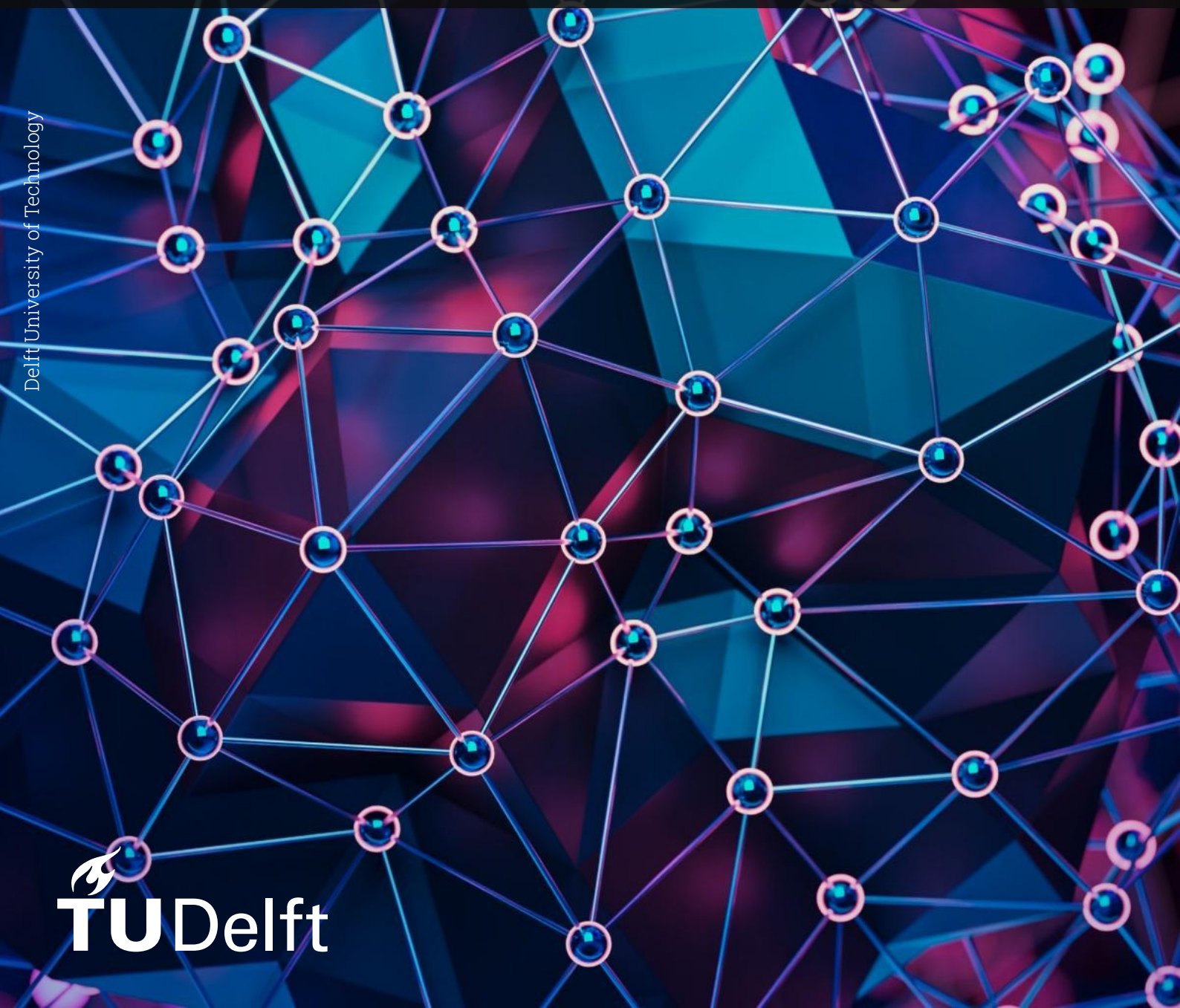
Cluster-Driven Risk Classification

Adapting Car Insurance Risk Models through Zip Code and License Plate Clustering

Faculty EEMCS, Delft University of Technology
Alyssa Wijker



TU Delft

Cluster-Driven Risk Classification

Adapting Car Insurance Risk Models through Zip Code and License Plate Clustering

by

Alyssa Wijker

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Thursday July 25, 2024 at 15:30.

Student number:	4816870	
Project duration:	January 8, 2024 – July 25, 2024	
Thesis committee:	Dr. N. Parolya,	TU Delft, daily supervisor
	Dr. N. V. Budko,	TU Delft, responsible supervisor
	T. R. de Wit,	Achmea, external supervisor

This document has been modified to protect confidential information.

An electronic version of this thesis is available at http://repository.tudelft.nl/.
Cover image adapted from https://www.itransition.com/machine-learning/statistics.

TU Delft

achmea

Preface

This thesis has been submitted for the degree of Master of Science in Applied Mathematics at Delft University of Technology, with a specialisation in Financial Engineering. Academic supervision was provided by Dr. Nestor Parolya. The research was conducted at Achmea under the supervision of Tim de Wit (Pricing Actuary), within the "Schade Bedrijven Actuariaat" department, part of the Business Finance Team.

First and foremost, I would like to thank Dr. Parolya. His feedback and insights were invaluable in the writing of this thesis. Our meetings consistently motivated me to broaden my perspective and inspired me to be thorough.
I would like to extend my gratitude to my colleagues at Achmea for their kindness, genuine interest in my thesis, and valuable suggestions during my team presentation. I especially want to thank my supervisor at Achmea, Tim de Wit. His expertise as an actuary helped me bridge the gap between the theoretical aspects of this thesis and the financial (actuarial) industry. I am also grateful for his enthusiasm, gezelligheid, and for making me feel at home at Achmea.
I want to express my gratitude to my friends and my boyfriend for their support and willingness to listen to my breakthroughs and challenges over the past seven months.
I would like to thank my sister for her encouragement, support, and believing in me. While I was working on this thesis, she was busy with her internship project and I am grateful that we could support each other through challenges both during this time and in life overall. Last, but definitely not least, I would like to thank my parents for their unconditional love and support. From making sure I knew the capitals of the provinces for my "topo toets" in primary school to supporting me throughout the past few months during my master's thesis, they have been with me every step of my academic journey. I am forever grateful.

Alyssa Wijker
Delft, July 2024

Abstract

Insurance premiums are determined based on policyholders' risk profiles; higher risk profiles result in higher premiums. Applying uniform premiums across all risk profiles can lead low-risk policyholders to seek cheaper alternatives, leaving insurers with primarily high-risk clients whose premiums may not cover expected losses. To address this, insurers categorize policyholders through risk classification. This thesis focuses on enhancing risk classification for claim frequency models in company vehicle insurance, specifically for "WAM" (covers damage caused to the car of others) and "ARD" (covers collision damage to one's own car).

Cluster analysis, underutilized in actuarial science due to mixed data types, can group policyholders for risk classification. By incorporating these clusters as risk factors in claim frequency GLMs, variable combinations are taken into account, enabling more personalized premium pricing. Through clustering zip codes and license plates, this thesis aims to refine risk classification by using K-prototypes (clustering based on the similarity in distance to the centroids) and spectral clustering (clustering based on the spectrum (i.e. eigenvalues) of the Laplacian matrix). K-prototypes is selected because it is among the most commonly utilized techniques and its implementation is required for spectral clustering. Spectral clustering, chosen for its effectiveness with large, non-linearly separable datasets, requires observation reduction by using *U-SPEC* due to high storage requirements.

To evaluate the clustering results, actuarial experts assess the sensicality, and the clusters are integrated into the GLMs to evaluate impacts on model metrics (deviance, AICc, and BIC). Spectral clustering outperforms K-prototypes (in this context) and improves risk classification for the ARD dataset. The WAM clusters do not improve the current GLM and thus cannot be used to predict the claim frequencies. Furthermore, they do not provide additional information that would be beneficial for other purposes. Nonetheless, despite the spectral WAM clusters not improving the GLM, the spectral clustering technique shows potential for application to other insurance datasets.
The group of (significant) spectral zip code clusters of the ARD dataset is the only set of clusters where all are stable with respect to time, allowing them to be directly incorporated into the GLM. All other clusters may also be included in the GLM, provided that the time dependency of the variables used for the clustering is carefully considered. Furthermore, when reducing the number of observations, the clustering results remain stable up to a certain point.

This thesis introduces methods for handling mixed data in clustering, including a customized distance measure combining the Euclidean, Hamming, and Gower's distances. Moreover, this thesis explores observation reduction techniques and their implications concerning high dimensional clustering, topics that haven't been studied in the context of license plate and zip code clustering before. The number of clusters after the application of observation reduction techniques, is determined by the number of informative eigenvectors corresponding to the isolated eigenvalues of the Laplacian matrix. This approach has not been used in license plate and zip code clustering contexts and, prior to this thesis, informative eigenvectors were only used to establish an upper bound for the number of clusters.

Contents

Nomenclature

Abbreviations

Abbreviation	Definition
AIC	Akaike Information Criterion
AICc	Akaike Information Criterion corrected
ARD	Aanrijdingsdekking in Dutch (i.e. the part of the complete casco that covers the damage the policyholder causes to their own car by collision)
ARI	Adjusted Rand Index
BIC	Bayesian Information Criterion
ESS	Error Sum of Squares
GLM	Generalized Linear Models
KT	License Plate
P.O. Box	Post Office Box
TPL	Third Party Liability
U-SPEC	Ultra-Scalable Spectral Clustering
WAM	Wettelijke Aansprakelijkheids Dekking in Dutch (equivalent to *TPL*)
ZC	Zip Code

1 Introduction

1.1. Motivation

Premiums are calculated by insurance companies according to the risk profiles of policyholders; those with higher risk profiles are required to pay higher premiums compared to low-risk policyholders. If an insurer were to apply a uniform premium across all risk profiles, low-risk policyholders would likely choose another insurance company that offers lower premiums. This would result in the insurer only attracting high-risk profiles and, as a consequence, the premiums paid by high-risk profiles might be insufficient to cover the expected cost of insured losses. [80] This is called adverse selection. [30]

To solve this problem, insurance companies categorize policyholders into various risk levels determined by their risk profiles, a process referred to as risk classification. [42] The increasing competition in the insurance industry forces companies to improve the analysis of their policyholders' risk profiles and with extensive data in the car insurance portfolio, insurers can create advanced models and algorithms to set premiums that align with the specific risk associated with each policyholder.

When classifying risk, the claim frequency and claim severity are modeled separately. Here, the claim frequency refers to the number of claims per unit of time, with the unit of time corresponding to the period for which premiums have been paid, known as the exposure. On the other hand, claim severity represents the average cost per claim. [43] Since the claim frequency tends to be more stable than the claim severity (due to the limited possible observations, namely 0 and 1), it can be calculated more accurately. [70] Therefore, this thesis will focus on the risk classification of claim frequency models of car insurance.

Cluster analysis is a widely utilized technique in statistical data analysis and machine learning that aims to reveal group structures within datasets. This method involves grouping objects in a manner that maximizes heterogeneity between the resulting clusters, while simultaneously maximizing homogeneity among observations classified within each cluster. In actuarial applications, clustering methods can be valuable for creating groups of policyholders, thereby enhancing customer segmentation and thus improving risk classification. However, clustering techniques remain underutilized in actuarial science. This is largely due to the mixed data (numerical, categorical, and ordinal) that is used in this field while many clustering techniques rely on the Euclidean distance between numerical data points to measure similarity. [41]

This thesis aims to improve the risk classification of the claim frequency models by applying clustering techniques on zip codes and license plates.

1.2. Contribution of this thesis

In 1997, Williams and Huang introduced the application of the K-means algorithm in the actuarial field to identify policyholders with a high claims ratio within a motor vehicle insurance portfolio and Huang later expanded the K-means algorithm to handle datasets containing both numerical and categorical variables in 1998. [78] [34] Furthermore, Jamotton et al. modified the K-means algorithm to cluster ordinal data in 2023, but the paper does not consider categorical data and the numerical data must be converted to ordinal data for this algorithm to be effective. This is also the only paper that applies spectral clustering to an insurance portfolio. [41] So, to date, no algorithm has been developed that effectively clusters numerical, categorical, and ordinal data. Furthermore, no research paper has employed spectral clustering on zip codes and license plates for the risk classification in car insurance.

Also, De Bont performed one-dimensional clustering of zip codes based on their claim frequencies in 2022, and Esposito clustered zip codes with the condition that those within the same cluster must be contiguous on a map in 2019. [9] [79] Lastly, up till now, no license plates have been clustered to improve the risk classification of car insurance.

The contribution of this thesis is that it introduces two modified algorithms: one based on the K-means method and the other on spectral clustering. These algorithms are designed to effectively handle mixed data consisting of numerical, categorical, and ordinal variables. Notably, this research will pioneer the clustering of license plates to enhance the risk classification in car insurance and it marks the first application of spectral clustering to both zip codes and license plates. For this multi-dimensional clustering, there is no requirement that zip codes in the same cluster must be contiguous. Lastly, this thesis will explore observation reduction techniques and their implications concerning high-dimensional clustering (i.e. scenarios where the number of features is comparable to or greater than the number of observations), topics that have not been studied in the context of license plate and zip code clustering before. [8]

1.3. Research objective and research questions

In summary, **the objective** of this thesis can be described as follows:

Improve the risk classification of the claim frequency models of two coverages, namely "WAM" ("wettelijke aansprakelijkheidsverzekering" in Dutch) and "ARD" ("aanrijdingsverzekering" in Dutch), of a car insurance product (significantly) by clustering zip codes and license plates and using these clusters as risk factors in the models.

In order to achieve this goal, the following **research question** has to be answered:

How can zip codes and license plates be clustered in such a way that, by using these clusters as risk factors, the risk classification of the claim frequency models of two coverages ("WAM" and "ARD") of a car insurance product is improved significantly?

What the "WAM" and "ARD" coverages entail, will be discussed in the next chapter.

The research question consists of the following eleven **sub-questions** (the answer to each question can be found in the section between the brackets):

1. What modifications should be done to the dataset before the clustering techniques can be applied? (Section 4.1)
2. What clustering techniques should be used for the problem at hand? (Section 3.3)
3. How should the optimal number of clusters be determined (for each method)? (Section 4.2)
4. How should the different clustering techniques be validated and their results be compared? (Section 4.4)
5. Which of the clustering techniques performs the best for the problem at hand? (Section 5.2 and Chapter 6)
6. How does the use of clustering techniques affect the risk classification of the claim frequency models? (Section 5.2)

7. What observation reduction techniques should be used for the problem at hand? (Section 4.3)
8. How should the different observation reduction techniques be validated and their results be compared? (Section 4.4)
9. What is the effect of using observation reduction techniques? (Section 5.3)
10. How stable are the clustering results with respect to the time? (Section 5.3)
11. What are the ethical implications of this research? (Section 7.2)

Before these questions are answered, some distance measures and the fundamentals of graph theory (that are required for this project) are outlined in Chapter 2. Next, the problem setting is introduced in more detail in Chapter 3; it describes the context in which the research is conducted, discusses the downsides of the current approach to risk classification, provides an overview of related approaches to solving the problem, and proposes the solution that will be explored in this thesis. Chapter 4 contains an overview of the methodology for the data preparation, clustering, observation reduction, and evaluation. The empirical results and the evaluations of the clustering techniques can be found in Chapter 5. The thesis is concluded in Chapter 6 and Chapter 7 describes the limitations of the findings of this research, proposes notions for further research, and reflects on the ethical considerations.

In the appendix, at the end of the report, detailed results can be found.

2 Preliminaries

Before describing the problem in more detail in Chapter 3, this chapter dives into the mathematical concepts required for the project. Section 2.1 outlines the distance measures that are used for the clustering methods and Section 2.2 explains the fundamentals of graph theory that are necessary for the implementation of the spectral clustering algorithm.

2.1. Distance measures

This section provides explanations of the distance measures that are used for the clustering methods. Each subsection addresses a specific measure; the Euclidean distance which is used for numerical values is described in Subsection 2.1.1, the Hamming distance for the categorical (nominal) values is discussed in Subsection 2.1.2, and Subsection 2.1.3 outlines the calculation of Gower's distance for categorical (ordinal) values.
Note that all three distance measures are symmetric i.e. for all points p and q holds that $d(p,q) = d(q,p)$ (where $d(p,q)$ is the distance between p and q).

2.1.1. Euclidean distance

The distance between numerical data points is calculated with the Euclidean distance which is equal to the length of the line segment between the data points. [66] First the formula of the distance is given for points in a one dimensional space. After that, the formula of the distance for points in a multi-dimensional space is stated.

One dimensional space

The distance between two points on the real line is determined by the absolute value of the numerical difference in their coordinates. [29] In other words, if p and q represent two points on the real line, the Euclidean distance between them ($d_E(p,q) \in \mathbb{R}$) is equal to:

$$d_E(p,q) = |p-q| = \sqrt{(p-q)^2}$$

Multi-dimensional space

For points $p = (p_1, \ldots, p_n)^\top$ and $q = (q_1, \ldots, q_n)^\top$ in an n-dimensional real space, the Euclidean distance ($d_E(p,q) \in \mathbb{R}$) is equal to:

$$d_E(p,q) = \sqrt{(p_1-q_1)^2 + \cdots (p_n-q_n)^2} = \|p-q\|$$

The last expression is referred to as the Euclidean norm. [66] Note that this formula is used for the data in this report since the numerical data points are multi-dimensional.

The Euclidean distance between two categorical values is undefined (e.g. (the absolute value of) the numerical difference between the color red and blue is unknown). Instead, for categorical values, the distance measures outlined in the following two subsections are used.

2.1.2. Hamming distance

The Hamming distance is equal to the number of mismatches between categorical data points. For example, the Hamming distance between "012345" and "022546" is equal to 3 and the Hamming distance between [red, BMW, Germany] and [yellow, Porsche, Germany] is 2. This means that for two data points p and q with n categorical variables, $0 \leq d_H(p,q) \leq n$ and $d_H(p,q) \in \mathbb{N}$ (where d_H is the Hamming distance). [69]
Note that the Hamming distance will only be used for the categorical (nominal) variables since the Hamming distance does not take the order of categorical (ordinal) variables into account. For example, the Hamming distance is equal to 1 for urbanization levels of 1 and 2, but also for urbanization levels of 1 and 5. Therefore, Gower's distance is applied to the categorical (ordinal) variables.

2.1.3. Gower's distance

Let $\mathbf{X} = \{x_{ij}\}$ be a data matrix with m rows (i.e. data points) and n columns (i.e. variables). Gower's similarity $G(j,k) \in \mathbb{R}$ between data points j and k (regardless of the data types) is equal to:

$$G(j,k) = \frac{\sum_{i=1}^{n} w_{ijk} s_{ijk}}{\sum_{i=1}^{n} w_{ijk}} \tag{2.1}$$

Where $w_{ijk} \in [0,1]$ is the weight of data points j and k and variable i, and $s_{ijk} \in \mathbb{R}$ is the similarity score of data points j and k for variable i. [59]

This means that Gower's distance (i.e. Gower's dissimilarity) $d_G(j,k) \in \mathbb{R}$ between data points j and k is equal to:

$$d_G(j,k) = 1 - G(j,k) \tag{2.2}$$

$G(j,k)$ is defined for numerical and categorial (both nominal and ordinal) variables. [59] However, for this project, Gower's distance will be used as a distance measure for ordinal variables *only*.
In order to calculate $d_G(j,k)$ for ordinal variables, all x_{ij} have to be ranked first. An example of this ranking is shown in Figure 2.1. The first row shows the environment-friendliness of a car (with "A" being the most eco-friendly) for eight data points and the second row shows the ranking of this variable for the eight observations. Note that the number of possible states of the environment friendliness variable is less than the number of data points (since $4 < 8$). Thus, in the ranking of objects, ties cannot be avoided: objects having the same score will take the same position in the ordering. Row 3 shows the partially ranked variable that is converted to ranks by computing the following value:

$$\text{Number of variables with a lower partial rank} + 1 + \frac{\text{Number of objects that have the same partial rank} - 1}{2}$$

Lastly, the T value in the final row of the figure is equal to the number of objects that have the same rank score. [59]

Data point	1	2	3	4	5	6	7	8
Environment friendliness	A	C	A	A	D	A	C	B
Partially ranked variable	1	3	1	1	4	1	3	2
Partially ranked variable converted to ranks	2.5	6.5	2.5	2.5	8	2.5	6.5	5
T Value	4	2	4	4	1	4	2	1

Figure 2.1: This figure shows an example of the ranking of the environment-friendliness variable (with "A" being the most eco-friendly) for eight data points. The partially ranked variables, partially ranked variables converted to ranks, and T values are shown.

After all x_{ij} have been replaced by their ranks $r_{ij} \in \mathbb{R}_{>0}$, then $w_{ijk} \in \{0,1\}$ and $s_{ijk} \in \mathbb{R}$ can be calculated for ordinal variables in the following way:

$$w_{ijk} = \begin{cases} 0 & \text{if } x_{ij} \text{ or } x_{ik} \text{ is unknown} \\ 1 & \text{if both } x_{ij} \text{ and } x_{ik} \text{ are known} \end{cases}$$

$$s_{ijk} = \begin{cases} 1 & \text{if } r_{ij}{=}r_{ik} \\ 1 - \frac{|r_{ij}-r_{ik}|-(T_{ij}-1)/2-(T_{ik}-1)/2}{\max\{r_i\}-\min\{r_i\}-(T_{i,\max}-1)/2-(T_{i,\min}-1)/2} & \text{otherwise} \end{cases}$$

Here $T_{ij} \in \mathbb{N}$ is the number of objects which have the same rank score for variable i as data point j (including j). Furthermore, $T_{i,\max} \in \mathbb{N}$ is the number of objects that have the maximum rank ($\max\{r_i\}$) and $T_{i,\min} \in \mathbb{N}$ is the number of objects which have the minimum rank ($\min\{r_i\}$). [59]

Gower's distance for ordinal variables can then be obtained by substituting w_{ijk} and s_{ijk} in Equations 2.1 and 2.2. [59]

2.2. Fundamentals of graph theory

This section covers the fundamentals of graph theory that are necessary for the implementation of the spectral clustering algorithm. In Subsection 2.2.1 some basic graph notation is introduced and Subsection 2.2.2 discusses various methods of constructing similarity graphs.

2.2.1. Graph notation

Definition 2.2.1 (Graphs)**.** A **graph** G is a pair of sets (V, E) where V is non-empty and E is a subset of the set $\{\{v_i, v_j\} : v_i, v_j \in V, v_i \neq v_j\}$ of all two-element subsets of V. The set V is known as the set of **vertices** and the set E as the set of **edges**. [7]

Definition 2.2.2 (Adjacency matrix)**.** Two vertices v_i, v_j in a graph $G(V, E)$ are called **adjacent** if $v_i v_j \in E$ and **nonadjacent** if $v_i v_j \notin E$. The **adjacency matrix** A is a matrix where $A_{i,j}$ is 1 if the i-th vertex is adjacent to the j-th vertex and 0 otherwise. [7]

For the spectral clustering algorithm, $G(V, E)$ (consisting of n vertices) is assumed to be an undirected graph (i.e. the edges have no specified direction assigned to them). [26] Furthermore, it is assumed that G is weighted; each edge between two vertices v_i and v_j is allocated a non-negative weight $w_{ij} \geq 0$. [51]

Definition 2.2.3 (Weighted adjacency matrix)**.** The **weighted adjecency matrix** of $G(V, E)$ is equal to $W = (w_{i,j})_{i,j=1,\ldots,n}$ with w_{ij} as defined before. If v_i and v_j are non-adjacent, $w_{ij} = 0$. Furthermore, as G is undirected, it holds that $w_{ij} = w_{ji}$. [51]

Definition 2.2.4 (Degree of a vertex)**.** The **degree** of vertex $v_i \in V$ is equal to $d_i = \sum_{j=1}^{n} w_{ij}$. Note that the sum only runs over all vertices adjacent to v_i, as for non-adjacent vertices v_k, $w_{ik} = 0$. [51]

Definition 2.2.5 (Degree matrix)**.** The **degree matrix** D is a diagonal matrix where the degrees $d_1, \ldots, d_n$ are placed on the diagonal. [51]

For subset $B \subset V$, the shorthand notation $i \in B$ is used to denote the set of indices $\{i | v_i \in B\}$. Lastly, the size of subset B is represented by $|B|$, specifying the number of vertices in the subset. [51]

2.2.2. Similarity graphs

By constructing similarity graphs, the local neighborhood relationships between the data points are modeled. To transform a given set $x_1, \ldots, x_n$ of data points with pairwise similarities $s_{ij} \in \mathbb{R}$ (see Figure 2.2a for an example) to a similarity graph, the following techniques can be used. [51]

- **The ϵ-neighborhood graph**. All points whose pairwise distances are smaller than ϵ are connected by an edge. Since the distances between connected points are generally within the same scale (at most ϵ), weighting the edges would not add additional information about the data to the graph. Therefore, the ϵ-neighborhood graph is typically regarded as an unweighted graph. [51] Figure 2.2b shows the ϵ-neighborhood graph of the example in Figure 2.2a for $\epsilon = 0.5$.
- **k-nearest neighbor graphs**. Vertex v_i is linked with v_j if v_j is among the k-nearest neighbors of v_i (based on distance). This definition yields a directed graph since the neighborhood relationship is not symmetric. There are two methods to convert this graph into an undirected one. [51]

The first approach is to disregard the edges' directions, meaning v_i and v_j are connected with an undirected edge if either v_i is among the k-nearest neighbors of v_j *or* vice versa. The resulting graph is known as the *k-nearest neighbor graph*. [51]
The second method links vertices v_i and v_j if *both* v_i is among the k-nearest neighbors of v_j and v_j is among the k-nearest neighbors of v_i. This yields the *mutual k-nearest neighbor graph*. [51]
In both cases, after connecting the appropriate vertices, the edges are weighted by the similarity of their endpoints. [51] Figures 2.2c and 2.2d respectively show the k-nearest neighbor graph and *mutual* k-nearest neighbor graph of the example for $k = 2$.

- **The fully connected graph**. All points that have a positive similarity with each other are connected. The edges are weighted by s_{ij}. [51] Figure 2.2e shows the fully connected graph of the example.

Note that for the similarities between the data points, the distance measures of Section 2.1 can be used (since $s_{ij} = 1 - d_{i,j}$).

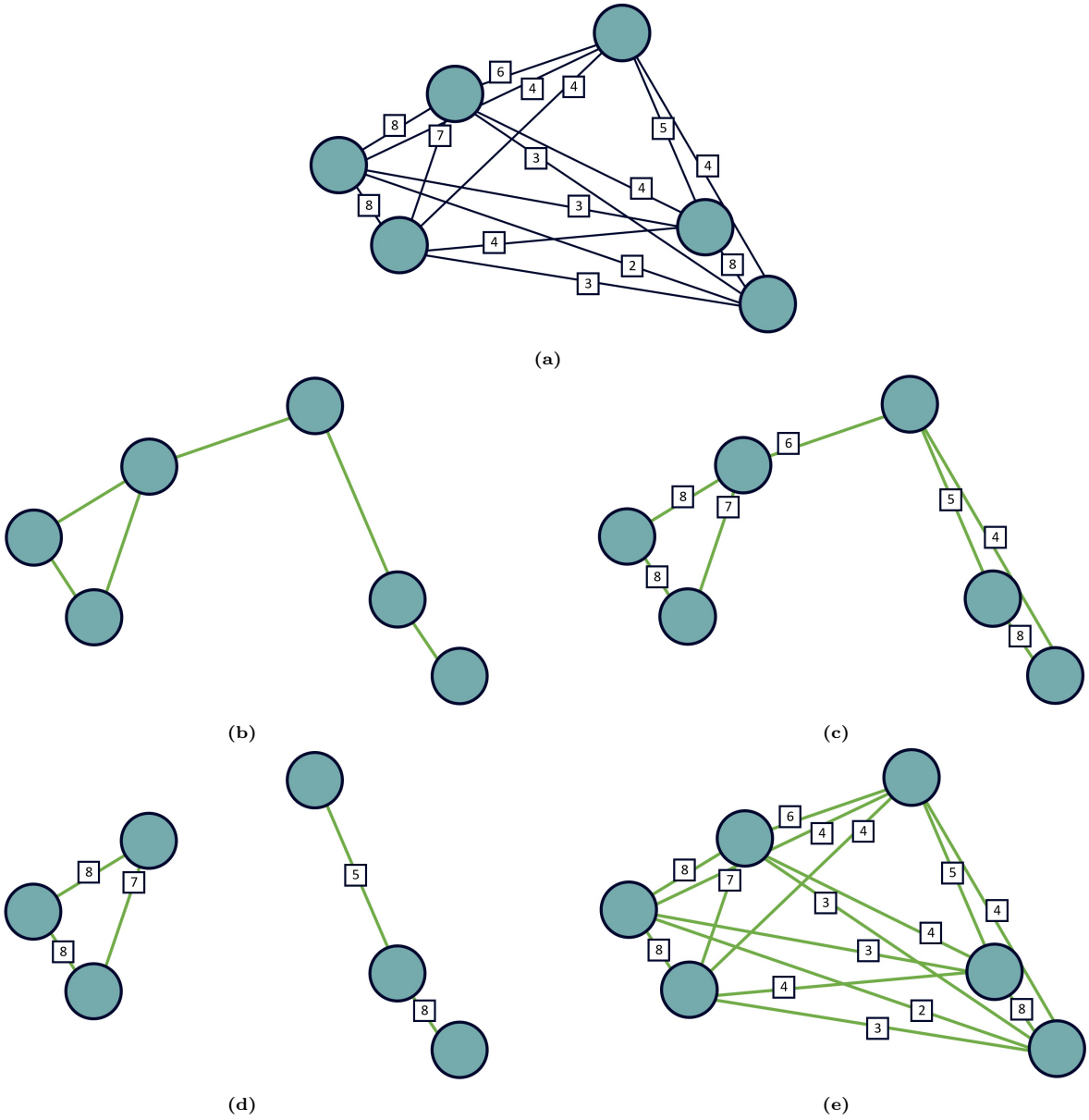


Figure 2.2: This figure shows: **(a)** an example of six data points (blue circles) with their pairwise similarities multiplied by 10 (black squares), **(b)** Its ϵ-neighborhood graph ($\epsilon = 0.5$), **(c)** Its k-nearest neighbor graph ($k = 2$), **(d)** Its mutual k-nearest neighbor graph ($k = 2$), **(e)** Its fully connected graph.

3

Problem setting

This chapter contains an overview of the problem setting. Section 3.1 describes the context in which the research is conducted; it explains how (car) insurance works, summarizes the various kinds of car insurance coverages, and describes the data used for this project. Section 3.2 discusses the downsides of the current approach to risk classification and provides some important considerations when coming up with a new approach. Lastly, Section 3.3 outlines related approaches to solving the problem and Section 3.4 concludes the chapter by proposing the solution that will be explored in this thesis.

3.1. Context description

3.1.1. What is (car) insurance?

An insurance contract (called a policy) is an agreement between the policyholder (i.e. the insured) and the insurer. This contract provides financial protection against the occurrence of a certain event such as damage, loss, or illness. In order to provide this protection, the policyholder pays specified premiums to the insurance company and in return the company provides a guarantee of compensation if the event occurs. Therefore, insurance can be seen as transferring risk from the insured to the insurer. [44]

Insurance can be categorized into two groups: life insurance, and non-life (or general) insurance. Life insurance pays out an insured amount to the beneficiary or beneficiaries whenever the policyholder deceases within the term of the policy, at the expiration date when the policyholder is alive, or both depending on the policy. [74] Due to the long policy period, life insurance is often seen as an investment. [67] On the other hand, non-life insurance covers aspects that are unrelated to human life and includes coverage for properties like homes and vehicles, as well as health and travel insurance. Moreover, it provides financial protection against losses inflicted by another person (such as theft and accidents), and by natural disasters (such as floods, fires, and environmental events). Non-life insurance plans are often of shorter term than life insurance plans. [74]

Car insurance is a type of non-life insurance that provides protection against financial losses caused by an accident or by other damage to a vehicle (e.g. theft, vandalism, etc.). What kind of protection the car insurance exactly offers, depends on the type of coverage. [45] This will be further discussed in the next subsection.

3.1.2. Types of car insurance coverages

In the Netherlands, there are three types of car insurance coverages:

Third-Party Liability

Third-party liability TPL insurance ("wettelijke aansprakelijkheidsverzekering" in Dutch) covers the damage that the policyholder causes to the car and property of others. This means that the damage

the policyholder causes to one's own car is not covered. [4] TPL is required by law to protect other people. [55]

Third-Party Liability + Limited Casco

This coverage consists of the TPL coverage and of limited casco; damage to the car of others *and* part of the damage to the policyholder's car is covered. This includes damage to the policyholder's car as a result of theft, fire, collision with an animal, broken windows, storm, and hail. Damage the policyholder causes to one's own car is not covered. [4]

All Risk (Third-Party Liability + Complete Casco)

The all risk coverage consists of the TPL coverage and of complete casco: damage to the car of others *and* damage to the policyholder's car (even if this is the policyholder's fault) is covered. [4]

This thesis focuses on two coverages: TPL, and the part of the complete casco that covers the damage the policyholder causes to their own car by collision. From here on out, these two coverages are referred to by their Dutch abbreviations: "WAM" and "ARD".
By considering these two coverages separately, a level of homogeneity between the individuals can be guaranteed since the variations across groups based on their insurance choices will be eliminated. Moreover, each coverage protects the policyholder against a specific risk. By studying these coverages separately and only considering the factors belonging to those coverages, the risks connected to them can be predicted more accurately.

3.1.3. Available data

For this project, the data connected to company cars is considered. These cars can, for example, be used for the transport of goods, as taxis, as ambulances, etc. Moreover, as mentioned previously, only data of cars with the WAM and/or ARD coverage are taken into account which results in two datasets (one per coverage).
The data has been provided by Achmea and spans over ten years. Figure 3.1 shows the first five rows of the ARD dataset (dummy values are displayed and some columns are ommitted). Every entry in a dataset corresponds to an individual policyholder throughout an exposure period. The policyholder's data is recorded at the start of the policy period and remains constant throughout the exposure duration. However, if there is a change in any characteristic, a new record is generated for the policyholder. This also means that if there is a claim, a new record is created. Therefore, the data of a policyholder can be split into claim records and no claim (i.e. policy) records. For this project, only the claim records are taken into account since these records contain the most information regarding the risk for which the coverage offers protection. However, the claim frequencies are extracted from the entire dataset (claim and policy records) in order to evaluate the clustering methods at a later stage. How these claim frequencies are calculated, will be explained in Section 3.2.

License plate	Zip code	Record	Policy number	...	Number of claims	Claim cost	Start date	End date	Number of insured years
48MNR9	1628BR	Policy	123456789	...	0	0	10APR2012	07MAY2012	0.071038
98GTR0	3012MN	Policy	987654321	...	0	0	03MAR2013	01JAN2014	0.830601
98GTR0	3012MN	Policy	987654321	...	0	0	01JAN2014	01JAN2015	1
98GTR0	3012MN	Claim	987654321	...	1	7500	01JAN2015	10JAN2016	0
98GTR0	3012MN	Policy	987654321	...	0	0	01JAN2015	10JAN2016	1

Figure 3.1: This figure shows the first five rows of the ARD dataset (dummy values are displayed and some columns are ommitted).

Data classes and dimensions

The data of the (claim records of) policyholders can be grouped into five data classes: policy related, claim related, geographical, vehicle, and insured company related data. These classes and examples per class are shown in Figure 3.2.
In order to cluster license plates, only data connected to the license plates can be used. Therefore, only the vehicle data is used for this clustering. For the same reason, for the clustering of the zip codes, only the geographical data is used.

Note that the vehicle and geographical data remain constant throughout the policy length. For example, the car brand connected to a license plate will stay the same over time. Therefore, only the most recent claim records are taken into account for the clustering.

Data classes	Examples
Policy related	Policy number, premium paid per year, type of coverage
Claim related	Number of claim free years, number of claims, claim date
Geographical data (zip code)	Urbanisation level, social class, province
Vehicle data (license plate)	Car weight, car brand, fuel type
Insured company related data	Company activities, legal entitity, company sector

Figure 3.2: This figure shows the five classes into which the data can be grouped. Examples of characteristics are also shown per class.

By only considering the most recent claim records, 60,083 and 39,311 rows of unique license plates (of the ARD and WAM datasets respectively) are available for the clustering. In the same way, 33,903 and 25,049 rows of unique zip codes (of the ARD and WAM datasets respectively) are available. Moreover, the vehicle data consists of 114 characteristics (i.e. columns), and the geographical data of 158.

Considerations for the data

Before the data connected to the license plates and zip codes can be clustered, the following matters need to be considered:

- **P.O. box zip codes**. Some of the cars are registered under a zip code associated with a post office box (i.e. P.O. box). This means that all of the geographical data is connected to the zip code of that P.O. box and *not* to the zip code of the company/house (where the car is parked most often). Therefore, the characteristics of the geographical data do not contain any information regarding the risk for which the coverage offers protection. Using these P.O. box zip codes alongside the zip codes of companies/houses can thus negatively impact the clustering.
- **(Multi)collinearity**. Multicollinearity occurs when two or more characteristics are highly linearly related (see Definition 3.1.1). For example, in the vehicle data the features X_1 = Days to export and X_2 = Months to export are almost perfectly collinear since the equation in Definition 3.1.1 always approximately holds with $\lambda_0 = 0$, $\lambda_1 = 1$, $\lambda_2 = -12$, and $c = 0$. In this case, using both features for the clustering instead of one does not provide extra information. Furthermore, using both features puts more weight on a "time to export" feature. This influences the clustering negatively.

  Definition 3.1.1 (Multicollinearity)**.** Variables $X_1, \cdots, X_n$ are said to be perfectly **multicollinear** if there exist $\lambda_0, \cdots, \lambda_n \in \mathbb{R}$ such that

  $$\lambda_0 + \lambda_1 X_{1i} + \cdots + \lambda_n X_{ni} = c, (c \in \mathbb{R})$$

  holds for every ith and jth variable X_{ji}. [49]

  In this thesis, multicollinearity is assessed by examining pairwise correlations $r_{x,y}$ between variables x and y. If the correlation exceeds 0.95, one of the features is removed from the dataset. The formula for the correlation coefficient will be provided in Subsection 4.1.1.
  A correlation of $r_{x,y} = 0.95$ was chosen as the threshold for detecting multicollinearity. This decision stems from the guideline for the VIF (Variance Inflation Factor), where multicollinearity is considered high if $VIF = \frac{1}{1-r_{x,y}^2} > 10$. This translates to $1 - r_{x,y}^2 < \frac{1}{10}$, implying that $r_{x,y} > \sqrt{1 - \frac{1}{10}} \approx 0.95$. [37]
- **Missing values**. For various policyholders, some of the characteristics are unknown. These missing values can be imputed with multiple different methods or the corresponding policyholder (or characteristic) can be deleted altogether.
- **Non-standardized data**. The different columns (i.e. characteristics) of the datasets differ in their ranges. For example, the price of the car can range from 0 to 200,000 while the number

of car doors only ranges from 1 to 8. Since most clustering techniques (including K-means) are based on the distance between data points, a clustering can be completely dominated by a column such as the car price. It is therefore crucial to standardize the data prior to the clustering. [76]

- **Mixed data**. The vehicle and geographical datasets contain numerical and categorical (both nominal and ordinal) features. Categorical data is information that is divided into groups. This can further be broken down into nominal data (categories without an inherent order or structure) and ordinal data (categories with a specific order or understood structure through the categorical names). Numerical data refers to data in the form of numbers. In this case, just as with ordinal data, there is a specific order. Moreover, numerical data has equal spacing (e.g. the distance between 1 and 2 is equal to the distance between 2 and 3). Ordinal data does not have this equal spacing (e.g. it is impossible to say if urbanisation levels of 1 and 2 have the same distance between them as levels 2 and 3). [72] Figure 3.3 shows an overview of the different data types (including examples). Furthermore, Figures 3.4 and 3.5 show some examples of characteristics with their corresponding data types for the vehicle and geographical data respectively.
  Most clustering techniques can only be applied to numerical data. Therefore, these techniques need to be modified before they can be used for the datasets of this project.

Section 4.1 describes how the five matters of consideration described in this subsection are resolved.

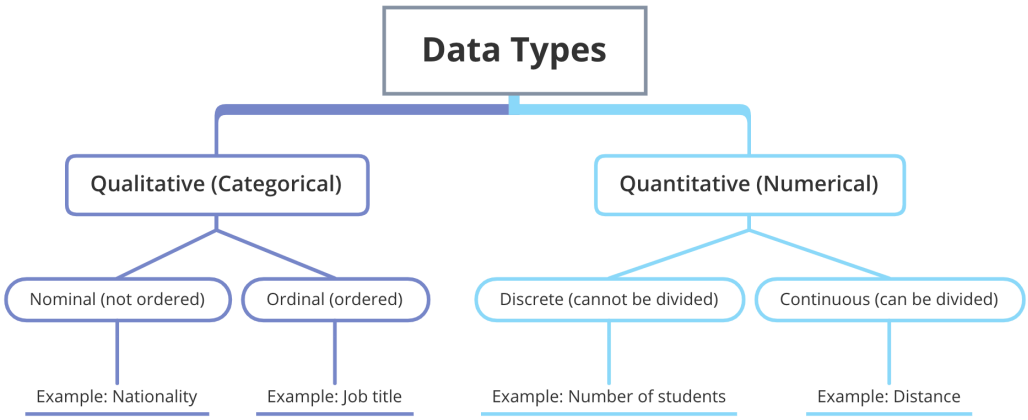


Figure 3.3: This figure shows an overview of the different data types (including examples).

Data name	Description	Type of data
kenteken	License plate	String
gewicht_def	Car weight	Integer
klasse_gewicht	Class of car weight	Categorical (Ordinal)
ouderdom	Car age	Integer
klasse_ouderdom	Class of car age	Categorical (Ordinal)
merk_landherkomst	Country of origin of car	Categorical
merk_def3	Car brand	Categorical
brandstof_def	Fuel type	Categorical
aantal_zitplaatsen	Number of seats	Integer
vermogen_def	Car power (hp)	Integer
klasse_vermogen	Class of car power	Categorical (Ordinal)
klasse_topsnelheid	Class maximum speed	Categorical (Ordinal)
ind_zuiningheid	Class of environment friendliness	Categorical (Ordinal)

Figure 3.4: This figure shows some examples of characteristics with their corresponding data types for the vehicle data.

Data name	Description	Type of data
postcode	Zip code	Categorical
ind_urbanisatie	Urbanisation level	Categorical (Ordinal)
ind_provincie	Province	Categorical
postcode_eerste_positie	First number of zip code	Categorical
ind_sociale_klasse	Social class	Categorical (Ordinal)
ind_nietwall	Level of non-western immigrants	Categorical (Ordinal)
ind_perc_zak_auto	Level of percentage of company cars	Categorical (Ordinal)
ind_TYPE_WON	Type of home	Categorical
ind_corop	Index COROP (cluster of municipalities in same province)	Categorical
ind_bedr_prak_woningen	Level of homes used partly as companies/practices	Categorical (Ordinal)
ind_nielsen	Nielsen area (e.g. 1 is AMS, ROT, The Hague, 5 ZL, NB, LB)	Categorical
ind_FINTYPE	Financial type of habitants (e.g. active borrowers, financial professionals etc.)	Categorical

Figure 3.5: This figure shows some examples of characteristics with their corresponding data types for the geographical data.

3.2. The problem

As mentioned in the Introduction, the research question of this thesis is as follows:

How can zip codes and license plates be clustered in such a way that, by using these clusters as risk factors, the risk classification of the claim frequency models of two coverages ("WAM" and "ARD") of a car insurance product is improved significantly?

This poses the following question: how can the current risk classification of the claim frequency models of the WAM and ARD coverages be improved? In other words: what are the downsides to the current approach of this risk classification? These questions are answered in Subsection 3.2.1. Moreover, Subsection 3.2.2 provides some important considerations when coming up with a new approach.

3.2.1. The current approach

Calculating the premium

The premium of a policy is determined by three fundamental components; operational expenses, profits, and variable costs. [58] For this project, the variable costs (i.e. the pure premiums) are of interest. The pure premium is defined as the expected cost of all the claims a policyholder is anticipated to file throughout a coverage period. [18] In order to explain the calculation of the pure premium, the computations of the claim frequency F_i and claim severity S_i of policyholder i are discussed first.

As explained in the Introduction, the claim frequency $F_i \in \mathbb{R}_{\geq 0}$ of policyholder i is equal to the number of claims $N_i \in \mathbb{N} \cup \{0\}$ per unit of time for which premiums have been paid $t_i \in \mathbb{R}_{>0}$ (referred to as exposure). [9] Thus,

$$F_i = \frac{N_i}{t_i} \tag{3.1}$$

Note that the exposure t_i is measured in fractions of years (e.g. for an exposure of a month, $t_i = 1/12$) since not all policies last an entire year. [9]

The claim severity $S_i \in \mathbb{R}_{\geq 0}$ of a policyholder i is equal to the average cost per claim. [9] So,

$$S_i = \frac{L_i}{N_i}$$

Where $L_i \in \mathbb{R}_{\geq 0}$ is equal to the total loss over a time period t_i of policyholder i. [9]

As mentioned previously, the pure premium $p_i \in \mathbb{R}_{\geq 0}$ is defined as the expected cost of all the claims a policyholder is anticipated to file throughout a coverage period. Therefore, the expected pure premium

of policyholder i is equal to the expected claim frequency F_i multiplied by the expected average cost per claim S_i. [9] So,

$$\mathbb{E}(p_i) = \mathbb{E}(F_i) \cdot \mathbb{E}(S_i)$$

Assuming that the claim frequency and claim severity are independent, these two variables can be modeled separately. This approach, known as the frequency-severity method, enables a more in-depth understanding of the underlying factors contributing to the frequency and severity of claims. Additionally, the method permits the selection of distinct distributions for claim frequency and severity which is useful considering that the claim frequency tends to align with a Poisson distribution, while the claim severity leans towards a Gamma distribution. [9]

In order to predict the expected claim frequency and severity, Generalized Linear Models (GLMs) are often employed on historical data. However, since it is infeasible to determine individual premiums for each policyholder, pure premiums are computed per risk level, utilizing estimates of the response variables. [9]

This thesis will focus on the claim frequency models since the frequency tends to be more stable than the claim severity (due to the limited possible observations, namely 0 and 1) and thus can be calculated more accurately. [70]
The variables of the GLM used for the expected frequencies, can be seen as risk factors. Which risk factors are used in the current approach will be explained next.

Risk factors used in the claim frequency GLM

For the current approach, the GLM of the claim frequencies depends solely on risk factors established to have a significant impact on claim frequencies, as evidenced by historical data. For example, when using the ARD dataset, it becomes clear that the number of claim-free years significantly affects the average claim frequency negatively which can be seen in Figure 3.6a. Therefore, this variable is used as a risk factor in the GLM of the claim frequencies. Note that the number of claim-free years can be negative in Figure 3.6a since five years are subtracted each time a claim is filed. [36]
On the other hand, in Figure 3.6b, it is evident that the percentage of company cars in a zip code area does not have a significant effect on the average claim frequency. Therefore, this variable is not considered a risk factor in the GLM.
Lastly, there is a significant effect of the province of the zip code on the average claim frequency (see Figure 3.6c) (e.g. Noord-Holland has a higher average claim frequency than a more rural province such as Drenthe). However, even though the province is considered a risk factor in the GLM, the variations in claim frequencies within the same province are substantial. For instance, Amsterdam and a village like Egmond, both located in Noord-Holland, differ significantly in claim frequencies.

So, variables lacking a direct impact on claim frequencies are excluded as risk factors. Consequently, the effect of combinations involving these variables is lost. For example, the percentage of company cars in the zip code area might have a significant effect on the claim frequency when combined with the province (e.g. a higher percentage of company cars in Noord-Holland leads to a higher claim frequency, yet the same percentage of company cars in Zeeland leads to an equally lower claim frequency, causing the variable to be overlooked in the current approach).
Applying clustering techniques allows for the consideration of the impact of variable combinations. If it becomes evident a variable is still insignificant with regard to the claim frequency, it can be omitted at a later stage by applying dimensionality reduction techniques. Moreover, clustering license plates and zip codes enables a more personalized approach to premium pricing for individuals (an aspect not feasible with the current approach, as mentioned earlier). This tailored premium leads to a more precise customer segmentation, resulting in a more representative premium. Such a premium is crucial as explained in the Introduction.

The next subsection provides some important considerations for the new approach (i.e. when applying clustering techniques).

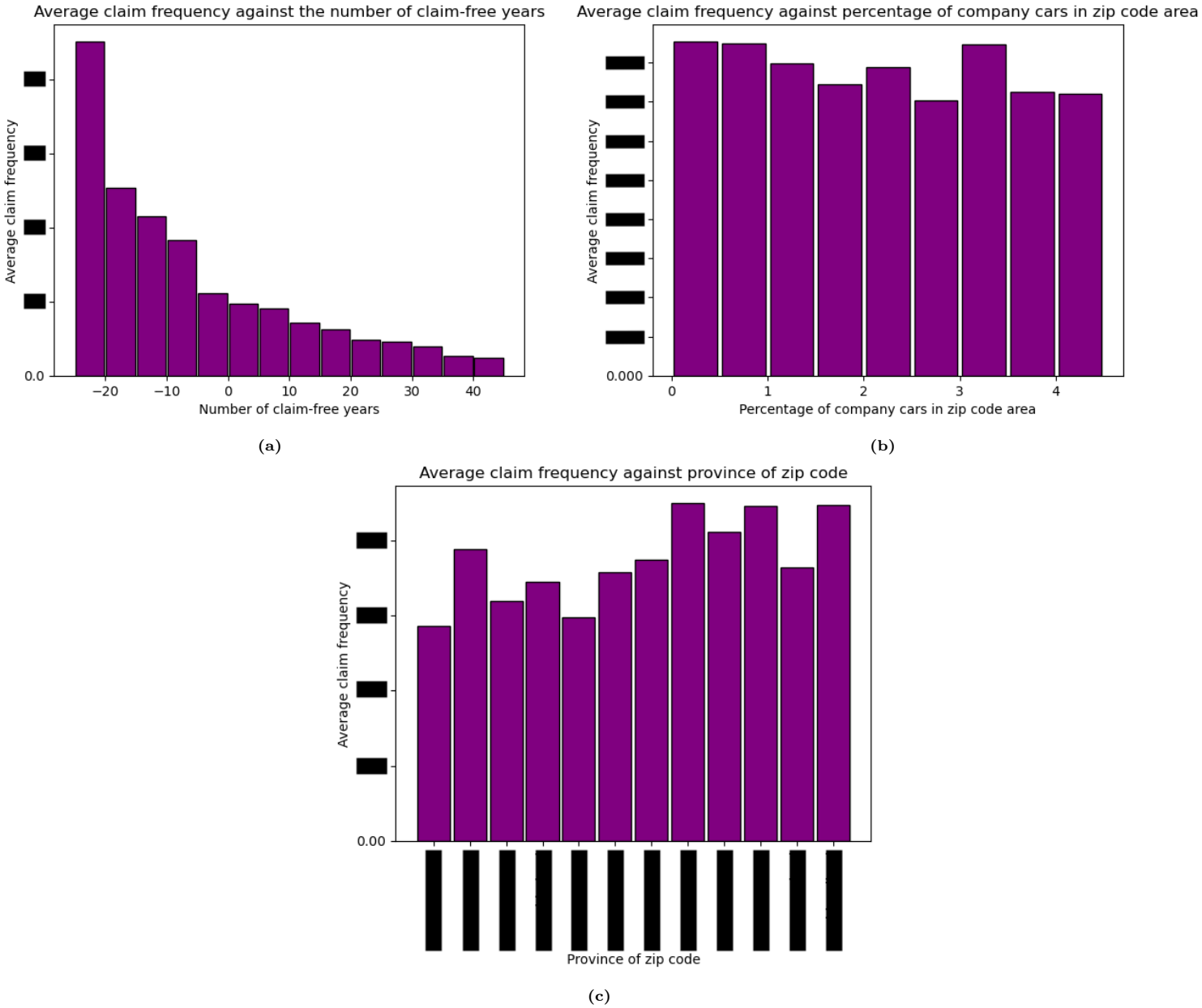


Figure 3.6: This figure shows (for the ARD dataset) bar plots of average claim frequencies against three different features: **(a)** Number of claim-free years, **(b)** Percentage of company cars in zip code area, **(c)** Province of zip code.

3.2.2. Important considerations for a new approach

In the attempt to improve the current risk classification of the frequency models by applying clustering techniques to the datasets, the following considerations need to be taken into account:

- **Mixed data**. As mentioned in Subsection 3.1.3, the vehicle and geographical data of the ARD and WAM datasets contain numerical, categorical (nominal), and categorical (ordinal) features. However, many clustering techniques are designed for numerical data, as they rely on distance measures, which are typically only defined for numerical data points. Therefore, these techniques need to be modified before they can be used for the datasets of this project.
- **Significance**. Similar to the variables in the current approach, in order to be used as risk factors in the GLM, the resulting clusters are required to have a significant effect on the average claim frequency. In other words, if all clusters result in approximately the same average frequencies, the current risk classification cannot be improved since the clusters do not contain any information

concerning the risks for which the coverages offer protection. So, the significance of the clusters with respect to the claim frequency can be regarded as a measure of success of the clustering.

- **Explainability**. As clustering license plates and zip codes allows for a more personalized premium pricing approach, policyholders will experience individualized impacts from the tailored premium they will have to pay. Therefore, there is a requirement for a certain level of explainability in the clustering process; it should be clear how the clustering techniques work and what their limitations are. Note that it is not of interest why specific zip codes and license plates are assigned to certain clusters.

The next section outlines existing approaches to clustering mixed data. After that, Section 3.4 provides a summary of the modified clustering techniques that will be applied in this thesis and it explains how the three considerations described in this subsection will be taken into account.

3.3. Related approaches

Cluster analysis is a widely utilized technique in statistical data analysis and machine learning that aims to reveal group structures within datasets. This method involves grouping objects in a manner that maximizes heterogeneity between the resulting clusters, while simultaneously maximizing homogeneity among observations within each cluster. [41]
Clustering finds applications in various domains, one of which is customer segmentation; through clustering, distinct customer groups can be identified based on their preferences, behavior, or demographics, allowing personalized marketing strategies and recommendations. Other examples of applications of clustering include image recognition, fraud detection, and data compression. [63] Across a wide array of scientific disciplines (ranging from statistics, computer science, and biology to social sciences and psychology), researchers consistently strive to gain an initial understanding of empirical data. This is often achieved by employing clustering techniques to identify groups exhibiting "similar behavior." [51]

Although there are many different clustering approaches, this thesis focuses on centroid-based and connectivity-based methods. Centroid-based clustering assigns data points to groups based on their similarity in distance to the centroid (i.e. average) of their clusters. Its objective is to minimize the sum of distances between each data point and the centroid of its allocated cluster. [61] An example is shown in Figure 3.7. Connectivity-based clustering (or hierarchical clustering) relies on the concept that each object is linked to its neighbors based on their proximity distance, indicating their degree of relationship. The stronger the connection between two data points, the higher the likelihood that they belong to the same cluster. [28] [3] An example of connectivity-based clustering is shown in Figure 3.8.

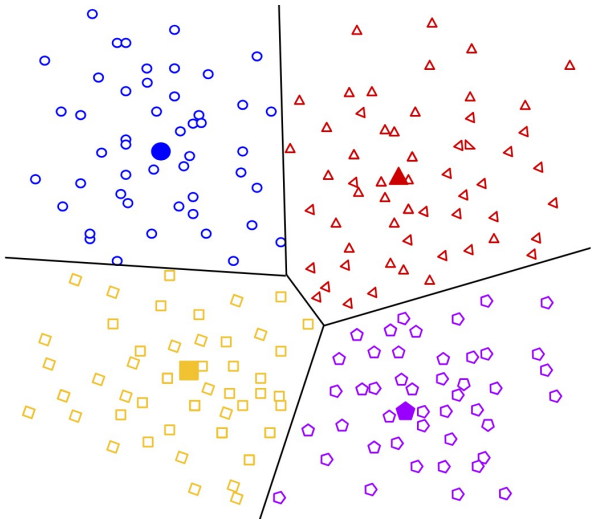


Figure 3.7: This figure shows an example of the centroid-based clustering approach. [28]

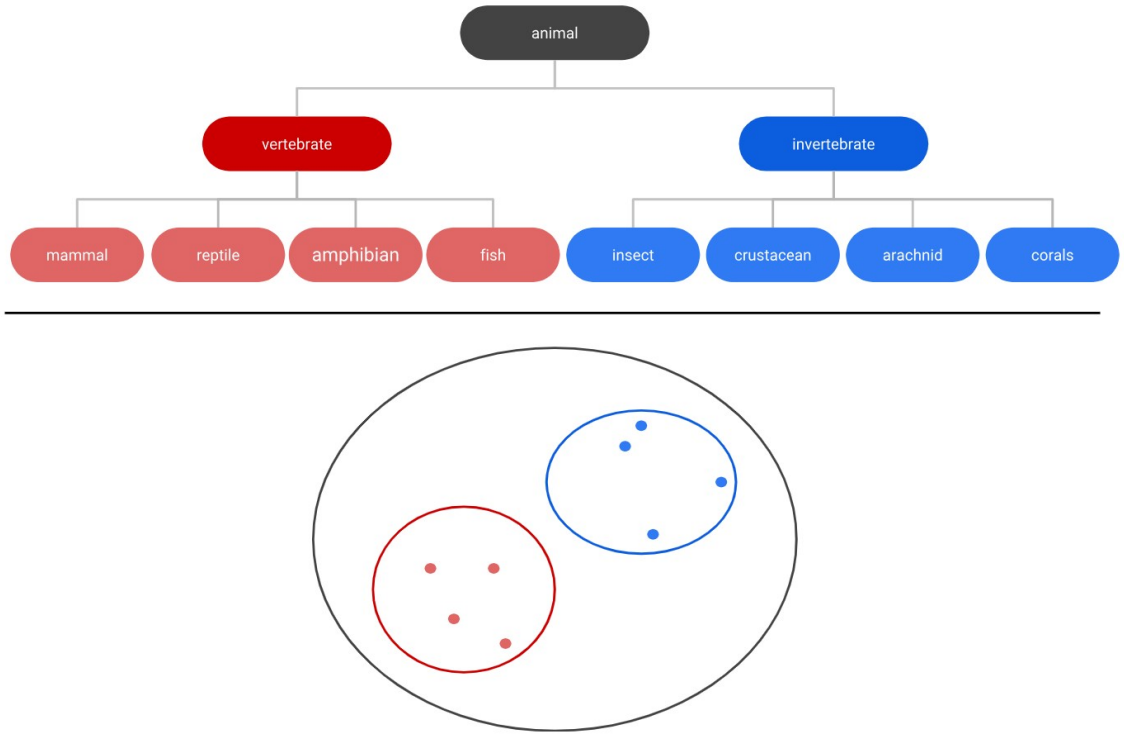


Figure 3.8: This figure shows an example of the connectivity-based clustering approach. [28]

Clustering techniques can also be categorized into unsupervised and supervised methods. Unsupervised clustering techniques operate on unlabeled data (i.e. no output data is utilized) and thus can discover hidden patterns. Therefore, claim frequencies are not taken into account for the clustering. In contrast, supervised clustering techniques operate on labeled data, thus taking claim frequencies into account. [17]

For this thesis, clusters are formed using the vehicle and geographical data of claim records. However, even if the claim frequency of a policyholder is unknown (such as with a new customer), it should still be possible to assign them to a cluster (post-formation) based on the defining characteristics of these clusters. Therefore, the clusters must remain independent of claim frequencies, and thus, the focus of this thesis lies on two unsupervised clustering techniques: K-means (centroid-based) and spectral clustering (connectivity-based). The following two subsections provide explanations of these two methods and present overviews of the benefits and drawbacks of the techniques.

K-means clustering

K-means is a widely used clustering algorithm that assigns data points to groups based on their similarity in distance to the centroid (i.e. average) of their clusters. Its objective is to minimize the sum of squared distances between each data point and the centroid of its allocated cluster. [61]
This method was selected because it is among the most commonly utilized techniques and its implementation is required for spectral clustering. [16]

The main benefits of using the K-means method to cluster data are:

- **Efficiency**. K-means clustering is renowned for its efficiency, characterized by its linear time complexity. This means that large datasets can be handled effectively. [16]
- **Simplicity**. One of the main benefits of K-means clustering lies in its simplicity; it is relatively straightforward to implement and enables the identification of unknown data groups within complex datasets. [16]
- **Flexibility**. K-means clustering is a flexible algorithm that can easily accommodate changes. For example, it can incorporate custom distance metrics and initialization methods. [16]

The main drawbacks of using the K-means method to cluster data are:

- **Sensitivity to outliers**. In K-means clustering, outliers have the potential to distort the cluster centroids, thereby resulting in inaccurate clustering outcomes. [16]
- **Dependence on initialization of centroids**. The initial positions of the centroids can have a significant impact on the final clustering outcomes. [16]
- **Inability to handle non-linearly separable data**. The K-means algorithm cannot cluster non-convex and non-linearly separable data since it is assumed that all clusters are spherical and possess the same variance. Non-linearly separable data is data that cannot be separated into the correct clusters with a linear line (see the right plot of Figure 3.9 for an example). [81] [16]

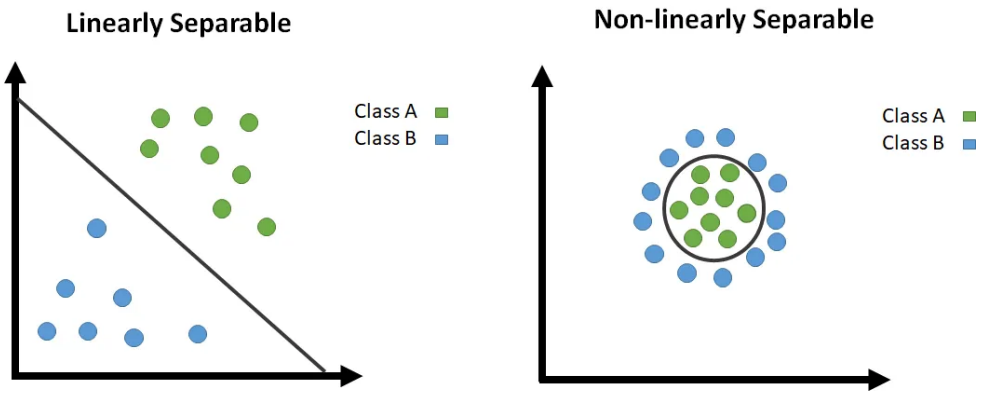


Figure 3.9: This figure shows examples of linearly and non-linearly separable data. [1]

Spectral clustering

Rather than directly clustering the data in the input space, spectral clustering involves constructing a similarity graph where nodes represent data points and edges symbolize similarities between the points. The algorithm then utilizes the spectral properties of the graph, specifically the eigenvalues and eigenvectors of the graph's Laplacian matrix, to project the data into a lower-dimensional space. In this transformed space, traditional clustering techniques, like K-means, can be applied more effectively. [81] Since spectral clustering relies on the connectivity of data points rather than their distances, unlike K-means, it can cluster non-convex and non-linearly separable data. Hence, spectral clustering was selected due to the high likelihood that the WAM and ARD datasets are non-linearly separable. Another reason for choosing this technique is that the datasets are relatively large and spectral clustering performs dimensionality reduction when nodes are mapped to a low-dimensional space. This technique will therefore require less memory and a shorter run time. [20]

The main benefits of using the spectral clustering method are:

- **Scalibility**. The algorithm can handle large data sets since the data is projected into a lower-dimensional space. [51]
- **Ability to handle non-linearly separable data**. As mentioned before, unlike the K-means method, this technique can cluster non-linearly separable data. [51]

The main drawbacks of using the spectral clustering method are:

- **Relatively slow**. Spectral clustering is a relatively slow algorithm compared to other clustering methods such as K-means due to the construction of a similarity graph. [21]
- **Less simple to explain**. The spectral clustering algorithm relies on the spectral properties of the graph which makes it less intuitive to explain compared to the K-means method. [21]
- **Dependence on initialization of centroids in K-means step**. Since the spectral clustering algorithm includes a K-means step, the final clustering outcomes depend on the initial positions of the centroids. [21]

How the K-means and spectral clustering algorithms exactly work and what practical details have to be considered will be explained in Chapter 4. The next three subsections outline existing approaches to handling mixed data for clustering in general, for K-means clustering, and for spectral clustering.

Handling mixed data for clustering in general

The majority of clustering algorithms can only handle data that is exclusively numerical or exclusively categorical. [46] Nevertheless, there exist various methods to enable them to handle mixed data:

- **Convert numerical values to categorical ones**. Numerical variables can be converted into categories through a process known as discretization. This involves dividing the numerical variable into N intervals, the values are then labeled as categories based on the interval in which they fall. [46]
  Downsides: Aside from the loss of information, the challenge with this approach lies in selecting the correct discretization method; in many cases, variables lack natural groupings. The uncertainty in choosing an appropriate discretization method for each variable presents a problem since this choice directly impacts the performance of clustering algorithms. [46]
- **Convert categorical values to numerical ones**. Similarly, categorical variables can be converted to numerical ones by one hot encoding; for each unique category, a new binary variable is created and added to the dataset. After this, each one-hot encoded variable is standardized. [46]
  Downsides: It is important to note that for some categories many new binary variables have to be generated (e.g. if there are twenty different car brands in the dataset, twenty variables are added). Therefore, when clustering the data, this method will require a substantial amount of memory and a relatively large run time. Furthermore, when one hot encoding ordinal variables, the order/structure of the values will be lost. [46]
- **Gower's distance**. Gower's distance is a similarity measure for two data points that contain both numeric and categorical variables. It employs distinct similarity measures for each data type: the Euclidean distance for numerical data, the Jaccard distance for categorical (nominal) data, and for ordinal data, the variables are initially sorted, followed by applying the Manhattan distance with an adjustment for ties. The resulting similarity scores for each data type are then combined to produce an overall similarity score between two data points. [46] Further details regarding the Euclidean distance and Gower's distance can be found in Chapter 2.
  Downsides: This method lacks flexibility in choosing different similarity measures. Moreover, it is not possible to assign weights to the similarity measures of the various data types. This poses a problem because, for instance, the Jaccard distance is equal to the number of matching categories between two data points divided by the total number of categories. As a result, it has a maximum value of 1, while the Euclidean distance lacks an upper bound. This leads to a dominance of the Euclidean distance in the overall similarity measure. A situation that could be resolved by allowing weights to be assigned to the similarity measures of the data types. [38]
- **Cluster Ensemble Based Mixed Data Clustering**. An overview of this method's algorithm is shown in Figure 3.10. First, the data is split into two sub-datasets: one that exclusively contains the numerical variables and one that exclusively contains the categorical variables. Next, existing clustering algorithms designed for the two different data types, are applied. For example, K-means is employed for the pure numerical dataset, and K-modes is used for the pure categorical set (how the K-modes algorithm works will be explained in "Handling mixed data for K-means"). The clustering results of the numerical and categorical data can be seen as categories and a categorical clustering algorithm is applied to these two categories to obtain the final clusters. [32]
  Downsides: A drawback of this method is the loss of information; by clustering the numerical and categorical variables separately, the relationships between numerical and categorical values are not taken into account in the clustering process. [32]

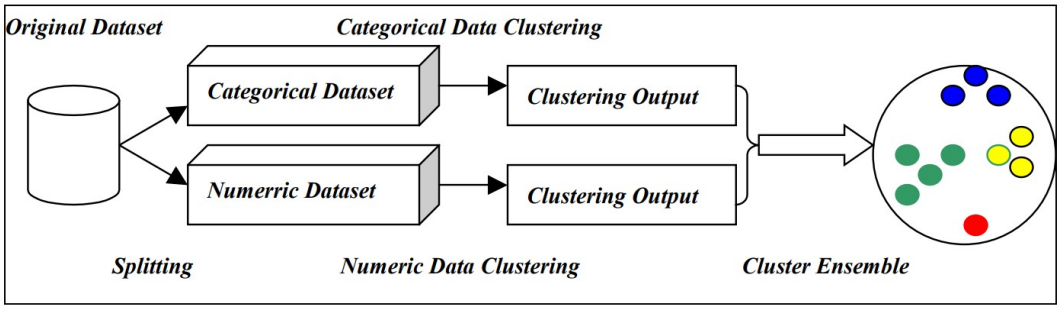


Figure 3.10: This figure shows an overview of the Cluster Ensemble Based Mixed Data Clustering algorithm. [32]

Handling mixed data for K-means

When the dataset contains categorical variables, two problems arise when applying K-means:

1. Since the K-means algorithm relies on the Euclidean distance to quantify the similarities between data points, computing these similarities is impossible when the dataset contains categorical variables. [46]
2. Calculating the centroids with the means of the points in the clusters is not possible when dealing with categorical data. For example, the average between the car brands Audi and BMW cannot be determined. [34]

In order to handle mixed data, the K-means clustering algorithm can be modified in the following ways:

- **K-modes**. K-modes is a clustering algorithm similar to K-means. However, it is designed to handle pure categorical data. [34] It solves the two previously mentioned problems as follows:

  1. Rather than computing distances between numerical values, it calculates the number of mismatches between categorical values. [61] This metric is referred to as the Hamming distance. For example, the Hamming distance between "012345" and "022546" is equal to 3. [69] Further details regarding the Hamming distance can be found in Chapter 2.
  2. Moreover, instead of calculating the centroids with the averages of the data points in the clusters, the modes are computed. [34] The mode represents the value that appears most frequently within the cluster. [31] Note that this means that the mode is not always unique. For instance, the mode of set $\{[a,b],[a,c],[c,b],[b,c]\}$ can be either $[a,b]$ or $[a,c]$. [34]

  Downsides: The K-modes algorithm is designed to *exclusively* handle categorical data. [34]
- **K-prototypes**. K-prototypes is a clustering algorithm that merges K-means and K-modes to handle datasets containing mixed data (i.e. both numerical and categorical data). [39] It solves the two previously mentioned problems in the following way:

  1. Similar to Gower's distance, this method employs distinct similarity measures for each data type. However, it also assigns weights to both types of features and calculates a weighted distance metric to determine cluster allocations. [61]
  2. The centroids are calculated by using the means for numerical attributes and the modes for categorical ones. [39]

Handling mixed data for spectral clustering

Since the spectral clustering algorithm relies on the Euclidean distance to compute the similarity matrix, obtaining this matrix is impossible when the dataset contains categorical variables. In order to handle mixed data, the spectral clustering algorithm can be modified as follows:

Modified spectral clustering. For each data type, a similarity matrix is constructed by using distinct similarity measures. Next, the rows of the matrices are scaled to be between 0 and 1. Similar to K-prototypes, the total similarity matrix is then calculated by taking the weighted sum of the similarity matrices. Note that, since the total similarity matrix exclusively consists of numerical values, K-means can still be applied for the clustering in the lower-dimensional space. [53]

3.4. Proposed solution

Out of all the approaches described in the previous section, the **K-prototypes** and **modified spectral clustering** algorithms are the only ones that do not present any immediate and obvious drawbacks. Therefore, these two methods are employed in this thesis. The full mathematical formulations and more detailed explanations of these techniques can be found in the next chapter. This chapter also discusses how ordinal data is taken into account for the two methods.

As explained in Section 3.2, in order to improve the current risk classification of the frequency models by applying clustering techniques to the datasets, the following matters should be taken into account: the resulting clusters should be significant with respect to the claim frequency, and the clusters should be explainable.
Significance. Since K-prototypes and the modified spectral clustering algorithms are unsupervised clustering techniques, the claim frequencies are not taken into account for the clustering. [17] Therefore, it is impossible to guarantee significance beforehand.
Explainability. To guarantee explainability, the next chapter will discuss how the clustering techniques work and what their limitations are in full detail. Furthermore, every parameter choice for the clustering will be explained and, if possible, a meaningful and understandable interpretation/description will be provided for each cluster.

4

Methodology

In this chapter, an overview of the methodology is provided. Section 4.1 describes the pre-processing steps that have to be executed prior to the application of the clustering techniques. Information on the clustering methods and observation reduction techniques of this project can be found in Sections 4.2 and 4.3 respectively. Lastly, Section 4.4 discusses how the performances of the clustering techniques are evaluated and Section 4.5 explains how the stability of the techniques (with respect to the time and number of observations) are assessed.

4.1. Pre-processing

Prior to the application of the clustering techniques, data preparation is essential. The process of this preparation is detailed in Subsection 4.1.1.
In Subsections 4.1.2 and 4.1.3, two groups of sample datasets are introduced: "Sample datasets A" and "Sample datasets B". Sample datasets A are used to gain insight into the performances of the K-prototypes and modified spectral clustering algorithms by evaluating these techniques across various 2D datasets. On the other hand, sample datasets B are employed to evaluate two observation reduction techniques.

4.1.1. Data preparation

As explained in Subsection 3.1.3, before the data connected to the license plates and zip codes can be clustered, the following matters need to be considered: P.O. box zip codes, (multi)collinearity, missing values, non-standardized data, and mixed data. For the ARD and WAM datasets, the process of data preparation can be divided into the following steps:

1. **Data extraction**. First, the relevant data is extracted from the ARD and WAM datasets. As explained in Subsection 3.1.3, this means that, after the claim frequencies are calculated with Formula (3.1), the most recent claim records are selected. The obtained datasets (one for each of the two coverages) are then split up into vehicle and geographical data for the clustering of license plates and zip codes respectively, resulting in four datasets in total: ARD vehicle data, WAM vehicle data, ARD geographical data, and WAM geographical data.
2. **Delete P.O. box zip codes**. Approximately five percent of the zip codes within the ARD and WAM geographical datasets belong to P.O. boxes. As explained in Subsection 3.1.3, using the P.O. box zip codes alongside the zip codes of the companies/houses can negatively impact the clustering. Therefore, the P.O. box zip codes are deleted from the geographical datasets.
3. **Delete redundant variables**. Next, redundant features are deleted. For instance, "PROV" and "provincie" are removed as "ind_provincie" already denotes the province corresponding to the zip code. Similarly, within the vehicle datasets, the "class of car weight" variable is eliminated, given the existence of the "car weight" variable.
4. **Delete features that only have one unique value**. Variables that only display one unique value across all license plates will be deleted from the vehicle datasets as these features do not

influence the clustering results. For the same reason, variables that only showcase one unique value across all zip codes are removed from the geographical datasets.

5. **Handle missing values**. For various policyholders, some of the values for the features are unknown. These missing values are handled in the following way:

   - Missing values within categorical features are addressed by creating a new category labeled as "Unknown".
   - For missing values within numerical features, an iterative imputer is employed. However, in order to calculate the error of the imputation, 5% of the known observations are deleted for each variable beforehand. Here the known observations refer to all observations that are not missing.
     For the imputation of missing values, the "Iterative imputer" package of "sklearn" is used. This package imputes missing values by modeling each feature with missing values as a function of other features. [64] For example, as will be explained in step 6, the lower and upper bounds for the number of gears of the car are highly correlated. So, the missing values of the lower bound variable can be imputed by using the correlation with the upper bound variable and vice versa.
     The imputed values of the deleted known observations are compared to the actual values and the relative errors that are obtained are close to 10% for each dataset. Lastly, the iterative imputer is employed on the complete dataset (without the elimination of known observations). The outcome is used in the following steps.

6. **Check for (multi)collinearity**. As explained in Definition 3.1.1, multicollinearity occurs when two or more variables are highly linearly correlated. Incorporating multicollinear features for the clustering, rather than utilizing a singular one, does not yield additional information. Furthermore, employing such features puts more weight on their influence, thereby affecting the clustering negatively. For the data preparation, the multicollinearity is assessed among numerical features, categorical features, and numerical-categorical feature pairs by using the following approaches:

   - To check the multicollinearity among numerical variables, the correlation $r_{x,y} \in [-1, 1]$ is computed between every pair of features $x, y \in \mathbb{R}^N$ with Equation 4.1.

$$r_{x,y} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}} \tag{4.1}$$

     Here $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ and $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ are the means of features x and y respectively and N is equal to the number of observations. [14]
     If the correlation is greater than 0.95, one of the two features is removed from the dataset (see Subsection 3.1.3 for the reasoning behind this cut-off). For example, the lower and upper bounds for the number of gears of the car possess a correlation of 0.9999909. Therefore, the lower bound variable is deleted.

   - To check the multicollinearity among categorical variables, the Chi-squared test is combined with Cramer's V. [71] First, Pearson's Chi-squared value $\chi^2 \in \mathbb{R}_{\geq 0}$ is calculated with the following formula:

$$\chi^2(x, y) = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{4.2}$$

     Here the i's are the possible combinations of values belonging to the categorical features x and y, $O_i \in \mathbb{N} \cup \{0\}$ is the number of observations that have combination i, $E_i \in \mathbb{N} \cup \{0\}$ is the expected number of observations that have combination i, and $n \in \mathbb{N}$ is the total number of observations (i.e. the product of the number of unique values of x and of y). [40] [68]
     Figure 4.1 shows an example of the O_i's for the categorical features "Car brand" and "Fuel type". Such a table is referred to as a contingency table.

Car Brand	Fuel Type			Total
	Gas	Electric	Hybrid	
Audi	40	20	10	**70**
Mini	30	20	0	**50**
Fiat	50	10	10	**70**
Total	**120**	**50**	**20**	**190**

Figure 4.1: This figure shows an example of the O_i's for the categorical features "Car brand" and "Fuel type". Such a table is referred to as a contingency table.

Here $n = 9$ and, $E_1 = 190 \cdot \mathbb{P}(\text{Fuel Type} = \text{Gas}) \cdot \mathbb{P}(\text{Car Brand} = \text{Audi}) = 190 \cdot 120/190 \cdot 70/190 = 120 \cdot 70/190 \approx 44.211$. In the same manner, the other E_i's are computed. Lastly, χ^2 is calculated with Formula 4.2.

The null and alternative hypotheses of the Chi-squared test are as follows:
H_0: There is no relationship between categorical variables x and y.
H_1: There is a relationship between categorical variables x and y.
The p-value of the test is calculated by examining the right tail probability of χ^2 under the Chi-squared distribution with degrees of freedom $k \in \mathbb{N} \cup \{0\}$. Note that in the example of Figure 4.1, $k = (3-1) \cdot (3-1) = 4$. A p-value less than 0.05 is considered to be statistically significant, in which case the null hypothesis is rejected. So, if the p-value of x and y is less than 0.05, then there is a relationship between x and y. [40]

The p-values are calculated for all possible pairs of categorical variables in the dataset. If H_0 is rejected for features x and y, Cramer's $V \in [0, 1]$ is computed to obtain the correlation between the variables in the following way:

$$V(x, y) = \sqrt{\frac{\chi^2(x, y)/n}{\min\{k-1, r-1\}}}$$

Where $n \in \mathbb{N}$ is the number of observations, and $k, r \in \mathbb{N}$ are the number of unique values for x and y respectively. [71] Note that both correlation equations $r_{x,y}$ and $V(x, y)$ are symmetric (i.e. $r_{x,y} = r_{y,x}$ and $V(x, y) = V(y, x)$).

It is important to highlight that, for example, the "Car brand" variable has a correlation close to 1 with the "Car's country of origin" variable (e.g. BMW's always have Germany as their country of origin). Nonetheless, removing the "Car's country of origin" variable from the dataset could lead to overlooking certain relationships in the clustering (e.g. potentially missing the presence of a cluster exclusively composed of German cars). So, in contrast to the procedure applied to numerical variables, if the correlation V exceeds 0.95, none of the features are removed from the dataset. However, these correlations can offer explanations for the clustering outcomes.

- The multicollinearity among numerical-categorical feature pairs is not checked due to the complexity of computing their correlations. [60] Furthermore, if such a correlation were to surpass 0.95, no features could be deleted for the same reason as explained earlier for categorical variables. Lastly, in that case, it would be unclear whether the numerical or the categorical variable should be removed from the dataset.

7. **Standardize the data.** The variables of the datasets differ in their ranges (see Subsection 3.1.3 for an example). Since most clustering techniques (including K-means) are based on the distance between data points, a clustering can be completely dominated by a variable that possesses a wide range of values. Therefore, it is crucial to standardize each feature x prior to the clustering by using the following formula:

$$x_{standardized} = \frac{x - \mu}{\sigma}$$

Here, $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ is the mean of feature x and $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$ is the standard deviation of x. N is equal to the number of observations. [76]
Note that only numerical features are standardized since the categorical variables do not differ significantly in their ranges.

8. **Determine significance of the features.** As explained in Section 3.2.1, the resulting clusters are required to have a significant effect on the average claim frequency. Therefore, features that do not affect the claim frequency, can be removed from the dataset prior to the clustering. The significance of each variable is assessed by applying regularized linear regression models.
Linear regression is an algorithm for regression that assumes a linear relationship between inputs and the target variable. Regularized linear regression is an extension of linear regression that incorporates penalties into the loss function during training, thereby promoting simpler models characterized by smaller coefficient values. [10]
The significance of each variable in the dataset is assessed by employing regularized linear regression models (here the features serve as inputs and the claim frequency as the target variable). If the coefficient of a variable is approximately zero according to the regression model, the feature is insignificant with respect to the claim frequency and can thus be removed from the dataset. [10]
The following three regularized linear regression models are used:

   - *Lasso.* Lasso includes an L1-norm penalty which leads to the reduction of coefficients for input variables that contribute minimally to the prediction task. The penalty allows coefficient values to be equal to zero, in which case the variables are regarded as insignificant. [10]
   - *Ridge.* Ridge incorporates an L2-norm penalty that also shrinks the coefficients for the input variables that contribute minimally to the prediction task. However, in contrast to Lasso regression, coefficient values cannot be equal to zero. Therefore, variables are insignificant if their coefficients are *approximately* zero. [11]
   - *Elastic net.* Elastic net regression is a combination of Ridge and Lasso regression; it includes both the L1- and L2-norm penalty functions. When the coefficient of a variable is zero or approximately zero, the feature is regarded as insignificant. [12]

   Only when a feature is insignificant according to all three models, it is deleted from the dataset.

4.1.2. Sample datasets A

Sample datasets A are used to gain insight into the performances of the K-prototypes and modified spectral clustering algorithms by evaluating these techniques across various 2D datasets. These datasets are shown in Figure 4.2. The first two plots show non-linearly separable data; the first plot illustrates data distributed in a manner resembling two noisy circles sharing the same center, while the second plot portrays data distributed like two noisy moons. As explained in Section 3.3, it is expected that the spectral clustering algorithm clusters these two non-linearly separable datasets more accurately than the K-prototypes method. In the third plot, three sample datasets are depicted, each distributed according to the Gaussian distribution with varying variances. The fourth and fifth plot show the same matter, but according to the anisotropic, and Gaussian distributions (same variances) respectively. Next, the sixth plot displays three datasets: one following the Gaussian distribution and two distributed in a manner resembling two noisy moons. The data in this plot is an example of data that is “on different scales”, i.e. the distances between data points are different in different regions of the space. [51] The data in the final plot is homogeneously distributed and an example of a ‘null’ situation: there is no good clustering. Note that the sixth plot shows 1000 data points while the others display 500.

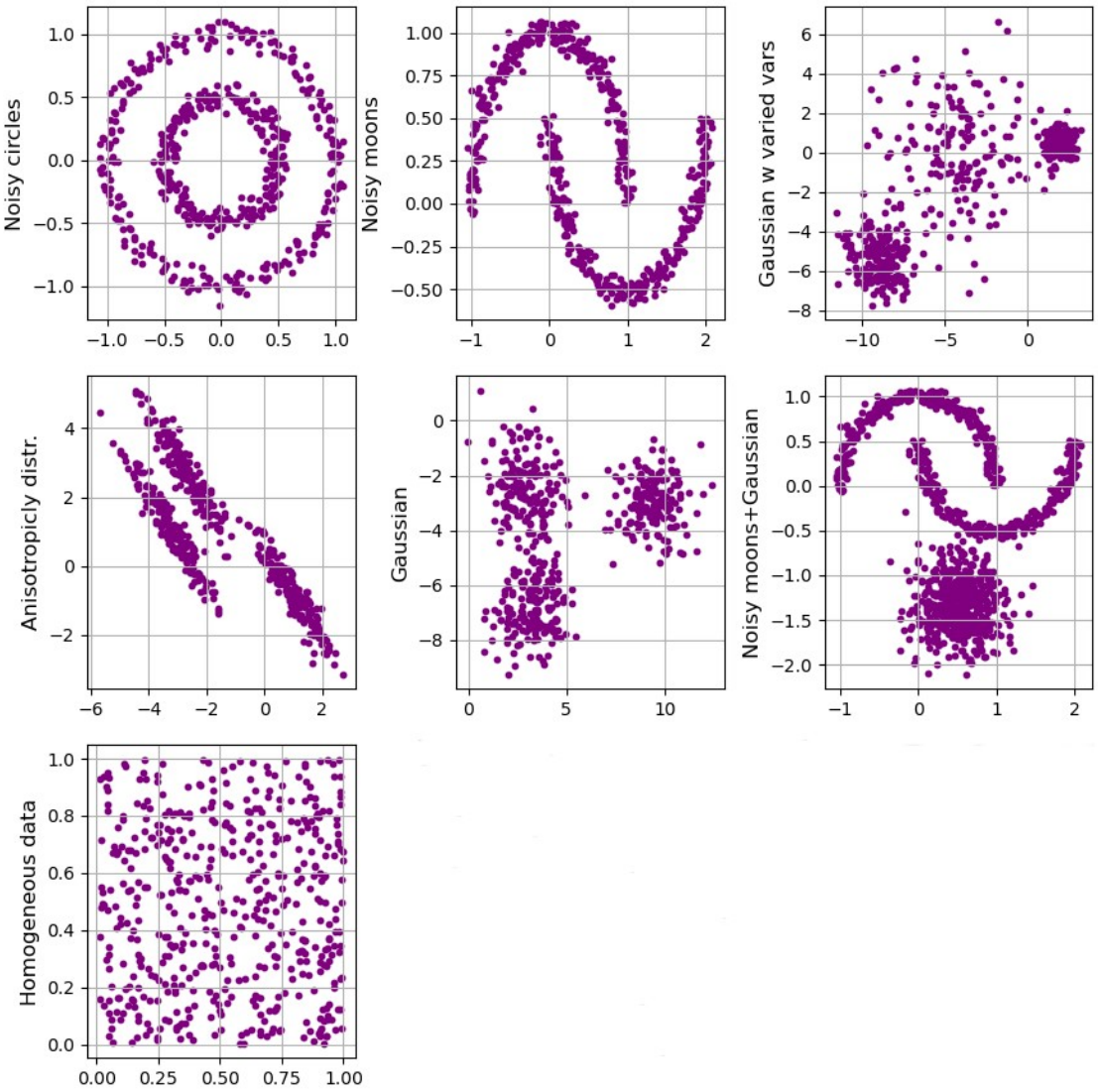


Figure 4.2: This figure shows the plots of sample datasets A. From left-to-right and top-to-bottom: noisy circles, noisy moons, Gaussian distributions with varied variances, anisotropic distributions, Gaussian distributions, noisy moons and a Gaussian distribution, and homogeneous data.

4.1.3. Sample datasets B

Sample datasets B are employed to evaluate the two observation reduction techniques of Section 4.3. Three datasets are created: one consisting of a small number of clusters (namely 3), another with an intermediate number of clusters (namely 7), and a third with a large number of clusters (namely 10). Each of these clusters is normally distributed with a random standard deviation between 0.75 and 1. Moreover, each dataset comprises 400 data points with 400 dimensions.

4.2. Clustering

As explained in Section 3.3, in this thesis, the focus lies on two unsupervised clustering techniques: K-means (centroid-based) and spectral clustering (connectivity-based). Subsections 4.2.1 and 4.2.2 discuss these two techniques respectively; the algorithms are explained and practical details are provided. Moreover, the modified versions of the algorithms, that are able to handle mixed data (as explained in Section 3.3), will be introduced.

4.2.1. K-means clustering

Standard algorithm

K-means clustering aims to minimize the sum of squared distances between each data point $x_i \in \mathbb{R}^m$ (with m dimensions) and the centroid $c_k \in \mathbb{R}^m$ of its allocated cluster C_k. [61] This objective can be mathematically formulated as:

$$\min_{C_1,\ldots,C_K,\ c_1,\ldots,c_K} \sum_{k=1}^{K} \sum_{i \in C_k} d_E(x_i, c_k)^2. \tag{4.3}$$

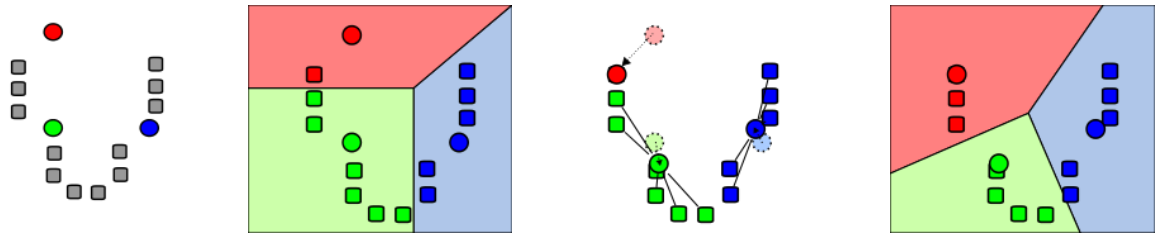
Here $K \in [2, n]$ is equal to the total number of clusters and d_E is the Euclidean distance as defined in Chapter 2. Furthermore, $\sum_{k=1}^{K} \sum_{i \in C_k} d_E(x_i, c_k)^2$ is referred to as the error sum of squares (i.e. ESS). [61]
In order to accomplish the objective of Equation 4.3, Algorithm 1 is followed.

Algorithm 1 The K-means algorithm

Input: $x_1, \ldots, x_n$ data points to be clustered
K number of clusters

Initialize centroids $c_1, \ldots, c_K$
while ESS improves **do**
  Assign x_i to cluster $k = \text{argmin}_j d_E(x_i, c_j)^2$
  for $k = 1, \ldots, K$ **do**
    Update $c_k = \frac{1}{|C_k|} \sum_{i \in C_k} x_i$
  end for
end while

Figure 4.3 shows an example of the demonstration of the standard K-means algorithm as described in Algorithm 1 where $K = 3$. In Figure 4.3a, three initial centroids are randomly generated within the domain of the data. Next, Figure 4.3b displays the creation of the three clusters by assigning each data point to the cluster of which the centroid is closest. This means that data point x_i is assigned to cluster k for which $d_E(x_i, c_k)^2$ is minimized. In Figure 4.3c the centroids are recalculated by averaging all data points within each cluster. The steps in Figure 4.3b and 4.3c are repeated until convergence of the ESS is achieved. Lastly, Figure 4.3d shows the final result of the K-means clustering algorithm. [77]



(a) **(b)** **(c)** **(d)**

Figure 4.3: This figure shows an example of the demonstration of the standard K-means algorithm: **(a)** Three initial centroids, denoted as $K = 3$, are randomly generated within the domain of the data, **(b)** K clusters are formed by assigning each data point to the cluster of which the centroid is closest, **(c)** The centroids are recalculated by averaging all data points within each cluster, **(d)** Steps b and c are repeated until convergence of the ESS is achieved. [77]

Prior to the implementation of the K-means algorithm, two practical details have to be considered:

- **Initialization of the centroids**. The K initial centroids can be randomly selected from the dataset, a method prone to volatility, as the resulting clusters heavily rely on these randomly selected centroids. [52]
  Alternatively, the centroids can be initialized using the K-means++ algorithm. This method involves choosing one centroid uniformly from the data points, and then selecting all other centroids

from the dataset such that the probability of selecting a point as a centroid is directly proportional to its distance to the nearest previously chosen centroid. [52]

- **Determine the number of clusters K**. The elbow plot can be used to determine the number of clusters K for K-means. To construct this plot, the K-means algorithm is executed for $K = 2$ up to $K = K_{\max}$, and the error sum of squares (ESS) values are plotted against K. As K increases, the ESS decreases since the distances from each data point to the nearest centroids decrease with the availability of additional centroids. However, the marginal benefit of adding clusters drops for each new cluster, resulting in an elbow-shaped pattern in the plot. $K_{\max}$ is selected to be sufficiently large to ensure that the elbow shape can be observed. [25]
  An example of an elbow plot is shown in Figure 4.4. Around $K = 3$, the ESS ceases to decrease significantly. Therefore, in this example, three clusters should be created. [73]

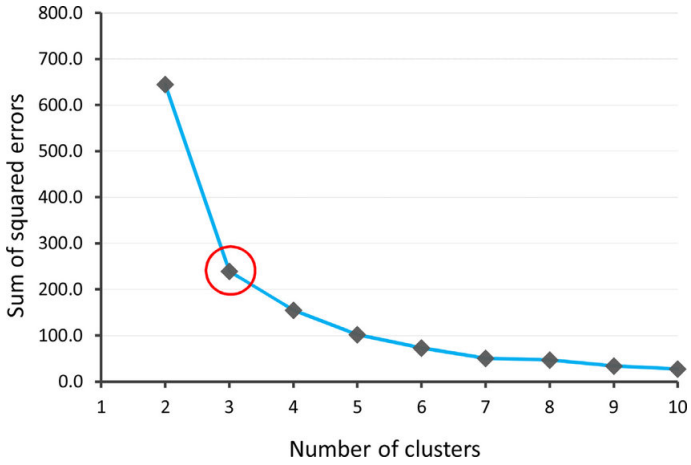


Figure 4.4: This figure shows an example of an elbow plot. In this case, the ESS ceases to decrease significantly around $K = 3$. [73]

K-modes

K-modes is a clustering algorithm similar to K-means, designed specifically for handling categorical data. Rather than relying on the Euclidean measure d_E to calculate the distance between data points, the Hamming distance d_H (for nominal data) or Gower's distance d_G (for ordinal data) is used. [34] See Chapter 2 for detailed explanations of these distance measures.
Furthermore, instead of calculating the centroids based on the averages of the data points in the clusters, the modes (i.e. the values that appear most frequently within the clusters) are computed. With these two modifications, Algorithm 1 can be used for the implementation of K-modes. Additionally, the determination of the number of clusters can be carried out by using the same method as employed for K-means. [34]

For K-modes, the centroids can be initialized with one of the following two methods:

- **The Huang method** initializes centroids by considering the frequencies of categorical attributes. First, a centroid is randomly chosen from the data points. Next, the method selects data points for additional centroids, ensuring they differ from the already chosen centroids by at least a specified threshold in terms of their attribute frequencies. [39]
- **The Cao method** aims to select initial centroids that are well separated from each other and is similar to the K-means++ method, but designed for handling categorical data. It evaluates the overall data distribution and computes a density measure for each point considering both the number of points that are close to it (in terms of a defined distance) and the categorical attributes. Points with higher density measures are more likely to be chosen as initial centroids. This way, the method aims to identify centroids that accurately reflect the density structure of the data. Despite being slower than the Huang method, the Cao method is typically preferred for its robustness (since it considers the distribution of the data). [39]

K-prototypes

K-prototypes merges K-means and K-modes to handle datasets containing mixed data (i.e. numerical and categorical (nominal) data). The following distance measure $d(p,q) \in \mathbb{R}_{\geq 0}$ is used:

$$d(p,q) = (1-\alpha) \cdot d_E(p_i, q_i) + \alpha \cdot d_H(p_j, q_j)$$

Here $\alpha \in [0,1]$ is the weight assigned to the categorical (nominal) distance measure. p_i and q_i are the numerical parts of p and q respectively and p_j and q_j the categorical (nominal) parts. For example, for p =(2,3,"Porsche") and q =(4,5,"Mercedes"), it holds that $p_i = (2,3)$, $q_i = (4,5)$, p_j =("Porsche") and q_j =("Mercedes"). [39]
Moreover, the centroids are calculated by using the means for numerical attributes and the modes for categorical ones. With these two adjustments, Algorithm 1 can be used for the implementation of K-prototypes. The number of clusters is determined in the same manner as for the K-means method and the initialization of the centroids follows the same procedures as for K-modes. [39]

Since the datasets of this thesis contain mixed data (i.e. numerical, categorical (nominal), and categorical (ordinal)), a modified version of the K-prototypes algorithm is implemented. The following distance measure $d(p,q) \in \mathbb{R}_{\geq 0}$ is used:

$$d(p,q) = (1-\alpha-\gamma) \cdot d_E(p_i, q_i) + \alpha \cdot d_H(p_j, q_j) + \gamma \cdot d_G(p_l, q_l)$$

Here $\gamma \in [0,1]$ is the weight assigned to the categorical (ordinal) distance measure, and p_l and q_l are the categorical (ordinal) parts of p and q respectively.
For this thesis, α en γ are set equal to the fractions of categorical (nominal) and categorical (ordinal) features with respect to the total number of variables. Furthermore, the number of clusters is determined by using the elbow plot, and the centroids are initialized with the Cao method (since it is more robust than the Huang method).

4.2.2. Spectral clustering

Standard algorithm

As explained in Section 3.3, rather than directly clustering the data in the input space, spectral clustering involves constructing a similarity graph where nodes represent data points and edges symbolize similarities between the points. The algorithm then utilizes the spectral properties of the graph, specifically the eigenvalues and eigenvectors of the graph's Laplacian matrix, to project the data into a lower-dimensional space. In this transformed space, traditional clustering techniques, like K-means, can be applied more effectively. [81]

$G(V,E)$ is assumed to be an undirected and weighted graph with weighted adjacency matrix $W \in \mathbb{R}^{n \times n}$ where $w_{i,j} = w_{j,i} \geq 0$. The **unnormalized Laplacian matrix** $L \in \mathbb{R}^{n \times n}$ of the graph is given by:

$$L = D - W$$

Where $D \in \mathbb{R}^{n \times n}$ is the degree matrix as defined in Section 2.2. [51]
The unnormalized Laplacian matrix is symmetric and positive semi-definite. Therefore, all its eigenvalues are real and non-negative. [51]

Algorithm 2 describes the unnormalized spectral clustering algorithm.

Algorithm 2 The unnormalized spectral clustering algorithm

Input: $x_1, \ldots, x_n$ data points to be clustered
$K \in [1,n]$ number of clusters

Compute the similarity matrix S and the unnormalized Laplacian $L = D - W$
Construct a matrix H whose columns are the eigenvectors corresponding to the K minimal eigenvalues of L.
Use K-means to cluster the rows of H into $C_1, \ldots, C_K$

In some scenarios, the unnormalized variant of the spectral clustering algorithm does not produce the most desired results because the structure of the graph is dominated by a few nodes with the largest degree $D_{i,i}$. In that case, the normalized version of spectral clustering is applied. [51] The **normalized Laplacian matrix** $\bar{L} \in \mathbb{R}^{n \times n}$ of the graph is given by:

$$\bar{L} = D^{-1/2} L D^{-1/2} = I - D^{-1/2} W D^{-1/2}$$

The normalized Laplacian matrix is also symmetric and positive semi-definite and therefore its eigenvalues are real and non-negative. [51] Algorithm 3 describes the normalized spectral clustering algorithm.

Algorithm 3 The normalized spectral clustering algorithm

Input: $x_1, \ldots, x_n$ data points to be clustered
$K \in [1, n]$ number of clusters

Compute the similarity matrix S and the normalized Laplacian $\bar{L} = I - D^{-1/2} W D^{-1/2}$
Construct a matrix H whose columns are the eigenvectors corresponding to the K minimal eigenvalues of $\bar{L}$.
Use K-means to cluster the rows of H into $C_1, \ldots, C_K$

Prior to the implementation of the spectral clustering algorithm, five practical details have to be considered:

- **Determine the number of clusters K**. In contrast to K-means/K-prototypes, the ESS can rise when K increases because more eigenvectors are incorporated into H in that case. Consequently, clustering the rows into K clusters requires consideration of more features, specifically K features which can result in a higher ESS. Therefore, instead of the elbow plot, for spectral clustering the eigengap heuristic is used to determine the number of clusters K. This method involves selecting K such that $\lambda_1 \ldots \lambda_K$ are very small, but λ_{K+1} is relatively large. For example, in Figure 4.5, the number of clusters is equal to four since there is a large gap between λ_4 en λ_5. [51]

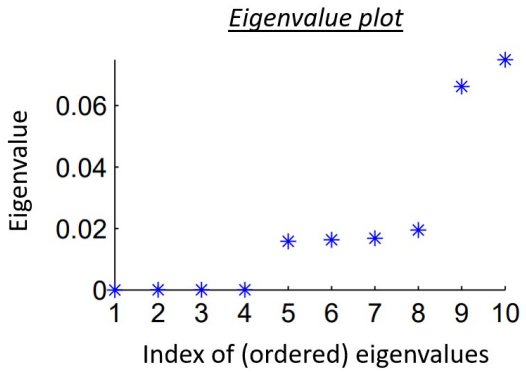


Figure 4.5: This figure shows an example of an eigenvalue plot. In this case, the eigengap is visible after $K = 4$. [51]

- **Type of similarity graph**. Out of the similarity graphs that were described in Section 2.2, in this thesis, the k-nearest neighbor graph is applied to obtain the weighted adjacency matrix W out of similarity matrix S. This decision stems from the graph's ability to connect points "on different scales", its simplicity in implementation, its tendency to yield a sparse adjacency matrix W, and its resilience against inappropriate parameter choices compared to other similarity graph types. [51] For this thesis, the number of nearest neighbors $k \in \mathbb{N}$ in the graph is set equal to $\lfloor n^{1/2} \rfloor$ (as suggested by [50]) where $n \in \mathbb{N}$ is the number of data points and $\lfloor \cdot \rfloor$ denotes the floor function.
- **Choice of similarity function**. The local neighborhoods created by the similarity function have to be "meaningful". This means that the points that are considered to be "very similar" by the similarity function are also closely related within the context of the data's application. [51] Which similarity function is employed for this thesis, is explained under "Modified spectral clustering".

- **Unnormalized or normalized Laplacian?** The unnormalized Laplacian only minimizes between-cluster similarity while the normalized variant also maximizes within-cluster similarity. Furthermore, the unnormalized Laplacian is less suitable for high-dimensional problems and poses consistency issues regarding the resulting eigenvectors (utilized in H) when the parameters in the similarity function are altered. [51] [15] Therefore, the normalized Laplacian will be used for the spectral clustering of this thesis.
- **Method for computing the eigenvectors**. Since the k-nearest neighbor graph is used, the Laplacian matrix will be sparse. [51] Therefore, the "scipy.sparse.linalg" package can be utilized; the (sparse) LU decomposition enables the computation of the eigenvalues and eigenvectors.

Modified spectral clustering

The spectral clustering method is modified to handle mixed data. To do so, for each data type, a similarity matrix is constructed by using distinct similarity measures; $1 - d_E$ for numerical features, $1 - d_H$ for categorical (nominal) ones, and $1 - d_G$ for categorical (ordinal) variables. Next, the rows of the matrices are scaled to be between 0 and 1. The total similarity matrix S is then calculated by taking the weighted sum of the three similarity matrices S_E, S_H, and S_G. That is;

$$S = (1 - \alpha - \gamma) \cdot S_E + \alpha \cdot S_H + \gamma \cdot S_G$$

Here $\alpha \in [0, 1]$ and $\gamma \in [0, 1]$ are the weights assigned to the categorical (nominal) and categorical (ordinal) similarity matrices respectively. Similar to the parameters in the K-prototypes distance measure, these parameters are set equal to the fractions of the categorical (nominal) and categorical (ordinal) features with respect to the total number of variables.
Note that, since the total similarity matrix exclusively consists of numerical values, K-means can be applied for the clustering in the lower-dimensional space. [53]

4.3. Observation reduction techniques for spectral clustering

The weighted adjacency matrix W (utilized in the spectral clustering algorithm) denotes the degree or weight of adjacency between two vertices. Thus, for a dataset containing n data points, W is an $n \times n$ matrix. However, in datasets with a large number of observations—such as those examined in this thesis, each comprising over 25 thousand rows—storing such a matrix requires tens of gigabytes and proves inefficient for computations. [35] To address this challenge, two distinct observation reduction techniques are employed; the random removal technique (discussed in Subsection 4.3.1) and the Ultra-Scalable Spectral Clustering ($U\text{-}SPEC$) algorithm (introduced in Subsection 4.3.2). Lastly, Subsection 4.3.3 explains some of the practical considerations that are necessary when applying observation reduction techniques and Subsection 4.3.4 provides the updated versions of the normalized spectral clustering algorithm (i.e. Algorithm 3) according to the two observation reduction techniques.

4.3.1. Random removal

For the random removal method, observations are randomly deleted from the dataset, after which spectral clustering is applied to the remaining data. Every omitted data point can then be assigned to one of the resulting clusters based on the minimal squared distance to the cluster's centroid. This process thus involves a combination of spectral clustering and K-prototypes.

The main benefits of reducing the observations through random removal lie in the method's efficiency and simplicity (compared to $U\text{-}SPEC$). Nevertheless, this technique tends to lack robustness; the resulting clusters are often unstable since the output depends on the quality of the remaining data points (i.e. those that were not deleted). [35]

4.3.2. Ultra-Scalable Spectral Clustering (U-SPEC)

The Ultra-Scalable Spectral Clustering ($U\text{-}SPEC$) technique was developed in 2020 and consists of the following three phases:

- **Phase 1: Hybrid representative selection**. In the first phase, a hybrid representative selection strategy is applied to choose a subset of data points (i.e. representatives) for the clustering. This

strategy seeks to find a balance between the efficiency of random selection (that was described in Subsection 4.3.1) and the effectiveness of a K-prototypes-based selection. [35]

- **Phase 2: Approximation of k-nearest representatives**. In the second phase, a coarse-to-fine method is implemented to effectively approximate the k-nearest representatives for each data point. Furthermore, a sparse adjacency sub-matrix is created for the n data points and the p representatives. [35]
- **Phase 3: Bipartite graph partitioning**. Lastly, in the third phase, the sub-matrix of phase 2 is treated as a bipartite graph. Such graphs feature a vertex set V that can be divided into two non-empty subsets A and B (i.e. $A \cup B = V$ and $A \cap B = \emptyset$), where each edge connects one vertex from A to one from B. [23] The bipartite graph can be partitioned to acquire the final spectral clustering result. [35]

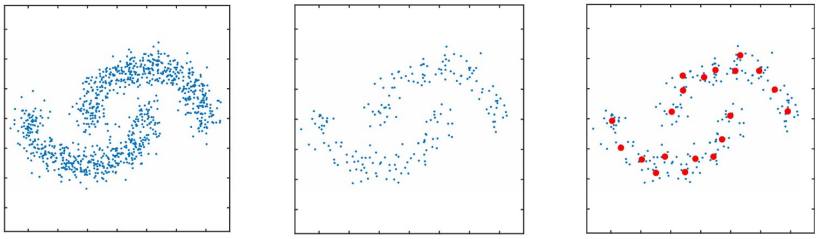
These three phases of U-$SPEC$ will be further explained in the following subsections.

Phase 1: Hybrid representative selection

The hybrid representative selection strategy is shown in Figure 4.6 (here Figure 4.6a displays a sample dataset of two noisy moons). First, a set of $p' \in \mathbb{N}$ candidate representatives is randomly sampled such that $p < p' \ll n$ (shown in Figure 4.6b). Then, on the p' candidates, the K-prototypes method is applied to acquire $p \in \mathbb{N}$ clusters (shown in Figure 4.6c). The p cluster centers are used as the set of representatives. This set is denoted as:

$$\mathcal{R} = \{r_1, r_2, \cdots, r_p\}$$

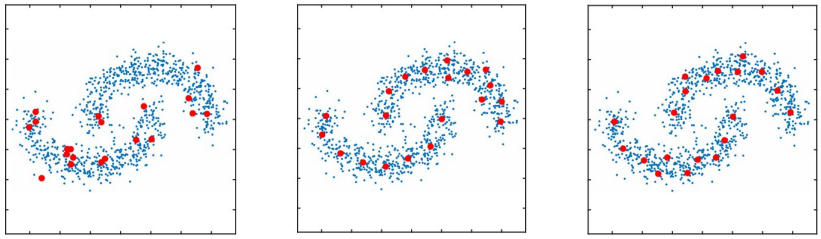
where r_i is the i-th representative in $\mathcal{R}$. [35]
Note that the number of candidates p' should be substantially larger than p to supply enough candidates while still keeping p' significantly smaller than n for large datasets. For this thesis, $p' = 10p$ (as suggested by [35]).



(a) (b) (c)

Figure 4.6: This figure shows the hybrid representative selection strategy: **(a)** A sample dataset of two noisy moons, **(b)** A set of p' candidate representatives is randomly sampled such that $p < p' \ll n$, **(c)** On the p' candidates, the K-prototypes method is applied to acquire p clusters. The centers of these clusters are used as the set of representatives. [35]

Figure 4.7 shows that the set of representatives generated by the hybrid selection (Figure 4.7c) more accurately reflects the data distribution compared to the random selection (Figure 4.7a) while requiring significantly less computational cost than the K-prototypes-based selection approach (Figure 4.7b).



(a) (b) (c)

Figure 4.7: This figure shows a comparison of the representatives (in red) of the sample dataset produced by the following three methods: **(a)** Random selection, **(b)** K-prototypes-based selection, **(c)** Hybrid selection (i.e. the method of the first phase). [35]

Phase 2: Approximation of k-nearest representatives

The core concept behind the k-nearest representative approximation is to first find the nearest region, then the nearest representative (denoted as r_l) within that region, and finally the k-nearest representatives in the neighborhood of r_l. This k-nearest representatives approximation is shown in Figure 4.8. To efficiently implement the procedure, the following two pre-processing steps are required:

- *Pre-step 1.* With K-prototypes, the set of representatives is grouped into $z_1 \in \mathbb{N}$ clusters (referred to as rep-clusters $\mathcal{RC} = \{rc_1, rc_2, \cdots, rc_{z_1}\}$) (see Figure 4.8b). [35] Here $z_1 \ll p$ and, for this thesis, $z_1 = \lfloor p^{1/2} \rfloor$ (as suggested by [35]).
- *Pre-step 2.* For each representative in $\mathcal{R}$, its k'-nearest neighbors are identified and stored. Here $k' = 10k$ and $k = \lfloor p^{1/2} \rfloor$ as explained in Subsection 4.2.2. [35]

For each data point x_i in the (complete) dataset, the k-nearest representatives are identified according to the following three steps:

- *Step 1.* Find the nearest rep-cluster to x_i, denoted as rc_j (see Figures 4.8c and 4.8d). [35]
- *Step 2.* Find the nearest representative to x_i inside the rep-cluster rc_j, denoted as r_l (see Figures 4.8e and 4.8f). [35]
- *Step 3.* Out of r_l and its k'-nearest neighbors, find the k-nearest representatives to x_i (see Figures 4.8g and 4.8h). [35]

Lastly, after obtaining the k-nearest representatives for each data point, a sparse $n \times p$ adjacency sub-matrix B can be created. This matrix comprises k non-zero entries for each row, resulting in a total of $n \cdot k$ non-zero entries. [35]

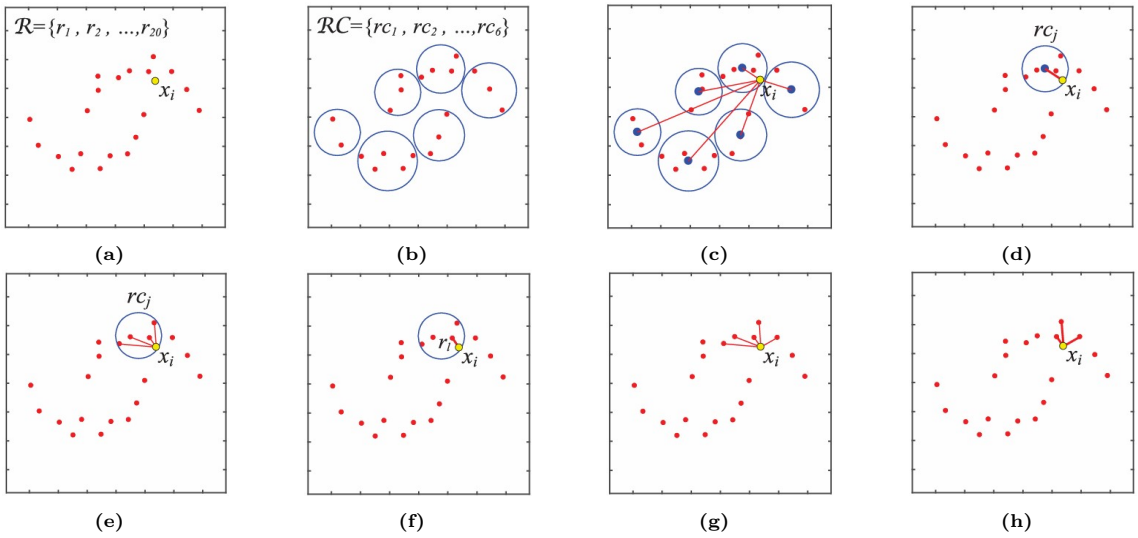


Figure 4.8: This figure shows the procedure of the k-nearest representatives approximation: **(a)** The representative set $\mathcal{R}$ and a data point x_i (from the complete dataset), **(b)** The representatives partitioned into $z_1 = 6$ rep-clusters with K-prototypes, **(c)** The distances between x_i and the rep-cluster centers are computed, **(d)** The nearest rep-cluster rc_j is selected, **(e)** The distances between x_i and all the representatives in rc_j are computed, **(f)** The nearest $r_l \in rc_j$ is selected, **(g)** The distances between x_i and the representatives in the k'-nearest neighborhood of r_l are computed, **(h)** The approximate k-nearest representatives of x_i are obtained. [35]

Phase 3: Bipartite graph partitioning

The n objects in the (complete) dataset $\mathcal{X}$ and the p representatives in the set $\mathcal{R}$ are part of the bipartite graph $G = \{\mathcal{X}, \mathcal{R}, B\}$ where $\mathcal{X} \cup \mathcal{R}$ is the node set and B is the adjacency sub-matrix that reflects the relationship between $\mathcal{X}$ and $\mathcal{R}$. [35] G is a bipartite graph since each edge connects one vertex from $\mathcal{X}$ to one from $\mathcal{R}$ if they are adjacent and because $\mathcal{X} \cup \mathcal{R} = V$ and $\mathcal{X} \cap \mathcal{R} = \emptyset$ (as the representatives are the centroids of the K-prototypes-based selection and are unlikely to coincide with actual data points). [23]

G can also be viewed as a general graph with $n+p$ nodes and the following $(n+p)\times(n+p)$ adjacency matrix:

$$W = \begin{bmatrix} 0 & B^T \\ B & 0 \end{bmatrix}$$

Given this adjacency matrix W, the normalized Laplacian $\bar{L}$ can be constructed and thus spectral clustering can be performed according to Algorithm 3. However, [35] proposes an alternative method that utilizes a smaller graph $G_{\mathcal{R}}$, which results in a less computationally intensive solution for the eigenproblem of the Laplacian. Here, the graph $G_{\mathcal{R}} = \{\mathcal{R}, W_{\mathcal{R}}\}$ (with p nodes) has node set $\mathcal{R}$ and adjacency matrix $W_{\mathcal{R}} = B^T D_{\mathcal{X}}^{-1} B$ where $D_{\mathcal{X}} \in \mathbb{R}^{n\times n}$ is a diagonal matrix with the sum of the i-th row of B as its (i,i)-th entry. It has been proven by Li et al. that solving the eigenproblem of the Laplacian with adjacency matrix W on graph G is equivalent to solving it with $W_{\mathcal{R}}$ on graph $G_{\mathcal{R}}$. [35]
Let the first K eigenpairs for the eigenproblem with $W_{\mathcal{R}}$ be denoted as $\{(\lambda_i, v_i)\}_{i=1}^{K}$ with $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_K < 1$ and for the eigenproblem with W as $\{(\gamma_i, u_i)\}_{i=1}^{K}$ with $0 = \gamma_1 \leq \gamma_2 \leq \cdots \leq \gamma_K < 1$. Then, by Li et al., the following equations hold:

$$\gamma_i(2-\gamma_i) = \lambda_i$$

$$u_i = \begin{bmatrix} h_i \\ v_i \end{bmatrix}$$

$$h_i = \frac{1}{1-\gamma_i} T v_i$$

Where $T = D_{\mathcal{X}}^{-1} B$ is referred to as the transition probability matrix. [35]
Note that u_i is an $(n+p)\times k$ matrix, so the first n rows (corresponding to the n objects) can be used to construct the rows of H in Algorithm 3, upon which K-means is applied to obtain the final clustering result. [35]

Compared to the random removal method, *U-SPEC* is less efficient and intuitive. However, the main benefits of the method lie in its robustness. Therefore, it is expected that *U-SPEC* will yield more accurate clustering results. To check this hypothesis, in Subsection 5.1.2, both observation reduction techniques (namely the random removal method and *U-SPEC*) will be evaluated after applying them to sample datasets B.

4.3.3. Practical details

All practical details regarding the parameter choices of the random removal method and *U-SPEC* (such as p and z_1) have been explained in the previous two subsections. However, the observation reduction techniques yield high-dimensional datasets; the number of features/dimensions is comparable to or greater than the number of observations. According to [15], the number of isolated eigenvalues does not necessarily match the number of clusters in this case. Therefore, the conventional method of relying on the eigengap heuristic to determine the number of clusters, as explained in Subsection 4.2.2, cannot be applied when employing observation reduction techniques. Instead, the number of informative eigenvectors is used to estimate the number of clusters. [15] The methodology for obtaining these informative eigenvectors, will be explained by means of an example.

Example of determining the number of clusters

The high-dimensional sample dataset from [15] is used to provide a demonstration of the process for acquiring the informative eigenvectors. The dataset consists of 512 data points, each having 2048 dimensions, conforming to a Gaussian distribution and grouped into three clusters. This results in a ratio of 1/4 between data points and dimensions.
Figure 4.9 shows a histogram of the eigenvalues of the normalized Laplacian of the sample dataset. Note that the Laplacian is multiplied by n to improve the interpretability of the x-axis. Furthermore, the largest eigenvalue is close to n and thus omitted to retain the visibility of the histogram. In the plot, "Eigval. 2" therefore refers to the second largest eigenvalue of the Laplacian.

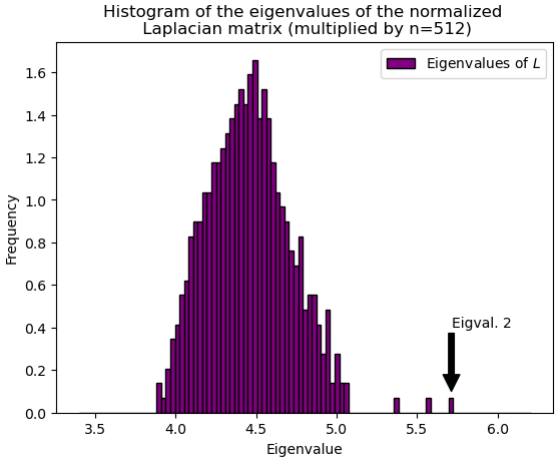


Figure 4.9: This figure shows a histogram of the eigenvalues of the normalized Laplacian (multiplied by $n = 512$) of the sample dataset. "Eigval. 2" refers to the second largest eigenvalue.

In Figure 4.10, the eigenvectors corresponding to the four isolated eigenvalues of the histogram are depicted. Notably, eigenvector 1, associated with the largest eigenvalue, retains information, evidenced by a heightened volatility between the two dotted black lines. Similarly, eigenvector 2 is informative, with entries predominantly greater before the second dotted line and eigenvector 4 also carries information, with predominantly higher entries between the two dotted lines. However, eigenvector 3 is non-informative since it does not exhibit any trends regarding the values of its entries.
So, there are three informative eigenvectors and thus it is estimated that the sample dataset consists of three clusters, which aligns with the true number of clusters. Moreover, it is apparent that the conventional method of relying on the eigengap heuristic, which returned four clusters, would have led to an overestimation.
Lastly, note that the number of informative eigenvectors serves as an upper bound for the number of clusters, implying that the actual number of clusters might theoretically be even smaller. [15]

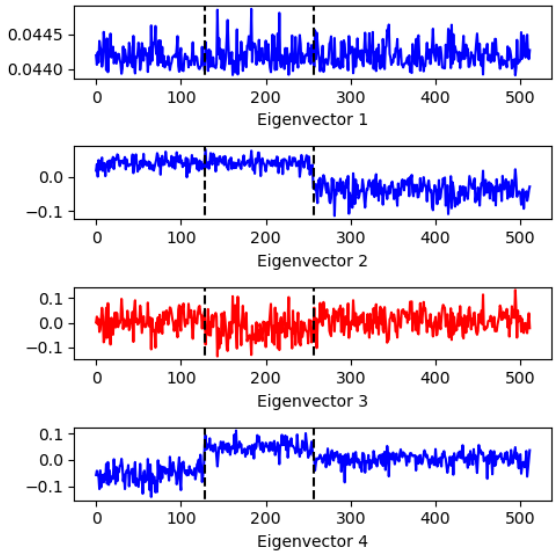


Figure 4.10: This figure shows the eigenvectors corresponding to the four isolated eigenvalues of the histogram. Eigenvector 3 (shown in red) is non-informative.

Number of columns of H

According to the number of informative eigenvectors, the rows of H in Algorithm 3 should be grouped into $K = 3$ clusters with K-means. However, there remains uncertainty regarding whether the columns

of H are the eigenvectors corresponding to the $K = 3$ *or* $K' = 4$ minimal eigenvalues of $\bar{L}$. Here K' represents the number of isolated eigenvalues (i.e. the number of clusters according to the eigengap heuristic). Nevertheless, for the sample dataset, the ARI (which will be explained in the next section) is higher when K' is used (ARI= 0.922) compared to K (ARI= 0.916). The same holds for other ratios between the data points and dimensions (e.g. 1 and 4). Therefore, in Algorithm 3, the columns of H should be the eigenvectors corresponding to the K' minimal eigenvalues and K-means is used to group the rows of H into K clusters.

For both observation reduction techniques, the number of clusters is equal to the number of informative eigenvectors. Furthermore, for the *U-SPEC* method, the eigenvectors associated with the Laplacian of $W_{\mathcal{R}}$—representing a high-dimensional scenario—are used instead of those linked to the Laplacian of W.

4.3.4. Updated versions of the normalized spectral clustering algorithm

The updated versions of the normalized spectral clustering algorithm (i.e. Algorithm 3) according to the random removal technique and the *U-SPEC* method are described in Algorithms 4 and 5 respectively.

Algorithm 4 The updated normalized spectral clustering algorithm with random removal

Input: $x_1, \ldots, x_n$ data points to be clustered
$m < n$ number of data points to be removed

Randomly remove m observations from the dataset

Compute the similarity matrix S and the normalized Laplacian $\bar{L} = I - D^{-1/2}WD^{-1/2}$
Extract the number of isolated eigenvalues of $\bar{L}$ and set K' equal to this number
Obtain the eigenvectors of the isolated eigenvalues and set K equal to the number of informative eigenvectors
Construct a matrix H whose columns are the eigenvectors corresponding to the K' isolated eigenvalues of $\bar{L}$.
Use K-means to cluster the rows of H into $C_1, \ldots, C_K$

for $k = 1, \ldots, K$ **do**
Calculate the cluster centroids c_k according to K-prototypes
end for
Assign each of the m omitted data points x_i to cluster $k = \text{argmin}_j d(x_i, c_j)^2$

Algorithm 5 The updated normalized spectral clustering algorithm with *U-SPEC*

Input: $x_1, \ldots, x_n$ data points to be clustered
p' number of candidate representatives
p number of representatives

Perform Phase 1 (Hybrid representative selection) and Phase 2 (Approximation of k-nearest representatives) of *U-SPEC* with p and p'

Calculate W and $W_{\mathcal{R}}$ according to Phase 3 (Bipartite graph partitioning)
Compute the normalized Laplacian $\bar{L} = I - D^{-1/2}W_{\mathcal{R}}D^{-1/2}$
Extract the number of isolated eigenvalues of $\bar{L}$ and set K' equal to this number
Obtain the eigenvectors of the isolated eigenvalues and set K equal to the number of informative eigenvectors
Calculate u_i of Phase 3 by using the eigenvectors corresponding to the K' isolated eigenvalues of $\bar{L}$.
Construct a matrix H whose rows are the first n rows of u_i. This means that there are K' columns.
Use K-means to cluster the rows of H into $C_1, \ldots, C_K$

4.4. Evaluation techniques

To evaluate the performances of the clustering techniques, several methods can be employed. [19] [65] This subsection discusses four of these methods.

4.4.1. Adjusted Rand index (ARI)

The Rand index quantifies the similarity between two data clusterings, Cl_1 and Cl_2. Cl_1 represents the clustering output generated by the clustering technique (e.g. K-prototypes or modified spectral clustering), while Cl_2 denotes the actual clustering. This means that the Rand index can only be used as an evaluation technique if the actual clusters are known. Thus, it can only assess the clustering techniques for the sample datasets A (and observation reduction techniques for sample datasets B) and not for the ARD and WAM datasets. [62]

The Rand index is computed as follows;

$$\text{Rand index} = \frac{a+b}{a+b+c+d} \tag{4.4}$$

Where the number of pairs of data points that belong to the *same* cluster in Cl_1 and the *same* cluster in Cl_2 are denoted by a, those that belong to *different* clusters in Cl_1 and *different* clusters in Cl_2 are equal to b, those that belong to the *same* cluster in Cl_1 and *different* clusters in Cl_2 are represented by c, and those that belong to *different* clusters in Cl_1 and the *same* cluster in Cl_2 are equal to d.
Note that the Rand index lies between 0 and 1. [62]

The *adjusted* Rand index (i.e. ARI) is the corrected-for-chance version of the Rand index. It is calculated as follows:

$$ARI = \frac{\text{Rand index} - \mathbb{E}(\text{Rand index})}{\max(\text{Rand index}) - \mathbb{E}(\text{Rand index})} = \frac{\text{Rand index} - \mathbb{E}(\text{Rand index})}{1 - \mathbb{E}(\text{Rand index})}$$

To obtain $\mathbb{E}$(Rand index), a contingency table similar to Figure 4.1 is created. The rows represent the clusters of the clustering output Cl_1, while the columns represent those of Cl_2. The entry in row i and column j is equal to the number of data points that cluster i of Cl_1 and cluster j of Cl_2 have in common. The expected Rand index can then be calculated with Formula 4.4 and the procedure outlined in step 6 of Subsection 4.1.1. [57]
Note that the adjusted Rand index ranges from -1 to 1, and that, as mentioned previously, the ARI is only used to assess the clustering techniques for the sample datasets A and observation reduction techniques for sample datasets B (and not for the ARD and WAM datasets). [57]

4.4.2. Error sum of squares (ESS)

The ESS is commonly utilized for evaluating clustering techniques. [56] However, as mentioned in Subsection 4.2.1, as the number of clusters increases, the ESS decreases since the distances from each data point to the nearest centroids decrease with the availability of additional centroids. Therefore, this measure can only effectively evaluate clustering techniques if the same number of clusters is used for each method, which is not assumed in this thesis.
Furthermore, the ESS becomes unreliable when the data is non-linearly separable. An example of this is shown in Figure 4.11; the distances from each data point to the nearest centroids are smaller in the left plot. Thus, the ESS is smaller for the left plot, despite the right plot potentially representing a more accurate clustering. Since the ARD and WAM datasets are most likely non-linearly separable, the ESS cannot be used to evaluate the clustering techniques.

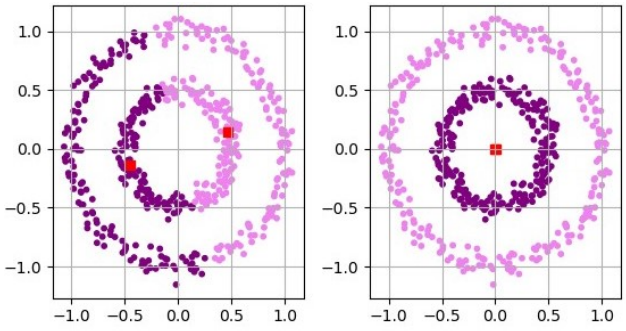


Figure 4.11: This figure shows an example of two clusterings of non-linearly separable data; a case in which the *ESS* is an unreliable measure. The red squares represent the centroids.

4.4.3. Sensicality of the clusters

Many other evaluation measures (such as the Hubert index, Silhouette coefficient, and gap statistic) are also unreliable when the data is non-linearly separable. [65] Therefore, in order to evaluate the clustering techniques for the ARD and WAM datasets, experts in the actuarial field will judge the degree to which the clusters are meaningful. The more logical the clustering output of a technique, the more accurate the clustering.
Note that, beforehand, the adjusted Rand index is used to evaluate the clustering techniques for sample datasets A (since the correct clusters are known). The results provide insights into how the techniques will perform on the ARD and WAM datasets.

4.4.4. Comparison with the current risk classification approach

Besides evaluating the clustering results based on their sensicalities, their impact on the risk classification of the claim frequency models is also analyzed. This involves incorporating the clusters as additional risk factors into the existing claim frequency GLM, while keeping the original risk factors. For example, if for the ARD dataset, the spectral clustering algorithm identifies three clusters for the license plates (i.e. "KT") and four for the zip codes (i.e. "ZC"), then seven boolean variables are introduced as new risk factors: Cluster_Spec_KT_0 to Cluster_Spec_KT_2, and Cluster_Spec_ZC_0 to Cluster_Spec_ZC_3. Each license plate and its corresponding zip code are then assigned values for these boolean variables based on the clustering model outputs. This approach can also be applied for the K-prototypes method.

For this modified GLM model, the coefficients of all parameters, including those used in the current GLM, are estimated. The model summary is then compared to that of the existing GLM claim frequency model by using the evaluation metrics that are described in the following four subsections.

Standard error

The standard error of a parameter in the model summary is a measure of the uncertainty around the estimated difference in claim frequency with respect to the null-cluster. For this thesis, the null-cluster is set equal to the cluster containing the most exposure (as defined in Subsection 3.2.1).
The standard error is equal to $se = \frac{\sigma}{\sqrt{n}}$ where n is the sample size. σ is calculated by using the formula of σ that was stated in Section 4.1.1 for the estimated difference in claim frequency c. The standard error percentage is then computed by taking $\frac{se}{c} \cdot 100\%$. If a parameter's standard error percentage is below 50, it significantly differs from the null-cluster and thus impacts the model, indicating that this variable is useful for predicting risk in the GLM model. [75]
Note that a standard error percentage of 50% implies that $2 \cdot se = c$. Therefore, a standard error percentage of 50% corresponds to an estimated difference in claim frequency that is equal to two standard errors. Furthermore, for a two-sided confidence interval, this is approximately equivalent to a p-value of 0.05. [48] [2] So, if the standard error percentage of a cluster is greater than 50%, $c < 2 \cdot se$, and thus p-value> 0.05. Therefore, the null hypothesis, which states that the cluster does not significantly differ from the null-cluster, cannot be rejected.

Deviance

The unit deviance $D(y, \mu) \in \mathbb{R}_{\geq 0}$ is a bivariate function that satisfies the following conditions:

- $D(y, y) = 0$
- $D(y, \mu) > 0, \forall y \neq \mu$

The total deviance $D_{tot}(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) \in \mathbb{R}_{\geq 0}$ of a model with predictions $\hat{\boldsymbol{\mu}}$ of observations $\boldsymbol{y}$ is the sum of the unit deviances, i.e. $D_{tot}(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = \sum_i D(y_i, \hat{\mu}_i) = 2(l(\boldsymbol{y}; \boldsymbol{y}) - l(\hat{\boldsymbol{\mu}}; \boldsymbol{y}))$ where l denotes the log likelihood. The smaller the total deviance of the model, the better the fit. [33]

AICc

Similar to the deviance, the Akaike Information Criterion (AIC) is a goodness-of-fit statistic. However, unlike the deviance, the AIC also addresses the risk of overfitting by favoring simpler models since increasing the number of parameters in the model almost always enhances its fit.
The AIC value of a model can be calculated with the following formula:

$$AIC = 2k - 2\log(\hat{\mathcal{L}})$$

Here k is equal to the number of estimated model parameters and $\hat{\mathcal{L}}$ refers to the maximized value of the likelihood function for the model. The smaller the AIC value, the better the model. [82]

To evaluate the modified GLM model, the AICc (AIC corrected) is compared to that of the existing GLM claim frequency model. The measure is computed as follows:

$$AICc = AIC + \frac{2k^2 + 2k}{n - k - 1}$$

Where n is the number of data points. This adjustment corrects the AIC for small sample sizes. A lower AICc value indicates a better model. [54]

BIC

The Bayesian Information Criterion (BIC) is similar to the AIC, but applies a larger penalty for the number of parameters when the sample size exceeds seven. Given that the license plate and zip code datasets contain more than seven data points, the BIC will put a greater penalty on the model parameters than the AIC. Moreover, unlike the AIC, the BIC is consistent. [13]
The BIC is calculated as follows:

$$BIC = k \cdot \log(n) - 2\log(\hat{\mathcal{L}})$$

The smaller the BIC value, the better the model. [13]

4.5. Stability of the techniques

4.5.1. Time stability

To obtain the clustering results of both methods, data from the past ten years is used. However, features related to license plates and zip codes can change over time. For instance, cities may develop and undergo changes in urbanization, average income, and education and more cars become sustainable resulting in higher ratings of eco-friendliness. Therefore, the time stabilities of the K-prototypes and spectral clustering results are evaluated.
This is done by implementing interaction terms in the modified GLM (described in Subsection 4.4.4) that combine the binary cluster variable and the policy year. Interaction terms are added one at a time to test if, for every policy year, the same effect on the claim frequency can be observed for each cluster. If the effects over time are random (i.e. there is no trend) and if the claim frequencies fluctuate around a mean, the cluster is considered stable over time. Therefore, the risk factor of such a cluster remains included in the model. If a cluster is unstable over time, it should be removed from the GLM.

4.5.2. Stability with respect to the number of observations

The stabilities of the *U-SPEC* clustering results with respect to the number of observations are evaluated. This helps determine the applicability of this clustering method to smaller datasets, such as those for different coverages or insurance products. To assess the stability, the number of candidate representatives p' is varied as p' is the subset of the total data points to which *U-SPEC* is applied. Keeping other variables (such as the number of representatives p and the number of rep-clusters z_1) constant ensures accurate comparison.

p' will range from 2000 to 200 since 2000 is the initial number of points used to obtain the spectral clustering results that are compared to those of K-prototypes. Besides, more points would require excessive runtime. 200 is the minimum as $p' \geq p = 200$. The specific values of p' are: 2000 (used twice to check consistency), 1500, 1000, 500, and 200. For each value, the rand index is computed relative to $p' = 2000$ and box plots of the claim frequencies are created for all clusters.

Note that the stabilities of the K-prototypes' results are not evaluated, as *U-SPEC* is of more interest since it performs better than K-prototypes, which will be shown in the next chapter.

In the upcoming chapter, the results of the methods outlined in this chapter are discussed.

5 Results

In this chapter, the results of the experiments, that were described in Chapter 4, are discussed. In Section 5.1 the results of sample datasets A and sample datasets B are evaluated, while in Section 5.2 the license plate and zip code clustering results of the ARD and WAM datasets are assessed. Lastly, in Section 5.3 the stability of the cluster results (with respect to time and the number of observations) is analyzed.

5.1. Sample datasets results

In this section, the results of the sample datasets are discussed. Subsection 5.1.1 delves into the findings of sample datasets A; the performances of the K-prototypes and modified spectral clustering algorithm are assessed by evaluating these techniques across various 2D datasets (see Figure 4.2). Subsection 5.1.2 focuses on the results of sample datasets B; the two observation reduction techniques (i.e. random removal and $U\text{-}SPEC$) are assessed.

5.1.1. Results of sample datasets A

Figure 5.1 shows the results of the K-prototypes and modified spectral clustering algorithms for sample datasets A. The number of clusters for each dataset is determined by using the elbow plot for K-prototypes and the eigenvalue plot for modified spectral clustering. In the figure, it can be seen that the modified spectral clustering algorithm outperforms the K-prototypes method, especially for non-linearly separable datasets (e.g. the noisy circles and noisy moons sample data).
In the case of the homogeneously distributed data, there is no good clustering. For the spectral clustering method, the eigenvalue plot shows a single cluster, correctly indicating that the data should not be clustered. On the other hand, K-prototypes still groups the data, but the resulting clusters lack meaningful interpretation. This occurs because the elbow shape cannot be observed at $k = 1$ since the ESS cannot be calculated for $k = 0$. Therefore, the K-prototypes method cannot determine when effective clustering is impossible.

Figure 5.2 shows a table of the adjusted Rand indices for each sample dataset with the K-prototypes and spectral clustering methods. It can be seen that, except for two cases (specifically, the Gaussian and homogeneously distributed data), the spectral clustering algorithm yields a higher ARI and thus outperforms the K-prototypes method. In the case of the Gaussian distributed data, both algorithms produce the same adjusted Rand index, while for the homogeneously distributed data, the ARI cannot be computed since there is no correct clustering (which is a prerequisite for the index calculation). However, as mentioned prior, spectral clustering correctly does not cluster the data in this case.

In conclusion, for sample datasets A, the modified spectral clustering algorithm outperforms the K-prototypes method, especially for non-linearly separable data (as was predicted and explained in Section 3.3). Due to the high likelihood that the WAM and ARD datasets are non-linearly separable, it is thus expected that using the modified spectral clustering algorithm will result in a better clustering

for the vehicle and geographical datasets. It is also expected that the spectral clustering method will indicate when effective clustering is not possible.

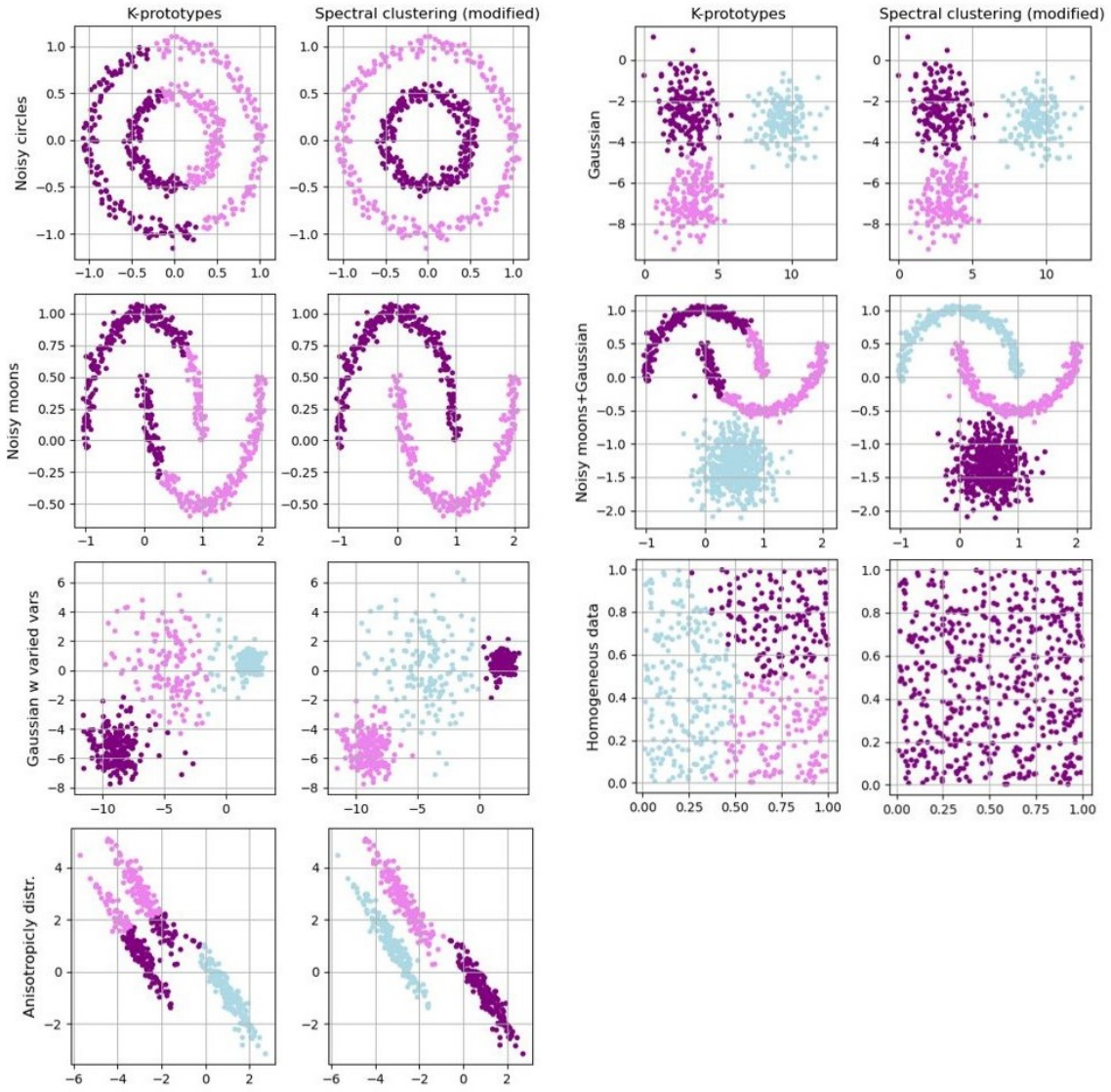


Figure 5.1: This figure shows the results of the K-prototypes and modified spectral clustering algorithms for sample datasets A.

	Adjusted Rand index	
Dataset	**K-Prototypes**	**Spectral Clustering**
Noisy circles	-0.00186	1
Noisy moons	0.237	1
Gaussian w varied vars	0.787	0.896
Anisotropicly distr.	0.558	0.994
Gaussian	0.970	0.970
Noisy moons + Gaussian	0.787	0.990
Homogeneous	-	-

Figure 5.2: This figure shows a table of the adjusted Rand indices for each sample dataset with the K-prototypes and spectral clustering methods.

5.1.2. Results of sample datasets B

Figure 5.3 shows a table of the adjusted Rand indices of sample datasets B for both of the observation reduction techniques (i.e. *U-SPEC* and random removal) after removing 320 data points. For the random removal method, it also shows the maximum number of data points that can be deleted from sample datasets B while maintaining an ARI of 1.
Each dataset in sample datasets B consists of 400 data points, with the condition for *U-SPEC* being $p \leq p' \ll n = 400$. However, setting $p' = 100$, for instance, results in $p = \frac{1}{10} \cdot p' = 10$ and $z_1 = \lfloor p^{1/2} \rfloor = 1$. This means that in phase 2 of *U-SPEC*, there is only one rep-cluster ($z_1 = 1$), leading to an insufficient number of candidates to approximate the k-nearest neighbors in Figure 4.8. Thus, the relationship between p', p, and n outlined in Subsection 4.3.2 ($10p = p' \ll n$) is altered when applying *U-SPEC* to sample datasets B. Specifically, $p' = 200$ and $p = 80$. z_1, k, and k' can then be calculated as described in Subsection 4.3.2. So, the second and third phases of *U-SPEC* are applied to $p = 80$ out of the 400 data points (i.e. 320 data points are removed), resulting in an observation-to-dimension ratio of $\frac{80}{400} = \frac{1}{5}$.

Figure 5.3 shows that, for the *U-SPEC* method, the ARI rises as the number of clusters increases. Conversely, for the random removal technique, the ARI climbs as the number of clusters decreases. It can also be seen that in sample dataset B, where the clusters are well-defined, nearly all of the 400 data points can be removed while maintaining an ARI of 1 when the number of clusters is small (in this case, 3). This statement does not hold for an intermediate and large number of clusters.

	U-SPEC	Random removal	
Dataset (#Clusters)	ARI with 80 datapoints	Max #datapoints removed for ARI=1	ARI with 80 datapoints
Small (3)	0.574	372	1
Intermediate (7)	0.765	15.2	0.966
Large (10)	0.815	13.1	0.867

Figure 5.3: This figure shows a table of the adjusted Rand indices of sample datasets B for both of the observation reduction techniques (i.e. *U-SPEC* and random removal) after removing 320 data points. For the random removal method, it also shows the maximum number of data points that can be deleted from sample datasets B while maintaining an ARI of 1.

Despite the random removal technique outperforming the *U-SPEC* method in terms of ARI across all three cluster count scenarios, the *U-SPEC* method is utilized for the WAM and ARD datasets. This decision is based on the following four reasons:

1. **Small dataset**. The *U-SPEC* method was designed for clustering *large* datasets. However, with only 400 data points in sample datasets B, the dataset size is relatively small. Considering that the WAM and ARD datasets are larger, it is anticipated that *U-SPEC* will outperform the random removal method in those cases.
2. **Relationships of the parameters**. As mentioned prior, due to the relatively small size of the dataset, the selected parameters may not adhere to the relationships outlined in Subsection 4.3.2 (e.g. it does not hold that $p' \ll n$). This could affect the performance of the *U-SPEC* method. However, given that the WAM and ARD datasets are larger, it is anticipated that the parameters will align with the relationships specified in Subsection 4.3.2 and therefore *U-SPEC* will outperform the random removal method in these instances.
3. **Well-defined clusters**. The clusters in sample datasets B are clearly defined as they are artificially generated. Consequently, many data points can be removed while still preserving an ARI of 1 when the number of clusters is small. Moreover, this might lead to higher ARI scores for the random removal technique compared to *U-SPEC*. However, the WAM and ARD datasets lack clearly defined clusters. Thus, *U-SPEC* will most likely outperform the random removal method in these cases.
4. **Small z_1 and p**. Due to the small size of sample datasets B, both the number of representatives ($p = 80$) and the number of rep-clusters ($z_1 = \lfloor p^{1/2} \rfloor = \lfloor 80^{1/2} \rfloor = 8$) are small. This can negatively impact the performance of the *U-SPEC* method. However, considering that the WAM and ARD

datasets are larger, resulting in larger values for p and z_1, it is expected that *U-SPEC* will outperform the random removal method in these scenarios.

So, it is expected that the *U-SPEC* method will outperform the random removal technique for the WAM and ARD datasets. Therefore, the spectral clustering results of the next section are obtained by applying *U-SPEC* as the observation reduction technique.

5.2. Clustering results

In this section, the clustering results are evaluated. For the ARD dataset, Subsection 5.2.1 discusses the license plate clustering results, while Subsection 5.2.2 addresses the outcomes of the zip code clustering. Additionally, Subsection 5.2.4 provides a summary of the WAM dataset results, with the detailed results available in Appendix A. Sections 5.2.3 and 5.2.5 compare the modified GLMs with the current claim frequency GLMs for the ARD and WAM datasets respectively.

5.2.1. License plate clustering of the ARD dataset

K-prototypes method

For the ARD license plate dataset, the K-prototypes algorithm was run with values of K ranging from 2 to 20 to create the elbow plot shown in Figure 5.4. The elbow shape can be observed around $K = 9$, indicating that the optimal number of clusters for the license plates ("KT") in the ARD dataset is nine.

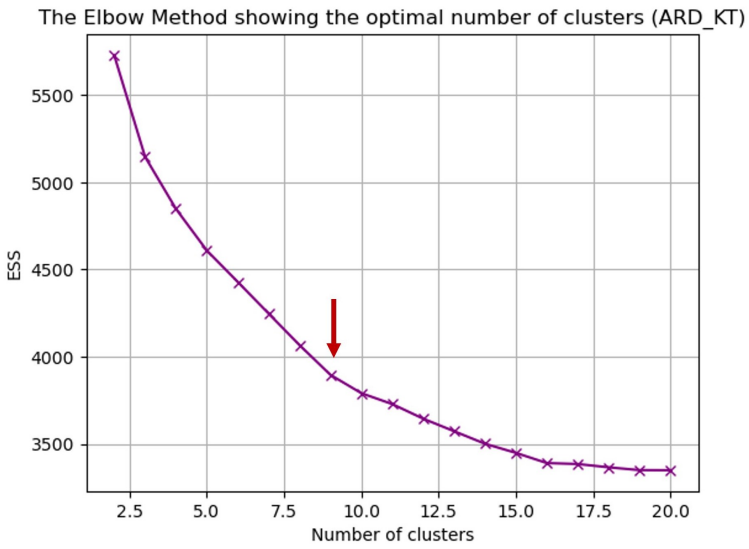


Figure 5.4: This figure shows the elbow plot for the ARD license plate dataset. The elbow shape can be observed around $K = 9$, indicating that the optimal number of clusters for the license plates ("KT") in the ARD dataset is nine.

Each license plate is assigned to one of $K = 9$ clusters with K-prototypes. Figure 5.5 displays a table of the cluster centroids, with most columns omitted for brevity. This table will be used to describe the clusters later in this subsection.

Cluster	Weight	Age	...	Class top speed	Index eco-friendliness
0	-1.08	0.116	...	1	A
1	0.0896	1.06	...	5	C
2	-2.98	0.546	...	Unknown	Unknown
3	0.318	0.367	...	2	A
4	0.793	-0.0481	...	5	Unknown
5	1.26	1.21	...	5	Unknown
6	-0.127	-0.808	...	4	D
7	-0.00836	-0.182	...	3	Unknown
8	0.469	-0.980	...	Unknown	Unknown

Figure 5.5: This figure shows a table of the cluster centroids, with most columns omitted for brevity.

Figure 5.6 shows the box plots of each cluster regarding the claim frequency. The distinct averages and variations observed in these box plots indicate that the clusters are significant in terms of claim frequency, making them suitable for inclusion as risk factors in the GLM. An Analysis of variance (ANOVA) (a statistical test used to evaluate the difference between the means of more than two groups) can be used to make sure the means differ. [47] However, when incorporating the clusters as additional risk factors in the GLM, the standard errors are already used to assess the significance of each cluster.

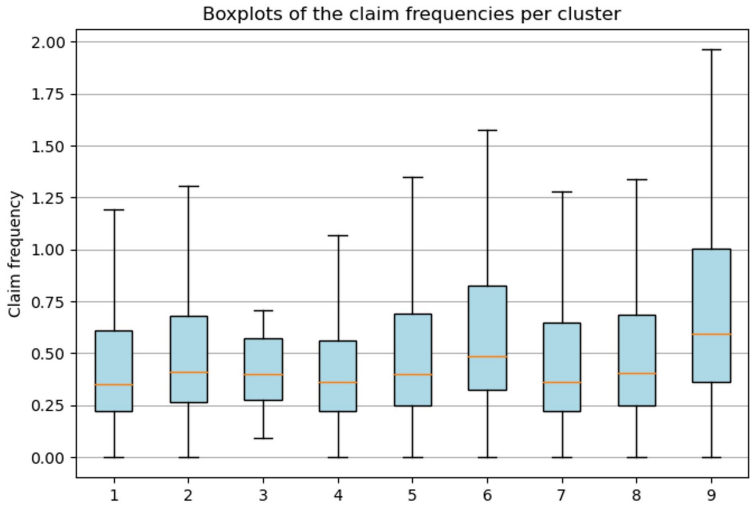


Figure 5.6: This figure shows the box plots of each cluster regarding the claim frequency. It can be observed that every cluster has a distinct average and variation in its box plot.

Figure 5.7 presents a table of the descriptions of all clusters derived from Figure 5.5. For each cluster, the number of license plates, average claim frequency, and the class of average claim frequency are also provided. The descriptions of the clusters can be summarized as follows:

- **Cluster 0**: Affordable “middle of the road” cars.
- **Cluster 1**: New, high-power, and high-speed cars.
- **Cluster 2**: Small vehicles with two or three wheels.
- **Cluster 3**: Low-power, low-speed cars.
- **Cluster 4**: Expensive, large, high-power cars that are not eco-friendly.
- **Cluster 5**: Expensive, and new electric vehicles.
- **Cluster 6**: Old, inexpensive, non-eco-friendly cars.
- **Cluster 7**: Expensive “middle of the road” cars.
- **Cluster 8**: Cars with a lot of unknown data.

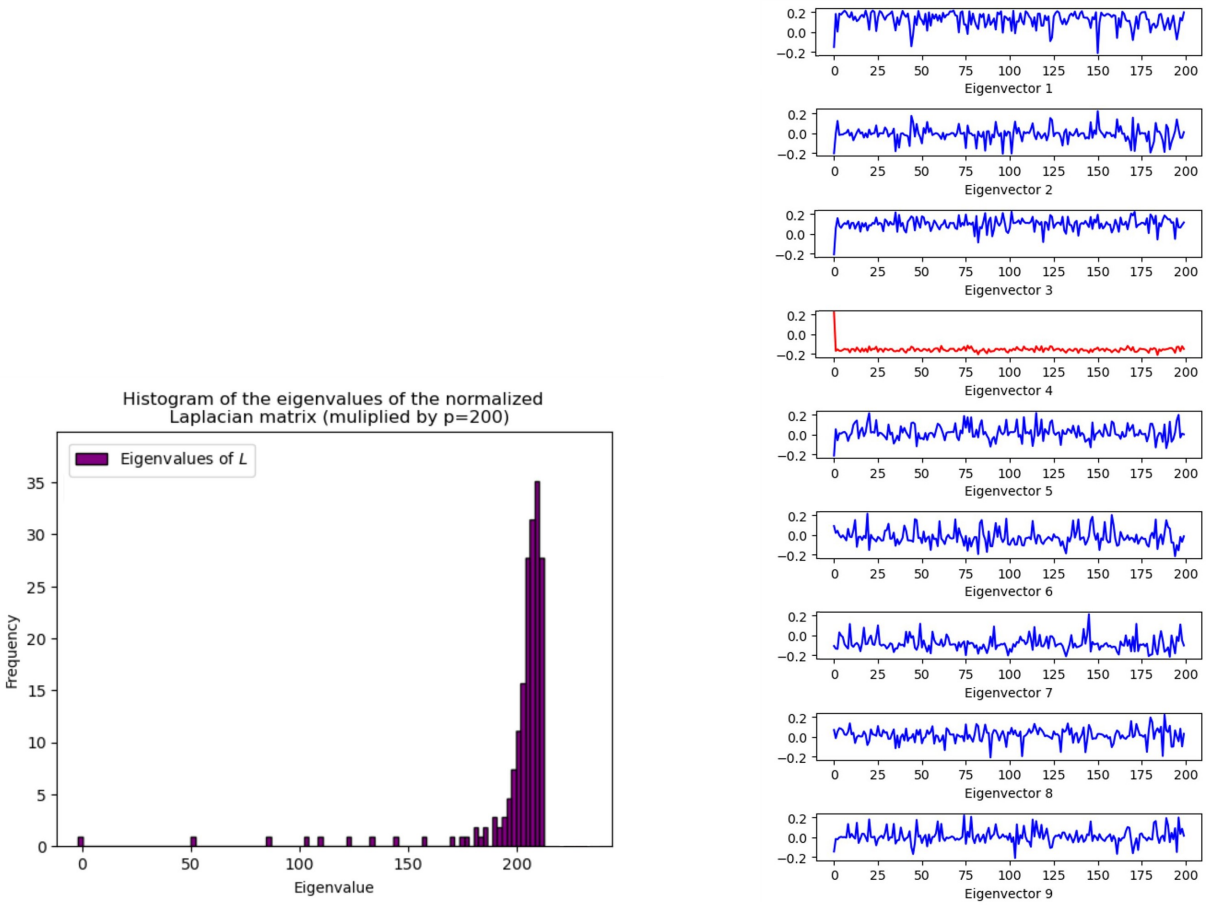
Cluster	Description	#License Plates	Claim frequency	
			Average	Class
0	Contains most license plates. Cheapest cars, smallest cylinder volume, lowest class of top speed, and smallest length. Eco-friendly. Lot of VW Polo's and hatchbacks. "Middle of the road" cluster with cluster 7, but lower acf.	[redacted]	0.555	Low
1	New cars with a lot of power. Most common brands are BMW and Audi (also contains a lot of VW Golf). High class top speed, and high lower bound of number of gears.	[redacted]	0.605	Intermediate
2	European vehicle category L5 & L1 (small vehicles with 2 or 3 wheels). A lot of unknown data (e.g. usage and eco-friendliness). Most common brand is Piaggio.	[redacted]	0.983	***Highest***
3	Not a lot of power. Most common brands are Toyota, Lexus, Mitchubishi etc.. Lot of SUV's and station wagons. Hybrids with a low class of top speed and semi-automatic gear box.	[redacted]	0.485	***Lowest***
4	Least eco-friendly cars, most cylinders, most power, expensive cars with a large wheelbase. Most common brand is Mercedes-Benz.	[redacted]	0.618	Intermediate
5	Heaviest and newest vehicles (doesn't have the most cylinders). Expensive, electric, and high class top speed. Most common brand is Tesla.	[redacted]	0.690	High
6	Least eco-friendly (with cluster 4), lot of seats, relatively old. MPV cars with catalytic converters. Relatively cheap. Most common brands: Opel, Renault, Peugeot.	[redacted]	0.575	Low
7	Contains most license plates (with cluster 0). Also "Middle of the road" cluster, but more expensive, greater length, and bigger cylinder volume. Eco-friendly and intermediate class of top speed. Lot of station wagons and Volvo's.	[redacted]	0.605	Intermediate
8	Lot of unknown data (e.g. "Name of car"). Thus, similar to cluster 2, but contains (older) cars.	[redacted]	0.899	High

Figure 5.7: This figure shows a table of the descriptions for all clusters. The number of license plates, average claim frequency, and class of average claim frequency are also provided for every cluster.

Evaluation: According to experts in the actuarial field, the clusters make sense. For example, vehicles with two or three wheels tend to have a higher average claim frequency, and cars like Toyota and Lexus in cluster 3 are considered safer, resulting in a lower claim frequency. However, cluster 2, with just eighteen license plates, poses a challenge. As it exhibits the highest average claim frequency, including cluster 2 as a risk factor would raise premiums for these eighteen vehicle owners, which is not feasible. Additionally, due to its small size, cluster 2 is unlikely to remain stable over time, as discussed further in Section 5.3.

Modified spectral clustering method

The ARD license plate dataset consists of $n = 60,083$ data points. Therefore, the $U\text{-}SPEC$ method was applied with $n \gg p' = 2000$ to ensure a sufficient number of rep-clusters z_1. This implies that $p = \frac{1}{10} \cdot 2000 = 200$, allowing z_1, k, and k' to be calculated as described in Subsection 4.3.2. Note that these parameters yield a high-dimensional dataset as the ratio of dimensions to observations is equal to $81/p = 81/200 = 0.405$. Therefore, the number of clusters is determined with the method outlined in Subsection 4.3.3.
Figure 5.8a shows a histogram of the eigenvalues of the normalized Laplacian, multiplied by $p = 200$, with $U\text{-}SPEC$. In Figure 5.8b, the eigenvectors corresponding to the nine isolated eigenvalues of this histogram are depicted. Since eigenvector 4 is non-informative, there are eight informative eigenvectors, indicating that the optimal number of clusters is *at most* eight.



(a) **(b)**

Figure 5.8: This figure shows, for the ARD license plate dataset: **(a)** a histogram of the eigenvalues of the normalized Laplacian (multiplied by $p = 200$), **(b)** the eigenvectors corresponding to the nine isolated eigenvalues of the histogram (eigenvector 4 is non-informative and shown in red).

The *U-SPEC* algorithm is completed with eight clusters and Figure 5.9 shows the box plots of each of these clusters regarding the claim frequency. The distinct averages and variations observed in these box plots indicate that the clusters are significant in terms of claim frequency, making them suitable for inclusion as risk factors in the GLM.

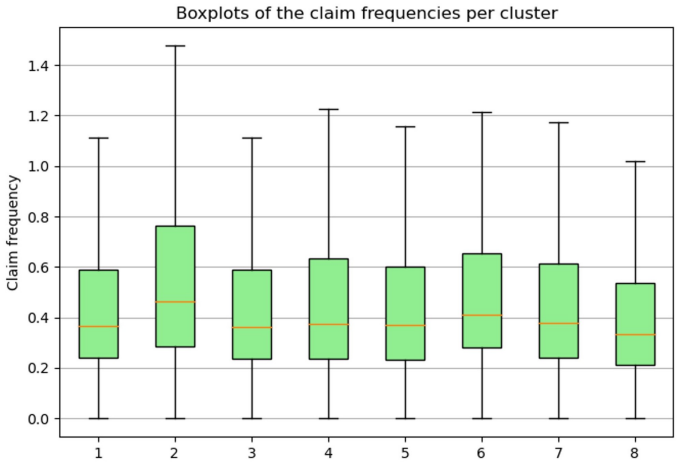


Figure 5.9: This figure shows the box plots of each cluster regarding the claim frequency. It can be observed that every cluster has a distinct average and variation in its box plot.

Figure 5.10 displays the *ordered* box plots of each cluster's claim frequency for both the K-prototypes (in blue) and spectral clustering (in green) methods. For cluster 2, spectral clustering shows greater variation in claim frequency. However, for clusters 4, 7, and 8, the variation is greater with K-prototypes. Therefore, it can be concluded that the K-prototypes clusters generally exhibit greater variation in claim frequency, indicating that spectral clustering more effectively maximizes the homogeneity among observations within the same cluster.
Note that the cluster containing vehicles with two or three wheels was omitted since spectral clustering did not produce this cluster.

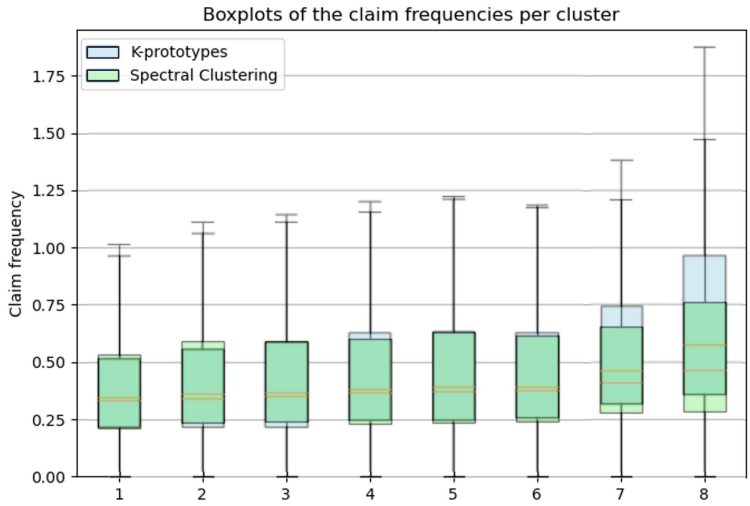


Figure 5.10: This figure shows the *ordered* box plots of each cluster's claim frequency for both the K-prototypes (in blue) and spectral clustering (in green) methods.

Figure 5.11 presents a table of the descriptions of all clusters. For each cluster, the number of license plates, average claim frequency, and the class of average claim frequency are also provided. The descriptions of the clusters can be summarized as follows:

- **Cluster 0**: Expensive, large, high-power cars that are not eco-friendly.
- **Cluster 1**: Old cars with a lot of unknown data (including small vehicles with two or three wheels).
- **Cluster 2**: "Middle of the road" cars.
- **Cluster 3**: Large eco-friendly cars.
- **Cluster 4**: Small low-power, low-speed, and eco-friendly cars.
- **Cluster 5**: Expensive, high-speed, and new electric vehicles.
- **Cluster 6**: Old, inexpensive, and heavy cars.
- **Cluster 7**: Light, low-power, most affordable cars.

Cluster	Description	#License Plates	Claim frequency	
			Average	Class
0	Most power and most cylinders. Heaviest, most expensive, and longest cars. Most gears (lower bound), least eco-friendly cars, and highest class top speed. Most common brand is BMW (also Volvo and Audi) and a lot of SUV's.	[redacted]	0.459	Low
1	Oldest cars, average price, and a lot of unknowns (e.g. class of top speed, eco-friendliness and brand). This cluster also includes a lot of small vehicles with 2 or 3 wheels.	[redacted]	0.592	***Highest***
2	Light cars and average age, price, power, and eco-friendliness. Highest class of top speed and the most common brand is Volkswagen. "Middle of the road" cluster.	[redacted]	0.450	Low
3	Most seats, average price, and long cars. Low emmision and very eco-friendly. This is the only cluster that has combi cars (instead of SUV or hatchbacks) as the most common car type. Most common brand is Mercedes-Benz.	[redacted]	0.488	Intermediate
4	Lowest class of top speed, least amount of seats, least amount of power, and very eco-friendly. Most common brand is Toyota.	[redacted]	0.470	Intermediate
5	Newest cars, second most power, most doors, second most expensive. Most common car model is Tesla model 3, very eco-friendly, and has the highest top speed with cluster 1. Most common brand is Volkswagen (also a lot of Mercedes-Benz and Tesla).	[redacted]	0.516	High
6	Old and heavy cars, not a lot of cylinders, cheap, and a low class of top speed. Most common brand is Volkswagen.	[redacted]	0.483	Intermediate
7	Light cars, least amount of cilinders, least amount of power, and cheapest cars. Average eco-friendliness and class of top speed. Most common brand is KIA.	[redacted]	0.421	***Lowest***

Figure 5.11: This figure shows a table of the descriptions for all clusters. The number of license plates, average claim frequency, and class of average claim frequency are also provided for every cluster.

Evaluation: Similar to the K-prototypes clusters, the clusters generated by the spectral clustering method are considered logical by experts in the actuarial field. For instance, vehicles with lots of unknown variables and small vehicles with two or three wheels exhibit a higher average claim frequency, while lighter and cheaper cars demonstrate lower frequencies, possibly due to the owners' decreased likelihood of filing insurance claims for such vehicles. Furthermore, since there is no cluster containing just eighteen license plates, the *U-SPEC* clusters appear more feasible than those of K-prototypes. For the other clusters, it is difficult to determine which method produces more logical groups. Therefore, a quantitative comparison is provided in Subsection 5.2.3.

5.2.2. Zip code clustering of the ARD dataset

K-prototypes method

For the ARD zip code dataset, the K-prototypes algorithm was run with values of K ranging from 2 to 13 to create the elbow plot shown in Figure 5.12. The elbow shape is observed around $K = 4$, indicating that the optimal number of clusters for the zip codes ("ZC") in the ARD dataset is four.

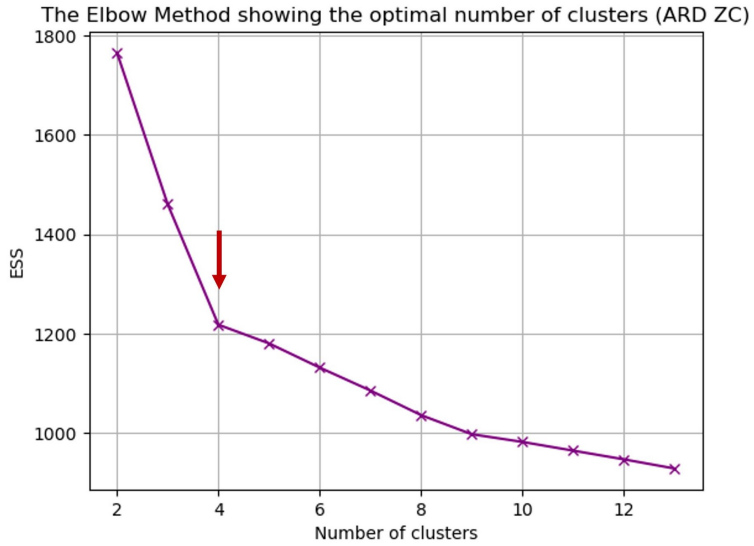


Figure 5.12: This figure shows the elbow plot for the ARD zip code dataset. The elbow shape is observed around $K = 4$, indicating that the optimal number of clusters for the zip codes ("ZC") in the ARD dataset is four.

Figure 5.13 shows the box plots of each cluster regarding the claim frequency. The distinct averages and variations observed in these box plots once again indicate that the clusters are significant in terms of claim frequency, making them suitable for inclusion as risk factors in the GLM.

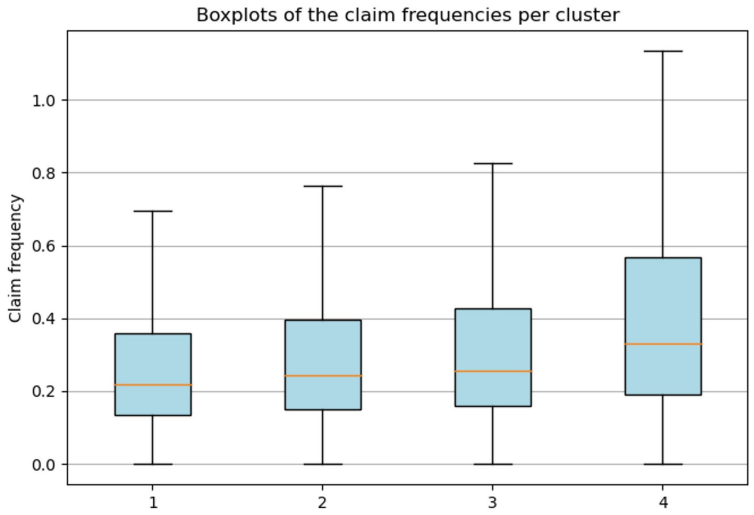


Figure 5.13: This figure shows the box plots of each cluster regarding the claim frequency. It can be observed that every cluster has a distinct average and variation in its box plot.

Figure 5.14 presents a table of the descriptions of all clusters. For each cluster, the number of zip codes, average claim frequency, and the class of average claim frequency are also provided. The descriptions of the clusters can be summarized as follows:

- **Cluster 0**: Rural areas.
- **Cluster 1**: Rich suburbs with a high level of education.
- **Cluster 2**: Regions characterized by newer houses and elderly residents.
- **Cluster 3**: Urban areas with a high population density.

Cluster	Description	#Zip codes	Claim frequency	
			Average	Class
0	Contains most zip codes, fintype: advice-sensitive families, geotype: conservative benefactors, hometype: older home owners. Low urbanisation and large distances to supermarkets and banks. Older home owners, intermediate social class, and least amount of non-western immigrants. Intermediate income and age. Has the biggest cars (with cluster 1) and the most motor cycles.	[redacted]	0.275	***Lowest***
1	On each zip code least amount of people, fintype: financial professionals, geotype: intellectual culture lovers, hometype: unknown. Intermediate urbanisation and highest social class. Highest percentage of company cars, income, education and couples with (older) kids. Most owner occupied houses, households with two earners, investors, and “Glossy” readers. Biggest cars (with cluster 0) and households of these zip codes spend the most on groceries.	[redacted]	0.300	Low
2	Fintype: interested conservatives, geotype: social believers, hometype: older home owners. Intermediate urbanisation, elder couples without kids, intermediate social class, least amount of non-western immigrants (with cluster 0), and intermediate income and education. Newest homes (new=70s), most single elderly people, and highest age (of people). Chance of changing health insurance lowest and least sensitive to switching car brands.	[redacted]	0.330	Intermediate
3	On each zip code most amount of people, fintype: active borrowers, geotype: average sport fans, hometype: single apartment tenants. Highest urbanisation, most young single people, lowest social class, most non-western immigrants, lowest incomes, lowest education, and oldest houses. Lowest age (of people), smallest living spaces, most borrowers, highest chance of changing health insurance, most sensitive to switching car brands, and highest risk of defaulting. Spends the least on daily groceries, not a lot of car ownership, and oldest, smallest and cheapest cars.	[redacted]	0.440	***Highest***

Figure 5.14: This figure shows a table of the descriptions for all clusters. The number of zip codes, average claim frequency, and class of average claim frequency are also provided for every cluster.

Figure 5.15 shows the four-digit zip code clusters of the Netherlands generated by the K-prototypes method. This map is created by taking the mode of the six-digit zip code clusters over each four-digit region. The darker the color in the map, the higher the average claim frequency of the corresponding cluster.
It is evident that cluster 0 encompasses the largest region on the map and that all large cities belong to the (urban) cluster 3. Furthermore, rich suburban places like “Het Gooi” (near Hilversum) are classified under cluster 1.

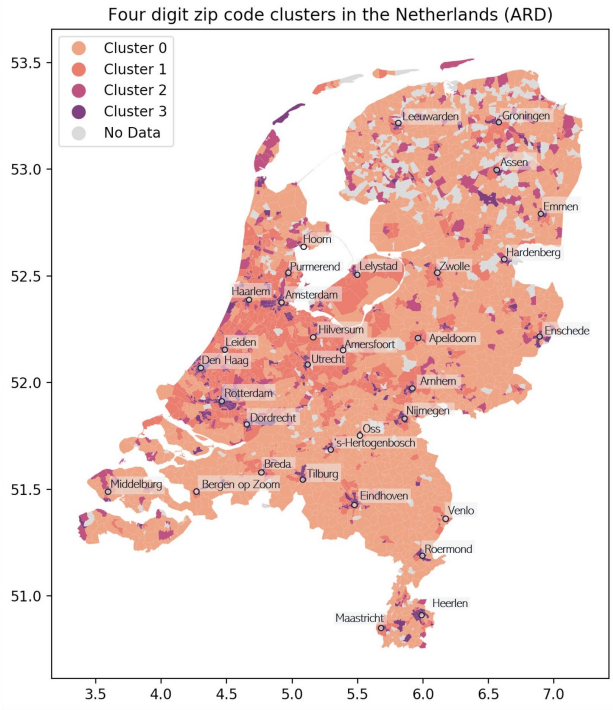


Figure 5.15: This figure shows the four-digit zip code clusters of the Netherlands generated by the K-prototypes method. This map is created by taking the mode of the six-digit zip code clusters over each four-digit region.

Figure 5.16 shows the zip code clusters of Amsterdam produced with the K-prototypes method. It is apparent that densely populated areas, like the city center, are classified under cluster 3, while upscale neighborhoods such as those surrounding the canals (i.e. "Grachtengordel West") and "Oud-Zuid" belong to cluster 1. Notably, cluster 0 is absent from this map.

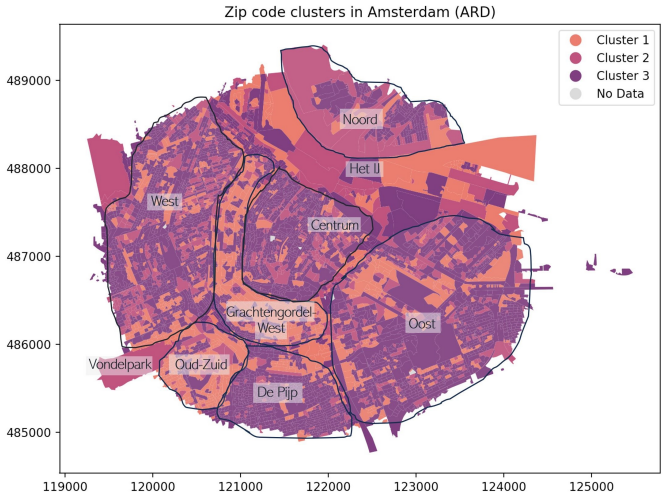


Figure 5.16: This figure shows the zip code clusters of Amsterdam produced with the K-prototypes method.

Lastly, Figure 5.17 displays the zip code clusters of Amsterdam and its surrounding area. Cities like Amstelveen and the Bijlmer are grouped into cluster 3, while smaller towns like Badhoevedorp are categorized under cluster 1.

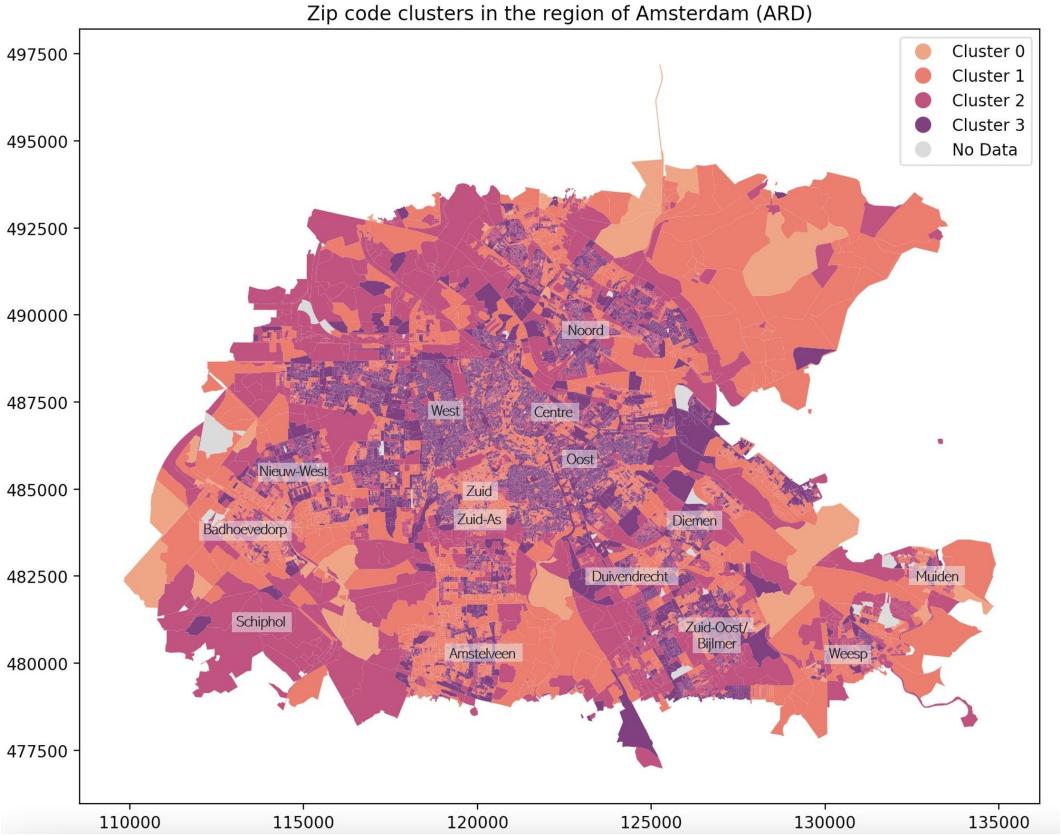
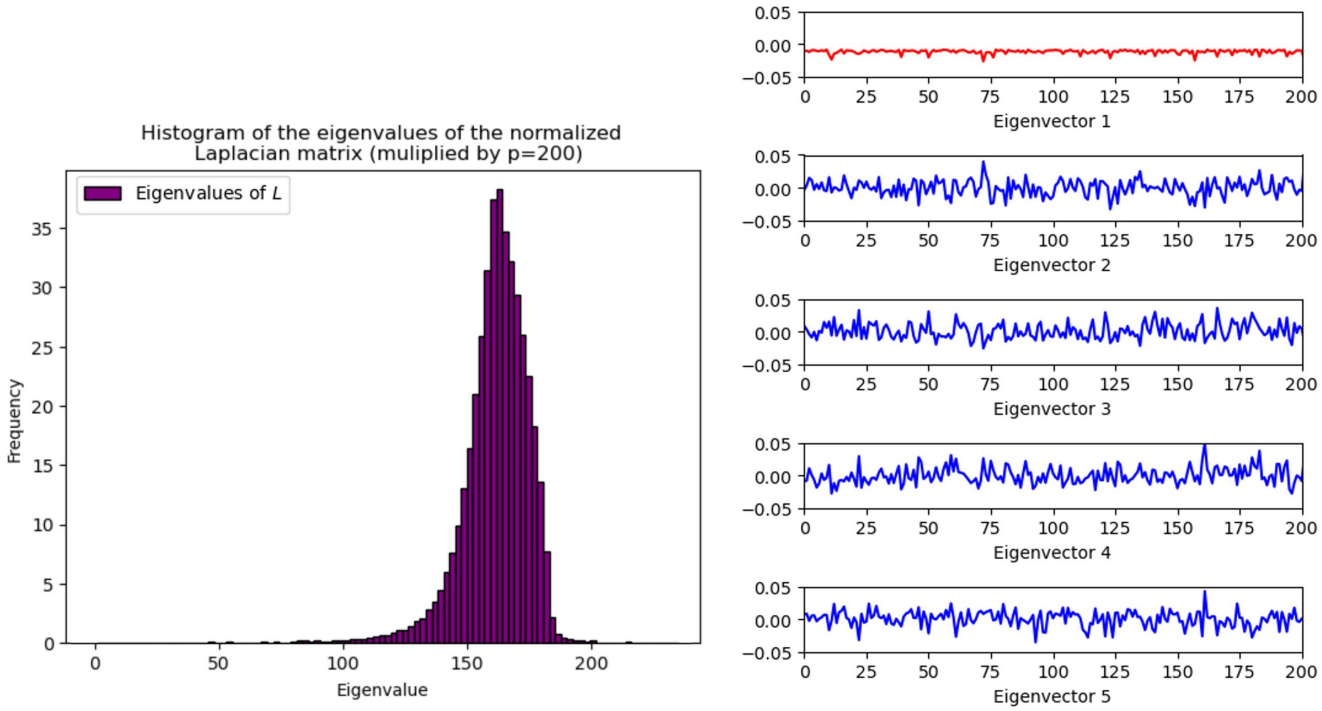


Figure 5.17: This figure shows the zip code clusters of Amsterdam and its surrounding area.

Evaluation: According to experts in the actuarial field, the clusters make sense. For example, cars belonging to urban areas tend to exhibit a higher average claim frequency compared to those belonging to rural areas.

Modified spectral clustering method

The ARD zip code dataset consists of $n = 32,217$ data points. Therefore, the *U-SPEC* method was again applied with $n \gg p' = 2000$ to ensure a sufficient number of rep-clusters z_1. Moreover, p, z_1, k, and k' are as defined for the ARD license plate dataset. These parameters yield a high-dimensional dataset since the ratio of dimensions to observations is equal to $114/p = 114/200 = 0.57$. Therefore, the number of clusters is again determined with the method outlined in Subsection 4.3.3.
Figure 5.18a shows a histogram of the eigenvalues of the normalized Laplacian, multiplied by $p = 200$, with *U-SPEC*. In Figure 5.18b, the eigenvectors corresponding to the five isolated eigenvalues of this histogram are depicted. Since eigenvector 1 is non-informative, there are four informative eigenvectors, indicating that the optimal number of clusters is *at most* four.



(a) **(b)**

Figure 5.18: This figure shows, for the ARD zip code dataset: **(a)** a histogram of the eigenvalues of the normalized Laplacian (multiplied by $p = 200$), **(b)** the eigenvectors corresponding to the five isolated eigenvalues of the histogram (eigenvector 1 is non-informative and shown in red).

The *U-SPEC* algorithm is completed with four clusters and Figure 5.19 shows the box plots of each of these clusters regarding the claim frequency. The distinct averages and variations observed in these box plots indicate that the clusters are significant in terms of claim frequency, making them suitable for inclusion as risk factors in the GLM.

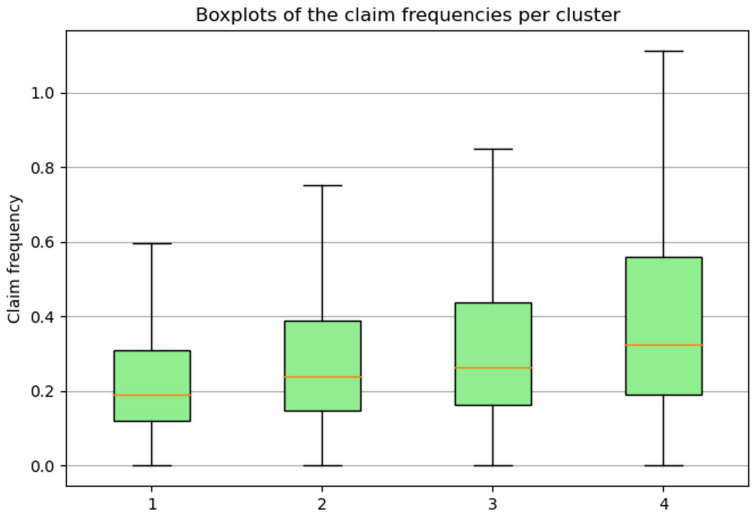


Figure 5.19: This figure shows the box plots of each cluster regarding the claim frequency. It can be observed that every cluster has a distinct average and variation in its box plot.

Figure 5.20 displays the box plots of each cluster's claim frequency for both the K-prototypes (in blue) and spectral clustering (in green) methods. For all clusters, the variation in claim frequency is greater with K-prototypes. Therefore, it can be concluded that spectral clustering more effectively maximizes the homogeneity among observations within the same cluster.

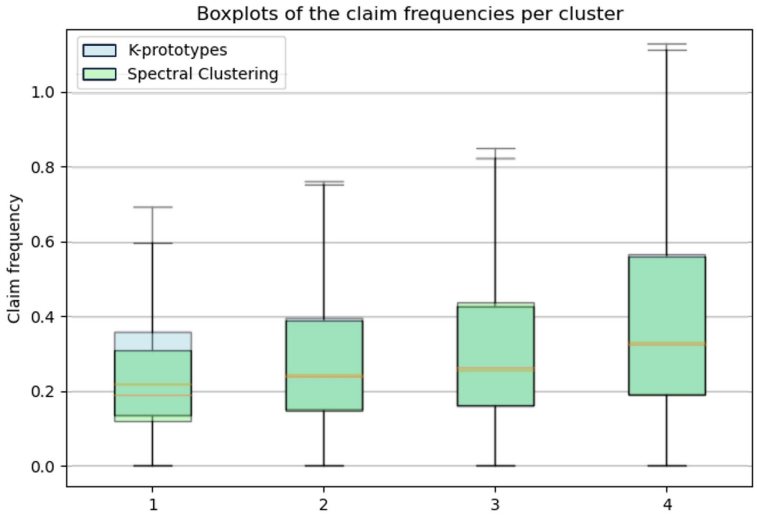


Figure 5.20: This figure shows the box plots of each cluster's claim frequency for both the K-prototypes (in blue) and spectral clustering (in green) methods.

Figure 5.21 presents a table of the descriptions of all clusters. For each cluster, the number of zip codes, average claim frequency, and the class of average claim frequency are also provided. The descriptions of the clusters can be summarized as follows:

- **Cluster 0**: Rural areas in the northern part of the country.
- **Cluster 1**: Rich suburbs with a high level of education.
- **Cluster 2**: Rural areas in the southern part of the country.
- **Cluster 3**: Urban areas with a high population density.

Cluster	Description	#Zip codes	Claim frequency	
			Average	Class
0	Fintype: advice-sensitive families, geotype: creative environment lovers, hometype: older home owners. Low urbanisation and large distances to supermarkets and banks. Older home owners, intermediate social class, and least amount of non-western immigrants. Intermediate income, education and age. Has the biggest and oldest cars, the greatest living area, and the most motor cycles. Highest WOZ value and the most common vacation region is north.	[redacted]	0.236	***Lowest***
1	Contains most zip codes, fintype: financial professionals, geotype: intellectual culture lovers, hometype: unknown. Intermediate urbanisation and age, and highest social class. Highest percentage of company cars, income, education and couples with (older) kids. Most owner occupied houses, households with two earners, investors, and "Glossy" readers. Newest cars and households spend the most on groceries. Lowest chance of defaulting and highest living area after cluster 0. WOZ is unknown and the most common vacation region is middle.	[redacted]	0.297	Intermediate
2	Fintype: advice sensitive families, geotype: conservative benefactors, hometype: older home owners. Same social class and income as cluster 0, but newer houses, a lower WOZ value, a smaller living area, a higher average age, and less investors. Moreover, cars drive less km than in cluster 0, and smaller distances, smaller cars, and less educated people. Not a lot of non-western immigrants and a low urbanisation. The most common vacation region is south.	[redacted]	0.332	Intermediate
3	Fintype: active borrowers, geotype: average sport fans, hometype: single apartment tenants. Highest urbanisation, lowest social class, most non-western immigrants, lowest incomes, lowest education, and oldest houses. Lowest age (of people), smallest living spaces, most borrowers, highest chance of changing health insurance, most sensitive to switching car brands, and highest risk of defaulting. Spends the least on daily groceries, not a lot of car ownership, and oldest, smallest and cheapest cars. The most common vacation region is north.	[redacted]	0.433	***Highest***

Figure 5.21: This figure shows a table of the descriptions for all clusters. The number of license plates, average claim frequency, and class of average claim frequency are also provided for every cluster.

Figure 5.22 shows the four-digit zip code clusters of the Netherlands generated by K-prototypes and spectral clustering. Notably, clusters 1 and 3 appear nearly identical on both maps. However, cluster 2 from the K-prototypes map disappears in the spectral clustering one. Instead, the rural cluster 0 from K-prototypes is divided into two separate clusters for *U-SPEC*: a rural north cluster (cluster 0) and a rural south cluster (cluster 2).

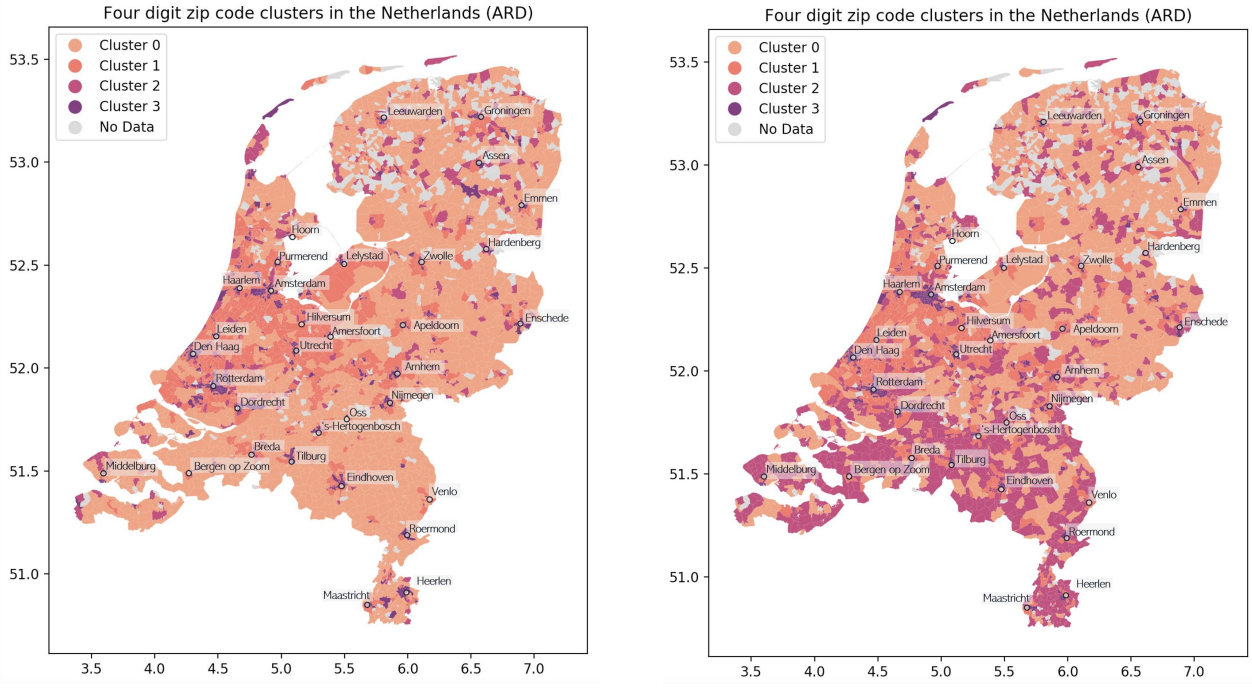
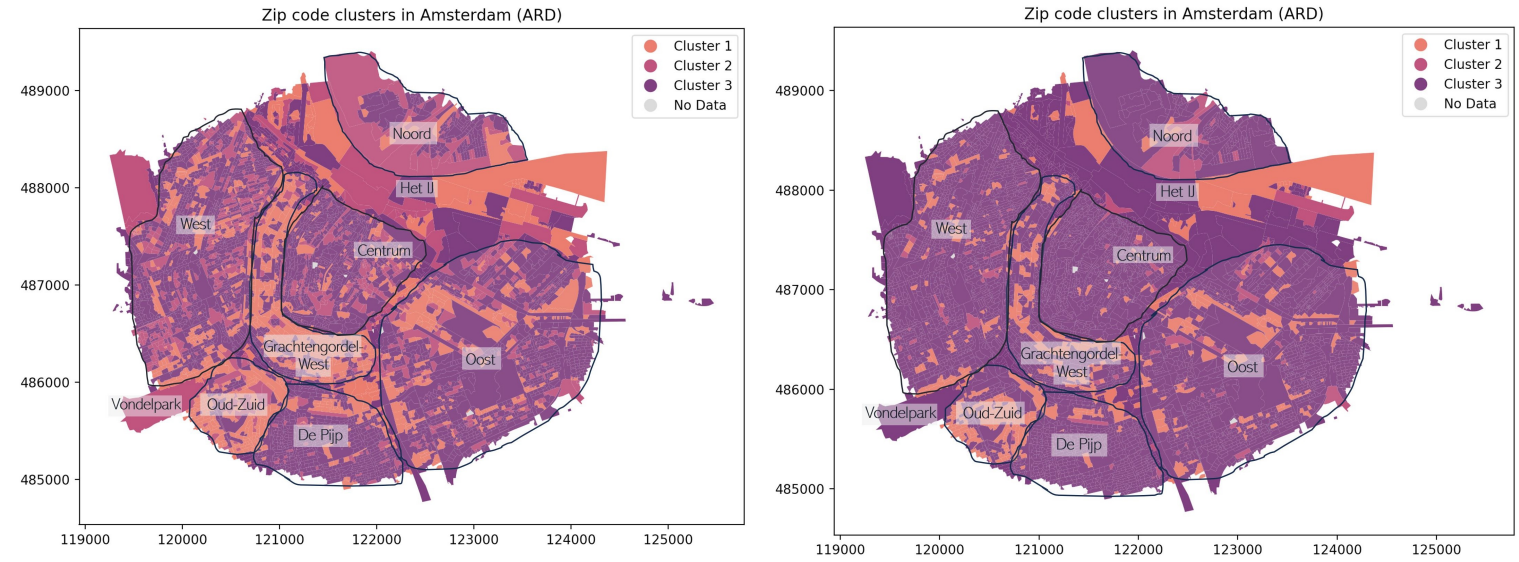


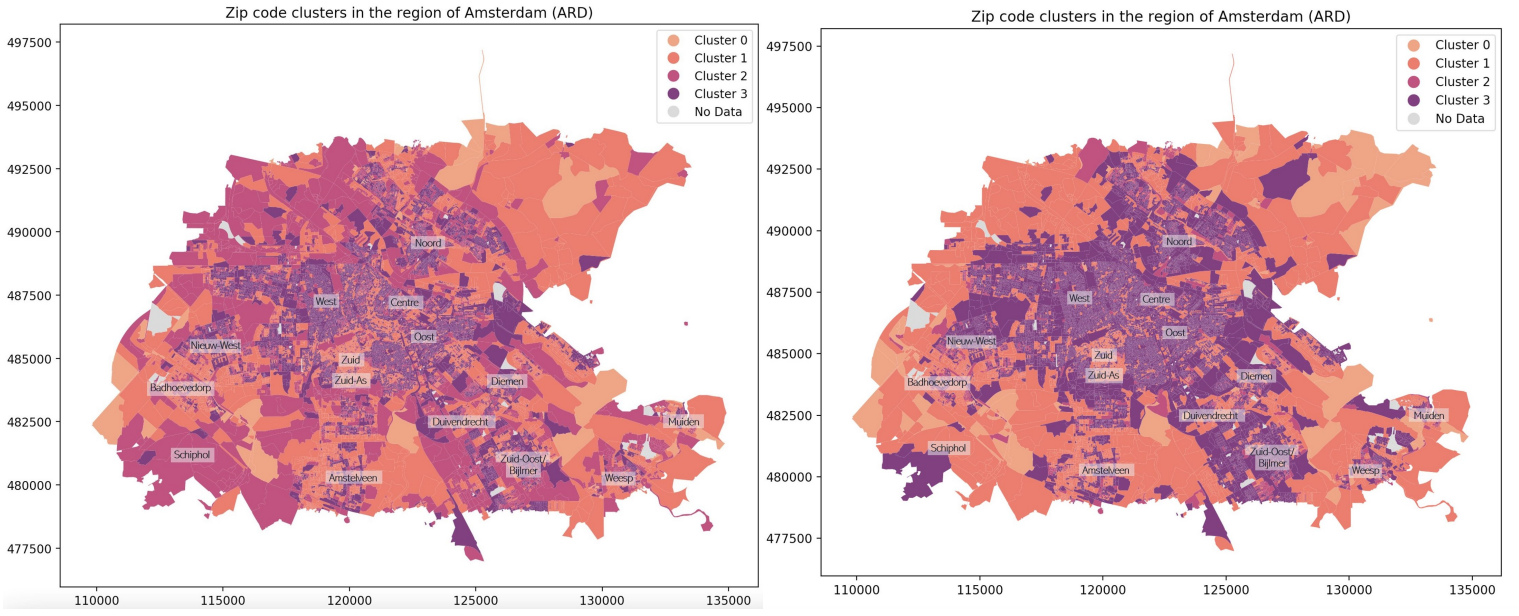
Figure 5.22: This figure shows the four-digit zip code clusters of the Netherlands generated by: **(a)** K-prototypes, **(b)** Spectral clustering.

Figure 5.23 shows the zip code clusters of Amsterdam produced with K-prototypes and spectral clustering, while Figure 5.24 extends this comparison to Amsterdam and its surrounding area. The spectral clustering algorithm seems to produce more homogeneous clusters. For instance, in Figure 5.23, the "West" area exhibits more noise with the K-prototypes method compared to spectral clustering. Nevertheless, upscale neighborhoods such as "Grachtengordel West" and "Oud-Zuid" belong to cluster 1 for both methods.



(a) (b)

Figure 5.23: This figure shows the zip code clusters of Amsterdam produced with: **(a)** K-prototypes, **(b)** Spectral clustering.



(a) (b)

Figure 5.24: This figure shows the zip code clusters of Amsterdam and its surrounding area produced with: **(a)** K-prototypes, **(b)** Spectral clustering.

Evaluation: Similar to the K-prototypes clusters, the clusters generated by the spectral clustering method are considered logical by experts in the actuarial field. For example, cars belonging to urban areas tend to exhibit a higher average claim frequency compared to those belonging to rural areas. Furthermore, in practice, the claim frequency is higher in the south of the Netherlands than in the north, just as Figure 5.22b shows. Nevertheless, it is difficult to determine which method produces more logical groups. Therefore, a quantitative comparison is provided in Subsection 5.2.3.

5.2.3. Comparison with the current risk classification approach for ARD

As explained in Subsection 4.4.4, the clusters are incorporated as additional risk factors in the existing claim frequency GLM. For both clustering methods, the standard errors of the parameters are given and the deviance, AICc, and BIC are provided.

K-prototypes method

Figure 5.25 shows a table of the estimated differences in claim frequencies with respect to the null clusters for both the license plate ("KT") and zip code ("ZC") clusters of the ARD dataset for K-prototypes. The standard errors and standard error percentages are also displayed. It is evident that, since the standard error percentages of license plate clusters 1 and 2 exceed 50%, these clusters do not significantly differ from the null-cluster (here license plate cluster 0) and thus do not impact the model. Therefore, these two clusters are excluded from the risk prediction in the GLM model.

Name	Value	Standard Error	Standard Error (%)
Cluster_K_KT (0)	/	/	/
Cluster_K_KT (1)	0.0371	0.02219	59.8
Cluster_K_KT (2)	-0.4325	0.38854	89.8
Cluster_K_KT (3)	0.0802	0.02986	37.2
Cluster_K_KT (4)	0.1629	0.02099	12.9
Cluster_K_KT (5)	0.1714	0.03150	18.4
Cluster_K_KT (6)	0.0497	0.01881	37.9
Cluster_K_KT (7)	0.2054	0.01891	9.2
Cluster_K_KT (8)	0.2501	0.02392	9.6
Cluster_K_ZC (0)	0.0277	0.01297	46.8
Cluster_K_ZC (1)	/	/	/
Cluster_K_ZC (2)	0.0518	0.01505	29.0
Cluster_K_ZC (3)	0.1343	0.01714	12.8

Figure 5.25: This figure shows a table of the estimated differences in claim frequencies with respect to the null clusters for both the license plate ("KT") and zip code ("ZC") clusters for K-prototypes. The standard errors and standard error percentages are also displayed. Green and red percentages indicate significant and insignificant clusters respectively.

Figure 5.26 shows a table comparing the deviance, AICc, and BIC of the current GLM to those of the GLM with the *significant* K-prototypes clusters as risk factors (thus excluding license plate clusters 1 and 2). The modified GLM shows a lower deviance, indicating a better fit. However, both the AICc and BIC of the modified GLM are higher than that of the current GLM. Therefore, incorporating the K-prototypes clusters does not improve the current risk classification.

	GLM with K-prototypes clusters	Current GLM
Deviance	[redacted]	[redacted]
AICc	291,878.4	291,407.6
BIC	293,081.8	292,683.7

Figure 5.26: This figure shows a table comparing the deviance, AICc, and BIC of the current GLM to those of the GLM with the (significant) K-prototypes clusters as risk factors.

Modified spectral clustering method

Figure 5.27 shows a table of the estimated differences in claim frequencies with respect to the null clusters for both the license plate and zip code clusters of the ARD dataset for modified spectral clustering. The standard errors and standard error percentages are also displayed. It is evident that, since the standard error percentage of zip code cluster 0 exceeds 50%, this cluster does not significantly differ from the null-cluster (here zip code cluster 1) and thus does not impact the model. Therefore, this cluster is excluded from the model.

Name	Value	Standard Error	Standard Error (%)
Cluster_Spec_KT (0)	-0.2325	0.01824	7.9
Cluster_Spec_KT (1)	/	/	/
Cluster_Spec_KT (2)	-0.7639	0.02982	3.9
Cluster_Spec_KT (3)	-0.0549	0.01768	32.2
Cluster_Spec_KT (4)	-0.1525	0.01914	12.6
Cluster_Spec_KT (5)	-0.1120	0.02371	21.2
Cluster_Spec_KT (6)	-0.1901	0.01813	9.5
Cluster_Spec_KT (7)	-0.2029	0.01797	8.8
Cluster_Spec_ZC (0)	-0.0057	0.01545	271.3
Cluster_Spec_ZC (1)	/	/	/
Cluster_Spec_ZC (2)	0.0988	0.01277	12.9
Cluster_Spec_ZC (3)	0.1252	0.01619	12.9

Figure 5.27: This figure shows a table of the estimated differences in claim frequencies with respect to the null clusters for both the license plate and zip code clusters for modified spectral clustering. The standard errors and standard error percentages are also displayed. Green and red percentages indicate significant and insignificant clusters respectively.

Figure 5.28 shows a table comparing the deviance, AICc, and BIC of the current GLM to those of the GLM with the *significant* spectral clusters as risk factors. The modified GLM shows a lower deviance, indicating a better fit. Furthermore, both the AICc and BIC of the modified GLM are lower than that of the current GLM. Therefore, incorporating the spectral clusters improves the current risk classification.

	GLM with spectral clusters	Current GLM
Deviance	[redacted]	[redacted]
AICc	291,124.6	291,407.6
BIC	292,491.3	292,683.7

Figure 5.28: This figure shows a table comparing the deviance, AICc, and BIC of the current GLM to those of the GLM with the (significant) spectral clusters as risk factors.

5.2.4. Clustering of the WAM dataset

In this subsection, a summary is provided of the clustering results of the WAM dataset. The detailed results are available in Appendix A.

For the license plates, both the K-prototypes and spectral clustering results are considered logical by experts in the actuarial field. Furthermore, the K-prototypes clusters generally exhibit greater variation in claim frequency, indicating that spectral clustering more effectively maximizes the homogeneity among observations within the same cluster. Nevertheless, it is difficult to determine which method produces more logical groups. Therefore, a quantitative comparison is provided in Subsection 5.2.5.

The zip code clusters generated by the spectral clustering method are considered logical by experts in the actuarial field, whereas the K-prototypes clusters lack practical relevance. Therefore, it can be

concluded that spectral clustering produces more accurate clusters than K-prototypes in this case. Additionally, the K-prototypes clusters generally exhibit greater variation in claim frequency. A quantitative comparison is provided in the next subsection.

5.2.5. Comparison with the current risk classification approach for WAM

As explained in Subsection 4.4.4, the clusters are incorporated as additional risk factors in the existing claim frequency GLM. For both clustering methods, the standard errors of the parameters are given and the deviance, AICc, and BIC are provided.

K-prototypes method

Figure 5.29 shows a table of the estimated differences in claim frequencies with respect to the null clusters for both the license plate ("KT") and zip code ("ZC") clusters of the WAM dataset for K-prototypes. The standard errors and standard error percentages are also displayed. Since the standard error percentage of zip code cluster 4 exceeds 50%, the cluster does not significantly differ from the null-cluster (here zip code cluster 0) and thus does not impact the model. When clusters 0 and 4 are combined, the standard error percentage drops below 50% (see Figure 5.29) and thus is included in the risk prediction in the GLM model.
Note that all other zip code clusters are significant with respect to the null-cluster. Therefore, cluster 4 can be combined with any of the other clusters, not just cluster 0.

Name	Value	Standard Error	Standard Error (%)
Cluster_K_KT (0)	/	/	/
Cluster_K_KT (1)	0.0876	0.02200	25.1
Cluster_K_KT (2)	0.3742	0.02258	6.0
Cluster_K_KT (3)	0.1929	0.02084	10.8
Cluster_K_ZC (0), (4)	/	/	/
Cluster_K_ZC (1)	0.0817	0.02015	24.7
Cluster_K_ZC (2)	-0.0597	0.01767	29.6
Cluster_K_ZC (5)	0.1424	0.03031	21.3

Figure 5.29: This figure shows a table of the estimated differences in claim frequencies with respect to the null clusters for both the license plate ("KT") and zip code ("ZC") clusters for K-prototypes. The standard errors and standard error percentages are also displayed. Green and red percentages indicate significant and insignificant clusters respectively.

Figure 5.30 shows a table comparing the deviance, AICc, and BIC of the current GLM to those of the GLM with the *significant* K-prototypes clusters as risk factors. The modified GLM shows a lower deviance, indicating a better fit. However, both the AICc and BIC of the modified GLM are higher than that of the current GLM. Therefore, incorporating the K-prototypes clusters does not improve the current risk classification.

	GLM with K-prototypes clusters	Current GLM
Deviance	[redacted]	[redacted]
AICc	198,298.4	196,843.1
BIC	198,859.4	197,332.5

Figure 5.30: This figure shows a table comparing the deviance, AICc, and BIC of the current GLM to those of the GLM with the (significant) K-prototypes clusters as risk factors.

Modified spectral clustering method

Figure 5.31 shows a table of the estimated differences in claim frequencies with respect to the null clusters for both the license plate and zip code clusters of the WAM dataset for modified spectral clustering. The standard errors and standard error percentages are also displayed. Again, since the standard error percentage of zip code cluster 4 exceeds 50%, the cluster does not significantly differ

from the null-cluster (here zip code cluster 1) and thus does not impact the model. When clusters 0 and 4 are combined, the standard error percentage drops below 50% (see Figure 5.31) and thus is included in the risk prediction in the GLM model.

Name	Value	Standard Error	Standard Error (%)
Cluster_Spec_KT (0)	0.9576	0.02060	2.2
Cluster_Spec_KT (1)	0.6792	0.02172	3.2
Cluster_Spec_KT (2)	0.5032	0.02181	4.3
Cluster_Spec_KT (3)	/	/	/
Cluster_Spec_ZC (0), (4)	0.1145	0.01730	15.1
Cluster_Spec_ZC (1)	/	/	/
Cluster_Spec_ZC (2)	0.1739	0.01906	11.0

Figure 5.31: This figure shows a table of the estimated differences in claim frequencies with respect to the null clusters for both the license plate and zip code clusters for modified spectral clustering. The standard errors and standard error percentages are also displayed. Green and red percentages indicate significant and insignificant clusters respectively.

Figure 5.32 shows a table comparing the deviance, AICc, and BIC of the current GLM to those of the GLM with the *significant* spectral clusters as risk factors. The modified GLM shows a lower deviance, indicating a better fit. Furthermore, the deviance is lower for the GLM with spectral clusters than with K-prototypes clusters (see Figure 5.30). Both the AICc and BIC of the modified spectral GLM are higher than that of the current GLM. Therefore, incorporating the spectral clusters does not improve the current risk classification.

	GLM with spectral clusters	Current GLM
Deviance	[redacted]	[redacted]
AICc	199,713.4	196,843.1
BIC	200,262.4	197,332.5

Figure 5.32: This figure shows a table comparing the deviance, AICc, and BIC of the current GLM to those of the GLM with the (significant) spectral clusters as risk factors.

Thus, although the spectral clusters improved the GLM for the ARD dataset, neither the K-prototypes clusters nor the spectral clusters enhanced the current GLM for the WAM dataset. One possible explanation is that the claim frequencies for the ARD dataset are influenced by features related to license plates and zip codes. For instance, ARD coverage is not legally required, so claims are primarily made by owners of new and expensive cars, who typically live in rich suburbs. Conversely, WAM coverage is mandatory, and claim frequencies are more closely tied to driver characteristics (e.g. age and gender) than to license plates and zip codes. Therefore, incorporating license plate and zip code clusters into the GLM for the WAM dataset worsens its performance.

5.3. Stability of the results

In this section, the stability of the cluster results of the previous section are analyzed. Subsection 5.3.1 discusses the time stability of the results, while Subsection 5.3.2 treats the stability of the results with respect to the number of observations.

5.3.1. Time stability of the results

As explained in Subsection 4.5.1, to evaluate the time stabilities of the results, interaction terms of the binary cluster variables and the policy year are added to the modified GLM of Subsection 5.2.3. These terms are added one at a time to test if, for every policy year, the same effect on the claim frequency can be observed for each cluster. If the effects over time are random (i.e. there is no trend) and if the claim frequencies fluctuate around a mean, the cluster is considered stable over time. Therefore, the risk factor of such a cluster remains included in the model. If a cluster is unstable over time, it should

be removed from the GLM.
Note that the time stability of the clusters is not a performance metric for the clustering techniques themselves. Instead, it assesses the applicability of the clusters in the GLM and, consequently, their suitability for premium pricing. For this pricing, it is crucial that the estimated claim frequencies remain stable over time and are predictive for the upcoming years.

ARD dataset

K-prototypes method:
Figure 5.33 shows the changes in claim frequency over the policy years per license plate cluster for the ARD dataset and with the K-prototypes method on the left y-axis. The yellow bars in the figure represent the exposures for each policy year (as defined in Subsection 3.2.1), expressed as a percentage of the total exposure on the right y-axis. Clusters 4 and 7 are the only clusters that do not exhibit any trends in their claim frequencies *and* that fluctuate around a mean frequency over time. Therefore, these two clusters are the only ones suitable for inclusion in the GLM for K-prototypes.
Note that cluster 0 is the null-cluster. Therefore, it does not display changes in claim frequency over time. Additionally, cluster 2 only contains eighteen license plates, making it sensitive to changes in policy year due to the limited data points available per year. As a consequence, this cluster shows significant volatility.

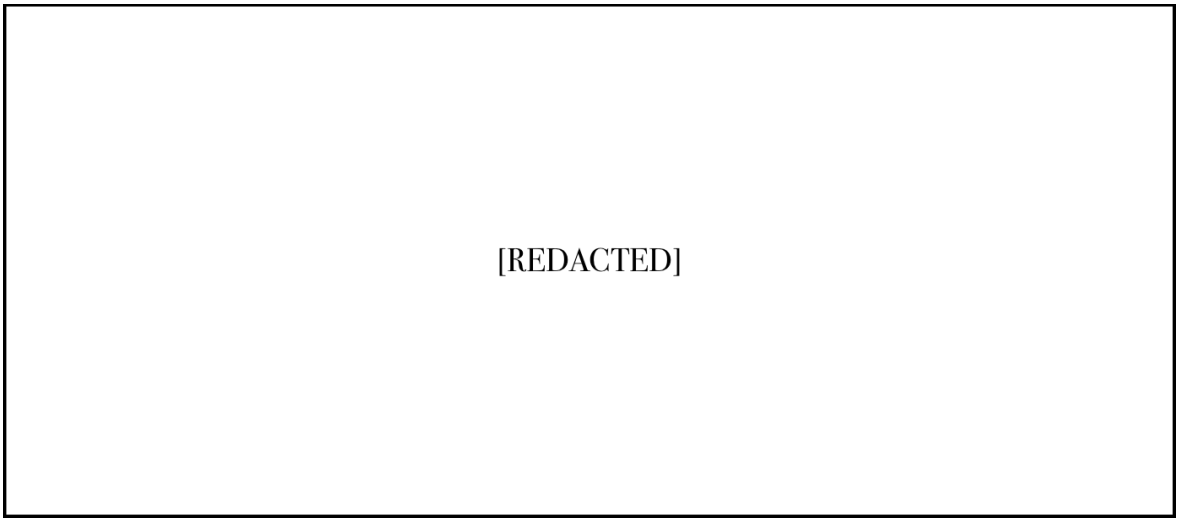


Figure 5.33: This figure shows the changes in claim frequency over the policy years per license plate cluster for the ARD dataset with the K-prototypes method on the left y-axis. The yellow bars represent the exposures for each policy year, expressed as a percentage of the total exposure on the right y-axis.

Figure 5.34 shows the changes in claim frequency over the policy years per *zip code* cluster for the ARD dataset and with the K-prototypes method. Cluster 3 is the only cluster that appears to have a stable effect over time. Therefore, this cluster is the only one suitable for inclusion in the GLM for K-prototypes.

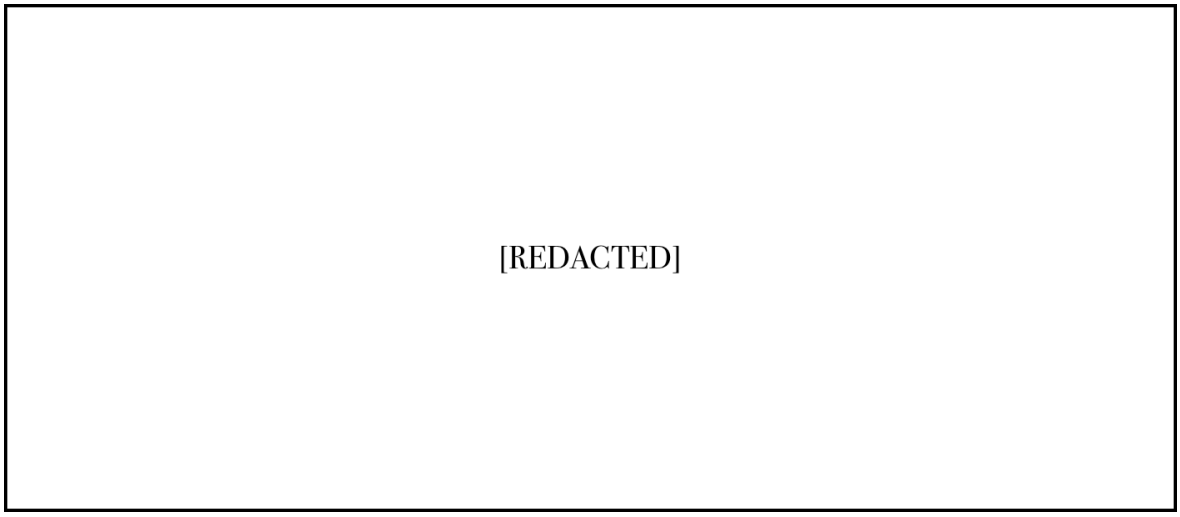


Figure 5.34: This figure shows the changes in claim frequency over the policy years per zip code cluster for the ARD dataset with the K-prototypes method on the left y-axis. The yellow bars represent the exposures for each policy year, expressed as a percentage of the total exposure on the right y-axis.

Modified spectral clustering method:
Figure 5.35 shows the changes in claim frequency over the policy years per license plate cluster for the ARD dataset and with the modified spectral clustering method. None of the eight clusters appear stable over time as they exhibit downward trends. Therefore, all clusters should be excluded from the GLM for spectral clustering. However, it is worth noting that the license plate clusters are more stable with the spectral clustering method compared to the K-prototypes method. Thus, the spectral license plate clusters can be included in the GLM, provided that the time dependency of the variables used for the clustering is carefully considered. This topic will be further discussed in Chapter 7.

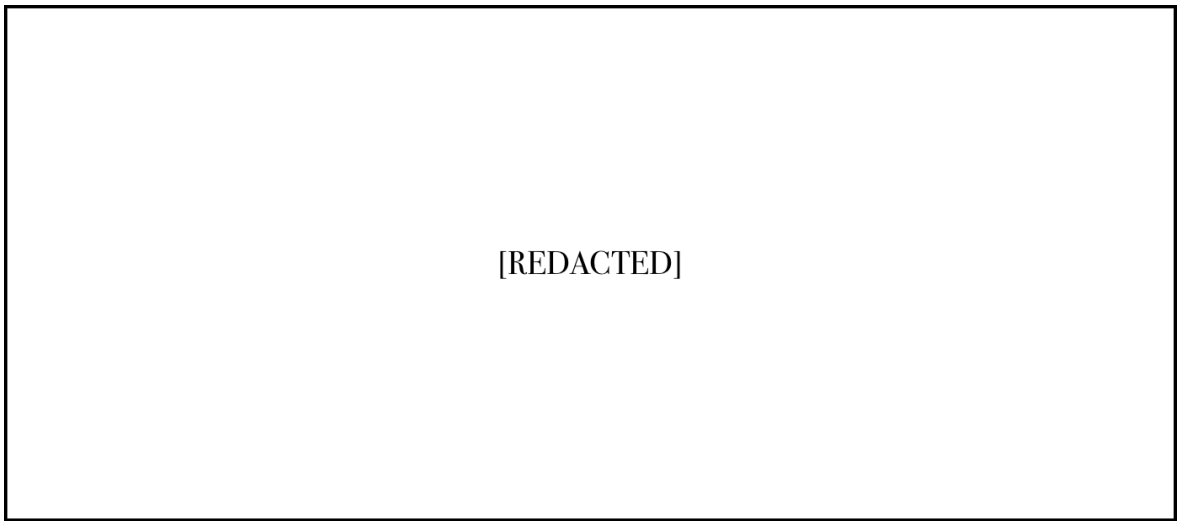


Figure 5.35: This figure shows the changes in claim frequency over the policy years per license plate cluster for the ARD dataset with the modified spectral clustering method on the left y-axis. The yellow bars represent the exposures for each policy year, expressed as a percentage of the total exposure on the right y-axis.

Figure 5.36 shows the changes in claim frequency over the policy years per zip code cluster for the ARD dataset and with the modified spectral clustering method. Clusters 2 and 3 are the only clusters that appear to have a stable effect over time. Therefore, these clusters are the only ones suitable for inclusion in the GLM for spectral clustering.
Note that cluster 0 was considered insignificant with respect to cluster 1 (i.e. the null-cluster), as discussed in Subsection 5.2.3. This is also evident in the graph since cluster 0 fluctuates around the null-cluster over time.

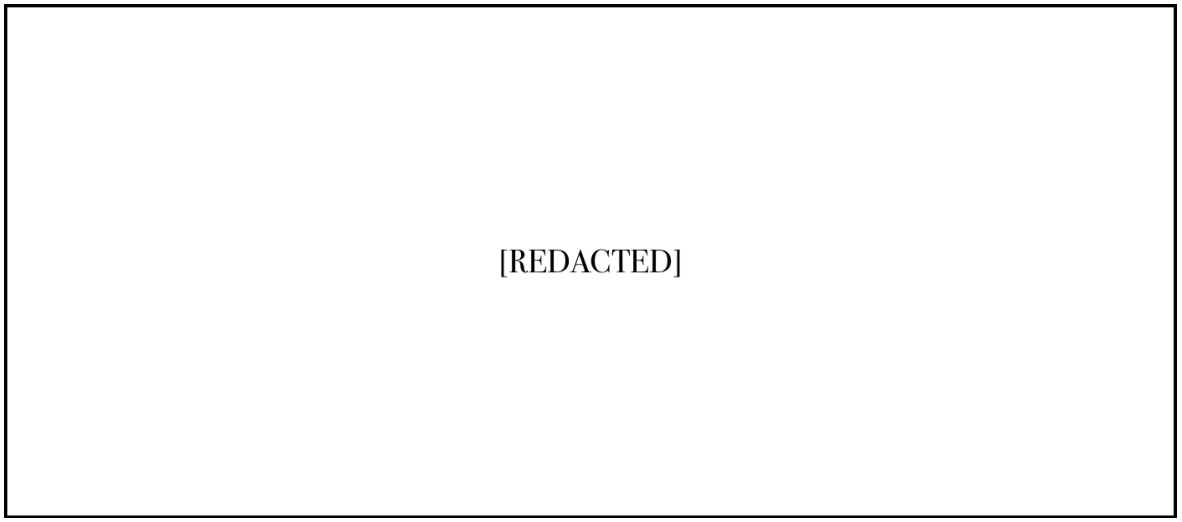


Figure 5.36: This figure shows the changes in claim frequency over the policy years per zip code cluster for the ARD dataset with the modified spectral clustering method on the left y-axis. The yellow bars represent the exposures for each policy year, expressed as a percentage of the total exposure on the right y-axis.

WAM dataset

K-prototypes method:
Since the K-prototypes method does not improve the current GLM and since the spectral clustering method outperforms K-prototypes (as explained in Appendix A and Chapter 6), the time stability of the K-prototypes clusters is not evaluated for the WAM dataset.

Modified spectral clustering method:
Figure 5.37 shows the changes in claim frequency over the policy years per license plate cluster for the WAM dataset and with the modified spectral clustering method. None of the four clusters appear stable over time since they do not fluctuate around a mean frequency. Therefore, all clusters should be excluded from the GLM for spectral clustering.

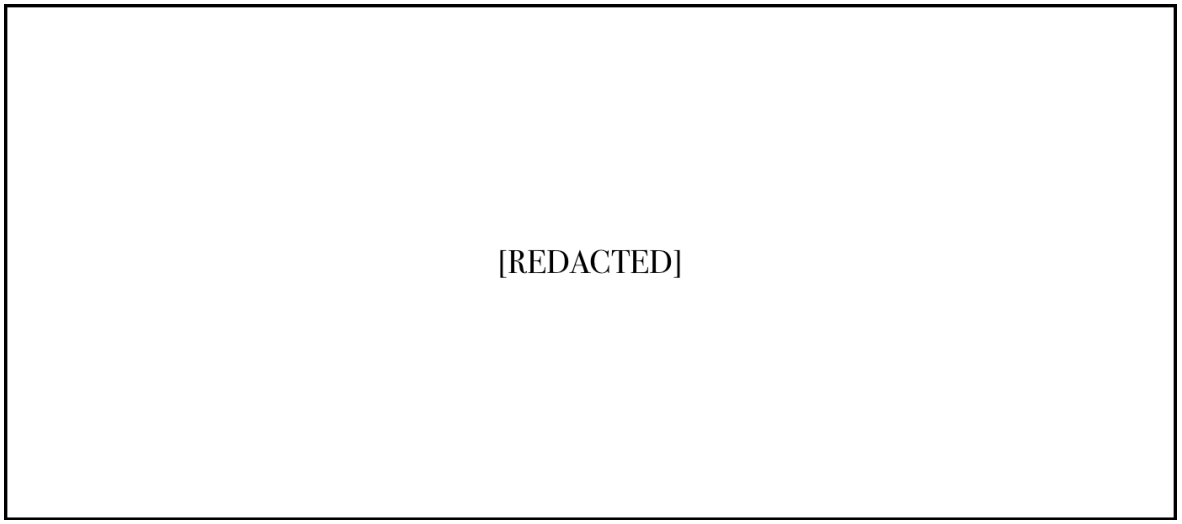


Figure 5.37: This figure shows the changes in claim frequency over the policy years per license plate cluster for the WAM dataset with the modified spectral clustering method on the left y-axis. The yellow bars represent the exposures for each policy year, expressed as a percentage of the total exposure on the right y-axis.

Figure 5.38 shows the changes in claim frequency over the policy years per zip code cluster for the WAM dataset and with the modified spectral clustering method. As explained in Subsection 5.2.5, clusters 0 and 3 are combined to form a new cluster 0 since cluster 3 did not significantly differ from

cluster 0. None of the clusters in Figure 5.38 appear to have a stable effect over time. Therefore, all clusters should be excluded from the GLM for spectral clustering. Lastly, note that the license plate clusters are more stable over time than the zip code clusters for the WAM dataset.

[REDACTED]

Figure 5.38: This figure shows the changes in claim frequency over the policy years per zip code cluster for the WAM dataset with the modified spectral clustering method on the left y-axis. The yellow bars represent the exposures for each policy year, expressed as a percentage of the total exposure on the right y-axis.

5.3.2. Stability of the results with respect to the number of observations

ARD dataset

License plate data:
Figure 5.39 shows the Rand indices (relative to the $U\text{-}SPEC$ clusters with $p' = 2000$) over various values of p' for the ARD license plate dataset. The indices lie between 0.75 and 0.79 for all p' values. Additionally, the two distinct clustering results for $p' = 2000$, used as a consistency check, yield an index of 0.79. Therefore, the clustering results remain relatively consistent across different values of p'.

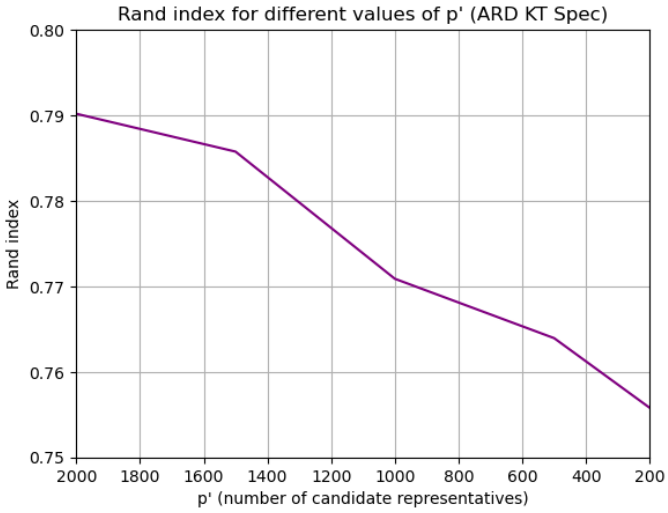


Figure 5.39: This figure shows the Rand indices (relative to the $U\text{-}SPEC$ clusters with $p' = 2000$) over various values of p' for the ARD license plate dataset.

Figure 5.40 shows the box plots of the claim frequencies for each p' value and across the different license plate clusters in the ARD dataset. For most clusters, the box plots for each p' value have similar means and variations in claim frequency. However, cluster 7 exhibits notable differences in its

box plots for different p' values, especially when $p' = 200$. This implies that when there is already significant variation in the box plot for $p' = 2000$, the box plots will differ considerably for other p' values. Therefore, for cluster 7, the *U-SPEC* algorithm is likely to generate different clusters for each p' value.

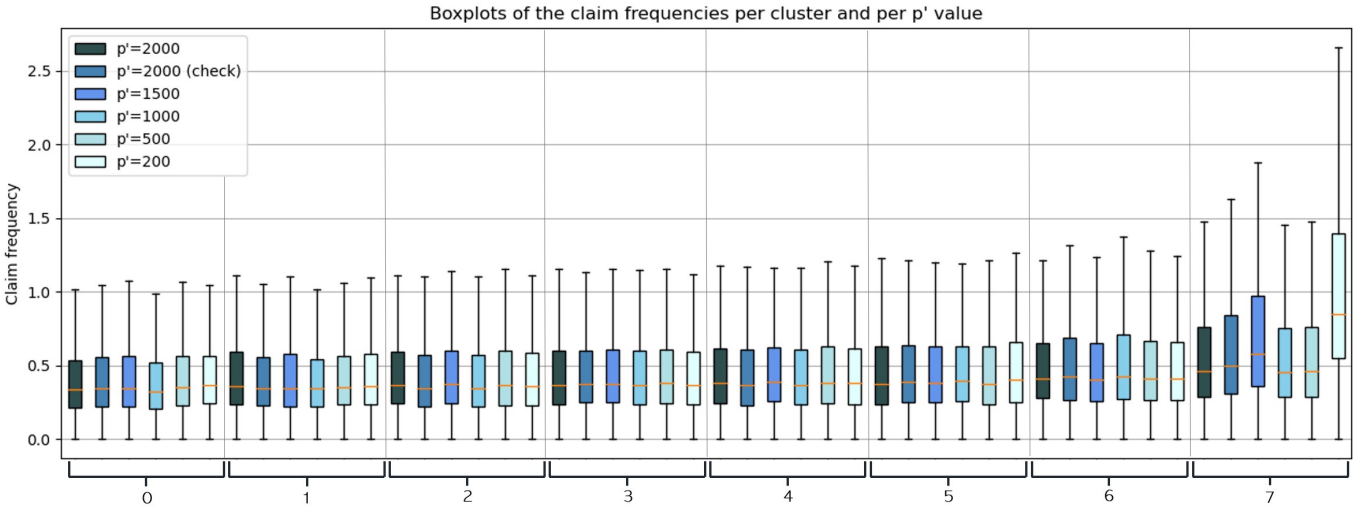


Figure 5.40: This figure shows the box plots of the claim frequencies for each p' value and across the different license plate clusters in the ARD dataset.

Zip code data:
Figure 5.41 shows the Rand indices (relative to the *U-SPEC* clusters with $p' = 2000$) over various values of p' for the ARD zip code dataset. The indices lie between 0.76 and 0.89 for all p' values. Additionally, the two distinct clustering results for $p' = 2000$, used as a consistency check, yield an index of 0.89. Therefore, the clustering results remain relatively stable across different values of p'. However, for $p' = 200$, there is a significant drop in the Rand index, suggesting that the clusters differ more from those produced with $p' = 2000$ compared to other values of p'.

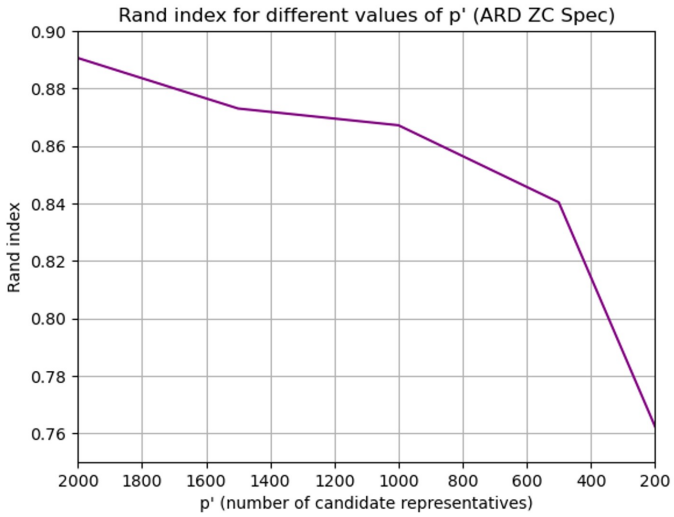


Figure 5.41: This figure shows the Rand indices (relative to the *U-SPEC* clusters with $p' = 2000$) over various values of p' for the ARD zip code dataset.

Figure 5.42 shows the box plots of the claim frequencies for each p' value and across the different zip code clusters in the ARD dataset. For all clusters, the box plots for each p' value have similar means and variations in claim frequency. Nevertheless, cluster 2 appears to be the least stable with respect to p', indicating that this cluster is most likely to differ when p' is changed.

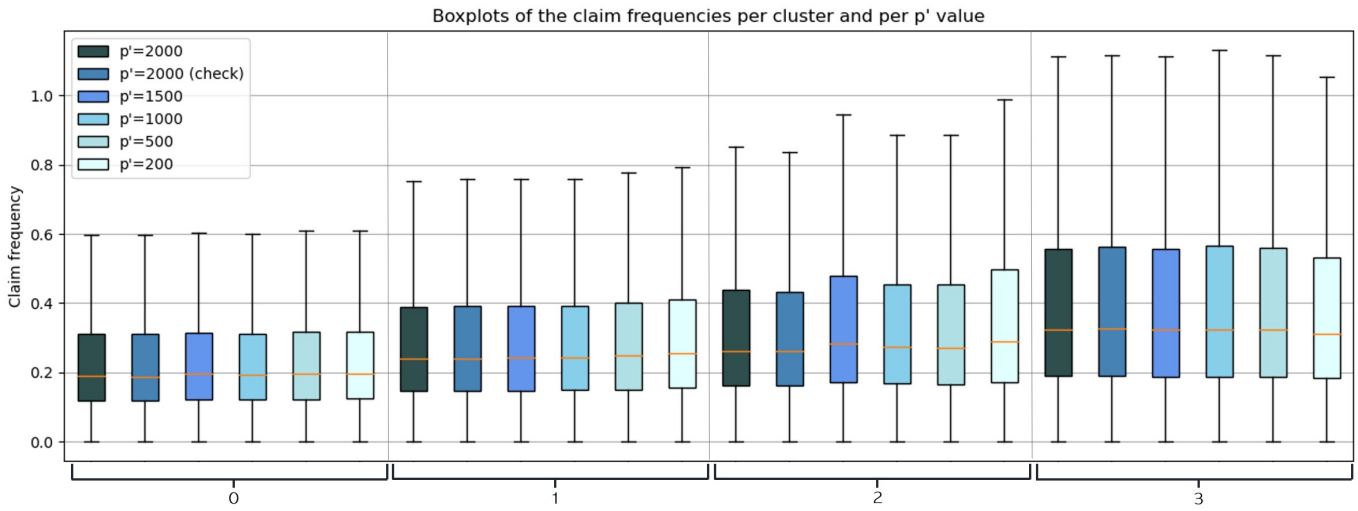


Figure 5.42: This figure shows the box plots of the claim frequencies for each p' value and across the different zip code clusters in the ARD dataset.

Figure 5.43 shows the four-digit zip code clusters of the Netherlands generated by *U-SPEC* with the different values of p'. Clusters 1 and 3 remain stable across varying numbers of observations, while cluster 2 becomes less prevalent in the south as p' decreases. Note that cluster 2 was anticipated to be the most likely to differ when p' is altered as previously discussed. Furthermore, the clustering results for $p' = 200$ resemble those of K-prototypes.

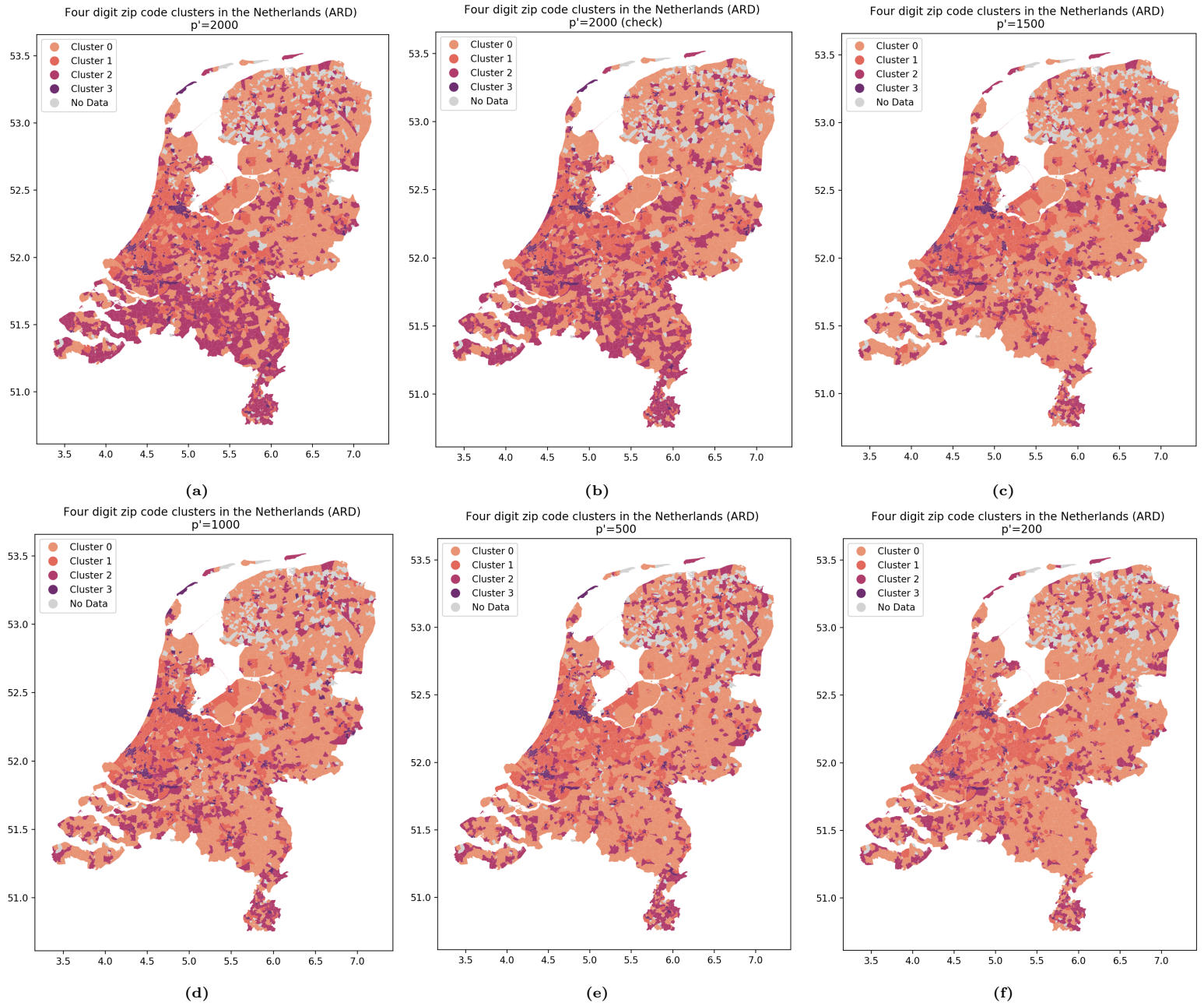


Figure 5.43: This figure shows the four-digit zip code clusters of the Netherlands generated by *U-SPEC* with: **(a)** $p' = 2000$, **(b)** $p' = 2000$ (to check the consistency), **(c)** $p' = 1500$, **(d)** $p' = 1000$, **(e)** $p' = 500$, **(f)** $p' = 200$.

To identify the value of p' at which the cluster results "break" (in this case when cluster 2 no longer appears in the same areas as for $p' = 2000$), the four-digit zip code cluster maps are recreated for p' values ranging from 2000 to 1500 in steps of 100. However, Figure 5.44e shows that for $p' = 1900$, cluster 2 is dispersed across the entire map rather than being concentrated in the southern part of the Netherlands as it is for $p' = 2000$. Therefore, the clustering results have already broken for $p' > 1900$, and thus the plots for $p' < 1900$ are not generated. Instead, maps of the clusters created by *U-SPEC* with $p' = 1950$ and $p' = 1975$ are created and shown in Figures 5.44d and 5.44c.
Figure 5.44 illustrates that the clustering breaks somewhere between $p' = 1975$ (for which cluster 2 is concentrated in the south, similar to the map with $p' = 2000$) and $p' = 1950$ (for which cluster 2 is more spread out over the entire map). Given that the difference between 1975 and 2000 is only 25 data points, it can be concluded that the clustering results are most reliable when $p' = 2000$. Using fewer observations will lead to different clusters.

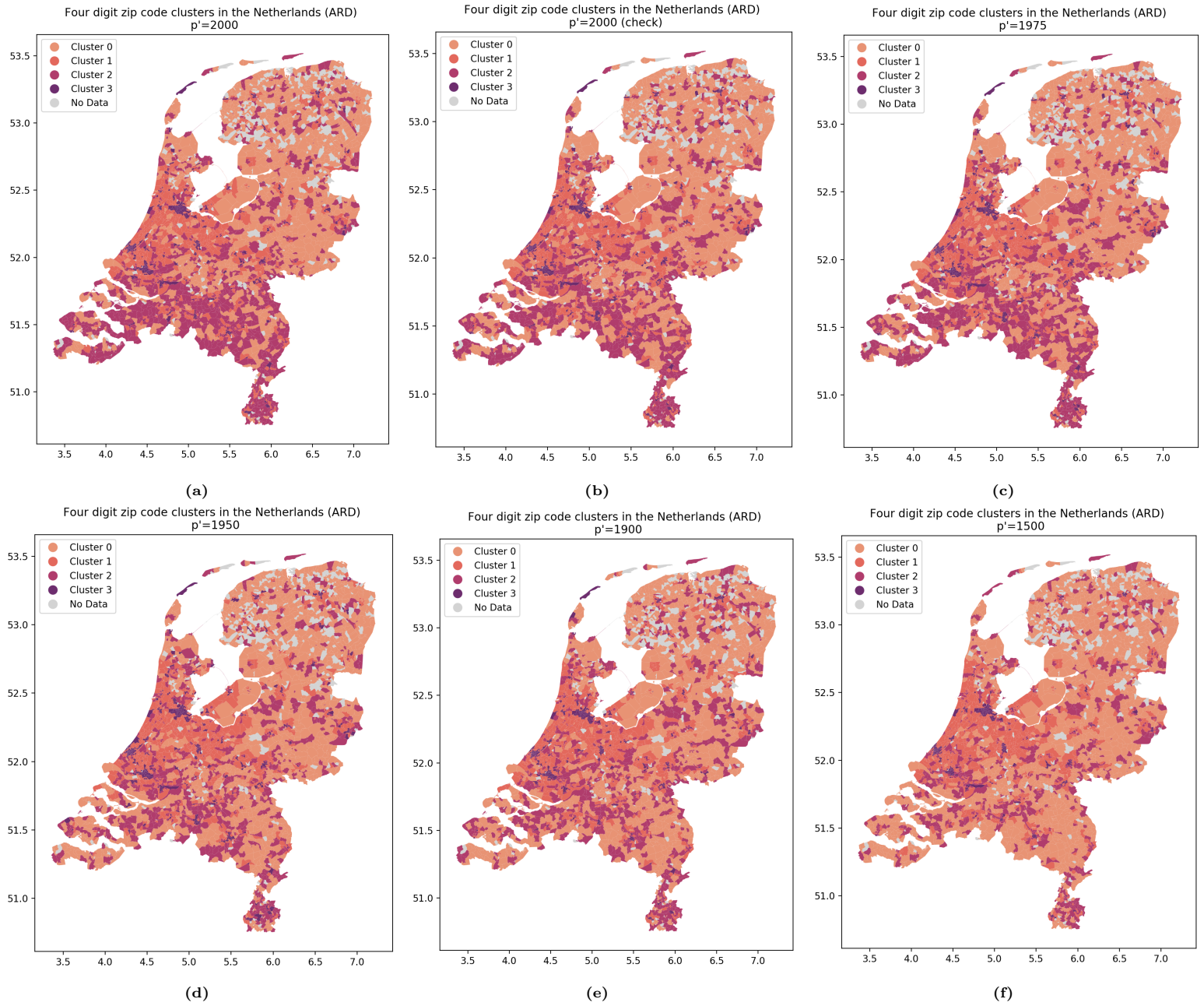


Figure 5.44: This figure shows the four-digit zip code clusters of the Netherlands generated by *U-SPEC* with: **(a)** $p' = 2000$, **(b)** $p' = 2000$ (to check the consistency), **(c)** $p' = 1975$, **(d)** $p' = 1950$, **(e)** $p' = 1900$, **(f)** $p' = 1500$.

WAM dataset

License plate data:
Figure 5.45 shows the Rand indices (relative to the *U-SPEC* clusters with $p' = 2000$) over various values of p' for the WAM license plate dataset. The indices lie between 0.64 and 0.66 for all p' values and remain relatively constant for different values of p'. Additionally, the two distinct clustering results for $p' = 2000$, used as a consistency check, yield an index of 0.644.

Note that the Rand indices of the ARD license plate dataset are higher than those of the WAM license plate dataset. This indicates that the clustering results for the ARD dataset remain more consistent across different values of p'. This could be attributed to the cars in the ARD (claims) dataset being more alike to each other than those in the WAM (claims) dataset. WAM coverage, being mandatory, encompasses a wide variety of cars. On the other hand, ARD coverage is optional, with

claims predominantly made by owners of new or expensive vehicles who may be less concerned about maintaining a claim-free record. As a result, the ARD dataset is likely more homogeneous, leading to similar subsets of cars being selected regardless of the p' value and thus ensuring more stable clustering results.

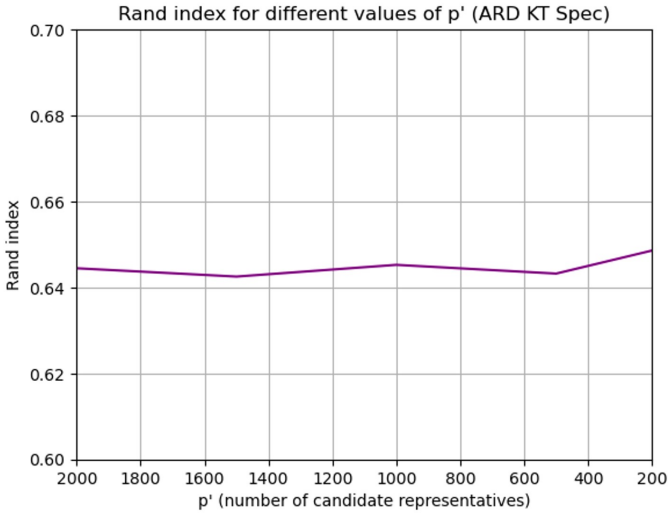


Figure 5.45: This figure shows the Rand indices (relative to the *U-SPEC* clusters with $p' = 2000$) over various values of p' for the WAM license plate dataset.

Figure 5.46 shows the box plots of the claim frequencies for each p' value and across the different license plate clusters in the WAM dataset. Cluster 3 exhibits the most notable differences in its box plots for different p' values. Similar to the ARD dataset, this implies that when there is already significant variation in the box plot for $p' = 2000$, the box plots will differ considerably for other p' values. Therefore, for cluster 3, the *U-SPEC* algorithm is likely to generate different clusters for each p' value.
For the same reason that was previously discussed, the differences in means and variations of the box plots are greater for the WAM dataset compared to the ARD dataset.

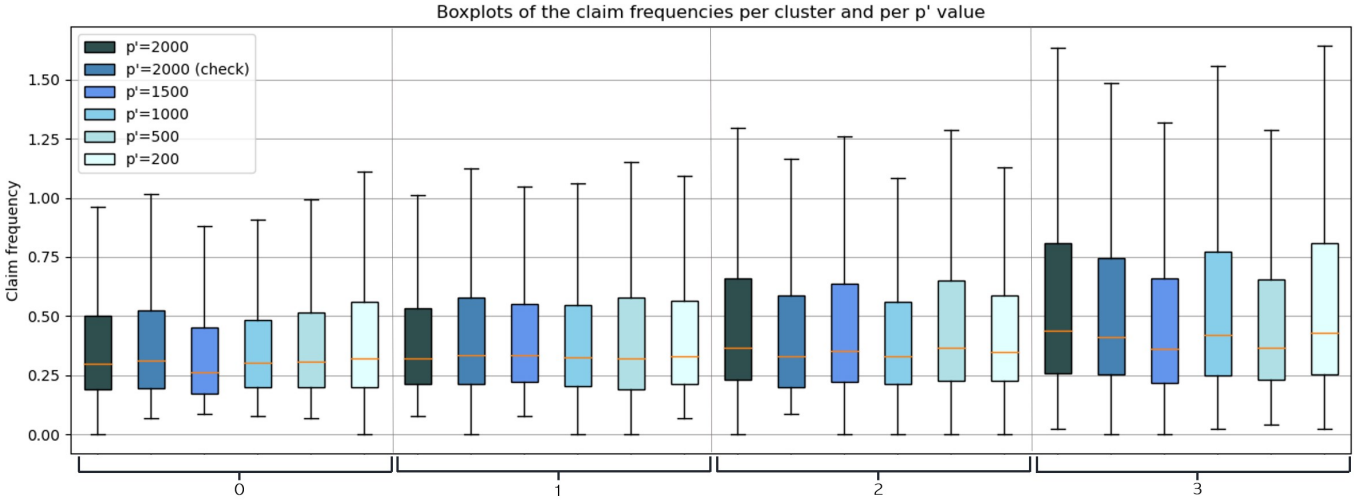


Figure 5.46: This figure shows the box plots of the claim frequencies for each p' value and across the different license plate clusters in the WAM dataset.

Zip code data:
Figure 5.47 shows the Rand indices (relative to the *U-SPEC* clusters with $p' = 2000$) over various values of p' for the WAM zip code dataset. The indices lie between 0.74 and 0.87 for all p' values.

Additionally, the two distinct clustering results for $p' = 2000$, used as a consistency check, yield an index of 0.87. Therefore, the clustering results remain relatively stable across different values of p'. However, for $p' = 200$, there is a significant drop in the Rand index, suggesting that the clusters differ more from those produced with $p' = 2000$ compared to other values of p'.

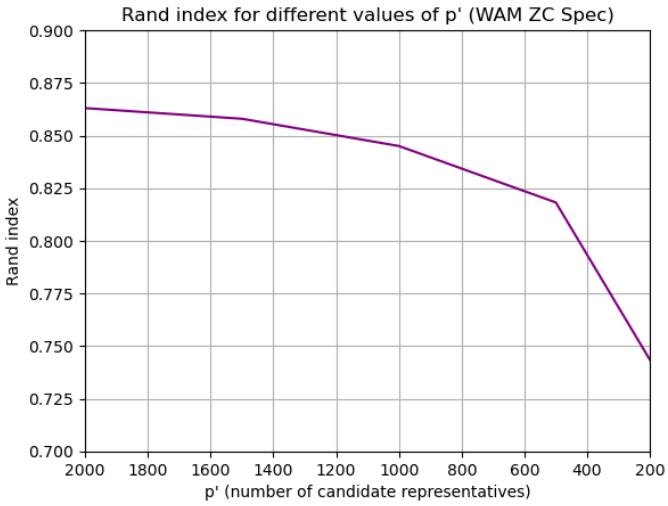


Figure 5.47: This figure shows the Rand indices (relative to the *U-SPEC* clusters with $p' = 2000$) over various values of p' for the WAM zip code dataset.

Figure 5.48 shows the box plots of the claim frequencies for each p' value and across the different zip code clusters in the WAM dataset. For all clusters, the box plots for each p' value have similar means and variations in claim frequency. Nevertheless, clusters 0 and 2 appear to be the least stable with respect to p', indicating that these clusters are most likely to differ when p' is changed.

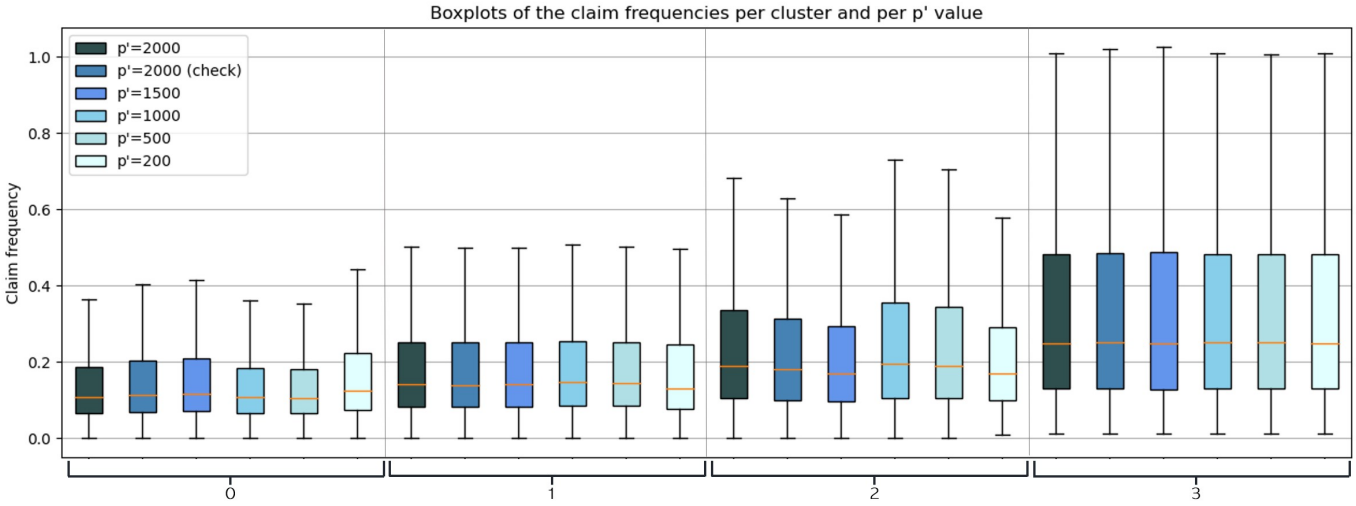


Figure 5.48: This figure shows the box plots of the claim frequencies for each p' value and across the different zip code clusters in the WAM dataset.

Figure 5.49 shows the four-digit zip code clusters of the Netherlands generated by *U-SPEC* with the different values of p'. Clusters 1 and 3 remain stable across varying numbers of observations, while cluster 2 appears in different locations as p' decreases. Note that cluster 2 was anticipated to be the most likely to differ when p' is altered as previously discussed. It can be observed that the cluster results "break" somewhere between $p' = 1000$ (for which all clusters are similar to the map with $p' = 2000$) and $p' = 500$ (for which cluster 2 is concentrated in the south). Using fewer observations will lead to different clusters. Note that for $p' = 200$, cluster 2 becomes prevalent in the "Randstad" region, which

includes the Netherlands' four largest cities (Amsterdam, Rotterdam, The Hague, and Utrecht), their suburbs, and the towns in between.

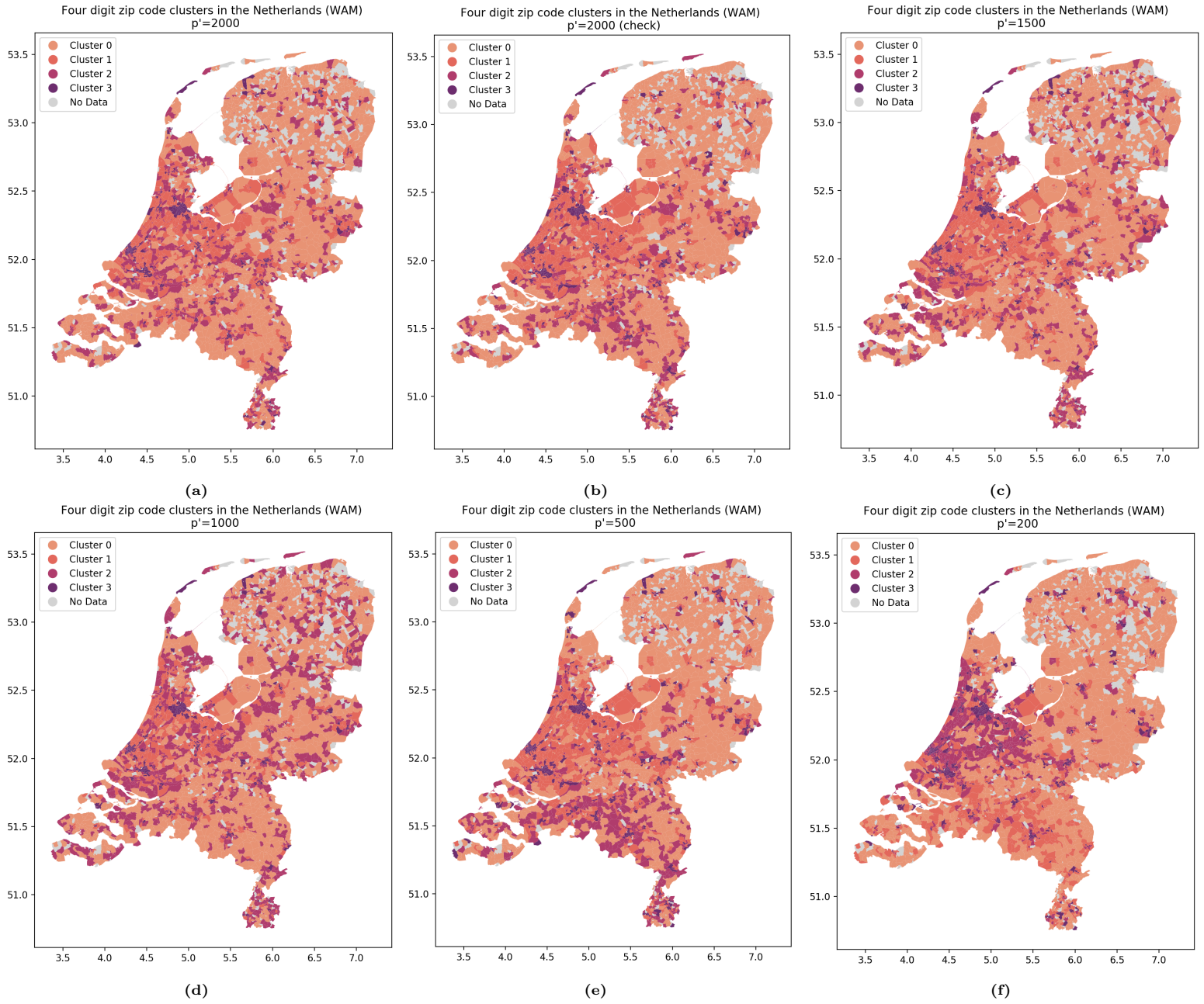


Figure 5.49: This figure shows the four-digit zip code clusters of the Netherlands generated by *U-SPEC* with: **(a)** $p' = 2000$, **(b)** $p' = 2000$ (to check the consistency), **(c)** $p' = 1500$, **(d)** $p' = 1000$, **(e)** $p' = 500$, **(f)** $p' = 200$.

6

Conclusion

The objective of this thesis was described as follows:

Improve the risk classification of the claim frequency models of two coverages, namely "WAM" and "ARD", of a car insurance product (significantly) by clustering zip codes and license plates and using these clusters as risk factors in the models.

Before the clustering techniques were applied, the dataset was pre-processed by;

1. Extracting the relevant data.
2. Deleting the P.O. box zip codes.
3. Deleting redundant variables.
4. Deleting features that only have one unique value.
5. Handling missing values.
6. Checking for (multi)collinearity.
7. Standardizing the data.
8. Determining the significance of the features.

The centroid-based K-prototypes and the connectivity-based (normalized) spectral clustering were employed to cluster the WAM and ARD datasets. K-prototypes was selected because it is among the most commonly utilized techniques and its implementation is required for spectral clustering. Spectral clustering, on the other hand, was selected for its capability to effectively cluster non-linearly separable and relatively large data sets.
Due to the high storage requirements of spectral clustering for large datasets, observation reduction techniques such as random removal and *U-SPEC* can be applied. Initially, these techniques were compared using a sample dataset; despite the random removal technique yielding a higher Adjusted Rand Index in this instance, *U-SPEC* was chosen as the observation reduction method for spectral clustering of the ARD and WAM datasets. This decision was based on *U-SPEC* demonstrating superior performance compared to random removal for large datasets, especially under certain parameter settings and for clusters with less distinct boundaries. Although the sample dataset did not exhibit these characteristics, both the ARD and WAM datasets possessed all of them.
For K-prototypes, the number of clusters was found by analyzing the elbow plot, while for spectral clustering, this was determined by the number of informative eigenvectors corresponding to the isolated eigenvalues of the Laplacian matrix (after observation reduction). Both clustering techniques utilized a distance measure that combined distances between numerical, categorical, and ordinal data points into a weighted sum. Lastly, spectral clustering employed the k-nearest neighbor graph as its similarity graph.

To evaluate the clustering techniques for the ARD and WAM datasets, experts in the actuarial field judged the degree to which the clusters were meaningful. Additionally, the clusters were incorporated into the existing GLM as risk factors, and their impacts on deviance, AICc, and BIC were assessed. Figure 6.1 shows a table that illustrates the sensicality and compares the deviance, AICc, and BIC of

the GLM with clusters as risk factors against the current GLM (for both datasets and both clustering techniques). Based on the results presented in this table, spectral clustering outperforms K-prototypes for both datasets. Furthermore, the spectral clusters of the ARD dataset are the only ones that improve the current risk classification of the claim frequency models, as evidenced by decreased AICc and BIC values. The WAM clusters do not improve the current GLM and thus cannot be used to predict the claim frequencies. Furthermore, they do not provide additional information that would be beneficial for other purposes. Nonetheless, despite the spectral WAM clusters not improving the GLM, the spectral clustering technique shows potential for application to other insurance datasets.

		Logical (yes/no)	Deviance (up/down)	AICc (up/down)	BIC (up/down)
ARD Dataset	K-Prototypes	Yes	Down	Up	Up
	Spectral Clustering	Yes	Down	Down	Down
WAM Dataset	K-Prototypes	Not for zip codes	Down	Up	Up
	Spectral clustering	Yes	Down	Up	Up

Figure 6.1: This figure shows a table that illustrates the sensicality and compares the deviance, AICc, and BIC of the GLM with clusters as risk factors against the current GLM (for both datasets and both clustering techniques).

In the context of this thesis, the spectral clustering technique outperforms the K-prototypes technique for several reasons:

- According to the box plots of the clusters, the spectral clusters exhibit less variation in claim frequency. So, spectral clustering more effectively maximizes the homogeneity among observations within the same cluster.
- Spectral clustering shows superior performance for non-linearly separable datasets, as evidenced by a sample dataset, and the ARD and WAM datasets are likely non-linearly separable.
- Spectral clustering yields more homogeneous clusters, reducing their randomness. This is not only apparent in the zip code clustering maps but also in the case where K-prototypes formed a cluster containing only eighteen license plates.
- Due to observation reduction techniques, fewer data points are required for spectral clustering, making the method applicable to datasets with fewer observations, such as those of other coverages.
- Based on a sample dataset, spectral clustering accurately identifies situations where no meaningful clustering exists, whereas K-prototypes does not.
- In the case of the ARD dataset, spectral clustering improves the current GLM, while for the WAM dataset, the spectral clusters make more sense than the K-prototypes ones.

The group of (significant) spectral zip code clusters of the ARD dataset was the only set of clusters where all were stable with respect to time, allowing them to be directly incorporated into the GLM. All other clusters may also be included in the GLM, provided that the time dependency of the variables used for the clustering is carefully considered, as explained in the next chapter.
When reducing the number of observations, the Rand Index remains relatively high for all spectral clustering results. Furthermore, for the ARD zip code data the spectral clustering "breaks" when the number of observations drops from 1975 to 1950, while for the WAM zip code data, this occurs at 1000 observations.
The ethical implications of this research are discussed in the next chapter.

This thesis pioneered the clustering of license plates to enhance the risk classification in car insurance and it marked the first application of spectral clustering to both zip codes and license plates. To achieve this, modifications were made to K-means and spectral clustering methods to effectively handle mixed data, including the introduction of a unique distance measure. This measure is the weighted sum of the Euclidean distance for numerical variables, Hamming distance for categorical variables, and Gower's distance for ordinal variables.
Moreover, this thesis explored observation reduction techniques and their implications concerning high dimensional clustering, topics that haven't been studied in the context of license plate and zip code

clustering before. The number of clusters after the application of observation reduction techniques, was determined by the number of informative eigenvectors corresponding to the isolated eigenvalues of the Laplacian matrix. This approach had not been used in license plate and zip code clustering contexts and, prior to this thesis, informative eigenvectors were only used to establish an upper bound for the number of clusters.

All of the sub-questions have been answered and the objective of this thesis was partially achieved; the risk classification of the claim frequency model of the ARD coverage of a car insurance product was (significantly) improved by clustering zip codes and license plates and using these clusters as risk factors in the model. This was accomplished through spectral clustering. For the WAM coverage, the risk classification did not improve with the clusters of this thesis. Notions for further research aimed at improving the risk classification of the WAM coverage, can be found in the next chapter.

7

Discussion

This chapter discusses the limitations of the findings and proposes notions for further research in Section 7.1. An ethical framework is provided in Section 7.2.

7.1. Limitations of the findings & notions for further research

In this section, the limitations of the findings of this research are discussed and notions for further research are proposed.

Discussion of the observation reduction technique

First, certain elements of the observation reduction technique (i.e. *U-SPEC*) can be modified for future research. In this thesis, $p' = 2000$ (and thus $p = \frac{1}{10} \cdot 2000 = 200$) was selected to ensure a sufficient number of rep-clusters z_1. Increasing p' would result in longer run times; currently, the *U-SPEC* algorithm takes three hours to run for the license plate ARD data and 30 hours for the zip code ARD data (due to the higher number of ordinal features requiring more time to compute Gower's distance). Similar run times are observed for the WAM dataset. Therefore, while a larger p' could enhance stability in terms of the number of observations (see Figure 5.45, where increasing p' could yield a Rand index close to 1 for the p' value used to check consistency), it also increases the runtime. Thus, for future research, p' could be increased, but optimizing the code for efficiency is crucial. If optimization is not feasible, using a more powerful computer or implementing parallel computing across multiple cores/processors is necessary. Note that as p' increases, p will also increase as $p = \frac{p'}{10}$.
In addition to random removal and *U-SPEC*, future research could explore other observation reduction techniques. For example, similar to the Cao method used to initialize centroids for K-prototypes by ensuring they are well separated (see Subsection 4.2.1), a similar approach could be employed to select a subset of observations, thereby reducing the overall number of observations.
Lastly, in this thesis, the number of informative eigenvectors (corresponding to the isolated eigenvalues of the Laplacian matrix) was used to determine the number of clusters, rather than serving as the initially intended upper bound. Future research could investigate and theoretically prove the consistency of the number of informative eigenvectors being equal to the number of clusters. Additionally, in this thesis, the number of informative eigenvectors was determined by visual inspection of plots. Future research could develop a more objective method, such as assessing the white noise and stationarity of the eigenvectors.

Potential modifications for further research

For further research, alternative distance measures or different weights (i.e. α and γ) could be explored. For instance, Chebyshev's distance, which computes the maximum difference between two vectors, could be applied to numerical features. For categorical features, Jaccard's distance, which measures the similarity between two sets by comparing their unions and intersections, could be used. [22]
In this thesis, the k-nearest neighbors graph was chosen as the similarity graph for spectral clustering

because of its ability to connect points across different scales, its ease of implementation, its tendency to produce a sparse adjacency matrix W, and its resilience to unsuitable parameter choices compared to other types of similarity graphs. For future research, other similarity graphs, such as the fully connected graph discussed in Chapter 2, could be investigated.
Lastly, beyond K-prototypes and spectral clustering, other clustering methods could be explored. For example, hierarchical clustering methods (e.g. "AGNES" and "DIANA") which are also connectivity-based techniques like spectral clustering, could be investigated. [6] Furthermore, this thesis focused on centroid-based and connectivity-based methods, but future research could look into other techniques, such as density-based methods. An example of such a method is "DBSCAN", which identifies core samples in high-density regions and expands clusters from these points. This method can be applied in this case since the data has irregular shapes and there is no prior knowledge about the number of clusters. [27]

Possible extensions of this research

Possible extensions of this research could include improving the WAM clustering. As mentioned in Subsection 5.2.5, for the WAM coverage, claim frequencies are more closely tied to driver characteristics (e.g. age and gender) than to license plates and zip codes (which are more relevant to ARD coverage). Therefore, incorporating license plate and zip code clusters into the GLM for the WAM dataset worsens its performance. Instead, in the future, clusters can be created for the WAM dataset based on driver characteristics. However, since the dataset comprises company cars where multiple drivers may be associated with a single vehicle, this specific approach may not be suitable for the WAM data of this thesis. Nevertheless, for future research, this clustering method can be explored by focusing on private (i.e. non-company) car WAM coverage datasets.
The spectral license plate and zip code clusters of the ARD dataset improved the current GLM. However, as explained in Subsection 5.3.1, these spectral license plate clusters are not stable with respect to time. Therefore, moving forward, the time dependency of the variables used in the license plate clustering should be carefully considered. Variables showing excessive time dependence, such as the "APK" date (periodic vehicle inspection date) variable, should be transformed into time-independent forms (for instance, by calculating the number of days between the contract's start date and the APK date).
Moreover, in the future, it can be investigated whether all variables positively contribute to the clustering or if certain features could be eliminated through dimensionality reduction. This approach could also lead to faster run times.
Lastly, for this thesis, the ARD and WAM coverage datasets of company cars were clustered. Future research can explore the extent to which other types of coverages, insurances (e.g. fire or storm insurance), and private (non-company) vehicle datasets can be clustered. Moreover, the clustering techniques combined with the distance measure described in this thesis can be applied to any mixed dataset requiring grouping. For instance, this approach could be used to create groups based on the risk of defaulting on loans (i.e. credit risk) or for customer segmentation in supermarkets to enhance targeted marketing strategies and personalized recommendations. Lastly, since K-prototypes identified a distinct group comprising eighteen ARD license plates linked to vehicles with two or three wheels, this clustering technique holds potential for detecting anomalous behavior, such as identifying suspicious transactions related to fraud or money laundering.

7.2. Ethical framework

In this section, the ethical implications of this research are discussed.

Transparency and explainability

In the insurance industry, it is crucial to ensure that the methods and data used for clustering are transparent and understandable to stakeholders, such as policyholders, to maintain trust and accountability. This transparency is also essential for complying with regulatory standards. In other words, the premium pricing method and the associated risk classification technique should be transparent and explainable, clearly describing how the clustering techniques work and which variables are considered. For this thesis, the clustering techniques were thoroughly explained, the cluster descriptions were pro-

vided, and the variables considered in the analysis were outlined.
Moving forward, it would be beneficial to provide non-technical explanations of the clustering algorithms. This approach would enable stakeholders and regulators without a mathematical background to understand how these methods work. Analogies and visualizations could explain how data is grouped and show the significance of the variables used (e.g. with SHAP values). [5] Through feedback, these explanations can be improved to ensure clarity and comprehension.

Discrimination and bias

There is a risk that certain groups of people are unfairly targeted or disadvantaged based on patterns found in the license plate and zip code data. For example, in this thesis, variables such as "non-western immigrants" are used to obtain the clustering outcomes and the clusters with higher claim frequencies exhibit higher concentrations of non-western immigrants, resulting in potentially higher premiums for these groups. To ensure less discrimination and bias in the clustering algorithms, these types of variables should not be taken into account.
It is worth noting that removing these variables does not eliminate bias entirely. For example, certain car brands may be more frequently driven by women than men, which could lead to partiality in the models. However, it is impossible to remove all variables that might introduce even a small amount of bias. Nevertheless, being aware of each variable's impact concerning bias is essential.

Bibliography

[1] Ahmed, M. (2021). *Linear Support Vector Machines Explained.* Medium. Retrieved February 29, 2024, from https://linguisticmaz.medium.com/support-vector-machines-explained-8804cac06883

[2] Alma Better. (n.d.). *Confidence Intervals and Margin of Error.* Alma Better. Retrieved June 26, 2024, from https://www.almabetter.com/bytes/tutorials/applied-statistics/confidence-intervals-and-margin-of-error

[3] AnalytixLabs. (2022). *What is Clustering in Machine Learning: Types and Methods.* AnalytixLabs. Retrieved March 26, 2024, from https://www.analytixlabs.co.in/blog/types-of-clustering-algorithms/

[4] ANWB. (n.d.). *Autoverzekering kiezen.* ANWB. Retrieved February 26, 2024, from https://www.anwb.nl/verzekeringen/autoverzekering/autoverzekering-kiezen

[5] Awan, A.A. (2023). *An Introduction to SHAP Values and Machine Learning Interpretability.* Datacamp. Retrieved July 4, 2024, from https://www.datacamp.com/tutorial/introduction-to-shap-values-machine-learning-interpretability

[6] Bindal, S. (2021). *Clustering in ML – Part 4: Connectivity Based Clustering.* Applied Singularity. Retrieved June 28, 2024, from https://appliedsingularity.com/2021/07/27/clustering-in-ml-part-4-connectivity-based-clustering/

[7] Bishnoi, A. (2022). *Graph Theory* [Lecture Notes].

[8] Bobbitt, Z. (2021). *What is High Dimensional Data? (Definition & Examples).* Statology. Retrieved April 29, 2024, from https://www.statology.org/high-dimensional-data/

[9] De Bont, D. (2022). Geographical risk in the Dutch car insurance: A data-driven approach to measure regional effects on the claim frequency. [Unpublished manuscript]. Faculty of Behavioural, Management and Social sciences, UTwente.

[10] Brownlee, J. (2021). *How to Develop LASSO Regression Models in Python.* Machine Learning Mastery. Retrieved March 25, 2024, from https://machinelearningmastery.com/lasso-regression-with-python/

[11] Brownlee, J. (2020). *How to Develop Ridge Regression Models in Python.* Machine Learning Mastery. Retrieved March 25, 2024, from https://machinelearningmastery.com/ridge-regression-with-python/

[12] Brownlee, J. (2020). *How to Develop Elastic Net Regression Models in Python.* Machine Learning Mastery. Retrieved March 25, 2024, from https://machinelearningmastery.com/elastic-net-regression-in-python/

[13] Brownlee, J. (2020). *Probabilistic Model Selection with AIC, BIC, and MDL.* Machine Learning Mastery. Retrieved June 10, 2024, from https://machinelearningmastery.com/probabilistic-model-selection-measures/.

[14] Calkins, K.G. (2005). *Correlation Coefficients.* Andrews. Retrieved March 21, 2024, from https://www.andrews.edu/~calkins/math/edrm611/edrm05.htm#:~:text=Correlation%20coefficients%20whose%20magnitude%20are,can%20be%20considered%20highly%20correlated.

[15] Couillet, R., & Liao, Z. (2023). *Random Matrix Methods for Machine Learning.* [Unpublished manuscript]. Cambridge University.

[16] DataRundown. (n.d.). *K-Means Clustering: 7 Pros and Cons Uncovered*. DataRundown. Retrieved March 28, 2024, from https://datarundown.com/k-means-clustering-pros-cons/

[17] Delua, J. (2021). *Supervised vs. Unsupervised Learning: What's the Difference?*. IBM. Retrieved March 1, 2024, from https://www.ibm.com/blog/supervised-vs-unsupervised-learning/

[18] Denuit, M., Maréchal, X., Pitrebois, S., & Walhin, J.(2007). *Actuarial modelling of claim counts: Risk classification, credibility and bonus-malus systems*. John Wiley & Sons.

[19] Desgraupes, B. (2017). *Clustering Indices*. Cran R-Project. Retrieved April 2, 2024, from https://cran.r-project.org/web/packages/clusterCrit/vignettes/clusterCrit.pdf

[20] Doshi, N. (2019). *Spectral clustering*. Towards Data Science. Retrieved February 29, 2024, from https://towardsdatascience.com/spectral-clustering-82d3cff3d3b7

[21] Ellis, C. (2022). *When to use spectral clustering*. Crunching the Data. Retrieved April 16, 2024, from https://crunchingthedata.com/when-to-use-spectral-clustering/

[22] Eskandar, S. (2023). *Exploring Common Distance Measures for Machine Learning and Data Science: A Comparative Analysis*. Medium. Retrieved June 28, 2024, from https://medium.com/@eskandar.sahel/exploring-common-distance-measures-for-machine-learning-and-data-science-a-comparative-analysis-ea0216c93ba3

[23] Faudree, R. (2003). *Encyclopedia of Physical Science and Technology* (3rd ed., pp 15-31). Academic Press.

[24] Finley, T., & Joachims, T. (2005). Supervised clustering with support vector machines. *Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005)*. http://dx.doi.org/10.1145/1102351.1102379

[25] Geeks for Geeks. (2023). *Elbow Method for optimal value of k in KMeans*. Geeks for Geeks. Retrieved March 28, 2024, from https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/

[26] Geeks for Geeks. (2023). *What is Undirected Graph? | Undirected Graph meaning*. Geeks for Geeks. Retrieved March 29, 2024, from https://www.geeksforgeeks.org/what-is-unidrected-graph-undirected-graph-meaning/

[27] Geeks for Geeks. (2023). *Different Types of Clustering Algorithm*. Geeks for Geeks. Retrieved June 28, 2024, from https://www.geeksforgeeks.org/different-types-clustering-algorithm/

[28] Google Developers. (2022). *Clustering Algorithms*. Google Developers. Retrieved March 26, 2024, from https://developers.google.com/machine-learning/clustering/clustering-algorithms

[29] Gorthy, S. (2021). *Euclidean Distance*. Medium. Retrieved March 11, 2024, from https://srikgorthy.medium.com/euclidean-distance-8fae145ef5f3

[30] Hayes, A. (2023). *Adverse Selection: Definition, How It Works, and The Lemons Problem*. Investopedia. Retrieved February 26, 2024, from https://www.investopedia.com/terms/a/adverseselection.asp

[31] Hayes, A. (2024). *Mode: What It Is in Statistics and How to Calculate It*. Investopedia. Retrieved March 1, 2024, from https://www.investopedia.com/terms/m/mode.asp

[32] He, Z., Xu, X., & Deng, S. (2005). Clustering Mixed Numeric and Categorical Data: A Cluster Ensemble Approach. https://doi.org/10.48550/arXiv.cs/0509011

[33] Howell, E. (2021). *Saturated Models, Deviance and the Derivation of Sum of Squares*. Medium. Retrieved June 4, 2024, from https://towardsdatascience.com/saturated-models-deviance-and-the-derivation-of-sum-of-squares-ee6fa040f52

[34] Huang, Z. (1998). Extensions to the K-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery, 2*(2), 283–304.

[35] Huang, D., Wang, C., Lai, J., Wu, J., & Kwoh, C. (2020). Ultra-Scalable Spectral Clustering and Ensemble Clusterings. *IEEE Transactions On Knowledge and Data Engineering*, 32(6).

[36] Interpolis. (n.d.). *Negatieve schadevrije jaren*. Retrieved February 29, 2024, from https://www.interpolis.nl/verzekeren/autoverzekering/schadevrije-jaren/negatieve-schadevrije-jaren#:~:text=Heb%20je%20meer%20schadevrije%20jaren,beginnend%20bestuurder%20schade%20hebt%20geclaimd.

[37] Investopedia Team. (2024). *Variance Inflation Factor (VIF)*. Investopedia. Retrieved July 11, 2024, from https://www.investopedia.com/terms/v/variance-inflation-factor.asp

[38] Jadeja, M. (2022). *Jaccard Similarity Made Simple: A Beginner's Guide to Data Comparison*. Medium. Retrieved February 29, 2024, from https://medium.com/@mayurdhvajsinhjadeja/jaccard-similarity-34e2c15fb524

[39] Jahangiry, P. (2024). *Module 12- Clustering* [Power Point slides]. Github. https://github.com/PJalgotrader/Machine_Learning-USU/blob/main/Lectures%20and%20codes/Module%2012-%20Clustering/Module%2012-%20Clustering.pdf

[40] Jain, R. (2020). *Correlation between Categorical Variables*. Medium. Retrieved March 22, 2024, from https://medium.com/@ritesh.110587/correlation-between-categorical-variables-63f6bd9bf2f7

[41] Jamotton, C., Hainaut, D., & Hames, T. (2023). *Insurance Analytics with Clustering Techniques* (LIDAM Discussion Paper ISBA Issue 2023-02). UC Louvain, ISBA, LIDAM.

[42] Kagan, J. (2023). *Insurance Risk Class: Definition and Associated Premium Costs*. Investopedia. Retrieved January 22, 2024, from https://www.investopedia.com/terms/i/insurance-risk-class.asp

[43] Kagan, J. (2023). *Frequency-Severity Method: Definition and How Insurers Use It*. Investopedia. Retrieved January 22, 2024, from https://www.investopedia.com/terms/f/frequencyseverity-method.asp

[44] Kagan, J. (2023). *Insurance: Definition, How It Works, and Main Types of Policies*. Investopedia. Retrieved February 26, 2024, from https://www.investopedia.com/terms/i/insurance.asp

[45] Kagan, J. (2023). *Auto Insurance: Definition, How It Works, Coverage Types & Costs*. Investopedia. Retrieved February 26, 2024, from https://www.investopedia.com/terms/a/auto-insurance.asp

[46] Keany, E. (2021). *The Ultimate Guide for Clustering Mixed Data*. Medium. Retrieved February 29, 2024, from https://medium.com/analytics-vidhya/the-ultimate-guide-for-clustering-mixed-data-1eefa0b4743b

[47] Kenton, W. (2024). *What Is Analysis of Variance (ANOVA)?*. Investopedia. Retrieved June 4, 2024, from https://www.investopedia.com/terms/a/anova.asp#:~:text=Key%20Takeaways-,Analysis%20of%20variance%20(ANOVA)%20is%20a%20statistical%20test%20used%20to,should%20equal%20close%20to%201..

[48] Koehrsen, W. (2018). *Statistical Significance Explained*. Medium. Retrieved June 20, 2024, from https://towardsdatascience.com/statistical-significance-hypothesis-testing-the-normal-curve-and-p-values-93274fa32687

[49] Kooijman, I.J. (2021). Signature-based model recognition in financial time series. [Unpublished manuscript]. Faculty EEMCS, Delft University of Technology.

[50] Kumar, A. (2020). *KNN Algorithm: When? Why? How?*. Medium. Retrieved May 2, 2024, from https://towardsdatascience.com/knn-algorithm-what-when-why-how-41405c16c36f

[51] Von Luxburg, U. (2007). A Tutorial on Spectral Clustering. *Statistics and Computing*, 17(14).

[52] Mayo, M. (2022). *Centroid Initialization Methods for k-means Clustering*. KD-Nuggets. Retrieved March 28, 2024, from https://www.kdnuggets.com/2020/06/centroid-initialization-k-means-clustering.html

[53] Mbuga, F., & Tortora, C. (2021). Spectral Clustering of Mixed-Type Data. *Stats 2022, 5*(1), 1-11. https://doi.org/10.3390/stats5010001

[54] Meng, L. (2017). *STAT 410 - Linear Regression (Lecture 14)*. Rice. Retrieved June 10, 2024, from https://bpb-us-e1.wpmucdn.com/blogs.rice.edu/dist/e/8375/files/2017/08/Lecture14-rixnqf.pdf.

[55] Netherlands Enterprise Agency, RVO. (n.d.). *Third-party liability insurance for motor vehicles*. Retrieved February 26, 2024, from https://business.gov.nl/regulation/vehicle-insurance/

[56] ODSC. (2018). *Unsupervised Learning: Evaluating Clusters*. Medium. Retrieved April 2, 2024, from https://odsc.medium.com/unsupervised-learning-evaluating-clusters-bd47eed175ce

[57] OECD AI. (n.d.). *Adjusted Rand Index (ARI)*. OECD AI. Retrieved April 2, 2024, from https://oecd.ai/en/catalogue/metrics/adjusted-rand-index-%28ari%29

[58] Ohlsson, E., & Johansson, B.(2010). *Non-life insurance pricing with generalized linear models*. Springer.

[59] Podani, J. (1999). Extending Gower's General Coefficient of Similarity to Ordinal Characters. *Taxon*, 48(2), 331-340. http://dx.doi.org/10.2307/1224438

[60] Ramakrishnan, R. (2021). *An Alternative to the Correlation Coefficient That Works For Numeric and Categorical Variables*. Medium. Retrieved March 22, 2024, from https://rviews.rstudio.com/2021/04/15/an-alternative-to-the-correlation-coefficient-that-works-for-numeric-and-categorical-variables/#:~:text=Further%2C%20if%20either%20variable%20of,the%20point%2Dbiserial%20correlation%20coefficient.

[61] Reddy, Y. (2023). *K-means, kmodes, and k-prototype*. Medium. Retrieved February 29, 2024, from https://medium.com/@reddyyashu20/k-means-kmodes-and-k-prototype-76537d84a669

[62] Saraswat, M. (n.d.). *Practical Guide to Clustering Algorithms & Evaluation in R*. Hacker Earth. Retrieved April 2, 2024, from https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/clustering-algorithms-evaluation-r/tutorial/

[63] Sarvandani, M. (2023). *Top 17 applications of clustering in machine learning*. Medium. Retrieved March 26, 2024, from https://medium.com/@mohamadhasan.sarvandani/top-applications-of-clustering-in-machine-learning-d202f73d6dce

[64] Scikit-learn. (n.d.). *sklearn.impute.IterativeImputer*. Scikit-learn. Retrieved March 22, 2024, from https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html

[65] Somashekara, M.T., & Manjunatha, D. (2014). Performance Evaluation of Spectral Clustering Algorithm using Various Clustering Validity Indices. *International Journal of Electronics Communication and Computer Engineering*, 5(6).

[66] Tabak, J. (2008). *Geometry: The Language of Space and Form* (1st ed.). Facts On File, Incorporated.

[67] Team Acko. (2023). *Non-Life Insurance Policy: Types, Features, Benefits & Importance*. Retrieved February 26, 2024, from https://www.acko.com/general-info/non-life-insurance/

[68] The University of Memphis. (n.d.). *Correlation Between Continuous & Categorical Variables*. Medium. Retrieved March 22, 2024, from http://www.ce.memphis.edu/7012/L17_CategoricalVariableAssociation.pdf

[69] Tokareva, N.(2015). *Bent Functions: Results and Applications to Cryptography*. Academic Press.

[70] Tufvessona, O., Lindströmband, J., & Lindström, E. (2019). Spatial statistical modelling of insurance risk: a spatial epidemiological approach to car insurance. *Scandinavian Actuarial Journal, 2019-6*, 508-522. https://doi.org/10.1080/03461238.2019.1576146

[71] Ucar, K.T. (2023). *How to Calculate the Correlation Between Categorical and Continuous Values*. Medium. Retrieved March 22, 2024, from https://medium.com/@ktoprakucar/how-to-calculate-the-correlation-between-categorical-and-continuous-values-dcb7abf79406

[72] University of Adelaide. (n.d.). *Types of Data in Statistics: Numerical vs Categorical Data*. Retrieved February 27, 2024, from https://online.adelaide.edu.au/blog/types-of-data

[73] Unsupervised Learning. (n.d.). *Clustering Algorithms: K-Means*. Unsupervised Learning. Retrieved March 28, 2024, from https://aiplanet.com/learn/unsupervised-learning/clustering-analysis-and-techniques/927/clustering-algorithms-k-means

[74] Vishakha, R. (2023). *Life Insurance Vs. General Insurance*. Forbes. Retrieved February 26, 2024, from https://www.forbes.com/advisor/in/life-insurance/life-insurance-vs-general-insurance/

[75] Watts, V. (2022). *Introduction to Statistics* (1st ed.). Fanshawe College Pressbooks.

[76] Whitfield, B. (2023). *When and Why to Standardize Your Data*. Built In. Retrieved March 21, 2024, from https://builtin.com/data-science/when-and-why-standardize-your-data

[77] Wikipedia. (2024). *K-Means Clustering*. Wikipedia. Retrieved March 28, 2024, from https://en.wikipedia.org/wiki/K-means_clustering

[78] Williams, G. & Huang, Z. (1997). Mining the knowledge mine: The hot spots methodology for mining large real world databases. *Australian Joint Conference on Artificial Intelligence, 10*, 340-348. http://doi.org/10.1007/3-540-63797-4_87

[79] Xie, S.,& Esposito, E. X. (2019). Defining geographical rating territories in auto insurance regulation by spatially constrained clustering. *Risks, ISSN 2227-9091, MDPI, Basel, Vol. 7, Iss. 2*, 1-20. https://doi.org/10.3390/risks7020042

[80] Yang, Y., Qian, W., & Zou, H. (2015). Insurance Premium Prediction via Gradient Tree-Boosted Tweedie Compound Poisson Models. https://doi.org/10.48550/arXiv.1508.06378

[81] Yehoshua, R. (2023). *Spectral Clustering*. Medium. Retrieved February 29, 2024, from https://medium.com/@roiyeho/spectral-clustering-50aee862d300

[82] Zajic, A. (2022). *What Is Akaike Information Criterion (AIC)?*. BuiltIn. Retrieved June 4, 2024, from https://builtin.com/data-science/what-is-aic

A

Clustering results of the WAM dataset

A.1. License plate clustering of the WAM dataset

A.1.1. K-prototypes method

For the WAM license plate dataset, the K-prototypes algorithm was run with values of K ranging from 2 to 20 to create the elbow plot shown in Figure A.1. The elbow shape can be observed around $K = 4$, indicating that the optimal number of clusters for the license plates ("KT") in the WAM dataset is four.

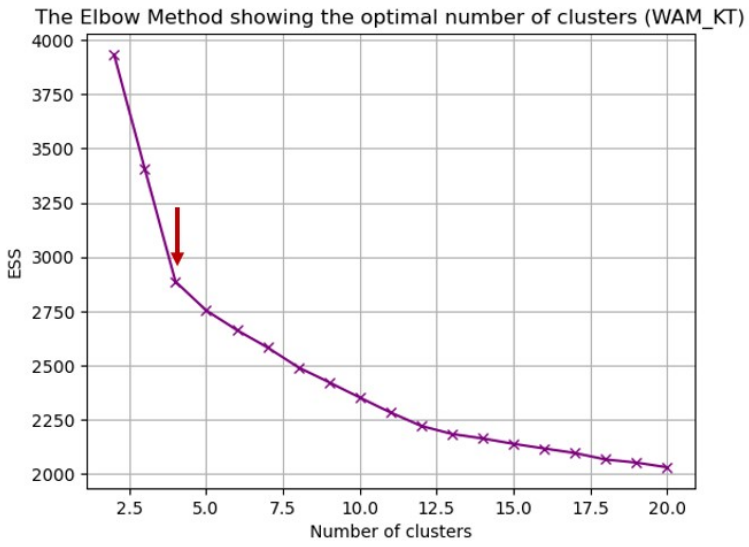


Figure A.1: This figure shows the elbow plot for the WAM license plate dataset. The elbow shape can be observed around $K = 4$, indicating that the optimal number of clusters for the license plates ("KT") in the WAM dataset is four.

Figure A.2 shows the box plots of each cluster regarding the claim frequency. The distinct averages and variations observed in these box plots indicate that the clusters are significant in terms of claim frequency, making them suitable for inclusion as risk factors in the GLM.

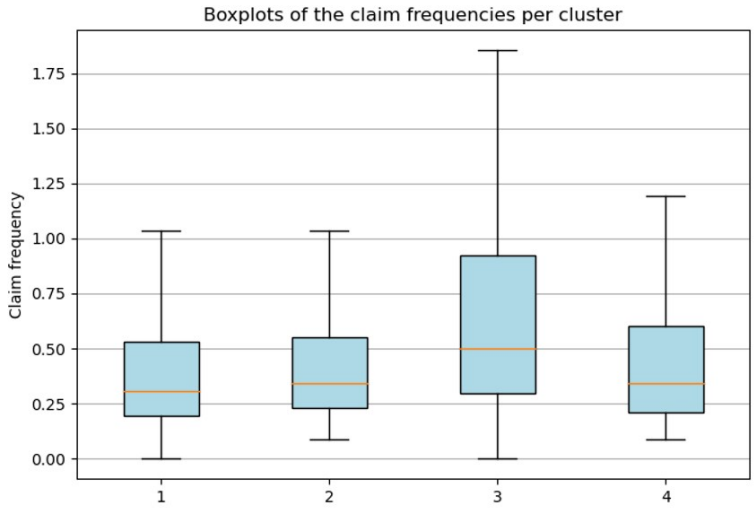


Figure A.2: This figure shows the box plots of each cluster regarding the claim frequency. It can be observed that every cluster has a distinct average and variation in its box plot.

Figure A.3 presents a table of the descriptions of all clusters. For each cluster, the number of license plates, average claim frequency, and the class of average claim frequency are also provided. The descriptions of the clusters can be summarized as follows:

- **Cluster 0**: Light, affordable, low-power, and eco-friendly cars.
- **Cluster 1**: New, expensive, high-power, and high-speed cars.
- **Cluster 2**: Cars with a lot of unknown data (including small vehicles with two or three wheels and vans).
- **Cluster 3**: Eco-friendly "middle of the road" cars.

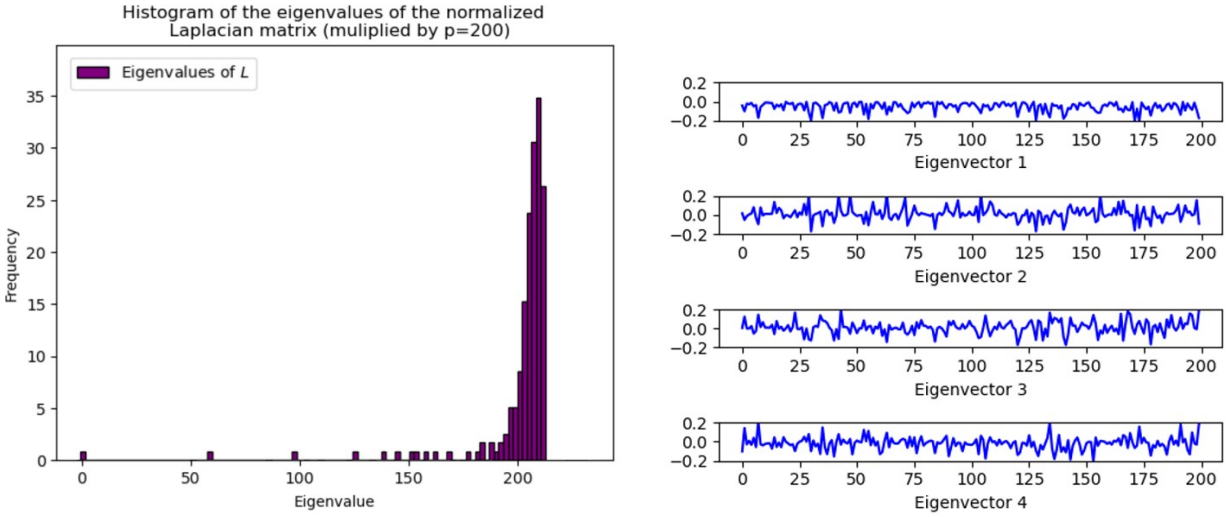
Cluster	Description	#License Plates	Claim frequency	
			Average	Class
0	Contains the most license plates. Lowest weight, least number of seats, least amount of power, and cheapest cars. Only cluster where most cars don't have turbo and has a low class of top speed. Eco-friendly cars, hatchbacks, and a lot of Polo's. Most common brand is Volkswagen.	[redacted]	0.436	***Lowest***
1	New cars with a lot of power. Most common brand is BMW (also contains a lot of Mercedes-Benz). High class top speed, and high lower bound of number of gears. Most expensive cars and the only cluster where most cars have automatic transmission.	[redacted]	0.449	Low
2	Contains the least amount of license plates. Old cars and a lot of unknown data (e.g. usage and eco-friendliness). Only cluster that contains European vehicle category L (small vehicles with two or three wheels) and N (vans). The most common brand is Mercedes-Benz as a result of the many vans (the most common commercial name is Merc. Sprinter). The fuel that is used the most is Diesel.	[redacted]	0.726	***Highest***
3	"Middle of the road" cluster (e.g. in terms of weight, power, and price). Combi cars that are eco-friendly. Most common brand is Volvo and most common fuel Diesel.	[redacted]	0.495	Intermediate

Figure A.3: This figure shows a table of the descriptions for all clusters. The number of license plates, average claim frequency, and class of average claim frequency are also provided for every cluster.

Evaluation: According to experts in the actuarial field, the clusters make sense. For example, vehicles with two or three wheels, vans, and vehicles with a lot of unknown data (i.e. vehicles in cluster 2) tend to have a higher average claim frequency. On the other hand, lighter cars with less power (such as the cars in cluster 0) are considered safer, resulting in a lower claim frequency.

A.1.2. Modified spectral clustering method

The WAM license plate dataset consists of $n = 39,311$ data points. Therefore, the *U-SPEC* method was applied with $n \gg p' = 2000$ to ensure a sufficient number of rep-clusters z_1. Moreover, p, z_1, k, and k' are as defined for the ARD license plate dataset. Note that these parameters yield a high-dimensional dataset as the ratio of dimensions to observations is equal to $81/p = 81/200 = 0.405$. Therefore, the number of clusters is determined with the method outlined in Subsection 4.3.3.
Figure A.4a shows a histogram of the eigenvalues of the normalized Laplacian, multiplied by $p = 200$, with *U-SPEC*. In Figure A.4b, the eigenvectors corresponding to the four isolated eigenvalues of this histogram are depicted. Since all eigenvectors are informative, the optimal number of clusters is *at most* four.



(a) **(b)**

Figure A.4: This figure shows, for the WAM license plate dataset: **(a)** a histogram of the eigenvalues of the normalized Laplacian (multiplied by $p = 200$), **(b)** the eigenvectors corresponding to the four isolated eigenvalues of the histogram.

The *U-SPEC* algorithm is completed with four clusters and Figure A.5 shows the box plots of each of these clusters regarding the claim frequency. The distinct averages and variations observed in these box plots indicate that the clusters are significant in terms of claim frequency, making them suitable for inclusion as risk factors in the GLM.

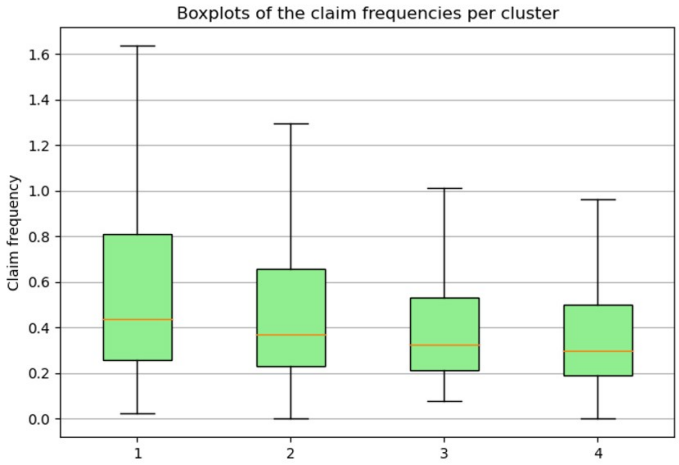


Figure A.5: This figure shows the box plots of each cluster regarding the claim frequency. It can be observed that every cluster has a distinct average and variation in its box plot.

Figure A.6 displays the *ordered* box plots of each cluster's claim frequency for both the K-prototypes (in blue) and spectral clustering (in green) methods. For cluster 3, spectral clustering shows greater variation in claim frequency. However, for all other clusters, the variation is greater with K-prototypes. Therefore, it can be concluded that the K-prototypes clusters generally exhibit greater variation in claim frequency, indicating that spectral clustering more effectively maximizes the homogeneity among observations within the same cluster.

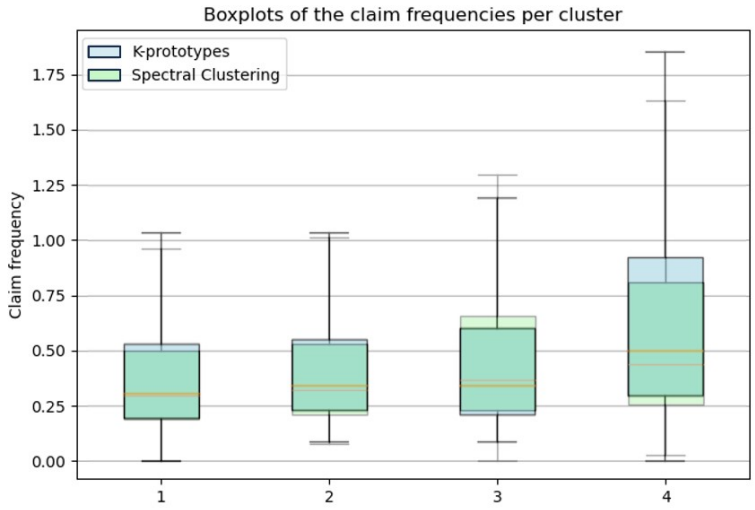


Figure A.6: This figure shows the *ordered* box plots of each cluster's claim frequency for both the K-prototypes (in blue) and spectral clustering (in green) methods.

Figure A.7 presents a table of the descriptions of all clusters. For each cluster, the number of license plates, average claim frequency, and the class of average claim frequency are also provided. The descriptions of the clusters can be summarized as follows:

- **Cluster 0**: Old cars with a lot of unknown data (including small vehicles with two or three wheels and vans).
- **Cluster 1**: Old high-speed cars that are not eco-friendly.
- **Cluster 2**: Heavy, new, eco-friendly, high-speed, and high-power automatic cars.
- **Cluster 3**: Light, affordable, low-power, and eco-friendly cars.

Cluster	Description	#License Plates	Claim frequency	
			Average	Class
0	Average weight and power. Oldest cars and second cheapest. A lot of unknowns (e.g. eco-friendliness, class of top speed, emission, brand, and car type). A lot of Volkswagens and (delivery) vans.	[redacted]	0.648	***Highest***
1	Average weight and second oldest. Has the most seats and cylinders. Least eco-friendly, but a high class of top speed. Combi cars and the most common brand is Volkswagen (a lot of Golfs).	[redacted]	0.534	Intermediate
2	Only cluster that has automatic as the most common transmission. Heaviest and newest cars with the most power. Most doors and most expensive cars that are eco-friendly and have a high class of top speed. Combi cars and the most common brand is Mercedes-Benz (also a lot of BMW's and Audi's).	[redacted]	0.430	Low
3	Lightest cars with the least amount of power, seats and doors. Cheapest cars with the least amount of cylinder volume and number of gears (lower bound). Low class of top speed, eco-friendly, and the only cluster that has no turbo as the most common turbo index. Hatchbacks and the most common brand is Toyota (also a lot of Fiat's and Peugeot's).	[redacted]	0.407	***Lowest***

Figure A.7: This figure shows a table of the descriptions for all clusters. The number of license plates, average claim frequency, and class of average claim frequency are also provided for every cluster.

Evaluation: Similar to the K-prototypes clusters, the clusters generated by the spectral clustering method are considered logical by experts in the actuarial field. For instance, vehicles with lots of unknown variables exhibit a higher average claim frequency, while lighter and cheaper cars demonstrate lower frequencies, possibly due to the owners' decreased likelihood of filing insurance claims for such vehicles. It is difficult to determine which method produces more logical groups. Therefore, a quantitative comparison is provided in Subsection 5.2.5.

A.2. Zip code clustering of the WAM dataset

A.2.1. K-prototypes method

For the WAM zip code dataset, the K-prototypes algorithm was run with values of K ranging from 2 to 13 to create the elbow plot shown in Figure A.8. The elbow shape is observed around $K = 5$, indicating that the optimal number of clusters for the zip codes ("ZC") in the WAM dataset is five.

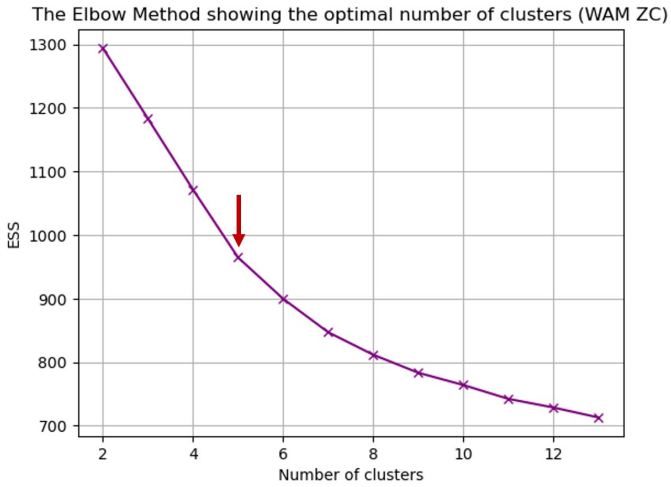


Figure A.8: This figure shows the elbow plot for the WAM zip code dataset. The elbow shape is observed around $K = 5$, indicating that the optimal number of clusters for the zip codes ("ZC") in the WAM dataset is five.

Figure A.9 shows the box plots of each cluster regarding the claim frequency. The distinct averages and variations observed in these box plots once again indicate that the clusters are significant in terms of claim frequency, making them suitable for inclusion as risk factors in the GLM.

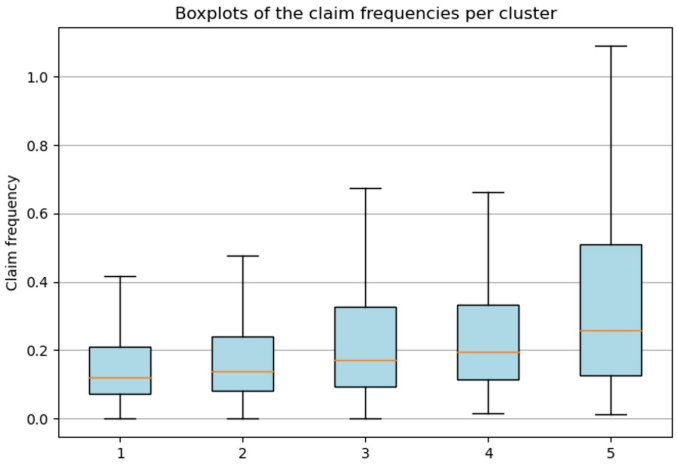


Figure A.9: This figure shows the box plots of each cluster regarding the claim frequency. It can be observed that every cluster has a distinct average and variation in its box plot.

Figure A.10 presents a table of the descriptions of all clusters. For each cluster, the number of zip codes, average claim frequency, and the class of average claim frequency are also provided. The descriptions of the clusters can be summarized as follows:

- **Cluster 0**: Rural areas in the southern part of the country and Flevoland.
- **Cluster 1**: Rural areas in the northern part of the country.
- **Cluster 2**: "Middle of the road" areas.
- **Cluster 3**: Rich suburbs with a high level of education.
- **Cluster 4**: Urban and rural areas characterized by a low education level and social class.

Cluster	Description	#Zip codes	Claim frequency	
			Average	Class
0	Contains most zip codes, fintype: advice-sensitive families, geotype: intellectual culture lovers, hometype: detached house villagers. A lot of families with older children. Low urbanisation, high social class, average income, and high education. Average age is intermediate, has the most motor cycles and people here drive the most km.	[redacted]	0.158	***Lowest***
1	On each zip code least amount of people, fintype: financial professionals, geotype: intellectual culture lovers, hometype: older homeowners. A lot of elderly single people. Low urbanisation, high social class, and average income and education. Low number of non-western immigrants, newest homes (70s), and highest average age. Not many money borrowers, lowest chance at switching health insurance, and lowest risk of defaulting. People drive the least km.	[redacted]	0.181	Low
2	Fintype: unknown, geotype: unknown, hometype: unknown. A lot of unknown data such as age of inhabitants and houses. Average urbanisation, social class, income, and education. High chance of switching car brands.	[redacted]	0.246	Intermediate
3	Fintype: financial professionals, geotype: intellectual culture lovers, hometype: terrace house families. A lot of families with older kids. High social class, income, percentage of company cars, and education. Average age, old houses, a lot of investors and a lot of glossy readers.	[redacted]	0.260	Intermediate
4	Fintype: passive laymen, geotype: passive minima, hometype: single apartment tenants. Mix of intermediate and high urbanisation. Low social class, average income, and average age. Low education and a lot of money borrowers (not many investors or savers). Old and small cars, high chance of switching car brands, highest risk of defaulting, and youngest moms when having their firstborn.	[redacted]	0.400	***Highest***

Figure A.10: This figure shows a table of the descriptions for all clusters. The number of zip codes, average claim frequency, and class of average claim frequency are also provided for every cluster.

Figure A.11 shows the four-digit zip code clusters of the Netherlands generated by the K-prototypes method. This map is created by taking the mode of the six-digit zip code clusters over each four-digit region. The darker the color in the map, the higher the average claim frequency of the corresponding cluster.
It is evident that cluster 0 encompasses the largest region on the map and that cluster 4 appears in both urban and rural areas.

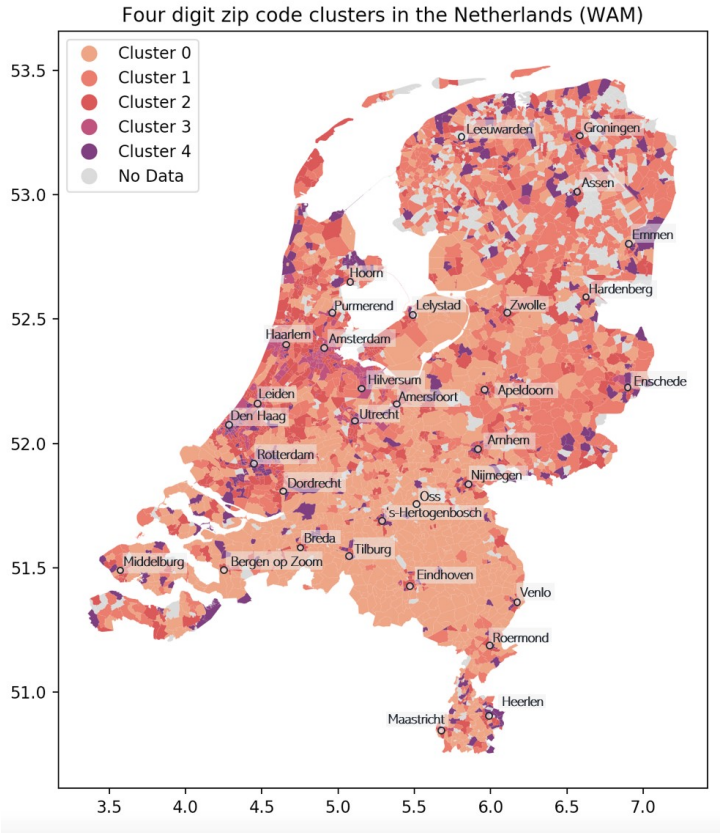


Figure A.11: This figure shows the four-digit zip code clusters of the Netherlands generated by the K-prototypes method. This map is created by taking the mode of the six-digit zip code clusters over each four-digit region.

Figure A.12 shows the zip code clusters of Amsterdam produced with the K-prototypes method. It is apparent that densely populated areas, like the city center, are classified under the "middle of the road" cluster (i.e. cluster 2), while upscale neighborhoods such as those surrounding the canals (i.e. "Grachtengordel West") and "Oud-Zuid" belong to cluster 3. Notably, cluster 0 is absent from this map.

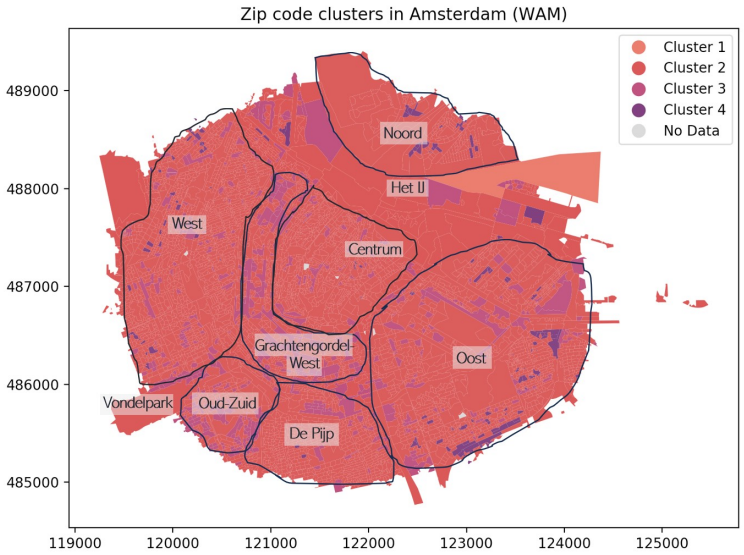


Figure A.12: This figure shows the zip code clusters of Amsterdam produced with the K-prototypes method.

Lastly, Figure A.13 displays the zip code clusters of Amsterdam and its surrounding area. Cluster 2 encompasses the largest region and most cities are categorized under cluster 3. Economically disadvantaged neighborhoods like the Bijlmer, Nieuw-West, and Noord are grouped into cluster 4.

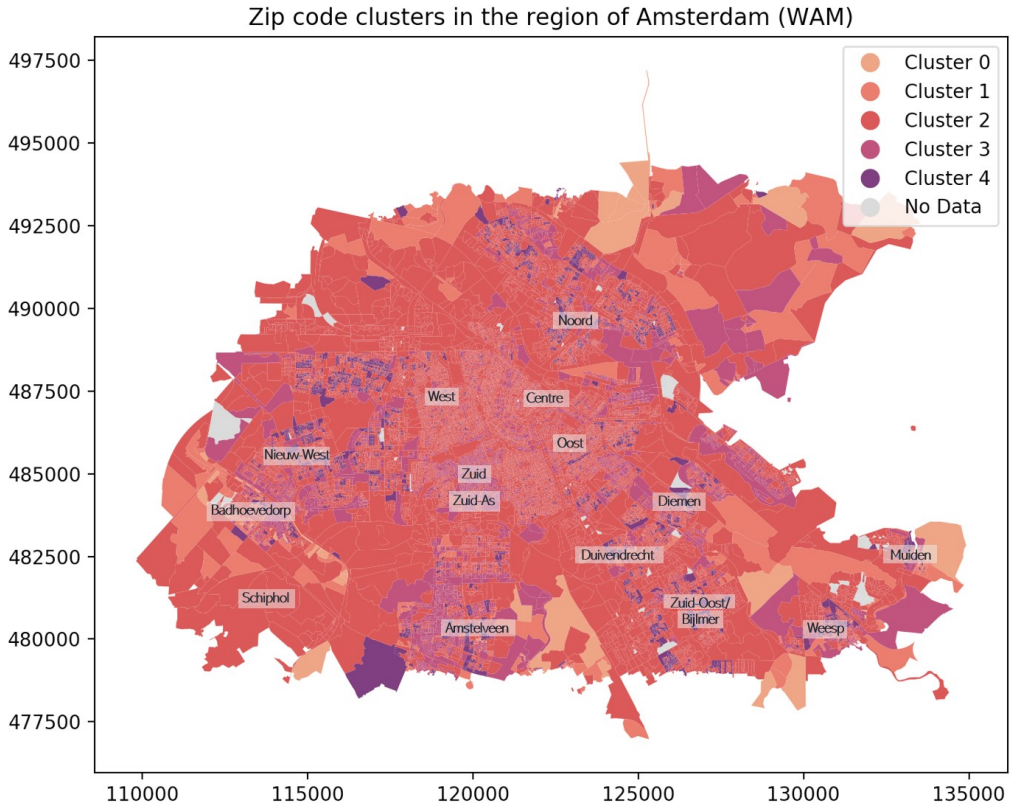
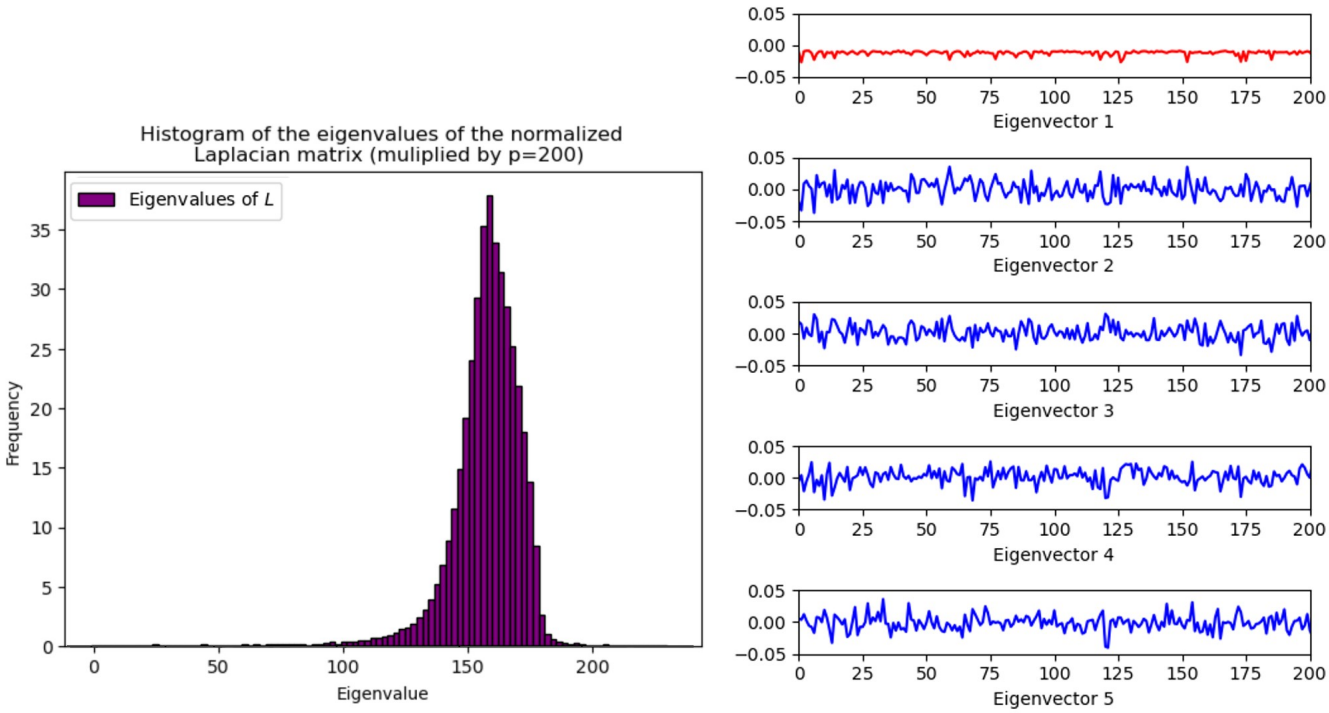


Figure A.13: This figure shows the zip code clusters of Amsterdam and its surrounding area.

Evaluation: According to experts in the actuarial field, the clusters lack practical relevance. For example, in Figure A.12, rich suburbs such as "Grachtengordel West" and "Oud-Zuid" exhibit higher claim frequencies than other parts of Amsterdam, which contradicts real-world data. Furthermore, Figure A.11 shows areas with high claim frequencies in the northern part of the Netherlands, a pattern that does not align with actual observations.

A.2.2. Modified spectral clustering method

The WAM zip code dataset consists of $n = 23,838$ data points. Therefore, the $U\text{-}SPEC$ method was again applied with $n \gg p' = 2000$ to ensure a sufficient number of rep-clusters z_1. Moreover, p, z_1, k, and k' are as defined for the ARD license plate dataset. These parameters yield a high-dimensional dataset since the ratio of dimensions to observations is equal to $114/p = 114/200 = 0.57$. Therefore, the number of clusters is again determined with the method outlined in Subsection 4.3.3.
Figure A.14a shows a histogram of the eigenvalues of the normalized Laplacian, multiplied by $p = 200$, with $U\text{-}SPEC$. In Figure A.14b, the eigenvectors corresponding to the five isolated eigenvalues of this histogram are depicted. Since eigenvector 1 is non-informative, there are four informative eigenvectors, indicating that the optimal number of clusters is *at most* four.



(a) (b)

Figure A.14: This figure shows, for the WAM zip code dataset: **(a)** a histogram of the eigenvalues of the normalized Laplacian (multiplied by $p = 200$), **(b)** the eigenvectors corresponding to the five isolated eigenvalues of the histogram (eigenvector 1 is non-informative and shown in red).

The *U-SPEC* algorithm is completed with four clusters and Figure A.15 shows the box plots of each of these clusters regarding the claim frequency. The distinct averages and variations observed in these box plots indicate that the clusters are significant in terms of claim frequency, making them suitable for inclusion as risk factors in the GLM.

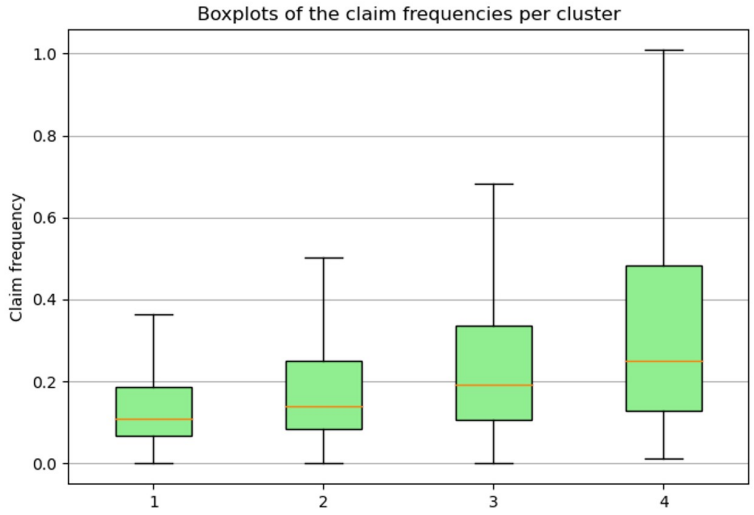


Figure A.15: This figure shows the box plots of each cluster regarding the claim frequency. It can be observed that every cluster has a distinct average and variation in its box plot.

Figure A.16 displays the box plots of each cluster's claim frequency for both the K-prototypes (in blue) and spectral clustering (in green) methods. For cluster 2, spectral clustering shows greater variation in claim frequency. However, for clusters 1 and 4, the variation is greater with K-prototypes. Therefore, it can be concluded that the K-prototypes clusters generally exhibit greater variation in

claim frequency, indicating that spectral clustering more effectively maximizes the homogeneity among observations within the same cluster.
Note that the "middle of the road" cluster of K-prototypes (i.e. cluster 2) was omitted since spectral clustering did not produce this cluster.

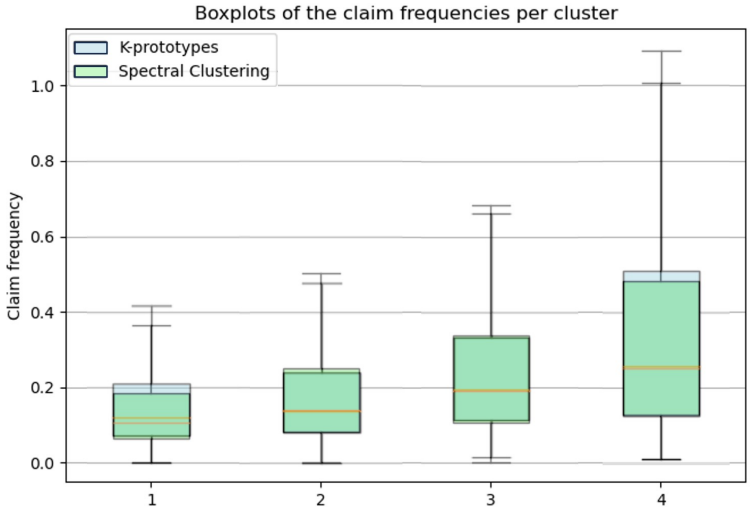


Figure A.16: This figure shows the box plots of each cluster's claim frequency for both the K-prototypes (in blue) and spectral clustering (in green) methods.

Figure A.17 presents a table of the descriptions of all clusters. For each cluster, the number of zip codes, average claim frequency, and the class of average claim frequency are also provided. The descriptions of the clusters can be summarized as follows:

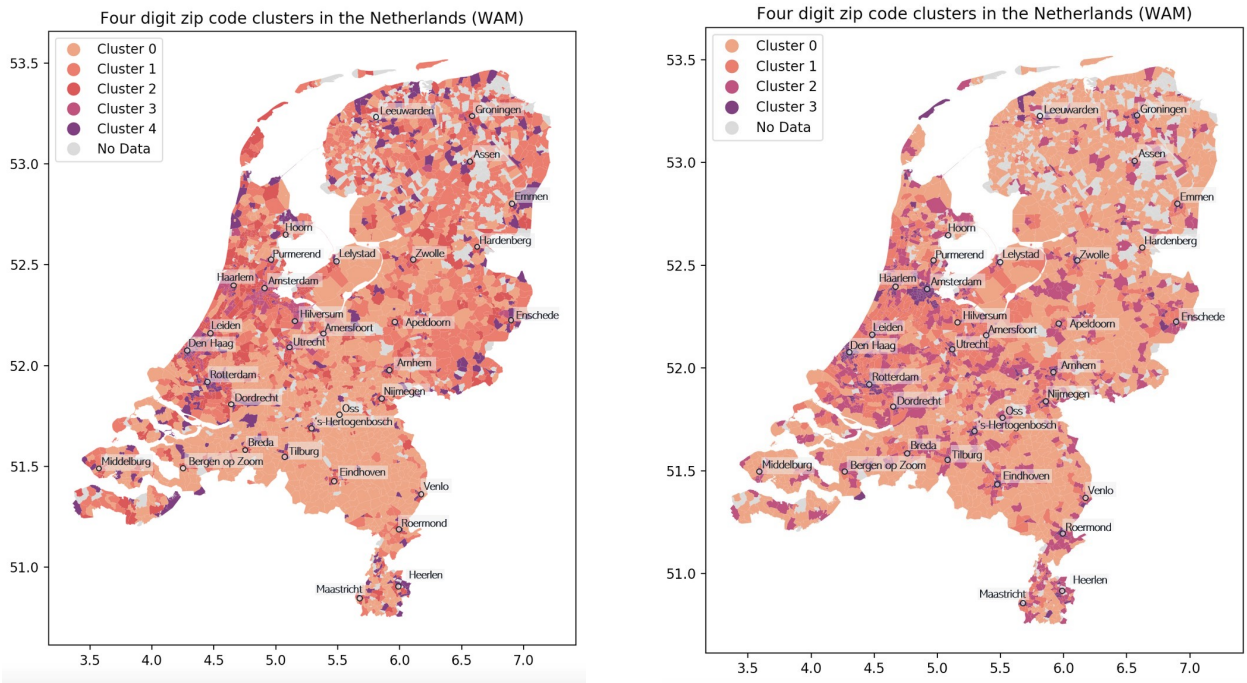
- **Cluster 0**: Rural areas.
- **Cluster 1**: Rich suburbs with a high level of education.
- **Cluster 2**: Regions characterized by newer houses and elderly residents.
- **Cluster 3**: Urban areas with a high population density.

Note that, while these summaries are identical to those of the K-prototypes zip code clusters from the ARD dataset, the cluster distributions differ in the maps.

Cluster	Description	#Zip codes	Claim frequency	
			Average	Class
0	Fintype: advice-sensitive families, geotype: creative environment lovers, hometype: older home owners. Lowest urbanisation and large distances to supermarkets and banks. Intermediate social class and age and the lowest income with cluster 3. Oldest and biggest cars and highest WOZ-value, highest house area, and the most motor cycles.	[redacted]	0.142	***Lowest***
1	Contains most zip codes, fintype: financial professionals, geotype: intellectual culture lovers, hometype: unknown. Intermediate urbanisation, highest social class, income, and education. Highest percentage of company cars, most expensive and biggest cars, and the most dual-income households. A lot of unknown (e.g. WOZ value, average age, and chance of switching health insurance).	[redacted]	0.188	Low
2	Fintype: advice sensitive families, geotype: conservative benefactors, hometype: older home owners. Same urbanisation level as cluster 1, newest houses (70s), and intermediate social class and income. Highest average age and a lot of older couples without kids.	[redacted]	0.261	Intermediate
3	Fintype: active borrowers, geotype: passive minima, hometype: single apartment tenants. Highest urbanisation, youngest people, lowest social class, and most non-western immigrants. Lowest average income, least dual-income households, smallest house area, and least amount of car owners. Most amount of borrowers, highest chance of switching health insurance, and highest chance of defaulting.	[redacted]	0.367	***Highest***

Figure A.17: This figure shows a table of the descriptions for all clusters. The number of license plates, average claim frequency, and class of average claim frequency are also provided for every cluster.

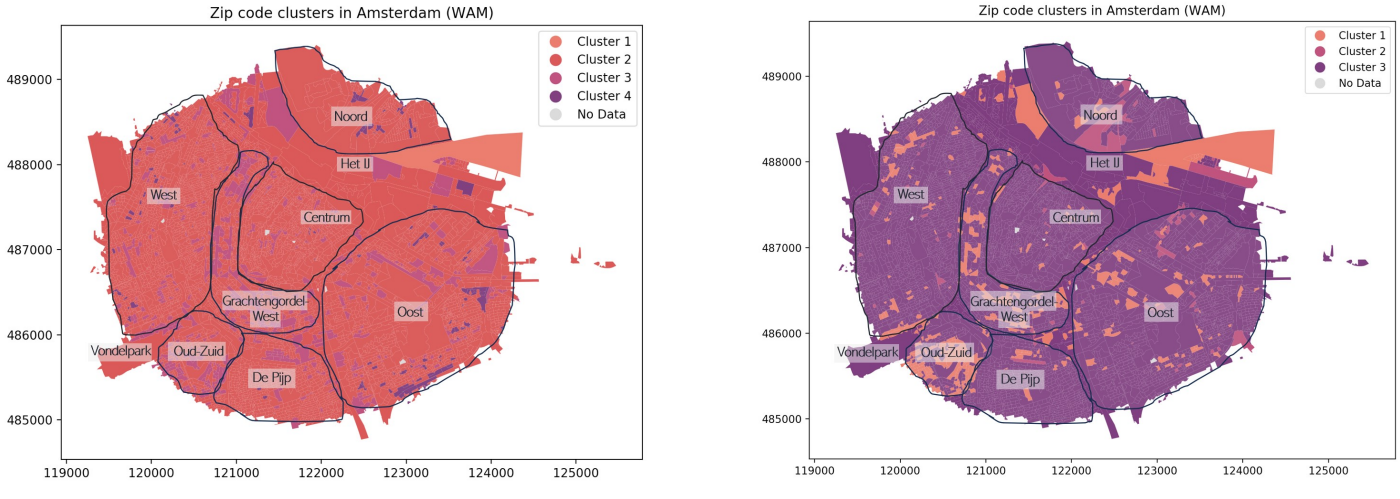
Figure A.18 shows the four-digit zip code clusters of the Netherlands generated by K-prototypes and spectral clustering. Notably, the clusters differ between the two methods.
Furthermore, the clusters in Figure A.18b bear resemblance to the spectral zip code clusters of the ARD dataset depicted in Figure 5.22b. However, for the WAM dataset, cluster 2 is less concentrated in the southern region and is more widely distributed across the country.



(a) **(b)**

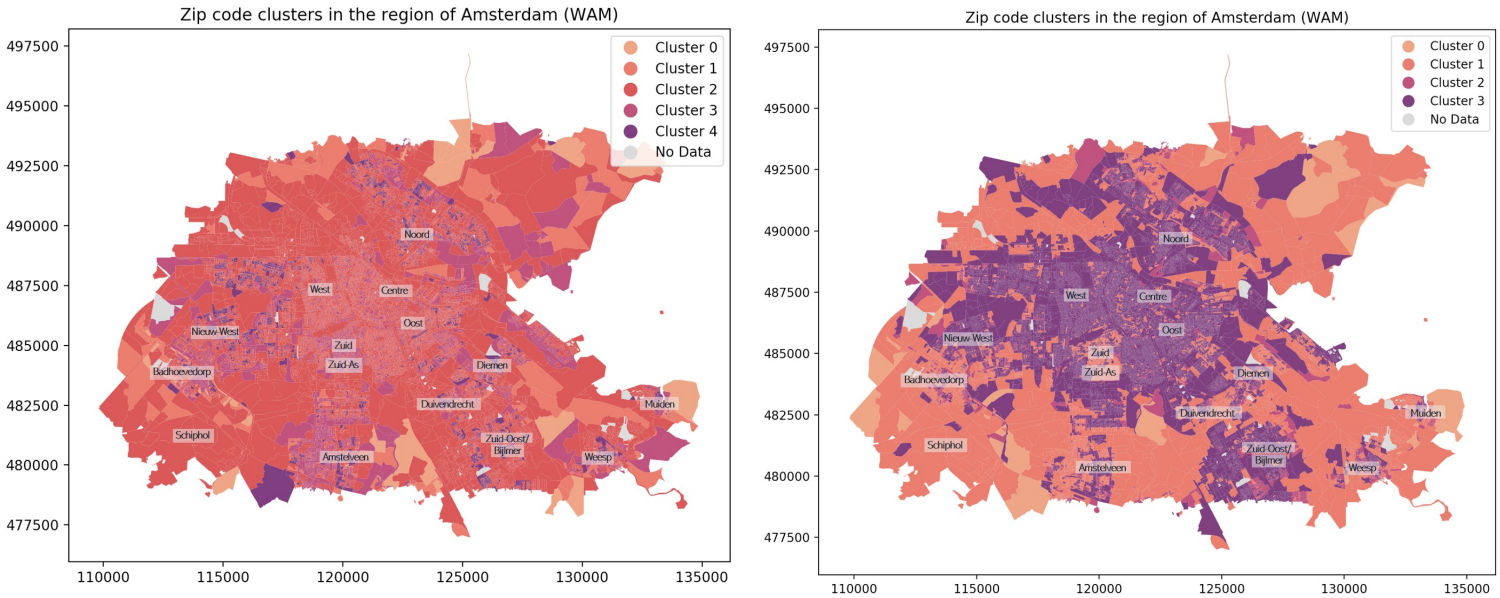
Figure A.18: This figure shows the four-digit zip code clusters of the Netherlands generated by: **(a)** K-prototypes, **(b)** Spectral clustering.

Figure A.19 shows the zip code clusters of Amsterdam produced with K-prototypes and spectral clustering, while Figure A.20 extends this comparison to Amsterdam and its surrounding area. The spectral clustering algorithm appears to produce the same zip code clusters as those of the ARD dataset with the spectral clustering method (i.e. those shown in Figures 5.23b and 5.24b).



(a) **(b)**

Figure A.19: This figure shows the zip code clusters of Amsterdam produced with: **(a)** K-prototypes, **(b)** Spectral clustering.



(a) **(b)**

Figure A.20: This figure shows the zip code clusters of Amsterdam and its surrounding area produced with: **(a)** K-prototypes, **(b)** Spectral clustering.

Evaluation: The clusters generated by the spectral clustering method are considered logical by experts in the actuarial field. For example, cars belonging to urban areas tend to exhibit a higher average claim frequency compared to those belonging to rural areas.
Furthermore, as explained in the previous subsection, the K-prototypes clusters lack practical relevance. For example, with the K-prototypes technique, rich suburbs such as “Grachtengordel West” and “Oud-Zuid” exhibit higher claim frequencies than other parts of Amsterdam, which contradicts real-world data. The spectral clustering results correctly suggest that these richer suburbs have lower claim frequencies than the rest of Amsterdam. Therefore, it can be concluded that spectral clustering produces more accurate clusters than K-prototypes in this case. Nevertheless, a quantitative comparison is provided in Subsection 5.2.5.