learning. The choice of the type of supervision depends on the purpose of the model and the data available. Figure 2.24 shows an algorithms guide based on the python scikit-learn package. The selection is based on the type of tasks and the number of data points available.

For each type of supervision, several algorithms are available. Some of the most relevant algorithms for supervised learning are k-Nearest Neighbors, Linear Regression, Logistic Regression, Support Vector Machines (SVMs), Decision Trees, Random Forests. Details and data pre-processing requirements for the main algorithms for machine learning are summarised in Figure 2.25.

A widely used programming for machine learning is scikit-learn (Pedregosa et al., 2019). The scikit-learn module integrates several machine learning algorithms, both applicable to supervised and unsupervised problems. Through a user-friendly application programming interface (API), scikit-learn enables seamless integration of machine learning tools into Python projects (Buitinck et al., 2013).

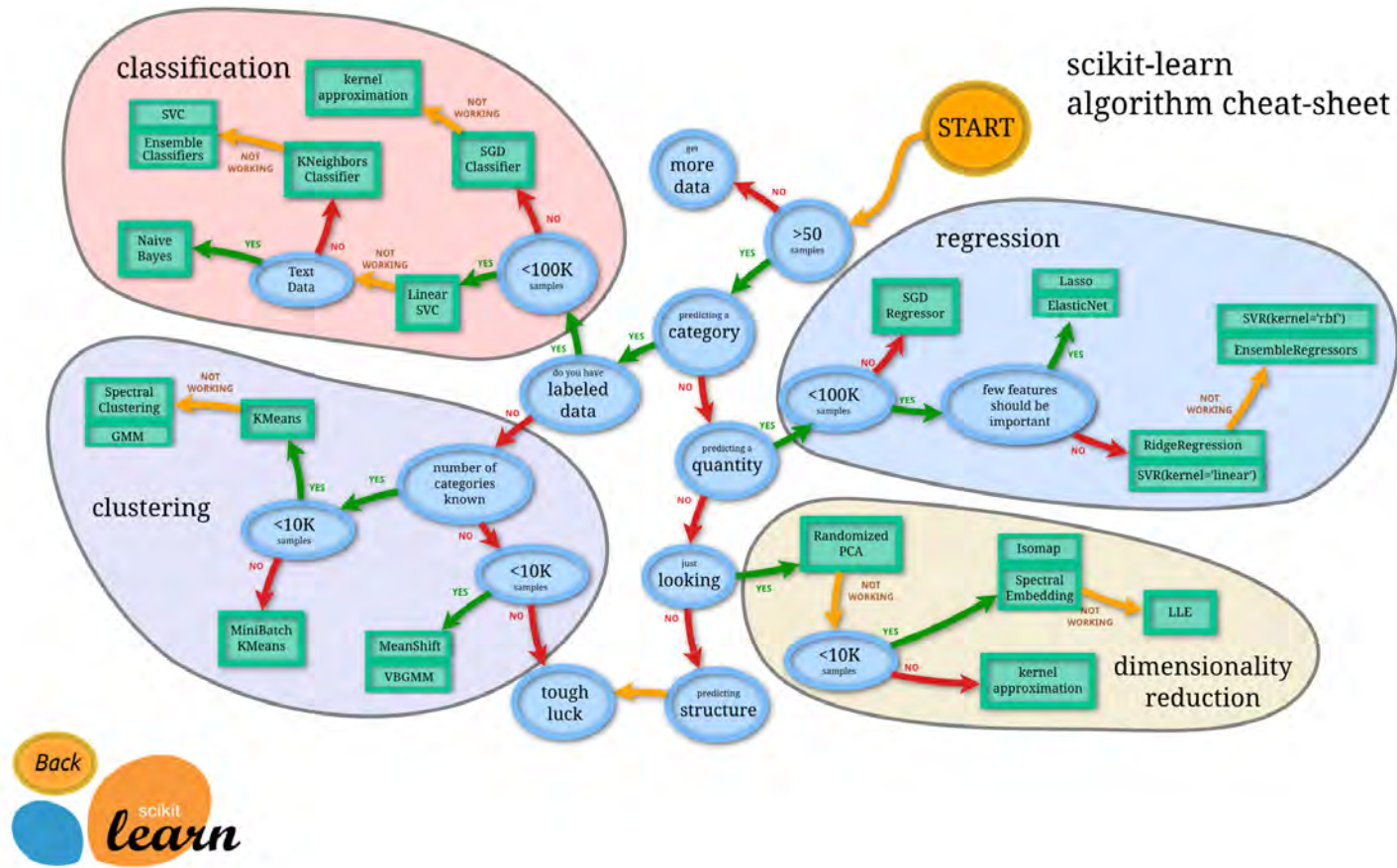


Figure 2.24: How to choose a machine learning algorithm? (Pedregosa et al., 2019)

Table A.1: A summary of models and some of their characteristics

Model	Allows $n < p$	Pre-processing	Interpretable	Automatic feature selection	# Tuning parameters	Robust to predictor noise	Computation time
Linear regression [†]	×	CS, NZV, Corr	✓	×	0	×	✓
Partial least squares	✓	CS	✓	○	1	×	✓
Ridge regression	×	CS, NZV	✓	×	1	×	✓
Elastic net/lasso	✓	CS, NZV	✓	✓	1–2	×	✓
Neural networks	✓	CS, NZV, Corr	×	×	2	×	×
Support vector machines	✓	CS	×	×	1–3	×	×
MARS/FDA	✓		○	✓	1–2	○	○
K -nearest neighbors	✓	CS, NZV	×	×	1	○	✓
Single trees	✓		○	✓	1	✓	✓
Model trees/rules [†]	✓		○	✓	1–2	✓	✓
Bagged trees	✓		×	✓	0	✓	○
Random forest	✓		×	○	0–1	✓	×
Boosted trees	✓		×	✓	3	✓	×
Cubist [†]	✓		×	○	2	✓	×
Logistic regression*	×	CS, NZV, Corr	✓	×	0	×	✓
{LQRM}DA*	×	NZV	○	×	0–2	×	✓
Nearest shrunken centroids*	✓	NZV	○	✓	1	×	✓
Naïve Bayes*	✓	NZV	×	×	0–1	○	○
C5.0*	✓		○	✓	0–3	✓	×

[†]regression only *classification only

Symbols represent affirmative (✓), negative (×), and somewhere in between (○)

- CS = centering and scaling
- NZV = remove near-zero predictors
- Corr = remove highly correlated predictors

Figure 2.25: Overview of main machine learning algorithms (Kuhn & Johnson, 2013)

2.8.2 Linear regression

A linear regression model is a linear model that assigns weights to each input feature and computes the sum to make a prediction. A constant parameter called the bias or intercept (θ_0) is added into the equation. The equation for a linear regression model prediction is shown in Equation 2.1.

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n \quad (2.1)$$

where: \hat{y} is the predicted value

n is the number of features

x_i is the i^{th} feature value

θ_j is j^{th} model parameter

Using a vectorized form, Equation 2.1 can be written as the dot product of θ and x as shown in Equation 2.2.

$$\hat{y} = \theta \cdot x \quad (2.2)$$

where: θ is the model's parameter vector (including θ_1 to θ_n and the bias term θ_0)

x is the instance's feature vector (including all the terms from x_1 to x_n)

2.8.3 Logistic regression

A logistic regression model is similar to a linear regression model as it computes a weighted sum of the input features, however, returns a binary prediction. It is defined as follow:

$$\hat{y} = \begin{cases} 0 & \text{if } \hat{p} < 0.5 \\ 1 & \text{if } \hat{p} \geq 0.5 \end{cases} \quad (2.3)$$

where: \hat{y} is the prediction

\hat{p} is the probability defined as

$$\hat{p} = \sigma(\mathbf{x}^T \boldsymbol{\theta}) \quad (2.4)$$

where: \mathbf{x} is vector of input features

$\boldsymbol{\theta}$ is the model parameter vector

σ is the logistic function defined as

$$\sigma(t) = \frac{1}{1 + e^{-t}} \quad (2.5)$$

2.8.4 Support Vector Machine (SVM)

SVMs algorithms can be linear or nonlinear. Equation 2.6 the properties of a classifier prediction using a separating hyperplane:

$$\hat{y} = \begin{cases} 0 & \text{if } \mathbf{w}^T \mathbf{x} + b < 0 \\ 1 & \text{if } \mathbf{w}^T \mathbf{x} + b \geq 0 \end{cases} \quad (2.6)$$

where: \hat{y} is the prediction

\mathbf{w} is the the feature weights vector

\mathbf{x} is the instance

b is the bias term

Equation 2.6 works well for data sets that are linearly separable. This method can be generalized to adopt nonlinear boundaries by enlarging and transforming the feature space. Details on the implementation of nonlinear SVM can be found Hastie et al. (2009).

2.8.5 Decision trees

One of the main advantages of adopting decision trees algorithm is the high interpretability. This is possible by inspecting the steps within model through two-dimensional trees (Molnar, 2020).

The development of a decision tree relies on two main steps. First, the space of all possible values should be divided into non-overlapping regions. For recursive binary splits for classification problems, Gini index and the entropy (here denoted G and D respectively) are commonly used as the primary measures. These are defined as follow:

$$G = \sum_{k=1}^K \hat{p}_{mk}(1 - \hat{p}_{mk}) \quad (2.7)$$

$$D = - \sum_{k=1}^K \hat{p}_{mk} \log(\hat{p}_{mk}) \quad (2.8)$$

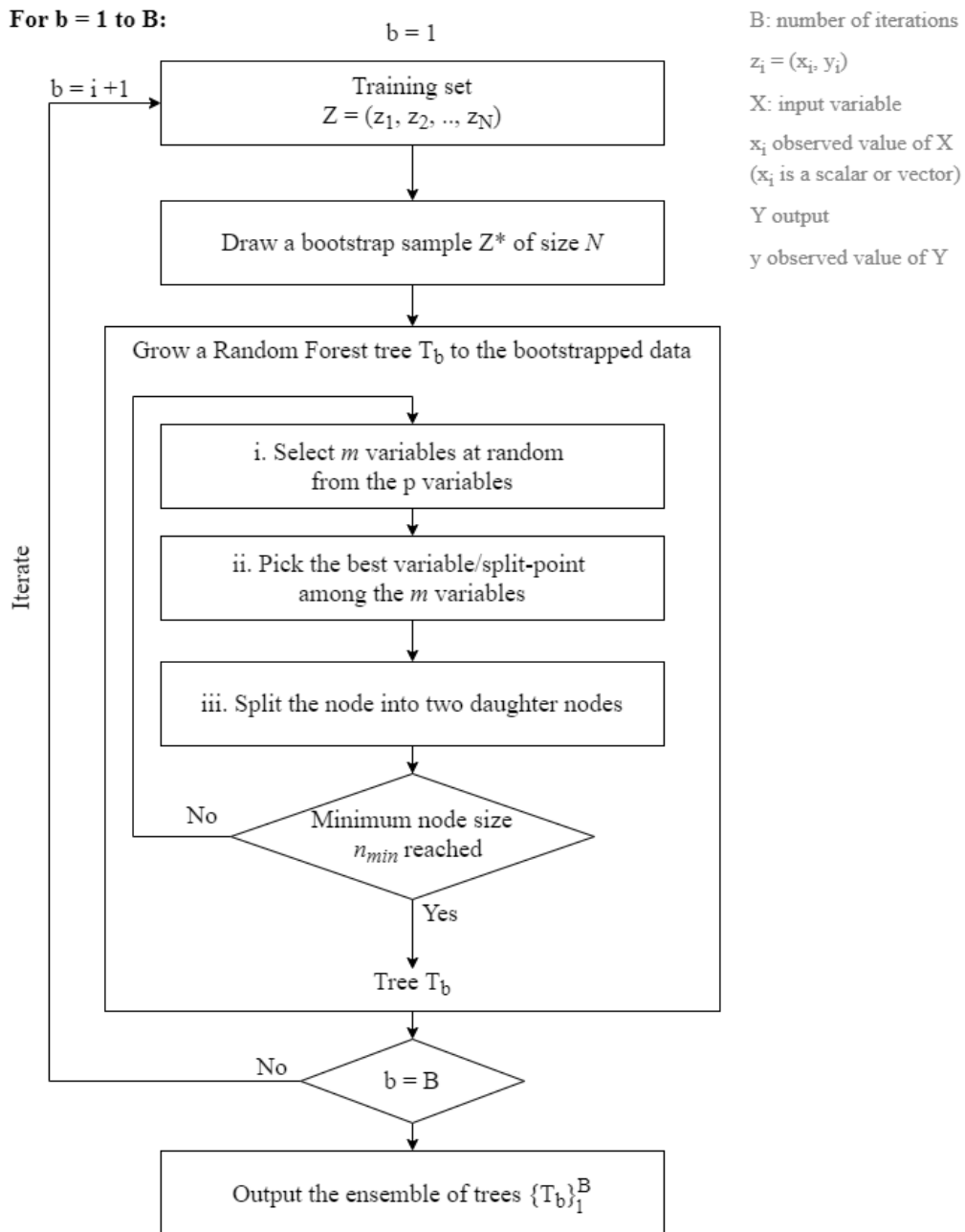
where p is the proportion of training observations in the m th region from the k th class. Second, the mean of the response values for the training observations is used to make prediction for observation within the same region. The process is then iterated to grow the tree (Hastie et al., 2009; James et al., 2013).

2.8.6 Random forest

Random Forest is an ensemble method that averages the aggregate of predictions from individual Decision Trees (Breiman, 2001). This has the benefit of reducing the variance of the error rather than the bias, thus it typically will produce more accurate prediction beyond the training data set than using a single decision tree model. Figure 2.26 shows the main steps of the random forest algorithm. Interested readers are directed to Hastie et al. (2009) for more details related to the mathematical background behind the random forest algorithm.

2.9 Model evaluation

A key step in developing a machine learning model is selecting the optimum algorithm. This can be a challenge with the large range of algorithms available. Data scientists have developed a range of objective metrics to aid model performance evaluation and hyperparameters (parameter of an algorithm) tuning. For classification models, the hyperparameters command the trade-off between precision and recall as well as the trade-off between bias and variance (Burkov, 2020).



To make a prediction at a new point x:

Regression: $\hat{f}_{rf}^B(x) = \frac{1}{B} \sum_{b=1}^B T_b(x)$

Classification: Let $\hat{c}_b(x)$ be the class prediction of the b^{th} Random Forest tree. Then $\hat{c}_{rf}^B(x) = \text{majority vote}(\hat{c}_b(x))_1^B$

A bootstrap sample is a random sample of the data taken with replacement. Each bootstrap sample is the same size as the original data. A bootstrap sample can contain multiple occurrences of an identical data point. The model performance is evaluated on the samples not selected by the bootstrap (Efron and Tibshirani 1986).

Figure 2.26: Overview of the Random Forest algorithm, adapted from (Hastie et al., 2009) and (Efron & Tibshirani, 1986)

Model performance - Confusion matrix

		PREDICTED CLASS		
		Positive	Negative	
ACTUAL CLASS	Positive	True Positive (TP)	False Negative (FN) <i>Type II error</i>	Recall $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) <i>Type I error</i>	True Negative (TN)	Specificity $\frac{TN}{(TN + FP)}$
		Precision $\frac{TP}{(TP + FP)}$	Negative Predictive Value $\frac{TN}{(TN + FN)}$	Accuracy $\frac{TP + TN}{(TP + TN + FP + FN)}$

Figure 2.27: Details of a confusion matrix

Damage prediction model

32

2.9.1 Performance evaluation of classification models

The accuracy of a model is important, but it is not the only performance measure for classification models. Other important metrics include,

- precision (also called positive predictive value),
- recall (also called sensitivity),
- F-score,
- cost-sensitive accuracy, and
- the area under the receiver operating characteristic (ROC) curve (AUC).

Figure 2.27 shows a typical confusion matrix. It is a useful tool for reporting on the accuracy of a predictive model, it tables the actual class (as rows) against the model prediction (as columns) This separates the prediction outcomes as true positives, true negatives, false positive, and false negative.

These definitions enable two additional measures: precision and recall which are as defined in equations 2.9 and 2.10. Precision indicates the accuracy of the positive predictions, and recall expresses the true positive rate as a ratio of the positive predictions over all actual positive values.

$$precision = \frac{TruePositives}{TruePositives + FalsePositives} \quad (2.9)$$

$$recall = \frac{TruePositives}{TruePositives + FalseNegatives} \quad (2.10)$$

The F_1 score is a single combined representative metric of precision and recall developed to simplify the performance comparison. The F_1 score, as defined in equations 2.11 and 2.12, is the harmonic mean of precision and recall. It is high only if both recall and precision are high. A high F_1 score thus indicates a good model performance.

$$F_1score = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} \quad (2.11)$$

which can also be re-written as,

$$F_1score = \frac{TruePositives}{TruePositives + \frac{FalseNegatives + FalsePositives}{2}} \quad (2.12)$$

2.9.2 Good fit, overfitting, underfitting

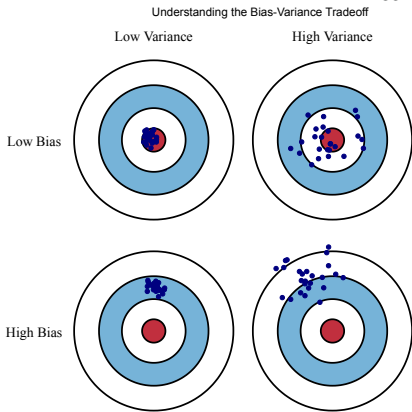
The performance of a machine learning model relates to its ability to learn, from a training set, and generalise predictions on unseen data (test data) (Hastie et al., 2009). To achieve this objective, it is important to find a balance between the training error and the prediction error (generalisation error). This is known as the bias-variance trade-off (Burkov, 2020). Figure 2.28a illustrates the concepts of bias and variance using points on a target. Each point symbolises an individual realisation of the model (Fortmann-Roe, 2012). In this illustration, the variance is represented by the scattering of the data. The lower the variance, the lower the scatter in the data. The bias is expressed by the distance to the middle of the target. The closer the points are to the bullseye, the lower the bias.

Figure 2.28b symbolically presents the bias-variance trade-off as a function of the model complexity and prediction error. More complex models bring a lower bias on the training set, however, also induce more variance in the model. Ideally, both low bias and low variance are desirable (James et al., 2013). Figure 2.28b makes it clear that both measures typically cannot be minimised simultaneously, as a model that performs well on the training set will have difficulties to generalise thus leading to a higher prediction error for the test set. The task of the model developer is thus to select a model and

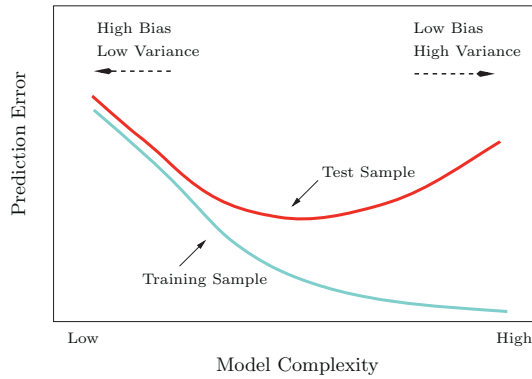
11/5/2020

38

2. Overview of Supervised Learning



(a) Graphical illustration of bias and variance (Fortmann-Roe, 2012)

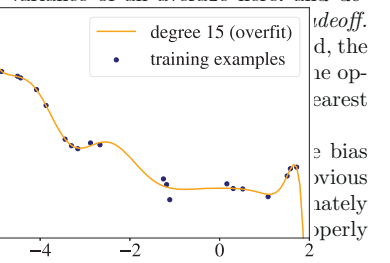
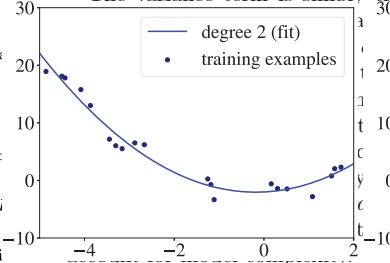
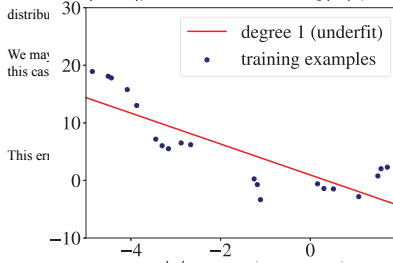


(b) Test and training error as a function of model complexity.

1.3 Mathematical Definition

after Hastie, et al. 2009

If we denote the variable we are trying to predict as Y and our covariates as X , we may assume that there is a relationship relating one to the other such as $Y = f(X) + \epsilon$ where the error term ϵ is normally distributed.



Underfitting

Good fit

Overfitting

(c) Graphical representation of underfitting, good fit, and overfitting (Burkov, 2020)

2 An Illustrative Example: Voting

Let's undertake a simple model building task. We wish to create a model for the percentage of people who will vote for a Republican president in the next election. As models go, this is conceptually trivial and much simpler than what people commonly envision when they think of modeling. In this case, the data clearly illustrate the difference between bias and variance. A straightforward, if flawed (as we will see below), way to build this model would be to call 50 numbers from the phone book, call each one and ask the responder who they planned to vote for in the next election. Imagine we got the following results:

Voting Republican	Voting Democratic	Non-Respondent	Total
13	16	21	50

Figure 2.26 graphically illustrates the concept of 'good fit', 'underfitting', and 'overfitting' for a set of points.

The red linear curve has a low variance but a high bias which lead to a large prediction error for new instances. In other words, the model is underfitting.

Contrary, the yellow curve has a low bias but a high variance. The model will perform well on the training set but is not able to generalise properly for new instances, the model is overfitting. Finally, the blue curve captures the overall trend of the data while remaining simple enough. The model generalise well for new instances.

Several solutions to overfitting are possible:

- the model is too complex for the data. Very tall decision trees or a very deep neural network often overfit;
- there are too many features and few training examples; and
- you don't regularize enough.

Several solutions to overfitting are possible:

- use a simpler model. Try linear instead of polynomial regression, or SVM with a linear kernel instead of RBF, or a neural network with fewer layers/units;
- reduce the dimensionality of examples in the dataset;

2.10 Interpretability of machine learning models

2.10.1 Background

Apart from making predictions, machine learning can derive insights, identify relationships between input variables, and/or find patterns in the data that may be hidden from conventional analysis (Lee, 2018). Depending on the aim and purpose of the machine learning model, obtaining correct prediction only may be satisfactory. Recommender systems are an example where the emphasis is cast on the results (recommendations) rather than on the paths that led to it. However, recent applications of machine learning showed that interpretability could help the end-user (Honegger, 2018). Human interpretability of the predictions is closely related to model trust (Miller, 2019). Opacity in the way predictions are made often leads to mistrust in the model, making its application, implementation and acceptance more difficult (Molnar, 2020).

Model interpretability is achievable in two main ways. It could come from the possibility for humans to understand the parameters of the algorithm (intrinsic interpretability). This is for example the case for linear regression which remains interpretable due to its simple structure. Table 2.2 shows a list of intrinsically interpretable models. They are sometimes called 'white-box models' as they are relatively straight forward to be understood by humans. However, many current ML algorithms are too difficult to be directly interpreted by humans. For example, it is not straightforward to rationalise the actions and decisions of artificial neural network (ANN) models. Moreover, the best performing models can be a combination of multiple algorithms (ensemble models) making the prediction process even more complicated to follow. Models that are not easily understood by human are referred to as 'black box models'. For complex models, interpretability could come from methods that analyse the machine learning model after it has been trained (post hoc methods). Developing post hoc solutions to make complex model decisions understandable to humans remains a topical research endeavour (Du et al., 2020; Molnar, 2020; Ribeiro et al., 2016a, 2016b, 2018).

Table 2.2: List of some intrinsic interpretable machine learning algorithms, after (Molnar, 2020)

Algorithm	Linear	Monotone	Interaction	Task
Linear regression	Yes	Yes	No	regression
Logistic regression	No	Yes	No	classification
Decision trees	No	Some	Yes	classification and regression
Naive Bayes	No	Yes	No	classification
k-nearest neighbors	No	No	No	classification and regression

2.10.2 SHapley Additive exPlanations (SHAP)

SHAP is a methodology originally conceived in game theory for computing the contribution of model features to explain the prediction of a specific instance (Lundberg & Lee, 2017). The SHAP methodology has latter been extended to the interpretation of tree-based machine learning algorithms (Lundberg et al., 2018). It can be used to rank the importance of the model features.

The permutation of features is another technique used to find the importance of features in tree-based models. The feature importance is determined through the influence on the model prediction error of the feature's values permutation. A high model prediction error indicates a feature with significant importance. To be able to rank the features by importance, the process is repeated for each feature in the model (Fisher et al., 2018).

SHAP relies on the weight of feature attribution rather than on the study of the decrease in model performance. It is thus more robust than the feature importance using permutation (Lundberg et al., 2018; Molnar, 2020).

Figure 2.29 presents an example of a SHAP summary plot showing the feature importance for the Census income data set. The Census Income data set, obtained from the UCI machine learning repository (Dua & Graff, 2019), entails demographic data and information whether a person makes less or more than US\$50k annually. Lundberg et al (2020) used twelve attributes, numerical and categorical, to train a ML model using gradient boosting (LightGBM). SHAP was then applied on the ML model to obtain the

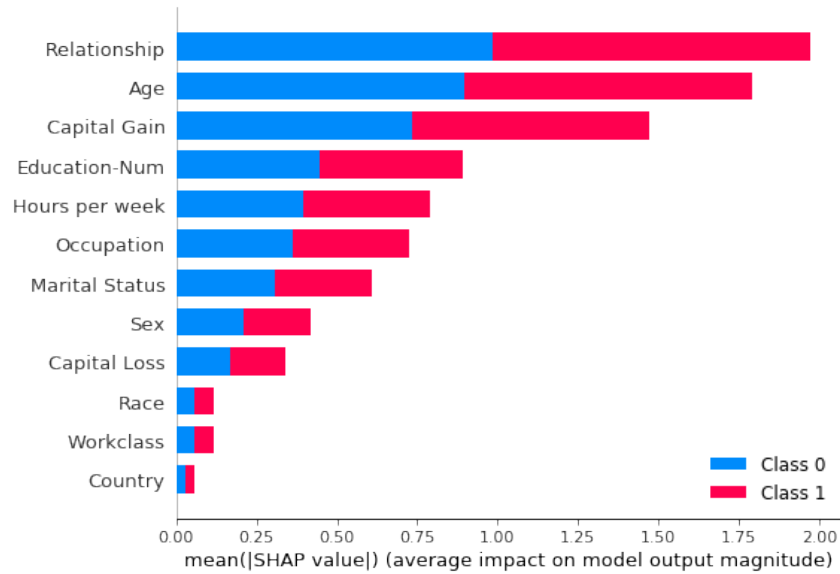


Figure 2.29: Feature importance computed using SHAP (Lundberg, 2020)

feature importance. The attributes “relationship”, “age,” and “capital gain” stood out as the most significant, impacting the most the likelihood of a person to make more than US\$50k per year.

Development of a universal damage data collection framework

This chapter proposes a new paper-based tool which addresses the need for a global yet detailed universal methodology for building damage assessment post-earthquakes. Currently, to make use of the data sets from around the world, significant effort is required to decode the data which often have unique local and regional context and bias. The struggle beings at data collection where there is a lack of consistent methodology and definitions that can adequately cover the regional nuance. This new form is based on the GEM taxonomy v2.0 and the European Macroseismic Scale EMS-98 and is flexible enough to be used anywhere in the world.

The recent Puebla earthquake on 19th September 2017 led to significant building damage in Mexico City and the state of Morelos. A team from New Zealand assessed damage throughout the capital and trialled the new paper form along a significantly affected street, Calle La Morena. This chapter presents the case study which showcase the use of a comprehensive damage data and buildings characteristics visualisation. This highlights the correlation between the damage and the building features, and leads to better comprehension of the damage drivers.

3.1 Introduction

One of the earliest contemporary earthquake reconnaissance and observation for scientific account was that by Robert Mallet (1810-1881). Mallet spent one month in Italy following the 1857 great Neapolitan earthquake (Ferrari & McConnell, 2005), and he collected meaningful data and documented his finding in the report “The first principles of observational seismology” (Mallet, 1862). Approximately fifty years later, the U.S. experienced one of the worst natural disasters in its history with the 1906 San Francisco earthquake. The earthquake and following fire killed more than 3,000 persons and left 400,000 persons homeless (The U.S. National Archives and Records Administration, 2017). More importantly, before the event, San Francisco was the premier city of the West Coast. After 1906, population abandoned San Francisco, and Los Angeles quickly outgrew San Francisco in the following decade. Los Angeles is now the financial centre of California and is five times as large as San Francisco (Jones, 2014). The San Francisco earthquake and other significant earthquakes such as the 1923 Kanto and 1995 Kobe Earthquakes highlighted the importance of the economic and social consequences of earthquake disasters. It pointed to the need for improving the resilience of our cities and urban environment.

Purpose of modern earthquake engineering is to mitigate damage in buildings and infrastructures to reduce the impact of earthquakes on society. Earthquake risk reduction is a multidisciplinary risk management exercise and is not restricted to improving the science behind more accurate seismic hazard prediction. It encompasses enhancing the understanding of the exposure and vulnerability of our built environment. Then, applying appropriate risk management measures such as Avoid-Control-Accept-Transfer to minimise the related losses. Earthquake risk reduction requires knowledge from seismology, structural and geotechnical engineering as well as psychology and economics. The most valuable resource in damage mitigation and loss prediction for future earthquakes is empirical data and lessons from past events. In structural engineering fields, the perishable building performance data on damage and undamaged buildings are invaluable in identifying failure causes and damage patterns. These observations can enable engineers, planners and officials to adjust the current setting

to improve future social and economic outcomes. The most direct lever for this is via improving structural design standards and seismic loss prediction models.

3.2 Improved paper form based on the GEM Building Taxonomy v2.0

The improved post-earthquake building damage assessment form (herein noted as the form) proposed by this project was designed with universal applicability in mind, in line with the GEM objective. Care was taken to develop a standard and consistent definition of building features to ensure past, current and future data are comparable. It is the first damage assessment form that combines and aligns the GEM Building Taxonomy v2.0 (Brzev et al., 2013) and EMS-98 (Grünthal, 1998) in one paper form. While the European Macroseismic Scale EMS-98 was primarily developed for European countries, engineers used the EMS-98 damage scale to assess building damage all over the world: in Italy (Borg et al., 2010; Del Gaudio et al., 2017), in France/Spain (Monfort et al., 2011), in Mexico (Juarez-Garcia et al., 2004), in New Zealand (Cattari et al., 2015; Fikri et al., 2018; Stirling et al., 2015). The GEM Building Taxonomy is appropriate to consistently describe and classify buildings worldwide (L. Allen et al., 2015). It was already employed in several projects (Global Earthquake Model (GEM), 2015; Silva et al., 2018; Wieland et al., 2015). Unlike most existing damage survey forms, the newly developed paper form allows for the collection of non-structural components seismic performance data. It combines the nonstructural damage observation form developed by (Taghavi & Miranda, 2003) and the non-structural building components taxonomy developed by Porter (2005). It includes informative sketches on the type of seismic design obtained from the glossary of GEM taxonomy (L. Allen et al., 2013) to facilitate the assessment of the structural system. A copy of the paper form is included in Appendix B.3.

The improved paper form comprises of six mains sections.

- Section one collects information on the assessor and general building information such as building location (address and GPS coordinates).
- Section two records building information: type of occupancy, number of stories, building position within a block, date of construction, building shape.

- Section three records the general description of building damage including location and extent based on the EMS-98 (Grünthal, 1998). This section has been deliberately placed on the first page to facilitate the data post-processing and directly enable to identify the level of damage without having to go through the entire form. This section also introduces essential definitions that assist the assessment of building elements.
- Section four focuses on structural elements. It records details about the building material and the lateral-load resisting system (LLRS) in each direction, and any structural irregularity in the structure (plan and vertical).
- Section five records information on the building exterior attributes (e.g. roof, façade), flooring systems, foundation system and ground condition.
- Section six is dedicated to recording observations on non-structural components and some building content. A space is reserved on the final page for a sketch of the building and it prompts the assessor to record reference and captions for any photographs taken.

3.2.1 Field trial of the improved paper form

An international Learning From Earthquake (LFE) team trialled the improved form on site in Mexico City following the 19th September 2017 Puebla earthquake. The team included local and foreign researchers, graduate and undergraduate students. The experience in deploying the improved paper form was generally positive. It recorded greater detail on the building and the damage, it enabled consistent damage grading and was commented to be easy to use. The definition of the building damage according to EMS-98 was well received and understood by local assessors, as this was also used in other local efforts. One of the advantages of EMS-98 is that the documents are available in full and short form in English, French, Spanish, and Chinese (Grünthal, 1998). Experience in the field showed that training is necessary for assessors to become familiar with assessment forms. The alignment of the new paper form with GEM taxonomy also made the reporting of the information in GEM – Direct Observation Tools (Jordan et al., 2014) straightforward.

The new form was designed with an emphasis on non-structural components, to collect empirical data on the seismic performance of non-structural components. However, the experience from the field trial is that internal inspection is not always possible due to difficulties in obtaining access to private properties, the limited time available and other health and safety concerns. An alternative and refinement for future missions could be to send a questionnaire on non-structural components to owners of damaged building. Previous research showed that satisfactory results could be achieved if the form is self-explanatory (Taghavi & Miranda, 2003).

3.2.2 Future improvements

Assessors appreciated the simplified sketches depicting damage and structural categories included in the form. Some assessors suggested more pictures and complementary explanation to be included in the paper form (e.g. explanation on short columns, cripple walls, torsional eccentricity). Future refinement will need to consider the trade-off between improved explanation and increasing the form length. Version 1.2.0 of the IDCT Android app (March 2018) was available in Spanish. Translation of the paper form into Spanish and other languages used in earthquake-prone countries is recommended and could make the assessment process more accessible. Due to internal access difficulty mentioned previously, it may be argued that the non-structural components section in the current form should be reduced to recording obvious and critical non-structural component damage.

The table in Appendix C.2 shows a detailed comparison of the features of the assessment form based on the GEM Building Taxonomy v2.0 versus the local form used in Mexico. A significant advantage of the paper form based on the GEM Building Taxonomy v2.0 is the direct integration in the GEM data type. Any assessments completed with the IDCT tool produce outputs that are consistent with the data required for the GEM exposure and consequences databases, and thus directly usable by any software and tool developed by the GEM community (Global Earthquake Model (GEM), 2014; Pagani et al., 2014). Interested readers about the possibilities of GEM tools are directed to The OpenQuake-engine User Manual (Pagani et al., 2019) as well as the OpenQuake Risk Modeller's Toolkit - User Guide (Global Earthquake Model (GEM),

2018). The broad terminology of GEM Building Taxonomy gives high flexibility in attribute definitions. The scope goes beyond the building industry and is transferable for loss calculation and estimation for the insurance industry.

3.3 Case study: Calle La Morena

3.3.1 Building assessment

The assessors conducted detailed inspections on twenty-five buildings in the western part of Calle La Morena. These inspections formed a representative case study for buildings with broad variations of features and wide extent of damage ranging from no damage to total collapse. The buildings are located in the geotechnical zone III a (Gobierno del Distrito Federal Mexico, 2004), south of the most damaged neighbourhoods of La Condesa and La Roma. It was not possible to access the inside of the buildings thus the assessments were limited to exterior inspections only. Nevertheless, the exterior inspections provided meaningful data on the building characteristics and extent of damage of the structural elements. In the Calle La Morena, the building assessor assessed damage with the improved GEM paper tool. The assessor then transferred the data in a digital form using the GEM Windows tool (Jordan et al., 2014), as shown in Figure 3.1a. The software links the damage assessment as well as photographs taken on site to a geotagged data point, which can be exported as a kmz file (Figure 3.1b) or as a .shp - Shapefile.

3.3.2 Statistical findings

Figure 3.2b shows the overall damage extent distribution for the assessed buildings, and Figure 3.2c presents the same distribution as categorised by number of storeys. Out of the twenty-five buildings assessed, fifteen suffered no damage, ten experienced at least slight damage and one building collapsed. The damage scale followed EMS-98, and a consistent colour code was used for all subsequent graphs in this chapter. The colour code ranged from beige for buildings with no damage to dark red for collapsed buildings.

Figure 3.3 plots the damage grade distribution categorised by the number of stories and building occupancy. It shows that five to eight storeys building were the most

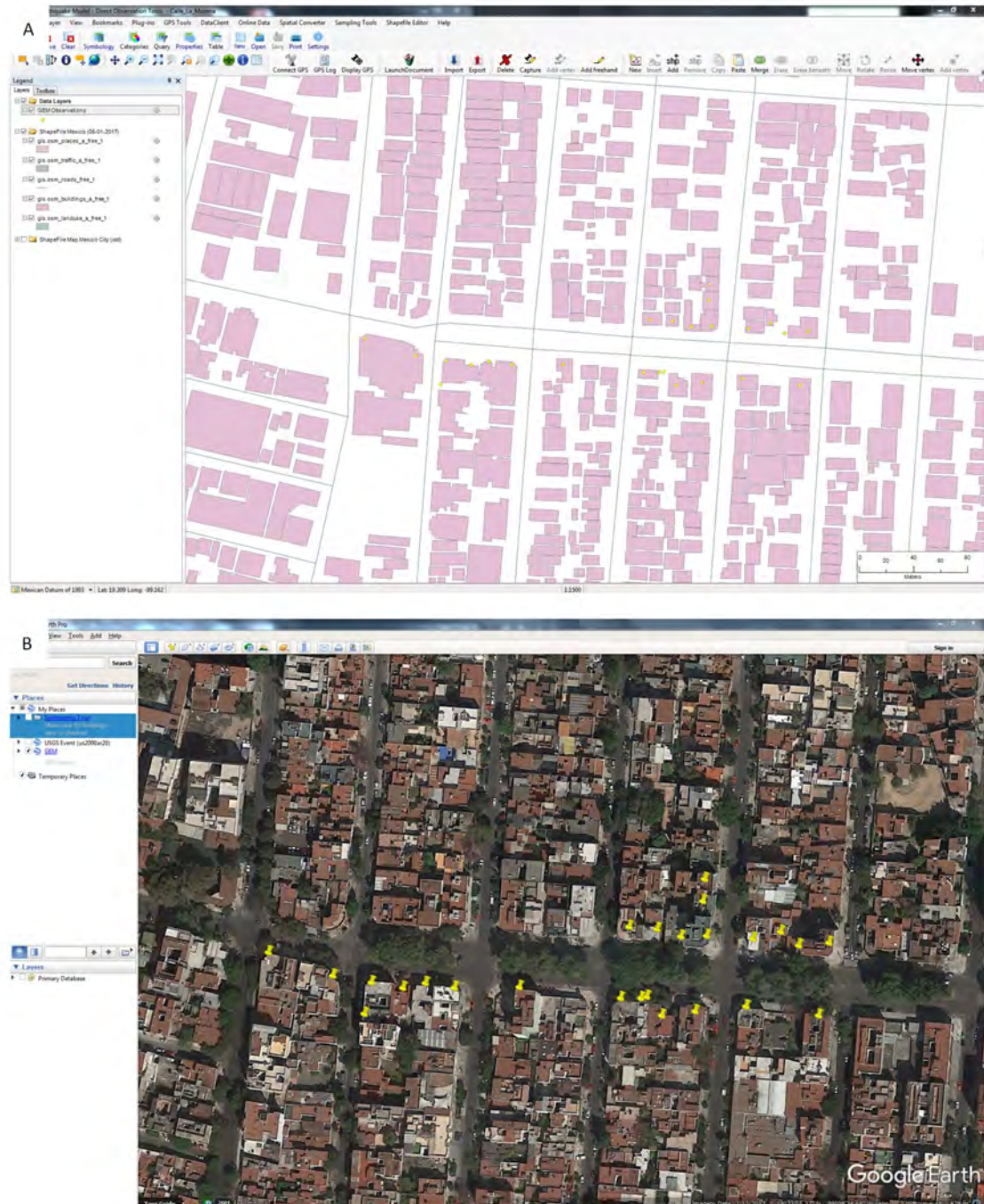


Figure 3.1: Screenshots of the GEM IDCT software. (a) Depicting building boundaries as available from shapefiles. (b) Location of the 25 buildings assessed.

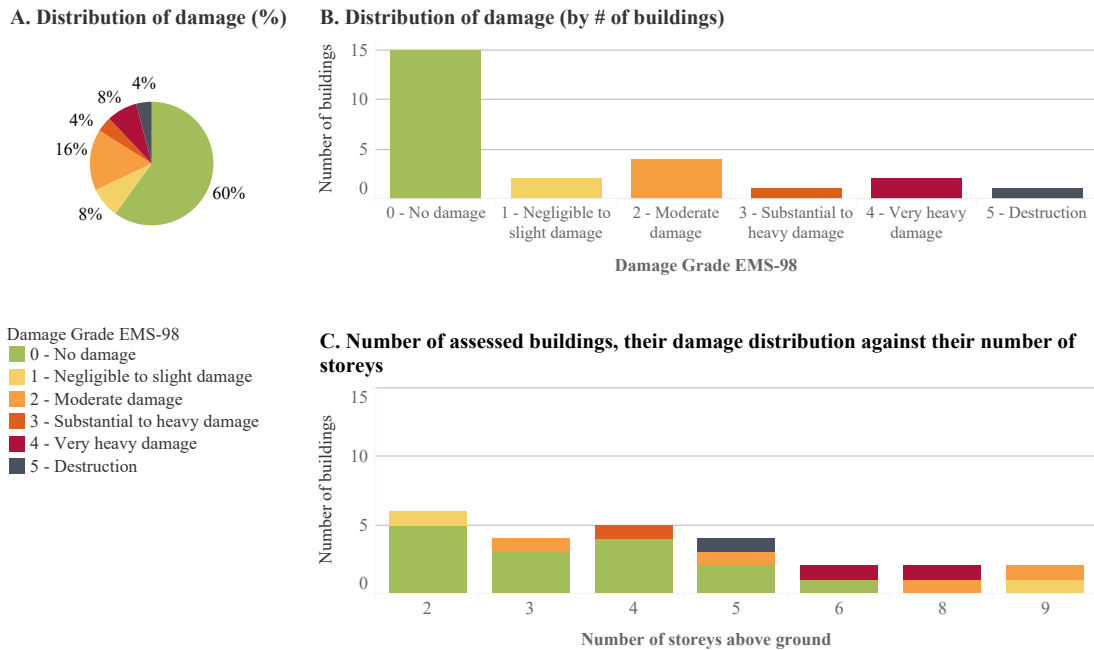


Figure 3.2: Distribution of damage categories for the 25 buildings studied in Calle La Morena. Damage categories as per EMS-98.

severely damaged. Of the assessed buildings, 60% are residential (RES) and 36% are mixed-used, mostly residential and commercial (MIX1).

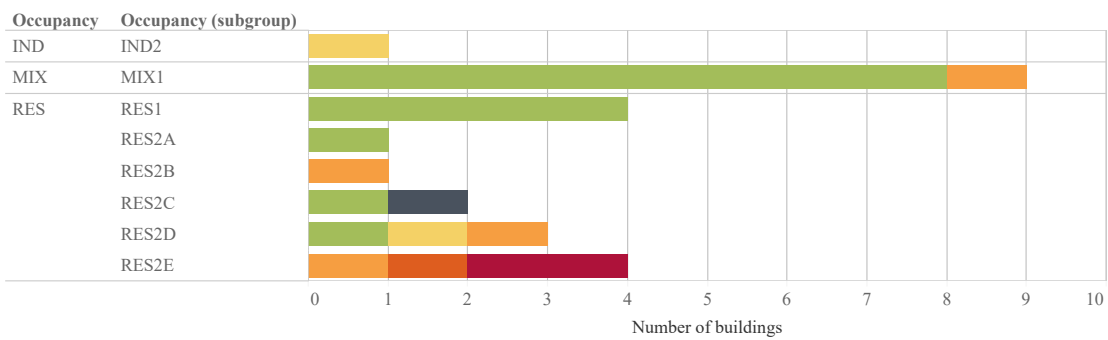
Buildings are classified into bins according to their plan shape and their position as per the GEM Building taxonomy definitions in Figure 3.4. 84% of the buildings assessed had two adjoining structures and 16% have adjoining building on one side only. The plan shape observed in Calle La Morena Street were diverse. Whilst conclusive statement cannot be made due to the small sample size, it appears buildings with a solid rectangular plan suffered the least damage. By contrast, H-shaped buildings experienced very heavy damage.

Figure 3.4 depicts the damage grade distribution classified by building material and lateral load resisting system. In the case study sample, 50% of the building are concrete buildings and 10% are masonry buildings. In 40% of the cases, it was not possible to determine the material of the structural elements. It was even more challenging when attempting to identify the lateral load resisting system in the field. Overall, it was not possible to identify the lateral load resisting system for 70% of the buildings in the case study. Like other reconnaissance missions, this experience highlighted the difficulty in obtaining accurate classification from external observations. This is oftentimes made

A. Occupancy and number of storeys



B. Occupancy (detailed)



OCCUPANCY (GEM Building Taxonomy v2.0)

IND Industrial

MIX Mixed use

IND2 Light industrial

MIX1 Mostly residential and commercial

RES Residential

RES1 Single dwelling

RES2B 3-4 Units

RES2D 10-19 Units

RES2A 2 Units (duplex)

RES2C 5-9 Units

RES2E 20-49 Units

Figure 3.3: (a) Building categorized by occupancy and number of storey. (b) Detail of the building occupancy.

more difficult with the lack of official authority of scientific data collection teams and health and safety concerns. The difficulty of defining the building structural system and material of the structural elements from external observation increases the inaccuracy and variability in the data. Having access to building plans would significantly improve the situation. For some buildings in Mexico City outside the scope of this case study, building plans and structural drawing were available. For most of these cases it was possible to determine the structural system and the material of the structural elements. This highlights the importance of preparation and collecting inventory data prior to a disaster, and archiving them so that the information is accessible immediately after a shock event.

Heavy damage and collapse, damage grades 3 to 5, were concentrated in buildings with an irregular structure (see Figure 3.6). For those buildings with an irregular structure, Figure 3.7 details the type of building irregularity observed in the vertical and horizontal directions. Figure 3.7a highlights that irregular buildings with a soft-storey

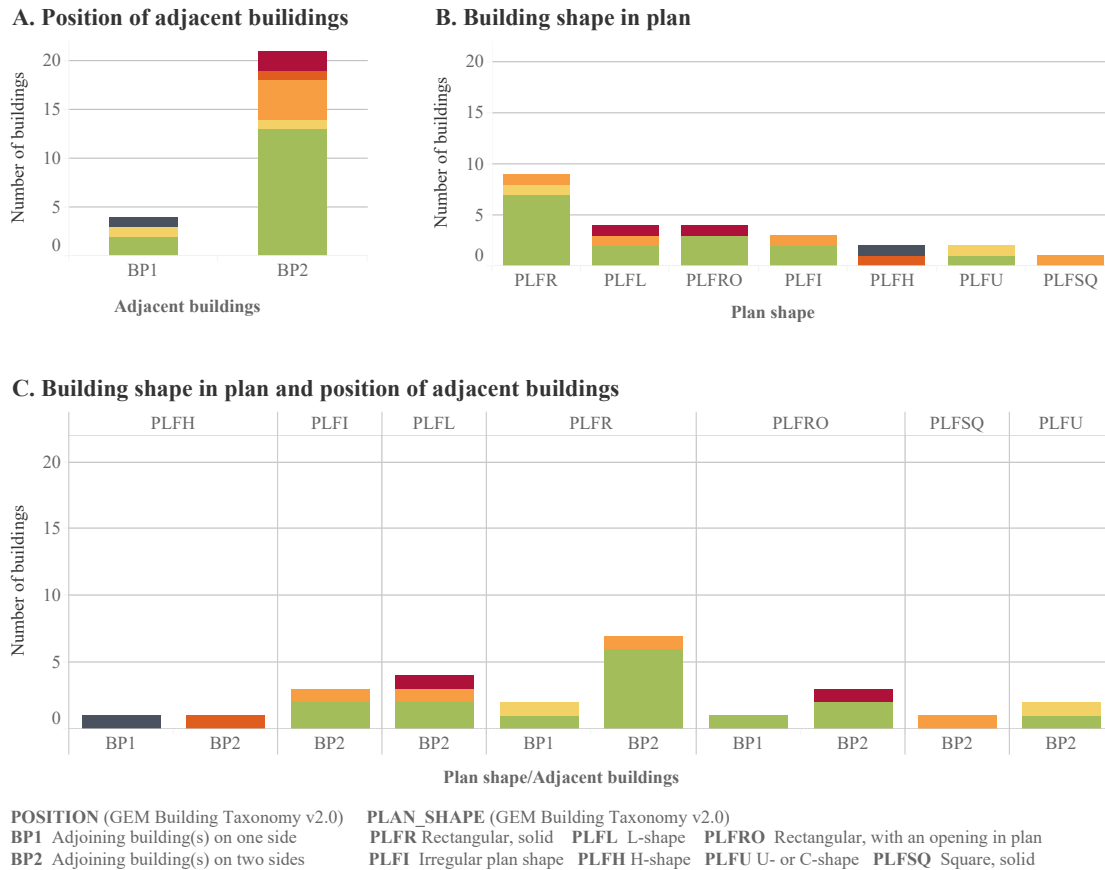


Figure 3.4: Damage grade distribution categorised by adjoining buildings and plan shape. (a) Configuration of adjoining buildings. (b) Plan shape of each building. (c) Combination of plan shape and building position.

suffered larger damage. Figure 3.7b shows that higher damage grades are also linked to buildings having torsion eccentricities and re-entrant corners.

3.3.3 Examples of observed damage

The LFE mission provided excellent opportunity to observe many common classic building failure mechanisms. Figure 3.8 presents a collection of noteworthy examples. Figure 3.8a depicts an example of a corner building with torsion eccentricity. The collapsed building had a re-entrant corner (REC). From our limited case study samples, all substantial and heavily damaged buildings exhibited torsion issues. 30% of the vertically irregular structures experienced pounding (POP) (an example is shown in Figure 3.8b) and 60% had soft story (SOS) failure (an example is shown in Figure 3.8c).

During the assessment following the Puebla earthquake, the team saw several cases of failed columns. Assessors observed shear cracks, as shown in Figure 3.9a or total column



Figure 3.5: Damage grade distribution categorised by adjoining buildings and plan shapes. Material of structural system in (a) longitudinal direction and (b) transversal direction. Lateral Load Resisting system in (c) longitudinal direction and (d) transversal direction.

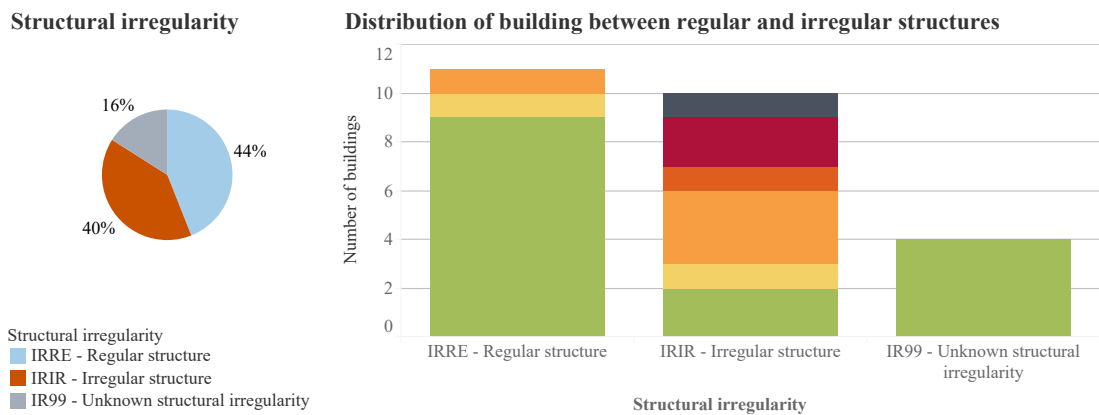
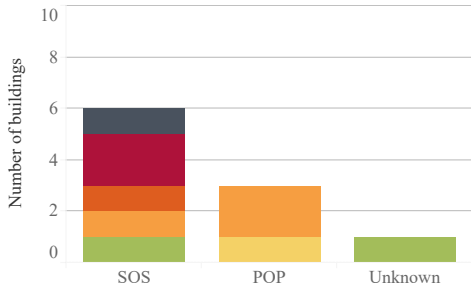


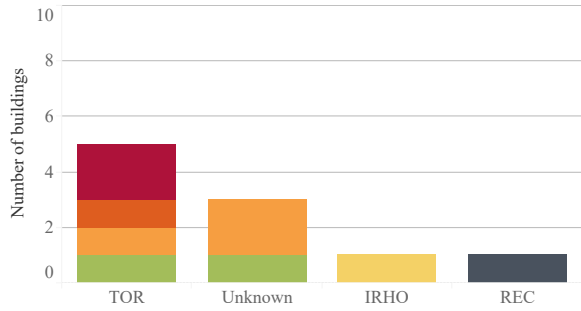
Figure 3.6: Distribution of damage for regular and irregular structures

A. Principal vertical irregularity



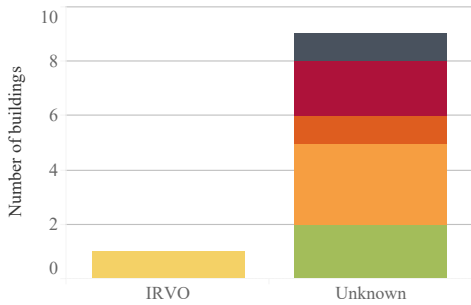
Vertical structural irregularity (GEM Building Taxonomy v2.0)
 SOS Soft storey
 POP Pounding potential

B. Principal horizontal irregularity



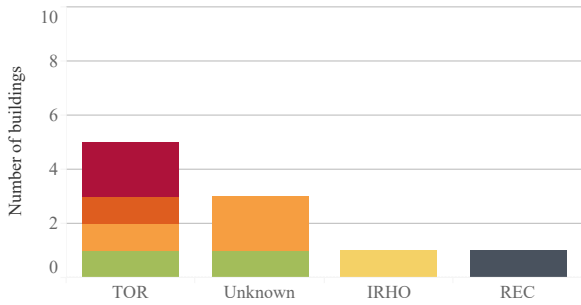
Horizontal irregularity (GEM Building Taxonomy v2.0)
 TOR Torsion eccentricity IRHO Other plan irregularity
 REC Re-entrant corner

C. Secondary vertical irregularity



Vertical structural irregularity (GEM Building Taxonomy v2.0)
 IRVO Other vertical irregularity

D. Secondary horizontal irregularity



Horizontal irregularity (GEM Building Taxonomy v2.0)
 TOR Torsion eccentricity IRHO Other plan irregularity
 REC Re-entrant corner

Figure 3.7: Distribution of damage for irregular structures. Principal (a) vertical and (b) horizontal irregularities. Secondary (c) vertical and (d) horizontal irregularities.

failure with longitudinal reinforcement bar buckling due to insufficient confinement from widely spaced stirrups (Figure 3.9b).



Figure 3.8: Representative damage in La Morena. (a) Corner building with torsion eccentricity. (b) Pounding failure. (c) Soft storey in the ground floor.



Figure 3.9: (a) Failure of the column in shear. (b) Buckling of the longitudinal reinforcement bars.

3.4 Conclusion

Following the 2017 Puebla earthquake in Mexico, a new paper form based on the GEM Building Taxonomy v2.0 was trialled on twenty-five buildings in Calle La Morena. Each category has a clear structure aligned with the GEM Building Taxonomy v2.0 and the damage grade is expressed according to the European Macroseismic Scale EMS-98 (Grünthal, 1998). The form also included indicative building feature sketches and explanation of the damage grades to aid the assessors in the identification of the structural system.

The buildings assessed were diverse by their number of stories, structural systems, occupancy, and position in a block. Building assessors collected relevant building characteristics such as the number of stories, the building occupancy, the position of the building, and the damage grade. Building damage ranged from no damage to total collapse.

The trial using the new form based on the GEM Building Taxonomy v2.0 was generally positive. Experience gathered in the field showed that the definition of building features according to the GEM taxonomy provided consistency in the data collected. The new form simplified the data extraction as data were easily exported into a GEM framework. The GEM structure simplifies the universal understanding while allowing for regional specificities.

Earthquake reconnaissance missions provide critical data for improving damage predictions models and quantitative insight into actual building performance. This case study is of interest to understand the factors and correlation of factors leading to vulnerability. This is assisted by visualising the damage distribution data individually by each building characteristic. The case study also highlighted the varying seismic performance of buildings located in close proximity with similar seismic demand.

Development of a machine learning model for building damage prediction in the Roma and Condesa neighbourhoods - Mexico City , Mexico

This chapter presents the development of a seismic damage prediction model for the Roma and Condesa neighborhoods using machine learning techniques. It details a framework suitable for working with future post-earthquake observation data. The machine learning model uses building damage information collected following the 2017 Puebla, Mexico earthquake event and structural characteristics from 237 buildings located in the Roma and Condesa neighbourhoods in Mexico City. These neighbourhoods are of particular interest due to the availability of seismic records captured by nearby recording stations, and detailed historic information from when the neighbourhoods were affected by the 1985 Michoacán earthquake.

This chapter investigates four algorithms for machine learning classification. Random forest was found to be the best performing algorithm, achieving a 67% prediction accuracy. The study of feature importance for the random forest reveals that the building

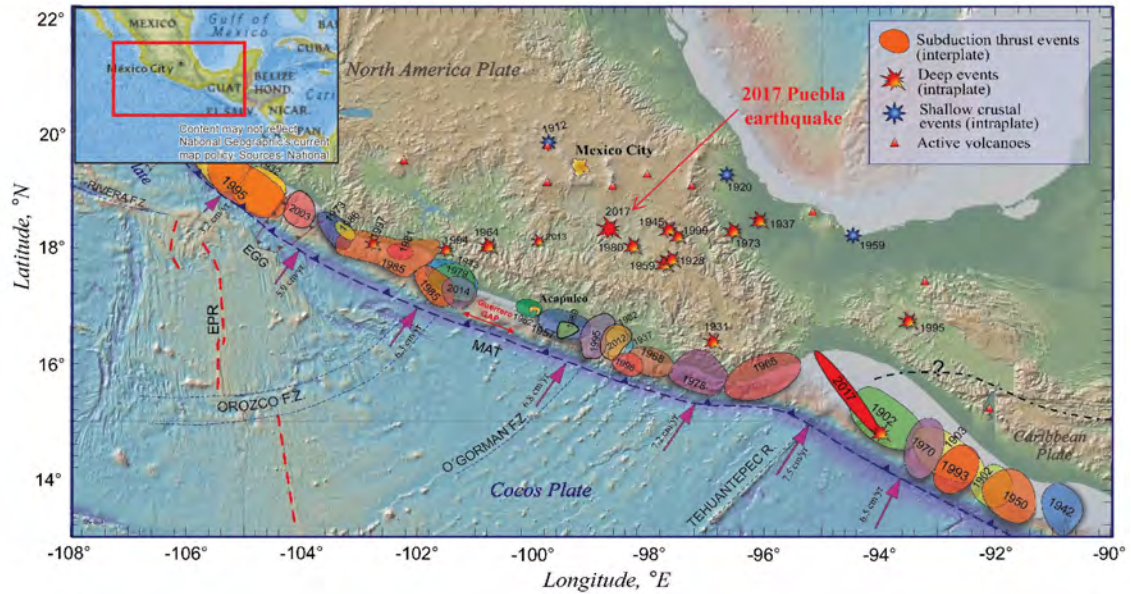


Figure 4.1: Localisation of historical seismic events in Mexico (Servicio Sismológico Nacional (SSN), 2017)

location, seismic demand and building height, in descending order, influence a building post-earthquake outcome the most.

4.1 Introduction

On Tuesday, 19 September 2017, at 1:14 pm local time (18:14:38 UTC) a magnitude M_w 7.1 earthquake struck the central part of Mexico (United States Geological Survey (USGS), 2017). The epicenter of this intraplate earthquake was located approximately 120 km away from downtown Mexico City (Jaimes, 2017) as shown in Figure 4.1. The combination of a strong and deep earthquake at a moderate distance with a soft soil basin led to severely damaged building in Mexico City (Mayoral, Asimaki, et al., 2019). The 2017 Puebla Mexico earthquake occurred on the exact anniversary of the 1985 Michoacán earthquake. On 19 September 2017, the earthquake early warning alarm went off twice, once for the annual drill at 11:00 am and the second for the earthquake warning at 1:14 pm local time (R. Allen et al., 2017).

Following the earthquake, locals and international teams assessed buildings damage across Mexico City (Colegio de Ingenieros Civiles de México (CICM), 2017b; Díaz et al., 2017; Galvis et al., 2017; Roeslin et al., 2020; Roeslin et al., 2018; Weiser et al., 2017).

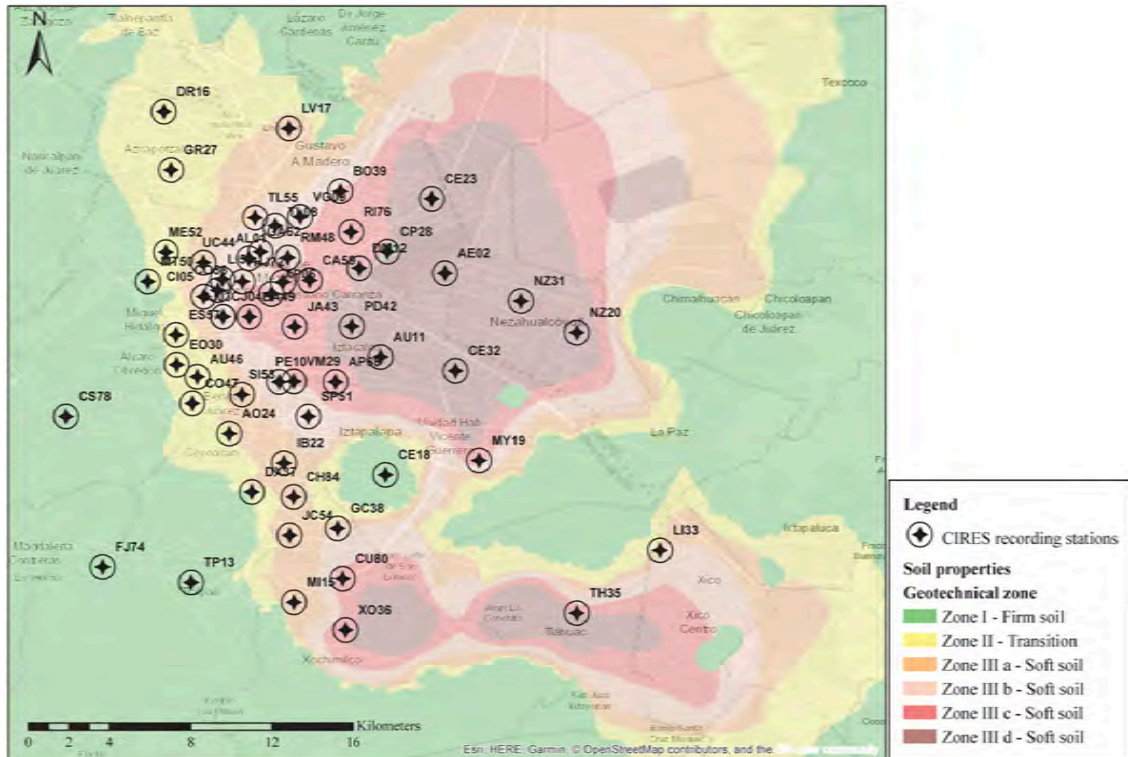


Figure 4.2: Location of the CIREs recording stations (Centro de Instrumentación y Registro Sísmico (CIREs), 2017) over a map of Mexico City

4.2 Soil characteristics and seismic recording in Mexico

In 1985 prior to the Michoacán earthquake, there were nine recording stations operating in Mexico City (Gomez-Bernal & Saragoni, 2002). The Mexican government has since expanded and upgraded the strong motion recording network, and at the time of the 2017 Puebla earthquake, the Mexican Instrumentation and Seismic Record Center operated 61 stations (Figure 4.2). The raw data of the accelerograms are publicly available online (Centro de Instrumentación y Registro Sísmico (CIREs), 2017).

Mexico City is well known for its basin site effect due to its geological formation origins (Mayoral, Asimaki, et al., 2019). A summative source of information on the local soil condition is the local building code. The code entails six different geotechnical zones based on the characteristics: Firm Zone (Zone I), Transition Zone (Zone II), Lakebed Zone (Zones III a, b, c and d) (Gobierno del Distrito Federal Mexico, 2004).

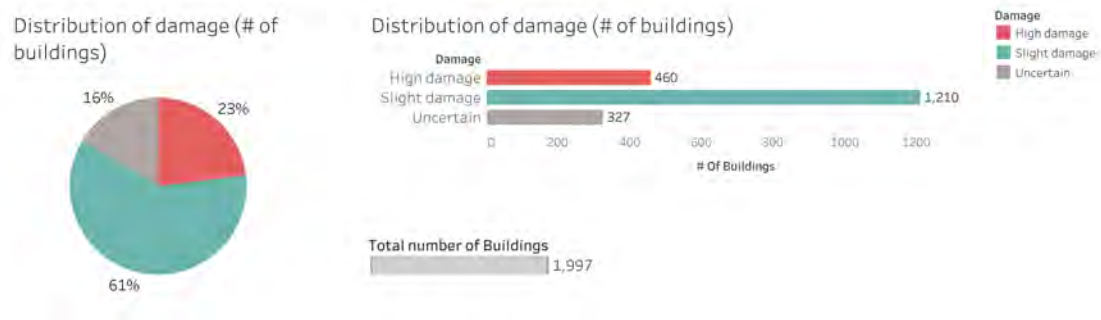


Figure 4.3: Preliminary classification of inspected structures according to the level of damage, after (Colegio de Ingenieros Civiles de México (CICM), 2017a)

4.3 Building damage and damage distribution

The 19 September 2017 Puebla earthquake induced significant building damage in Mexico City and led to the collapse of 46 buildings (Galvis et al., 2017). Engineers from the Colegio de Ingenieros Civiles de Mexico (CICM) assessed 1,997 buildings throughout Mexico City. As shown in Figure 4.3, 23% of the buildings presented a high risk due to structural damage (Colegio de Ingenieros Civiles de México (CICM), 2017a). EERI's Building Damage Sampling Team (BDST) conducted damage survey for 713 damaged and undamaged buildings, located around recording stations. Statistics show that severe damage was concentrated in buildings having 4 to 8 stories above ground. Buildings with concrete frame and masonry infill were the most affected. The BDST observed significant variations in the damage state of buildings subjected to the same seismic demand thus emphasizing the influence of the building structural and geometric characteristics (Weiser et al., 2017). The Puebla-Morelos earthquake also pointed the deficiencies of buildings constructed prior to the 1987 change in design code (Colegio de Ingenieros Civiles de México (CICM), 2017a; Weiser et al., 2017). Different ground conditions and soil periods led to significant variations in the damage state of buildings throughout Mexico City. Building damage was characterized by the importance of local site effects which led to an increase in the seismic demand for building with structural periods between 0.8s and 1.6s (4 to 8 stories in height) (Mayoral, Asimaki, et al., 2019).

In addition to the broad building surveys by the local authority, an university UAM Azcapotzalco, Mexico City team conducted detailed block-by-block building damage surveys of over 300 buildings (237 within Roma and Condesa neighbourhoods), with level of damage recorded against the European Macroseismic scale EMS-98 (Grünthal,

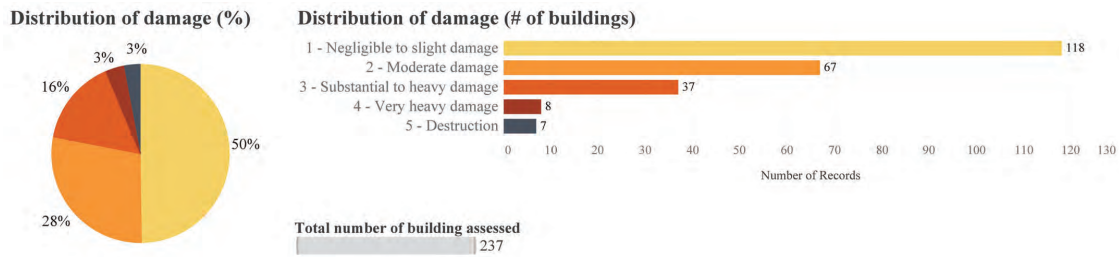


Figure 4.4: Damage severity of the surveyed buildings located in the Roma and Condesa neighborhoods

1998). These surveys collected detailed building data based on a streamlined taxonomy specific to the local building stock. Thus to enable comparison and potential combination to observations from other events worldwide, the raw building information were converted and aligned to the GEM Building Taxonomy v2.0 (Brzev et al., 2013; Roeslin et al., 2018). The GEM Building Taxonomy v2.0 was chosen as it is flexible and has been designed to accommodate local variations sensitively worldwide (L. Allen et al., 2015).

Figures 4.4 and 4.5 present damage statistics generated using the data collected by the UAM team in the Roma and Condesa neighborhoods. Figure 4.4 shows that 50% of the building assessed experienced slight damage, 28% moderate damage, and 22% substantial damage and above. Figure 4.5a shows buildings categorized by the number of storeys. It can be seen that 88% of the buildings assessed were lower than ten storeys. Figure 4.5b gives the building year. Unfortunately, the construction year is unknown for almost two-fifth of the buildings. Nevertheless, for the 144 buildings with information on the date of construction or retrofit, 94% were constructed before 1985. Figure 4.5c and 4.5d present the buildings categorized by material type and lateral load resisting system (LLRS), respectively. Most of the buildings assessed had reinforced concrete main structural system, but similarly as it was described in chapter 3, information on LLRS is difficult to establish and in this case it was missing for most than half of the buildings assessed.

Figure 4.6 shows a map of the surveyed buildings in the Mexico City urban area. The survey focused on the Roma and Condesa neighborhoods (Figure 4.7a and Figure 4.7b), as firstly, building damage was densely concentrated in these areas, and secondly these areas had been subjected to ongoing work and scrutiny since 1985 (Arellano-Mendez et al., 2004)). The new data would enable verification of existing seismic vulnerability

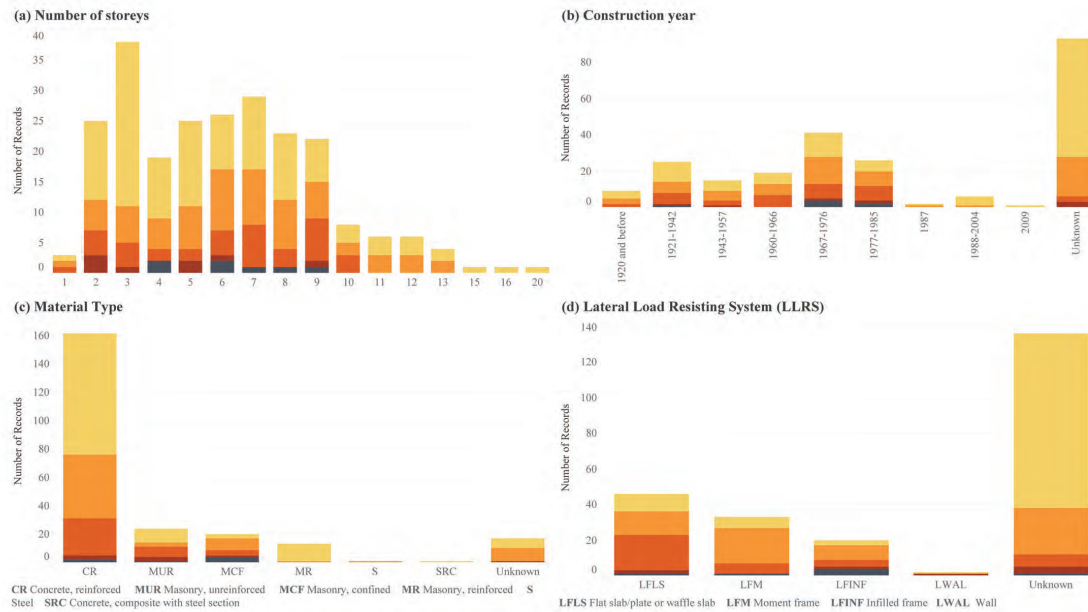


Figure 4.5: Statistics on the UAM building damage database: categorised by (a) number of storeys, (b) construction year, (c) material type, (d) lateral load resisting system

assessments, enable comparison on the efficacy of various policy interventions, and provide valuable data on the effectiveness of different seismic retrofit techniques.

4.4 Machine learning model development

The approach for the development of a machine learning damage prediction model for this study follows the framework suggested by Géron (2019), as depicted in Figure 2.20 in section 2.6.5.

4.4.1 Problem framing

The objective is to develop a machine learning model to estimate seismic building damage in the Roma and Condesa neighbourhoods in Mexico City. The machine learning model is trained on merged data, including building damage obtained from post-event observations and seismic demand from recording stations, making the model development a supervised learning problem.

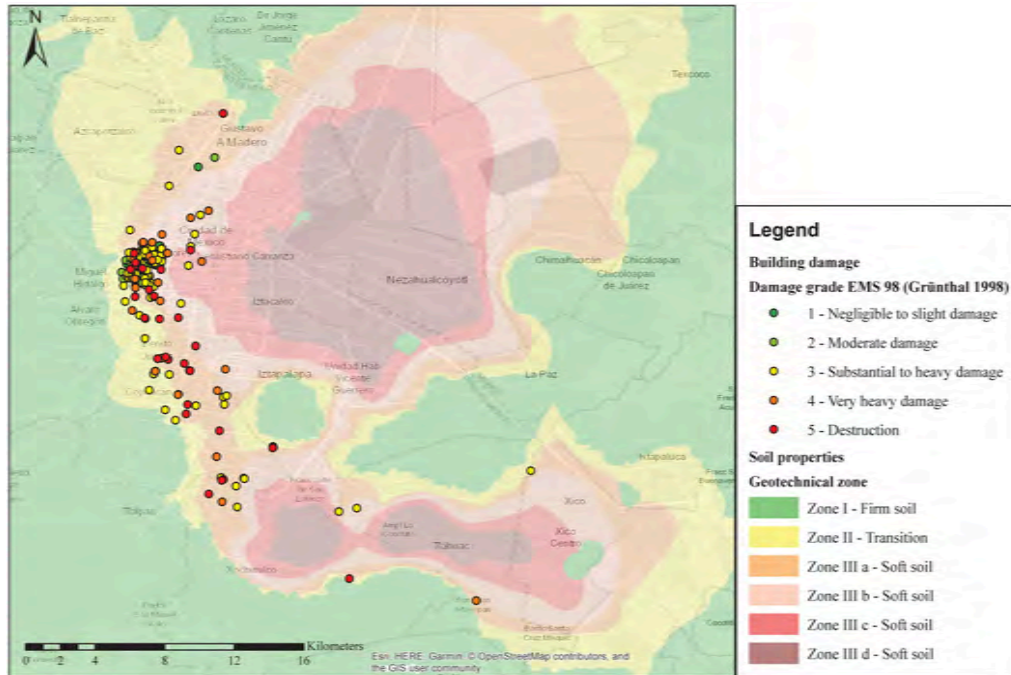
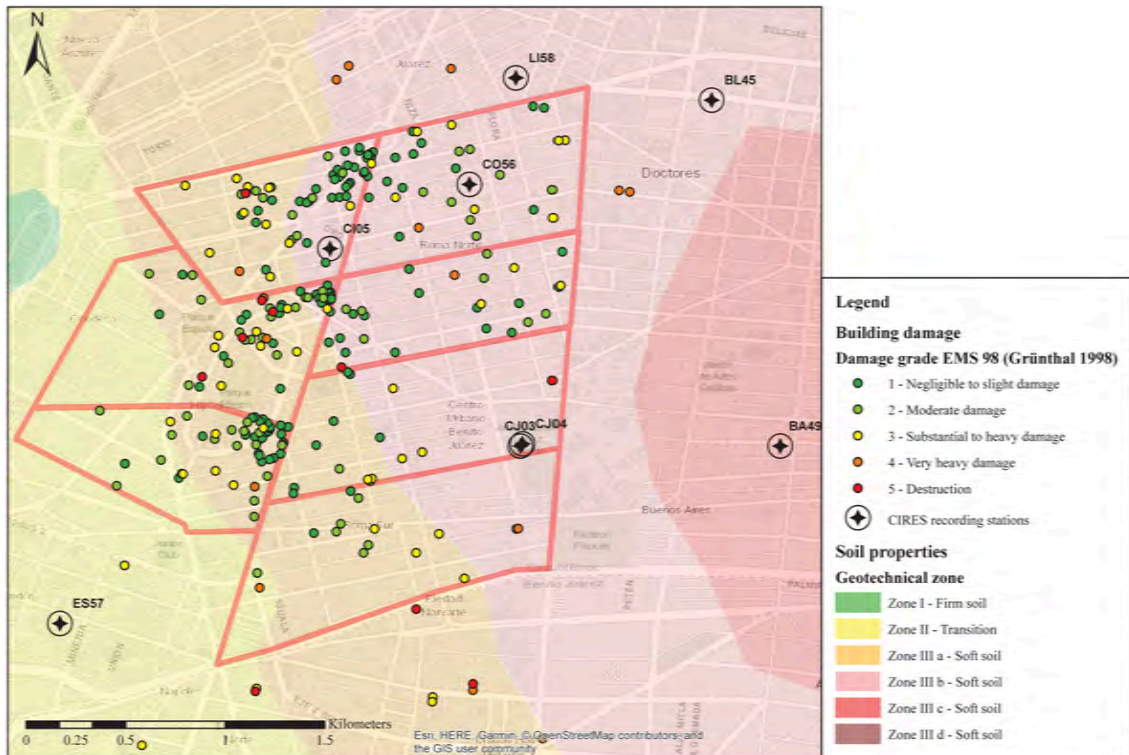


Figure 4.6: Locations of the buildings assessed by the UAM team superimposed over a map of the geotechnical zones of Mexico city

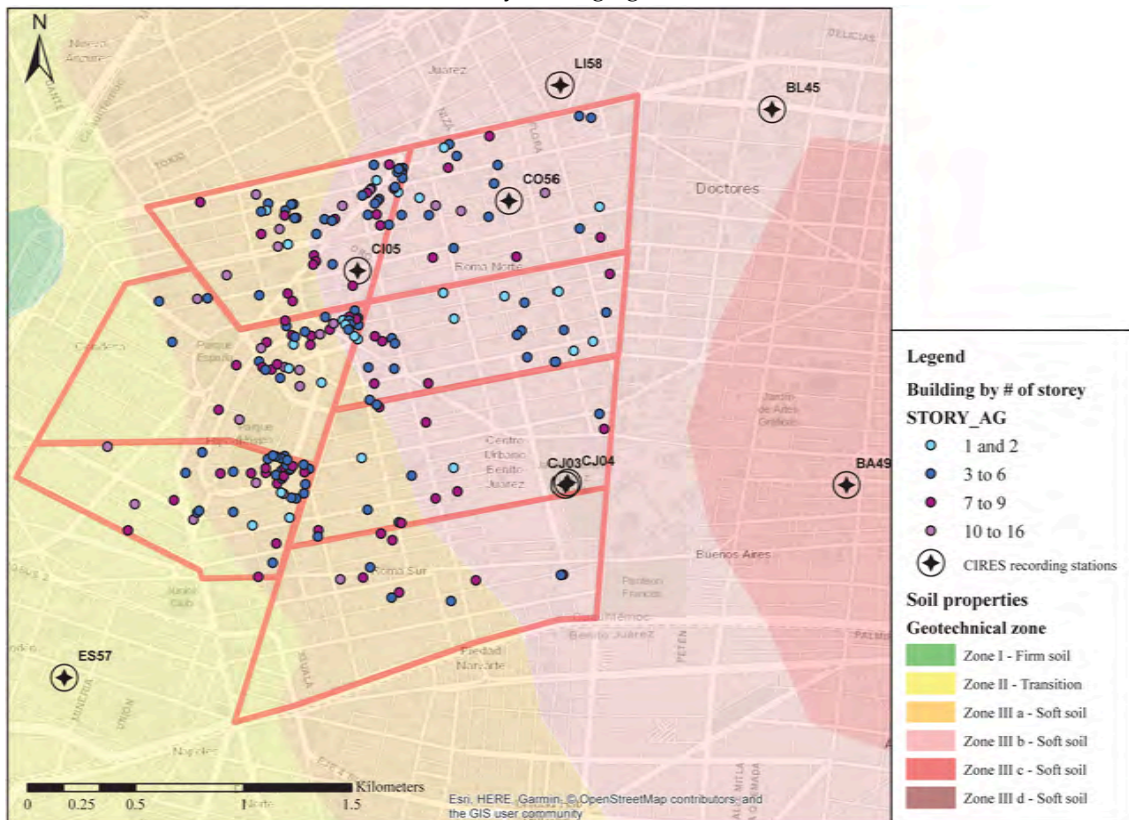
4.4.2 Data collection

Before the development of the machine learning model, it is essential to understand and be familiar with the data available. Figure 4.4 presents an overview of the damage distribution for the 237 buildings located in the Roma and Condesa neighbourhoods. The raw database includes the following features:

- the building location (latitude and longitude),
- the damage state of the building according EMS-98,
- the construction material,
- the type of lateral load resisting system,
- the floor system type,
- the number of stories above ground,
- the date of construction or retrofit, and
- information about the presence of any plan irregularity.



(a) By damage grade



(b) By number of stories

Figure 4.7: Buildings assessed by the UAM team in the Roma and Condesa neighbourhoods

Table 4.1: PGA and max spectral acceleration for the CIRES stations

Station	Geotech. zone	max S_A N00E [g]	max S_A N90W [g]	PGA N00E [g]	PGA N90W [g]	Average PGA [g]
BA49	IIIc	0.653	0.436	0.091	0.116	0.103
BL45	IIIb	0.335	0.394	0.105	0.117	0.111
CI05	IIIb	0.486	0.475	0.116	0.117	0.116
CJ03	IIIb	0.471	0.574	0.114	0.100	0.107
CJ04	IIIb	0.464	0.568	0.127	0.099	0.113
CO56	IIIb	0.453	0.337	0.112	0.117	0.114
ES57	IIIa	0.306	0.335	0.072	0.086	0.079
LI58	IIIb	0.367	0.417	0.098	0.092	0.095
MT50	I	0.188	0.270	0.048	0.060	0.054
UC44	IIIa	0.396	0.489	0.128	0.128	0.128
AL01	IIIb	0.361	0.374	0.120	0.111	0.115

However, not all features are always recorded for each data entry and procedures to address these are explained in section 4.4.4 on data preparation.

A key input for the machine learning model is the seismic demand that each building experienced leading to the damage observed. For the model development, this study utilises spectral acceleration as the main parameter to characterize the seismic demand on buildings. Nevertheless, PGA is included as well so to ascertain its relevance and importance in the final machine learning model. Table 4.1 shows the PGA and max S_A values recorded at the CIRES strong motion stations located near the area of interest. As the orientation of the main resisting building frames is unknown or is not recorded, it is not possible to identify the governing earthquake component. Thus, the average of the north-south (N00E) and East-West (N90W) component PGA are adopted for this study. Figure 4.8 shows PGA values for the area of interest. The values have been derived using an inverse distance weighted (IDW) interpolation. It can be seen that that PGA is highest around station CI05.

From a structural dynamic standpoint, the seismic demand depends on the interactions of a building's dynamic characteristics, the site condition, and the magnitude

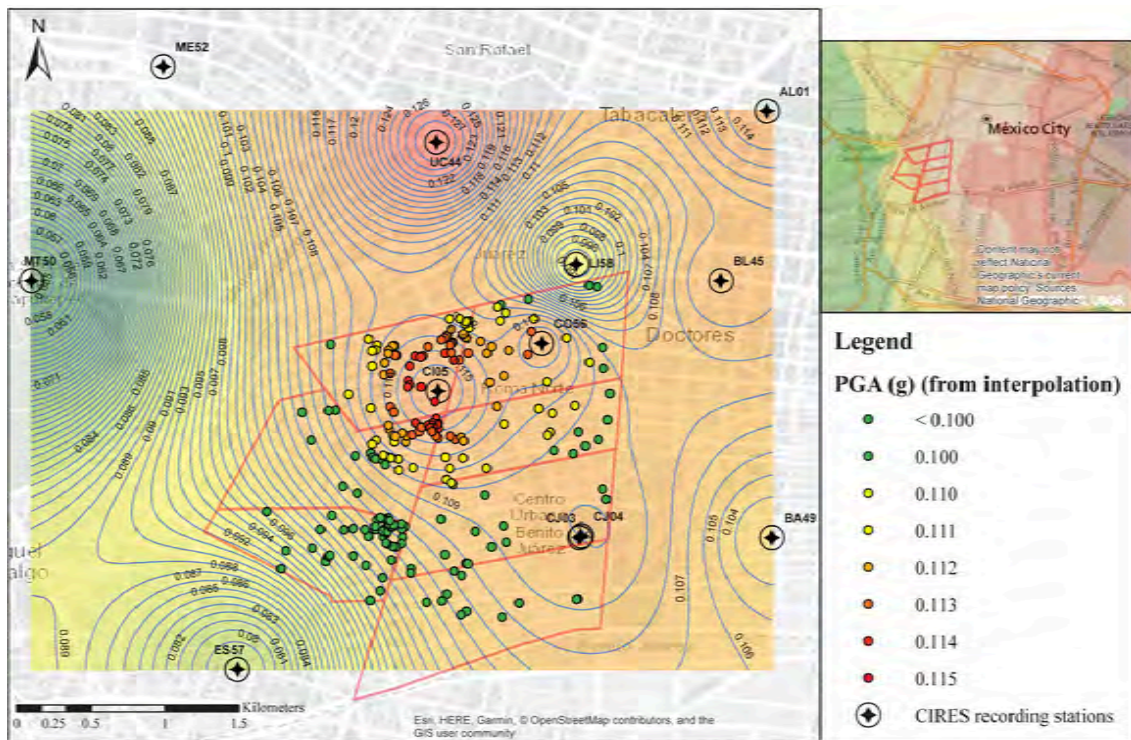


Figure 4.8: Inverse distance weighted (IDW) interpolation of the PGA values between the CIRES recording stations

and frequency content of the ground motion. A simple approach to account for this is to characterize the ground motion by its response spectrum, and quantifying the interaction by a building's corresponding first mode spectral acceleration.

To achieve this for the dataset, the raw accelerograms from CIRES were filtered and processed to obtain the response spectrum for each ground motion station. To estimate the dynamic characteristics for the 237 buildings in the Roma and Condesa neighbourhoods, the buildings' fundamental natural periods were estimated using the empirical formula developed by Muria-Vila and Gonzalez-Alcorta (1995) for the Mexican building stock. This is shown in Equation 4.1. This simple equation only requires the number of stories and the building structural system type, both of which are available from the post-earthquake detailed building surveys.

$$T_1 = a * N \quad (4.1)$$

where: T_1 is the fundamental period of vibration of the building

a is a coefficient taken from Table 4.3

N is the number of storeys above ground.

Table 4.3: Coefficient a for building period T_1 in Mexico City, after (Muria-Vila & Gonzalez-Alcorta, 1995)

Building type	Firm soil	Soft soil
Frames	0.100	0.126
Frames and walls	0.063	0.102
Masonry	0.040	0.073

An illustrative example of this process for a 12 storeys building is shown in Figure 4.9. It was not possible to define the natural period for 19 buildings in the dataset due to the insufficient information regarding their structural system and material.

4.4.3 Data exploration

After the inclusion of additional information from other databases as explained in the previous section, it results in fifteen features for model development as shown in Table 4.5.

4.4.4 Data preparation

The UAM team assessed building damage according to the EMS-98 five-step scale, from negligible to slight damage (category 1) to destruction (category 5). Indicative guidance on the interpretation is shown in Figure 4.10 as it appeared in the data collection form provided to the assessors. Figure 4.11a shows that the upper classes occur infrequently leading to a class imbalance and difficulties for the machine learning model to accurately learn for classes 2 to 5 as the number of occurrence in each class is very limited. First versions of the model aiming to predict five damage classes led to unsatisfactory model performance. Thus, the data was pre-processed with categories 2, 3, 4, and 5 combined in a common category. This left the damage prediction as a binary target where "0" represents negligible to slight damage and "1" moderate damage and above (Figure 4.11b).

Figure 4.12 shows the pair-plots between selected numerical variables. From the plots, it can be seen that the the natural period is predictably highly linearly correlated

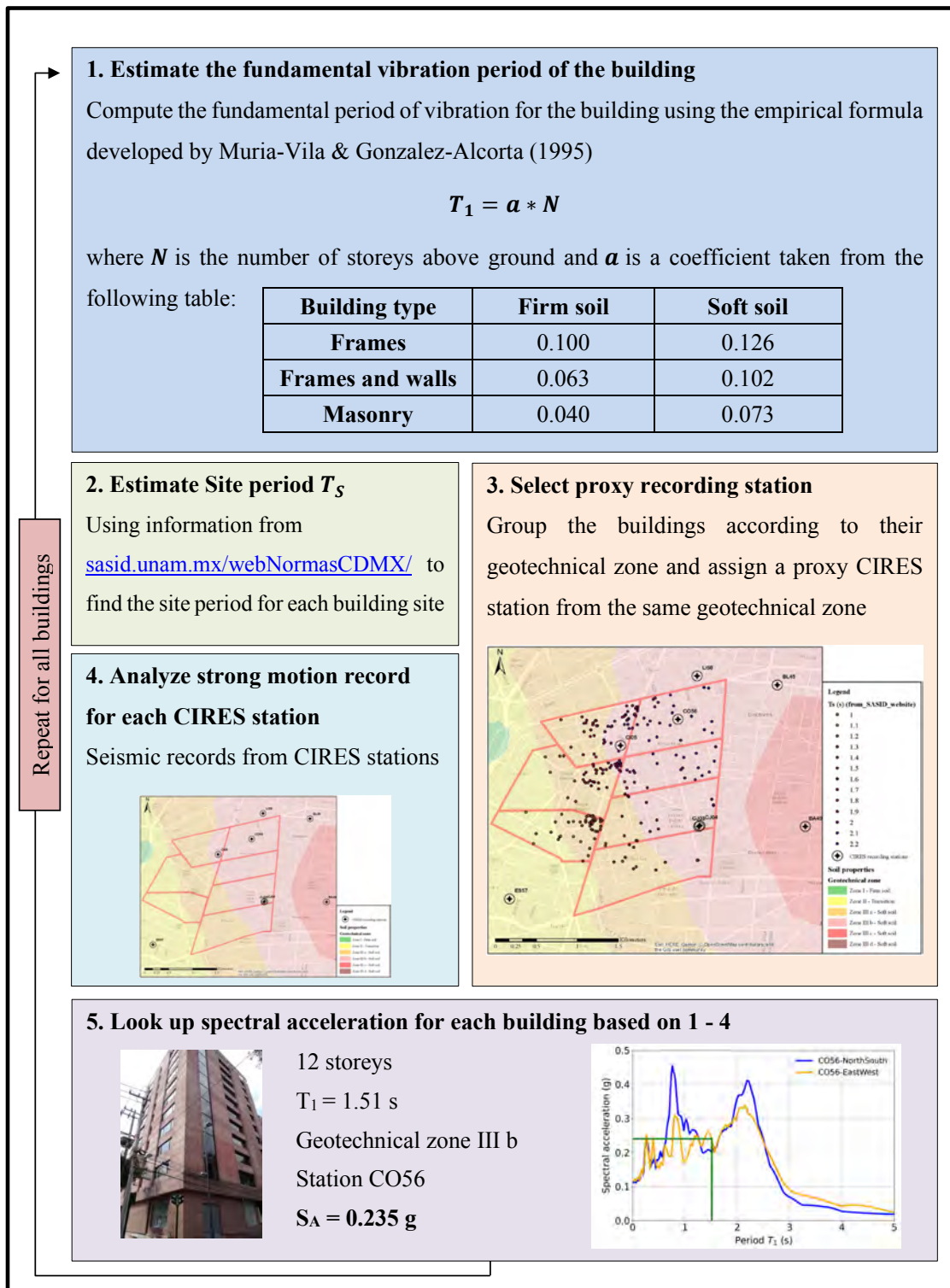


Figure 4.9: Steps to derive the seismic demand for each building

Table 4.5: Features present in the final dataset (after addition of supplementary information)

Feature name	Feature	Source	Feature type
Building	Building address	Collected on site	Text
Latitude	Building latitude	Extracted from building address	Numerical
Longitude	Building longitude	Extracted from building address	Numerical
Damage EMS98	Damage EMS 98	Assessed on site	Categorical
MAT TYPE	Material type	Observed on site	Categorical
LLRS	Type of lateral load-resisting system	Observed on site	Categorical
FLOOR TYPE	Floor system type	Observed on site	Categorical
STORY AG	Number of storeys above ground	Observed on site	Numerical
YR BUILT	Date of construction or Retrofit	Observed on site	Numerical
STR HZIR P	Plan irregularity	Observed on site	Binary
Geotech Zone	Geotechnical zone	Extracted from 2004 Mexico design codes	Categorical
Natural period (calculated)	Building natural period (s)	Calculated based on Equation 4.1	Numerical
Ts (from SASID website)	Site period	Obtained from website	Numerical
PGA (interpolated)	Peak ground acceleration	Interpolated, based on CIRES data	Numerical
S_A (from similar station)	S _A (T ₁) Spectral acceleration of T ₁ for the structure	Obtained from a CIRES station	Numerical

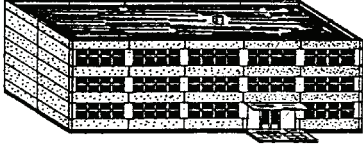
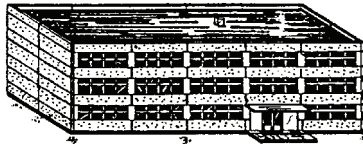

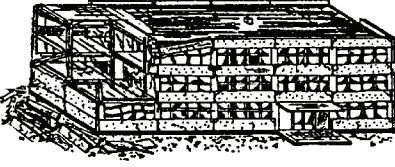

Classification of damage to buildings of reinforced concrete	
	<p>Grade 1: Negligible to slight damage (no structural damage, slight non-structural damage)</p> <p>Fine cracks in plaster over frame members or in walls at the base. Fine cracks in partitions and infills.</p>
	<p>Grade 2: Moderate damage (slight structural damage, moderate non-structural damage)</p> <p>Cracks in columns and beams of frames and in structural walls. Cracks in partition and infill walls; fall of brittle cladding and plaster. Falling mortar from the joints of wall panels.</p>
	<p>Grade 3: Substantial to heavy damage (moderate structural damage, heavy non-structural damage)</p> <p>Cracks in columns and beam column joints of frames at the base and at joints of coupled walls. Spalling of concrete cover, buckling of reinforced rods. Large cracks in partition and infill walls, failure of individual infill panels.</p>
	<p>Grade 4: Very heavy damage (heavy structural damage, very heavy non-structural damage)</p> <p>Large cracks in structural elements with compression failure of concrete and fracture of rebars; bond failure of beam reinforced bars; tilting of columns. Collapse of a few columns or of a single upper floor.</p>
	<p>Grade 5: Destruction (very heavy structural damage)</p> <p>Collapse of ground floor or parts (e. g. wings) of buildings.</p>

Figure 4.10: Detailed description of the damage grade (Grünthal, 1998)

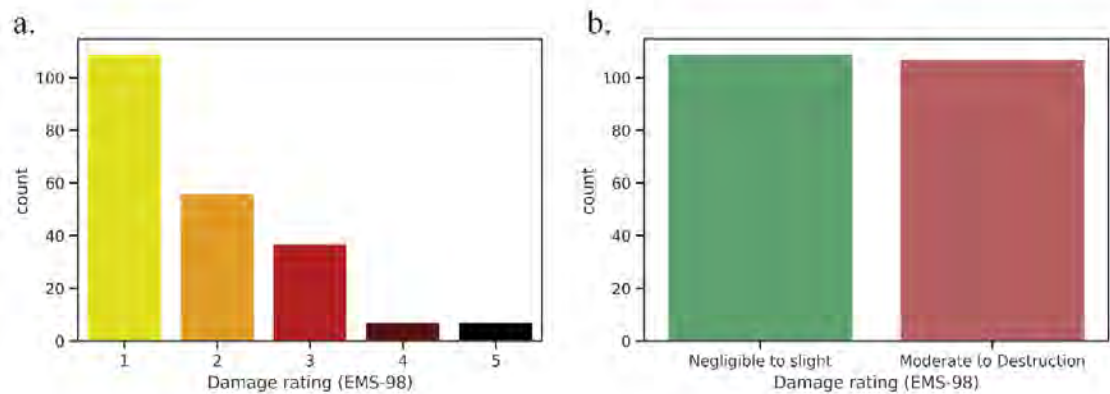


Figure 4.11: (a) Number of training data point available as five damage classes; (b) Data points distribution after transforming target feature as a binary damage class

with the number of stories. This correlation is confirmed by a high Pearson correlation coefficient as shown in Figure 4.13. Thus, the feature 'Natural period(calculated)' is removed as justified in section 2.7.2.

An important step of data exploration is also to review the quality of the data, especially missing values. Figure 4.14 highlights the availability of the different features across the 216 buildings. Table 4.6 presents a detailed overview of the available instances and missing values for each variable. Upon scrutinizing the result, 115 (53%) buildings do not have Lateral Load Resisting System (LLRS) data, 79 buildings (37%) are missing construction year (YR BUILT) data, and 67 buildings (31%) are missing floor type (FLOOR TYPE) data. Thus, with over 50% of the dataset missing the LLRS values, a decision was made to remove it from the model.

Common feature engineering methodologies to fill-up missing construction year and type of floors were trialled. This included back-fill or forward-fill, replacement with the mean, use of the median or mode, and even the training of a machine learning model (k-nearest neighbours). If none of the solutions provides a satisfactory output, the instances with missing value can be taken out if sufficient data remain available in the database. A machine learning model using k-nearest neighbours was trained to fill-up missing values for the construction year. And missing building year information could be guessed through expert knowledge. However, none of the fill-up techniques improved the overall model prediction accuracy, and it risks introducing bias in the dataset. Consequently, construction year and type of floors were removed from the model

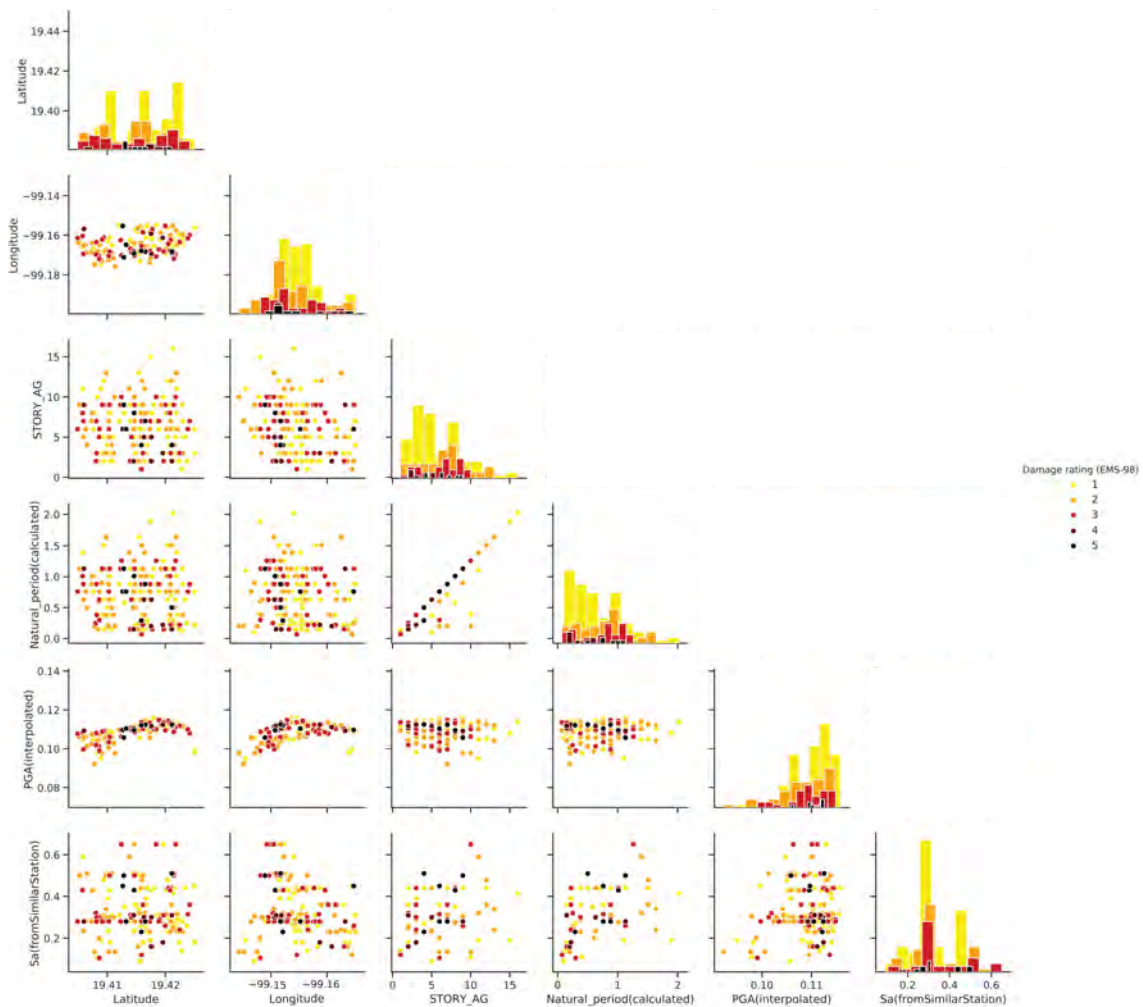


Figure 4.12: Pair plots showing the relationship between the variables: number of stories, natural period, PGA, and S_A . The hue represents the damage grade

in favour of maintaining the integrity of the modelling process.

Our processed database ultimately had five numerical features: 'Latitude', 'Longitude' 'Damage EMS 98 (target feature)', 'STORY AG', ' S_A ' and three categorical features: 'MAT TYPE', 'STR HZIR P', 'Geotech Zone'. One-hot encoding was applied to the three categorical features. This led to four binary variables/columns for material type (concrete, masonry unreinforced, masonry confined, masonry reinforced), two for torsion eccentricity (no torsion, torsion) and three for the geotechnical zone (Zone II, Zone III a, Zone III b). The results are shown graphically in Figure 4.15. However, converting n categories to n binary columns leads to unnecessary redundant information, in other words, the state of the final column can be inferred by the values of the other columns in the same grouping. Thus, one category per feature was removed leaving n-1 columns to improve performance and the prediction accuracy of the algorithm.

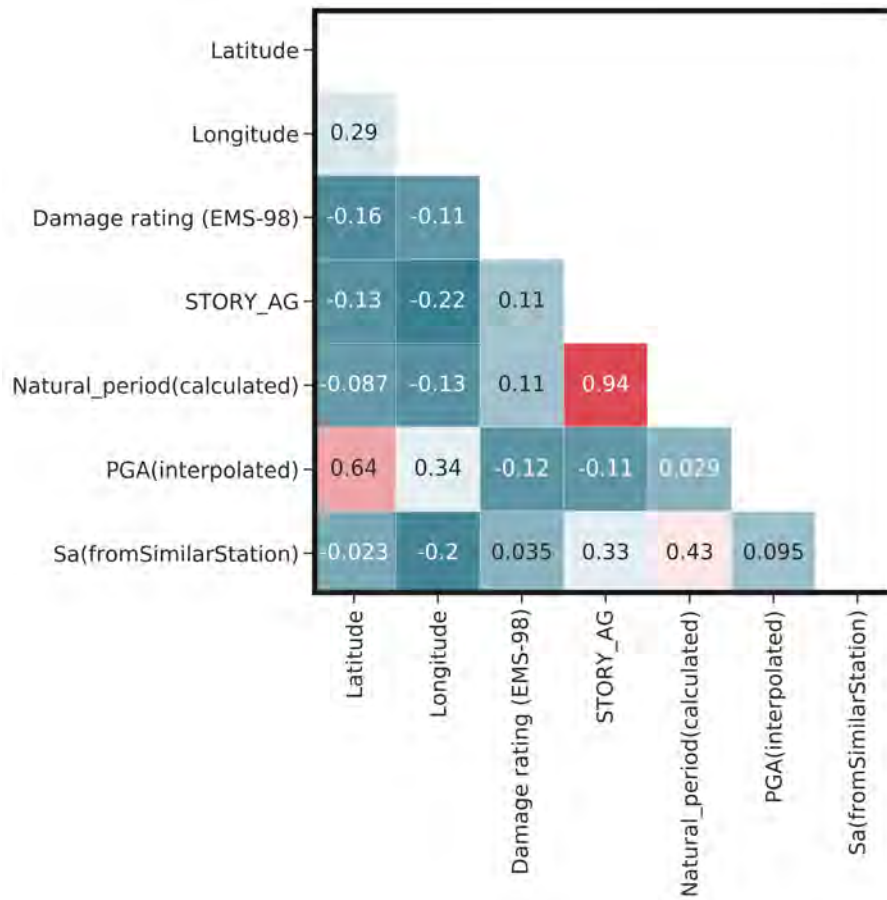


Figure 4.13: Pearson correlation coefficient before pre-processing of the database

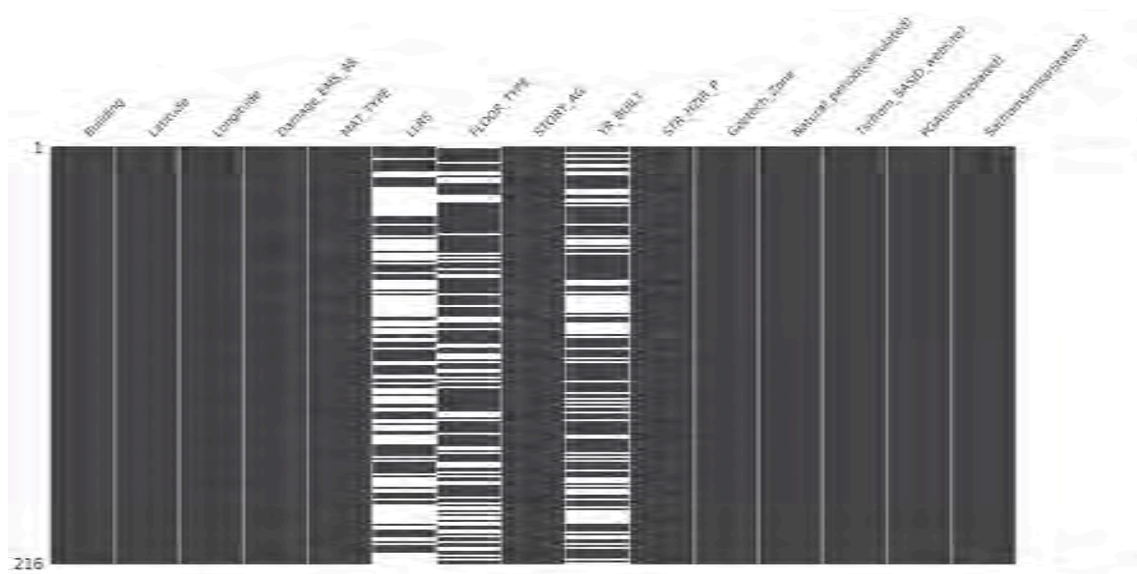


Figure 4.14: Graphical representation of missing values (on the raw database)

Table 4.6: Number of values available for each feature

Features	# of records	# of missing values	% of missing values
Building address	216	0	0%
Building latitude	216	0	0%
Building longitude	216	0	0%
Damage EMS98	216	0	0%
Material type	216	0	0%
Type of lateral load-resisting system	101	115	53%
Floor system type	149	67	31%
Number of storeys above ground	216	0	0%
Date of construction or Retrofit	137	79	37%
Plan irregularity	216	0	0%
Geotechnical zone	216	0	0%
Peak ground acceleration	216	0	0%
Seismic demand – Spectral acceleration	216	0	0%

	0	1	2	3	4	5	6	7	8	9
MAT_TYPE_CR	1	1	0	0	0	1	1	1	1	1
MAT_TYPE_MCF	0	0	0	0	0	0	0	0	0	0
MAT_TYPE_MR	0	0	1	0	0	0	0	0	0	0
MAT_TYPE_MUR	0	0	0	1	1	0	0	0	0	0
STR_HZIR_P_No	1	0	1	1	1	1	1	1	1	0
STR_HZIR_P_TOR	0	1	0	0	0	0	0	0	0	1
Geotech_Zone_II	0	0	0	0	0	0	0	0	0	0
Geotech_Zone_IIIa	0	0	0	0	0	0	0	1	0	0
Geotech_Zone_IIIb	1	1	1	1	1	1	1	0	1	1

Figure 4.15: Overview of categorical features after one-hot encoding

Before training a model, it is important to split the data and leave a part of the data untouched that can later be used to evaluate the prediction accuracy of the model on previously unseen data. Two sets, a training set and a testing set were created. The database of 216 buildings was split in 75% / 25% thus having 162 and 54 in the training and testing set respectively.

4.4.5 Model selection and training

For this study, the aim of the machine learning model is to predict possible damage for new input data (features). The collected building damage data entails both the features

(building and seismic characteristics) as well as the target (building damage), thus leading itself as a supervised learning problem. Logistic regression, SVM, decision trees, and random forests are algorithms that are suitable to perform supervised classification tasks. The process of the model fitting to the training set was realised using built-in function from scikit-learn (Cournapeau, 2007).

While the target prediction is important, the ‘decision process’ or path taken by the algorithm to classify the damage state of the building is also of interest. This favoured logistic regression and decision tree as they are intrinsic interpretability. SVM and random forest algorithms, while not intrinsically interpretable, were nonetheless trialled. Random forest algorithm was eventually chosen as the algorithm for this study as it provided the best predictive performance. To allow for human model interpretation and the look-through characteristics, we applied post-hoc interpretation methods to extract feature importance.

4.5 Model prediction

4.5.1 Prediction performance

Table 4.7 presents the prediction performance in terms of precision, recall, F_1 score, and overall model accuracy for the four algorithms trialled. Random forest model delivered the best prediction accuracy followed by decision tree. The random forest model achieved an F_1 score of 0.65 for the category 0 and 0.68 for the category 1, with a precision of 0.68 and 0.66 and a recall of 0.63 and 0.70 for the category 0 and 1 respectively. This means that when the model claims a building is category 0, it is correct 68% of the time and it correctly predicts 63% of the building within category 0. Figure 4.16 shows the confusion matrices for the model using random forest.

4.6 Feature importance Random Forest algorithm

Applying post-hoc methods to the most promising random forest model allowed the relative influence of different input features to be evaluated. The feature importance can be analysed using Shapley values. Figure 4.17 shows the SHAP feature importance of random forest computed on the test set. Each row represents a variable according to the

Table 4.7: Damage prediction accuracy for logistic regression, support vector machine, decision trees, and random forest models

Algorithm	Set	Prediction targets	Precision	Recall	F ₁ score	Accuracy
Logistic regression	Train Set	Category 0	0.69	0.68	0.69	0.69
		Category 1	0.68	0.69	0.68	
		Accuracy on the train set				
	Test Set	Category 0	0.67	0.59	0.63	0.65
		Category 1	0.63	0.70	0.67	
		Accuracy on the test set				
Support vector machine (SVM)	Train Set	Category 0	0.72	0.87	0.78	0.76
		Category 1	0.83	0.65	0.73	
		Accuracy on the train set				
	Test Set	Category 0	0.59	0.74	0.66	0.61
		Category 1	0.65	0.48	0.55	
		Accuracy on the test set				
Decision tree	Train Set	Category 0	0.87	1.00	0.93	0.93
		Category 1	1.00	0.85	0.92	
		Accuracy on the train set				
	Test Set	Category 0	0.64	0.78	0.70	0.67
		Category 1	0.71	0.56	0.63	
		Accuracy on the test set				
Random forest	Train Set	Category 0	1.00	1.00	1.00	1.00
		Category 1	1.00	1.00	1.00	
		Accuracy on the train set				
	Test Set	Category 0	0.71	0.56	0.63	0.67
		Category 1	0.66	0.78	0.70	
		Accuracy on the test set				

feature importance in the random forest model. They are classified in a decreasing order such that the most important feature appears at the top.

For this case study, the three main features influencing model decision are the longitude (east-west location of the building), the latitude (north-south location of the building), and PGA (interpolated). From an engineering point of view, it might be evident that the longitude and latitude stand out first as seismic demand varies depending on the building location. Previous studies pointed out the influence of site effects in Mexico City Mayoral, Roman, et al. (2019) and Mayoral et al. (2017) reported the significant spatial variability of ground motions recordings in the east-west direction for the area

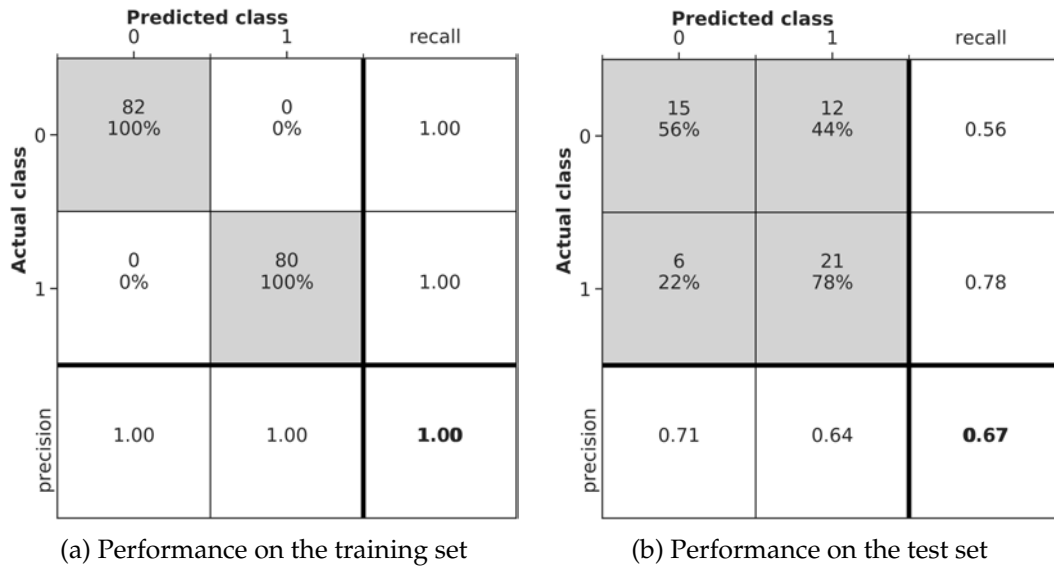


Figure 4.16: Performance of the Random Forest (RF) algorithm

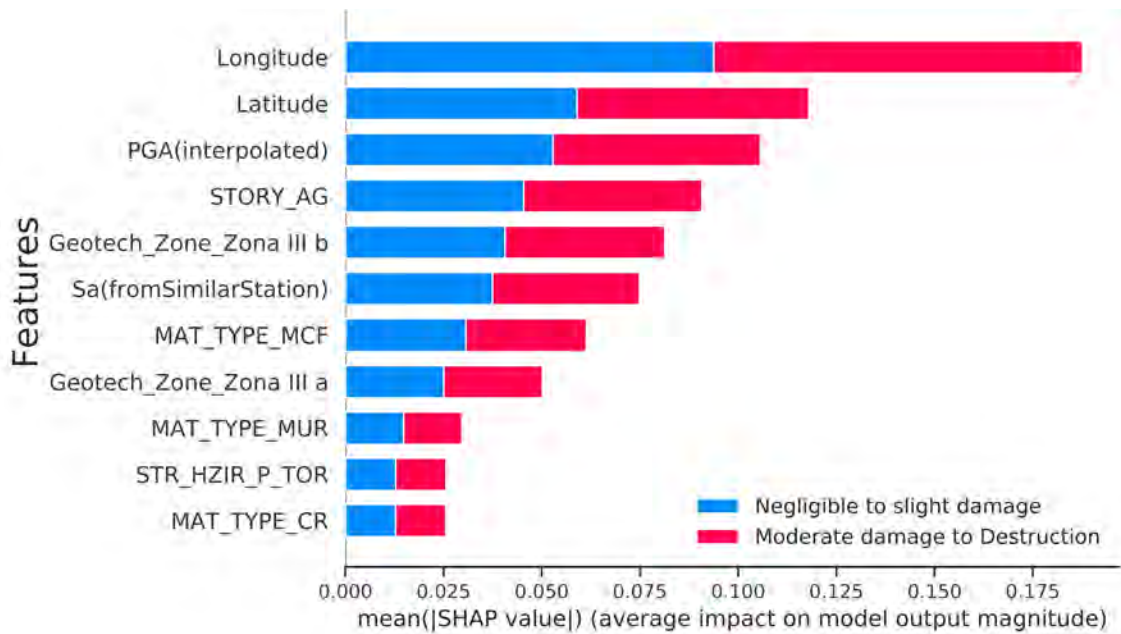


Figure 4.17: Feature importance based on Shapley value

encompassing the Roma and Condesa neighbourhoods. The feature importance results reported here thus correspond with the high variability of conditions in the east-west direction for the Roma and Condesa neighbourhoods. Future models encompassing soil condition as a variable may provide further insights on the east-west variability of the studied area.

4.7 Conclusion

This chapter proposed a new analysis technique of analysing and utilising post-earthquake damage survey data by applying machine learning. This was successfully applied to the detailed building survey data following the 2017 Puebla earthquake in the Roma and Condesa neighborhoods in Mexico City. The final random forest model was 67% accurate in predicting damage to the building stock. The ML model enabled key building features contributing to damage to be assessed through the post-hoc method, which can also be useful in aiding future policy development for seismic risk mitigation. Furthermore, the machine learning model can be readily further applied to develop rapid regional building performance estimates, for this event and potentially for future events. This case study highlighted the importance of consistent input data and data pre-processing. It also demonstrated a framework for developing machine learning models from earthquake reconnaissance data, and useful metrics for evaluating the success of the modelling. The experience showed that input data is key. A larger data set with more buildings for each of the damage class would offer new opportunities (e.g. increasing the damage prediction resolution) and significantly improve the model accuracy.

Opportunities also exist in future reconnaissance missions and future building assessments to include information on losses captured in insurance claims. This then would enable researchers to extend the methodology for the development of seismic loss prediction models.

Data integration for the development of a seismic loss prediction model using EQC's residential claims database

This chapter documents a detailed investigation into the available residential insurance claims data provided by EQC for this research. It also presents the time-consuming merging process of aggregating various information from multiple databases. The process draws heavily upon using Geographic Information System (GIS) techniques. The experience highlights that built-in functions in current off-the-shelf software do not lead to a satisfactory output. This project relies on the use of Land Information New Zealand (LINZ) spatial data as an intermediary to enable the data merging. This approach leads to a reduction of available sample point but it has been otherwise shown to be generally effective. The resulting merged data set is sufficiently large for developing a machine learning model.

5.1 Introduction

In 2010-2011, New Zealand experienced the most damaging earthquakes in its history, known as the Canterbury earthquake sequence (CES). It led to extensive damage to Christchurch buildings, infrastructure and its surroundings; affecting commercial and residential buildings. The direct economic losses represented 20% of New Zealand's GDP in 2011. Owing to New Zealand's particular insurance structure, the insurance sector contributed to over 80% of losses for a total of more than NZ\$31 billion. Over NZ\$11 billion of the losses arose from residential building claims and were covered either partially or entirely from the NZ government backed Earthquake Commission (EQC) EQcover insurance scheme. In the process of resolving the claims, EQC collected detailed financial loss data, post-event observations, and building characteristics for each of the approximately 434,000 claims lodged following the CES. This coincided with the effort by the very active NZ earthquake engineering community, which exploited the event and collected extensive data on the ground shaking levels, soil conditions, and liquefaction occurrence throughout wider Christchurch, as a large scale outdoor experiment.

5.2 Residential building loss data: EQC claims data set

Following the changes brought by the Earthquake Commission Amendment Act 2019, EQC permitted access to the claims database for research purposes only on request. This study uses the March 2019 version of the EQC claims database. Over 95% of the insurance claims for the CES have been settled by that time. However, revision of the event apportionment is still subjected to review at that time, meaning that the division of the cost between events and thus also between EQC and the private insurers is not finalised.

The EQC claims data set contains 62 data features. The data set contains the information such as the date of the event, the opening and closing date of a claim, a unique property number, and the claim amount for the building, content and land. Among the 62 variables, the database also includes building characteristics. However, not all meta-data were collected in every instance and this led to incomplete data. As shown in Figure 5.1, the original EQC database has up to 85% of the values missing

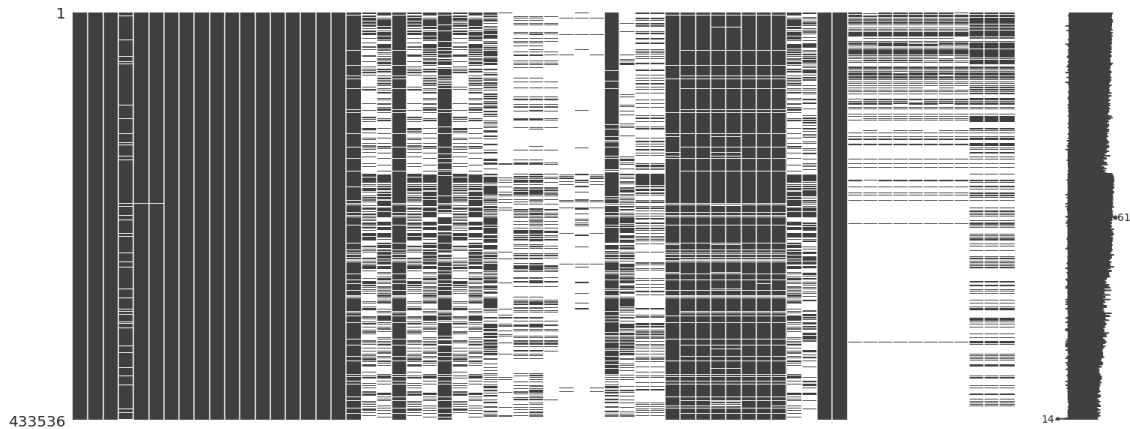


Figure 5.1: Graphical overview of the raw data in the EQC claims database for the Canterbury earthquake sequence. The columns represent attributes and the rows examples. White areas represent missing values. Column 4 represents the PortfolioID, columns 5 and 6 the longitude and latitude respectively.

for critical features regarding the building characteristics (e.g. construction year, primary construction material, number of stories). Furthermore, the building characteristics may be subjective to individual assessor’s visual observation. The scarce information for building characteristics combined with the necessity to have full data for key variables led to the need to add information from other sources.

5.3 Data collection from additional databases

To overcome missing information various alternate databases have been merged with the EQC claims data set. The following describes attributes and information available in each database.

5.3.1 Sourcing building characteristics

The RiskScape ‘New Zealand Building’ inventory data set (NIWA & GNS Science, 2015) has been adopted by this project to deliver critical information on buildings characteristics. The ‘New Zealand Building’ inventory collected building asset information for use within the RiskScape software (NIWA & GNS Science, 2017). This data set contains detailed engineering and other information for every building in New Zealand. Table 5.1 shows an overview of selected attributes available in the RiskScape database.

Table 5.1: Overview of selected features in the RiskScape data set

Attribute	Attribute categories
Longitude	NZTM coordinates (Easting)
Latitude	NZTM coordinates (Northing)
Construction Type	1: Reinforced Concrete Shear Wall; 2: Reinforced Concrete Moment Resisting Frame; 3: Steel Braced Frame; 4: Steel Moment Resisting Frame; 5: Light Timber; 6: Tilt Up Panel; 7: Light Industrial; 8: Advanced Design; 9: Brick Masonry; 10: Concrete Masonry; 11: Unknown Residential; 12: Unknown Commercial
Deprivation Index	DI 1 (least deprived) to DI 10 (most deprived)
Floor Area	Numerical value in m ²
Floor Type	1: Timber; 2: Concrete slab
Footprint Area	Numerical value in m ²
Roof Cladding Class	1: Clay/Concrete Tile; 2: Concrete Slab; 3: Membrane; 4: Metal Tile; 5: Other – Heavy; 6: Other – Light; 7: Sheet Metal
Storeys	Number of storeys
Use Category	1: Residential Dwellings; 2: Commercial – Business; 3: Commercial – Accommodation; 4: Industrial - Manufacturing, Storage; 5: Industrial - Chemical, Energy, Hazardous; 6: Fast Moving Consumer Goods; 7: Government; 8: Territorial Authority/Civil Defence; 9: Lifeline Utilities; 10: Police; 11: Hospital, Clinic; 12: Fire Station; 13: Community; 14: Education; 15: Resthome; 16: Religious; 17: Forestry, Mining; 18: Farm; 19: Lifestyle; 20: Parking; 21: Clear Site; 22: Other
Wall Cladding Class	1: Weatherboard; 2: Stucco, Roughcast; 3: Corrugated Iron; 4: Plastic; 5: Fibre Cement Sheet; 6: Fibre Cement Plank; 7: Reinforced Concrete; 8: Concrete Masonry; 9: Brick; 10: Glass; 11: Curtain Wall Glazing; 12: Sheet Metal; 13: Other Sheet – Combustible; 14: Other Sheet - Non-Combustible; 15: Other
Year of Construction	1800 - 2016

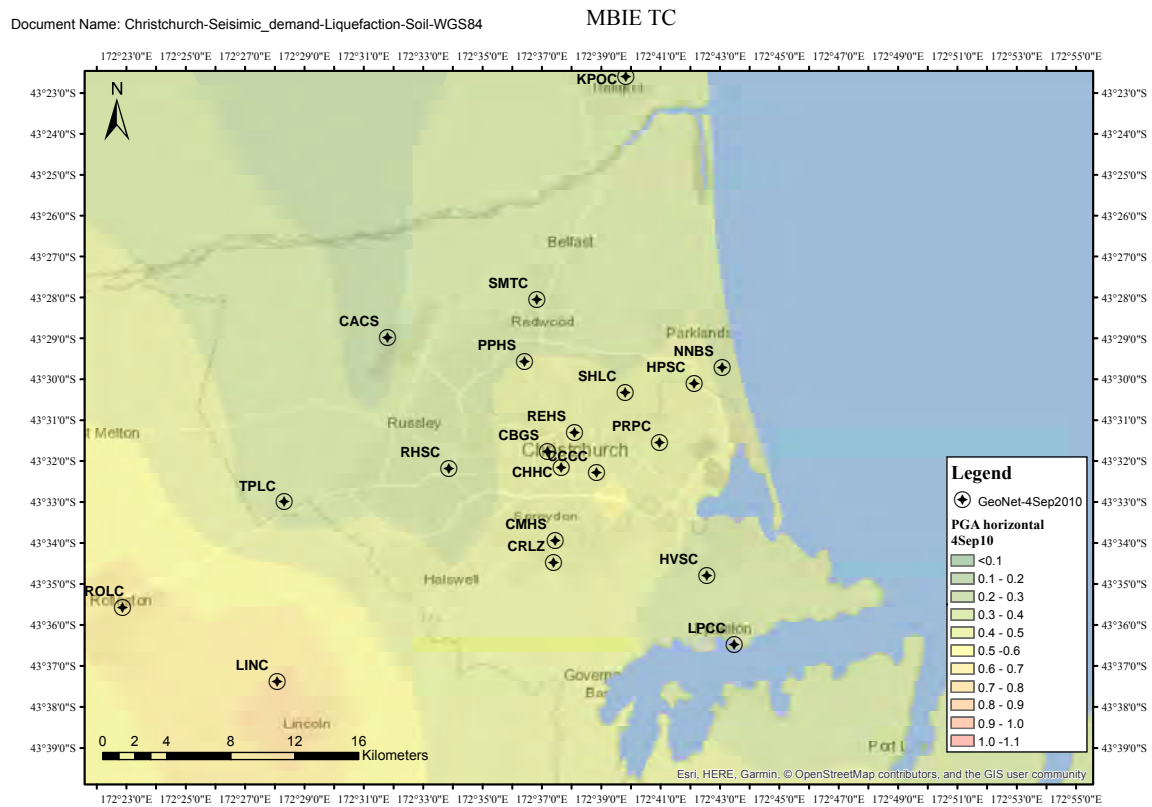


Figure 5.2: Location of the GeoNet recording stations in Christchurch (black dots) and interpolated PGA contours for the 4 September 2010 earthquake

5.3.2 Seismic demand

A key input for the damage prediction model is the seismic demand for each individual building. This project utilises recordings from the GeoNet strong motion database which contains recordings from large earthquakes (M_w 3.5 to 7.8) that occurred in New Zealand between 1968 and 2016 (GeoNet, 2012; Kaiser et al., 2017; Van Houtte et al., 2017). GeoNet freely provided strong motion seismograph recordings of all events in the CES as recorded at 14 recording stations located throughout Christchurch. Whilst there are many possible metrics to describe the seismic demand, this study focuses on using summary data such as peak ground acceleration (PGA). The GeoNet data was interpolated for all Christchurch for four of the critical events for this study using inverse distance weighted (IDW) in ArcMap (Esri, 2019). Figure 5.2 presents an example of a map layer interpolated using ArcMap.

5.3.3 Liquefaction occurrence

During the CES, extensive liquefaction occurred during four events: 4 September 2010, 22 February 2011, 13 June 2011, and 23 December 2011. The liquefaction and related land damage was the most significant during the 22 February 2011 event. The extent of land damage from liquefaction as well as the PGA contours for the four aforementioned events are shown on Figure 5.3. The location and severity of the liquefaction occurrence was based on interpretation from on-site observations and LIDAR surveys. Similar maps showing the severity of the observed liquefaction are available to download as .kmz file from the New Zealand Geotechnical Database (NZGD) (Earthquake Commission (EQC) et al., 2012). Figure 5.4 shows a map of the liquefaction occurrence for the 22 February 2011 event.

The Land damage and liquefaction vulnerability due to the CES has been extensively studied. The interested reader is directed to the report from J. Russell and van Ballegooy (2015).

5.3.4 MBIE Technical categories

Following the CES, the Ministry of Business Innovation & Employment (MBIE) and the Canterbury Earthquake Authority (CERA) introduced land classifications and zones to aid foundations repair and rebuild decisions. CERA delimited a “Red Zone” where the land has been so extensively affected during the CES and that it is expected to experience poor land performance in future events. Accordingly, the construction of any buildings is outlawed in the Red Zone (see Figure 5.5). Similarly, “Green Zone” designates where the construction of residential buildings can take place. The “Green Zone” has been subdivided into three technical categories (TC): TC1, TC2, and TC3 depending on the possible future land damage that can occur. For each category, Table 5.2 details the future land performance expectation from liquefaction and related foundation requirements. Figure 5.5 shows the extent of the three TC zones in Christchurch.

5.3.5 Soil conditions

The soil properties influence the seismic demand and response of structures and effect the liquefaction occurrence. This study uses soil information from the Land

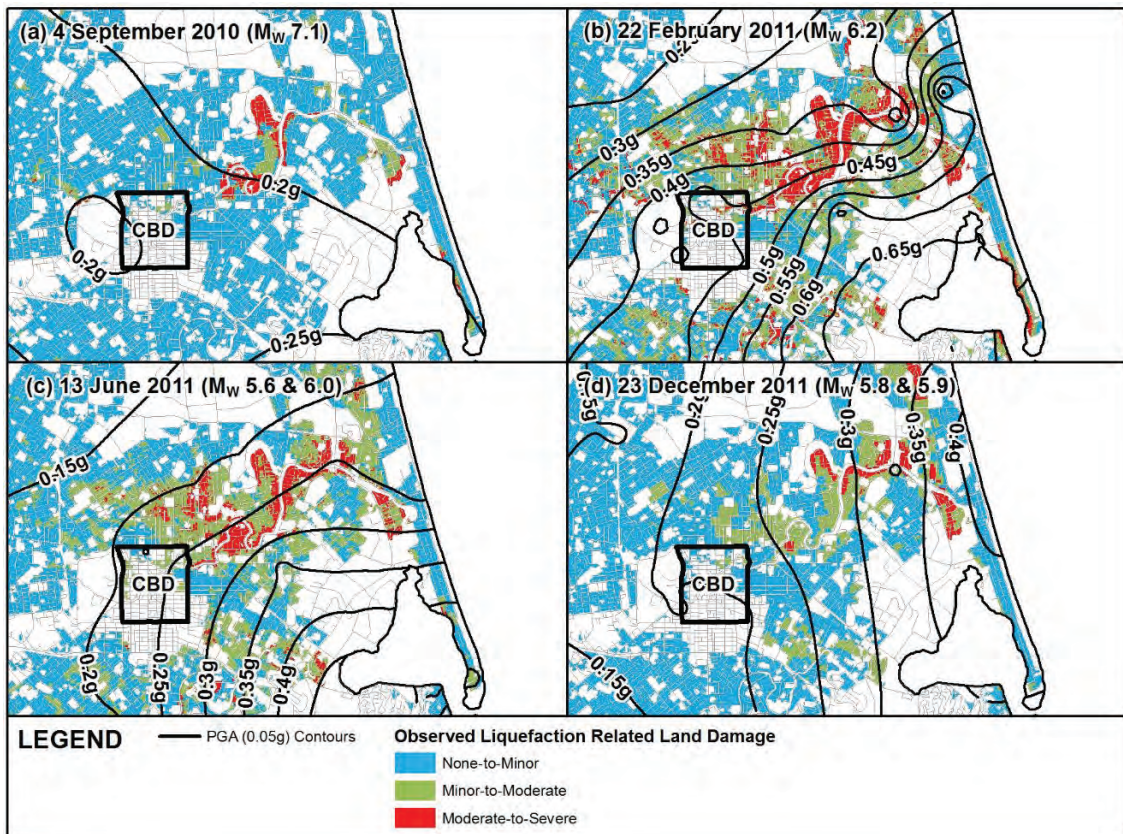


Figure 3.6: Maps showing the inferred levels of earthquake shaking and the observed land damage for urban residential properties in Christchurch after the (a) 4 September 2010, (b) 22 February 2011, (c) 13 June 2011 and (d) 23 December 2011 earthquakes.

In Figure 3.6 the contour lines for the June 2011 and December 2011 are the estimated PGA contour lines for the main earthquake events on those dates. These do not capture the influence of the PGAs associated with the foreshocks of these events which are relevant to the liquefaction-related damage observed. This is discussed in more detail in Section 3.7.3.

Urban Christchurch has experienced approximately 500 year return period levels of earthquake shaking for one or more of the four main earthquakes. The exceptions to this are the north-western suburbs (i.e. Avonhead, Belfast, Bishopton, Brooklands, Bryndur, Burnside, Casebrook, Ilam, Kaiapoi, Northcote, Spencerville, Styx and Upper Riccarton) which experienced approximately 100 year return period levels of earthquake shaking.

Foundation Technical Category (TC)	Future Land Performance/ Foundation Criteria
<p>TC1 Mapped Liquefaction Related Land Damage</p> <p>As a result of the earthquakes and related insurance claims for land damage with EQC, extensive land damage evaluations were undertaken by geotechnical engineers and engineering geologists. These evaluations characterised the extent and severity of liquefaction related land damage after each of the main earthquakes.</p>	<p>Future land damage from liquefaction is unlikely, and ground settlements from liquefaction effects are expected to be within normally accepted tolerances.</p>
<p>TC2</p> <p>Liquefaction related land damage mapping of residential properties was carried out immediately after the September 2010, February 2011, and June 2011 earthquakes to assess the extent and severity of the surface effects of liquefaction. The mapping was supplemented by interpretation of aerial photography after each earthquake to help identify areas where liquefaction ejecta occurred, but which were not mapped.</p>	<p>Liquefaction damage is possible in future large earthquakes. Shallow geotechnical investigations may be required. Suspended timber floor or cantilever slab foundation options can be used.</p>
<p>TC3</p> <p>Liquefaction related land damage mapping of residential properties was carried out immediately after the September 2010, February 2011, and June 2011 earthquakes to assess the extent and severity of the surface effects of liquefaction. The mapping was supplemented by interpretation of aerial photography after each earthquake to help identify areas where liquefaction ejecta occurred, but which were not mapped.</p>	<p>Liquefaction damage is possible in future large earthquakes. Deep geotechnical investigations may be required. High require specific engineering input for foundations.</p>

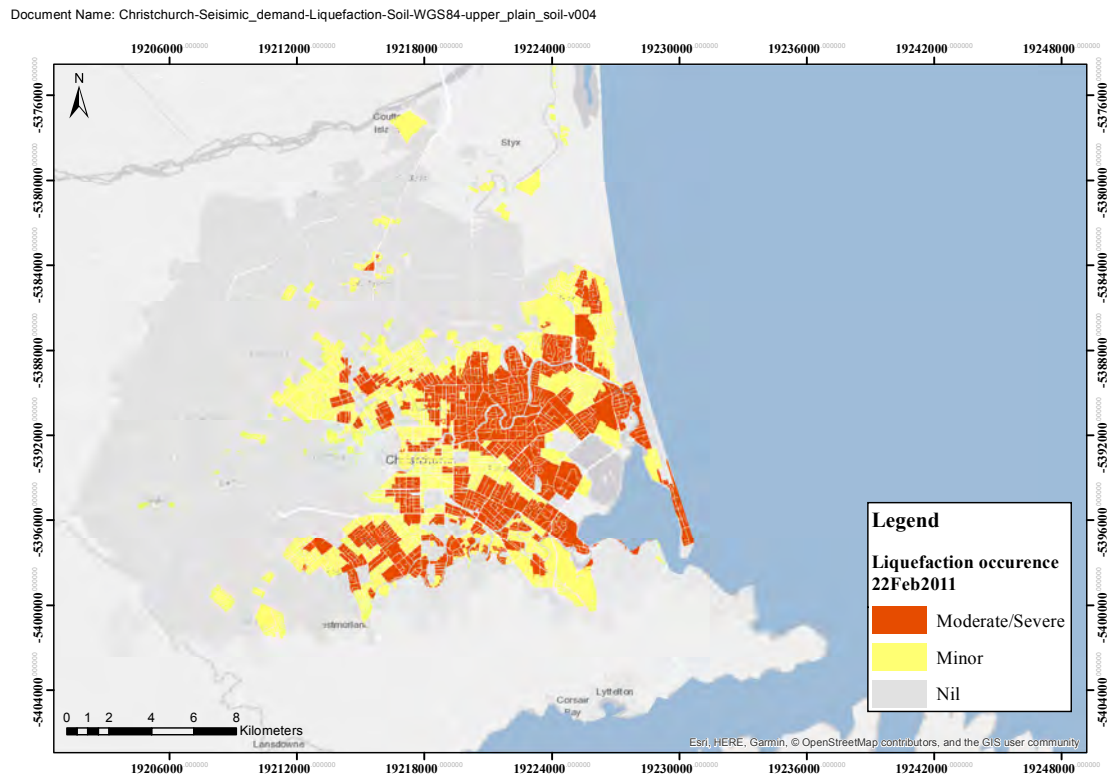


Figure 5.4: Liquefaction occurrence for 22 February 2011, data from (Earthquake Commission (EQC) et al., 2012)

Information New Zealand (LINZ) (Land Information New Zealand (LINZ), 2019) and Land Resource Information Systems (LRIS) (Land Resource Information Systems (LRIS), 2014). LRIS publishes a databases that provides topographical and soil conditions for the Christchurch region (Land Resource Information Systems (LRIS), 2010). Figure 5.6 shows a map of the soils for Christchurch and detailed soil descriptions can be found in Appendix D. Information on the soil database related to the soil phase, texture, depth and slope class can be found in Kear et al. (1967) and Cox (1978).

5.4 Feature extraction/selection

5.4.1 Extract EQC residential building claims related to the CES

The EQC insurance claims data set is organised according to the event date when the damage is purported to have stemmed from. It is thus necessary to extract and organise the data that are directly related to the CES. The CES started on 4 September 2010 and ended on 23 December 2011. It includes main events (4 September 2010 (M_w 5.9), 22

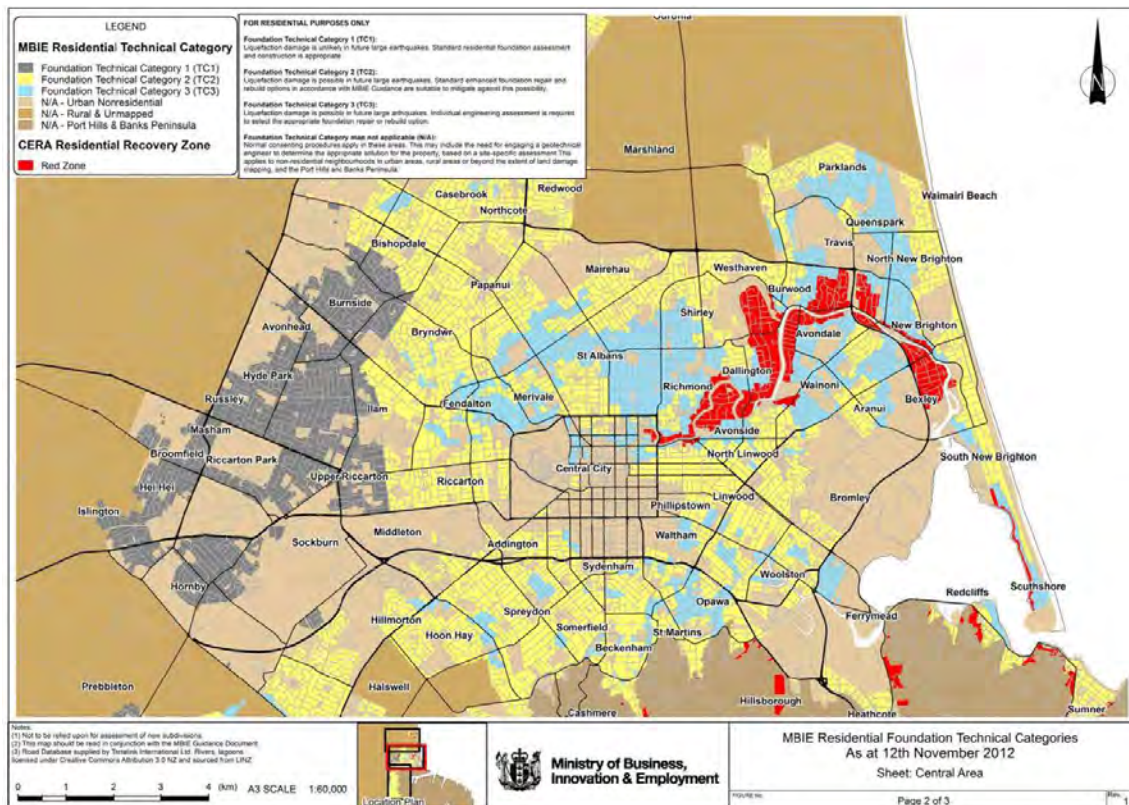


Figure 5.5: Map of CERA “Red Zone” and MBIE Residential Technical Category (Ministry of Business Innovation & Employment (MBIE), 2012)

February 2011 (M_w 6.2), 13 June 2011 (M_w 6.0), 23 December 2011 (M_w 5.8 & 5.9)) followed by multiple aftershocks. After filtering for data between the 4 September 2010 and the 23 December 2011 in the EQC claims data set, it results in 76 earthquake events for which claims have been lodged.

Removing rows with missing information for key variables

In its raw format the EQC data set has 433,536 rows for the 76 earthquake events of the CES. As shown in Figure 5.1 values for the PortfolioID and building coordinates are missing for 21,114 and 327 instances respectively. As these features are critical to identifying a building and thus any subsequent merging process, all rows where either the PortfolioID and/or building coordinates are missing are removed. After this, the claims data set pertaining is reduced to 412,418 instances.

Some buildings were damaged in multiple events during the CES thus leading to several claims being submitted for the same building throughout the CES. Figure 5.7 shows the number of claims and the number of properties for each earthquake event

Document Name: Christchurch-Seismic_demand-Liquefaction-Soil-WGS84-upper_plain_soil-v004

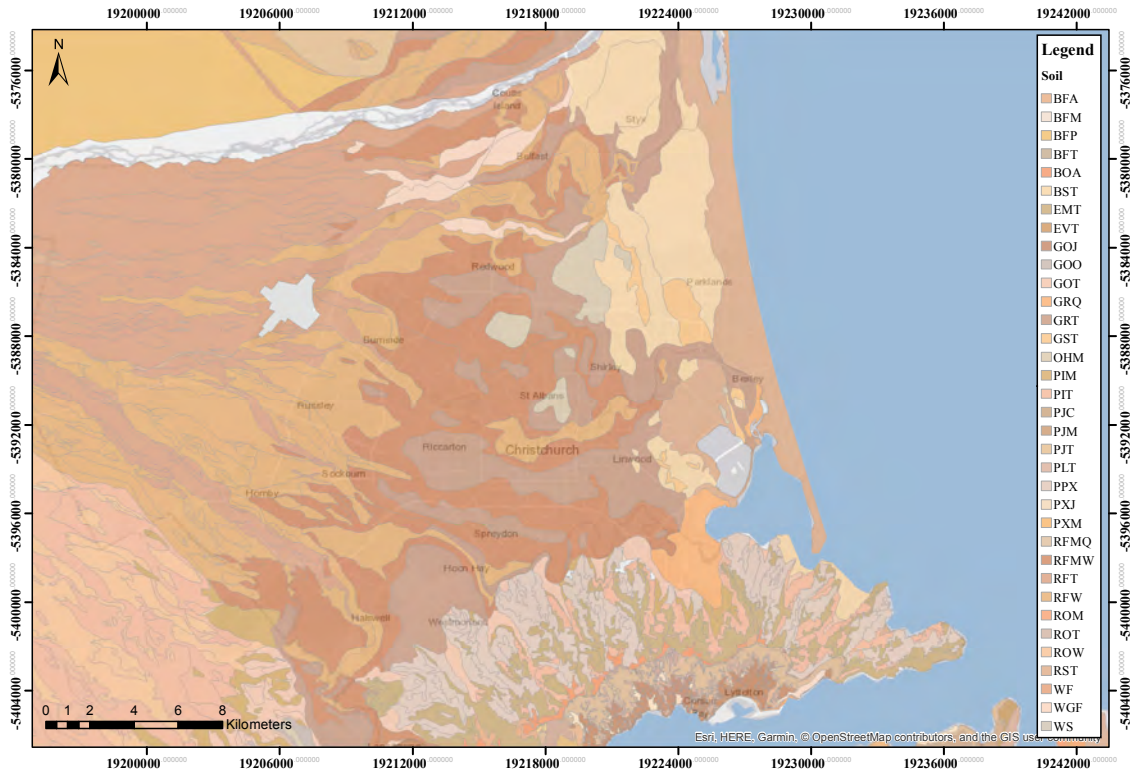


Figure 5.6: Map showing the NZSC soil order classification in Christchurch (layer obtained from (Land Resource Information Systems (LRIS), 2010)). Information of the soil codes can be found in Table D.1 in Appendix D.

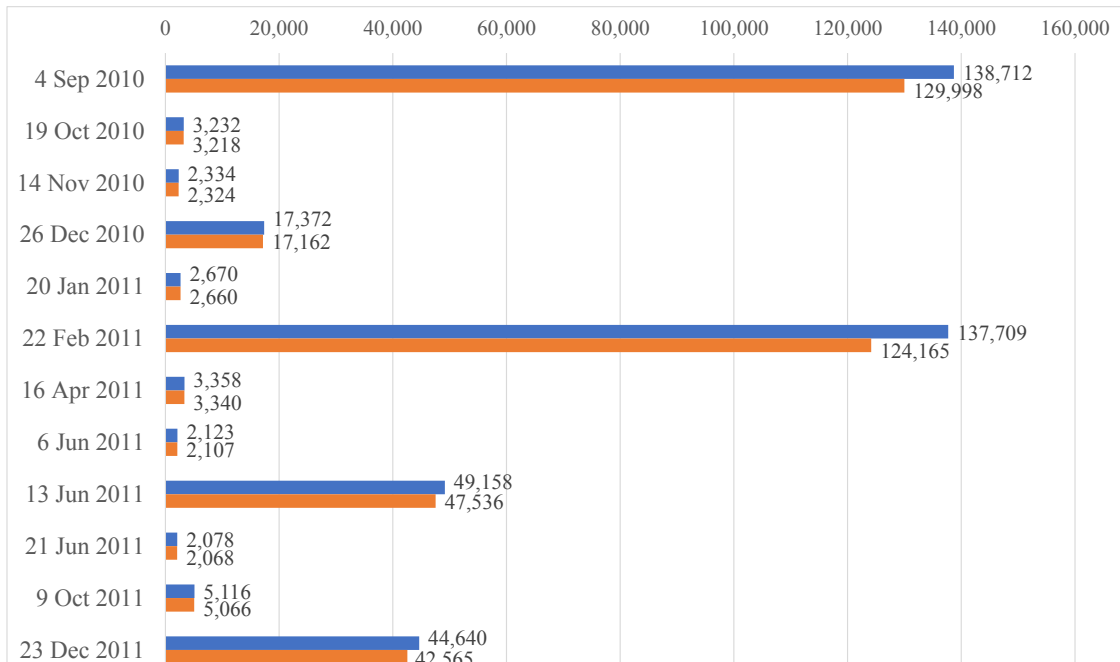
in the CES. For each earthquake, a comparison of the count of the number of claims versus the number of properties reveals that the number of properties is always lower. This indicates that in some instances there are multiple claims for the same building in a particular earthquake event.

Rearrange feature order

The raw EQC's claims database is claim centric. This means that one row of data corresponds to one claim. However, the total damage to a property can consist of multiple claims or multiple rows of data filed at different dates, particularly due to the nature of multiple events in the CES. Thus, it is later necessary to transform the database into a property centric layout. To ease the data manipulation in the pre-processing steps, the variable 'PortfolioID' is moved as the first column of the EQC data set and thus as principle identifier for each instance.

Export data for each earthquake event

Following the aforementioned steps above, the claims data related to each specific



Process



Step 1: Extract EQC data

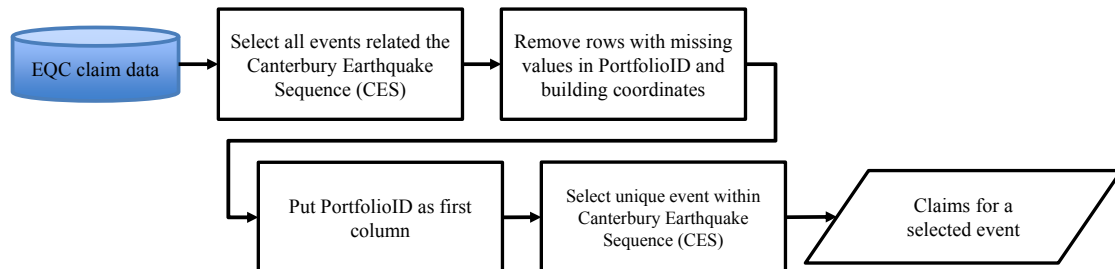


Figure 5.8: Steps to extract EQC data for a unique event in the Canterbury earthquake sequence (CES)

earthquake event is re-exported into separate files. This simplifies the merging of information from external database described in section 5.3, as the seismic demand and liquefaction occurrence data are collated according to specific events in the CES. Figure 5.8 shows an overview of the processing steps from the raw EQC claims data to data for a selected event.

5.4.2 Select claim status

The EQC claim database entails an attribute that indicates the current status of the claim. Not all the claims included in the database are settled. For some assessment might still be in progress or not even started. To obtain a data set with consistent claim values representative of the actual building loss, it is necessary to select claims that have been settled. Figure 5.9 shows the number of claims related to the 4 September 2010 event pertaining to each of the claims status categories. 51.5% of the claims are settled and the payment is completed. However, 26% have not been assessed or settled yet or have been declined. 15.2% have been settled on a different claim related to the identical property.

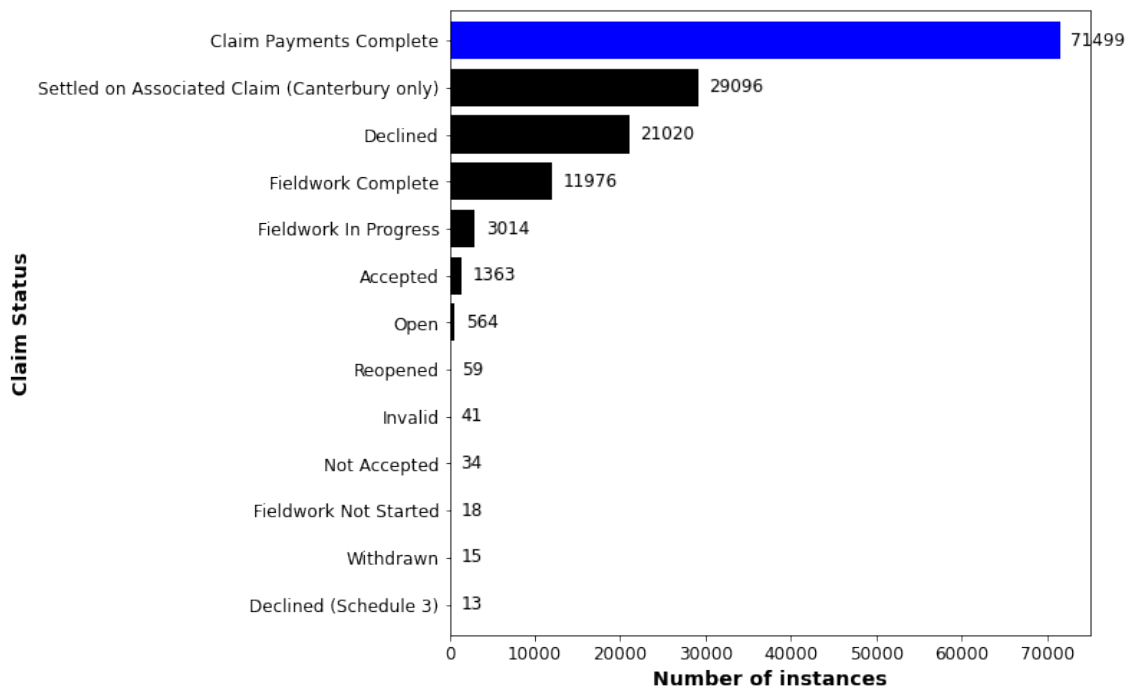
For training a supervised machine learning model, it is necessary to have “clean” data with known target attributes. Thus, only claims with “Claim Payments Complete” status are used for the ML model training. Figure 5.10 shows the number of instances remaining for each important earthquake event in the CES.

5.5 Overview of the data merging

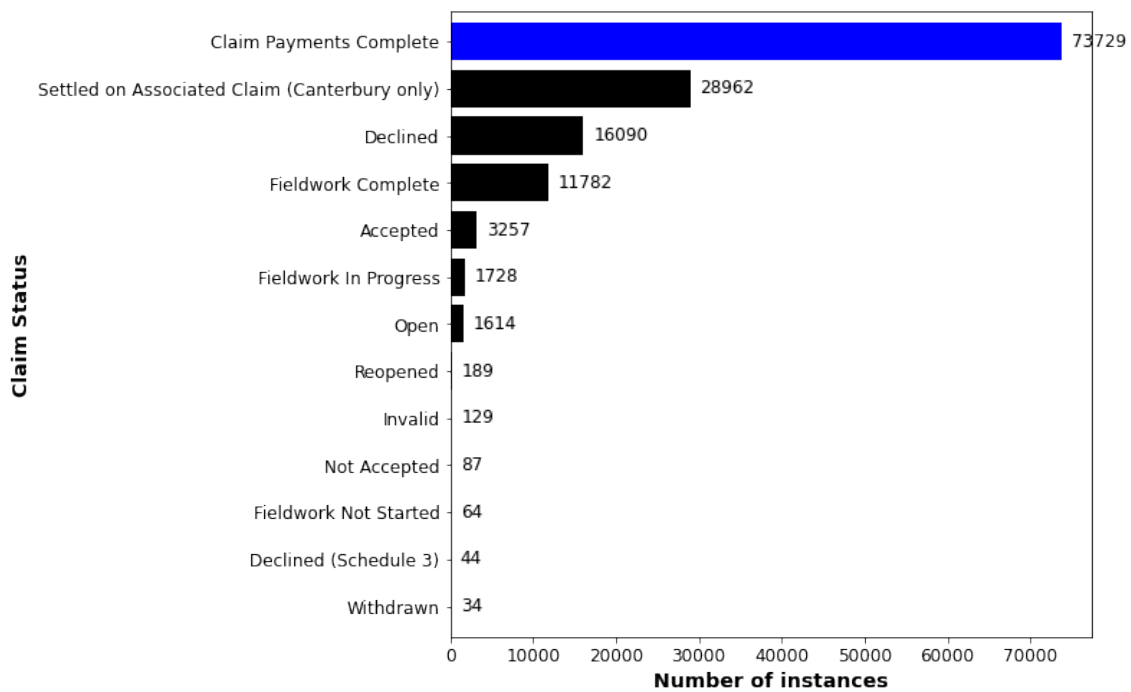
Figure 5.11 shows a schematic overview of the information that are merged on EQC’s claims data set for this project. The RiskScape data set delivered key buildings characteristics. GeoNet delivered key seismic demand through interpolated PGA. The Canterbury maps on liquefaction susceptibility and the NZGD maps on liquefaction and lateral spreading observation provided information on liquefaction occurrence observed after the 4 September 2010, 22 February 2011, 13 June 2011, and 23 December 2011 earthquakes. Finally, the LRIS soil map for the upper plains and downs of Canterbury delivered technical information on the type of soil present in the different areas of Christchurch. The specific challenges for each of the data merging steps are outlined in section 5.6.

5.6 Merging building characteristics from RiskScape with EQC residential claims

Most of the machine learning algorithms require data to be complete across all instances. It was thus decided to supplement the EQC claims data set with building characteristics



(a) 4 September 2010



(b) 22 February 2011

Figure 5.9: Number of instances for each category in the attribute ClaimStatus

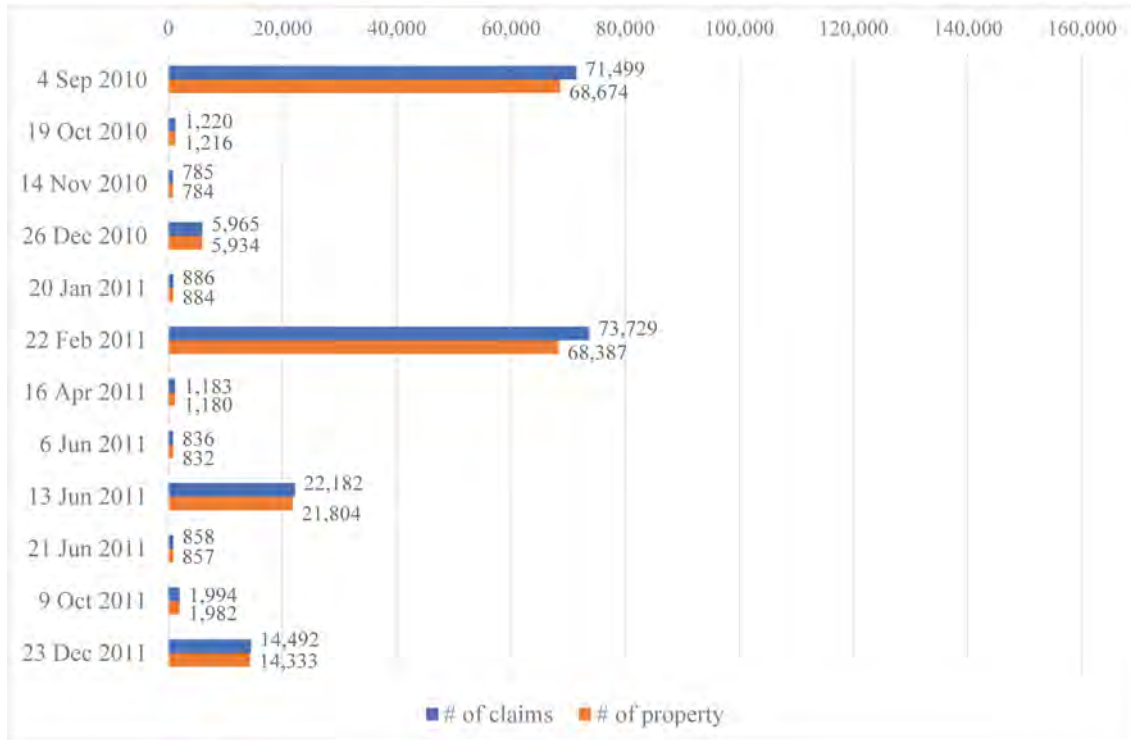


Figure 5.10: Number of claims and property for events in the CES after filtering for ClaimStatus. Only events with more than 1,000 instances prior to cleansing are shown.

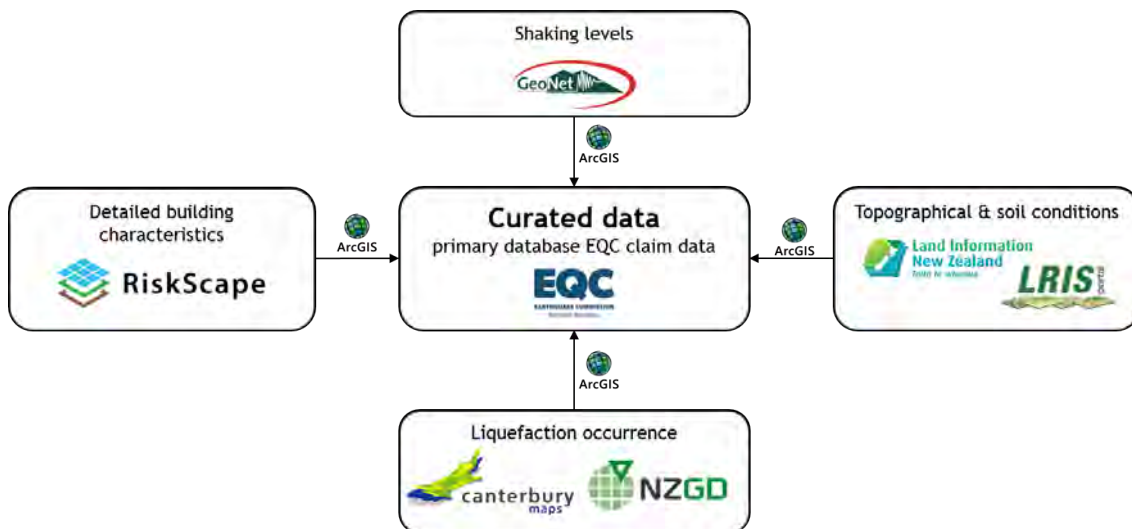


Figure 5.11: Schematic overview of the merging of information on top of EQC's claims data

from the RiskScape database. This was not straight forward due to the absence of a common attribute between the data sets and the non-exact matching of the coordinates between RiskScape data points and the location of EQC claims.

5.6.1 Initial merging attempts

Spatial join function embedded in GIS software

The first merging attempt used the spatial join function built-in into geographic information system (GIS) software. However, as EQC data points are fixed to the cardinal point for each street address, while RiskScape data points are fixed to the exact coordinates of each physical building, the software was not able to merge EQC claims data to the RiskScape buildings.

Spatial nearest neighbour

The second merging attempt used a spatial nearest neighbour join (NNJoin) function (Tveite, 2019). However, the RiskScape data set contains information for main dwelling as well as secondary buildings (e.g. garages, garden sheds). Thus, in some instances, the spatial nearest neighbour join led to incorrect merging, and merging of multiple building characteristics to one EQC claim.

In other cases, the spatial nearest neighbour join assigned incorrect RiskScape buildings to the EQC claims. It was found that sometimes the neighbouring buildings can be closer than the correct main dwelling to the EQC claims data point at the street address.

Figure 5.12 shows a sample comparison between the EQC claims data point locations and the actual locations of the buildings taken from the RiskScape data set. From the map, it can be seen that the points from the two data sets are not close to each other. For some property, it can also be observed that the EQC claims data set entails two points, meaning that multiple claims have been lodged throughout the CES for that particular property.

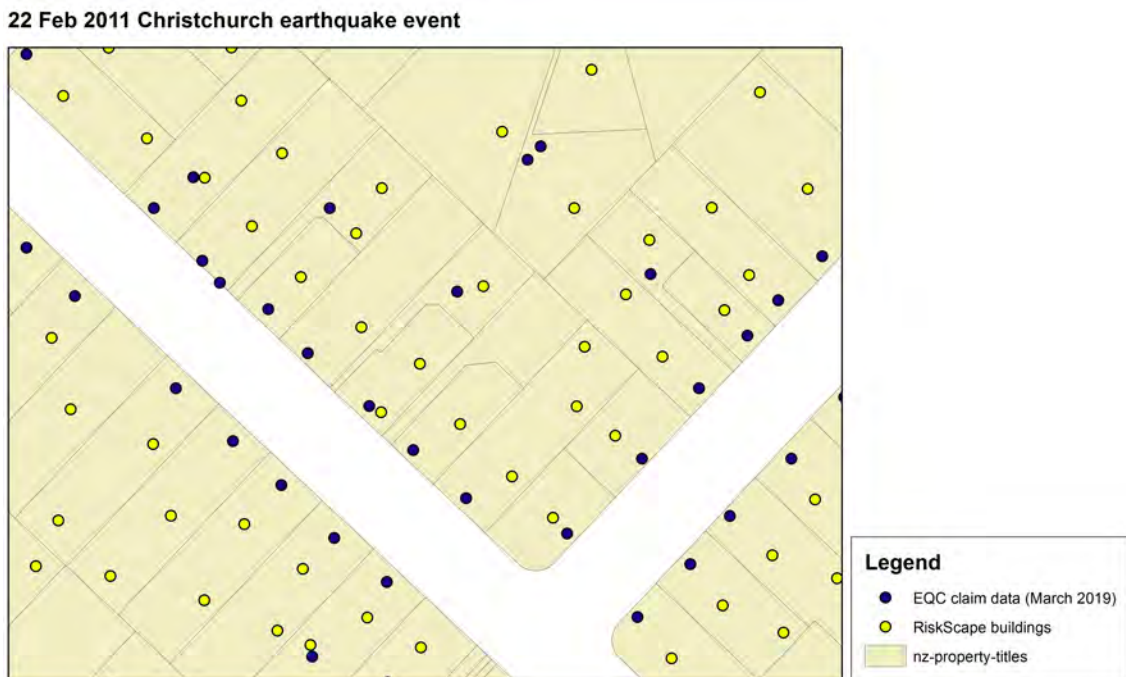


Figure 5.12: Comparison of the spatial location of the EQC claims data points and the RiskScape buildings

5.6.2 Alternate approach: constraining the merging using property boundaries

An alternate approach is to use the Land Information New Zealand (LINZ) NZ Property Titles data set (Land Information New Zealand (LINZ), 2020a) as an intermediary to constrain the merging process between EQC and RiskScape within property boundaries. However, given multiple RiskScape points and potentially multiple EQC claims data points can be present within a property, GIS approach alone will not be sufficient. The final adopted methodology used the building street address to enable data matching. However, the LINZ NZ Property Titles did not directly include information about the street address. This was instead available in the LINZ NZ Street Address data set (Land Information New Zealand (LINZ), 2020b). Thus, it was necessary to merge the LINZ NZ Street Address data (points) with the LINZ NZ Property Titles (polygons) before being able to use the street address information related to a property.

5.6.3 Merge LINZ NZ Street Address with LINZ NZ Property Titles

Figure 5.13 shows an overview of the process to combine the LINZ NZ Street Address (points) and LINZ NZ Property Titles (polygons). Both layers were merged using

Step 2.1: Merge LINZ NZ street address with LINZ property tiles

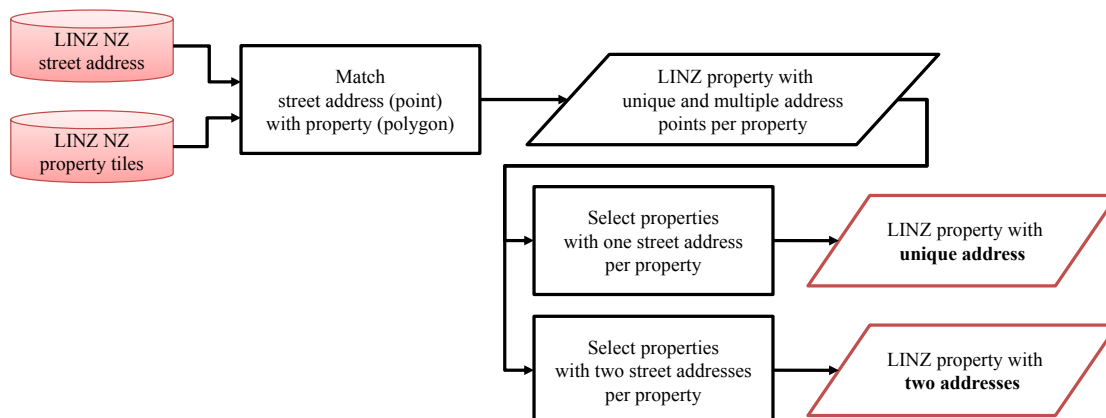


Figure 5.13: Merging of LINZ NZ street address with LINZ NZ property tiles

ArcMap (Esri, 2019). The merging created polygons having street address information. However, some of the property titles do not have a matching street address (see Figure 5.14). Filtering was thus performed to select only the properties containing street address point(s). Despite the filtering, some limitations related to the LINZ data sets remained. Figure 5.15 shows a satellite image of properties in Christchurch where some properties entail multiple street addresses within the same property outlines. This issue seems common case for apartments or properties that were recently subdivided. It was found that 89% of the LINZ property titles have one LINZ street address point per property, 7% have two address points, and 4% have three points or more. It should be reminded that the objective is to use property boundaries as intermediary mean to merge EQC claims with RiskScape building information. Properties with multiple street addresses would lead to inconsistencies in merging with claims, possibly associating the wrong building that lies within the same property boundary.

In order to train a machine learning model for building loss prediction, it is critical to have a training data set with reliable information. As no automatic solution was deemed suitable to address the issue with properties having three LINZ street address points or more, it has been decided to focus only on property titles with one and two street address point(s) per property. Highlighted in green in Figure 5.16, the map shows examples of properties having only one street address within each property boundary.

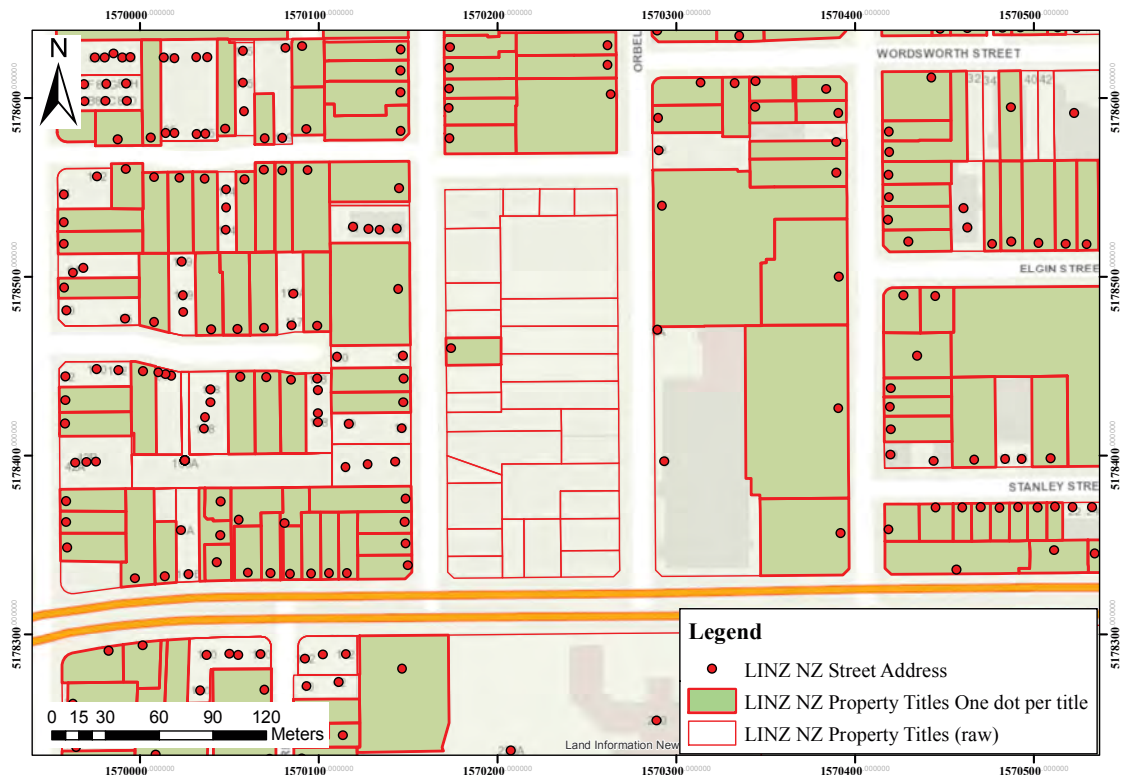


Figure 5.14: Map of an urban block in Christchurch overlaid with the LINZ NZ Street Address and LINZ NZ Property Titles layers. This highlights some Property Titles do not have a matching LINZ NZ Street Address.

5.6.4 Merge RiskScape with LINZ for instances with unique street address per property

The RiskScape database contains information for residential buildings as well as secondary buildings (e.g. external garages, garden shed). Therefore, some properties contain multiple RiskScape points within a LINZ property title (Figure 5.17). All RiskScape points present in a property were merged to LINZ street address for now. The merging used the “spatial join” function in ArcMap (Esri, 2019). It led to the street address information being added to all RiskScape points within a property.

5.6.5 Filtering the LINZ RiskScape data set for primary dwelling data

After LINZ street address was added to all RiskScape points within a property from the previous step, the data set still needs to be filtered to remove points associated with secondary buildings to align with the EQC claims data set, which relates to residential dwellings only.

Document Path: D:\ArcGIS\Christchurch-Merging_all_data-NZGD_2000-v002.mxd



Figure 5.15: Satellite image of urban blocks in Christchurch overlaid with the LINZ NZ Street Address and LINZ NZ Property Titles layers. The polygons with a bold red border represent LINZ NZ property titles having only one street address.

Figure 5.18 shows an overview of the filtering process. The first step checks if an address is unique. If an address appears only once within the merged RiskScape data set with LINZ property information, it can be used without further processing. If an address appears multiple times, it means that the property matched with multiple RiskScape points and thus needs to be filtered.

An exploratory analysis of the RiskScape data revealed that some of the building characteristics are assigned to the incorrect point within a property. As an example, the building characteristics related to the house were assigned to the RiskScape point located at the position of the garage and vice-versa. Filtering RiskScape points on the location would lead to incorrect building properties being merged to EQC claims. However, the RiskScape database includes two variables related to the building size (i.e. the building floor area and building footprint). For property containing only two RiskScape points and under the assumption that the principal dwelling is the building with the largest floor area and footprint on a property, it was possible to filter the data to retain RiskScape information related to main dwelling only. After successful filtering, these instances were

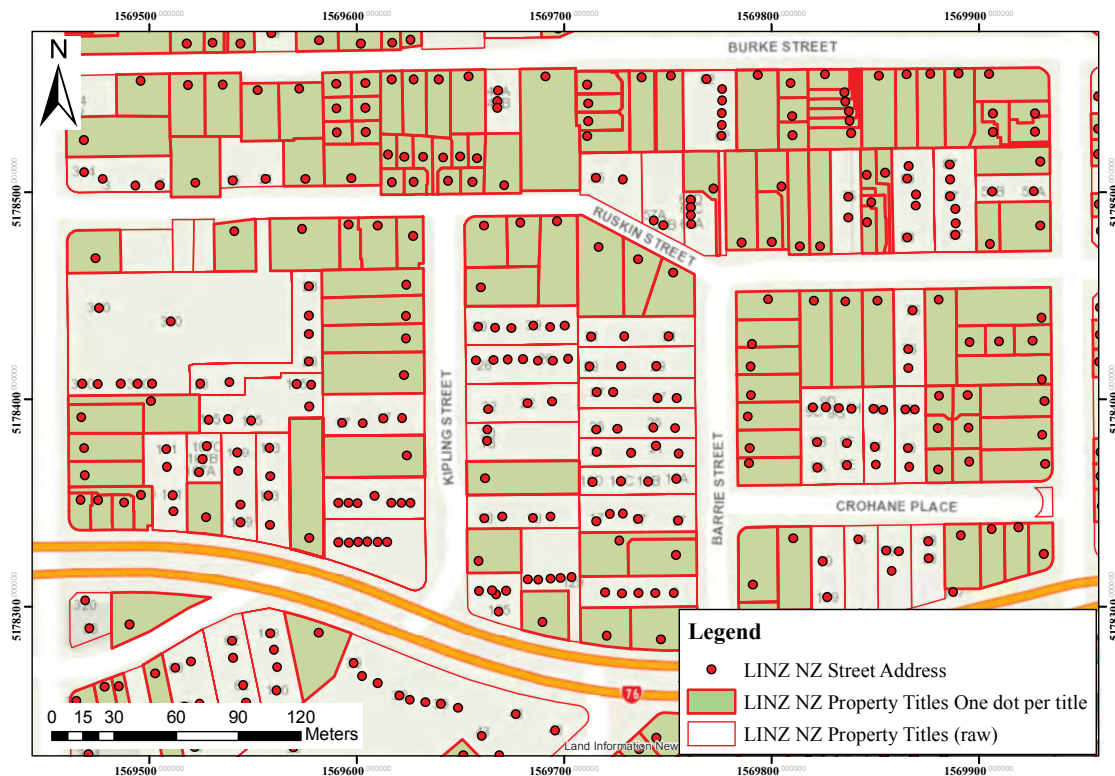


Figure 5.16: Map of urban blocks in Christchurch overlaid with the LINZ NZ Street Address and LINZ NZ Property Titles layers. The green polygons with a bold red border represent the selected LINZ NZ property titles having only one street address.

grouped with the RiskScape instances having one point per property title. The result is a RiskScape data set with street addresses containing residential buildings only.

Some of the properties have three or more RiskScape points (see Figure 5.19). Automatic filtering of the data using the largest building floor area is unreliable for those instances. In the aim of retaining only trusted data, it has been decided to discard such instances with one street address and more than three RiskScape instances in a property.

5.6.6 Properties with two street addresses and one or multiple RiskScape instances

7% of the LINZ property titles have two street address points. As the number of instances used to train a supervised machine learning model often affects the model accuracy, it was attempted to retrieve instances that were not collected via the previously mentioned approach. Nevertheless, the philosophy here followed was to put emphasis on the

Document Path: D:\ArcGIS\Christchurch-Merging_all_data-NZGD_2000-v002.mxd



Figure 5.17: Satellite view of an urban block in Christchurch with RiskScape points and selected LINZ NZ Property Titles

Process



Step 3: Filtering RiskScape

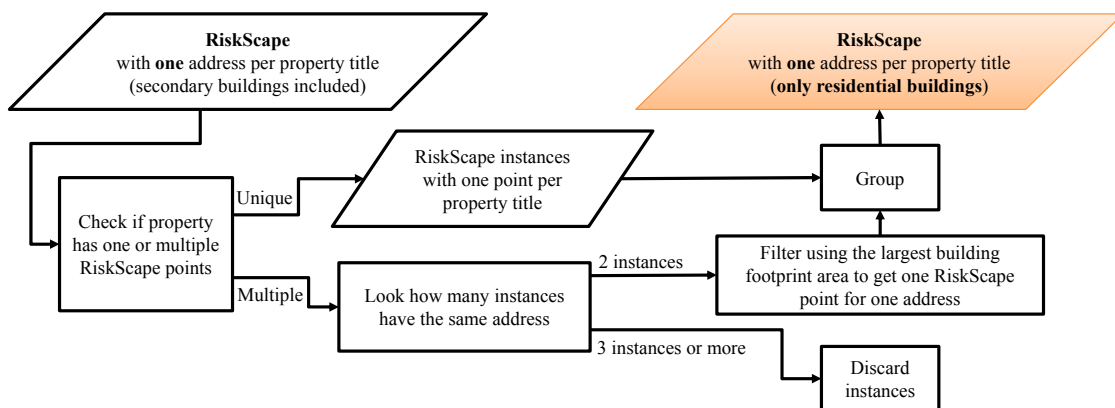


Figure 5.18: Steps to filter RiskScape data including secondary buildings to RiskScape data with residential buildings only

Selected area around 76 Grassmere Street, Papanui, Christchurch

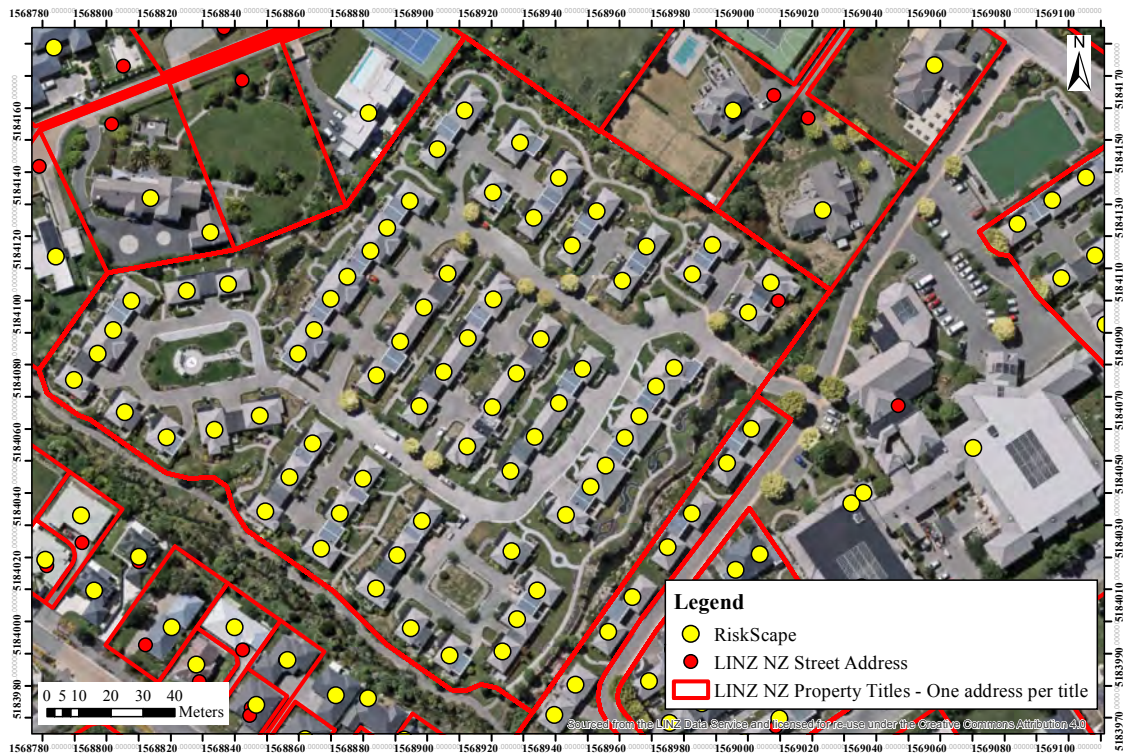


Figure 5.19: Property with one LINZ street address but multiple RiskScope points

quality of the data rather than the number of points. A specific approach to retain some of the instances having two LINZ street addresses per properties is explained here.

Figure 5.20 and Figure 5.21 show properties having two LINZ street addresses but a different number of RiskScope points per property. The property showed in Figure 5.20 has two RiskScope points. A merging approach similar to the one previously where the RiskScope point is selected using the largest building floor area is impractical as both of the RiskScope points relate to a residential dwelling for which an EQC claim has been lodged. From Figure 5.20 it can also be seen that a spatial nearest neighbour alone would lead to unsatisfactory result as both the LINZ street address points are located closer to the RiskScope point representing the house near the street. Figure 5.21 shows properties having two LINZ street addresses and multiple RiskScope points pertaining to houses as well as secondary buildings. An adequate selection of RiskScope points is required in order to have a data set entailing residential buildings only.

Based on the complexity presented by the examples above, the effort is focused on retaining the case when there are two LINZ street addresses and two Riskscape points in the same properties. The steps are as follows. First, select the LINZ street

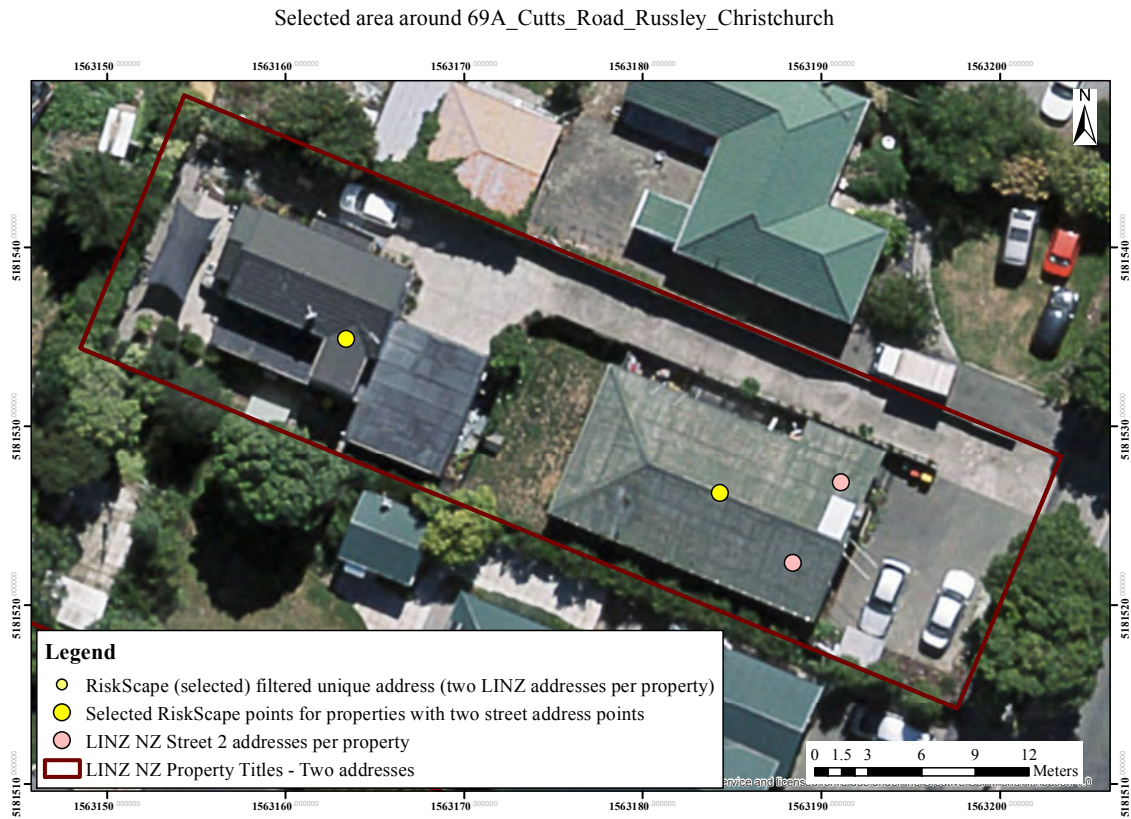


Figure 5.20: Property with two LINZ NZ street address points and two RiskScape points

Figure 5.21: Property with two LINZ NZ street address points and three RiskScape points



Figure 5.22: Neighbouring properties having two LINZ NZ street addresses and two RiskScope points each leading to issues with the “spatial join - closest”

address points for property tiles with two street addresses and the RiskScope points located within properties having two addresses. Merge the points using the “spatial join - closest” function embedded within the ArcMap software (Esri, 2019). This leads to each RiskScope points being assigned the closest LINZ street address point.

In some cases, multiple RiskScope points were assigned to the same LINZ street address point, which created duplicate instances. Figure 5.22 shows an example for neighbouring properties both having two LINZ street address points and two RiskScope points. For one of the LINZ street address point, the “spatial join - closest” function merged the LINZ point to the closest RiskScope point but also wrongly assigned another RiskScope point from the neighbouring property, thus creating a duplicate instance. These cases were removed. Following the selection of RiskScope points merged to their unique single LINZ points, the data was appended to the previous RiskScope data set.

5.6.7 LINZ and RiskScape merging for instances with unique and double street address(es) per property

Table 5.3 summarises the merging steps depending on the number of LINZ street address points and RiskScape points per LINZ property title. Three combinations were retained:

1. a direct selection for properties with one street address and one RiskScape point,
2. a selection with filtering using the largest floor area for properties with one street address and two RiskScape points, and
3. “spatial join - closest” combined with filtering to retain only the resulting instances with two LINZ street addresses and two RiskScape points.

To obtain clean data only, properties entailing three or more street address points or RiskScape building were discarded. While the current selection approach is conservative, it ensured each EQC claim can automatically be assigned to the corresponding residential building using the street address. For cases with multiple street addresses or residential buildings within the same property, a manual assignment of RiskScape points to LINZ street address points would enable the inclusion of more instances. However, this was impractical and insignificant for retaining only 4.04% of the overall LINZ property titles.

5.6.8 Merge EQC claims data with street addresses

The overall merging process of EQC claims points to LINZ street address points is similar to the process merging RiskScape to LINZ. The limitations related to the combination of the LINZ NZ street address data with the LINZ NZ property titles apply here as well. Hence, it was only possible to merge EQC claims to street address for points contained within LINZ property titles with one street address and to some extent retain claims for properties with two street addresses per title.

5.6.9 Multiple EQC instances

The EQC claims were treated separately for each event in the CES (4 September 2010, 22 February 2011, 13 June 2011, 23 December 2011). However, even after the selection of

Table 5.3: Overview of the action taken depending on the number of LINZ NZ street address and RiskScape point present per LINZ NZ property title

LINZ NZ street address	RiskScape	Action
1 point per LINZ property title	1 point per property title	Direct selection
1 point per LINZ property title	2 points per property title	Select the RiskScape point with the largest building floor area
1 point per LINZ property title	3 or more points per property title	Discarded
2 points per LINZ property title	1 point per property title	The automatic selection and filtering did not retain those instances as it could not differentiate this specific case
2 points per property LINZ title	2 points per property title	Retain these instances based on "spatial join - closest" combined with filtering.
2 points per property LINZ title	3 or more points per property title	Discarded
3 or more points per LINZ property title	Any configuration	Discarded

the claims related to an event, some properties still entailed multiple EQC instances. Two reasons were found.

1. First, for properties with two street addresses and two RiskScape points, the claims needed to be assigned to the corresponding building, similar to that described in section 5.6.6 using "spatial join - closest". This process led to some EQC claims being merged to another street address, albeit to a more lesser extent than for RiskScape points as EQC claims points are located closer to LINZ street address point.
2. Second, for properties with one street address, there were several EQC points when multiple claims were lodged for the same event.

In its original structure, the EQC claims data set is claim centric meaning that each row represents a claim that was lodged for one of the earthquake events during the CES. The aim, however, is to develop a machine learning model for loss prediction on

Step 4: Merge EQC claim with RiskScape using address

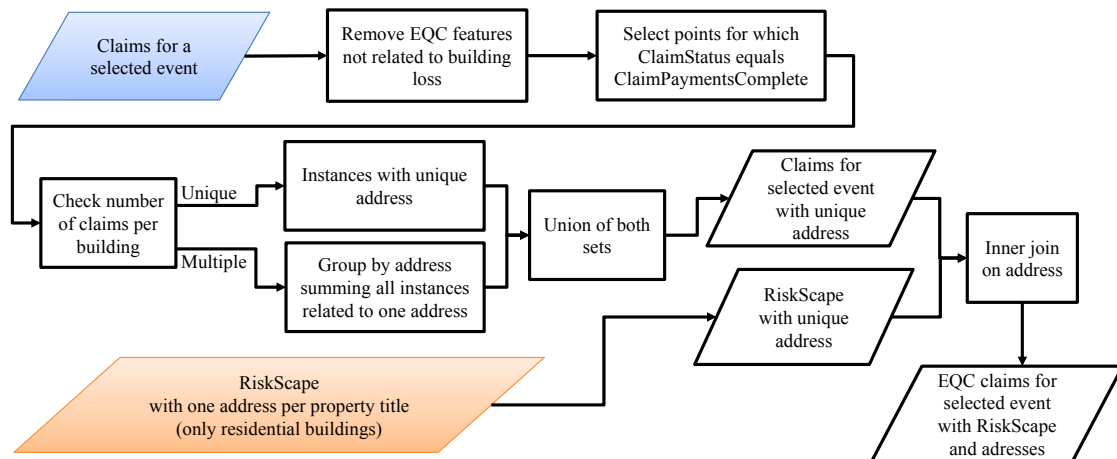


Figure 5.23: Steps to merge EQC and RiskScape using the LINZ NZ street address

a building by building basis. Therefore it is necessary to have training data where each row relates to one building only. The transformation was achieved by grouping multiple instances pertaining to the same building (instances having an identical address).

The claims data points were grouped by street address, and then the claim values were summed to obtain the overall losses that a building experienced for the selected earthquake event. This transforms the EQC data set to become property centric. This new layout facilitates the understanding of the number of events which affected each residential building and enables mapping of the necessary information on the building characteristics (from RiskScape), seismic demand, liquefaction and soil conditions for a considered event.

5.6.10 Merge EQC claims with RiskScape

Once the LINZ NZ street address information added to RiskScape and EQC, these data sets were merged in Python using the street address as a common field. Figure 5.23 shows a schematic of the overall merging process.

5.7 Add the seismic demand, liquefaction, and soil conditions information to EQC claims database

The final step of preparing the EQC claims data was to add information related to the seismic demand, the liquefaction occurrence, the location of MBIE Technical categories,

Step 5: Add additional information

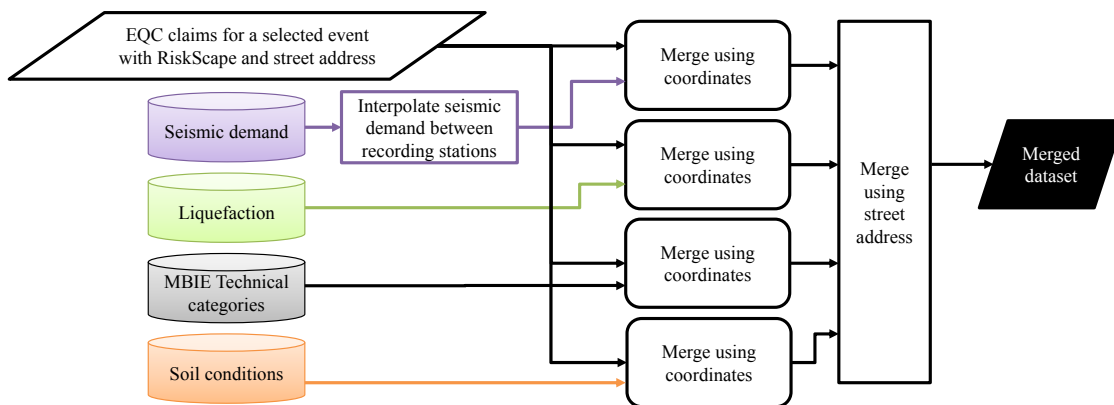


Figure 5.24: Steps to add the seismic demand, the liquefaction occurrence, the location of MBIE Technical categories, and the soil conditions on top of EQC claims

and the soil conditions. This was achieved within ArcMap (Esri, 2019) by importing each of the data set as a separate GIS layer. The information contained within each GIS layer was merged with the EQC claims previously combined with RiskScape as explained in section 5.6. Finally, using the street address as a common attribute, the information was combined in one merged data set. Figure 5.24 shows a schematic overview of the process.

5.8 The number of usable data points through the data merging process

Figure 5.25 shows the evolution of the number of instances for the 4 September 2010 and 22 February 2011 after each step in the merging process. In its original form, the EQC raw data set entails 145,000 claims for the 4 September 2010 and 144,300 claims for 22 February 2011. The first step was to clean the raw data by removing the instances missing information for the PortfolioID and building coordinates. This led to the loss of 6,300 points for 4 September 2010 and 6,600 for 22 February 2011.

The second step retained only the instances for which the claims payment was completed. The extraction of those claims induced the largest drop of the number of points with a loss of 48.5% for the instances related to 4 September 2010 and 46.5% for 22 February 2011 leaving only 71,500 and 73,700 claims for each respective event.

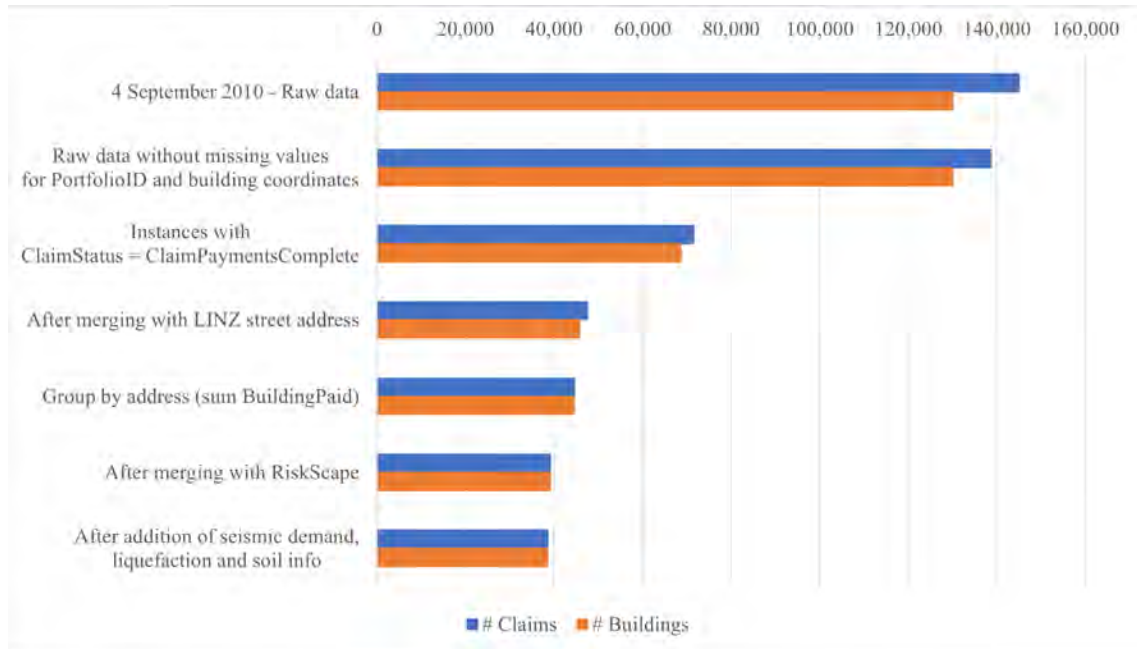
The third step merged the extracted claims with the LINZ NZ street addresses. This merging further induced a large drop in the number of instances. 33.5% of the claims

for 4 September 2010 and 25.9% for 22 February 2011 were lost because the correct street address could not be attached to a claim.

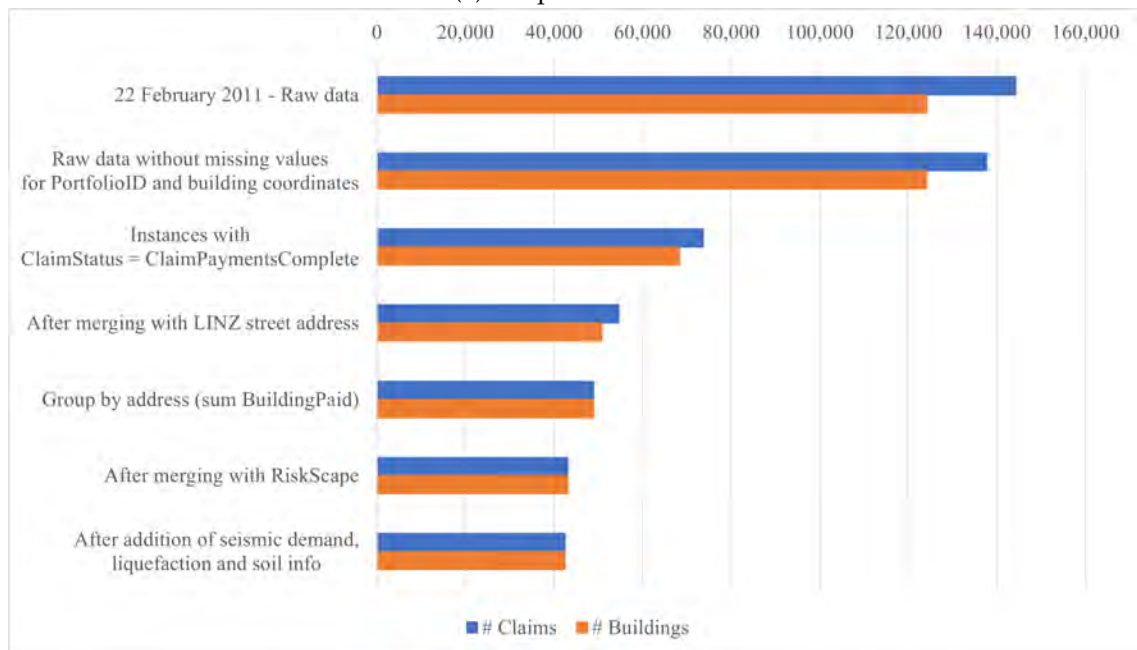
The fourth step transformed the data from a claim centric data set to a property centric data set. This reduced the number of claims from 47,500 to 44,500 for 4 September 2010 and from 54,600 to 49,000 for 22 February 2011. However, no information was lost during this transformation process. Multiple claims related to the same building were aggregated together to obtain the overall losses for the building for a selected earthquake event.

The fifth step, merging the data with RiskScape, led again to a loss of information as some of the instances could not be matched to the correct point from the RiskScape data set. 12% of the instances were lost for both 4 September 2010 and 22 February 2011.

Finally, the sixth step, dealing with the addition of information related to the seismic demand, liquefaction occurrence, and soil information led to the loss of 1.4% and 1.5% of the instances leaving 38,607 and 42,486 points for 4 September 2010 and 22 February 2011 respectively.



(a) 4 September 2010



(b) 22 February 2011

Figure 5.25: The number of data points after each processing step for event on 4 September 2010 and 22 February 2011

5.9 Conclusion

An examination of the raw data showed that the EQC claims data set has up to 85% of key building characteristics attributes missing. This chapter presented an approach to supplement information for the building attributes, seismic demand, liquefaction occurrence and soil type on top of EQC claims data set. Several challenges were encountered during the merging process mainly due to the fact that it was difficult to identify buildings across multiple data sets as there were no unique identifiers. Approaches using built-in functions from GIS software or based on spatial nearest neighbours led to unsatisfactory results as the claims were not located in the vicinity of the RiskScape points. The proposed solution used LINZ NZ property titles and LINZ NZ street address as an intermediary for the merging process. It was not possible to retain all the instances as some of the properties contained multiple street address points which led to difficulties assigning the RiskScape buildings to the correct address. The proposed solution worked well for instances related to properties having a unique LINZ street address, and properties with one or two RiskScape buildings, matching the street address and EQC claims.

The difficulties in the merging process highlighted the need for a unique identifier for each residential building to allow for better integration of the information from multiple data sources. Although the merging induced a drop in the number of instances, the most significant loss of points came from the selection of the claim status. The final merged data sets include 38,600 instances for 4 September 2010 and 42,400 points for 22 February 2011. These merged data sets have EQC's claims and additional corresponding attributes, and they will be used as inputs for developing seismic loss prediction models for residential buildings in New Zealand in the next two chapters.

Data pre-processing and model development of a seismic loss prediction model for residential buildings - Christchurch, New Zealand

This chapter presents the development of a machine learning model for the seismic loss prediction of residential buildings in Christchurch, New Zealand. It presents the main features available in the merged data set, details the distribution of the data, and documents the filtering process. The chapter then explains the processing of the target attribute and the processing of BuildingPaid from a numerical into a categorical variable. Subsequently, it describes the attribute selection, discusses the reasons for the non-inclusion of some attributes, and highlights the importance of attribute preparation. Finally, the chapter documents the algorithm selection, training, and model evaluation.

6.1 Introduction

The merging process of residential building claims from the 4 September 2010 and 22 February 2011 events led to 38,600 and 42,400 instances respectively. Each instance contains EQC information on the building losses (capped) enriched with data related to the building characteristic, seismic demand, liquefaction occurrence, and the soil type. For supervised machine learning algorithms to “learn” from a large number of instances, the input data has to satisfy specific requirements (e.g. no missing values). It was thus necessary to pre-process the data, deleting instances with missing information, and filtering the categorical data to remove any outliers that would influence the prediction performance and affect the model ability to generalise. Once the data pre-processed, the model attributes were selected, and the data set was split into a training and validation set. The machine learning algorithms were then trained using the training set. Additional efforts were applied to the model training for the 4 September 2010 event as the target variable had shown a significant imbalance between the categories. Several algorithms such as linear regression, decision tree, support vector machine (SVM), and random forest were then applied. Their prediction accuracy were evaluated and compared. The algorithm leading to better prediction performance was retained. Particular attention was also paid to the human interpretability of the model. Intrinsically interpretable algorithms had been preferred. More complex algorithms were applied in combination with post hoc methods to allow for interpretability as this study aims to develop a ‘grey-box’ model where the intermediate steps can be followed. Such a ‘grey-box model’ allows modellers to look through and validate the predictions at various key intermediate steps. It later also enables different stakeholders to extract relevant information that matters to them.

6.2 Feature filtering

Before fitting a machine learning model to a data set, it is necessary to remove any instance with missing value as many of the machine learning algorithms are unable to make predictions with missing features.

EQC variables

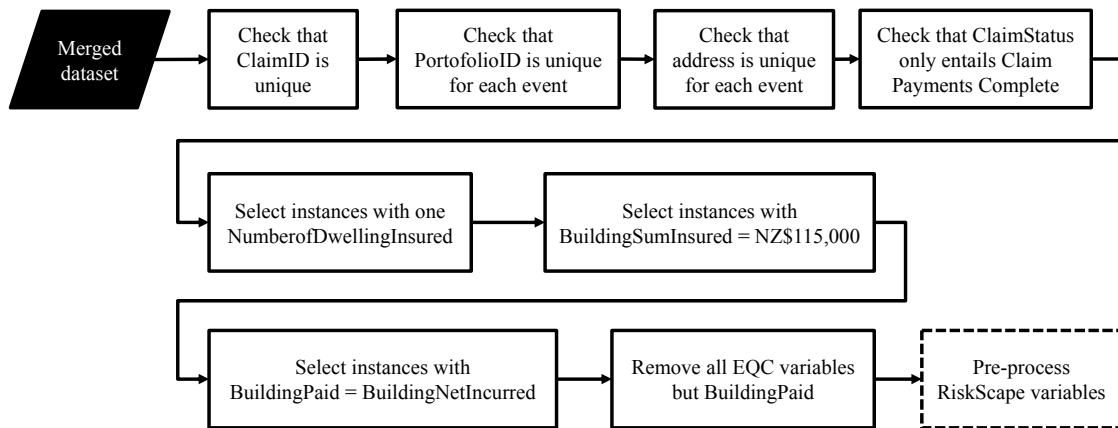


Figure 6.1: Overview of the steps for the filtering of the EQC features

Underrepresented categories within attributes are also carefully examined. Categories with few instances introduce challenges for the machine learning algorithms as the model will have difficulties to “learn” and generalise for the particular category. In some cases where the meaning is not changed, it is possible to combine instances from different categories. However, whenever a combination of multiple classes is not possible, categories entailing a few instances are removed.

The following section presents the filtering process for the EQC and RiskScope attributes. The objective is to retain as many information while ensuring that each category contains sufficient number of instances to obtain the best possible model accuracy.

6.2.1 EQC attributes

Figure 6.1 outlines the process for verifying the key EQC features within the previously merged data set. To begin, the merged data set was inspected to ensure that the ClaimID, Portfolio ID, and street address are unique for each instance, and that the data set contains only claims for which the payment was completed.

Number of Dwelling Insured

The EQC claims data set contains an attribute specifying the number of dwelling insured on a claim. Figure 6.2 shows the number of instances having one, two or more dwelling assigned to a claim. While 93% of the claims for 4 September 2010 and 89% for 22 February 2011 event are for single dwellings, 7% and 11% have zero dwelling for 4

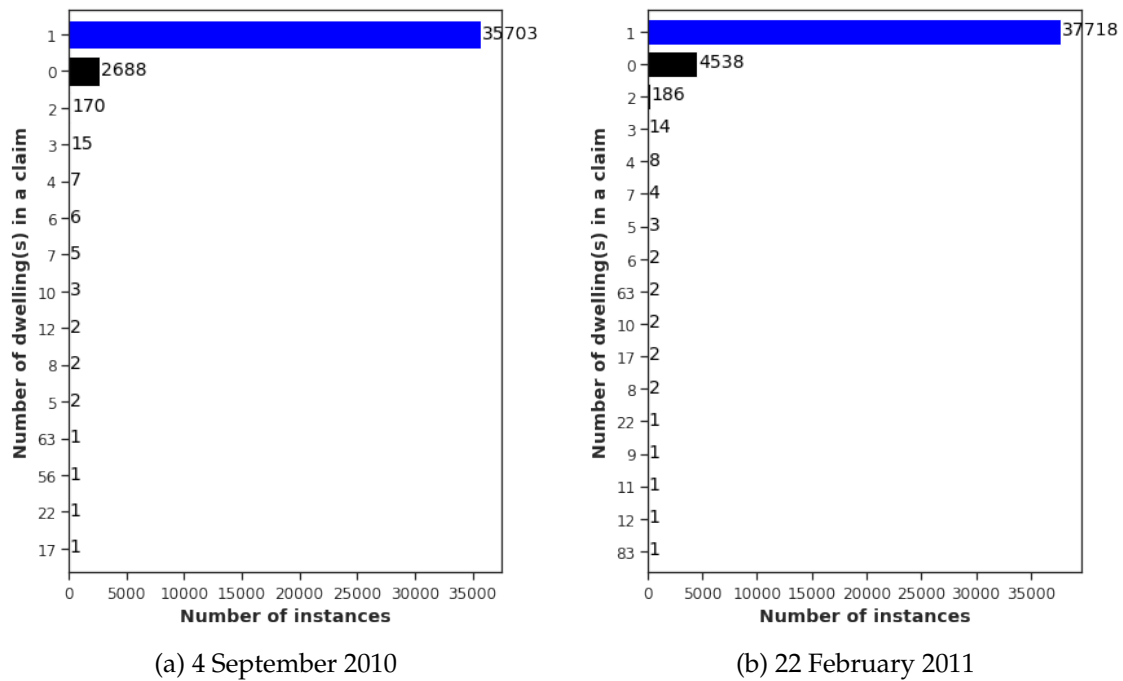


Figure 6.2: Number of instances for each value in Number of Dwelling insured

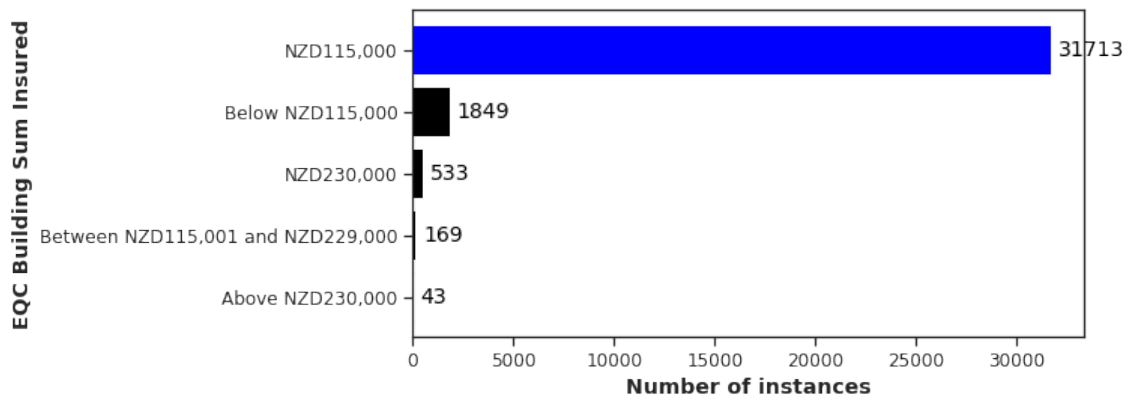
September 2010 and 22 February 2011 respectively. For both of the data sets, less than 1% have two or more dwellings. To avoid any possible issue with the division of the claim value between the multiple buildings, only claims related to one dwelling are retained.

EQC Building Sum Insured

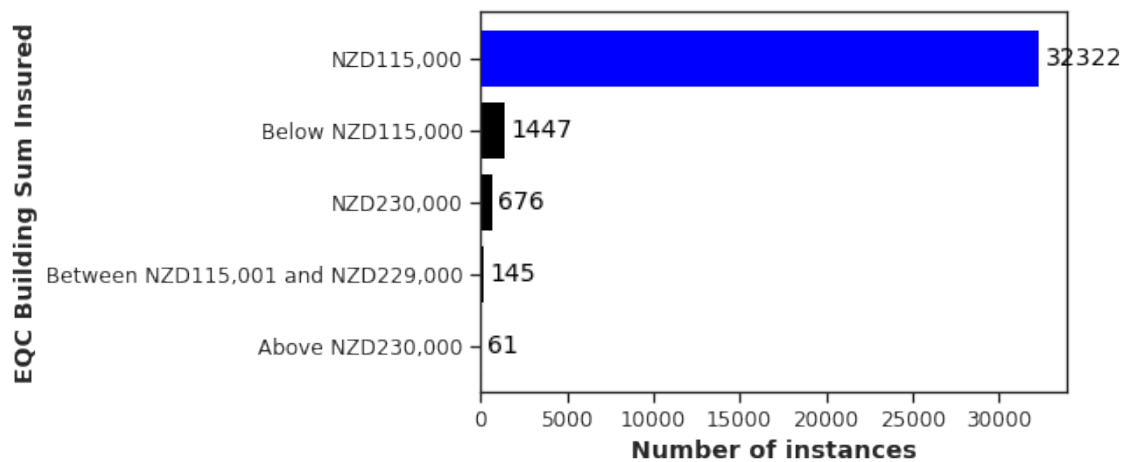
At the date of the CES, EQC provided a maximum cover of NZ\$100,000 (+ GST) or NZ\$115,000 for a residential dwelling for each natural event (Earthquake Commission (EQC), 2019b). In the EQC data set, a numerical attribute gives the maximum cover related to the claim lodged. For examination, the data was binned into five categories as shown in Figure 6.3. For both the 4 September 2010 and 22 February 2011 event, more than 92% of the claims relate to buildings having a cover of NZ\$115,000. Nevertheless, some instances show a maximum cover above the NZ\$115,000 threshold despite only properties with after one dwelling is retained from the previous step. Conversely, about 5% of the instances for both events are below the maximum cover. To ensure data integrity for the machine learning model, only the instances with exactly NZ\$115,000 maximum cover were selected.

Building Paid = Building Net Incurred

The EQC data set contains two attributes ('Building Paid' and 'Building Net Incurred')



(a) 4 September 2010



(b) 22 February 2011

Figure 6.3: Number of instances for EQC Building Sum Insured (categorised)

related to the payments for building damage. For claims marked as completed and closed those attributes should be equal. Despite the selection of settled claims only, 'Building Paid' were not exactly equal to 'Building Net Incurred' for 1% of the instances (see Figure 6.4). Those instances were removed to ensure that 'Building Paid', which will be used as the target variable for the machine learning model, are reliable final loss value for each building.

6.2.2 RiskScape attributes

Building Use Category

RiskScape has an attribute which specifies the use category of a building. While EQC insured residential buildings only, the RiskScape 'Use Category' shows that some of the buildings have a different main purpose (see Figure 6.5). To ensure the claims related to

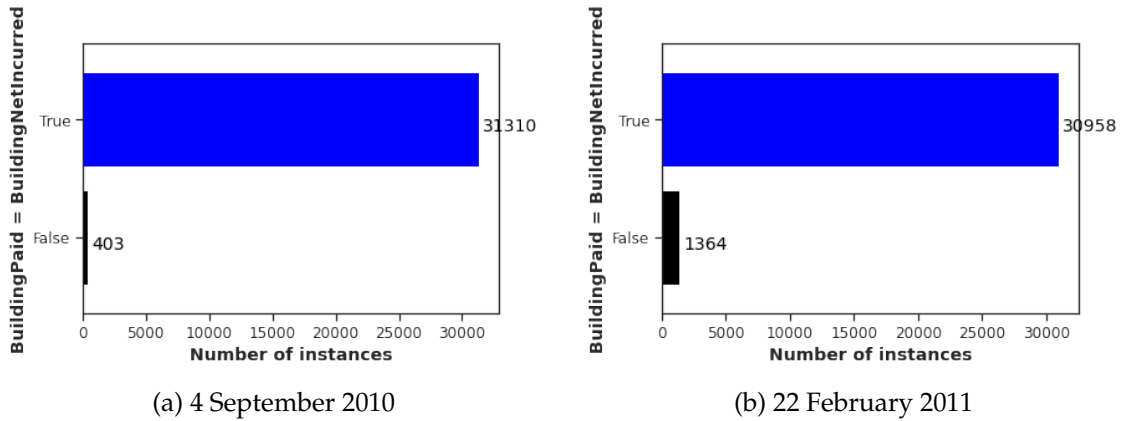


Figure 6.4: Number of instances for which Building Paid equals Building Net Incurred

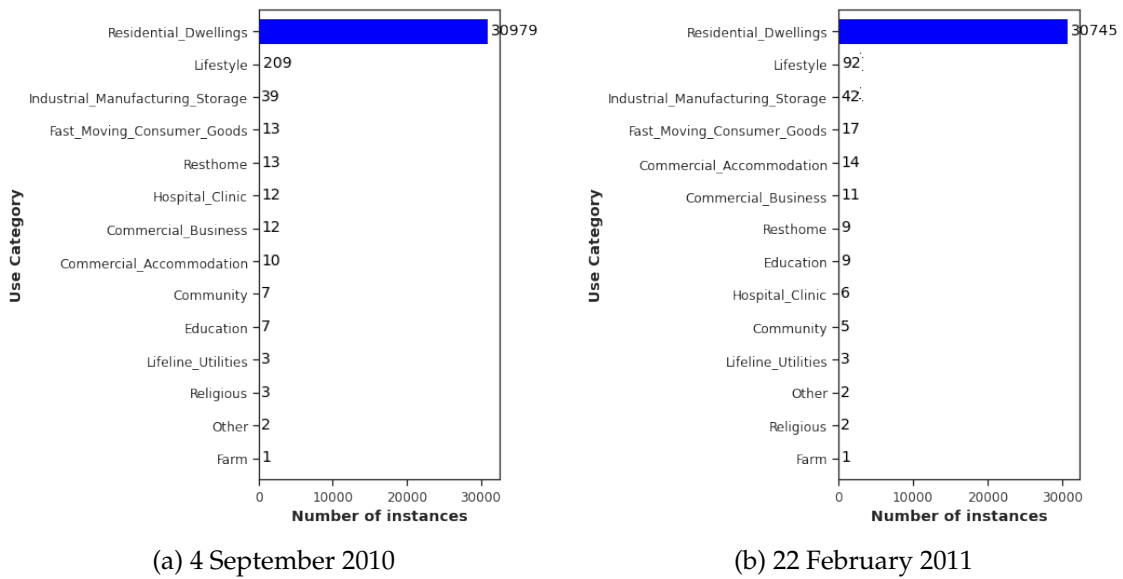


Figure 6.5: Number of instances per 'Use Category'

residential properties only, all instances not having residential dwellings 'Use Category' were discarded.

Building Floor area

An examinations of the building floor area revealed the presence of outliers, with values reaching up to 3,809 sqm for a house (see Figure 6.6). A filtering threshold was thus set at 1,000 sqm to remove the outliers. This led to a minimal loss of the instances (0.1%) but eliminated the outliers. The distribution of the building floor area below 1,000 sqm is shown in Figure 6.7.

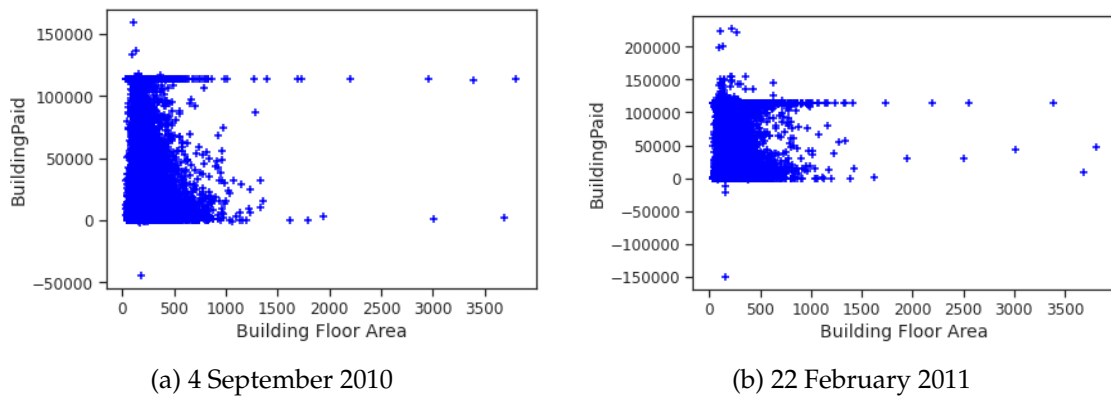


Figure 6.6: Distribution of Building Paid against Building Floor area

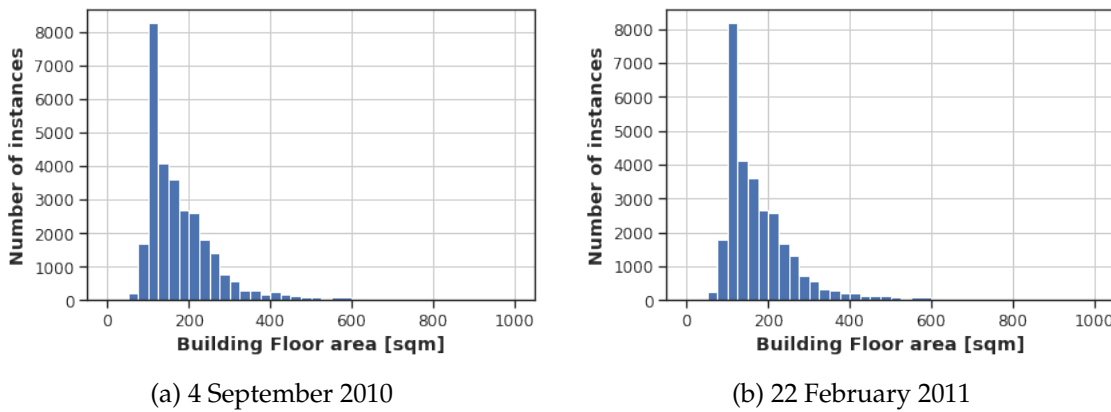


Figure 6.7: Distribution of Building Floor area (selected below 1,000 sqm)

Construction Type

Figure 6.8 shows the number of instances for each construction type in the merged data set. Light timber buildings is the most prevalent construction type, with 26,414 buildings and 26,387 buildings for 4 September 2010 and 22 February 2011, respectively. Conversely, steel braced frame, light industrial, reinforced concrete (RC) moment-resisting frame, and tilt-up panel only appear in very few instances. Given that these categories have less than 100 instances, it is unlikely machine learning models can make correct predictions for those construction types. It was thus decided that these are not within our scope and to filter out those underrepresented categories.

Selected, along with light timber dwellings were buildings where main construction type is classified as RC shear wall, concrete masonry, and brick masonry. While the latter category only entails 347 and 371 instances for 4 September 2010 and 22 February 2011 respectively, it was deemed necessary from an engineering point of view to retain brick masonry as possible construction type in the model.

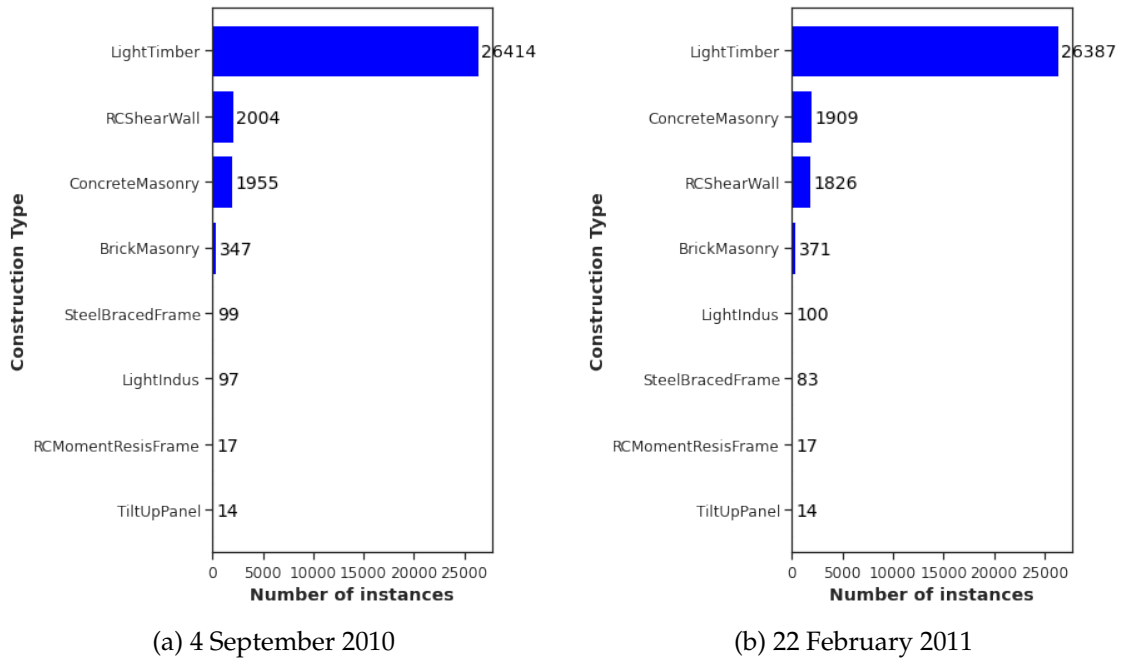


Figure 6.8: Number of instances for each Construction Type category

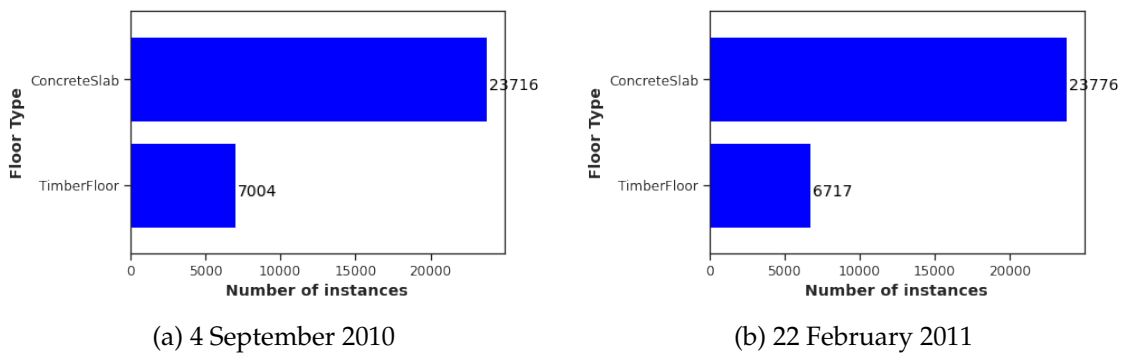


Figure 6.9: Number of instances for each category of Floor Type

Floor Type

Figure 6.9 shows the number of instances per building floor type. The attribute has two categories: concrete slab and timber floor. Sufficient instances are present in both categories such that no filtering was required.

Deprivation Index

The deprivation index attribute in the RiskScape data set describes the relative affluence of the neighbourhood where the building is located. Figure 6.10 shows the number of instances per deprivation index category. Nine of the ten categories are well represented. Only the category for the deprivation index 10 (most deprived) has a lower 279 instances for 4 September 2010 and 316 for 22 February 2011. Nevertheless, all data was kept

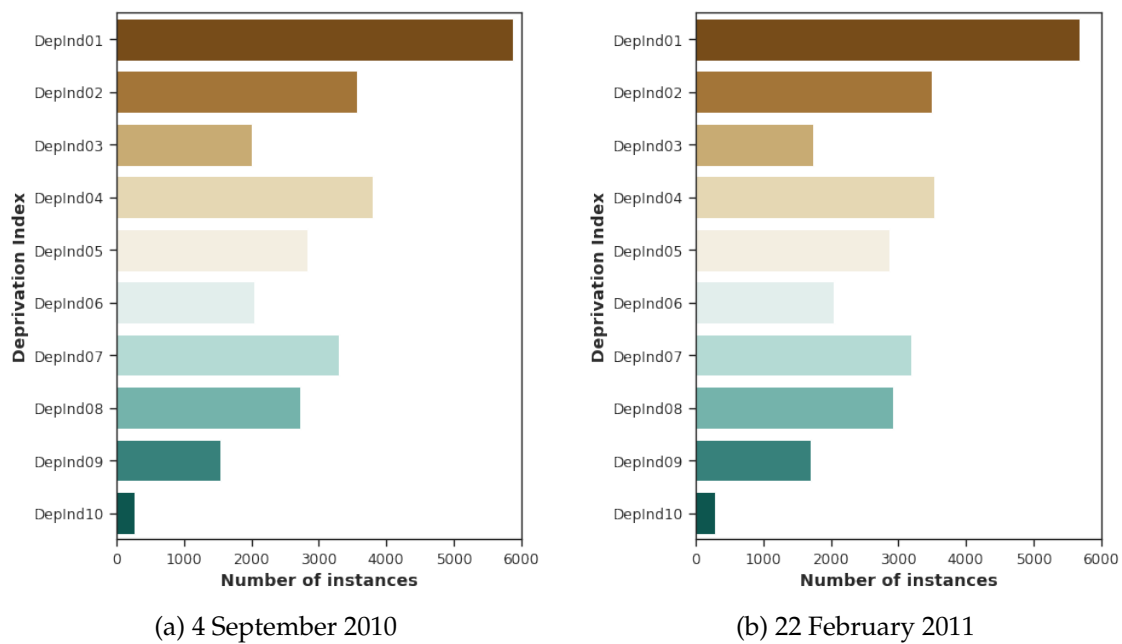


Figure 6.10: Number of instances for each category of Deprivation index (DepInd01 = least deprived to DepInd10=Most deprived)

in order to capture the full possible range of values related to the deprivation index attribute.

Wall Cladding

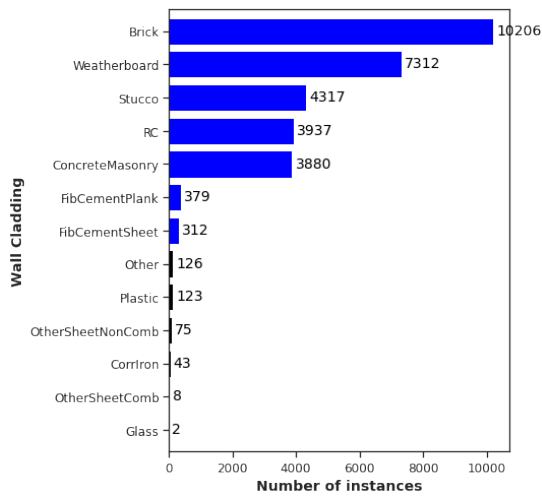
The RiskScape wall cladding attribute has fifteen categories. Thirteen of those are present in the data set for 4 September 2010 and 22 February 2011 (see Figure 6.11). Eight of these are under-represented: fibre cement sheet, fibre cement plank, plastic, other sheet – combustible and non-combustible, corrugated iron, glass, and other. The instances with ‘fibre cement sheet’ and ‘fibre cement plank’, were combined together and retained within the model, while the remaining categories were discarded.

Roof Cladding

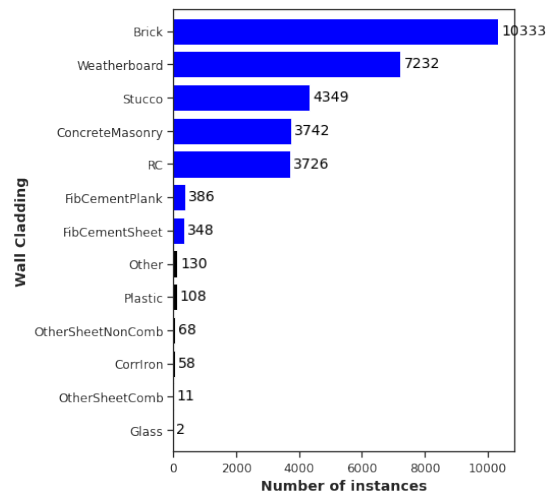
The attribute roof cladding differentiates seven types of roof material. Figure 6.12 number of instances per category for 4 September 2010 and 22 February 2011. The three most common roof material are sheet metal, clay/concrete tile, and metal tiles. There are insufficient entries for ‘other heavy’ and ‘concrete slab’ and these categories were thus discarded.

Soil Type

Figure 6.13 shows the number of instances per soil type. To ensure that the machine

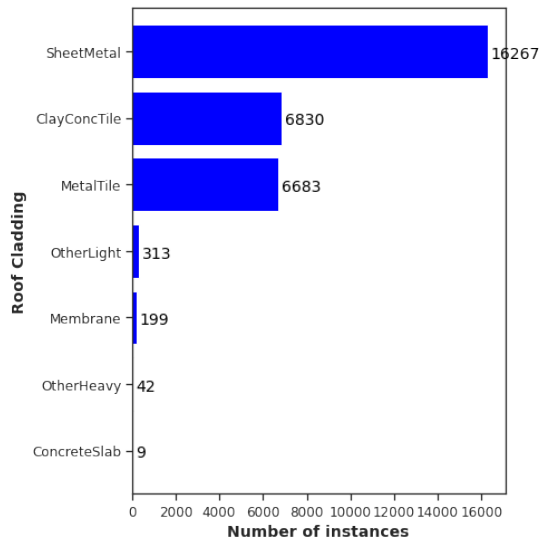


(a) 4 September 2010

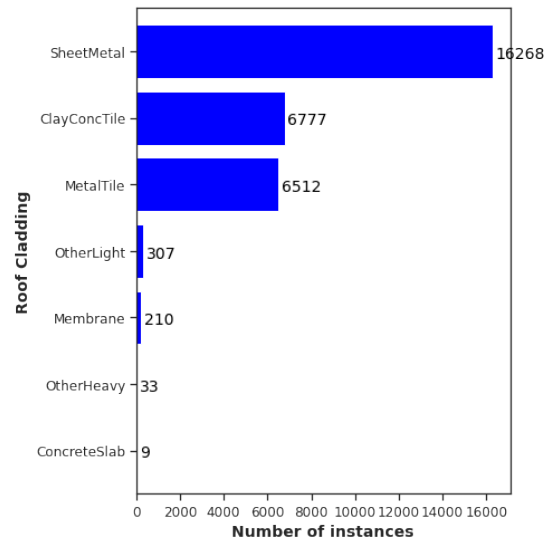


(b) 22 February 2011

Figure 6.11: Number of instances for each category of Wall Cladding



(a) 4 September 2010



(b) 22 February 2011

Figure 6.12: Number of instances for each category of Roof Cladding

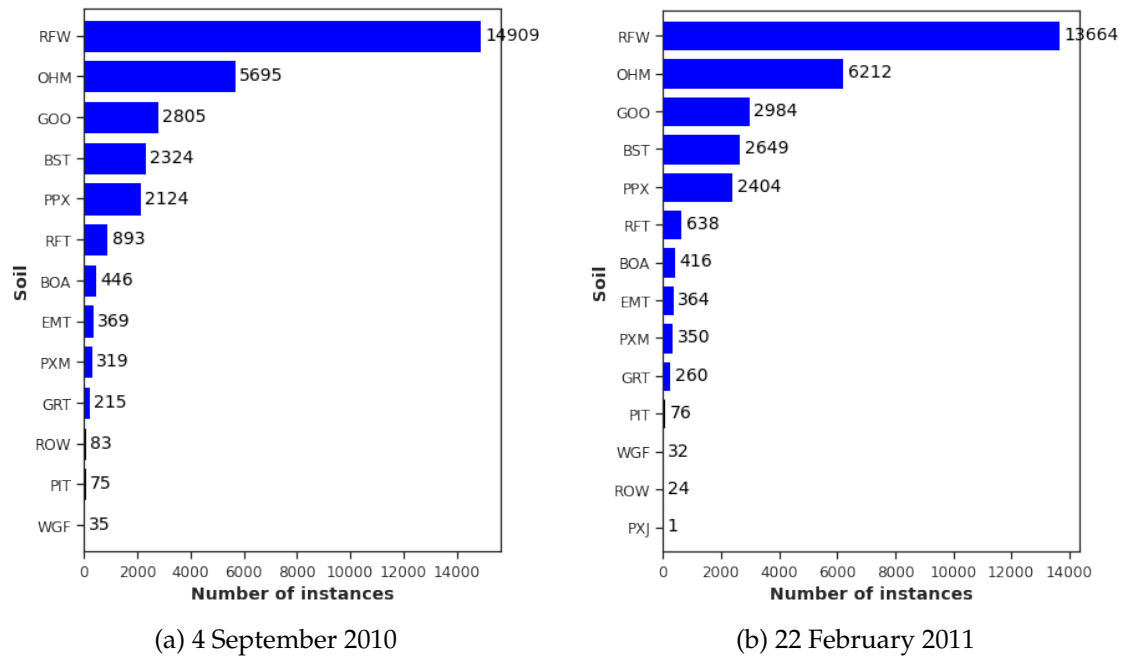


Figure 6.13: Number of instances for each soil category

learning can generalise, the soil types having less than a hundred instances for 4 September 2010 or 22 February 2011 were removed.

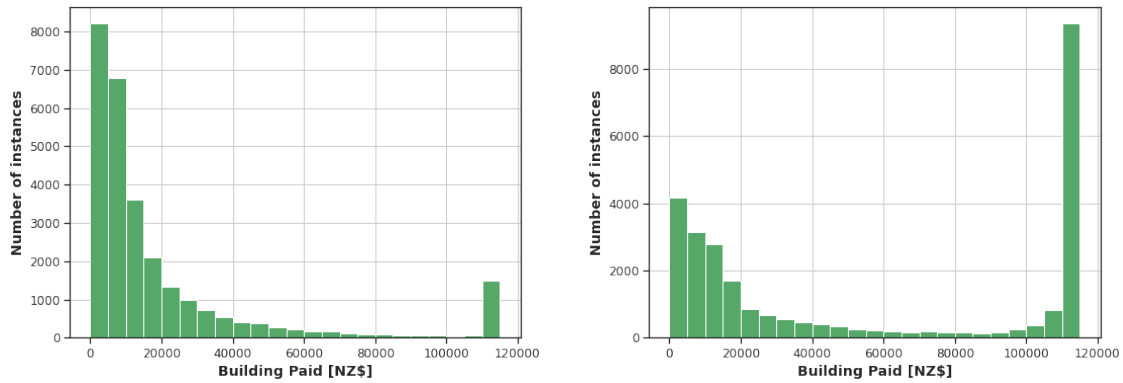
6.2.3 Filtering of the target attribute: Building Paid

Despite the data set having been previously filtered in section 6.2.1 to only have 'EQC Building Sum Insured' as exactly NZ\$115,000, some data instances still had 'BuildingPaid' greater than NZ\$115,000. It was even discovered that some instances had negative values.

It was thus decided to only include data instances with BuildingPaid between NZ\$0 and NZ\$115,000. Figure 6.14 shows the distribution of Building Paid within the selected range for 4 September 2010 and 22 February 2011. The overall distribution is relatively similar for both events with many buildings below NZ\$20,000, few claims between NZ\$20,000 and NZ\$110,000, and many instances close to the NZ\$115,000 cap. However, the 22 February 2011 event led to more claims that reached the maximum cover.

6.2.4 Evolution of the number of points during the feature filtering

Figure 6.15 presents a graphical overview of the further data verification through feature filtering after the filtering from database merging. Figure 6.16 shows the evolution of the



Step 6: Data pre-processing



RiskScape & additional variables

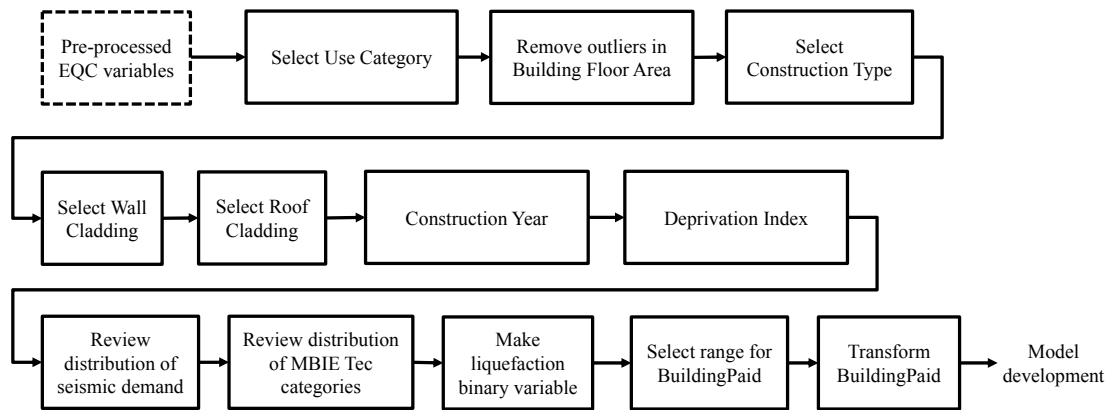


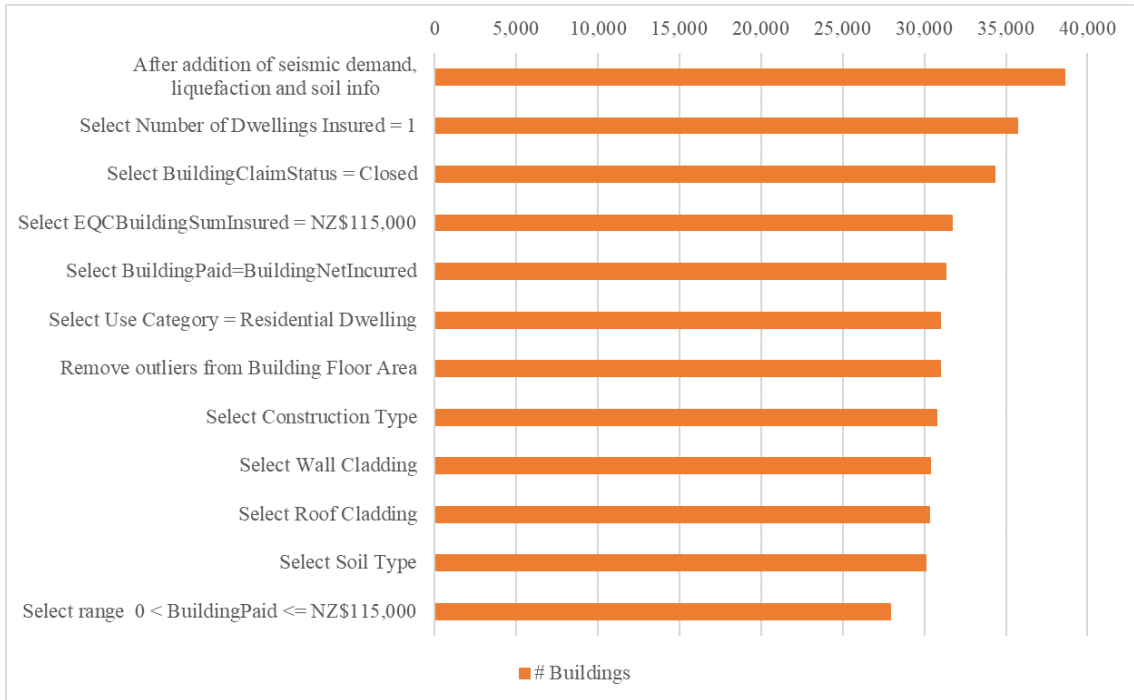
Figure 6.15: Overview of the filtering steps for the RiskScape attributes

number of instances after each filtering operation for 4 September 2010 and 22 February 2011 data. 27,932 instances for 4 September 2010 and 27,479 instances for 22 February 2011 remain following feature filtering.

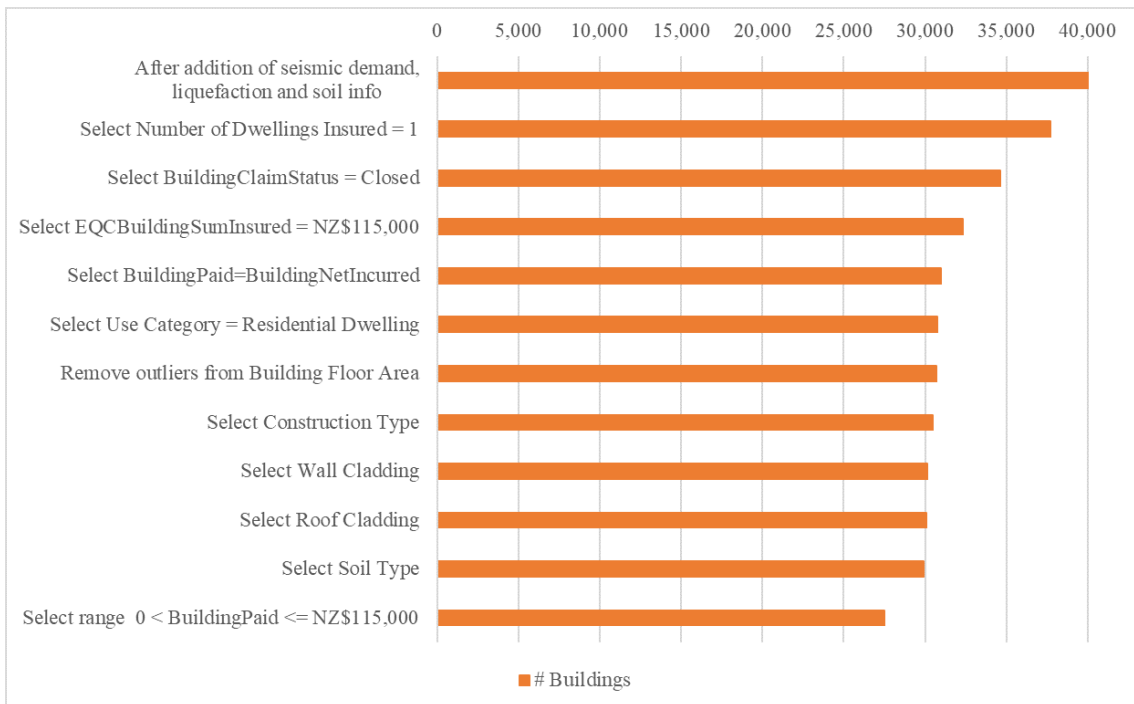
6.3 Processing of the target attribute

6.3.1 Building loss ratio or Building Paid

An attempt was made to create a ‘Building Loss Ratio’ target attribute, a concept similar to the Building Damage Ratio (BDR) proposed by J. Russell and van Ballegooy (2015). The ‘Building Loss Ratio’ is the ratio of ‘Building Paid’ over the ‘Modelled Dwelling Value’, where ‘Building Paid’ is the total cash paid by EQC to date and the ‘Modelled Dwelling Value’ is the dwelling value in NZ\$ modelled by EQC. Using the ‘Building



(a) 4 September 2010



(b) 22 February 2011

Figure 6.16: Evolution of the number of instances after each feature filtering step

Loss Ratio' rather than 'Building Paid' creates a nondimensionalised target feature that is independent of the residential property value. However, the 'Modelled Dwelling Value' attribute was frequently missing for many residential buildings, further there is great uncertainty in this attribute. It was thus decided to simply use Building Paid only as the target attribute.

6.3.2 Cap

At the time of the CES in 2010-2011, EQC's liability was capped to the first NZ\$100,000 (+GST) (NZ\$115,000) of building damage. Costs above this cap were borne by private insurers if building owners previously subscribed to adequate insurance coverage. Private insurer could not disclose information on private claims settlement, leaving the claims database for this study soft-capped at NZ\$115,000 for properties with over NZ\$100,000 (+GST) damage.

6.3.3 Transform BuildingPaid to a categorical attribute

In the original EQC claims data set, 'Building Paid' is a numerical attribute. Initial modelling attempts using of 'BuildingPaid' as a numerical target variable produced poor model predictions in terms of both accuracy or ability for generalisation. 'Building Paid' was thus transformed into a categorical attribute. Rather than defining the category thresholds randomly, thresholds for the cut-offs were chosen according to the EQC definitions related to limits for cash settlement, the Canterbury Home Repair Programme, and the maximum coverage provided (Earthquake Commission (EQC), 2019a).

Any instances with less than and equal to NZ\$11,500 is classified as the category 'low', reflecting the limit of initial cash settlement consideration.

Next, while the maximum EQC building sum insured was at NZ\$115,000, it was found that many instances that were over cap showed a Building Paid value close to but not exactly at NZ\$115,000. In consultation with the risk modelling team at EQC, the threshold for the category 'Over Cap' was set at NZ\$113,850 as this represents the actual cap value, nominal cap value minus 1% excess.



Figure 6.17: Schematic overview of the thresholds for the transformation of Building Paid from a categorical to a numerical attribute

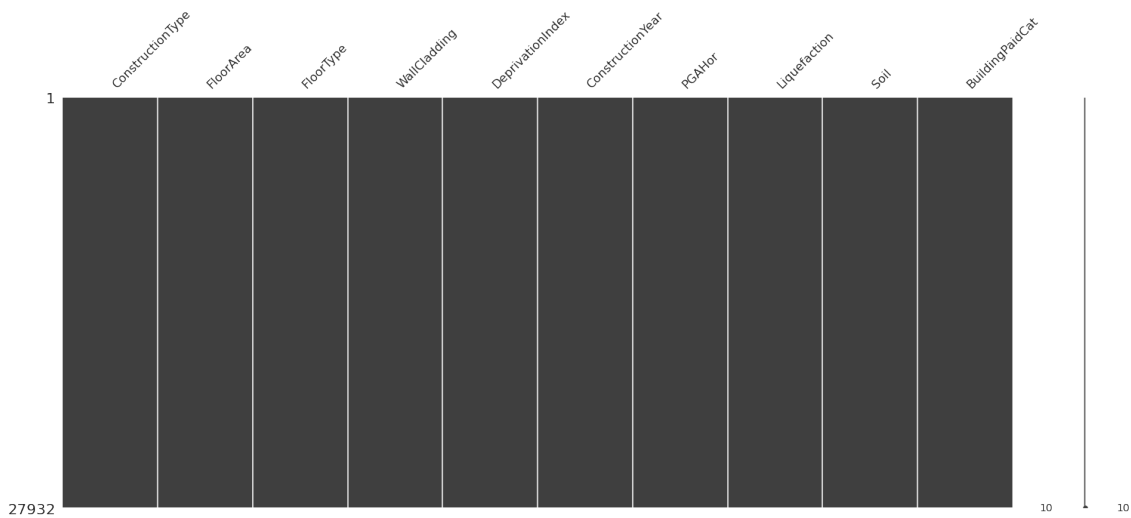


Figure 6.18: Selected attributes for the model using data from 4Sep2010

Instances with Building Paid values between NZ\$11,500 and NZ\$113,850 were subsequently assigned the category 'medium'. Figure 6.17 shows a schematic overview of the thresholds used to transform Building Paid from a numerical attributes into three categorical attributes.

6.4 Attribute selection

Nine attributes plus the target variable Building Paid Category were selected for the model development. Figure 6.18 shows a graphical overview of the attributes selected and the number of instances for the pre-processed data. None of the columns in Figure 6.18 show any white rows confirming that the data is complete with no missing value for all the instances.

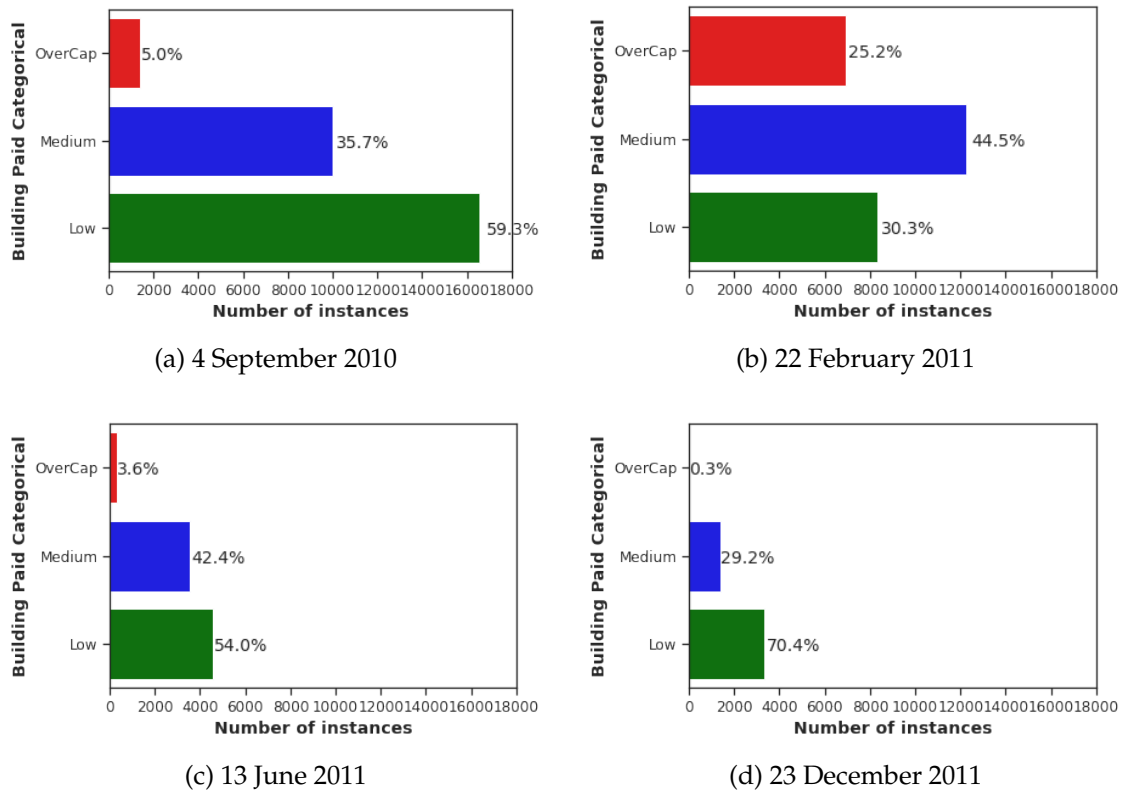


Figure 6.19: Number of instances in Building Paid categorical in the filtered data set

6.4.1 Target attribute: Building Paid - Categorical

As described in section 6.3.3, the numerical attribute 'Building Paid' was transformed into three categories depending on the value of 'Building Paid'. Figure 6.19 shows the number of instances in each category for the four main events of the CES.

6.4.2 Liquefaction

Widespread liquefaction occurred mainly during the 22 February 2011 and 13 June 2011 events. The 4 September 2010 and 23 December 2011 events experienced liquefaction but to a limited extent (see Figure 5.3). Figure 6.20 shows the number of claims for which buildings experienced liquefaction, or not, for the main events in the CES.

6.4.3 PGA

Figure 6.21 shows the distribution of the PGA values for the main events in the CES. The majority of the residential buildings experienced less than 0.60 g of PGA for all events. Only during 22 February 2011 and 13 June 2011, a small number of buildings

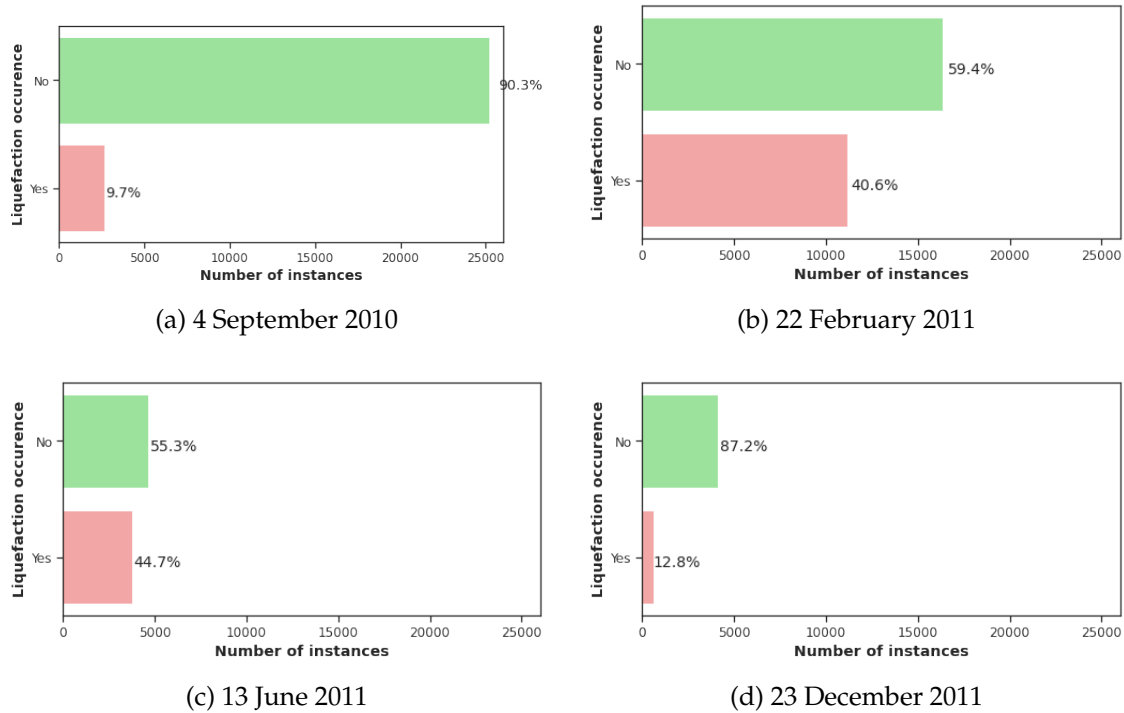


Figure 6.20: Number of instances in the filtered data set which experienced liquefaction

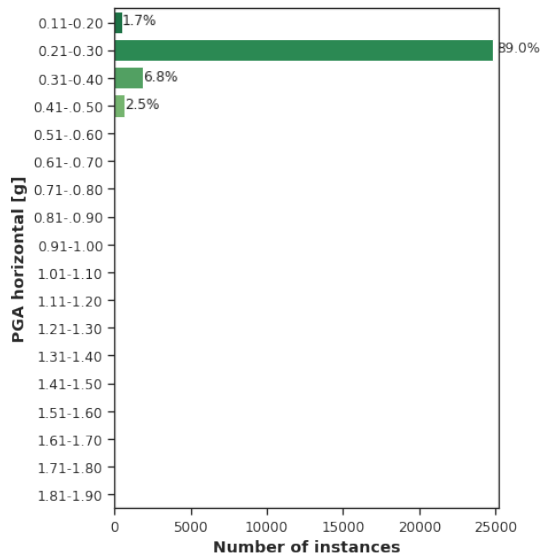
experienced higher accelerations. The 4 September 2010 event and 23 December 2011 event generated PGA levels between 0.10 g and 0.50 g, however, for both events, over 90% of the buildings experienced PGA values below 0.30 g (see Figure 6.21a and Figure 6.21d). The 22 February 2011 and 13 June 2011 led to PGA values up to 1.34 g and 1.88 g respectively (see Figure 6.21b and Figure 6.21c). For the 22 February 2011, 86% of the damaged building were in the range 0.21 g to 0.60 g, while for the 13 June 2011 84% of the residential buildings for which a claim has been lodged experienced values below 0.40 g.

6.4.4 Construction Type

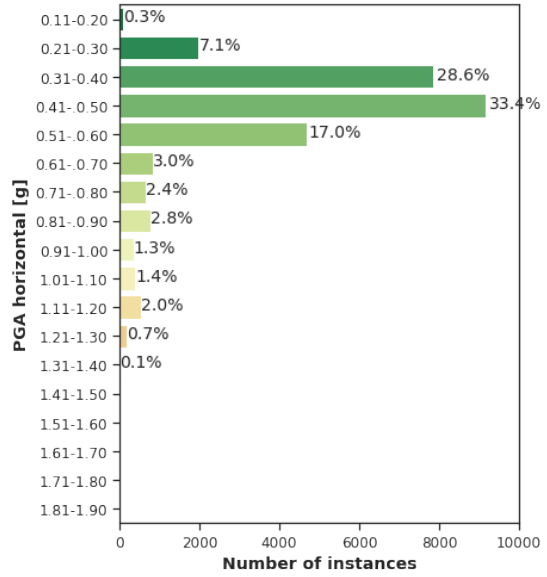
Figure 6.22 shows the number of instances per construction type. On average, 86.1% of the residential buildings damaged are made out of light timber, 6.5% are RC shear wall buildings, 6.3% are concrete masonry houses, and 1.1% are brick masonry buildings.

6.4.5 Building Floor Area

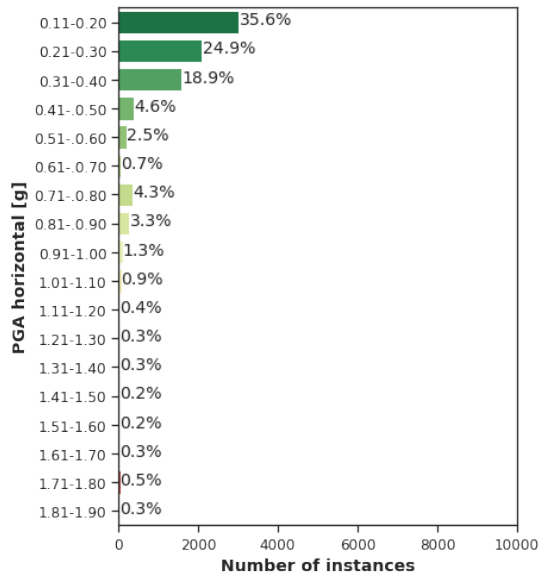
Figure 6.23 shows the distribution of houses by building floor area.



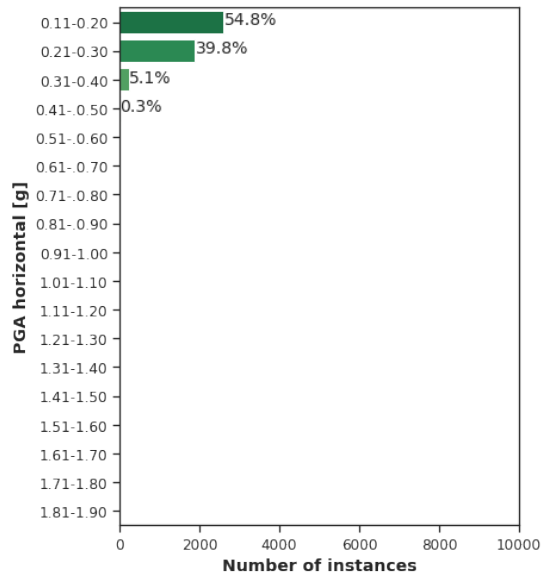
(a) 4 September 2010



(b) 22 February 2011

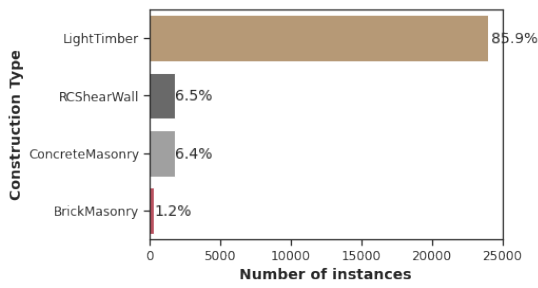


(c) 13 June 2011

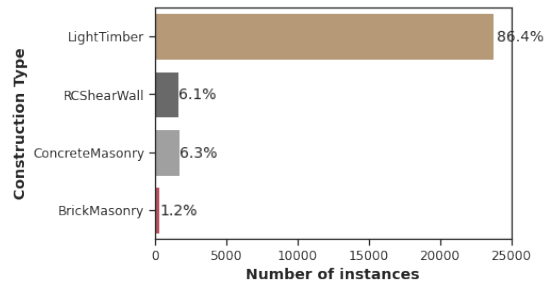


(d) 23 December 2011

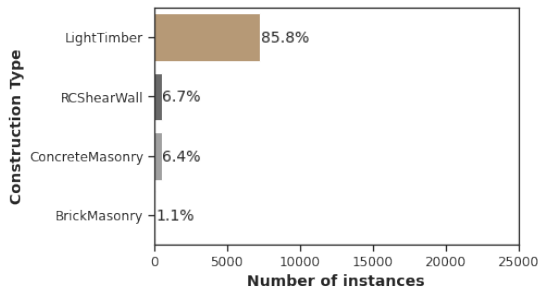
Figure 6.21: Distribution of PGA in the filtered data set



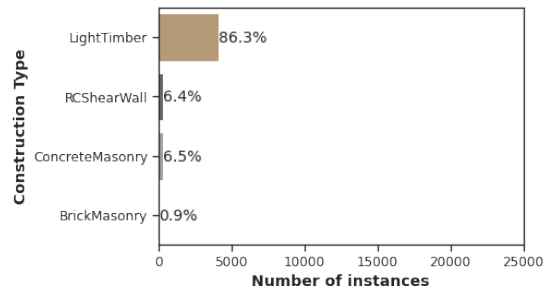
(a) 4 September 2010



(b) 22 February 2011

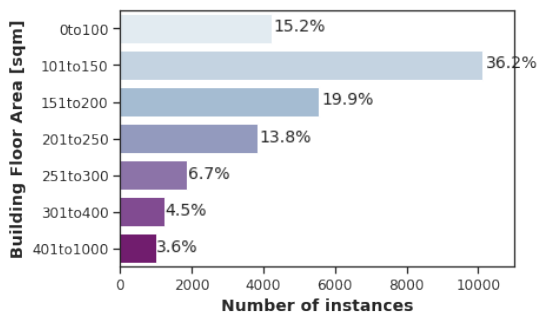


(c) 13 June 2011

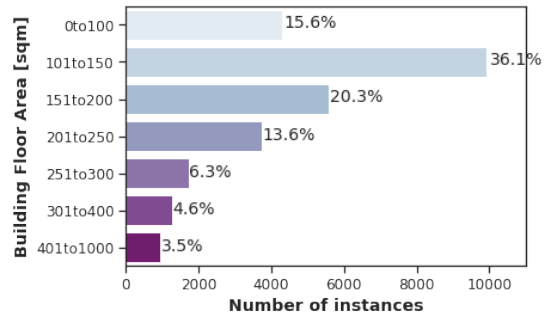


(d) 23 December 2011

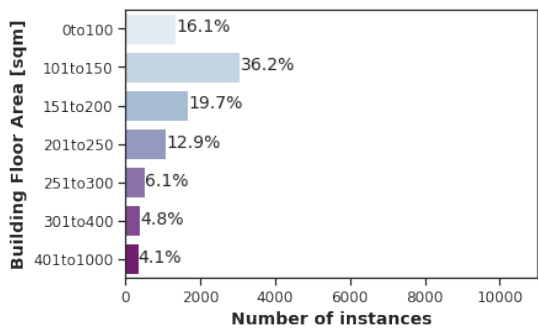
Figure 6.22: Number of instances per Construction Type in the filtered data set



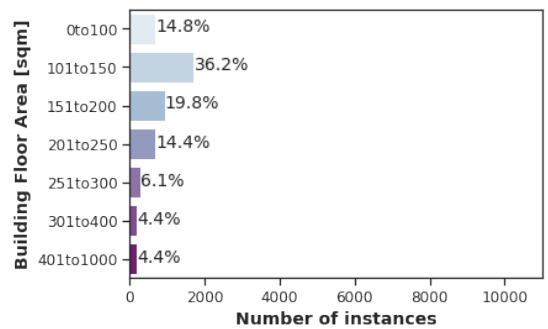
(a) 4 September 2010



(b) 22 February 2011



(c) 13 June 2011



(d) 23 December 2011

Figure 6.23: Number of instances by Building Floor Area in the filtered data set

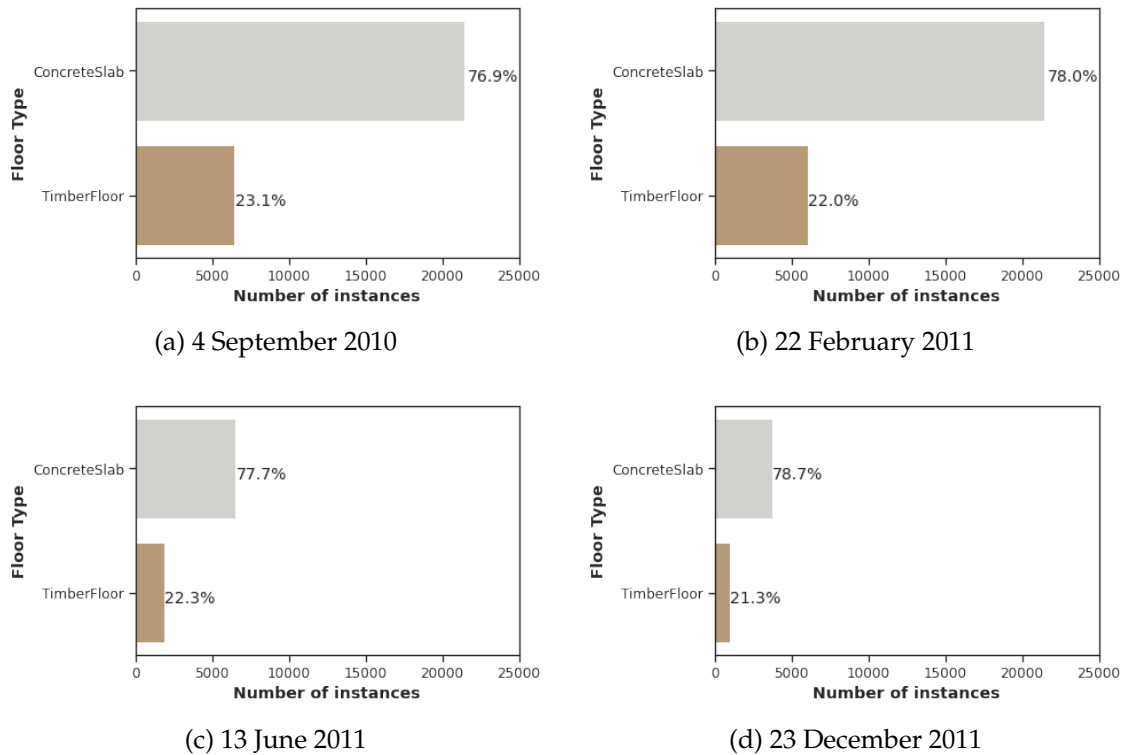


Figure 6.24: Number of instances by Building Floor Area in the filtered data set

6.4.6 Floor Type

Figure 6.24 shows the number of building having a timber floor or concrete slab as floor type. On average 77.8% of the damaged buildings have a concrete slab and 22.2% a timber floor.

6.4.7 Wall Cladding

Figure 6.25 shows the number of instances by wall cladding class. With 34% on average, brick is the most common wall cladding type. Weatherboard the second cladding type represent 23.5% of the instances. Stucco, reinforced concrete, and concrete masonry account for 14.1%, 12.8% and 13.2% respectively. Fibre cement (plank and sheet) are found in less than 3% of the buildings.

6.4.8 Deprivation Index

Appendix E shows a map of the deprivation index in urban Christchurch. The neighbourhoods east of Christchurch's CBD are recorded the most deprived. Figure 6.26

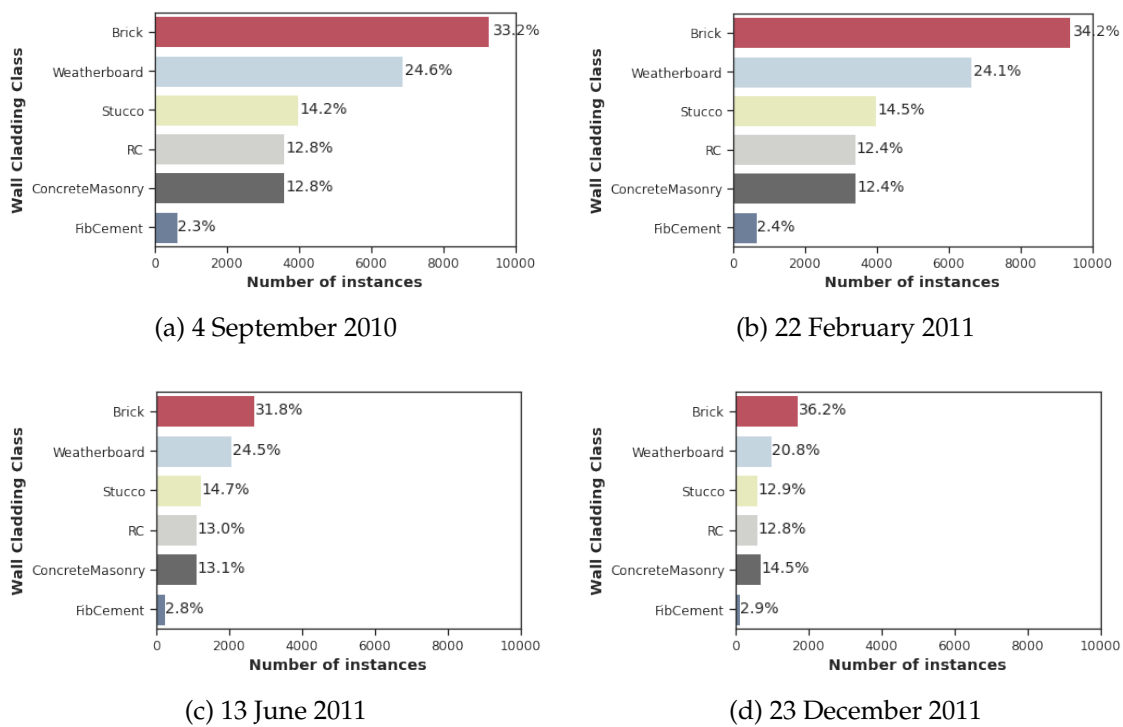


Figure 6.25: Number of instances by Building Floor Area in the filtered data set

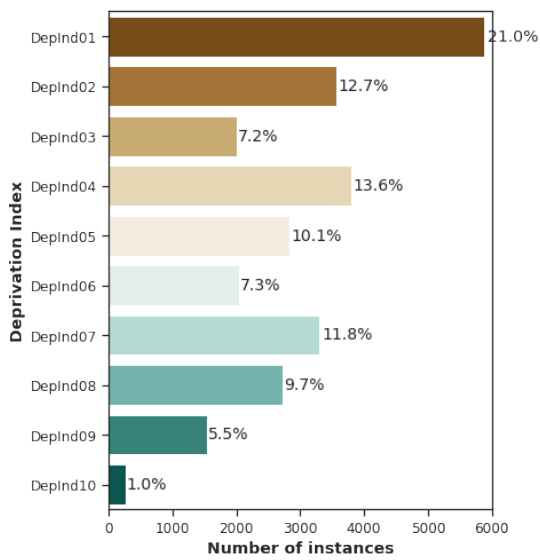
shows the distribution of the deprivation index of the buildings for the four key events in the CES. A key to note is that distribution amongst deprivation index remains similar.

6.4.9 Construction year

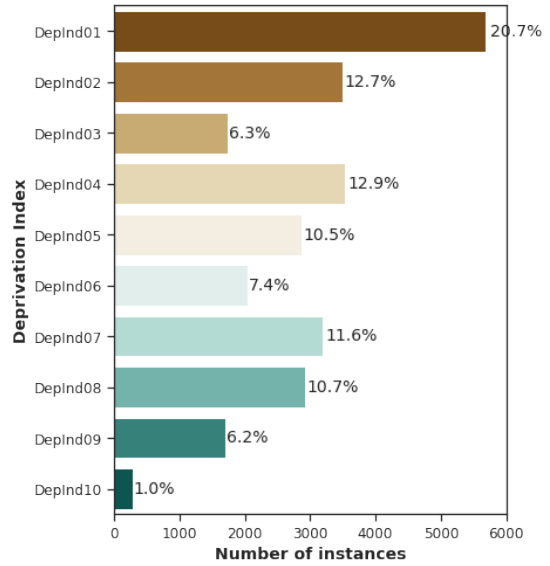
Figure 6.27 shows the number of instances per period of construction. The oldest building dates from 1888. The majority of the dwellings were constructed after 1950, with approximately 28% built between 1953 and 1972, 20% between 1973 and 1992, and 26% constructed after 1993.

6.4.10 Soil type

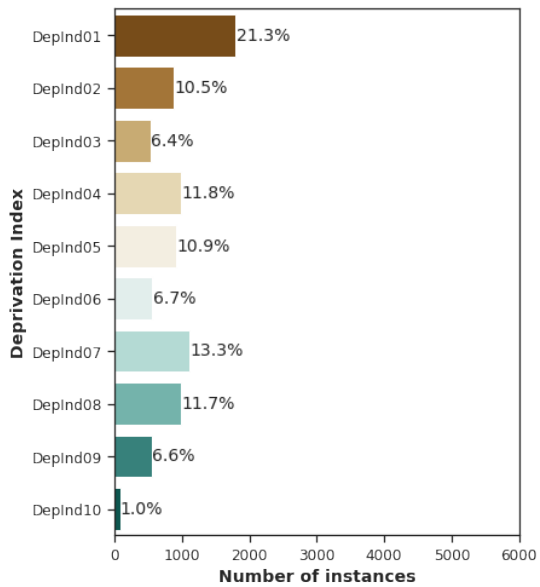
Figure 6.28 shows the number of buildings classified by soil code according to the soil map for the Upper Plains and Downs of Canterbury (Land Resource Information Systems (LRIS), 2010). The soil type recent fluvial (RFW), organic humic (OHM), gley orthic (GOO), brown sandy (BST), and pallic perch-gley (PPX) are the most represented. RFW is the most common with 47.6% of the building built on recent fluvial soil. OHM follows with 19.6% of the instances constructed on this soil type. GOO, BST, and PPX, the next represented categories account for 10.1%, 9.0%, and 6.9% respectively.



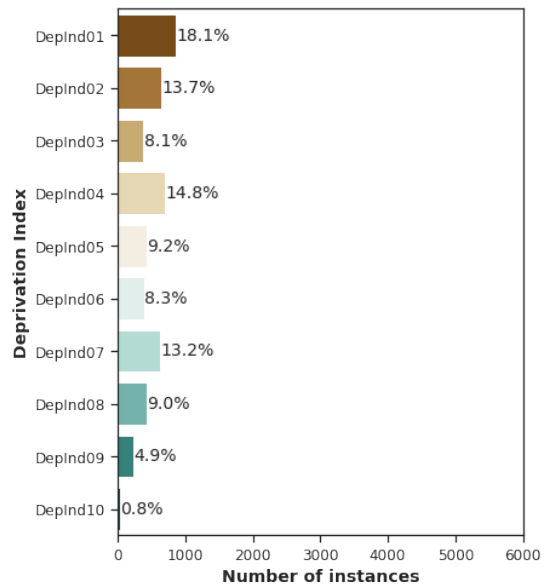
(a) 4 September 2010



(b) 22 February 2011

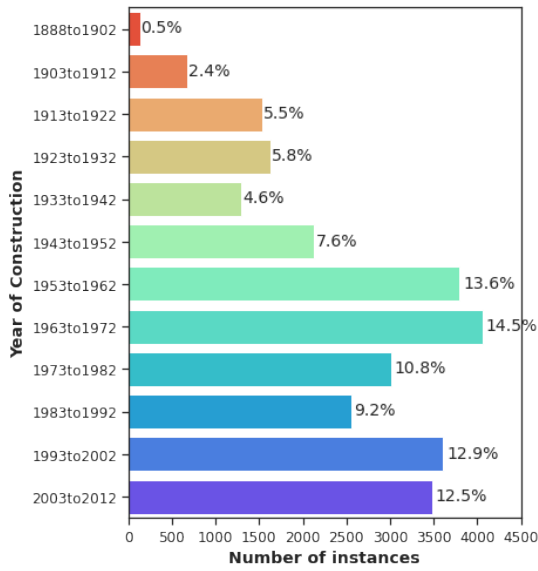


(c) 13 June 2011

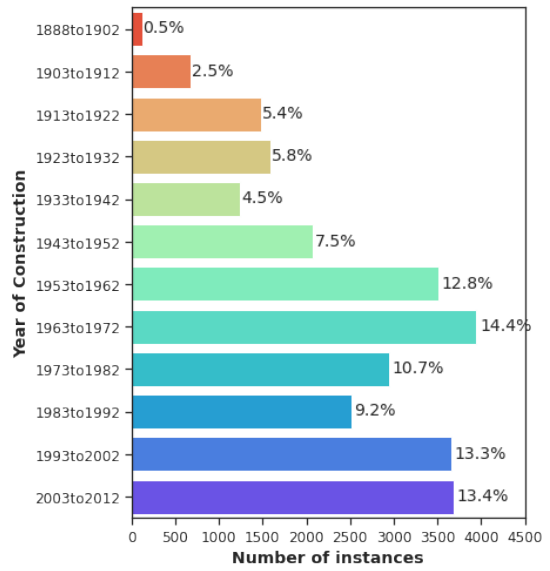


(d) 23 December 2011

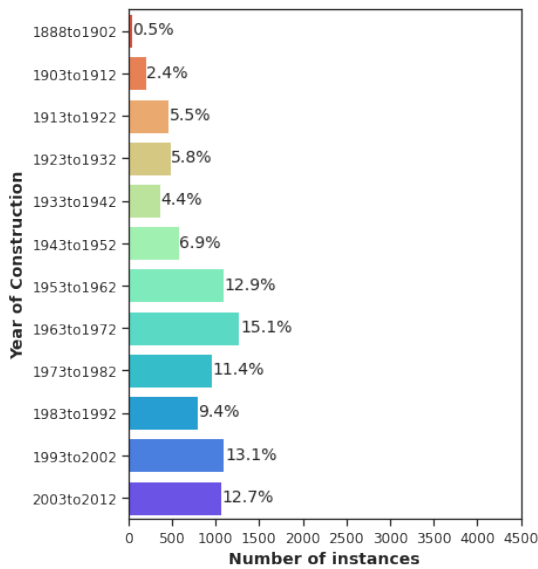
Figure 6.26: Number of instances by Deprivation Index in the filtered data set



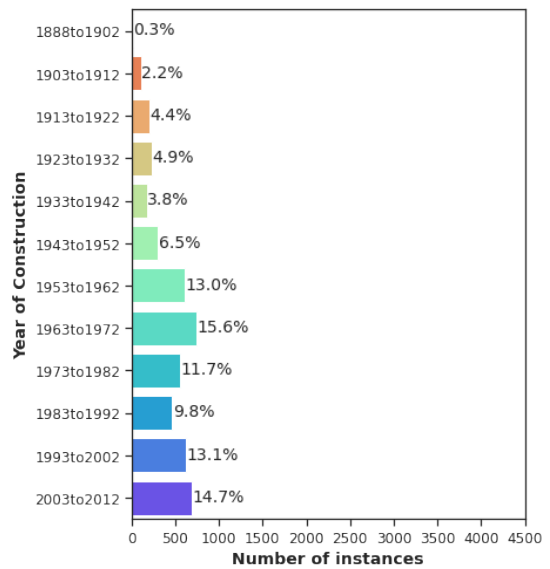
(a) 4 September 2010



(b) 22 February 2011

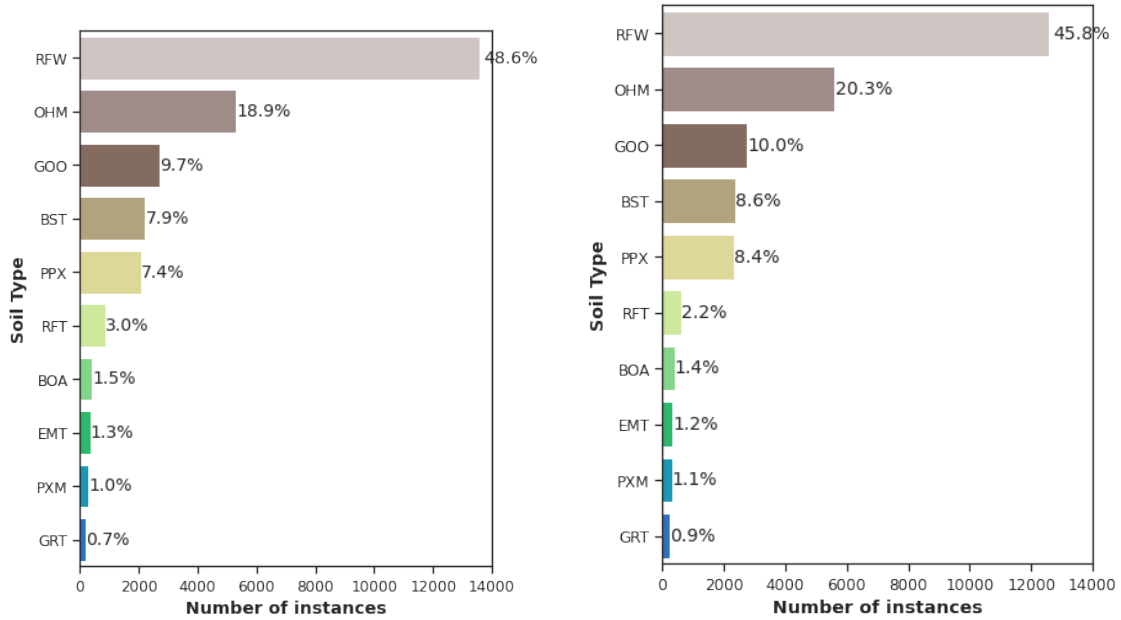


(c) 13 June 2011



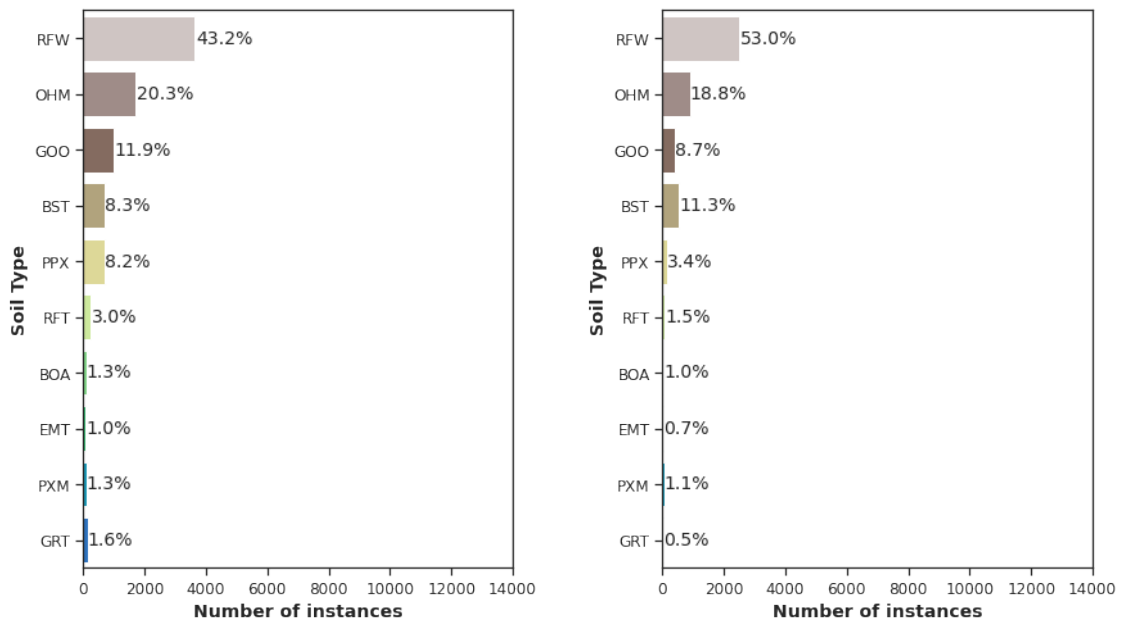
(d) 23 December 2011

Figure 6.27: Number of instances by construction period in the filtered data set



(a) 4 September 2010

(b) 22 February 2011



(c) 13 June 2011

(d) 23 December 2011

Figure 6.28: Number of instances per soil type in the filtered data set

6.4.11 Latitude and Longitude

Information on the building coordinates (latitude and longitude) are available in the merged data set. For some machine learning prediction models, latitude and longitude might deliver useful information. For house price prediction for instance, building coordinates can convey background information such as the walkability and the desirability of the neighbourhood (i.e. wealth of the neighbourhood, quality of the local elementary school). Retaining latitude and longitude in such models might thus incorporate hidden insights and increase the predictive accuracy (Géron, 2019; Ma et al., 2020).

In a house prediction model, as the house is stationary, the latitude and longitude attribute convey the same background information for each model over time (i.e. the latitude and longitude attributes have the same meaning today as for future house price predictions). However, in earthquake engineering, the epicentre location is changing for different earthquakes such that for each event, the background information related to the attributes latitude and longitude is not the same. The latitude and longitude attributes are thus not retained in this model. Notably, information such as seismic demand, liquefaction occurrence, and soil conditions are captured directly through other attributes that are generalisable for all earthquake events.

6.5 Attribute preparation

6.5.1 Training, validation, and test set

For machine learning, the data is split into three distinct sets, the training, validation (or development), and test set. Figure 6.29 shows a schematic of the splitting and their use in the development of the seismic loss model. The training and validation sets are coming from the same data set using 80% of the data for training and 20% for validation. The 4 September 2010 pre-processed data has 27,932 instances. Thus, there are 22,345 instances in the training set and 5,587 instances in the validation set. The 22 February 2011 entails 27,479 total instances, thus leading to 21,983 examples in the training set and 5,496 in the validation set.

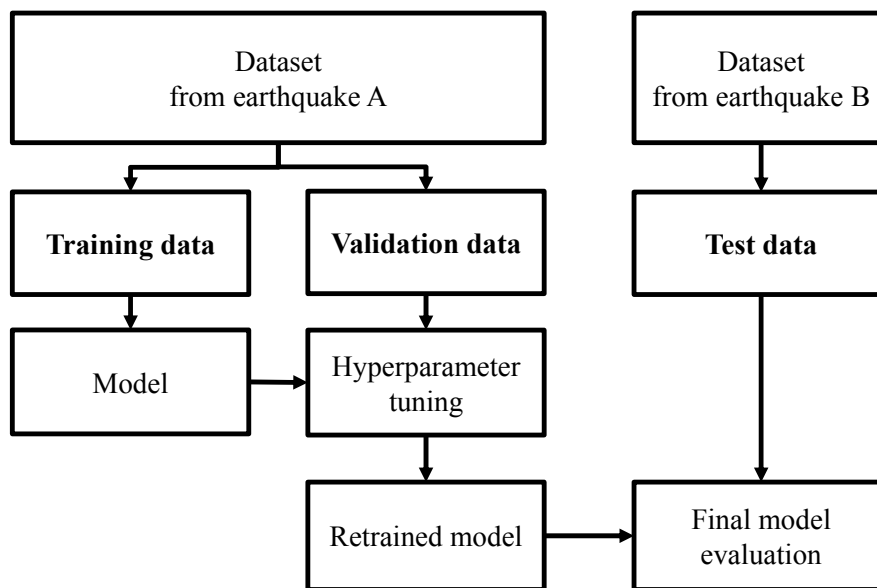


Figure 6.29: Overview of the training, validation, and test data sets and their usage in the development of a machine learning seismic loss model for Christchurch

Unlike the 'traditional approach' where the test set is held out from the same data as the training and validation set, the test set here employed comes from another earthquake. Testing the model using data from another earthquake in the CES (pre-processed in the same way as the training and validation set) enables to evaluate the model capacity to generalise to other events. Thus changing the earthquake from which the input and test data set comes from, it is possible to study multiple combinations and find the model which works the best for the entire CES. This approach to evaluate the model performance using test data from another earthquake event is explained in section 7.2.

6.5.2 Handling categorical features

Categorical attributes are transformed into binary arrays for adoption by machine learning algorithms. For the model in this study, strings in categorical features were first transformed into an ordinal integer using the scikit-learn Ordinal encoder. Once converted to integers, the scikit-learn One Hot Encoder was used to encode the categorical features as one-hot numeric array.

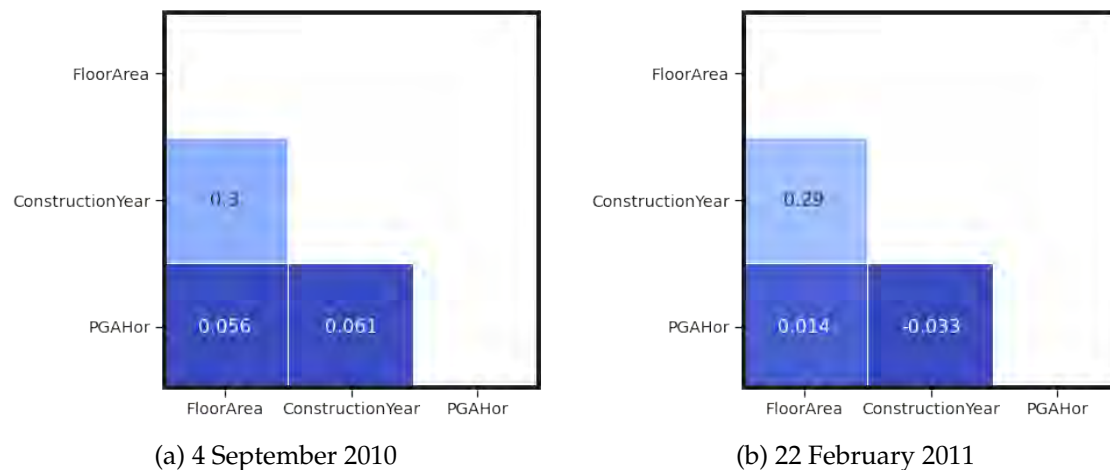


Figure 6.30: Correlation matrix between the numerical features

6.5.3 Handling numerical features

Numerical features are checked against each other for correlation prior to the machine learning training. If two features are correlated, best practice is to remove one of them. Figure 6.30 presents correlation matrices showing the Pearson correlation coefficients between the numerical attributes. It confirmed that there is no significant linear correlation between selected numerical attributes.

The numerical data is also normalised prior to the training process according to best practice. This step is called feature scaling. The most common feature scaling techniques are min-max scaling (also called normalisation) and standardisation. Both these techniques are implemented in scikit-learn. In this study, a min-max scaling (normalisation) approach was used to scale the numerical features.

6.5.4 Addressing class imbalance

Figure 6.19a shows the number of instances for each category in the target variable Building Paid for the 4 September 2010 data. While the categories 'low' and 'medium' have respectively 16,558 and 9,970 instances, the category 'over cap' has only 1,404 instances. The 'over cap' category is thus the minority class with a significant difference in the number of instances compared to the two other categories. Training a machine learning algorithm using the data in this form would lead to poor modelling performance for the over cap category. Thus, before training the model, the imbalanced-learn Python toolbox (Lemaitre et al., 2017a) was applied to address the class imbalance.

The toolbox encompasses several under-sampling and over-sampling techniques, however not all of them apply to multi-class problem (see Figure 2.23 in section 2.7.5). The over-sampling and under-sampling techniques suitable for multi-class problem were trialled (random over-sampling (ROS), cluster centroids (CC) or random under-sampling (RUS)). For the 4 September 2010 data, ROS delivered the best results regarding the overall model predictions as well as the prediction for the minority class over cap.

6.6 Algorithm selection and training

The model is trained using the merged data set which included information on the model attributes as well as the target attribute 'BuildingPaidCat', thus making the training a supervised learning task. Given the nine attributes selected for the model development (see Figure 6.18), the objective of the model is to predict if a building will fall within the category 'low', 'medium', or 'over cap' (expressed via the target variable 'BuildingPaidCat') thus leading to a categorical model for three classes. Several machine learning algorithms can perform supervised learning task for categories (e.g. logistic regression, support vector machine (SVM), artificial neural networks (ANN)). Those algorithms differentiate themselves by their complexity. More complex algorithms can develop more detailed models with a potential improved prediction performance, but complex algorithms are also more prone to overfitting.

For this study, the prediction performance is an essential metric. Nevertheless, the human interpretability of the model is also of significant interest as the goal is to produce a 'grey-box' model enabling for the derivation of insights. Unfortunately, not all algorithms are intrinsically interpretable. Table 2.2 in section 2.10 shows a list of some intrinsic interpretable machine learning algorithms. In this project, logistic regression, decision trees, SVM and random forest classification algorithms were trialled. Once the model is trained, the hyperparameters of the algorithms were tuned (see Figure 6.29). The performance of each model is presented in the following section.

6.7 Model evaluation

Table 6.1 shows the performance of logistics regression, SVM, decision trees, and random forest machine learning models trained and validated with data from 4 September 2010. Random forest stood out for its overall prediction performance. It achieved an accuracy of 0.62 on the validation set despite over-sampling of the underrepresented class and careful consideration of overfitting via tuning of the hyperparameters. Deemed as the best performing algorithm, random forest was thus selected for the machine learning model in this study.

Table 6.2 presents the performance of the Random Forest algorithm for training and validation sets from the 4 September 2010, 22 February 2011, 13 June 2011, and 23 December 2011. Despite the thorough attribute filtering, attribute selection, attribute preparation, and model development addressing class imbalance and carefully checking for under- and over-fitting, the prediction accuracy of the Random Forest algorithm on the validation set for all the four events did not exceed 0.63.

There are numerous possible reasons for the limited model performance. For three of the four events, there is a significant class imbalance between the classes of the target variable with over cap instances being mostly underrepresented. Despite the use of the Python imbalance toolbox to address the imbalance, having more instances in the over cap category would be beneficial. Similarly, having more direct information collected on-site about the building characteristics would improve the completeness of the EQC data set, which could benefit the model performance.

Figure 6.18 shows the target variable and nine selected model attributes. These attributes were selected based on domain knowledge as possible features that could affect the building losses. There may be other attributes that were not considered in this study that have direct and indirect effects on the value of a claim. Machine learning can only generalise and make new predictions if the selected model attributes are relevant to the problem. It is thus possible that the inclusion of additional attributes might be beneficial to the overall model accuracy.

Table 6.1: Model evaluation for logistic regression, SVM, decision trees, and Random Forest for the 4 September 2010 data

Algorithm	Set	Prediction targets	Precision	Recall	F ₁ score	Accuracy
Logistic regression	Train Set	Low	0.71	0.62	0.66	0.55
		Medium	0.48	0.42	0.45	
		Over cap	0.15	0.47	0.22	
		Accuracy on the train set				
	Validation Set	Low	0.71	0.63	0.67	0.55
		Medium	0.50	0.43	0.46	
		Over cap	0.14	0.46	0.21	
		Accuracy on the validation set				
Support vector machine (SVM)	Train Set	Low	0.59	0.60	0.60	0.63
		Medium	0.60	0.53	0.56	
		Over cap	0.69	0.75	0.72	
		Accuracy on the train set				
	Validation Set	Low	0.72	0.58	0.64	0.52
		Medium	0.48	0.44	0.46	
		Over cap	0.12	0.50	0.20	
		Accuracy on the validation set				
Decision tree	Train Set	Low	0.58	0.60	0.59	0.62
		Medium	0.58	0.48	0.53	
		Over cap	0.68	0.77	0.73	
		Accuracy on the train set				
	Validation Set	Low	0.71	0.58	0.64	0.52
		Medium	0.48	0.43	0.45	
		Over cap	0.12	0.48	0.19	
		Accuracy on the validation set				
Random forest	Train Set	Low	1.00	1.00	1.00	1.00
		Medium	1.00	1.00	1.00	
		Over cap	1.00	1.00	1.00	
		Accuracy on the train set				
	Validation Set	Low	0.67	0.78	0.72	0.61
		Medium	0.50	0.40	0.44	
		Over cap	0.27	0.12	0.17	
		Accuracy on the validation set				

Table 6.2: Model evaluation for Random Forest model for 4Sep2010, 22Feb2011, 13June2011, and 23Dec2011

Algorithm	Set	Prediction targets	Precision	Recall	F ₁ score	Accuracy
Random forest 4Sep2010	Train Set	Low	1.00	1.00	1.00	1.00
		Medium	1.00	1.00	1.00	
		Over cap	1.00	1.00	1.00	
	Accuracy on the train set					
	Validation Set	Low	0.67	0.78	0.72	0.61
		Medium	0.50	0.40	0.44	
		Over cap	0.27	0.12	0.17	
Accuracy on the validation set						
Random forest 22Feb2011	Train Set	Low	1.00	1.00	1.00	1.00
		Medium	1.00	1.00	1.00	
		Over cap	1.00	1.00	1.00	
		Accuracy on the train set				
	Validation Set	Low	0.62	0.51	0.56	0.54
		Medium	0.51	0.68	0.58	
		Over cap	0.55	0.32	0.41	
Accuracy on the validation set						
Random forest 13Jun2011	Train Set	Low	1.00	1.00	1.00	1.00
		Medium	1.00	1.00	1.00	
		Over cap	1.00	1.00	1.00	
		Accuracy on the train set				
	Validation Set	Low	0.58	0.65	0.61	0.53
		Medium	0.47	0.43	0.45	
		Over cap	0.20	0.06	0.10	
Accuracy on the validation set						
Random forest 23Dec2011	Train Set	Low	1.00	1.00	1.00	1.00
		Medium	1.00	1.00	1.00	
		Over cap	1.00	1.00	1.00	
		Accuracy on the train set				
	Validation Set	Low	0.70	0.84	0.77	0.63
		Medium	0.24	0.13	0.17	
		Over cap	0.00	0.00	0.00	
Accuracy on the validation set						

6.8 Conclusion

This chapter described the necessary data pre-processing and the machine learning development process. First, the EQC features were filtered to retain only claims that have been settled and related to single dwellings only. Outlier instances and categories with too few instances were removed and regarded beyond scope as those could have negatively affected the model performance. The target variable BuildingPaid was filtered and transformed into a categorical variable according to the thresholds of payments defined by EQC. Once filtered, the new categorical attribute was selected for the model development. Special attention was applied to ensure attributes selected do not related to any earthquake in particular to maximise the chance of generalisation. The data set was then divided into a training and a validation set. Categorical and numerical features were put in a form useable by machine learning algorithms. A solution to address class imbalance was introduced. Finally, several supervised algorithms for classification were selected, trained, and their performance evaluated. Following the thorough data preparation process and careful model development, random forest was deemed the best performing algorithm which achieved an accuracy of 0.61 on the 4 September 2010 validation set.

Model testing and knowledge extraction from the seismic loss prediction model for Christchurch residential buildings

This chapter documents the testing of the machine learning model developed in the past two chapters. The model generalisation and prediction performance are tested against data from other main events of the CES. The chapter also presents insights derived from the machine learning model.

7.1 Introduction

Seismic damage and loss models are developed to estimate consequences from probable future earthquake events. A key consideration is the ability of the model to be generalised for unknown events. Thus, each machine learning model developed previously using training data from one earthquake in the CES is tested against data from the three other main events. Besides the prediction, the development of the machine learning model also reveals relationships between the model attributes. The relationship of numerical variables is analysed and presented via pairplots. Then the relationships of Building Paid

with each model attribute are studied. The influence of liquefaction is carefully examined as it has a strong influence on the distribution of the building losses especially for the 22 February 2011 event. Finally, a post-hoc method (SHAP) is applied to the Random Forest algorithm to derive the features importance of the model.

Results show that PGA is selected by machine learning as the most important feature for each model. For the 22 February 2011 event, the liquefaction occurrence stands out as the second important feature. This is particularly novel as the machine learning process only analysed empirical data and it delivers these insights without any prior engineering knowledge about loss mechanisms or physics. These findings corroborate current earthquake engineering knowledge. This highlights the benefits of interpretable machine learning to derive new actionable insights.

7.2 Model testing on another event in the CES

The previous chapter presented the model development and training for each of the main earthquake events in the CES (4 September 2010, 22 February 2011, 13 June 2011, and 23 December 2011). The random forest algorithm performed the best for all the events. Figure 6.29 in section 6.5.1 showed the process for the model development. Each model is tested here on instances from the three other main events of the CES.

Figure 7.1a and Figure 7.1d show the confusion matrix for the random forest model for the 4 September 2010 and 22 February 2011 validated on the same event respectively. Figure 7.1b shows the confusion matrix for the random forest model developed with the 4 September 2010 data and tested on the 22 February 2011 instances. Figure 7.1c presents the confusion matrix for the random forest model developed with the 22 February 2011 data and tested on the 4 September 2010 instances. For each confusion matrix, the diagonal in the green area represents the correct predictions. The top integer numbers in each of the upper left boxes display the number of instances predicted, and the percentage in the bottom rows represent that instance as a percentage of the population. The closest the value on the diagonal sum to 100%, the better the prediction. Mistakenly predicted instances are shown off the diagonal. Figure 2.27 in section 2.9.1 summarises the interpretation of a confusion matrix.

		Predicted			Recall
		Low	Medium	OverCap	
Actual	Low	2596 46.47%	682 12.21%	42 0.75%	0.78
	Medium	1153 20.64%	798 14.28%	49 0.88%	
	OverCap	123 2.20%	111 1.99%	33 0.59%	
Precision		0.67	0.50	0.27	0.61

(a) 4 September 2010 model tested on 4 September 2010

		Predicted			Recall
		Low	Medium	OverCap	
Actual	Low	6290 22.89%	1800 6.55%	236 0.86%	0.76
	Medium	7713 28.07%	4056 14.76%	469 1.71%	
	OverCap	3709 13.50%	2845 10.35%	361 1.31%	
Precision		0.36	0.47	0.34	0.39

(b) 4 September 2010 model tested on 22 February 2011

		Predicted			Recall
		Low	Medium	OverCap	
Actual	Low	13652 48.88%	2892 10.35%	14 0.05%	0.82
	Medium	6272 22.45%	3686 13.20%	12 0.04%	
	OverCap	900 3.22%	502 1.80%	2 0.01%	
Precision		0.66	0.52	0.07	0.62

(c) 22 February 2011 model tested on 4 September 2010

		Predicted			Recall
		Low	Medium	OverCap	
Actual	Low	816 14.85%	803 14.61%	35 0.64%	0.49
	Medium	439 7.99%	1820 33.11%	199 3.62%	
	OverCap	29 0.53%	1050 19.10%	305 5.55%	
Precision		0.64	0.50	0.57	0.54

(d) 22 February 2011 model tested on 22 February 2011

Figure 7.1: Confusion matrices for the random forest algorithm

Table 7.1 to Table 7.4 present the full data on the performance of the random forest models trained on the claims data from one event and tested on the other CES events. Comparing the performance of each model on the different main earthquakes in the CES, the model for the 22 February 2011 event stood out. Despite the limited performance on the validation set, the model trained with data from the 22 February 2011 event achieved the best performance on the three other mains events (taking into account accuracy and the F_1 -score for each class) hinting at a better model generalisability. The larger sample size, especially for the category “over cap” (see Figure 6.19b), and better distribution between the categories of BuildingPaid, which did not require the application of sampling techniques to overcome class imbalance, might have influenced the higher generalisation ability of the model trained on data from the 22 February 2011 event.

Table 7.1: Random forest model for the 4 September 2010 tested on the 22 February 2011, 13 June 2011, and 23 December 2011

Algorithm	Set	Prediction targets	Precision	Recall	F ₁ score	Accuracy
Random forest	Train Set 4Sep2010	Low	1.00	1.00	1.00	1.00
		Medium	1.00	1.00	1.00	
		Over cap	1.00	1.00	1.00	
		Accuracy on the train set				
	Validation Set 4Sep2010	Low	0.67	0.78	0.72	0.61
		Medium	0.50	0.40	0.44	
		Over cap	0.27	0.12	0.17	
		Accuracy on the validation set				
	Test Set 22Feb2011	Low	0.36	0.76	0.48	0.39
		Medium	0.47	0.33	0.39	
		Over cap	0.34	0.05	0.09	
		Accuracy on the test set				
	Test Set 13Jun2011	Low	0.57	0.64	0.60	0.50
		Medium	0.45	0.37	0.40	
		Over cap	0.04	0.05	0.05	
		Accuracy on the test set				
Test Set 23Dec2011	Low	0.70	0.72	0.71	0.58	
	Medium	0.29	0.26	0.27		
	Over cap	0.00	0.00	0.00		
	Accuracy on the test set					

Table 7.2: Random forest model for the 22 February 2011 tested on the 4 September 2010, 13 June 2011, and 23 December 2011

Algorithm	Set	Prediction targets	Precision	Recall	F ₁ score	Accuracy
Random forest	Train Set 22Feb2011	Low	1.00	1.00	1.00	1.00
		Medium	1.00	1.00	1.00	
		Over cap	1.00	1.00	1.00	
		Accuracy on the train set				
	Validation Set 22Feb2011	Low	0.62	0.51	0.56	0.54
		Medium	0.51	0.68	0.58	
		Over cap	0.55	0.32	0.41	
		Accuracy on the validation set				
	Test Set 4Sep2010	Low	0.66	0.82	0.73	0.62
		Medium	0.52	0.37	0.43	
		Over cap	0.07	0.00	0.00	
		Accuracy on the test set				
	Test Set 13Jun2011	Low	0.62	0.49	0.55	0.52
		Medium	0.46	0.59	0.52	
		Over cap	0.12	0.11	0.11	
		Accuracy on the test set				
	Test Set 23Dec2011	Low	0.71	0.80	0.75	0.62
		Medium	0.29	0.20	0.24	
		Over cap	0.00	0.00	0.00	
		Accuracy on the test set				

Table 7.3: Random Forest model for the 13 June 2011 tested on the 4 September 2010, 22 February 2011, and 23 December 2011

Algorithm	Set	Prediction targets	Precision	Recall	F ₁ score	Accuracy
Random forest	Train Set 13Jun2011	Low	1.00	1.00	1.00	1.00
		Medium	1.00	1.00	1.00	
		Over cap	1.00	1.00	1.00	
		Accuracy on the train set				
	Validation Set 13Jun2011	Low	0.58	0.65	0.61	0.53
		Medium	0.47	0.43	0.45	
		Over cap	0.20	0.06	0.10	
		Accuracy on the validation set				
	Test Set 4Sep2011	Low	0.60	0.71	0.65	0.54
		Medium	0.39	0.33	0.36	
		Over cap	0.04	0.00	0.01	
		Accuracy on the test set				
	Test Set 22Feb2011	Low	0.35	0.65	0.46	0.40
		Medium	0.46	0.44	0.45	
		Over cap	0.36	0.02	0.03	
		Accuracy on the test set				
Test Set 23Dec2011	Low	0.71	0.76	0.73	0.61	
	Medium	0.29	0.24	0.26		
	Over cap	0.00	0.00	0.00		
	Accuracy on the test set					

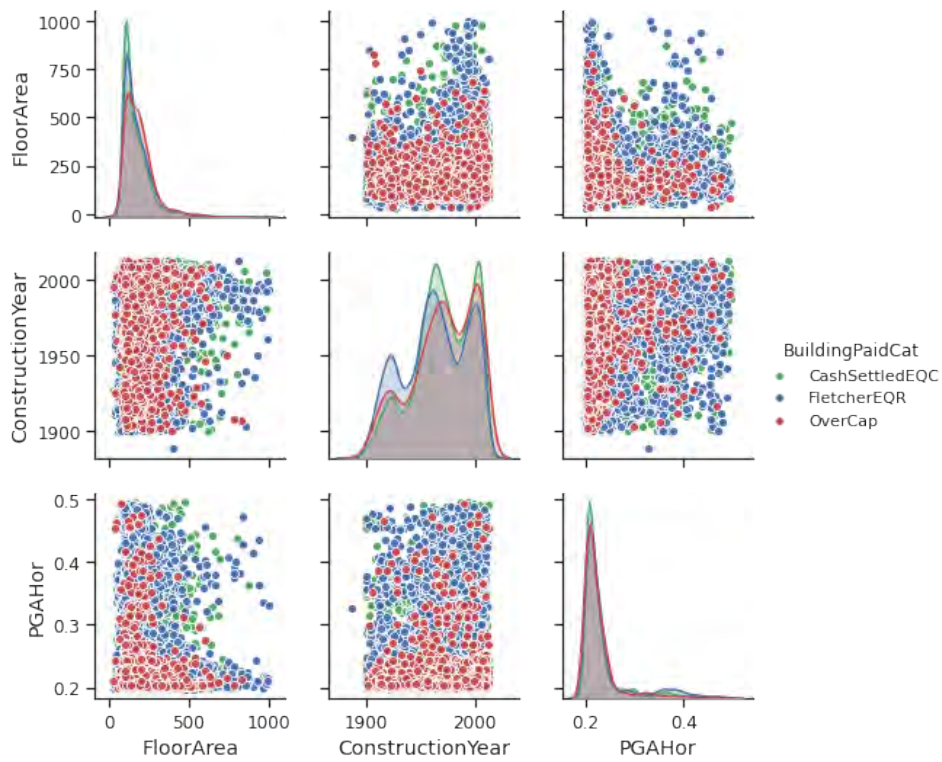
Table 7.4: Random forest model for the 23 December 2011 tested on the 4 September 2010, 22 February 2011, and 13 June 2011

Algorithm	Set	Prediction targets	Precision	Recall	F ₁ score	Accuracy
Random forest	Train Set 22Feb2011	Low	1.00	1.00	1.00	1.00
		Medium	1.00	1.00	1.00	
		Over cap	1.00	1.00	1.00	
		Accuracy on the train set				
	Validation Set 23Dec2011	Low	0.70	0.84	0.77	0.63
		Medium	0.24	0.13	0.17	
		Over cap	0.00	0.00	0.00	
		Accuracy on the validation set				
	Test Set 4Sep2010	Low	0.59	0.90	0.71	1.00
		Medium	0.37	0.10	0.16	
		Over cap	1.00	1.00	1.00	
		Accuracy on the test set				
	Test Set 22Feb2011	Low	0.31	0.94	0.46	0.32
		Medium	0.45	0.08	0.13	
		Over cap	0.08	0.00	0.00	
		Accuracy on the test set				
Test Set 13Jun2011	Low	0.55	0.89	0.68	0.54	
	Medium	0.46	0.13	0.21		
	Over cap	0.00	0.00	0.00		
	Accuracy on the test set					

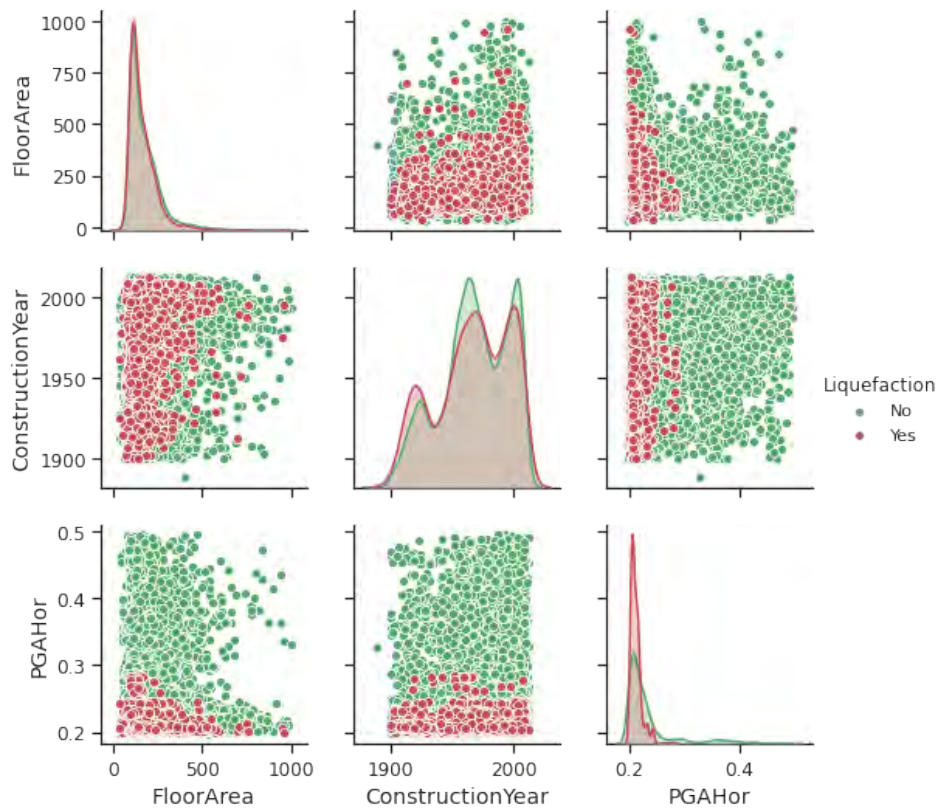
7.3 Relationship between numerical variables

Among the nine input model attributes (see Figure 6.18), three are numerical: Floor Area, ConstructionYear, and PGA. Figure 7.2 to Figure 7.5 show the relationship between the numerical attributes for the four key events in the CES. The figures are colour coded to highlight actual Building Paid Categories and a green-red colour code is used to denote the presence of liquefaction. The graphs on the diagonal present the data distribution as a layered kernel density estimate (KDE).

As shown by the graphs in the upper left corner and middle, the distribution of the floor area and construction year is relatively similar between the four main events. Of particular interest is the graph in the lower right corner showing the distribution of PGA for each earthquake event. In combination with Figure 6.20, it is clear that for 4 September 2010 and 23 December 2011 events, most of the buildings suffered PGA values lower than 0.40 g. Figure 7.5b appears to show a change in the peak PGA required to trigger liquefaction during the 23 December 2011 earthquake. Interestingly for the 4 September 2010 and the 13 June 2011 events, the over cap claims were not necessarily associated with the buildings that experienced the highest PGA values (see Figure 7.2a and Figure 7.4a respectively).

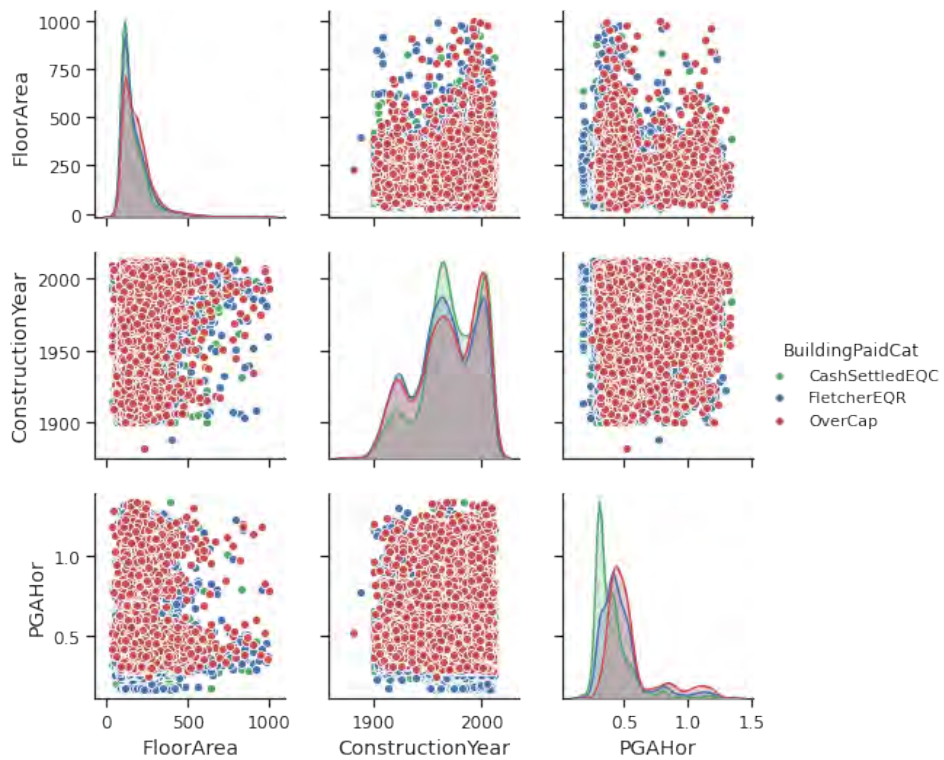


(a) Numerical attributes and Building Paid Categorical

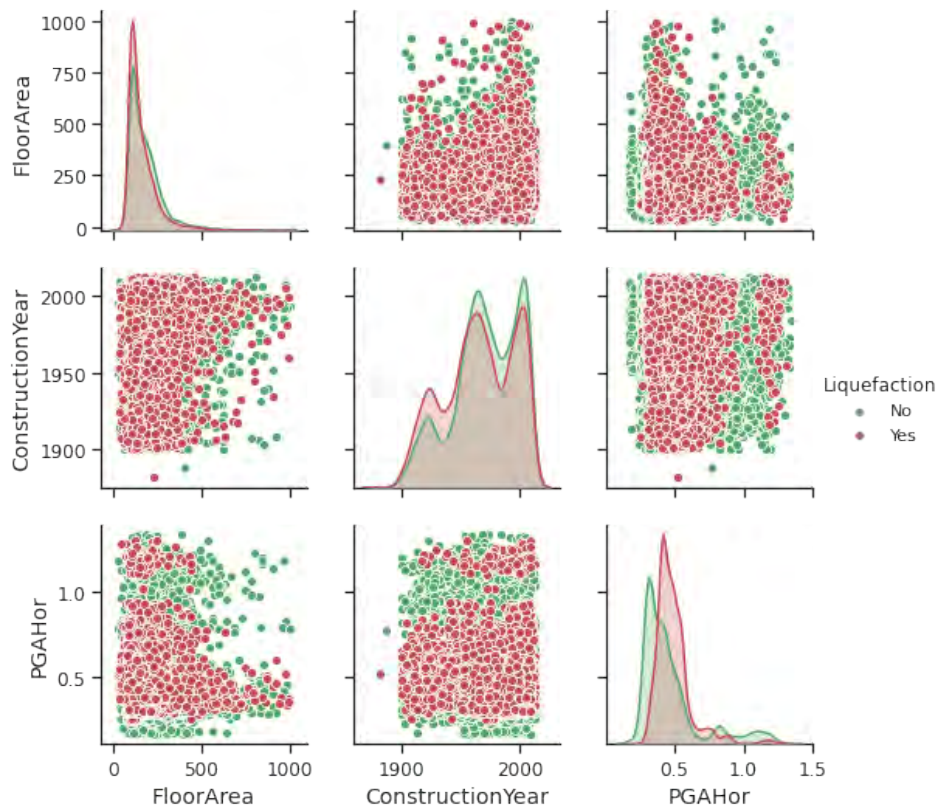


(b) Numerical attributes and liquefaction

Figure 7.2: Pairplots for the numerical attributes - 4 September 2010

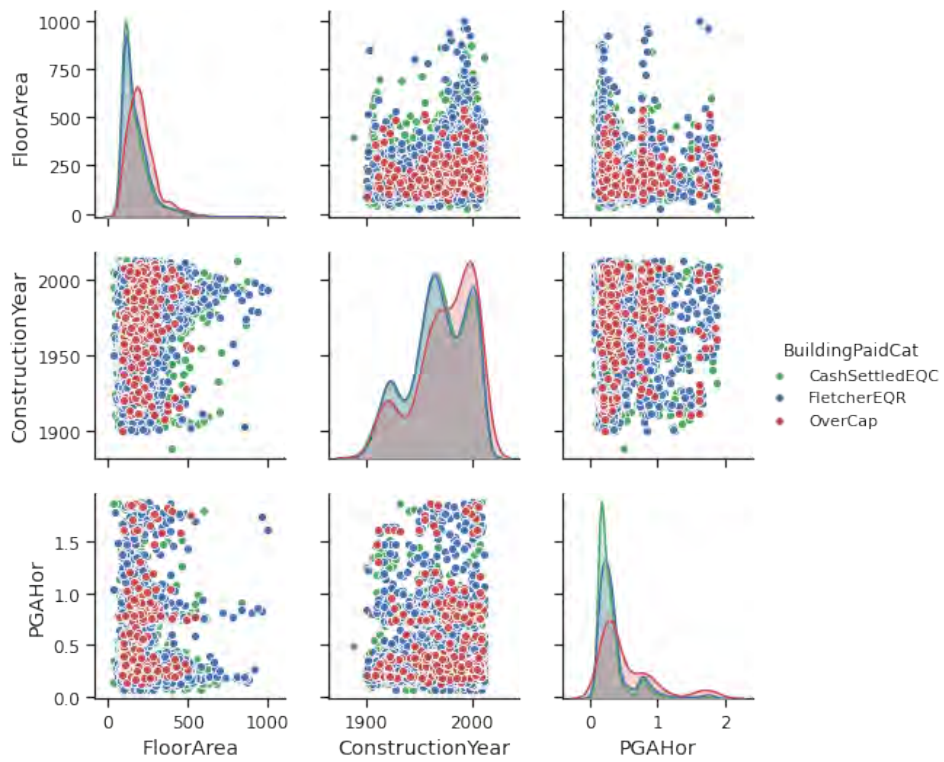


(a) Numerical attributes and Building Paid Categorical

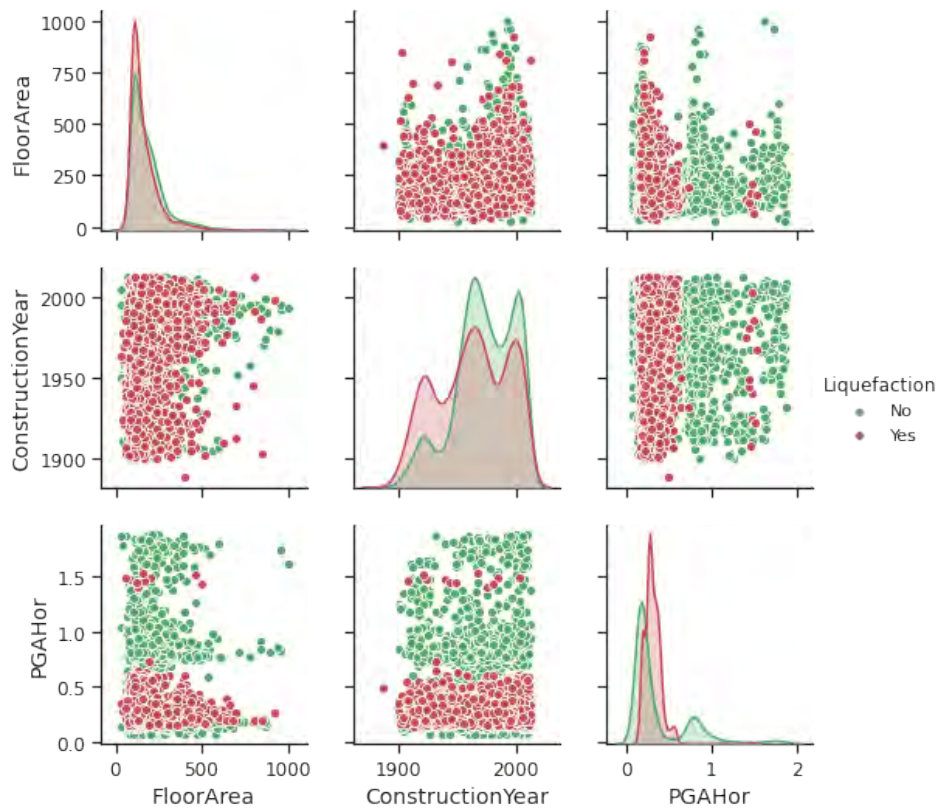


(b) Numerical attributes and liquefaction

Figure 7.3: Pairplots for the numerical attributes - 22 February 2011

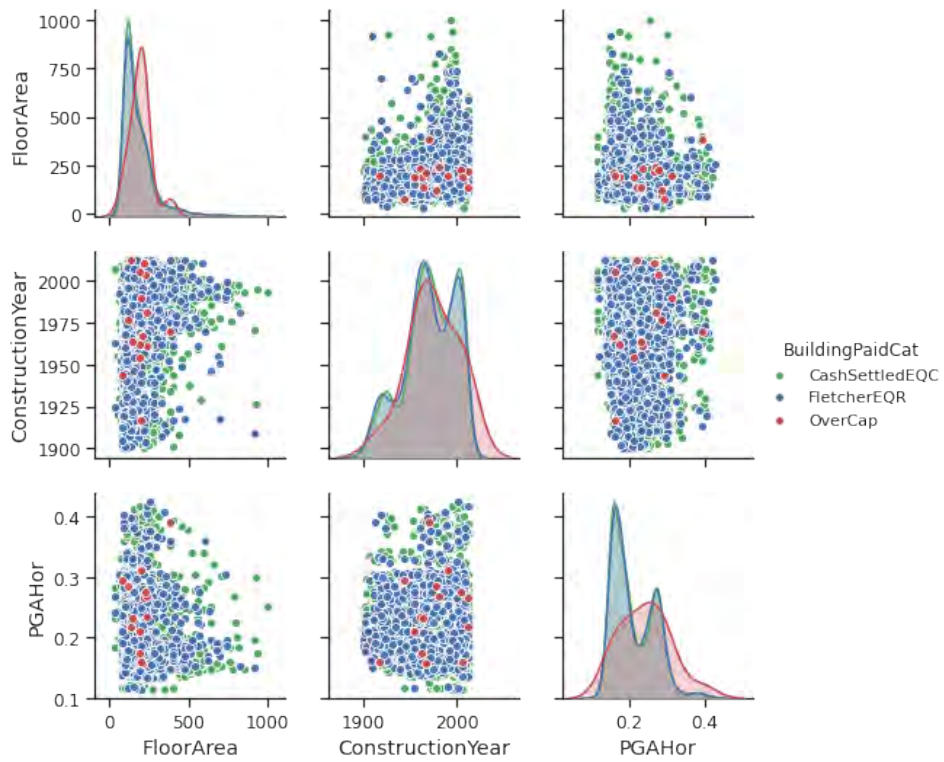


(a) Numerical attributes and Building Paid Categorical

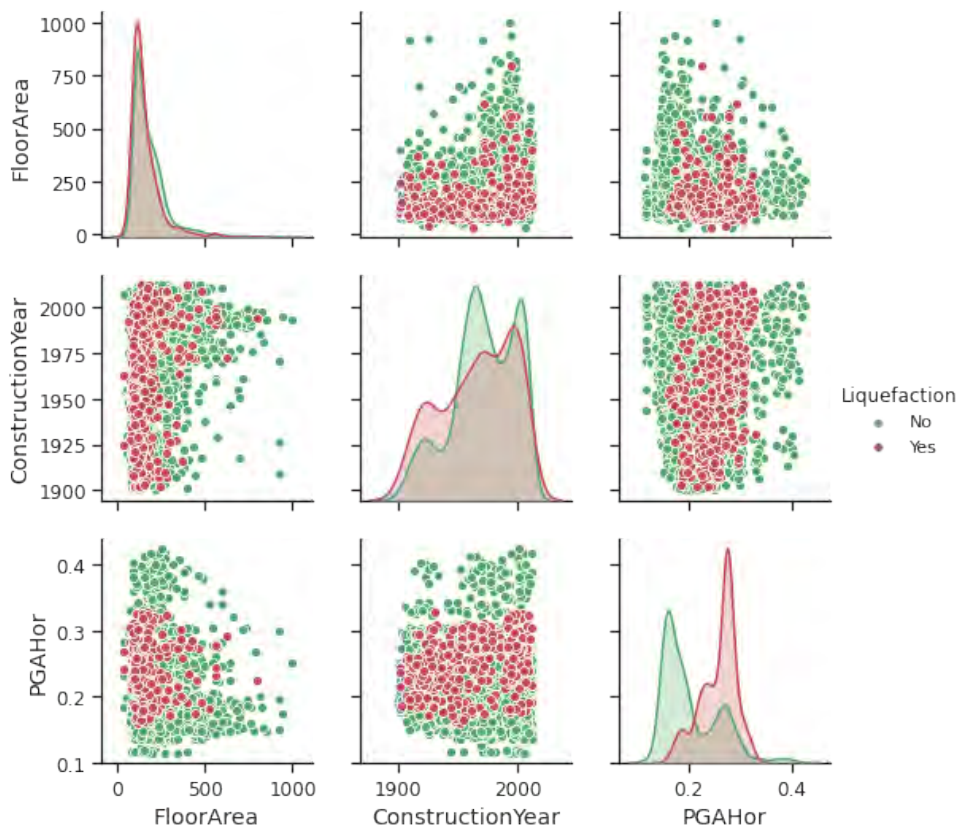


(b) Numerical attributes and liquefaction

Figure 7.4: Pairplots for the numerical attributes - 13 June 2011



(a) Numerical attributes and Building Paid Categorical



(b) Numerical attributes and liquefaction

Figure 7.5: Pairplots for the numerical attributes - 23 December 2011

7.4 Feature importance from the random forest model

7.4.1 SHAP feature importance

The SHapley Additive exPlanations (SHAP) post-hoc method was applied on the random forest models for analysing the relative influence of the different input variables. Figure 7.6 to 7.9 show the SHAP feature importance for the random forest models trained on the four key events in the CES in chronological order.

Table 7.5 summarises the five most important features for each model. PGA stands out as being the most important feature for all models. The construction year and the floor area of the building appear in the top five most important features for all events, albeit at a different rank depending on the event.

The liquefaction occurrence is second for 22 February 2011 model and fourth for 4 September 2010 model. The soil type Recent Fluvial (RFW) also plays a significant role for the 4 September 2010, 22 February 2011, and 13 June 2011.

Table 7.5: Five most important features according to the SHAP values for the random forest model

Feature rank	4 Sep 2010	22 Feb 2011	13 Jun 2011	23 Dec 2011
1	PGA	PGA	PGA	PGA
2	Construction Year	Liquefaction	Floor Area	Floor Area
3	Floor Area	Soil - RFW	Construction Year	Construction Year
4	Liquefaction	Floor Area	Soil - RFW	Floor Type – Concrete slab
5	Soil - RFW	Construction Year	Deprivation Index	Soil - PPX

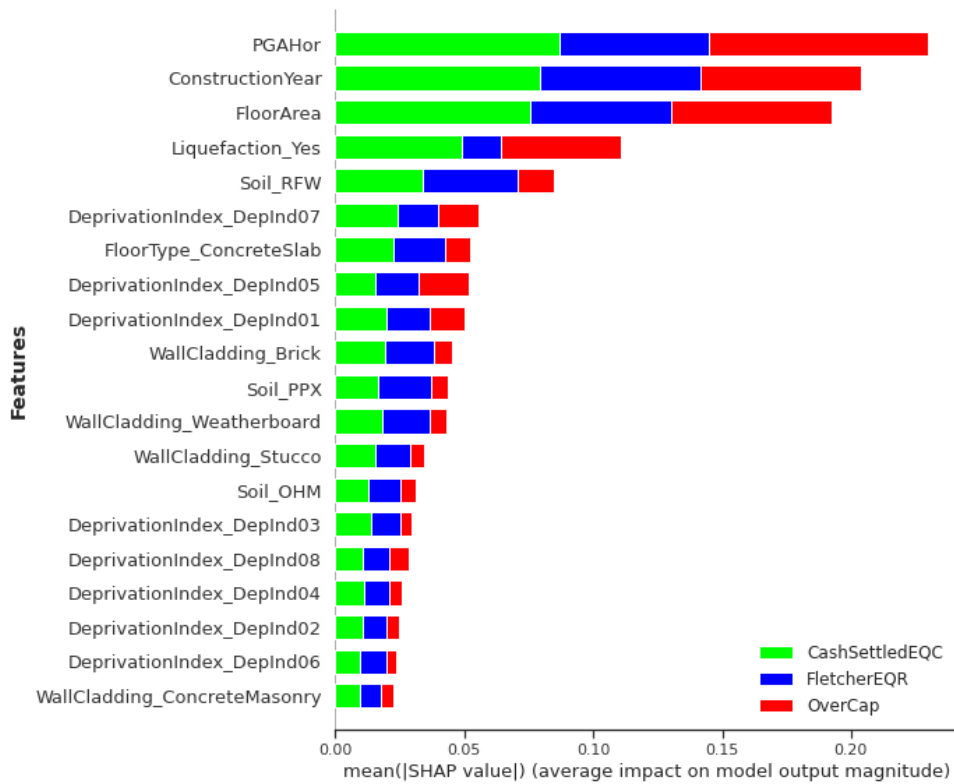


Figure 7.6: SHAP feature importance for the random forest model (4 September 2010)

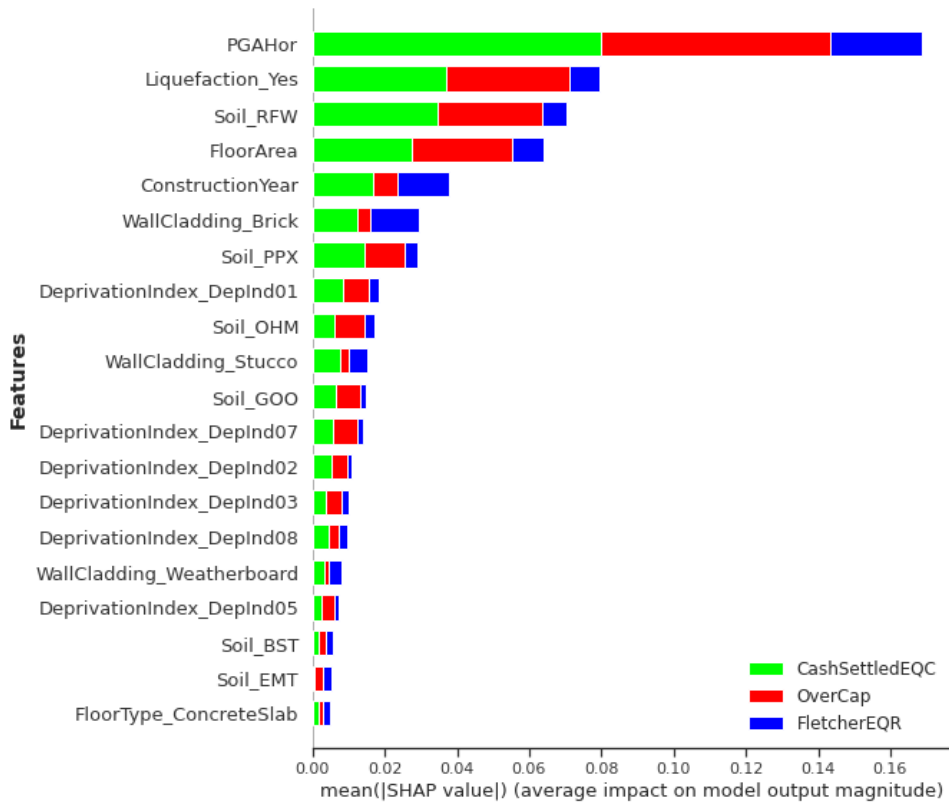


Figure 7.7: SHAP feature importance for the random forest model (22 February 2011)

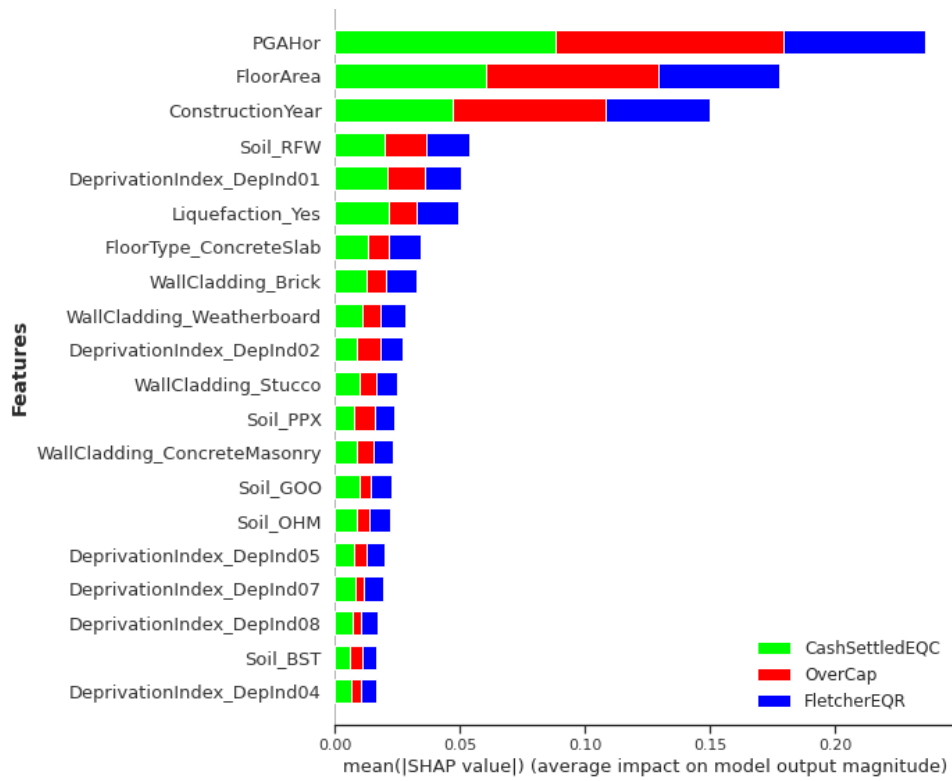


Figure 7.8: SHAP feature importance for the random forest model (13 June 2011)

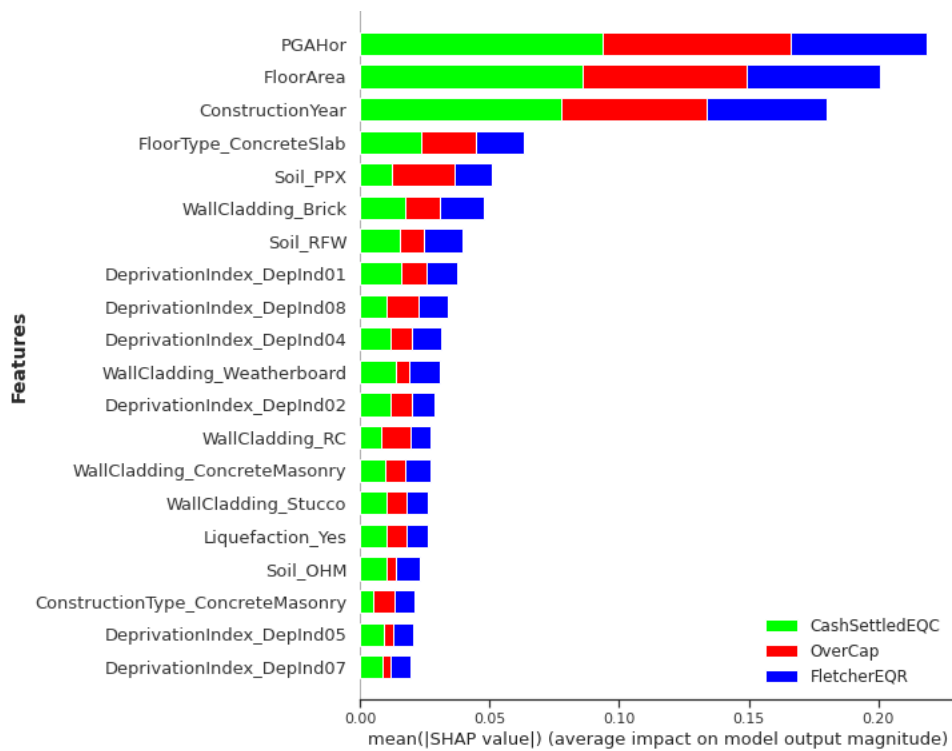


Figure 7.9: SHAP feature importance for the random forest model (23 Dec 2011)

7.4.2 Discussion of the results

The study of the feature importance of the machine learning models seems to distinguished two types of event: shaking dominated events (4 September 2010, 13 June 2011, and 23 December 2011) and liquefaction dominated (22 February 2011). The influence of PGA on the residential building losses is highlighted for the all the key events of the CES. This validates the probabilistic seismic loss estimation methodology which relies on PGA and the spectral acceleration at selected periods as intensity measures (IM) as the key input. It is satisfying to observe that machine learning, which has no physical understanding or prior knowledge related to building damage and loss, is capable of capturing the importance of PGA from empirical data alone.

For the shaking dominated events the three main important features are similar (PGA, building year of construction, and building floor area) only in a different order. Any feature following those three main ones show a limited importance.

The year of construction appears second for the 4 September 2010 event and appears of significant importance for the 13 June 2011 and 23 December 2011 events. However, it is only fifth for the 22 February 2011 event (see Figure 7.7). This seems to highlight the importance of the construction year for shaking dominated events. It is possible that the feature `ConstructionYear` captures information related to the evolution of the seismic codes.

For the 22 February 2011, PGA significantly stands out. It is followed by the liquefaction occurrence and soil type pointing out the importance of liquefaction on building damage. It thus seems that the damage and losses due to the 22 February 2011 event were driven by liquefaction. This result corroborates the findings from previous studies, which highlighted the influence of liquefaction on building damage (Rogers et al., 2015; J. Russell & van Ballegooy, 2015).

7.5 Conclusion

This chapter presented the testing of the machine learning model previously developed and findings extracted from machine learning model. The model testing showed that despite the limited model accuracy on the validation set, the model developed using the 22 February 2011 claims data was the one that fitted best the 4 September 2010 and 23 December 2011 events. Despite the relatively limited prediction accuracy on unseen data, this made the model trained on the 22 February 2011 data the more generalisable.

The interaction between the numerical variables of the model was studied. Pairplots presented the relationship between PGA, the floor area, and the year of construction. Then, the SHAP method was applied to obtain the feature importance from the random forest models developed for the four main events of the CES. For all the models, PGA stood out as the most important feature. The building floor area, the construction year, and soil were among the five most important features for the four key events. The SHAP method also extracted the liquefaction as the second most important feature for the 22 February 2011 model, thus highlighting the influence of liquefaction on building losses for this event.

Conclusions

This study aimed to enhance scientists and engineers ability to make prediction on the impact of earthquakes. This led to endeavours which included the examination of existing frameworks for post-earthquake data collection in generating better empirical data, and the potential use of machine learning models as a new rapid and adaptive tool to transform empirical seismic damage and loss data impact insights.

To this end, this research i) reviewed current assessment forms for building damage, ii) explored current data science techniques especially machine learning, iii) proposed a new paper form for the building damage assessment based on the GEM building taxonomy v2.0, iv) applied machine learning to seismic damage data of building collected following the 2017 Puebla earthquake, v) merged additional information on top of EQC's claims data set for the 2010-2011 Canterbury earthquake sequence, vi) developed a seismic loss prediction model for residential buildings in Christchurch, and vii) presented findings and insights generated from the application of data science techniques to the previously merged data set and machine learning model.

8.1 Post-earthquake damage data collection

Chapter 2 reviewed the current practice in seismic damage assessment. It surveyed post-earthquake damage collection forms that are currently used for the assessment of seismic building damage. It found that most of the forms were developed for a

geographical region in particular and focused on structural building damage and often omitted non-structural components.

To allow for better flexibility in the damage assessment between regions and offer the opportunity to collect information about non-structural elements, a new paper form was developed and introduced in Chapter 3. This new paper form is based on the GEM Building taxonomy v2.0 and expresses seismic building damage according to the European Macroseismic Scale EMS-98. It also allowed for recording non-structural component damage observations.

Following the 2017 Puebla earthquake, the form was trialled on twenty-five buildings in Calle La Morena, Mexico City. The use of the European Macroseismic Scale EMS-98 and GEM Building taxonomy v2.0 brought consistency in the building damage data collected on site. Building damage statistics were presented via dashboards to aid comprehensive damage data understanding and facilitating the derivation of useful insights.

8.2 Machine learning for the seismic damage prediction for residential buildings in the Roma and Condesa neighbourhoods, Mexico City

Chapter 4 presented a case study for the development of a damage prediction model using machine learning. It used empirical data collected following the 2017 Puebla Mexico earthquake. Building characteristics and the damage grade was available for 237 buildings located in the Roma and Condesa neighbourhoods in Mexico City. The building damage was complemented with information on the seismic demand derived from recording stations. The data was then pre-processed to be usable by machine learning algorithms. Four machine learning algorithms were applied: logistic regression, support vector machine, decision tree, and random forest. Random forest, the best performing algorithm, achieved 67% prediction accuracy.

In order to derive insights from the model, a post-hoc method was applied to the random forest algorithm. The SHAP feature importance highlighted that the building

location (latitude and longitude), the PGA, and the building height are the parameters that most influenced the model output.

8.3 Machine learning for the seismic loss prediction for residential buildings in Christchurch, New Zealand

Chapter 5 to Chapter 7 presented the development of a seismic loss prediction model using the EQC insurance claims data of residential buildings. The data set entails more than 433,500 claims pertaining to the 2010-2011 Canterbury earthquake sequence. It was found that not all the instances related to claims were settled or approved. Additionally, critical information related to the building characteristics were missing for more than 85% of the instances. This led to the necessity to complement these attributes with information from additional databases such as the RiskScape New Zealand Building inventory. A new approach to merge the claims data with the building information had to be found as the RiskScape data set and EQC data set do not share a common field apart from the geographic location (latitude and longitude).

The final proposed approach used the LINZ NZ property titles and LINZ NZ street address as an intermediary for the data integration. It was then possible to constrain the merging to property boundaries. While this approach improved the quality of the merged data as instances related to neighbouring properties were excluded, limitations still remained. One of the limitations relates to the RiskScape data set as it included residential dwelling as well as secondary buildings. This was eventually overcome through additional logic including the use of property boundaries and footprint area data. Following the merging between EQC and RiskScape, additional information related to the seismic demand, liquefaction occurrence, and soil conditions were added.

Once merged, the data set was filtered and pre-processed to enable the application of machine learning. The target attribute, *BuildingPaid*, was transformed from a numerical to a categorical variable. The model attributes were selected. After the preparation of the categorical and numerical features, four machine learning models were trained using data from four key events in the CES. Random forest achieved the best performance accuracy for all events. The prediction accuracy for random forest ranged from 0.53 on

the 13 June 2011 event validation set to 0.63 on the 23 December 2011 event validation set. Possible reasons for the limited model accuracy include the limited information in the raw data set, class imbalance, and the selection of model attributes.

The models were then assessed for their ability to generalise through testing on data pertaining to another event in the CES. The model for the 22 February 2011 only reached 0.54 on the validation set but generalised best for other key events in the CES. The better generalisation for the 22 February 2011 model might be related to size of the earthquake event. With a larger sample size for the over-cap category it was not necessary to apply oversampling techniques to address the class imbalance before the application of machine learning.

Chapter 7 presented insights derived from the merged data set and machine learning model. The random forest models were analysed for feature importance using the SHAP post-hoc methodology. PGA was the most important feature for the four models developed across the different machine learning models. This validates current probabilistic approaches for the seismic damage and loss assessment such as the PBEE where PGA or the spectral acceleration at selected period can be used as intensity measure. The feature importance delivered insights related to the parameters that affect building losses the most.

Machine learning applied to seismic damage and loss data delivers valuable insights. It enables to make use of empirical data to obtain useful insights without the need of complex and time-consuming methodologies. The insights are not only useful for the insurance sector but also for engineers, risk managers, emergency planners, and government to better understand critical damage drivers.

8.4 Current challenges in the application of machine learning for the prediction of seismic damage and loss

In recent year, the application of machine learning to various tasks has grown significantly. This growth can to some extent be linked to the increase in computer power and advance in research related to machine learning algorithms which enable the application of machine learning to more complex tasks. However, one of the key points

for the successful application of machine learning is the availability of data. A machine learning algorithm can only properly “learn” from data if there are enough instances in the training set. In many problems with complex relationships within the model attributes, the larger the data set, the better the model performance as the algorithm has seen enough examples to “learn” and generalise the relationships between the model attributes.

With 27,932 instances available for model training for the 4 September 2010 event, 27,479 for the 22 February 2011 event, 6,736 for the 13 June 2011 event, and 4,743 for the 23 December 2011 event the data sets in this study can be considered as “large” from a civil engineering perspective. Nevertheless, machine learning models, especially the ones relying on sophisticated algorithms, may require significantly more data to properly execute and be able to generalise.

The EQC claims data for residential building was initially collected for insurance purposes and not with the intent to develop of a seismic loss prediction model through machine learning. For best machine learning performance, data must not have missing features. This was not always possible due to the nature of earthquake events, and also there is doubt on the reliability on data collected on-site. The data integration with other database enabled the addition of building characteristics as well as additional information related to the seismic demand, liquefaction occurrence, and soil conditions.

The issues faced during the merging of the EQC data set with RiskScape building characteristics and LINZ information highlighted the need for an improved solution to identify each building in New Zealand. It is believed that the establishment of a unique building identifier common to several databases will introduce consistency, thus opening new opportunities for the application of data science techniques and the derivation of insights.

At the time of the CES, EQC only provided building coverage up to NZ\$100,000 (+GST) which led to the EQC data set being capped at NZ\$115,000. Losses above the NZ\$115,000 threshold were covered by private insurers, given that the building owner subscribed to appropriate private insurance. Any detail for building loss above NZ\$115,000 was not available for this study. The access to data from private insurances would enlarge the range of BuildingPaid giving more information on the buildings which

suffered significant losses.

The prediction accuracy also depends on the attributes present in the model. Attributes that might have influenced the value of the building losses might not be present in the current model. Thus, it would be of interest to study the influence of additional attributes on the model performance.

8.5 Future work and opportunities

Chapter 6 highlighted the importance of data pre-processing. A more in depth analysis of the actual value of BuildingPaid might bring an improved model performance. Taking into account apportionment between the events in the CES would provide a more accurate allocation of loss to each event and enable to capture more details about over cap instances.

To mitigate issues related to sequential damage throughout the CES, the data could be segregated by geographical area where the majority of damage occurred for each event. This might lead to a "cleaner" train set and thus might deliver more accurate predictions.

In order to better understand generalisation errors, different test sets might be employed in future work. Possibilities for other test sets include: holdouts by geographical area, soil type, year of building construction, and random sampling.

In addition to opportunities related to the available data, new attributes could be included. Once developed, a machine learning pipeline can be retrained with limited efforts. This facilitates future studies employing different combinations of building parameters. Any attribute that is deemed impactful on the building loss could be added in the model. Different attributes and their influences on building losses could be tested. For example, the introduction of additional parameters related to properties and social factors might deliver an improved model accuracy as well as new insights. The importance of the new attribute could then be studied via the feature importance of the random forest model.

Finally, the machine learning model can also be retrained whenever new claim data becomes available from a future earthquake. The new data could yield improvement in the accuracy of the loss prediction model.

Appendices

Loss databases

Table A.1: Overview of the main loss databases, adapted from (Integrated Research on Disaster Risk, 2014)

	EM-DAT	NatCatSERVICE	Sigma explorer	GLIDE	DesInventar	SHELDUS
Owner	Centre for Research on the Epidemiology of Disasters (CRED), Université Catholique de Louvain, Belgium	Munich Re, Germany	Swiss Re, Switzerland	Asian Disaster Reduction Center (ADRC), Japan	Varies by country	Hazards and Vulnerability Research Institute (HVRI), University of South Carolina, USA
Web link	emdat.be	NatCatSERVICE	sigma-explorer.com	glidenumber.net	desinventar.org	sheldus.org
Spatial Coverage	Global	Global	Global	Global	National	National
Spatial Resolution	Country	Country	Country	Country	County, municipality	U.S. county
Data sources	U.N agencies, IFRC, World Bank, reinsurers, press, news agencies	Property claims service, insurance clients, U.N agencies, World Bank, press		U.N agencies, IFRC, World Bank, reinsurers, press, news agencies	U.N agencies, weather services, geological services, press	U.S. National Climatic Data Center, National Geophysical Data Center, U.S. Geological Survey (USGS)
Recording Thresholds	≥10 fatalities, ≥100 affected, declaration of state of emergency, or call for international assistance			≥10 fatalities, ≥100 affected, declaration of state of emergency, or call for international assistance	≥1 human loss or ≥US\$1 in economic loss	≥1 human loss or ≥US\$1 in economic loss
Hazard coverage						
Geophysical	✓	✓	✓	✓	✓	✓
Hydrological	✓	✓	✓	✓	✓	✓
Meteorological	✓	✓	✓	✓	✓	✓
Climatological	✓	✓	✓	✓	✓	✓
Biological	✓			✓	✓	
Technological	✓		✓	✓	✓	
Loss indicators						
Fatalities	✓	✓	✓		✓	✓
Aggregated economic loss	✓	✓	✓		✓	
Insured loss		✓	✓			

**Forms for the evaluation of seismic
building damage**

B.1 ATC-20 Detailed Evaluation Safety Assessment Form

ATC-20 Detailed Evaluation Safety Assessment Form

Inspection

Inspector ID: _____

Affiliation: _____

Inspection date and time: _____ AM PM

Final Posting from page 2

- Inspected
 Restricted Use
 Unsafe

Building Description

Building name: _____

Address: _____

Building contact/phone: _____

Number of stories above ground: _____ below ground: _____

Approx. "Footprint area" (square feet): _____

Number of residential units: _____

Number of residential units not habitable: _____

Type of Construction

- | | |
|---|---|
| <input type="checkbox"/> Wood frame | <input type="checkbox"/> Concrete shear wall |
| <input type="checkbox"/> Steel frame | <input type="checkbox"/> Unreinforced masonry |
| <input type="checkbox"/> Tilt-up concrete | <input type="checkbox"/> Reinforced masonry |
| <input type="checkbox"/> Concrete frame | <input type="checkbox"/> Other: _____ |

Primary Occupancy

- | | | |
|---|---------------------------------------|-------------------------------------|
| <input type="checkbox"/> Dwelling | <input type="checkbox"/> Commercial | <input type="checkbox"/> Government |
| <input type="checkbox"/> Other residential | <input type="checkbox"/> Offices | <input type="checkbox"/> Historic |
| <input type="checkbox"/> Public assembly | <input type="checkbox"/> Industrial | <input type="checkbox"/> School |
| <input type="checkbox"/> Emergency services | <input type="checkbox"/> Other: _____ | |

Evaluation


Investigate the building for the conditions below and check the appropriate column. There is room on the second page for a sketch.


	Minor/None	Moderate	Severe	Comments
Overall hazards:				
Collapse or partial collapse	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	_____
Building or story leaning	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	_____
Other _____	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	_____
Structural hazards:				
Foundations	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	_____
Roofs, floors (vertical loads)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	_____
Columns, pilasters, corbels	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	_____
Diaphragms, horizontal bracing	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	_____
Walls, vertical bracing	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	_____
Precast connections	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	_____
Other _____	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	_____
Nonstructural hazards:				
Parapets, ornamentation	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	_____
Cladding, glazing	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	_____
Ceilings, light fixtures	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	_____
Interior walls, partitions	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	_____
Elevators	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	_____
Stairs, exits	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	_____
Electric, gas	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	_____
Other _____	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	_____
Geotechnical hazards:				
Slope failure, debris	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	_____
Ground movement, fissures	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	_____
Other _____	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	_____

General Comments: _____

Continue on page 2

B.2 GEM paper based assessment tool

	Project		Date dd/mm/yyyy		Page 1 of 2														
	Completed by		Time hh:mm																
Phone: (std) no	Email:		Struct Eng.	Architect	Building Official	Civil Eng.													
Basic Building Metadata		Build No.	GPS	Longitude (X) <small>Decimal degrees</small>	Latitude (y) <small>Decimal degrees</small>														
Address (building location)		City:		State:		Zip/PCode													
		No. stories above grade:		between	Exact	Circa	Min	Max											
		No. Stories below grade:		between	Exact	Circa	Min	Max											
		Ht. of G.F. above grade (m)		between	Exact	Circa	Pre	Post											
Direction Main Facade		Unkn.	Perpendicular to street	Parallel to street	No. Occupants														
Occupancy Type		Un-known	Mixed Use	Residential	Comm/ Public	Industrial	Agriculture	Assembly	Govmt.	Other	Education								
Position within block		Detached Building	Adjoining building(s) on one side	Adjoining building(s) on two sides	Adjoining building(s) on three sides	Slope in Degrees													
Occupancy Details		Un-known	Single Unit	Multi Unit (no)	unkn	2 units	3-4 units	5-9 units	10-19units	20-49units	50+ units	Temp Lodge	Institut housing	Mobile Home	Agric	Unkn	Store	Animl	Process
		Retail/trade	Wholesale/storage	Office/prof/tec	Medical	Entertainment	Public building	Covered Parking	Industry	Heavy	Light	Bus stn	Rail stn	Airport					
		Assembly	Recr. Leis.	Relig. Gath.	Arena	Cinema Conc.	Govmt. gen.	Emerg.S services.	Education	Pre-sch.	School	Coll/Uni class/off	Coll/Uni lab/fac.	Other/ Mixed use					
Lateral Load Resisting System Type (LLRS) LONGITUDINAL		Unknown	LLRS Material Type	Unknown	LLRS Material Technology		Unknown												
None		Ductility	Steel	Other metal	Riveted	Welded	Bolted												
Moment frame	Unknown	Masonry	Reinforced	Unreinforced	Confined	Adobe	Stone	Rubble	Dressed										
Infilled Frame	Ductile	Fired clay unknown	Solid	Hollow Brick	Hollow Block	Concrete block unknown	Solid	Hollow	Other										
Braced Frame	Non-ductile	Concrete	Unknown Reinforce.	Composite With steel section	Reinforced	Un-Reinforced	Unknown	Cast in Place	Cast in Place - pre stressed										
Post and Beam	Base isolation and or energy diss. dev	Pre-cast	Pre-cast pre-stressed																
Wall	Earth	Unknown	Reinforced	Unreinforced	Rammed	Wet	Other												
Dual frame-wall sys.	Flat slab/plate/waffle	Infilled	Wood	Unknown	Light	Heavy	Solid	Wattle&daub	Bamboo	Other									
Hybrid	Other	Other																	
Lateral Load Resisting System Type (LLRS) TRANSVERSE		Unknown	LLRS Material Type	Unknown	LLRS Material Technology		Unknown												
None		Ductility	Steel	Other metal	Riveted	Welded	Bolted												
Moment frame	Unknown	Masonry	Reinforced	Unreinforced	Confined	Adobe	Stone	Rubble	Dressed										
Infilled Frame	Ductile	Fired clay unknown	Solid	Hollow Brick	Hollow Block	Concrete block unknown	Solid	Hollow	Other										
Braced Frame	Non-ductile	Concrete	Unknown Reinforce.	Composite With steel section	Reinforced	Un-Reinforced	Unknown	Cast in Place	Cast in Place - pre stressed										
Post and Beam	Base isolation and or energy diss. dev	Pre-cast	Pre-cast pre-stressed																
Wall	Earth	Unknown	Reinforced	Unreinforced	Rammed	Wet	Other												
Dual frame-wall sys.	Flat slab/plate/waffle	Infilled	Wood	Unknown	Light	Heavy	Solid	Wattle&daub	Bamboo	Other									
Hybrid	Other	Other																	
Structural Irregularity Primary		Unknown	Regular	Irregular	Vertical	Soft storey	Cripple wall	Short column	Other	Pounding potential	Setback	Change in vert. struct	Other	Plan-	Torsion eccentricity	Re-entrant C.	Other		
Structural Irregularity Secondary		Unknown	Regular	Irregular	Vertical	Soft storey	Cripple wall	Short column	Other	Pounding potential	Setback	Change in vert. struct	Other	Plan-	Torsion eccentricity	Re-entrant C.	Other		

 GEM <small>GLOBAL EARTHQUAKE MODEL</small> <small>DIRECT OBSERVATION TOOLS</small>	Project		Date dd/mm/yyyy		Page 2 of 2		
	Completed by		Time hh:mm				
Phone: (std) no	Email:	Struct Eng.	<input type="checkbox"/>	Architect	<input type="checkbox"/>		
		Building Official	<input type="checkbox"/>	Civil Eng	<input type="checkbox"/>		
Basic Building Metadata		Build No.	GPS	Longitude (X) <small>Decimal degrees</small>	Latitude (y) <small>Decimal degrees</small>		
Address (building location)		City:		State:	Zip/PCode		
		No. stories above grade: between <input type="checkbox"/>		Exact <input type="checkbox"/>	Circa <input type="checkbox"/>	Min <input type="checkbox"/>	Max <input type="checkbox"/>
		No. Stories below grade: between <input type="checkbox"/>		Exact <input type="checkbox"/>	Circa <input type="checkbox"/>	Min <input type="checkbox"/>	Max <input type="checkbox"/>
		Ht. of G.F. above grade (m) between <input type="checkbox"/>		Exact <input type="checkbox"/>	Circa <input type="checkbox"/>	Pre <input type="checkbox"/>	Post <input type="checkbox"/>
No. Dwellings		Year of Construct/Retrofit: between <input type="checkbox"/>		Exact <input type="checkbox"/>	Circa <input type="checkbox"/>		
Direction Main Facade		Unkn. <input type="checkbox"/>	Perpendicular to street <input type="checkbox"/>	Parallel to street <input type="checkbox"/>	No. Occupants		
Occupancy Type		Un-known <input type="checkbox"/>	Mixed Use <input type="checkbox"/>	Residential <input type="checkbox"/>	Comm/Public <input type="checkbox"/>		
		Industrial <input type="checkbox"/>	Agriculture <input type="checkbox"/>	Assembly <input type="checkbox"/>	Govmt. <input type="checkbox"/>		
		Other <input type="checkbox"/>	Education <input type="checkbox"/>				
Position within block		Detached Building <input type="checkbox"/>	Adjoining building(s) on one side <input type="checkbox"/>	Adjoining building(s) on two sides <input type="checkbox"/>	Adjoining building(s) on three sides <input type="checkbox"/>		
Occupancy Details		Un-known <input type="checkbox"/>	Single Unit <input type="checkbox"/>	Multi Unit (no) <input type="checkbox"/>	unkn <input type="checkbox"/>		
		2 units <input type="checkbox"/>	3-4 units <input type="checkbox"/>	5-9 units <input type="checkbox"/>	10-19units <input type="checkbox"/>	20-49units <input type="checkbox"/>	
		50+ units <input type="checkbox"/>	Temp Lodge <input type="checkbox"/>	Institut housing <input type="checkbox"/>	Mobile Home <input type="checkbox"/>	Agric <input type="checkbox"/>	
		Retail/trade <input type="checkbox"/>	Wholesale/storage <input type="checkbox"/>	Office/prof/tec <input type="checkbox"/>	Medical <input type="checkbox"/>		
		Entertainment <input type="checkbox"/>	Public building <input type="checkbox"/>	Covered Parking <input type="checkbox"/>	Industry <input type="checkbox"/>		
		Heavy <input type="checkbox"/>	Light <input type="checkbox"/>	Bus stn <input type="checkbox"/>	Rail stn <input type="checkbox"/>		
		Airport <input type="checkbox"/>	Unkn <input type="checkbox"/>	Store <input type="checkbox"/>	Animl <input type="checkbox"/>		
		Process <input type="checkbox"/>					
Lateral Load Resisting System Type (LLRS) LONGITUDINAL		Unknown <input type="checkbox"/>	LLRS Material Type	Unknown <input type="checkbox"/>	LLRS Material Technology		
		Unknown <input type="checkbox"/>	Unknown <input type="checkbox"/>	Unknown <input type="checkbox"/>	Unknown <input type="checkbox"/>		
None <input type="checkbox"/>		Ductility	Steel <input type="checkbox"/>	Other metal <input type="checkbox"/>	Riveted <input type="checkbox"/>		
		Welded <input type="checkbox"/>	Bolted <input type="checkbox"/>				
Moment frame <input type="checkbox"/>		Unknown <input type="checkbox"/>	Masonry <input type="checkbox"/>	Reinforced <input type="checkbox"/>	Unreinforced <input type="checkbox"/>		
		Ductile <input type="checkbox"/>	Confined <input type="checkbox"/>	Adobe <input type="checkbox"/>	Stone <input type="checkbox"/>		
		Non-ductile <input type="checkbox"/>	Unreinforced <input type="checkbox"/>	Fired clay unknown <input type="checkbox"/>	Solid <input type="checkbox"/>		
		Concrete <input type="checkbox"/>	Unknown Reinforce. <input type="checkbox"/>	Concrete block unknown <input type="checkbox"/>	Solid <input type="checkbox"/>		
		Composite With steel section <input type="checkbox"/>	Reinforced <input type="checkbox"/>	Unreinforced <input type="checkbox"/>	Hollow <input type="checkbox"/>		
		Un-Reinforced <input type="checkbox"/>	Unknown <input type="checkbox"/>	Cast in Place <input type="checkbox"/>	Hollow <input type="checkbox"/>		
		Other <input type="checkbox"/>	Unknown <input type="checkbox"/>	Cast in Place - pre stressed <input type="checkbox"/>	Other <input type="checkbox"/>		
Post and Beam <input type="checkbox"/>		Base isolation and or energy diss. dev <input type="checkbox"/>	Concrete <input type="checkbox"/>	Unknown Reinforce. <input type="checkbox"/>	Composite With steel section <input type="checkbox"/>		
		Earth <input type="checkbox"/>	Unknown <input type="checkbox"/>	Reinforced <input type="checkbox"/>	Unreinforced <input type="checkbox"/>		
		Pre-cast <input type="checkbox"/>	Pre-cast pre-stressed <input type="checkbox"/>				
Wall <input type="checkbox"/>		Dual frame-wall sys. <input type="checkbox"/>	Earth <input type="checkbox"/>	Unknown <input type="checkbox"/>	Reinforced <input type="checkbox"/>		
		Flat slab/plate/waffle <input type="checkbox"/>	Infilled <input type="checkbox"/>	Wood <input type="checkbox"/>	Unknown <input type="checkbox"/>		
		Hybrid <input type="checkbox"/>	Other <input type="checkbox"/>	Other <input type="checkbox"/>	Light <input type="checkbox"/>		
		Other <input type="checkbox"/>	Other <input type="checkbox"/>	Other <input type="checkbox"/>	Heavy <input type="checkbox"/>		
		Other <input type="checkbox"/>	Other <input type="checkbox"/>	Other <input type="checkbox"/>	Solid <input type="checkbox"/>		
		Other <input type="checkbox"/>	Other <input type="checkbox"/>	Other <input type="checkbox"/>	Wattle&daub <input type="checkbox"/>		
		Other <input type="checkbox"/>	Other <input type="checkbox"/>	Other <input type="checkbox"/>	Bamboo <input type="checkbox"/>		
		Other <input type="checkbox"/>	Other <input type="checkbox"/>	Other <input type="checkbox"/>	Other <input type="checkbox"/>		
Lateral Load Resisting System Type (LLRS) TRANSVERSE		Unknown <input type="checkbox"/>	LLRS Material Type	Unknown <input type="checkbox"/>	LLRS Material Technology		
		Unknown <input type="checkbox"/>	Unknown <input type="checkbox"/>	Unknown <input type="checkbox"/>	Unknown <input type="checkbox"/>		
None <input type="checkbox"/>		Ductility	Steel <input type="checkbox"/>	Other metal <input type="checkbox"/>	Riveted <input type="checkbox"/>		
		Welded <input type="checkbox"/>	Bolted <input type="checkbox"/>				
Moment frame <input type="checkbox"/>		Unknown <input type="checkbox"/>	Masonry <input type="checkbox"/>	Reinforced <input type="checkbox"/>	Unreinforced <input type="checkbox"/>		
		Ductile <input type="checkbox"/>	Confined <input type="checkbox"/>	Adobe <input type="checkbox"/>	Stone <input type="checkbox"/>		
		Non-ductile <input type="checkbox"/>	Unreinforced <input type="checkbox"/>	Fired clay unknown <input type="checkbox"/>	Solid <input type="checkbox"/>		
		Concrete <input type="checkbox"/>	Unknown Reinforce. <input type="checkbox"/>	Concrete block unknown <input type="checkbox"/>	Solid <input type="checkbox"/>		
		Composite With steel section <input type="checkbox"/>	Reinforced <input type="checkbox"/>	Unreinforced <input type="checkbox"/>	Hollow <input type="checkbox"/>		
		Un-Reinforced <input type="checkbox"/>	Unknown <input type="checkbox"/>	Cast in Place <input type="checkbox"/>	Hollow <input type="checkbox"/>		
		Other <input type="checkbox"/>	Unknown <input type="checkbox"/>	Cast in Place - pre stressed <input type="checkbox"/>	Other <input type="checkbox"/>		
Post and Beam <input type="checkbox"/>		Base isolation and or energy diss. dev <input type="checkbox"/>	Concrete <input type="checkbox"/>	Unknown Reinforce. <input type="checkbox"/>	Composite With steel section <input type="checkbox"/>		
		Earth <input type="checkbox"/>	Unknown <input type="checkbox"/>	Reinforced <input type="checkbox"/>	Unreinforced <input type="checkbox"/>		
		Pre-cast <input type="checkbox"/>	Pre-cast pre-stressed <input type="checkbox"/>				
Wall <input type="checkbox"/>		Dual frame-wall sys. <input type="checkbox"/>	Earth <input type="checkbox"/>	Unknown <input type="checkbox"/>	Reinforced <input type="checkbox"/>		
		Flat slab/plate/waffle <input type="checkbox"/>	Infilled <input type="checkbox"/>	Wood <input type="checkbox"/>	Unknown <input type="checkbox"/>		
		Hybrid <input type="checkbox"/>	Other <input type="checkbox"/>	Other <input type="checkbox"/>	Light <input type="checkbox"/>		
		Other <input type="checkbox"/>	Other <input type="checkbox"/>	Other <input type="checkbox"/>	Heavy <input type="checkbox"/>		
		Other <input type="checkbox"/>	Other <input type="checkbox"/>	Other <input type="checkbox"/>	Solid <input type="checkbox"/>		
		Other <input type="checkbox"/>	Other <input type="checkbox"/>	Other <input type="checkbox"/>	Wattle&daub <input type="checkbox"/>		
		Other <input type="checkbox"/>	Other <input type="checkbox"/>	Other <input type="checkbox"/>	Bamboo <input type="checkbox"/>		
		Other <input type="checkbox"/>	Other <input type="checkbox"/>	Other <input type="checkbox"/>	Other <input type="checkbox"/>		
Structural Irregularity Primary		Unknown <input type="checkbox"/>	Regular <input type="checkbox"/>	Irregular <input type="checkbox"/>	Vertical		
		Soft storey <input type="checkbox"/>	Cripple wall <input type="checkbox"/>	Short column <input type="checkbox"/>	Other <input type="checkbox"/>		
		Pounding potential <input type="checkbox"/>	Setback <input type="checkbox"/>	Change in vert. struct <input type="checkbox"/>	Other <input type="checkbox"/>		
		Torsion eccentricity <input type="checkbox"/>	Re-entrant C. <input type="checkbox"/>	Other <input type="checkbox"/>	Plan.		
Structural Irregularity Secondary		Unknown <input type="checkbox"/>	Regular <input type="checkbox"/>	Irregular <input type="checkbox"/>	Vertical		
		Soft storey <input type="checkbox"/>	Cripple wall <input type="checkbox"/>	Short column <input type="checkbox"/>	Other <input type="checkbox"/>		
		Pounding potential <input type="checkbox"/>	Setback <input type="checkbox"/>	Change in vert. struct <input type="checkbox"/>	Other <input type="checkbox"/>		
		Torsion eccentricity <input type="checkbox"/>	Re-entrant C. <input type="checkbox"/>	Other <input type="checkbox"/>	Plan.		

**B.3 New paper form for the seismic assessment of building
based on GEM Building Taxonomy v2.0**

Seismic Assessment based on GEM Building Taxonomy v2.0


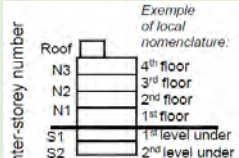
Project _____
 Inspection time _____
 Start (local time) _____ : _____ (hh/mm) Inspection duration _____ (min)
 Completed by _____
 Function Structural eng. Building official

Date ____/____/____ (dd/mm/yyyy)
 Areas inspected
 Exterior and interior
 Exterior only
 Architect
 Student

General building information

Building name _____ Street name and nb _____
 Neighborhood _____ City _____ Zip/Pcode _____
 State _____ Country _____
 Coordinates _____ Longitude X _____ Latitude Y _____

Building information

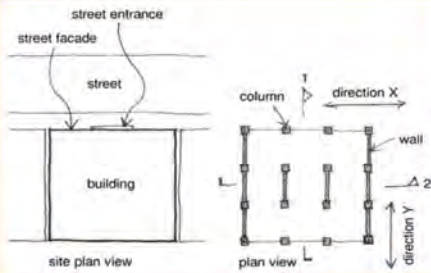
<p>OCCUPANCY</p> <p>Unknown occupancy type</p> <p>Residential</p> <p>Residential, unknown type</p> <p>Single dwelling</p> <p>Multi-unit, unknown type</p> <p>2 Units (duplex)</p> <p>3-4 Units</p> <p>5-9 Units</p> <p>10-19 Units</p> <p>20-49 Units</p> <p>50+ Units</p> <p>Temporary lodging</p> <p>Institutional housing</p> <p>Mobile home</p> <p>Informal housing</p> <p>Commercial and public</p> <p>Commercial and public, unknown type</p> <p>Retail trade</p> <p>Wholesale trade and storage (warehouse)</p> <p>Offices, professional/technical services</p> <p>Hospital/medical clinic</p> <p>Entertainment</p> <p>Public building</p> <p>Covered parking garage</p> <p>Bus station</p> <p>Railway station</p> <p>Airport</p> <p>Recreation and leisure</p>	<p>Mixed use</p> <p>Mixed, unknown type</p> <p>Mostly residential and commercial</p> <p>Mostly commercial and residential</p> <p>Mostly commercial and industrial</p> <p>Mostly residential and industrial</p> <p>Mostly industrial and commercial</p> <p>Mostly industrial and residential</p> <p>Industrial</p> <p>Industrial, unknown type</p> <p>Heavy industrial</p> <p>Light industrial</p> <p>Assembly</p> <p>Assembly, unknown type</p> <p>Religious gathering</p> <p>Arena</p> <p>Cinema or concert hall</p> <p>Other gatherings</p> <p>Government</p> <p>Government, unknown type</p> <p>Government, general services</p> <p>Government, emergency response</p> <p>Education</p> <p>Education, unknown type</p> <p>Pre-school facility</p> <p>School</p> <p>College/university, offices and/or classrooms</p> <p>College/university, research facilities and/or labs</p> <p>Other occupancy type</p>	<p>DATE OF CONSTRUCTION OR RETROFIT</p> <p>Year unknown</p> <p>Exact date of construction or retrofit</p> <p>Upper and lower bound for the date of construction or retrofit: between _____ and _____</p> <p>Latest possible date of construction retrofit</p> <p>Approximate date of construction or retrofit</p>
<p>BUILDING POSITION IN A BLOCK</p> <p>Unknown building position</p> <p>Detached building</p> <p>Adjoining building(s) on one side</p> <p>Adjoining building(s) on two sides</p> <p>Adjoining buildings on three sides</p> <p>Corner building</p> <p>BUILDING HEIGHT</p> <p>Height in meters _____ m</p> <p>Building tilt _____</p>	<p>NUMBER OF STOREY</p> <p>Number of storeys unknown</p> <p>Number of storeys above ground</p> <p>Range of number of storeys above ground _____</p> <p>Exact number of storeys above ground _____</p> <p>Approximate number of storeys above ground _____</p> <p>Number of storeys below ground</p> <p>Number of storeys below ground unknown</p> <p>Range of number of storeys below ground _____</p> <p>Exact number of storeys below ground _____</p> <p>Approximate number of storeys below ground _____</p> <p>Height of ground floor level above grade _____</p> <p>Height of ground floor level above _____</p> <p>Range of height of ground floor level above grade _____</p> <p>Exact height of ground floor level above grade _____</p> <p>Approximate height of ground floor level above grade _____</p> <p>Slope of the ground _____</p> <p>Slope of the ground unknown _____</p> <p>Slope of the ground _____</p>	<p>SHAPE OF THE BUILDING PLAN</p> <p>Unknown plan shape</p> <p>Square, solid</p> <p>Square, with an opening in plan</p> <p>Rectangular, solid</p> <p>Rectangular, with an opening in plan</p> <p>L-shape</p> <p>Curved, solid (e.g. circular, elliptical ovoid)</p> <p>Curved, with an opening in plan</p> <p>Triangular, solid</p> <p>Triangular, with an opening in plan</p> <p>Polygonal, solid (e.g. trapezoid, pentagon, hexagon)</p> <p>Polygonal, with an opening in plan</p> <p>E-shape</p> <p>H-shape</p> <p>S-shape</p> <p>T-shape</p> <p>U- or C-shape</p> <p>X-shape</p> <p>Y-shape</p> <p>Irregular plan shape</p> <p>Other (please sketch the building shape below)</p> <div style="text-align: right;">  </div>
<p>Example of local nomenclature:</p> 		<p>EXPOSURE AND CONSEQUENCES</p> <p>Number of Day Occupants _____</p> <p>Number of Night Occupants _____</p> <p>Number of Transit Occupants _____</p> <p>Number of Dwelling _____</p> <p>Plan Area (m²) _____</p> <p>Replacement cost (per m²) _____</p> <p>Number of Fatalities _____</p> <p>Number of Injured _____</p> <p>Number Missing _____</p>

General building damage

<p>Grade 1: Negligible to Slight Damage (<10%)</p> <p>Grade 2: Moderate Damage (10-30%)</p> <p>Grade 3: Substantial to Heavy Damage (30-60%)</p> <p>Grade 4: Very Heavy Damage (60-90%)</p> <p>Grade 5: Destruction (>90%)</p>	<p>Damage to building of reinforced concrete</p> <p>Grade 1: Negligible to slight damage (no structural damage, slight non-structural damage)</p> <p>Fine cracks in plaster over frame members or in walls at the base.</p> <p>Fine cracks in partitions and infills.</p> <p>Grade 2: Moderate damage (slight structural damage, moderate non-structural damage)</p> <p>Cracks in columns and beams of frames and in structural walls.</p> <p>Cracks in partition and infill walls; fall of brittle cladding and plaster.</p> <p>Falling mortar from the joints of wall panels.</p> <p>Grade 3: Substantial to heavy damage (moderate structural damage, heavy non-structural damage)</p> <p>Cracks in columns and beam column joints of frames at the base and at joints of coupled walls. Spalling of concrete cover, buckling of reinforced rods. Large cracks in partition and infill walls, failure of individual infill panels.</p> <p>Grade 4: Very heavy damage (heavy structural damage, very heavy non-structural damage)</p> <p>Large cracks in structural elements with compression failure of concrete and fracture of rebars; bond failure of beam reinforced bars; tilting of columns.</p> <p>Collapse of a few columns or of a single upper floor.</p> <p>Grade 5: Destruction (very heavy structural damage)</p> <p>Collapse of ground floor or parts (e. g. wings) of buildings.</p>	<p>Damage to masonry buildings</p> <p>Grade 1: Negligible to slight damage (no structural damage, slight non-structural damage)</p> <p>Hair-line cracks in very few walls. Fall of small pieces of plaster only. Fall of loose stones from upper parts of buildings in very few cases.</p> <p>Grade 2: Moderate damage (slight structural damage, moderate non-structural damage)</p> <p>Cracks in many walls. Fall of fairly large pieces of plaster.</p> <p>Partial collapse of chimneys.</p> <p>Grade 3: Substantial to heavy damage (moderate structural damage, heavy non-structural damage)</p> <p>Large and extensive cracks in most walls.</p> <p>Roof tiles detach. Chimneys fracture at the roof line; failure of individual non-structural elements (partitions, gable walls).</p> <p>Grade 4: Very heavy damage (heavy structural damage, very heavy non-structural damage)</p> <p>Serious failure of walls; partial structural failure of roofs and floors.</p> <p>Grade 5: Destruction (very heavy structural damage)</p> <p>Total or near total collapse.</p>
<p>Is there localised damage?</p> <p>No</p> <p>Yes (if yes please precise floor nb, location)</p> <p>Floor number: _____</p> <p>Location: _____</p>		

Structural System

Direction		
Direction X (Longitudinal)	Unknown direction	
	Parallel to street	
Direction Y (Transverse)	Unknown direction	
	Perpendicular to street	



Material		MAT_TYPE Material type		MAT_TECH Material technology	
L	T	L	T	L	T
	Unknown material		Unknown concrete technology		
	Concrete, unknown reinforcement		Cast-in-place concrete		
	Concrete, unreinforced		Precast concrete		
	Concrete, reinforced		Cast-in-place prestressed concrete		
	Concrete, composite with steel section		Precast prestressed concrete		
	Steel		Steel, unknown		
			Cold-formed steel members		
			Hot-rolled steel members		
			Steel, other		
	Metal (except steel)				
	Masonry, unknown reinforcement		Masonry unit, unknown		
	Masonry, unreinforced		Adobe blocks		
	Masonry, confined		Stone, unknown technology		
	Masonry, reinforced		Dressed stone		
			Fired clay unit, unknown type		
			Fire clay bricks		
			Concrete blocks, unknown type		
			Concrete block, solid		
			Concrete blocks, hollow		
			Masonry unit, other		
	Other				
	Material other				

Lateral-load resisting system	
L	T
	LLRS lateral-load resisting system
	Unknown lateral load-resisting system
	No lateral load-resisting system
	Moment frame
	Infilled frame
	Braced frame
	Post and Beam
	Wall
	Dual frame-wall system
	Flat slab/plate or waffle slab
	Infilled flat slab/plate or infilled waffle slab
	Hybrid lateral load-resisting system
	Other lateral load-resisting system

Structural Irregularity

STR_IRREG Regular or irregular		STR_HZIR_P		STR_HZIR_S		STR_VEIR_P		STR_VEIR_S	
L	T	L	T	L	T	L	T	L	T
	Unknown structural irregularity		Plan irregularity - primary				Vertical structural irregularity - primary		
	Regular structure		Plan irregularity - secondary				Vertical structural irregularity - secondary		
	Irregular structure								

The user can choose a maximum of two vertical and two plan irregularities for a building. However, if a building has two irregularities of the same type (plan/vertical), the user needs to prioritize them by identifying the primary irregularity first and the secondary irregularity next.

Structural system (Severity of Damage)

Negligible to Slight Damage (<10%)
Moderate Damage (10-30%)
Substantial to Heavy Damage (30-60%)
Very Heavy Damage (60-90%)
Destruction (>90%)



Exterior Attributes

Roof		ROOF_COVMAT Roof covering		ROOFSYSTYP Roof system type	
L	T	L	T	L	T
	Unknown roof shape		Unknown roof covering		Roof material, unknown
	Flat		Concrete roof without additional covering		Concrete roof, unknown
	Pitched		Heavy roof covering (slate, stone slab/clay or concrete tile)		Cast-in-place beamless reinforced concrete roof
	Sawtooth		Light roof covering (metal tile, shingle, membrane)		Cast-in-place beam-supported reinforced concrete roof
	Curved		Vegetative/Earthen roof covering		Precast concrete roof with concrete topping
	Complex regular		Solar panelled roofs		Masonry roof, unknown
	Complex irregular		Roof covering, other		Composite masonry and concrete roof system
	Roof shape, other				Metal roof, unknown
					Metal beams or trusses supporting light roofing
					Metal roof beams supporting precast concrete slabs
					Composite steel roof deck and concrete slab
					Wooden roof
					Fabric roof
					Roof material, other

Exterior walls/Façade

Unknown material of exterior walls
Concrete exterior walls
Concrete and masonry
Masonry exterior walls
Steel and masonry
Glass exterior walls
Metal exterior walls
Vegetative exterior walls
Wooden exterior walls
Plastic/vinyl exterior walls, various
Cement-based boards for exterior walls
Other

Exterior walls (Severity of Damage)	
L	T
	Negligible to Slight Damage (<10%)
	Moderate Damage (10-30%)
	Substantial to Heavy Damage (30-60%)
	Very Heavy Damage (60-90%)
	Destruction (>90%)

FLOOR_MAT Floor system material

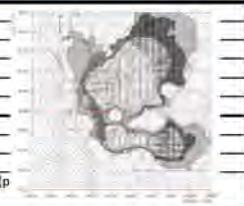
No elevated or suspended floor material (single-storey building)
Floor, material, unknown
Concrete floor, unknown
Cast-in-place beamless reinforced concrete floor
Cast-in-place beam-supported reinforced concrete floor
Precast concrete floor with reinforced concrete topping
Precast concrete floor without reinforced concrete topping
Masonry floor, unknown
Metal floor, unknown
Metal beams, trusses, or joists supporting light flooring
Metal floor beams supporting precast concrete slabs
Composite steel deck and concrete slab
Composite cast-in-place RC and masonry floor system
Wooden floor, unknown
Floor material, other
FLOOR_CONN Floor connections
Floor-wall diaphragm connection unknown
Floor-wall diaphragm connection not provided
Floor-wall diaphragm connection present

Foundation system

FOUNDN_SYS	
L	T
	Unknown foundation system
	Shallow foundation, with lateral capacity
	Shallow foundation, no lateral capacity
	Deep foundation, with lateral capacity
	Deep foundation, no lateral capacity
	Foundation, other

Soil type

Very soft clay
Silt or clay
Granular loose
Granular compact
Rock
Zone Mexican codes
Firm soil (Zone I)
Transition (Zone II)
Soft-soil (Zone IIIa, b, c or d) (p)



Non-structural elements

NON-STRUCTURAL WALL/ PARAPETS		
Does the building have this component?	<input type="checkbox"/> No <input type="checkbox"/> Yes (if yes, please fill the following information)	Repair price
Material	Type of damage	
Cast in place concrete	Parapet cracking	
Hollow concrete block	Parapet crushing	
Solid concrete block	Parapet locally falling out	
Hollow fired clay block	Parapet collapsed	
Solid brick	Non-structural wall cracking	
Sesimic performance features	Non-structural wall crushing	
No reinforcement	Non-structural wall locally falling out	
Steel reinforcement	Non-structural wall collapsed	
		Total cost to repair this component \$

EXTERIOR WINDOWS/GLAZING		
Does the building have this component?	<input type="checkbox"/> No <input type="checkbox"/> Yes (if yes, please fill the following information)	Repair price
Material	Type of damage	
Wood frame	Cracking	
Steel frame	Frame distortion	
Aluminium frame	Fall out	
Other:	Other:	
		Total cost to repair this component \$

STAIRS		
Does the building have this component?	<input type="checkbox"/> No <input type="checkbox"/> Yes (if yes, please fill the following information)	Repair price
Material	Type of damage	
Prefabricated steel	Non structural damage, local steel yielding	
Precast concrete	Local concrete cracking, localized concrete spalling	
Cast-in-place concrete	Localized steel yielding	
Other:	Buckling of steel, weld cracking	
Sesimic performance features	Extensive concrete cracking, concrete crushing	
Particular connection detailing	Extensive concrete cracking, concrete crushing, buckling of rebar	
	Loss of live load capacity. Connection and or weld fracture	
	Loss of live load capacity. Extensive concrete crushing, connection failure	
	Loss of live load capacity	
		Total cost to repair this component \$

WALL PARTITION		
Does the building have this component?	<input type="checkbox"/> No <input type="checkbox"/> Yes (if yes, please fill the following information)	Repair price
Material	Type of damage	
Gypsum with metal stud	Non structural damage, local steel yielding	
Gypsum with wood studs	Local concrete cracking, localized concrete spalling	
Gypsum + Wallpaper	Localized steel yielding	
Gypsum + Ceramic Tile	Buckling of steel, weld cracking.	
Sesimic performance features	Extensive concrete cracking, concrete crushing	
Fixed above	Extensive concrete cracking, concrete crushing, buckling of rebar	
Lateral braced above	Loss of live load capacity. Connection and or weld fracture	
Fixed below	Loss of live load capacity. Extensive concrete crushing, connection failure	
	Loss of live load capacity	
		Total cost to repair this component \$

DOORS		
Does the building have this component?	<input type="checkbox"/> No <input type="checkbox"/> Yes (if yes, please fill the following information)	Repair price
Material	Type of damage	
Wood	Cracking	
Metal	Crushing	
	Frame distortion	
		Total cost to repair this component \$

SUSPENDEED CEILINGS		
Does the building have this component?	<input type="checkbox"/> No <input type="checkbox"/> Yes (if yes, please fill the following information)	Repair price
Type	Type of damage	
Vertical support only	5 % of ceiling grid and tile damage	
Vertical and Lateral support	30% of ceiling grid and tile damage	
Sesimic performance features	50% of ceiling grid and tile damage	
Braced	Dropped acoustical tile	
Unbraced	Perimeter damage	
	Separation of runners	
		Total cost to repair this component \$

NON-SUSPENDEED CEILINGS		
Does the building have this component?	<input type="checkbox"/> No <input type="checkbox"/> Yes (if yes, please fill the following information)	Repair price
Material	Type of damage	
Plasterboard	Cracking	
Drywall	Local spalling	
Other:	Collapse	
	Other:	
		Total cost to repair this component \$

FLOOR FINISHES		
Does the building have this component?		<input type="checkbox"/> No <input type="checkbox"/> Yes (if yes, please fill the following information)
		Repair price
Material	Type of damage	
Stone	Falling from building	
Tile	Damaged panels and connections	
Glass	Crush	
Other:	Other:	
		Total cost to repair this component \$

ELEVATORS		
Does the building have this component?		<input type="checkbox"/> No <input type="checkbox"/> Yes (if yes, please fill the following information)
		Repair price
Type	Type of damage	
Traction geared	Controller anchorage failed, and or machine anchorage failed, and or motor generator anchorage failed, and or governor anchorage failed, and or rope guard failures.	
	Rail distortion, and or intermediate bracket separate and spread, and or counterweight bracket break or bend, and or car bracket break or bend, and or car guide shoes damaged, and or counterweight guide shoes damaged, and or counterweight frame distortion, and or tail sheave dislodged and/or twisted	
	Cab stabilizers bent, or cab walls damaged, or cab doors damaged.	
	Cab ceiling damaged.	
	Damaged controls.	
Hydraulic	Damaged vane and hoist-way switches, and or bent cab stabilizers, and or damaged car guide shoes.	
	Damaged entrance and car door, and or flooring damage.	
	Oil leak in hydraulic line, and or hydraulic tank failure.	
		Total cost to repair this component \$

PIPING		
Does the building have this component?		<input type="checkbox"/> No <input type="checkbox"/> Yes (if yes, please fill the following information)
		Repair price
Material	Type of damage	
Plastic	Minor leakage at flange connections	
Copper	Pipe Break	
Steel	Other:	
Aluminium		
Cast iron		
Other:		
Sesimic performance features		
Braced		
Unbraced		
		Total cost to repair this component \$

FIRE PROTECTION		
Does the building have this component?		<input type="checkbox"/> No <input type="checkbox"/> Yes (if yes, please fill the following information)
		Repair price
Type	Type of damage	
Wet pipe	Spraying & Dripping Leakage at joints	
Dry pipe	Joints Break - Major Leakage	
Other:		
		Total cost to repair this component \$

HEATING SYSTEMS		
Does the building have this component?		<input type="checkbox"/> No <input type="checkbox"/> Yes (if yes, please fill the following information)
		Repair price
Material	Type of damage	
Unit heater	Sliding	
Boiler	Overturning. Broken/bent bolts	
Other:	Broken gas and exhaust lines	
Sesimic performance features		
Unanchored	Loss of function	
Anchored	Other:	
		Total cost to repair this component \$

COOLING SYSTEMS		
Does the building have this component?		<input type="checkbox"/> No <input type="checkbox"/> Yes (if yes, please fill the following information)
		Repair price
Material	Type of damage	
Packaged chiller	Sliding	
Rooftop air cond.	Overturning. Broken/bent bolts	
Other:	Leaking refrigerant	
Sesimic performance features		
Unanchored	Loss of function	
Anchored	Other:	
		Total cost to repair this component \$

DUCTS		
Does the building have this component?		<input type="checkbox"/> No <input type="checkbox"/> Yes (if yes, please fill the following information)
		Repair price
Material	Type of damage	
Plastic	Sliding	
Steel	Overturning	
Other:	Other:	
Sesimic performance features		
Unanchored		
Anchored		
		Total cost to repair this component \$

LIGHTING		
Does the building have this component?	<input type="checkbox"/> No <input type="checkbox"/> Yes (if yes, please fill the following information)	Repair price
Type	Type of damage	
Incandescent	Disassembly of rod system at connections with horizontal light fixture, low cycle fatigue failure of the threaded rod, pullout of rods from ceiling assembly.	
Neon	Loss of function	
Other:	Other:	
Sesimic performance features		
Non seismic		
Seismically rated		Total cost to repair this component \$

POWER GENERATORS		
Does the building have this component?	<input type="checkbox"/> No <input type="checkbox"/> Yes (if yes, please fill the following information)	Repair price
Type	Type of damage	
Diesel generator	Damaged, inoperative. Pipes and nozzles damaged.	
Petrol generator	Anchorage failure.	
Other:	Damaged, inoperative but anchorage is OK. Pipes and nozzles damaged.	
Sesimic performance features	Damaged, inoperative. Drive shaft misalignment.	
Unanchored	Anchorage failure & Equipment damaged beyond repair.	
Anchored	Damaged, inoperative. Minor electrical damage, e.g., failed relay.	
	Damaged, inoperative but anchorage is OK	
	Damaged, inoperative. Exhaust line disconnected at expansion bellows.	
	Damaged, inoperative. Exhaust line disconnected at expansion bellows.	
	Equipment is damaged and inoperative but anchorage is OK.	
	Other:	Total cost to repair this component \$

TANKS		
Does the building have this component?	<input type="checkbox"/> No <input type="checkbox"/> Yes (if yes, please fill the following information)	Repair price
Material	Type of damage	
Plastic	Pipe break	
Metal	Tank or vessel rupture	
Other:	Other:	
Sesimic performance features		
Unanchored		
Anchored		Total cost to repair this component \$

FIXED FURNISHINGS		
Does the building have this component?	<input type="checkbox"/> No <input type="checkbox"/> Yes (if yes, please fill the following information)	Repair price
Type	Type of damage	
Fixed artwork	Only sliding, no damage	
Fixed casework	Cracks and crushing. Minor damage.	
Other:	Damaged, loss of function	
Sesimic performance features	Other:	
		Total cost to repair this component \$

BOOKCASE		
Does the building have this component?	<input type="checkbox"/> No <input type="checkbox"/> Yes (if yes, please fill the following information)	Repair price
Material	Type of damage	
Wood	Bookcase slides. Some content fall over. No damage to the bookcase	
Metal	Book case falls over and contents are scattered. Likely damage to bookcase.	
Other:	Other:	
Sesimic performance features		
Unanchored laterally		
Anchored laterally		Total cost to repair this component \$

FILING CABINET		
Does the building have this component?	<input type="checkbox"/> No <input type="checkbox"/> Yes (if yes, please fill the following information)	Repair price
Type	Type of damage	
Wood	Sliding. Some content fall over. No damage to the bookcase	
Metal	Filing cabinet falls over and contents are scattered. Likely damage to file cabinet.	
Other:	Other:	
Sesimic performance features		
Unanchored		
Anchored		Total cost to repair this component \$

DESK		
Does the building have this component?	<input type="checkbox"/> No <input type="checkbox"/> Yes (if yes, please fill the following information)	Repair price
Type	Type of damage	
Wood	Sliding. Some content fall over. No damage to the bookcase	
Metal	Filing cabinet falls over and contents are scattered. Likely damage to file cabinet.	
Other:	Other:	
Sesimic performance features		
Unanchored		
Anchored		Total cost to repair this component \$

DESKTOP COMPUTER UNITS		
Does the building have this component?	<input type="checkbox"/> No <input type="checkbox"/> Yes (if yes, please fill the following information)	Repair price
Type	Type of damage	
<input type="checkbox"/> Fixed artwork	Only sliding, no damage	
<input type="checkbox"/> Fixed casework	Cracks and crushing. Minor damage.	
<input type="checkbox"/> Other:	Damaged, loss of function	
Sesimic performance features	Other:	Total cost to repair this component \$
<input type="checkbox"/>		

TV SETS		
Does the building have this component?	<input type="checkbox"/> No <input type="checkbox"/> Yes (if yes, please fill the following information)	Repair price
Type	Type of damage	
<input type="checkbox"/> Cathodic panel display	Only sliding, no damage	
<input type="checkbox"/> Flat panel display	Cracks and crushing. Minor damage.	
<input type="checkbox"/> Other:	Damaged, loss of function	
Sesimic performance features	Other:	Total cost to repair this component \$
<input type="checkbox"/> Unanchored		
<input type="checkbox"/> Anchored		

Building owner contact information:

Name of contractors/subcontractors involved in the repairs:

Building sketch

Pictures (please indicate the name of the first and last picture taken for this building)

APPENDIX

C

GEM Building Taxonomy v2.0

C.1 Overview of the GEM building taxonomy v2.0

Table C.1: Overview of the GEM building taxonomy v2.0 categories

Attribute group	GEM attribute reference	Attribute	Attribute levels	Example options	Example	
Structural system	1	Direction	Direction of the building	Building principal axis parallel or perpendicular to street	Direction (longitudinal) Parallel to street Enrique Rebsamen	
	2	Material of LLRS	Material type (Level 1)	Concrete, steel, metal, masonry, earth, wood	Material type: Concrete reinforced	
			Material type (Level 2)	Cast-in place, precast, cold-formed, hot-rolled	Material technology: Cast-in place concrete	
			Material properties (Level 3)	No mortar, mud mortar, cement mortar	Material properties: Unknown	
	3	Lateral Load Resisting System (LLRS)	Type of lateral load-resisting system (Level 1)	Moment frame, infilled frame, braced frame	Type of lateral load-resisting system: Infilled frame	
System ductility (Level 2)			Ductile, non-ductile	System ductility: Unknown		
Generic building information	4	Building height	Height expressed in number of storeys	Number of storeys above ground, below ground, height of ground floor level above grade	Exact number of storeys above ground: 8	
	5	Date of Construction or Retrofit	Construction or retrofit completed	Exact date, approximate date, latest possible date of construction or retrofit	Approximate date of construction or retrofit: 1980	
		Occupancy	Building occupancy class - general (Level 1)	Residential, Commercial and public, mixed use, industrial	Building occupancy: Residential	
			Building occupancy class - detail (Level 2)	Single dwelling, 10-19 units, retail trade, school	System ductility: Unknown	
Exterior Attributes	7	Building Position within a block		Detached building, adjoining buildings	Corner building	
	8	Shape of the Building Plan	Plan shape (footprint)	Rectangular solid, rectangular with an opening in plan, L-shaped	Plan shape: L-Shape	
	9	Structural Irregularity	Regular or irregular (Level 1)			Irregular structure
			Plan irregularity or vertical irregularity (Level 2)	Plan irregularity - primary, Plan irregularity - secondary	Two horizontal structural irregularities, one vertical irregularity	
Type of irregularity (Level 3)			Torsion eccentricity, re-entrant corner, soft storey, cripple wall, short column, pounding potential, setback, change in vertical structure	Torsion eccentricity (primary horizontal), re-entrant corner (secondary horizontal), soft storey (primary vertical)		
10	Exterior walls	Exterior walls	Concrete, glass, vegetative exterior wall	Concrete and masonry		
Roof/ Floor/ Foundation	11	Roof	Roof shape (Level 1)	Flat, pitched, curved	Roof shape: Flat	
			Roof covering (Level 2)	Concrete roof, without additional covering, membrane roof covering	Roof covering: Unknown	

Continuation of Table C.1					
Attribute group	GEM attribute reference	Attribute	Attribute levels	Example options	Example
Roof/ Floor/ Foundation			Roof system material (Level 3)	Masonry roof, concrete roof	Roof system material: Unknown
			Roof system type (Level 4)	Vaulted masonry, cast-in-place beamless reinforced concrete roof	Roof system type: Unknown
			Roof connections (Level 5)	Roof tie-down present	Roof connections: Unknown
	12	Floor	Floor system material (Level 1)	Masonry floor, concrete floor	Floor system material: Unknown
			Floor system type (Level 2)	Shallow-arched masonry floor, wooden floor	Floor system type: Unknown
			Floor connections (Level 3)	Floor-wall diaphragm connection present	Floor connections: Unknown
13	Foundation System	Foundation System	Shallow/deep foundation, with lateral or no lateral capacity	Foundation System: Unknown	

C.2 Comparison of the GEM assessment methodology

Table C.2: Comparison of the GEM assessment methodology vs. the local Mexican procedure

Categories	Assessment based on the GEM Building Taxonomy	Assessment following the local Mexican procedure
Language of survey	English	Spanish
Building taxonomy	Based on GEM Building Taxonomy v2.0. Applicable to any region in the world	No reference to a standard international recognised building taxonomy. Assessment follows general questions about the building
Damage scale	European Macro-seismic scale EMS-98	European Macro-seismic scale EMS-98
Structural system	Distinguish between lateral and transverse direction of the building. Possible to define two building lateral-load resisting systems for two principle directions of the building	Only one structural system type noted, no information collected for different building directions
Structural system material	Taxonomy captures a wide range of building materials available worldwide (e.g. earth or bamboo are included) Confusion exist about classifying concrete frames with masonry infills as either concrete or masonry	General definition of the material (concrete, masonry, steel). Good level of details for concrete sub types
Structural regularity	Definition of horizontal and vertical irregularities with specific terms such as torsion eccentricity, re-entrant corner, soft storey	Subjective and non-specific definition (good, intermediate, bad)
Non-structural elements	Capture of damage to non-structural elements. Assessment form is highly detailed and complex	Assess general level of damage of exterior (e.g. windows, façade, balcony) and internal (e.g. partition walls, ceilings, lamps) non-structural elements
Tool for data collection	Paper form, Android mobile app and a windows software	Paper form or online form (require internet connection)
Data processing	Data Model is aligned to the GED4GEM and GEMECD data structures. IDCT Mobile Tools can directly populate both the GEM exposure and consequences databases	Data processing and exporting has to be done manually or translation via external tools
Methodology validation	GEM building taxonomy validated by an EERI team which described it as "highly functional, robust and able to describe different buildings around the world"	Tool specific to Mexico. At the time of writing, no validation from external peer review available
Output/ Export of the data collected	CSV file, Google Earth-compatible kmz file, Shapefile, Direct upload to the Global Exposure Database (GED)	Excel file

APPENDIX **D**

Soil code

Table D.1: Soil code (Land Resource Information Systems (LRIS), 2010)

Code	Order	Group	Series
BFA	Brown	Firm	Glenroy, Summit
BFM	Brown	Firm	Glenroy, Kakahu, Mt Somers, Okuku, Skipton+Kakahu
BFP	Brown	Firm	Lismore, Lismore+Pahau
BFT	Brown	Firm	Gorge
BOA	Brown	Orthic	Ashwick, Hororata, Lyndhurst, Lyndhurst+Ruapuna, Rapaki, Ruapuna, Staveley
BST	Brown	Sandy	Halkett, Halkett+Eyre, Halkett+Templeton+Eyre, Waikuku
EMT	Melanic	Mafic	Cashmere, Evans
EVT	Melanic	Vertic	Waiareka
GOJ	Gley	Orthic	Waterton, Waterton+Temuka
GOO	Gley	Orthic	Taitapu, Temuka+Waterton+Windermere, Waimairi, Willowby, Windermere
GOT	Gley	Orthic	Coopers-Creek, Haylands, Horotane, Temuka, Temuka+Windermere, Willowby
GRQ	Gley	Orthic	Motukarara
GRT	Gley	Recent	Taitapu, Taitapu+Kaiapoi, Taitapu+Motukarara, Taitapu+Waikuku
GST	Gley	Sandy	Aranui complex
OHM	Organic	Humic	Waimairi, Windermere
PIM	Pallic	Immature	Heathcote, Wakanui, Wakanui+Pahau, Wakanui+Templeton, Wakanui+Temuka
PIT	Pallic	Immature	Clifton, Glasnevin, Kiwi, Mayfield+Hororata, Paparua, Scarborough, Taiko, Templeton, Templeton+Eyre, Templeton+Halkett, Templeton+Taitapu, Templeton+Wakanui
PJC	Pallic	Argillic	Glenmark+Amberley, Glenmark+Waipara
PJM	Pallic	Argillic	Lowcliffe, Lowcliffe+Templeton+Waterton, Pahau, Pahau+Darnley
PJT	Pallic	Argillic	Amberley, Darnley, Darnley+Ashley, Darnley+Mayfield, Darnley+Pahau, Mayfield, Mayfield+Darnley
PLT	Pallic	Laminar	Hatfield, Hatfield+Lismore
PPX	Pallic	Perch-gley	Ashley, Ashley+Mairaki, Oxford, Takahe
PXJ	Pallic	Fragic	Waipara, Waipara+Amberley
PXM	Pallic	Fragic	Mairaki, Mairaki+Ashley
RFMQ		Fluvial Recent	Greenpark
RFMW	Recent	Fluvial	Kaiapoi, Kaiapoi+Taitapu, Kaiapoi+Waimakariri
RFT	Recent	Fluvial	Rangitata, Rangitata+Fereday, Rangitata+Kaiapoi, Rangitata+Selwyn, Rangitata+Te Kakahi, Selwyn, Selwyn+Kaiapoi, Selwyn+Rangitata, Taumutu, Taumutu+Taitapu
RFW	Recent	Fluvial	Rakaia, Rakaia+Kaiapoi, Rakaia+Taitapu, Rakaia+Waimakariri, Waimakariri, Waimakariri over Templeton, Waimakariri over Templeton+Rakaia, Waimakariri+Rakaia
ROM	Recent	Orthic	Kaiapoi, Springburn, Springburn+Wakanui
ROT	Recent	Orthic	Highbank, Kowai

Continuation of Table D.1			
Code	Order	Group	Series
ROW	Recent	Orthic	Barrhill, Eyre, Eyre+Halkett, Eyre+Templeton, Glasnevin, Paparua, Terrace scarp
RST	Recent	Sandy	Fereday, Fereday+Rangitata, Fereday+Waimakariri, Kairaki, Taylors Mistake
WF	Raw	Fluvial	River Bed+Rangitata, River Bed+Wakanui+Coopers Creek
WGF	Raw	Gley	Te Kakahi, Te Kakahi+Rangitata
WS	Raw	Sandy	Coastal sand and gravel

NZ Deprivation Index 2013

Christchurch

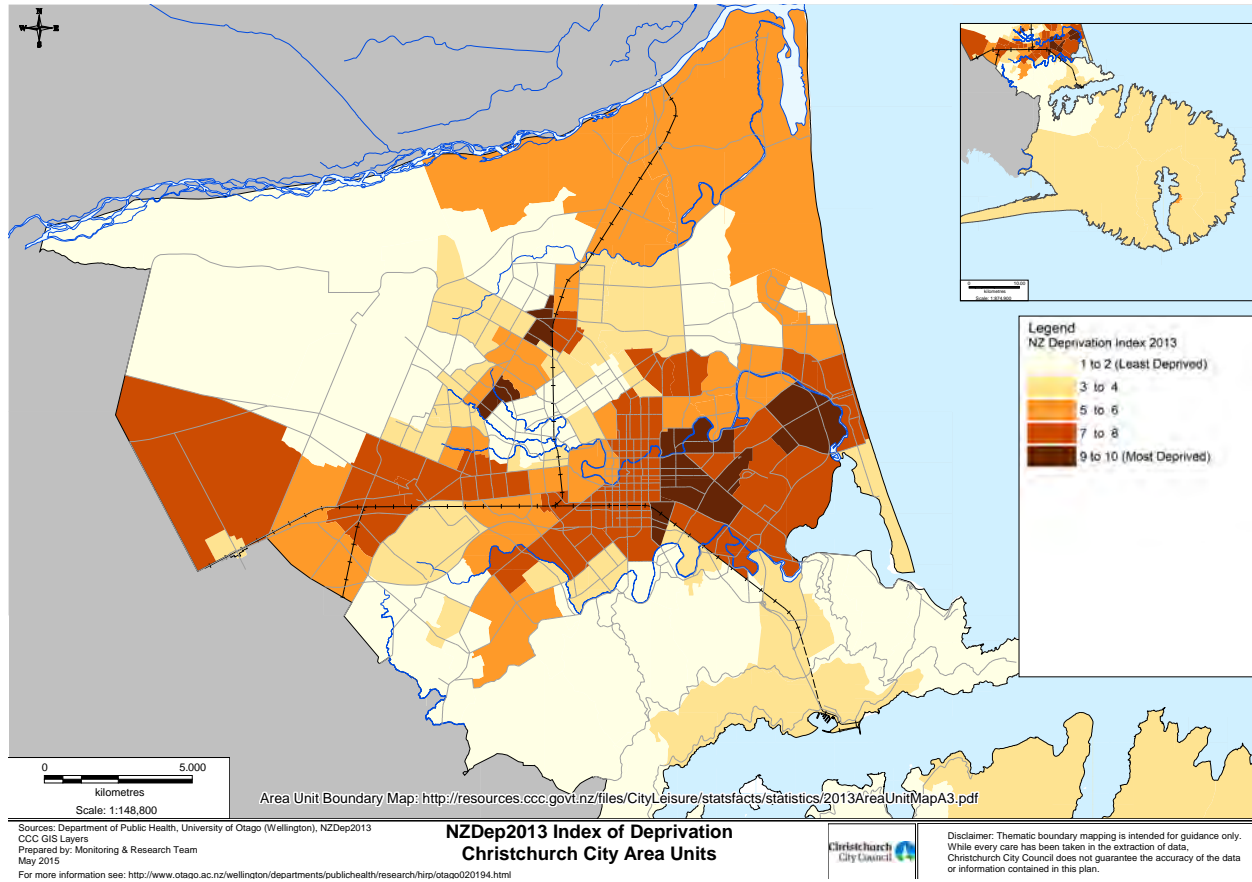


Figure E.1: NZDep2013 Index of Deprivation Christchurch City Area Units (Christchurch City Council, 2015)

References

- Allen, L., Brzev, S., Charleson, A., Scawthorn, C., & Silva, V. (2015). GEM Building Taxonomy – an open global building classification system. *2015 NZSEE Conference*.
- Allen, L., Charleson, A., Brzev, S., & Scawthorn, C. (2013). Glossary for GEM taxonomy. <https://taxonomy.openquake.org/>
- Allen, R., Cochran, E., Huggins, T., Miles, S., & Otegui, D. (2017). Earthquake early warning in Mexico City. http://learningfromearthquakes.org/2017-09-19-puebla-mexico/images/2017_09_19_Puebla_Mexico/pdfs/4-Earthquake_Early_Warning_Mex_Recon-Briefing.pdf
- Applied Technology Council. (1989). *ATC-20 Procedures for Postearthquake Safety Evaluation of Buildings* (tech. rep.). Applied Technology Council (ATC). Redwood City, California.
- Applied Technology Council (ATC). (1995). *ATC-20-2 Addendum to the ATC-20 Postearthquake Building Safety Evaluation Procedures* (tech. rep.). Applied Technology Council (ATC). Redwood City, California.
- Applied Technology Council (ATC). (1996a). *ATC-32 Improved Seismic Design Criteria for California Bridges: Provisional Recommendations* (tech. rep.). Applied Technology Council. Redwood City, CA, USA.
- Applied Technology Council (ATC). (1996b). *ATC-40 Seismic Evaluation and Retrofit of Concrete Buildings* (tech. rep. November 1996). Applied Technology Council (ATC). Redwood City, California.
- Arellano-Mendez, E., Juarez-Garcia, H., & Gomez-Bernal, A. (2004). Vulnerabilidad Sísmica de la Colonia Roma, Ciudad de Mexico. *XIV Congreso Nacional de Ingeniería Estructural*, 1–23.

- Aslani, H., & Miranda, E. (2005). Fragility assessment of slab-column connections in existing non-ductile reinforced concrete buildings. *Journal of Earthquake Engineering*, 9(6), 777–804. <https://doi.org/10.1080/13632460509350566>
- Badillo-Almaraz, H., Whittaker, A. S., & Reinhorn, A. M. (2007). Seismic Fragility of Suspended Ceiling Systems. *Earthquake Spectra*, 23(1), 21–40. <https://doi.org/10.1193/1.2357626>
- Baggio, C., Bernardini, A., Colozza, R., Corazza, L., Bella, M., Di Pasquale, G., Dolce, M., Goretti, A., Martinelli, A., Orsini, G., Papa, F., & Zuccaro, G. (2007). *Field Manual for post-earthquake damage and safety assessment and short term countermeasures (AeDES)* (tech. rep.). Joint Research Centre. Ispra, Italy. <https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/field-manual-post-earthquake-damage-and-safety-assessment-and-short-term-countermeasures>
- Barber, M. (2014). Data science concepts you need to know! <https://towardsdatascience.com/introduction-to-statistics-e9d72d818745>
- Bellagamba, X., Lee, R., & Bradley, B. A. (2019). A neural network for automated quality screening of ground motion records from small magnitude earthquakes. *Earthquake Spectra*, 35(4), 1637–1661. <https://doi.org/10.1193/122118eqs292m>
- Bevere, L., & Balz, G. (2012). *Lessons from recent major earthquakes* (tech. rep.). Swiss Reinsurance Company Ltd. Zurich, Switzerland. <https://www.swissre.com/institute/library/Expertise-Publication-lessons-from-recent-major-earthquakes.html>
- Borg, R., Indirli, M., Rossetto, T., & Kouris, L. (2010). L'Aquila earthquake April 6th, 2009: The damage assessment methodologies. *COST ACTION C26: Urban Habitat Constructions under Catastrophic Events - Proceedings of the Final Conference*, (September), 557–564.
- Bradley, A. B. (2010). Epistemic Uncertainties in Component Fragility Functions. *Earthquake Spectra*, 26(1), 41–62. <https://doi.org/10.1193/1.3281681>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Broccardo, M., Esposito, S., & Stojadinovic, B. (2016). Application of the PEER-PBEE Framework for Probabilistic Resilience Assessment of a Structural System.

- International Symposium on Sustainability and Resiliency of Infrastructure*, (November). <https://doi.org/10.13140/RG.2.2.20215.42406>
- Brown, P. C., & Lowes, L. N. (2007). Fragility functions for modern reinforced-concrete beam-column joints. *Earthquake Spectra*, 23(2), 263–289. <https://doi.org/10.1193/1.2723150>
- Brzev, S., Scawthorn, C., Charleson, A., Allen, L., Greene, M., Jaiswal, K., & Silva, V. (2013). *GEM Building Taxonomy Version 2.0* (tech. rep.). GEM Foundation. Pavia, Italy. <https://doi.org/10.13117/GEM.EXP-MOD.TR2013.02>
- Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., Vanderplas, J., Joly, A., Holt, B., & Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 108–122. <http://arxiv.org/abs/1309.0238>
- Burkov, A. (2020). *Machine Learning Engineering*. True Positive Inc.
- Burton, H. V., Deierlein, G., Lallemand, D., & Lin, T. (2016). Framework for Incorporating Probabilistic Building Performance in the Assessment of Community Seismic Resilience. *Journal of Structural Engineering*, 142(8), C4015007. [https://doi.org/10.1061/\(ASCE\)ST.1943-541X.0001321](https://doi.org/10.1061/(ASCE)ST.1943-541X.0001321)
- Cattari, S., Ottonelli, D., Pinna, M., Lagomarsino, S., Clark, W., Giovinazzi, S., Ingham, J., Marotta, A., Liberatore, D., Sorrentino, L., Leite, J., Lourenco, P., & Goded, T. (2015). Preliminary results from damage and vulnerability analysis of URM churches after the Canterbury earthquake sequence 2010-2011. *Proceeding of the 2015 New Zealand Society for Earthquake Engineering (NZSEE) Conference*, 10.
- Centro de Instrumentación y Registro Sísmico (CIRES). (2017). Búsqueda de registros. http://www.cires.org.mx/registro_busquedas_sismo_es.php
- Christchurch City Council. (2015). NZDep2013 Index of Deprivation Christchurch City Area Units. <https://www.ccc.govt.nz/assets/Documents/Culture-Community/Stats-and-facts-on-Christchurch/Facts-Stats-and-Figures/Maps-Directory/map-deprivation-quintile-area-unit.pdf>

- Christophersen, A., Hainzl, S., Gerstenberger, M. C., Rhoades, D. A., & Smith, E. G. C. (2013). *The Canterbury sequence in the context of global earthquake statistics* (tech. rep. No. 196). GNS Science. Wellington, New Zealand.
- Colegio de Ingenieros Civiles de México (CICM). (2017a). *Resumen preliminar de danos de los inmuebles inspeccionados por las brigadas del CICM del sismo del 19/09/2017* (tech. rep.). Colegio de Ingenieros Civiles de México (CICM). Mexico City, Mexico. <https://www.smie.org.mx/archivos/eventos/mantengase-informado/2017-noviembre-resumen-preliminar-danos-inmuebles-inspeccionados-brigadas-cicm.pdf>
- Colegio de Ingenieros Civiles de México (CICM). (2017b). Sismo 19 de septiembre. <https://www.sismosmexico.org/>
- Cornell, C. A., & Krawinkler, H. (2000). Progress and Challenges in Seismic Performance Assessment. *PEER Center News*, 3(2). <https://apps.peer.berkeley.edu/news/2000spring/performance.html>
- Cournapeau, D. (2007). scikit-learn. <https://scikit-learn.org/stable/>
- Cox, J. E. (1978). *Soils and agriculture of part Paparua County, Canterbury, New Zealand*. (tech. rep.). Soil Bureau, DSIR. Wellington, New Zealand. <https://iris.scinfo.org.nz/document/9201-soil-bureau-bulletin-34-soils-and-agriculture-of-part-paparua-county/>
- Cutfield, M. R. (2015). *Advanced methods for performance-based seismic loss assessment and their application to a base isolated and conventional office building* (Doctoral dissertation). The University of Auckland. The University of Auckland. <http://hdl.handle.net/2292/27716>
- Del Gaudio, C., De Martino, G., Di Ludovico, M., Manfredi, G., Prota, A., Ricci, P., & Verderame, G. M. (2016). Empirical fragility curves from damage data on RC buildings after the 2009 L'Aquila earthquake. *Bulletin of Earthquake Engineering*, 15(4), 1425–1450. <https://doi.org/10.1007/s10518-016-0026-1>
- Del Gaudio, C., Ricci, P., Verderame, G. M., & Manfredi, G. (2017). Urban-scale seismic fragility assessment of RC buildings subjected to L'Aquila earthquake. *Soil Dynamics and Earthquake Engineering*, 96(March), 49–63. <https://doi.org/10.1016/j.soildyn.2017.02.003>

- Deloitte Access Economics. (2015). *Four years on: Insurance and the Canterbury Earthquakes* (tech. rep. February). Deloitte Access Economics. <https://www.vero.co.nz/documents/newsroom/deloitte-vero-four-years-on-insurance-canterbury-earthquakes-report-february-2015.pdf>
- Dhakal, R. (2011). *Structural design for earthquake resistance: past, present and future* (tech. rep.). Canterbury Earthquake Royal Commission. Christchurch, New Zealand. <http://canterbury.royalcommission.govt.nz/documents-by-key/2011-09-2753>
- Díaz, A., Murren, P., & Walker, S. (2017). *Preliminary observations in the aftermath of the September 19, 2017 Puebla-Morelos earthquake* (tech. rep.). Skidmore, Owings & Merrill LLP. San Francisco, CA.
- Drayton, M. J., & Verdon, C. L. (2013). Consequences of the Canterbury earthquake sequence for insurance loss modelling. *Proceeding of the 2013 New Zealand Society for Earthquake Engineering Conference*, 1–7. http://db.nzsee.org.nz/2013/Paper_44.pdf
- Du, M., Liu, N., & Hu, X. (2020). Techniques for interpretable machine learning. *Communications of the ACM*, 63(1), 68–77. <https://doi.org/10.1145/3359786>
- Dua, D., & Graff, C. (2019). *UCI Machine Learning Repository*. University of California, School of Information; Computer Science. <http://archive.ics.uci.edu/ml>
- Earthquake Commission (EQC). (2017). Budget Announcement : EQC levy to increase. <https://www.eqc.govt.nz/news/budget-announcement-eqc-levy-to-increase>
- Earthquake Commission (EQC). (2019a). *Briefing to the Public Inquiry into the Earthquake Commission: Canterbury Home Repair Programme* (tech. rep. 24 June 2019). Earthquake Commission (EQC). Wellington, New Zealand. https://www.eqc.govt.nz/sites/public_files/documents/Inquiry/7.%20Canterbury%20Home%20Repair%20Programme%20Briefing%20rs.pdf
- Earthquake Commission (EQC). (2019b). EQC Insurance. <https://www.eqc.govt.nz/what-we-do/eqc-insurance>
- Earthquake Commission (EQC). (2019c). EQC welcomes Act changes and gets ready to respond. <https://www.eqc.govt.nz/news/eqc-welcomes-act-changes-and-gets-ready-to-respond>

- Earthquake Commission (EQC). (2019d). International reinsurers continue to provide cover to EQC. <https://www.eqc.govt.nz/news/international-reinsurers-continue-to-provide-cover-to-etc>
- Earthquake Commission (EQC). (2019e). The Natural Disaster Fund. <https://www.eqc.govt.nz/about-etc/our-role/ndf>
- Earthquake Commission (EQC), Ministry of Business Innovation and Employment (MBIE), & New Zealand Government. (2012). New Zealand Geotechnical Database (NZGD). <https://www.nzgd.org.nz/Default.aspx>
- Earthquake Engineering Research Institute (EERI), & International Association for Earthquake Engineering (IAEE). (2000). World Housing Encyclopedia (WHE). <http://www.world-housing.net/>
- Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1(1), 54–75. <https://doi.org/10.1214/ss/1177013815>
- Eleftheriadou, A. K., & Karabinis, A. I. (2008). Damage Probability Matrices Derived From Earthquake Statistical Data. *Proceedings of the 14th World Conference on Earthquake Engineering, October 12-17, 2008*.
- Ellingwood, B. R., Celik, O. C., & Kinali, K. (2007). Fragility assessment of building structural systems in Mid-America. *Earthquake Engineering & Structural Dynamics*, 36(13), 1935–1952. <https://doi.org/10.1002/eqe.693>
- Elwood, K. J., Comerio, M., Cubrinovski, M., Davis, C. A., Johnston, D., O'Rourke, T., & Pampanin, S. (2014). Preface. *Earthquake Spectra*, 30(1), 7–9. <https://doi.org/10.1193/8755-2930-30.1.FMI>
- Esri. (2019). ArcGIS Desktop 10.7.1.
- Federal Emergency Management Agency (FEMA). (1997). *FEMA 273 - NEHRP Guidelines for the Seismic Rehabilitation of Buildings* (tech. rep.). Applied Technology Council (ATC). Redwood City, CA, USA.
- Federal Emergency Management Agency (FEMA). (2000). *FEMA 356 - Prestandard and Commentary for the Seismic Rehabilitation of Buildings* (tech. rep. No. 1). Applied Technology Council (ATC). Redwood City, CA, USA.

- Federal Emergency Management Agency (FEMA). (2018). *FEMA P-58-6: Guidelines for Performance-Based Seismic Design of Buildings* (tech. rep. December). Federal Emergency Management Agency (FEMA). Washington, D.C. <https://femap58.atcouncil.org/documents/fema-p-58/28-fema-p-58-6-guidelines-for-design/file>
- Feltham, C. (2011). Insurance and reinsurance issues after the Canterbury earthquakes. <https://www.parliament.nz/en/pb/research-papers/document/00PlibCIP161/insurance-and-reinsurance-after-canterbury-earthquakes>
- Ferrari, G., & McConnell, A. (2005). Robert Mallet and the 'Great Neapolitan earthquake' of 1857. *Notes and Records of the Royal Society*, 59(1), 45–64. <https://doi.org/10.1098/rsnr.2004.0076>
- Fikri, R., Dizhur, D., Walsh, K., & Ingham, J. (2018). Seismic performance of Reinforced Concrete Frame with Masonry Infill buildings in the 2010/2011 Canterbury, New Zealand earthquakes. *Bulletin of Earthquake Engineering*, (0123456789), pp 1–21. <https://doi.org/10.1007/s10518-018-0476-8>
- Fisher, A., Rudin, C., & Dominici, F. (2018). All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *arXiv*, (6). <http://arxiv.org/abs/1801.01489>
- Fortmann-Roe, S. (2012). Understanding the Bias- Variance Tradeoff. <http://scott.fortmann-roe.com/docs/BiasVariance.html>
- Foulser-Piggott, R., Bevington, J., & Vicini, A. (2014). *End-to-end demonstration of the Inventory Data Capture Tools* (tech. rep.). GEM Foundation. Pavia, Italy. <https://storage.globalquakemodel.org/media/publication/DATA-CAPTURE-GEM-EndtoEnd-IDCT-Demo-201406-V01.pdf>
- Foulser-Piggott, R., Vicini, A., Verrucci, E., Bevington, J., & Shelley, W. (2013). *IDCT Mobile Tools - Field Test Reports* (tech. rep.). GEM Foundation. Pavia, Italy.
- Galvis, F., Miranda, E., Heresi, P., Dávalos, H., & Silos, J. R. (2017). *Preliminary Statistics of Collapsed Buildings in Mexico City in the September 19 , 2017 Puebla-Morelos Earthquake* (tech. rep. October). EERI. <http://www.learningfromearthquakes.org/2017-09-19-puebla-mexico/?id=65:preliminary-statistics-of-collapsed-buildings-in-mexico-city-in-the-september-19-2017-puebla-morelos-earthquake>

- GeoNet. (2010). M 7.2 Darfield (Canterbury) Sat, Sep 4 2010. <https://www.geonet.org.nz/earthquake/story/3366146>
- GeoNet. (2011). M 6.2 Christchurch Tue, Feb 22 2011. <https://www.geonet.org.nz/earthquake/3468575>
- GeoNet. (2012). GeoNet strong-motion FTP site. <ftp://ftp.geonet.org.nz/strong/processed/>
- Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* (2nd ed.). O'Reilly Media.
- Global Earthquake Model (GEM). (2013). GitHub - GEM Direct Observation Tools for Android. <https://github.com/gem/DirectObservationToolsForAndroid>
- Global Earthquake Model (GEM). (2014). OpenQuake - Getting started. <https://www.globalquakemodel.org/oq-getting-started>
- Global Earthquake Model (GEM). (2015). South America Risk Assessment (SARA) Project. <https://sara.openquake.org/>
- Global Earthquake Model (GEM). (2018). OpenQuake Risk Modeller's Toolkit - User Guide. <https://docs.openquake.org/oq-irmt-qgis/v3.1.0/#>
- Gobierno del Distrito Federal Mexico. (2004). Gaceta oficial del Distrito Federal - Normas Técnicas Complementarias para Diseño por Sismo.
- Gomez-Bernal, A., & Saragoni, R. (2002). Respuesta Dinámica en suelos estratificados durante terremotos. *Proceedings of the VIII Jornadas Chilenas de Sismología e Ingeniería Antisísmica, CD Rom, paper No. 111, April 25-27, 2002.*
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. <http://www.deeplearningbook.org/>
- Greater Christchurch Group - Department of the Prime Minister and Cabinet. (2017). *Whole of Government Report: Lessons from the Canterbury Earthquake Sequence* (tech. rep.). Greater Christchurch Group, Department of the Prime Minister and Cabinet. Christchurch, New Zealand. <https://www.dPMC.govt.nz/sites/default/files/2017-07/whole-of-government-report-lessons-from-the-canterbury-earthquake-sequence.pdf>
- Grünthal, G. (1998). *European Macroseismic Scale 1998 (EMS-98)* (tech. rep.). European Seismological Commission (ESC). Luxembourg. <https://www.gfz-potsdam.de/>

- en / section / seismic-hazard-and-risk-dynamics / data-products-services / ems-98-european-macroseismic-scale/
- Gunay, M., & Mosalam, K. (2012). PEER performance based earthquake engineering methodology, revisited. *Proceeding of the 15 World Conference on Earthquake Engineering*. http://www.iitk.ac.in/nicee/wcee/article/WCEE2012_5606.pdf
- Günay, S., & Mosalam, K. M. (2013). PEER Performance-Based Earthquake Engineering Methodology, Revisited. *Journal of Earthquake Engineering*, 17(6), 829–858. <https://doi.org/10.1080/13632469.2013.787377>
- Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of machine learning research*, 1157–1182. <https://dl.acm.org/doi/10.5555/944919.944968>
- Haibo He, & Garcia, E. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284. <https://doi.org/10.1109/TKDE.2008.239>
- Hansson, S. O. (2018). Risk. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Fall 2018). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/fall2018/entries/risk/>
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.
- Honegger, M. (2018). *Shedding Light on Black Box Machine Learning Algorithms* (Doctoral dissertation). Karlsruhe Institute of Technology, Germany. <http://arxiv.org/abs/1808.05054>
- Hutchinson, T. C., & Ray Chaudhuri, S. (2006). Simplified expression for seismic fragility estimation of sliding-dominated equipment and contents. *Earthquake Spectra*, 22(3), 709–732. <https://doi.org/10.1193/1.2220637>
- Insurance Council of New Zealand (ICNZ). (2019). Canterbury Earthquakes. <https://www.icnz.org.nz/natural-disasters/canterbury-earthquakes/>
- Integrated Research on Disaster Risk. (2014). *Peril Classification and Hazard Glossary* (tech. rep.). Integrated Research on Disaster Risk (IRDR). Beijing, China. https://www.irdrinternational.org/knowledge_pool/publications/173

- Jaimes, M. A. (2017). Sismo del 19 de septiembre de 2017 M7.1, Puebla-Morelos (in Spanish). http://www.learningfromearthquakes.org/2017-09-19-puebla-mexico/images/2017_09_19_Puebla_Mexico/pdfs/UNAM_Presentation.pdf
- Jaiswal, K., & Wald, D. (2008). *Creating a Global Building Inventory for Earthquake Loss Assessment and Risk Management* (tech. rep.). U.S. Geological Survey (USGS). Reston, Virginia. <https://doi.org/10.3133/ofr20081160>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: With applications in R*. Springer.
- Jones, L. (2014). Lucy Jones Advises LA on Resiliency and 'The Big One'. *The Planning Report*, (November 2014). <https://www.planningreport.com/2014/11/07/lucy-jones-advises-la-resiliency-and-big-one>
- Jordan, C. J., Adlam, K., Laurie, K., Shelley, W., & Bevington, J. (2014). *User guide: Windows tool for field data collection and management* (tech. rep.). Global Earthquake Model (GEM). Pavia, Italy. <https://doi.org/10.13117/GEM.DATA-CAPTURE.TR2014.04>
- Juarez-Garcia, H., Gomez-Bernal, A., Arellano-Mendez, E., & Sordo Zabay, E. (2004). Seismic Vulnerability Assessment for Colonia Roma in Mexico City. *proceedings of the 13th World Conference on Earthquake Engineering*. https://www.iitk.ac.in/nicee/wcee/article/13_945.pdf
- Juran, J. M., & Godfrey, A. B. (1999). *Juran's Quality Handbook* (Fifth edit).
- Juran, J. M. (1951). *Quality Control Handbook* (First). McGraw-Hill Book Company.
- Juran, J. M., & De Feo, J. A. (2010). *Juran's Quality Handbook* (Sixth, Vol. 1). McGraw-Hill Education - Europe.
- Kaiser, A., Van Houtte, C., Perrin, N., Wotherspoon, L., & Mcverry, G. (2017). Site Characterisation of GeoNet Stations for the New Zealand Strong Motion Database. *Bulletin of the New Zealand Society for Earthquake Engineering*, 50(1), 39–49. <https://doi.org/10.5459/bnzsee.50.1.39-49>
- Kam, W. Y., Pampanin, S., & Elwood, K. (2011). Seismic performance of reinforced concrete buildings in the 22 February Christchurch (Lyttelton) earthquake. *Bulletin of the New Zealand Society for Earthquake Engineering*, 44(4), 239–278. <https://doi.org/10.5459/bnzsee.44.4.239-278>

- Kear, B. S., Gibbs, H. S., & Miller, R. B. (1967). *Soils of the Downs and Plains Canterbury and North Otago New Zealand* (tech. rep.). Soil Bureau. Wellington, New Zealand.
- Kiani, J., Camp, C., & Pezeshk, S. (2019). On the application of machine learning techniques to derive seismic fragility curves. *Computers & Structures*. <https://doi.org/10.1016/j.compstruc.2019.03.004>
- Kim, J. J., Elwood, K. J., Marquis, F., & Chang, S. E. (2017). Factors Influencing Post-Earthquake Decisions on Buildings in Christchurch, New Zealand. *Earthquake Spectra*, 33(2), 623–640. <https://doi.org/10.1193/072516EQS120M>
- King, A., Middleton, D., Brown, C., Johnston, D., & Johal, S. (2014). Insurance: Its Role in Recovery from the 2010–2011 Canterbury Earthquake Sequence. *Earthquake Spectra*, 30(1), 475–491. <https://doi.org/10.1193/022813EQS058M>
- Kiureghian, A. D. (2005). Non-ergodicity and PEER's framework formula. *Earthquake Engineering & Structural Dynamics*, 34(13), 1643–1652. <https://doi.org/10.1002/eqe.504>
- Kotu, V., & Deshpande, B. (2019). *Data Science* (Vol. 2). Morgan Kaufmann.
- Kovačević, M., Stojadinović, Z., Marinković, D., & Stojadinović, B. (2018). Sampling and machine learning methods for a rapid earthquake loss assessment system. *Proceeding of the 11th U.S. National Conference on Earthquake Engineering*.
- Krawinkler, H. (2005). *Van Nuys Hotel Building Testbed Report: Exercising Seismic Performance Assessment* (tech. rep.). Pacific Earthquake Engineering Research Center (PEER). Berkley, California. https://peer.berkeley.edu/sites/default/files/peer_511_krawinkler_testbed.pdf
- Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer New York. <https://doi.org/10.1007/978-1-4614-6849-3>
- Lallemant, D., Kiremidjian, A., & Burton, H. (2015). Statistical procedures for developing earthquake damage fragility curves. *Earthquake Engineering & Structural Dynamics*, 44(9), 1373–1389. <https://doi.org/10.1002/eqe.2522>
- Land Information New Zealand (LINZ). (2019). LINZ Data Services. <https://data.linz.govt.nz/>
- Land Information New Zealand (LINZ). (2020a). LINZ Data Service – NZ Property Titles. <https://data.linz.govt.nz/layer/50804-nz-property-titles/>

- Land Information New Zealand (LINZ). (2020b). LINZ Data Service – NZ Street Address. <https://data.linz.govt.nz/layer/53353-nz-street-address/>
- Land Resource Information Systems (LRIS). (2010). Soil map for the Upper Plains and Downs of Canterbury. <https://lris.scinfo.org.nz/layer/48157-soil-map-for-the-upper-plains-and-downs-of-canterbury/>
- Land Resource Information Systems (LRIS). (2014). LRIS Portal. <https://lris.scinfo.org.nz/>
- Lee, K.-F. (2018). *AI Superpowers: China, Silicon Valley, and the New World Order*. Houghton Mifflin Harcourt.
- Lemaitre, G., Nogueira, F., & Aridas, C. K. (2017a). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research*, 18(1), 559–563.
- Lemaitre, G., Nogueira, F., & Aridas, C. K. (2017b). Welcome to imbalanced-learn documentation! <https://imbalanced-learn.org/stable/>
- Lundberg, S. M. (2020). SHAP (SHapley Additive exPlanations) repository. <https://github.com/slundberg/shap>
- Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2018). Consistent Individualized Feature Attribution for Tree Ensembles. *2017 ICML Workshop*. <http://arxiv.org/abs/1802.03888>
- Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In I. G. Garnett, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (pp. 4765–4774). Curran Associates, Inc.
- Ma, T., Avati, A., Katanforoosh, K., & Ng, A. (2020). Deep Learning [CS229 Lecture Notes - Spring 2020]. <http://cs229.stanford.edu/syllabus.html>
- Maio, R., & Tsionis, G. (2015). *Seismic fragility curves for the European building stock: review and evaluation of analytical fragility curves* (tech. rep.). European Commission. Ispra, Italy. <https://doi.org/10.2788/586263>
- Mallet, R. (1862). *Great Neapolitan Earthquake of 1857 - The first principles of observational seismology* (Chapman and Hall London, Ed.). Royal Society of London.

- Mangalathu, S., & Burton, H. V. (2019). Deep learning-based classification of earthquake-impacted buildings using textual damage descriptions. *International Journal of Disaster Risk Reduction*, 36(February). <https://doi.org/10.1016/j.ijdr.2019.101111>
- Mangalathu, S., Hwang, S. H., Choi, E., & Jeon, J. S. (2019). Rapid seismic damage evaluation of bridge portfolios using machine learning techniques. *Engineering Structures*, 201(October), 109785. <https://doi.org/10.1016/j.engstruct.2019.109785>
- Mangalathu, S., & Jeon, J.-S. (2019). Machine Learning–Based Failure Mode Recognition of Circular Reinforced Concrete Bridge Columns: Comparative Study. *Journal of Structural Engineering*, 145(10), 04019104. [https://doi.org/10.1061/\(ASCE\)ST.1943-541X.0002402](https://doi.org/10.1061/(ASCE)ST.1943-541X.0002402)
- Mangalathu, S., Sun, H., Nweke, C. C., Yi, Z., & Burton, H. V. (2020). Classifying earthquake damage to buildings using machine learning. *Earthquake Spectra*, 36(1), 183–208. <https://doi.org/10.1177/8755293019878137>
- Mayoral, J. M., Asimaki, D., Tepalcapa, S., Wood, C., Roman-de la Sancha, A., Hutchinson, T., Franke, K., & Montalva, G. (2019). Site effects in Mexico City basin: Past and present. *Soil Dynamics and Earthquake Engineering*, 121(March), 369–382. <https://doi.org/10.1016/j.soildyn.2019.02.028>
- Mayoral, J. M., Hutchinson, T. C., & Franke, K. W. (2017). *Geotechnical Engineering Reconnaissance of the 19 September 2017 Mw 7.1 Puebla-Mexico City Earthquake: Version 2.0* (tech. rep. GEER-055). Geotechnical Extreme Events Reconnaissance Association (GEER). Berkeley, California. <https://doi.org/10.18118/G6JD46>
- Mayoral, J. M., Roman, A., De La Rosa, D., Alcaraz, M., & Rivas, R. (2019). Key findings and observations following the September 19th, 2017 Mw 7.1 Puebla-Mexico City earthquake. In F. Silvestri & N. Moraci (Eds.), *Earthquake geotechnical engineering for protection and development of environment and constructions* (pp. 833–844). CRC Press. <https://doi.org/10.1201/9780429031274>
- Middleton, D. A. (2002). EQC's use of computer modelling in a catastrophe response. *Proceedings of the 2002 Annual Conference of the New Zealand Society for Earthquake Engineering*. <https://www.nzsee.org.nz/db/2002/Paper31.PDF>

- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38. <https://doi.org/10.1016/j.artint.2018.07.007>
- Ministry of Business Innovation & Employment (MBIE). (2012). *Repairing and rebuilding houses affected by the Canterbury earthquakes* (tech. rep.). Ministry of Business Innovation & Employment (MBIE). Wellington, New Zealand. <https://www.building.govt.nz/building-code-compliance/canterbury-rebuild/repairing-and-rebuilding-houses-affected-by-the-canterbury-earthquakes/>
- Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.
- Mitrani-Reiser, J. (2007). *An ounce of prevention: probabilistic loss estimation for performance-based earthquake engineering* (Doctoral dissertation). California Institute of Technology. <http://thesis.library.caltech.edu/2207/>
- Moehle, J., & Deierlein, G. (2004). A framework methodology for performance-based earthquake engineering. *Proceeding of the 13th World Conference on Earthquake Engineering*.
- Molnar, C. (2020). *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable* (28 Dec 2020). <https://christophm.github.io/interpretable-ml-book/>
- Monfort, D., Lantada, N., Goula, X., Barbat, A. H., Negulescu, C., Pujades, L., Susagna, T., Valcarcel, J., & Rodríguez, J. (2011). Generación de escenarios de riesgo sísmico en una zona transfronteriza del Pirineo. *Congreso Nacional de Ingeniería Sísmica*, (May), 1–9. <http://hdl.handle.net/2117/14832>
- Munich RE. (2019). NatCatSERVICE. <http://natcatservice.munichre.com/>
- Muria-Vila, D., & Gonzalez-Alcorta, R. (1995). Propiedades dinamicas de edificios de la ciudad de Mexico. *Revista de Ingenieria Sismica*, 51, 25–45.
- New Zealand Parliament. (2019). Earthquake Commission Amendment Act 2019. https://www.parliament.nz/en/pb/bills-and-laws/bills-proposed-laws/document/BILL_77657/earthquake-commission-amendment-bill
- New Zealand Society for Earthquake Engineering (NZSEE). (2009). *Guidelines for Building Safety Evaluation During a State of Emergency* (tech. rep.). New Zealand Society for Earthquake Engineering (NZSEE). Wellington, New Zealand. <http://www.nzsee.org.nz/Guidelines/BuildingSafetyEvaluationAug09.pdf>

- NIWA, & GNS Science. (2015). RiskScape - Asset Module Metadata. https://wiki.riskscape.org.nz/images/7/75/New_Zealand_Building_Inventory.pdf
- NIWA, & GNS Science. (2017). RiskScape. <https://www.riskscape.org.nz/>
- O'Rourke, T. D., Jeon, S. S., Toprak, S., Cubrinovski, M., Hughes, M., van Ballegooy, S., & Bouziou, D. (2014). Earthquake response of underground pipeline networks in Christchurch, NZ. *Earthquake Spectra*, 30(1), 183–204. <https://doi.org/10.1193/030413EQS062M>
- Oxford English Dictionary (OED) Online. (2010). Definition of risk. <https://www.oed.com/view/Entry/166306>
- Pagani, M., Monelli, D., Weatherill, G., Danciu, L., Crowley, H., Silva, V., Henshaw, P., Butler, L., Nastasi, M., Panzeri, L., Simionato, M., & Vigano, D. (2014). OpenQuake Engine: An Open Hazard (and Risk) Software for the Global Earthquake Model. *Seismological Research Letters*, 85(3), 692–702. <https://doi.org/10.1785/0220130087>
- Pagani, M., Silva, V., Rao, A., Simionato, M., & Gee, R. (2019). The OpenQuake-engine User Manual. *Global Earthquake Model (GEM) Open-Quake Manual for Engine version 3.7.1.*, 183 pages. <https://docs.openquake.org/manuals/OpenQuake%20Manual%203.7.pdf>
- Pareto, V. (1906). *Manuale di economia politica con una introduzione alla scienza sociale.*
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2019). Choosing the right estimator. https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html
- Poland, C., Hill, J., Sharpe, R., Soulages, J., & Structural Engineers Association of California (SEAOC). Office of Emergency Services. (1995). *Vision 2000: Performance Based Seismic Engineering of Buildings* (tech. rep.). Structural Engineers Association of California (SEAOC). Sacramento, CA.
- Porter, K. (2003). An overview of PEER's performance-based earthquake engineering methodology. *Ninth International Conference on Applications of Statistics and Probability in Civil Engineering (ICASP9)*, 973–980. <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.538.4550&rep=rep1&type=pdf>

- Porter, K. (2005). *A Taxonomy of Building Components for Performance-Based Earthquake Engineering* (tech. rep. September). Pacific Earthquake Engineering Research Center (PEER). Berkeley, California. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.90.1198&rep=rep1&type=pdf>
- Porter, K. (2020). A Beginner's Guide to Fragility, Vulnerability, and Risk. *University of Colorado Boulder*, (16 January 2021), 139 pp. <http://www.sparisk.com/pubs/Porter-beginners-guide.pdf>
- Porter, K., Kennedy, R., & Bachman, R. (2007). Creating Fragility Functions for Performance-Based Earthquake Engineering. *Earthquake Spectra*, 23(2), 471–489. <https://doi.org/10.1193/1.2720892>
- Porter, K., & Kiremidjian, A. (2001). *Assembly-Based Vulnerability of Buildings and Its Uses in Seismic Performance Evaluation and Risk Management Decision-Making* (Doctoral dissertation). Stanford University. John A Blume Earthquake Engineering Center. <http://purl.stanford.edu/qf102hx9901>
- Potter, S. H., Becker, J. S., Johnston, D. M., & Rossiter, K. P. (2015). An overview of the impacts of the 2010-2011 Canterbury earthquakes. *International Journal of Disaster Risk Reduction*, 14, 6–14. <https://doi.org/10.1016/j.ijdrr.2015.01.014>
- Prati, R. C., Batista, G. E., & Monard, M. C. (2009). Data mining with unbalanced class distributions: Concepts and methods. *Proceedings of the 4th Indian International Conference on Artificial Intelligence, IICAI 2009*.
- Raschka, S., & Mirjalili, V. (2019). *Python Machine Learning* (Third edit). Packt Publishing.
- Reyners, M., Eberhart-Phillips, D., & Martin, S. (2014). Prolonged Canterbury earthquake sequence linked to widespread weakening of strong crust. *Nature Geoscience*, 7(1), 34–37. <https://doi.org/10.1038/ngeo2013>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016a). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. <http://arxiv.org/abs/1602.04938>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016b). Model-Agnostic Interpretability of Machine Learning. *2016 ICML Workshop on Human Interpretability in Machine Learning*. <http://arxiv.org/abs/1606.05386>

- Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-Precision Model-Agnostic Explanations. *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI'18)*, 1527–1535.
- Roeslin, S., Juárez-García, H., Elwood, K., Dhakal, R., & Gómez-Bernal, A. (2020). The September 19th, 2017 Puebla, Mexico earthquake. *Bulletin of the New Zealand Society for Earthquake Engineering*, 53(3), 150–172. <https://doi.org/10.5459/bnzsee.53.3.150-172>
- Roeslin, S., Ma, Q. T. M., & García, H. J. (2018). Damage Assessment on Buildings Following the 19th September 2017 Puebla, Mexico Earthquake. *Frontiers in Built Environment*, 4, 18. <https://doi.org/10.3389/fbuil.2018.00072>
- Rogers, N., van Ballegooy, S., Williams, K., & Johnson, L. (2015). Considering Post-Disaster Damage to Residential Building Construction - Is Our Modern Building Construction Resilient? *Proceedings of 6th International Conference on Earthquake Geotechnical Engineering*.
- Rojahn, C., Heintz, J. A., Hortacsu, A., & McLane, T. (2015). *FEMA P-154 Rapid Visual Screening of Buildings for Potential Seismic Hazards: A Handbook* (tech. rep.). Applied Technology Council. Redwood City, California. https://www.fema.gov/sites/default/files/2020-07/fema_earthquakes_rapid-visual-screening-of-buildings-for-potential-seismic-hazards-a-handbook-third-edition-fema-p-154.pdf
- Rojahn, C., Sharpe, R. L., Scholl, R. E., Kiremidjian, A. S., & Nutt, R. V. (1985). *ATC-13 Earthquake Damage Evaluation Data for California* (tech. rep.). Applied Technology Council (ATC), Federal Emergency Management Agency (FEMA). Redwood City, California.
- Rosser, J., Morley, J. G., & Vicini, A. (2014). User guide: Android mobile tool for field data collection. *GEM Technical Report*, 26. <https://doi.org/10.13117/GEM.DATA-CAPTURE.TR2014.03>
- Russell, J., & van Ballegooy, S. (2015). *Canterbury Earthquake Sequence: Increased Liquefaction Vulnerability assessment methodology* (tech. rep.). Tonkin & Taylor Ltd. Auckland, New Zealand. <https://www.eqc.govt.nz/ILV-engineering-assessment-methodology>
- Russell, S., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach* (Fourth). Pearson.

- Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, 3(3), 210–229. <https://doi.org/10.1147/rd.33.0210>
- Servicio Sismológico Nacional (SSN). (2017). *Reporte Especial - Sismo del día 19 de Septiembre de 2017, Puebla-Morelos (M 7.1) (in Spanish)* (tech. rep.). Servicio Sismológico Nacional UNAM. Mexico City, Mexico. http://www.ssn.unam.mx/sismicidad/reportes-especiales/2017/SSNMX_rep_esp_20170919_Puebla-Morelos_M71.pdf
- Shephard, R. B., Spurr, D. D., & Walker, G. R. (2002). The Earthquake Commission's earthquake insurance loss model. *Proceedings, 2002 Annual Conference of the New Zealand Society for Earthquake Engineering*. <https://www.nzsee.org.nz/db/2002/Paper32.PDF>
- Silva, V. (2019). Uncertainty and correlation in seismic vulnerability functions of building classes. *Earthquake Spectra*, 35(4), 1515–1539. <https://doi.org/10.1193/013018eqs031m>
- Silva, V., Crowley, H., Jaiswal, K., Acevedo, A. B., Pittore, M., & Journey, M. (2018). Developing a Global Earthquake Risk Model. *Proceedings of the 16th European Conference on Earthquake Engineering*, (July).
- Silva, V., Pagani, M., Schneider, J., & Henshaw, P. (2019). Assessing Seismic Hazard and Risk Globally for an Earthquake Resilient World. *Contributing Paper to GAR 2019*, 24 p. <https://www.unisdr.org/we/inform/publications/65866>
- Stirling, M. W., Gerstenberger, M., Goded, T., & Ries, W. (2015). *Macroseismic Intensity Assessment for the M6.2 2011 Christchurch Earthquake* (tech. rep. GNS Science Consultancy Report 2015/26). GNS Science. Lower Hutt, New Zealand.
- Sun, H., Burton, H. V., & Huang, H. (2020). Machine Learning Applications for Building Structural Design and Performance Assessment: State-of-the-Art Review. *Journal of Building Engineering*. <https://doi.org/10.1016/j.jobe.2020.101816>
- Taghavi, S., & Miranda, E. (2003). *Response Assessment of Nonstructural Building Elements* (tech. rep.). Pacific Earthquake Engineering Research Center (PEER). Berkeley, California, Pacific Earthquake Engineering Research Center. https://peer.berkeley.edu/sites/default/files/0305_s._taghavi_e._miranda_.pdf

- Tang, J., Alelyani, S., & Liu, H. (2014). Feature selection for classification: A review. *Data classification: Algorithms and applications* (pp. 37–64). CRC Press.
- The U.S. National Archives and Records Administration. (2017). San Francisco Earthquake, 1906. <https://www.archives.gov/legislative/features/sf>
- Tveite, H. (2019). QGIS NNJoin plugin. <https://github.com/havatv/qgisnnjoinplugin>
- UCLouvain, & Guha-Sapir, D. (2020). EM-DAT: The Emergency Events Database. <https://www.emdat.be/>
- United Nations. (2016). *Report of the open-ended intergovernmental expert working group on indicators and terminology relating to disaster risk reduction* (tech. rep. December). United Nations General Assembly (UNGA). New York. <https://digitallibrary.un.org/record/852089?ln=en>
- United Nations Office for Disaster Risk Reduction (UNDRR). (2015). Hazard. <https://www.undrr.org/terminology/hazard>
- United Nations Office for Disaster Risk Reduction (UNDRR). (2019). *Global Assessment Report on Disaster Risk Reduction (GAR19)* (tech. rep.). United Nations Office for Disaster Risk Reduction (UNDRR). Geneva, Switzerland. <https://gar.undrr.org/>
- United States Geological Survey (USGS). (1999). What is the "Ring of Fire"? https://www.usgs.gov/faqs/what-ring-fire?qt-news_science_products=0#qt-news_science_products
- United States Geological Survey (USGS). (2017). September 19, 2017 - Magnitude 7.1 Earthquake in Mexico. <https://www.usgs.gov/news/magnitude-71-earthquake-mexico>
- Van Houtte, C., Bannister, S., Holden, C., Bourguignon, S., & Mcverry, G. (2017). The New Zealand strong motion database. *Bulletin of the New Zealand Society for Earthquake Engineering*, 50(1), 1–20. <https://doi.org/10.5459/bnzsee.50.1.1-20>
- Villar-Vega, M., & Silva, V. (2017). Assessment of earthquake damage considering the characteristics of past events in South America. *Soil Dynamics and Earthquake Engineering*, 99, 86–96. <https://doi.org/10.1016/j.soildyn.2017.05.004>
- Weiser, D., Hunt, J., Jampole, E., & Gobbato, M. (2017). *EERI Earthquake Reconnaissance Team Report: M7 A product of the EERI Learning From Earthquakes Program* (tech. rep. February). Earthquake Engineering Research Institute (EERI). Oakland,

- California. http://www.learningfromearthquakes.org/2017-09-19-puebla-mexico/images/2017_09_19_Puebla_Mexico/pdfs/MexicoCity_EERI_LFEReport_M7.1PueblaMexicoEarthquake.pdf
- Wieland, M., Pittore, M., Parolai, S., Begaliev, U., Yasunov, P., Tyagunov, S., Moldobekov, B., Saidiy, S., Ilyasov, I., & Abakanov, T. (2015). A Multiscale Exposure Model for Seismic Risk Assessment in Central Asia. *Seismological Research Letters*, 86(1), 210–222. <https://doi.org/10.1785/0220140130>
- Xie, Y., Ebad Sichani, M., Padgett, J. E., & DesRoches, R. (2020). The promise of implementing machine learning in earthquake engineering: A state-of-the-art review. *Earthquake Spectra*, 33. <https://doi.org/10.1177/8755293020919419>
- Yang, T. Y. (2013). Assessing seismic risks for new and existing buildings using performance-based earthquake engineering (PBEE) methodology. *Handbook of seismic risk analysis and management of civil infrastructure systems* (pp. 307–333). Woodhead Publishing. <https://doi.org/http://dx.doi.org/10.1533/9780857098986.3.307>
- Yang, T. Y., Moehle, J., Stojadinovic, B., & Der Kiureghian, A. (2009). Seismic Performance Evaluation of Facilities: Methodology and Implementation. *Journal of Structural Engineering*, 135(10), 1146–1154. [https://doi.org/10.1061/\(ASCE\)0733-9445\(2009\)135:10\(1146\)](https://doi.org/10.1061/(ASCE)0733-9445(2009)135:10(1146))
- Zhang, Y., Burton, H. V., Sun, H., & Shokrabadi, M. (2018). A machine learning framework for assessing post-earthquake structural safety. *Structural Safety*, 72, 1–16. <https://doi.org/10.1016/j.strusafe.2017.12.001>