Rick Willems

Developing a diagnostic assessment tool to evaluate damage in buildings

Rick Willems
Student number 4195639
Master Student Civil Engineering TU Delft
Track Building Engineering
Specialisation Structural Design
R.P.H.Willems@student.tudelft.nl

Photo on cover by author
'Damage' caused by overloading due to accidental loads by lightning
The Hollywood Tower Hotel
Walt Disney Studios

Developing a diagnostic assessment tool to evaluate damage in buildings

By

Rick Willems

in partial fulfilment of the requirements for the degree of

Master of Science
in Civil Engineering

at the Delft University of Technology,
to be defended publicly on 20th of May 2022.

Thesis committee:	Dr. ir. K.C. Terwel	TU Delft, Building Engineering
	Prof. dr. Ir. J.G. Rots	TU Delft, Structural Mechanics
	Dr. D.M.J. Tax	TU Delft, Mathematics and Computer Science
	Ir. P.A. Korswagen	TU Delft, Structural Mechanics

Preface

Dear reader, this is my master thesis report. I am Rick Willems, a master student in civil engineering following the track building engineering with specialisation structural design at the TU Delft. When looking for a subject for this thesis, I wanted to find it in the forensic civil engineering field. It is a niche field in civil engineering, where one needs a broad knowledge to solve the puzzle of what was the cause of damage. Meanwhile, you can really help people in need.

The topic of this thesis offered me an opportunity to read many damage reports and thus I gained an inside on how forensic engineering is applied in practice. Also, I read discussions between forensic experts and I saw the problems that arise during assessment. I hope this gives me a head start in my career as forensic civil engineer.

Another one of my interests is making computer programs and tools. This thesis gives me an opportunity to execute my hobby in creating program scrips. However, computer programming is not a part of my education. So I had to catch up on a lot of pattern recognition topics. Nevertheless, it was fun to explore this, for me, new field of science.

Both subjects, forensic civil engineering and computer programming, are combined in this thesis. With the help of a large number of damage reports and pattern recognition technics, I was able to generate relations between damage factors and damage causes. Unfortunately, not all relations were reliable or acceptable with respect to with literature. Therefore, in practice the found relations are very limited in their use.

One of the main findings is that, besides the *earthquake load*, *the age of a building* and *trees* has a significant influence on the occurrence of earthquake damage. Another finding is that hindered deformation mostly occurred on the inside of a building. More findings about relations can be found in the main report. There it is explained that regression analyses were difficult to execute. Also, unequal settlement causes was hard to analyse.

I want to thank the committee for their support. Also thanks to Irma Laponder for encouraging me in my English writing and Johan Schmal for patiently correcting my grammar mistakes in most of my texts. At last but not least, everyone that kept me company during the lonely Covid-19 lockdowns. As an example I would like to mention Paul Sprangers and Tim Hinssen. Their podcast, Kleine Boodschap about everything related with the Efteling, kept my mind from wandering off during the weeks of tedious work when 1850+ reports of damage cases were converted into a database.

Rick Willems
Leidschendam, April 2022

Disclaimer	**EN:** This report is a master thesis project for the master Building Engineering, of the Faculty Civil Engineering and Geosciences at TU Delft. This research primarily had an educational purpose. Therefore, one should be reluctant to draw general conclusions from this report. **NL:** Dit rapport is een master afstudeerproject in het kader van afronding van de master Building Engineering van de faculteit Civiele Techniek en Geowetenschappen van TU Delft. Het onderzoek heeft voornamelijk een onderwijskundig doel. Men dient derhalve terughoudend te zijn met het trekken van algemene conclusies op basis van deze studie.

Table of contents

Abstract

Mining activities at the Groningen gas field are causing earthquakes which result in building damage. This has started a discussion on what type of damage is caused by earthquakes. These discussions are typical in the forensic engineering field, especially in complex cases. The problem is the lack of regulations, in terms of standardizations and uniformization.

To solve that problem, experts can be provided with an independent tool which can contribute to the investigation of the cause of building damage. The tool can help to indicate potential damage causes. This will support the findings of experts. Also, it can draw attention to overlooked damage causes.

The tool is based on relations found in a database of damage cases that have been determined earlier. The database consists of damage cases in the Groningen province. Not all available damage cases were incorporated in this thesis, because processing the damage reports to a database was a labour-intensive job. The analysed dataset consists of 1830 damage cases in 49 buildings. Experts were able to determine the cause of damage in 1180 of these damage cases, which results in a ratio of 64.4% known cases. Only the known cases where applied in the analysis. The buildings were located in seven different areas in the province of Groningen.

Each analysed damage case consists of a damage cause and a description. A description has been structured in 191 characteristics. These characteristics have been categorised into three types: building characteristics, context characteristics and damage characteristics. Building characteristics say something about the function, materials and size of the building. Context characteristics explain the sub soil, vibration sources and external forces in the surrounding of the building. Damage characteristics describe how damage is presented in terms of position, location and shape of damage.

Whether the found relations can be deployed in practice, depends on how useful those relations are. Useful is defined as reliable and meaningful. Reliable is how a found pattern performs according to a test, mostly measured in terms of accuracy or coefficient of determination. Meaningful is whether the found relations are logical to be explained by literature or plausible damage situations. The pattern recognition can introduce some relations and can provide them with a reliability value. However, if the relations are not explainable, they do not mean anything for use in practice.

The relations in these data were found by deploying pattern recognition methods. Two algorithms were utilized as a pattern recognition method: decision tree and linear regression. A decision tree algorithm splits the data into groups by applying thresholds on case characteristics. These thresholds can be made visual in a decision tree figure. Linear regression tries to obtain a target value by means of a linear relation of characteristics. Therefore, the linear regression algorithm determines the slope value of each characteristic.

Classification analyses were done with decision trees on six damage cause categories. The results of that type of analysis were capable of determining if or which damage was caused by a certain cause. Linear regression was performed in order to find regression relations where the technical attributability of a damage cause could be calculated for each case. In the more complex task of regression analysis, only three damage cause categories were suitable for finding a relation.

To determine whether damage was caused by earthquakes, earthquake load in terms of PGV is an important characteristic. Also, the age of a building and trees has a possible significant influence on the occurrence of earthquake damage, according to the found pattern. A relation between those last two characteristics and earthquake damage is not described in literature. Besides that, this decision tree pattern seems to be the most useful pattern for in practice.

Another interesting finding is that hindered deformation mostly occurred at the inside of a building. Combined with other characteristics, a pattern on this damage cause performed with the highest score in this thesis. It has an accuracy of 77%. This means that 77% of the cases in the test set were correctly predicted by the produced classification decision tree. However, the found relation with the characteristics is not always explainable or meaningful so as to be applied in practice. More conclusions of classification analysis are shown in Table 1.

Summary of classification conclusions

Damage cause category	Reliability	Meaningful	Conclusion (useful)
A.1.1 Initial insufficient resistance to bear loads	51% accuracy	Damage located above opening **Not:** Distance to road	**No**, Not explainable context characteristics.
A.3.5 Overloading due to vibrations by earthquakes	61% accuracy	PGV load **Not:** relative influence tree	**Cautious use,** Good performance of PGV split
B.1.1 Initial hindered deformations	77% accuracy	Damage located inside **Not:** age of building	**No,** better meaningful relations possible.
B.2.1 Initial imposed deformations	69% accuracy	Concrete foundation material **Not:** longest side of building	**No,** illogical characteristics
C.1.1 Unequal settlements with equal loads	43% accuracy	Presence of trees and roots growth	**No,** not reliable.
C.X Unequal settlements	34% accuracy	Construction year **Not:** possible accumulation of rain or snow	**No,** not reliable.

Table 1 Summary of classification conclusions. First column is the damage cause on which pattern recognition is trained. Second column describes how reliable the found pattern is in terms of accuracy. The third column mentions which characteristics in the pattern are logical to explain and which are not. If the pattern produced useful relations is explained in the last column.

The presented findings above are classification relations. Regression analyses were difficult to execute. A desired positive coefficient of determination (R^2) could not be reached without subjective interference in the pattern recognition. The best regression result was obtained on damage caused by earthquakes. It had a R^2 of 0.48. Which means that 48% of the data was describable in a linear relation. More conclusions of regression analysis are shown in Table 2.

Summary of regression conclusions

Damage cause category	Reliability	Meaningful	Conclusion (useful)
A.3.5 Overloading due to vibrations by earthquakes	0.48 Coefficient of determination	PGV load **Not:** relative influence	**Cautious use,** Good performance of PGV
B.X Deformations	0.28 Coefficient of determination	Horizontal crack **Not:** building renovations	**No,** Low reliability
C.X Unequal settlements	0.18 Coefficient of determination	PGV load **Not:** Settlement differences over the length of the building	**No,** Low reliability

Table 2 Summary of regression conclusions. First column is the damage cause on which pattern recognition is trained. Second column describes how reliable the found pattern is in terms of accuracy. The third column mentions which characteristics in the pattern are logical to explain and which are not. If the pattern produced useful relations is explained in the last column.

It has been interesting to study the relation between characteristics and damage causes. However, the results are not of decisive importance. The building and context characteristics supported by literature were not always selected or applied properly by the pattern recognition. Also, the potential of damage characteristics was not recognized by the algorithms. Nonetheless, the results of earthquake related damage seem promising. They even indicate characteristics which may be worth investigating more closely.

1 Introduction

This chapter starts with the background of this thesis. After that, the motive and relevance of this project is discussed, followed by the goal, research questions and the methodology. It ends with the definition of key terms.

1.1 Background

The Groningen gas field was discovered in 1959. A few years later in 1963, the first gas was extracted from the Groningen soil (NLOG, 2016). These mining activities are causing earthquakes (Evers, 2019). Since 2003, the residents in the north of the Netherlands felt an increase of the earthquakes in terms of numbers and severeness (GBB, 2021). The people living in the north of the Netherlands worried about the consequences of the trembling. Deltares predicted that buildings can experience light damage and a few might show moderate damage due to vibrations of earthquakes (de Lange, Oostrom, Dortland, Borsje, & Richemont, 2011). According to law, mining activities that may cause damage to buildings or infrastructure are not permitted (Rijksoverheid, 2002, p. Artikel 13 lid d). If damage still occurs that is related to mining activities, the damage has to be compensated by the mining company. This started a discussion on what damage was caused by earthquakes.

One of the discussion points was if damage caused by earthquakes could be limited to a contour in where it might be caused by earthquakes [schadeafhandelingsgebied] (Verbeek, 2017). Consequently, the NAM (Nederlandse Aardolie Maatschappij) stated that damage outside of this contour could not be caused by earthquakes. Research on this discussion has been done by different organisations and these researches were commissioned by different bodies. Arcadis was the first research company commissioned by the mining company NAM. They took a sample of damage cases in the area around Steenbergen to check if they were caused by mining (Kok, de Jong, & Kastelein, 2015). Later, Witteveen+Bos did a research with a larger sample group of damage cases (Salet & Bruurs, 2017). The NCG (national coordinator Groningen) urgently requested the NAM to commissioned this research. Finally**,** the NCG asked TU Delft to do a research to clarify causes of damage in the province of Groningen (van Staalduinen, Terwel, & Rots, 2018, p. 5). They all were searching for relations between the mining activities and characteristics of damage.

1.2 Motivation and relevance

To investigate the relations between damage and causes, it is interesting to known how experts judge building damage. This also gives an answer on how discussions arise about the cause of damage in general.

When investigating the damage in general, experts first exclude impossible causes and compose hypotheses of plausible causes (falsification). Next, they test the hypotheses to determine the cause of the damage (verification) (Borsje & de Richemont, 2011, p. 4). This investigation is mostly influenced by their experience and theoretical knowledge (De Vent, 2011, p. 34). As experts do not all have the same level of experience and expertise, they may arrive at different damage causes when evaluating the same case (Braakman, 2021). Also the theoretical knowledge is not always identical to the unique appearance of each damage. Because of that, especially in complex cases, there is often a discussion between different parties about the cause of damage (Van Maanen Winters, Vuggen, & Quaedvlieg, 2007).

Discussions between experts can be reduced by standardization and uniformization of damage investigations. These regulations could be composed with a large number of known damage cases. However, the occurrence of similar damage cases is very exceptional. Historical damage cases can perhaps be traced back to relations. These patterns can experts administer during their falsification and hopefully also include in their verification phase. This can reduce the discussion about the cause of damage.

1.3 Goal

The intended result of this research is to produce a tool that could determine cause of damage based on a collection of information that is relevant for forensic civil engineering experts. This collection of relevant information is called *(case) characteristics* in this thesis. The output of such tool depends on what data is provided for analysis. The TU Delft provided the data. This data consist of characteristics and technical attributability per cause for every investigated damage case (van Staalduinen, Terwel, & Rots, 2018). Based on the provided data, the following output could be produced by the tool: What cause was accountable for the damage and how much is a cause technical accountable for the damage?

Relations between characteristics and damage were needed to make the intended tool. This relations were discovered with pattern recognition algorithms. Therefore, the main question in this thesis will be:

What useful relations are possible between case characteristics and damage causes?

Where useful is defined in terms of meaningful and reliability.

1.4 Research questions

Which characteristics are important when investigating damage?
It is important to know which case characteristics are used when investigating damage in general. For instance, the company name of the internet provider is not relevant when examining damage. However, the type of foundation does have relevance for damage research.

Why are these characteristics important in determining the cause of damage?
A reliable way of evaluating the result of the algorithms is to check if the outcome is expected. For example, earthquake damage must be the result of an earthquake load.

Which case characteristics are available for analysis?
Due to various reasons, not all characteristics were available for this research. Some of them were incomplete, whereas others were missing completely in the provided dataset.

What to do with the characteristics that are not available for analysis?
Handling the missing data is important for the reliability of the thesis outcome. Missing data could restrict the intended tool, because the missing data is relevant in determining the cause of damage. Therefore, an analysis is done of the question what to do with characteristics that are missing or are incomplete.

What adjustments must be made to the characteristics in order to make them available for the algorithms?
The provided data must be prepared for processing because they have mostly been delivered alphabetically whereas algorithms only allow for numerical input. For instance, for the answer to the question if a basement is present in a building, the options are ‘yes’ or ‘no’. This alphabetical form must be transformed into a numerical form of ‘1’ for yes, or ‘-1’ for no.

Which causes may be revealed when investigating damage?
Damage can occur because of various causes. Several researchers have tried to categorize the causes (De Vent, 2011, p. 26). In this thesis, a choice has been made a to the type of categorization to be

applied, based on the provided dataset. It is also explained how this categorization has been developed.

Which of these causes are available for analysis?
Not all causes were included in the provided data set. This tends to restrict the application of the tool. For instance, the tool cannot evaluate damage that is related to severe weather conditions if this cause is not available for analysis. Consequently, the missing causes cannot be indicated with a tool.

How do pattern recognition algorithms arrive at relations between characteristics and damage?
A thorough understanding of the working of pattern recognition algorithms is important in order to execute them correctly. Also, it provides awareness of the variables of an algorithm method. This variables can lead to more reliable results.

How to evaluate the reliability of the found relations?
Any algorithm can come up with relations, but the reliability of the relations is not always sufficient. There are different ways of evaluating the relations. They will be discussed later in the report.

How to evaluate the relations for their meaningfulness?
Algorithms can indicate unexpected characteristics as an important factor for a cause of damage. These characteristics are not always relevant for the specific cause. For example, the wind direction of the façade is not a logical characteristic for a cause of overloading. This will result in unmeaningful relations. Therefore, it is important to determine what an unmeaningful relation is as well as what to do with it.

1.5 Methodology

This thesis consists of two phases (Figure 1). The data used to find the relations are collected and discussed in the first phase. The relation finding algorithms were explained and applied in the second phase.

1.5.1 Phase one

Pattern recognition algorithms were used to find relations between case characteristics and causes of damage. These algorithms need instances of damage cases that have already been determined. The cases implemented in this thesis were come from the NCG (National Coördinator Groningen). They commissioned TU Delft to do research in the north of The Netherlands (van Staalduinen, Terwel, & Rots, 2018). The characteristics covered by TU Delft were discussed in the first half of chapter 2 in this thesis. Furthermore, an explanation of the methods applied by TU Delft to collected, stored and used characteristics was given in that chapter. However, the notation of these characteristics was not always been practicable or present for relation finding algorithms. Therefore, these characteristics have been modified before continuing to the second phase. These modifications were also explained in the first half of chapter 2. This method of discussing the characteristics was been repeated for the review of the causes of damage in the second half of chapter 2.

1.5.2 Phase two

The deployed relation finding algorithms were explained in chapter 3. On the advice of the committee the decision tree analysis and the linear regression have been used as relation finding algorithms. Chapter 3 explains how to calculate and evaluate the accuracy of the applied algorithms. These calculations have been made in chapters 4 (Results of classification analysis – Decision tree), 5 (Results of regression analysis – Decision tree) and 6 (Results of regression analysis – Linear Regression). The concept of the design cycle has been carried out to improve the calculations step by step. As a result these chapters consist of multiple calculation runs. The final results are discussed in chapter 7 and the conclusions are presented in chapter 8.

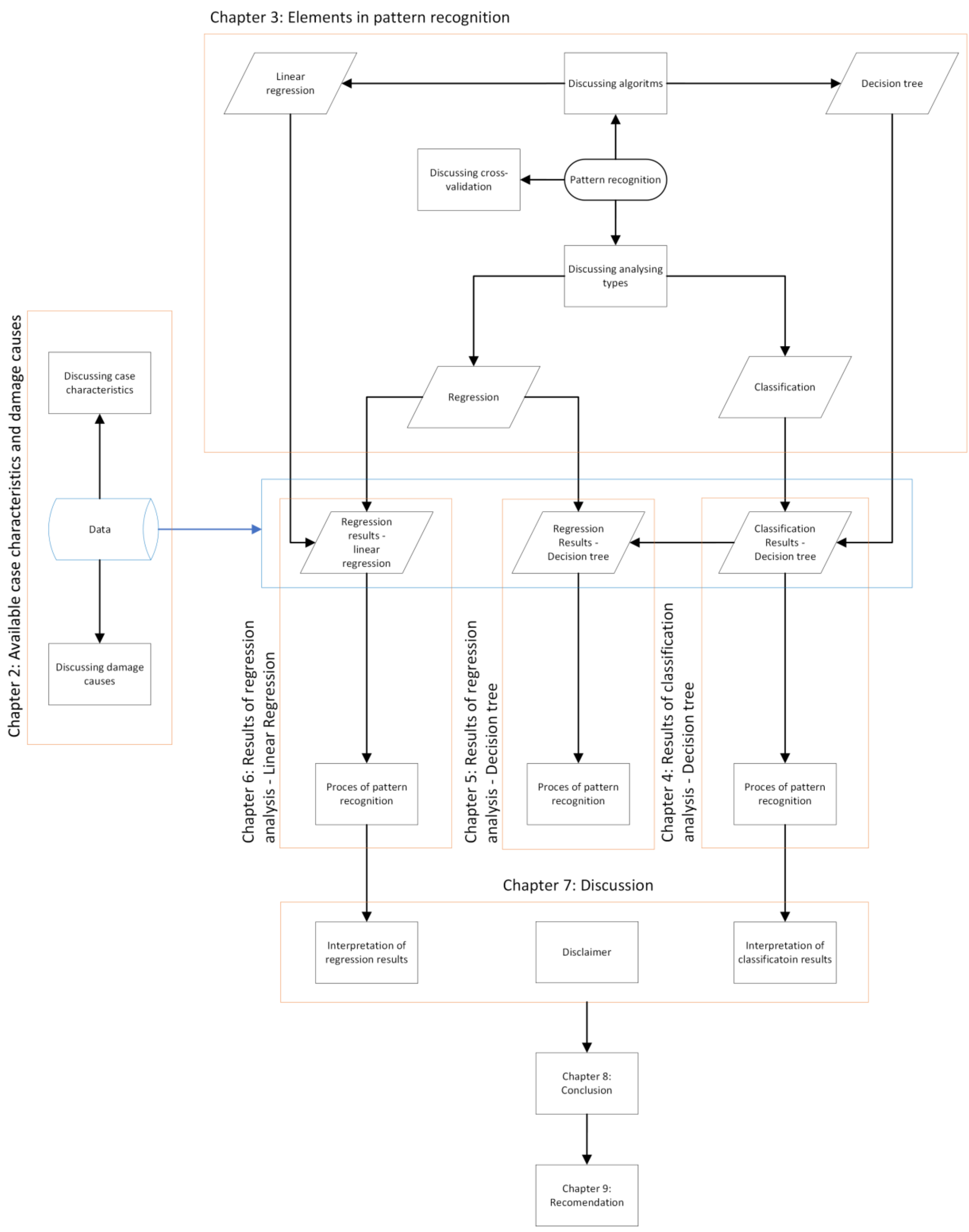


Figure 1 Flowchart of the report layout. Phase 1 is the data collection at the left side of the flowchart. Phase two is the right side of the flowchart.

1.6 Definitions of key terms

Damage

There is a wide variety of definitions of damage in the literature. These are mostly based on the context of the research or book. In this thesis, damage is interpretated in terms of structural damage. Van Herwijnen indicates that economical damage is also a part of forensic engineering (van Herwijnen, 2009, p. 11), but this aspect is not a part of this thesis. De Vent takes a more coordinating approach and notated damage as a manifestation of a defect (De Vent, 2011, p. 253). According to TNO, damage is limited to cracks (Borsje & de Richemont, 2011, p. 3). However, Witteveen+Bos define damage as: cracks, spalling, displacement or skewness and leaks (Salet & Bruurs, 2017, p. 17). Staalduinen et al. regard damage as visible manifestation that a building structure or part can no longer perform as it should (Terwel & Schipper, 2018, p. 2). According to them, damage manifests itself in practice as cracks, permanent deformations and permanent displacements. As this thesis is based on their work, their definition of damage is adopted in this thesis.

Damage is a visual manifestation of a lack of performance of a building structure or parts of it.

(Terwel & Schipper, 2018, p. 2)

Pattern

The term 'pattern' is sometimes confusing in this thesis, because two science fields are brought together and each of them employs 'pattern' in a different context. According to the Mathematics and Computer Science field, pattern is a description of a class in terms of measurements of features of an object (Webb, 2002, p. 2). For example, the pattern of a coffee mug (class) is a cylinder with a C shaped handle (features). This is in contrast with the Civil Engineering field, where the term pattern is defined as layout of a crack (Aeneas Media, 2019). For instance, cracks that follow a map's shape pattern, like # (Concrete Construction Staff, 2001). Because of this variation in definition, the term pattern is avoided as much as possible.

When this term is unavoidably applied, it is always placed in the context of the corresponding field. For instance, the definition of the Mathematics and Computer Science field is meant when the phrase 'pattern recognitions algorithms' is used. Whereas pattern in 'crack pattern' is defined in accordance with the Civil Engineering field.

Pattern recognition or *relations*

Description of a class in terms of measurements of features of an object (Webb, 2002, p. 2). For example, the pattern of a coffee mug (class) is a cylinder with a C shaped handle (features).

Crack pattern

"A characteristic combination of cracks with different crack directions." (De Vent, 2011, p. 254) In other words, crack pattern is the layout of a crack (Aeneas Media, 2019). For instance, cracks that follows a map's shape pattern, like # (Concrete Construction Staff, 2001).

Damage pattern

One of the characteristics used in this research is damage pattern according to De Vent. De Vent developed a tool that can categorize the appearance of damage, based on a typical combination of symptoms and conditions (De Vent, 2011, p. 253). For example, a tilted building is damage pattern 60 and a spalling masonry brick is damage pattern 1. The difference with *crack pattern* is that *damage pattern* includes all types of structural damage, and is not restricted to crack damage only.

Feature
Variables stated by the researcher are important for classification. These variables are called features in the mathematics and computer science field. In other words, data used for pattern recognition partly consist of measurements of features (Webb, 2002, p. 2). For example, temperature in weather forecast is a feature that can be measured in Celsius or Fahrenheit. However, the term feature appeared too meaningless in this thesis, based as it was on the provided feedback. It lacked the context of the Civil engineering field. Also, it was inadequate due to technical language of referenced literature. Therefore, the term characteristic is used instead of feature.

Characteristic
A characteristic is an element in a description of a damage case. Technical inspections and bureau studies are necessary to collect all information of each characteristic of a damage case (Salet & Bruurs, 2017, p. 8). For example, 1950 is a measurement of the characteristic *construction year*. In the Mathematics and Computer Science field characteristic is defined as *feature* in pattern recognition (see *feature*).

Unfortunately, the relevant literature is not consistent with the terminology of characteristic. De Vent introduces *context condition* as a term for characteristic. De Vent defines it as a condition that influences the occurrence of damage in a prerequisite or contributory way (De Vent, 2011, p. 254). However, the term *context condition* conflicts with *context characteristic* as meant in surrounding information. Witteveen+Bos applied the term *information* instead of characteristic. They define it as a collection of information that is relevant for the assessment of damage by experts (Salet & Bruurs, 2017, p. 17). Other literature does not adopt a unique term for variables of information about a damage case. For example: Ratay refers to background information and records of field activities (Ratay, 2010, p. 6.7); Borsje et al writes about information of the building, surrounding factors and damage inventory (Borsje & de Richemont, 2011, pp. 5 - 9); and TU Delft defines data collection as building characteristics, context characteristics and damage characteristics (Terwel & Schipper, 2018, p. 2). Since this thesis is based on the information of Staalduinen et al., the term characteristic is derived from TU Delft description.

Other literature suggests the term hazard parameter (Korswagen, Jonkman, & Terwel, 2019). The Peak Ground Velocity (PGV), for instance, is a hazard parameter for hazard *earthquakes*. Although this can be an adequate terminology for variables of information about a damage case, not all characteristics are covered by this term. For example, the dimension of a building does not have a connection with a hazard. Also, not all characteristics can be categorized as parameter. Such as the presence of a basement. It is an indicator for the hazard *unequal settlements*, but it has no sliding scale that can be interpreted in a parameter.

In the end, characteristics are defined as a collection of:

- information that has an influence on the occurrence of damage in a contributory or prerequisite way and,
- a description of damage itself.

This collection of information is interesting for experts in their judgement of damage.

Case
A report of one single occurrence of damage is called case. This consists of characteristics and a judgement by experts about the cause of damage. Buildings can have multiple cases if more than one occurrence of damage is present.

Cause
The reason why damage occurs is called a cause. Reasons consist of multiple layers. For example, if a car causes damage to a building by driving into it. Is the reason of damage the collision with an object, or the incompetent driving skill of the driver? Also, a bad road design could be a reason for cars hitting buildings. What about a mechanical malfunction of a vehicle? Therefore, TU Delft introduced the term technical attributability.

Technical attributability (technische toerekenbaarheid)
Terwel et el. describes the attributability as the share of an individual cause (Terwel & Schipper, 2018, p. 6). Staalduinen et al adopt a more complete definition of technical attributability: 'The judgement of the extent to which a cause has led to exceeding the allowed stresses in building or structure.' [translated] (van Staalduinen, Terwel, & Rots, 2018, p. 6). In the example of a car hitting a building a collision with an object is the (direct) cause of exceeding the allowed stresses in building or structure.

2 Available case characteristics and damage causes

Pattern recognition algorithms were utilized to find relations between case characteristics and damage causes. These algorithms requires damage cases that have already been analysed. The data in this thesis was provided by NCG as part of the research of TU Delft (van Staalduinen, Terwel, & Rots, 2018). However, the provided data needs to be prepared in order to be processed by the algorithms. That is because the algorithms only allowed numerical input and the provided data was mostly in plain text. In this chapter, it is explained how plain text from the provided data is converted into data that could be handled by the algorithms.

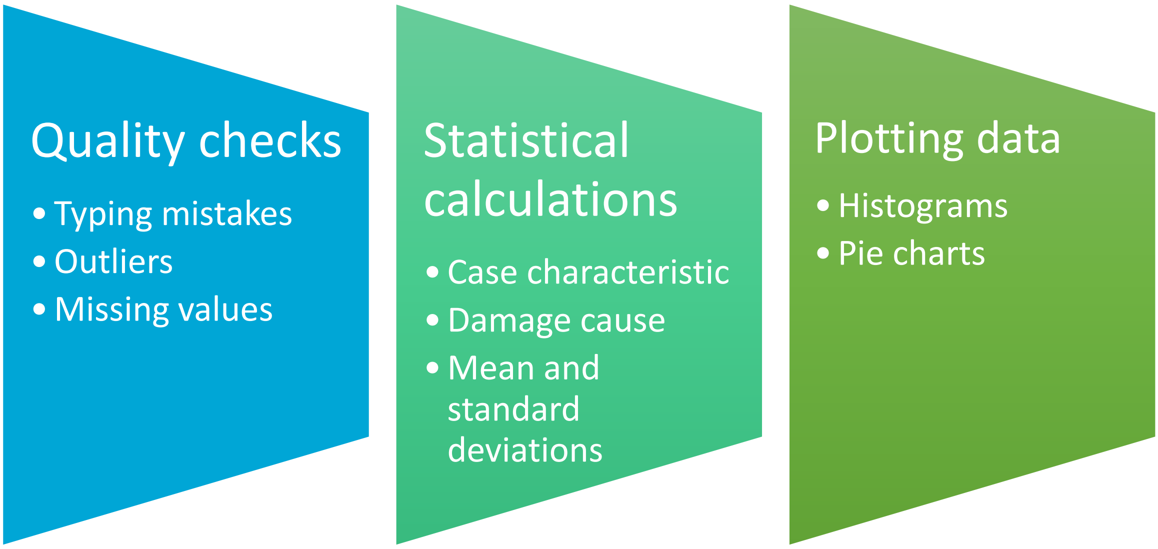


Figure 2 Applied initial data analysis (IDA)

While that is done, Webb suggests to do an initial data analysis (IDA). This provides data understanding and a good description of it. Also, it can give vital clues to which methods are suitable for analysis. This IDA consists of three parts: quality checking of the data, statistical summary calculations and gaining feel for the data by generating plots (Webb, 2002, p. 446).

The quality checking of the data is distinguished with three factors (Figure 2 - left). First factor, errors of measurements can degrade the data quality. Those malfunctions can occur in recording equipment transcriptions. In the case of this thesis, it can occur during a manual transformation of textual information in numerical data. A typing mistake is made quickly, or essential information in the report can be overlocked easily in this process. These errors are difficult to detect. The second factor of data quality checks are the presence of outliers. Those outliers are an indication of inconsistency in the data. Therefore, it is a signal for the previous mentioned errors in the data. The last factor of quality checking data is the appearance of missing data or values. It can arise in multiple ways, but it is important for the analysing of this quality check to known why and how this missing data appears (Webb, 2002, p. 447).

Statistical summary calculations might provide essential directions to which characteristics are important in the pattern recognition analysis (Figure 2 - middle). These statistical calculations should be done for each class and characteristic individually and on the whole data set. Commonly deployed statistics are mean and standard deviations calculations. Relative comparisons between that values can help with finding clues out of the data (Webb, 2002, p. 447).

Plotting data in histograms of individual characteristics or scatterplots of pairwise combinations of these can help with providing an insight of data (Figure 2 - right). These visual techniques are useful for exploratory data analysis (Webb, 2002, p. 447). However, not every characteristic is suitable for these types of plotting. Binary characteristics and categorised variables are presented better in pie charts, instead of histograms.

Those three parts of IDA is done in the appendix, where it is discussed how the case characteristics were collected from the data. While executing this IDA, other important questions are also answered. These questions help to justify the application of case characteristic. These issues are discussed in the following paragraphs.

2.1 Important characteristics for damage analysis

Before carrying out a IDA, it is essential to known which characteristics are important when investigating damage. Characteristics are a collection of information that is relevant for forensic civil engineering experts. According to the research of TU Delft, there are three types of characteristics: building characteristics, context characteristics and damage characteristics (See Figure 3). Building characteristics say something about the function, materials and size of the building. Context characteristics say something about sub soil, vibrations sources and external forces in the surrounding of the building. Damage characteristics say something about how damage occurs in terms of position, location and shape of damage. A complete list of all characteristics could be find in the appendix (van Staalduinen, Terwel, & Rots, 2018, p. 41)

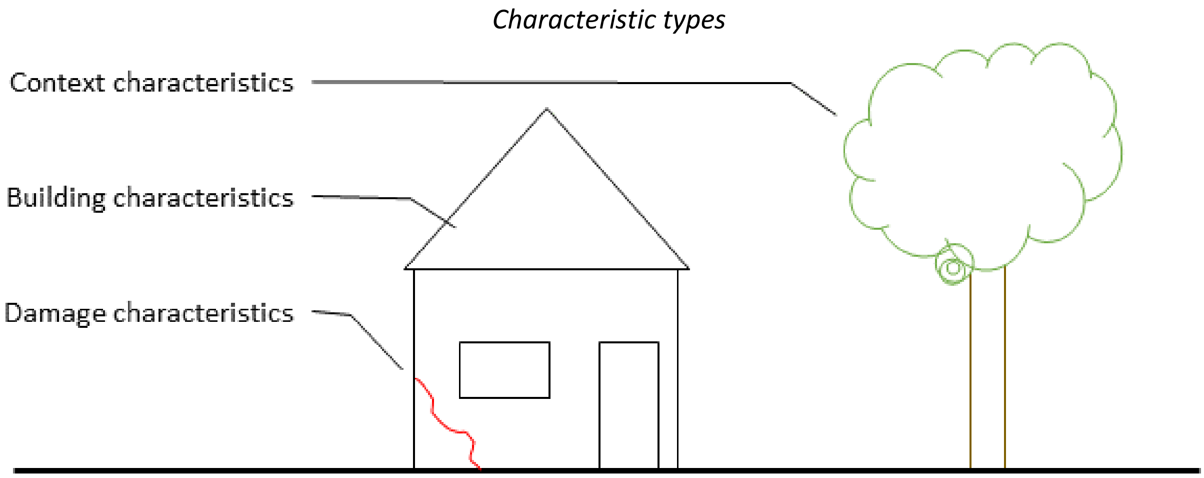


Figure 3 Visualisation of the three different case characteristic types.

2.2 Influence of characteristics – falsification

Not all chase characteristics are relevant for every damage cause category. Literature provides an hint at which causes a characteristic could be relevant. This falsification is interesting during the pattern recognition, because it is difficult to obtain a result at some analysis. In those cases, a well-founded selection of characteristics can help to improve the results.

2.3 Transformation of provided characteristics into algorithm excepting data

The characteristics consist mostly on 'yes' or 'no' questions. For an algorithm it is necessary to transform this characteristics in numeric value that it can calculate with. An example of this is the classification of the damage. This could be: *light, medium* or *severe*. This is easy to transform in: L=1, M=2 and S=3. The higher the classification, the higher the value of this characteristic. Another example is the direction of the crack: *vertical, horizontal* or *diagonal.* In this case there are three different features: *vertical* (yes/no)*, horizontal* (yes/no) or *diagonal* (yes/no). In other words, a higher value of the vertical characteristic is not a horizontal crack. In the appendix is explained how every characteristic is stored in the data set.

2.4 Handling missing data

Missing data is a common phenomenon when dealing with real world data. Examples of these are incomplete questionnaires and missing measurements of a medical patient. Also, data can be missing in the available data set. These missing datapoints need to be adapt, otherwise the used algorithm cannot produce meaningful and reliable relations. According to Webb, missing data can be handled in tree ways.

Removing missing data

		Damage cases					
		#1	#2	#3	~~#4~~	#5	#6
Case characteristics	A	Yes	Nee	Yes	~~Yes~~	No	Yes
	B	Yes	Nee	Yes	~~No~~	Yes	No
	C	No	Yes	Yes	~~?~~	Yes	No
	D	2.0	3.4	1.2	~~3.6~~	4.0	1.0
	E	0.1	0.5	0.1	~~1~~	0.1	0.9
	F	10	20	15	~~30~~	10	80
	G	Yes	No	No	~~Yes~~	No	No
Damage cause		B.1.1	B.2.3	B.1.1	~~C.2.1~~	B.X	C.X

		Damage cases					
		#1	#2	#3	#4	#5	#6
Case characteristics	A	Yes	Nee	Yes	Yes	No	Yes
	~~B~~	~~Yes~~	~~?~~	~~Yes~~	~~No~~	~~Yes~~	~~No~~
	C	No	Yes	Yes	No	Yes	No
	D	2.0	3.4	1.2	3.6	4.0	1.0
	E	0.1	0.5	0.1	1	0.1	0.9
	F	10	20	15	30	10	80
	G	Yes	No	No	Yes	No	No
Damage cause		B.1.1	B.2.3	B.1.1	C.2.1	B.X	C.X

Figure 4 Two option to omit missing data from the data set. Left, delete the complete damage case. Right, delete the characteristic completely.

The first one is to omit the missing data from the dataset. There are two options to do this. The first option is to delete the complete damage case where the data is missing from (Figure 4 left). The second option is removing the characteristic completely (Figure 4 right). The first option is suitable if the amount of missing data is very small. When the characteristic is removed it may lead throwing away potentially useful information. How useful the information is depends on why a characteristic is important when analysing damage. Another deciding factor of when removing the data is the spread of the data (Webb, 2002, p. 413). For example, the type of foundation is very important when analysing damage. However, the proportion of missing data within the characteristic 'foundation type' does not matter anymore if all known foundation types are all the same. In that case it is better to remove the characteristic completely.

Replacing missing data

		Damage cases					
		#1	#2	#3	#4	#5	#6
Case characteristics	A	Yes	Nee	Yes	Yes	No	Yes
	B	Yes	Nee	Yes	No	Yes	No
	C	No	Yes	Yes	No	Yes	No
	D	2.0	3.4	1.2	?	4.0	1.0
	E	0.1	0.5	0.1	1	0.1	0.9
	F	10	20	15	30	10	80
	G	Yes	No	No	Yes	No	No
Damage cause		B.1.1	B.2.3	B.1.1	C.2.1	B.X	C.X

		Damage cases					
		#1	#2	#3	#4	#5	#6
Case characteristics	A	Yes	Nee	Yes	Yes	No	Yes
	B	Yes	Nee	Yes	No	Yes	No
	C	No	Yes	Yes	No	Yes	No
	D	2.0	3.4	1.2	3.6	4.0	1.0
	E	0.1	0.5	0.52	1	0.1	0.9
	F	10	20	15	30	10	80
	G	Yes	No	No	Yes	No	No
Damage cause		B.1.1	B.2.3	B.1.1	C.2.1	B.X	C.X

Figure 5 Non removing options of handling missing data. Left, proceeding analysis. Right, substitute with average value.

The second way on how to handle missing data is by using only available data (Figure 5 left). The suitability of this way of handling depends on the type of analysis performing. Decision tree analyses are not capable of handling missing data (Buitinck, et al., 2020). Another analyses that could be done is the calculation of the mean. It is done by only calculating the mean of the available data. This still can give poor results, because they are based on different numbers of samples, or different numbers of damage cases (Webb, 2002, p. 413).

The third way on how missing date could be handled is by substitute the missing values (Figure 5 right). The simplest and crudest method is to use the mean value of the characteristic (Webb, 2002, p. 414). Other options are to use the median or the most frequent one (Pedregosa, et al., 2011).

The choice between the different ways of handle with missing data depends on: "How much is missing, why data is missing, whether the missing values can be recovered" (Webb, 2002, p. 413). When discussing the characteristics one by one, an answer is given to this questions. This will lead to an answer on what to do with the missing data.

2.5 Causes of building damage

All possible damage causes are discussed in this section. For making a forensic building damage tool, it is important to know what the limitations are. One of the limitations is that not all causes are covered in this model, because not every cause can be found in the data (Figure 9). Therefore, this chapter starts with a literature review about generally accepted categorisation of damage causes. Next, damage cases are presented that are suitable for the analysis. Consequently, the limitation of the tool are described at the end of this chapter.

Cause of damage can be divided in three main categories. De Vent made this classification based on how loads work on buildings (De Vent, 2011, p. 74). These three main damage cause categories were: overloading, hindered dimensional changes and settlement. Every main category can be divided in variations of damage causes. However, TNO used a slightly different categorisation: loads, deformations and unequal settlements in sub soil (Borsje & de Richemont, 2011, p. 11). Witteveen+Bos used the categorisation of Borsje et al. with some adaptions of variations (Borsje & de Richemont, 2011, p. 18). This version was further developed by Staalduinen et al. and has been deployed in this thesis.

The list of possible causes are presented in appendix 2 (van Staalduinen, Terwel, & Rots, 2018, p. 31)

2.5.1 Explanation A.X cause category – Loads

A structure can bear a load if the forces on a building are in equilibrium with the resistance of the structure (De Vent, 2011, p. 86). This equilibrium can be disrupted in two ways. First, resistance is too low for loads (Figure 6 left). This is categorised in damage cause sublevel A.1.X as 'insufficient resistance'. Second, loads on the structure are too high for resistance. This is categorized in sublevels A.2.X, A.3.X and A.4.X as 'overloading'.

Failure to meet design principles is mostly the reason why structures have not sufficient resistance to bear the load (van Staalduinen, Terwel, & Rots, 2018, p. 30). This could occur during three moments in the lifetime of a building: Initial, when design or execution faults were made during construction; in the user phase, when faults were made during renovation; or at the end of the lifetime, when materials were degraded and therefore can no longer resist a load.

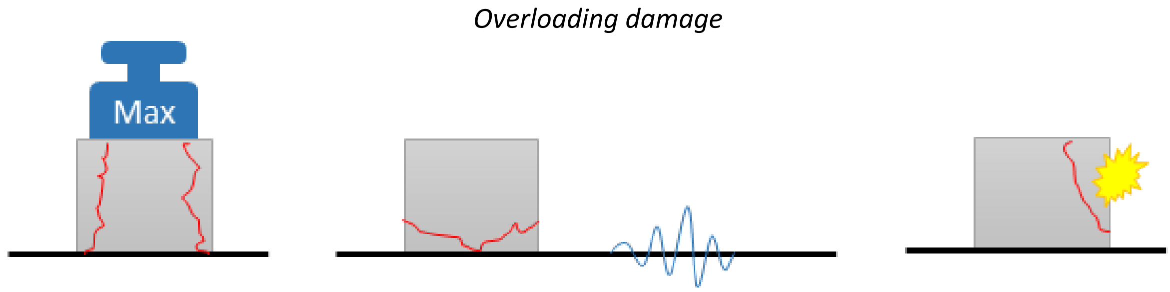


Figure 6 Examples of overloading. Left, overloading due to insufficient resistance or unexpected higher loads. Middle, overloading due to dynamic loading. Right, overloading due to incidental loads.

The construction is overloaded when the loads are higher than the expected design loads (Borsje & de Richemont, 2011, p. 15). There are three causes that lead to overloading. First, when buildings are (over)used (Figure 6 left). This type of overusing could occur in three situations: when a building is utilized too much in normal use, e.g. like bridges that undergo heavier traffic loads over the years; when expected usage of building is changed, e.g. an office area is transited into a storage; or renovations can turn into overloading situation, e.g. the placement of walls or swimming pools.

Secondly, dynamic loads that result in vibrations can lead to overloading of the structure (Figure 6 middle). This type of dynamic loads are generating stress changes in materials (van Staalduinen, Terwel, & Rots, 2018, p. 15). When the impact of those vibrations are high enough, they can lead to damage or even collapse of buildings (De Vent, 2011, p. 92). Dynamic loads can come from, traffic (road or rail), construction activities, industry or earthquakes.

Furthermore, incidental loads can end in overloading of structures (Figure 6 right). These loads are described as accidental or exceptional loads (van Staalduinen, Terwel, & Rots, 2018, p. 33) (Terwel & Schipper, 2018, p. 5). This type of overloading is recognised mostly by local smash-up of materials (Borsje & de Richemont, 2011, p. 15). Incidental loads could come from, collision with an object, explosions or severe weather.

2.5.2 Explanation B.X cause category – Deformations

Damage could occur because of deformations. Mostly, these deformations are small and relative slow movements, but large enough to cause damage. The dimension change of materials could be the consequences of variation in humidity or temperature. Especially porous materials are sensitive for deformations, like masonry, timber and concrete (Borsje & de Richemont, 2011, p. 15). Damage from deformations takes place for two reasons.

Deformations

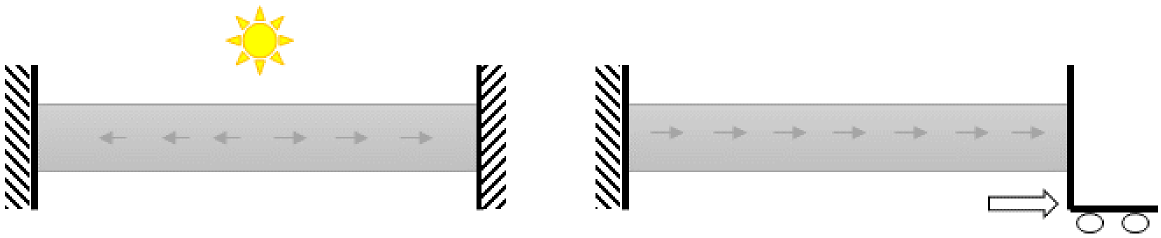


Figure 7 Examples of deformations. Left, hindered deformations. Right, imposed deformations.

First, a building element wants to expand, but it cannot expand (Figure 7 left). This is called hindered deformations (Salet & Bruurs, 2017, p. 41). An example of this are locked-in railway tracks heated by the sun. The tracks become longer, but because they were locked-in, they have no space to accommodate this deformation. As a result of that, the railway tracks can show damage (Scarpas, Erkens, Dollevoet, & Houben, 2015).

The other reason why deformations occur is because of imposed deformations. A building element has to expand, but it does not want to (Figure 7 right). To put it more formally, a building element cannot follow the deformations of another building element (Salet & Bruurs, 2017, p. 42). This could lead to damage or breaking of the imposed deformed element.

Hindered deformation damage could be prevented with sufficient design and execution of structures. Dilatations can accommodate these movements. Also, well-constructed connections can ware off damage from deformations (van Staalduinen, Terwel, & Rots, 2018, p. 34). Ones again, these structural design or execution faults can occur during three moments in the lifetime of a building. It could arise during the construction phase, the usage phase (renovations) or the terminal phase (degradation of materials) (van Staalduinen, Terwel, & Rots, 2018, p. 33).

In the same manner, imposed deformations could be prevented with good measurements. Commonly, imposed deformations are the consequences of deformations of other construction elements. However, not only construction elements are responsible for inducing imposed deformations. Growing roots and branches from trees could exert pressure on materials (van Staalduinen, Terwel, & Rots, 2018, p. 33).

Damage from deformations is found frequently in stone materials, like cracks that are appearing in cement, concrete or calcium-silicate bricks. They originate during the construction phase, where the stone material hardens and dry out. Consequently, the stone material shrinks a little bit. This process is called shrinkage. Shrinkage damage can occur if this is not accommodated for in the design of the structure. There are different shrinkage mechanisms possible: chemical shrinkage, volume of material that changed during the hardening process; plastic shrinkage, volume reduction during the plastic phase of materials; temperature or thermal shrinkage, deformations being caused by cooling or heating; hygrometric shrinkage, materials that were absorbed or drained of fluid; or dehydration or hydraulic shrinkage, liquid draining with permanent consequences by contraction of the pores in the material (van Staalduinen, Terwel, & Rots, 2018, p. 34).

2.5.3 Explanation C.X cause category – Unequal settlements

Unequal settlements can cause damage to a building, because sub soil is not equally supporting the building foundation anymore. Consequently, forces in the structure have to follow another load path. Also, displacement can manifest itself in parts of the building if unequal settlements occur (De Vent, 2011, p. 76). In both cases, it can result in damage. Staalduinen et al indicated 17 situations, divided over three subcategories that can develop in unequal settlements under the foundation of buildings (van Staalduinen, Terwel, & Rots, 2018, p. 31).

Although they were very specific with those situations, it was in practise not possible to distinguish a cause of unequal settlements other than that it was noted. Therefore, all situations that develop unequal settlements were combined as one cause. This was noted as C.X. One exception on this was the situation of autonomous settlements with unchanged loads. Nevertheless, the 17 situations that can develop unequal settlements will still be discussed in the next paragraphs, because they are essential for the understanding why some characteristics are important for the analysis.

The basis of soil mechanics is needed to explain the principle of settlement. Soil consists of small particles. The space between the particles, the pores, are filled with ground water. When soil is loaded, the forces of a load are initially shared between particles and water (Verruijt, 2018, p. 32). Water will leave pores over a length of time, because it is pressured out of the pores by the newly introduced load (van Staalduinen, Terwel, & Rots, 2018, p. 34). Consequently, the particles have to take a larger share of the total load. That will result in an increase of the stresses in the particles. This will lead to a rearrangement of the system of the particle skeleton, where particles roll and slide over each other (Verruijt, 2018, p. 98). This soil deformation over a length of time ends in settlements, also called creep. The duration of this process is depending on multiple factors (Verruijt, 2018), but mostly summed up in a consolidation coefficient (van Staalduinen, Terwel, & Rots, 2018, p. 34).

Unequal settlement

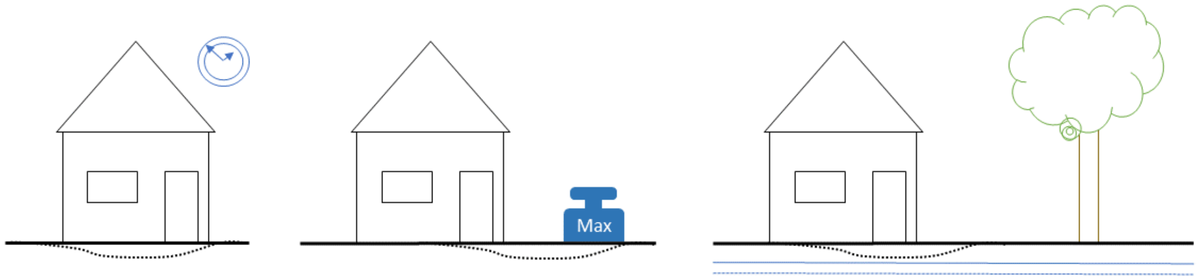


Figure 8 Examples of unequal settlements. Left, autonomous settlements. Middle, load change. Right, change in ground water level.

The first subcategory of situations that can develop unequal settlements are unchanging loads with autonomous settlements (Figure 8 left). Autonomous settlements are the long-time soil deformation initialled by constructing the building. Unequal soil properties or unequal foundation loads are responsible for the unequal part of this settlement (van Staalduinen, Terwel, & Rots, 2018, p. 34).

The second subcategory of situations that can develop unequal settlements are load changes in the subsoil (Figure 8 middle). This works more or less the same as the previous subcategory. However, the origin of the load change is not from constructing the building, but external factors in the surrounding of a building that are causing stress changes in the soil. Those external factors are mostly construction or building activities that have a significant influence on the stresses in the sub soil such as excavations and elevations, e.g. constructing of rail and road ways; or building activities, like a construction of a new building nearby or an extension of the building itself (van Staalduinen, Terwel, & Rots, 2018, p. 34).

The last subcategory of situations that can develop unequal settlements are changes in the subsoil (Figure 8 right). As previously mentioned, ground water that flows out the sub soil pores is an essential part of ground settlements. Therefore, changes of the sub soils ground water level (GWL) have an influence in that process. GWL can change because of various reasons like trees roots that take in water from the sub soil or pumps that retrieve water to accommodate work activities in the surrounding (Salet & Bruurs, 2017, p. 231). Other changes in sub soil that can develop unequal settlements has to do with the arrangement of the particles in the sub soil. Vibrations in here can rearrange those particles. There are various sources that can induce vibrations, such as traffic, work activities and earthquakes (Borsje & de Richemont, 2011, p. 13).

2.6 Availability of causes within the provided data

Unfortunately, not all causes were available for analysing, because they were not mentioned as cause in the data or the concurrency of a cause was very rare.

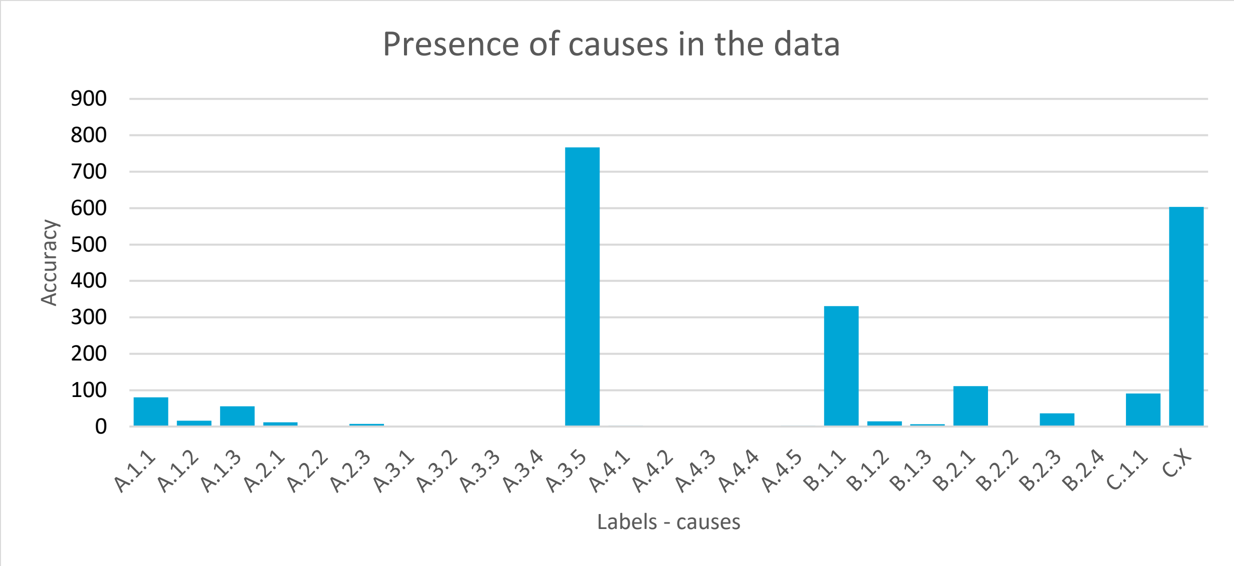


Figure 9 The number of times a cause category was mentioned in the dataset.

The figure above shows how many times the specific damage cause was mentioned in the data. The most mentioned cause is A.3.5 (vibration loads because of earthquakes). That is followed by C.X (unequal settlements) and B.1.1 (initial hindered deformations). On the other hand could be seen that some causes never were mentioned as reason for damage. These causes are: A.2.2, A.3.1, A.3.2, A.3.3, A.3.4, A.4.2, A.4.3, A.4.4, B.2.2 and B.2.4 (see appendix 4). Consequently, those causes are no longer part of this thesis research.

It is not very surprising that these causes are very rare in the data. They describe unique situations that hardly occur. For example, overloading by an explosion (A.4.2) is not something that should happen often. Also, this are situations where damage is repaired before they could exanimated by the experts of the TU Delft. For example, damage from a storm (A.4.4) is mostly directly repaired by the owner, like the replacing of rooftiles. Other directly repaired damage causes are related to building renovations (B.2.2). There, damage is often directly repaired by the construction workers.

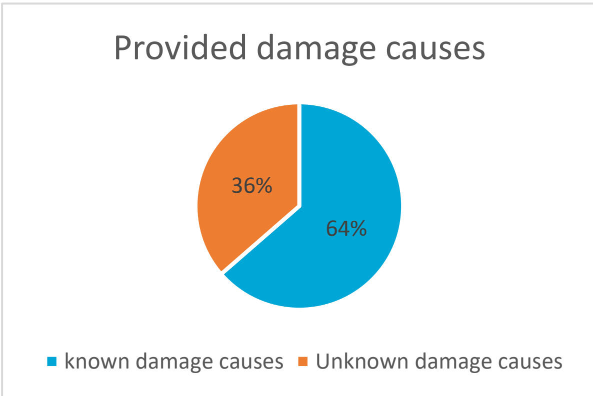


Figure 10 The cause of damage was not provided for all damage cases. Only the damage cases with a known cause were incorporated in this thesis.

There is also a batch of causes of which presence in the data was very rare. For those rare causes, pattern recognition is difficult to apply, or even not impossible. When a cause has sufficient presence in data is determined by the number of characteristics (Webb, 2002, p. 445). Webb stated in this general guideline that five - ten times more mentions per cause are needed for pattern recognition than applied features measurements. At the initial phase of this thesis, the number of features is about 200 characteristics. That will implies that 1000 – 2000 mentioned per cause are needed. With the current size of data, pattern recognition is not possible because of insufficient amount of mentions per cause. With a meaningful pre-selection of 20 characteristics, 100 – 200 mentions are needed per cause. As a result of that, six causes remain available for analysing. Those are:

- A.1.1 (Initial insufficient resistance to bear loads), present in 22 of the 41 buildings;
- A.3.5 (Overloading due to vibrations by earthquakes), present in 39 of the 41 buildings;
- B.1.1 (Initial hindered deformations), present in 36 of the 41 buildings;
- B.2.1 (Initial imposed deformations), present in 26 of the 41 buildings;
- C.1.1 (unequal settlements with equal loads), present in 6 of the 41 buildings;
- C.X (Unequal settlements), present in 38 of the 41 buildings.

The damage cases from unequal settlements with equal loads (C.1.1) are also included in the overall damage cause category unequal settlements (C.X).

3 Elements in pattern recognition

In general, the aim of pattern recognition is to derive a system that can distinguish and classify objects (Beyerer, Richter, & Nagel, 2017, p. xix). In our case, the opt system needs to distinguish the cause of damage based on case characteristics. This pattern recognition seems an easy task in daily practice. For example, we need to distinguish eatable objects from non-eatable objects (Beyerer, Richter, & Nagel, 2017, p. v). However, pattern recognition is a sophisticated assignment for computer programs (Patch & Smalley, 2004, p. 1). This sophisticated assignment is made easier by dividing the task into smaller problems.

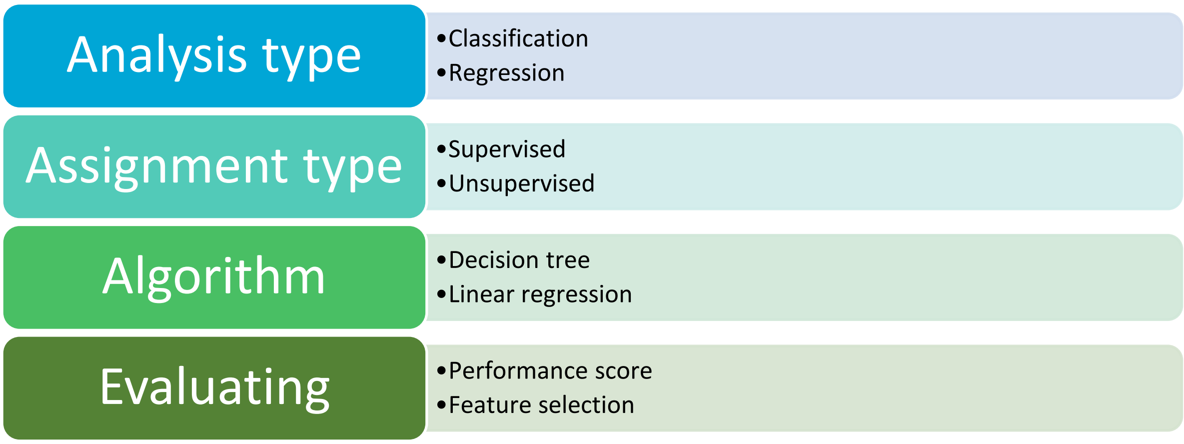


Figure 11 Dividing pattern recognition in smaller problems.

The first division is made in type of analysis. There are two types: classification and regression. Classification analysis sorts things into their own groups according to their characteristics (Cambridge University Press, 2021). For example, when LEGO bricks are sorted according to their shape and colour. In the case of this thesis, classification pattern recognition tries to find results in terms of which cause category caused a damage. On the other hand, regression says something about the change in the state of things (Cambridge University Press, 2021). This type of analysis gives results in quantifiable outcome. An example of regression pattern recognition is a weather forecast that tries to predict the temperature or how much rain will fall the following days. In this thesis, regression analysis is suitable for predicting the technical attributability of damage. In other words, how much a specific cause category has contributed to the damage of a building.

The second division made in order to make the sophisticated pattern recognition assignment easier is the division between supervised and unsupervised analysis. The opt results of a pattern recognition are known in supervised analysis (Webb, 2002, p. 6). The algorithm is trained to patterns based on provided results. For example, pattern recognition on handwriting seeks to fit the shape of a line to a letter or number. Unsupervised analyses cluster the data based on similarity of characteristics (Gevrilova, 2020). In the case of an unsupervised handwriting analysis, the algorithm clusters pen lines together that are more or less the same, but it cannot connect them to a letter or number. Supervised analyses are particularly suitable for the intended result of this thesis.

The third division made in order to make the sophisticated pattern recognition assignment easier is by defining an algorithm. There are different types of algorithms to choose from: Bayes' rule and estimations of the class-conditional densities, like normal-based models and nonparametric approaches; or a discriminant function approach, like linear discriminant functions and tree-based methods (Webb, 2002, p. 27). The committee advised to try two types of algorithms on this dataset:

linear regression and *decision tree.* Further explanation on why those algorithms are chosen is clarified after these algorithms are explained, later in this chapter.

The last division made in order to make the sophisticated pattern recognition assignment easier is by separating the problem of evaluating the results of pattern recognition. The performance of the algorithm can be evaluated with various methods. These methods mainly consist of scoring rate calculations, different ways of cross-validations and feature selection. These topics will be discussed later in this chapter.

3.1 Linear regression

This algorithm is based on finding a linear relation between characteristics and causes. The relation is based on the equation (Harrison, 2016):

$$\hat{y} = mx + b \tag{3-1}$$

In which:

$\hat{y}$ is the predicted cause parameter of damage,
m is the slope of the best fit line,
x is a parameter of a characteristic,
b is the y-intercept constant.

If there are sufficient data points between x and y, it is possible to calculate the m and b in the equation. In the presented example, there is only one characteristic (x) in equation (3-1) which linear regression tries to fit on the cause (y). In this case, the slope (m) and y-intercept can be calculated with (Harrison, 2016):

$$m = \frac{\bar{x} \times \bar{y} - \overline{xy}}{(\bar{x})^2 - \overline{x^2}} \tag{3-2}$$

$$b = \bar{y} - m\bar{x} \tag{3-3}$$

In which:

$\bar{x}$ is the average of a characteristic parameter,
$\bar{y}$ is the average of a damage train cause parameter,
$\overline{x^2}$ is the average of squared characteristic parameter (x^2),
$\overline{xy}$ is the average of the multiplication of a characteristic parameter (x) and a damage train cause parameter (y).

The calculation of the slope and y-intercept become more complicated when more characteristics are involved in linear regression pattern recognition. The method of least squares is used in linear regression algorithms to calculate slope (m) and y-intercept (b) in these cases. They try to minimize the sum of the squares of the vertical distance between the data points and the line (Stewart, 2012, p. 26), in this case the sum of the squares of the damage cause parameter. This (ordinary) least squares method is mathematically defined in the form of (Pedregosa, et al., 2011, pp. scikit-learn.org/stable/modules/linear_model.html):

$$\min_{m} \|\hat{y}(m, x)\|_2^2 \tag{3-4}$$

In which:

y is the train cause parameter of damage.

The mathematical notation of the linear regression algorithm is in the form of (Pedregosa, et al., 2011, pp. scikit-learn.org/stable/modules/linear_model.html):

$$\hat{y}(m, x) = m_0 + m_1 x_1 + \cdots + m_p x_p \tag{3-5}$$

In which:
p is the number of characteristics present in the data.

3.2 *Decision tree*

Decision tree methods can be applied to classification analysis as well as regression analysis. The algorithm is suitable for supervised learning. The intended result is a structured set of simple decision rules, derived from data, which predicts target value variables (Pedregosa, et al., 2011, pp. scikit-learn.org/stable/modules/tree.html). A decision consists of a comparison between a value of one specific characteristic and a threshold value (Webb, 2002, p. 226). These simple decisions are structured in a trees roots shape.

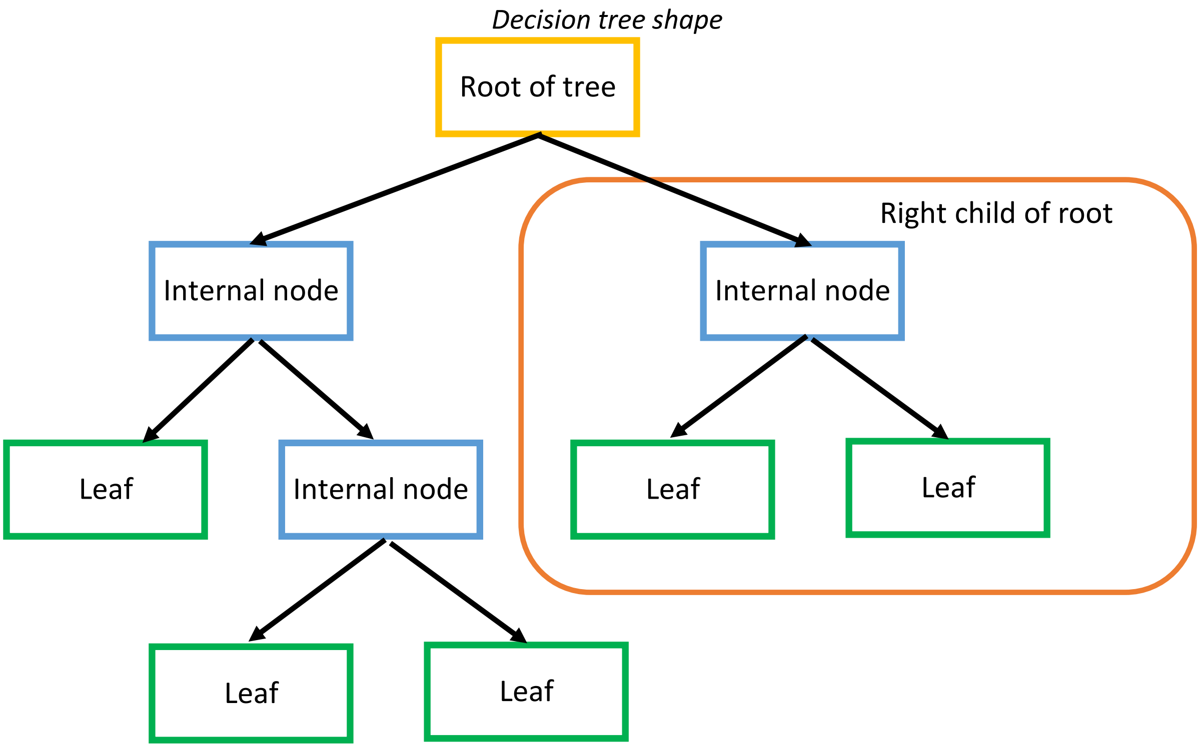


Figure 12 The shape of a decision tree.

The shaping of decision trees starts with the top node. This starting point node is called the root of the tree (Figure 12 yellow). The decision at this root dictates if the pattern proceeds to the internal node on the left or on the right (Figure 12 blue). They are called the left child and right child respectively. The right child is chosen if a specific characteristic value exceeds a threshold value (Figure 12 orange), otherwise the left child is chosen. At this internal node, a new decision takes place about the direction in which the pattern proceeds to a next internal node. After proceeding through all internal nodes, a decision tree ends in a terminal node, also called leaf (Figure 12 green), in which a predicted value is assigned to a pattern (Webb, 2002, p. 226). In this thesis, a predicted value can be a cause of damage when carrying out a classification analysis, or it can be a technical attributability of one damage cause when carrying out a regression analysis. Decisions within the tree are based on case characteristics.

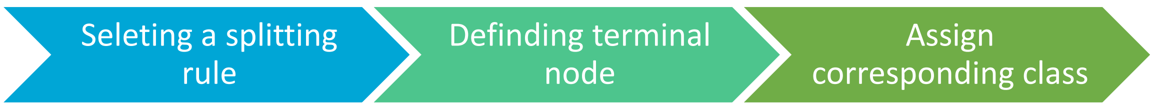


Figure 13 The three steps of creating a decision tree.

The creation of a decision tree consists of three steps (Figure 13). First, case characteristics are selected by means of an associated threshold that will be applied to divide data at each node. This is called selecting a splitting rule. Next, the terminal nodes are defined. A decision tree algorithm may continue splitting until all train samples in a leaf belong to the same classification class. However, this may result in a large tree with a large amount of patterns. There is very little chance that two or more damage cases have exactly the same pattern. This inconvenient splitting until pure class membership is achieved is called overfitting. Several stopping rules that reduce the risk of overfitting are discussed later in this chapter. The final step in the creation of a decision tree is a relatively straightforward assignment. In this step, terminal nodes are assigned to their corresponding damage cases (Webb, 2002, p. 228).

There are various splitting rules and decision tree algorithms. Basically all methods try to determine the best split in a set of items during each step (Rokach & Maimon, 2005). Multiple criteria can define the best split. Webb describes a criteria function of Breiman et al. (1984) and the Gini criterion (Webb, 2002, p. 232), where Pedregosa et al. also introduces criteria such as misclassification and half poisson deviance (Pedregosa, et al., 2011, pp. scikit-learn.org/stable/modules/tree.html) Which splitting criteria are to be applied depends on the decision tree algorithm that is applied and the type of pattern recognition analysis that is executed. The applied Python package in this thesis, Scikit-Learn, uses an optimised version of the CART algorithm (Buitinck, et al., 2020).

Several stopping rules are available in the Sckit-Learn Python package to reduce the risk of overfitting. The stopping rules are designed to stop unnecessary and unwanted splitting before reaching single pure leaves.

The first rule is setting the maximum depth of the tree. The depth of a decision tree is defined by the number of nodes in the longest pattern. An algorithm stops when the set maximum depth is reached instead of continuing splitting until the level of pure leaves.

Stopping rules

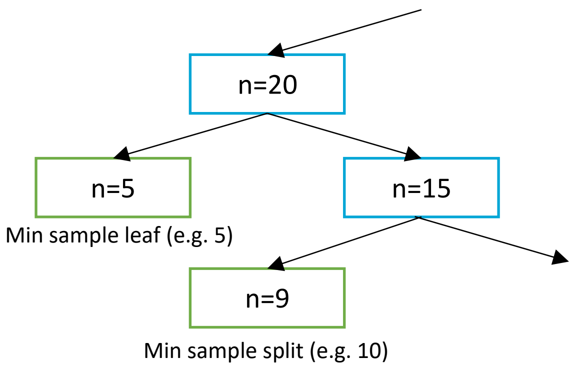


Figure 14 Examples of stopping rules.

The second stopping rule sets a minimum sample split (Figure 14). An algorithm will stop splitting when the number of samples in the split is equal to or less than the selected minimum number of required samples for a split. For example, when a minimum sample split is set to ten, whereas a pattern only contains nine impure damage cases that are to be split by an internal node yet to be designed, it creates a leaf instead of an internal node.

A third splitting rule sets a minimum number of samples for a leaf (Figure 14). An algorithm will only create an internal node the number of samples in both the right and left child are at least equal to the required number of samples set by this rule. For example, each leaf has at least five damage cases at the end of the training phase if the number of minimum required samples of a leaf was set to five (Pedregosa, et al., 2011, pp. scikit-learn.org/stable/modules/tree.html).

Other stopping rules are also possible within Scikit-Learn, such as setting a minimum weight required for a leaf or a maximum number of features when considering a split, but the splitting rules discussed above seems sufficient to prevent overfitting (Pedregosa, et al., 2011, pp. scikit-learn.org/stable/modules/tree.html).

3.3 Motive for chosen pattern recognition method

Generally speaking, the choice for a pattern recognition method is a trial and error process. Recommendation on which algorithm to apply on a specific problem does not exist in literature. In practice, smart choices for a method are based on experiences from the past. However, there are still some arguments to apply the advised algorithms, which are discussed above.

There are a few reason for applying the decision tree for this problem. The first reason is that decision tree methods are already deployed for forensic engineering purposes (De Vent, 2011). Secondly, the method of execute splitting rules in decision tree is already utilized in civil engineering practice in form of falsification criteria (van Staalduinen, Terwel, & Rots, 2018). Also, decision trees are better in classification analyses than other algorithms, because mostly all other methods are not as well equipped in handling with discrete values as this proposed algorithms does. This also another reason why decision tree seems a well fitted algorithm for this data. The provided data consists mainly of discrete values such as if a basement is present or not. The use of splitting rules in this method are accommodate those discrete values, whereas other algorithms still handle them as continuous values.

Decision tree methods are not suited well for regression analysis, therefore linear regression is a reasonable basis for supervised analysis. Decision tree tries to discrete continuous labels values. For example, the continuous value of the technical attributability is split by decision tree method in groups of 10 to 20%. As a result of that, some precision is lost in the pattern, as it does not matter if it is 1% or 9%. While this precision is important for earthquake related damage. Another reason to apply linear regression is that it is not expected that a characteristic has an optimum value where damage occurs. It is predicted that increasing continuous values will result in more damage. For example, higher loads will result in more damage. There is not an optimum where after increasing loads will result in less change of damage. Therefore, it seems logical to start with searching for linear relations rather than parabolic or goniometric relations. As final reason for linear regression, it is a simple and robust model. This is necessary since the provided data has noise in it. It has a lot of characteristics and a clear pre-described relations is missing. When dealing with such data, a easy to understand robust algorithm is an option.

Also for both methods, the result are easy to interpreted. Decision tree methods can be checked by visual follow through the decision tree. The patterns are well structed and it is easy to apply them. For linear regression, influence of individual features (characteristics) are clear presented in the result in form of their individual weights. That gives a picture on how much and which way a single characteristic will have an effect on a result. Interpretation of the result is more difficult to do with other algorithms, because the relations and patterns have a more complex nature.

3.4 Evaluating the performance of a pattern

The evaluation of a found pattern or relation is carried out with an independent data set. This independent data set is obtained by splitting the original data in a train set and a test set (Webb, 2002, p. 445). The former is used to train the algorithm and the latter is used to test the found relation of the algorithm (Figure 15). How this split can be made will be explained later, but first it is explained how a performance score can be calculated with the help of the test data. There are many possible scoring parameters (Pedregosa, et al., 2011, pp. scikit-learn.org/stable/modules/model_evaluation.html). Which one to apply depends on the type of pattern recognition that is carrying out. In this thesis, two types of scoring parameters are applied. One type of scoring is implemented in classification and the other is used for regression analysis.

Random split

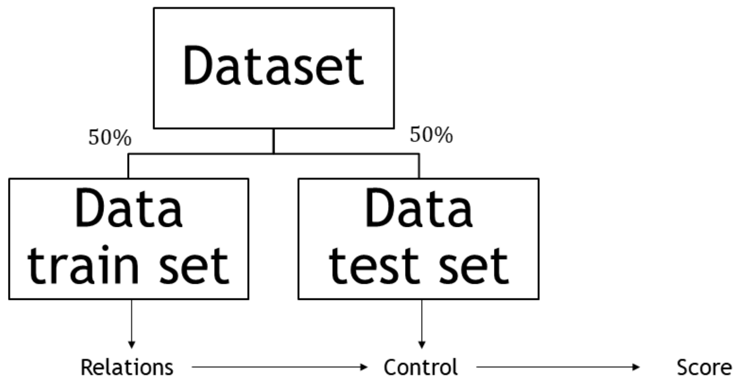


Figure 15 Simple single random split

The *accuracy score* is applied as a scoring parameter for classification pattern recognition analysis. This scoring parameter provides a proportion of correctly predicted classifications. The prediction is made by applying the found relations of a data train set on the data of the test set. The correct predictions are counted and divided by the total number of performed tests (Pedregosa, et al., 2011, pp. scikit-learn.org/stable/modules/model_evaluation.html#accuracy-score). The range of an accuracy score is between zero and one, where one is a perfect fit score. This accuracy scoring parameter is defined in mathematical form as:

$$\text{accuracy}\ (y, \hat{y}) = \frac{1}{n_{cases}} \sum_{i=0}^{n_{cases}-1} 1(\hat{y}_i = y_i) \tag{3-6}$$

In which:
y is the train cause parameter of damage,
n_{cases} is the number of cases in the test set,
$1(x)$ is an indicator function.

The *coefficient of determination (R²)* is applied as a scoring parameter for regression pattern recognition analysis. This scoring parameter compares the predicted value with the actual mean value. It does that by dividing the difference between the predicted and actual value by the difference between the actual and the mean of the actual value. This procedure is explained in mathematical form by (Pedregosa, et al., 2011, pp. scikit-learn.org /stable /modules /model_evaluation. html #r2-score):

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3-7}$$

In which:
$\bar{y}$ is the average of the train cause parameter.

The minimum of R^2 is infinite. If $R^2 = 0$, there is no improvement compared with the mean value. If $R^2 > 0$, it is an improvement compared with the mean value. In other words, the closer to 1.0, the more reliable the result. The square roots in equation (3-7) will result penalising larger differences heavier than smaller differences. (Harrison, 2016, pp. pythonprogramming.net/r-squared-coefficient-of-determination-machine-learning-tutorial)

3.4.1 Cross validation

There are various options for splitting the available data in a training set and a test set. One option is to take a random sample from the total data as training set (Webb, 2002, p. 445). Harrison proposes to make an equal 50-50 split between training and test data (Harrison, 2016)(See Figure 15). However, there are a few downsides to this splitting method. First, a large proportion of the data is not used for training. As a result, not all data are efficiently applied when searching for patterns (Webb, 2002, p. 445). Secondly, a random split of cases cannot provide a required independent test set. Damaged buildings often have multiple damage cases. Since all damage cases in a building have the same building and context characteristics, this may result in test data that are largely the same as the train data. Therefore, the test data are no longer independent.

The first problem, the inefficient processing of all data, can be solved by employing cross-validation. The difference with the previously presented split option is that in cross--validation the data are split multiple times. This method calculates the score of the pattern recognition by taking n-x samples in the train set and validating these against the 'x' remaining sample. The set of remaining samples is called the validation set. This is repeated 'k' times using a different composition of the train data each time (Webb, 2002, p. 254). Each repetition is called a fold. Figure 16 shows a cross validation with four folds (k = 4) and a validation size of 25% (x = 0.25). This version of cross validation is called a K-fold (Pedregosa, et al., 2011, pp. scikit-learn.org/stable/modules/cross_validation.html#k-fold).

K-fold cross-validation

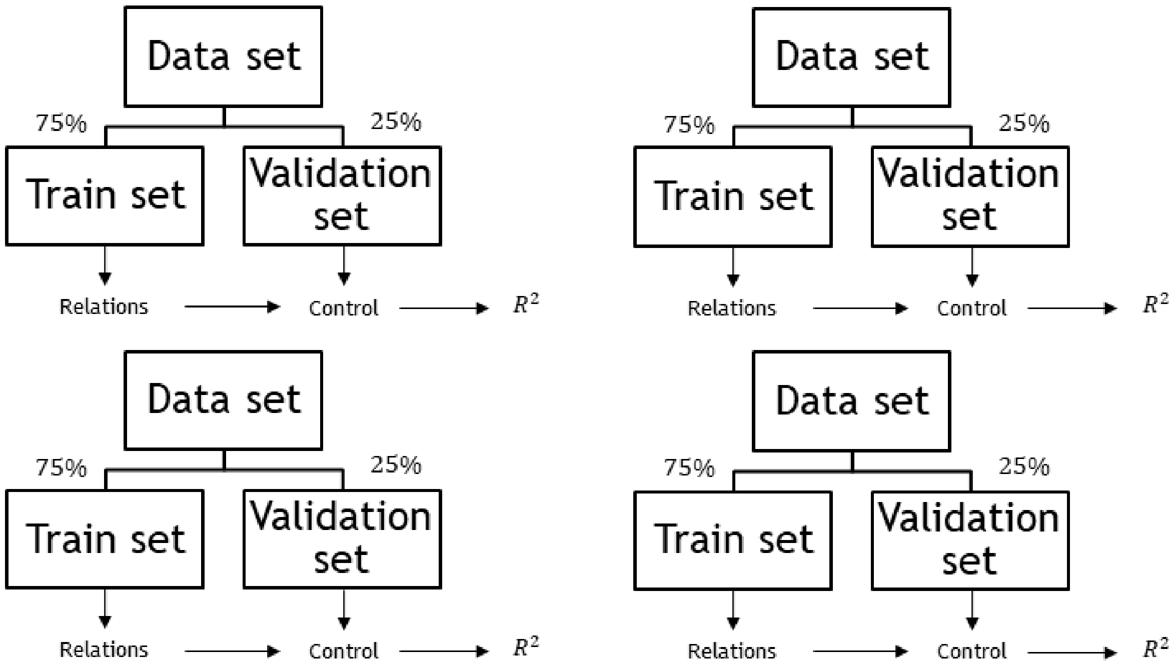


Figure 16 Example of cross validation k-fold

A common version of cross validation is the Leave-One-Out strategy (LOO). The validation set consists of one case instead of a combination of multiple cases (Webb, 2002, p. 410). It is actually a k-fold method in which the number of folds equals the number of available cases. The LOO cross validation is an efficient way of applying the data, because only one case is missing in every training set. However, the k-fold method is in generally advised by most authors if the computational execution of a large data set is too expensive. (Pedregosa, et al., 2011, pp. scikit-learn.org/stable/modules/cross_validation.html#leave-one-out).

LOGO Cross-Validation

	Group 1	Group 2	Group 3	Group 4	Group 5
Fold 1	Validation	Train	Train	Train	Train
Fold 2	Train	Validation	Train	Train	Train
Fold 3	Train	Train	Validation	Train	Train
Fold 4	Train	Train	Train	Validation	Train
Fold 5	Train	Train	Train	Train	Validation

Table 3 An example of a Leave-One-Group-Out cross-validation method (LOGO)

The second problem, of not gaining an independent test set when applying a split, can be overcome by grouping similar cases in a test set. In this way, the cases from one building are grouped together. They are all placed, as a group, in a train set or in a validation set, but cases in one group are never split between the two sets. This principle of grouped data can be applied to the same cross-validation methods explained earlier. For example, grouped data in the leave-one-out strategy are called the Leave-One-Group-Out method (LOGO) (Pedregosa, et al., 2011, pp. scikit-learn.org/stable/modules/cross_validation.html#leave-one-group-out).

GSS Cross-Validation

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8	Group 9	Group 10
Fold 1	Val	Train	Val	Train	Train	Train	Train	Train	Train	Train
Fold 2	Train	Val	Train	Train	Val	Train	Train	Val	Train	Train
Fold 3	Train	Train	Train	Val	Train	Train	Train	Train	Val	Train
Fold 4	Val	Train	Train	Train	Val	Train	Train	Train	Train	Train
Fold 5	Train	Train	Train	Train	Train	Val	Val	Train	Train	Train
Fold 6	Train	Train	Val	Train	Train	Train	Train	Train	Train	Val

Table 4 An example of a Group Shuffle Split cross-validation method (GSS)

Pedregosa et al. suggests yet another cross validation method. This method is called Group Shuffle Split (GSS). It is almost similar to a grouped k-fold (See Table 5), where multiple groups can be implemented in one validation group. However, the order of the groups are randomly shuffled in GSS (See Table 4Table 4 An example of a Group Shuffle Split cross-validation method (GSS)

). This is preferable for the provided database, because buildings are ordered according to the area. This area ordering influences the earthquake load. Another additional benefit in applying GSS instead of LOGO is the control over the cross validation. With GSS, it is possible to set the number of groups (buildings) in one data validation set. It is also possible to control the number of folds that has to be made by GSS strategy (Pedregosa, et al., 2011, pp. scikit-learn.org/stable/modules/generated/sklearn.model_selection.GroupShuffleSplit.html). However, this type of control of the cross-validation should not have too much influence on the score of the pattern recognition, because the type of applied splitting does not change the pattern finding algorithm itself. It only has an influence on the evaluation on the algorithm.

GKF Cross-Validation

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8	Group 9	Group 10
Fold 1	Val	Val	Train	Train	Train	Train	Train	Train	Train	Train
Fold 2	Train	Train	Val	Val	Train	Train	Train	Train	Train	Train
Fold 3	Train	Train	Train	Train	Val	Val	Train	Train	Train	Train
Fold 4	Train	Train	Train	Train	Train	Train	Val	Val	Train	Train
Fold 5	Train	Train	Train	Train	Train	Train	Train	Train	Val	Val

Table 5 An example of a Group-K-Fold cross-validation method (GKF) with 5 folds.

Another advantage of cross-validation is that it provides an possibility to optimize certain parameters. This validation helps to obtain a well performing pattern. However, a standalone data set is necessary to evaluate the performance of this pattern. This independent data set is called a test set.

Pattern recognition

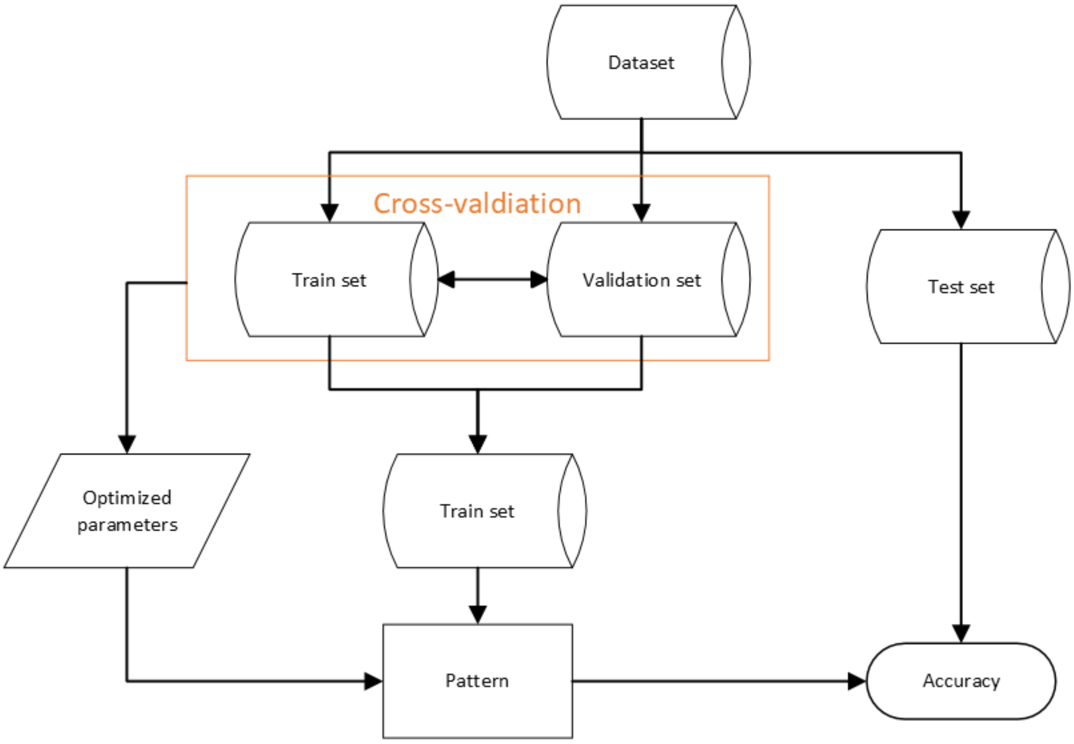


Figure 17 Flowchart of pattern recognition.

The process of pattern recognition is shown in the figure above. The provided data is split in three sets: train set, validation set and test set. The first two are combined in an earlier explained cross-validation. This will provide optimized parameters for a pattern. That pattern is evaluated on its performance with an test set.

3.4.2 Importance of characteristics

Besides scoring with cross-validation, feature importance is another pattern recognition evaluation method. Computer scripts are capable of calculating the importance of a feature. Therefore it is possible to say how much each characteristic has on the outcome of a pattern. The importance say something about if a characteristic is important for a pattern, but it does not say how a characteristic influence a pattern.

Statistical analysis of these characteristic importance's is available in combination with cross validation. Basically there are three possible outcomes of this statistical analysis. Option one, a characteristic is important in determine the cause of a damage, when all feature importance's from a cross validation score relatively high compared with the other characteristics. The second option is that a characteristic has no influence on a cause. This will occur when all cross validation steps agree on relatively low characteristic importance. The last option is that there is too much noise in a characteristic. The cross validation steps do not lead to a consensus about the importance of a characteristic. This could indicate an underperforming pattern recognition algorithm.

3.5 The class imbalance problem

When classes in a dataset differs a lot, it could influence the result of pattern recognition. This class imbalance problem causes the algorithm to focus on the dominant class. This will result in incorrect pattens for the not dominant classes. Imbalance means that there is a difference in class probabilities (Japkowicz & Sthephen, 2002).

There are a few solutions for this problem (Badr, 2019), but the following option was advised as the easiest to implement, practical and delivering an expected result. The imbalance solution for classification decision trees consists of two adaptions in the pattern recognition task: (1) using the balanced accuracy rate during evaluation, and (2) using weighted samples in the training and the decision tree.

First, the calculation of the accuracy score is adjusted. In this adjusted calculation, both classes are equally contributing to an accuracy score. This is done by allocating a weight to each damage case of a validation set. This value is called a sample weight. The sum of all sample weights of each class will be equal to 1.0. For example, a building has ten damage cases and two of them are caused by cause 'X'. Without balance corrections, each damage case will be weighted as 0.1. With balance corrections, the two 'X' causes will be weighed as 0.5 each. While the other eight damage cases will be weighted as 0.125 each. This will result in equal contribution of each class in the calculation of the accuracy. The sample weight for the validation set is calculated as:

$$sw_i = \frac{1}{n_i} \tag{3-8}$$

In which:
sw_i is the calculated sample weight for class i,
n_i is the number of samples of class i.

Secondly, the design of the decision tree is adjusted according to the frequencies of the classes to which the pattern is directed. This can be done by enabling a feature in the Decision Tree Classification python package of Sklearn. The class weight of each damage case is then adjusted to the frequency of the class in the form of (Buitinck, et al., 2020):

$$cw = \frac{n_s}{(n_c \times n_y)} \tag{3-9}$$

In which:
cw is the calculated class weight,
n_s is the number of samples in train set,
n_c is the number of classes in train set,
n_y is the number of samples of each class.

4 Results of classification analysis – Decision tree

In this chapter, it is explained how decision tree classification analysis has been carried out and how the results of those patterns can be interpreted. The chapter will start with a description of the process on how the patterns (relations) are obtained. After that, the results of each damage cause pattern is discussed.

4.1 The process of obtaining a classification decision tree with pattern recognition

The process starts with an description of an initial result. The improvements to make the basic result more reliable are explained step by step in the rest of this chapter. Also the problems that occurred during this pattern recognition analysis are defined, including the path that was followed to solve those problems.

All classes at ones

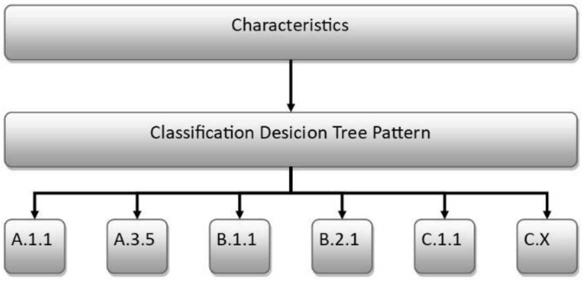


Figure 18 Classification done to all causes

Individual classes

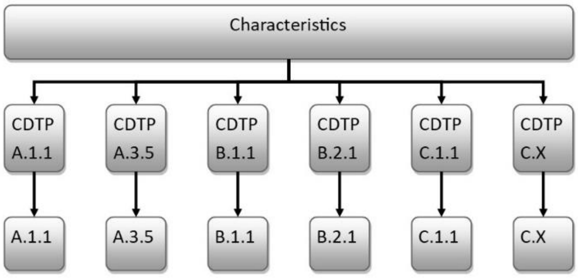


Figure 19 Classification done to each cause individual

Classification can be applied all causes at once, or to each cause individually. In the first option, a pattern will be constructed that can answer the question to which class an object belongs. In this thesis, it answers what the cause of building damage is (see Figure 18). In the second option, patterns will be assembled for each class separately. Here, each pattern answers the question whether a damage case belongs to the class or not.

In this research, it checks if damage was or was not caused by cause 'X' (see Figure 19). Although the first option (Figure 18) is more practical for experts in the field, it was not possible to construct such a pattern. A few problems occurred during the execution of classification to all causes. The first one was that the results were not understandable enough to give an insight in which characteristics were important for each cause. The second problem was that almost all damage cases were caused by multiple damage causes. This means that two cases that are exactly the same belong to multiple cause classes. This will make supervised pattern recognition learning very difficult when a pattern is trained to place an object (damage case) in multiple classes. Therefore, classification is applied to each cause individually.

4.1.1 Initial basic classification results

The basic classifier is a supervised decision tree algorithm without any application of stopping rules, cross-validation restrictions or data adjustments. As a first step, Group Shuffle Split (GSS) cross-validation was applied to calculate a score for the performance of this algorithm. Initially, the GSS variables were arbitrarily set to 41 folds and a test size of 0.25. In this case, the number of folds is the same as the number of buildings. A test size of 0.25 will result in ¼ of the buildings being placed in a validation set. An accuracy determination was deployed as performance scoring parameter, because this scoring type is suitable for classification.

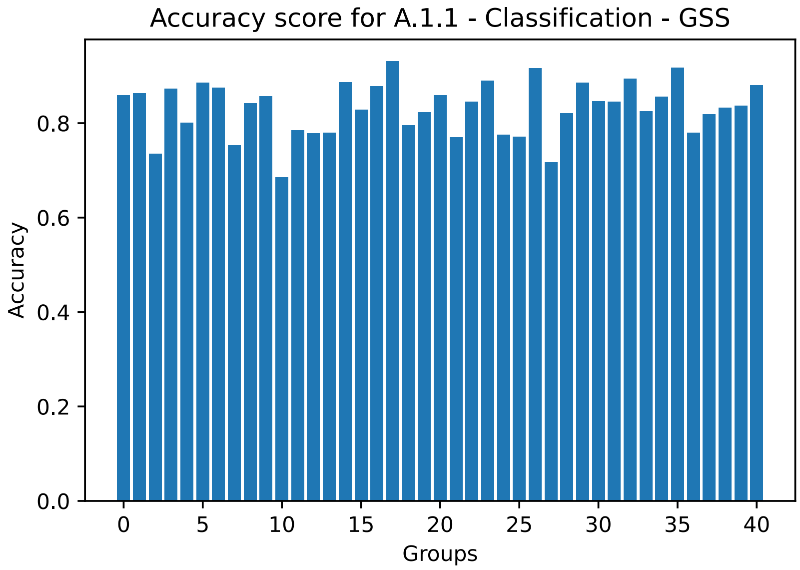


Figure 20 Cross-validation result of basic run on classification pattern recognition on damage cause Initial insufficient resistance to bear loads (A.1.1)

Figure 20 shows the accuracy score for every GSS cross-validation set on the basic initial pattern recognition run on classification on damage cause *initial insufficient resistance to bear loads* (A.1.1). In other words, this figure shows the accuracy of a pattern that can determine if damage was caused by cause A.1.1 with answering options 'yes' or 'no'. The horizontal x-axis displays each of the 41 validation groups. The accuracy can be read form the vertical y-axis. It can be concluded from Figure 20 that the basic pattern is correct on average in 83% of the cases.

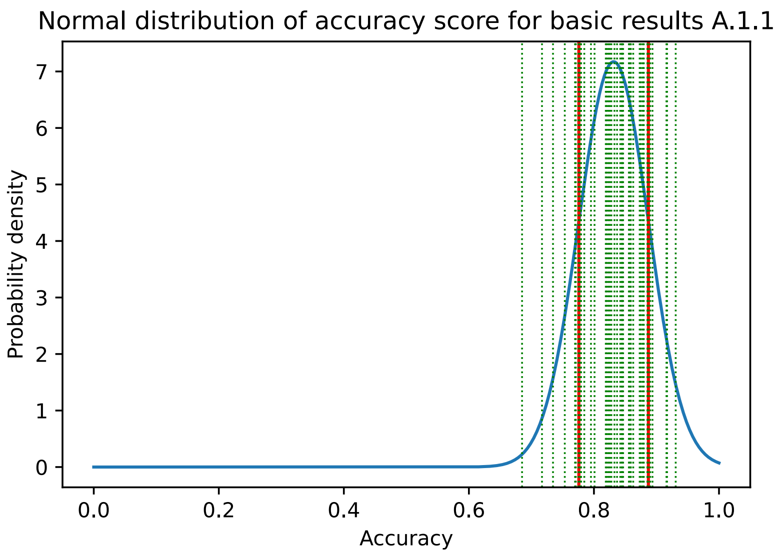


Figure 21 Normal distribution of cross-validation of basic run on classification pattern recognition on damage cause initial insufficient resistance to bear loads (A.1.1).

The information from Figure 20 can be made more understandable when it is represented in a normal distribution. This is shown in Figure 21. Every bar from Figure 20 is placed as a vertical dotted green line on the horizontal x-axis in Figure 21 with an accuracy scale from zero to one. The mean and the standard deviation are derived from this information. The red lines in Figure 21 indicate the standard deviation relative to the mean. Finally, the normal is calculated according to equation 4-1 (Stewart, 2012, p. 572) and plotted in Figure 21 as a blue line.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)} \tag{4-1}$$

In which:
$f(x)$ is normal distribution or probability density function,
σ is the standard deviation,
μ is the mean.

The sharper a normal distribution is, the more consensus there is between the different cross-validation results of the accuracy of the pattern. By definition, the area underneath the normal distribution is always equal to one. This will set the height of the normal distribution. Figure 21 shows that there is consensus about the accuracy score of classification pattern A.1.1 (initial insufficient resistance to bear loads) of 83%. There are no outliers in the data which could pollute or introduce noise the produced pattern.

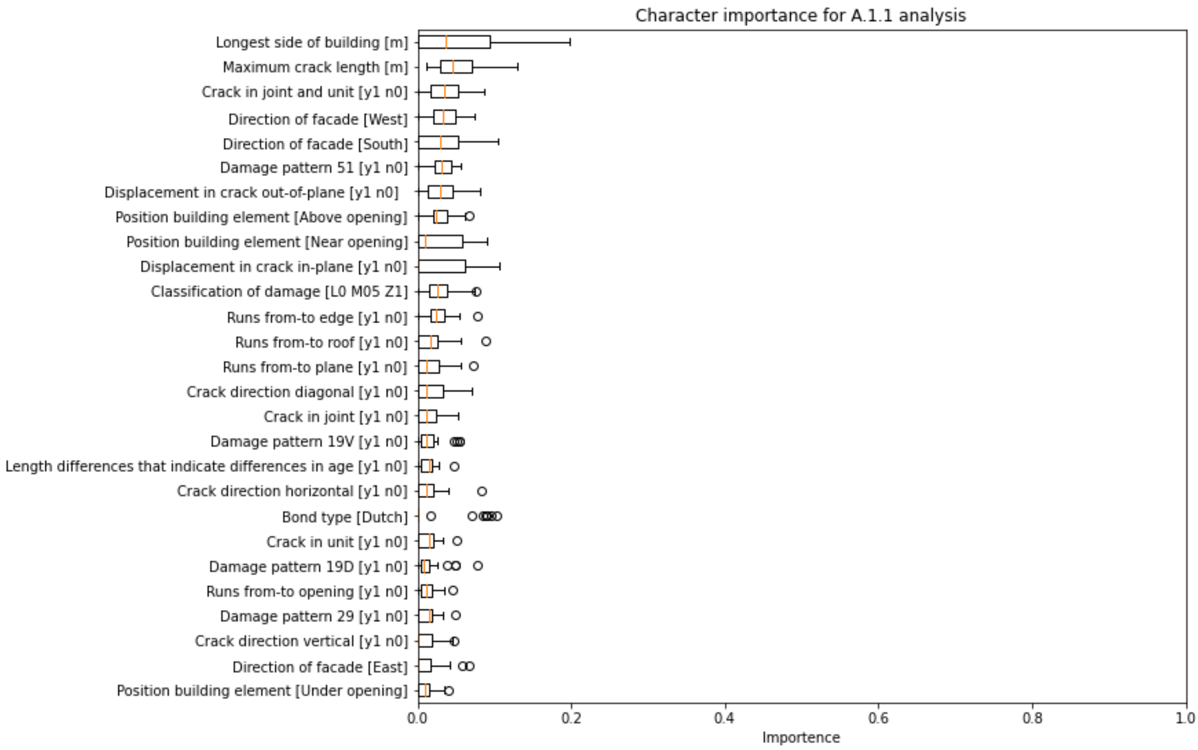


Figure 22 Characteristic importance according to decision tree classification to initial insufficient resistance to bear loads (A.1.1) with GSS cross-validation.

As mentioned in chapter 3.4.2, another way of evaluating the performance of a pattern is examining the importance of each characteristic. The characteristic importance is calculated in every validation group in a cross-validation. This range of importance for each characteristic can be visualized by a box-and-whisker plot, shortly boxplot. The red vertical line in a boxplot is the median, it is the point where 50% of the data is larger than the rest of the data. The box around this red line is the range in which 50% of the data is located. The T-shaped lines arising from the boxes are the minimum and maximum points of the data. Sometimes outliers are present in the data. These datapoints are made visual with small circles. (Dekking, Kaaikamp, Lopuhaä, & Meester, 2005, p. 236)

The boxplots for a basic classification run on A.1.1 (initial insufficient resistance to bear loads) are published in Figure 22. The importance is displayed on the horizontal x-axis. The vertical y-axis is a representation of all case characteristics with an average higher than 1%. From the figure, it can be concluded that a Dutch bond type in the façade certainly has the most influence on the outcome of the question whether damage has something to do with cause A.1.1 or not, according to a basic decision tree classification run. Note, the measure of the importance on the x-axis tells to what degree a characteristic is important for a pattern, but it does not say in what way a characteristic influences a pattern.

4.1.2 Recalculation of a characteristic importance and characteristic selection

As mentioned above, not all 200 characteristics shown in Figure 22 are analysed, but only those characteristics with more than 1% average importance. Per definition, the sum of all characteristics importance of one validation set should be equal to 100%. A second run has been done, with only those selected 1%+ characteristics, to ensure the sum of all importance's is still 1.0 as illustrated in Figure 22. In addition, this second run also improves the performance of the pattern recognition, because irrelevant characteristics are no longer part of the analysis.

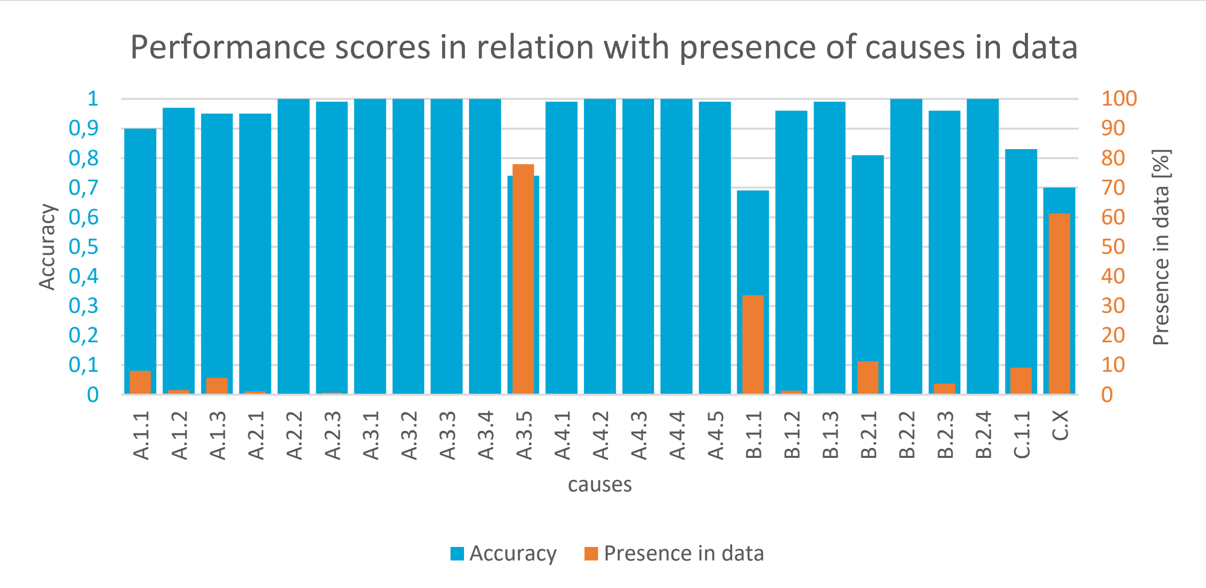


Figure 23 Relation between performance of imbalanced classes with presence of classes in data for LOGO cross-validation. Available damage causes are: Initial insufficient resistance to bear loads (A.1.1), Overloading due to vibrations by earthquakes (A.3.5), Initial hindered deformations (B.1.1), initial imposed deformations (B.2.1), Unequal settlements with equal loads (C.1.1) and Unequal loads (C.X)

Improved accuracy scores of all causes are gathered in Figure 23. They are plotted with blue bars with a scale on the left vertical y-axis. The improvements consist of a second run, as discussed earlier, and another cross-validation method is applied in order to calculate the accuracy. The applied cross-validation is the Leave-One-Group-Out method as explained in chapter 3. For example, the initial average accuracy of cause A.1.1 from Figure 20 and Figure 21 was 83%, and now it has been improved to 90% as can be read blue bar on the left of Figure 23.

Also in Figure 23, the presence of the causes are plotted in orange bars with a scale corresponding on the right vertical y-axis. The presence is defined as the relative number of times a cause category was mentioned as cause of damage. The sum of those causes is more than 100%, because damage generally falls under multiple cause categories. The presence in the data is calculated as follows:

$$P_y = \frac{n_y}{n_{known}} \times 100\% \tag{4-2}$$

In which:
P_y is the presence in the data of cause category y,
n_y is the number of times cause category y was mentioned as cause of damage,
n_{known} is the number of times a cause was determined in the database.

4.1.3 Class imbalance problem

Figure 23 indicates a relation between the performance of the pattern and the presence of a cause in the data. For example, patterns with very rare classes perform almost always correctly, like A.4.5 (Overloading due to accidental loads by lighting). However, more frequently mentioned cause classes perform relatively less accurately, like C.X (unequal settlements). This relation becomes more visible by including the excluded causes from chapter 2 (see Figure 23). Those causes perform very accurately, but the hypothesis is that the percentage of the times the pattern was off is the same as the percentage of presence in the data. The imbalance of the cause classes problem is the basis for the suspected relation in Figure 23 (see chapter 3).

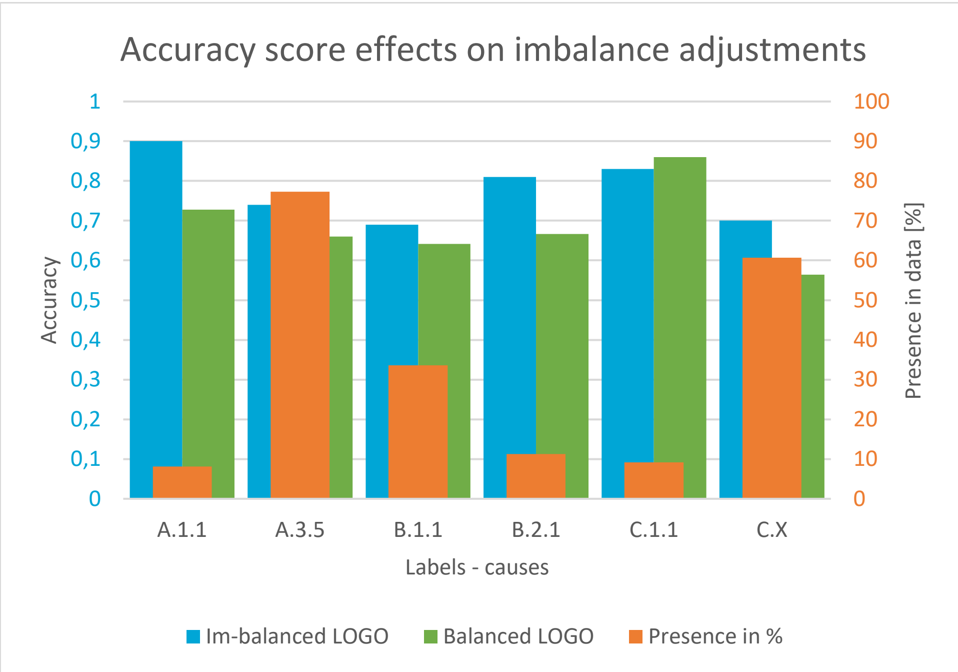


Figure 24 Visualisation of the effects due to imbalance adjustments. damage causes are: Initial insufficient resistance to bear loads (A.1.1), Overloading due to vibrations by earthquakes (A.3.5), Initial hindered deformations (B.1.1), initial imposed deformations (B.2.1), Unequal settlements with equal loads (C.1.1) and Unequal loads (C.X).

The consequences for the accuracy score due to the imbalance adjustments are shown in Figure 24. The former imbalanced accuracy scores are displayed in blue bars with a scale on the left vertical y-axis. The balanced accuracy scores are on the same axis with green bars. In orange it is displayed how many times a cause was mentioned as damage with a scale on the right vertical y-axis. In contrast with Figure 23, Figure 24 does not indicate a relation between the class frequency and the accuracy score.

Performance scores of each cross-validation fold on a C.1.1

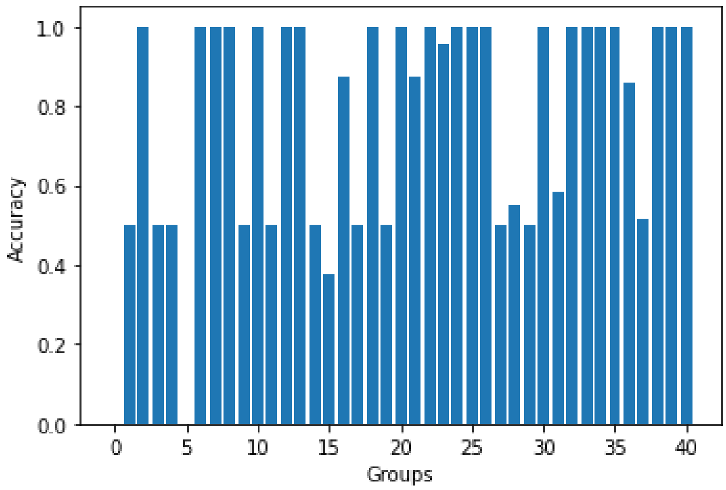


Figure 25 Intermediate results of accuracy scores after imbalance corrections for damage cause C.1.1 (unequal settlements with equal loads) with LOGO cross validation.

The formerly deployed cross-validation method was the Leave-One-Group-Out (LOGO) cross-validation. For that method, a validation set only consists of damage cases of one building. As a result, the variance of classes in a validation set is restricted. An example of this is shown in Figure 25. Here an accuracy score is plotted in blue bars, with a scale on the y-axis, for every LOGO cross-validation step for a classification pattern to cause C.1.1. In other words, the numbers on the x-axis represent the buildings in the data. Noticeable is that the variance in accuracy is limited to mostly 0.5 or 1.0. The hypothesis is that the buildings which score 0.5 accuracy only have one case of C.1.1 cause. If this single cause had not correctly been determined by the pattern, the accuracy will score 0.5 because of the imbalance adjustment. The problem with that is that the validation set size is limiting the preciseness of the accuracy score calculations.

The problem of limited validation set size can be solved by enlarging the validation set size. Therefore, a Group-Shuffle-Split (GSS) cross-validation has been carried out instead of a LOGO. In GSS, the number of buildings in one validation set can be controlled while the number of folds can be held equal to the number in the LOGO. Keeping the number of folds equal is done to provide a clear comparison of the consequences of different sizes of a validation set.

How many folds the GSS has to execute depends on the desired preciseness of an accuracy score. After consulting the committee, a preciseness of at least 10% to 20% is desired for accuracy calculations. This can be realized by including at least five to ten damage cases of the desired cause in one validation set. To rephrase this, at least five damage cases with damage cause *unequal settlements with equal loads* (C.1.1) have to be in one validation set when evaluating a pattern for damage cause C.1.1. However, it is desirable to have ten positive damage cases in one validation set. In the evaluation of pattern A.3.5 (earthquakes), for example, 18 of the 41 buildings did not comply with this minimum presence in the validation set rule. Eventually three buildings in one validation set were needed to meet the requirement of the desired preciseness.

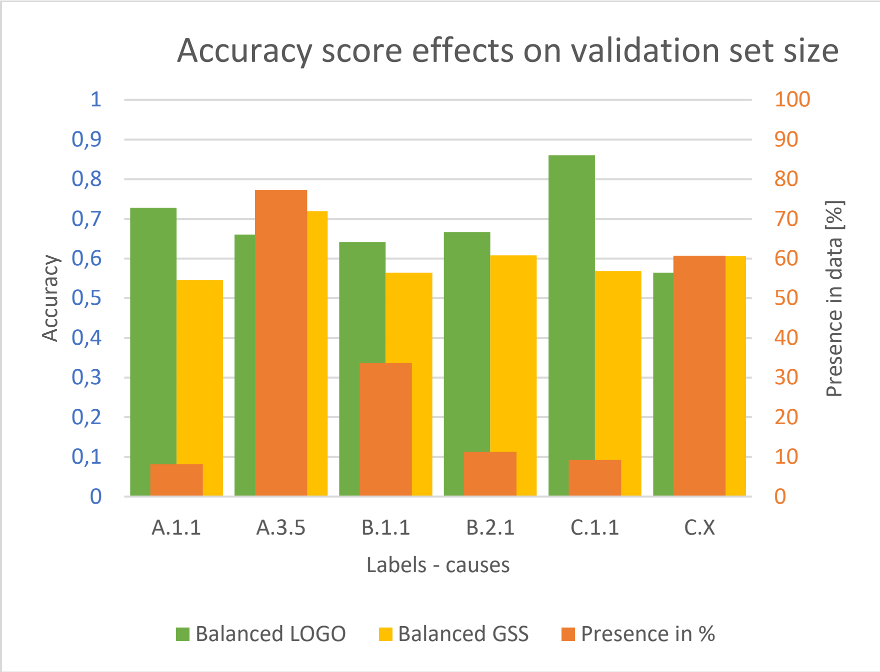


Figure 26 visualisation of the effects on the accuracy score due to larger validation set for preciseness requirements. damage causes are: Initial insufficient resistance to bear loads (A.1.1), Overloading due to vibrations by earthquakes (A.3.5), Initial hindered deformations (B.1.1), initial imposed deformations (B.2.1), Unequal settlements with equal loads (C.1.1) and Unequal loads (C.X).

Figure 26 shows the effects on the accuracy score when a validation set size is changed in order to meet the preciseness requirements of at least 10% to 20%. The green bars are the accuracy scores with insufficient validation size from Figure 24. The adjusted accuracy scores are plotted in yellow bars. Both green and yellow bars have a scale on the left vertical y-axis. The presence in the data is displayed in orange bars for comparison. It has a scale on the right vertical y-axis. According to Figure 26, all pattern causes have an accuracy around 60%. Also, a relation between the presence of a cause in the data and an accuracy score is not evident any more.

4.1.4 Is data size limiting the accuracy?

Issues discussed above have all in common that the size of the data is limiting the accuracy. For example, rare causes in the data base cannot be analysed with pattern recognition. Also, the required number of classes in a validation set is limiting the volume of available damage cases for training a pattern. For this reason, extra data are added to the data set.

Original dataset	Large dataset
1549 damage cases total	1830 damage cases total
985 (63.6%) known cases	1180 (64.4%) known cases
41 buildings	49 buildings
6 area	7 areas

The extra data consist of eight buildings with 281 damage in total. This increases the dataset with 18% to 1830 damage cases. Of those cases, it was possible to determine the cause in 1180 damage cases.

To put it in percentages, the relative amount of known damage causes increased with 0.8 percentage points to 64.4% (Figure 10).

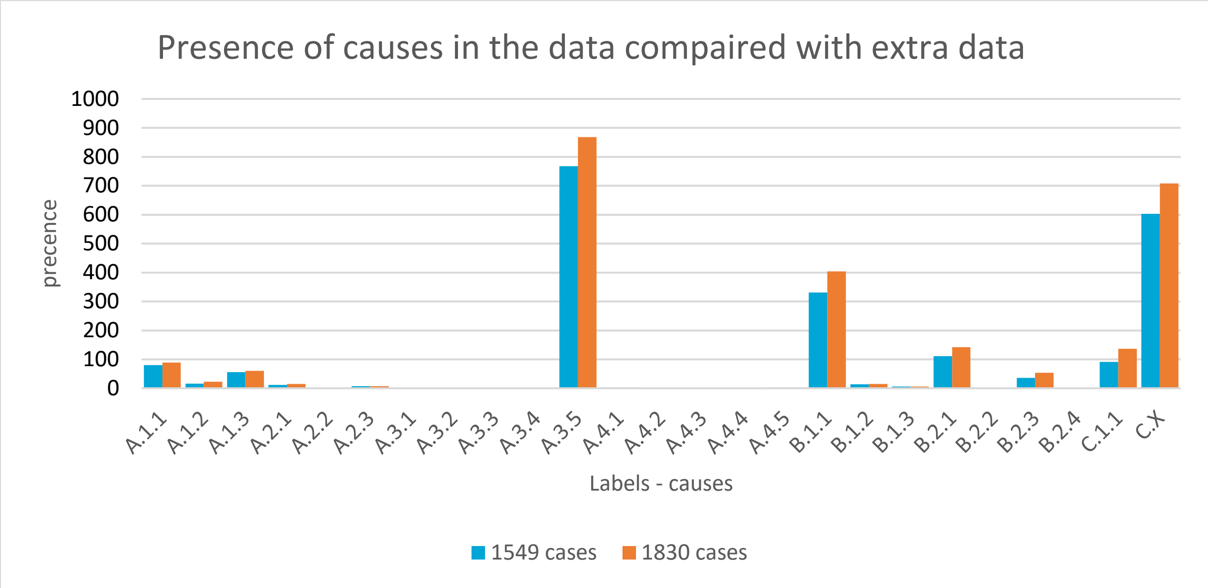


Figure 27 Visualisation of the effects on the absolute number of mentioned damage cause due to adding of extra cases. The damage causes are: Initial insufficient resistance to bear loads (A.1.1), Overloading due to vibrations by earthquakes (A.3.5), Initial hindered deformations (B.1.1), initial imposed deformations (B.2.1), Unequal settlements with equal loads (C.1.1) and Unequal loads (C.X).

The consequences of adding extra data to the number of times a cause is mentioned is shown in Figure 27. The horizontal x-axis represents all possible damage causes as discussed in chapter 2. The number of times a cause is present in the dataset is scaled on the vertical y-axis. blue bars originate display the original smaller dataset with 1549 damage cases. The orange bars are the larger extra data set with 1830 cases.

The guideline explained in chapter 2 is applied to check if more than the six damage causes mentioned at page 14 can be incorporated in the pattern recognition analysis. The earlier excluded damage causes do not exceed the requirement stated in that chapter. Therefore, it could be concluded from Figure 27 that rejected cases in chapter 2 do not have enough mentions for the pattern recognition analysis. For example, the number of times damage cause category B.2.3 was mentioned as damage cause is increased from 36 to 54 mentions, which is still below the guideline sated by Webb.

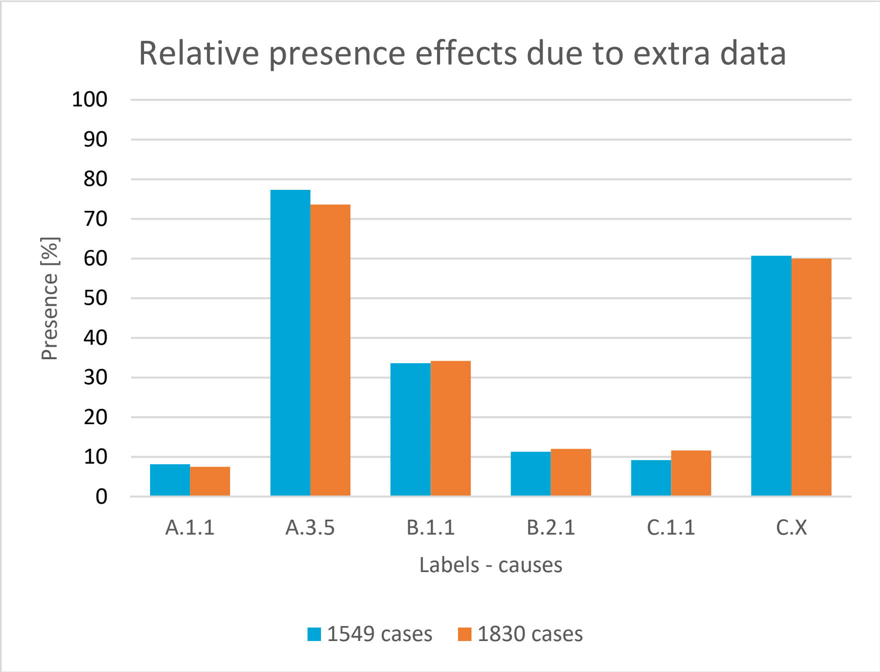


Figure 28 Visualisation of the effect on adding extra data, expressed in percentage of total data. The damage causes are: Initial insufficient resistance to bear loads (A.1.1), Overloading due to vibrations by earthquakes (A.3.5), Initial hindered deformations (B.1.1), initial imposed deformations (B.2.1), Unequal settlements with equal loads (C.1.1) and Unequal loads (C.X).

Figure 28 is based on the information of Figure 27, but the vertical scale is changed into relative scale in percentage and only the relevant causes are displayed. As Figure 27 indicates that the absolute number of these causes increases, Figure 28 demonstrates that the relative values not change significantly. That can indicate that the database is sufficient large. Adding extra data will probably lead to constant determination of damage cases. This may suggest that extra data will not lead radical changes in patterns.

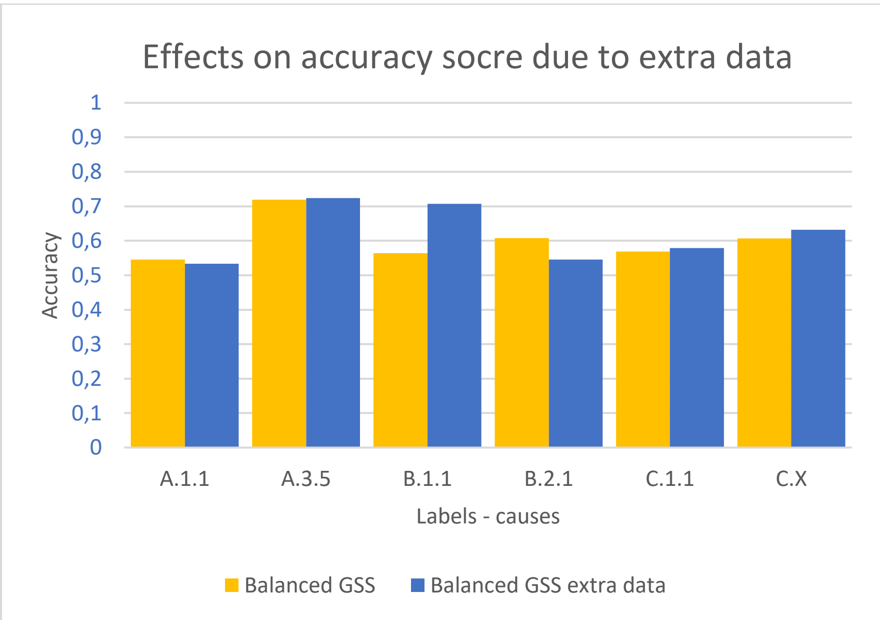


Figure 29 Bar plots on the effects on accuracy scores due to adding of extra data. The damage causes are: Initial insufficient resistance to bear loads (A.1.1), Overloading due to vibrations by earthquakes (A.3.5), Initial hindered deformations (B.1.1), initial imposed deformations (B.2.1), Unequal settlements with equal loads (C.1.1) and Unequal loads (C.X).

The consequences of adding extra data on the accuracy score is plotted at Figure 16 for the six remaining damage cause categories. The height of the yellow bars indicate the accuracy score of the original 1549 damage cases data set. These are the same values as the yellow bars in Figure 26. The blue bars display the accuracy score with a larger data set of 1830 damage cases.

As Figure 29 shows, most of the damage cause patterns did not improve significantly due to a larger data set. Only B.1.1 (initial hindered deformations) improved. Whereas B.2.1 (initial imposed deformations) decreased in accuracy. It is difficult to say that the improvement of B.1.1 is because it received 42 more damage cases than B.2.1. Damage cause A.3.5 (overloading due to vibrations by earthquakes) received 70 more damage cases than B.2.1 and its accuracy did not improve. Also in relative terms, B.1.1 was enlarged with 22.1%. Which is almost exact the average relative increase of all damage cause categories (23.7%).

Both damage causes (B.1.1 and B.2.1) have a relation with deformation. One is dealing with hindered deformations and the other with imposed deformations. For experts, it is sometimes difficult to choose between the two damage cause options. Therefore it is interesting to see that there is a significant difference in accuracy between the both options, because experts take in account the same indicators or characteristics to determine that deformations is the cause of damage. Apparently, the pattern recognition did that not. It found differences between the two deformation options.

The data size is limiting the accuracy for some damage cause categories. These categories, about deformations, are a difficult to determine damage cause. Therefore, more damage cases are needed to find a pattern in the data. However, the addition of extra data is a time consuming manual process. For that reason, more damage cases are not added to the data set.

4.1.5 Overlapping groups in the validation sets

However, the applied cross-validation method (GSS) will result in overlapping groups in the validation sets. This will occur when the validation set size is being made larger, while the number of folds is kept equal. To accommodate that, GSS will incorporate one group (one building) in multiple validation sets (Table 4). This repeatedly validating with one building is not desirable. Therefore, a comparison is made with another cross-validation method where a group is applied only once in a validation set. This cross-validation method is the Group-K-Fold method (Table 5). In GKF, the number of folds can be set to what a cross-validation has to execute. Since this will reduce the amount of validation sets, it consequently enlarges the number of buildings (groups) in one validation set (Pedregosa, et al., 2011).

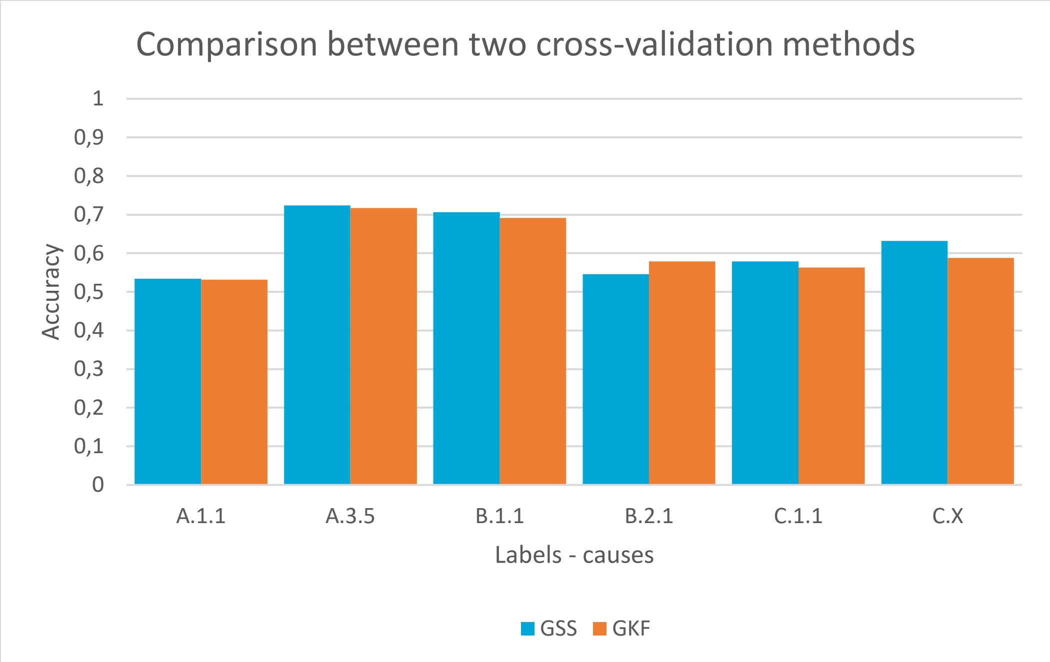


Figure 30 End results of classification with decision tree method, where GSS and GKF are compared. The damage causes are: Initial insufficient resistance to bear loads (A.1.1), Overloading due to vibrations by earthquakes (A.3.5), Initial hindered deformations (B.1.1), initial imposed deformations (B.2.1), Unequal settlements with equal loads (C.1.1) and Unequal loads (C.X).

The consequences of overlapping groups in validation sets is investigated with Figure 30. The blue bars are the GSS extra data accuracy scores from Figure 29. Overlapping groups are present in these values. The orange bars display the accuracy scores with a GKF cross-validation method. Overlapping groups are not present in these results. The accuracy score does not significantly change between the two methods. Therefore, it could be concluded that overlapping groups have no influence on the accuracy score with these research data.

Cross-validation settings for GSS and GKF

	A.1.1	**A.3.5**	**B.1.1**	**B.2.1**	**C.1.1**	**C.X**
Buildings in one validation set, according to GSS	11	3	5	8	15	3
Buildings in one validation according to GKF	6-8	1-3	1-4	4-6	11-13	1-3
Number of folds according to GSS	49	49	49	49	49	49
Number of folds according to GKF	7	26	17	9	4	23

Table 6 Validation size set and number of folds setup for end results of decision tree classification

The numbers in the table above show which setting of the applied cross-validation was needed for the pattern recognition analysis. For example, GKF cross-validation needed one to three buildings in one validation set with 26 folds of an analysis on pattern A.3.5 (earthquakes). The number of buildings in one validation set are a range in GKF, because sometimes one building contains enough damage cases to fulfil the requirements. It could be concluded that the number of buildings in one validation set does not differ a lot between the two methods. However, the GKF method needed on average two to three groups fewer than a GSS analysis. The difference in the number of folds is at least 50% between the two.

4.1.6 Applied stopping rules

As mentioned in chapter 3, stopping rules are an imported tool to prevent overfitting of a pattern. These rules have constantly been applied in all pattern calculation runs mentioned above in order to find the most optimal accuracy score of each individual pattern. Since discussion of stopping rule settings at each intermediate calculation step would not contribute to the readability, only the settings of the end result are discussed now.

Multiple splitting rules are possible. However, only two of them were applied in decision tree settings, because they provided the best results. Also, these rules are easy to understand, which will help to apply them correctly. The adopted rules are the *maximum depth* and the *minimum sample leaf*. Practically speaking, *minimum split size* has a similar effect on a pattern as minimum sample leaf. Therefore, both rules were not applied simultaneously. Moreover, the minimum sample leaf was better to be controlled in order to get the best optimal accuracy score.

According to Pedregosa et al., finding the optimal setting for stopping rules is a trial and error process. The following procedure tends to be the most practical method. The initial default step starts with a pattern without any stopping rules. This will provide the depth of a decision tree. This depth will be needed as initial setting for a maximum depth. The starting point for minimum sample leaf is one. The second step in this process is decrease the maximum depth with intervals of five. For example, a starting point of a maximum depth of 25 will be decreased to 20, 15, 10 and 5. The accuracy score should increase until a maximum depth is near an optimum. Further decreasing the maximum depth will also decrease the accuracy score. The next step is exploring values around the decision tree depth found in the previous step. This exploration implies is finetuning the optimal settings in order to reach the highest accuracy score. Ones the optimal maximum depth has been established, the same procedure is followed with the minimum sample leaf while a maximum depth is set on an earlier found optimal value. However, the minimum sample leaf is now increased instead of decreased in intervals of five (eq. 1-5-10-15-20-25-etc.). Optimizing a minimum sample leaf first and optimizing the maximum depth afterwards would not make a difference to the end results. The best stopping rules settings are shown in the Appendix 5: Results classification Runs – Decision tree.

4.2 Results classification analysis

The results of the classification analysis with decision tree algorithm are presented here. The complete results are placed in the appendix. The presentation of the results will start with the accuracy scores. It is followed by one decision trees of an interesting damage cause.

4.2.1 Calculating the accuracy of a damage cause pattern

In the paragraphs above, the average accuracy was presented of the cross-validation folds. As mentioned in chapter 3, an standalone data set is necessary to obtain an independent performance score. This test set was gathered by applying the extra data set from paragraph 4.1.4 as a test set. It is an independent data set from a different area in the province of Groningen. The train set consists of the original data set of 1549 damage cases. The accuracy of each damage cause pattern is presented in the figure below (Figure 31).

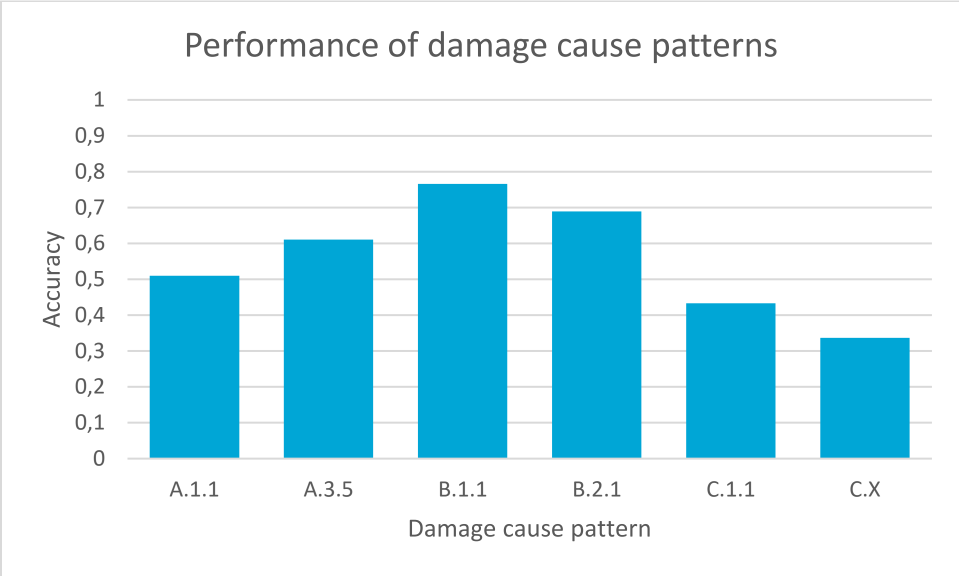


Figure 31 End results of the damage cause patterns in terms of their performance score in accuracy form. The damage causes are: Initial insufficient resistance to bear loads (A.1.1), overloading due to vibrations by earthquakes (A.3.5), initial hindered deformation (B.1.1), initial imposed deformations (B.2.1), unequal settlements with equal loads (C.1.1) and unequal settlements (C.X).

4.2.2 Decision tree results

All decision tree results are placed in the appendix. However, one interesting causes are placed here. This is a decision tree to determine if an earthquake was the cause of damage.

Classification decision tree for overloading due to vibrations by earthquakes (A.3.5)

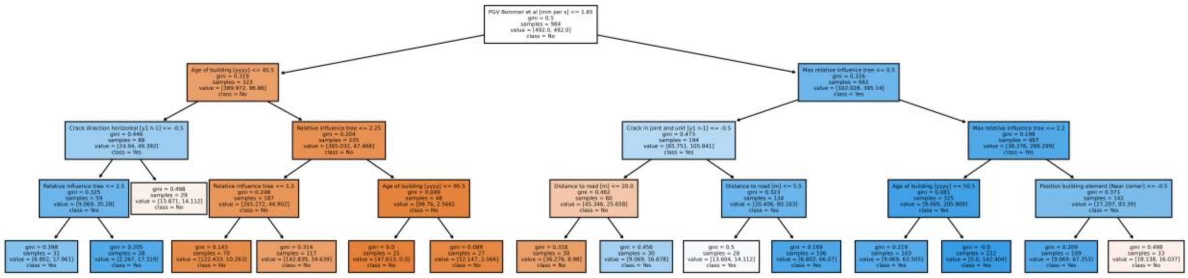


Figure 32 A classification decision tree for damage cause 'overloading due to vibrations by earthquakes (A.3.5)

5 Results of regression analysis – Decision tree

The technical attributability of damage is determined with regression analysis. In this chapter, the decision tree method is applied as a pattern recognition algorithm. It first starts with a description on how is searched for a regression model with decision tree pattern recognition. The chapter ends with the discussion of a obtained result.

5.1 The process of obtaining a regression decision tree with patter recognition

Similarly as classification analysis, a pattern is searched for each damage cause individually (see Figure 19). The reasons for this are the same as previously mentioned arguments. In addition to these argumentations, determining technical attributability offers more options than classification analysis. This would make a regression decision tree to all damage causes at once impractically large and difficult to interpret.

Not all suitable damage cases are complete. Some damage causes could not be included in the analysis because of various reasons. This results in damage cases where the technical attributability does not sum up to 100%.

5.1.1 Outliers in the regression analysis

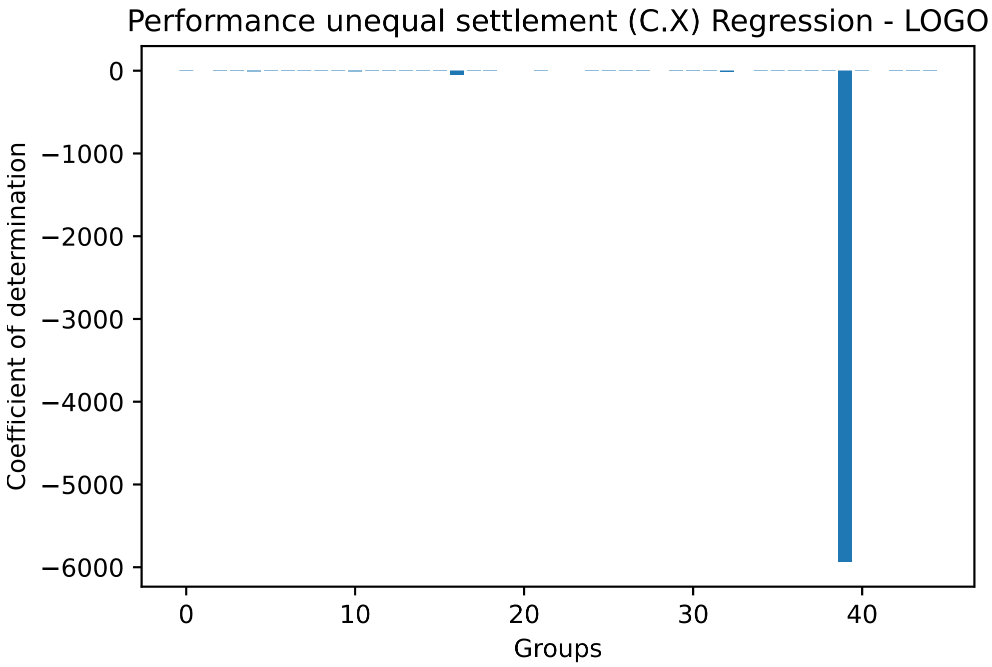


Figure 33 Initial cross-validation regression result for C.X (damage cause unequal settlement). Cross-validation done with Leave One Group Out strategy. Each group is presented on the horizontal x-axis and there performance score, coefficient of determination (R^2) is placed on the vertical y-axis. Feature selection is based on the results of the classification analysis.

The labels in this regression analysis differ a lot compared to classification analysis. This will lead, among other things, to a relatively low performance score in terms of the coefficient of determination (see chapter 3). An example of this is shown in Figure 33. Here, the performance score of one building differs significantly compared with the other buildings. This type of outlier buildings is excluded from further regression pattern recognition, because they will pollute any result. The buildings that are excluded and their explanatory graphs are placed in the appendix.

5.1.2 Characteristic selection based on classification results

Once the outliers have been excluded, it is still necessary to direct the pattern recognition to reach a positive coefficient of determination (R^2). This is done by reducing the number of characteristics. One option of reducing characteristics is applying the results of the classification pattern recognition. A selection of characteristics was made based on the importance on the classification results. This selected list of characteristics is implemented for regression analysis. In classification analysis, only characteristics were integrated with an importance of 1% or higher for the second run. This threshold was set arbitrarily and is now further investigated. Together with the stopping rules, this threshold value too is optimised in order to get the best accuracy score of a pattern. It is expected that this further investigation will falsify characteristics that induce noise in a pattern.

5.1.3 Equalizing train set and validation set

The characteristics selection based on the classification results were not sufficient to reach positive R^2 values. One problem could be that the domain of the train set differs a lot from the domain of the validation set. This could result in a pattern that works in the domain of the train set, but not in the domain of the validation set. For example, if the validation set has only values of high technical attributability and the train set lacks those values, the pattern recognition was not able to produce a pattern for the damage cases in the validation set. This could lead to negative values of R^2.

To assess and solve this problem, the average of the technical attributability of the train set is compared with the average of the technical attributability of the validation set. Enlarging the validation set size in the cross-validation will equalize those values. For this reason, the applied cross validation is a Group K-Fold method (GKF). In this method, the number of folds can be set between two and the number of buildings in the available data. The size of the validation set becomes larger when the number of folds becomes smaller. The averages of both sets come closer together when the number of folds is reduced.

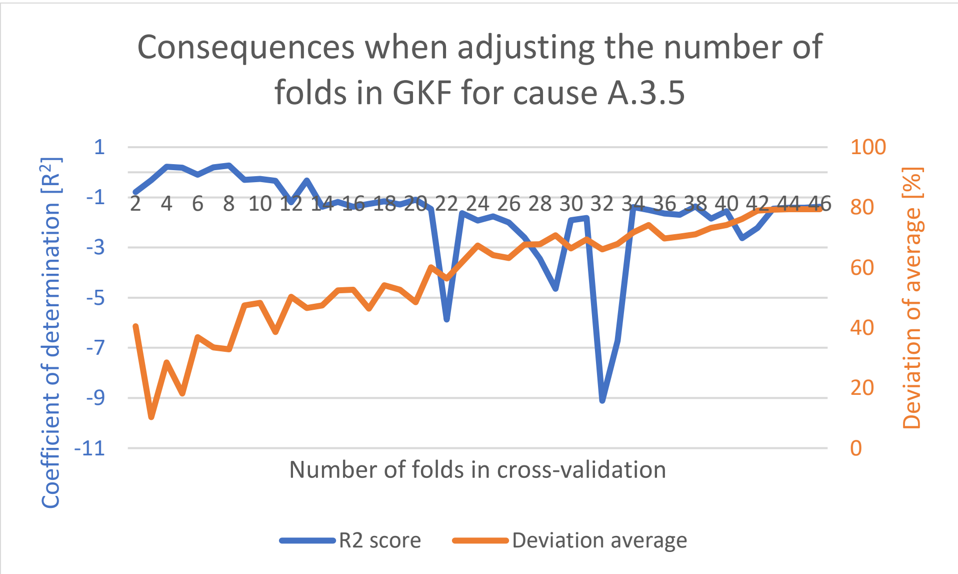


Figure 34 Results for regression on cause 'vibration due to earthquakes' (A.3.5) with decision tree method. GKF cross-validation is applied with a range of all possible fold settings.

An example of such fold adjustment is shown in Figure 34. It is a visualisation of the effect on the coefficient of determination and the average deviation. The number of folds in a GKF cross-validation is presented on the horizontal x-axis. It starts at the minimum limit of 2 folds and ends at the maximum available groups of 46 folds. On the left vertical y-axis, the coefficient of determination is given in blue.

The deviation of the average between the validation and train sets is shown in orange on the right vertical y-axis. For example, a regression pattern on A.3.5 (overloading due to vibrations by earthquakes) with a GKF of eight folds has an R^2 of 0.27 and the deviation of average between validation and train set is 32.8%. It could be concluded that increasing the number of folds, decreases the R^2 score and increases the deviation of average.

The deviation of average is calculated according to the following equation:

$$\mathrm{DfA}_i = \frac{|\bar{y}_{\text{validation}} - \bar{y}_{\text{train}}|}{\bar{y}_{\text{train}}} \times 100\% \qquad (5\text{-}1)$$

In which:

$\bar{y}_{\text{validation}}$ is the average of the technical attributability a validation set,
$\bar{y}_{\text{train}}$ is the average of the technical attributability a train set,

Equation 5-1 calculates the deviation of every fold in the cross-validation individually. For example, if 3 folds were executed, there are 3 different validation sets and therefore 3 different deviation of averages. The mean of those three deviations of averages was calculated in order to come to one presentable value. So correctly speaking, the values presented in Figure 34 are the means of the deviation of average of each cross-validation step.

The choice of the number of folds for the pattern recognition analysis is arbitrary by nature, but it is based on the lowest possible deviation of average in combination with an acceptable coefficient of determination. Another objective that was taken in consideration is the intention to choose the highest possible number of folds. This is desirable to obtain reliable statistical values of the results, such as the calculation of the main and the standard deviation. The motivation of the chosen number for folds per damage cause are placed in the appendix with their supporting figures like Figure 34.

Figure 35 Consequences of optimizing the number of folds in the GKF cross-validation on the coefficient of determination.

The regression results with the decision tree algorithm are presented in Figure 35. The results are in the form of coefficient of determination (R^2). This is an indication of how well the data can be described with a decision tree. Only positive values are relevant, with a maximum of 1.0. The analysed causes are: Initial insufficient resistance to bear loads (A.1.1), overloading due to vibrations by earthquakes (A.3.5), initial hindered deformations (B.1.1), initial imposed deformations (B.2.1), unequal settlements with equal loads (C.1.1) and unequal settlements (C.X). The blue bars indicate the results without adjustment of the number of folds applied in the GKF cross-validation. The regression results with optimized folds are presented with orange bars. For example, the initial result for A.3.5 was a R^2 of -1.4 and it was +0.28 after folds optimalisation. It could be concluded that reducing the number of folds, in order to equalize the train and validation sets, has a positive effect on the R^2. However, the overall result after folds adaption is not sufficient, since almost all R^2 results are still negative.

5.1.4 Characteristic selection based on civil engineering insight

The previous results were based on the characteristics selected by the classification algorithm. However, not all selected characteristics were logical in civil engineering practice. Those groundless characteristics are removed in the following regression pattern recognition analysis based on the falsification done in the appendix of chapter 2.

Figure 36 Results regression tree pattern recognition with GKF cross-validation. It is a visualisation of the consequences when the characteristics are selected based on civil engineering insight.

Figure 36 shows the consequences of selecting characteristics based on civil engineering insight. It is an updated version of Figure 35, where the grey bars represent the R^2 score for pattern recognition with explainable characteristics for each damage cause. For example, the R^2 score for A.1.1 is improved from -0.22 to -0.16. This is the first time in this thesis that the results of the pattern recognition are steered by a logical selection of characteristics in the expectation to obtain better results. Whereas some cause categories improve their R^2, others show a decrease of their score. This was not expected. A smart selection of characteristics was meant to help the pattern recognition. It was not to worsen the score. More detailed results, including a visualisation of each decision tree, is placed in the appendix.

Running a pattern recognition is a labour-intensive process. Each variable is explored in order to optimize the result. The stopping rules must be re-analysed each time an adaptation of the data is done. Since it is expected that decision tree algorithms are not well qualified for regression analyses and these methods are more labour-intensive than linear regression, the results now produced are declared as end results of regression with decision tree algorithm.

5.2 *Relation between vibration damage by earthquakes and Peak Ground Velocity*

It is interesting to know what the relation is between the Peak Ground Velocity (PGV) due to earthquakes and the damage caused by the vibration of earthquakes. Firstly, it is interesting because the provided data was produced with the intention to define this relation. Also, it is a check on the expected maximum reachable result. This is because, during the assessment of the damage cases of the data, a pattern was included to determine the damage caused by earthquakes (van Staalduinen, Terwel, & Rots, 2018). If the pattern recognition works, it should be able to find this pattern in the data.

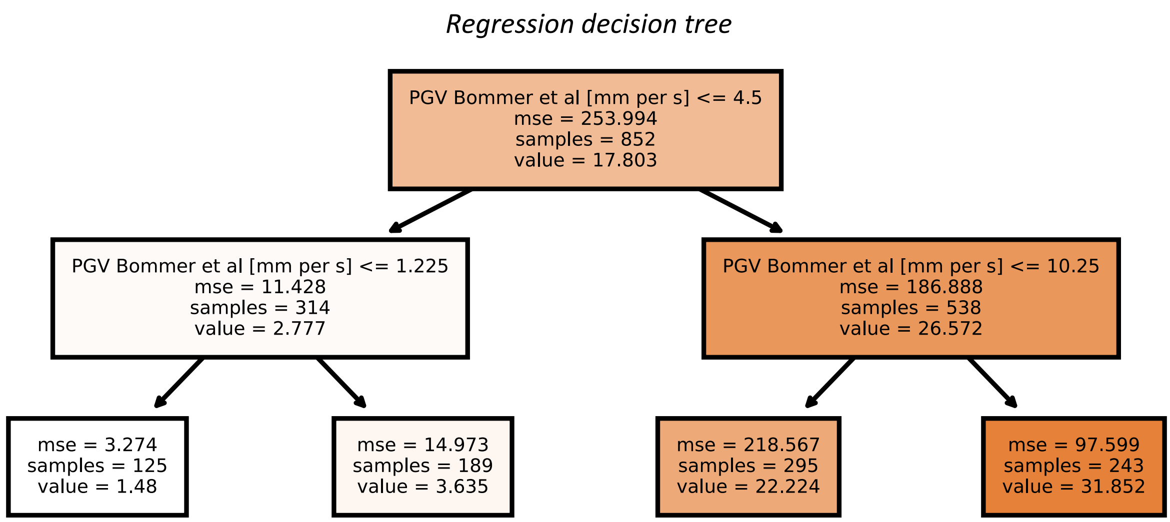


Figure 37 Decision tree pattern for A.3.5 with only PGV as characteristic. Cross-validation GKF: 8 folds, 1830 cases. Stopping rules: max depth 1, min sample leave 1. Performance score with average of cross-validation.

The R^2 of this regression pattern of damage cause category A.3.5 is 0.39[1]. Simplified, 39% of the data can be explained with a decision tree pattern. This pattern is shown in Figure 37. The pattern starts at the top with the tree's root. The first line in this box is a splitting rule to which damage cases are compared. If the splitting is correct in comparison with the evaluating damage case, the pattern proceeds to the left. In case of a damage case exceeding the splitting rule, the pattern proceeds to the right. Eventually, a pattern ends at the bottom with a leaf. The technical attributability assigned to a damage case is placed at the last line of this leaf as *value*. For example, if a damage case has a PGV of 3 mm/s, the technical attributability of this pattern is 3.6%. The number of sample cases a leaf was trained on is presented in the middle row as *samples*. Not all trainings samples can correspond exactly to the assigned value. The MSE (mean square value) explains how the samples in the leaf relate to this value (Dekking, Kaaikamp, Lopuhaä, & Meester, 2005, p. 305). The MSE value has a domain between 0 and infinity. Better relations are indicated with smaller values (Deval, 2020).

[1] Calculated as average of cross-validation on 1830 damage cases. An analysis on 1549 damage cases and 280 damage cases as test set would result in a R^2 of -23.9.

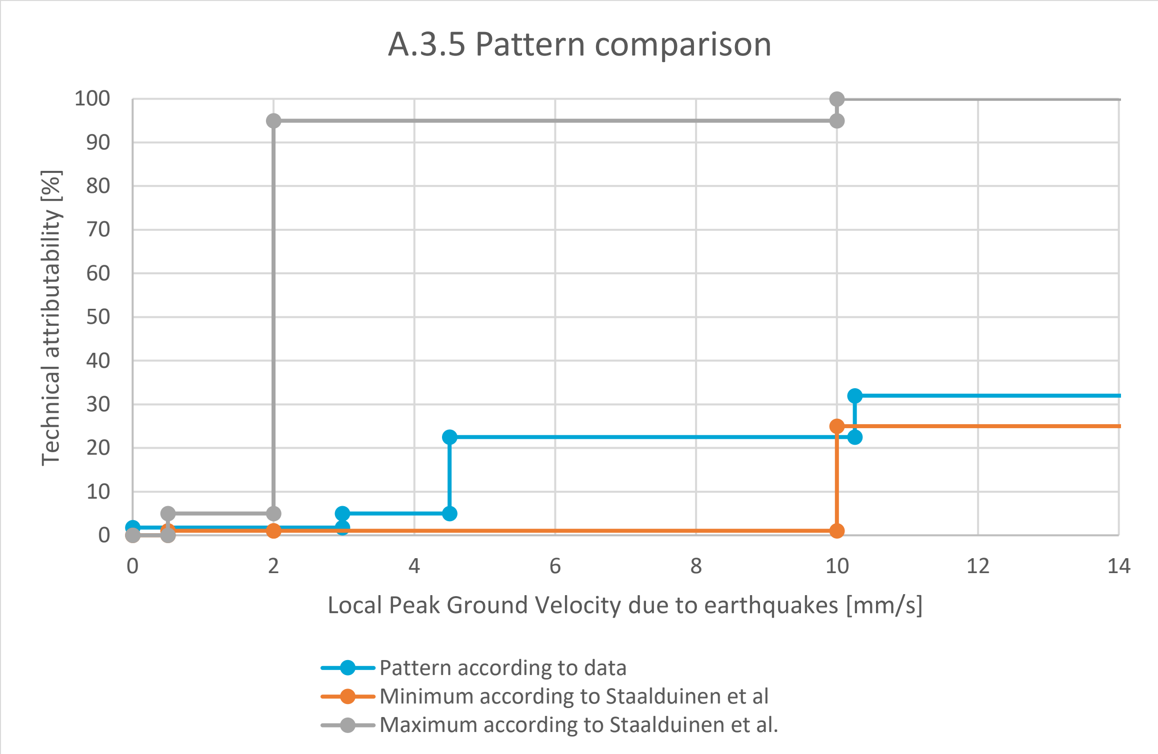


Figure 38 A comparison between the induced pattern in the data by Staalduinen et al. and the produced pattern by decision tree pattern recognition.

A comparison between the decision tree in Figure 37 and the induced pattern by Staalduinen et al is visualized in Figure 38. The blue line is the pattern from Figure 37. The pattern of Staalduinen et al consists of a range in which the technical attributability should be located. The minimum of this domain is represented with the orange line. The maximum possible technical attributability according to Staalduinen et al is shown with the grey line. For example, if damage appeared in a building which was loaded by an earthquake vibration of 4.0 mm/s (PGV), the produced decision tree pattern predicts that the technical attributability is 5%. Whereas staalduinenen et al gives a range between 1% and 95%.

Some conclusions could be made from this result. First, a coefficient of determination around the 0.40 is probably the highest reachable score with the given data. It is a result of a predetermined pattern within the data. It is expected that other relations do not reach higher than 0.40, because they do not have a predetermined pattern in the data. Secondly, the pattern produced by the decision tree seems to follow the Staalduinen et al pattern. It has almost the same thresholds. For example, both patterns have a threshold around the 10 mm/s. Also, it follows the same direction. The technical attributability of both patterns increase with an increase of PGV. However, the produced decision tree pattern seems to follow the lower limit of the Staalduinen et al pattern more closely than the higher limit. Consequently, the higher technical attributability causes will not be found with the produced pattern. This explains the high MSE values for the higher PGV leaves in contrast with the lower PGV leaves.

The analysis above have been done with the large dataset to incorporate the most possible data. That will obtain the best possible result. Therefore, there was no test set available to calculate an independent performance score. The presented scores are the average scores of the cross-validation.

6 Results of regression analysis – Linear Regression

The results of a linear regression are presented in this chapter. It first starts with the process on how these patterns are generated. After that, the results are presented and discussed on if they can be applied in practice.

6.1 The process of obtaining a linear regression model with pattern recognition

In using linear regression, a tool is made which can predict the technical attributability of damage causes. The process to produce these results differs from the earlier two chapters above. The former methods were aimed at avoiding interference with the results. This should prevent steering the results into generally accepted paths. This not interfering of results may lead to new insights in the civil forensic engineering field. However, following the algorithm closely is only possible if the produced patterns perform sufficiently. That was not the case for linear regression. Any result without characteristic selection from outside resulted in a (large) negative coefficient of determination. For example, a characteristic selection based on a result of a well performing classification decision tree of A.3.5 would end up in a coefficient of determination of -0.13.

Positive values of R^2 could only be reached if the characteristic selection was closely monitored and dictated. Consequently, some decisions in this process are made arbitrarily based on the understanding of the damage cause, the insight of the characteristics and the feedback of the algorithm. Therefore, it is difficult to report on this arbitrary process, since the result depends on which decision was made at which moment. Also, the large range of possibilities, variables and options makes the presented result one result out of many. One could easily reproduce the presented results, but also come up with other results that perform better or worse.

Eventually, it was possible to produce three patterns with positive R^2 values. These three patterns will be discussed later in this chapter. The chapter starts with a description of the process that was executed before the linear regression could begin. Some adjustments to the data had to be made in order to process the data easily by means of the algorithm.

6.1.1 Normalisation of characteristics

Linear regression fits characteristic values in a linear equation. While doing this, it treats the values as they are. This has some influence on the assigned slope of a characteristic. For example, if the construction year (1950) and the length of a crack (0.8 m) both have the same influence on a technical attributability, it will result in different assigned slopes. See the next calculation example:

$$\begin{aligned}\hat{y} &= m_{\text{construction year}} \times x_{\text{construction year}} + m_{\text{crack length}} \times x_{\text{crack length}} + b\\ &= m_{\text{construction year}} \times 1950 + m_{\text{crack length}} \times 0.8 + b\\ &= 0.000513 \times 1950 + 1.25 \times 0.8 + b\\ &= 1 + 1 + b\end{aligned}$$

Although both characteristics have the same influence, it cannot be interpreted from the characteristic slopes. Therefore, every characteristic is scaled between 0.0 and 1.0. This will lead to characteristic slopes that can be compared to each other. The minimum value of a characteristic is transformed to 0 and the maximum value of a characteristic is converted to 1.0. If the construction year has, for example, a range between 1800 and 2000, it would be normalized to a range between 0.0 and 1.0 with the following formula:

$$x_{\text{feature construction year}} = 0.005 \times x_{\text{characteristic construction year}} + -9$$

In which:

$x_{\text{feature construction year}}$ is the normalized characteristic ‘construction year’,
$x_{\text{characteristic construction year}}$ is the real value of characteristic ‘construction year’.

The slope (u) and intercept point (c) of this normalization can be calculated for each characteristic with the following equations:

$$u = \frac{1}{\max - \min} \tag{6-1}$$

$$c = -\max \times \frac{1}{\min - \max} \tag{6-2}$$

In which:

$\max$ is the maximum of a characteristic,
$\min$ is the minimum of a characteristic.

6.1.2 Equitizing of characteristics

Some characteristics have values that are generally close to each other with a few exceptions. An example of is the crack length (see Figure 39 left). Most lengths are less than 4 meters, but a few cracks are more than 12 meters long. When characteristics like this are normalized, the measured values of this characteristic are located in a limited part of the 0.0 to 1.0 scale. Training a linear regression algorithm with such data is difficult, because of the limited variance of a characteristic. The algorithm has to work hard if this value is important for the calculation of technical attributability. It will result in characteristic slopes which are relatively large in comparison with the expected influence.

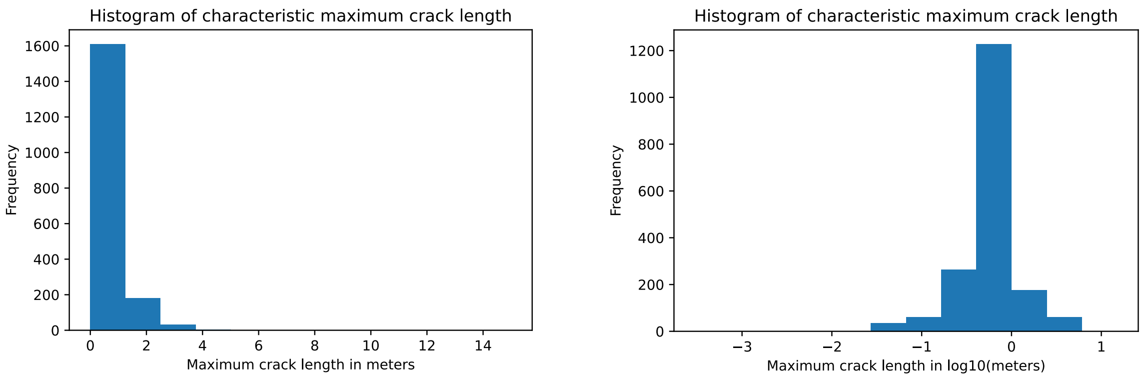


Figure 39 Histograms of the characteristic ‘maximum crack length’. Left figure is without processing. Right figure is with log scale processing. Bin width is based on the work of Sturges and Amer [1926] (Dekking, Kaaikamp, Lopuhaä, & Meester, 2005, p. 211)

The concentrated datapoints characteristics are converted in a log10 scale before they are normalized and processed by the pattern recognition algorithm. An example of pre-scaling data is shown in Figure 39. Consequently, the values in a characteristic can be handled better by linear regression. Characteristics that undergo this log10 transformation are:

- Area of building [m^2]
- Longest side of building [m]
- Shortest side of building [m]
- Distance to road [m]
- PGV Bommer et al [mm/s]
- Maximum crack width[mm]
- Maximum crack length [m]

6.1.3 Building outliers removal

Similarly as in decision tree regression analysis (chapter 5), some outlier buildings had to be removed in order to reach a positive coefficient of determination values. the leave-one-group-out cross-validation method has been applied to indicate buildings that stand out of the rest. However, there are some downsides of this cross-validation method. First, some damage cause categories only occur once or two times in a building. The calculation of the R^2 in these cases is either not possible or it could be far off. Secondly, one outlier damage case in a building could provoke the removal of other damage cases in the same building, since all building cases are removed. Lastly, the removal of outlier buildings is stopped when all other buildings are visual on an algorithm performance graph. this is an arbitrarily chosen point in this thesis. It is possible to continue to remove buildings with the smallest R^2 value until the point where a few buildings which are more or less the same remain. However, that is not the intended goal of this thesis. The aim is to produce a tool which can be applied on as many buildings as possible. Which buildings are excluded and their explanatory graphs is recorded in the appendix.

6.2 Results linear regression

Linear regression pattern recognition performs better with a reduced number of characteristics. The experience with the algorithm hints at an optimum between ten and five characteristics. The process of finding a positive R^2 value starts with selecting a long list of interesting characteristics in relation with the chosen damage cause category. Based on the initial performance of the result, characteristics with less indicated influence on the technical attributability are removed. These characteristics are marked with a slope around zero, by the algorithm. Also, characteristics which are at unexplainable locations are removed. For example, if characteristic *age of the building* has a negative slope for cause *aging*, it is removed from the list. A new run is done each time a characteristic is removed. A characteristic is placed back in the list if the coefficient of determination did not improve. This trial and error process is continued until the best possible result is reached.

The analysis have been done with the large dataset to incorporate the most possible data. That will obtain the best possible result. Therefore, there was no test set available to calculate an independent performance score. The presented scores are the average scores of the cross-validation.

It was not possible to produce a pattern for cause category A.1.1 (Initial insufficient resistance to bear loads). The only logical related characteristic is the possibility of snow accumulation. However, no damage case of A.1.1 was indicated as the result of rain or snow build up on the roof.

All other results are incorporated in the discussion, because it will increase the readability of the report.

7 Discussion

The results above are discussed in this chapter. It starts with the interpretation of the classification results from chapter 4. The regression results are interpreted from paragraph 7.7. These discussion of the results are structured on each damage cause. The chapter ends with a disclaimer. There, uncertainties of the results are explained and discussed.

7.1 *Interpretation of the classification results to cause Initial insufficient resistance to bear loads (A.1.1)*

The found relations by the pattern recognition with damage cause initial insufficient resistance to bear loads is discussed in this sub-chapter. The relations are considered on their reliability, if they are explainable and if it is meaningful for practice.

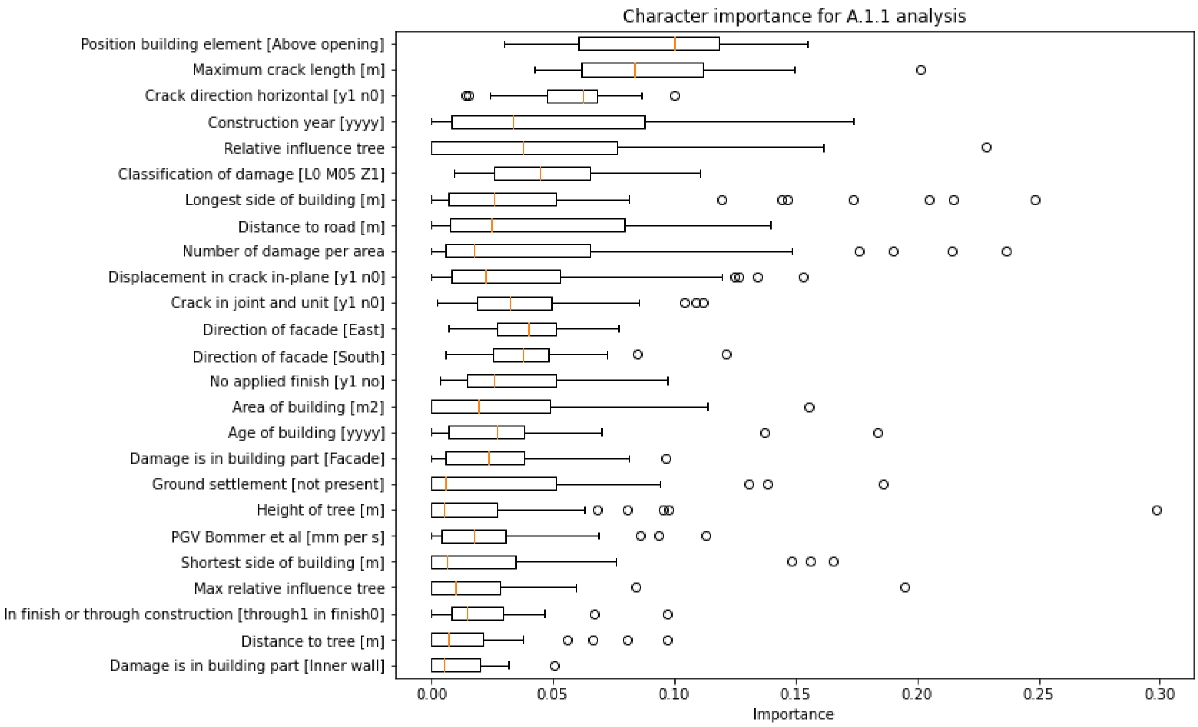


Figure 40 Importance of each characteristic for a classification analysis on damage cause initial insufficient resistance to bear loads, represented in boxplot figures. Obtained by GSS cross-validation with large dataset of 1830 damage cases.

For damage cause *Initial insufficient resistance to bear loads* (A.1.1), it was possible to find a relation with an accuracy score of 51% (Figure 31). It was difficult for the algorithm to find a simple pattern. Therefore, the produced decision tree is the largest of all other damage causes. As a result, drawing conclusions form individual characteristics is difficult. Also, the lack of theoretical key characteristics from literature for this cause does not help. This is represented by the pattern as well. A single characteristic is not donated significantly more important than others.

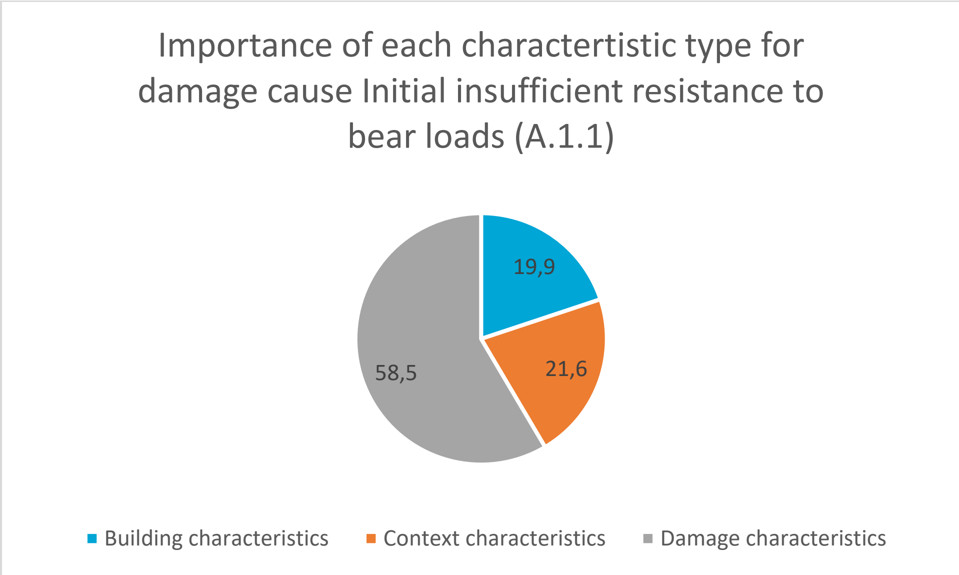


Figure 41 Characteristic types distributed by their importance for classification determination of damage cause A.1.1. Obtained by GSS cross-validation with large dataset of 1830 damage cases.

The importance of damage characteristics is 58.5% (Figure 41). Of those characteristics, damage located above openings was the most important (Figure 40). Which is explainable in general, because stresses are higher above openings. If the higher stresses are not sufficiently incorporated in the design, it may result in damage due to insufficient resistance. An example of an inadequate construction is the missing of lintels above openings in older buildings (Salet & Bruurs, 2017, p. 25). The decision tree indeed indicates that the presence of damage located above openings is the result of damage cause category A.1.1.

Damage above openings

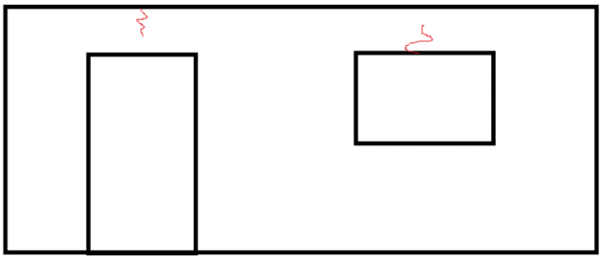


Figure 42 Damage above openings is an indication of initial insufficient resistance to bear loads (A.1.1), according to pattern recognition analysis.

Building characteristics and context characteristics are each important for around 20% in the relation (Figure 41). Most important building characteristic is the year of construction. That could be explained as older building may not be designed or built sufficient enough. An example of this is mentioned in the text above, where lintels were not applied by default in older construction methods. This relation is demonstrated in the pattern where the age of building is older than 127.5 years, the damage is most likely the cause of initial insufficient resistance to bear loads.

Age of building

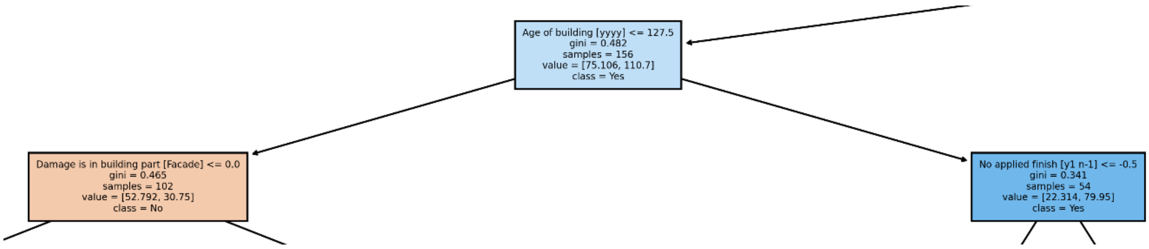


Figure 43 Internal node in classification decision tree of cause A.1.1. Inhere, building characteristics 'Age of building' has a threshold of 127.8 years. This threshold is exceeded if a building is older than 127.5 years. In that case, the pattern proceeds to the right.

The distance to a road seems to be an important context characteristic. However, it is an unexplainable characteristic in the context of this damage cause. Vibration of road traffic cannot lead to initial insufficient resistance. This is also displayed in the patterns, as there seems to be no clear relation with distance to roads (Figure 44). Damage with roads further than 5.5 m away are not likely caused by A.1.1, while roads closer than 9.5 m are most likely caused by A.1.1. Another unexplainable characteristic is the earthquake load on the building. However, the influence of this characteristic is small.

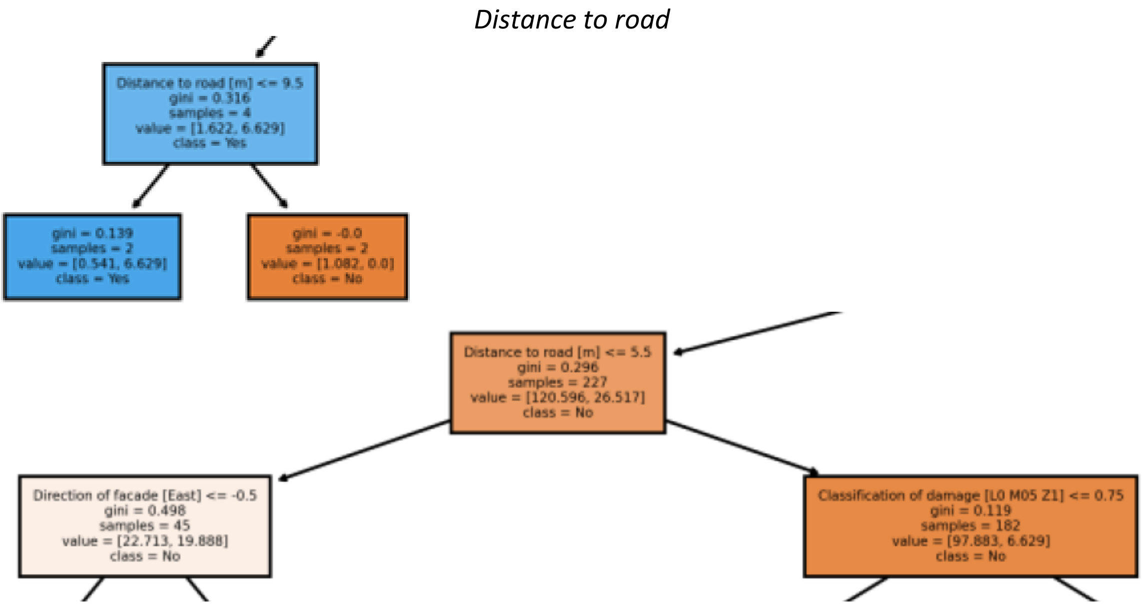


Figure 44 Internal nodes in classification decision tree of cause A.1.1. where building characteristics 'distance to road' has a multiple mentions and thresholds. Obtained by GSS cross-validation with large dataset of 1830 damage cases.

Concluding, the pattern performs with 51% accuracy. Wherein the found relations do not give a clear classification relation between case characteristics and damage cause initial insufficient resistance to bear loads. The decision tree is large and defuse. It does not provide a structured relation, but it was only able to make a pattern by finding the exceptions. That result in a large decision tree without a clear pattern. There are some explainable damage and building characteristics. However, the found context related characteristics are not explainable. Therefore, the produced relations cannot meaningfully to be applied in a tool for practice use.

7.2 *Interpretation of the classification result to cause Overloading due to vibrations by earthquakes (A.3.5)*

The found relations by the pattern recognition with damage cause *overloading due to vibrations by earthquakes* is discussed in this sub-chapter. The relations are considered on their reliability if they are explainable and if it is meaningful for practice.

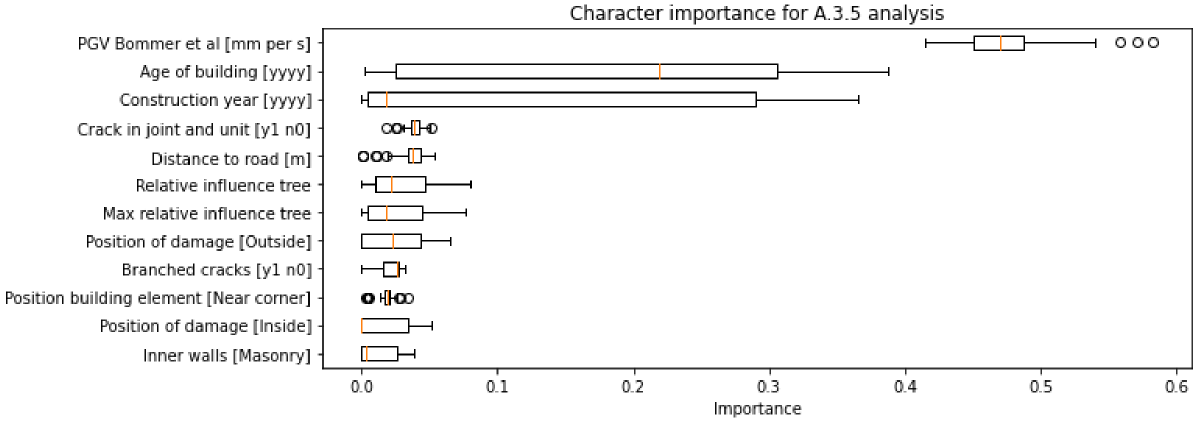


Figure 45 Importance of each characteristic for a classification analysis on damage cause 'overloading due to vibrations by earthquakes', represented in boxplot figures. Obtained by GSS cross-validation with large dataset of 1830 damage cases.

For damage cause category *overloading due to vibrations by earthquakes (A.3.5)* it was possible to find a classification relation with an accuracy score of 61% (Figure 31). The found relations shows a clear picture of characteristics which are important for the determination if a damage is partly caused by earthquakes. The three most important characteristics are the earthquake load in peak ground velocity, the age of the building and the construction year (Figure 45).

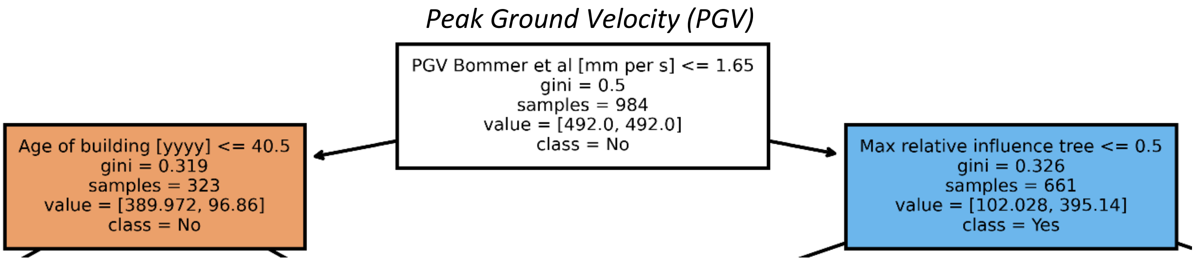


Figure 46 Start node in classification decision tree of cause A.3.5, where context characteristic 'PGV Bommer et al' has a threshold of 1.65 mm/s. Damage in buildings where this earthquake load was smaller as the threshold will proceed to the left child.

The pattern recognition donates a peak ground velocity (PGV) of 1.65 mm/s as threshold. Damage with values lower than this threshold are most likely not caused by overloading due to vibrations of earthquakes. In practice, a threshold of 2 mm/s is applied in the investigation if damage is caused by vibrations of earthquakes (Instituut Mijnbouwschade Groningen, 2021). It is an internationally accepted limit and this threshold is intensively supported by literature.

Although the produced threshold is 17.5% of the threshold applied in Groningen, it is still a defendable result. First, the range of occurred PGV's is much larger, with a maximum of around 40mm/s. Secondly, in Groningen a chance of exceedance of 1% is enforced, while the pattern recognition has shown a PGV with a chance of exceedance of 25%. Lastly, the threshold of Groningen is based on literature, while the value of 1.65 mm/s is partly based on real damage cases. Also, the applied threshold in literature is not a hard threshold in the provided data. A relation between a PGV value and earthquake damage had a limit of 0.5 mm/s, where any damage with lower PGV values could not be attributed to earthquakes. Damage with PGV values between 0.5 and 2.0 mm/s stood a very small chance of being caused by earthquakes, according to the pre-determined relation (van Staalduinen, Terwel, & Rots, 2018, p. 151). Therefore, a lower found threshold by the pattern recognition is possible or acceptable.

The PGV threshold is low, compared to the literature. However, it does not mean that the produced pattern excludes any earthquake damage with PGV values lower as 1.65 mm/s. To be ahead of the text below, if the building is younger than 40.5 years, it is most likely that the damage is partly caused by earthquakes.

Threshold performance

	PGV under threshold	PGV above threshold
Not earthquake damage	Correct	Incorrect
Earthquake damage	Incorrect	Correct

Table 7 How the correctness of a prediction is assessed.

An analysis has been made on the relation between a PGV threshold and the correctness of a prediction with this threshold. Table 7 shows how the correctness of a prediction is assessed. A prediction is correct if the PGV of a damage case was under a set threshold while an earthquake was not the cause of damage. Also, a prediction is correct if the PGV of a damage case was above as set threshold while an earthquake was the cause of damage.

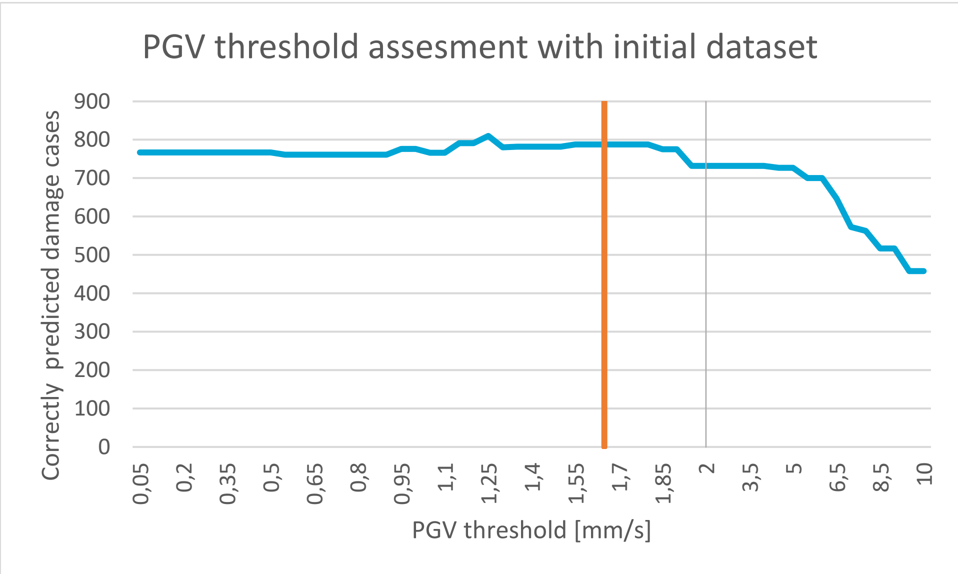


Figure 47 Threshold assessment with absolute data. The blue line indicates how many damage cases have been correctly predicted with a PGV threshold of the x-axis. The orange line is the PGV threshold advised by the pattern recognition. The vertical grey line shows a change in interval of the x-axis.

The set threshold can be treated as a parameter. That is done in Figure 47. For example, if the PGV threshold is set to 2.0 mm/s, 732 damage cases were assessed correctly. Also, if the threshold was set to 0.2 mm/s, 767 damage cases were assessed correctly. From this figure could be concluded that it does not matter if the PGV threshold is somewhere between 0.0 mm/s and 5.0 mm/s.

However, the applied dataset is an imbalanced dataset. Therefore, conclusions from this figure would be wrong. The number of damage cases with earthquake damage in this data set is 767. That is equal to the number of correct predicted damage cases with a threshold of 0.5 mm/s or smaller. The other 217 unrelated earthquake damage cases are incorrect predicted with a threshold of 0.5mm/s. When increasing the threshold, the number of correctly predicted damage cases will increase, because more unrelated earthquake damage will be predicted correctly. However, the increase of this prediction size is relatively small, because the size of unrelated earthquake damage is small compared with the earthquake damage cases. Further increasing the set threshold will lead to an decrease of correct predicted damage cases, because more earthquake damage is predicted incorrectly (see Table 7). This decrease is relatively large, because the number of earthquake damage in the dataset is large.

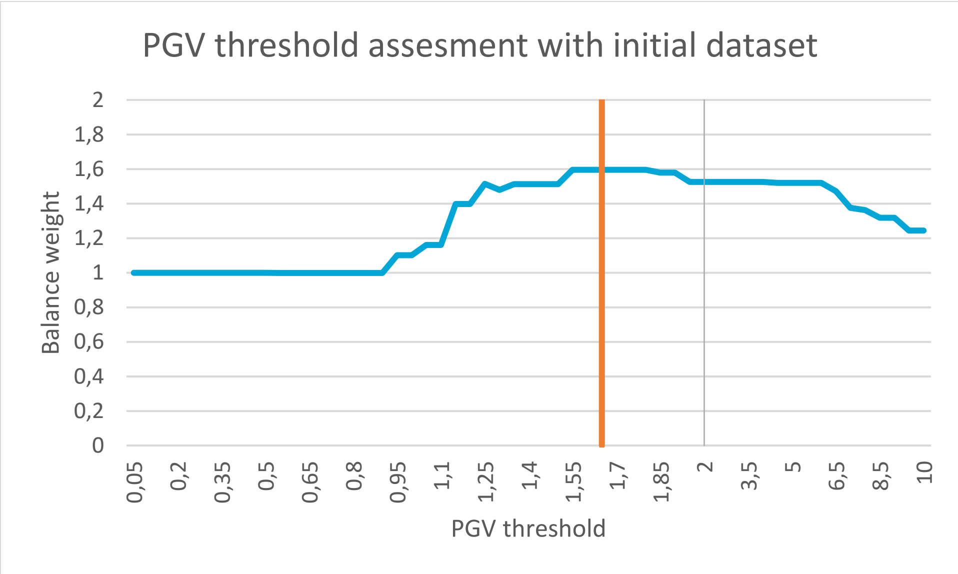


Figure 48 Threshold assessment with balanced data. The blue line indicates how many damage cases have been correctly predicted with a PGV threshold of the x-axis. The orange line is the PGV threshold advised by the pattern recognition. The vertical grey line shows a change in interval of the x-axis.

The applied dataset is balanced in Figure 48. All earthquake damage has a balanced weight of 1.0. This is represented in the horizontal line at 1.0 between threshold 0.0 mm/s and 0.9 mm/s. All other unrelated earthquake damage has a balanced weight of 1.0. This means that if all damage cases are predicted correctly (earthquake damage + not earthquake damage = 1 + 1 = 2), the blue line in Figure 48 would be at balance weight level 2.0.

Now it does become clear that it does matter if the PGV threshold is 0.0 mm/s or 5.0 mm/s. The threshold that results in the most correct predictions is at 1.65 mm/s. This optimal threshold is equal as what the pattern recognition advised.

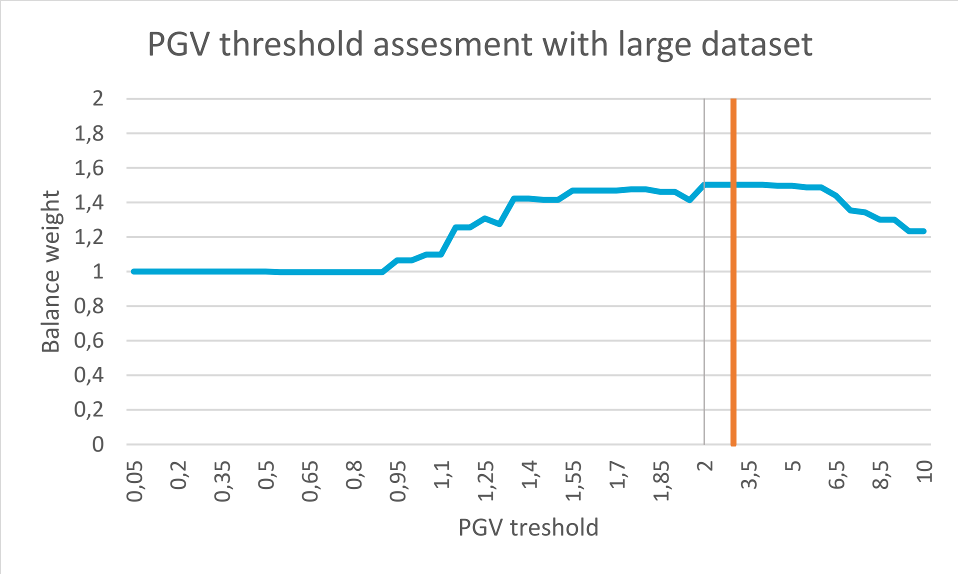


Figure 49 Threshold assessment with balanced large dataset. The blue line indicates how many damage cases have been correctly predicted with a PGV threshold of the x-axis. The orange line is the PGV threshold advised by the pattern recognition. The vertical grey line shows a change in interval of the x-axis.

A reason why the presented threshold of 1.65 mm/s differs from the in practice accepted 2.0 mm/s could be that the applied dataset in Figure 48 is too small. The consequences of enlarging the dataset is shown in Figure 49. In here, the validation set is incorporated in the initial dataset to enlarge the data with almost 20%. This will result in a new optimal threshold of almost 3.0 mm/s (2.975 mm/s to be exact). Therefore, enlarging the dataset has an significant influence on the threshold. It could be concluded that the applied initial dataset is too small to get an acceptable result.

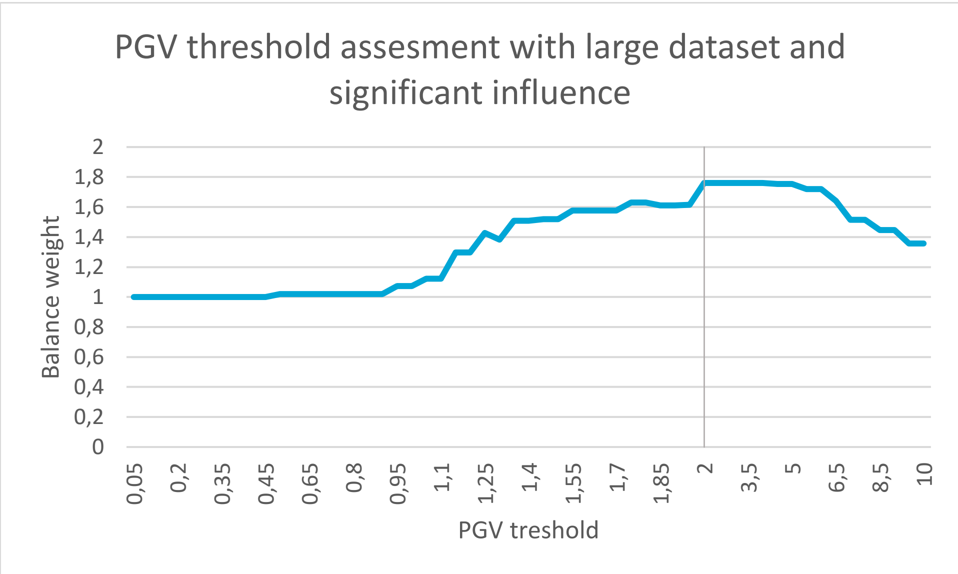


Figure 50 Threshold assessment with balanced large dataset and removed trigger function of earthquakes. The blue line indicates how many damage cases have been correctly predicted with a PGV threshold of the x-axis. The vertical grey line shows a change in interval of the x-axis.

Another reason why the found threshold differs from the with literature supported threshold of 2.0 mm/s could be that the trigger function of earthquakes could occur with PGV values below the threshold of 2.0 mm/s. The trigger function of earthquakes is that damage can occur without vibrations, but the earthquake initiates the occurrence of damage. In other words, damage would appear without the vibrations. However, an earthquake gives the last push what starts the process of exceeding the resistance stress levels. That trigger function is represented in the data with a technical attributability of 1%. Damage with this level of technical attributability are replaced by none related earthquake damage in Figure 50. Removing the trigger function has an influence on the height of the balance weight, but it does not change the optimal threshold. In other words, the threshold will almost stay the same, but it will be more successful in splitting the data for earthquake damage classification.

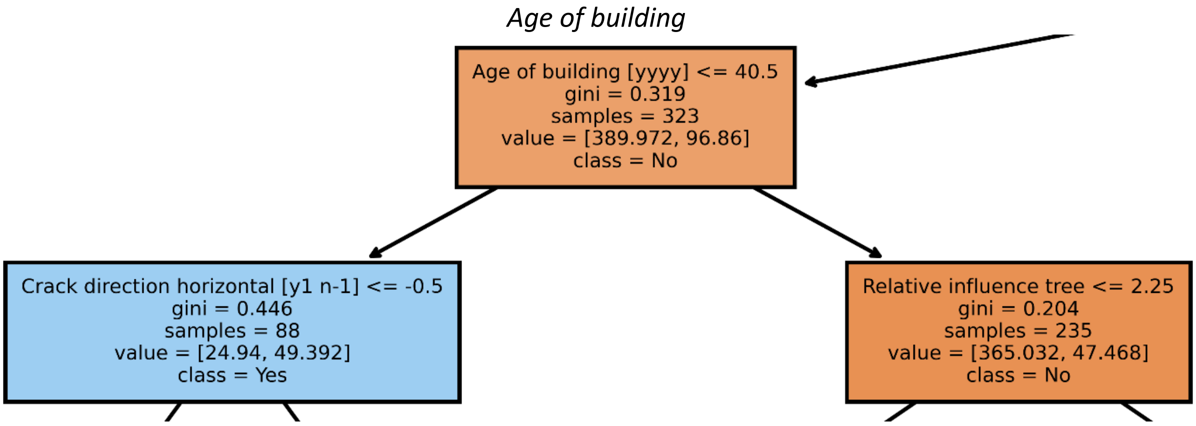


Figure 51 Internal node in classification decision tree of cause A.3.5, where building characteristic 'age of building' has a threshold of 40.5 years. Damage in a building younger as 58 years fulfil the splitting rule. Therefore the pattern proceeds to the left child.

The other two important characteristics are closely related to each other. The age of the building is also includes the moment of first detection of damage. Damage in buildings older than 40.5 years (building age <= 40.5) is most likely not caused by earthquakes (Figure 51). This is explainable because older buildings generally have lighter timber floors. This will make older buildings less vulnerable for induced inertia forces of earthquakes. Also, there is an possibility that the age of a building is an indication for other damage cause category, like damage occurring because of initial insufficient resistance to bear loads or autonomous settlements.

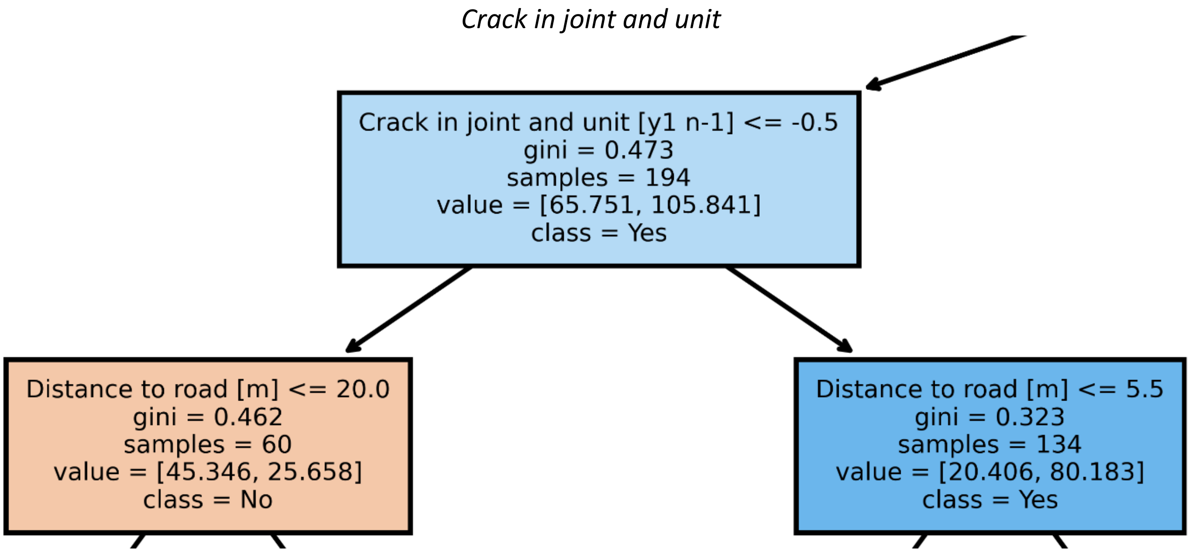


Figure 52 Internal node in classification decision tree of cause A.3.5, where damage characteristics 'crack in joint and unit' splits the data. If a crack is not in both joint and unit, it has a value of -1. Therefore it fulfils the threshold and the pattern proceeds to the left child.

An important damage characteristic is damage located in joint and unit. For example, if there is a crack in the stone and mortar material of a masonry wall. It is explainable that vibrations loosen the connection between those two materials. However, hard evidence for this relation is missing in the literature. According to the found pattern, cracks that runs through both the joint and the unit material are most likely caused partly by earthquakes. This is the case when the PGV is more than 2.975 mm/s. The likelihood of earthquake damage is even higher if the PGV is more as 7.0 mm/s. The relation is too specific to find this in literature. However, the found pattern does not seem to be an exception, because it splits 15% of the evaluated damage cases (Figure 52).

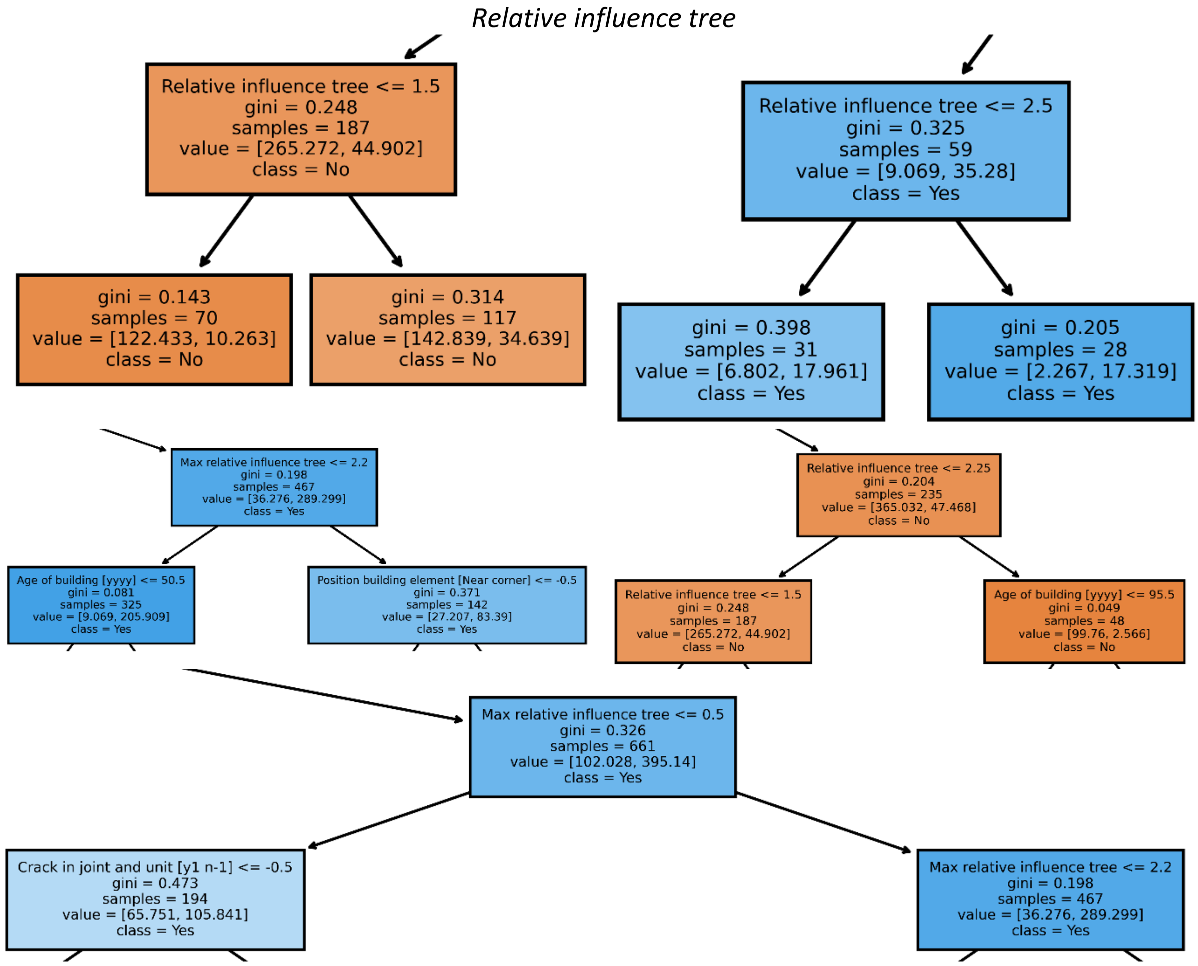


Figure 53 Internal nodes in classification decision tree of cause A.3.5, where context characteristic 'influence tree' as multiple mentions and thresholds.

The classification result has some unexplainable context characteristics. First, it indicates an influence of trees on earthquake damage. Relations between trees and earthquakes are not known in the literature. Splitting rules in the produced decision tree only provide certainty of the result, but it will not decide the outcome of the classification. There are patterns where the presence of trees will result in earthquake damage (Figure 53). However, the pattern is less sure when large trees are located close to buildings. A suggestion could be that trees are reinforcing the ground, like trees on hills prevent mudslides. In some cases, this could reduce the earthquake vibrations in buildings. In other cases, it could enlarge earthquake vibrations in buildings. This dynamic behaviour of soil is a complicated science field.

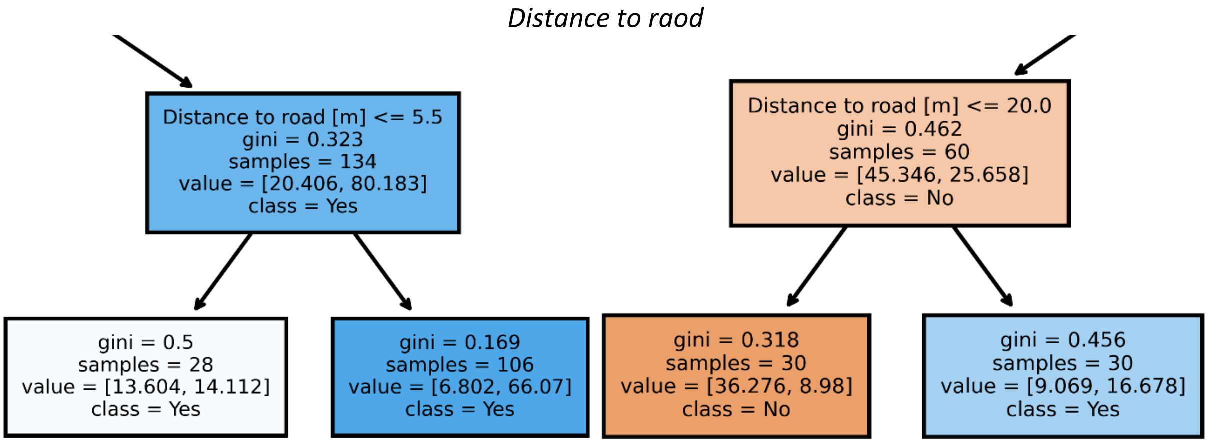


Figure 54 Internal nodes in classification decision tree of cause A.3.5, where building characteristic 'Distance to road' has multiple mentions and thresholds.

The second context characteristic which is difficult to explain is the distance between a building and a road. Roads can cause vibrations like earthquakes. Similar to tree influence, the distance to a road will provide more certainty about the answer, rather than it has a decisive influence on it (Figure 54 left). Which is explainable, as vibrations from roads and vibrations from earthquakes are difficult to keep separate from each other. When the road is located further away, the chance is higher that earthquakes partly caused the damage.

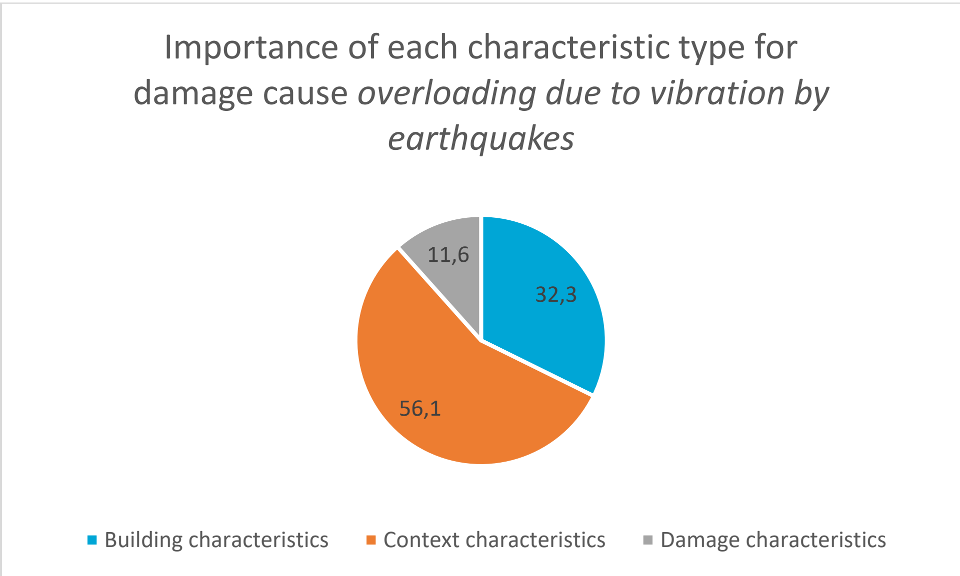


Figure 55 Characteristic types distributed by their importance of classification analysis of damage cause A.3.5. Obtained by GSS cross-validation with large dataset of 1830 damage cases.

Remarkable is that damage characteristics are less important (11.6%) in relation to the other characteristic types (building characteristics 32.3% and context characteristics 56.1%). Therefore the layout and location of damage are less important for the determination of earthquake causes than other characteristics. This seems logical as the trigger function of earthquakes can result in a large variation of damage layout and locations.

Concluding, the decision tree is able to assign earthquake damage in an reliable and explainable way. Therefore, the found relation is meaningful to be applied in practice. However, cautious use is advised as some characteristics are irrelevant for earthquake damage.

7.3 *Interpretation of the classification result to cause Initial hindered deformations (B.1.1)*

The found relations by the pattern recognition with damage cause *initial hindered deformations (B.1.1)* is discussed in this sub-chapter. The relations are considered on their reliability, if they are explainable and if it is meaningful for practice.

Position of damage is inside

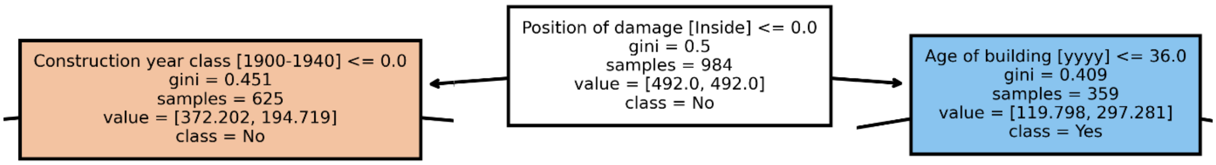


Figure 56 Internal node in classification decision tree of cause B.1.1, where damage characteristic 'position of damage (inside)' splits the data. If a crack is located inside a building, it has a value of 1.0. Therefore it does not fulfil the threshold and the pattern proceeds to the right child.

The Initial hindered deformations (B.1.1) classification decision tree, with an accuracy of 77% (Figure 31), has a successful starting point (the root) where the data are split by damage characteristic *position of damage*. Damage located on the inside of a building is most likely caused by B.1.1 (Figure 56). Because it is such a successful split, it is the most important characteristic in this relation (Figure 57). That hindered deformation occurs at locations inside a building is explainable. Materials and construction layouts, such as the construction of inner walls and applied finishes, are less homogeneous inside buildings. However, hindered deformations may still occur at outside locations.

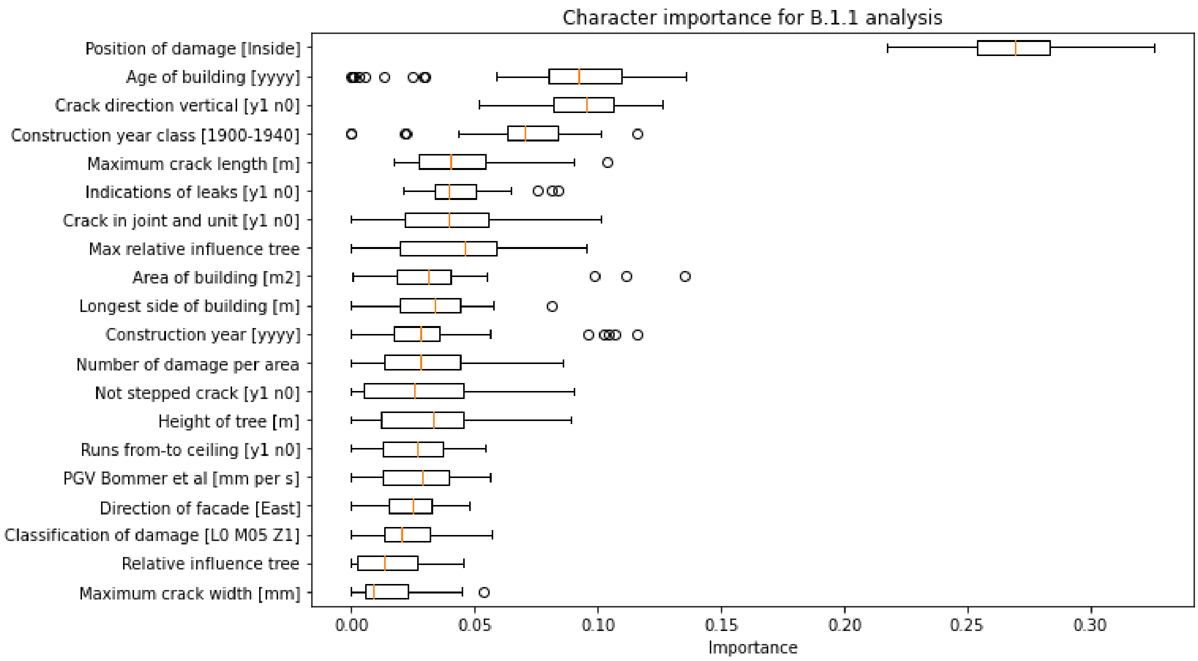


Figure 57 Importance of each characteristic for a classification analysis on damage cause 'initial hindered deformations', represented in boxplot figures. Obtained by GSS cross-validation with large dataset of 1830 damage cases.

Other important characteristics in classification analysis are: the age of the building, whether a crack is vertical and whether the construction year is between 1900 and 1940 (Figure 57). Crack direction is difficult to interpret, as it depends on small local details. According to the patterns in the decision tree, a vertical crack is most likely caused by hindered deformations.

It is sometimes difficult to read an apparent relation from a decision tree. Examples of this are the relations dealing with the construction year of a building. It is applied by the pattern recognition in various ways:

- If the building is older than 96 years, damage is most likely **not** caused by B.1.1,
- If the building is newer than 137 years, damage is most likely caused by B.1.1,
- If the building is newer than 147.5 years, damage is most likely **not** caused by B.1.1,
- If the building is newer than 100.5 years, damage is most likely **not** caused by B.1.1,
- If the building is older than 106.5 years, damage is most likely caused by B.1.1,
- If the building is older than 15.5 years, damage is most likely caused by B.1.1,
- If the building is newer than 36 years, damage is most likely caused by B.1.1
- Constructed between 1900 and 1940 is most likely **not** caused by B.1.1,

A possible explanation could be that building methods and materials changed over the years which has consequences for the occurrence of initial hindered deformation damage.

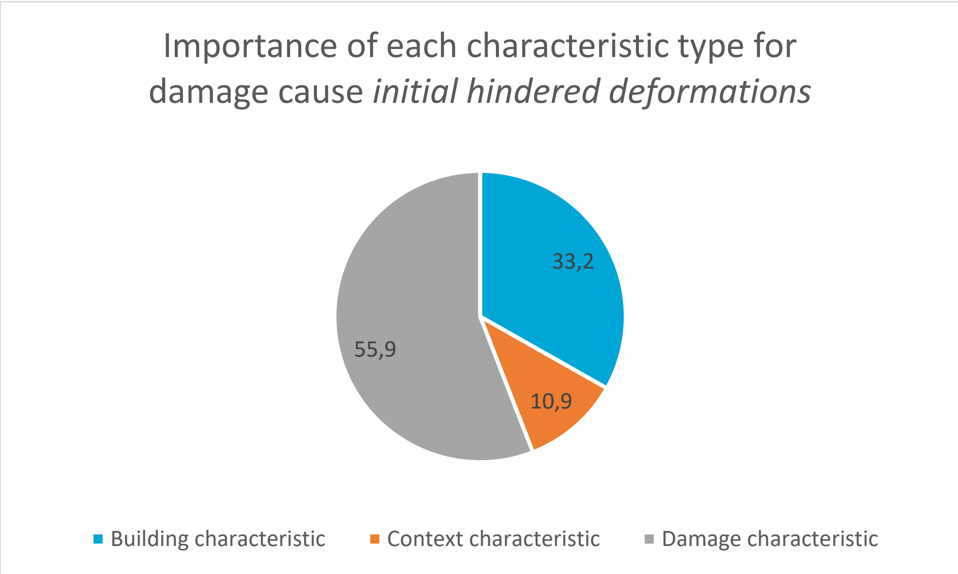


Figure 58 Characteristic types distributed by their importance of classification determination of damage cause B.1.1. Obtained by GSS cross-validation with large dataset of 1830 damage cases.

The distribution of the characteristic types according to their importance (Figure 58) favours the found relation. Hindered deformations are mostly the cause of building characteristics such as applied materials and connections. Context characteristics are less important for the developing of deformation damage. This theory is presented in the pattern recognition by giving building characteristic an importance of 33.2%, while context characteristics are getting 10.9%. The remaining 55.9% is put down to damage characteristics.

Concluding, the numbers show that the found pattern is reliable. It has a accuracy of 77% and the standard deviations within the characteristics seem limited (Figure 57). Moreover, it is possible to explain the selected characteristics with hypotheses of damage situation. However, there are better explainable hypotheses possible and the relations that are proposed by the pattern are diffuse. Therefore, the decision tree is not meaningful for application in practice.

7.4 *Interpretation of the classification result to cause Initial imposed deformations (B.2.1)*

The found relations by the pattern recognition with damage cause *initial imposed deformations* (B.2.1) is discussed in this sub-chapter. The relations are considered on their reliability, if they are explainable and if it is meaningful for practice.

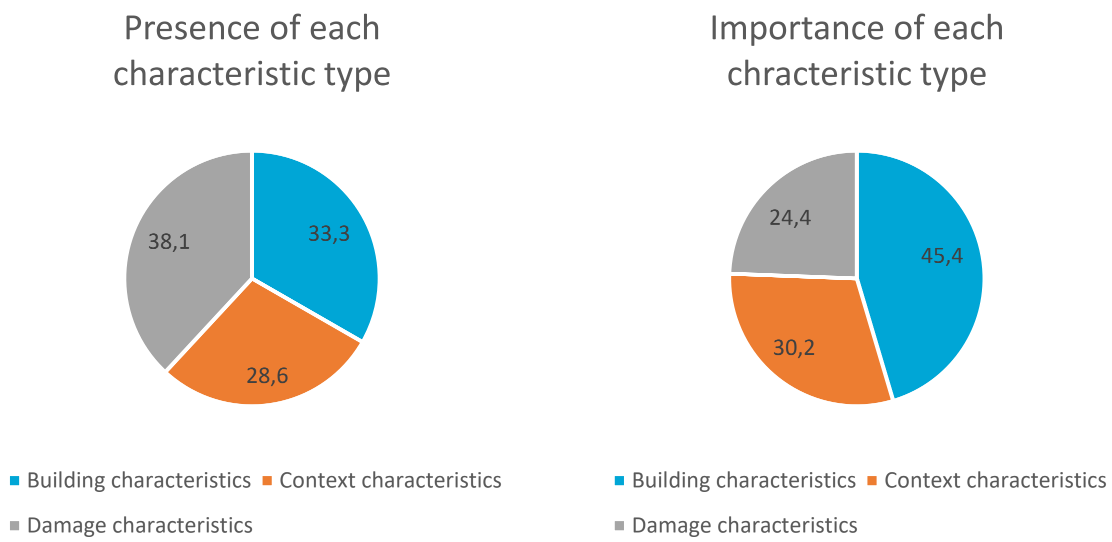


Figure 59 Characteristic types distributed by their presence (left) and importance (right) of classification analysis of damage B.2.1. Obtained by GSS cross-validation with large dataset of 1830 damage cases.

The classification analysis for initial imposed deformations (B.2.1) have an accuracy of 69% (Figure 31). In the presence distribution (Figure 59 left), the shares between the different characteristic types is almost evenly spread, despite the fact that the pattern recognition was provided with more damage characteristics than building or context characteristics. If the influence of each characteristic in the found pattern is incorporated in the characteristic presence (Figure 59 right), building characteristics seem to be the most important type with 45.4%. It is followed by context characteristics with 30.2%.

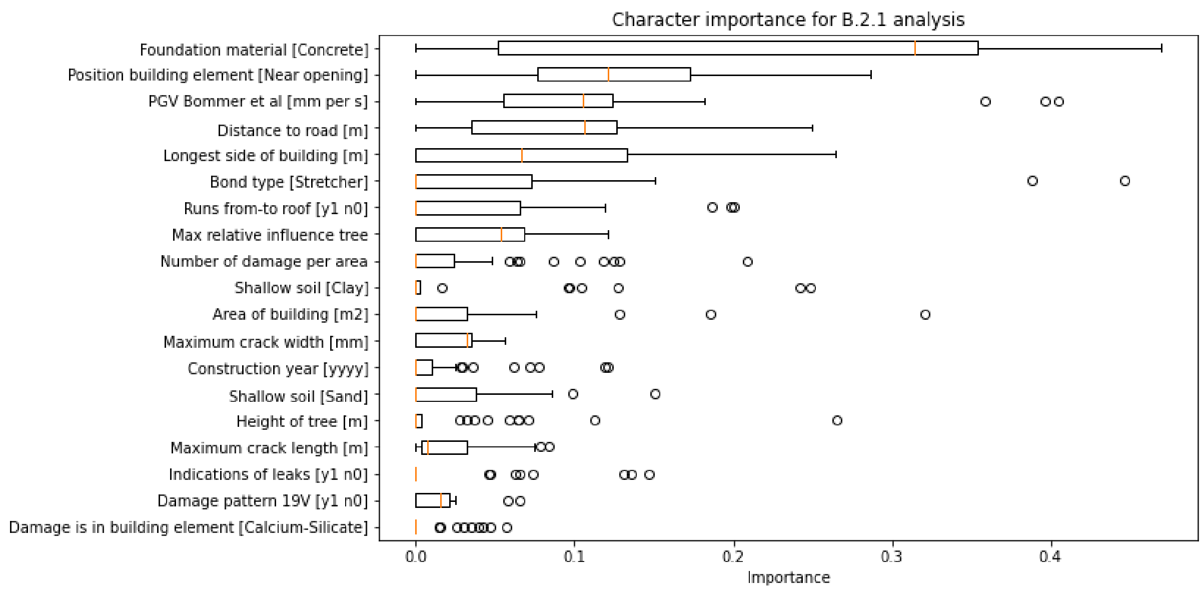


Figure 60 Importance of each characteristic of a classification analysis on damage cause 'initial imposed deformations', represented in boxplot figures. Obtained by GSS cross-validation with large dataset of 1830 damage cases.

It seems logical that building characteristics are important for initial imposed deformations. Differences in materials and inadequate connections can cause deformation related damage. Examples indicated by the pattern recognition are: the foundation material, the size of the building and the type of masonry bond (Figure 60).

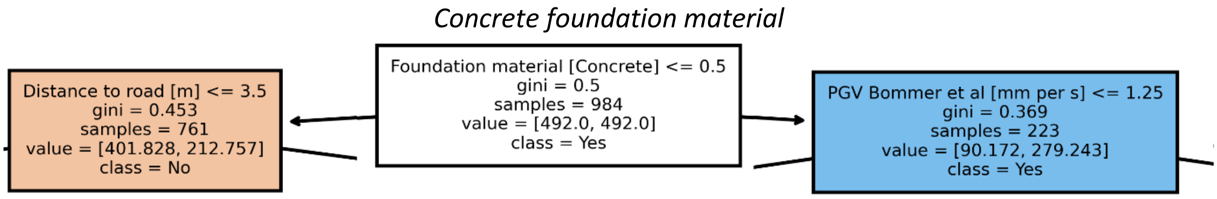


Figure 61 Start node in classification decision tree of cause B.2.1, where building characteristic 'foundation material [concrete]' splits the data. If the foundation is made out concrete, it has a value of 1.0. Therefore it does not fulfil the threshold and proceeds to the right child.

For the characteristic foundation material, the distribution of the importance is rather large compared with the other characteristics (Figure 60). So it cannot be indicated as a key characteristic in the determination of B.2.1 as the cause of damage. However, the pattern states that buildings with a concrete foundation are more vulnerable for imposed deformation damage (Figure 61).

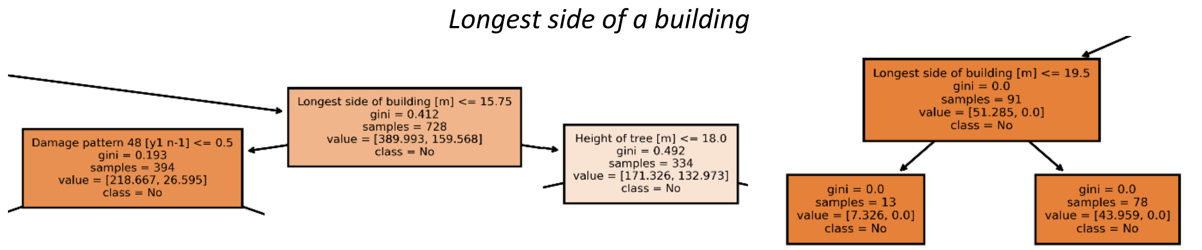


Figure 62 Internal nodes in classification decision tree of cause B.2.1, where building characteristic 'longest side of building' has a threshold at 15.75 m and at 19.5 m. For example, if the longest site of a building is 20 m, the pattern proceeds than to the right child.

The size of a building seems explainable. Bigger buildings have large construction elements which may deform more than smaller elements. That is also represented in the decision tree. However, two patterns show an opposite relation with the longest side of the building. Both of them have a different threshold. The first one explains that it is most likely damage in shorter buildings is not caused by imposed deformations (Figure 62 left). While the second pattern presents an relation without consequence (Figure 62 right).

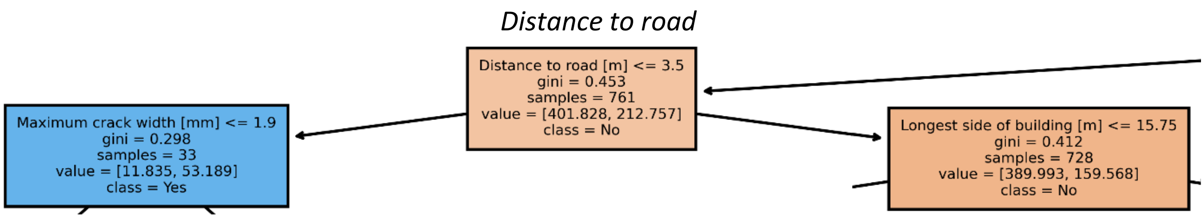


Figure 63 Internal node in classification decision tree of cause B.2.1, where context characteristic 'Distance to road' has a threshold at 3.5m. A pattern with a distance to a road smaller than 3.5 meter proceeds to the left child.

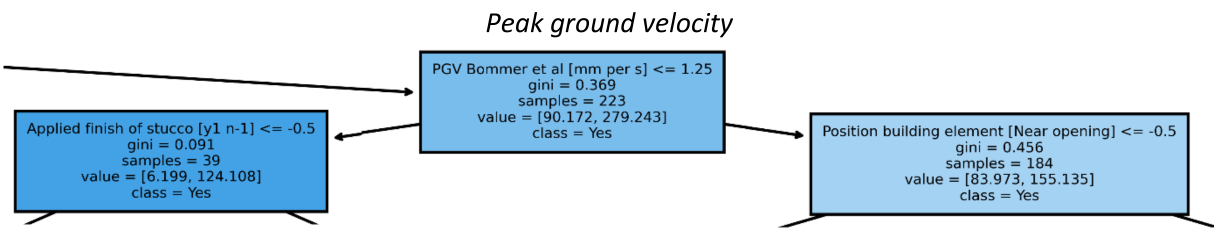


Figure 64 Internal node in classification decision tree of cause B.2.1, where context characteristic 'PGV Bommer et al' has a threshold at 1.25 mm/s. Damage in buildings where this earthquake load was smaller as the threshold will proceed to the left child.

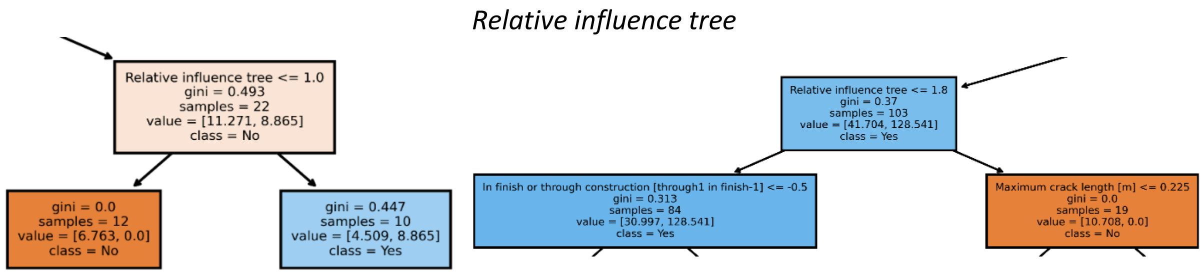


Figure 65 Internal node in classification decision tree of cause B.2.1, where context characteristic 'relative influence tree' has a threshold at 1.8. Damage in buildings where the relative influence is larger as 1.8 will not fulfil that threshold and the pattern proceeds therefore to the left child.

According to the found relation, important context characteristics are the peak ground velocity (Figure 63), the distance to a road (Figure 64) and the influence of trees (Figure 65). The first two seems illogical as they cannot explain imposed deformations of building elements. Vibrations of earthquakes may weaken connections. These loose connections can, for example, result in instable roofs which will press on facades. On the other hand, the influence of trees is a logical context characteristic for imposed deformations. Tree roots can move building elements which may end in deformation damage. This relation is made clearly visible by a pattern. Most probably, the damage is not caused by imposed deformations when trees are not higher than 1.8 times the distance to the tree (Figure 65).

Another explanation for these illogical characteristics lies in the method that is applied during the investigation of damage cases in the provided data. Imposed deformation presents itself with minor damage. Therefore, it was difficult for the experts to determine if earthquake vibration plays a part in the appearance of damage. Especially if the trigger function of earthquake tremors is taken into account. This could result in a relation between a PGV characteristic and imposed deformation damage. However, B.2.1 damage may also occur in buildings which have no expectance earthquake load.

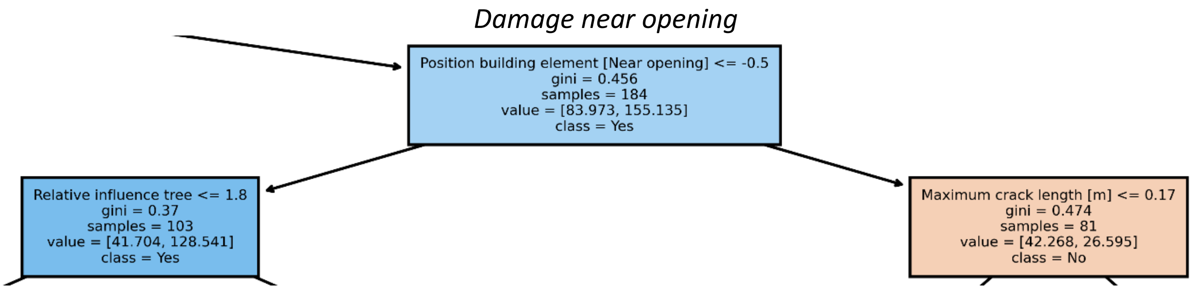


Figure 66 Internal node in classification decision tree of cause B.2.1, where damage characteristic 'position building element [near opening]' splits the data. If a damage is not located near an opening, it as a value of -1.0. Therefore, it does fulfil the threshold and the pattern proceeds to the left child.

An indicated important damage characteristic is damage that is located near an opening (Figure 60). A window or door may also deform because of humidity and temperature differences. Also, openings lead to local decreasing of load resistance. Which is a likely place for damage to occur. However, the produced decision tree suggests otherwise. It proposes that it is most likely that damage near openings is not caused by imposed deformations (Figure 66). Damage location near corners is better explainable as imposed deformation. Unfortunately, corner characteristics were not selected by pattern recognition (Figure 60).

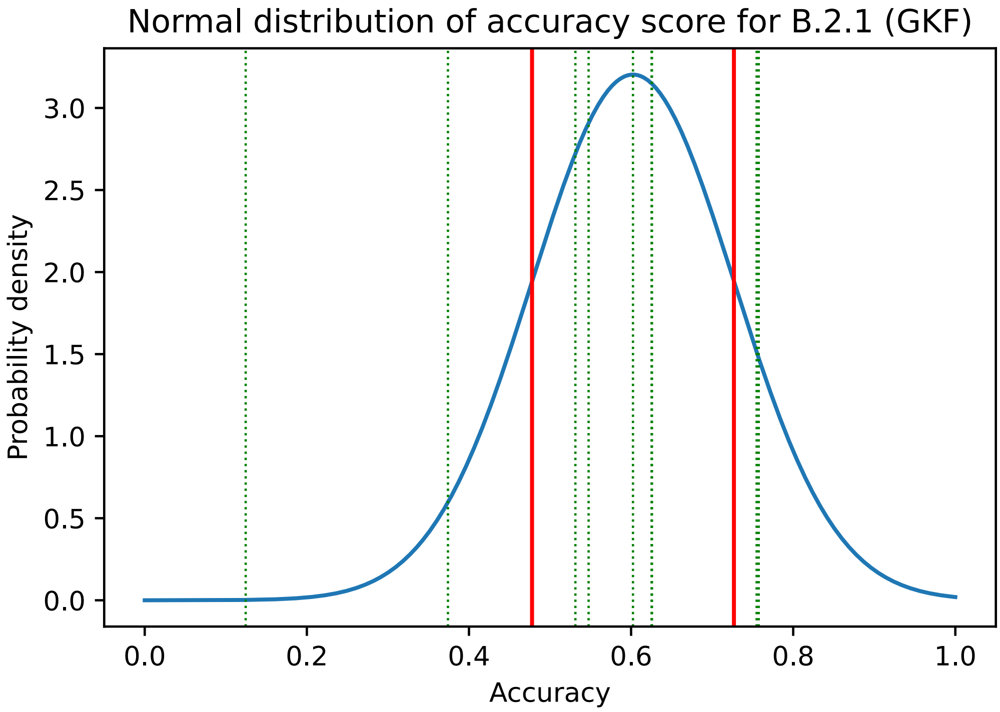


Figure 67 Normal distribution of cross-validation classification on damage cause B.2.1. Every green dotted line is the accuracy score of one step in cross-validation process. The blue line indicates a normal distribution of those green lines. Red lines shows the standard deviation of the normal distribution. Obtained by GKF cross-validation with the original dataset of 1549 damage cases.

It must also be pointed out that there is a large difference between the various cross-validations. The accuracy scores differ a lot (a standard deviation of 12.5% on average 60.2% accuracy, see Figure 67). Also the importance of each selected characteristics is not constant for every cross-validation (Figure 60). This will mean that every single run from a cross-validation can produce a completely different decision tree. The reason behind this could be that the provided data is too small in size for this analysis. That makes it difficult to draw a conclusion about the relation between a characteristic and damage cause initial imposed deformations. Therefore, a conclusion could be that the found relation with this decision tree is not reliable enough. For that reason, it cannot be applied meaningful in practice.

7.5 *Interpretation of the classification result to cause Unequal settlements with equal loads (C.1.1)*

The found relations by the pattern recognition with damage cause *unequal settlements with equal loads (C.1.1)* is discussed in this sub-chapter. The relations are considered on their reliability, if they are explainable and if it is meaningful for practice.

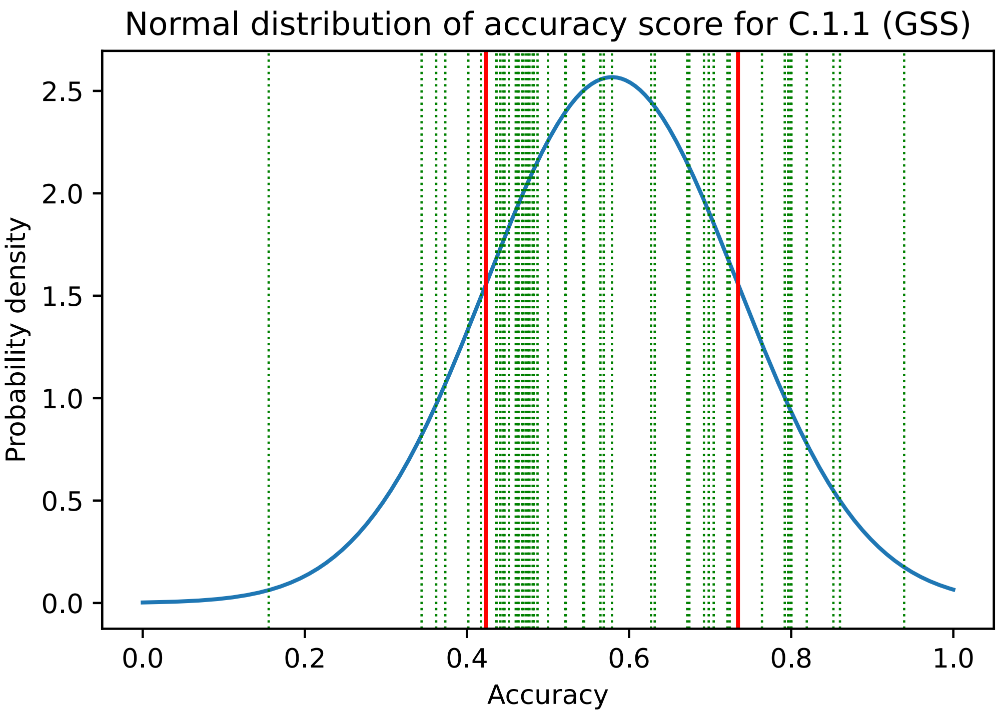


Figure 68 Normal distribution of cross-validation classification on damage cause C.1.1. Every green dotted line is the accuracy score of one step in cross-validation process. The blue line indicates a normal distribution of those green lines. Red lines shows the standard deviation of the normal distribution. Obtained by GSS cross-validation with large dataset of 1830 damage cases.

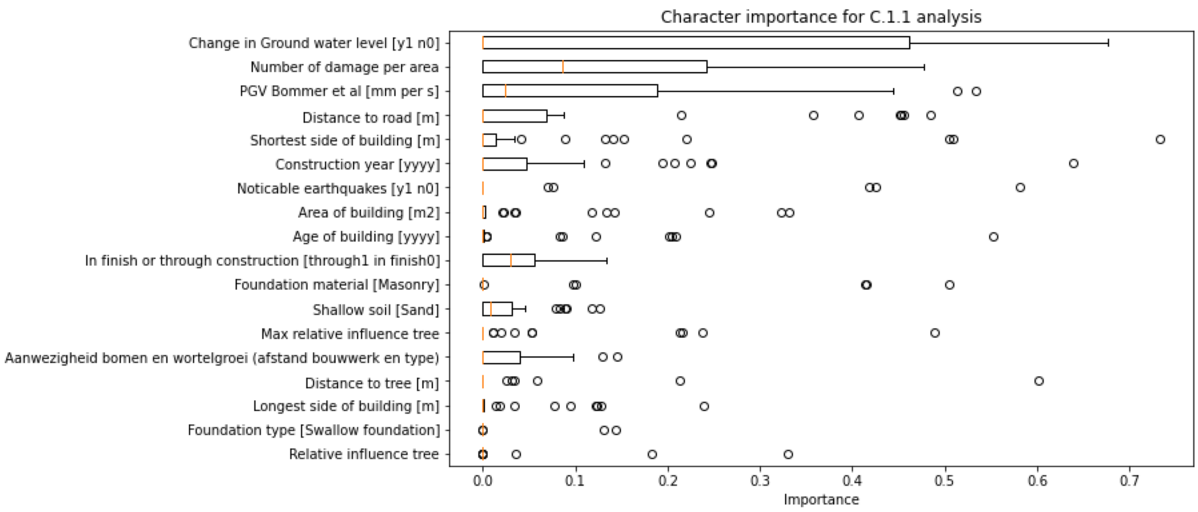


Figure 69 Importance of each characteristic for a classification analysis on damage cause unequal settlements with equal loads, represented in boxplot figures. Obtained by GSS cross-validation with large dataset of 1830 damage cases.

Classification pattern recognition could produce a relation for damage cause unequal settlements with equal loads (C.1.1) with an accuracy of 43% (Figure 31). However, the variation between the cross-validations is large (Figure 68). The accuracy score has a standard deviation of 15% and the most important characteristics have a standard deviation of more than the average value (a mean of 0.21, with a std. of 0.26). Moreover, the boxplot calculations show that outliers in the characteristics importance are spread out over the whole range (Figure 69). Therefore, the found pattern is not trustable, as it various a lot within the provided data.

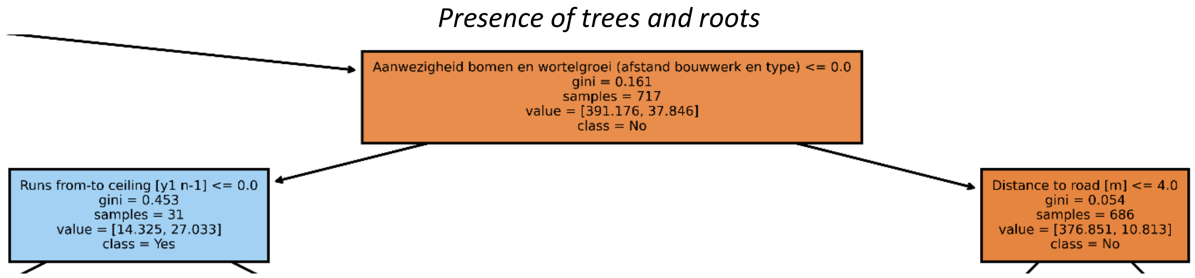


Figure 70 Internal node in classification decision tree of cause C.1.1, where context characteristic 'presence of trees and root growth' split the data. If a damage is in a building with close by trees, it as a value of 1.0. Therefore it does not fulfil the threshold and the pattern proceeds to the right child.

Still, some characteristics in the decision tree are logical. The change in ground water level is explainable, as it can induce settlement differences while building loads stay the same. Also the presence of trees could have an effect on settlements. However, it has a relative small influence on the result according to the produced cross-validation. Moreover, the predicted relation states that the presence of trees near a building is an indication that it is most likely that damage is not the cause of unequal settlements with equal loads. It could be explained as the influence on soil conditions by trees as another damage cause category (C.3.8).

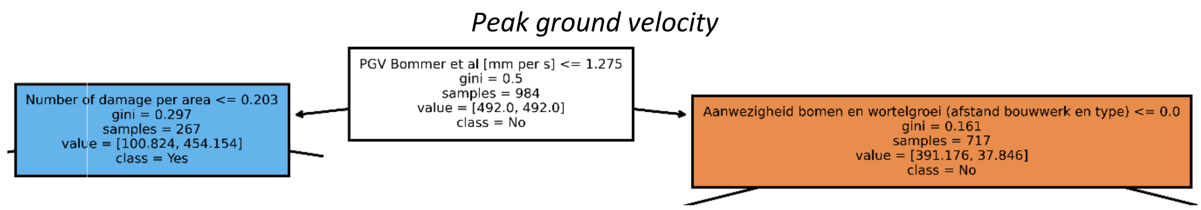


Figure 71 Start node in classification decision tree of cause C.1.1, where context characteristic 'PGV Bommer et al' split the data. If a damage is in a building with a PGV higher than 1.275 mm/s, it does not fulfil the threshold and the pattern proceeds to the right child.

The earthquake load seems unexplainable as third most important characteristic for a regression model on damage cause unequal settlements with equal loads. It is even the start node (root) of the decision tree (Figure 71). However, it is most likely that damage with a peak ground velocity above the threshold is not the cause of unequal settlements with equal loads. The threshold is almost the same value as determined in the earthquake regression analysis (1.65mm/s). This supports the suggestion that the pattern rules out any earthquake damage with this split. Therefore, the choice for that characteristic in this decision tree pattern is defensible.

As stated above, the construction of the provided data may have an influence on the found relation between PGV and unequal settlement with equal loads damage. Earthquake damage mostly occurs in combination with other damage types. It is possible that earthquakes trigger small soil settlements. This can result in a relation between that characteristic and C.1.1 damage, which was picked up by the pattern recognition.

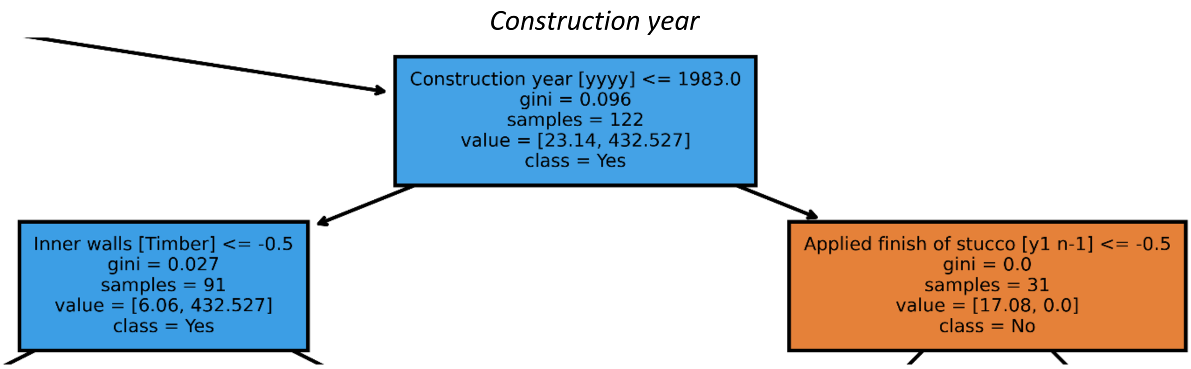


Figure 72 Internal node in classification decision tree of cause C.1.1, where building characteristic 'construction year' has a threshold at 1983. If a damage is in a building with a construction year later as 1983, it does not fulfil the threshold and the pattern proceeds to the right child.

An successful split in the decision tree is made by the characteristic construction year. It has a threshold at 1983 (Figure 72). Older buildings seem to be more vulnerable for unequal settlements with equal loads. An explanation for this could be that younger buildings have a better designed foundation as older buildings.

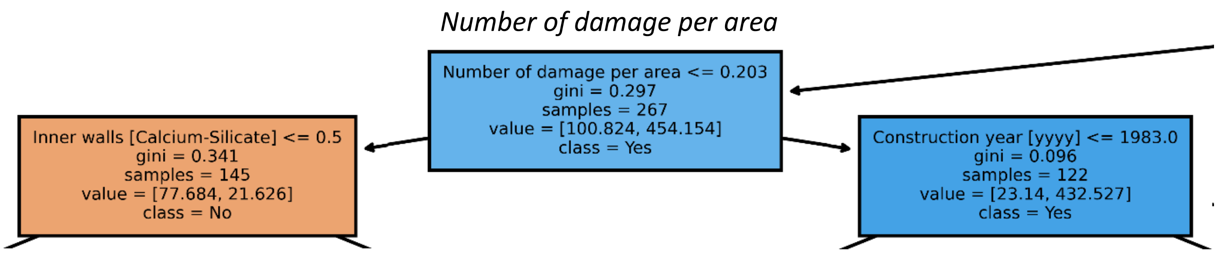


Figure 73 Internal node in classification decision tree of cause C.1.1, where building characteristic 'number of damage per area' has a threshold at 0.203 #/m². If a damage is in a building with more than 0.203 damage per m², it does not fulfil the threshold and the pattern proceeds to the right child.

According to the boxplot figures (Figure 69), the number of damage cases per area is an important characteristic in this pattern recognition. This could be explained as unequal settlements has an effect on the whole building. Other damage causes such as overloading and deformations are the result of local stress disturbances. Whereas unequal settlements have an impact on the stresses in the hole construction. This results in multiple damage cases in the building, as shown in the decision tree (Figure 73).

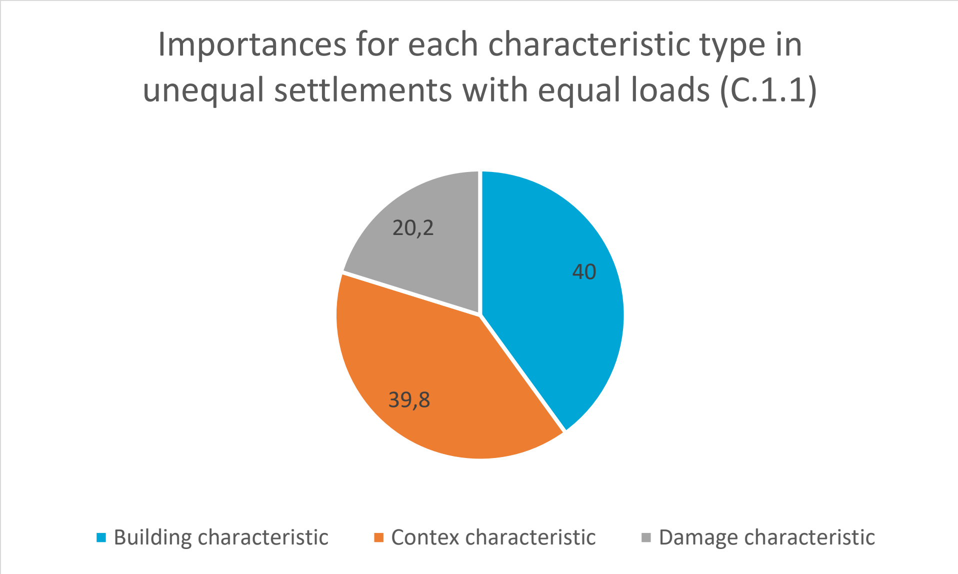


Figure 74 Characteristic types distributed by their importance of classification analysis of damage cause C.1.1. Obtained by GSS cross-validation with large dataset of 1830 damage cases.

When comparing the different characteristics types with each other, it is noticeable that damage characteristics are the least present in the decision tree. Apparently, the layout and location of damage is of less importance for the determination whether it is caused by unequal settlements with the same loads as other characteristic types. So it explains that C.1.1 damage can occur in different forms. That seems logical, as damage cause unequal settlements with equal loads is damage which can appear at all weak locations in a building.

As earlier concluded, the found relation is not reliable. Therefore it is not meaningful to introduce this decision tree pattern in practice.

7.6 *Interpretation of the classification result to cause Unequal settlements (C.X)*

The found relations by the pattern recognition with damage cause *unequal settlements* (C.X) is discussed in this sub-chapter. The relations are considered on their reliability, if they are explainable and if it is meaningful for practice.

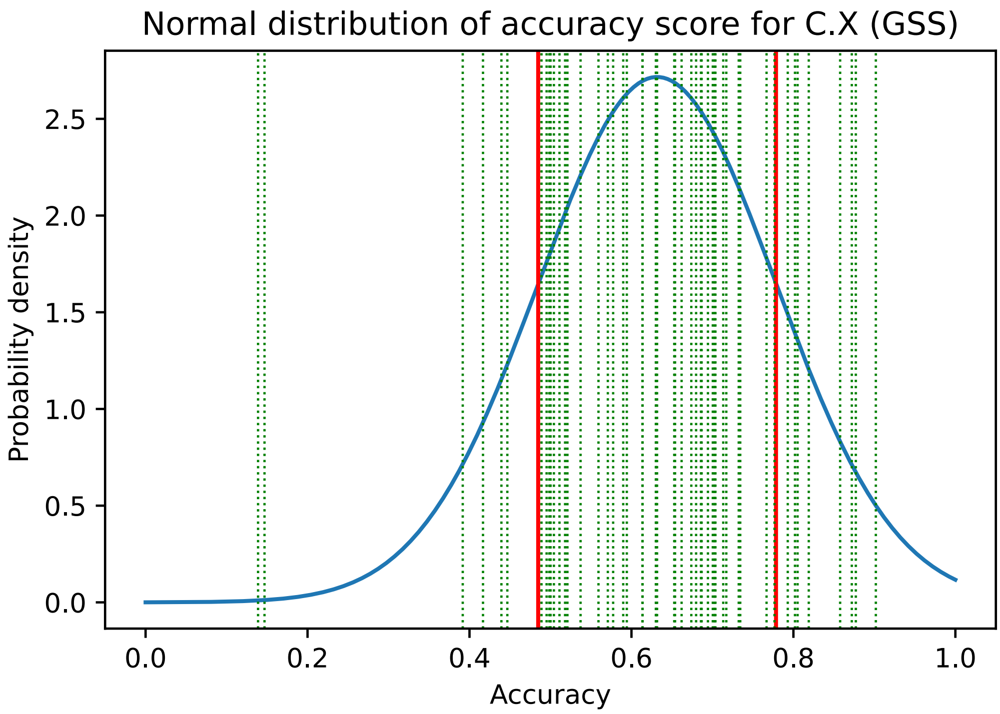


Figure 75 Normal distribution of cross-validation classification on damage cause C.X. Every green dotted line is the accuracy score of one step in cross-validation process. The blue line indicates a normal distribution of those green lines. Red lines shows the standard deviation of the normal distribution. Obtained by GSS cross-validation with large dataset of 1830 damage cases.

Classification pattern recognition is performed on damage cause category unequal settlements with an accuracy of 34% as result (Figure 31). In the cross-validation, this score varies between 14% and 90% and it has an standard deviation of 14% (Figure 75). At a first glace on the accuracy score, the found pattern achieves bad and good results. On average however, it seems to fulfil its job in deciding if damage was caused by unequal settlements.

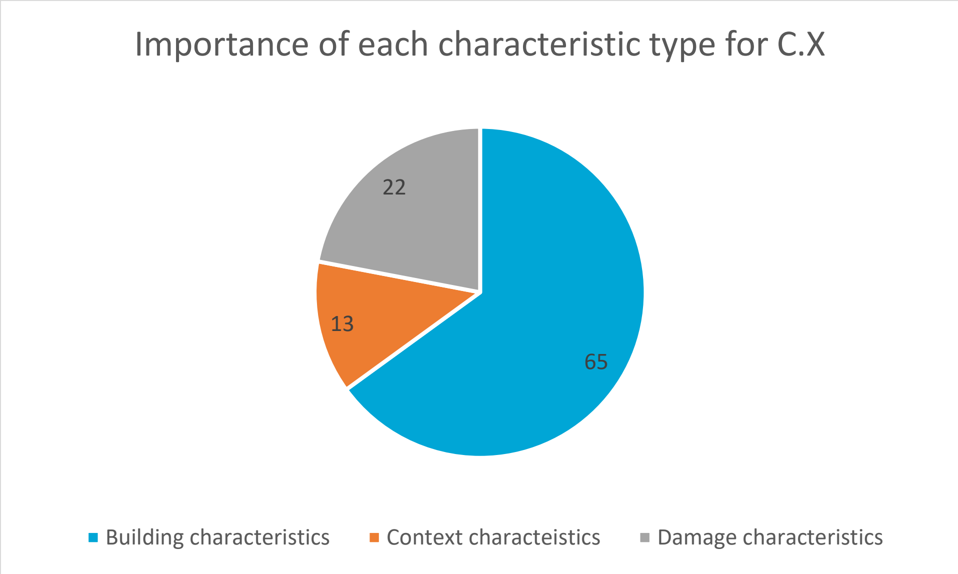


Figure 76 Characteristic types distributed by their importance of classification analysis of damage cause C.X. Obtained by GSS cross-validation with large dataset of 1830 damage cases.

According to the found decision tree, the most important characteristic type is building characteristics with 65% (Figure 76). It is followed by damage characteristics with 22%. The least important type is the context characteristics. This was unexpected as unequal settlements are mostly influenced by surrounding factors near the building, such as the change in ground water level and the behaviour of subsoil conditions. However, the pattern recognition has a preference for building characteristics. Apparently, the way a soil is loaded by a building has more consequences on unequal settlements than how a soil behaves. Although, not all the measurements on soil conditions were always present in the data. Also, the foundation building characteristics and soil context characteristics influence each other. If a building undergoes unequal settlements, is it a result of a bad foundation or poor soil conditions?

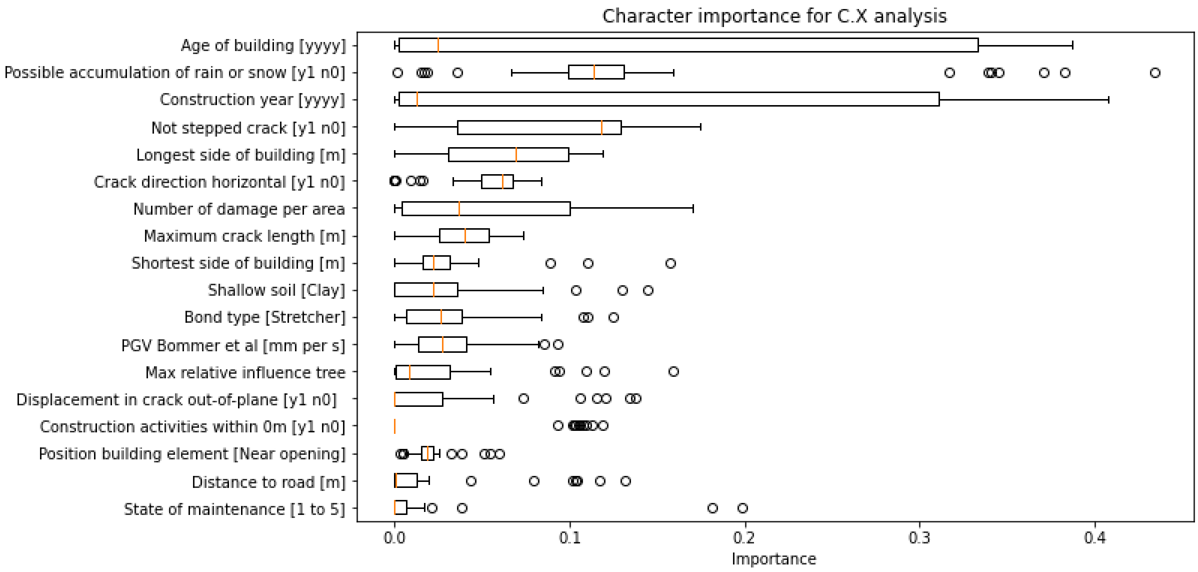


Figure 77 Importance of each characteristic for a classification analysis on damage cause 'unequal settlements', represented in boxplot figures. Obtained by GSS cross-validation with large dataset of 1830 damage cases.

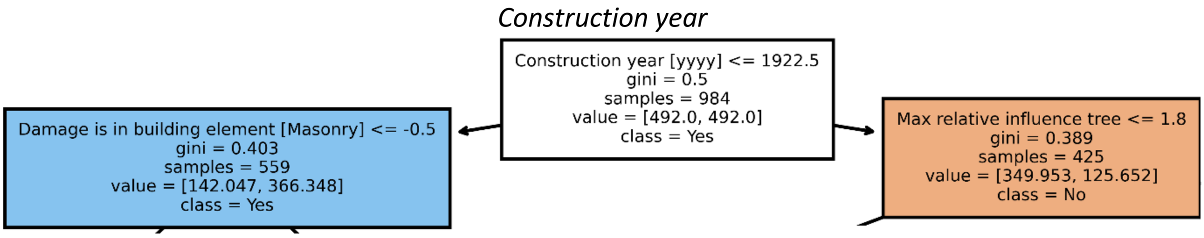


Figure 78 Start node in classification decision tree of cause C.X, where context characteristic 'construction year' has a threshold of 1922.5. Damage in buildings with an higher construction year will proceed to the left child.

The most important characteristic indicated by the algorithm is the age of the building (Figure 77). It could be explained, as subsoil under younger buildings had more research done on it, when a building was designed than older buildings. Back in time, the soil investigations were not as advanced as now. That is indeed the relation proposed by the pattern recognition (Figure 78). At the root of the tree (starting point), the data are split at a construction year of 1922.5. It is most likely that damage in buildings that are younger than the threshold not caused by unequal settlements. The decision tree continues splitting to rule out some exceptions, but the split at a construction year of 1922.5 is almost sufficient to figure out which damage was caused by C.X. However, this characteristic has a standard deviation of 17% on a mean of importance at 17% (Figure 77). This means that the age split is not always consistent within the provided data and it is therefore not reliable.

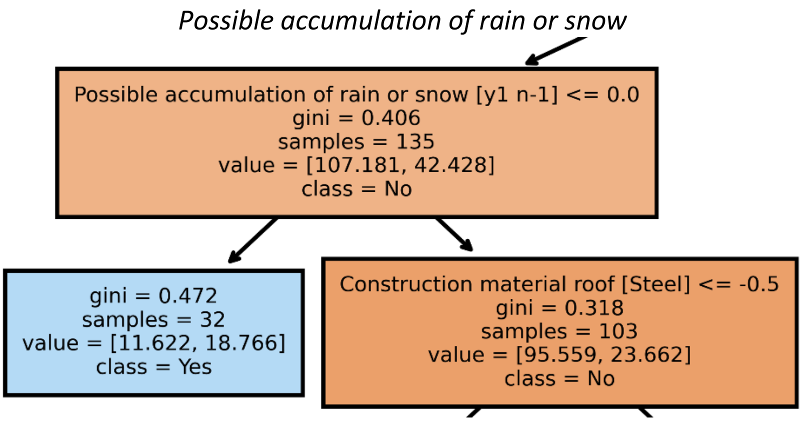


Figure 79 Internal node in classification decision tree of cause C.X, where building characteristics 'possible accumulation of rain or snow' splits the data. If snow or rain can accumulate on a roof, it has a value of 1. Therefore it does not fulfils the threshold and the pattern proceeds to the right child.

A more reliable but less important characteristic is the possibility of rain and snow to accumulating on the roof of the building (Figure 79). This characteristic was incorporated in the data to check if structures could be overloaded by snow or rain loads. Therefore it was unexpected that this characteristic was selected by the algorithm as indication for damage cause unequal settlements. The characteristic was checked for flat roofs by reviewing satellite images and drawing during inspections. A suggestion could be that flat roofs have a force flow in buildings that have another effect on the foundation than buildings with a sloped roof.

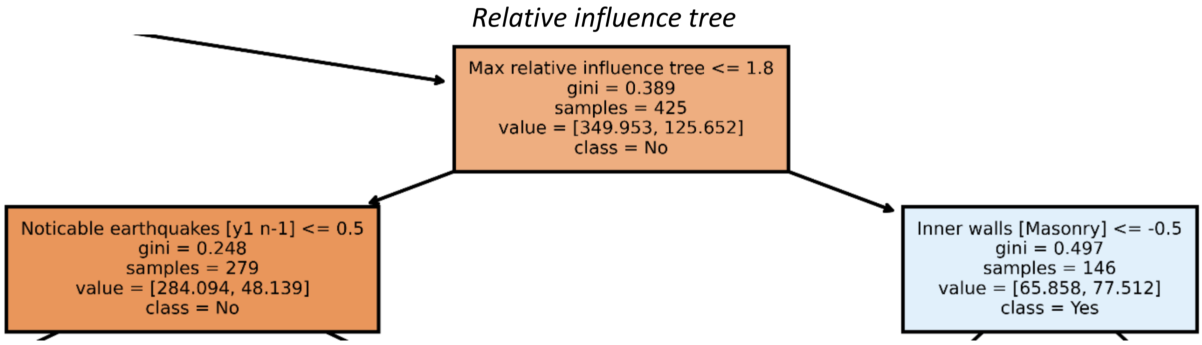


Figure 80 Internal node in classification decision tree of cause C.X, where context characteristics 'Max relative influence tree has a threshold at 1.8. Damage in buildings with an higher tree influence will proceed to the right child.

The last noticeable case characteristic in the decision tree is the relative influence of a tree. That is a ratio between the hight of a tree and the distance to a building. A higher ratio express higher trees and closer to buildings. In the case of Figure 80, this characteristic has an threshold of 1.8. It means that trees with an hight of 1.8 times the distance to a buildings has an influence on the occurrence of damage cause category 'unequal settlements'. Staalduinen et al applied a threshold of 1.0 as falsification. The found relation indicates that a more strict falsification threshold is possible. That trees as an influence on unequal settlements is logical, as they have an influence on the ground water level. In the appendix are more examples explained on how trees could have an influence on unequal settlements.

All remaining characteristics have a small influence on the outcome of the decision tree. The algorithm applies these characteristics in order to be more certain about the result. In other words, it implements them to obtain more purified leaves.

Concluding, the expected context characteristics are missing in the found relation. The pattern recognition favours the building characteristics. Also, the used characteristics are not always explainable with hypotheses. Moreover, this decision tree is not the most reliable pattern in this thesis. Therefore, the relations in the decision tree are not meaningful enough to be applied in practice. However, it may support experts in their investigations and research to unequal settlement damage.

7.7 *Interpretation of the regression result to cause Overloading due to vibrations by earthquakes (A.3.5)*

Characteristic	Group 1	Group 2	Group 3	Group 4	Group 5	Mean
Peak ground velocity [mm/s]	55.18	52.64	52.61	50.87	53.05	52.87
Max relative influence tree	13.94	14.22	11.67	14.82	16.83	14.29
Branched cracks [y1 n0]	3.1	2	2.26	2.53	1.61	2.3
Position of damage [Inside]	-2.46	-2.87	-2.01	-1.47	-3.18	-2.4
Age of building [yyyy]	-6.81	-5.62	-6.04	-12.09	-5.46	-7.2

Table 8 Result of cross-validation (group 1 - 5) with mean result on linear regression of an earthquake damage cause pattern. This table shows the slope of each characteristic of the linear equation.

It was possible to generate a linear regression pattern for A.3.5 (Overloading due to vibrations by earthquakes) with a positive coefficient of determination of 0.48. The most important characteristic is context characteristic Peak Ground Velocity (PGV). This is expected and explainable for earthquake related damage.

Similar to classification, the age of the building has a negative relation with earthquake damage (Table 8). That seems unexplainable, because older buildings are have weakened over the years. Maybe this weakening will allow vibrations to go through the building without damage. Hypotheses in favour of the found relation are that older buildings have lighter timber floors as newer buildings with concrete floors. This will make older buildings less vulnerable for induced inertia forces of earthquakes. Lastly, there is a possibility that the age of a building is an indication another damage cause category, like damage occurring because of shrinkage or autonomous settlements.

Another unexplainable characteristic is the influence of nearby trees (Table 8). An estimation of the height and distance of nearby trees seems to have a relation with earthquake related damage. A possible explanation could be that tree roots under a building will contribute to the transfer of ground vibrations to the building.

The only two damage characteristics are branched cracks and the position of damage. Both have relatively less influence on the outcome. Most likely a branched crack will rather be the result of earthquake damage, although the presence of branched cracks in the data was only 5%. Damage located on the outside of a building is more often caused by earthquake than damage located on the inside.

Concluding, the found relation can help with improving earthquake damage investigations. The new insights show that other characteristics than PGV could be important in the determination on the share of earthquakes in damage. However, those characteristics are not evident enough to be applied in practice without caution.

7.8 *Interpretation of the regression result to cause Deformations (B.X)*

It was very difficult to construct a pattern for cause categories B.1.1 and B.2.1. Even for experts, it is difficult to assign a technical attributability to hindered and imposed deformations when deformations have a part in the occurrence of damage. Sometimes experts determine the cause of damage as one of those two, without assigning one specific. For that reason, all damage cases that had a relation with deformations are placed together in one group as B.X. That results in a pattern with a positive coefficient of determination of 0.28.

Characteristics	Group 1	Group 2	Mean
Crack direction horizontal [y1 n0]	4.77	11.51	8.14
Damage is in building part [Ceiling]	5.95	-6.64	-0.34
Position of damage [Inside]	-13.92	-12.82	-13.37
Damage is in building element [Masonry]	-11.83	-24.4	-18.11
Building renovations [y1 n0]	-27.63	-33.78	-30.7
Maximum crack width [mm]	-24.38	-60.16	-42.27

Table 9 Results of cross-validation (group 1 & 2) with mean result on linear regression of damage category deformation (B.X). This table shows the slope of each characteristic of the linear equation.

For deformation damage, building characteristics are important indications. Only one of the selected building characteristics is the characteristic that indicates if renovations were carried out. This characteristic is marked with a negative slope (see Table 9). This means that when renovations were carried out on a building, it will reduce the technical accountability of damage cause deformations. This is the opposite of what was expected. During renovations, other new materials will be introduced to the building. This may lead to new deformation damage. However, renovation can also lead to overloading damage causes (A.X). Therefore, the found relation could be right as renovations may also lead to other damage causes. Another reason could be that deformation damage was repaid during the renovations. A as result of that, it was not detected by at the time of the investigation.

All other characteristics in table 3 fall in the category of damage characteristics. The only characteristic with a positive slope is damage with a horizontal crack. This is in contrast with the classification analysis of initial hindered deformations (B.1.1), where vertical cracks were an indication characteristic for this type of damage. Evaluation of the crack direction on its relevancy is difficult, as the direction of the crack is determined by local details. Deformation damage can result in both vertical cracks as horizontal cracks. In the found relation, horizontal cracks are in favour for deformation damage. This seems to be more in line with the literature (De Vent, 2011). The same holds good for damage located at inside places, where classification models are in contradiction of what is presented here in regression.

However, a combination of crack direction and damage location is in agreement between both analysis types. If damage is located at the outside of a building and it is not vertical (a horizontal crack) it is most likely not the result of deformation damage according to the produced patterns. That is according to what is observed in practice. Long outside facades with deformation damage have most likely vertical cracks. Horizontal deformation cracks are mostly found at inside building locations, such as cracks connections between ceilings and walls. Therefore, this could be an explanation for the on first hand contradictory between classification and regression analysis.

The remaining damage characteristics of Table 9 are explainable as being a part of a deformation damage regression tool. Ceilings are subjectable to deformation damage, as beams deform under loads or as they are subjected to other circumstances. Damage in masonry materials are more likely caused by overloading causes. Similar for maximum crack width, as narrow cracks are most likely caused by creep or shrinkage.

Concluding, the found relations are not always explainable with hypotheses. There are also some contradictory relation with earlier classification analysis. However, these regression results are supported with literature. Still, it is not enough to become a meaningful relation to be applied in practice.

7.9 *Interpretation of the regression result to cause Unequal settlements (C.X)*

The variance of the technical attributability of cause C.1.1 was too small to create a linear regression model.

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Mean
Maximum crack length [m]	41.12	43.96	47.92	35.29	52.8	47.26	44.72
Runs from-to floor	8.08	5.03	7.68	7.38	2	13.04	7.2
Runs from-to opening	5.08	4.88	6.11	6.9	4.49	5.32	5.46
Ground settlement [not present]	4.57	3.26	2.41	5.01	0.46	4.45	3.36
Crack direction horizontal	-3.22	-4.14	-6.1	0.2	-4.26	-1.54	-3.18
Shallow soil [Clay]	-2.41	-2.36	-4.56	-7.38	-0.43	-2.11	-3.21
Branched cracks	-5.96	-5.79	-4.89	-7.76	-5.53	-4.92	-5.81
Runs from-to roof	-6.29	-8.53	-8.97	-9.78	-3.27	-8.25	-7.51
Change in Ground water level	-7.74	-10	-10.25	-5.26	-13.09	-8.88	-9.2
Deep soil [Clay]	-8.54	-9.12	-8.96	-12.22	-9.17	-10.74	-9.79
Settlement differences in building	-23.36	-22.28	-23.27	-16.34	-25.39	-23.1	-22.29
PGV Bommer et al [mm per s]	-40.64	-42.79	-38.85	-40.91	-39.91	-40.59	-40.61

Table 10 Results of cross-validation (group 1 -6) with mean result on linear regression of damage category unequal settlements (C.X). This table shows the slope of each characteristic of the linear equation.

It was possible to deliver a linear regression pattern for unequal settlements (C.X) with a positive coefficient of determination of 0.18. This is the lowest score of the three presented models. The slopes of each characteristic for damage cause unequal settlements is shown in Table 10. A positive slope means that it will contribute to the technical attributability of damage cause C.X. Whereas negative slopes will reduce an unequal settlement technical attributability.

The most remarkable finding is that characteristic 'Settlement differences over the length of the building' (shorted in Table 10 to settlement differences in building) has a negative slope. This is unexplainable. It was expected that settlement differences established with measurement of bed joint levelling would indicate damage cause unequal settlements. A few reasons why the pattern recognition takes this false conclusion are the following. First, the interpretation per measurement could be different per expert. There was no a threshold of the maximum slope where experts have to declare a settlement difference. Buildings always have some settlement differences. However, the transformation from an analogous maximum slope value to a discrete statement of the presence of settlement differences was not transparent and did not always runs in the same way. Secondly, settlement differences over the length of the building were spotted in 91% of the buildings while the damage cause unequal settlements was mentioned in 60% of the damage cases. So it could be concluded that not all settlement differences will lead to damage in the C.X category.

There are more characteristics with unexplainable slopes. Clay is one of the reasons why unequal settlements occur. Also, a change in ground water level is an argument for damage cause unequal settlements. Moreover, the chance that damage is caused by C.X is very small if no ground settlements were detected. The linear regression results in illogical influences of those characteristics.

The only context characteristic that seems plausible is the earthquake load measured in peak ground velocity (PGV Bommer et al [mm/s]). Higher values of this characteristic will indicate that earthquake damage causes are more likely.

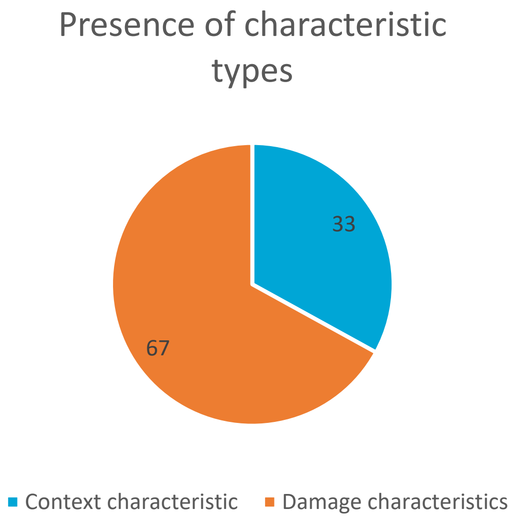


Figure 81 Presence of characteristic types in linear regression results to damage cause unequal settlement (C.X)

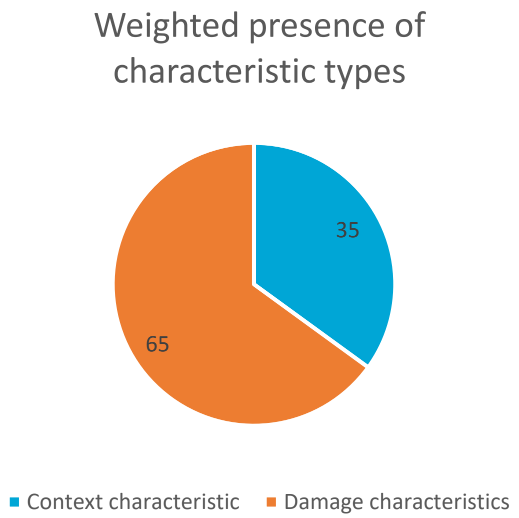


Figure 82 Presence of characteristic types in linear regression to damage cause unequal settlement (C.X), adjusted to their absolute mean slope.

When examining the different characteristic types, building characteristics are not noted in the linear regression result of damage cause unequal settlement (C.X.). According to the found pattern, building related characteristics have no influence in the occurrence of C.X. causes. Although building dimensions and weight have an effect on soil settlements, it is more likely that context related characteristics like soil buildup have an impact on unequal settlements. Context characteristics make up 33% of the regression results. The other 67% are the damage characteristics. These percentages will hardly change if they are adjusted to the characteristic slope.

Concluding, the found relations are not explainable with hypotheses. It is even contradictory of what was expected. Therefore, the found linear regression is not meaningful for practice.

7.10 Disclaimer

Within this research, there are two factors that could make the presented results uncertain. First, the way in which the provided data is created. Secondly, the calculation of the reliability scores.

The provided data were created by different experts. Although several measures were applied to increase the reliability of the results (like a structured approach with instructions and using the four eye principle) it cannot be fully excluded that there might be slight differences in the way the analysis was performed by different experts. Therefore, there might be slight inconsistencies in the damage reports.

Also, inconsistencies in the information gathering process had an influence on the presented analysis. Although this process was well structured, each expert had a method of his own of describing case characteristics. Especially damage characteristic were difficult to collect consistently from the data. Not all damage cases consist of pictures or clear sketches. For example, crack appearance characteristics were not always noted, such as in case a crack was stepped or contained branches. Also, some data were missing or were not stored as expected in reports. These data had to be determined afterwards with open source material or were found at other locations in the report.

Another problem with the provided data is that unintentional incorporated relations are found with the pattern recognition. This could lead to false evidence of wrong relations between characteristics and damage cause. For example, damage in which earthquakes had a trigger function. This might lead to unrelated earthquake causes with the PGV characteristic. However, these damage causes may easily occur at location without an earthquake load. Therefore, these relations are not useful for practice.

The last issue in relation to the provided data regards the research location. The data only consist of Dutch buildings located in Groningen. Therefore, the found relation and produced tools are only applicable for buildings located in that area.

The second factor of uncertainty is the way how the results of the pattern recognition are evaluated. The presented reliability scores for regression analysis are calculated as average of the cross-validation results. A fair calculation would be done with an independent dataset. Unfortunately, the construction of this independent dataset would take too much time, because it is an extensive manual task to construct this validation set and it fell out the scope of this graduation research. Note that the classification analysis are evaluated with an independent test set.

8 Conclusion

In this chapter, an answer is formulated on the main question of this thesis: Which useful relations are possible between case characteristics and damage causes? In here, case characteristics are elements in a description of a damage case. It could be information about a building or its surrounding, but it also includes an explanation of damage itself. Damage causes are categorised in 40 categories.

A few useful relations were found between case characteristics and damage cause. However, it was not possible to produce useful relations for all damage cause categories. Due to the limited available data, pattern recognition was able to produce relations in six damage cause categories.

To determine if a found relation is useful, it must be reliable and meaningful. A summary of this is made in Table 11 for classification analysis and Table 12 for regression analysis. In these tables, each damage cause pattern is judge on how reliable and meaning they are. If both judgements are positive about the found pattern, the conclusion about how useful the pattern is should also be positive (last column in table).

8.1 Classification relations between case characteristics and damage cause

Classification analysis finds relations between characteristics and a certain cause category. It indicates the cause of damage based on a selection of characteristics. It does not say how much that category contributes to the occurrence of damage. It only tells if a cause category is the cause of damage.

Damage caused by *initial insufficient resistance to bear loads* (A.1.1) did not have a clear relation with characteristics. Pattern recognition was able to produce a classification decision tree. A performance test shows that this pattern is accurate in 51% on the testcases. According to that pattern, damage located above openings in walls is an indication of this damage cause. Also, older buildings seem more vulnerable to insufficient resistance damage.

Damage caused by *overloading due to vibrations by earthquakes* (A.3.5) had a significant relation with the peak ground velocity (PGV). Pattern recognition was able to produce a decision tree. A performance test shows this pattern is accurate in 61% on the testcases. According to that decision tree, earthquakes contributes in most cases to the occurrence of A.3.5 damage if it had a PGV higher than 1.65 mm/s. Also, buildings newer than 40.5 years are vulnerable to earthquake related damage. It is possible to apply the decision cautiously in practices.

Damage caused by *initial hindered deformations* (B.1.1) had a significant relation with the location of damage. Pattern recognition was able to produce a classification decision tree with an accuracy of 77%. Which means that the pattern was right in 77% of the test cases. According to that decision tree, B.1.1 damage was often recognised with damage located at the inside of a building.

Damage caused by *initial imposed deformations* (B.2.1) did not have a clear relation with characteristics, because the results were not reliable. Although the found decision tree was right in 69% of the test cases, the results between different pattern recognition runs was rather large (a standard deviation of 12,1% on the accuracy). This can indicates that the provided data is too small in size. Also, it could mean that there is no relation between characteristics and this cause category possible.

Damage caused by *unequal settlements with equal loads* (C.1.1) did not have a clear relation with characteristics, because the results were not reliable. The found decision tree was right in 43% of the test cases. Also, the results between the different pattern recognition runs was rather large (a mean of 0.21, with a std. of 0.26). This may indicate that the provided data is too small in size. Also, it could mean that there is no relation between characteristics and this cause category possible.

Damage caused by *unequal settlements* (C.X) did not have a significant relation with characteristics, because the results were not reliable. It predicted only the right result in 33% of the predictions. It indicates that buildings from before 1922 contribute to the occurrence of unequal settlement damage. Also, trees signal that unequal settlements play a part in the appearance of damage. However, characteristics were not always explainable. Also, the reliability of this pattern is very low, compared with the others (33%). Therefore, the relations in this pattern are not useful for practice.

Classification conclusions

Code	Damage cause	Reliability	Meaningful	Conclusion (useful)
A.1.1	Initial insufficient resistance to bear loads	51% accuracy	Damage located above opening, building age **Not:** Distance to road	**No**, Too big decision tree, not clear relations, not explainable context characteristics.
A.3.5	Overloading due to vibrations by earthquakes	61% accuracy	PGV load, age of building, cracks in joint and unit. **Not:** relative influence tree, distance to road	**Cautious use,** good performance of PGV split, other characteristics could be explainable. Reliability high, but could be better.
B.1.1	Initial hindered deformations	77% accuracy	Damage located inside **Not:** age of building	**No,** better meaningful relations possible. However, highly reliable.
B.2.1	Initial imposed deformations	69% accuracy	Concrete foundation material, relative influence of trees, damage near openings **Not:** longest side of building, PGV load, distance to road	**No,** large differences between various cross-validations, illogical characteristics
C.1.1	Unequal settlements with equal loads	43% accuracy	Presence of trees and roots growth, PGV load, construction year, number of damage cases per area	**No,** not reliable.
C.X	Unequal settlements	34% accuracy	Construction year, Relative influence of trees. **Not:** possible accumulation of rain or snow	**No,** not reliable.

Table 11 Summery of classification conclusions

8.2 Regression relations between case characteristics and damage cause

Regression relations indicate how much a characteristic has contributed to the occurrence of a certain damage cause. Building damage rarely appears because of just one reason. Mostly, multiple causes have contributed to manifestation of damage. The share of an individual cause in the occurrence of damage is called technical attributability. The produced regression relations predict this technical attributability for a certain damage cause.

Prediction of the technical attributability was difficult for the pattern recognition, because of the large number of options in this regression analysis. To obtain a result for this thesis, the regression analysis were heavily steered by manually selecting characteristics for this pattern recognition. Despite this interference, it was only possible to achieve sufficient patterns for three damage cause categories.

Damage caused by *overloading due to vibrations by earthquakes* (A.3.5) had a significant relation with the peak ground velocity (PGV). In a lower amount, trees located near a building also had a positive influence on the contribution of earthquake in the occurrence of damage. Another relation is that older buildings will reduce the share of damage cause A.3.5 in damage. The linear regression was applied in order to generate these relations. 48% of the data was describable with this linear regression (a coefficient of determination of 0.48). Despite that, these found relation can help with improving earthquake damage investigations. In that sense, the produced relations are useful.

Damage caused by *deformations* (B.X) did not have a clear relation with characteristics, because the found relations were not always explainable or logical. Moreover, the linear regression model had a relative low coefficient of determination of 0.28. This means that 28% of the data could be described with the applied model. Therefore, the generate relations are not useful for practice.

Damage caused by *Unequal settlements* (C.X) did not have a clear relation with characteristics, because the proposed characteristics were not always explainable or logical. Moreover, the linear regression model had a low coefficient of determination of 0.18. This means that 18% of the data could be described with the applied model. Therefore, the generate relations are not useful for practice.

Regression conclusions

Code	Damage cause	Reliability	Meaningful	Conclusion (useful)
A.3.5	Overloading due to vibrations by earthquakes	0.48 Coefficient of determination*	PGV load, age of building, outside damage **Not:** relative influence tree, branched cracks	**Cautious use,** Good performance of PGV and age of building.
B.X	Deformations	0.28 Coefficient of determination*	Horizontal crack, ceilings, masonry, crack width **Not:** building renovations	**No,** Low reliability, not meaningful relations.
C.X	Unequal settlements	0.18 Coefficient of determination*	PGV load **Not:** Settlement differences over the length of the building. Clay soil, Change in Ground water level,	**No,** Low reliability, not meaningful relations

Table 12 Summery of regression conclusions (*not determined with an independent test set)

9 Recommendations

There are a few options to improve the pattern recognition on damage cases.

The benefit of incorporating more damage cases in the data is explained in chapter 4. Both for the pattern recognition in general (see 4.1.4) as for one specific characteristic (see 0). The conclusion of both analysis was that enlarging the data has an influence on the result. In the case of the PGV threshold, it will become more in line with values from literature if more data is integrated. There are more data cases available, but the manual inventively process took too much time for this master thesis. Concluding, better results could be obtained if this data is applied in the dataset.

Be more consistent in storing data about building damage. As discussed in the chapter above, the provided data were sometimes inconsistent in the information gathering process. This may have an influence on the results.

Incorporate more damage cases from outside the Groningen area. The data only consist of building damage where earthquake loads were present. Building damage cases without earthquake loads can make the found relation applicable for more areas. Also it provides a better insight on PGV related damage causes.

Be more selective in characteristics. The selection of characteristics was kept broad in order not to rule out any unanticipated relation between characteristics and damage cause. With the findings of this report, next pattern recognition research could be more focused on a smaller selection of characteristics.

Explore other types of pattern recognition methods for regression analysis. Maybe other relation finding methods are better at finding regression relations between characteristics and damage causes. Mathematical progressive relations would be advisable to oscillating or parabolic relations, because it is not expected that an optimum exists in characteristics. An increase of an characteristic should always lead to an increase of damage. Also in the selection of an algorithm, it needs to be able to handle both discrete and continuous values. Most characteristics are discrete values. However, the most important values are continuous values, according to the now applied algorithms.

10 References

Aeneas Media. (2019, Januari 24). *Scheurpatroon*. (Aeneas Media) Retrieved Oktober 19, 2021, from BetonLexicon: https://www.betonlexicon.nl/S/Scheurpatroon/

Akkar, S., & Bommer, J. J. (2007, April). Empirical Prediction Equations for Peak Ground Velocity Derived from. *Bulletin of the Seismological Society of America, 92*(2), 511-530. doi:10.1785/0120060141

Anzai, Y. (1989). *Pattern Recognition and Machine Learning.* London, England: Academic Press Limited. Retrieved from https://tudelft.on.worldcat.org/oclc/25410182

Badr, W. (2019, Febuari 22). *Having an Imbalanced Dataset? Here Is How You Can Fix It.* (T. Editors, Editor, & Towards Data Science Inc.) Retrieved 12 21, 2021, from Towards data science: https://towardsdatascience.com/having-an-imbalanced-dataset-here-is-how-you-can-solve-it-1640568947eb

Beyerer, J., Richter, M., & Nagel, M. (2017). *Pattern Recognition: Introduction, Features, Classifiers and Principles.* Berlin, Germany: Walter de Gruyter GmbH. Retrieved from https://ebookcentral-proquest-com.tudelft.idm.oclc.org/lib/delft/reader.action?docID=5159257

Bommer, J. J., Dost, B., Edwards, B., Stafford, P. J., van Elk, J., Doornhof, D., & Ntinalexis, M. (2016, Februarie). Developing an Application-Specific Ground-Motion Model for Induced Seismicity. *Bulletin of the Seismological Society of America, 106*(1), 158-173. doi:10.1785/0120150184

Bommer, J. J., Stafford, P. J., & Ntinalexis, M. (2016). *Empirical Ground-Motion Prediction Equations for Peak Ground Velocity from Small-Magnitude Earthquakes in the Groningen Field Using Multiple Definitions of the Horizontal Component of Motion.* Imperial College (London). Assen: NAM. Retrieved from https://nam-feitenencijfers.data-app.nl/download/rapport/8b292dcd-7507-4067-b0d1-5564e407c9d1?open=true

Borsje, H., & de Richemont, S. (2011). *Methodiek voor onderzoek naar de oorzaak van gebouwschade - versie 2.* Technical Sciences. Delft: TNO.

Braakman, T. (2021, maart 21). *Bevingsschadeloket twijfelt aan deskundigen: 1500 melders moeten wachten.* Retrieved september 2021, 28, from RTVNoord: https://www.rtvnoord.nl/nieuws/803822/Bevingsschadeloket-twijfelt-aan-deskundigen-1500-melders-moeten-wachten

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., . . . Holt , B. (2020). *1.10. Decision Trees*. Retrieved 8 2021, 17, from Scikit-learn: https://scikit-learn.org/stable/modules/tree.html

Cambridge University Press. (2021). *Cambridge Dictionary*. Retrieved 11 16, 2021, from Cambridge Advanced Learner's Dictionary & Thesaurus: https://dictionary.cambridge.org/dictionary/english/

Concrete Construction Staff. (2001, 1 21). *Troubleshooting: Map Cracking*. (Zonda Media) Retrieved 11 29, 2021, from Concrete Construction: https://www.concreteconstruction.net/how-to/repair/troubleshooting-map-cracking_o de Lange, G., Oostrom, N., Dortland, S., Borsje, H. (., & Richemont, S. (. (2011). *Gebouwschade Loppersum.* Delft: Deltares. Retrieved September 30, 2021, from https://www.groninger-bodem-beweging.nl/wp-content/uploads/2018/04/gebouwschadeloppersum.pdf

De Vent, I. A. (2011). *Structural damage in masonry.* TU Delft. Delft: TU Delft. Retrieved from http://resolver.tudelft.nl/uuid:ced08aa2-c2e8-4aa8-9559-870bf12d5ba5

Dekking, F. M., Kaaikamp, C., Lopuhaä, H. P., & Meester, L. E. (2005). *A Modern Introduction to Probability and Statistics: Understanding Why and How.* London, United Kingdom: Springer-Verlag.

Deval, S. (2020, 8 8). *Mean Squared Error – Explained | What is Mean Square Error?* Retrieved 2 2, 2022, from Great learing: https://www.mygreatlearning.com/blog/mean-square-error-explained/

Edwards, J. (2005, Feb). Cracking in historic masonry and surface finishes: Conservation approaches and solutions. *Journal of Building Appraisal, 1*(1), 20-33. doi:10.1057/palgrave.jba.2940003

Emmit, S., & Gorse, C. A. (2014). *Barry's Introduction to construction of buildings* (3e ed.). Chichester, West Sussex, united kingdom: John Wiley & Sons. Retrieved 2021, from https://ebookcentral-proquest-com.tudelft.idm.oclc.org/lib/delft/reader.action?docID=1662769&ppg=4

Evers, L. (Director). (2019). *Drie veelgestelde vragen over aardbevingen door gaswinning* [Motion Picture]. Nederland: YouTube. Retrieved septermber 30, 2021, from https://www.youtube.com/watch?v=rwW-2ma0ycU

GBB. (2021, 1 1). *Geschiedenis*. (Groninger Bodem Beweging) Retrieved September 30, 2021, from Groninger Bodem Beweging: https://groninger-bodem-beweging.nl/gbb/geschiedenis/

Geurts, C. P. (2017). *Comparison between TNO sensor data and GMPE-PGV values.* TNO.

Gevrilova, Y. (2020, 10 21). *Pattern Recognition and Machine Learning in Simple Words*. Retrieved 11 16, 2021, from Serokell: https://serokell.io/blog/pattern-recognition-in-simple-words

Harrison. (2016, April 10). *Practical Machine Learning Tutorial with Python Introduction*. (PythonProgramming.net) Retrieved 2020, from Python Programming: https://pythonprogramming.net/machine-learning-tutorial-python-introduction/

Instituut Mijnbouwschade Groningen. (2021, Mei 17). *Bevingen en trillingssterktes in Groningen, hoe zit dat?* Retrieved from instituut Mijnbouwschae Groningen: https://www.schadedoormijnbouw.nl/nieuws/bevingen-en-trillingssterktes-in-groningen-hoe-zit-dat

Japkowicz , N., & Sthephen, S. (2002, November 15). The class imbalance problem: A systematic study. *Intelligent Data Analysis, Vol 6*(No. 5), 429-449. doi:10.3233/IDA-2002-6504

Kahneman, D. (2013). *Thinking, fast and slow* (First paperback edition ed.). New York, USA: Farrar, Straus and Giroux,.

KNMI. (n.d.). *Seismische meetstations*. (KNMI) Retrieved 3 17, 2021, from Koninklijk Nederlands Meteorologisch Instituut: https://www.knmi.nl/kennis-en-datacentrum/uitleg/seismische-meetstations

Kok, J., de Jong, B., & Kastelein, J. (2015). *Onderzoeksgebied 'Steenbergen' (gebouwschade in relatie tot NAM-Activieteiten Langelo).* Arcadis. Rotterdam: Arcadis Nederland BV. Retrieved Oktober 1, 2021, from https://www.nam.nl/nieuws/2015/onderzoeksresultaten-schademeldingen-steenbergen/_jcr_content/par/textimage.stream/1468338387059/14c611154301a965996650702a8bc4be1efea17d/main-report-research-steenbergen.pdf

Korswagen, P., Jonkman, S., & Terwel, K. (2019, Januari). Probabilistic assessment of structural damage from coupled multi-hazards. (M. Stewart, Ed.) *Structural Safety*(76), pp. 135-148. doi:https://doi.org/10.1016/j.strusafe.2018.08.001

Malomo, D., DeJong, M., & Penna, A. (2019, 12 12). Influence of Bond Pattern on the in-plane Behavior of URM Piers. (P. B. Lourenço, & P. Roca, Eds.) *International Journal of Architectural Haritage*. doi:10.1080/15583058.2019.1702738

Mishra, G. (2016). *Types of Bounds in Brick Masonry Wall Construction and their Uses*. (The Constructor) Retrieved 6 30, 2021, from The Constructor Building ideas: https://theconstructor.org/building/types-bonds-brick-masonry-flemish-english-wall/11616/

Mohajari, N., Assouline, D., Guiboud, B., Gudmundsson, A., & Scartezzini, J.-L. (2017, 12 28). A city-scale roof shape classification using machine learning for solarenergy applications. *Renewable Energy, 121*, 81-93. doi:10.1016/j.renene.2017.12.096.

NLOG. (2016, augustus 9). *Groningen gasveld*. (Ministerie van Economische Zaken en TNO-AGE) Retrieved september 30, 2021, from Nederlandse Olie- en Gasportaal: https://www.nlog.nl/groningen-gasveld

Patch, K., & Smalley, E. (2004). *Pattern Recognition Technologies: Getting the Picture.* Boston: Technology Research News. Retrieved from https://ebookcentral-proquest-com.tudelft.idm.oclc.org/lib/delft/detail.action?docID=3411449

Patrovi, T., Fraundorfer, F., Azimi, S., Marmanis, D., & Reinartz, P. (2017, 5 31). Roof type selection based on Patch-based classification using deep learing for high resolution satellite imagery. *The International Archives of the Photogrammerty, Remote Sensing and Spatiol Information Sciences, XLII*(1/W1), 653-657. doi:10.5194/isprs-archives-XLII-1-W1-653-2017

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, É. (2011, 10). Scikit-learn: Machine Learning in Python. (M. Braun, Ed.) *Journal of Machine Learning Research (JMLR), 12*, 2825-2830.

Pruiksma, J., & Rózsás, Á. (2017). *Vibration levels at foundations of houses in Groningen due to induced earthquakes.* Delft: TNO. Retrieved 2021, from https://www.nationaalcoordinatorgroningen.nl/binaries/nationaal-coordinator-groningen/documenten/rapporten/2017/juni/30/rapport-vibration-levels/20170630+Rapport+Vibration+levels+at+foundations+of+houses+in+Groningen+due+to+induced+earthquakes.PDF

Ratay, R. T. (2010). *Forensic structrual engineering handbook* (2e ed.). New York, USA: McGraw-Hill.

Rijksoverheid. (2002). *Mijnbouwwet.* Den Haag: Rijksoverheid. Retrieved Oktober 1, 2021, from https://wetten.overheid.nl/jci1.3:c:BWBR0014168&z=2021-05-22&g=2021-05-22

Rokach, L., & Maimon, O. (2005, October 24). Top-down induction of decision trees classifiers - a survey. (V. Marik, Ed.) *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 35*(4), 476-487. doi:10.1109/TSMCC.2004.843247

Salet, T., & Bruurs, M. (2017). *Schadeonderzoek Groningen Buitengebied.* Deventer: Witteveen+Bos.

SBR. (2017). *SBR Trillingsrichtlijn A: Schade aan bouwwerken: 2017.* Delft: SBRCURnet. Retrieved from https://www.naa.nl/wp-content/uploads/2019/04/SBR-Trillingsrichtlijn-A-Schade-aan-bouwwerken-2017.pdf

Scarpas, A., Erkens, S. M., Dollevoet, R. P., & Houben, L. J. (2015). *College Dictaat CTB3320, Weg- en Railbouwkunde* (Vol. Deel 4. Constructief ontwerp van spoorwegen). (V. L. Markine, Ed.) Delft, Zuid-Holland, Nederland: Technische Univeriteit Delft.

Shah, S. A., Arshad, H., Muhammad , F., Syed, R. S., Mudasser, K. M., Sunera, I., . . . Waseem, M. (2019, 10 14). Sustainable Brick Masonry Bond Design and Analysis:An Application of a Decision-Making Technique. (D. Shi, Ed.) *Applied Sciences, 9*(20). doi:10.3390/app920431

Shrestha, J. K., Predhan, S., & Gautam, D. (2020, 6 1). In-plane behavior of various brick bonds in masonry walls. (D. H. Chen, & K. M. ElZahaby, Eds.) *Innovative Infrastructure Solutions, 2020*(5). doi:https://doi-org.tudelft.idm.oclc.org/10.1007/s41062-020-00306-x

Stewart, J. (2012). *Calculus, Early Transcendentals* (7th edition ed.). Toronto, Canada: Brooks/Cole Cengage Learing.

Terwel, K., & Schipper, R. (2018). Innovative ways of dealing with existing problems: how to reliably assess the cause of damage of masonry structures in an area with man-induced earthquakes? *40th.* Nantes: IABSE.

van Herwijnen, F. (2009). *Leren van instortingen.* (C. van Eldik, Ed.) Zoetermeer, Zuid Holland, Nederland: Bouwen met Staal.

Van Maanen Winters, F. M., Vuggen, F. A., & Quaedvlieg, H. M. (2007, 3 13). *Ingestorte balkons complex Patio Sevilla te Maasricht.* (G. Biesmans, Ed.) Retrieved September 28, 2021, from Raad voor de Rechtspraak: http://deeplink.rechtspraak.nl/uitspraak?id=ECLI:NL:RBMAA:2007:BA0574

van Staalduinen, P., Terwel, K., & Rots, J. (2018). *Onderzoek naar de oorzaken van bouwkundige schade in Groningen.* TU Delft, Faculteit Civiele Techniek en Geowetenschappen. Delft: TU Delft. Retrieved 2020, from https://d1rkab7tlqy5f1.cloudfront.net/News/2018/07_Juli/TU%20Delft/Onderzoek%20oorzaken%20bouwkundige%20schade%20Groningen%20-%20TU%20Delft%20CITG-CM-2018-01%2011%20juli%202018.pdf

Verbeek, I. (2017, April 7). *Foutje, bedankt NAM!* Retrieved Oktober 1, 2021, from Aardbevingen in Groningen: https://aardbevingeningroningen.net/2017/04/07/foutje-bedankt-nam/#more-4121

Verruijt, A. (2018). *An Introduction to Soil Mechanics* (30 ed.). (S. M. Hassanizadeh, Ed.) Cham, Zwiterland: Springer.

Vree, J. d. (n.d.). *Metselverbanden, voorbeelden*. (Joostdevree.nl) Retrieved 6 30, 2021, from JoostdeVree: https://www.joostdevree.nl/shtmls/metselverbanden_voorbeelden.shtml

Webb, A. (2002). *Statistical Pattern Recognition* (2e ed.). (P. Heinemann, Ed.) Chichester, West Sussex, England: John Wiles & Sons, Ltd. Retrieved from https://tudelft.on.worldcat.org/oclc/763160180

Wensen, A. v. (2012). *Bomen zijn soms een gevaar voor de fundering* . Nieuwegein: Kennis Centrum Aanpak Funderingsproblematiek. Retrieved Juli 27, 2021, from http://www.kcaf.nl/site/images/uploads_pdf/2012.03.19_Bomen_te_dicht_bij_de_funderin g.pdf

Appendix 1: Discussion of characteristics

All characteristics are discussed in this appendix. There are three main categories of characteristics: building characteristics, Context characteristics and damage characteristics. Building characteristics say something about the function, materials and size of the building. Context characteristics say something about sub soil, vibrations sources and external forces in the surrounding of the building. Damage characteristics say something about how damage occurs in terms of position, location and shape of damage.

Building characteristics

Building characteristics say something about the function, materials and size of the building. The possible building characteristics according to Staalduinen et al are: Constructions year, function, type of building, dimensions of the building, type of façade, type of roof, presence of a basement, foundation (type and material), supporting walls, inner walls, presence of dilatations, possible accumulation of rain and snow, indications of leaks, building renovations and state of maintenance.

Construction year

Construction year is the year a building is constructed.

Why construction year is important for damage investigation?

De Vent explains that the factor time is important, which is included in the building characteristic 'construction year'. According to her, damage is the result of a change. This changing event starts at a moment in time and has a certain duration (De Vent, 2011, p. 143). Elaborating on this, she proposed 3 ways to connect the factor time (construction year) with a description of damage:

- Damage has appeared after an event (De Vent, 2011, p. 144),
- Damage has appeared gradually, over a considerable period of time (De Vent, 2011, p. 145),
- Damage has appeared in the first years after construction (De Vent, 2011, p. 145)

Based on this, the underlying factor of the hazard parameter 'construction year' is the time between construction and the first appearance of the damage.

Another topic to describe the importance of the building characteristic 'construction year' is explained in the research of Witteveen+Bos. They suppose that damage will occur more often when buildings are older (Salet & Bruurs, 2017, p. 25). There are two reasons for this. Firstly, is that quality of the used materials is improved of the years. Secondly, the quality of constructing a building is improved. Both improvement are based on the experience of the contractors from earlier building projects. This assumption was tested by Salet et al with lintels damage. Their result fulfils the hypotheses that older buildings has more damage (Salet & Bruurs, 2017, p. 25). The same result was for hindered deformations (Salet & Bruurs, 2017, p. 34). Also Witteveen+Bos stated that more imposed deformation damage occurs when buildings are from a recent construction year, because the buildings elements are getting bigger. Bigger building element will give bigger creep deformations (Salet & Bruurs, 2017, p. 36). This assumption was also tested by Salet et al. Also this result fulfils the hypotheses that there is a relation between construction year and the occurrence of damage.

The TU Delft made a classification with the construction year:

- < 1900
- 1900-1940
- 1940-1970

- > 1970

This classification is based on the improvement of the construction and foundation technology over the different building periods. The main difference is the used foundation. The most buildings after 1970 where build on a reinforced concrete when applying a shallow foundation, where building before 1970 mostly used masonry as material. Also the soil survey was improved and a different type of façade material was used after 1970. Most building from before 1970 and certainly before 1940 have a basement (van Staalduinen, Terwel, & Rots, 2018, p. 181).

Could construction year be excluded from a damage cause category?

Based on the research of the TU Delft it is expected that the construction year divided in classes has an influence on the damage cause C, a change in soil condition.

How is construction year applied in the data?

In summary, the building characteristic 'construction year' has the following elements in the data:

- The time in years between construction of the building and the appearance of the damage [yyyy], shortened to "Age of building"
- The construction year [yyyy]
- The construction year classification only for damage cause category C (soil condition):
    - < 1900 [yes = 1, no = 0]
    - 1900-1940 [yes = 1, no = 0]
    - 1940-1970 [yes = 1, no = 0]
    - > 1970 [yes = 1, no = 0]

Visualisation of the applied characteristics 'construction year'

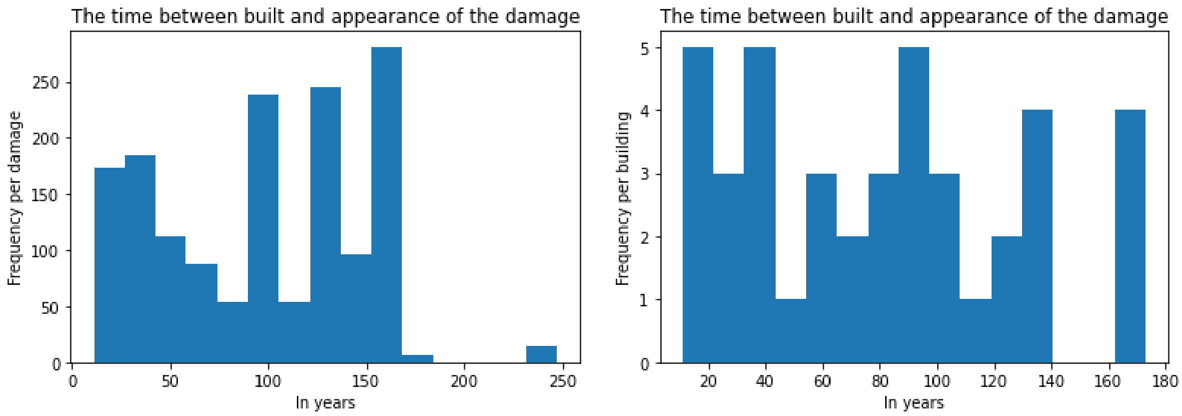


Figure 83 Histograms of characteristic 'the time between construction of the building and the appearance of the damage', no data is missing. Frequency per damage at the left, frequency per building at the right.

Some buildings have damage occurs at different moments in time. There are also buildings that were built in parts, with multiple construction years. This explains the differences between the two frequencies left and right at the x-axis.

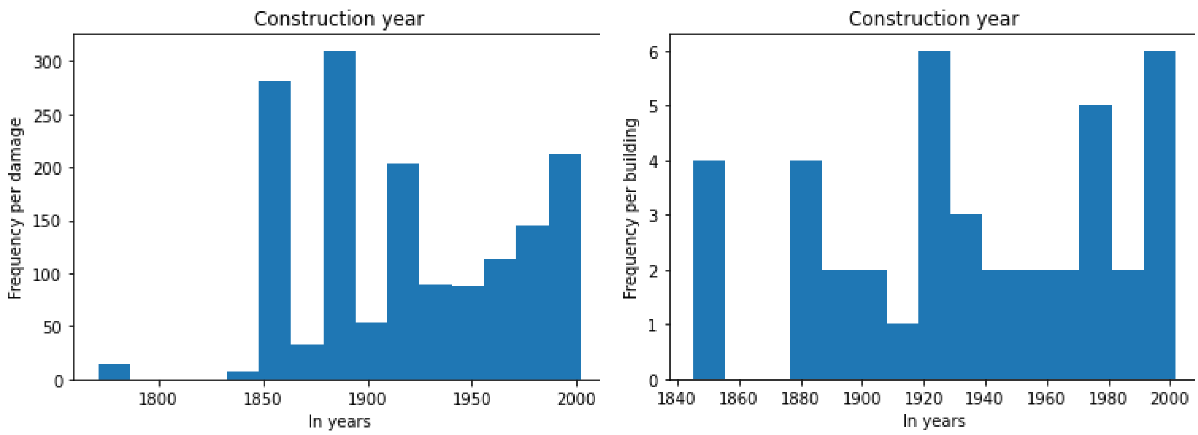


Figure 84 Histograms of characteristic 'construction year', no data is missing. Frequency per damage at the left, frequency per building at the right.

Some building are built in parts, with multiple construction years. This explains the differences between the two frequencies left and right at the x-axis.

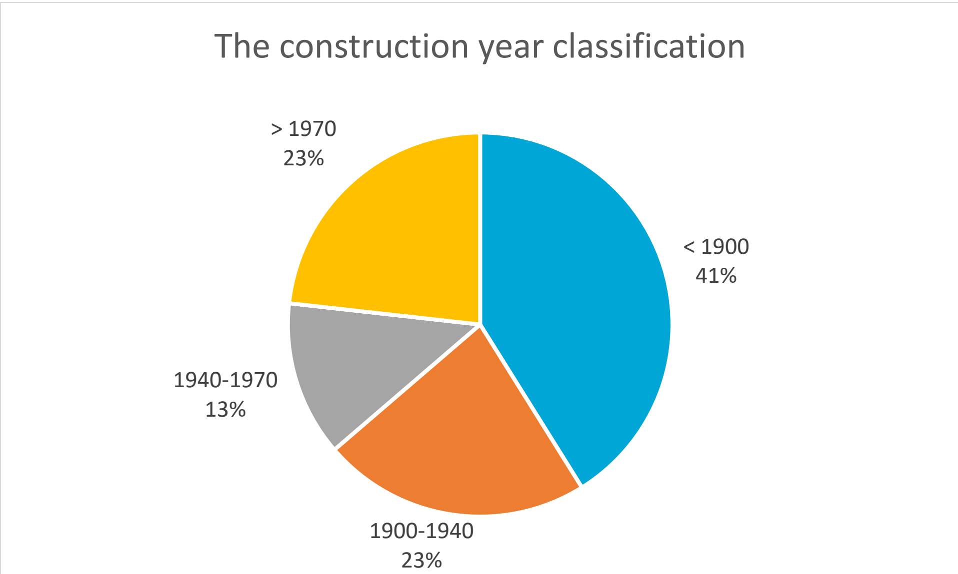


Figure 85 Spreading of characteristic construction year classification, no missing data.

Function

The function of a building is a description of how a building is used.

Why is the function of a building important for damage investigation?

When analysing damage, it is normal to collect the function of a building (Borsje & de Richemont, 2011, p. 5), but a relation between the function of a building and damage is not found by de Vent, Staalduinen et al or Salet et al. What they suppose is that the change of function could causing a damage: "However, it may also relate to changes in use of a building that involve change in use load…" (De Vent, 2011, p. 114). It would therefore be preferable to collect from the data if a change in use occurred during the lifetime of the building. Unfortunately this is often unknown and very rare in the available data.

How is the function of a building stored in the data?

Witteveen+Bos proposed the following answer possibilities on what type of function a building could have: Living, industry, sports, gathering, office, shopping, lodging, other, non-function and unknown (Salet & Bruurs, 2017, p. 79). Not all of this function are covered in the available data, because the TU

Delft research was concentration on houses. When searching though the data, the following types of functions are available as a characteristic:

- Living [yes = 1, no = 0]
- Farm [yes = 1, no = 0]
- Gathering [yes = 1, no = 0]
- Shed [yes = 1, no = 0]
- Industry [yes = 1, no = 0]

Visualisation of the applied characteristic 'function'

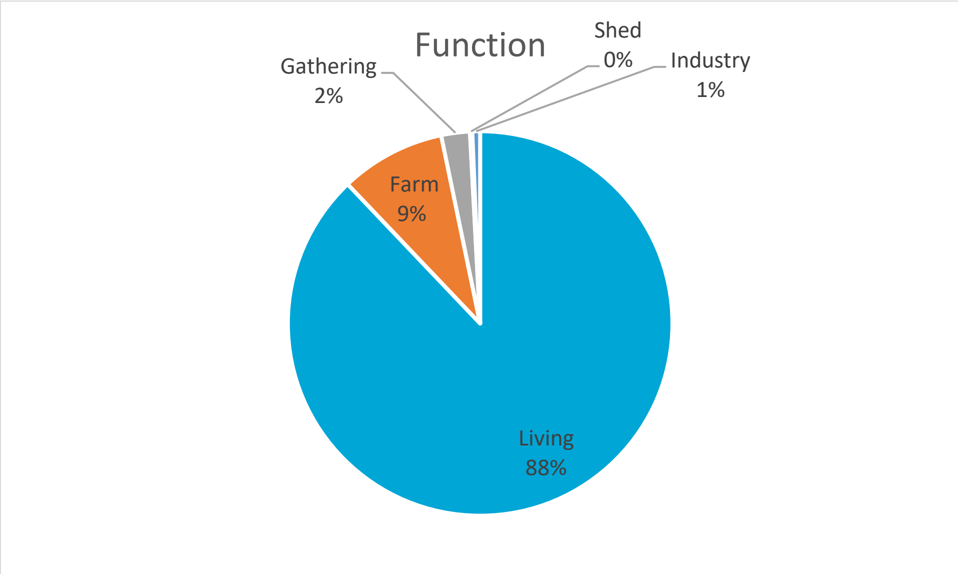


Figure 86 Spreading of characteristic function, no missing data.

Type of building

The building type explains the shape of the building.

Why is the building type important for damage investigation?

The TU Delft found a relation between building type and number of damages in a building (van Staalduinen, Terwel, & Rots, 2018, p. 196). Therefore it could be expected that this building characteristic may be a good indicator for a relation between characteristic and damage cause.

Witteveen+Bos concluded that lager buildings with more mass are less vulnerable for vibrations than relative smaller buildings (Salet & Bruurs, 2017, p. 204). For example, a detached house has more vibration damage than an large apartment building. Also, every building type has a typical roof construction. This will lead to different reaction against loads. Therefore, this could lead to different type of damage per building type. For example, semidetached houses and farmhouses are more vulnerable for horizontal forces from the roof (spatten van een dak) (Salet & Bruurs, 2017, p. 426). Which is in agreement with literature (Ratay, 2010).

Could building type be excluded from a damage cause category?

Based on this it is expected that the building characteristic 'type of building' has an influence on damage cause A.3.X Overloading because of vibrations.

How is the type of a building stored in the data?

Witteveen+Bos has the following building types: detached house, shed, semidetached, farmhouse, apartment, corner house, terraced house, farm, warehouse, sport hall, swimming pool, windmill, church, other and unknown (Salet & Bruurs, 2017, p. 79). TU Delft has the following available building types defined: detached house, terraced house, apartment(s) and semidetached house (van Staalduinen, Terwel, & Rots, 2018, p. 23). The following characteristic for type of building came across during the data gathering process:

- Detached house [yes = 1, no = 0]
- Semidetached house [yes = 1, no = 0]
- Terraced house [yes = 1, no = 0]
- Farm [yes = 1, no = 0]
- Apartment [yes = 1, no = 0]
- Public accessible [yes = 1, no = 0]
- Windmill [yes = 1, no = 0]
- Shed [yes = 1, no = 0]

Visualisation of the applied characteristic 'type of building'

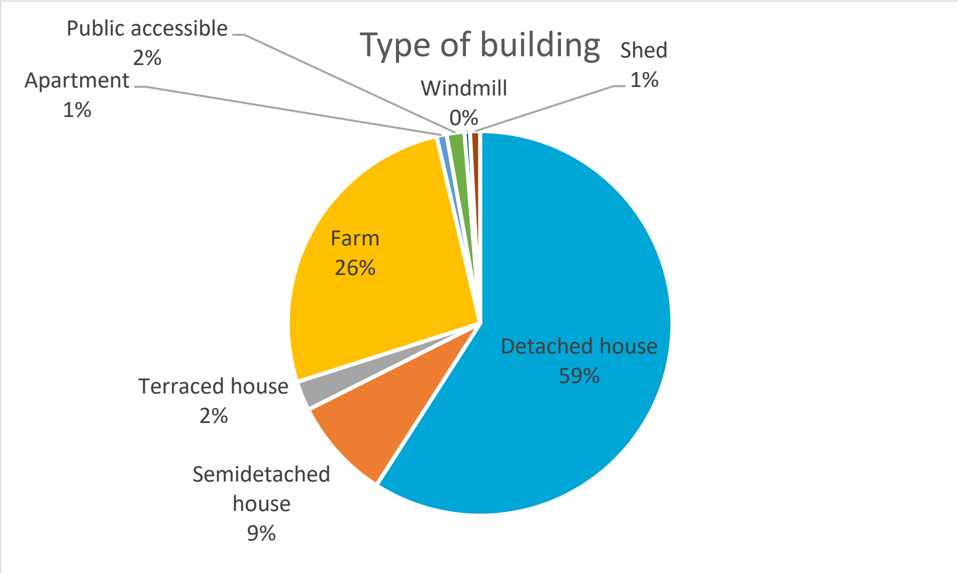


Figure 87 Spreading of characteristic type of building, no missing data.

Dimensions of the building

Measurements on the size of a building.

Why are the dimensions of a building important for damage investigation?

Staalduinen et al stated that one of the factors on damage caused by unequal settlements are the dimensions of the building (van Staalduinen, Terwel, & Rots, 2018, p. 137). It is difficult to formulate a if-then statement with this relation, because this is a relative relation. A building on a larger settlement area than itself, is expected to have less damage than a building on a smaller settlement area than itself. This is supposed by Witteveen+Bos (Salet & Bruurs, 2017, p. 70). Both reports described the dimensions of the building in terms of area.

Could the dimension of a building be excluded from a damage cause category?

Witteveen+Bos tried to find relations with the number of damages per area. They used this characteristic as an indicator (Salet & Bruurs, 2017, p. 23). They did not find a relation between this and earthquake vibrations (Salet & Bruurs, 2017, p. 38).

Another method on how the dimension of a building can lead to a damage is the size of the building component. It has an influence on the dimensional changes. Building materials experience dimensional changes caused for example by thermal or hydric influences. How bigger the material, how larger the deformation (De Vent, 2011, p. 102). Therefore it is expected that the dimensions of the building as an influence on the cause B.X (Deformations).

How are the dimensions of a building stored in the data?

The following possible answers on building characteristic 'dimensions of the building' are available from the data a characteristic:

- Area of building [m^2] for damage cause category B (deformations) & C (soil conditions)
- Number of damage per area [## / m^2]
- Dimensions of longest side of the building [m] for damage cause category B (deformations) & C (soil conditions)
- Dimensions of shortest side of the building [m] for damage cause category B (deformations) & C (soil conditions)

Visualisation of the applied characteristics 'dimensions of a building'

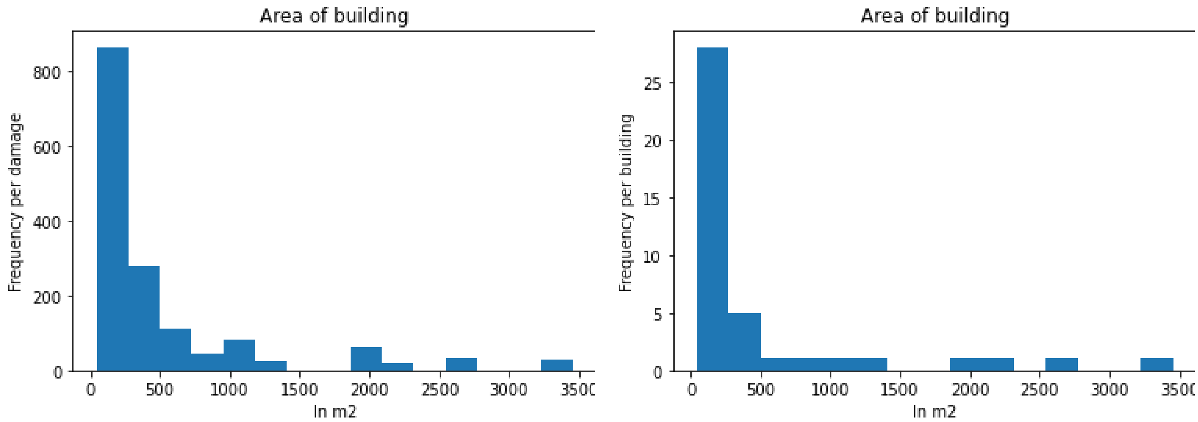


Figure 88 Histograms of characteristic 'area of building', no missing data. Frequency per damage at the left, frequency per building at the right.

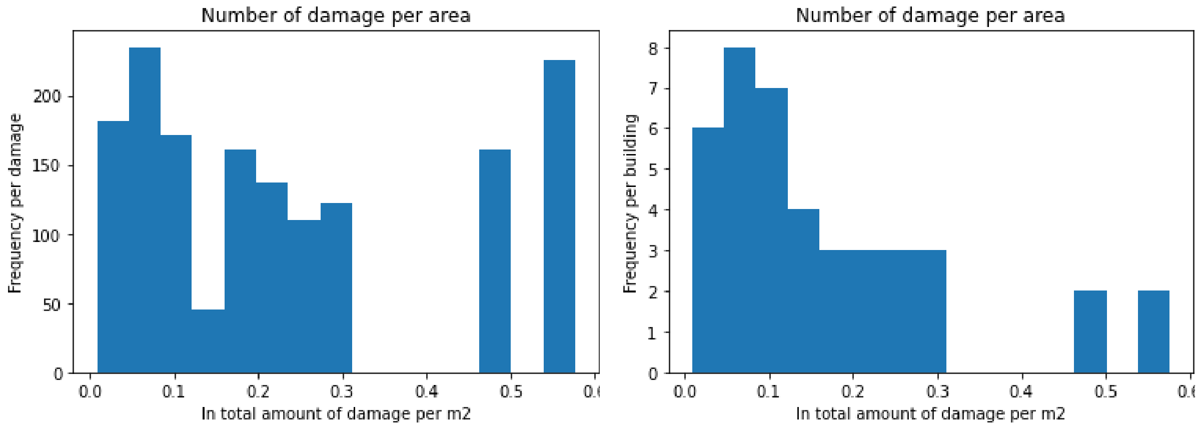


Figure 89 Histograms of the characteristic 'number of damage per area', no data is missing. Frequency per damage at the left, frequency per building at the right.

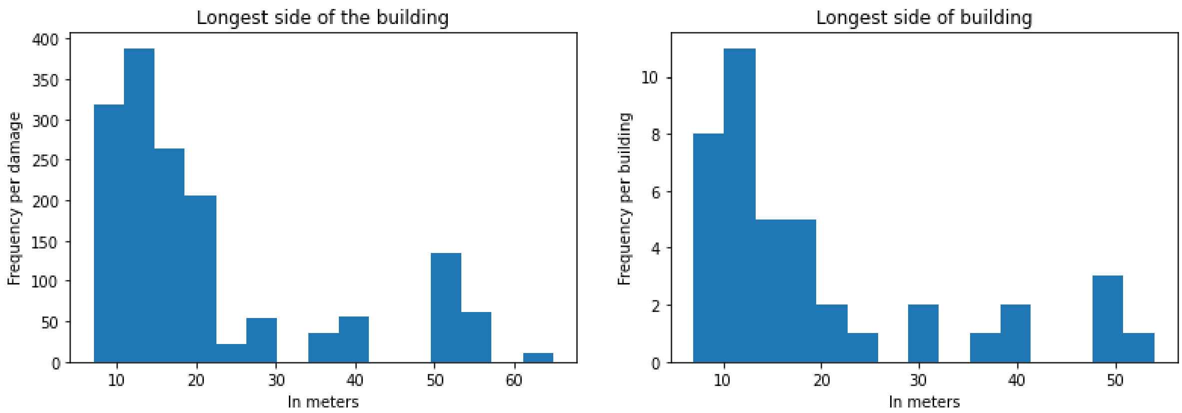


Figure 90 Histograms of characteristic 'longest side of the building, no data is missing'. Frequency per damage at the left, frequency per building at the right.

The x-axis differences between the two histograms. This has to do with if a building is connected to a relative significant larger shed. In that case, the dimensions of the building are adapted to the corresponding damage. This will result in different longest side of a building per building.

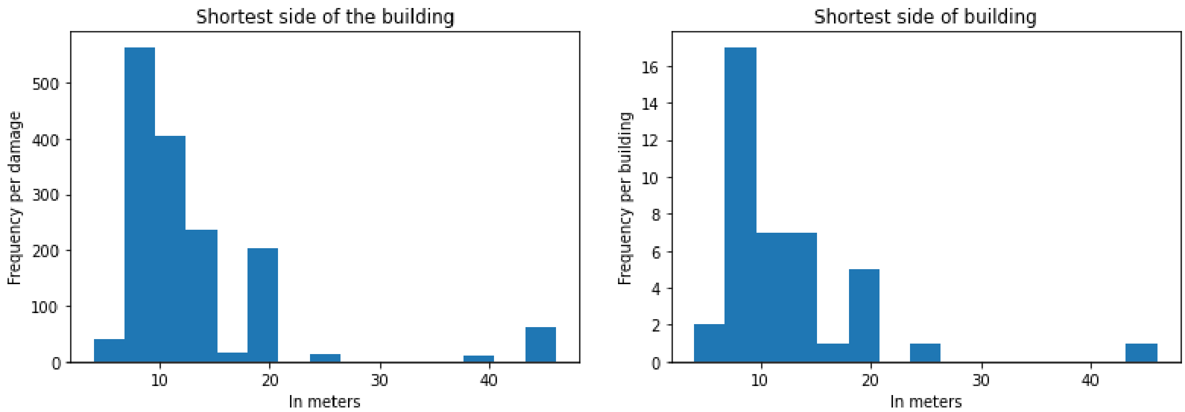


Figure 91 Histograms of characteristic 'shortest side of the building', no missing data. Frequency per damage at the left, frequency per building at the right.

The x-axis differences between the two histograms. This has to do with when a building is connected to a relative significant larger shed. In that case, the dimensions of the building are adapted to the corresponding damage. What will result in different shortest side of a building per building.

Type of façade

The description of the type of façade could be done in different ways.

First one is the used material. Concluding from the data there are two building materials used as a façade: masonry brickwork (99,8%) and timber. Timber is mostly used for sheds.

The second way to describe the type of façade is the thickness of the façade in terms the length of a brick (Emmit & Gorse, 2014, p. 172). A brick has a long side, the stretcher face, and a shorter side, the header face. If a wall is 1 brick thick (1 B), it means that the wall has a thickness of the stretcher length. Other façade thicknesses are: ½ B (header face), 1 ½ B and 2 B. It is unfortunately difficult to indicate the thickness of a façade without drawing of the building or partly deconstructing the façade.

The last way to describe the type of façade is the layout of the bricks, called the bond. There are 4 types of bond available in the data: Stretcher bond, Dutch bond, (English) cross bond and wild bond.

There are more type of masonry bond types in literature (Emmit & Gorse, 2014, p. 171) (Mishra, 2016) (Vree, sd), but this were the ones that came across in the data.

Why is the type of façade important for damage investigation?

The bound type does not matter when looking at the elastic behaviour of a wall, but there are differences between the bonds in the nonlinear phase. Bonds with more stretcher faces have a higher ultimate loading strength (Shrestha, Predhan, & Gautam, 2020). There are other properties of a masonry wall that have more influence on the strength of the wall (Shah, et al., 2019), but it could be concluding that the bond pattern may considerably effect on the behaviour of the wall (Malomo, DeJong, & Penna, 2019).

How is the type of façade stored in the data?

The following possible answers on building characteristic 'type of façade' are available in the data as a characteristic:

- Stretcher bond [yes = 1, no = 0],
- Dutch bond [yes = 1, no = 0],
- Cross bond [yes = 1, no = 0],
- Wild bond [yes = 1, no = 0].

Visualisation of the applied characteristic 'type of façade'

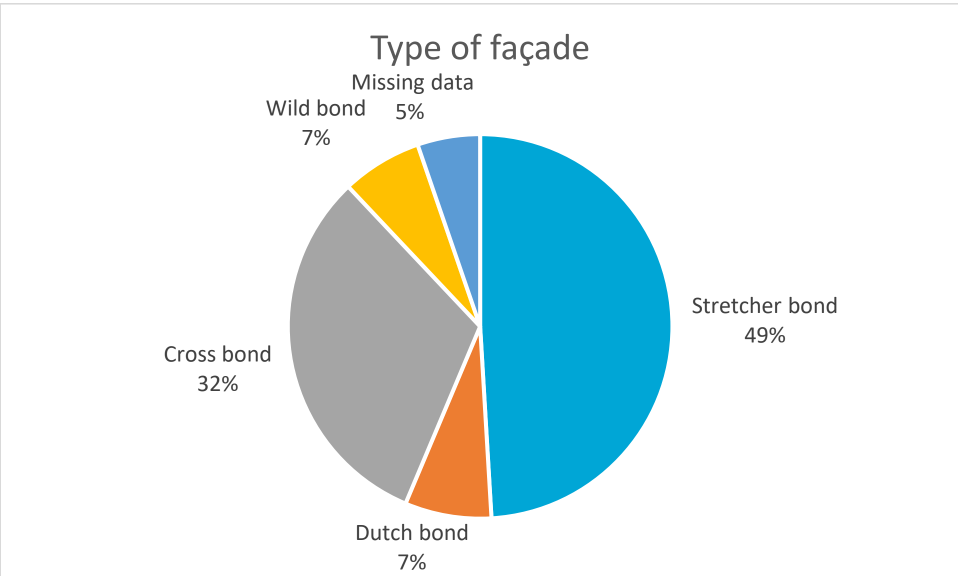


Figure 92 Spreading of characteristic type of façade, 5% of the data is missing.

What to do with the missing data?

Some data is missing because the provided pictures did not have a resolution high enough to see the header face out of the stretcher face stones. It is not possible to substitute the data with a mean, there is no mean of 'type of façade'. Two option remain: substitute the missing data with the most frequent one or ignore the missing data and proceed analyses. The impact of declaring a wrong type of façade is expected to be small, because it is expected that a relating between the type of façade and the damage cause is small. The relation didn't appear in the literature, although the cross bond with less stretcher faces would be a little bit less stronger than the stretcher bond. When substituting the missing data with stretcher bond, probability of a mistake is 48% and for cross bond is this 67%. For doing nothing and proceed as it is, the probability of a mistake is 100%. If risk is defined as impact times probability and the impact of all options is assumed the same, because it is that small, the option with the smallest risk is to substitute the missing data with stretcher bond.

Missing data in characteristic 'type of façade' is substitute with stretcher bond.

Type of roof

Ideally, this would by a description of the supporting construction of the roof. Unfraternally, this is unknown for almost every case in het database. What is known is the cover of the roof and sometimes the construction material of the roof, which was mostly timber.

Why is the type of roof important for damage investigation?

The type of roof is important for determine the cause of a damage, because there is a well-known fail mechanism involving the roof. In this mechanism, the horizontal forces from the roof are causing deformations or cracks in wall near de connection with the floors (De Vent, 2011, p. 136) & (Salet & Bruurs, 2017, p. 248).

Other fields of science made a categorisation on the layout of the roof. This categorisation is based on the diversity of roofs in the local area. In Munich city there were: flat, gable, half-hip, hip, pyramid, mansard and complex roofs (Patrovi, Fraundorfer, Azimi, Marmanis, & Reinartz, 2017). In the city of Geneva there were: Flat, shed, gabled, hipped, gambrel, mansard flat, mansard hipped, cross-hipped, cross-gabled, corner-hipped, corner-gabled, pyramidal and complex roofs (Mohajari, Assouline, Guiboud, Gudmundsson, & Scartezzini, 2017). Mohajeri et al reduced this 13 roofs shapes into 6 main roof-shape classes for a better presentable analysis. It results in: flat & shed, gable, hip & pyramidal, gambrel & mansard, cross and corner gable & hip and complex. Based on Partovi et al and Mohajeri et al is a selection made on the different layouts. This selection is based on the structural performances and consequences of the different layouts.

How is the type of roof stored in the data?

The following possible answers on building characteristic 'type of roof' are available in the data as a characteristic:

- Roof cover material
  - Rooftiles [yes = 1, no = 0],
  - Timber [yes = 1, no = 0],
  - Slates [yes = 1, no = 0],
  - Tar and gravel roof [yes = 1, no = 0],
  - Corrugated roofs [yes = 1, no = 0].
- Construction material roof
  - Timber [yes = 1, no = 0],
  - Steel [yes = 1, no = 0],
  - Concrete [yes = 1, no = 0],
  - Stone [yes = 1, no = 0].

Visualisation of applied characteristic 'type of roof'

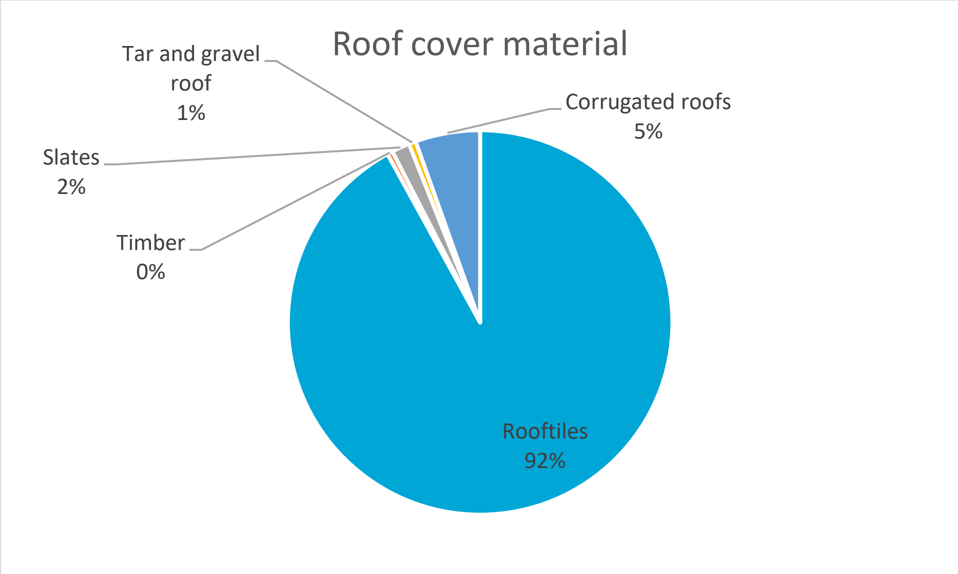


Figure 93 Spreading of characteristic 'Roof cover material', no missing data.

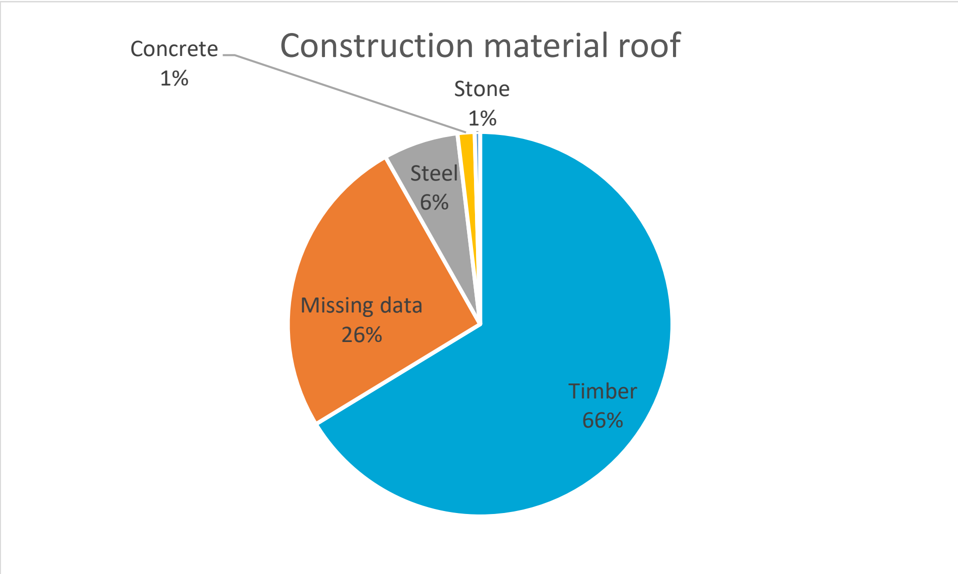


Figure 94 Spreading of characteristic 'Construction material roof', 26% of the data is missing.

What to do with the missing data?

Some data is missing in the characteristic *construction material roof*. The data is missing because: it was not mentioned in construction drawings, it was not possible to excess the roof construction for inspection or it was not mentioned in the case report. The part of missing data is too big to remove the corresponding damage cases. Two option remain: removing the characteristic or substitute the missing data. The impact of substitute the missing data is zero, according to input from the committee. Therefore, removing the characeteristic is a reasonable option.

The characteristic 'construction material roof' is removed from the analysis because too much data is missing.

Presence of a basement

If a room is located under the ground level, which is called a basement.

Why is the presence of a basement important for damage investigation?

A basement is an interruption in the foundation's building. This will result in a disruption of loads on the foundation. This can lead to damage. (Salet & Bruurs, 2017, p. 70 & 258) (van Staalduinen, Terwel, & Rots, 2018, p. 182).

How is the presence of a basement stored in the data?

The following possible answers on building characteristic 'presence of a basement' are available in the data as a characteristic:

- Presence of a basement [yes = 1, no = 0], only for damage cause category C (soil condition)

Visualisation of applied characteristic 'presence of a basement'

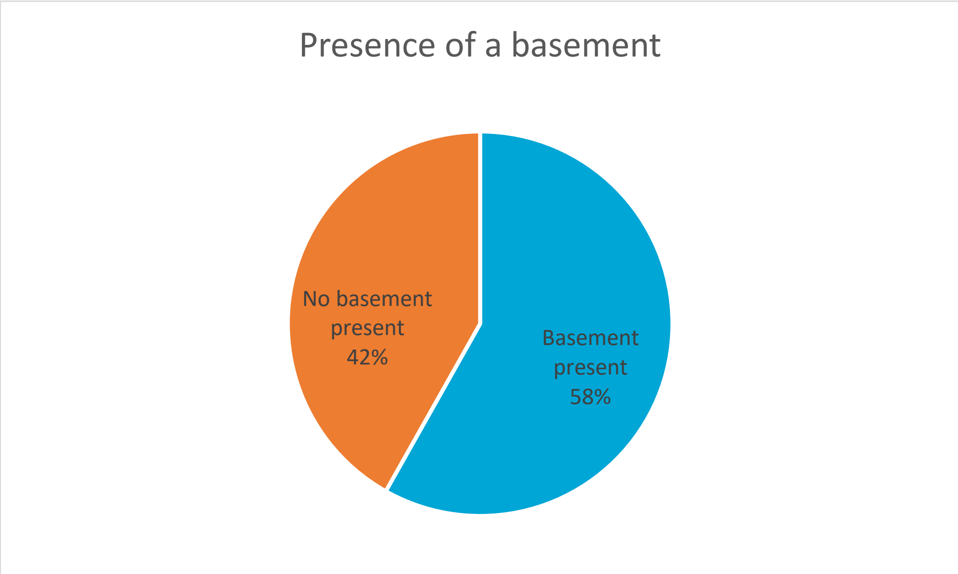


Figure 95 Spreading of characteristic 'Presence of a basement', no missing data.

Collecting this data from the case reports could be in violation of the 'WYSIATI' (What you see is all there is) rule for quality protection of data collection (Kahneman, 2013). If reports do not have mentions of a basement, it was notated as 'not present'. It is assumed that mentioning something that is not there is too obvious to do.

Foundation (type and material)

Description of the foundation in terms of type and material

Why is the foundation important for damage investigation?

The foundation transfers the loads of a building to the soil and is the a important topic in forensic engineering. The most common damage causes involving the foundation are: settlement, expansive soil, lateral movement and deterioration. Most of them fall in the damage category C.X (soil conditions), but A.3.5 (earthquakes) as also an influence. (Ratay, 2010, p. 16.2)

There a two main types of foundation: shallow and deep foundations. In the Netherlands, a swallow foundation is mostly a strip foundation and deep foundations are mostly driven piles. The applied material various between building, mainly because different building methods were utilized over the years.

How is the foundation stored in the data

The following possible answers on building characteristic 'foundation' are available in the data as a characteristic:

- Foundation type, only for damage category C.X (soil conditions) and A.3.5 (earthquakes)
  - Swallow foundation [yes = 1, no = 0],
  - Deep foundation [yes = 1, no = 0],
- Foundation material, only for damage category C.X (soil condition) and A.3.5 (earthquakes)
  - Concrete [yes = 1, no = 0],
  - Masonry [yes = 1, no = 0],
  - Timber [yes = 1, no = 0],
  - Calcium-Silicate [yes = 1, no = 0].

Visualisation of applied characteristic 'foundation'

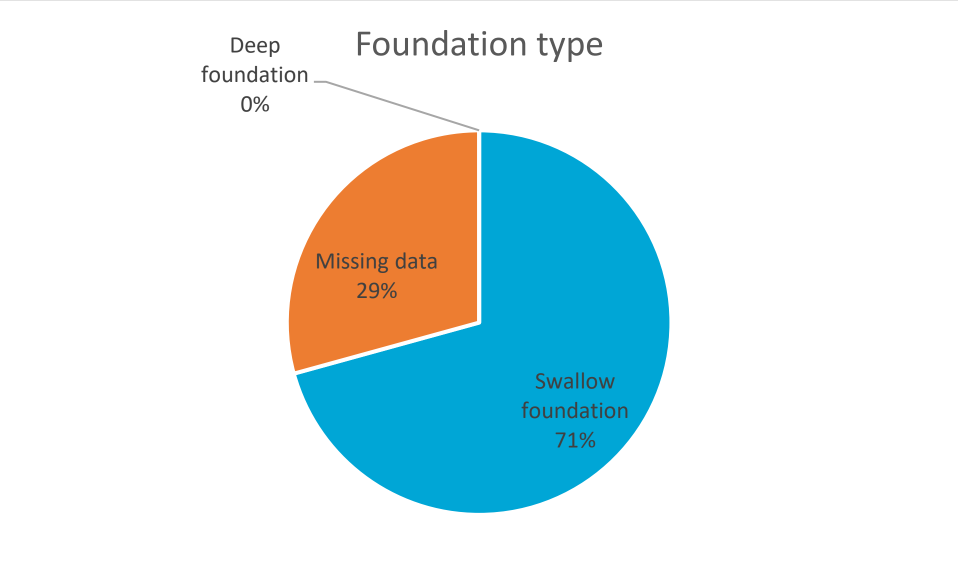


Figure 96 Spreading of characteristic 'foundation type', 29% of the data is missing.

What to do with the missing data?

Some data is missing because: the foundation type was not clear from the drawings, it was not investigate by excavate the foundation during inspection or it was not known by the owner. The impact of the foundation type on a damage case is big, according to literature (Ratay, 2010). The most desirably option would be to substitute the data, but the risk with that is too high, because it has a the large impact. The best option is to remove the characteristic, because too much data is missing. Also, when substitute the missing data with the most frequent, the characteristic has no influence on the pattern recognition anymore. In that situation, the characteristic 'foundation type' is 100% swallow foundation.

The characteristic 'foundation type' is removed from the analysis, because too much data is missing.

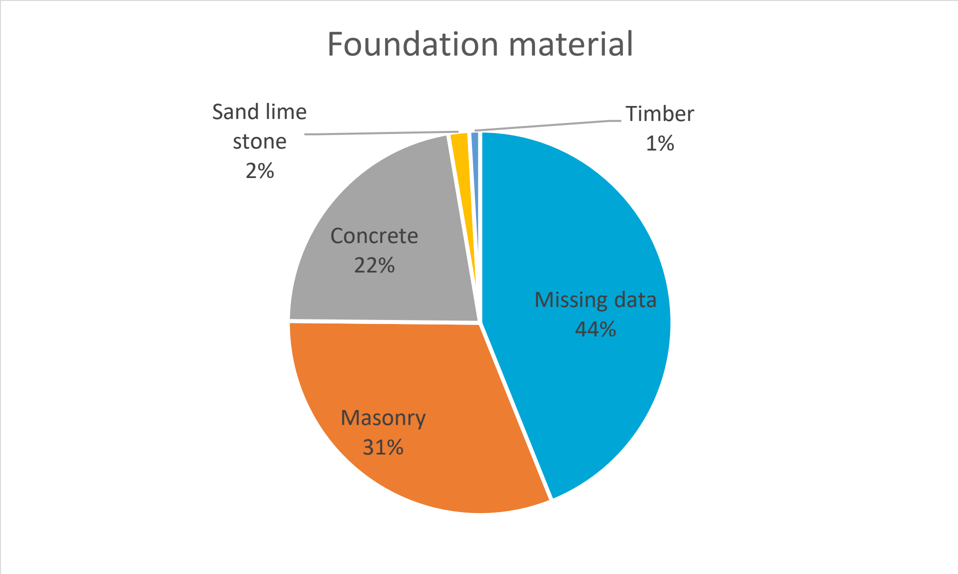


Figure 97 Spreading of characteristic 'foundation material', 31% of the data is missing.

Some data is missing in the characteristic 'foundation material'. This data is missing because drawings of buildings were missing or the foundation material was not written on it. Other ways to determine the foundation material would be to excavate a small hole down a façade. This was not always possible to do for the inspectors. The best option is to remove the characteristic from the analysis. Substitution of the missing data is difficult, because there is no significant most frequent used foundation material available.

The characteristic 'foundation type' is removed from the analysis because too much data is missing.

Supporting walls

Walls in a construction that transfers forces.

Why are supporting walls important for damage investigation?

The importance of supporting walls for forensic engineering does not need an explanation, although very little is said about it in literature. The most important is the relative layout of the supporting walls. A supporting wall can move or deform under its load. If this is hindered by other elements, like inner walls, it can cause damage (van Staalduinen, Terwel, & Rots, 2018, p. 150).

A description of the layout of each supporting wall in a building is difficult, because it is too unique, too much work and it is mostly unknown. Therefore, the description of supporting walls is reduced to the applied material.

How are supporting walls stored in the data?

The following possible answers on building characteristic 'supporting walls' are available in the data as a characteristic:

- Material of supporting walls
    - masonry [yes = 1, no = 0],
    - Timber [yes = 1, no = 0].

Visualisation of applied characteristic 'supporting walls'

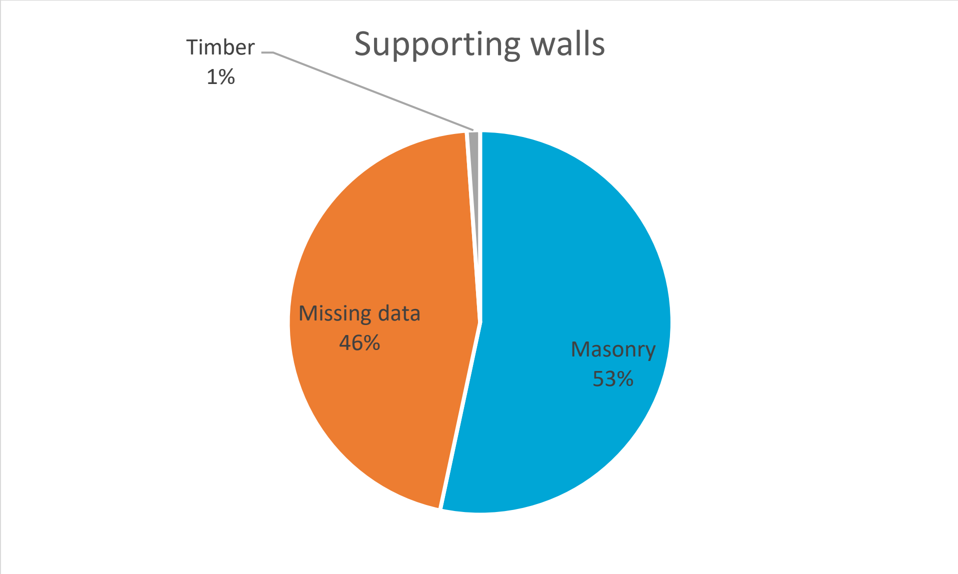


Figure 98 Spreading of the characteristic 'supporting walls',46% of the data is missing.

What to do with the missing data?

Some data is missing in the characteristic 'supporting walls', because it was not always clear what a supporting wall is and what material is used. Therefore, the best option is to remove this characteristic from the analysis.

The characteristic 'supporting walls' is removed from the analysis, because too much data is missing.

Inner walls

Wall in a building that not transfers forces.

Why are inner walls important for damage investigation?

A significant part of the damage cases are in an inner wall. For that reason, also the type of inner wall of a building is collected.

How are inner walls stored in the data?

The following possible answers on building characteristic 'inner walls' are available in the data as a characteristic:

- Material of inner walls,
  - Masonry [yes = 1, no = 0],
  - Sand lime stone [yes = 1, no = 0],
  - Timber [yes = 1, no = 0].

Visualisation of applied characteristic 'inner walls'

Figure 99 Spreading of the characteristic 'inner walls', 80% of the data is missing.

What to do with missing data?

Some data is missing in the characteristic 'inner walls', because it was not always clear what an inner wall was. Also, the applied material in a wall was not available. Therefore, the best option is to remove this characteristic from the analysis.

The characteristic 'inner walls' is removed from the analysis because too much data is missing.

Presence of dilatations

Dilatations are joints that can accommodate deformations.

Why are the presence of dilatations important for damage investigation?

Materials can deform for different reasons. One of this reasons is creep. There are 4 types of creep: chemical creep, plastic creep, temperature creep and hygrometric creep. These deformation, like creep, can cause a damage. This type of damage can be prevented by applying dilatations in the construction (van Staalduinen, Terwel, & Rots, 2018, p. 34).

How is the presence of dilatation stored in the data?

The following possible answers on building characteristic 'Presence of dilatations' are available in the data as a hazard parameter:

- Presence of dilatations [yes = 1, no = 0] only for damage cause category B.X (Deformations)

Visualisation of applied characteristic 'presence of dilatations'

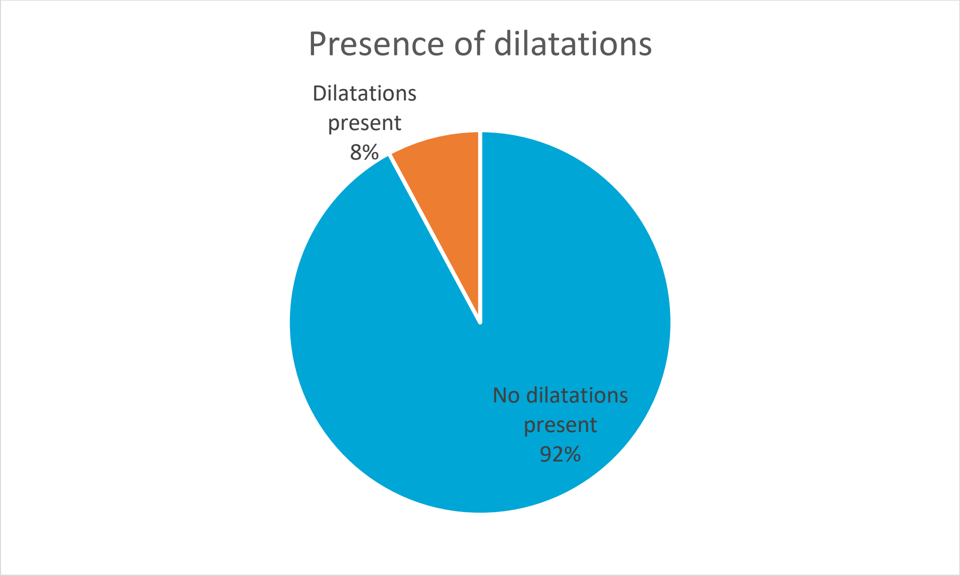


Figure 100 Spreading of characteristic 'presence of dilatations', no missing data.

Collecting this data from the case reports could be in violation of the 'WYSIATI' (What you see is all there is) rule for quality protection of data collection (Kahneman, 2013). If reports do not have mentions of a basement, it was notated as 'not present'. It is assumed that mentioning something that is not there is too obvious to do.

Possible accumulation of rain or snow

Snow or rain can be build up on flat roofs.

Why is possible accumulation of rain or snow important for damage investigation?

Roofs collapses because of rain or snow accumulation is a well-known damage case. Especially flat roofs are vulnerable for this type of damage (van Staalduinen, Terwel, & Rots, 2018, p. 33). One of the failure mechanisms is ponding (Ratay, 2010, p. 7.9), where flexible structural roof systems deflect under the weight of the rain- or snow water and form a small pond on the roof. When water drainage is not sufficient, this pond will grow in depth until eventually the roof will collapse.

How is the possible accumulation of rain or snow stored in the data?

The following possible answers on building characteristic 'Possible accumulation of rain and snow' are available in the data as a characteristic:

- Possible accumulation of rain and snow (presence of a flat roof) [yes = 1, no = 0] only for damage cause category A.X (Overloading)

Visualisation of applied characteristic 'possible accumulation of rain or snow

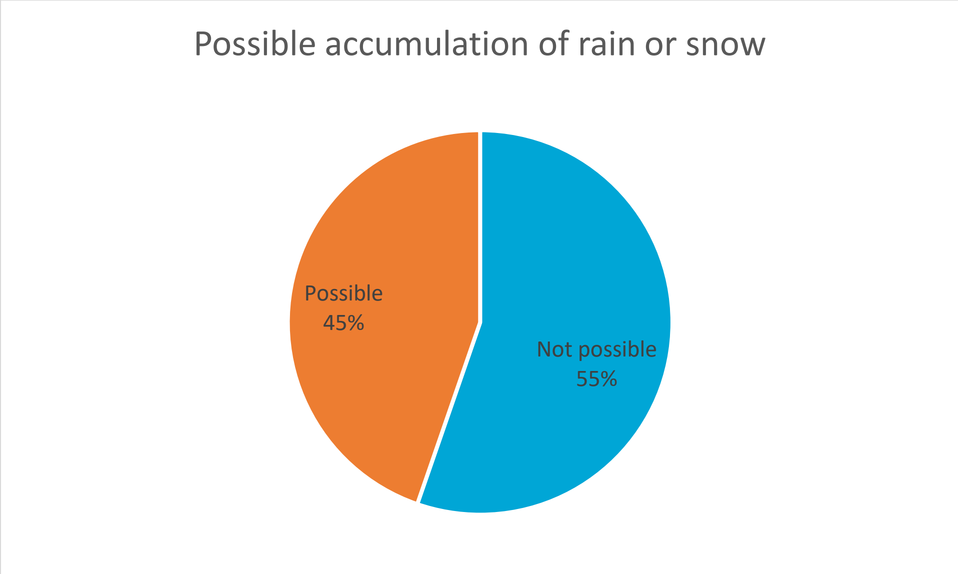


Figure 101 Spreading of characteristic 'possible accumulation of rain or snow', no missing data.

Indications of leaks

A leak is water at unwanted location.

Why is an indication of leaks important for damage investigation?

Leakage is an important factor in forensic engineering (Borsje & de Richemont, 2011, p. 3). A leak could damage a building in different ways or could be the result of a damage. It could damage a building by deterioration of construction materials, like timber (Ratay, 2010, p. 14.17) and steel (Ratay, 2010, p. 11.12) or by flooding the foundation what leads to settlements (Salet & Bruurs, 2017, p. 71). The other way around, a leakage could also be the result of hindered deformations of drainage of roof gutters (De Vent, 2011, p. 78).

How is an indication of leaks stored in the data?

The following possible answers on building characteristic 'Indications of leaks' are available in the data as a characteristic:

- Indications of leaks [yes = 1, no = 0].

Visualisation of applied characteristic 'indication of leaks'

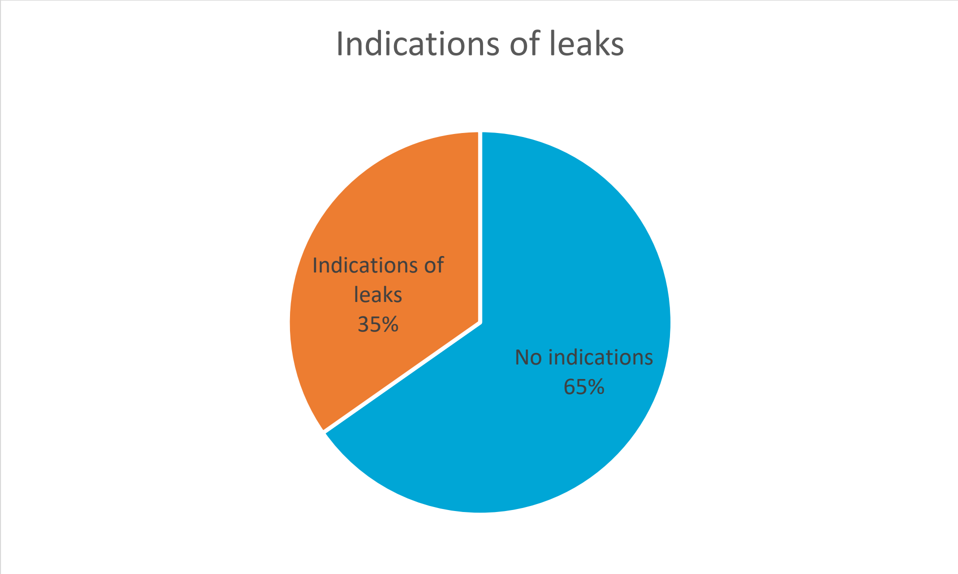


Figure 102 Spreading of characteristic 'indications of leaks', no missing data.

Collecting this data from the case reports could be in violation of the 'WYSIATI' (What you see is all there is) rule for quality protection of data collection (Kahneman, 2013). If reports do not have mentions of a basement, it was notated as 'not present'. It is assumed that mentioning something that is not there is too obvious to do.

Building renovations

Construction activities on a building, after a period of usage.

Why are building renovations important for damage investigations?

Building renovations could be explained as an active intervention in the loadbearing structure. This will result in a change of load path in the structure. This change could lead to damage in all categories (De Vent, 2011, p. 93). When the load path is changed, some forces may take a unintended route where the materials strength is insufficient, causing damage from cause category overloading (A.X). If building elements are changed for different ones, such as a different material or shape, the reaction on a load is different in the original situation. It could lead to another deformation, causing damage from cause category B.X (deformations). Also, home extensions will redistribute the forces on a foundation. This could lead to settlement differences within the building (Salet & Bruurs, 2017, p. 257)

How are building renovation stored in the data?

The following possible answers on building characteristic 'building renovations' are available in the data as a characteristic:

- Building renovations [yes = 1, no = 0].

Visualisation of applied characteristic 'building renovations'

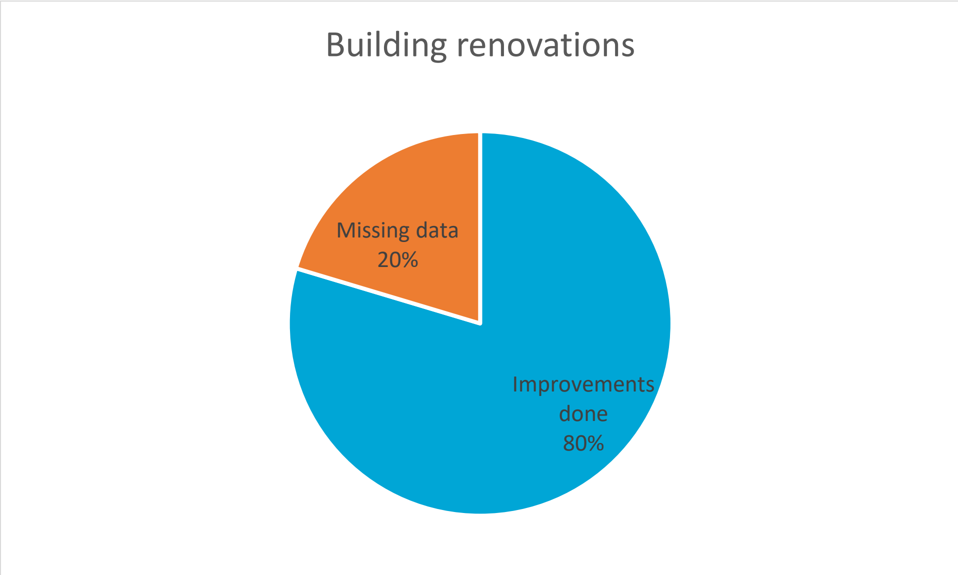


Figure 103 Spreading of characteristic 'building renovations', 20% of the data is missing

What to do with the missing data?

Some data is missing from the characteristic 'building renovations'. The data is missing because it was only mentioned what was improved to the building. It is not stated if no renovations have been made. Mostly because it is not known what previous owners of the building have done. The unknown cases are notated as missing data, according to the rule 'WYSIATI' (What you see is all there is) rule for quality protection of data collection (Kahneman, 2013). The best option is to substitute the missing data as 'no renovations done'. It is expected that only recent damage cases are in the database and that damage cases long before the memory of the owner are already repaired.

Missing data in characteristic 'building renovations' is substitute with no renovations done.

State of maintenance

How a building is kept in sufficient condition.

Why is the state of maintenance important for damage investigation?

Maintenance is necessary to control the natural deterioration of some materials. This is called scheduled maintenance, like repainting timber for protection against the weather. Other types of maintenance are corrective maintenance after a failure and condition-based maintenance at the end of a material life-time (Ratay, 2010, p. 9.17).

How is the state of maintenance stored in the data?

The judgement on the state of maintenance in the available data is subjective. It is mostly done from the office with images from Google Street View on a level of 'good' too 'insufficient' based on the cleanliness of the roof tiles, masonry walls and painted surfaces. A small motivation is given by the judgement, but it could different per reviewer.

The following possible answers on building characteristic 'state of maintenance' are available in the data as a characteristic:

- State of maintenance [1, 2, 3, 4 or 5] where,
    - Insufficient [= 1]
    - Sufficient [= 2]

- Pretty [= 3]
- Pretty good [= 4]
- Good [= 5]

Visualisation of applied characteristic 'state of maintenance'

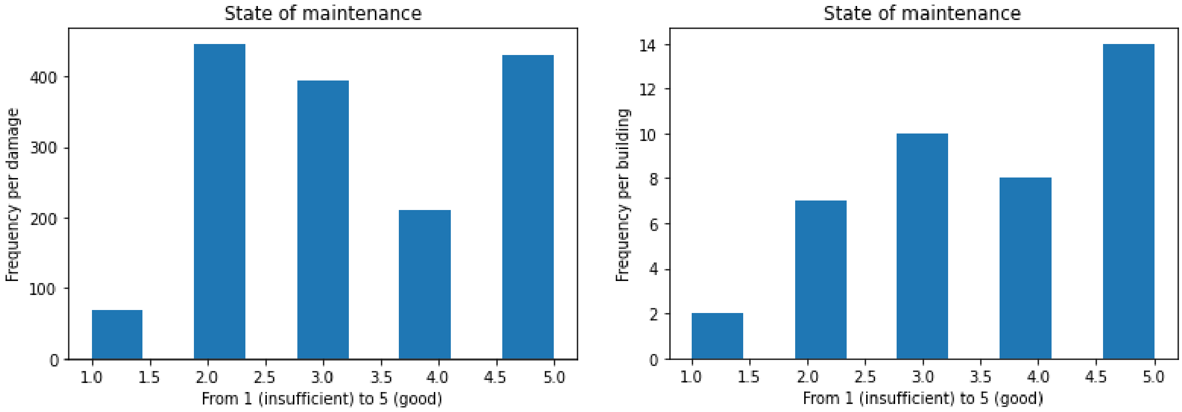


Figure 104 Histograms of characteristic 'state of maintenance', no missing data. Frequency per damage at the left, frequency per building at the right.

Context characteristics

Context characteristics tells something about the soil conditions, vibrations sources and external forces in the surrounding of the building. The context characteristics are: Soil build-up, changes on ground water level, presence of trees, ground settlement, presence of read traffic, presence of rail traffic, presence of construction activities, presence of industry activities, emergencies that have occurred and earthquakes (van Staalduinen, Terwel, & Rots, 2018, p. 41).

Soil build-up

The types of soil underneath a building.

Why is soil build-up important for damage investigation?

The type of soil underneath a building is determined based on the DINO loket, Geotop model and local hand boreholes or dynamic probing (van Staalduinen, Terwel, & Rots, 2018, p. 135). According to Verruijt is there a large variability in soil types. Even within a small area, a soil property could vary (Verruijt, 2018, p. 20). There is a unified classification system for soil categorisation: Gravel, Sand, Silt, Clay, Organic and Peat.

How is the soil build-up stored in the data?

The following types of soil were present in the Groningen area as a hazard parameter for context characteristic 'soil build-up':

- Shallow ground, only for damage cause category C (soil conditions):
    - Sand [yes = 1, no = 0],
    - Clay [yes = 1, no = 0],
    - Peat [yes = 1, no = 0].
- Deep ground, only for damage cause category C (soil conditions):
    - Sand [yes = 1, no = 0],
    - Clay [yes = 1, no = 0],
    - Peat [yes = 1, no = 0].

Visualisation of applied characteristic 'soil build-up'

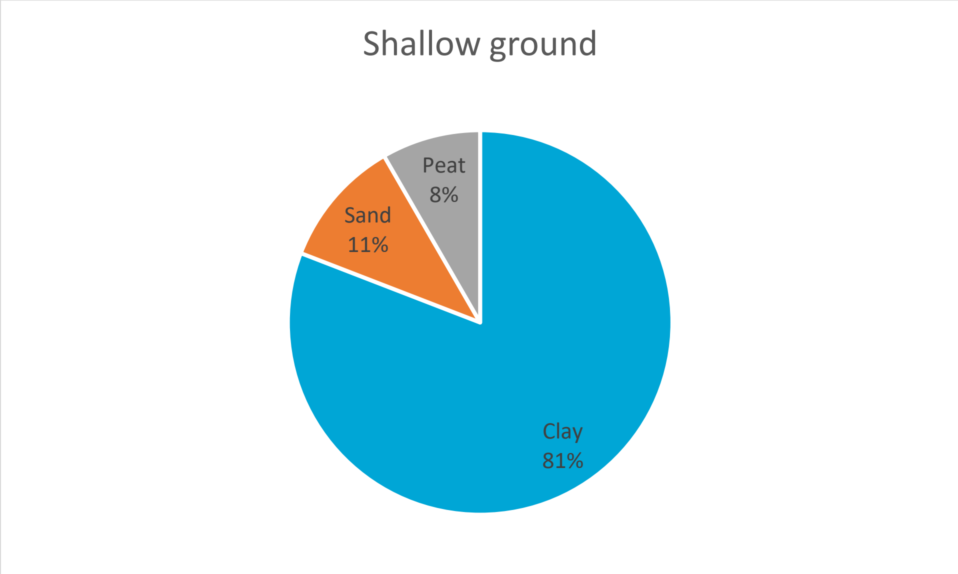


Figure 105 Spreading of characteristic 'shallow ground', no missing data.

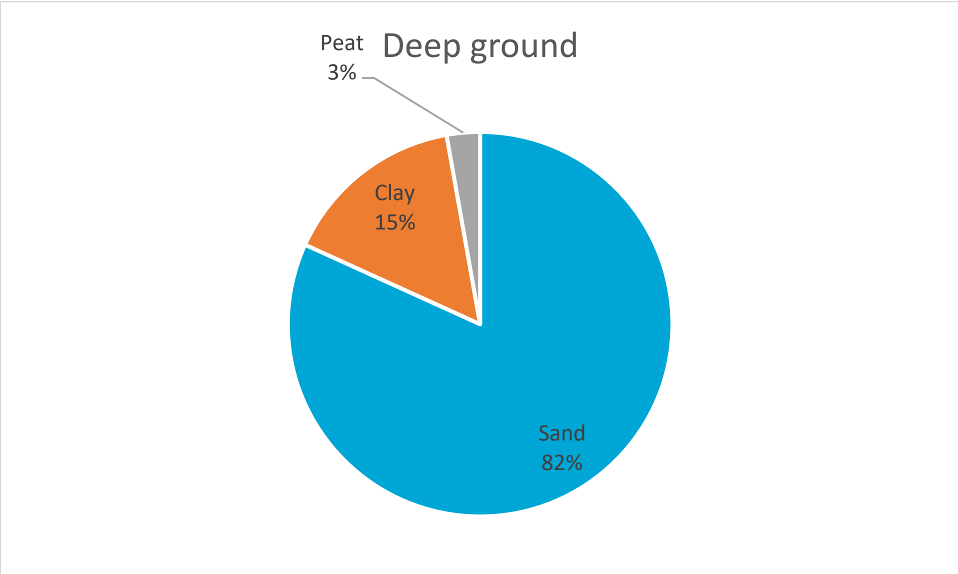


Figure 106 Spreading of characteristic 'Deep ground', no missing data.

Groundwater level

The depth when groundwater appears, measured from ground level.

Why is the groundwater level important for damage investigation?

There are different reasons why the groundwater level is important for forensic engineering. The change in groundwater level plays a part in it. A decline or increase of this level could result in a settlement of the ground level or peat oxidation (Salet & Bruurs, 2017, p. 137).

How is the groundwater level stored in the data?

The following possible answers on context characteristic 'groundwater level' are available in the data as a characteristic:

- Change in ground water level [yes = 1, no = 0] only for damage cause category C (soil conditions)

Visualisation of applied characteristic 'groundwater level'

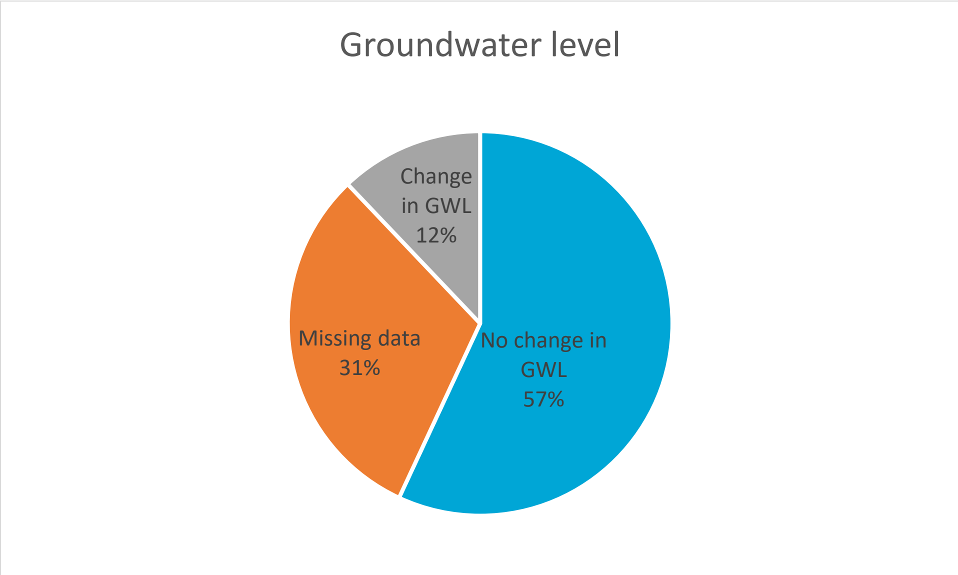


Figure 107 Spreading of characteristic 'ground water level', 31% of the data is missing.

What to do with the missing data?

Some data is missing in the hazard parameter 'change in ground water level', because historical information of the ground water level was not available at all location. The impact of a change in groundwater level is high and the probability of a mistake when substitute the data is significant. Therefore, when substituting the missing data, the risk of a mistake is relative big. Removing the characteristic or the missing data cases is also not desirable. The groundwater level will only change significantly if the waterboard decide to do so. Places where the groundwater level is unknown are locations it was not monitored by a waterboard. Therefore, the groundwater level at this places will not change.

Missing data in characteristic 'groundwater level' is substitute with no change in groundwater level.

Trees

Significant greenery that could have an influence on the building's construction.

Why are trees important for damage investigation?

There are two ways how a tree can cause damage at buildings. The roots of the tree plays a key part in both of them.

The first way is the (over)growth of the tree roots under foundation or façade. Salet et al calls this overloading (Salet & Bruurs, 2017, p. 146) as well as de Vent (De Vent, 2011, p. 85). Technically, tree roots are forcing parts of the building foundation away. Therefore, this damage cause is categories as imposed deformations (B.2) (van Staalduinen, Terwel, & Rots, 2018, p. 43).

The second way is by extraction of groundwater by the tree roots. This could cause settlements, what could result in a damage (van Staalduinen, Terwel, & Rots, 2018, p. 146) (Salet & Bruurs, 2017, p. 205). Therefor this damage cause is categorised as C.X (unequal settlements)

According to Salet et al, the influence area of trees which causes imposed deformation (B.2) is one times the width of the tree's crown. For unequal settlements (C.X), it is 3 times the crown's width (Salet & Bruurs, 2017, p. 243) (Wensen, 2012). In practice, the influence area is determined by measuring the height of the tree and the distance to the building (Salet & Bruurs, 2017, p. 66). According to the falsification of Staalduinen et al, there is no influence of tree roots growth if the distance to the building is larger than the height of the tree (van Staalduinen, Terwel, & Rots, 2018, p. 147). This suggest that the influence of a tree could descripted as the height of the tree divided by the distance to the building, as in:

$$R = \begin{cases} \frac{H}{d}, & H \geq d \\ H, & d = 0 \\ 0, & H < d \end{cases}$$

In which:

R is the relative influence of the tree,
H is the height of the tree in round meters,
d is the distance between the tree and the building in round meters.
Please note that this is own written formula. It is unknown if the formula is true. It is done to sum all factors of tree influence in one number for computer analyses.

How are trees stored in the data?

The following possible answers on context characteristic 'trees' are available in the data as characteristic:

- Presence of shrub [yes = 1, no = 0] only for B.X and C.X,
- Height of shrub [meters] only for B.X and C.X,
- Distance to shrub [meters] only for B.X and C.X,
- Relative influence shrub [m/m] only for B.X and C.X,
- Presence of tree [yes = 1, no = 0] only for B.X and C.X,
- Height of tree [meters] only for B.X and C.X,
- Distance to tree [meters] only for B.X and C.X,
- Relative influence tree [m/m] only for B.X and C.X.

Visualisation of applied characteristic 'trees'

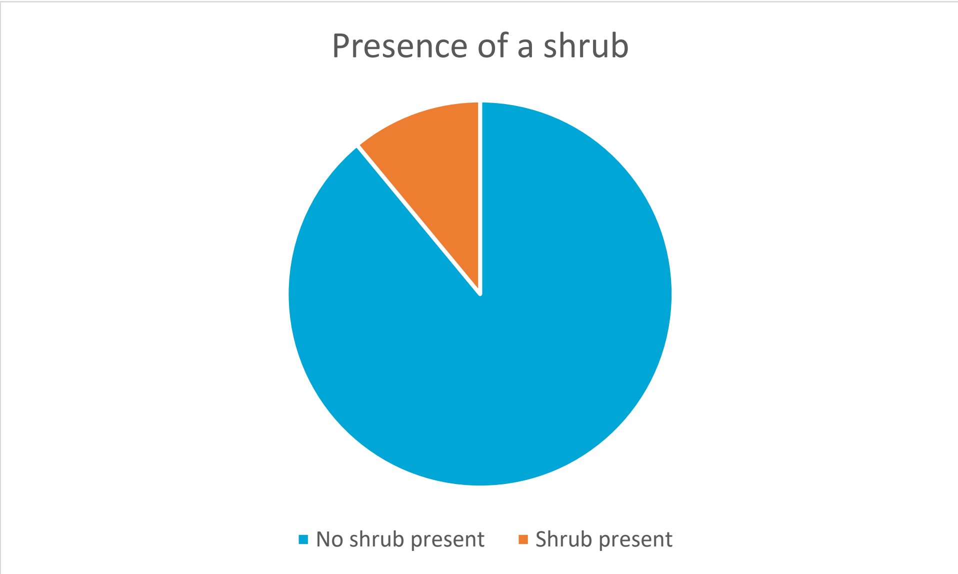


Figure 108 Spreading of characteristic 'presence of a shrub', no missing data.

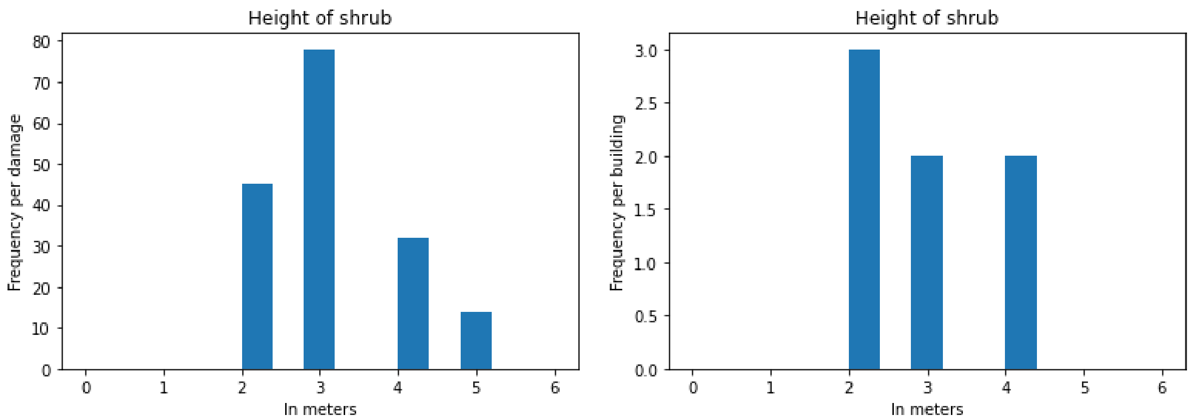


Figure 109 Histograms of characteristic 'height of shrub', no missing data. Frequency per damage at the left, frequency per building at the right.

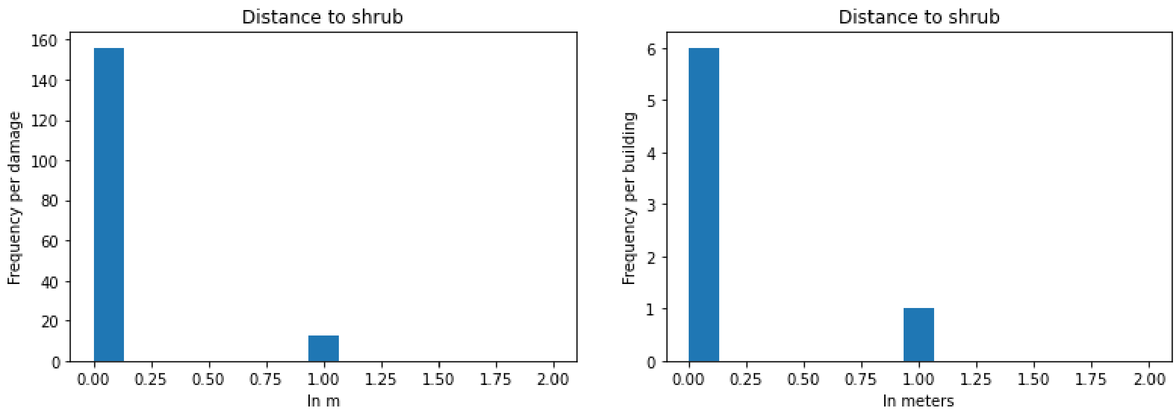


Figure 110 Histograms of characteristic 'distance to shrub', no missing data. Frequency per damage at the left, frequency per building at the right.

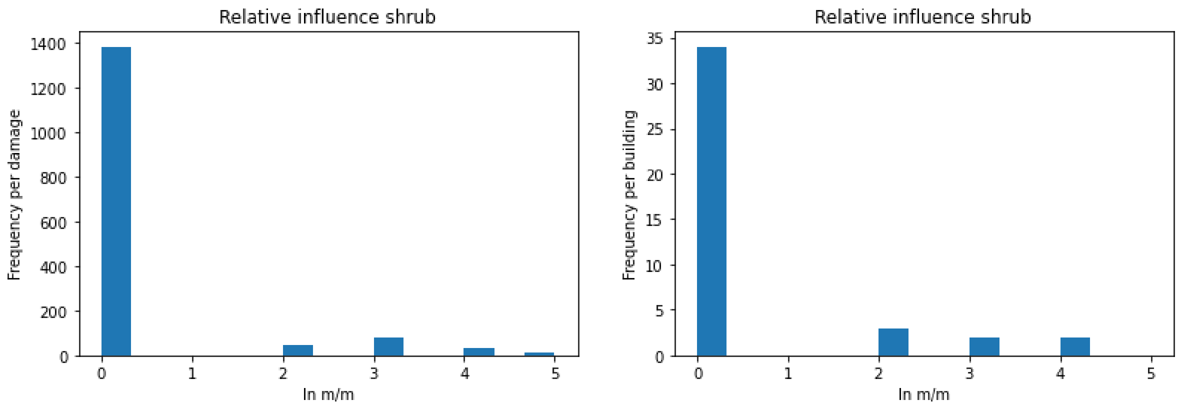


Figure 111 Histograms of characteristic 'relative influence shrub', no missing data. Frequency per damage at the left, frequency per building at the right.

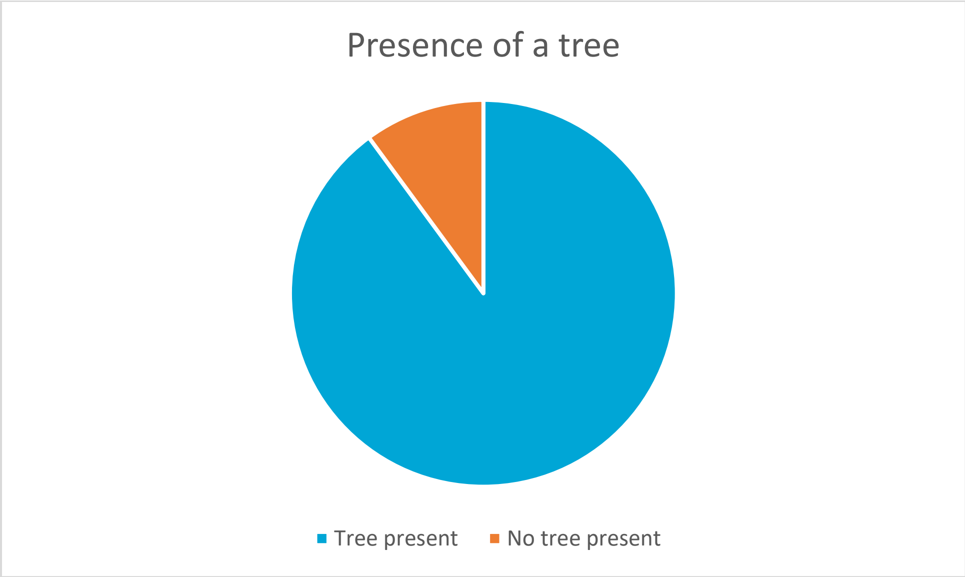


Figure 112 Spreading of characteristic 'presence of a tree', no missing data.

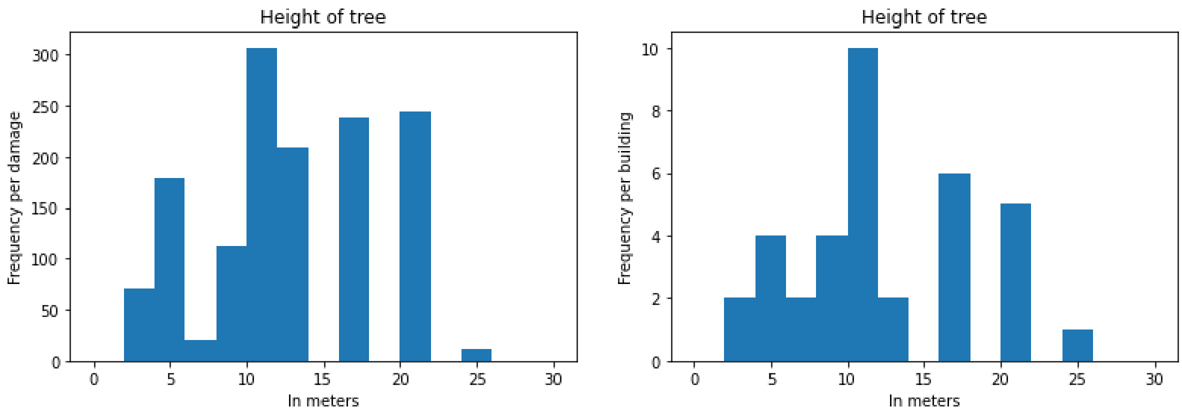


Figure 113 Histograms of characteristic 'height of tree', no missing data. Frequency per damage at the left, frequency per building at the right.

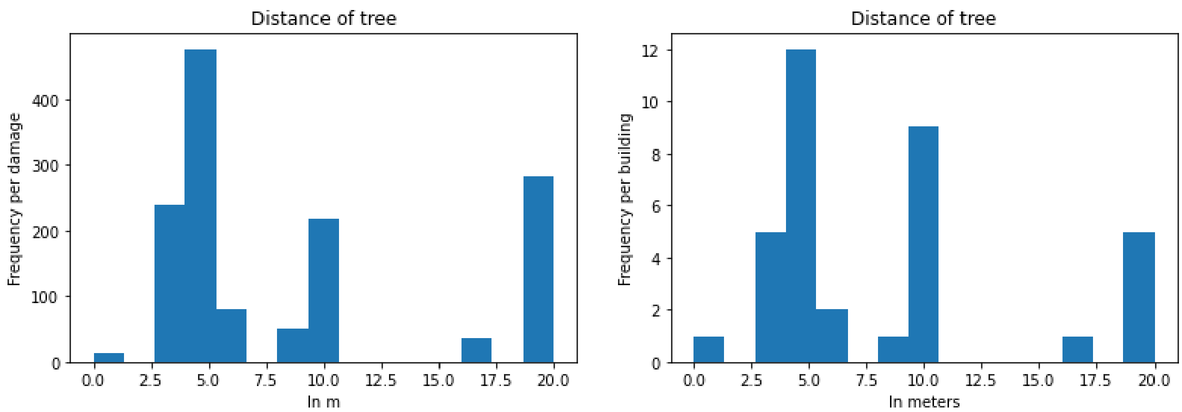


Figure 114 Histograms of characteristic 'distance of tree', no missing data. Frequency per damage at the left, frequency per building at the right.

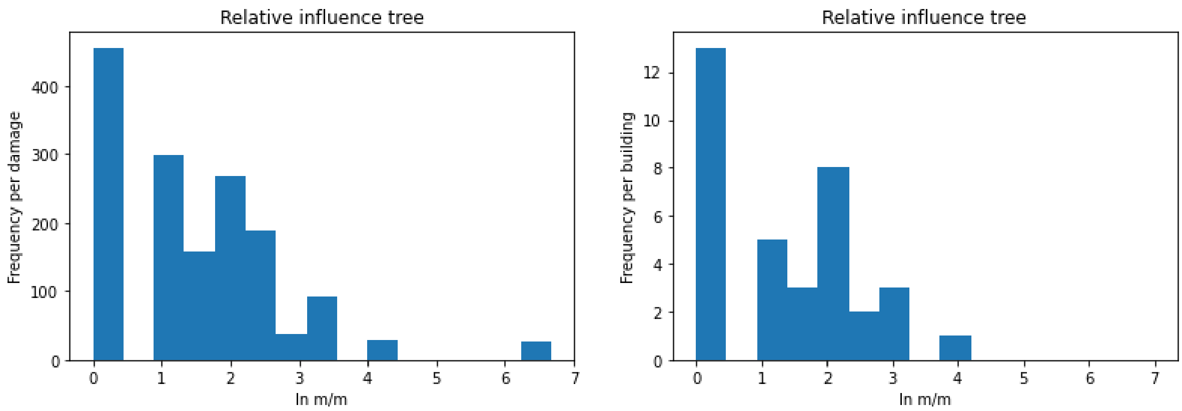


Figure 115 Histograms of characteristic 'relative influence tree', no missing data. Frequency per damage at the left, frequency per building at the right.

Ground settlement

Displacement of ground level.

Why is ground settlement important for damage investigation?

Ground settlement can be the result of different reasons: mining, change in groundwater level, change in loads, excavations or dragging by human or animal, etc. (van Staalduinen, Terwel, & Rots, 2018, p. 129)

How is ground settlement stored in the data?

The following possible answers on context characteristic 'ground settlement' are available in the data as a characteristic:

- Ground settlement decline [yes = 1, no = 0] only for damage cause category C (soil conditions),
- Ground settlement lift [yes = 1, no = 0] only for damage cause category C (soil conditions),
- No ground settlement [yes = 1, no = 0] only for damage cause category C (soil conditions).

Visualisation of applied characteristic 'ground settlement

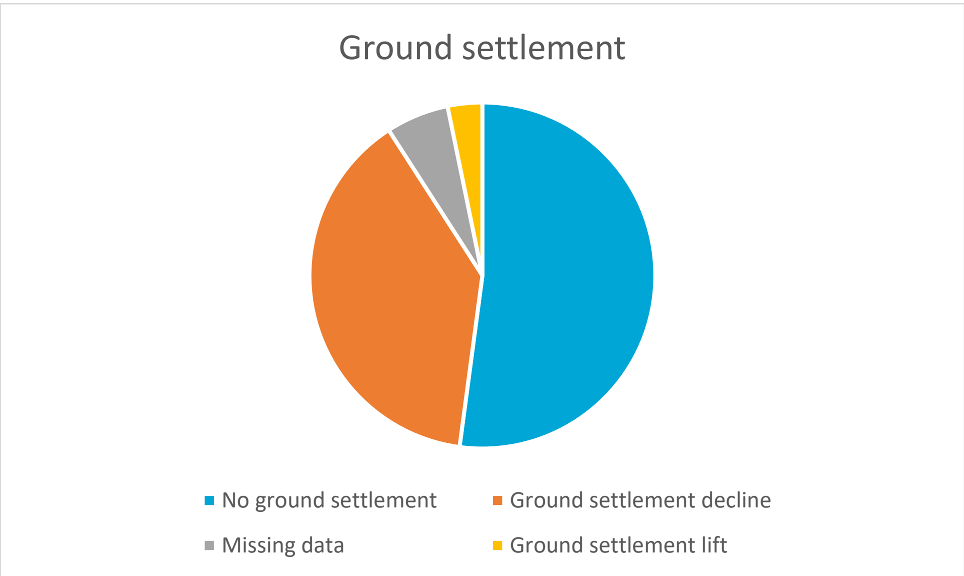


Figure 116 Spreading of characteristic 'ground settlement', 6% of the data is missing.

What to do with the missing data?

Some data is missing in the characteristic 'ground settlement' because: historical satellite measurements were not available and no other reports mentioned something about ground settlement. The unknown cases are notated as missing data, according to the rule 'WYSIATI' (What you see is all there is) rule for quality protection of data collection (Kahneman, 2013). The best option is to substitute the missing data as 'no ground settlement'. It is assumed that if noticeable settlements had occur, they would be mentioned by reports or by owners. It is also the most frequent option in this characteristic.

Missing data in characteristic 'ground settlement' is substitute with no ground settlement.

Traffic

Vehicles on roads or railways nearby buildings.

Why is traffic important for damage investigation?

According to Salet et al, road traffic can cause vibrations that could lead to damage due to overloading (Salet & Bruurs, 2017, p. 39). Also, the vibrations could lead to a stress change in the soil (Salet & Bruurs, 2017, p. 70). Staalduinen et al determine a falsification criteria based on speed bumps, maximum speed, road surface and road usage (van Staalduinen, Terwel, & Rots, 2018, p. 148). In practice there were no cases where road traffic was causing a damage (van Staalduinen, Terwel, & Rots, 2018, p. 135). Therefore, the measurement of road traffic is reduced to the distance between the road and the building. Parking spaces and garages close to building were not included in this measurement, because it is expected that this type of traffic will not cause vibrations large enough to cause damage. Exceptions to this were yards with heavy traffic, such as trucks and tractors.

Like road traffic, also rail traffic could cause vibrations that could lead to damage by overloading (Salet & Bruurs, 2017, p. 95). The vibrations of trains could lead to a change in stresses in the soil (Salet & Bruurs, 2017, p. 70). Staalduinen et al determine a falsification criteria based on the distance to the railway and the type of rail traffic (van Staalduinen, Terwel, & Rots, 2018, p. 148). In practice, there were no cases where rail traffic was causing damage (van Staalduinen, Terwel, & Rots, 2018, p. 135). Therefore, the measurement of rail traffic is reduced to the distance between the railway and the building, with a maximum of 100m.

How is traffic stored in the data?

The following possible answers on context characteristic 'traffic' are available in the data as a characteristic:

- Distance to the nearest public or intensively used road [m] only for A.3 (overloading by vibrations) and C.X (soil conditions),
- Distance to nearest railway, within 100m [m] only for A.3 (overloading by vibrations) and C.X (soil conditions).

Visualisation of applied characteristic 'traffic'

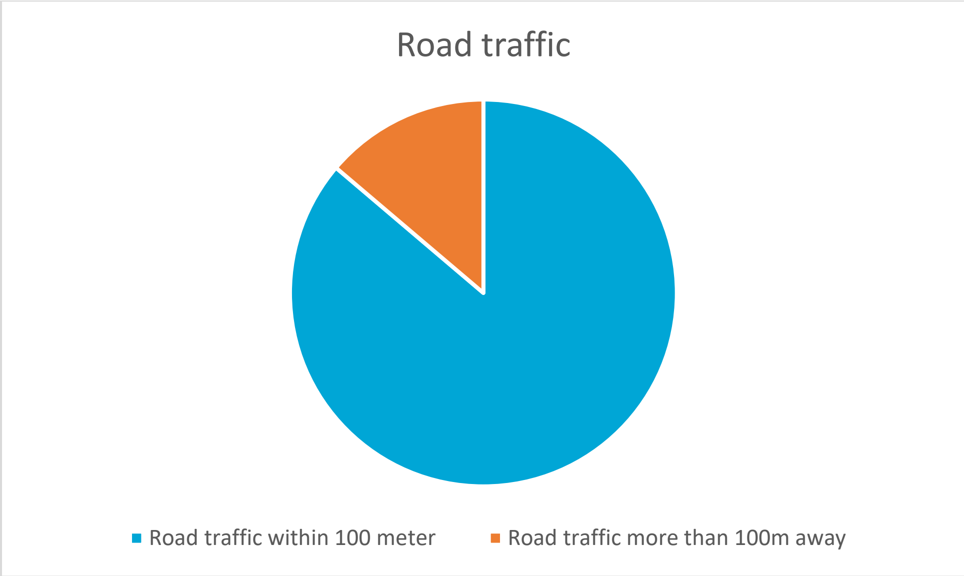


Figure 117 Spreading of characteristic 'Road traffic', no missing data.

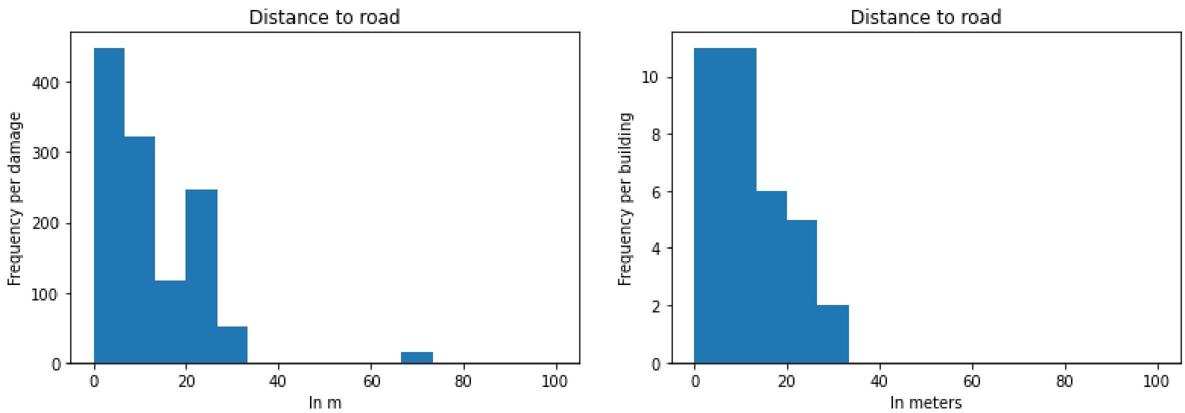


Figure 118 Histograms of characteristic 'distance to road', no data is missing. Frequency per damage at the left, frequency per building at the right.

What to do with the missing data?

The distance to the nearest road is not measured if a public or intensively used road is more than 100 meter away. This appears as missing data in the analysis. According to Staalduinen et al, a road more than 30-50 meters away from the building would not affect a damage (van Staalduinen, Terwel, & Rots, 2018, p. 148). For decision tree analysis, it is sufficient to substitute the missing data with a infinity large number, like 99999. For linear analysis, that is difficult, because it will influence the overall outcome. Replacing the missing data with the mean is not a good representation of the real world. It is possible to substitute the missing data with own measurements from Google Maps. The missing values can be recovered.

Missing data in characteristic 'road traffic' is substitute with recovered measurements.

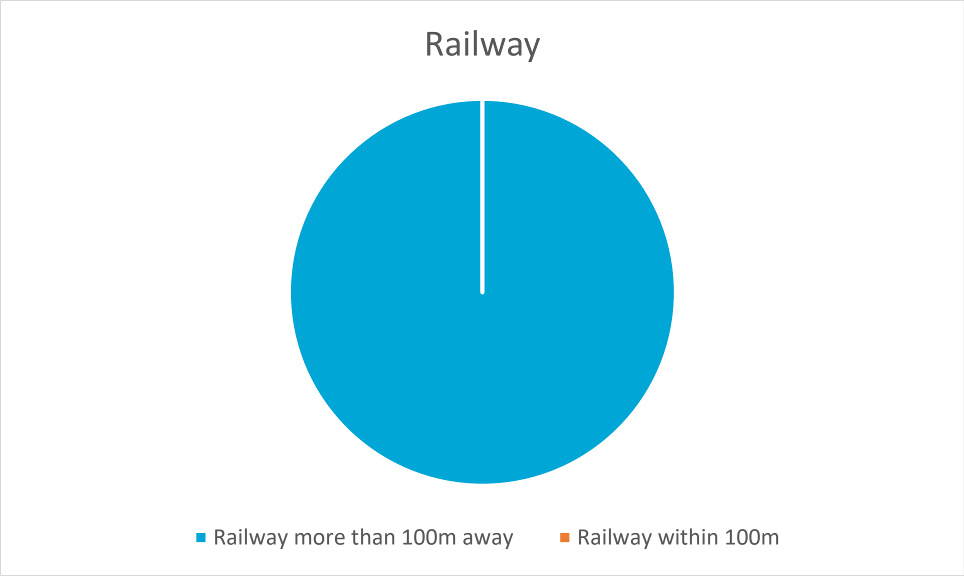


Figure 119 Spreading of characteristic 'Road traffic', no missing data.

The characteristic 'Railway' is removed from the data, because in zero of the damage cases was a railway within 100m.

Construction activities

Construction side was located near a building.

Why is construction activities important for damage investigation?

Construction activities could cause vibrations or a change in soil conditions that could lead to a damage (Salet & Bruurs, 2017, p. 242). Although Staalduinen et al indicated a influence area of 0, 20 and 100 meter (van Staalduinen, Terwel, & Rots, 2018, p. 41), they didn't give a falsification criteria for construction activities (van Staalduinen, Terwel, & Rots, 2018, p. 147). According to Salet et al, the influence area for pile driving is 15-20 meter and for vibrating of sheet piles 5-10 meter (Salet & Bruurs, 2017, p. 242). They based their calculations on the CUR 166 and SBR richtlijn deel A. This calculations are probably the reason for the division by Staalduinen et al of 0, 20 and 100 meter.

How is traffic stored in the data?

The following possible answers on context characteristic 'construction activities' are available in the data as a characteristic:

- Construction activities within 0 meters [yes = 1, no = 0] only for A.3 (overloading by vibrations) and C.X (soil conditions).
- Construction activities within 20 meters [yes = 1, no = 0] only for A.3 (overloading by vibrations) and C.X (soil conditions).
- Construction activities within 100 meters [yes = 1, no = 0] only for A.3 (overloading by vibrations) and C.X (soil conditions).

Visualisation of applied characteristic 'construction activieties'

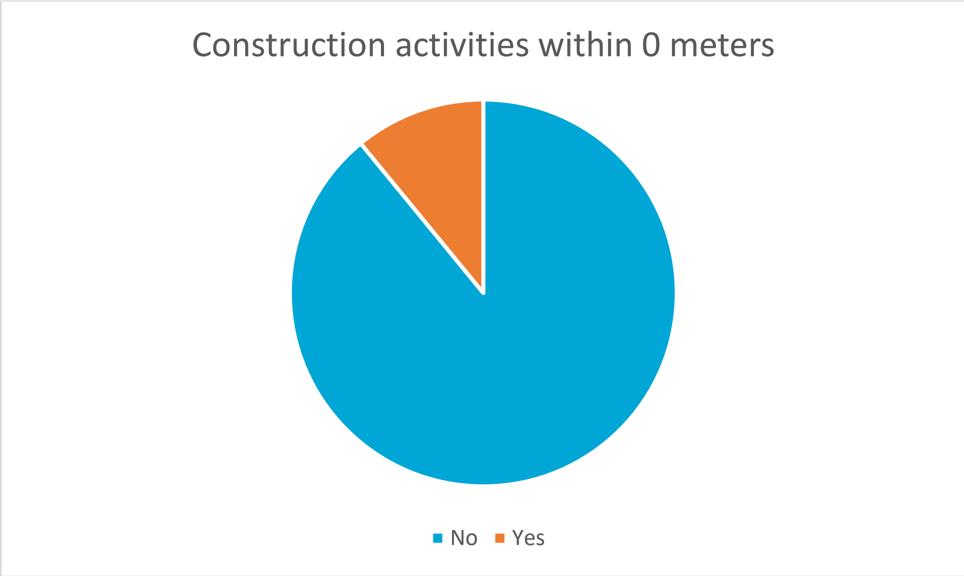


Figure 120 Spreading of characteristic 'construction activities within 0 meters', no missing data.

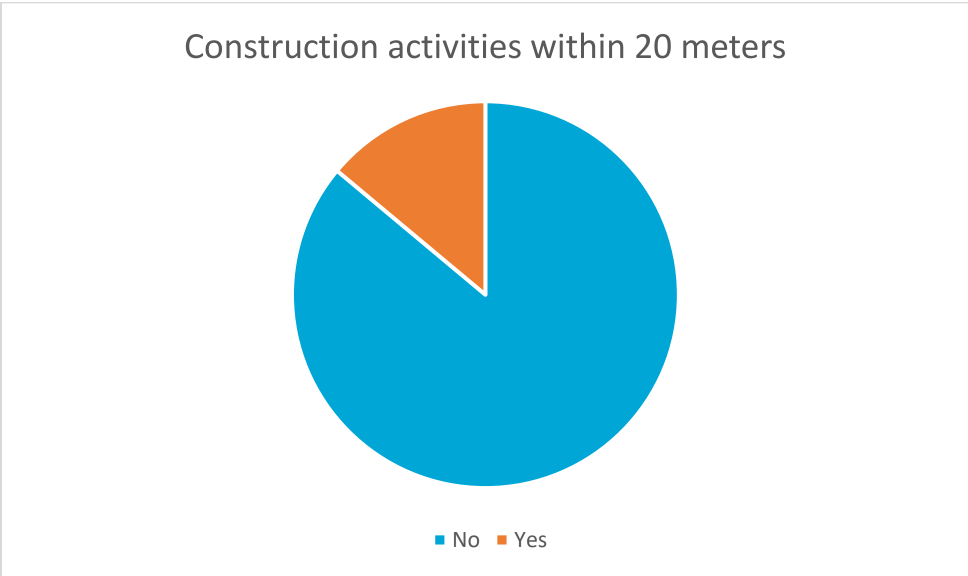


Figure 121 Spreading of characteristic 'construction activities within 0 meters', no missing data.

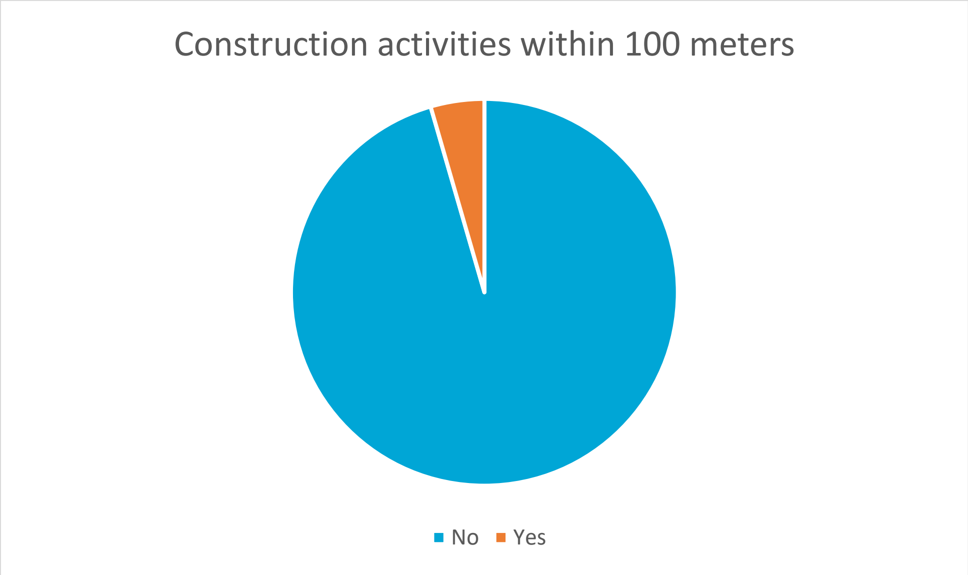


Figure 122 Spreading of characteristic 'construction activities within 0 meters', no missing data.

Collecting this data from the case reports could be in violation of the 'WYSIATI' (What you see is all there is) rule for quality protection of data collection (Kahneman, 2013). If reports do not have mentions of a basement, it was notated as 'not present'. It is assumed that mentioning something that is not there is too obvious to do.

Industrial activities

An industrial side was or is located near a building.

Why are industrial activities important for damage investigation?

As the same as construction activities, industrial activities could introduce vibrations and a change in soil conditions that could cause a damage (Salet & Bruurs, 2017, p. 242). It is difficult to define a falsification criteria for this, because of the large variation of possible activities that could cause vibrations or a change of soil condition (van Staalduinen, Terwel, & Rots, 2018, p. 147). As an initial baseline is chosen for a influence area of 100m (van Staalduinen, Terwel, & Rots, 2018, p. 41).

How are industrial activities stored in the data?

The following possible answers on context characteristic 'Industrial activities' are available in the data as a characteristic:

- Industrial activities within 100 meters [yes = 1, no = 0] only for A.3 (overloading by vibrations) and C.X (soil conditions).

Visualisation of applied characteristic 'industrial activities'

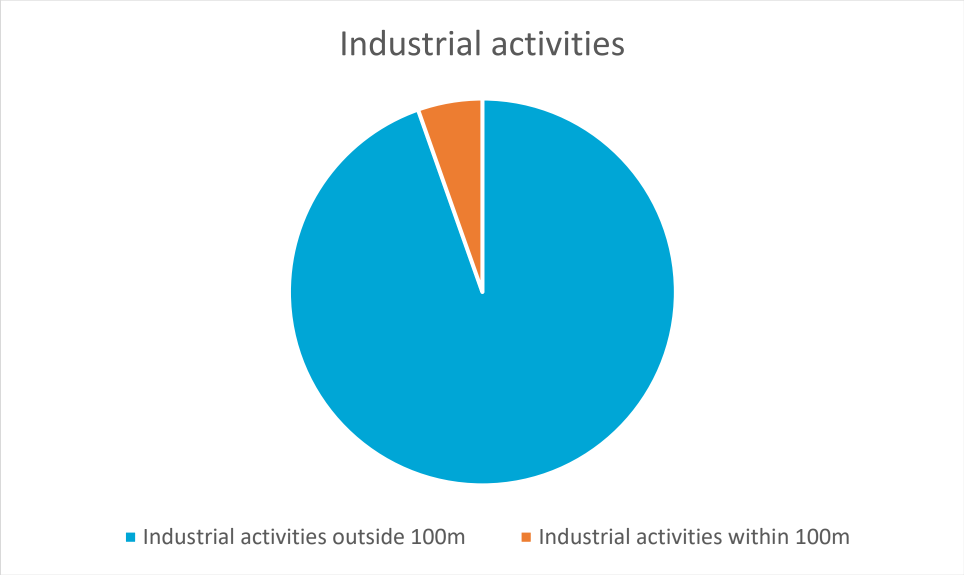


Figure 123 Spreading of characteristic 'industrial activities', no missing data.

Calamities

Unexpected situations that had an influence on the building's construction.

Why are calamities important for damage investigation?

Situations that are unexpected or very rare are calamities. Examples are: fire, explosions, lighting, collisions, floods, etc. (van Staalduinen, Terwel, & Rots, 2018, p. 29) It is expected that this unexpected situations can result in damage. The notations of a calamity is reduced to a discrete 'yes' or 'no' instead of a more elaborate explanation, because of the very rare occurrence of those situations.

How are calamities stored in the data?

The following possible answers on context characteristics 'calamities' are available in the data as a characteristic:

- Occurrence of a calamity [yes = 1, no = 0] only for A.4 (overloading by incidental loads).

Visualisation of applied characteristic 'calamities'

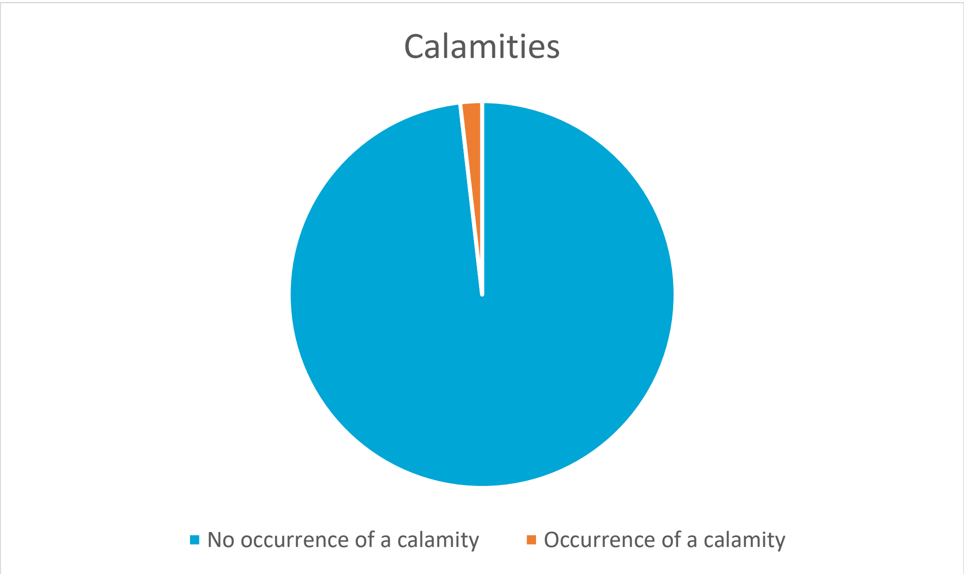


Figure 124 Spreading of characteristic 'calamities, no missing data.

Vibrations due to earthquakes

Sudden ground motions.

Why are vibrations due to earthquakes important for damage investigations?

Earthquakes can cause damage at a building in two ways. Direct, by vibrating the construction; or indirect, where the vibration in the ground cause a change in de soil stresses and therefore could impact the foundation of the building (van Staalduinen, Terwel, & Rots, 2018, p. 32 & 35).

How are vibrations due to earthquakes stored in the data?

There are different ways to measure an earthquake, with a variance of: instruments, units and (updated) models. In the data there are 2 types of meaningful earthquake vibration measurements available: the Bommer et al model GMM v3 at 25% and the Staalduinen model based on the previous mentioned one. For comparison also an actual updated model is included in the analyses, the GMM v6 at 1%. More explanation about what measurement, unit and model are used is explained in a separate appendix. That extra explanation about earthquakes is done because it is an important subject for the Groningen area.

The following possible answers on context characteristic 'Vibration due to earthquakes' are available in the data as a characteristic:

- PGV Bommer et al (GMM v3 at 25%) [mm/s] only for A.3.5 (overloading by vibrations of earthquakes) and C.X (soil conditions),
- PGV Staalduinen et al (GMM v3 at 25%) [mm/s] only for A.3.5 (overloading by vibrations of earthquakes) and C.X (soil conditions),
- PGV GMM v6 (at 1%) [mm/s] only for A.3.5 (overloading by vibrations of earthquakes) and C.X (soil conditions).

Visualisation of applied characteristic 'vibrations due to earthquakes'

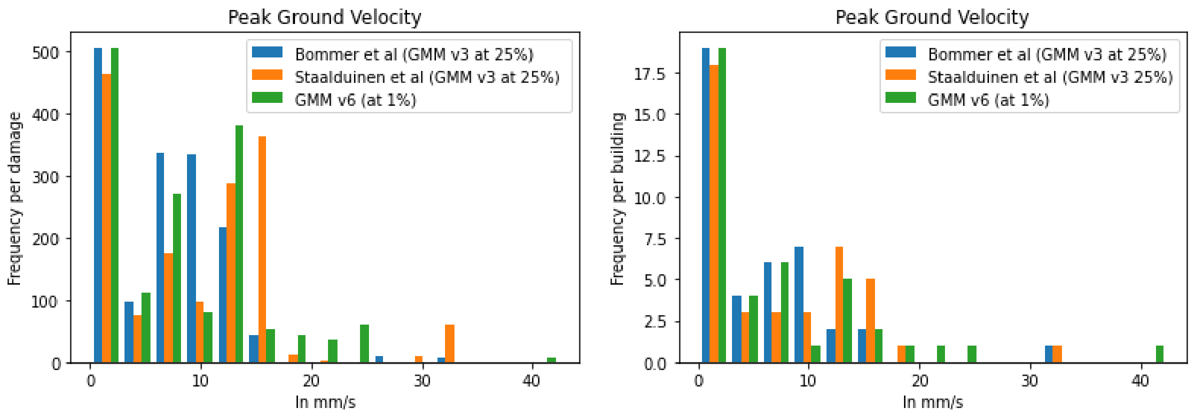


Figure 125 Histograms of characteristic 'vibrations due to earthquakes', no missing data. Frequency per damage at the left, frequency per building at the right.

Damage characteristics

Damage characteristics say something about how damage occurs in terms of position, location and shape of damage. Staalduinen et al defines the following damage characteristics: Position of damage, moment of first detection, displacement out-of-plane & in-plane, settlement difference, skewness of floor/building and damage classification. In the case of a crack the following is also collected: Displacement in the crack, crack width over the length of the crack, crack pattern, dirt, erosion, length differences that indicate differences in age, applied finish over crack, presence of repair work, direction, runs from-to, crack width, crack length, appearance and damage pattern according to de Vent. (van Staalduinen, Terwel, & Rots, 2018, p. 41).

Position of damage

As described by Terwel et al., it is sometimes difficult to determine the position of damage based on earlier damage reports (Terwel & Schipper, 2018). According to Staalduinen et al, the position of damage is described in terms of: building part, building element, location of building element, in finish or through construction and crack in unit or joint (van Staalduinen, Terwel, & Rots, 2018, p. 41).

Location

The position of a building part is done in 3 ways. First, if the damage is at the inside part or at the outside part of the building. Secondly, what is the type of building part, such as a wall, floor or ceiling. Third, at what side of the building this part is located. This last part is only done for facades. It is noted in terms of wind direction. In the data, the view direction is sometimes used, such as front view façade, rear façade and side façade. However, structural damage wise, the view direction has not an influence on the occurrence of damage, because the functional view of a façade cannot influence the forces construction. Moreover, the data will become more uniform when wind direction is applied instead of view direction. The wind direction may be a factor in damage by sun light radiation that could degrade building materials. Also, the sunlight could influence deformations of certain elements. A north east façade is noted as an east and north façade.

Material

The building element is interpreted as building material. Six building materials came a cross in the data: masonry, sand lime stone, concrete, timber, glass and plasterboard. A direct link between material and damage cause is unknown.

Relative location

The location of building element is interpreted as the relative location to openings and other building parts nearby. These connections are an obstacle for the forces in the element. Therefore, it is expected that this has an influence on damage. The word ‘nearby’ is carefully chosen, because this term does not describe the physical connection between a damage and obstacle. For example: a vertical crack parallel between two windows is nearby a window, but does not run between two windows. When damage is at window or door in the element, then it is notated as opening. Arbitrary is chosen for a division of: opening, door, corner, plane. Where opening includes windows and doors. Doors are taken separately, because there are damage causes of slamming doors and not from slamming windows (A.4.1 Overloading because of incidental loads by an impact of an object). Damage in plane means that there are no openings around. Another division is made in category openings with above an under: near opening (all openings), above opening, under opening.

Crack depth

Sometimes the damage is only at the surface of the element. Mostly is this the cause of hindered deformations of the surface finish, creep. In other cases, the damage is deeper and goes in the construction behind. It is difficult to determine if damage is at the surface, in the finish, of the element without deconstructing the element. Because of that, this determination is subjective to the observer.

Crack layout

There are two types of cracks in masonry elements. Cracks that runs through the joint in between the stones and cracks that runs through the stones, the unit. In the last case, the shape of the crack has no relation with the masonry bond. It still runs partly through the joint material from unit to unit, but the direction of the crack was not influenced by the bond. There are cases that they run through both, the joint material and the unit. This means that the crack over it length partly follows the joint material and partly not.

How is the position of damage stored in the data?

The following possible answers on damage characteristic 'position of damage' are available in the data as a charcteristic:

- Damage is inside the building [yes = 1, no = 0],
- Damage is outside the building [yes = 1, no = 0],
- Damage is in building part:
    - Building part façade [yes = 1, no = 0],
    - Building part inner wall [yes = 1, no = 0],
    - Building part roof [yes = 1, no = 0],
    - Building part floor [yes = 1, no = 0],
    - Building part ceiling [yes = 1, no = 0],
    - Building part basement wall [yes = 1, no = 0],
    - Building part outside-/garden wall [yes = 1, no = 0].
- Direction of façade:
    - North façade [yes = 1, no = 0] only for B.X (deformations),
    - East façade [yes = 1, no = 0] only for B.X (deformations),
    - South façade [yes = 1, no = 0] only for B.X (deformations),
    - West façade [yes = 1, no = 0] only for B.X (deformations).
- Building element:
    - Damage is in masonry [yes = 1, no = 0],
    - Damage is in calcium-silicate [yes = 1, no = 0],
    - Damage is in concrete [yes = 1, no = 0],
    - Damage is in timber [yes = 1, no = 0],
    - Damage is in glass [yes = 1, no = 0],
    - Damage is in plasterboard [yes = 1, no = 0].
- Position of building element:
    - Building element in plane [yes = 1, no = 0],
    - Building element near opening [yes = 1, no = 0],
    - Building element above opening [yes = 1, no = 0],
    - Building element under opening [yes = 1, no = 0],
    - Building element near door [yes = 1, no = 0] only for A.4.1 (Overloading because of incidental loads by an impact of an object).
    - Building element near corner [yes = 1, no = 0].
- In finish or through construction [through = 1, in finish = 0],
- Crack in joint [yes = 1, no = 0],
- Crack in unit [yes = 1, no = 0],
- Crack in joint and unit [yes = 1, no = 0].

Visualisation of applied characteristic 'Position of damage' & how it missing data is treated

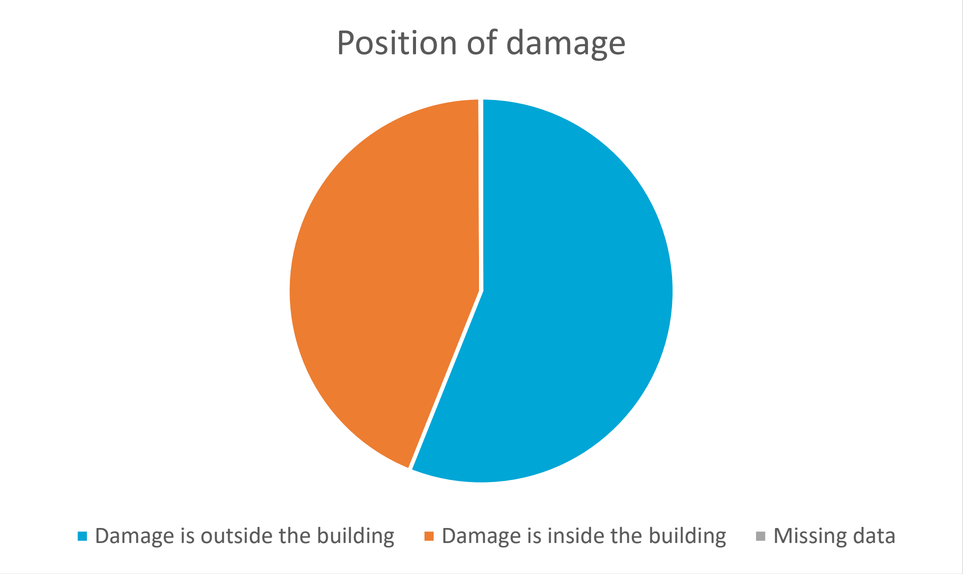


Figure 126 Spreading of characteristic 'position of damage (inside/outside)'. Data is missing in less than 5 damage cases.

Some data is missing in the characteristic 'position of damage (inside/outside)', because the location of the damage was unclear. When the location of inside/outside is unclear, the damage cause is definitely unclear or very difficult to determine. Considering the small group of missing data, the best option is to remove the missing data cases.

Missing data in characteristic 'position of damage' is removed with the corresponding damage cases.

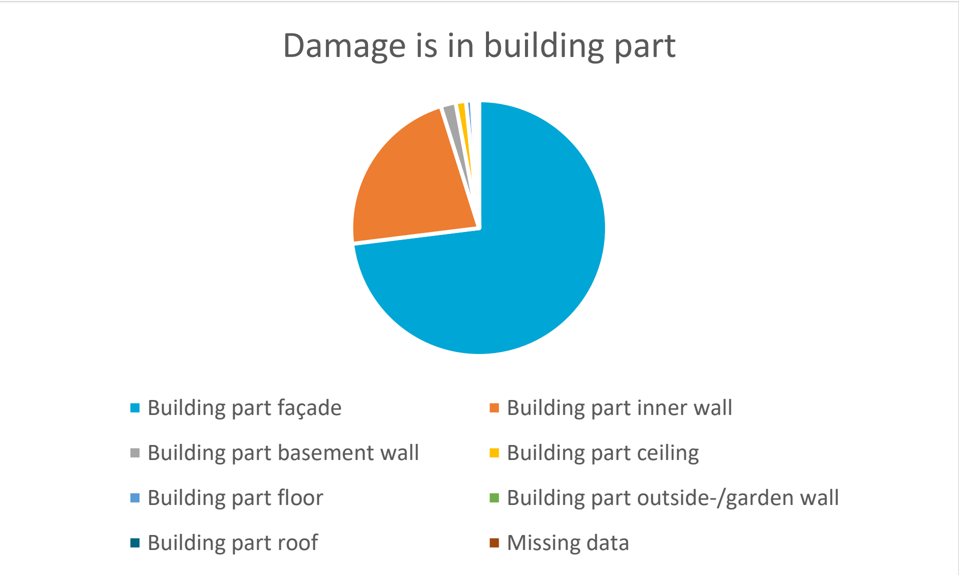


Figure 127 Spreading of characteristic 'damage is in building part'. Data is missing in less than 5 damage cases.

Some data is missing in the characteristic 'damage is in building part', because the location of the damage was unclear. When the location of the damage is unclear, the damage cause is definitely unclear or very difficult to determine. Considering the small group of missing data, the best option is to remove the missing data cases.

Missing data in characteristic 'position of damage' is removed with the corresponding damage cases.

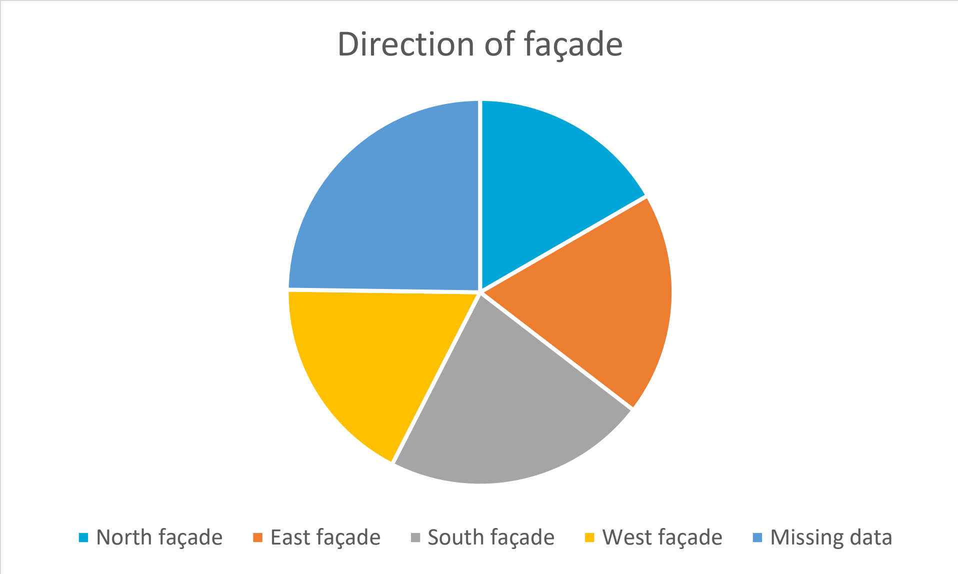


Figure 128 Spreading of characteristic 'direction of façade', 25% of the data is missing.

As earlier mentioned, a NE-façade is noted as a north and an east façade. This pollute the percentage calculations in Figure 128. Considering that, it is notable that the wind direction of the façade is almost equally spread in the data. Some data is missing in the characteristic 'direction of façade'. This is because of: the building part is not a façade (91.5% of missing data) or the location of the façade is unknown. Therefore, the most cases of the missing data part is a real representation. There is not always a wind direction possible. The way the data is stored, allows to continue with the analyses. Every direction is stored if it is there or not. The input are four times a zero if everything is not present and that is a real life representation.

Missing data in characteristic 'direction of façade' is allowed.

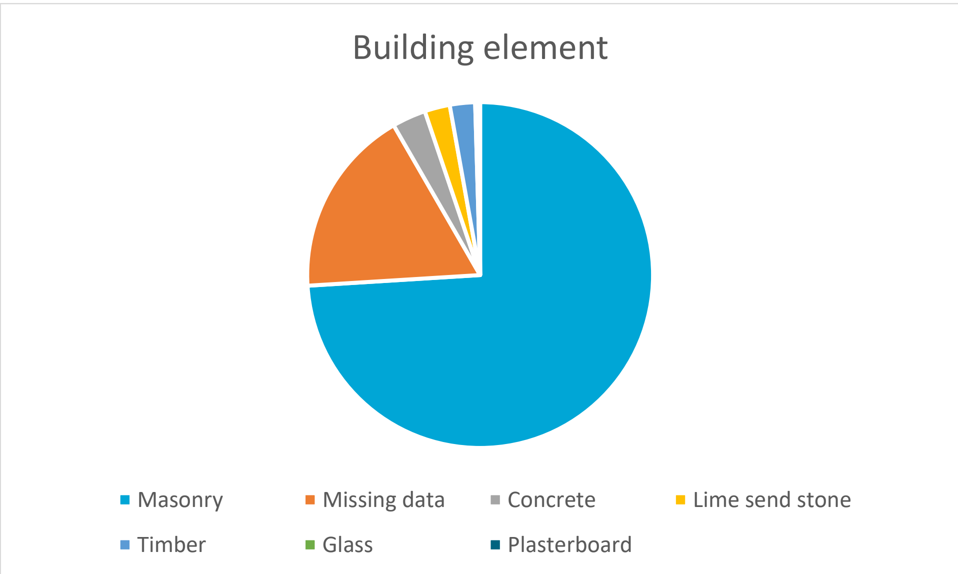


Figure 129 Spreading of characteristic 'Building element', 18% of the data is missing

Some data is missing in the characteristic 'building element'. A large part of the missing data are inside walls. The material of the inside walls are unknown because: it is not mentioned in reports, drawing or the wall is applied with a finish. The applied finish hindered the determination of the walls material. The missing data could be substitute with masonry, but this material is mostly used in outside façades and not for inner walls. Inner walls are mostly form timber with plasterboard or lime sand stone blocks. The missing data is substitute with lime sand stone, because it is the most frequent used material in inner walls according to this data.

Missing data in characteristic 'building element' is substitute with lime sand stone.

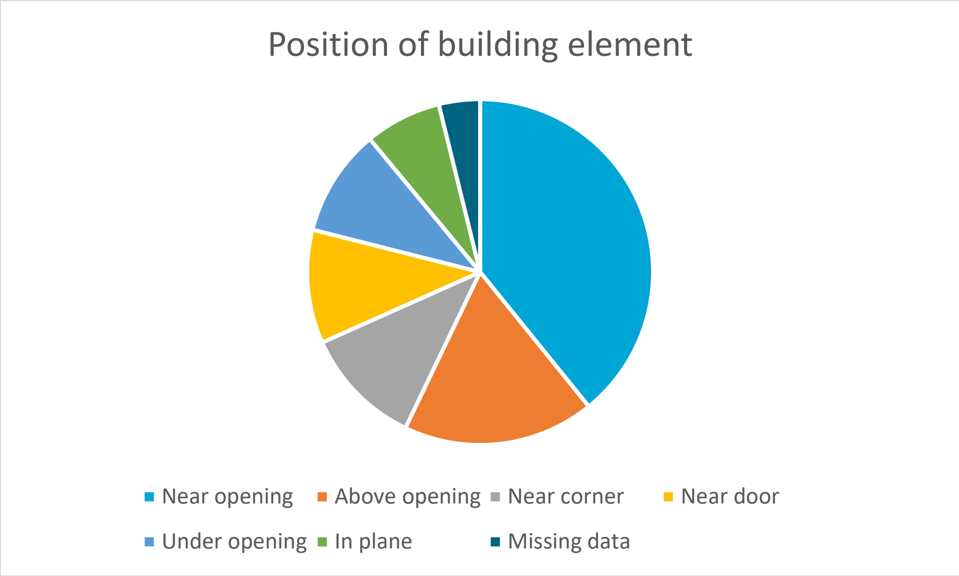


Figure 130 Spreading of characteristic 'position of building element', 4% of the data is missing.

Some data is missing in the characteristic 'position of building element'. This data is missing because the location of the damage is unclear. Photos, drawings or descriptions could make an indication of the location, but this was not done for those missing cases. An option would be to substitute the missing data with 'in plane', because no mentioning of a location could mean that nothing of interest was around the damage. That was often not the case. The location was truly not known. Therefore, the best option is to remove the missing data cases. Another argument is that if the location is unclear, it would be very difficult to determine the cause of the damage.

Missing data in characteristic 'position of damage' is removed with the corresponding damage cases.

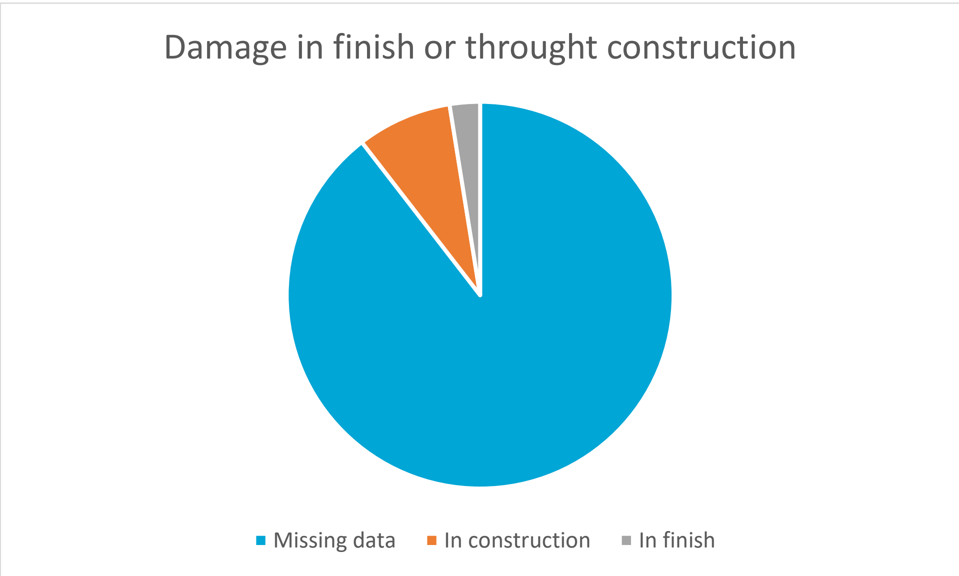


Figure 131 Spreading of characteristic 'damage in finish or through construction, 89% of the data is missing.

Some data is missing in the characteristic 'damage in finish or through construction'. This data is missing because: this characteristic was not specific mentioned in every report and it is very difficult for an inspector to collect this characteristic in a non-destructive way (removal of the applied finish). The best option is to remove this characteristic from the analysis, because a large part of the data is missing.

The characteristic 'damage in finish or through construction' is removed from the analysis because too much data is missing.

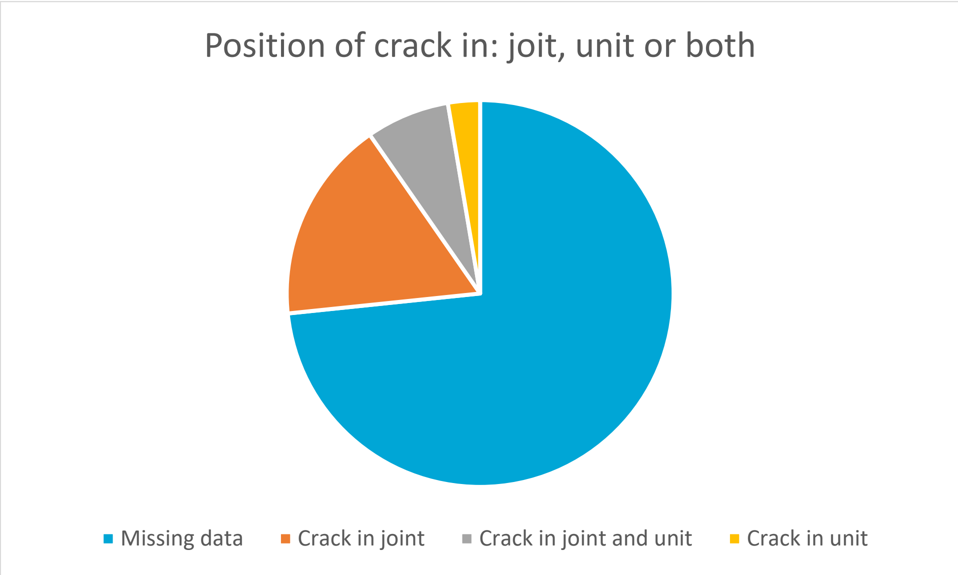


Figure 132 Spreading of characteristic 'position of crack in: joint, unit or both, 73% of the data is missing.

Some of the data in characteristic 'position of crack in: joint, unit or both' is missing. A large part of the data is missing because damage was not in a masonry wall or a applied finish hindered the view on joint and units of the damaged element. This characteristic was not specific stated in the reports and it had to be collected from photo's or drawings of the crack. Those pictures and drawing didn't have enough detail to collect this characteristic in most cases. Two options remain. The first option is to remove the characteristic because of too much missing data. The second option is to accept that some data is missing, because it does not exist. With not existing means that not all damage is in an element that consist of joints and units. Considering that 74% of damage is in a masonry wall and 56% of damage has no finish, too much of the data is missing to continue the analysis with this characteristic.

The characteristic 'position of crack in: joint, unit or both' is removed from the analysis because too much data is missing.

Moment of first detection

The moment in time damage was first noted.

The moment of first detection of damage is sometimes arbitrary determined. The problem is that not all damage is directly observed and noted by the owner of the building. Therefore, the date of the inspection where damage is first describe is notated as the moment of first detection. See Construction year on how this information is applied and why it is important for analysing damage.

How the moment of first detection is stored in the data.

The following possible answers on damage characteristic 'moment of first detection' are available in the data as a characteristic:

- First detection of damage [dd-mm-jjjj]. See Construction year on how this information is used.

Displacement out-of-plane & in-plane

Parts of the building could displace, mostly the cause of a change in soil conditions. Building parts could move out-of-plane, perpendicular to the building parts direction, or in-plane movement, a change in the direction of the façade.

How displacement out-of-plane & in-plane is stored in the data

The following possible answers on damage characteristic 'displacement out-of-plane & in-plane' are available in the data as a characteristic:

- Building part displacement out-of-plane [yes = 1, no = 0].
- Building part displacement in-plane [yes = 1, no = 0], only for damage cause category C.X (change in soil condition)

Visualisation of applied characteristic 'displacement out-of-plane & in-plane' and how missing data is treated

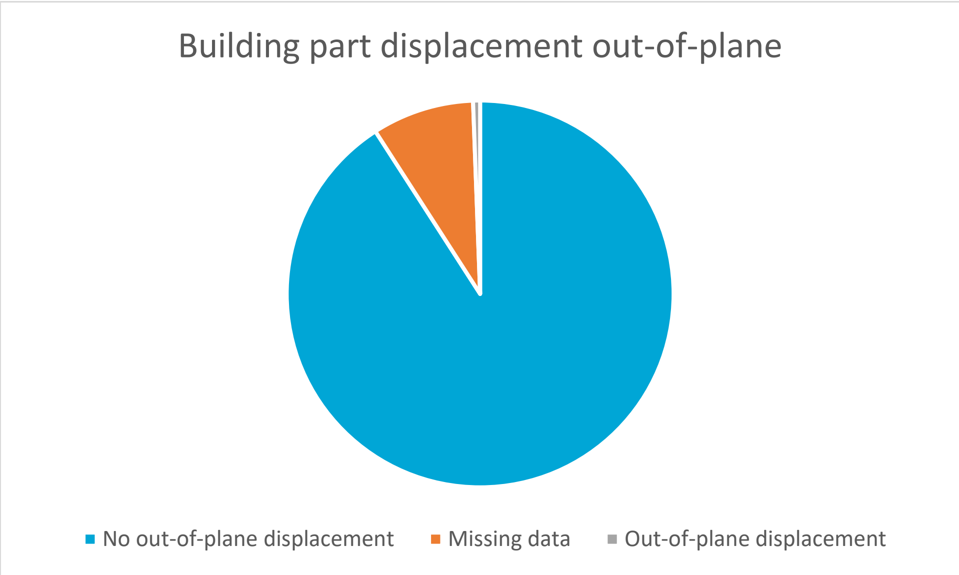


Figure 133 Spreading of characteristic 'building part displacement out-of-plane', 8% of the data is missing

Some of the data is missing in characteristic 'building part displacement out-of-plane'. This is because it was unclear in the reports if a building part was displaced. Also, the measurements for displacement determination were sometimes missing. The best option would be to substitute the missing data with the most frequent option, no out-of-plane displacement. Another argument to substitute missing data by a not occured event is that if there was an out-of-plane displacement, it would be noticed. The substitution could be in violation with the 'WYSIATI' rule for quality protection of data collection (Kahneman, 2013), but it is the best option for handling with missing data for this characteristic.

Missing data in characteristic 'building part displacement out-of-plane' is substitute with no out-of-plane displacement.

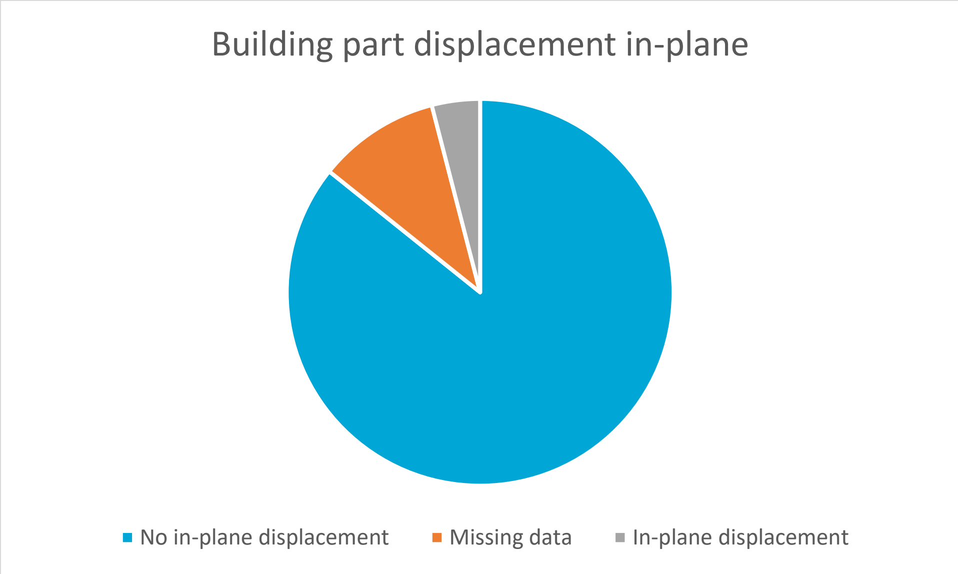


Figure 134 Spreading of characteristic 'building part displacement in-plane'.

Some of the data is missing in characteristic 'building part displacement in-plane. This is because it was unclear from the reports if a building part was displaced. Also, the measurements for displacement determination were missing sometimes. The best option would be to substitute the missing data with the most frequent option, no in-plane displacement. This option is by far the most frequent option. Another argument to substitute missing data by a not occured event is that if there was an in-plane displacement, it would be noticed. The substitution could be in violation with the 'WYSIATI' rule for quality protection of data collection (Kahneman, 2013), but it is the best option for handling with missing data for this characteristic.

Missing data in characteristic 'building part displacement in-plane' is substitute with no in-plane displacement.

Settlement differences over the length of the building

A settlement difference is a characteristic that is somewhere between a building characteristic and a damage characteristic. This characteristic says something about the state of the building in terms of settlement, which could be interpreted as damage. The difference between this characteristic and out-of-plane or in-plane displacement is that this is a characteristic of a building and the other one is a description of damage. Also a displacement could easily be seen, as a settlement difference needs a measurement of bedjoint levelling. Settlement differences over the length of the building are always caused by a change in soil conditions (C.X).

How are settlement differences over the length of the building stored in the data?

The following possible answers on damage characteristic 'Settlement differences over the length of the building' are available in the data as a characteristic:

- Settlement differences [yes = 1, no = 0], only for damage cause category C.X (change in soil condition)

Visualisation of applied characteristic 'settlement differences of the length of the building'

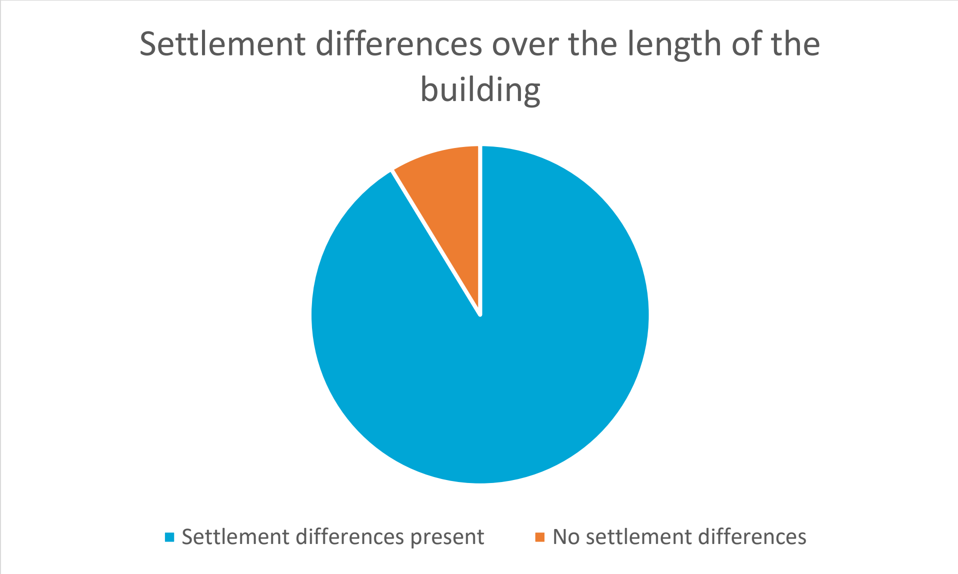


Figure 135 Spreading of characteristic 'Settlement differences over the length of the building', no missing data.

Skewness of floor/building

The following possible answers on damage characteristic 'skewness of floor/building' are available in the data as a characteristic:

- Skewness of floor/building [yes = 1, no = 0], only for damage cause category C.X (change in soil condition)

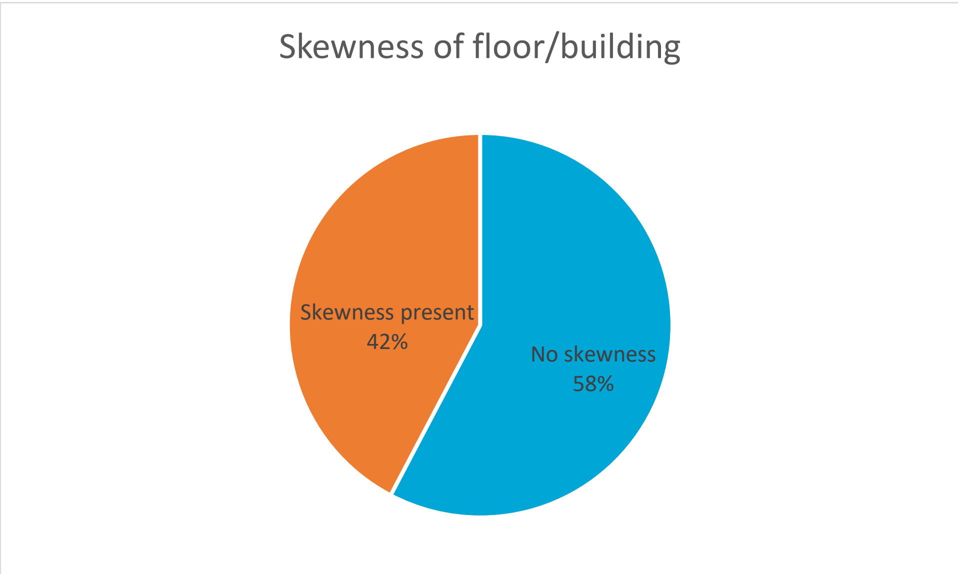


Figure 136 Spreading of characteristic 'skewness of floor/building', no missing data.

Displacement in the crack

A crack is the consequence of a differences in movement between the two sides of the crack. This movement could be in 2 directions: in-plane or out-of-plane. An in-plane movement is when the two sides from each other away or sliding along the edge over each other. An out-of-plane movement is when the two edge move perpendicular to the direction of the plane.

How the displacement in the crack is stored in the data

The following possible answers on damage characteristic 'displacement in the crack' are available in the data as a characteristic:

- Displacement in crack in-plane [yes = 1, no = 0],
- Displacement in crack out-of-plane [yes = 1, no = 0].

Visualisation of applied characteristic 'displacement in the crack'

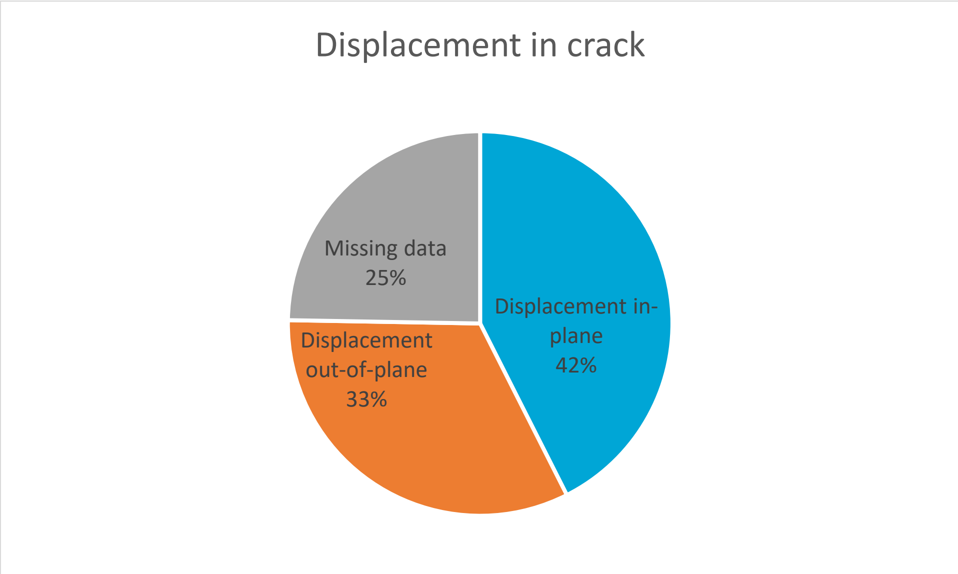


Figure 137 Spreading of the characteristic 'displacement in the cracks', 25% of the data is missing.

What to do with the missing data?

Some data is missing in the characteristic 'displacement in the crack'. It is very difficult to collect this characteristic from the data, because it is generally not noted during inspections or written in the reports. Sometimes pictures of damage helps with judging if a crack is displaced in- or out-of-plane. It is not possible to substitute the missing data with a mean, because there is not mean of this characteristic possible. Also, this characteristic has no overwhelming frequently noted option. Both options have more or less the same percentage. The proportion of the missing data is relative big, a quarter of the total amount of damage cases. Removing the missing data cases is a big loss of potential data. What stays over are 2 options: removing the characteristic from the analysis or continue the analysis with the missing data. The last option is possible, because of the way this two options of characteristic are stored in the data. Consequently it will mean that this characteristic is in 25% of the cases wrong, because there is always an answer on this characteristic in the case of a crack as damage.

The characteristic 'displacement in the crack' is removed from the analysis because too much data is missing.

Crack width over the length of the crack

The form of the crack width could say what type of stresses caused the crack. Tapered to one side are mostly caused by a moment, constant width are normal forces and tapered to both sides are mostly shear forces (De Vent, 2011, p. 122).

How is the crack width of the length of the crack stored in the data?

The following possible answers on damage characteristic 'displacement in the crack' are available in the data as a characteristic:

- Crack width is constant [yes = 1, no = 0],
- Crack width is variable [yes = 1, no = 0],
- Crack width is widen to one side [yes = 1, no = 0],
- Crack width is widen in the middle [yes = 1, no = 0].

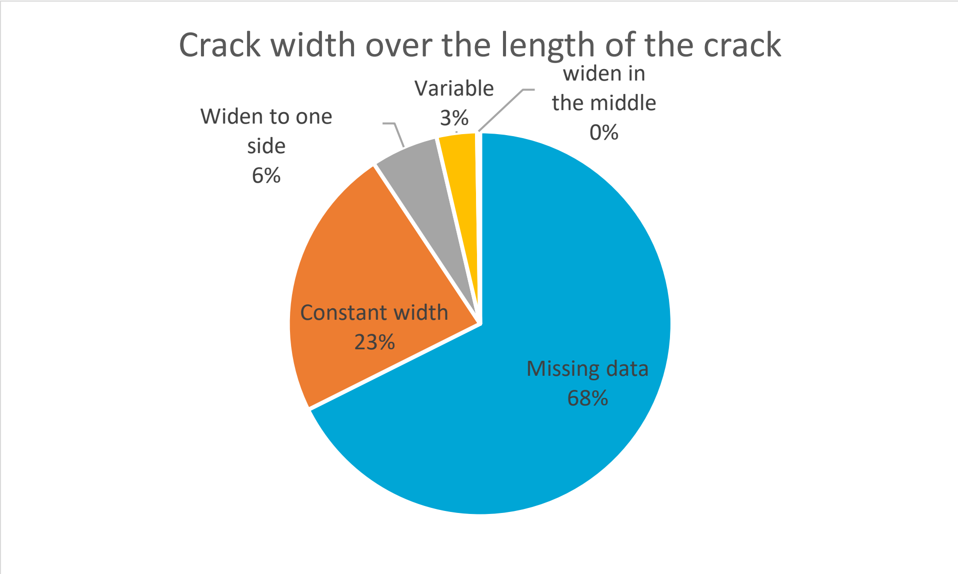


Figure 138 Spreading of characteristic 'crack width over the length of the crack, 63% of the data is missing.

What to do with the missing data?

Some data in characteristic 'crack width over the length of the crack' is missing in the data. This is because this characteristic is generally not noted by inspectors or written down in the reports. It is also impossible to collect this characteristic from sketches or drawing of the damage. Mostly the width over the length of the crack is only noted if something noticeable occurred. In other words, only a noticeable deviation of the crack width is mentioned in most cases. Otherwise nothing is stated about the crack width. This results in a large part of missing data. Proposed is to substitute the missing data with the constant crack width, because this is the most frequent one and it is the default crack width over the length.

Missing data in characteristic 'crack width over the length of the crack' is substitute with crack width is constant.

Crack pattern

A crack pattern is a typical of a crack. Staalduinen et al expected the following crack patterns based on the research of de Vent: # ,T, >---<, \---/, \|---|/,\|/, /|\, |/|, \/, /\, X and >. Unfortunately only in less than 5 damage cases could be categorised in one of these crack patterns. Therefor this damage characteristic is no longer part of this analysis.

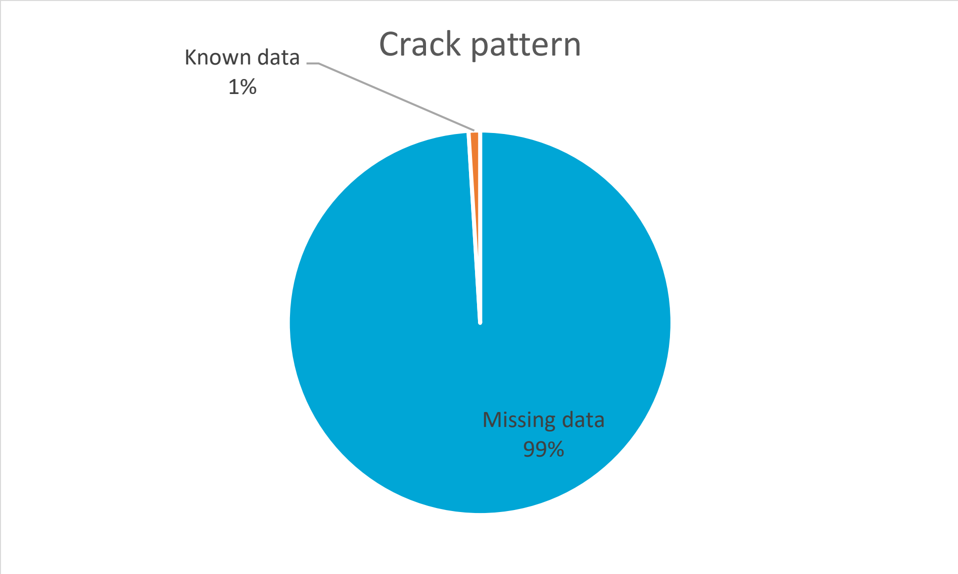


Figure 139 Spreading of the characteristic 'crack pattern', 99% of the data is missing.

The characteristic 'crack pattern' is removed from the analysis because too much data is missing.

Dirt

Dirt in cracks could be seen as an indication of aging, but it could also cause fatigue damage in stone materials. If a crack is filled with dirt and then want to close again, there is than a possibility that the crack can propagate (van Staalduinen, Terwel, & Rots, 2018, p. 89).

How is dirt stored in the data?

The following possible answers on damage characteristic 'dirt' are available in the data as a characteristic:

- Crack is polluted [yes = 1, no = 0].

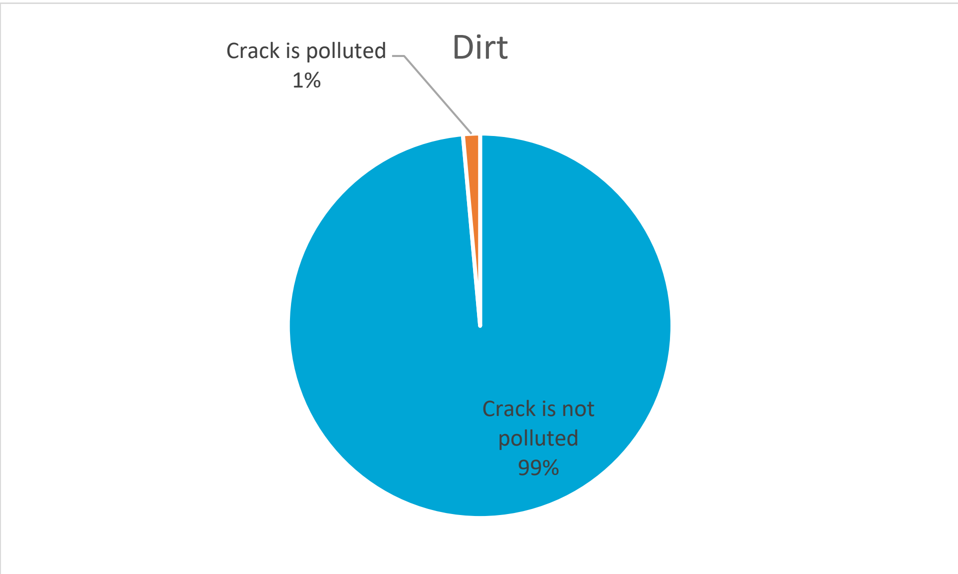


Figure 140 Spreading of the characteristic 'Dirt', no missing data.

Erosion

The erosion of crack edges are an indication of aging. If erosion is present, the crack is probably older. Some cracks could form in joint by degradation of the joint material. Erosion is an indication of this degradation proses (Edwards, 2005).

The following possible answers on damage characteristic ‘erosion’ are available in the data as a characteristic:

- Erosion of crack edges [yes = 1, no = 0].

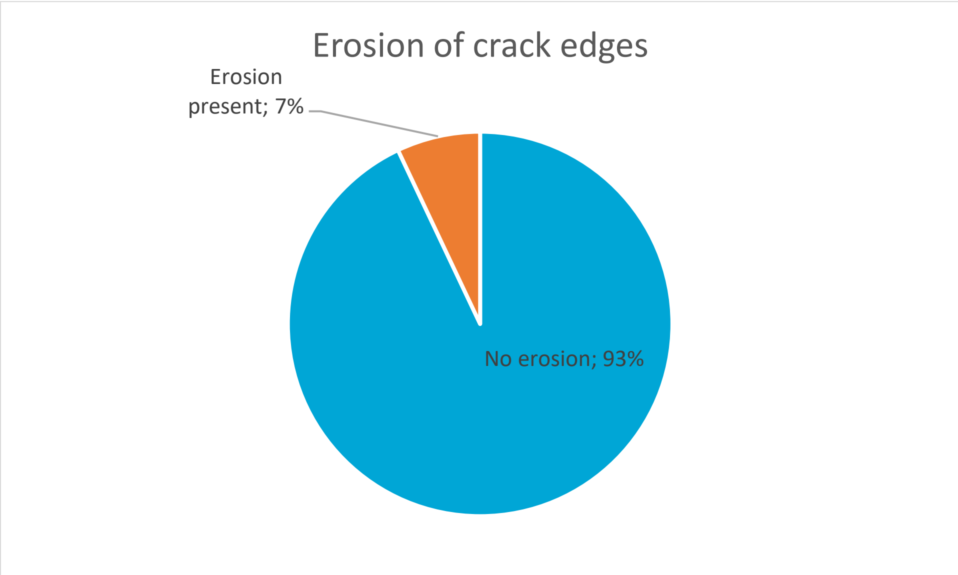


Figure 141 Spreading of characteristic 'erosion of crack edges, no data is missing.

Collecting this data from the case reports could be in violation of the ‘WYSIATI’ (What you see is all there is) rule for quality protection of data collection (Kahneman, 2013). If reports do not have mentions of a basement, it was notated as ‘not present’. It is assumed that mentioning something that is not there is too obvious to do.

Length differences that indicate differences in age

Differences in length that indicate differences in age was only found in 3 cases. Therefore, this characteristic is not included in the analysis.

Applied finish over crack

The applied finish over the crack has no influence on the forces through the construction. Therefore, it could be said that it has no influence on damage. It does not mean that damage cannot occur in the finish, it often does. It is expected that some finishes are more vulnerable than others. For example, plasterwork is more vulnerable for creep then wallpaper or paint.

How is the applied finish over cracks stored in the data?

The following possible answers on damage characteristic ‘applied finish over crack’ are available in the data as a characteristic:

- Applied finish of stucco [yes = 1, no = 0],
- Applied finish of plasterwork [yes = 1, no = 0],
- Applied finish of wallpaper [yes = 1, no = 0],
- Applied finish of tiles [yes = 1, no = 0],
- Applied finish of paint [yes = 1, no = 0],

- No applied finish [yes = 1, no = 0].

Visualisation of characteristic 'applied finish over crack'

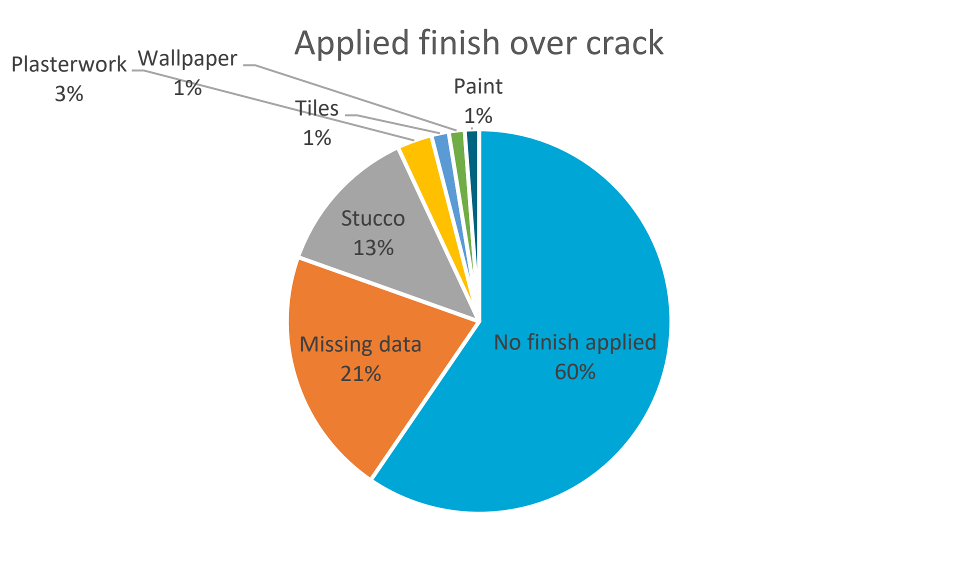


Figure 142 Spreading of the characteristic 'applied finish over the crack', 21% of the data is missing.

What to do with the missing data?

Some data in the characteristic 'applied finish over the crack is missing', because it was not always stated what the applied finish was. The most desirable option is to substitute the missing data. Substitute the missing data for 'no finish applied' for outside façades seems logical. For inside surfaces, stucco is the most frequent applied finish over a crack.

Missing data in characteristic 'applied finish over crack' is substitute with no applied finish for outside walls and stucco for inside walls.

Presence of repair work

Repair work is an indication of returning cracks.

How is the presence of repair work stored in the data?

The following possible answers on damage characteristic 'presence of repair work' are available in the data as a characteristic:

- Presence of repair work [yes = 1, no = 0].

Visualisation of characteristic 'presence of repair work'

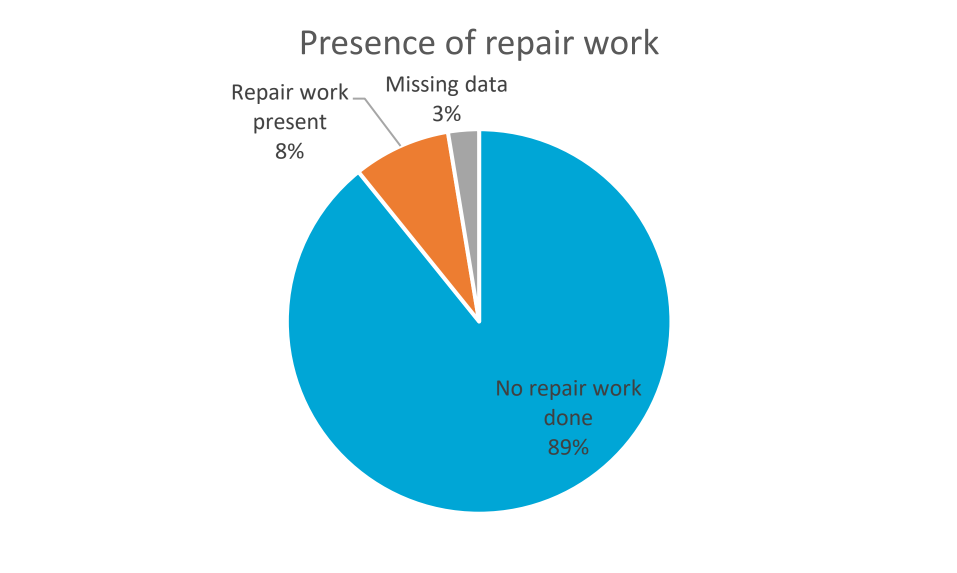


Figure 143 Spreading of characteristic 'presence of repair work', 3% of the data is missing.

What to do with the missing data?

Some data is missing in the characteristic 'presence of repair work', because it was not mentioned if repair work was done or not. It is not in cooperation with the 'WYSIATI' rule for quality protection of data collection (Kahneman, 2013), but it seems logical to substitute the missing data with the most frequent used option.

Missing data in characteristic 'presence of repair work' is substitute with no repair work present.

Crack direction

The direction of a crack is important for assessing damage. It is one of the key elements used by de Vent to determine possible causes of cracks. Analysing of a crack is not straightforward, because the relation between the load direction and the crack direction could be complicated (De Vent, 2011, p. 121). It could be said that the crack direction is a lead on how the forces flow in the construction.

How is the crack direction stored in the data?

The following possible answers on damage characteristic 'crack direction' are available in the data as a characteristic:

- Crack direction vertical [yes = 1, no = 0],
- Crack direction horizontal [yes = 1, no = 0],
- Crack direction diagonal [yes = 1, no = 0].

Visualisation of applied characteristic 'crack direction'

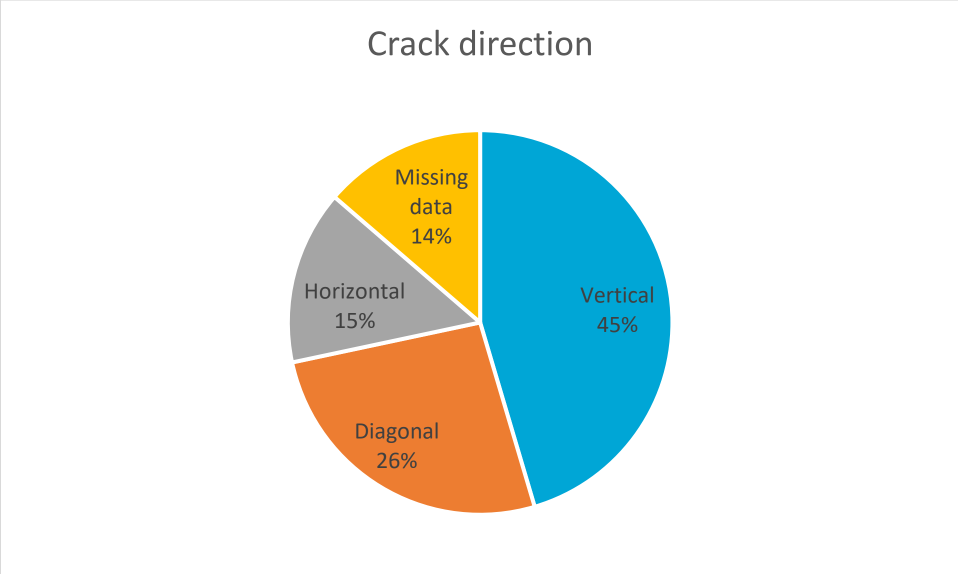


Figure 144 Spreading of characteristic 'crack direction', 14% of the data is missing.

What to do with the missing data?

Some data is missing in the characteristic 'crack direction'. There are three reasons why this data is missing. First, the crack direction was unclear, because the description of the crack is not comprehensive enough or the location of the crack is not found during inspection. This will mostly result in a unknown damage cause. Second, it was not possible to select a direction. Some damage is not a crack. Therefore, a crack direction was not always possible. Lastly, a crack was on a ceiling or at the floor. Also than it was not possible to determine a crack direction. It is sufficient to continue the analysis with the missing data. The missing data from the first reason will be removed from the dataset later in the process. The missing data from the other resons is a good representation of the damage situation.

Missing data in characteristic 'crack direction' is allowed.

Runs from-to

As indicated by Position of damage, the properties around the damage are important when assessing the damage. Windows, door and corner could have an influence on how forces flow through the construction. Every crack runs from something to something. A crack has a begin and an end. There is no definition on what point the crack starts and what the end point is. Is the begin point of a horizontal crack at the left side or at the right side? Therefore, damage has in this characteristic always 2 properties. For example, a crack between a window and a door has both 'window' as 'door' as runs from-to characteristic. Moreover, a crack between ground level and a place where nothing special is around has both 'ground level' as 'plane'. The difference between corner and edge is that a corner is a change of direction of the building element. With an edge the building element continuous in the same direction, but has a interrupting like a change in material or finishing. Forces react in another way on corners than edges.

How is runs form-to stored in the data?

The following possible answers on damage characteristic 'runs from-to' are available in the data as a characteristic:

- Runs from-to opening (includes: windows, window recess, window frame) [yes = 1, no = 0],
- Runs from-to door [yes = 1, no = 0],

- Runs from-to corner [yes = 1, no = 0],
- Runs from-to edge [yes = 1, no = 0],
- Runs from-to roof [yes = 1, no = 0],
- Runs from-to ceiling [yes = 1, no = 0],
- Runs from-to floor [yes = 1, no = 0],
- Runs from-to ground level [yes = 1, no = 0],
- Runs from-to plane [yes = 1, no = 0].

Visualisation of applied characteristic 'runs from-to'

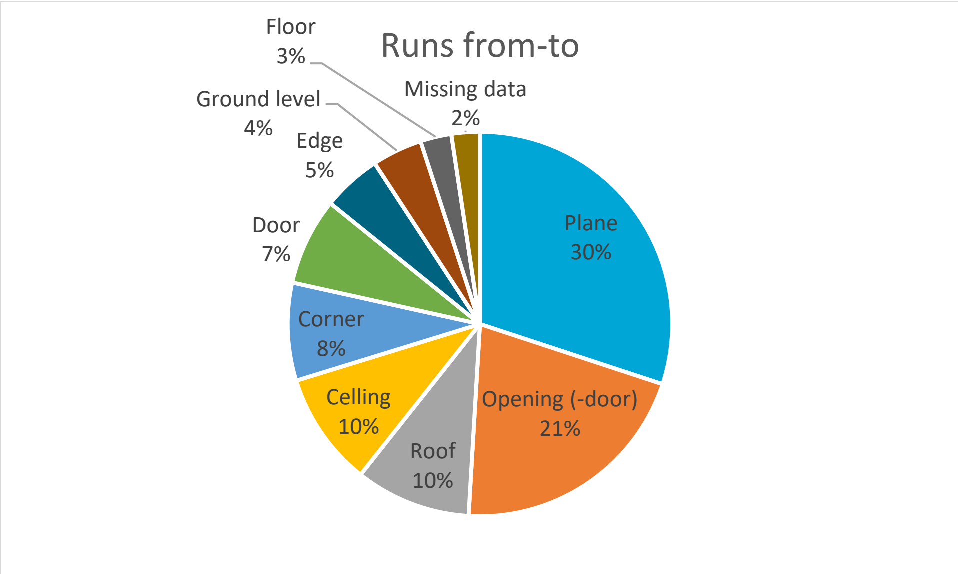


Figure 145 Spreading of characteristic 'runs from-to', 2% of the data is missing.

What to do with the missing data?

Some data is missing in the characteristic 'runs from-to'. Mostly, this data is missing because the location of the damage is not clear for the inspectors. If the location is unknown, also the cause will be unknown. Those unknown damage causes will be removed from the data. The last few missing data cases are there because the inspector was not clear enough with his description of damage or the damage has no 'runs from-to' characteristic. For example, skewness of a building or a displacement of a building part has no runs from-to property. The best option is to continue the analysis with the missing data.

Missing data in characteristic 'crack direction' is allowed.

Crack width

There is some discussion about the term crack width. De Vent argues that crack size is a better indication of measuring a crack (De Vent, 2011, p. 125). Crack width is still used in practise as level of damage (De Vent, 2011, p. 23) (van Staalduinen, Terwel, & Rots, 2018, p. 155). Therefore, it is interesting to collect this characteristic for this analyse. It is usually the measure the maximum crack width.

How is the crack width stored in the data?

The following possible answers on damage characteristic 'Crack width' are available in the data as a characteristic:

- Maximum crack width [mm].

Visualisation of applied characteristic 'crack width'

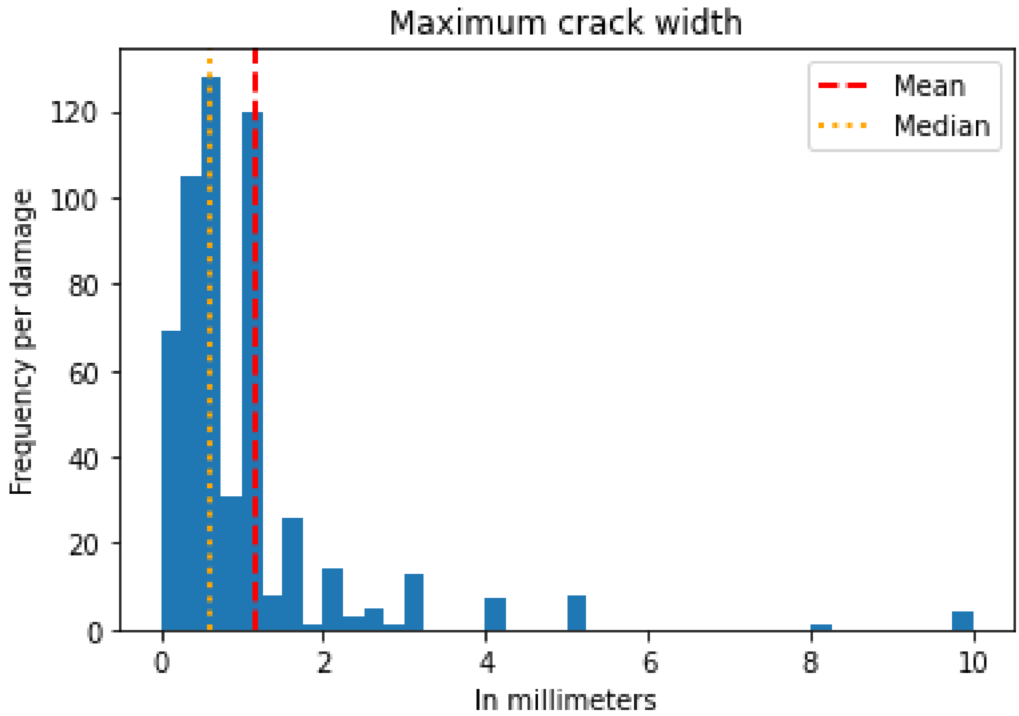


Figure 146 Histogram of characteristic 'maximum crack width, 64% of the data is missing.

What to do with the missing data?
Some data in the characteristic 'maximum crack width' is missing, because the location of the crack was unreachable for the inspector to measure the crack width. There are also cases where the measurement for the crack with was not a part of the inspection. Therefore, it was not collected. Another reason why some data is missing is that not all damage cases are a crack. It is possible to substitute the missing data. There are two options: the mean (1.16mm) or the median (0.6mm). According to Webb is the mean the best option, although the value of the median is expected to fit the best on the data.

Missing data in characteristic 'maximum crack width' is substitute with the mean of the known data, 1.16mm.

Crack length
As the same as crack width, the crack length is used to determine the level of damage (van Staalduinen, Terwel, & Rots, 2018, p. 155), but it is of less importance than the crack width (De Vent, 2011, p. 120).

How is the crack length stored in the data?
The following possible answers on damage characteristic 'Crack width' are available in the data as a characteristic:

- Maximum crack length [m].

Visualisation of applied characteristic 'maximum crack length'

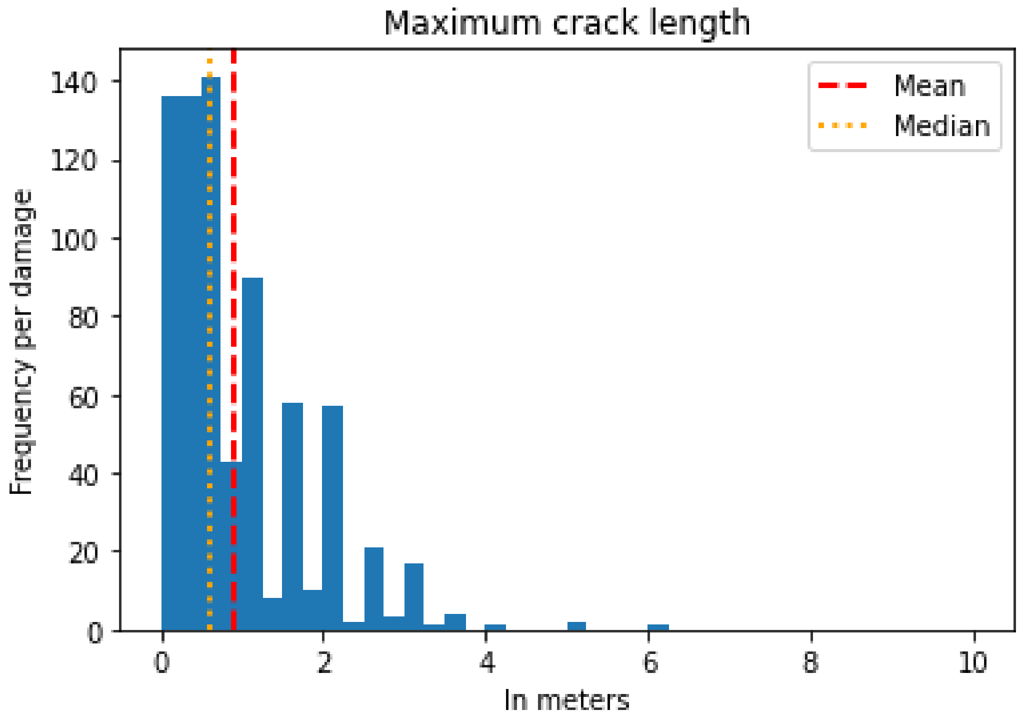


Figure 147 Histogram of hazard parameter 'maximum crack length, 52% of the data is missing.

What to do with the missing data?
Some data in the characteristic 'maximum crack length' is missing, because the crack was unreachable for the inspector to measure the crack length. There are also cases where the measurement of the crack length was not a part of the inspection. Therefore, it was not collected. Another reason why some data is missing is that not all damage cases are a crack. It is possible to substitute the missing data. There are two options: the mean (0.92 m) or the median (0.6 m). According to Webb is the mean the best option, although the value of the median is expected to fit the best on the data.

Missing data in characteristic 'maximum crack length' is substitute with the mean of the known data, 0.92m.

Appearance
A crack can appear in different forms and shapes. It is difficult to make a categorisation in this, for the same reason as explained in Crack pattern. It was possible to make a categorisation in a more macro scale.

How is the appearance of a crack stored in the data?
The following possible answers on damage characteristic 'appearance' are available in the data as a characteristic:

- Stepped crack [yes = 1, no = 0],
- Not stepped crack [yes = 1, no = 0],
- Hairline crack [yes = 1, no = 0],
- Branched cracks [yes = 1, no = 0].

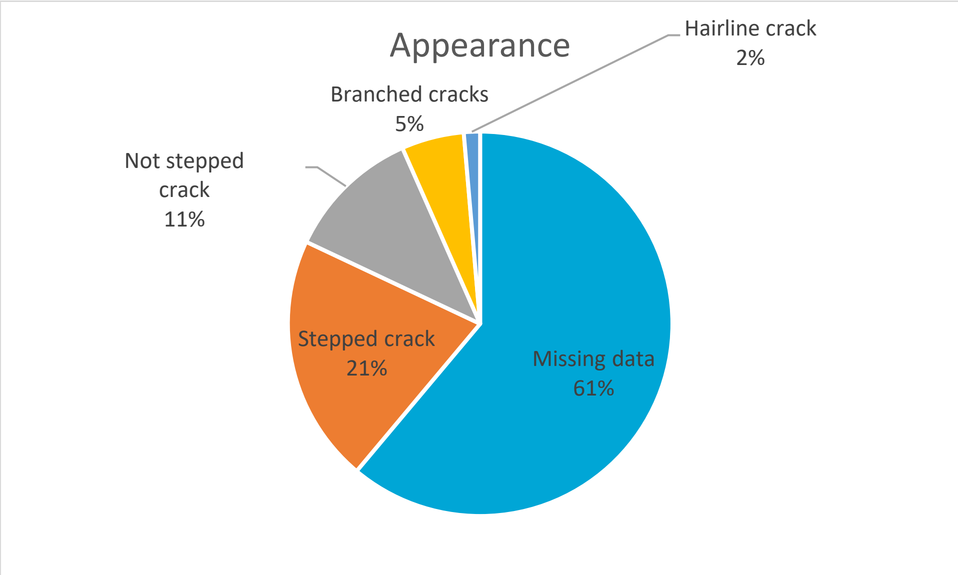


Figure 148 Spreading of the characteristic 'appearance', 61% of the data is missing.

What to do with the missing data?
A large part of the characteristic 'appearance' is missing in the data. This is because it was not always mentioned if a crack was stepped, branched or hairline. Some inspectors also mentioned if a crack was not stepped and this is shown in the data. Besides the notes of the inspector, if a drawing showed that the crack was stepped, it was collected as a stepped crack. It is assumed that if a crack is stepped, it was then collected as a stepped crack. Therefore, it is proposed to substitute the missing data as 'not stepped crack'. This is in violation of the 'WYSIATI' rule for quality protection of data collection (Kahneman, 2013), but is the best option to use the characteristic 'appearance' in the analysis.

Missing data in characteristic 'appearance' is substitute with not stepped crack.

Classification of damage
As described by de Vent, a damage classification system is necessary for good communication about the severity of the damage. This will help to indicate what type of repair actions is needed (De Vent, 2011, p. 23). There are many types of damage classification as thoroughly researched by de Vent. Staalduinen et al use a 3 step classification based on the crack width: light cracks ($w \leq 0.3$mm), medium cracks (0,3 mm < $w \leq$ 3,0 mm) and severe cracks ($w >$ 3,0 mm). A visual judgment on available photos is used when the crack width was not known.

How is the classification of damage stored in the data?
The following possible answers on damage characteristic 'classification of damage' are available in the data as a characteristic:

- Classification of damage [light=0, medium=0.5, large=1.0].

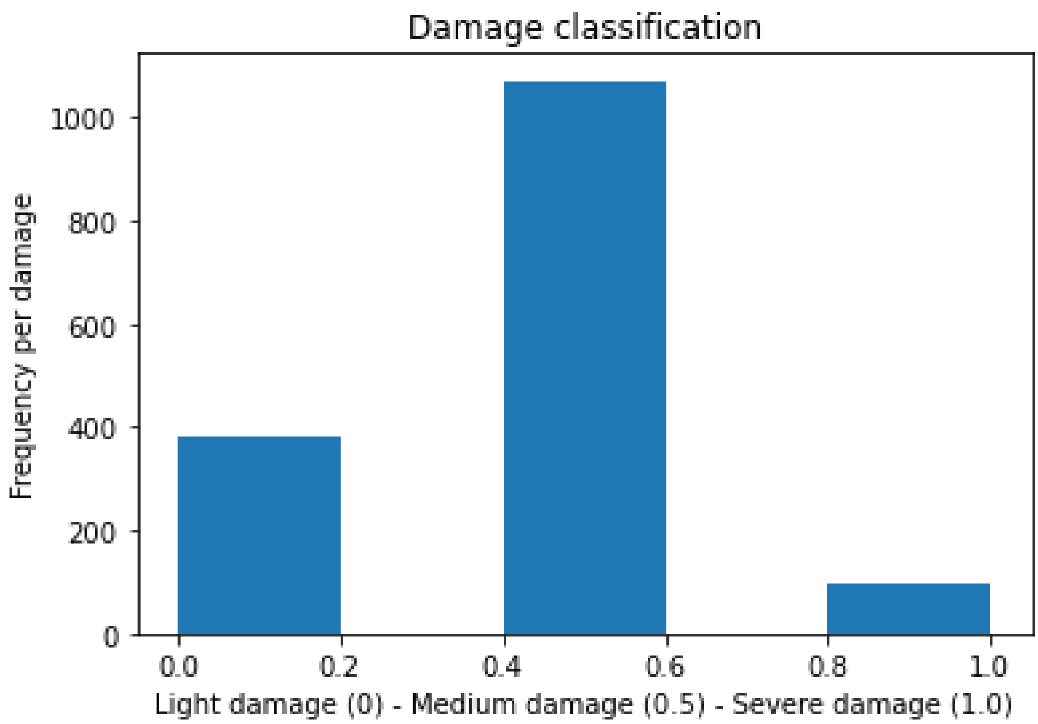


Figure 149 Histogram of characteristic 'classification of damage', data is missing in 5% of the cases.

What to do with the missing data?

Some of the data is missing because too little information about the damage was available for the classification. The crack width was not measured, photos of the damage were not included or the location of damage was not clear during inspection. Most of this cases will result in an unknown damage cause. Those unknown damage causes will be removed in a later moment. The best option for the rest of the missing data is to substitute it with a medium classification, because it is the most frequent classification type.

Missing data in characteristic 'classification of damage' is substitute with medium classification.

Damage pattern according to de Vent

The term *damage pattern* has different meanings in literature. It is not a fixed term. It is mostly used in the context of in the field of research. In this research, the meaning of the term *damage pattern* is based on the way this term is applied in the reverences reports. (De Vent, 2011) (van Staalduinen, Terwel, & Rots, 2018).

Damage pattern, a categorisation of damage solely based on the physical representation and relative location on the damage. For example, a horizontal crack in a masonry wall near an opening is damage pattern 19(H).

Staalduinen et al adjust the damage pattern classification of the de Vent. The adjusted classification is more in line with the analysed buildings in the Groningen area (van Staalduinen, Terwel, & Rots, 2018, p. 42).

How are damage patterns stored in the data?

The following possible answers on damage characteristic 'Damage pattern according to de Vent' are available in the data as a characteristic:

- Damage pattern 1 [yes = 1, no = 0] only for A.X and B.X (van Staalduinen, Terwel, & Rots, 2018, p. 44),
- Damage pattern 2 [yes = 1, no = 0]] only for A.X and B.X (van Staalduinen, Terwel, & Rots, 2018, p. 44),
- Damage pattern 19H [yes = 1, no = 0],
- Damage pattern 19D [yes = 1, no = 0],
- Damage pattern 19V [yes = 1, no = 0],
- Damage pattern 20 [yes = 1, no = 0],
- Damage pattern 21 [yes = 1, no = 0],
- Damage pattern 22 [yes = 1, no = 0],
- Damage pattern 23 [yes = 1, no = 0],
- Damage pattern 24 [yes = 1, no = 0],
- Damage pattern 29 [yes = 1, no = 0],
- Damage pattern 30 [yes = 1, no = 0] only for A.X and C.X (van Staalduinen, Terwel, & Rots, 2018, p. 44),
- Damage pattern 31 [yes = 1, no = 0],
- Damage pattern 33 [yes = 1, no = 0],
- Damage pattern 34 [yes = 1, no = 0] only for A.X and B.X (van Staalduinen, Terwel, & Rots, 2018, p. 44),
- Damage pattern 38 [yes = 1, no = 0],
- Damage pattern 41 [yes = 1, no = 0] only for A.X (van Staalduinen, Terwel, & Rots, 2018, p. 44),
- Damage pattern 48 [yes = 1, no = 0],
- Damage pattern 51 [yes = 1, no = 0],
- Damage pattern 55 [yes = 1, no = 0] only for A.X and B.X (van Staalduinen, Terwel, & Rots, 2018, p. 44),
- Damage pattern 60 [yes = 1, no = 0] only for B.X and C.X (van Staalduinen, Terwel, & Rots, 2018, p. 44).

Visualisation of applied characteristic 'crack pattern according to de Vent'

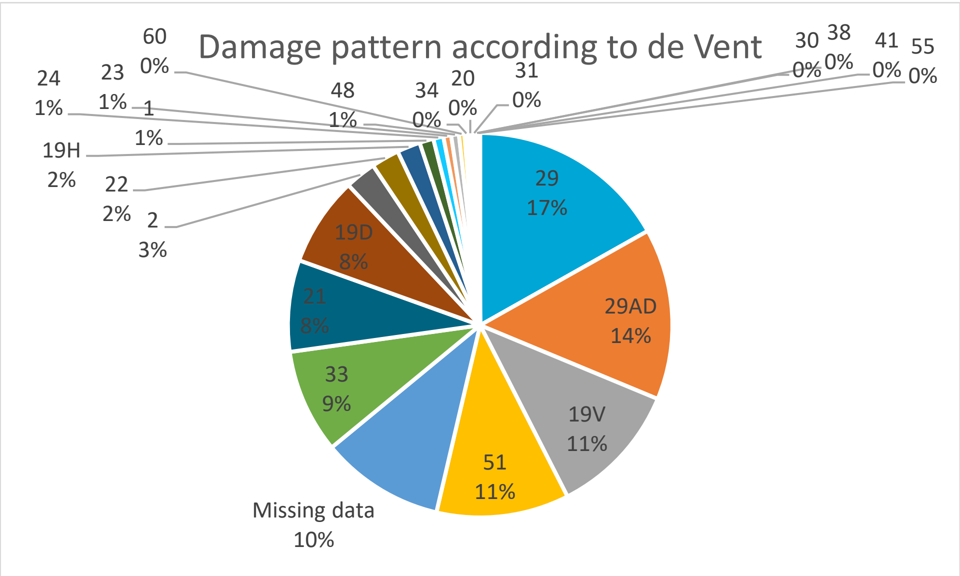


Figure 150 Spreading of the characteristic 'damage pattern according to the Vent, 10% of the data is missing.

What to do with the missing data?

Some data in the characteristic 'damage pattern according to de Vent' is missing because not all damage cases could be categorized in the typology of the Vent. Missing data could be interpreted as 'none of the above'. Therefore, it is proposed to continue with the missing data in de analysis.

Missing data in characteristic 'crack direction' is allowed.

List of all characteristics and summery of all discussed in this appendix above

	A	B	C	% missing data + remarks	
construction year					
Time between construction and damage	x	x	x	0	Linear
construction year	x	x	x	0	Linear
<1900			C.X	0	
19000-1940			C.X		
1940-1970			C.X		
>1970			C.X		
Function					
Living	x	x	x	0	
Farm	x	x	x		
Gathering	x	x	x		
Shed	x	x	x		
Industry	x	x	x		
Type of building					
Detached house	x	x	x	0	
Semidetached house	x	x	x		
Terraced house	x	x	x		
Farm	x	x	x		
Apartment	x	x	x		
Public accessible	x	x	x		
Windmill	x	x	x		
Shed	x	x	x		
Dimensions of the buildings					
Area			C.X	0	Log 10
number of damge / area	x	x	x	0	Linear
longest side		B.X	C.X	0	Log 10
shortest side		B.X	C.x	0	Log 10
Type of facade					
Stretcher bond	x	x	x	5%	Missing data is substitute with stretcher bond, because it is the most frequent one.
Dutch bond	x	x	x		
Flamish bond	x	x	x		
Cross bond	x	x	x		
Wild bond	x	x	x		
Type of roof					

roof cover material					
rooftiles	x	x	x	0	
Timber	x	x	x		
Slates	x	x	x		
Tar and gravel roof	x	x	x		
Corrugated roofs	x	x	x		
Construction material roof					
Timber	x	x	x	26%	Characteristic is removed from the analysis because of too much missing data
Steel	x	x	x		
Concrete	x	x	x		
Stone	x	x	x		
Roof layout					
Flat	x	x	x		
Shed	x	x	x		
Gable	x	x	x		
Hip	x	x	x		
Gambrel	x	x	x		
Mansard	x	x	x		
Corner	x	x	x		
Presence of a basement					
Presence of a basement			C.X	0	
Foundation					
Foundation type					
swallow foundation	A.3.5		C.X	29%	Characteristic is removed, because too much missing data.
Deep foundation	A.3.5		C.X		
Foundation material					
Concrete	A.3.5		C.X	44%	Characteristic is removed, because too much missing data.
Masonry	A.3.5		C.X		
Timber	A.3.5		C.X		
Sand lime stone	A.3.5		C.X		
Supporting walls					
Masonry	x	x	x	46%	Characteristic is removed, because too much missing data.
Timber	x	x	x		
Inner walls					
Masonry	x	x	x	80%	Characteristic is removed, because too much missing data.
Sand lime stone	x	x	x		
Timber	x	x	x		
Presence of dilatations					
Presence of dilatations		B.X	C.X	0	
Possible accumulation of ran and snow					
Possible accumulation of rain and snow	A.X			0	
Indication of leaks					
Indication of leaks	x	x	x	0	

Home improvemetns					
Home improvements	x	x	x	20%	Missing data is substitute with no improvements done
State of maintenance					
State of maintenance	x	x	x	0	
Soil build-up					
Shallow ground			C.X	0	
Deep ground			C.X	0	
Groundwater level					
Change of GWL			C.X	31%	Missing data is substitute with no change in ground water level, because is the most frequent one.
Trees					
Presence of shrub		B.X	C.X	0	
Height of shrub		B.X	C.X	0	
Distance to shrub		B.X	C.X	0	
Relative influance shrub		B.X	C.X	0	
Presence of tree		B.X	C.X	0	
Height of tree		B.X	C.X	0	
Distance to tree		B.X	C.X	0	
Relative influance tree		B.X	C.X	0	Linear
Ground settlement					
Ground settlement decline			C.X	6%	Missing data is substitute with no ground settlement, because it is the most frequent one
Ground settlement lift			C.X		
No ground settlement			C.X		
Traffic					
Distance to road	A.3		C.X	14%	Log 10 Missing data in characteristic 'road traffic' is substitute with 130
Distance to railway	A.3		C.X	0	Characteristic removed, because only one option was present
Construction activities					
Construction activities within 0 m	A.3		C.X	0	
Construction activities within 20 m	A.3		C.X	0	
Construction activities within 100 m	A.3		C.X	0	
Industrial activieties					
Industrial activities within 100 m	A.3		C.X	0	
Calamities					
Occurance of a calamity	A.4			0	
Vibrations due to earthquakes					
PGV Bommer et al	A.3.5		C.X	0	Log 10
PGV Staalduinen et al	A.3.5		C.X	0	
PGV GMM v6	A.3.5		C.X	0	

Appendix 1: Discussion of characteristics

Position of damage					
Damage is inside the building	x	x	x	0	
Damge is outside the building	x	x	x		
Damge in in building part					
Building part facade	x	x	x	<5	Missing data is removed with the corresponding damage cases
Building part inner wall	x	x	x		
Building part roof	x	x	x		
Building part floor	x	x	x		
Building part ceiling	x	x	x		
Building part basement wall	x	x	x		
Building part outside-/garden wall	x	x	x		
Direction of facade					
North facade		B.X		25%	Missing data is allowed, because characteristic is not always present.
East facade		B.X			
South facade		B.X			
West facade		B.X			
Building element					
Damage is in masonry	x	x	x	18%	Missing data is substitute with lime sand stone, because it is the most frequent one for inner walls (missing data)
Damage is in Calcium-Silicate	x	x	x		
Damage is in concrete	x	x	x		
Damage is in timber	x	x	x		
Damage is in glass	x	x	x		
Damgae is in plasterboard	x	x	x		
Position of building element					
Building element in plane	x	x	x	4%	Missing data is removed with the corresponding damage cases
Building element near opening	x	x	x		
Building element above opening	x	x	x		
Building element under opening	x	x	x		
Building element near door	A.4.1				
building elemen near corner	x	x	x		
In finish or through construction	x	x	x	89%	Characteristic is removed from the analysis because of too much missing data
crack in joint	x	x	x	73%	Characteristic is removed from the analysis because of too much missing data
crack in unit	x	x	x		
crack in joint and unit	x	x	x		
Moment of first detecton					

First dectection of damage	(see construction year)				Linear
Displacement out-of-plane & in-plane					
Building part displacement out-of-plane	x	x	x	8%	Missing data is substitute with 'no out-of-plane displacement, because it is the most frequent one
Building part displacement in-plane			C.X	10%	Missing data is substitute with 'no in-plane displacement, because it is the most frequent one
Settlement differences over the length of the building					
Settlement differences			C.X	0	
Skewness of floor/building					
Skewness of floor/building			C.X	0	
Displacement in crack					
Displacement in the crack in-plane	x	x	x	25%	Characteristic is removed from the analysis because of too much missing data
Displacement in the crack out-of-plane	x	x	x		
Crack width over the length of the crack					
Crack width is constant	x	x	x	68%	Missing data is substitute with 'constant width', because it is the most frequent one
Crack width is variable	x	x	x		
Crack width is widen to one side	x	x	x		
Crack width is widen to both sides	x	x	x		
Crack pattern					
					Characteristic is removed from the analysis because of too much missing data
Dirt					
Crack is polluted	x	x	x	0	
Erosion					
Erosion of crack edges	x	x	x	0	
Applied finish over crack					
Applied finish of stucco	x	x	x	21%	Missing data is substitute with 'no applied finish for outside walls and stucco for inside walls, because it are the most frequent ones
Applied finish of plasterwork	x	x	x		
Applied finish of wallpaper	x	x	x		
Applied finish of tiles	x	x	x		
Applied finish of paint	x	x	x		
No applied finish	x	x	x		
Presence of repair work					
Presence of repair work	x	x	x	3%	Missing data is substitute with 'no repair work present', because it is the most frequent one
Direction					
Crack direction vertical	x	x	x	14%	

Crack direction horizontal	x	x	x		Missing data is allowed, because characteristic is not always present.
Crack direction diagonal	x	x	x		
Runs from-to					
Runs from-to opening	x	x	x	2%	Missing data is allowed, because characteristic is not always present.
Runs from-to door	x	x	x		
Runs from-to corner	x	x	x		
Runs from-to edge	x	x	x		
Runs from-to roof	x	x	x		
Runs from-to ceiling	x	x	x		
Runs from-to floor	x	x	x		
Runs from-to ground level	x	x	x		
Runs from-to plane	x	x	x		
Crack width					
Maximum crack width	x	x	x	64%	Log10 Missing data is substitute with 1,249mm, because it is the mean of the known data.
Crack length					
Maximum crack length	x	x	x	52%	Log10 Missing data is substitute with 0,8835m, because it is the mean of the known data.
Appearance					
Stepped cracks	x	x	x	61%	Missing data substitute with 'not stepped crack', because it is the assumed default damage
Not stepped	x	x	x		
Hairline crack	x	x	x		
Branched cracks	x	x	x		
Classification of damage					
Classification of damage	x	x	x	5%	Missing data is substitute with medium classification, because it is the most frequent one
Damage pattern according to de Vent					
Damage pattern 1	A.X	B.X		10%	Missing data is allowed, because characteristic is not always present.
Damage pattern 2	A.X	B.X			
Damage pattern 19H	x	x	x		
Damage pattern 19D	x	x	x		
Damage pattern 19V	x	x	x		
Damage pattern 20	x	x	x		
Damage pattern 21	x	x	x		
Damage pattern 22	x	x	x		
Damage pattern 23	x	x	x		
Damage pattern 24	x	x	x		
Damage pattern 29	x	x	x		
Damage pattern 30	A.X		C.X		
Damage pattern 31	x	x	x		
Damage pattern 33	x	x	x		
Damage pattern 34	A.X	B.X			
Damage pattern 38	x	x	x		

Damage pattern 41	A.X				
Damage pattern 48	x	x	x		
Damage pattern 51	x	x	x		
Damage pattern 55	A.X	B.X			
Damage pattern 60		B.X	C.X		

Appendix 2: Earthquake characteristics selection

The TU Delft database includes 12 characteristic that are related to the context characteristic *vibrations by earthquake*. The context characteristic *vibrations by earthquake* is then over-represented in the analysis. The idee is that a better result would be obtained if it could be reduced to one variable. Another issue is that not all earthquakes characteristic are available when a user want to use the results of the analysis. The main questions are applied to choose between the 12 characteristics:

How meaningful is the found relation? > Which characteristic is the most meaningful?

How reliable is the found relation? > Which characteristic is the most reliable?

TNO P(v ≥ 0,5 mm s)
TNO P(v ≥ 1 mm/s)
TNO P(v ≥ 2 mm/s)
TNO P(v ≥ 5 mm/s)
TNO P(v ≥ 10 mm/s)
Bommer P(v ≥ 0,5 mm/s)
Bommer P(v ≥ 1 mm/s)
Bommer P(v ≥ 2 mm/s)
Bommer P(v ≥ 5 mm/s)
Bommer P(v ≥ 10 mm/s)
Max. Vibrations velocity Bommer [mm/s]
Max. Vibrations velocity P. van Staalduinen [mm/s]

Table 13 Twelve characteristic that are related to context characteristic; vibrations by earthquake

In the next paragraphs, it is explained how the TU Delft has obtained the 12 characteristics. This is done to give an answer on how reliable each characteristic is. Thereafter is described how the TU Delft applied the characteristic there research. That will give an answer on how meaningful the characteristic are.

Earthquake models

Models can nowadays quantify what the chance is on exceeding a specific acceleration level, given the magnitude of an earthquake at a given location compared to an epicentre (van Staalduinen, Terwel, & Rots, 2018, p. 55). The most important factor, when assessing the structural safety of buildings and infrastructure, is the ground acceleration (van Staalduinen, Terwel, & Rots, 2018, p. 55). In the assessment of the vibration load, the vibration velocity is the general test variable (SBR, 2017, p. 30).

Pruiksma and Rozsas did a research on how historical earthquakes can cause a vibration at the surface. They used the data from the TNO sensor network and the KNMI sensor network (van Staalduinen, Terwel, & Rots, 2018, p. 51) (Pruiksma & Rózsás, 2017, p. 2/71). Bommer et al did also a research on how historical earthquakes causes a vibration at the surface (Bommer, Stafford, & Ntinalexis, Empirical Ground-Motion Prediction Equations for Peak Ground Velocity from Small-Magnitude Earthquakes in the Groningen Field Using Multiple Definitions of the Horizontal Component of Motion, 2016). They used only the data of the KNMI sensor network (van Staalduinen, Terwel, & Rots, 2018, p. 51).

The KNMI network measures: magnitude, depth and location of epicentre. The KNMI data consist approximately of 1400 earthquakes. The measurements take place at the surface and in the ground, for example, at 200 m depth (van Staalduinen, Terwel, & Rots, 2018, p. 52). The TNO network is a sensor network to register vibrations in a building (van Staalduinen, Terwel, & Rots, 2018, p. 52). This sensor measure the acceleration of the vibration (Pruiksma & Rózsás, 2017, p. 10/70).

There are two models available:

- A model developed by Pruiksma and Rozsas of TNO, commissioned by NCG, with data of the TNO sensor network and the KNMI sensor network,
- A model developed by Bommer et al of the Imperial Collage London, commissioned by NAM, with data of the KNMI sensor network (Bommer, Stafford, & Ntinalexis, Empirical Ground-Motion Prediction Equations for Peak Ground Velocity from Small-Magnitude Earthquakes in the Groningen Field Using Multiple Definitions of the Horizontal Component of Motion, 2016).

There are two earthquake datasets available:

- KNMI network, with measurements in the ground (magnitude, depth and location of epicentre)
- TNO network, with measurements in buildings (vibration acceleration of a building).

Pruiksma and Rozsas model (TNO)

The approach of Pruiksma and Rozsas is to determine the distance between the epicentre and the measurement point, for earthquakes that were higher or equal to a magnitude of 2.5 (This were 5 earthquakes in total). Then they linked this distance to the measured peak value of the horizontal vibration velocity. This relation was then approached with the following model (model T1.1) (van Staalduinen, Terwel, & Rots, 2018, p. 56):

$$\ln(Sv) = d_1 + g(R) \tag{0-1}$$

$$(R) = d_2 \ln\left(\sqrt{{R_{epi}}^2 + d_3}\right) \tag{0-2}$$

In which:

Sv is the response quantity,
R_{epi} is the shortest distance to the epicentre,
$d_{1,2,3}$ are the model parameters.

It is also possible to use more than five earthquakes, this model is called V2 (van Staalduinen, Terwel, & Rots, 2018, p. 56). 14 earthquakes have been included for model V2 (Pruiksma & Rózsás, 2017, p. 17/70) (the table on (van Staalduinen, Terwel, & Rots, 2018, p. 57) is incorrect). Model V2 is explained in (Pruiksma & Rózsás, 2017, p. 23/70). The model of Pruiksma and Rozsas is not elaborate further in this report in anticipation of the outcome of Staalduinen et al (van Staalduinen, Terwel, & Rots, 2018, p. 66). They concluded to use the model of Bommer et al.

Bommer et al model (NAM)

The moddel of Bommer et al only operates with data of the KNMI. This GMM (ground-motion model) is called v3. Bommer et al also includes more earthquakes than Pruisma and Rozsas. Bommer et al had 22 earthquake events in total (Bommer, Stafford, & Ntinalexis, Empirical Ground-Motion Prediction Equations for Peak Ground Velocity from Small-Magnitude Earthquakes in the Groningen Field Using Multiple Definitions of the Horizontal Component of Motion, 2016, p. 2) (table 7.3 of (van

Staalduinen, Terwel, & Rots, 2018, p. 58) is incorrect). That are eight more earthquakes than the TNO V2 Model. That is because previous models only included measurements from B-stations. By also including the measurements of the G-stations in the analysis, more earthquakes could be included in the development of the model (Bommer, Stafford, & Ntinalexis, Empirical Ground-Motion Prediction Equations for Peak Ground Velocity from Small-Magnitude Earthquakes in the Groningen Field Using Multiple Definitions of the Horizontal Component of Motion, 2016, p. 2).

The measuring instruments of the B-stations are placed in a borehole. This instruments measure the vibrations in three directions. They are called Geofoons. Several Geofoons have been placed in one borehole at a vertical distance of 50 meter from each other. The depth of a borehole varies between 200 and 300 meters (KNMI, sd).

The measuring instruments of G-stations are mainly placed at the surface. This instruments measure the acceleration at the surface. Therefore, they are called accelerometers (KNMI, sd).

The use of the G-stations in the development of this model was possible because the influence of the V_{s30} (the 30-metre time-average shear-wave velocity in m/s) was found to be negligible (Bommer, Stafford, & Ntinalexis, Empirical Ground-Motion Prediction Equations for Peak Ground Velocity from Small-Magnitude Earthquakes in the Groningen Field Using Multiple Definitions of the Horizontal Component of Motion, 2016, p. 9). In other words: “The influence of the soil condition on the formulation of the vibration velocity was found to be very small.” (van Staalduinen, Terwel, & Rots, 2018, p. 57)

Bommer et al determine the PGV (peak ground velocity) at a requested location using the distance from the epicentre and the magnitude of the earthquake. This relation is found in 3 steps:

- determination of the magnitude-dependent distance saturation term R (van Staalduinen, Terwel, & Rots, 2018, p. 59),
- determination of the geometrical spreading term g(R) (Bommer, Stafford, & Ntinalexis, Empirical Ground-Motion Prediction Equations for Peak Ground Velocity from Small-Magnitude Earthquakes in the Groningen Field Using Multiple Definitions of the Horizontal Component of Motion, 2016, p. 9),
- determination of the local PGV.

Determination of the magnitude-dependent distance saturation term R

The distance to the origin of an earthquake is commonly defined by measuring the distance to the hypocenter, the epicentre or the Joyner-Boore distance (see Figure 151). The chosen distance depends on the purpose of the research and the height of the magnitude (Akkar & Bommer, 2007, p. 6). For the Groningen area, Bommer et al (2016, p. 9) defined the following equation to calculate distance:

$$R = \sqrt{R_{epi}^2 + [e^{(0.4233M - 0.6083)}]^2} \qquad (0\text{-}3)$$

Where R_{epi} is the distance to the epicentre and M the magnitude of the earthquake. The given constants in equation (0-3) are respectively c_5 and c_6 (Bommer, et al., 2016, p. 5).

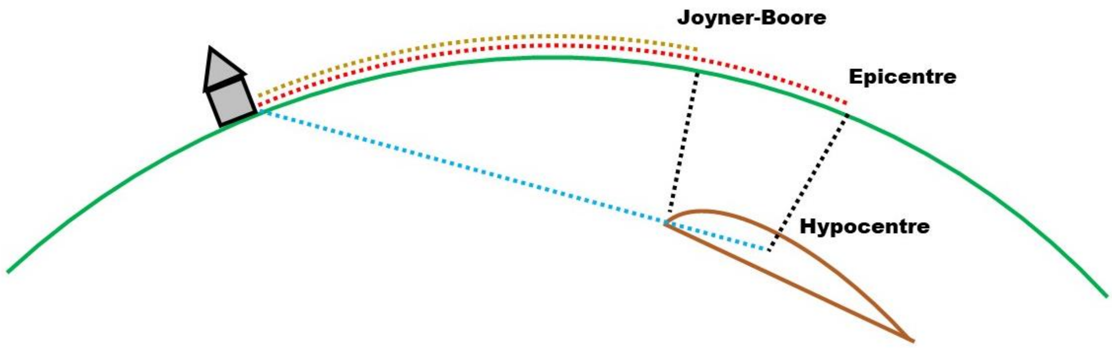


Figure 151 Distance to hypocentre (blue), distance to epicentre (red) and Joyner-Boore distance (Yellow).

Determination of the geometrical spreading term g(R)

In (Bommer, Stafford, & Ntinalexis, Empirical Ground-Motion Prediction Equations for Peak Ground Velocity from Small-Magnitude Earthquakes in the Groningen Field Using Multiple Definitions of the Horizontal Component of Motion, 2016, p. 9) is explained how the g(R) is determined.

$$\begin{aligned} g(R) &= c_4 \ln(R) && R \leq 6.32km \\ g(R) &= c_4 \ln(6.32) + c_{4a} \ln\left(\frac{R}{6.32}\right) && 6.32km < R \leq 11.62km \\ g(R) &= c_4 \ln(6.32) + c_{4a} \ln\left(\frac{11.62}{6.32}\right) + c_{4b} \ln\left(\frac{R}{11.62}\right) && R > 11.62km \end{aligned} \quad \text{(0-4)}$$

In which R is the modified distance of (0-3) and c_4, c_{4a} and c_{4b} are the model coefficients. These coefficients are determined with a regression analysis.

Determination of the local PGV

The peak ground velocity is determined with (Bommer, Stafford, & Ntinalexis, Empirical Ground-Motion Prediction Equations for Peak Ground Velocity from Small-Magnitude Earthquakes in the Groningen Field Using Multiple Definitions of the Horizontal Component of Motion, 2016, p. 9):

$$\ln(PGV) = c_1 + c_2 M + g(R) \quad \text{(0-5)}$$

In which:

PGV is the peak ground velocity in cm/s,
M Is the magnitude of the earthquake according to the data of the KNMI,
$g(R)$ Is the geometrical spreading term according to (0-4).

The model coefficients c_1 and c_2 are determined with a regression analysis.

Regression analysis and standard deviation

The model coefficients of (0-4) and (0-5) are determined with a maximum likelihood regression per earthquake (Bommer, et al., 2016, p. 5). This regression per earthquake event results in a standard deviation. It is an uncertainty of measurements of one earthquake. This is called an intra-event uncertainty (Salet & Bruurs, 2017, p. 98 | 148). The intra-event model coefficients combined in inter-event model coefficients for the Groningen area with a regression analysis. With this a standard deviation arises and is called the inter-event deviation (Salet & Bruurs, 2017, p. 98 | 148). The intra-event deviation and the inter-event deviation can be combined in a standard deviation for the entire model (Bommer, Stafford, & Ntinalexis, Empirical Ground-Motion Prediction Equations for Peak Ground Velocity from Small-Magnitude Earthquakes in the Groningen Field Using Multiple Definitions of the Horizontal Component of Motion, 2016, p. 11).

$$\sigma = \sqrt{\tau^2 + \phi^2} \tag{0-6}$$

In which:
σ is the total standard deviation of the entire model,
τ is the inter-event standard deviation (differences between earthquake events),
ϕ is the intra-event standard deviation (differences between measurements of one earthquake event).

Application of models in TU Delft research

The model of Bommer et al was chosen in the research of TU Delft (van Staalduinen, Terwel, & Rots, 2018, p. 66). To motivate the choice, both models were compared and it was concluded that (van Staalduinen, Terwel, & Rots, 2018, p. 63):

- Bommer et al found lower values for a R_{epi} distance of less than 20 km,
- Pruiksma and Rozsas found higher values for larger distance of R_{epi},
- The two models are generally quite similar.

The TNO itself has also investigated whether the TNO sensor network presents the same values as calculated by the Bommer et al model (Geurts, 2017). It should be noted in this regard that the model of Pruiksma and Rozsas is not only based on the TNO sensor network, but also on the KNMI sensor network data (Pruiksma & Rózsás, 2017, p. 2/71). They concluded that:

- The TNO sensor network is quite similar with the model of Bommer et al,
- Only a few measurements of the TNO sensor network are outside the 95% confidence interval compared to the Bommer et al model in Hellum, Wirdum and Scharmer,
- The TNO sensor network is systematically not over- or underestimating in comparison with the Bommer et al model,
- The average observed values of the TNO model are slightly higher than the average values form the Bommer at al model when the R_{epi} is less than 7km.

In the *Appendix 3: Determination of the horizontal peak velocity* is described what the different ways are to determine the horizontal peak velocity. The actual occurring peak values is used instead of the geometric average what is conventional in seismology, because it is a better estimation of the vibration velocity occurring at an arbitrary location near the Groningen field (van Staalduinen, Terwel, & Rots, 2018, p. 55). This means that the TU Delft has chosen to use the PGV_{Larger} (as $PGV_{max\ hor}$) and the PGV_{RotMax} (as $PGV_{max\ rot}$) in there research. It is unclear from the report and the available data which of the two was used in their assessment. The SBR vibration guideline states that the V_{Top} must be determined by taking the highest value of the two directions (SBR, 2017, p. 45). It is therefore plausible that the PGV_{Larger} was chosen.

Where Bommer et al developed their model with 22 earthquake events, the TU Delft applied this model to the complete database of the KNMI, consist of 1400 earthquakes (van Staalduinen, Terwel, & Rots, 2018, p. 66). When assessing a building, earthquake events before the construction year of the building and earthquake events after the damage occurred, are excluded from the model. This is evident from the available research data of the TU Delft.

As indicated earlier, the model used contains a standard deviation. This makes it possible to calculate the probability that a certain vibration velocity is exceeded. The TU Delft uses an exceedance probability of 25% (van Staalduinen, Terwel, & Rots, 2018, p. 95). In other words, a vibration velocity is calculated with a 25% chance of exceeding. In summary, TU Delft has used the vibration speed per building per earthquake with an exceedance probability of 25% based on the Bommer et al model with 1400 earthquakes measured by the KNMI.

The data from TU Delft shows they use also the model of Staalduinen to calculate the maximum vibration velocity. In this model, an S-curve graph is plotted in the points: P(v ≥ 0.5 mm/s), P(v ≥ 1.0 mm/s), P(v ≥ 2.0 mm/s), P(v ≥ 5.0 mm/s) and P(v ≥ 10.0 mm/s). The 25% probability of exceedance is then read from here. No literature or report is available for this method. Of the two PGV that were determined (Max. Vibrations velocity Bommer [mm/s] and Max. Vibrations velocity P. van Staalduinen [mm/s]), the maximum value of the two was used in the assessment. The maximum of those as always the Staalduinen PGV.

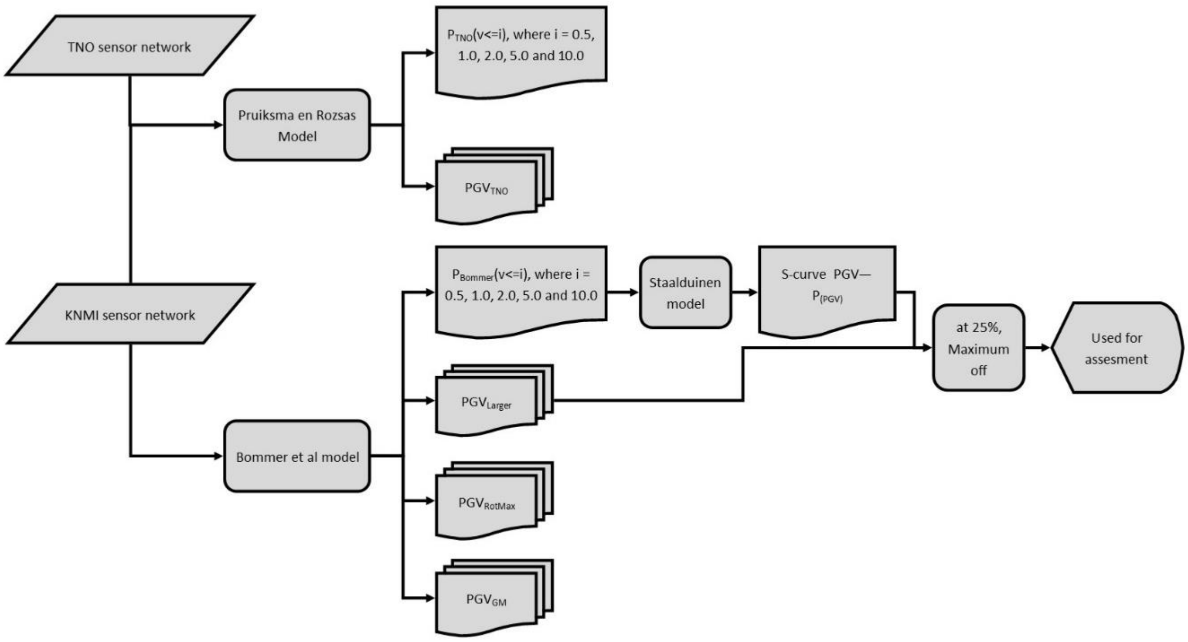


Figure 152 Visual overview of determination of PGV for assessment in case of building damage.

Characteristic choice regarding earthquakes

Which characteristic is the most meaningful?

The characteristic 'exceedance probability' are not meaningful for the assessment of damage. The chance that a vibration was larger than 5.0 mm/s has no direct link with an occurrence of damage. Literature explains that this is the acceleration of the vibration or the velocity of the vibration. That is why the characteristics *TNO P(v ≥ 0.5 mm/s)* up to and including *Bommer P(v ≥ 10 mm/s)* are not included in the analysis. Adding to this is that these characteristic are very specific for the TU Delft research and may not available when applying the results of this analysis.

Which characteristic is the most reliable?

What remains are the maximum ground velocity according to the Bommer et al model and the Staalduinen model. There are no references or reports found for the Staalduinenen model and therefore the Bommer et al model would be more reliable. In addition to that adds the Staalduinen model an extra fault to the result, because they try to fit a graph on the data. On the other side it is not clear which type of horizontal peak velocity is used in the Bommer et al model. Ultimately, the maximum of the two characteristics (Bommer and Staalduienen) was used in the research at TU Delft. The reasoning behind this is also not clear or referenced.

Appendix 2: Earthquake characteristics selection

Two options remain for choosing the characteristic:

- *PGV Bommer et al (GMMv3 at 25%) [mm/s]* , according to literature.
- *PGV Staalduinen et al (GMM v3 at 25%) [mm/s]*, according to application of the TU Delft.

The latter option is chosen, because it is also used in the dataset to be analyzed. This makes it more likely that relationships will be found in the data.

Appendix 3: Determination of the horizontal peak velocity

As indicated earlier, research into earthquake damage in Groningen uses the vibration velocity in the horizontal direction (van Staalduinen, Terwel, & Rots, 2018, p. 55 & 56). The Geofoons of the KNMI measure this horizontal peak velocity in two directions that are perpendicular to each other (KNMI, sd). Bommer et al explores 4 methods to calculate these two components in 1 peak horizontal direction, where PGV stands for 'peak ground velocity' (Bommer, Stafford, & Ntinalexis, Empirical Ground-Motion Prediction Equations for Peak Ground Velocity from Small-Magnitude Earthquakes in the Groningen Field Using Multiple Definitions of the Horizontal Component of Motion, 2016, p. 1). This 4 methods are:

- PGV_{GM} – Geometric mean value,
- PGV_{Larger} – The maximum of each component,
- PGV_{Pyth} – The Pythagoras theorem of each horizontal peak velocity per earthquake,
- PGV_{RotMax} – Largest peak on the velocity trace.

PGV_{GM} – Geometric mean value

The mean value of two vectors can be calculated by taking the square root of the product of the two vectors. In the equations below, the PGV_{NS} and the PGV_{EW} are the peak values during a vibration of two horizontally measured velocities, measured perpendicular to each other. The disadvantage of this method is that the maximum peak velocity is never found, because this is always averaged with a lower peak velocity in the other direction. There is even a theoretical possibility that when the vibration wave is purely in the direction of one of the components, the PGV_{GM} will be zero regardless of the height of the other peak velocity.

$$PGV_{GM} = \sqrt{PGV_{NS}PGV_{EW}} = \exp\left[\frac{\ln(PGV_{NS}) + \ln(PGV_{EW})}{2}\right] \qquad \text{(0-1)}$$

PGV_{Larger} – The maximum of each component

The maximum value of the two horizontal peak velocity components is selected in this method (Bommer, Stafford, & Ntinalexis, Empirical Ground-Motion Prediction Equations for Peak Ground Velocity from Small-Magnitude Earthquakes in the Groningen Field Using Multiple Definitions of the Horizontal Component of Motion, 2016, p. 4). The chance that a vibration wave will come purely from a direction of a horizontal measuring component is quite small. The disadvantage of this method is that the PGV is not calculated in the main direction, which will most likely be higher than the PGV of a horizontal component.

$$PGV_{Larger} = \max[PGV_{NS}, PGV_{EW}] \qquad \text{(0-2)}$$

PGV_{Pyth} – The Pythagorean theorem of each horizontal peak velocity per earthquake

Where previous two methods depend on the direction, this method calculates the horizontal peak velocity in the main direction using the Pythagorean theorem. The disadvantage of this method is that the horizontal peak velocity is used from both components. Those peak velocities do not have to occur simultaneously during an earthquake. Therefore a horizontal peak velocity is calculated in the main direction that may not have occurred (Bommer, Stafford, & Ntinalexis, Empirical Ground-

Motion Prediction Equations for Peak Ground Velocity from Small-Magnitude Earthquakes in the Groningen Field Using Multiple Definitions of the Horizontal Component of Motion, 2016, p. 8).

$$PGV_{Pyth} = \sqrt{{PGV_{NS}}^2 + {PGV_{EW}}^2} \tag{0-3}$$

PGV_{RotMax} – Largest peak on the velocity trace

The PGV_{RotMax} is an improved version of the PGV_{Pyth} that does not look at the maximum peak value occurring during a quake of a component, but the velocity of the two components that occurred at that moment. This prevents the PGV in the main direction from being calculated using two PGV in the component direction that did not occur at the same time during an earthquake. The theory of the PGV_{RotMax} was investigated by Watson-Lamprey et al. and then Bommer et al (Bommer, Stafford, & Ntinalexis, Empirical Ground-Motion Prediction Equations for Peak Ground Velocity from Small-Magnitude Earthquakes in the Groningen Field Using Multiple Definitions of the Horizontal Component of Motion, 2016, pp. 5-8) investigated how the different methods relate to each other.

$$PGV_{RotMax} = \max\left[\sqrt{V_{NS}(t)^2 + V_{EW}(t)^2}\right] \tag{0-4}$$

Appendix 4: Causes of damage

A. Loads	1. Insufficient resistance	1. Initial
		2. Building renovations
		3. Aging
	2. Overloading due to use	1. Normal use
		2. Change of use
		3. Building renovations
	3. Overloading due to vibrations	1. Road traffic
		2. Rail traffic
		3. Building activities
		4. Industrial activities
		5. Seismic activities
	4. Overloading accidental loads	1. Impact / collision
		2. Explosion
		3. Rain or Snow
		4. Storm
		5. Lighting
B. Deformations	1. Hindered	1. Initial
		2. Rebuilding / extension
		3. Aging / deterioration
	2. Imposed	1. Initial
		2. Rebuilding / extension
		3. Corrosion
		4. Tree roots or branches
C. Unequal settlements	1. Unchanging loads	1. Autonomous settlements
	2. Load changes on sub soil	1. Rebuilding / Extension
		2. Building activities
		3. Elevation
		4. Excavation
		5. Road
		6. Railway
	3. Changes in sub soil	1. Change of GWL (ground water level)
		2. Change of GWL for work activities
		3. Change of GWL by authorities
		4. Road traffic vibrations
		5. Rail traffic vibrations
		6. Work activities vibrations
		7. Earthquakes
		8. Change of GWL by trees
		9. Deep sub soil effects
		10. Natural changes of GWL

Appendix 5: Results classification Runs – Decision tree

These are the results of a classification run with Decision tree. The setting for most optimal results are shown in the table below.

	GSS settings			Stopping rules for first run		Stopping rules for second run		
Pattern	Folds	Validation size	Groups p. val. set	Max depth	Min samples leaf	Max depth	Min samples leaf	End result Accuracy
A.1.1	41	0.21	9	15	3	7	2	0.38
A.3.5	41	0.07	3	4	5	6	1	0.65
B.1.1	41	0.09	4	5	2	6	7	0.56
B.2.1	41	0.21	9	8	2	5	5	0.66
C.1.1	41	0.48	20	6	5	5	15	0.56
C.X	41	0.07	3	7	19	5	18	0.44

Table 14 Optimal settings for classification run with decision tree algorithm. 1549 damage cases are included in the cross-validation and 281 damage in the cases test set. GSS cross-validation is applied for pattern validation.

The first column in Table 14 indicates each analysed pattern. As explained in chapter 4, each damage cause has its own pattern (see Figure 19). The next three columns displays the best GSS cross-validation settings in order to construct a sufficient large validation set for accuracy calculations. For example, pattern A.1.1 has 41 folds (column folds) and 21% of the buildings are placed in a validation set (column Validation size). This will lead to 9 buildings in one validation set (column groups p. validation set). The next four columns describes the stopping rules settings for decision tree (explained in chapter 3). As discussed in chapter 4, a construction of a pattern consist of two runs. The first run provides a characteristic selection for the second run in order to reduce the amount of noise in the results. The last column shows the end result of each pattern calculated with an independent test set.

	GKF settings			Stopping rules for first run		Stopping rules for second run		
Pattern	Folds	Min val. size	Max val. size	Max depth	Min samples leaf	Max depth	Min samples leaf	End result Accuracy
A.1.1	7	5	7	15	6	16	1	0.51
A.3.5	27	1	2	5	26	4	18	0.61
B.1.1	11	1	5	4	7	7	10	0.77
B.2.1	7	5	7	6	5	6	2	0.69
C.1.1	4	9	11	5	2	5	1	0.43
C.X	16	1	4	7	20	5	29	0.33

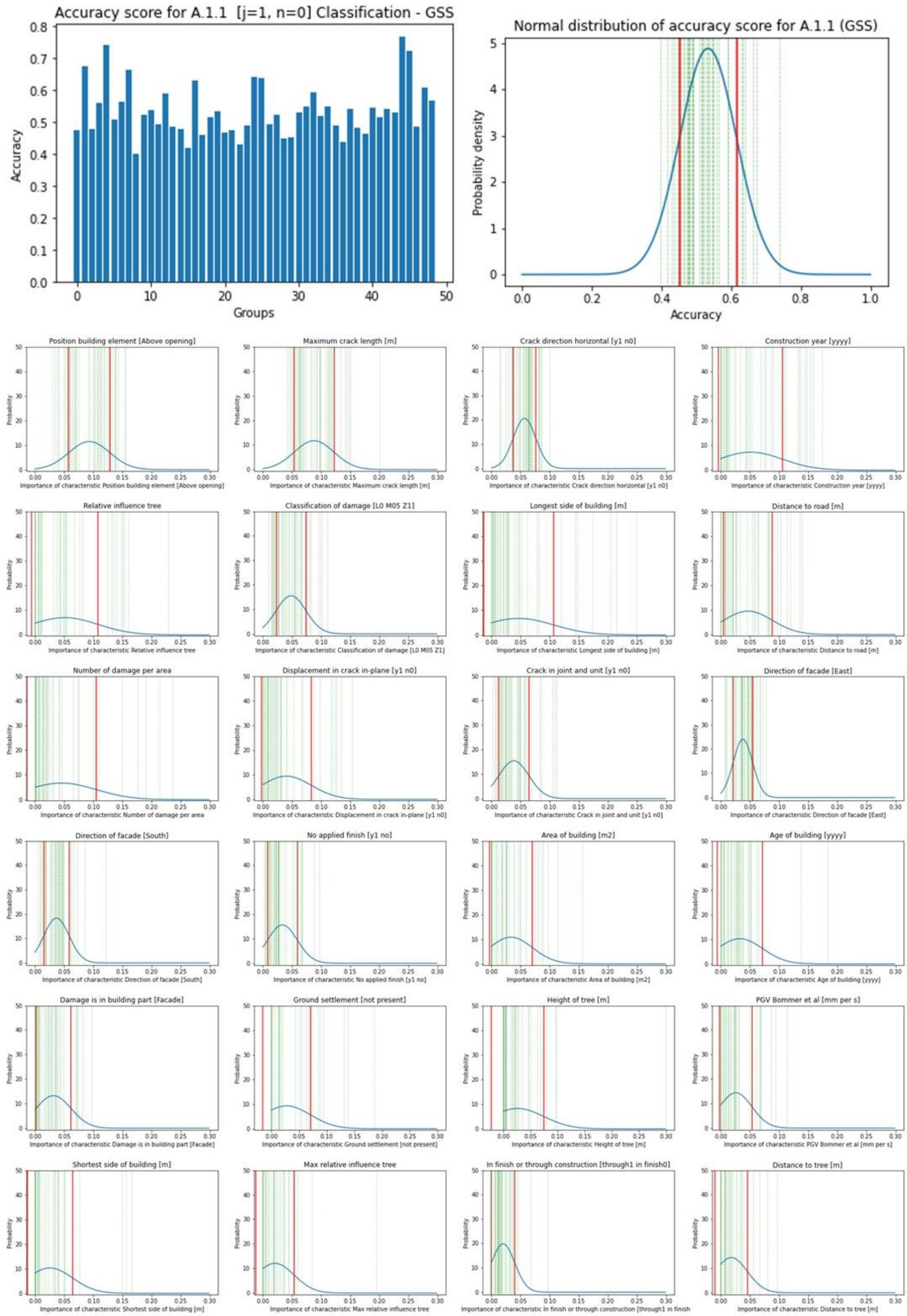
Table 15 Optimal settings for classification run with decision tree algorithm. 1549 damage cases are included in the analysis and 281 damage cases in the test set. GKF cross-validation is applied for pattern validation.

A GKF cross-validation was also done. The settings and result are shown in Table 15. This table has the same layout as Table 14, but has a few differences due to a different cross-validation. As mentioned in the report, the validation set has to have a large enough size. To obtain this desired size, the number of folds is set in the group-K-fold cross-validation (second column of Table 15). This result of a validation set size presented in the third and fourth column.

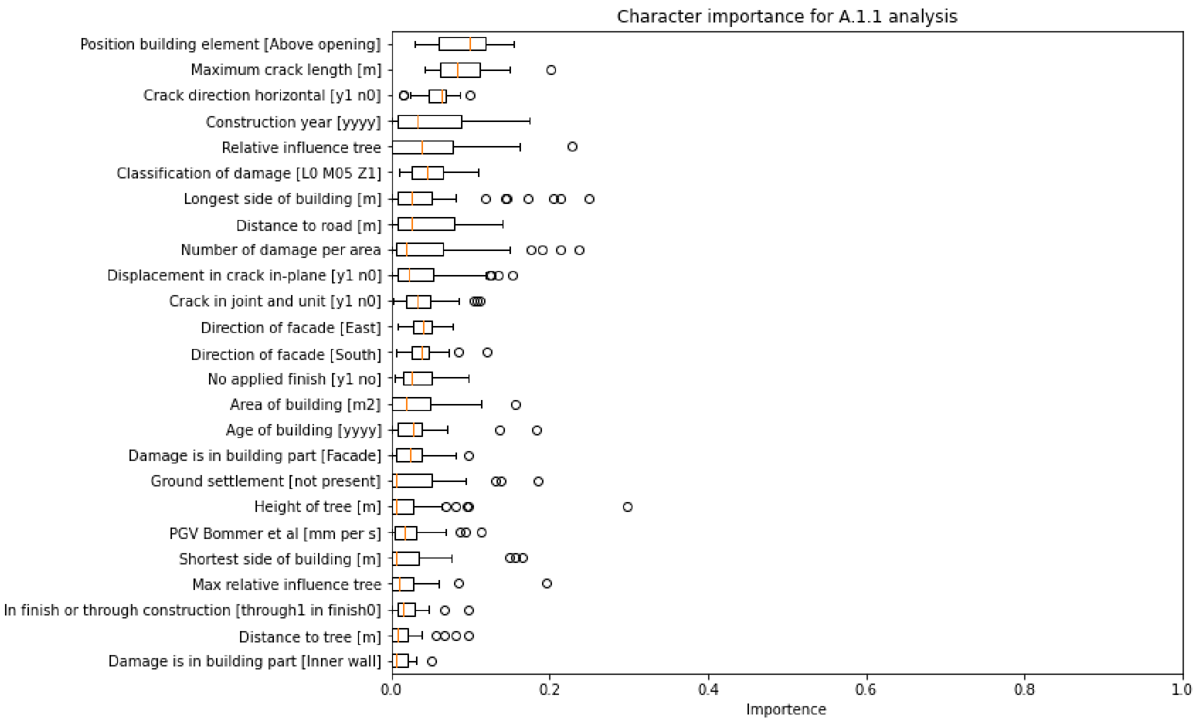
Appendix 5: Results classification Runs – Decision tree

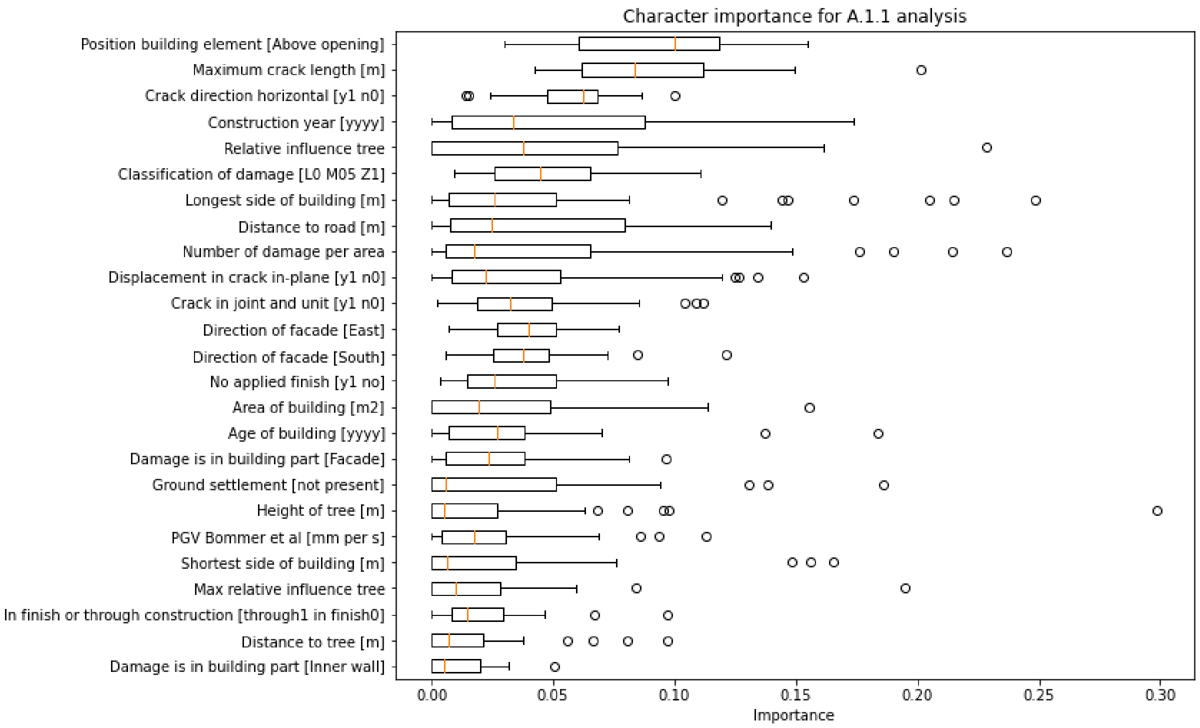
In the following pages, each pattern is made visual with the same graphs as in chapter 4. The blue bar plots are have the same layout as Figure 20. The figure left of the blue bar plot is equal to Figure 21. The smaller normal distribution plots underneath these figures are representations of the characteristic importance for that pattern. The construction of those smaller figures is also the same as Figure 21.

Pattern A.1.1 - Initial insufficient resistance to bear loads

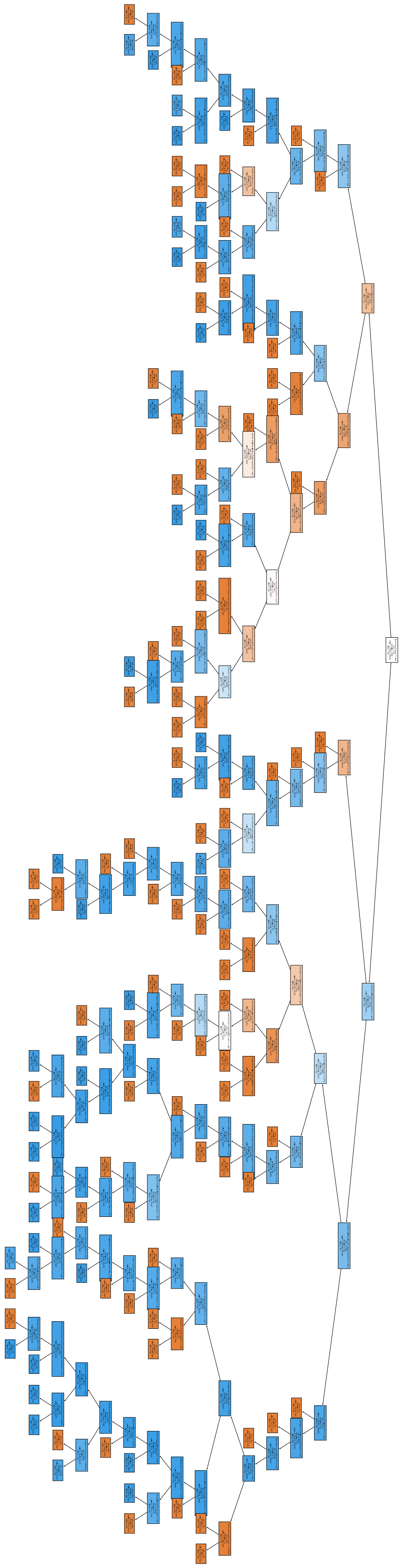


Appendix 5: Results classification Runs – Decision tree

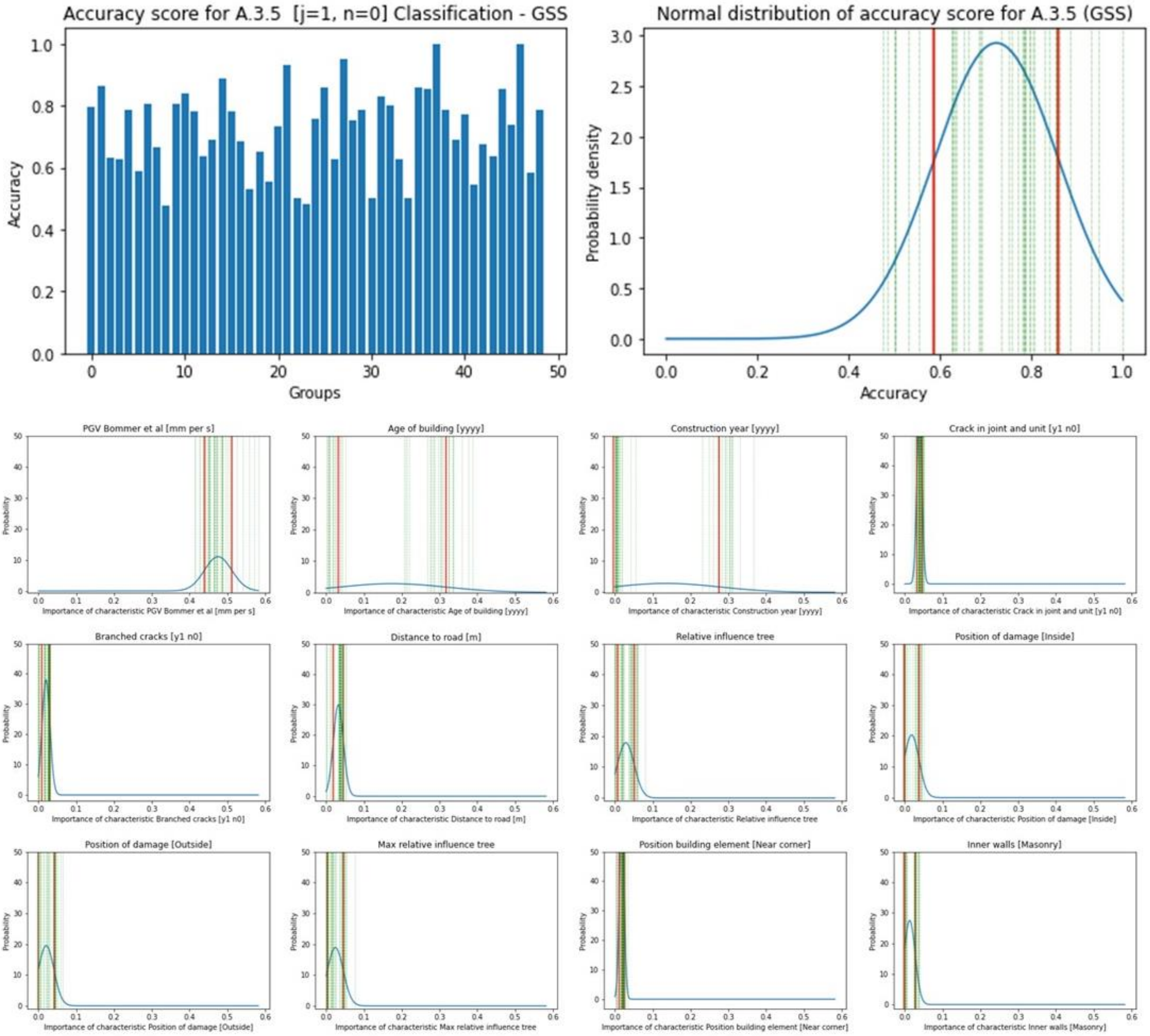




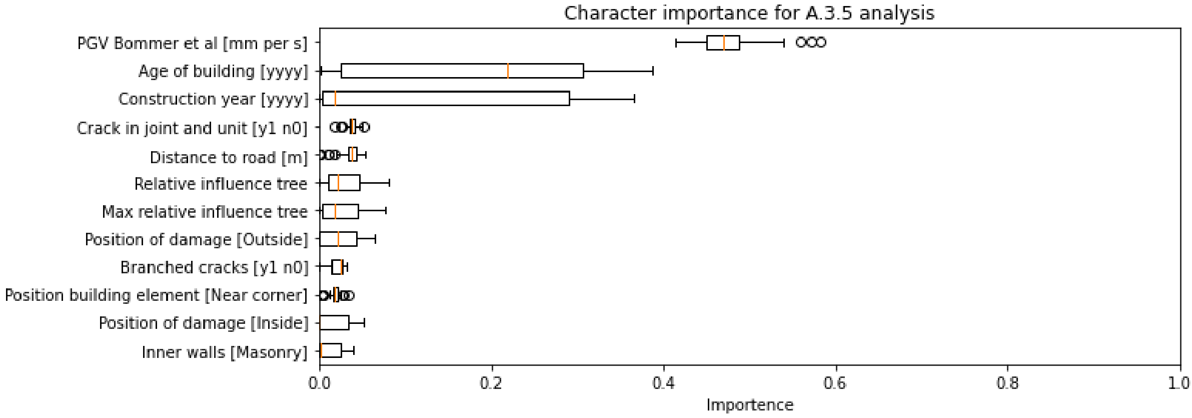
Appendix 5: Results classification Runs – Decision tree

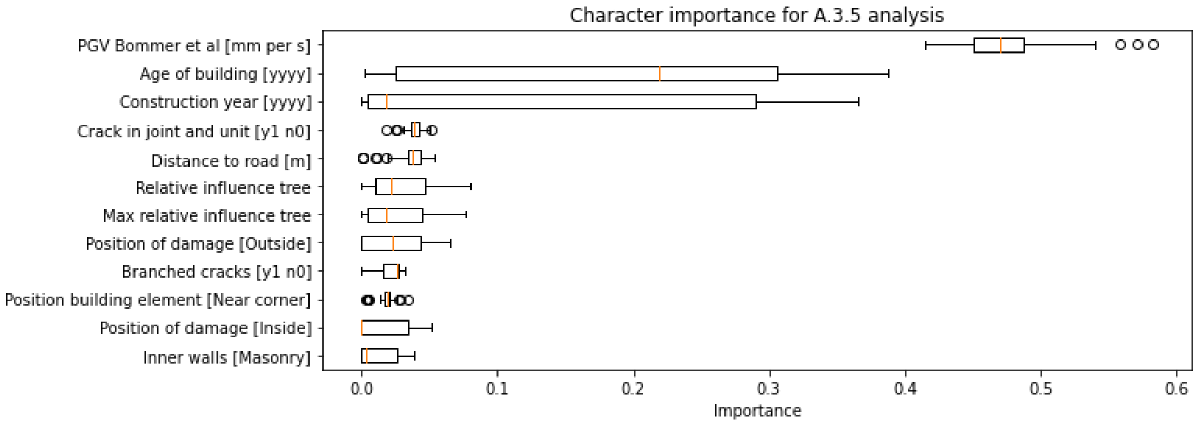


Pattern A.3.5 - Overloading due to vibrations by earthquakes

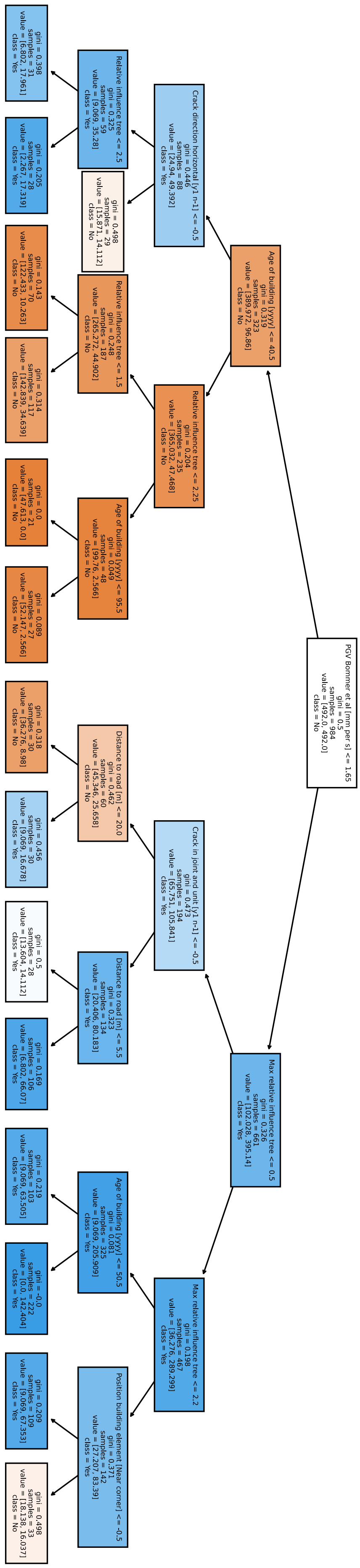


Appendix 5: Results classification Runs – Decision tree

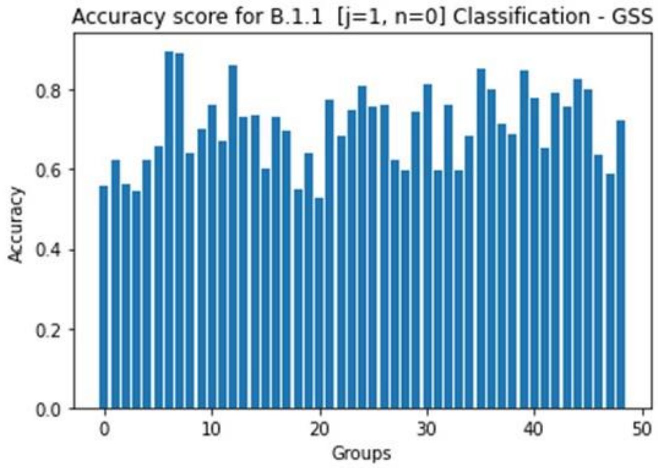


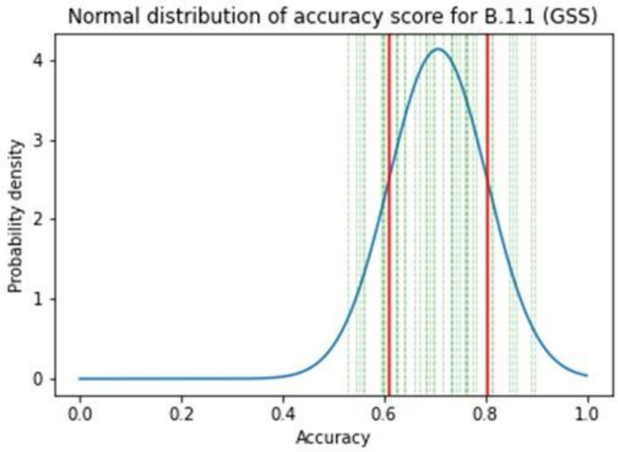


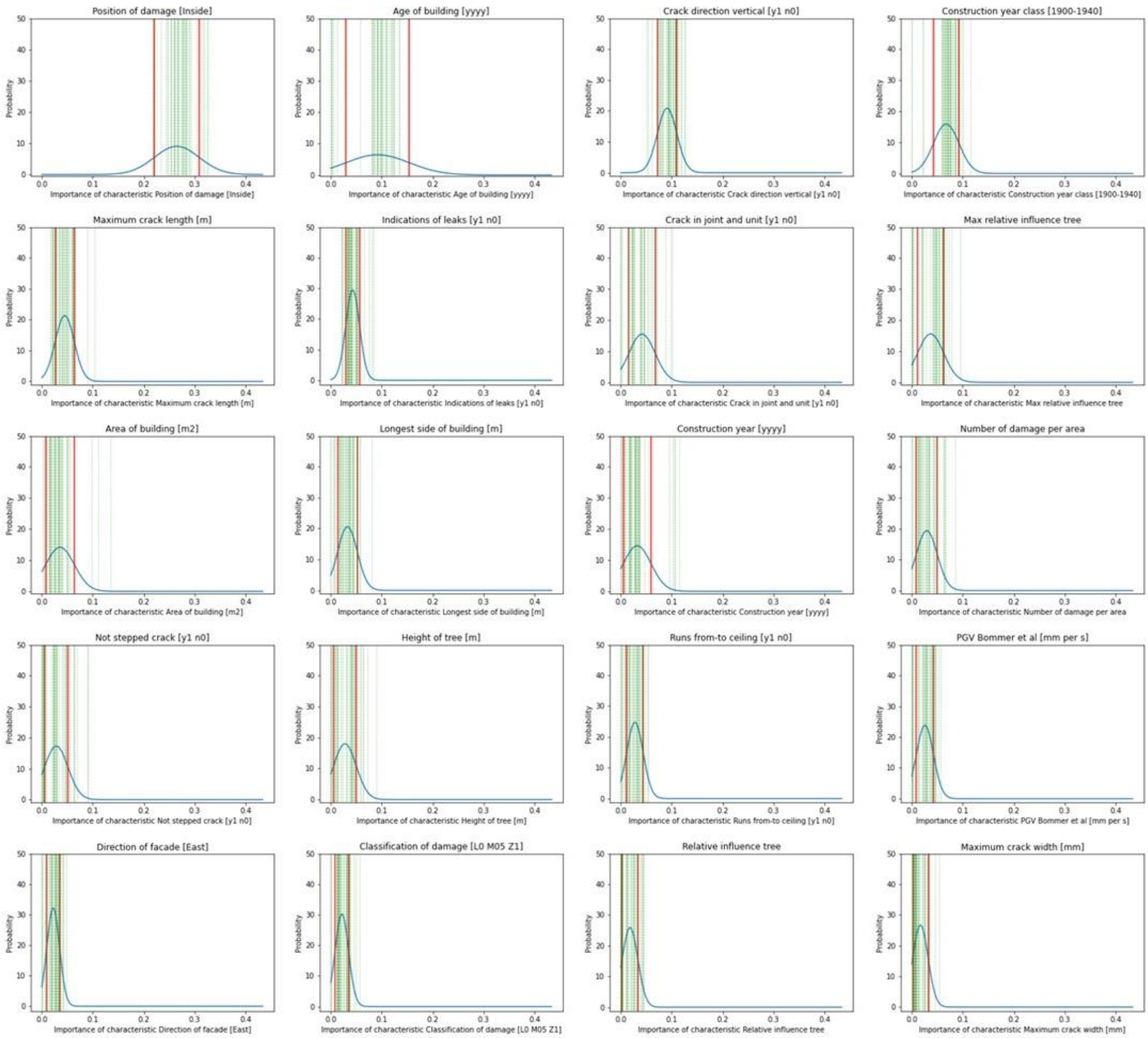
Appendix 5: Results classification Runs – Decision tree



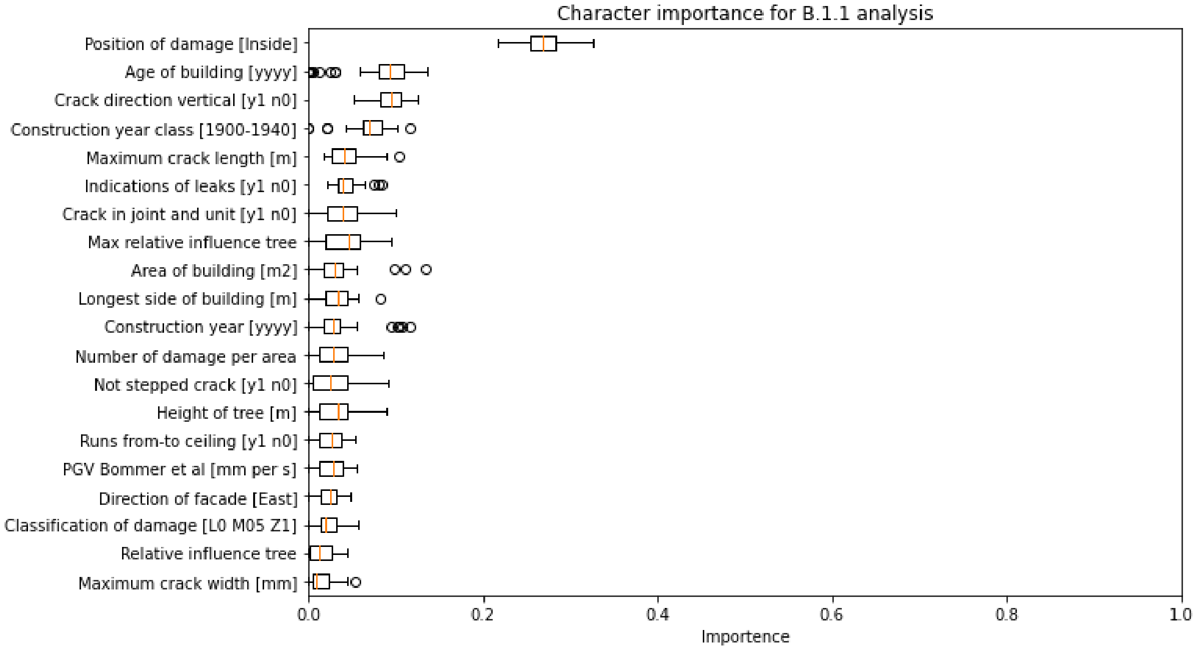
Pattern B.1.1 - Initial hindered deformations

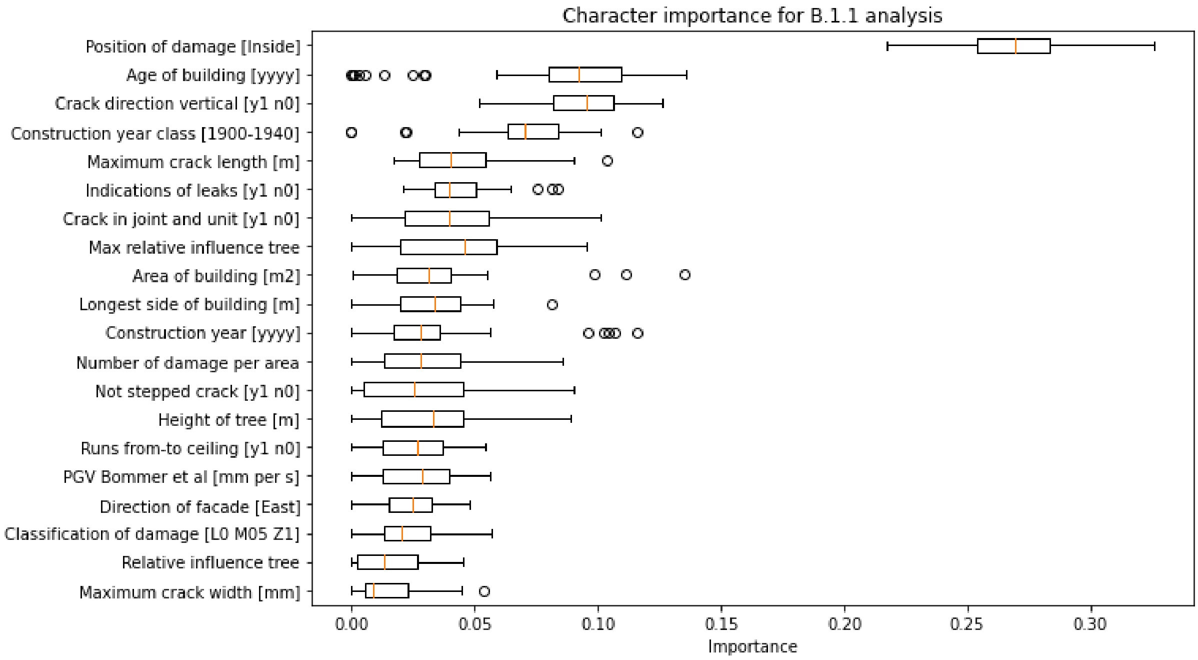




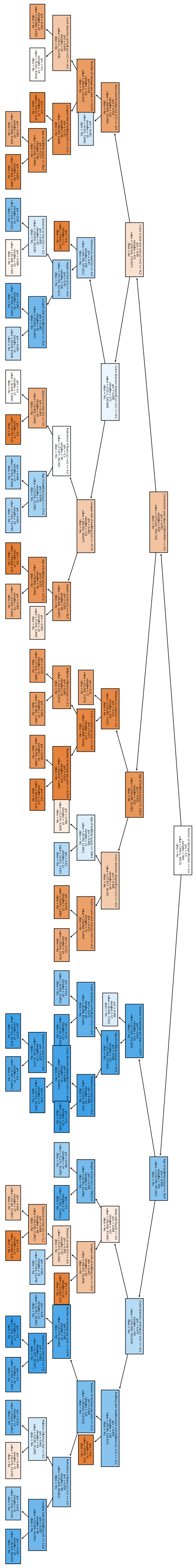


Appendix 5: Results classification Runs – Decision tree

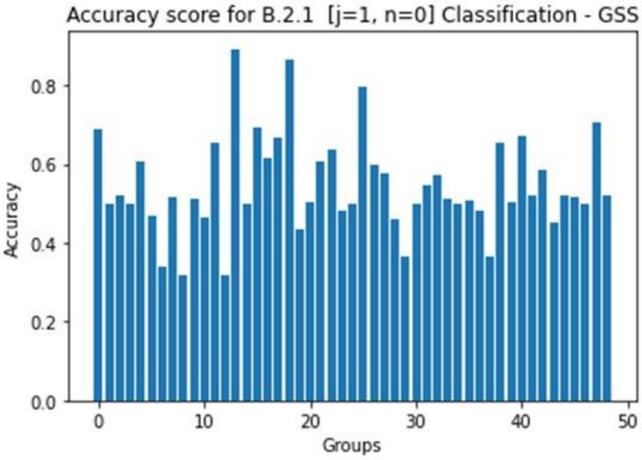


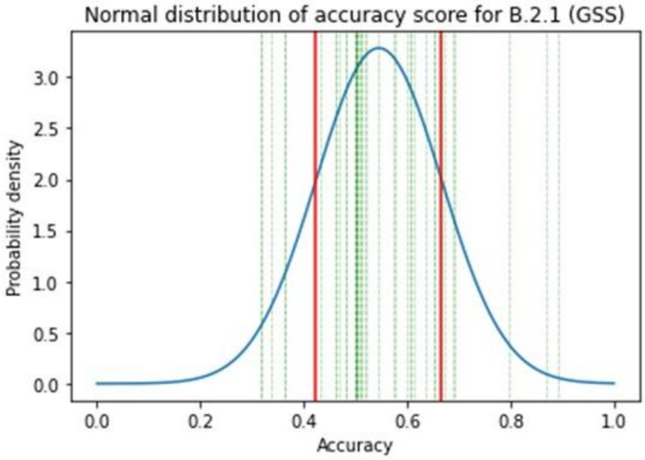


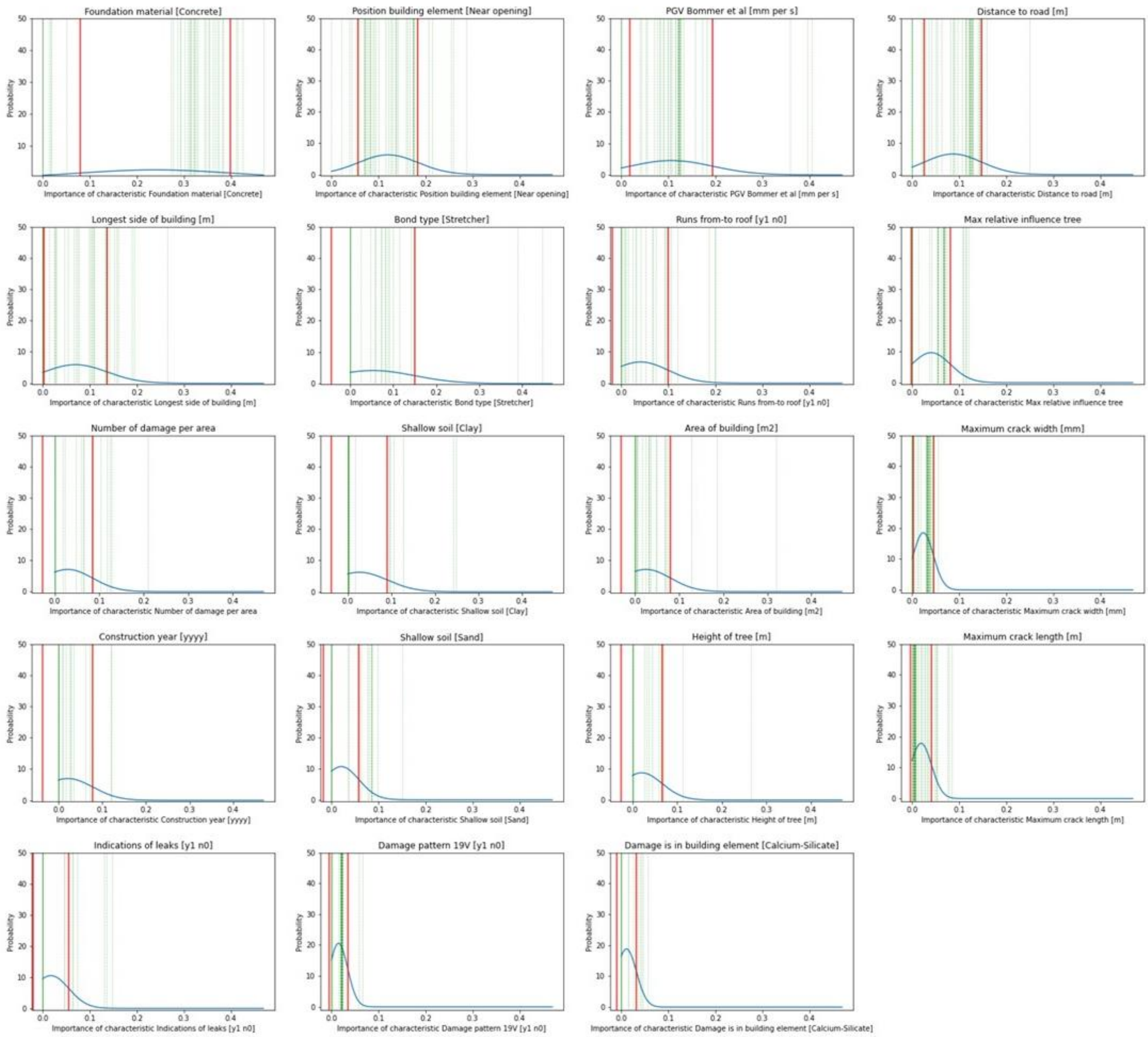
Appendix 5: Results classification Runs – Decision tree



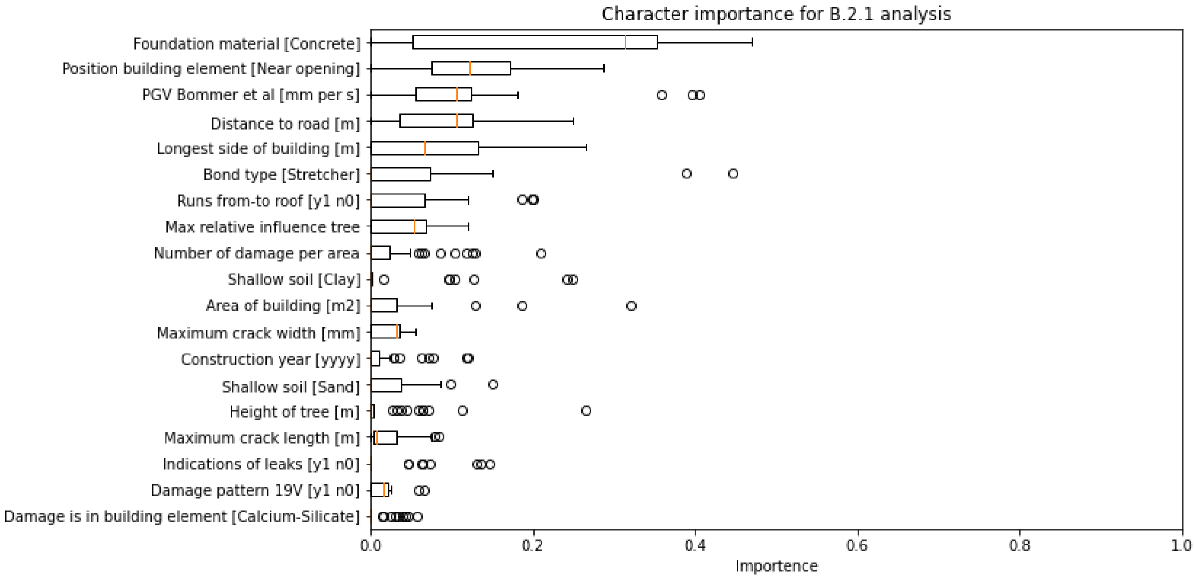
Pattern B.2.1 - Initial imposed deformations

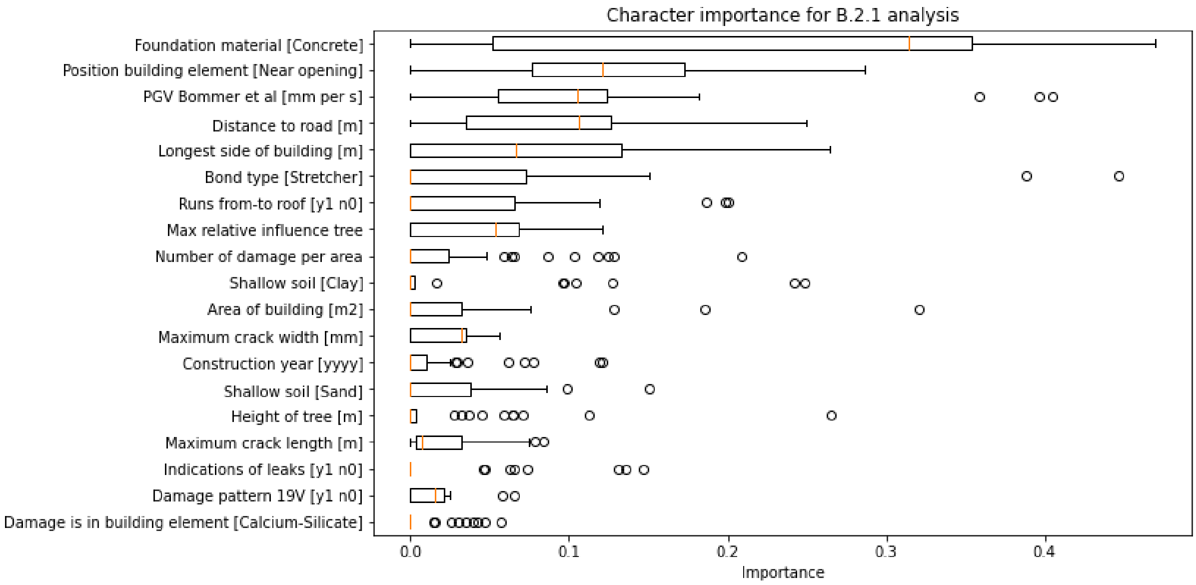




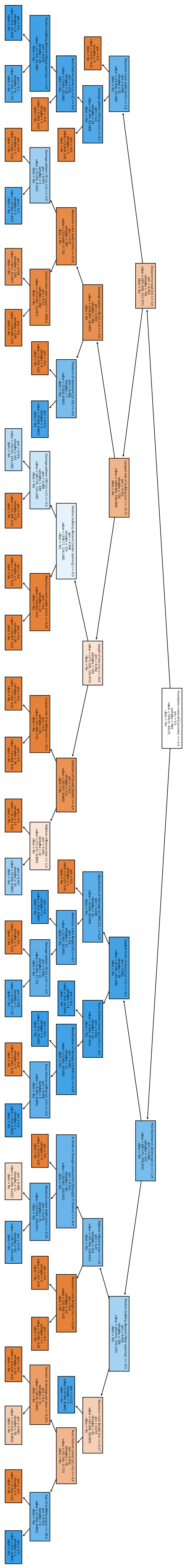


Appendix 5: Results classification Runs – Decision tree

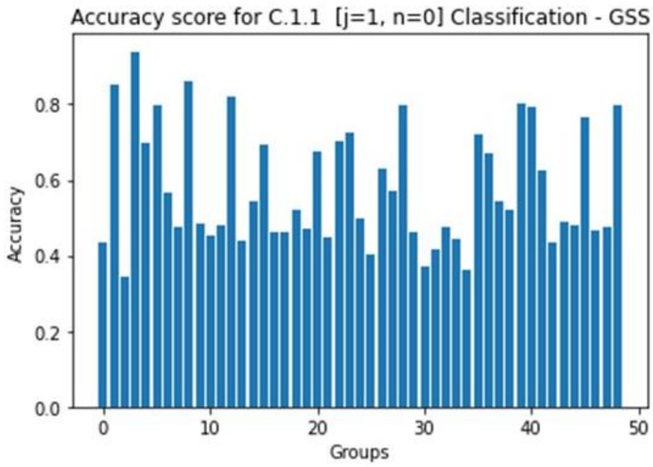


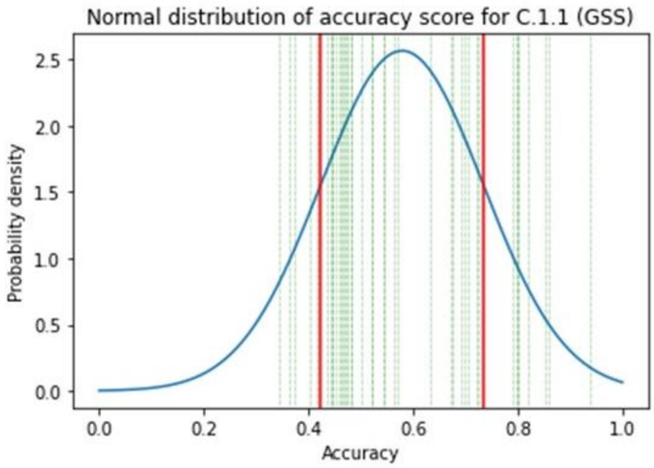


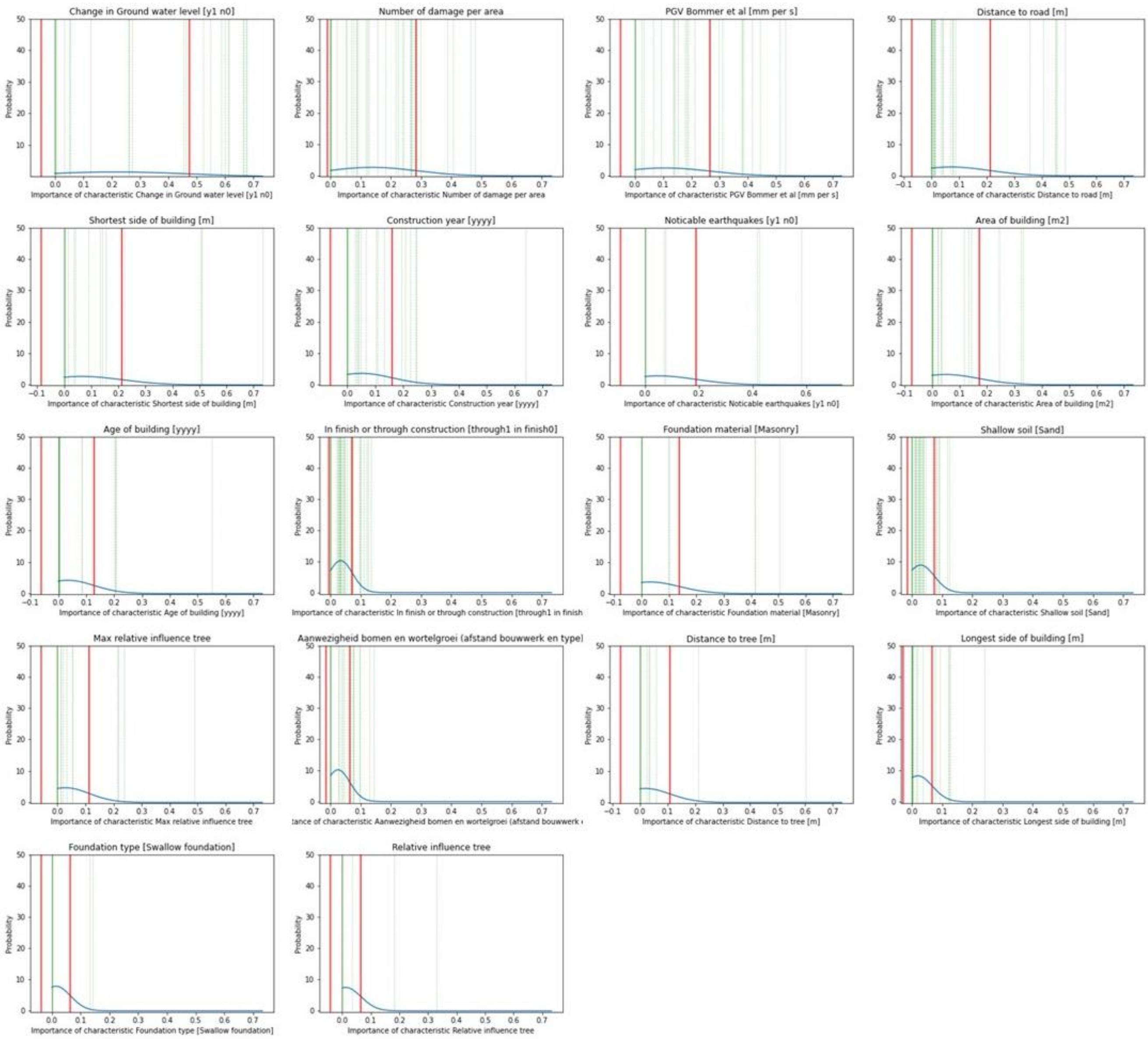
Appendix 5: Results classification Runs – Decision tree



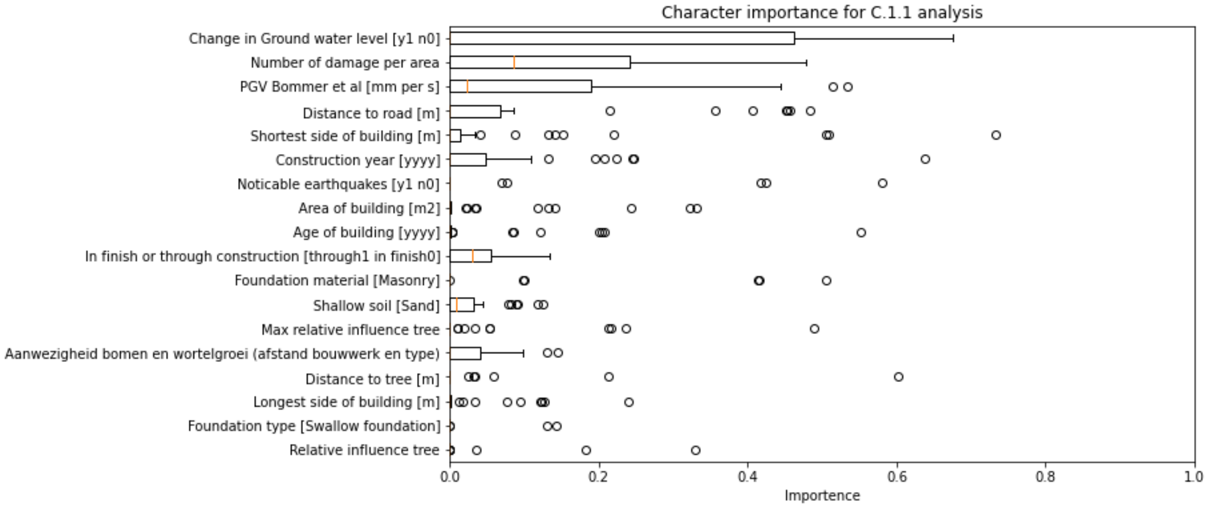
Pattern C.1.1 - Unequal settlements with equal loads

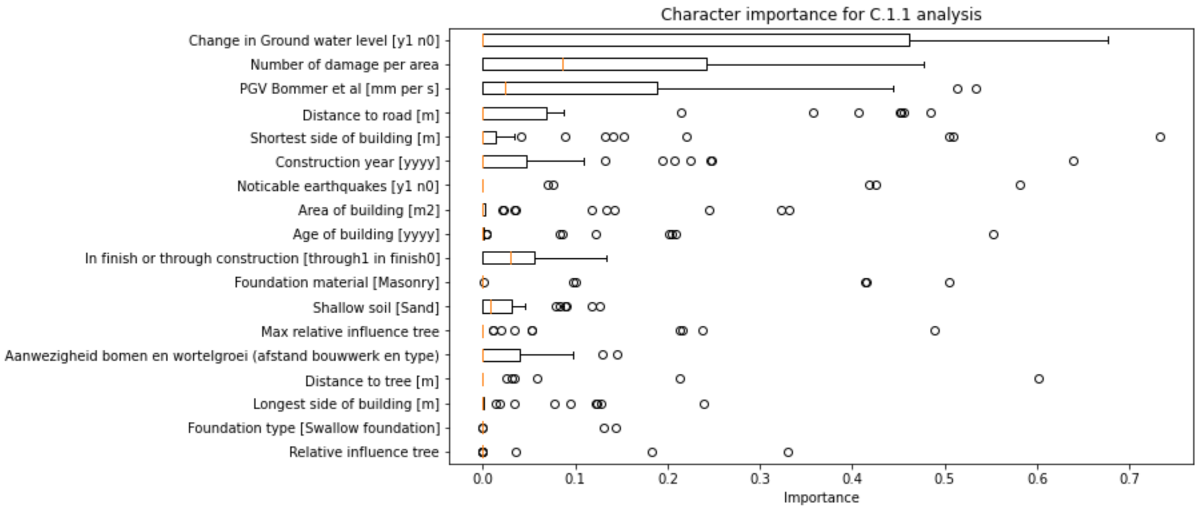




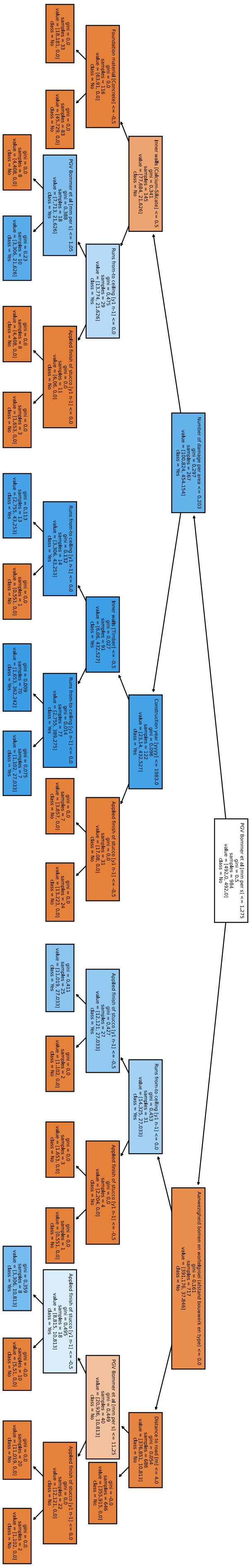


Appendix 5: Results classification Runs – Decision tree

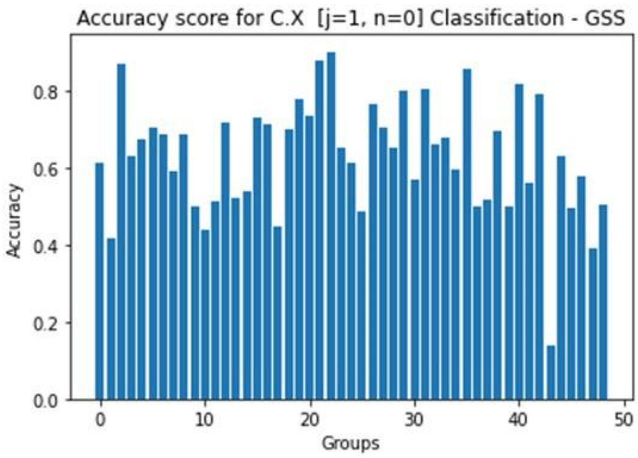


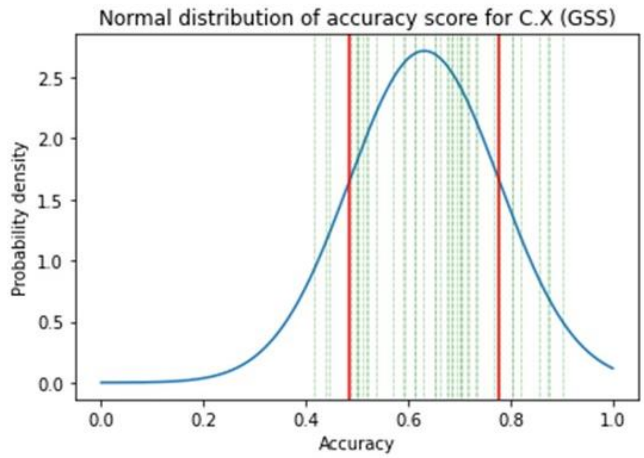


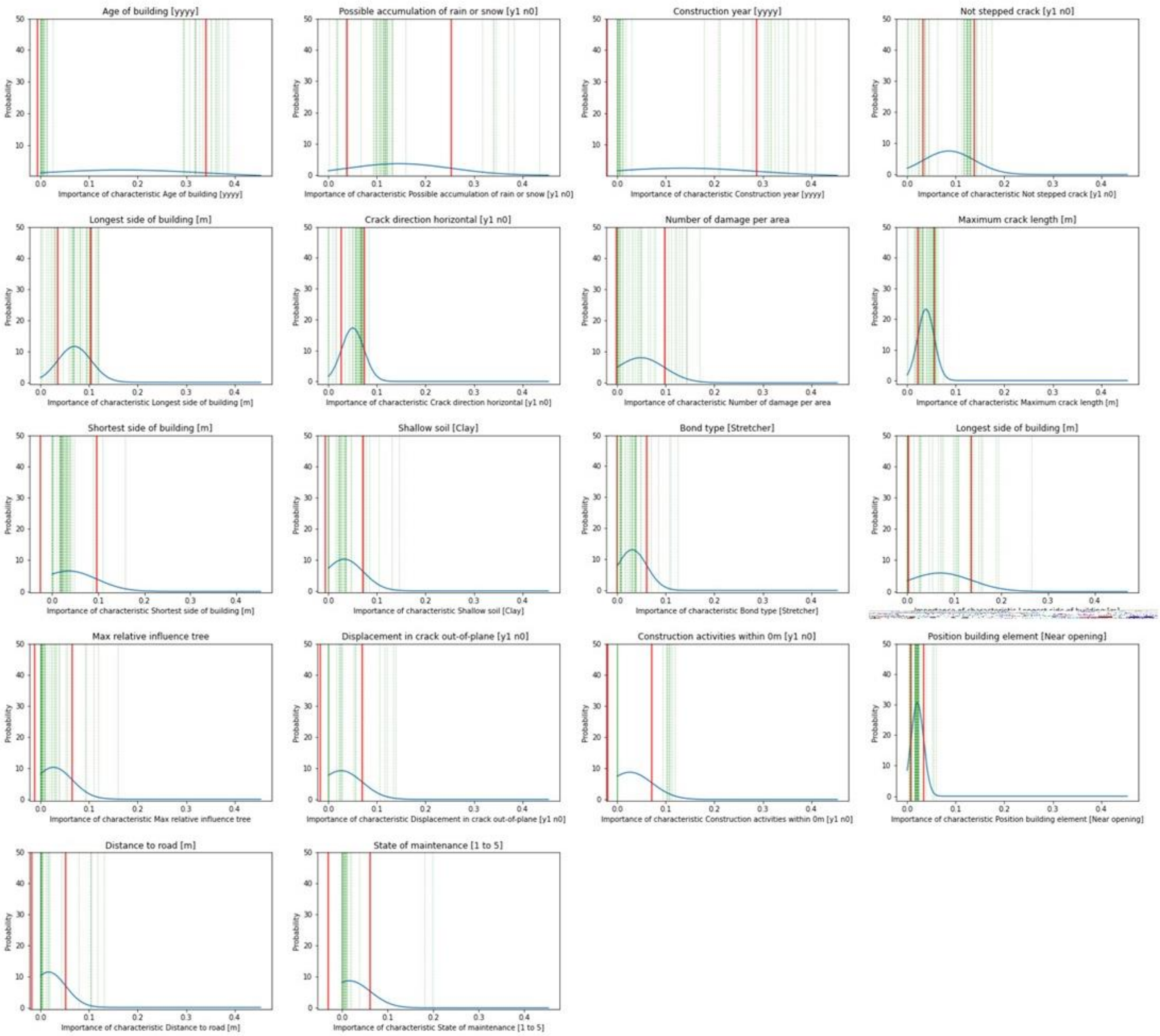
Appendix 5: Results classification Runs – Decision tree



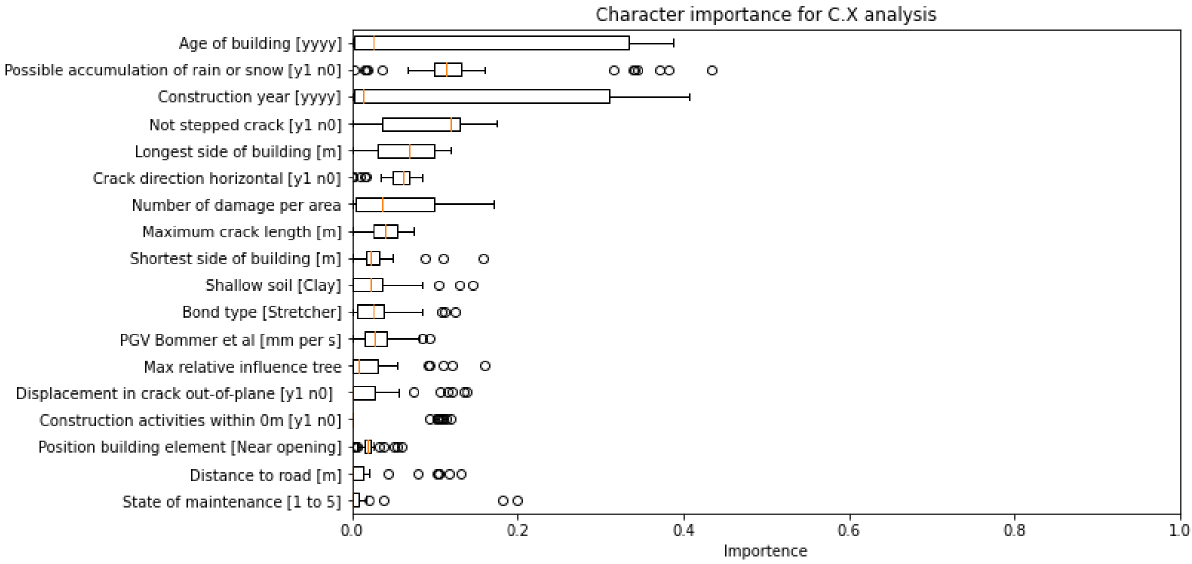
Pattern C.X - Unequal settlements

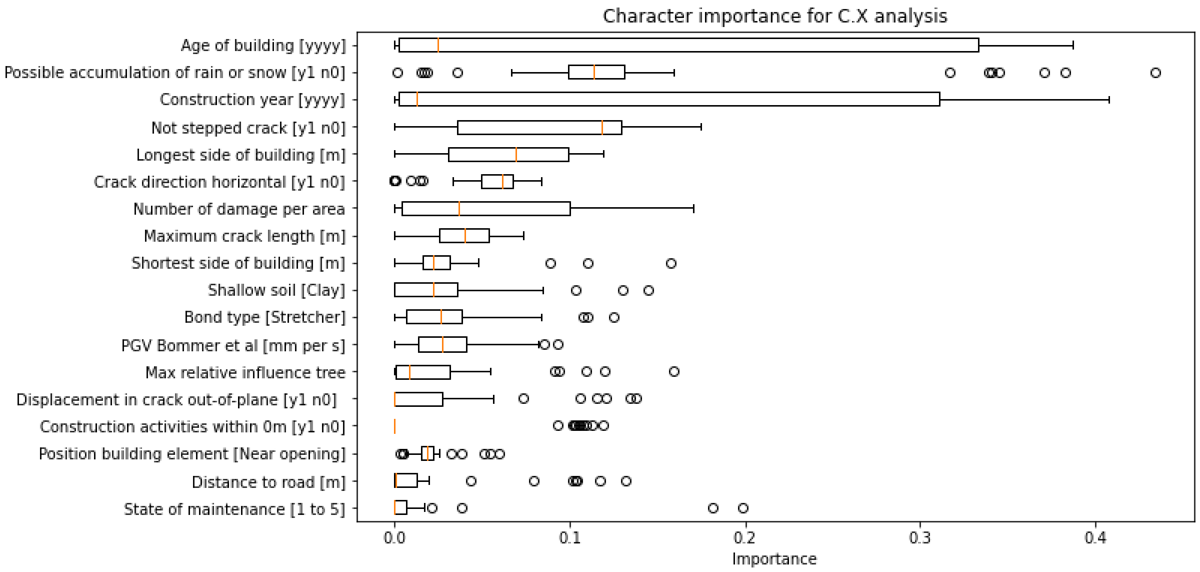




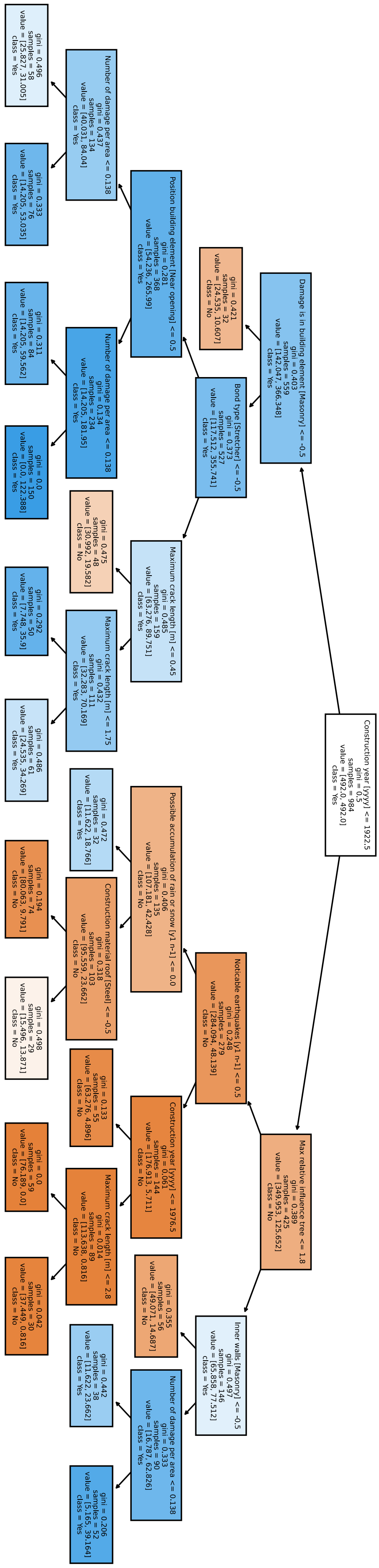


Appendix 5: Results classification Runs – Decision tree





Appendix 5: Results classification Runs – Decision tree



Appendix 6: Results Regression Runs – Decision tree

These are the results of a regression run with decision tree. The setting for most optimal results are shown in the table below.

Pattern	GKF settings				Stopping rules		End result coefficient of determination
	Removed buildings (outliers)	Folds	Available buildings for analysis	Number of buildings in one test set	Max depth	Min samples leaf	
A.1.1	45 25	3	22	6-8	2	11	-0.16
A.3.5	30	8	46	4-7	2	42	0.06
B.1.1	1 24 48 22	5	39	7-8	4	25	-0.32
B.2.1	4 28	3	29	9-11	2	14	-0.44
C.1.1	44	3	8	1-4	4	3	-0.04
C.X	44 46 48	9	42	1-6	5	29	-0.17

Table 16 Optimal settings for regression run with decision tree algorithm. 1850 damage cases are included in the analysis. GKF cross-validation is applied for coefficient of determination calculations.

The first column in Table 16 indicates each analysed pattern. As explained in chapter 5, each damage cause has its own pattern (see Figure 19). The next four columns displays the best GKF cross-validation settings in order to construct a sufficient large test set for accuracy calculations. For example, pattern A.1.1 has 3 folds (column folds). This will lead to 6 to 8 buildings in one test set. The next two columns describes the stopping rules settings for decision tree (explained in chapter 3). The last column shows the end result of each pattern.

Pattern A.1.1 – Initial insufficient resistance to bear loads

The first step of producing a regression model is to remove outliers from the data. This process is illustrated in the following figures.

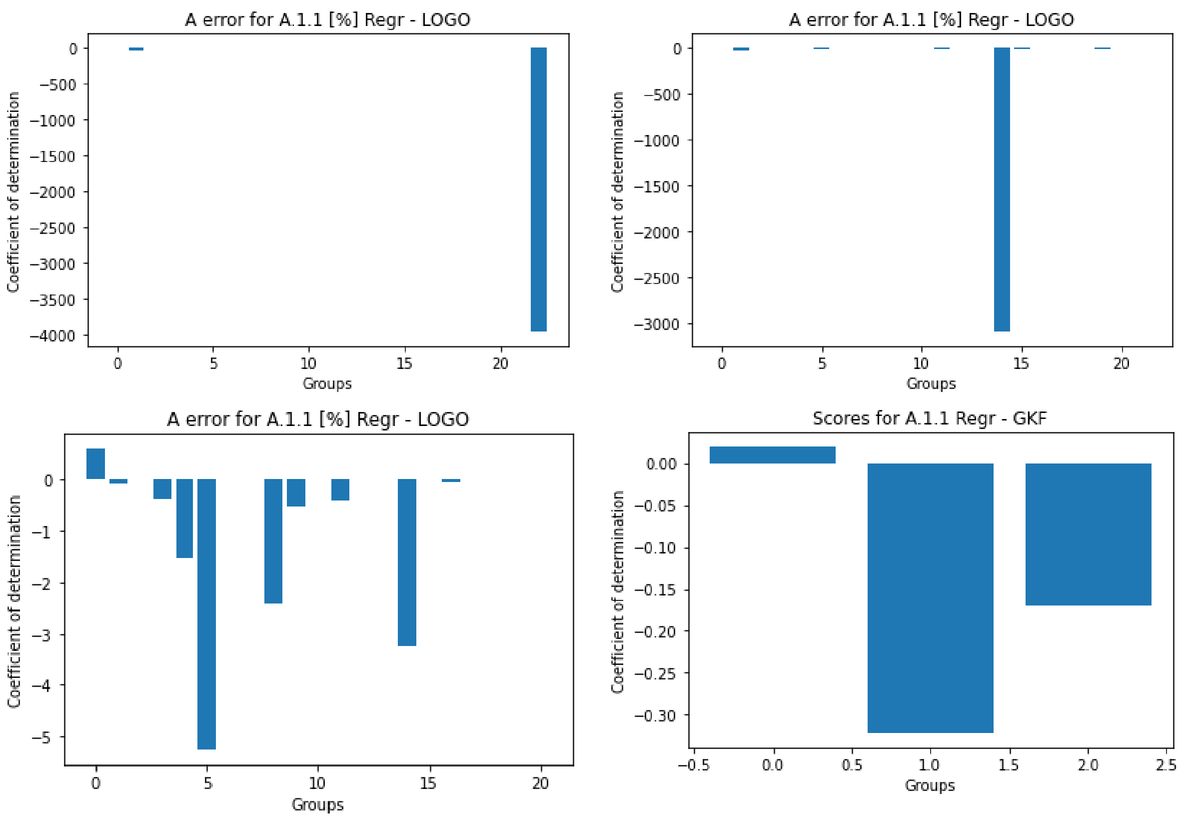


Figure 153 Coefficient of determination scores for decision tree regression to A.1.1

The initial start is at the upper left graph of Figure 153. Building 45 is an outlier in this graph. That outlier building is removed at stage 2 (upper right of Figure 153). Building 25 is an outlier in this graph. That outlier building is removed at stage 3 (lower left of Figure 153). The graph with the end scores is presented at the bottom right. The choice for three folds in the cross-validation is explained with the figure below.

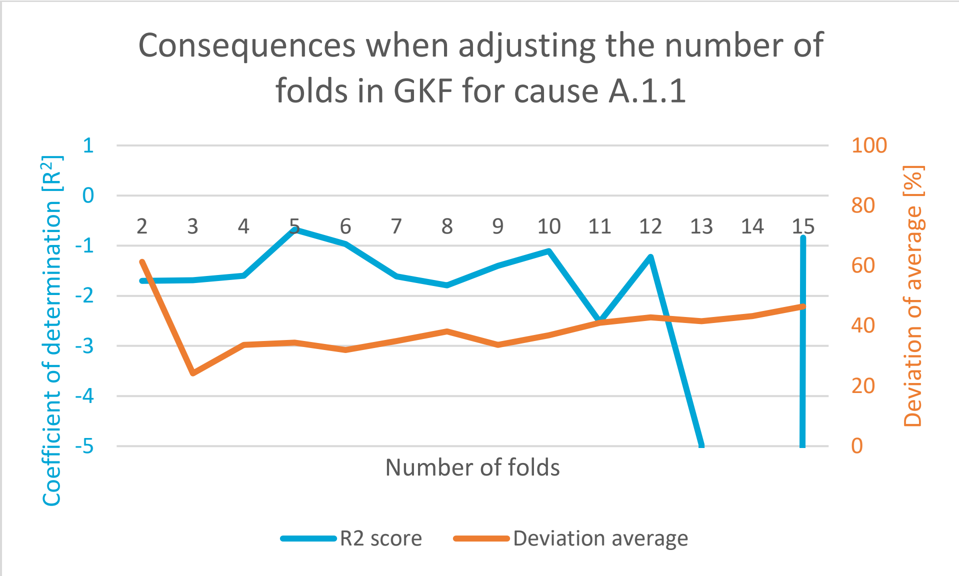


Figure 154 Results for regression on cause A.1.1 with decision tree method. GKF cross-validation is applied with a range of all possible fold settings.

The cross-validation is executed with 3 folds, after consulting Figure 154. It has the best deviation of average score.

	group1	*group2*	*group3*	*mean*	*std*
Number of damage per area	0,20	0	0,78	0,33	0,33
Construction year [yyyy]	0	0	0	0	0
Position building element [Above opening]	0,35	0	0	0,12	0,16
Displacement in crack in-plane [y1 n0]	0,45	0,83	0	0,43	0,34
Crack direction horizontal [y1 n0]	0	0	0,070	0,023	0,033
Maximum crack length [m]	0	0	0	0	0
Classification of damage [L0 M05 Z1]	0	0,17	0,15	0,11	0,078
Coefficient of determination [R^2]	0,020	-0,32	-0,17	-0,16	0,14
Deviation average [%]	32	5,5	34	24	13,2

Table 17 Output of importance of each characteristic, including the score and deviation average of each cross-validation.

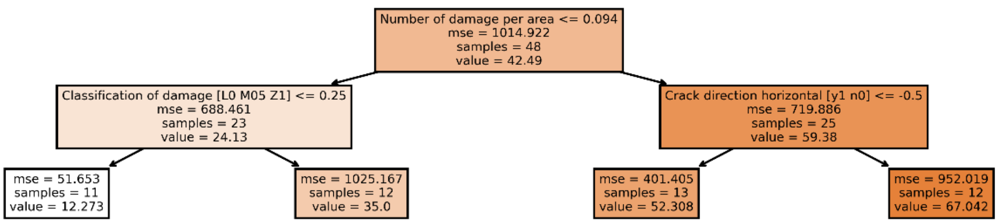


Figure 155 Regression decision tree for damage cause A.1.1.

Pattern A.3.5 - Overloading due to vibrations by earthquakes

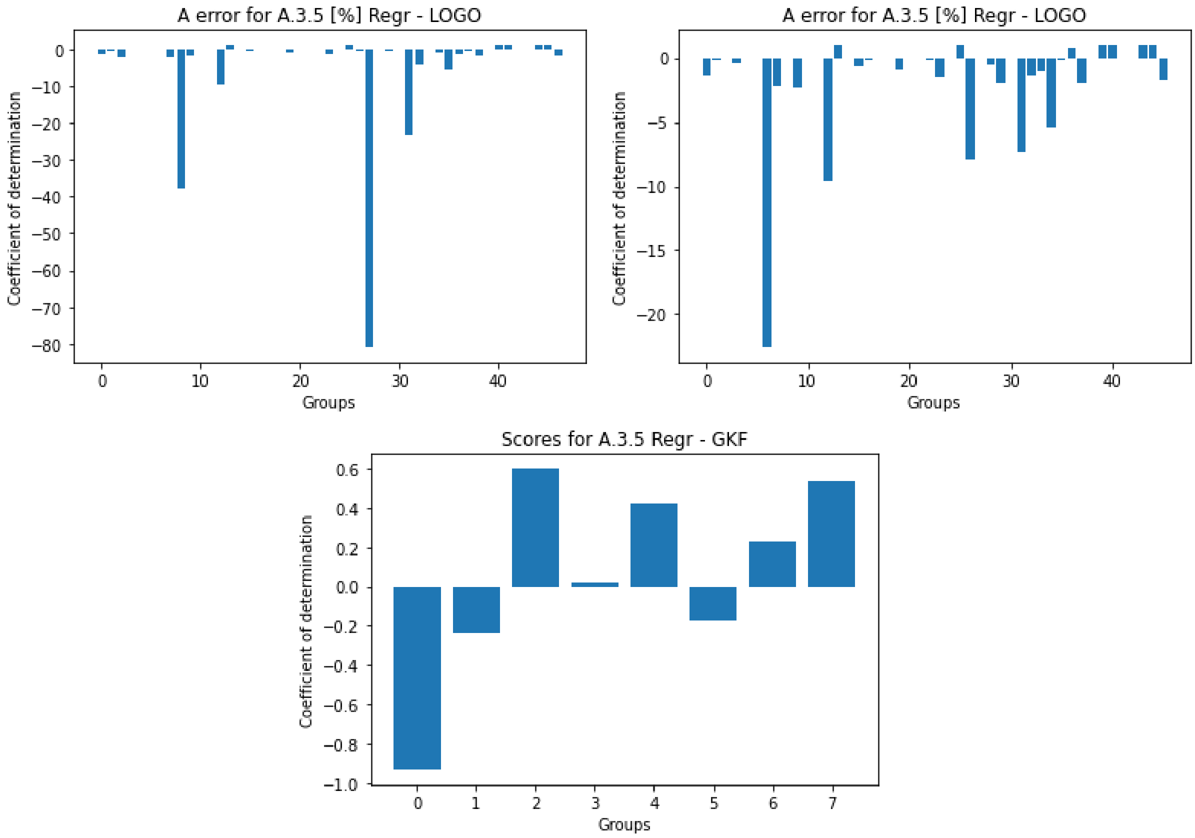


Figure 156 Coefficient of determination scores for decision tree regression to A.3.5

The initial start is at the upper left graph of Figure 156. Building 30 is an outlier in this graph. That outlier building is removed at the graph upper right of Figure 156. The graph with the end scores is presented at the bottom of Figure 156. The choice for three folds in the cross-validation is explained with the figure below.

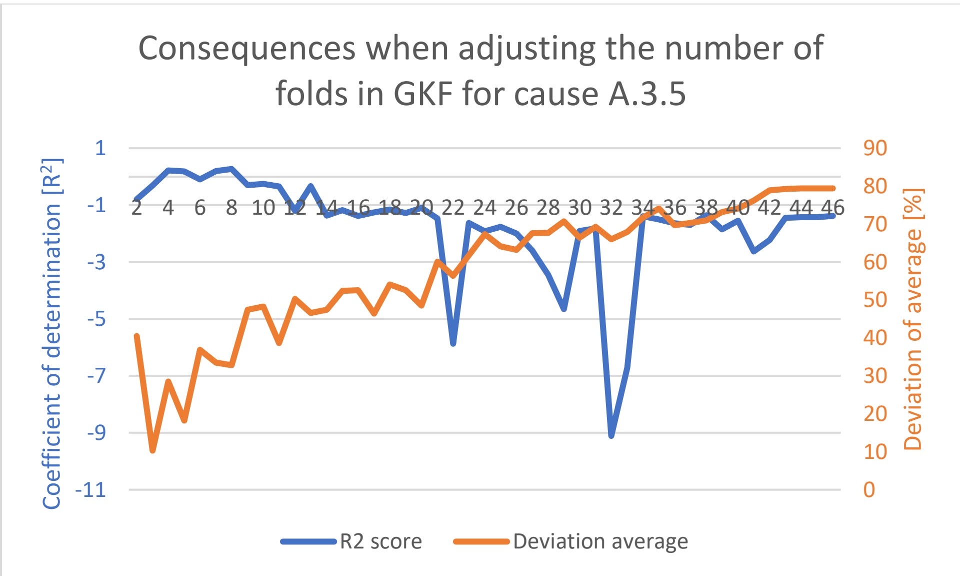


Figure 157 Results for regression on cause 'vibration due to earthquakes' (A.3.5) with decision tree method. GKF cross-validation is applied with a range of all possible fold settings.

The cross-validation is executed with 8 folds, after consulting Figure 157. It has a stable deviation of average score, while the R^2 score is positive and it is has the most available folds.

	group 1	*group 2*	*group 3*	*group 4*	*group 5*	*group 6*	*group 7*	*group 8*	*mean*	*std*
Age of building [yyyy]	0,0062	0,0049	0,0064	0,1228	0,0053	0,0067	0	0,0012	0,019	0,039
Construction year [yyyy]	0,15	0	0	0	0	0	0	0	0,02	0,051
Distance to road [m]	0	0	0	0	0	0	0,059	0	0,0074	0,02
PGV Bommer et al [mm per s]	0,84	1	0,99	0,88	0,99	0,99	0,94	1	0,95	0,059
Position of damage [Outside]	0	0	0	0	0	0	0	0	0	0
Crack in joint and unit [y1 n0]	0	0	0	0	0	0	0	0	0	0
Branched cracks [y1 n0]	0	0	0	0	0	0	0	0	0	0
scores	-0,93	-0,24	0,6	0,02	0,42	-0,17	0,23	0,54	0,06	0,47
Afwijking gemiddelde	48	39	51	41	2	24	13	44	33	17

Table 18 Output of importance of each characteristic, including the score and deviation average of each cross-validation

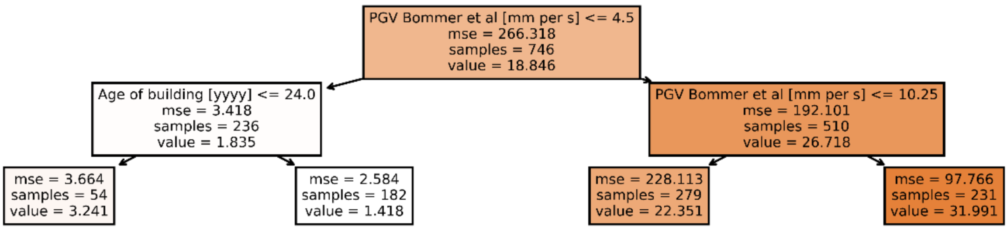


Figure 158 Regression decision tree for damage cause A.3.5

Pattern B.1.1 - Initial hindered deformations

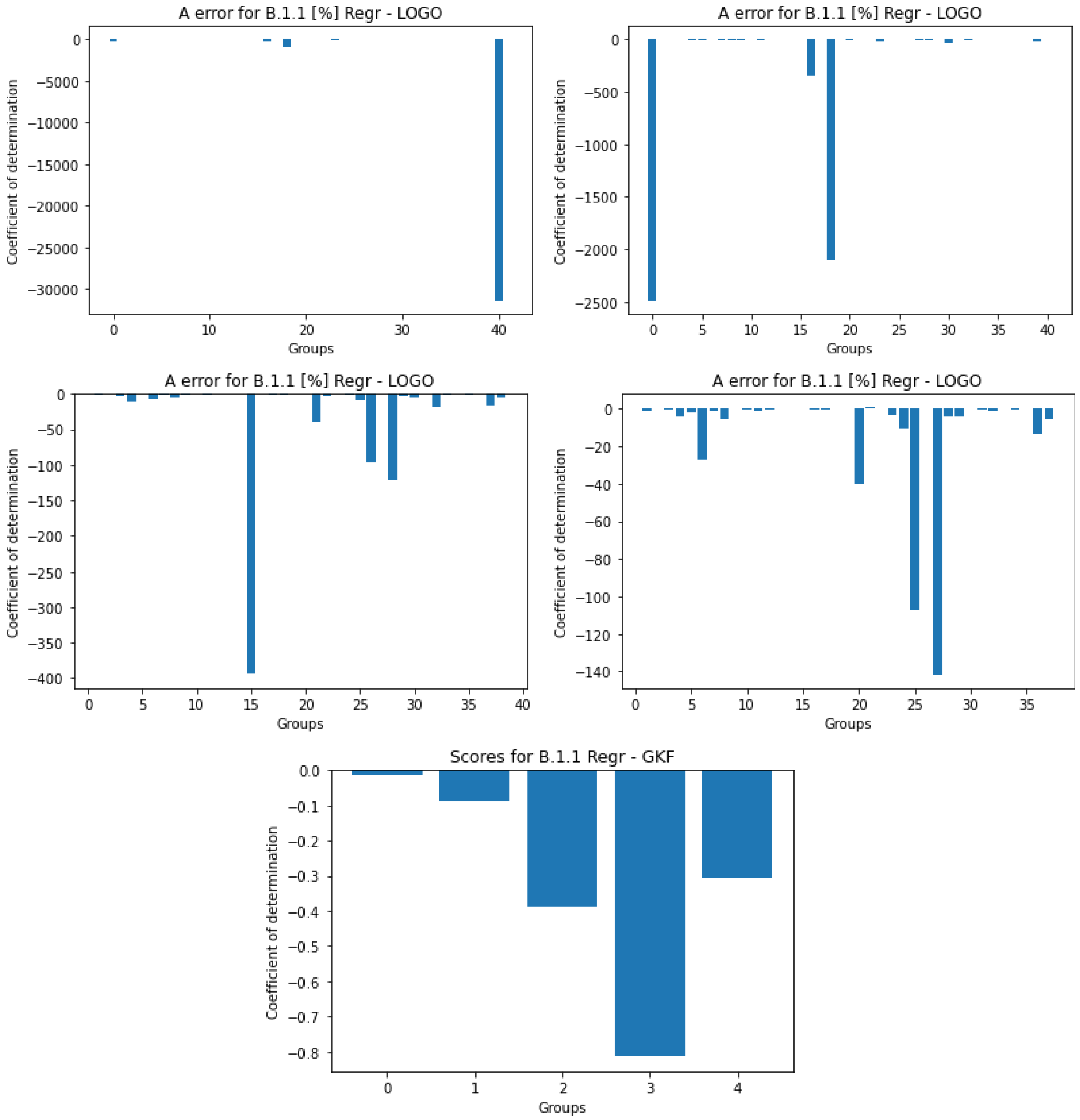


Figure 159 Coefficient of determination scores for decision tree regression to B.1.1

Buildings 1, 24, 48 and 22 were donated as outliers according to the upper graphs of Figure 159.

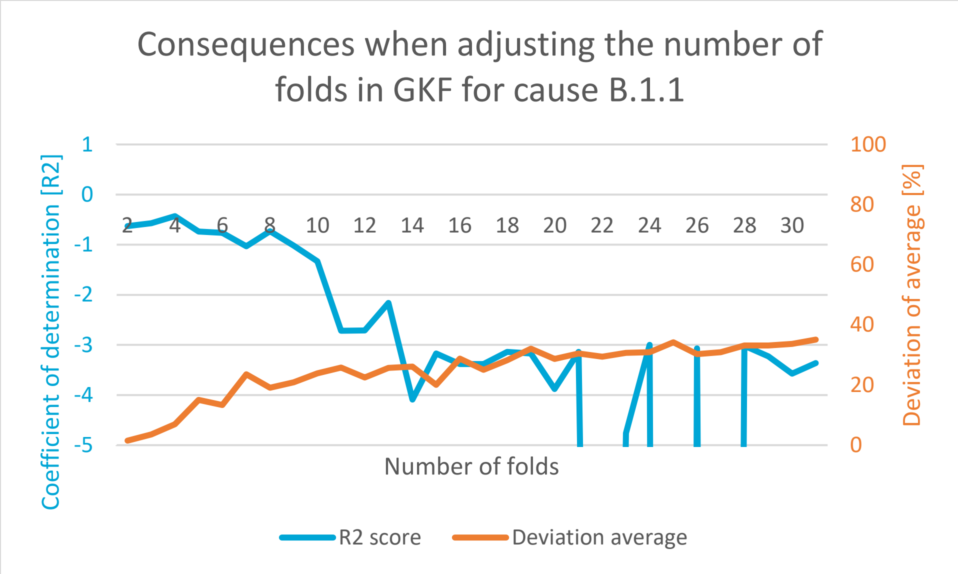


Figure 160 Results for regression on cause B.1.1 with decision tree method. GKF cross-validation is applied with a range of all possible fold settings.

The cross-validation is executed with 5 folds, after consulting Figure 160.

	group1	*group2*	*group3*	*group4*	*group5*	*mean*	*std*
Number of damage per area	0	0	0	0	0	0	0
Construction year [yyyy]	0,014	0,539	0,043	0,487	0,123	0,241	0,225
Area of building [m2]	0,151	0	0	0	0	0,03	0,061
Longest side of building [m]	0,05	0,154	0,422	0,032	0,265	0,185	0,145
Indications of leaks [y1 n0]	0	0	0	0	0	0	0
Height of tree [m]	0,43	0,24	0,36	0,15	0	0,23	0,15
Relative influence tree	0	0	0,076	0	0	0,015	0,03
Max relative influence tree	0	0	0	0,214	0	0,043	0,086
Position of damage [Inside]	0	0	0	0	0	0	0
Direction of facade [East]	0	0	0	0	0	0	0
Crack in joint and unit [y1 n0]	0	0	0	0	0	0	0
Crack direction vertical [y1 n0]	0,16	0	0,1	0,12	0,09	0,09	0,05
Runs from-to ceiling [y1 n0]	0	0	0	0	0	0	0
Maximum crack width [mm]	0	0	0	0	0	0	0
Maximum crack length [m]	0	0,066	0	0	0,037	0,021	0,027
Not stepped crack [y1 n0]	0,2	0	0	0	0,48	0,14	0,19
Classification of damage [L0 M05 Z1]	0	0	0	0	0	0	0
scores	-0,015	-0,089	-0,388	-0,814	-0,305	-0,322	0,281
Deviation average	29	8	7	7	24	15	9

Figure 161 Output of importance of each characteristic, including the score and deviation average of each cross-validation.

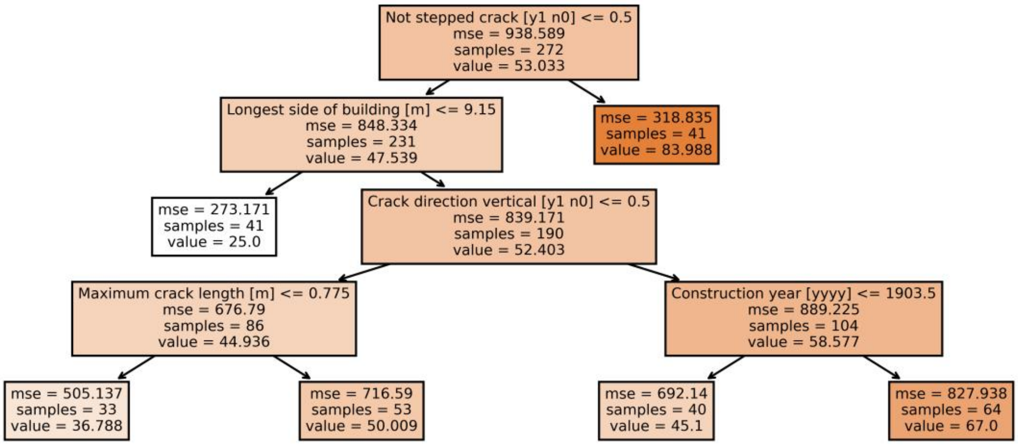


Figure 162 Regression decision tree for damage cause B.1.1.

Pattern B.2.1 - Initial imposed deformations

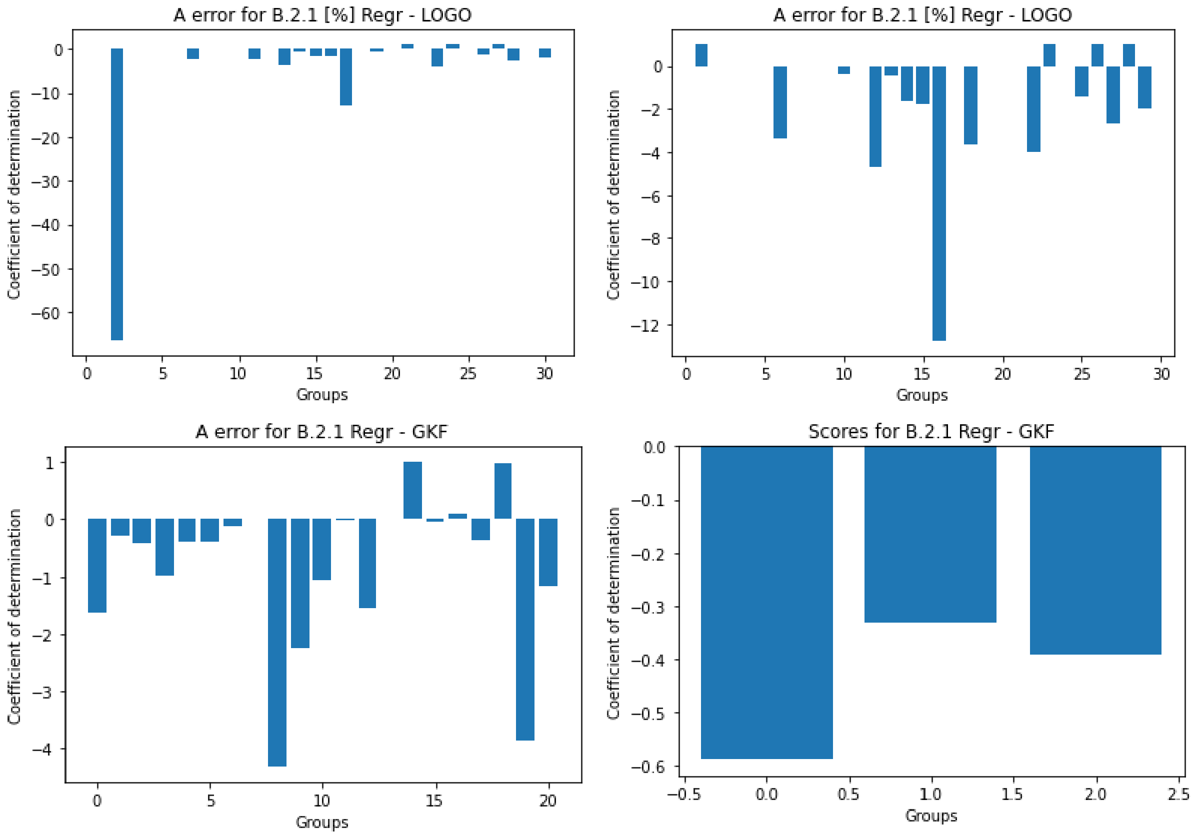


Figure 163 Coefficient of determination scores for decision tree regression to B.2.1

The initial start is at the upper left graph of Figure 163. Building 4 is an outlier in this graph. That outlier building is removed at stage 2 (upper right of Figure 163). Building 28 is an outlier in this graph. That outlier building is removed at stage 3 (lower left of Figure 163). The graph with the end scores is presented at the bottom right. The choice for three folds in the cross-validation is explained with the figure below.

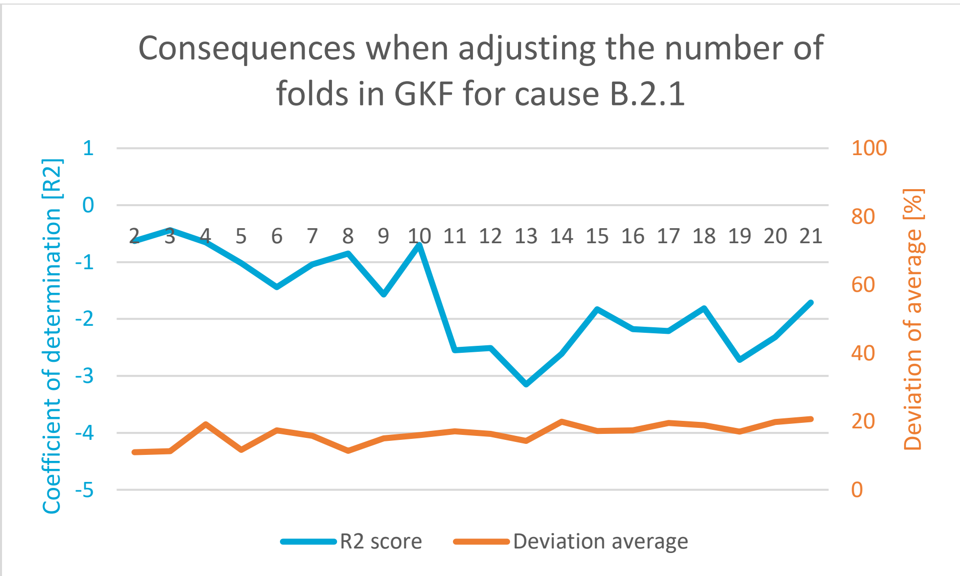


Figure 164 Results for regression on cause B.2.1 with decision tree method. GKF cross-validation is applied with a range of all possible fold settings.

The cross-validation is executed with 3 folds, after consulting Figure 164. It has the best deviation of average score with the highest initial coefficient of determination.

	group1	*group2*	*group3*	*mean*	*std*
Longest side of building [m]	0	0,77	0,79	0,52	0,37
Bond type [Stretcher]	0,38	0	0	0,13	0,18
Distance to road [m]	0,62	0,23	0,21	0,35	0,19
Position building element [Near opening]	0	0	0	0	0
Coefficient of determination [R^2]	-0,59	-0,33	-0,39	-0,44	0,11
Deviation average [%]	19	15	6	13	5

Table 19 Output of importance of each characteristic, including the score and deviation average of each cross-validation.

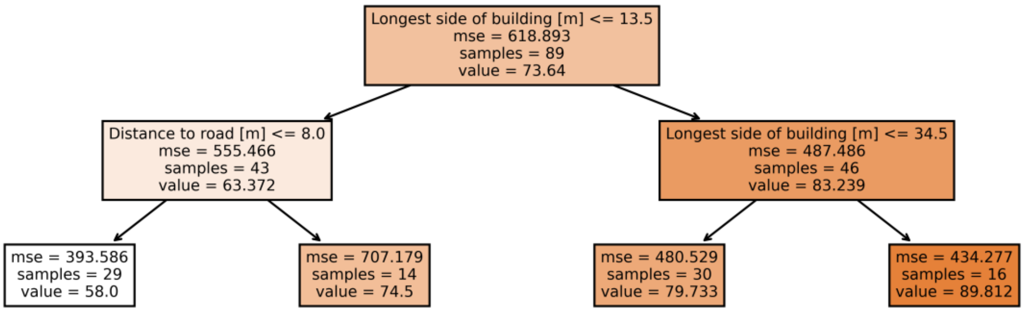


Figure 165 Regression decision tree for damage cause B.2.1.

Pattern C.1.1 – Unequal settlements with equal loads

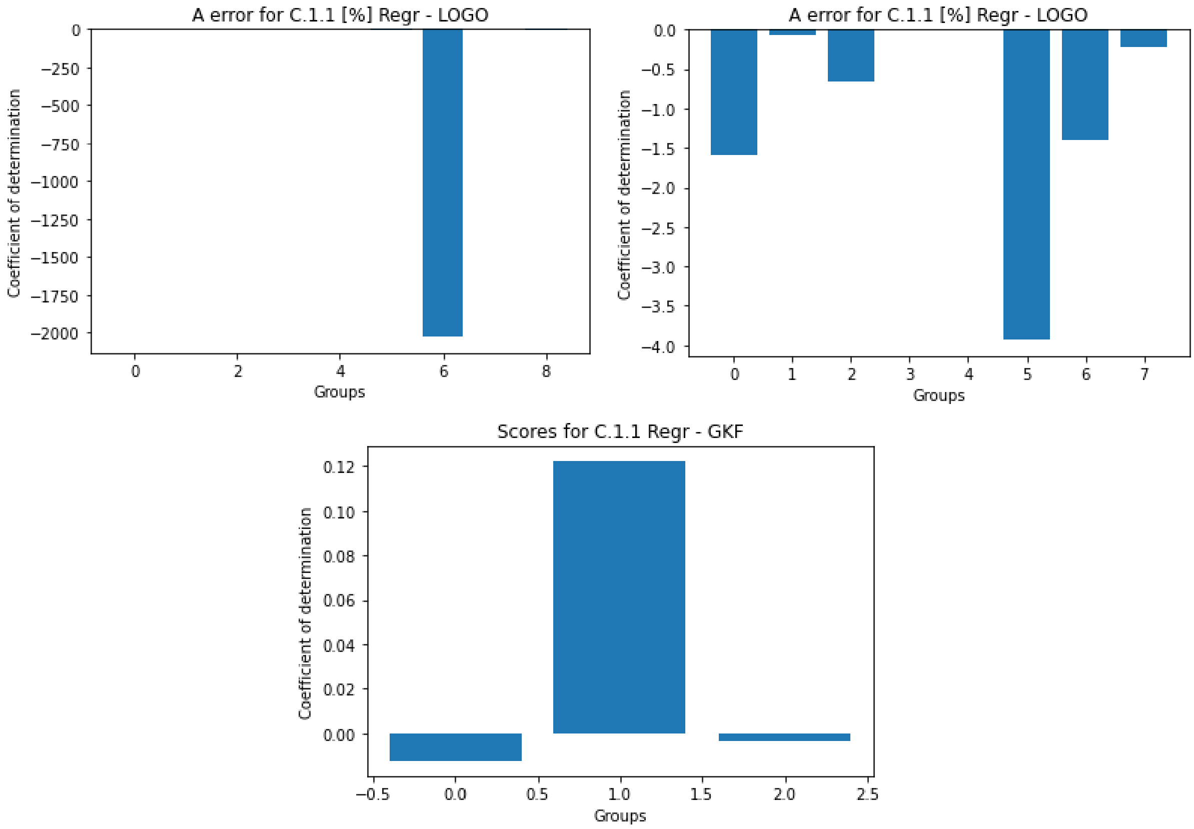


Figure 166 Coefficient of determination scores for decision tree regression to C.1.1

The initial start is at the upper left graph of Figure 166. Building 44 is an outlier in this graph. That outlier building is removed at stage 2 (upper right of Figure 166). The graph with the end scores is presented at the bottom. The choice for three folds in the cross-validation is explained with the figure below.

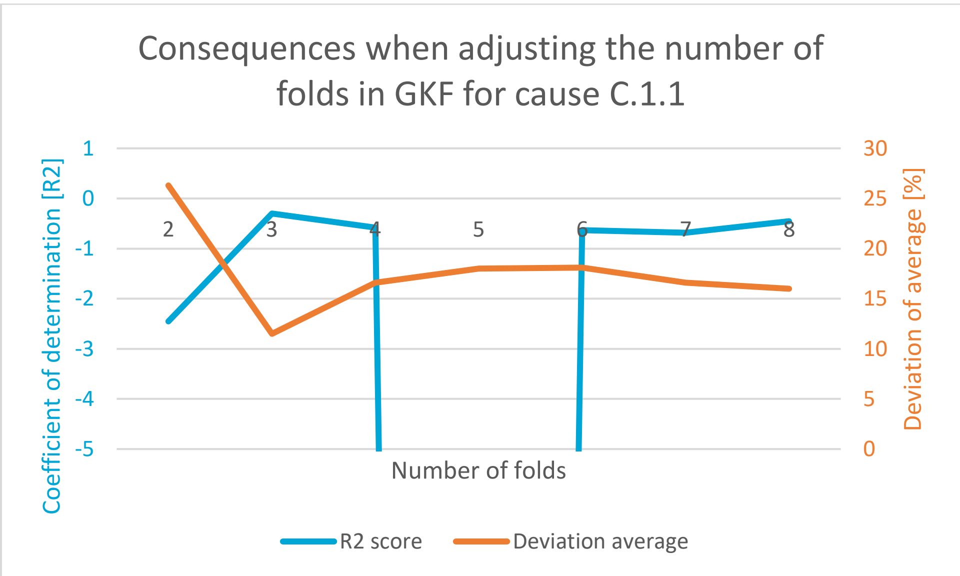


Figure 167 Results for regression on cause C.1.1 with decision tree method. GKF cross-validation is applied with a range of all possible fold settings.

The cross-validation is executed with 3 folds, after consulting Figure 167. It has the best deviation of average score. More folds seems unreasonable since only 8 buildings has damage within this category.

	group1	*group2*	*group3*	*mean*	*std*
Number of damage per area	0	0	0	0	0
Age of building [yyyy]	0	0	0	0	0
Construction year [yyyy]	0,713	0	0,004	0,239	0,335
Area of building [m2]	0	0,83	0,94	0,59	0,42
Shortest side of building [m]	0	0	0	0	0
Foundation material [Masonry]	0,002	0	0,055	0,019	0,026
Change in Ground water level [y1 n0]	0	0	0	0	0
Distance to road [m]	0	0	0	0	0
Noticable earthquakes [y1 n0]	0,28	0,17	0	0,15	0,12
PGV Bommer et al [mm per s]	0	0	0	0	0
In finish or through construction [through1 in finish0]	0,0033	0,0004	0,0053	0,003	0,002
Coefficient of determination [R^2]	-0,012	-0,286	0,193	-0,035	0,196
Deviation average	17	6	12	11	5

Table 20 Output of importance of each characteristic, including the score and deviation average of each cross-validation.

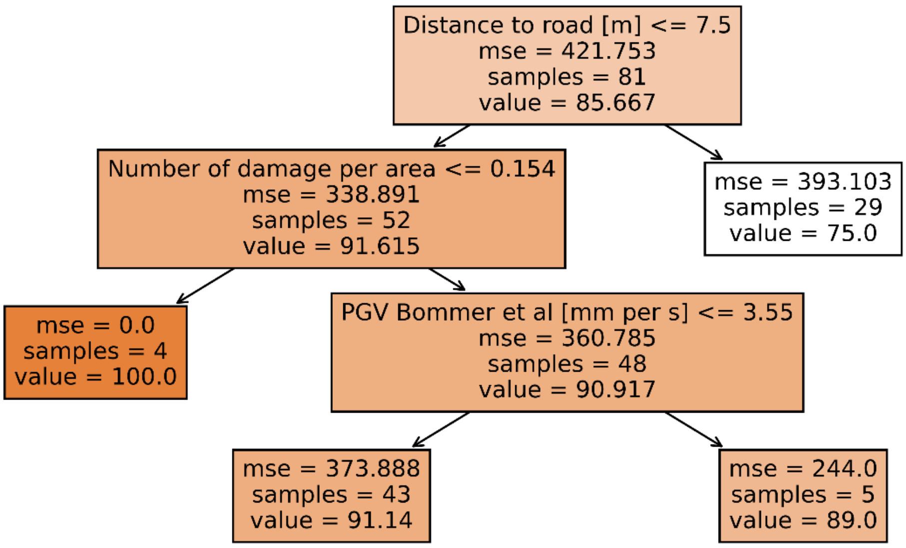


Figure 168 Regression decision tree for damage cause C.1.1.

Pattern C.X - Unequal settlements

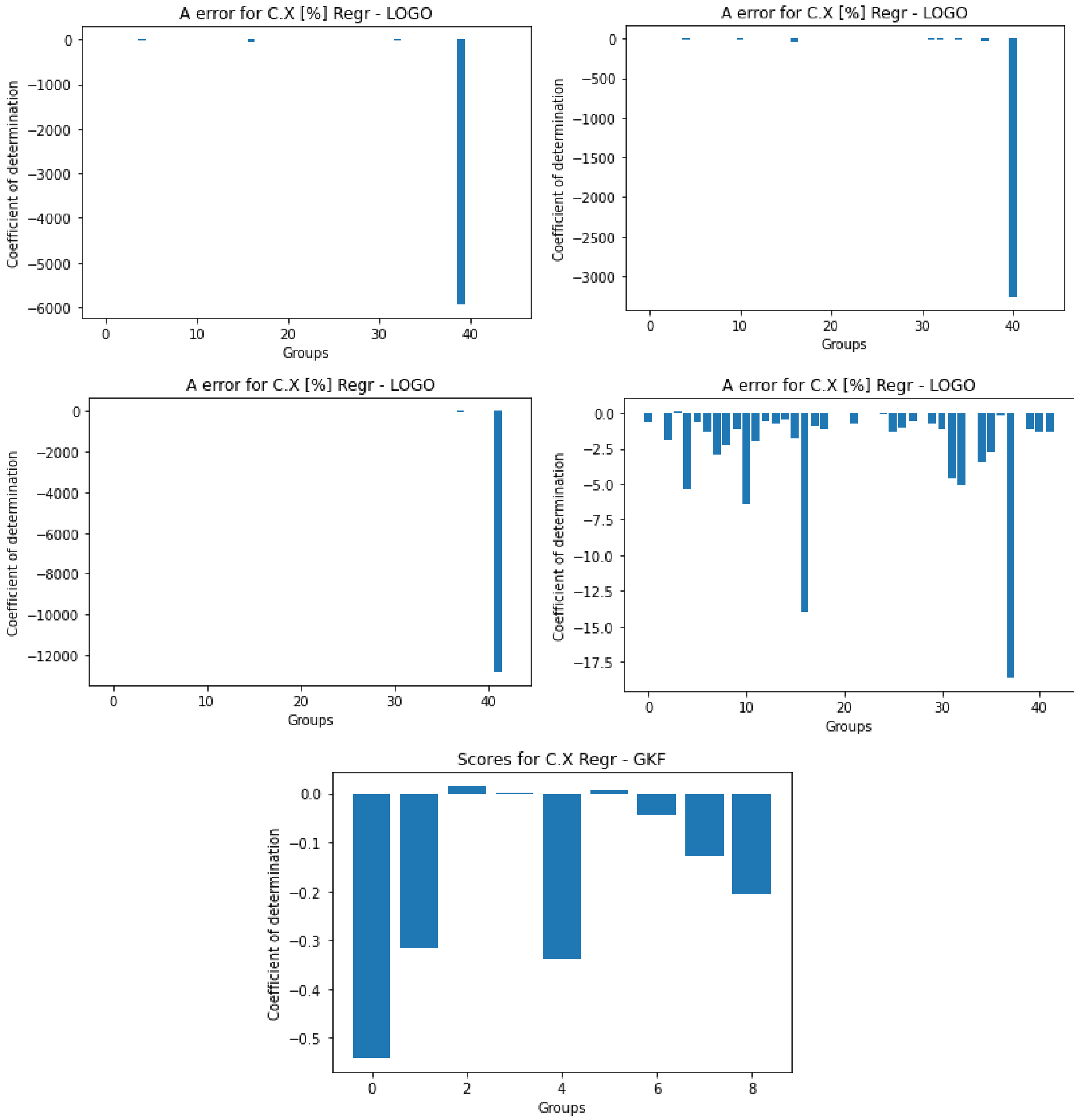


Figure 169 Coefficient of determination scores for decision tree regression to C.X

The initial start is at the upper left graph of Figure 169. Building 44 is an outlier in this graph. That outlier building is removed at stage 2 (upper right of Figure 169). Building 46 is an outlier in this graph. That outlier building is removed at stage 3 (lower left of Figure 153). The last outlier is visualised at the middle left graph. This outlier is building 48. The graph with the end scores is presented at the bottom. The choice for eight folds in the cross-validation is explained with the figure below.

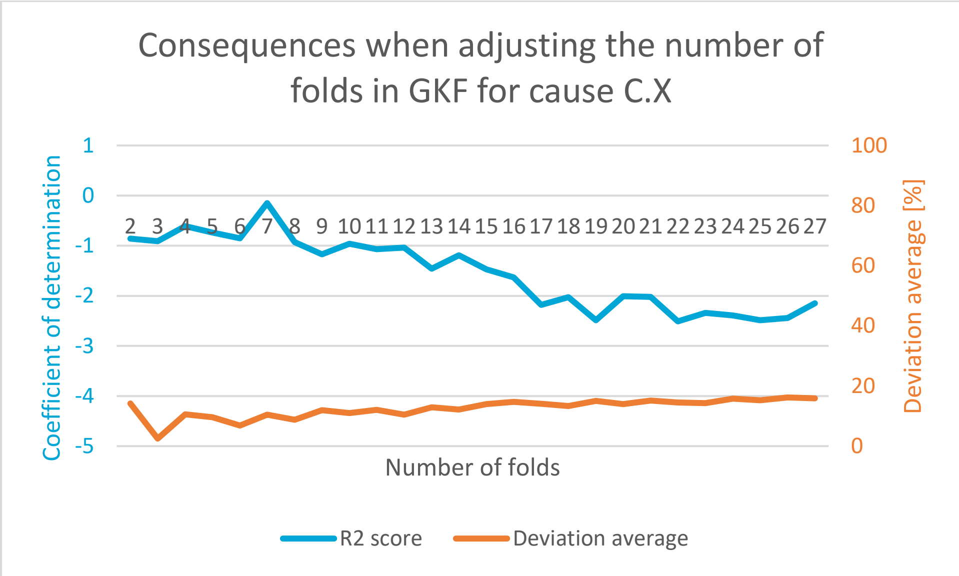


Figure 170 Results for regression on cause C.X with decision tree method. GKF cross-validation is applied with a range of all possible fold settings.

The cross-validation is executed with 9 folds, after consulting Figure 170. Deviation average is relatively constant. Consequently, the choice for the number of folds is not influenced by the deviation of average. So, the number of folds is more a arbitrary choice.

	group1	*group2*	*group3*	*group4*	*group5*	*group6*	*group7*	*group8*	*group9*	*mean*	*std*
Number of damage per area	0,041	0,099	0	0,037	0,046	0,051	0,005	0,007	0,079	0,041	0,032
Age of building [yyyy]	0	0	0,044	0	0	0,018	0	0	0,024	0,01	0,015
Construction year [yyyy]	0,054	0	0	0	0,009	0	0,016	0	0	0,009	0,017
Longest side of building [m]	0,032	0,068	0,061	0	0	0,038	0	0,02	0,128	0,039	0,04
Shortest side of building [m]	0	0	0	0	0,0054	0	0,0454	0	0	0,0056	0,0141
Bond type [Stretcher]	0	0	0	0,041	0,046	0,012	0	0	0	0,011	0,018
State of maintenance [1 to 5]	0	0	0	0,046	0	0,043	0	0	0	0,01	0,018
Shallow soil [Clay]	0	0	0	0	0	0	0	0	0	0	0
Max relative influence tree	0,18	0,25	0,2	0,31	0,24	0,03	0,29	0,25	0,18	0,21	0,08

Distance to road [m]	0,008	0	0,036	0	0	0,009	0,036	0,027	0	0,013	0,015
Construction activities within 0m [y1 n0]	0	0	0	0	0,0102	0	0	0	0	0,0011	0,0032
PGV Bommer et al [mm per s]	0,58	0,49	0,56	0,42	0,54	0,64	0,48	0,56	0,53	0,53	0,06
Position building element [Near opening]	0	0	0	0	0,0064	0,0074	0	0	0	0,0015	0,0029
Displacement in crack out-of-plane [y1 n0]	0	0	0	0	0	0	0,01	0	0	0,001	0,003
Crack direction horizontal [y1 n0]	0	0	0	0	0,031	0,028	0	0	0	0,007	0,012
Maximum crack length [m]	0,091	0,082	0,082	0,132	0,063	0,124	0,125	0,117	0,048	0,096	0,028
Not stepped crack [y1 n0]	0,013	0,014	0,02	0,018	0	0	0	0,011	0,008	0,009	0,007
Coefficient of determination [R^2]	-0,54	-0,32	0,02	0	-0,34	0,01	-0,04	-0,13	-0,21	-0,17	0,18
Deviation average	2	10	20	12	5	17	18	18	5	12	6

Figure 171 Output of importance of each characteristic, including the score and deviation average of each cross-validation.

Appendix 7: Linear regression

This appendix explains the settings applied to obtain the results of linear regression presented in chapter 6. This explanation is divided in three parts: the results of overloading due to vibrations of earthquakes (A.3.5), deformations (B.X) and unequal settlements (C.X).

Pattern A.3.5 – Overloading due to vibrations by earthquakes

Before a pattern is obtained, outliers are excluded from the analyses. The determination of those outliers is explained with the next graphs.

Outliers in linear regression pattern A.3.5

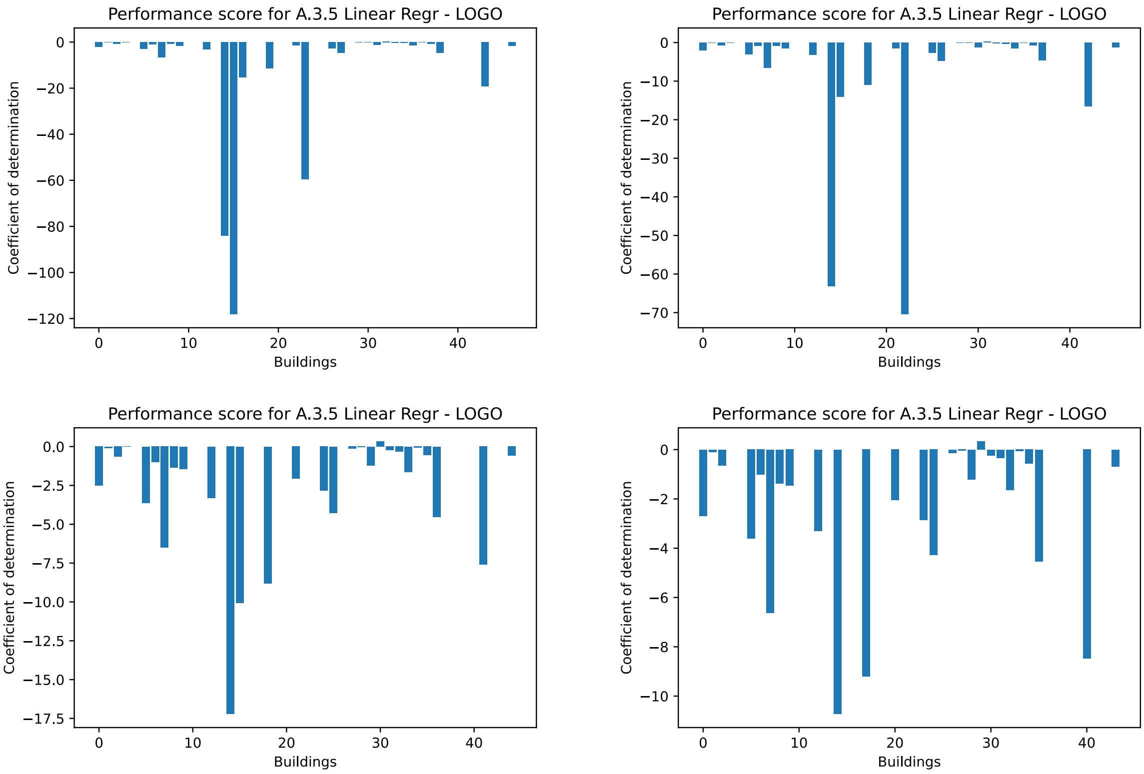


Figure 172 Localising the outliers for regression pattern A.3.5 (Overloading due to vibrations by earthquakes). Upper left, original output; upper right, building 18 removed; lower left, building 26 removed; lower right, building 16 removed.

According to Figure 172, the following buildings are indicated as an outlier: 18, 26 and 16. The next figure shows the consequences on the technical attributability. The blue parts shows where the removed buildings were located. It could be concluded from Figure 173 that the removed outliers had a low technical attributability.

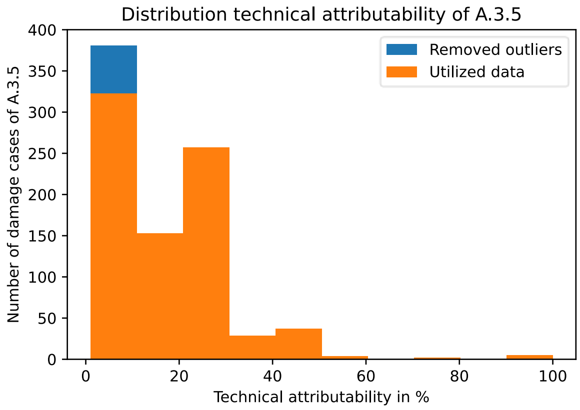


Figure 173 Consequences on the technical attributability if outliers are removed.

The following characteristics were applied: Age of building, (max) relative influence of trees, PGV according to Bommer et al, damage located at inside locations and the appearance fo brached cracks.

The applied number of folds in group-K-fold cross-validation is 5 (Figure 174). This resulted in a coefficient of determination of 0.48 (obtained as average of cross-validation). The mean of deviation of average between the train and validation set is 23.3 with a standard deviation of 10.1.

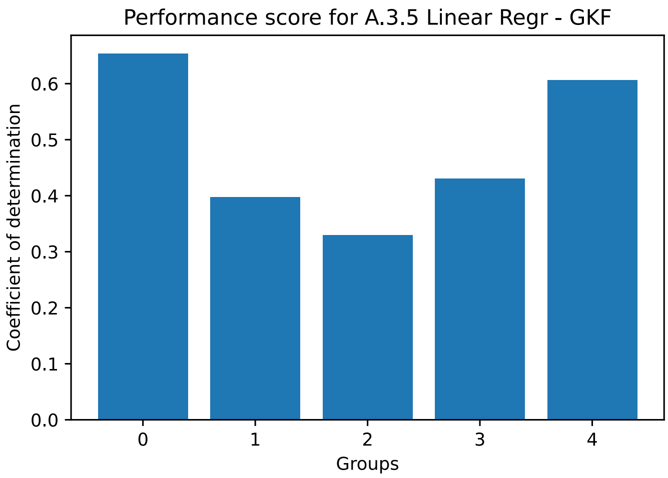


Figure 174 End results of linear regression pattern A.3.5 (overloading due to vibrations by earthquakes) in terms of performance scores of GKF cross-validation.

Linear regression pattern for overloading due to vibrations by earthquakes (A.3.5)

	Group 1	Group 2	Group 3	Group 4	Group 5	Mean	std
Age of building [yyyy]	-6.81	-5.62	-6.04	-12.09	-5.46	-7.2	2.49
Max relative influence tree	13.94	14.22	11.67	14.82	16.83	14.29	1.66
PGV Bommer et al [mm per s]	55.18	52.64	52.61	50.87	53.05	52.87	1.38
Position of damage [Inside]	-2.46	-2.87	-2.01	-1.47	-3.18	-2.4	0.61
Branched cracks [y1 n0]	3.1	2	2.26	2.53	1.61	2.3	0.5
y-intercept constant	-13.07	-11.11	-11.91	-11.4	-11.7	-11.84	0.68
Performance score (R2)	0.65	0.4	0.33	0.43	0.61	0.48	0.12
Difference train and val. set	23.77	13	37.6	11.48	30.85	23.34	10.08

Table 21 Variables in linear regression model for A.3.5 (overloading due to vibrations by earthquakes)

The table above presents the linear regression pattern for overloading due to vibrations by earthquakes. The upper half of the table shows the slope of each characteristics (rows) at each cross-validation fold (columns). The lower half of the table presents some model values, like the y-intercept of a linear regression and the performance score in terms of coefficient of determination.

Pattern B.X – Deformations

Before a pattern is obtained, outliers are excluded from the analyses. The following buildings are indicated as an outlier: 48, 24, 23, 44, 26, 19, 25, 33 and 30. The next figure shows the consequences on the technical attributability. The blue parts shows where the removed buildings were located. It could be concluded from Figure 175Figure 173 that the removed outliers had a high technical attributability.

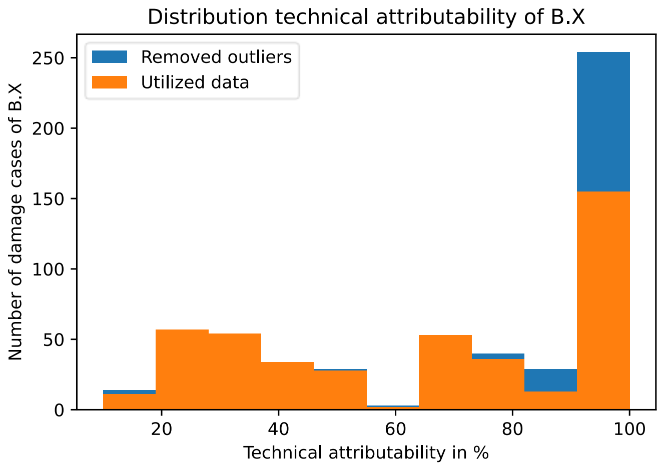


Figure 175 Consequences on the technical attributability if outliers are removed.

The following characteristics were applied: building renovations, damage is located at the inside of the building, damage in ceiling, damage in masnory material, horizontal crack direction and maximum crack width.

The applied number of folds in group-K-fold cross-validation is 2 (Figure 176). This resulted in a coefficient of determination of 0.28 (obtained as average of cross-validation). The mean of deviation of average between the train and validation set is 0.2 with a standard deviation of 0.00.

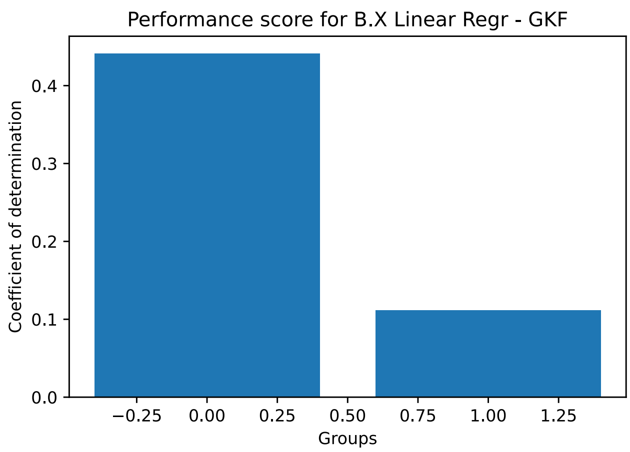


Figure 176 End results of linear regression pattern B.X (deformations) in terms of performance scores of GKF cross-validation.

Linear regression pattern for deformations (B.X)

	Group 1	Group 2	Mean	std
Building renovations [y1 n0]	-27,63	-33,78	-30,7	3,07
Position of damage [Inside]	-13,92	-12,82	-13,37	0,55
Damage is in building part [Ceiling]	5,95	-6,64	-0,34	6,29
Damage is in building element [Masonry]	-11,83	-24,4	-18,11	6,29
Crack direction horizontal [y1 n0]	4,77	11,51	8,14	3,37
Maximum crack width [mm]	-24,38	-60,16	-42,27	17,89
y-intercept constant	105,31	126,8	116,06	10,74
Performance score (R2)	0,44	0,11	0,28	0,16
Difference train and val. set	0,15	0,15	0,15	0

Table 22 Variables in linear regression model for B.X (deformations)

The table above presents the linear regression pattern for deformations. The upper half of the table shows the slope of each characteristics (rows) at each cross-validation fold (columns). The lower half of the table presents some model values, like the y-intercept of a linear regression and the performance score in terms of coefficient of determination.

Pattern C.X – Unequal settlement

Before a pattern is obtained, outliers are excluded from the analyses. The following buildings are indicated as an outlier: 3, 1, 33, 44, 48 and 46. The next figure shows the consequences on the technical attributability. The blue parts shows where the removed buildings were located. It could be concluded from Figure 177Figure 173 that the removed outliers had a high technical attributability.

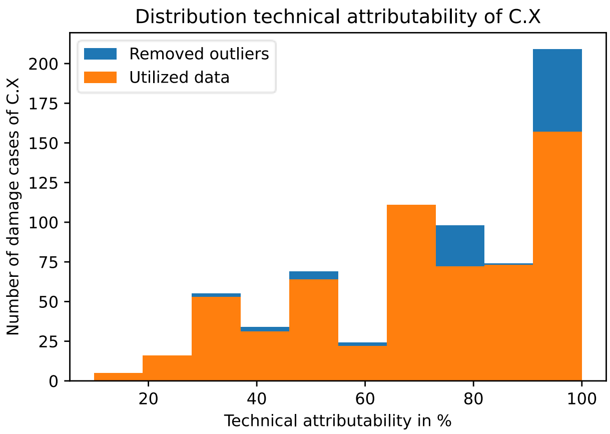


Figure 177 Consequences on the technical attributability if outliers are removed.

The following characteristics were applied: Clay in shallow soil, clay in deep soil, a change in ground water level, ground settlemnt, PGV according to Bommer et al, horizontal crack, runs from-to opening, runs from-to roof, runs from-to floor, maximum crack length, branched cracks, settlement differences over the leght of building.

The applied number of folds in group-K-fold cross-validation is 6 (Figure 178). This resulted in a coefficient of determination of 0.18 (obtained as average of cross-validation). The mean of deviation of average between the train and validation set is 7.7 with a standard deviation of 4.14.

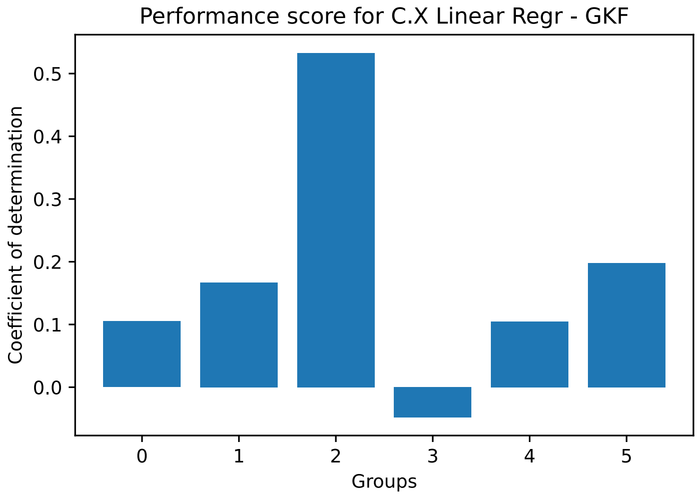


Figure 178 End results of linear regression pattern C.X (unequal settlements) in terms of performance scores of GKF cross-validation.

Linear regression pattern for unequal settlement (C.X)

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Mean	std
Shallow soil [Clay]	-2,41	-2,36	-4,56	-7,38	-0,43	-2,11	-3,21	2,22
Deep soil [Clay]	-8,54	-9,12	-8,96	-12,22	-9,17	-10,74	-9,79	1,29
Change in Ground water level [y1 n0]	-7,74	-10	-10,25	-5,26	-13,09	-8,88	-9,2	2,4
Ground settlement [not present]	4,57	3,26	2,41	5,01	0,46	4,45	3,36	1,56
PGV Bommer et al [mm per s]	-40,64	-42,79	-38,85	-40,91	-39,91	-40,59	-40,61	1,18
Crack direction horizontal [y1 n0]	-3,22	-4,14	-6,1	0,2	-4,26	-1,54	-3,18	2,03
Runs from-to opening [y1 n0]	5,08	4,88	6,11	6,9	4,49	5,32	5,46	0,81
Runs from-to roof [y1 n0]	-6,29	-8,53	-8,97	-9,78	-3,27	-8,25	-7,51	2,17
Runs from-to floor [y1 n0]	8,08	5,03	7,68	7,38	2	13,04	7,2	3,34
Maximum crack length [m]	41,12	43,96	47,92	35,29	52,8	47,26	44,72	5,54
Branched cracks [y1 n0]	-5,96	-5,79	-4,89	-7,76	-5,53	-4,92	-5,81	0,96
Settlement differences over length of building	-23,36	-22,28	-23,27	-16,34	-25,39	-23,1	-22,3	2,82
Intercept	87,09	87,51	85,07	90,94	83,72	83,21	86,25	2,62
scores	0,11	0,17	0,53	-0,05	0,1	0,2	0,18	0,18
Afwijking gemiddelde	8,55	7,86	5,2	15,02	8,59	1,25	7,74	4,14

Table 23 Variables in linear regression model for C.X (unequal settlements)

The table above presents the linear regression pattern for deformations. The upper half of the table shows the slope of each characteristics (rows) at each cross-validation fold (columns). The lower half of the table presents some model values, like the y-intercept of a linear regression and the performance score in terms of coefficient of determination.