

Faculté des sciences

Spatial Risk Zonation of Belgian Flood Reinsurance

A Clustering Approach to Catastrophe Model
Output

Author: Pablo Jérôme LETELIER DAUBY

Supervisor: Eddy VANBENEDEN

Readers: Michel DENUIT

Academic year 2025–2026

Master [120] en sciences actuarielles

Contents

Acknowledgements	4
Introduction	5
1 Catastrophe Models	7
1.1 The Need for Catastrophe Models: Limits of Traditional Actuarial Methods	7
1.2 The Anatomy of a Catastrophe Model	8
1.2.1 The Hazard Module	9
1.2.2 The Exposure Module	9
1.2.3 The Vulnerability Module	9
1.2.4 The Financial Module	10
1.3 Model Output: the Event Loss Table and Damage Distribution Table . . .	10
2 Geography and Hydrology of Belgium	11
2.1 Drainage Basins	12
2.1.1 The Scheldt Basin	12
2.1.2 The Meuse Basin	12
2.1.3 The Yser Basin	13
2.2 Zone Mapping of Basin and sub-Basin Membership	13
2.3 Basin Structure as the Foundation for Risk Zonation	14
3 Data Overview	15

3.1	Damage distribution Table	15
3.1.1	Descriptive statistics	16
3.1.2	Compositional structure and transformation	18
3.2	Event Loss Table	18
3.2.1	Descriptive statistics	19
3.3	Summary and motivation for clustering	21
4	Exceedance Probability	23
4.1	The Collective Risk Model Framework	24
4.2	Occurrence Exceedance Probability (OEP)	24
4.2.1	Derivation	25
4.2.2	Exceedance Frequency and Return Period	26
4.2.3	Exceedance Frequency Decomposition	27
4.3	Beta distribution	28
5	Clustering	30
5.1	Representative-based Clustering	30
5.1.1	K-Means	30
5.1.2	K-Medoids	31
5.1.3	Gaussian Mixture Models	31
5.2	Energy Distance Clustering	32
5.2.1	Hierarchical Energy Clustering	33
5.2.2	K-Groups	34
5.3	Clustering Evaluation	34
5.3.1	Internal Evaluation	35
5.3.2	External Evaluation	37
5.4	K -selection	38

5.5	Results	39
5.6	Conclusions	46
6	Reinsurance and Solvency II Applications	49
6.1	The Gross Solvency Capital Requirement	49
6.1.1	Portfolio and Cluster Level	49
6.2	XL Reinsurance as SCR Mitigation	50
6.2.1	Net SCR and Capital Relief	51
6.2.2	Ceded Annual Average Loss	52
6.2.3	Effect on the Exceedance Frequency Shares	53
6.3	Solvency II Use Test	54
6.4	Climate Change and Future Capital Needs	54
6.5	Conclusions	55
	General Conclusion	56

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Eddy Vanbeneden and Marc Melsen, for their guidance, availability, and support throughout the development of this thesis. Their expertise and passion for the subject made this work a genuine learning experience.

I also wish to thank my family and friends for their constant encouragement. A special word of gratitude goes to my father, mother and sister, who despite the distance never failed to make their support felt.

Editorial note: the redaction of this thesis was assisted by Claude, developed by Anthropic. All content, results, analyses, and conclusions are entirely the product of the author's own knowledge and work; the assistance was limited strictly to matters of written expression.

Introduction

State of the Art

The quantification of natural catastrophe risk has undergone a fundamental transformation over the past three decades. Traditional actuarial methods are structurally unsuited to catastrophe risks: a single flood event can simultaneously affect thousands of policies across a wide geographic area, violating the independence assumption and rendering purely experience-based frequency-severity models unreliable for tail estimation. The response to this limitation was the development of probabilistic catastrophe models in the late 1980s and early 1990s, accelerated by the devastating losses of Hurricane Andrew (1992), which exposed the industry's systematic underestimation of extreme event risk [20]. These models simulate large stochastic catalogues of physically plausible events, extending far beyond the historical record, and translate geophysical science into actuarially usable loss distributions through a modular architecture comprising hazard, exposure, vulnerability, and financial components.

A parallel development in the industry has been the adoption of standardised geographic risk segmentation frameworks, of which the most prominent is the CRESTA (Catastrophe Risk Evaluating and Standardizing Target Accumulations) system [10]. Established in 1977 as a joint initiative of major reinsurers, CRESTA zones generally follow administrative boundaries such as province boundaries or merged postal code areas, rather than hydrological divides. At the European level, more hydrologically grounded efforts exist. Flood hazard and risk mapping has matured across EU member states around river basin districts, with current practice now well established at strategic and river basin scales. As mentioned by European Commission et al. [13], however, regional and national hydrological models are potentially too coarse in resolution for local detail and decision making, which encourages insurers to develop their own internal models tailored to industry needs. Beyond the resolution issue, such internal models require loss distributions conditioned on the insurer's own exposure, something no public hazard map can provide regardless of its spatial granularity.

Contributions and Structure of this Work

This thesis develops a data-driven framework for the geographic zonation of Belgian inland flood risk, grounded in the output of a probabilistic catastrophe model rather than in administrative boundaries. The central innovation is the use of a *Damage Distribution Table* (DDT) as the empirical basis for clustering. For each simulated flood event, the DDT records the proportion of total portfolio losses allocated to each of the eleven Belgian zones, encoding the spatial footprint of the event directly from the model’s own loss estimates. Unlike approaches that group zones by geographic or administrative criteria, this allows a data-driven algorithm to recover the underlying hydrological structure of Belgian flood risk without imposing any prior zoning constraint.

The thesis is organised in six chapters. Chapter 1 introduces the historical development of catastrophe modelling and the architecture of the models, contextualising the two datasets used throughout: the Event Loss Table (ELT) and the DDT. Chapters 2 and 3 establish the geographical and empirical motivation for clustering. The former maps Belgium’s drainage basins onto its administrative provinces, demonstrating that flood-generating mechanisms differ fundamentally between the Scheldt and Meuse systems. The latter confirms that this hydrological structure is recoverable from the data itself. Chapter 4 provides the theoretical framework of the Occurrence Exceedance Probability curve and the exceedance frequency decomposition, which are needed to validate and interpret the clustering results actuarially. Chapter 5 is the methodological core, implementing and comparing five clustering families against both geometry-based and energy distance-based evaluation indices. Chapter 6 applies the resulting partition directly to the insurance industry, computing the Solvency Capital Requirement at portfolio and cluster level, analysing a Cat XL reinsurance treaty as a capital mitigation instrument, and connecting the framework to the Solvency II use test and climate change stress scenarios.

The data, code, and additional annexes supporting this thesis are provided in complementary files submitted alongside this document.

Chapter 1

Catastrophe Models

The quantitative assessment of natural catastrophe risk is today inseparable from the use of catastrophe models, which are probabilistic simulation frameworks that translate geophysical science into actuarially usable loss distributions. This chapter introduces the conceptual and technical foundations of catastrophe modelling in order to situate the data and methods used in the remainder of this thesis.

The roots of catastrophe modelling lie at the intersection of two independent traditions, combining the old-legacy practice of mapping insured exposure with the modern natural-hazard sciences, which developed standardised intensity metrics for seismology, meteorology, and hydrology [20]. These two strands converged decisively in the late 1980s, when a sequence of extraordinarily costly disasters, culminating in Hurricane Andrew (1992), exposed the inadequacy of purely experience-based pricing and reserving which forced the industry to adopt model-based loss estimation [20, p. 25]. The 1990s and 2000s brought rapid growth in computing power, enabling the on-premise deployment of catastrophe modelling platforms across the insurance and reinsurance industry [25]. More recently, cloud computing has further extended access to these tools, removing the need for local infrastructure. These developments collectively led to the modelling approach on which the present thesis relies, and whose outputs, the Event Loss Table and the Damage Distribution Table, are described in detail in Chapter 3.

1.1 The Need for Catastrophe Models: Limits of Traditional Actuarial Methods

Traditional property insurance pricing rests on the actuarial principle that expected losses can be estimated from historical experience, provided that the loss events are sufficiently numerous and approximately independent. This condition is satisfied for frequent, uncorrelated perils such as motor accidents or household fires, where the law of large numbers stabilises loss estimates across a large portfolio. Catastrophe events violate both assumptions: they are rare and spatially correlated, meaning that a single event can simultaneously damage thousands of policies [20].

The problem of data scarcity is particularly acute for the tail events that concern reinsurers most. Observational records of natural hazards are typically short relative to the time periods of interest to the insurance industry, which routinely extend to 10^2 – 10^3 years. Consequently, any purely empirical frequency-severity model calibrated on historical claims will systematically underestimate the probability and magnitude of extreme events [20].

A further complication is the correlated nature of catastrophe losses. When a major flood event occurs, losses across an insurer’s Belgian portfolio do not behave as independent draws from a common distribution; they are driven by the same meteorological event and may all be large simultaneously. This breaks the diversification logic underpinning traditional premium calculation and requires insurers to hold capital against scenarios that may never have been observed [20].

Catastrophe models address these limitations by simulating a large stochastic catalogue of physically plausible events. This includes events far more severe than any in the historical record and computing the implied loss distribution for a given portfolio [37]. The result is a probabilistic loss distribution that can inform pricing, capital allocation, and reinsurance purchasing in a way that pure experience-rating cannot.

1.2 The Anatomy of a Catastrophe Model

Despite the diversity of perils they cover, commercial catastrophe models share a common modular architecture. The literature consistently identifies four core components: a **hazard** module, an **exposure** (or inventory) module, a **vulnerability** module, and a **financial** (or loss) module, integrated within a software platform [25, 20]. Figure 1.1 summarises this architecture.

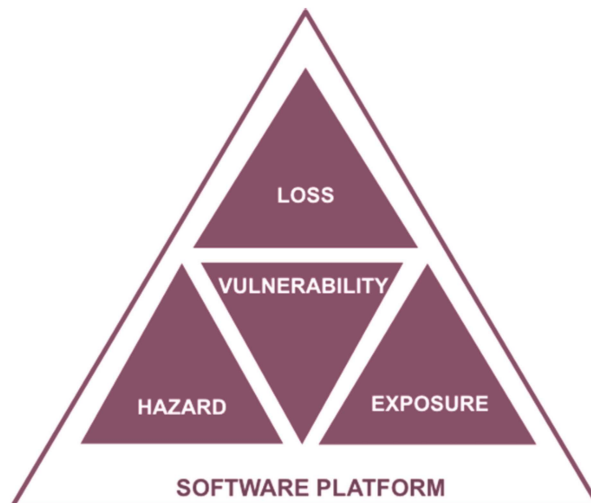


Figure 1.1: Schematic structure of a catastrophe model (hazard, exposure, vulnerability, and financial/loss). Jones et al. [25, Fig. 1.2]

1.2.1 The Hazard Module

The hazard module characterises the physical intensity of a peril across the geographic domain of interest. It operates in two stages. First, a *stochastic event generator* produces a large synthetic catalogue of plausible events, calibrated on historical observations but extending to time periods well beyond those in the observational record [37]. Because records rarely exceed a century, the stochastic module must extrapolate using extreme-value methods — typically Generalised Pareto Distributions fitted to the upper tail of the observed intensity distribution [37, p. 2310]. Second, each synthetic event is translated into a spatially distributed *hazard footprint*: a map of intensity across the modelled domain. For floods, the relevant metric is typically inundation depth, derived from a hydrological rainfall-runoff model coupled to a hydraulic model [25, 37].

1.2.2 The Exposure Module

The exposure module characterises the portfolio of assets at risk. The fundamental input is the *location* of each insured property, typically obtained through geocoding [20]. Beyond location, the module records characteristics that determine how a structure responds to a given hazard intensity. These structures refer to construction type, building age, number of stories, and occupancy class [20]. Where individual-level data are unavailable, insurers use industry exposure databases or proxy data such as impervious surface area to distribute aggregate sums insured across a geographic grid [25].

The spatial granularity of exposure data is a critical driver of model accuracy. Aggregating insured values to postcode level introduces uncertainty because it assumes a uniform spatial distribution of risk within each zone, which may be a poor approximation in heterogeneous urban environments.

1.2.3 The Vulnerability Module

The vulnerability module is the interface between physical hazard and financial loss. Damage functions are inherently uncertain. For flood, the relevant factors include not only water depth but also inundation duration, flow velocity, building construction type, and occupancy [37]. Empirical calibration is limited by the scarcity of post-event claims data at the extremes of the loss distribution, so vulnerability functions in the tail must rely heavily on engineering judgement and synthetic models [25]. Commercial models typically represent this uncertainty by sampling damage ratios from a beta distribution, whose parameters vary with hazard intensity, rather than applying a single deterministic curve [37].

1.2.4 The Financial Module

The financial module translates physical damage estimates into net insured losses by applying the policy conditions that govern each contract: deductibles, coverage limits, co-insurance clauses, and reinsurance structures. Across the stochastic event catalogue, the financial module aggregates per-event losses into portfolio-level metrics and produces the exceedance probability curves that are the primary deliverable of the modelling exercise [37]. The mechanics of these curves are described in Section 1.3.

1.3 Model Output: the Event Loss Table and Damage Distribution Table

For the purpose of this thesis, the older vintage inland flood model in RMS RiskLink for Europe was used. This model was among the first probabilistic catastrophe models designed to quantify both fluvial and pluvial flood risk across multiple European countries. Prior to the High Definition upgrades introduced around 2016, it operated at coarser spatial resolutions with more limited treatment of cross-border flood correlation and localised surface-water flooding. With this in mind, the analysis is based on relative modelling outputs per province in Belgium rather than absolute loss estimates.

The four-module pipeline described in Section 1.2 produces, for each simulated event, two complementary outputs that constitute the empirical foundation of this thesis. The **Event Loss Table** (ELT) records the occurrence rate, mean expected loss, and loss variability of each event, summarising the financial severity of the full stochastic catalogue at the portfolio level [25]. By running the ELT at the provincial level, a working table is derived that records, for each event, the proportion of total losses allocated to each of the eleven Belgian provinces, capturing the spatial footprint of the event across the territory. We refer to this derived table as the **Damage Distribution Table** (DDT) throughout the thesis for convenience, though this is not a standard industry term. Together, the ELT characterises the financial dimension of each event while the DDT characterises its geographic dimension. The structure, statistical properties, and role of both datasets in the analysis are described in detail in Chapter 3.

Chapter 2

Geography and Hydrology of Belgium

Belgium's 30,528 km² territory drains almost entirely northward to the North Sea through two large scale river basins, the Scheldt in the west and centre and the Meuse in the east and south, plus the smaller coastal Yser basin in the northwest. Minor drainage to the Rhine (via the Geul in Liège province) and to the Seine (a small headwater area near Momignies in Hainaut) accounts for a combined 847 km² of Walloon territory and is negligible at the national scale [18]. These basins differ in topography, geology, and runoff behaviour, and consequently in the nature of the flood events they generate. This chapter provides an overview of each basin and maps their extent onto the ten provinces and the Brussels Capital Region, establishing the hydrological context that motivates the risk-zonation approach developed throughout the thesis. Figure 2.1 maps the river basins of Belgium while Figure 2.2 maps the provinces and capital region of Belgium.

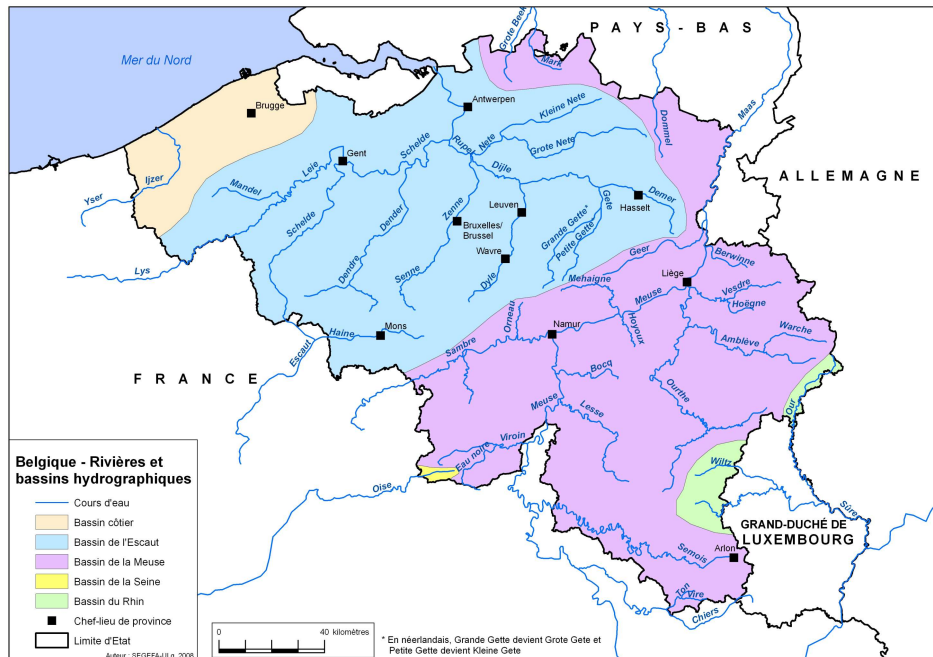


Figure 2.1: Belgian hydrological basins. SEGEFA [39].

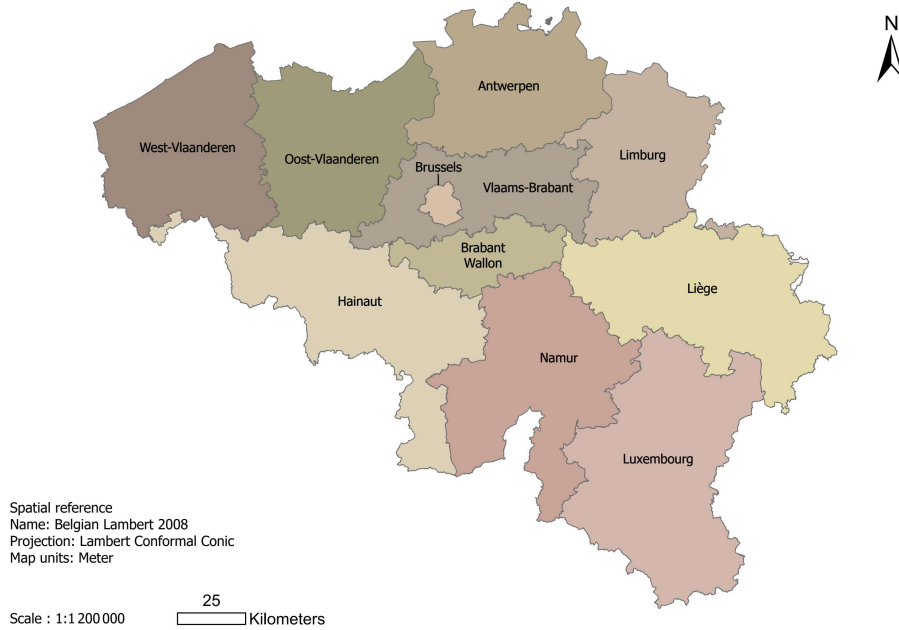


Figure 2.2: Administrative provinces and the Brussels Capital Region of Belgium. Elaboration using ArcGIS (Esri).

2.1 Drainage Basins

2.1.1 The Scheldt Basin

The Scheldt rises in northern France and flows approximately 350 km northward through Belgium before reaching the Dutch tidal estuary, draining a total basin of approximately 21,580 km² [23], with the large majority in Flanders. It is a lowland, rainfall-fed system in which floods develop from large-scale, sustained precipitation accumulations over the flat central plateau [42].

2.1.2 The Meuse Basin

The Meuse rises in France and flows roughly 950 km to the North Sea across five countries. Its Walloon portion covers 12,283.5 km², organised into eight sub-basins [18]. In contrast to the Scheldt, the Meuse is characterised by fast-onset, high-peak floods driven by its Ardennes sub-catchments, where steep valleys, thin soils, and high precipitation produce rapid runoff responses [24, 18]. The July 2021 event illustrates this potential, with rainfall accumulations exceeding the 200-year return level in the Vesdre basin, causing 39 fatalities and an estimated EUR 2.8 billion in economic damage [26]. Because the two basins respond to different meteorological precursors, they can flood independently, with markedly different spatial footprints [24].

2.1.3 The Yser Basin

The Yser flows 78 km northward through West-Vlaanderen before discharging into the North Sea at Nieuwpoort, draining approximately 1,101 km² [47]. Its low-lying coastal plain, with about 35% polder land at or near sea level, makes it hydrologically distinct from both the Scheldt and the Meuse [46], and Flanders recognises it as a separate river basin in its water management framework [42].

2.2 Zone Mapping of Basin and sub-Basin Membership

Administrative provinces and the Brussels Capital Region do not follow hydrological divides. Table 2.1 maps each administrative unit to its primary basin(s).

Province / Region	Federal Region	Primary basin(s) and sub-basin(s)
West Flanders	Flanders	Scheldt (Leie, Upper Scheldt), Yser
East Flanders	Flanders	Scheldt (Upper/Lower Scheldt, Leie, Dender)
Antwerp	Flanders	Scheldt (Rupel, Nete, Lower Scheldt)
Flemish Brabant	Flanders	Scheldt (Dyle-Zenne, Demer)
Limburg	Flanders	Scheldt (Demer), Meuse
Brussels	Capital Region	Scheldt (Senne / Zenne)
Walloon Brabant	Wallonia	Scheldt (Dyle-Gette, Senne)
Hainaut	Wallonia	Scheldt (Escaut-Lys, Haine, Dendre), Meuse (Sambre)
Namur	Wallonia	Meuse (Sambre, Lesse, upper and lower Meuse)
Liège	Wallonia	Meuse (Ourthe, Amblève, Vesdre, lower Meuse)
Luxembourg	Wallonia	Meuse (Ourthe, Semois-Chiers)

Table 2.1: Primary basin membership by Belgian province and the Brussels Capital Region, based on Gouvernement wallon [18] for Walloon sub-basins and Vlaamse overheid [42] for Flemish sub-basins.

Several units straddle basin boundaries or present particular hydrological complexity, a consequence of the fact that administrative provinces were drawn along political and historical lines rather than hydrological ones. Limburg drains mainly to the Scheldt via the Demer sub-basin but its eastern fringe borders the Meuse, exposing it to propagated Meuse floods from upstream Liège and Namur. Hainaut spans the Scheldt-drained central plateau in the north and the Sambre valley in the south, placing it at the interface of the two primary regimes. West-Vlaanderen is divided between Scheldt tributaries in the east and the Yser coastal system in the west [42, 18]. The absence of a one-to-one mapping between administrative provinces and hydrological basins means that these units cannot be unambiguously assigned to a single flood regime on the basis of geography alone. This constitutes an interesting additional challenge for the clustering analysis: rather than partitioning events by province label, the data-driven approach must recover the underlying hydrological structure from the damage patterns themselves, without imposing administrative boundaries as a constraint.

2.3 Basin Structure as the Foundation for Risk Zonation

The hydrological description developed in this chapter already provides a *prima facie* case for spatial risk zonation in Belgium, grounded in physical geography rather than data analysis. Three arguments follow from the basin overview above.

First, the flood-generating processes differ fundamentally between the Scheldt and Meuse systems. Scheldt events are broad, slow-rising, and correlated across Flemish and Brabant provinces, while Meuse events are sharp, fast, and concentrated in the Ardennes valleys of Liège, Namur, and Luxembourg [23, 24]. These are not variations of a single underlying mechanism but structurally distinct regimes driven by different topography, geology, and meteorological precursors. Pluvial flooding adds a further dimension: driven by localised convective precipitation rather than catchment-scale run-off, it is not confined to river valleys and can occur independently of basin membership, with urbanised areas being particularly exposed due to impervious surfaces and constrained drainage capacity [11, 8].

Second, the fluvial basins can flood independently [24]. Their loss distributions therefore have only partial temporal coincidence, as a catastrophic Meuse event need not coincide with any meaningful Scheldt event, and vice versa. Treating the Belgian portfolio as a single risk zone implicitly assumes a degree of spatial correlation that does not hold across basins, while understating the correlation that exists within each basin.

Third, the transition provinces, namely Limburg, Hainaut, and West Flanders, and the Yser coastal system constitute hydrological borderline cases that cannot be resolved by geography alone. Their dual basin exposure means they do not belong unambiguously to either primary regime.

Taken together, these three points establish that the Belgian territory is not hydrologically uniform and that a single risk zone is an inadequate representation for pricing and reinsurance purposes. The catastrophe model used in this thesis covers both fluvial and pluvial flood risk, though its primary structure reflects the basin-level fluvial component. Whether the ELT and DDT actually encode the distinctions outlined here, and how sharply, is an empirical question that the data analysis in Chapter 3 addresses.

Chapter 3

Data Overview

This chapter introduces the datasets used in the probabilistic flood risk assessment conducted in this thesis. The analysis relies on two complementary inputs provided by the catastrophe model: a Damage Distribution Table (DDT) across Belgian provinces and an Event Loss Table (ELT). The former describes how total event losses are spatially allocated across the eleven provinces, while the latter characterises the statistical properties of losses associated with each simulated flood event. Together, these datasets constitute the empirical foundation on which the risk quantification methodology developed in this thesis is built.

3.1 Damage distribution Table

Each row in the DDT corresponds to a unique flood event identified by an **EVENTID**, and the remaining columns represent the eleven selected zones — Antwerpen (ANT), Brabant Wallon (BRA), Brussels (BRU), Hainaut (HAI), Liège (LIE), Limburg (LIM), Luxembourg (LUX), Namur (NAM), Oost-Vlaanderen (OVL), Vlaams-Brabant (VLB), and West-Vlaanderen (WVL) — each recording the proportion of the total flood loss allocated to that province for that event. All row values sum to unity, which makes this a dataset of *compositional data*: observations whose components carry only relative meaning under a constant-sum constraint [19]. Figure 3.1 provides a first visual summary through a PCA of the selected zones proportions in both the raw and standard-scaled spaces. The geometry of the scatter already reflects the big picture of the underlying hydrological structure: the upper region of the standard-scaled PCA corresponds broadly to the *Scheldt* basin, while the lower-left corresponds to the *Meuse* basin.

Although, as its name suggests, the **Brussels Region** is not considered a province under the Belgian administrative system, we will refer it as such in the context of the DDT for the sake of simplicity.

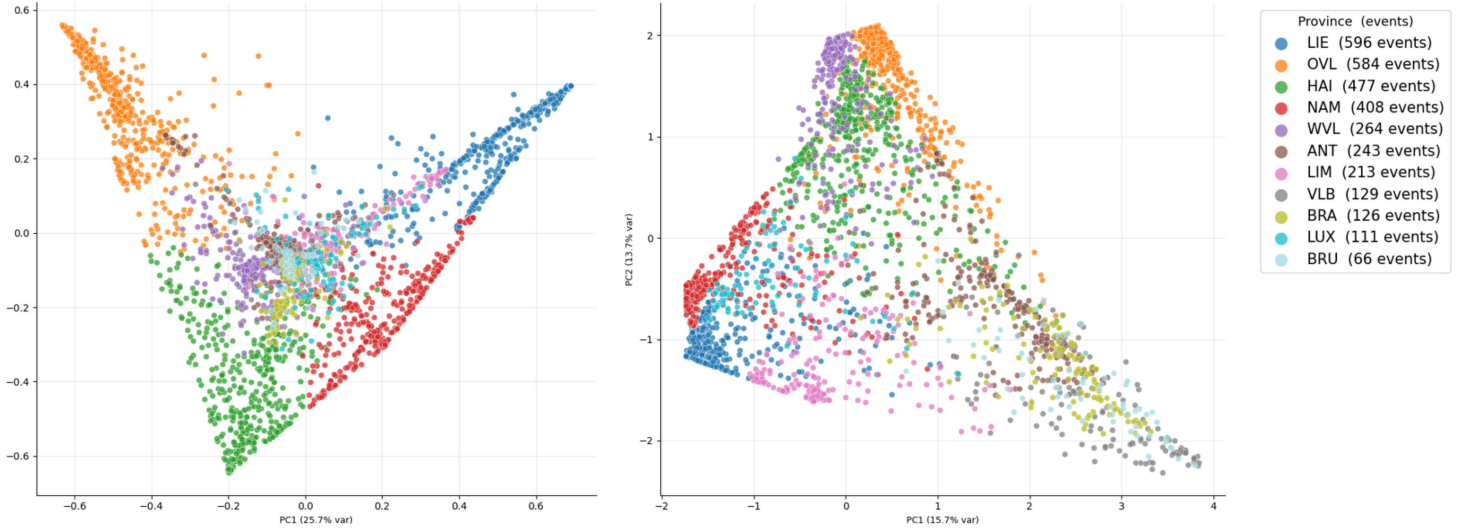


Figure 3.1: PCA of damage proportion data in raw space (left) and standard-scaled space (right), coloured by dominant zone.

3.1.1 Descriptive statistics

Table 3.1 reports the marginal descriptive statistics of the provincial loss proportions, and Figure A.3 (Annex) displays their distributions as boxplots. The distributions are highly right-skewed across all provinces, with medians close to zero for every province while the upper quantiles exhibit substantial variation. This heterogeneity manifests in two ways.

First, the Std/Mean ratio is lowest for HAI, LIE, NAM, OVL indicating that despite their higher absolute variability their losses are relatively more concentrated and predictable in proportion to their mean. A hydrological interpretation can be advanced, as these provinces cover the larger river corridors of the Meuse and Scheldt systems where damage is primarily driven by fluvial flooding and therefore more tightly linked to well-defined catchment-scale hydrological processes. The remaining provinces, with higher Std/Mean ratios, are associated with smaller watercourses and more urbanised catchments where pluvial flooding plays a greater role, introducing higher relative uncertainty in the damage patterns.

Second, the upper tail also varies considerably, as HAI, LIE, NAM, OVL, and WVL show p95 and p99 values approaching unity, indicating that they occasionally account for nearly all losses in a single event. This is consistent with the event dominance patterns in Figure 3.1, where the same high-variability provinces dominate a larger fraction of events. Conversely, BRA, BRU, and VLB dominate far fewer events, reflecting their smaller standard deviations and lighter upper tails.

A notable feature of the data is the high proportion of exact zeros. Table 3.2 reports the zero count and zero proportion for each province. BRU presents the highest sparsity (41.4% zeros), followed by BRA (18.4%) and LIM (13.8%). In contrast, HAI and OVL show much lower zero proportions (3.5% and 4.0% respectively). These zero counts reinforce the observation that flood events tend to concentrate in a small number of provinces rather

Statistic	ANT	BRA	BRU	HAI	LIE	LIM	LUX	NAM	OVL	VLB	WVL
Mean	0.073	0.043	0.024	0.144	0.177	0.077	0.046	0.120	0.155	0.053	0.086
Std	0.186	0.118	0.080	0.239	0.285	0.173	0.138	0.231	0.280	0.122	0.215
Std/Mean	2.548	2.744	3.333	1.660	1.610	2.247	3.000	1.925	1.806	2.302	2.500
p25	0.000	0.000	0.000	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000
p50	0.002	0.001	0.000	0.019	0.015	0.007	0.002	0.006	0.007	0.002	0.002
p75	0.024	0.017	0.006	0.185	0.229	0.062	0.019	0.083	0.127	0.035	0.027
p95	0.544	0.263	0.139	0.743	0.885	0.452	0.292	0.700	0.836	0.334	0.675
p99	0.905	0.650	0.439	0.976	0.991	0.942	0.831	0.948	0.965	0.610	0.981
Max	0.987	0.940	0.821	1.000	1.000	0.998	0.996	0.999	1.000	0.844	1.000

Table 3.1: Descriptive statistics of province loss proportions.

than distributing uniformly across the territory.

Province	ANT	BRA	BRU	HAI	LIE	LIM	LUX	NAM	OVL	VLB	WVL
Count	350	592	1331	113	306	444	321	319	129	249	338
Proportion	0.109	0.184	0.414	0.035	0.095	0.138	0.100	0.099	0.040	0.077	0.105

Table 3.2: Zero counts and zero proportions per province across all flood events.

Table 3.3 reports the conditional distribution of province loss proportions given that a specific province records a share above 25%. Each row represents the typical allocation of losses across all provinces when the conditioning province is significantly affected. The diagonal elements show that whenever a province exceeds the threshold it tends to account for the largest share of the total loss, generally above 50%, with the exception of the more central provinces BRA, BRU, and VLB.

The off-diagonal structure reveals patterns of co-occurrence that reflect the underlying hydrological network. When LIE is heavily affected, LIM, LUX, and NAM tend to absorb a substantial share of the remaining losses, consistent with their shared exposure along the *Meuse* river. Similarly, when OVL is above the threshold, HAI and WVL capture non-negligible proportions, reflecting their contributions to the *Scheldt* system tributaries from the Yser and Lys. The case of HAI is more complex: when it accounts for more than 25% of total damage, the residual losses are distributed among WVL, OVL, and NAM, which cannot be attributed to a single river basin since Hainaut is connected to both the Scheldt and the Meuse through the Sambre, as discussed in Section 2.2. In contrast, BRU, LIM, LUX, and WVL show limited spillover, with conditional proportions concentrated on themselves, indicating more localised risk patterns. In the case of WVL, this localisation can be directly attributed to the Yser basin.

Province	ANT	BRA	BRU	HAI	LIE	LIM	LUX	NAM	OVL	VLB	WVL
ANT	58.84	2.98	1.71	2.16	1.83	6.23	1.21	1.23	9.32	13.63	0.85
BRA	3.63	46.39	3.84	13.74	4.55	4.39	1.59	2.71	2.64	15.36	1.15
BRU	3.64	7.89	42.10	8.17	5.47	3.47	2.97	4.12	6.17	14.08	1.91
HAI	1.58	4.80	1.58	53.76	3.75	2.75	1.99	9.69	11.75	3.10	5.24
LIE	0.69	1.11	0.63	2.84	63.49	8.60	4.13	15.68	1.00	0.94	0.88
LIM	5.79	1.90	1.19	4.99	15.54	56.62	2.48	2.39	1.66	5.36	2.08
LUX	1.63	1.29	0.81	4.49	17.29	2.89	54.64	11.79	2.10	1.00	2.06
NAM	0.41	1.06	0.65	10.78	22.95	2.53	3.85	56.00	0.91	0.47	0.40
OVL	6.36	1.42	0.97	10.43	1.39	1.19	0.83	1.15	66.34	2.62	7.31
VLB	19.03	11.94	6.16	5.14	2.85	8.85	0.82	1.17	2.48	41.12	0.44
WVL	0.93	1.36	0.89	10.46	2.98	2.43	1.92	1.84	11.92	0.79	64.49

Table 3.3: Conditional loss proportions (%) given that the province loss share exceeds 25%. Diagonal entries correspond to the conditioning province itself.

3.1.2 Compositional structure and transformation

The geometry visible in Figure 3.1 — in particular the sharp angles in the raw space — reflects the simplex constraint inherent to compositional data: since all province proportions must sum to unity for each event, the data live on the $(n - 1)$ -dimensional simplex in $\mathbb{R}^n$. This constant-sum constraint induces a spurious negative correlation among components that is purely an artefact of the constraint, not a causal association [1].

Standard remedies within logratio analysis, such as the centred log-ratio (CLR) and isometric log-ratio (ILR) transformations proposed by Aitchison and Egozcue [2] (properly defined in [19]), aim to remove this constraint. In the present setting, however, applying these transformations does not yield satisfactory results: the CLR and ILR representations reduce interpretability in hydrological terms, producing PCA patterns that appear shuffled relative to the clearer structure in Figure 3.1 (see Figures A.1 and A.2 in Annex). This loss of interpretability propagates to the clustering partitions making inverse-transformed groups difficult to assign a meaningful hydrological reading. Therefore we choose to retain the standard-scaled simplex space for fitting the clustering algorithms, while remaining aware of the potential effects of spurious correlation.

3.2 Event Loss Table

Each row in the ELT corresponds to a unique flood event identified by an **EVENTID** consistent with that used in the damage distribution dataset. The **Rate** column expresses the annual occurrence rate λ_i of each event, while **Mean** provides the expected loss μ_i in euros. The variability of the loss estimate is further characterised by two standard deviation measures: **StdDevI**, which captures the idiosyncratic uncertainty component,

and **StdDevC**, which reflects the correlated uncertainty shared across events. Finally, **Exposure** represents the affected sums insured for the given event.

Each entry i in the ELT can be understood as a *compound distribution*, resulting from combining the frequency distribution of the number of occurrences N_i and the severity distribution of individual losses X_{ij} [25]. Following the collective risk model (Klugman et al. [28, Definition 9.1]) assumptions with $N_i \sim \text{Poisson}(\lambda_i)$ we have that the annual aggregate loss for event i is:

$$L_i = \sum_{j=1}^{N_i} X_{ij} \quad (3.1)$$

where the X_{ij} are i.i.d. random variables representing individual loss amounts. Under this compound Poisson framework, the expected annual loss (AAL) and variance of L_i are ([25, p. 31]):

$$\text{AAL}_i = \mathbb{E}[L_i] = \lambda_i \mu_i \quad (3.2)$$

$$\text{Var}(L_i) = \lambda_i \cdot \mathbb{E}[X_{ij}^2] = \lambda_i \left[(\sigma_i^I + \sigma_i^C)^2 + \mu_i^2 \right] \quad (3.3)$$

where $\sigma_i^I = \text{StdDevI}_i$ and $\sigma_i^C = \text{StdDevC}_i$ are additive in standard deviation space, so that $\text{Var}(X_{ij}) = (\sigma_i^I + \sigma_i^C)^2$ ([25, eq. 1.3]).

Since the per-occurrence loss ratio $Y_{ij} = X_{ij}/E_i$ is bounded within $[0, 1]$, it is naturally modelled using a Beta distribution [25]. Dividing by the deterministic constant E_i yields two scale-free quantities serving as direct inputs to the Beta parameterisation:

$$\text{MDR}_i = \mathbb{E}[Y_{ij}] = \frac{\mu_i}{E_i}, \quad \text{SDR}_i = \sqrt{\text{Var}[Y_{ij}]} = \frac{\sigma_i^I + \sigma_i^C}{E_i} \quad (3.4)$$

(see Section 4.3). For the annual aggregate damage ratio $T_i = L_i/E_i$, the same normalisation gives:

$$\mathbb{E}[T_i] = \lambda_i \cdot \text{MDR}_i, \quad \text{Var}(T_i) = \lambda_i \left[\text{SDR}_i^2 + \text{MDR}_i^2 \right] \quad (3.5)$$

3.2.1 Descriptive statistics

We focus on four quantities to characterise the risk profile of each event: the occurrence rate λ_i , the mean damage ratio MDR_i , the standard deviation of the damage ratio SDR_i , and the annual risk intensity $\mathbb{E}[T_i]$. Together, these capture frequency, expected severity, uncertainty, and overall annual risk contribution normalised by portfolio size. Table 3.4 reports their descriptive statistics; Figure 3.2 displays their spatial structure; and Table 3.5 presents the mean of each quantity partitioned by dominant province, as defined by Figure 3.1.

The distribution of λ_i is the most skewed of the four, with a median of $1.67 \times 10^{-4} \text{ yr}^{-1}$ and a mean of $2.72 \times 10^{-3} \text{ yr}^{-1}$ more than sixteen times larger. The portfolio is therefore dominated by rare events, with a small number of high-frequency events pulling the mean upward; the maximum of 1.00 yr^{-1} indicates that at least one event occurs annually. Figure A.4 reveals a discrete structure in the rate dimension where events groups at a small number of rate levels with wide variation in MDR within each group but but in

Statistic	λ_i	MDR _i	SDR _i	$\mathbb{E}[T_i]$
Min	1.05×10^{-5}	2.85×10^{-7}	8.67×10^{-7}	7.54×10^{-11}
p05	1.05×10^{-5}	1.99×10^{-5}	1.59×10^{-5}	4.32×10^{-9}
p25	1.11×10^{-5}	1.20×10^{-4}	4.89×10^{-5}	1.63×10^{-8}
Median	1.67×10^{-4}	5.37×10^{-4}	1.08×10^{-4}	5.05×10^{-8}
p75	4.00×10^{-4}	2.85×10^{-3}	3.31×10^{-4}	1.48×10^{-7}
p95	1.01×10^{-2}	1.88×10^{-2}	1.40×10^{-3}	5.81×10^{-7}
Mean	2.72×10^{-3}	3.36×10^{-3}	3.13×10^{-4}	1.55×10^{-7}
Max	1.00×10^0	6.40×10^{-2}	4.33×10^{-3}	7.60×10^{-6}

Table 3.4: Descriptive statistics of the four exposure-normalised risk quantities across all ELT events.

general, we observe an inverse relationship between both quantities. In the top-left panel of Figure 3.2, high rates concentrate in the upper-left region of the PCA space, which Figure 3.1 identifies as predominantly Hainaut. This is confirmed by Table 3.5, where Hainaut records the highest total occurrence rate ($\sum \lambda_i = 5.039 \text{ yr}^{-1}$) by a wide margin, driven by the large number of high-frequency events concentrated in that province. The remaining provinces show substantially lower aggregate rates.

Turning to MDR_i, its median is 5.37×10^{-4} , with a mean of 3.36×10^{-3} six times larger and a 5th-to-95th percentile range spanning three orders of magnitude. In the top-right panel of Figure 3.2, high MDR values concentrate in the left portion of the PCA space, corresponding in Figure 3.1 to Liège, Namur, and Oost-Vlaanderen. Table 3.5 confirms this: Liège has the highest mean MDR_i (9.07×10^{-3}), followed by Namur (6.37×10^{-3}) and Oost-Vlaanderen (3.31×10^{-3}), while Brussels records the lowest by a wide margin (9.33×10^{-5}).

Regarding SDR_i, Figure A.5 (Annex) confirms a strong linear relationship between SDR and MDR across events. The colour scale in that figure shows that this ratio is not constant: it decreases as MDR increases, meaning that events with higher expected damage have proportionally lower relative uncertainty, visible as a colour gradient along the linear trend. The same spatial pattern observed for MDR in Figure 3.2 is reproduced for SDR in the bottom-left panel, as expected from this proportionality.

Finally, the annual risk intensity $\mathbb{E}[T_i]$ has a median of $5.05 \times 10^{-8} \text{ yr}^{-1}$, a mean of $1.55 \times 10^{-7} \text{ yr}^{-1}$ three times larger, and a maximum of $7.60 \times 10^{-6} \text{ yr}^{-1}$ fifty times the mean. In the bottom-right panel of Figure 3.2, the highest values concentrate in the left and lower-left region of the PCA space, corresponding to the Walloon provinces in Figure 3.1. Table 3.5 shows that Luxembourg leads (3.12×10^{-7}), followed by Namur (2.59×10^{-7}) and Liège (2.20×10^{-7}). Hainaut, despite having the highest mean rate, ranks only fifth (1.66×10^{-7}) because its low MDR offsets its high frequency. The Flemish provinces rank consistently at the bottom, and Brussels is the lowest of all (8.80×10^{-8}).

Province	$\sum \lambda_i$	$\overline{\text{MDR}}_i$	$\overline{\text{SDR}}_i$	$\overline{\mathbb{E}[T_i]}$
HAI	5.039	4.93×10^{-4}	1.07×10^{-4}	1.66×10^{-7}
LIM	0.782	5.68×10^{-4}	1.03×10^{-4}	2.13×10^{-7}
LIE	0.723	9.07×10^{-3}	7.13×10^{-4}	2.20×10^{-7}
NAM	0.581	6.37×10^{-3}	5.74×10^{-4}	2.59×10^{-7}
LUX	0.502	6.73×10^{-4}	1.24×10^{-4}	3.12×10^{-7}
BRU	0.273	9.33×10^{-5}	4.97×10^{-5}	8.80×10^{-8}
WVL	0.234	4.85×10^{-4}	8.20×10^{-5}	7.34×10^{-8}
OVL	0.215	3.31×10^{-3}	3.15×10^{-4}	8.19×10^{-8}
ANT	0.172	8.93×10^{-4}	1.20×10^{-4}	3.74×10^{-8}
BRA	0.158	3.87×10^{-4}	9.44×10^{-5}	1.05×10^{-7}
VLB	0.064	4.31×10^{-4}	8.99×10^{-5}	4.80×10^{-8}

Table 3.5: Risk quantities by dominant province (events grouped by arg max of province proportions), sorted by $\sum \lambda_i$ in descending order.

3.3 Summary and motivation for clustering

In the ELT dataset, all four risk quantities (λ_i , MDR_i , SDR_i , $\mathbb{E}[T_i]$) exhibit strongly right-skewed distributions spanning multiple orders of magnitude as shown in Table 3.5. Frequency and severity are largely independent at the event level — MDR varies over several orders of magnitude within any given rate cluster (see Figure A.4 in Annex) — and their geographic signatures are dissociated: high-frequency events concentrate are mainly concentrated in Hainaut province, while high-severity events concentrate in Liège, Namur and Oost-Vlaanderen. The annual risk intensity $\mathbb{E}[T_i]$, which combines both dimensions, is dominated by Luxembourg, Namur, Liège and Limburg provinces.

In the DDT dataset, the province proportions are highly sparse and right-skewed, as confirmed by the descriptive statistics in Table 3.1, where medians are close to zero for every province while upper quantiles exhibit substantial variation. The high zero proportions reported in Table 3.2 further support this concentration, as most events attribute nearly all losses to a single province with the remainder distributed among spatially proximate provinces within the same hydrological basin. This pattern is confirmed by the conditional co-occurrence structure of Table 3.3, where provinces sharing the same basin consistently appear together when a dominant province is significantly affected, while provinces from distinct basins show limited co-occurrence.

These observations collectively establish two things. First, the Belgian flood portfolio is genuinely heterogeneous: events differ substantially in frequency, severity, geographic footprint, and annual risk contribution, and this heterogeneity cannot be reduced to a simple province label. Second, this heterogeneity is spatially structured, as the PCA scatter in Figures 3.1 & 3.2 shows that events with similar damage distributions occupy coherent regions of the compositional space whose geometry already mirrors the hydrological

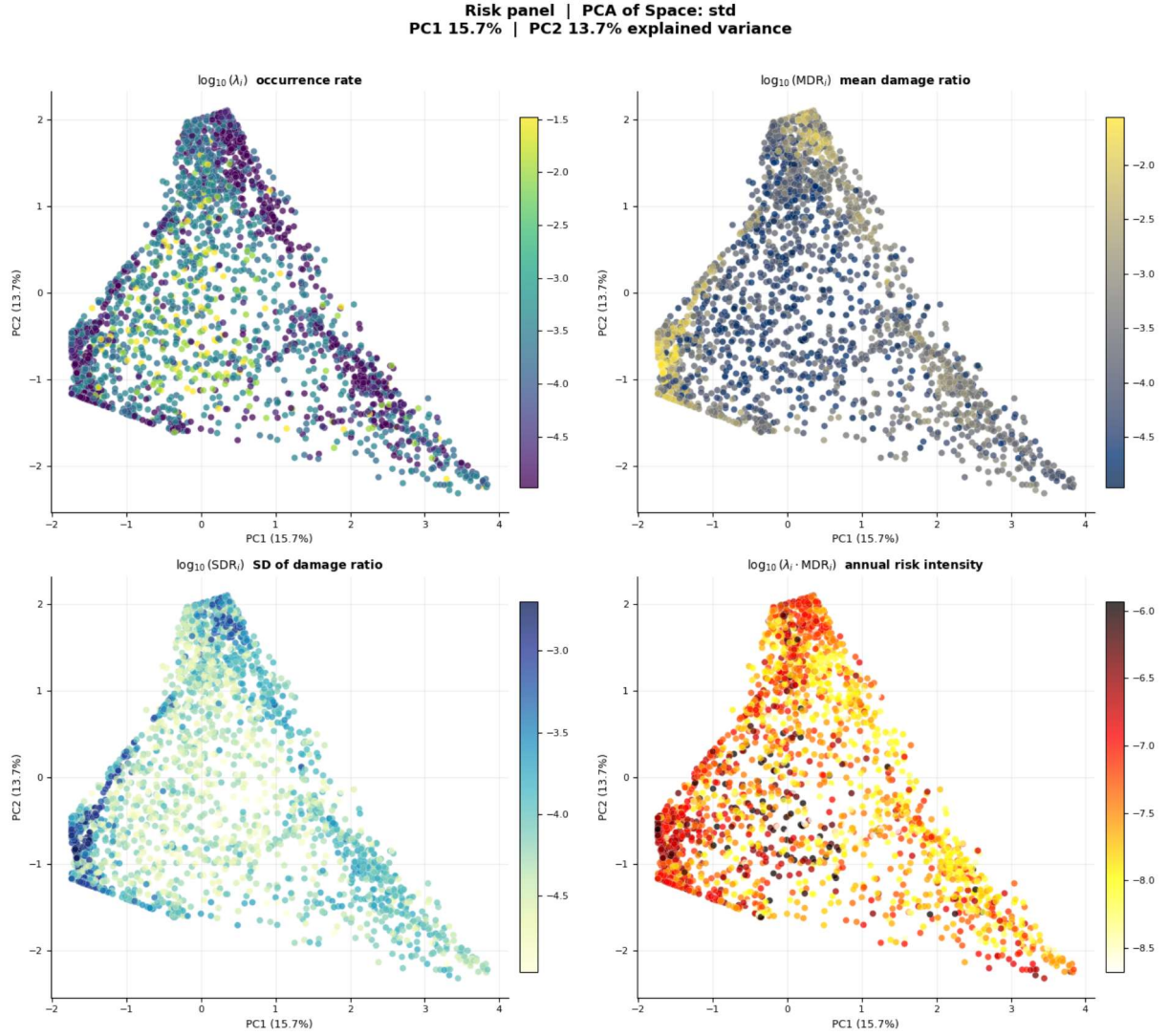


Figure 3.2: PCA scatter coloured by $\log_{10}(\lambda_i)$, $\log_{10}(\text{MDR}_i)$, $\log_{10}(\text{SDR}_i)$, and $\log_{10}(\mathbb{E}[T_i])$ (left to right, top to bottom).

basin structure. The boundaries between province-dominated groups are however diffuse and overlapping, making a single province assignment an insufficient summary of the compositional structure.

These empirical findings confirm the hydrological argument developed in Chapter 2: the Scheldt and Meuse basins operate through distinct physical processes, can flood independently, and expose different provinces to qualitatively different loss profiles. Geography and data converge on the same conclusion that risk zones defined at the national level are too coarse to capture the structure of Belgian flood risk, and together motivate the application of clustering techniques directly to the DDT. Rather than grouping events by their dominant province, clustering seeks to identify groups that share a coherent joint damage profile across all provinces simultaneously, with reasonable expectation that such groups will align with meaningful hydrological units. The extent to which this is the case, and how robustly it can be detected under different algorithms and distance metrics, is the central question of the analysis.

Chapter 4

Exceedance Probability

In the context of natural catastrophe risk management, one of the central challenges facing insurers and reinsurers is to determine the size and distribution of their portfolios potential losses Grossi and Kunreuther [20]. Exceedance probability (EP) curves provide a compact and tractable representation of the full loss distribution arising from a catastrophe model. Given an event loss table (ELT) produced by a catastrophe model one can derive the probability that losses exceed any given threshold l over a predefined future time period [22].

Two principal EP curves are produced by catastrophe models [25]:

- The **Occurrence Exceedance Probability (OEP)** curve, which gives the probability that the *largest occurrence loss* in a year exceeds a given threshold;
- The **Aggregate Exceedance Probability (AEP)** curve, which gives the probability that the *sum of all event losses* in a year exceeds a given threshold.

The OEP curve is particularly relevant for the pricing and structuring of per-occurrence reinsurance treaties (such as catastrophe excess-of-loss covers), while the AEP curve is more directly connected to questions of annual aggregate capital adequacy [25, p. 26]. In the remainder of this chapter, we focus primarily on the **OEP curve**, providing a rigorous probabilistic derivation and discussing its properties and practical computation from an event loss table. Beyond its role as a standard risk metric, the OEP curve will serve a dual purpose in this work. First, given a set of catastrophe events grouped into a cluster, the OEP curve summarises the tail risk profile of that cluster in a single, interpretable function. Second, and perhaps more importantly, the OEP curve acts as a validation tool for the clustering itself: if two clusters exhibit markedly different OEP curves, this is evidence that they capture genuinely distinct risk zones, characterised by different loss severities and exceedance probabilities. Conversely, similar OEP curves across clusters would suggest insufficient risk differentiation. This dual role — as both a risk measure and a diagnostic — motivates the central place of the OEP curve throughout the remainder of this thesis.

4.1 The Collective Risk Model Framework

In Section 3.2 we define the annual aggregate loss L_i for event i . In the context of n events their total aggregate loss becomes:

$$S := \sum_{i=1}^n L_i = \sum_{i=1}^n \sum_{j=1}^{N_i} X_{ij} \quad (4.1)$$

This construction model is called the individual risk model ([28, Definition 9.2]) not to be confused with the collective risk model ([28, Definition 9.1]) over which L_i is defined. However, the following result gives us the link between an individual risk model towards a collective risk model ([33, Theorem 3]):

Theorem 4.1 (Superposition of Compound Poisson Processes). *The aggregate loss S has the same distribution as the aggregate loss $\tilde{S} = \sum_{k=1}^N C_k$ from a single Collective Risk Model, where $N \sim \text{Poisson}(\Lambda)$ with*

$$\Lambda := \sum_{i=1}^n \lambda_i,$$

and the C_k are i.i.d. (also independent of N) with mixture distribution

$$F_c := \sum_{i=1}^n \frac{\lambda_i}{\Lambda} F_i.$$

where F_i are the cumulative distribution functions of severities $X_{i\bullet}$ from event i .

The practical value of the superposition property of compound Poisson processes extends well beyond theoretical elegance. As noted by Klugman et al. [28, p. 222], when a portfolio consists of several lines of business — or, in the catastrophe modelling context, several independent peril scenarios — each governed by its own compound Poisson model, one need not compute the aggregate loss distribution separately for each component and then derive the distribution of their sum. Instead, the Poisson parameters λ_i may simply be summed into a single rate $\Lambda = \sum_{i=1}^n \lambda_i$, and the individual severity distributions F_i replaced by their weighted mixture $\bar{F} = \sum_{i=1}^n \frac{\lambda_i}{\Lambda} F_i$. A single aggregate loss distribution calculation on this superposed model then suffices. In the ELT framework, this is precisely the operation performed when consolidating n scenarios into a single compound Poisson model.

4.2 Occurrence Exceedance Probability (OEP)

We begin by deriving the distribution of the maximum occurrence loss within the collective risk model framework. Subsequently, Pfeifer [33] shows that this result is consistent with the individual risk model perspective: provided that the individual and collective models are linked through the property stated in Theorem 4.1, the maximum occurrence loss under both frameworks shares the same distribution. In other words, the choice of model — individual or collective — is inconsequential for the computation of the OEP curve, as long as the linking conditions of Theorem 4.1 are satisfied.

4.2.1 Derivation

Let the individual losses in a year from all events be $C_1, C_2, \dots, C_N$ as defined in Section 4.1. Define the **Maximum occurrence Loss** as

$$M_{coll} := \max \{C_i \mid 1 \leq i \leq N\}.$$

The **Occurrence Exceedance Probability** is then defined as

$$\text{OEP}(l) := \mathbb{P}(M_{coll} > l) = 1 - \mathbb{P}(M_{coll} \leq l).$$

To derive $\mathbb{P}(M_{coll} \leq l)$ [22], note that the maximum loss does not exceed l if and only if *every* individual claim in the year does not exceed l . For a fixed number $N = n$ of i.i.d. claims,

$$\mathbb{P}(M \leq l \mid N = n) = \mathbb{P}(C_1 \leq l, \dots, C_n \leq l) = [F_c(l)]^n$$

Therefore, using the law of total probability,

$$\begin{aligned} \text{OEP}(l) &= 1 - \mathbb{P}(M_{coll} \leq l) \\ &= 1 - \sum_{n=0}^{\infty} \mathbb{P}(M_{coll} \leq l \mid N = n) \mathbb{P}(N = n) \\ &= 1 - \sum_{n=0}^{\infty} F_c(l)^n \cdot \mathbb{P}(N = n) \\ &= 1 - \mathbb{E}[F_c(l)^N]. \end{aligned} \tag{4.2}$$

This expression is entirely general. In the case $N \sim \text{Poisson}(\Lambda)$, we recognise the probability generating function of the Poisson distribution evaluated at $F_c(l)$ [33]:

$$\mathbb{E}[F_c(l)^N] = \varphi_N(F_c(l)) = e^{\Lambda(F_c(l)-1)},$$

so that

$$\mathbb{P}(M_{coll} > l) = 1 - \exp\{-\Lambda(1 - F_c(l))\}. \tag{4.3}$$

If we now define $M_{ind} := \max \{X_{ij} \mid 1 \leq j \leq N_i, 1 \leq i \leq n\}$ for X_{ij} defined in Section 3.2 we have (Pfeifer [33, Lemma 4]):

$$\mathbb{P}(M_{ind} > l) = 1 - \exp\left\{-\sum_{i=1}^n \lambda_i [1 - F_i(l)]\right\} = 1 - \exp\{-\Lambda[1 - F_c(l)]\} = \text{OEP}(l) \tag{4.4}$$

therefore $M_{ind} \sim M_{coll}$. This gives us a suitable expression where we can directly use the ELT table.

4.2.2 Exceedance Frequency and Return Period

It is convenient to introduce the **aggregate exceedance frequency**

$$\nu(l) := \sum_{i=1}^n \lambda_i [1 - F_i(l)] = \Lambda(1 - F_c), \quad (4.5)$$

so that equation (4.4) takes the compact form $\text{OEP}(l) = 1 - e^{-\nu(l)}$. The quantity $\nu(l)$ is the expected number of events per year whose loss exceeds l [25]. Since each event follows an independent Poisson process, the count of annual exceedances of l is Poisson-distributed with mean $\nu(l)$, and the OEP is the probability that this count is at least one:

$$\text{OEP}(l) = \mathbb{P}(N(l) \geq 1) = 1 - e^{-\nu(l)}. \quad (4.6)$$

This equivalence — $\mathbb{P}(\max \text{ loss} > l) \equiv \mathbb{P}(\text{at least one loss} > l)$ — is exact and follows directly from the definition of the maximum. Both $\nu(l)$ and $\text{OEP}(l)$ carry the same information, since one is a monotone transform of the other. However, $\nu(l)$ is a frequency measured in events per year and can exceed 1, whereas $\text{OEP}(l) \in [0, 1]$ is a proper annual probability. This makes OEP the natural quantity for reinsurance, where tail risk is expressed in terms of probabilities and, more importantly, their reciprocals: the **return periods**.

The return period T associated with a loss level l is the average number of years one expects to wait between successive exceedances of l . It is important to stress that this is a statement about a long-run average: a loss exceeding l with return period T does not mean that exactly one such exceedance will occur every T years. Rather, as noted in [41, p. 483], the probability of the threshold being exceeded is $1/T$ in *each and every* year, independently of past events. Formally, the return period is defined as the reciprocal of the annual exceedance probability (Humphreys [22]):

$$T(l) := \frac{1}{\text{OEP}(l)} = \frac{1}{\mathbb{P}(M > l)}. \quad (4.7)$$

Equivalently, $l(T) := \text{OEP}^{-1}(1/R)$ denotes the loss level that is exceeded on average once every T years, i.e. the quantile of $N(l)$ at level $1 - 1/R$. From equation (4.3), substituting the definition of R and taking logarithms on both sides of $1 - \text{OEP}(l) = \exp\{-\Lambda[1 - F_c(l)]\}$, we obtain the explicit relationship between return periods and threshold losses:

$$\ln\left(\frac{T}{T-1}\right) = \sum_{i=1}^n \lambda_i (1 - F_i(l)). \quad (4.8)$$

In practice, equation (4.8) is not inverted analytically. Instead, given a pre-specified grid of return periods $T_1 < T_2 < \dots < T_m$, the corresponding loss levels $l(T_1), l(T_2), \dots, l(T_m)$ are computed numerically by evaluating the right-hand side of (4.8) directly from the ELT. The OEP curve is then obtained as the set of pairs $\{(T_k, l(T_k))\}_{k=1}^m$.

4.2.3 Exceedance Frequency Decomposition

A key structural property of $\nu(l)$ is that it is additive. If the portfolio events are partitioned into K disjoint subsets $\mathcal{C}_1, \dots, \mathcal{C}_K$, then

$$\nu_*(l) = \sum_{k=1}^K \nu_k(l), \quad \nu_k(l) := \sum_{i \in \mathcal{C}_k} \lambda_i [1 - F_i(l)], \quad (4.9)$$

where $\nu_*(l) := \sum_{i=1}^n \lambda_i [1 - F_i(l)]$ denotes the portfolio-level aggregate exceedance frequency and $\nu_k(l)$ is the contribution of subset k alone. From now on we use the subscript $*$ to refer to portfolio-level quantities, reserving the subscript k for subset-level ones. This additivity follows directly from the superposition of independent Poisson processes and does *not* hold for OEP: since $1 - e^{-\nu_*} \neq \sum_k (1 - e^{-\nu_k})$ in general, the frequency $\nu_*(l)$ is the natural quantity for decomposing portfolio tail risk into contributions from subsets of events.

This motivates the definition of the **exceedance frequency share** of subset k :

$$\omega_k(l) := \frac{\nu_k(l)}{\nu_*(l)}, \quad \sum_{k=1}^K \omega_k(l) = 1 \quad \forall l. \quad (4.10)$$

Each $\omega_k(l) \in [0, 1]$ gives the fraction of the total expected annual exceedances of l attributable to subset k . Since it is normalised by the portfolio frequency $\nu_*(l)$, it measures the *relative* contribution of each subset to the annual exceedance at each loss level, independently of the absolute scale of ν_k . It should be noted that $\omega_k(l)$ measures a frequency contribution rather than a capital contribution: a subset with $\omega_k(l) \approx 1$ at a given loss level dominates the exceedance count but does not directly quantify the monetary share of losses. Nevertheless, a subset that drives the portfolio exceedance frequency at high loss levels is by construction the primary source of tail risk, making $\omega_k(T)$ a practically meaningful complement to the OEP curves for risk zonation purposes.

For reinsurance purposes it is more natural to express ω_k as a function of the return period rather than of the loss level. Since each partition $\mathcal{C}_1, \dots, \mathcal{C}_K$ for a given return period has its own level loss we choose the losses at total portfolio level as reference. Let $l_*^{-1}(T)$ denote the portfolio loss at return period T , defined implicitly by:

$$\text{OEP}_*(l_*^{-1}(T)) = \frac{1}{T}, \quad (4.11)$$

where $\text{OEP}_*(l) = 1 - e^{-\nu_*(l)}$ is the portfolio-level OEP curve. The **return-period exceedance frequency share** is then:

$$\omega_k(T) := \omega_k(l_*^{-1}(T)) = \frac{\nu_k(l_*^{-1}(T))}{\nu_*(l_*^{-1}(T))}. \quad (4.12)$$

The interpretation is as follows. The threshold $l_*^{-1}(T)$ is the loss level at which the *portfolio* experiences on average one exceedance every T years. At that same threshold, $\nu_k(l_*^{-1}(T))$ counts how many of those annual exceedances are generated by subset k alone,

and $\omega_k(T)$ expresses this as a fraction of the total $\nu_*(l_*^{-1}(T))$. Concretely, $\omega_k(T) = 0.9$ means that 90% of all annual exceedances of the portfolio T -year loss level are attributable to subset k .

The practical relevance of OEP curves and $\omega_k(T)$ will become apparent in Chapter 5, where the n events are partitioned into clusters representing distinct flood risk patterns. When such a partition exists, $\omega_k(T)$ provides a principled answer to the question of which cluster drives the portfolio OEP at each return period. In particular, a cluster with low event rate but heavy-tailed severity may contribute negligibly to $\nu_*(l)$ at moderate loss levels, yet have $\omega_k(T) \rightarrow 1$ as T grows, indicating that it entirely dominates the extreme tail. The behaviour of a portfolio that appears diversified at short return periods but is concentrated in a single cluster at long return periods — cannot be detected from the cluster OEP curves alone, since they are not additive, nor from raw ν_k values alone, since they are not normalised.

4.3 Beta distribution

It is worth noting that in Section 4.1, no distributional assumption was placed on the individual loss severities X_{ij} . The derivation of the OEP curve in Section 4.2 was therefore fully general, holding for any severity distribution with finite moments. We now specialise to a parametric family in order to make the exceedance probability tractable for numerical computation.

As discussed in Section 3.2, the losses X_{ij} are naturally bounded above by the total insured value of the portfolio. This constraint motivates a bounded support distribution, and the Beta distribution $\text{Beta}(\alpha, \beta)$ emerges as a natural and flexible choice for modelling the proportion of total insured value lost in a given event [25].

A random variable X follows a Beta distribution with parameters $\alpha, \beta > 0$, written $X \sim \text{Beta}(\alpha, \beta)$, if its probability density function is defined on $[0, 1]$ and given by [44]:

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)} = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad x \in [0, 1], \quad (4.13)$$

where $B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha + \beta)$ is the beta function and $\Gamma(\cdot)$ is the gamma function. The corresponding cumulative distribution function is $F(x; \alpha, \beta) = I(x; \alpha, \beta)$, where $I(\cdot; \alpha, \beta)$ denotes the regularised incomplete beta function, defined as [45]:

$$I(z; \alpha, \beta) := \frac{B(z; \alpha, \beta)}{B(\alpha, \beta)}, \quad z \in [0, 1], \quad (4.14)$$

and $B(z; \alpha, \beta) = \int_0^z t^{\alpha-1}(1-t)^{\beta-1} dt$ is the incomplete beta function. Note that $I(0; \alpha, \beta) = 0$ and $I(1; \alpha, \beta) = 1$, confirming that F is a proper CDF. The mean and variance are:

$$\mathbb{E}[X] = \frac{\alpha}{\alpha + \beta}, \quad \text{Var}(X) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}. \quad (4.15)$$

Given a loss X_{ij} from event i , we model it as

$$X_{ij} = Y_{ij} \cdot E_i, \quad (4.16)$$

where E_i is the total insured exposure value of event i and $Y_{ij} \sim \text{Beta}(\alpha_i, \beta_i)$ is the *per-occurrence loss ratio*. From Section 3.2 yields two scale-free quantities:

$$\text{MDR}_i := \mathbb{E}[Y_{ij}] = \frac{\mu_i}{E_i}, \quad \text{SDR}_i := \sqrt{\text{Var}(Y_{ij})} = \frac{\sqrt{\sigma_i^2 + \sigma_C^2}}{E_i}, \quad (4.17)$$

Matching these to the mean and variance of the $\text{Beta}(\alpha_i, \beta_i)$ distribution, one can solve explicitly for the Beta parameters via moment matching:

$$\alpha_i = \text{MDR}_i \left(\frac{\text{MDR}_i(1 - \text{MDR}_i)}{\text{SDR}_i^2} - 1 \right), \quad \beta_i = (1 - \text{MDR}_i) \left(\frac{\text{MDR}_i(1 - \text{MDR}_i)}{\text{SDR}_i^2} - 1 \right). \quad (4.18)$$

This moment-matching approach requires only the mean and variance of the individual losses — both of which are directly available from the ELT — making the Beta parameterisation straightforward to implement in practice.

With the Beta parameterisation of Y_{ij} established, the mixture distribution F_c of the superposed collective model introduced in Theorem 4.1 can now be made explicit. Recalling that $X_{ij} = Y_{ij} \cdot E_i$ and that F_i denotes the CDF of X_{ij} , we have:

$$F_i(l) = \mathbb{P}(X_{ij} \leq l) = \mathbb{P}\left(Y_{ij} \leq \frac{l}{E_i}\right) = I\left(\frac{l}{E_i}; \alpha_i, \beta_i\right), \quad (4.19)$$

where $I(\cdot; \alpha_i, \beta_i)$ is the regularised incomplete beta function. Substituting into equation (4.8), the OEP curve is then given explicitly in terms of the Beta parameters as:

$$\ln\left(\frac{T}{T-1}\right) = \sum_{i=1}^n \lambda_i \left(1 - I\left(\frac{l}{E_i}; \alpha_i, \beta_i\right) \right), \quad (4.20)$$

where α_i and β_i are determined from the ELT via the moment-matching equations above. Equation (4.20) is the fully explicit, computationally tractable form of the OEP curve under the Beta severity assumption, and it is this expression that is evaluated numerically in practice over a pre-specified grid of return periods.

Chapter 5

Clustering

This chapter presents the clustering methodology applied to the Damage Distribution Table in order to identify spatially coherent flood risk zones within the Belgian portfolio. Five algorithms are implemented and compared: K-means and K-Medoids (KM), Gaussian Mixture Models (GMM), Hierarchical Energy Clustering (HEC), and K-Groups Clustering (KGR). For each algorithm, we describe its theoretical foundations, its assumptions about cluster and the parametrisation choices relevant to the present dataset. We then address the question of the optimal number of clusters, discussing the internal validation metrics used to guide that decision. Finally, we present the results obtained under the selected configurations and evaluate how the resulting partitions respond to the risk-zonation objectives motivated in Chapters 2 and 3.

5.1 Representative-based Clustering

This section presents three classical clustering algorithms that are widely used in practice and serve as the baseline methods in this work. Throughout, $\mathcal{C} = \{C_1, \dots, C_K\}$ denotes a partition of n observations into K non-empty disjoint groups, with C_k the set of observations assigned to cluster k and $n_k = |C_k|$.

5.1.1 K-Means

K-Means partitions n observations into a pre-specified number of groups K by minimising the **within-cluster sum of squares** [21]:

$$\phi(C_1, \dots, C_K) = \sum_{k=1}^K \sum_{i \in C_k} \|\mathbf{x}_i - \mathbf{c}_k\|^2, \quad \mathbf{c}_k = \frac{1}{n_k} \sum_{i \in C_k} \mathbf{x}_i, \quad (5.1)$$

where $\mathbf{c}_k$ is the centroid of cluster C_k . Each cluster is therefore represented by its mean, and the objective rewards compact, spherical clusters in the feature space given its Euclidean nature.

5.1.2 K-Medoids

K-Medoids generalises K-Means to arbitrary dissimilarity measures by replacing the centroid with a *medoid*, the cluster member minimising the total distance to all other members of the cluster [27]:

$$\psi(C_1, \dots, C_K) = \sum_{k=1}^K \sum_{i \in C_k} d(\mathbf{x}_i, m_k), \quad m_k = \arg \min_{j \in C_k} \sum_{i \in C_k} d(\mathbf{x}_i, \mathbf{x}_j). \quad (5.2)$$

Since m_k is always an actual observation and the formulation requires no algebraic structure beyond the ability to evaluate $d(\cdot, \cdot)$, the algorithm is valid for any dissimilarity measure. This flexibility is the primary motivation for its use here: the compositional damage vectors may exhibit non-Euclidean geometry, and several distance measures that better respect the simplex structure are available. The algorithm used is PAM [38], and the following pairwise distances are considered: Manhattan (L_1), Cosine, Braycurtis, Canberra, Chebyshev, and Correlation, all available in `scikit-learn` [5].

5.1.3 Gaussian Mixture Models

Gaussian Mixture Models (GMM) provide a probabilistic generalisation of K-Means in which each cluster is represented by a multivariate Gaussian distribution rather than a single centroid. The data are assumed to arise from a mixture of a pre-specified number K of Gaussian components, with density:

$$p(\mathbf{x}) = \sum_{k=1}^K w_k \mathcal{N}(\mathbf{x}; \boldsymbol{\theta}_k, \boldsymbol{\Sigma}_k), \quad (5.3)$$

where $w_k > 0$ are the mixing weights with $\sum_k w_k = 1$, $\boldsymbol{\theta}_k$ is the mean vector and $\boldsymbol{\Sigma}_k$ is the covariance matrix of component k [32]. Unlike K-Means, GMM produces a soft assignment: each observation receives a posterior probability of belonging to each component, and a hard partition $\mathcal{C}$ is recovered by assigning each point to its most probable one. Parameters are estimated by maximum likelihood via the EM algorithm [12].

The key modelling choice is the covariance structure imposed on $\boldsymbol{\Sigma}_k$, which controls the shape and orientation of each cluster [7]. Four structures are considered:

- *Full*: each cluster has its own unrestricted covariance matrix, capturing correlations of any shape and orientation;
- *Tied*: all clusters share a single covariance matrix, allowing different locations but identical shapes;
- *Diagonal*: off-diagonal entries are set to zero, restricting each cluster to an axis-aligned ellipsoid with no inter-feature correlations;
- *Spherical*: a single scalar variance per cluster, producing hyperspherical clusters and recovering a soft probabilistic analogue of K-Means.

All three algorithms are sensitive to initialisation and converge only to a local optimum. To mitigate this, each run is initialised using a K-means++ scheme [3], and the full algorithm is repeated over multiple independent initialisations, retaining the solution with the lowest objective value. All implementations rely on the `scikit-learn` library [5].

5.2 Energy Distance Clustering

Energy distance is a metric that measures the distance between the distributions of random vectors, introduced by Rizzo and Székely [35]. As a statistic, it can be applied to measure the difference between a sample and a hypothesised distribution or the difference between two or more samples in arbitrary, not necessarily equal dimensions. The two energy-distance-based algorithms used in this work, HEC and K-Groups, share this common theoretical foundation.

Let X and Y be independent random vectors in $\mathbb{R}^d$ with distributions F and G respectively, and let X', Y' be independent copies of X and Y . For $\alpha \in (0, 2)$, the α -energy distance between F and G is defined as [35]

$$\mathcal{E}^{(\alpha)}(F, G) = 2 \mathbb{E}\|X - Y\|^\alpha - \mathbb{E}\|X - X'\|^\alpha - \mathbb{E}\|Y - Y'\|^\alpha \geq 0, \quad (5.4)$$

provided $\mathbb{E}\|X\|^\alpha < \infty$ and $\mathbb{E}\|Y\|^\alpha < \infty$. Rizzo and Székely [34] establish the following properties:

- (i) $\mathcal{E}^{(\alpha)}(F, G) \geq 0$ always;
- (ii) for $0 < \alpha < 2$, $\mathcal{E}^{(\alpha)}(F, G) = 0$ if and only if $F = G$;
- (iii) for $\alpha = 2$, $\mathcal{E}^{(\alpha)}(F, G) = 0$ if and only if $\mathbb{E}[X] = \mathbb{E}[Y]$.

Statements (i) and (ii), combined with the fact that the energy distance also satisfies the triangle inequality and symmetry axioms, establish that it is a proper metric on the space of distributions with finite α -moments, holding in arbitrary dimension and requiring no assumptions on distributional shape or covariance structure. Statement (iii) shows that the restriction $\alpha \in (0, 2)$ is essential for full distributional sensitivity: at $\alpha = 2$ the criterion collapses to a comparison of means only, losing the ability to detect differences in higher-order distributional properties. The standard energy distance corresponds to $\alpha = 1$.

Given independent samples $\mathbf{X} = \{x_1, \dots, x_{n_1}\}$ from F and $\mathbf{Y} = \{y_1, \dots, y_{n_2}\}$ from G , the empirical two-sample energy statistic is

$$\mathcal{E}_{n_1, n_2}^{(\alpha)}(\mathbf{X}, \mathbf{Y}) = \frac{2}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \|x_i - y_j\|^\alpha - \frac{1}{n_1^2} \sum_{i,j=1}^{n_1} \|x_i - x_j\|^\alpha - \frac{1}{n_2^2} \sum_{i,j=1}^{n_2} \|y_i - y_j\|^\alpha, \quad (5.5)$$

which is a consistent estimator of $\mathcal{E}^{(\alpha)}(F, G)$ [35]. The statistic is positive whenever the two samples are more dissimilar from each other than they are internally dispersed.

The energy distance extends naturally to the multi-sample setting via the Distance Components (DISCO) framework [34], a nonparametric analogue of the ANOVA decomposition of variance [17]. For K samples $\mathbf{X}_1, \dots, \mathbf{X}_K$ of sizes $n_1, \dots, n_K$ with $N = \sum_{k=1}^K n_k$, the total dispersion of the pooled sample decomposes as

$$T_\alpha = S_\alpha + W_\alpha, \quad (5.6)$$

where the within-sample dispersion is

$$W_\alpha = \sum_{k=1}^K \frac{n_k}{2} g_\alpha(\mathbf{X}_k, \mathbf{X}_k), \quad g_\alpha(A, B) := \frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} \|a - b\|^\alpha, \quad (5.7)$$

and the between-sample energy statistic is

$$S_\alpha = \sum_{1 \leq j < k \leq K} \frac{n_j n_k}{2N} \left(2 g_\alpha(\mathbf{X}_j, \mathbf{X}_k) - g_\alpha(\mathbf{X}_j, \mathbf{X}_j) - g_\alpha(\mathbf{X}_k, \mathbf{X}_k) \right). \quad (5.8)$$

The properties of S_α extend those of the two-sample case to the multi-sample setting [34]:

- (i) $S_\alpha(A_1, \dots, A_K) \geq 0$ always;
- (ii) for $0 < \alpha < 2$, $S_\alpha(A_1, \dots, A_K) = 0$ if and only if $A_1 = \dots = A_K$;
- (iii) at $\alpha = 2$, $S_2(A_1, \dots, A_K) = 0$ if and only if $A_1, \dots, A_K$ have equal means.

Statement (ii) makes S_α a consistent test statistic for distributional equality across groups, strictly more powerful than classical ANOVA [17] for detecting non-location distributional differences. Statement (iii) confirms that S_2 reduces exactly to the ANOVA between-group sum of squares, recovering the classical result as a special case.

5.2.1 Hierarchical Energy Clustering

The Hierarchical Energy Clustering algorithm (HEC) [40] is an agglomerative method that extends Ward's minimum variance method [43] by replacing the squared Euclidean distance between centroids with the energy distance of Section 5.2. As in Ward's method, HEC starts from n singleton clusters and successively merges the pair with the smallest cluster e -distance

$$e(A, B) = \frac{n_A n_B}{n_A + n_B} \mathcal{E}_{n_A, n_B}^{(\alpha)}(A, B) \geq 0, \quad (5.9)$$

until a single cluster remains, with distances updated at each step via the Lance-Williams recurrence [29]. The full merge sequence is represented as a dendrogram, from which any partition $\mathcal{C} = \{C_1, \dots, C_K\}$ for $K = 2, \dots, K_{\max}$ is obtained by cutting at the corresponding energy level without refitting.

The key property of HEC relative to Ward's method is statistical consistency: for $\alpha \in (0, 2)$, $e(A, B) > 0$ whenever A and B are drawn from different distributions, even if

those distributions share the same mean [40]. Ward’s method reduces to a comparison of centroids and cannot detect distributional differences beyond location, while HEC is sensitive to the full distribution of each cluster. This makes HEC particularly suited to the present problem, where clusters may differ in the shape and tail behaviour of their damage distributions rather than in their mean province profiles alone.

5.2.2 K-Groups

K-Groups [30] is a partitional clustering algorithm that generalises K-Means by replacing the squared distance-to-centroid criterion with an energy-distance criterion. Given a pre-specified number of groups K and a partition $\mathcal{C} = \{C_1, \dots, C_K\}$ of N observations, it minimises the total within-cluster energy dispersion:

$$W_\alpha(C_1, \dots, C_K) = \sum_{k=1}^K W_\alpha(C_k) = \sum_{k=1}^K \frac{1}{2n_k} \sum_{i \in C_k} \sum_{j \in C_k} \|\mathbf{x}_i - \mathbf{x}_j\|^\alpha, \quad (5.10)$$

Setting $\alpha = 2$ recovers exactly the K-Means objective [30]:

$$\phi^2(C_k) = \frac{1}{2n_k} \sum_{i \in C_k} \sum_{j \in C_k} \|\mathbf{x}_i - \mathbf{x}_j\|^2 = \sum_{i \in C_k} \|\mathbf{x}_i - \mathbf{c}_k\|^2, \quad (5.11)$$

so K-Means is a special case of K-Groups. For $\alpha \in (0, 2)$, the criterion is sensitive to differences in the full within-cluster distribution rather than just its spread around the mean, making K-Groups more robust to non-spherical and heavy-tailed cluster geometries. As with K-Means, the algorithm is sensitive to initialisation and converges only to a local optimum. To mitigate this, each run is initialised using a K-means++ scheme [3], and the full algorithm is repeated over multiple independent initialisations, retaining the solution with the lowest W_α .

The motivation for using K-Groups in this context mirrors that of HEC: flood damage distributions within a cluster may differ substantially in shape and tail behaviour, and an objective sensitive to the full distribution is preferable to one anchored to cluster centroids. In this work, both K-Groups and HEC are run across the same set of exponent values $\alpha \in \{0.4, 0.8, 1.0, 1.2, 1.6, 2.0\}$.

5.3 Clustering Evaluation

The evaluation of a clustering result in this work is designed to respond to four complementary purposes. **Hydrological coherence**: whether each cluster corresponds to a geographically coherent flood pattern, grounded in the hydrological structure of Belgium established in Chapter 2. **Geometrical**: whether clusters are compact and well-separated in the Euclidean feature space. **Distributional**: whether clusters differ in the full distribution of damage patterns, including shape, tail behaviour, and covariance structure. **Actuarial**: whether the partition produces clusters with meaningfully different risk profiles from a reinsurance perspective. No single index covers all four dimensions

simultaneously, and the framework combines internal quantitative indices, which address the geometrical and distributional purposes, with external visual diagnostics, which address the geographical and actuarial ones.

5.3.1 Internal Evaluation

Internal validation indices assess the quality of a clustering result using only the data and the partition itself, without any external reference. They measure properties such as compactness within clusters and separation between them, providing a quantitative and formally grounded basis for comparing algorithms and parameter configurations. The indices used in this work are the silhouette score, the Calinski–Harabasz index, and two energy distance-based metrics.

Silhouette Score

The silhouette score [36] measures how similar an observation is to its own cluster relative to the nearest competing cluster. For observation i , let $a(i)$ denote the mean distance to all other points in the same cluster, and $b(i)$ the mean distance to all points in the nearest other cluster. The silhouette coefficient of i is:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \quad (5.12)$$

with $s(i) \in [-1, 1]$. A value close to 1 indicates that the observation is well matched to its own cluster and poorly matched to neighbouring clusters. The overall silhouette score is the mean of $s(i)$ over all observations and is maximised over K to select the optimal number of clusters.

Calinski–Harabasz Index

The Calinski–Harabasz (CH) index [6] is the ratio of between-cluster dispersion to within-cluster dispersion, normalised by the respective degrees of freedom. Let $\mathbf{c}$ denote the grand centroid of all N observations, $\mathbf{c}_k$ the centroid of cluster C_k , and $n_k = |C_k|$. The between-cluster and within-cluster dispersions are respectively:

$$B_K = \sum_{k=1}^K n_k \|\mathbf{c}_k - \mathbf{c}\|^2, \quad W_K = \sum_{k=1}^K \sum_{i \in C_k} \|\mathbf{x}_i - \mathbf{c}_k\|^2. \quad (5.13)$$

The CH index is then:

$$\text{CH}(K) = \frac{B_K / (K - 1)}{W_K / (N - K)}. \quad (5.14)$$

This is directly analogous to the pseudo- F statistic of classical one-way ANOVA [17]: B_K plays the role of the between-group sum of squares, W_K the within-group sum of squares, and the degrees-of-freedom normalisation $(K - 1)$ and $(N - K)$ are identical.

Unlike ANOVA however, the groups are estimated from the data rather than fixed by experimental design, so no valid p -value can be derived and the index is used purely as a relative quality measure. A higher CH index indicates more compact and well-separated clusters.

DISCO F -ratio

The two indices above are geometry-based: they reward compact, spherical clusters and compare distances to centroids, making them sensitive exclusively to differences in cluster means. The Distance Components (DISCO) decomposition [34] addresses this limitation by grounding cluster evaluation in energy distance theory, which characterises equality of distributions rather than equality of means (see Section 5.2).

Given K clusters $C_1, \dots, C_K$ of sizes $n_1, \dots, n_K$ with $N = \sum_{k=1}^K n_k$, the total dispersion of the pooled sample decomposes additively as [34]:

$$T^\alpha = W^\alpha + S^\alpha, \quad (5.15)$$

where the within-cluster dispersion

$$W^\alpha = \sum_{k=1}^K W^\alpha(C_k) = \sum_{k=1}^K \frac{1}{2n_k} \sum_{i \in C_k} \sum_{j \in C_k} \|\mathbf{x}_i - \mathbf{x}_j\|^\alpha \quad (5.16)$$

is identical to the objective minimised by K-Groups (Section 5.2.2), and the between-cluster energy statistic is

$$S^\alpha = \sum_{1 \leq j < l \leq K} \frac{n_j n_l}{2N} \mathcal{E}_{n_j, n_l}^{(\alpha)}(C_j, C_l). \quad (5.17)$$

The **DISCO F -ratio** is then:

$$F_{N,\alpha} = \frac{S^\alpha / (K - 1)}{W^\alpha / (N - K)}. \quad (5.18)$$

It is consistent against all distributional alternatives for $0 < \alpha < 2$ and is maximised over K . $F_{N,\alpha}$ is the distributional analogue of the Calinski–Harabasz index: both adopt the same F -ratio structure decomposing total dispersion into between- and within-cluster components, and both are maximised by partitions with large separation and small internal scatter. Their difference is solely in what dispersion means. The CH index uses squared Euclidean distances from centroids, capturing only first-moment differences, while $F_{N,\alpha}$ replaces these with the energy statistic $\mathcal{E}_{n_j, n_l}^{(\alpha)}$, which vanishes if and only if the two distributions are identical [35, 34].

Normalised Energy Distance

To obtain a scale-free measure of pairwise separation between two clusters A and B , the normalised energy distance $H \in [0, 1]$ is defined as [35]:

$$H(A, B) = \frac{\mathcal{E}^{(\alpha)}(F_A, F_B)}{2 \mathbb{E} \|X - Y\|^\alpha} = 1 - \frac{\frac{1}{n_A^2} \sum_{i,j} \|a_i - a_j\|^\alpha + \frac{1}{n_B^2} \sum_{i,j} \|b_i - b_j\|^\alpha}{\frac{2}{n_A n_B} \sum_{i,j} \|a_i - b_j\|^\alpha}, \quad (5.19)$$

where $X \sim F_A$ and $Y \sim F_B$. From the properties established in Section 5.2, $H(A, B) = 0$ if and only if $F_A = F_B$, and $H(A, B) \rightarrow 1$ indicates maximal separation. Since the normalised energy distance is computed for every pair of clusters, the result is a $K \times K$ matrix of pairwise distributional dissimilarities.

This H -matrix complements the DISCO F -ratio by operating at a different level of resolution: whereas $F_{N,\alpha}$ aggregates between- and within-cluster energy dispersions into a single partition-level statistic, $H(C_j, C_l)$ isolates the distributional distance between each specific pair of clusters independently of the remaining partition structure. The two indices can diverge in informative ways, since the DISCO F -ratio weights each pairwise energy statistic by $n_j n_l / 2N$, so large clusters dominate the numerator regardless of how distinct smaller clusters may be. The H -matrix makes these asymmetries directly visible.

Comments

Compared with the geometry-based indices, both H and $F_{N,\alpha}$ are sensitive to the full distribution rather than to means alone, and therefore remain informative when clusters differ in covariance structure, skewness, or tail behaviour rather than in location. In this respect they complement rather than replace geometry-based indices: when both families agree, there is strong evidence of a robust partition; when they disagree, it is a signal that clusters may differ in higher-order distributional properties that centroid-based indices cannot detect.

5.3.2 External Evaluation

External validation complements the quantitative internal indices by assessing whether the clustering result is meaningful and interpretable in the context of the problem at hand. Rather than measuring geometric or distributional properties of the partition, it evaluates the clusters against domain knowledge and practical objectives. In this work, where the ultimate criterion is actuarial interpretability, three supervised visual diagnostics are used.

Mean Dominant Province Profile

The coordinate-wise mean of the compositional damage vectors assigned to each cluster reveals whether the partition recovers geographically coherent flood patterns and whether the resulting profiles are sufficiently distinct to justify separate treatment in the model. This diagnostic directly addresses the geographical purpose of the evaluation framework: a partition that does not recover the Meuse–Scheldt divide visible in Figure 3.1 would fail the most basic interpretability requirement regardless of its internal metric scores.

OEP Curves and Frequency Share

The OEP curve and frequency share ω_k of each cluster, defined in Chapter 4, provide a severity profile that an actuary can assess directly. Clusters should exhibit meaningfully different return-period loss levels for the partition to add value over a single-portfolio view. In particular, the ω_k decomposition reveals which cluster drives aggregate portfolio tail risk at each return period.

5.4 K -selection

As discussed in Sections 5.1 and 5.2, partition-based algorithms require a pre-specified number of clusters K . In practice, K is treated as a hyperparameter and selected by evaluating a quality index over a range of candidate values $K = 2, \dots, K_{\max}$ and retaining the value that maximises it. For this work, we consider the silhouette score and the DISCO F -ratio.

A methodological remark concerns how these indices are parameterised across two distinct roles. During K -selection, the indices are evaluated using parameters consistent with the fitting objective wherever possible: for K-Medoids variants, the silhouette score is computed using the same pairwise distance metric used during fitting, and for K-Groups and HEC the DISCO F -ratio is computed at the fitting α . When no such coherent parametrisation exists the default common space is used: Euclidean silhouette and DISCO F -ratio at $\alpha = 1$. For comparative reporting across all algorithms, this common metric space is adopted throughout, regardless of the distance metric or energy exponent used during fitting. This is a deliberate choice, as comparing algorithms in a common metric space is more informative than evaluating each under its own fitting geometry. The practical consequence is that algorithms optimised under non-Euclidean distances or $\alpha \neq 1$ may be structurally favoured by these reported indices, a limitation acknowledged and discussed further in Section 5.5.

The candidate range $K = 2, \dots, K_{\max}$ with consecutive integer steps exhausts all partitions up to the specified upper bound. The hierarchical energy clustering algorithm (Section 5.2.1) participates in the same evaluation framework but is treated differently, since HEC builds a complete dendrogram in a single fit and K is selected by evaluating the quality indices at each possible dendrogram cut rather than requiring a full refit for

each candidate value.

An initial run with $K_{\max} = 20$ was performed across all algorithms, a value chosen as sufficiently large relative to the number of provinces (11) to ensure that any meaningful partition structure would emerge without artificially constraining the search, while remaining computationally tractable. Inspection of Tables B.1 and B.2 reveals that the majority of algorithms select an optimal K above 10, particularly when silhouette is used as the selection criterion. This is a purely metric-driven K selection, since a finer partition always tends to increase geometric and distributional homogeneity within clusters at a significant cost to interpretability. Examining the corresponding cluster profiles confirms this pattern: at high K , clusters become dominated by a single province, making the partition essentially trivial and approaching the degenerate solution visible in Figure 3.1. Furthermore, the OEP curves and $\omega_k(T)$ reveal that capital risk is concentrated in a small number of clusters regardless of K , specially for high return periods. This suggests that the additional clusters introduced at high K capture noise rather than genuine risk structure.

The OEP curves and $\omega_k(T)$ distributions are particularly informative in this respect. Across all algorithms and values of K , three clusters consistently emerge with heavy-tailed loss distributions, ordered by decreasing tail importance and associated with flood events concentrated in Liège, Namur, and Oost-Vlaanderen respectively, while the remaining clusters exhibit substantially lighter tails. The same analysis has already been suggested in Section 3.2.1. This structure is observed consistently across algorithms, suggesting that three distinct tail risk behaviours are robustly encoded in the data regardless of the clustering method. This systematic pattern, combined with the interpretability cost of high K , motivates a more conservative upper bound. Setting $K_{\max} = 5$ is therefore adopted as the working constraint for the main analysis: three clusters capture the dominant risk drivers identified above, while a fourth and fifth cluster can accommodate more subtle secondary risk patterns and ensure that low-severity, geographically diffuse events are grouped together rather than inflating the cluster count with actuarially irrelevant partitions.

It should be emphasised that the choice of $K_{\max} = 5$ is a subjective judgement informed by the joint examination of quantitative indices and actuarial diagnostics rather than a formal optimisation criterion. A robust and defensible determination of $K_{\max}$ in general requires deep knowledge of the underlying data, iterative exploration of results, and domain expertise, all of which inform but cannot be reduced to any single metric.

5.5 Results

We now present the obtained results. Several relevant findings emerge that respond to different actuarial needs, and a proposal is made for choosing an algorithm variant given a particular scenario. The observations below are based on the systematic examination of all algorithm configurations listed in Tables 5.2 and 5.1, together with the diagnostic plots compiled in Annex.

Method	Time (s)	K	Silhouette $\uparrow$	CH $\uparrow$	DISCO $F \uparrow$
KM-euclidean	3.5	5	0.2664	539.00	350.36
KM-manhattan	3.7	5	0.2781	500.28	308.08
KM-cosine	3.6	5	0.2772	514.38	318.49
KM-braycurtis	3.8	5	0.2420	421.13	275.36
KM-canberra	3.9	5	0.2157	343.35	244.72
KM-chebyshev	4.8	5	0.2336	459.15	285.22
KM-correlation	4.7	5	0.2825	509.75	339.02
GMM-full	7.5	2	0.1255	334.50	247.05
GMM-tied	3.7	5	0.2669	499.66	348.81
GMM-diag	4.7	2	0.1246	332.28	246.30
GMM-spherical	4.6	5	0.2378	422.78	346.02
HEC- α 0.4	36.3	5	0.2095	394.98	324.51
HEC- α 0.8	28.2	5	0.2376	442.23	336.89
HEC- α 1.0	2.9	5	0.2540	494.45	321.74
HEC- α 1.2	28.7	5	0.2599	503.01	339.24
HEC- α 1.6	28.8	5	0.2616	492.27	332.17
HEC- α 2.0	27.3	5	0.2540	494.45	321.74
KGR- α 0.4	8.3	5	0.2602	476.74	356.57
KGR- α 0.8	8.2	5	0.2788	538.93	360.05
KGR- α 1.0	7.3	5	0.2820	544.12	360.45
KGR- α 1.2	7.4	5	0.2796	530.99	348.04
KGR- α 1.6	7.2	5	0.2825	538.91	342.77
KGR- α 2.0	7.4	5	0.2651	548.31	327.40

Table 5.1: Clustering evaluation metrics with K selected by maximising the Silhouette score ($K_{\max} = 5$). Arrows indicate the direction of improvement.

Hydrological plausibility as a necessary condition. The most fundamental finding, though not an unexpected one given the hydrological structure established in Chapter 2, is that two clusters can always be associated with the Meuse (dominated by Liège and Namur) and Scheldt (dominated by Oost-Vlaanderen, Hainaut, and West-Vlaanderen) basins respectively, regardless of the algorithm, distance metric, covariance structure, or energy exponent α . Both structures are primary in the sense that neither precedes the other: once the data space is partitioned, both basins emerge simultaneously as geographically interpretable groups. This universality serves less as a discovery than as a validation, confirming that all algorithms recover the dominant hydrological structure of the Belgian territory without exception. However, not all algorithms produce an equally clean separation. ECL- α 0.4 under DISCO F selection ($K = 2$), for instance, isolates Oost-Vlaanderen in a single cluster and groups the remainder together, resulting in non-negligible contributions from both Meuse and non-Meuse provinces within the same cluster (see Figure 5.2). KM-chebyshev under DISCO F selection ($K = 2$), by contrast, provides the clearest primary partition and is the recommended choice when only the Scheldt–Meuse separation is sought (see Figure 5.1).

For higher K , recovering the two primary basins is a necessary but not sufficient condition: algorithms that produce geographically incoherent partitions should be excluded

Method	Time (s)	K	Silhouette $\uparrow$	CH $\uparrow$	DISCO $F \uparrow$
KM-manhattan	2.9	4	0.2405	479.16	311.83
KM-cosine	2.9	4	0.2459	479.44	327.60
KM-braycurtis	3.8	3	0.1832	439.28	291.05
KM-chebyshev	3.6	2	0.1710	475.53	342.14
GMM-tied	4.3	3	0.1808	443.18	360.79
HEC- α 0.4	37.2	2	0.0934	354.46	319.40
HEC- α 0.8	30.3	3	0.1232	354.73	325.72
HEC- α 1.0	2.9	2	0.1677	461.43	332.44
KGR- α 0.4	7.3	2	0.1379	442.62	355.94
KGR- α 0.8	8.5	3	0.1797	442.42	362.60

Table 5.2: Clustering evaluation metrics with K selected by maximising the DISCO F -ratio restricted to algorithm variants whose selected K differs from that obtained under silhouette selection.

regardless of their internal metric scores. For example, KM-correlation achieves competitive internal metrics but produces a cluster nominally associated with the Yser basin (West-Vlaanderen, 52.4%) but also carries a non-negligible contribution from Luxembourg (18.2%), a province with no hydrological connection to the Yser catchment (see Figure 5.4). For the same actuarial need of isolating the Yser basin, HEC- α 1.6 resolves this inconsistency by allocating Luxembourg events to the Meuse cluster and concentrating West-Vlaanderen into a geographically coherent group as shown in Figure 5.3.

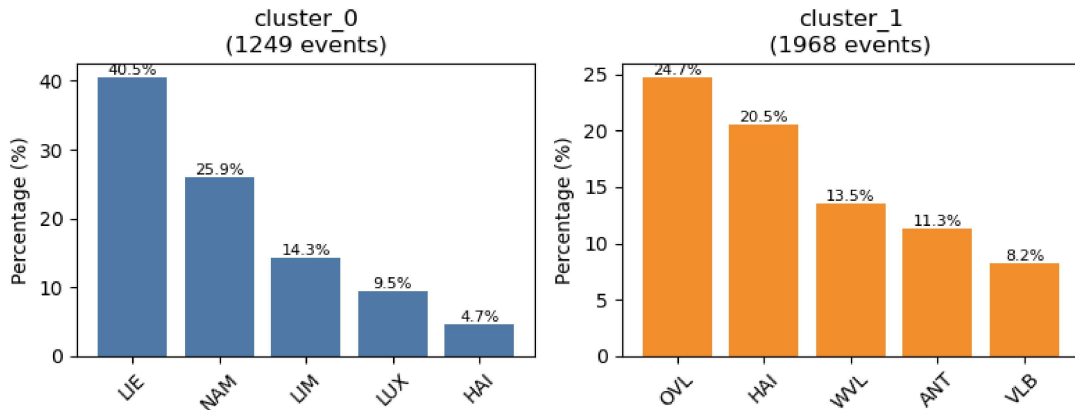


Figure 5.1: KM-chebyshev (DISCO F -selection) dominant provinces (top 5).

Scheldt partitions. Beyond the primary Meuse and Scheldt clusters, a third group typically emerges across algorithms, spanning Antwerpen, Vlaams-Brabant, Brabant Wallon, Brussels, Limburg, and occasionally Hainaut and West-Vlaanderen. This cluster presents a uniform distribution of damage patterns across the Scheldt basin provinces rather than a concentration in any single one. The uniformity of this third cluster is consistent with the low pairwise normalised energy distance H values observed relative to all other clusters: events with diffuse damage patterns are distributionally close to

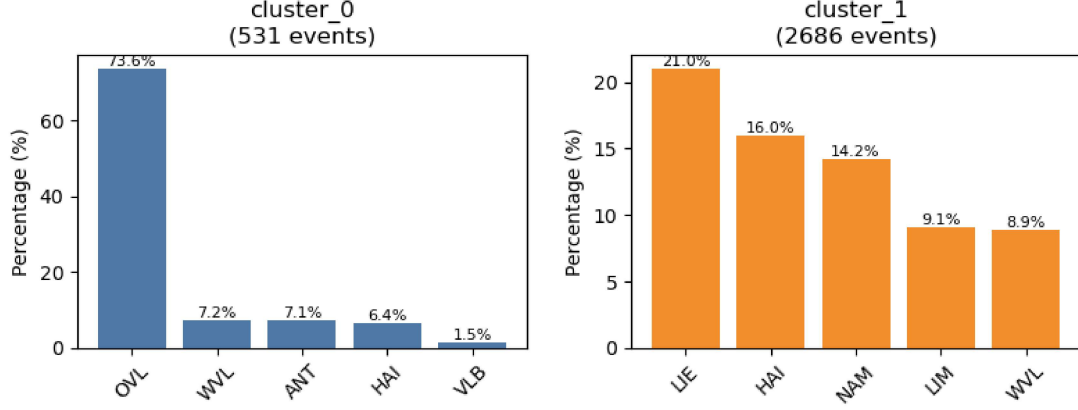


Figure 5.2: HEC- α 0.4 (DISCO F -selection) dominant provinces (top 5).

every other group, making it the least separated cluster in the partition. The group carries negligible tail risk across all configurations, as confirmed by OEP curves and $\omega_k(R)$ values actuarially indistinguishable from a low-severity benchmark. These events are numerous but non-dominant from a tail risk perspective, and their grouping into a single cluster is both statistically natural and practically desirable for identifying the main tail risks of the portfolio. We label this group the *low-risk Scheldt cluster*. In some configurations, provinces associated with the Meuse basin may also appear among its top five dominant provinces, though never as the most dominant one, in which case the label *agglomerative cluster* would be more precise, though this remains a labelling issue with no actuarial consequence. Notably, Oost-Vlaanderen is systematically distinguished from this group, maintaining its own flood pattern and concentrating the Scheldt tail risk in what we refer to as the *central Scheldt cluster*. Events extending into Hainaut and West-Vlaanderen may remain within the Oost-Vlaanderen cluster or be absorbed into the low-risk Scheldt cluster, an ambiguity typically resolved at higher K . Among algorithms producing this three-cluster structure, GMM-tied under DISCO F selection (Figures 5.5-5.7) cleanly identifies the tail risk of both primary basins while isolating the negligible-risk events in a dedicated third cluster, achieving a strong DISCO F -ratio and H values. KGR- α 0.8 achieves a marginally higher DISCO F (362.60) with practically identical evaluations, but computational efficiency, (re)insurance business communication, and widespread implementation in different programming languages.

At $K = 4$ and $K = 5$, additional splits emerge within the Scheldt basin, typically isolating the provinces identified in Section 2.2 as hydrologically ambiguous, namely Hainaut, West-Vlaanderen, and Limburg. Their emergence is statistically less stable and varies considerably across algorithms, reflecting sensitivity to the distance function rather than a robust risk structure. Since the tail risk remains concentrated in Oost-Vlaanderen regardless of these splits, pursuing this level of granularity is only warranted when a specific actuarial need justifies it; otherwise these additional clusters are of limited risk significance.

Meuse partition. The second type of split at higher K , and the only one carrying genuine risk zonation significance, separates the Meuse cluster into a Liège-dominated and

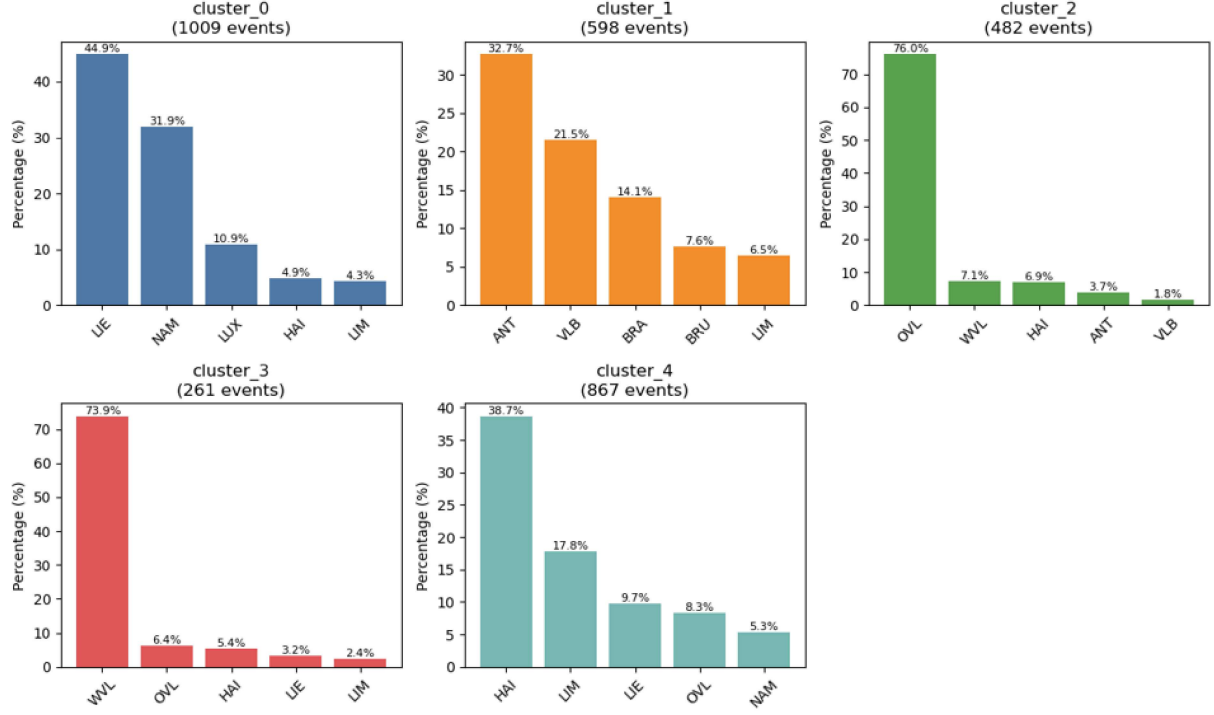


Figure 5.3: HEC- α 1.6 (silhouette) dominant provinces (top 5).

a Namur-dominated group. In terms of OEP curves, the two sub-clusters behave similarly up to return periods of approximately 200 years, beyond which the Liège-dominated cluster begins to yield materially higher losses (ex: Figure 5.8). The $\omega_k(R)$ decomposition confirms that beyond $R = 150$ years the Liège-dominated cluster consistently dominates and its share grows monotonically with R (ex: Figure 5.10). Whether to retain the Meuse as a single cluster or to separate it is a purely actuarial choice, relevant not only when the return period horizon exceeds 200 years but also when treaty exposures are geographically concentrated in one of the two regions. This split is not independent of the Scheldt sub-partitions described above: the Liège–Namur separation only appears at $K = 5$, so selecting an algorithm that produces it will simultaneously yield a segmented Scheldt basin. Multiple algorithms exhibit this partition, and among them KGR- α 1.0 offers the best trade-off across internal and external evaluations (see Figures 5.8-5.10). If simplicity and explainability are priorities, KM-euclidean also recovers the Liège–Namur split at $K = 5$ with strong evaluations and is arguably the easiest algorithm to communicate in a business context. In some configurations a further split isolating a Luxembourg-dominated group from the Meuse cluster is observed, though it carries negligible additional tail risk differentiation.

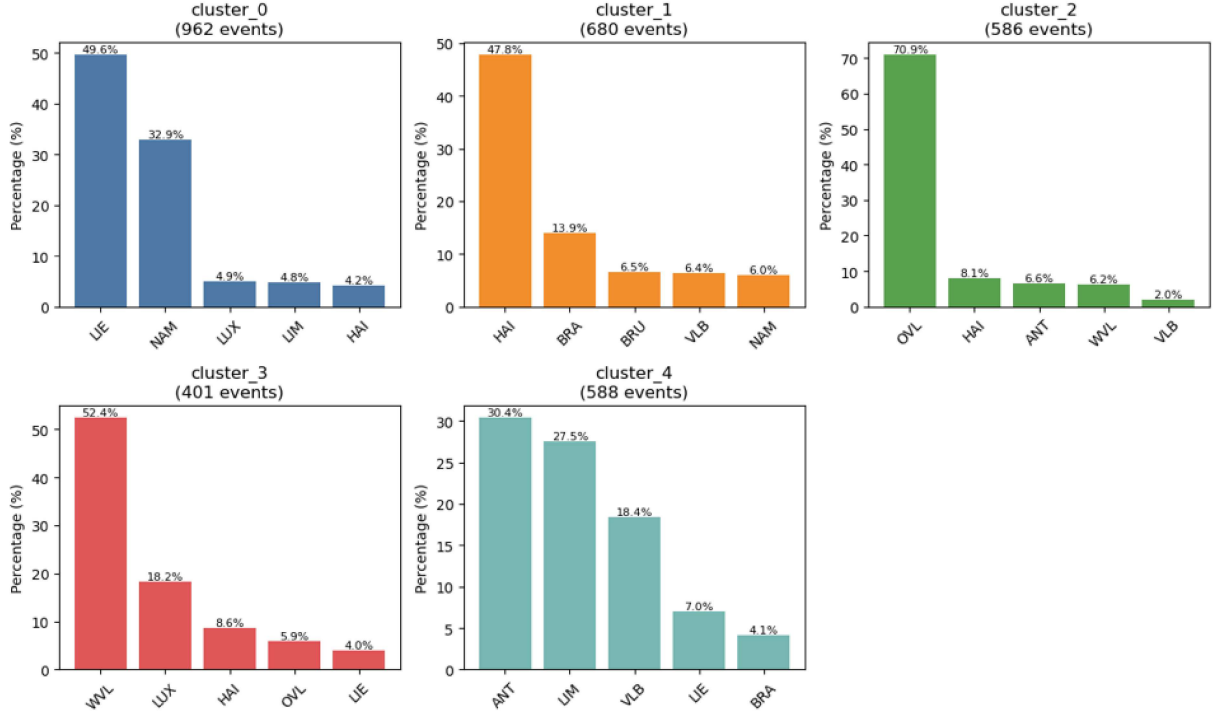


Figure 5.4: KM-correlation (DISCO F /silhouette -selection) dominant provinces (top 5).

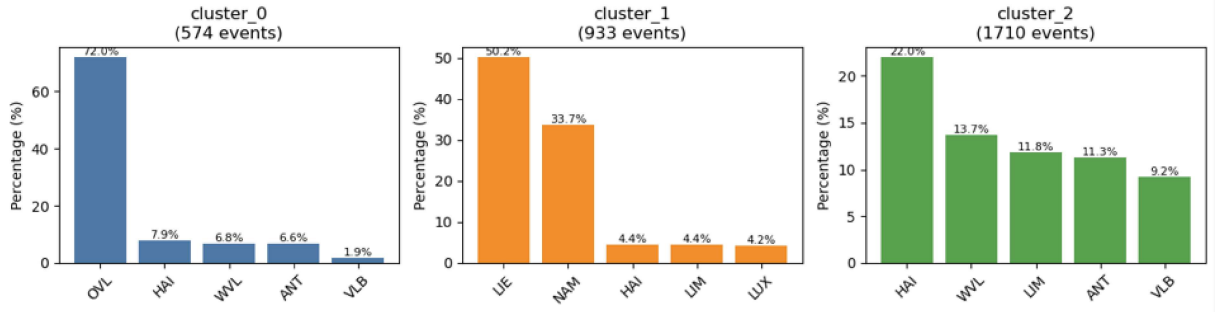


Figure 5.5: GMM-tied (DISCO F -selection) dominant provinces (top 5).

Remark. The results are presented following a linear logic of partitions through K , but this should not be understood as a strictly hierarchical structure: setting an algorithm to $K = 4$ and $K = 5$ does not guarantee that the clusters in $K = 5$ follow the structure observed at $K = 4$. As an exception, HEC algorithms are useful precisely for tracking how partitions emerge as K increases, since the dendrogram provides a direct reading of the merge sequence. Although the resulting hierarchies cannot be extrapolated directly to other algorithms, they offer a useful reference for understanding the order and relative stability of the splits observed across the full set of methods.

Risk hierarchy. The most important aggregate finding is that the Meuse basin is always the most capital-intensive catchment in the portfolio, and this dominance intensifies sharply at high return periods. Below 100 years, OEP curves remain comparable across clusters and multiple groups contribute meaningfully to the total exceedance frequency

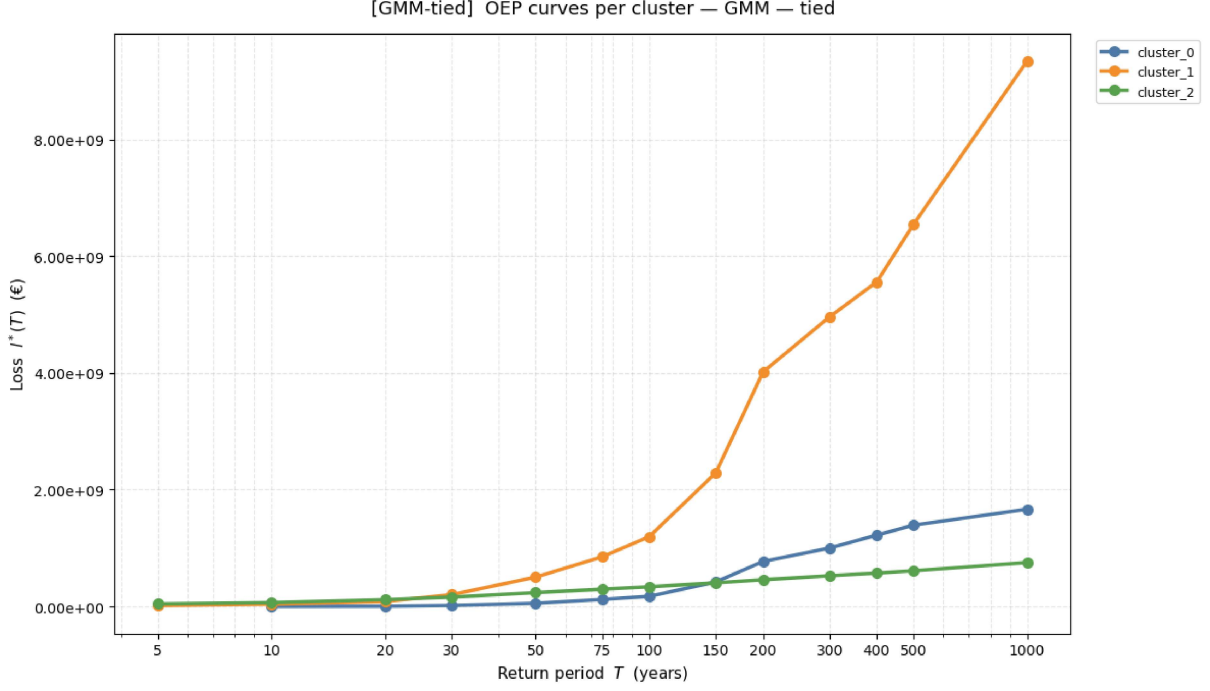


Figure 5.6: GMM-tied (DISCO F -selection) OEP curves.

$\nu_*(l)$. Beyond approximately $R = 150$ years, the exceedance frequencies of all non-Meuse clusters collapse toward zero, with the Oost-Vlaanderen cluster being the last to do so, while the OEP curves associated with the Meuse cluster grow significantly relative to all others. The tail risk hierarchy of the Belgian flood portfolio can therefore be summarised as: Liège, Namur, and Oost-Vlaanderen, corresponding hydrologically to the Meuse and central Scheldt basins (ex: GMM-tied Figures 5.5-5.7). Whether this dominance reflects a genuinely higher flood hazard, differences in flood protection infrastructure, or some combination of these factors lies beyond the scope of this work and represents a direction for further actuarial investigation.

Internal metrics, algorithms, and computational cost. As established in Sections 5.3.1, the DISCO F -ratio penalises K more strongly than the silhouette score, which is directly visible in the results: under DISCO F selection several algorithms select $K = 2$ or $K = 3$, while under silhouette selection the large majority converge to $K = 5$. K-Groups consistently achieves the highest DISCO F -ratio values and competitive silhouette and Calinski–Harabasz scores across all its variants, making it the strongest algorithm family on quantitative metrics alone. However, K-Means and GMM variants with similar evaluations are available and arguably preferable in practice: they are more widely implemented, better known in the actuarial and statistical communities, and offer clear advantages in terms of business communication and regulatory explainability. HEC variants, while theoretically grounded and particularly useful for visualising the sequential emergence of clusters through the dendrogram, are substantially more expensive computationally without producing internal metrics that compensate for this cost. GMM-full and GMM-diag underperform consistently across all metrics and both selection criteria, reflecting the difficulty of fitting unconstrained or diagonal covariance structures on this dataset.

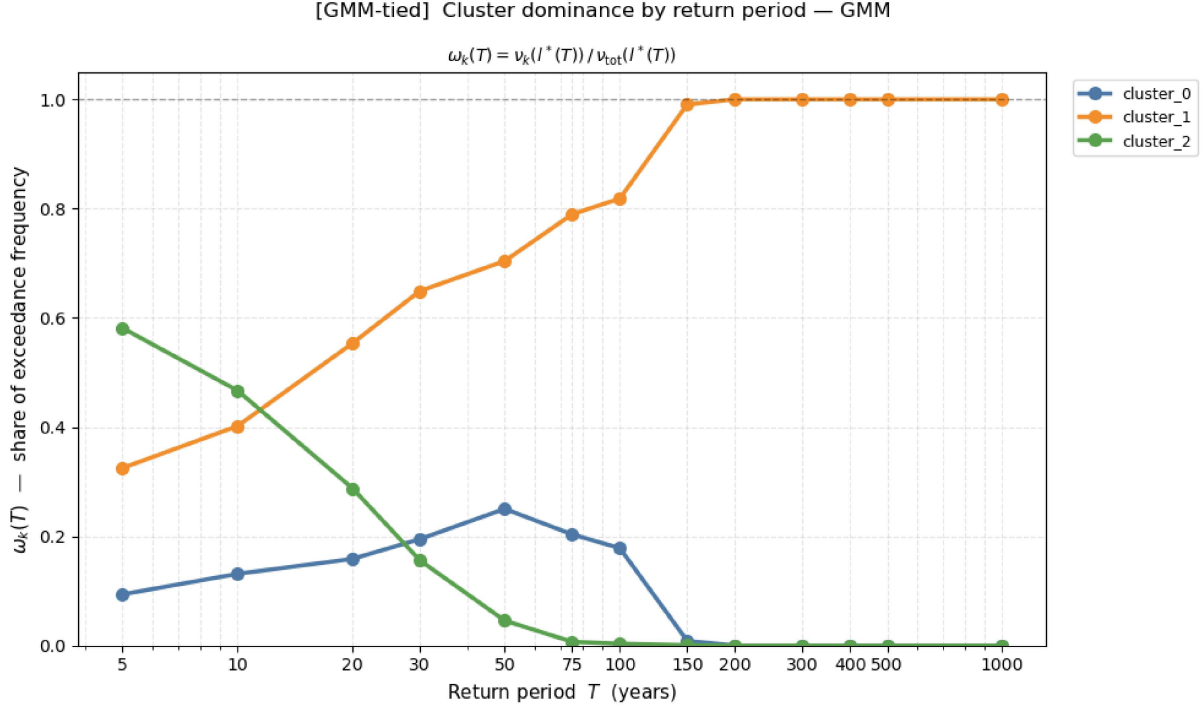


Figure 5.7: GMM-tied (DISCO F -selection) $\omega_K(R)$ values.

These results should be interpreted bearing in mind that all indices are reported in a common metric space, as discussed in Section 5.4: algorithms expected to benefit most from this alignment do not stand apart from the rest by a large margin, suggesting that the comparison is not severely distorted.

5.6 Conclusions

Acknowledging that no universally superior algorithm exists and that the optimal choice depends on the specific actuarial objective, GMM-tied with DISCO F -selection is the preferred algorithm for general-purpose zonation of the Belgian flood risk landscape. At $K = 3$ under DISCO F selection, it cleanly separates the Meuse, Scheldt, and low-risk Scheldt clusters, identifying the tail risk of both primary basins without introducing unnecessary fragmentation, while explicitly isolating the negligible-risk events in a dedicated third cluster and making the risk hierarchy of the portfolio immediately readable. The DISCO F -ratio is strong and the H value between the Meuse and Scheldt clusters indicates robust distributional separation. The relatively low H values involving the low-risk Scheldt cluster are consistent with the uniform distributional character of that group and do not reflect a failure of the partition. The widespread availability and interpretability of GMM across statistical software environments and the actuarial community represent a further practical advantage for business communication and regulatory reporting.

Regarding the K -selection criterion, the analysis supports the use of the DISCO F -ratio as the primary guide. As established in Section 5.4, the DISCO F -ratio penalises higher

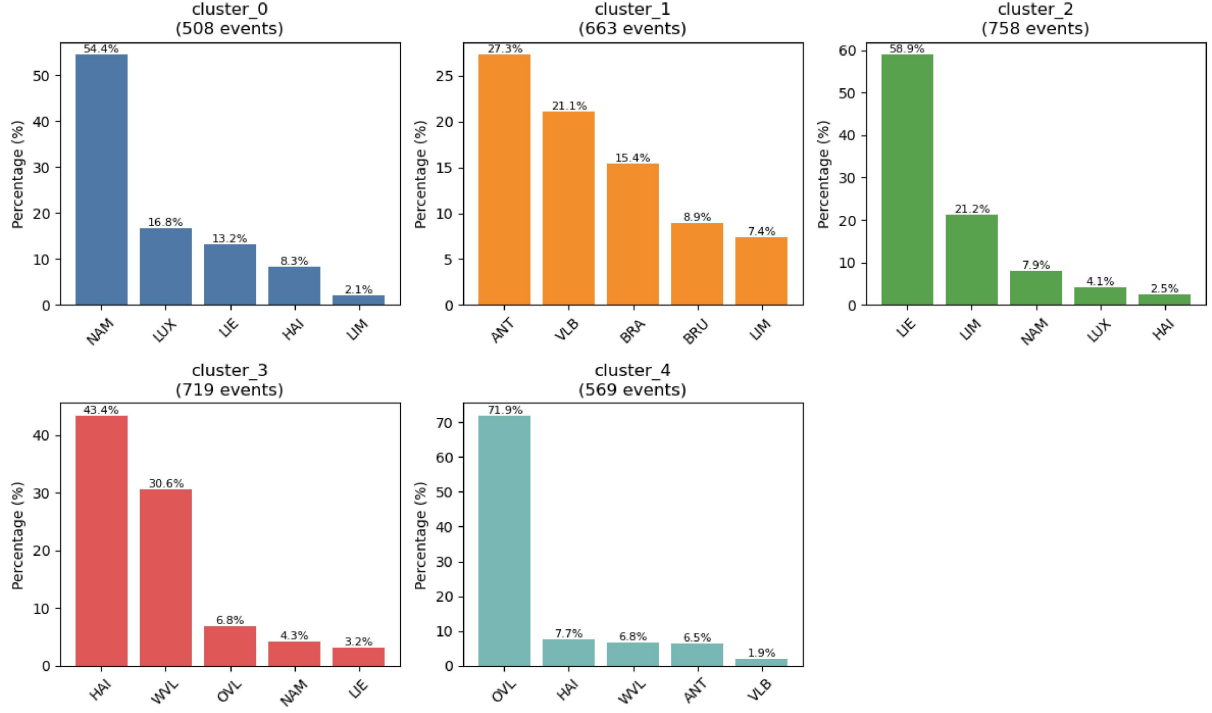


Figure 5.8: KGR- α 1.0 (DISCO F /silhouette -selection) dominant provinces (top 5).

values of K more strongly than the silhouette score, tending to favour more parsimonious partitions and avoiding the trivial high- K solutions observed in the preliminary runs with $K_{\max} = 20$. As an internal evaluation metric, its distributional character complements the geometry-based indices rather than replacing them: it captures differences in shape, tail behaviour, and covariance structure that centroid-based indices cannot detect, and is particularly informative when clusters are expected to differ in distributional structure rather than location alone. The joint use of DISCO F alongside geometric indices such as the silhouette score and the Calinski–Harabasz index proved effective in this analysis and is recommended in contexts where the distributional heterogeneity of clusters is a primary concern.

As for further research, several directions can be pursued. A number of methodological choices made in this work involve subjective judgements that depend largely on the objectives of the study. The primary choice was to introduce the concept of energy distance and its application to clustering, setting aside alternative validation metrics or clustering algorithms. In this context, the dendritic structure of HEC algorithms suggests that further research could be conducted into other hierarchical clustering methods as a tool for tracking the sequential emergence of partitions and, in the best-case scenario, yielding better results. Minor methodological choices, such as computing cluster dominance via the coordinate-wise mean of compositional damage vectors or the space in which the clustering and evaluation will take place, carry their own inconsistencies that have not been explored in depth for the sake of simplicity, but that are nonetheless worth investigating in future work.

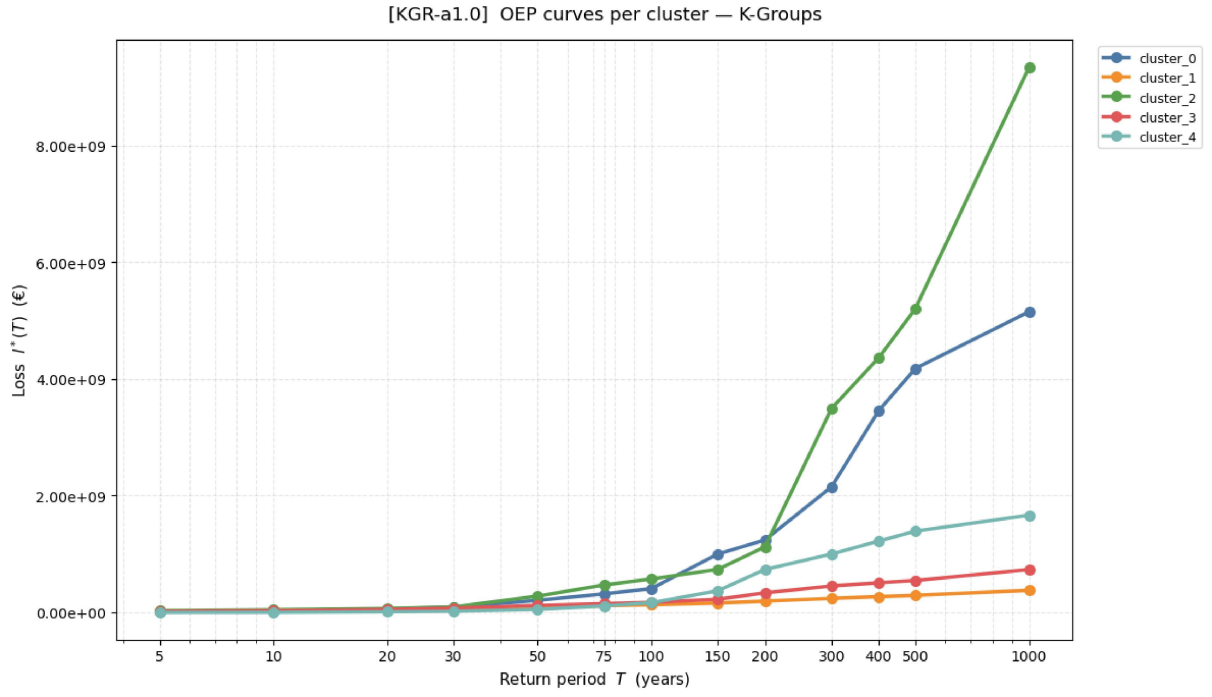


Figure 5.9: KGR- α 1.0 (DISCO F /silhouette -selection) OEP curves.

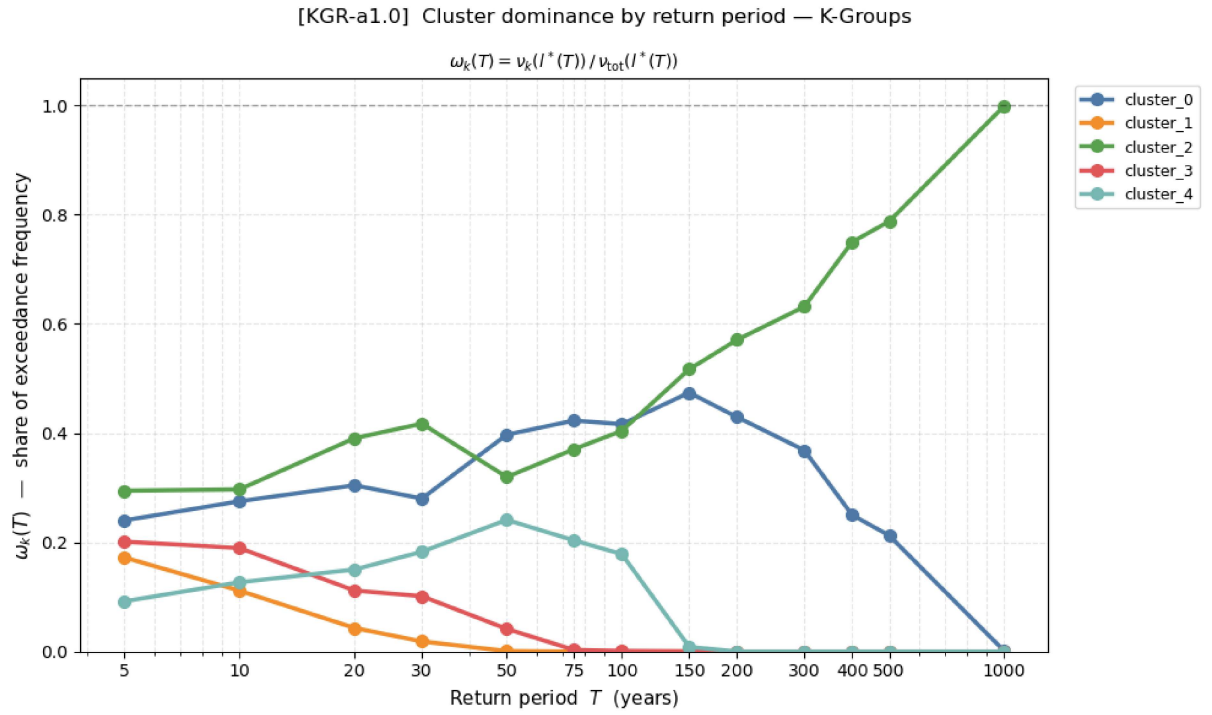


Figure 5.10: KGR- α 1.0 (DISCO F /silhouette -selection) ω_k values.

Chapter 6

Reinsurance and Solvency II Applications

The catastrophe model output and clustering methodology developed in the preceding chapters are applied in this chapter to the computation of the regulatory Solvency Capital Requirement under an internal model, and to the assessment of a Cat XL reinsurance treaty as a capital mitigation instrument. The Solvency II use test and the role of geographic granularity for climate change adaptation are also discussed. Throughout, the GMM-tied clustering result serves as the running numerical example, with three clusters labelled according to their dominant hydrological basin: Cluster 0 (Central Scheldt), Cluster 1 (Meuse), and Cluster 2 (Low-risk Scheldt), as illustrated in Figure 5.5. The developments apply unchanged to any clustering result.

6.1 The Gross Solvency Capital Requirement

Under the Solvency II internal model [16, 14], the Solvency Capital Requirement is the loss level that the maximum annual occurrence loss $M = \max_{i,j} X_{ij}$ exceeds with probability 0.5% over a one-year horizon, $\text{SCR}^g = \text{VaR}_{99.5\%}(M)$. From the return period relation of equation (4.8), the gross SCR is the unique solution to:

$$\nu(\text{SCR}^g) = \ln\left(\frac{200}{199}\right) \approx 0.005013, \quad (6.1)$$

where $\nu(l) = \sum_i \lambda_i [1 - F_i(l)]$ is the aggregate exceedance frequency of the relevant portfolio or sub-portfolio (equation (4.5)).

6.1.1 Portfolio and Cluster Level

At the portfolio level, $\nu = \nu_*$ in equation (6.1) gives SCR_*^g . The aggregate annual loss decomposes as $S = \sum_{k=0}^{K-1} S_k$, where each S_k is an independent compound Poisson

sub-aggregate with cluster exceedance frequency $\nu_k(l) = \sum_{i \in \mathcal{C}_k} \lambda_i [1 - F_i(l)]$, satisfying $\nu_*(l) = \sum_k \nu_k(l)$ (equation (4.9)). At the cluster level, substituting $\nu = \nu_k$ gives SCR_k^g , the capital cluster k would require as an isolated portfolio.

For the GMM-tied clustering, solving equation (6.1) numerically at each level yields:

- Portfolio: $\text{SCR}_*^g = 4,017,308,405 \text{ €}$
- Cluster 0 (Central Scheldt): $\text{SCR}_0^g = 769,994,161 \text{ €}$
- Cluster 1 (Meuse): $\text{SCR}_1^g = 4,017,308,401 \text{ €}$
- Cluster 2 (Low-risk Scheldt): $\text{SCR}_2^g = 456,343,755 \text{ €}$

The sum of standalone SCRs exceeds the portfolio SCR by $1,226,337,912 \text{ €}$ (30.5%). This diversification benefit reflects a certain degree of hydrological independence of the Meuse and Scheldt basins (Section 2.3). This is a consequence from S_k being independent, $\text{VaR}_p(S_0 + S_1 + S_2) \leq \sum_k \text{VaR}_p(S_k)$, with the inequality strict because a catastrophic Meuse year carries no direct implication for Scheldt losses in that same year.

To decompose the portfolio SCR across clusters without double-counting, the exceedance frequency share $\omega_k(T)$ of Section 4.2.3 is evaluated at $T = 200$:

$$\omega_k(200) = \frac{\nu_k(\text{SCR}_*^g)}{\nu_*(\text{SCR}_*^g)}, \quad \sum_{k=0}^{K-1} \omega_k(200) = 1. \quad (6.2)$$

The quantity $\omega_k(200)$ measures the proportion of the total expected annual exceedances of SCR_*^g attributable to cluster k ; it is a frequency-based allocation, not a proportion of losses. Figure 5.7 shows $\omega_k(T)$ across return periods. At $T = 200$, $\omega_0(200) \approx 0$, $\omega_1(200) = 1.000$, and $\omega_2(200) \approx 0$: the entire regulatory capital requirement is driven by Meuse flood scenarios, and no event in Clusters 0 or 2 produces an individual occurrence reaching SCR_*^g at the 200-year threshold.

6.2 XL Reinsurance as SCR Mitigation

A catastrophe excess-of-loss (Cat XL) treaty [25] with attachment point $R > 0$ and layer limit $L > 0$ transfers to the reinsurer the portion of each occurrence loss between R and $R + L$. The present analysis considers the unlimited layer ($L = \infty$), under which the reinsurer absorbs all losses above R with no upper cap. While such unlimited treaties are not used in private reinsurance markets they do exist in state-backed public-private partnerships for natural catastrophe protection, as illustrated by the Spanish Consorcio de Compensación de Seguros [9]. The determination of the optimal (R, L) pair in practice involves considerations—such as risk appetite, market conditions, and reinsurance pricing—that are beyond the scope of this work.

The net and ceded losses per occurrence are:

$$X_{ij}^{\text{net}} = \min(X_{ij}, R), \quad X_{ij}^{\text{ced}} = \max(X_{ij} - R, 0). \quad (6.3)$$

The annual ceded aggregate loss is $S^{\text{ced}} = S - S^{\text{net}} = \sum_k S_k^{\text{ced}}$, where $S_k^{\text{ced}} = \sum_{i \in \mathcal{C}_k} \sum_j X_{ij}^{\text{ced}}$. The independence structure across clusters is preserved since the ceding operation acts event-by-event and does not alter the Poisson occurrence processes.

6.2.1 Net SCR and Capital Relief

Under the unlimited treaty, the net survival function per occurrence is:

$$\mathbb{P}(X_{ij}^{\text{net}} > l) = \begin{cases} 1 - F_i(l) & l < R, \\ 0 & l \geq R, \end{cases} \quad (6.4)$$

so the net exceedance frequency satisfies, for any ν whether portfolio-level ν_* or cluster-level ν_k :

$$\nu^{\text{net}}(l) = \begin{cases} \nu(l) & l < R, \\ 0 & l \geq R. \end{cases} \quad (6.5)$$

The net SCR is the solution to $\nu^{\text{net}}(\text{SCR}^{\text{net}}) = \ln(200/199)$, and the capital relief is:

$$\Delta \text{SCR}(R, \infty) = \text{SCR}^g - \text{SCR}^{\text{net}} \geq 0. \quad (6.6)$$

Since $\nu^{\text{net}}(l) = 0$ for all $l \geq R$, any solution to the net SCR condition must lie in $(0, R)$, where ν^{net} coincides with ν . When $R \geq \text{SCR}^g$, the gross SCR threshold falls within this interval, so the net SCR condition is satisfied at the same point and $\Delta \text{SCR} = 0$, thus no capital is freed. When $R < \text{SCR}^g$, the gross threshold lies outside $(0, R)$ and the net SCR is pushed below it, giving $\text{SCR}^{\text{net}} = R$ and $\Delta \text{SCR} > 0$. The equivalence:

$$\Delta \text{SCR}(R, \infty) > 0 \iff R < \text{SCR}^g, \quad (6.7)$$

holds for any exceedance frequency ν .

Equation (6.7) shows that the retention R partitions the clusters into regimes according to whether R falls above or below each standalone SCR: clusters whose standalone SCR exceeds R obtain capital relief, while those whose standalone SCR lies below R are unaffected.

For the GMM-tied clustered portfolio, the standalone SCRs satisfy $\text{SCR}_2^g < \text{SCR}_0^g \leq \text{SCR}_1^g$. At $R = 1,000 \text{ M€}$, since $\text{SCR}_0^g = 770 \text{ M€} \leq R < \text{SCR}_1^g = 4,017 \text{ M€}$, equation (6.7) gives $\Delta \text{SCR}_1^g > 0$ while $\Delta \text{SCR}_0^g = \Delta \text{SCR}_2^g = 0$: only the Meuse cluster has its standalone SCR reduced, from $\text{SCR}_1^g = 4,017 \text{ M€}$ down to R . The standalone SCRs of Clusters 0 and 2 remain unchanged since their standalone SCRs lie below the retention. The portfolio and cluster results are:

- Portfolio: $\text{SCR}_*^{\text{net}} = R = 1,000,000,000 \text{ €}$, $\Delta \text{SCR}_* = 3,017,308,405 \text{ €}$
- Cluster 1 (Meuse): $\text{SCR}_1^{\text{g,net}} = R = 1,000,000,000 \text{ €}$, $\Delta \text{SCR}_1^g = 3,017,308,401 \text{ €}$
- Cluster 0 (Central Scheldt): $\text{SCR}_0^{\text{g,net}} = \text{SCR}_0^g = 769,994,161 \text{ €}$, $\Delta \text{SCR}_0^g = 0$
- Cluster 2 (Low-risk Scheldt): $\text{SCR}_2^{\text{g,net}} = \text{SCR}_2^g = 456,343,755 \text{ €}$, $\Delta \text{SCR}_2^g = 0$

6.2.2 Ceded Annual Average Loss

The expected annual ceded loss decomposes as:

$$\mathbb{E}[S^{\text{ced}}] = \sum_{k=0}^{K-1} \mathbb{E}[S_k^{\text{ced}}] = \sum_{k=0}^{K-1} \sum_{i \in \mathcal{C}_k} \lambda_i \mathbb{E}[\max(X_{ij} - R, 0)]. \quad (6.8)$$

Under the Beta severity parameterisation of Section 4.3, where $X_{ij} = Y_{ij} \cdot E_i$ with $Y_{ij} \sim \text{Beta}(\alpha_i, \beta_i)$, the per-event ceded mean can be written as:

$$\mathbb{E}[\max(X_{ij} - R, 0)] = \int_R^\infty \bar{F}_{X_{ij}}(x) dx = E_i \int_{R/E_i}^1 \bar{F}_{Y_{ij}}(y) dy, \quad (6.9)$$

where $\bar{F}_{X_{ij}}(x) = \mathbb{P}(X_{ij} > x)$ and $\bar{F}_{Y_{ij}}(y) = \mathbb{P}(Y_{ij} > y)$ are the survival functions of X_{ij} and Y_{ij} respectively, and the first equality uses the layer-cake representation of the expectation. This evaluates in closed form to:

$$\mathbb{E}[\max(X_{ij} - R, 0)] = E_i \bar{F}_{\alpha_i+1, \beta_i}\left(\frac{R}{E_i}\right) - R \bar{F}_{\alpha_i, \beta_i}\left(\frac{R}{E_i}\right), \quad (6.10)$$

where $\bar{F}_{\alpha, \beta}(x) = 1 - I(x; \alpha, \beta)$ is the complementary regularised incomplete beta function defined in Section 4.3.

For the GMM-tied portfolio at $R = 1,000 \text{ M€}$:

- $\mathbb{E}[S_0^{\text{ced}}] = 1,619,642 \text{ €}$
- $\mathbb{E}[S_1^{\text{ced}}] = 33,694,497 \text{ €}$
- $\mathbb{E}[S_2^{\text{ced}}] = 48,496 \text{ €}$
- $\mathbb{E}[S^{\text{ced}}] = 35,362,635 \text{ €}$

The cluster decomposition reveals an important asymmetry between the ceded AAL and the capital relief. From equation (6.7), Clusters 0 and 2 contribute nothing to the portfolio capital relief since $R > \text{SCR}_0^g$ and $R > \text{SCR}_2^g$. Yet their ceded AAL is non-zero. This asymmetry arises from the different nature of the two quantities: the capital relief is determined by the exceedance frequency $\nu_k(R)$ evaluated at the single threshold R , which is negligible for Clusters 0 and 2 at $R = 1 \text{ bn€}$; the ceded AAL is a linear integral of the survival function above R , accumulating contributions from the entire tail regardless of how small individual exceedance probabilities are. Events in Clusters 0 and 2 have non-zero probability of producing losses above R , they are simply too rare to move $\nu_k(R)$ to the level required by the SCR condition, and these small probabilities integrate to a positive ceded AAL via equation (6.10). The 4.7% of the ceded AAL attributable to Clusters 0 and 2 therefore represents a premium cost arising from the linearity of $\mathbb{E}[S^{\text{ced}}]$, with no corresponding capital benefit at the regulatory threshold.

The quantity $\mathbb{E}[S^{\text{ced}}]$ is the pure premium of the treaty in the actuarial sense, representing the expected annual cost transferred to the reinsurer before any loading for expenses,

profit, or capital. Determining a technically sound total premium from this pure premium requires further considerations including capital loading, market conditions, and cost allocation that are beyond the scope of this work. Jones et al. [25] provide an informative account of this process in the context of catastrophe reinsurance (Sections 2.12.3–2.12.5), illustrating that catastrophe modelling provides the pure premium as one input into a broader pricing decision.

6.2.3 Effect on the Exceedance Frequency Shares

After the treaty with $R < \text{SCR}_*^g$, the net SCR threshold equals R . Since $\nu_k^{\text{net}}(l) = \nu_k(l)$ for all $l < R$ (equation (6.5)), the net exceedance frequency shares at $T = 200$ are:

$$\omega_k^{\text{net}}(200) = \frac{\nu_k^{\text{net}}(R)}{\nu_*^{\text{net}}(R)} = \frac{\nu_k(R)}{\nu_*(R)} = \omega_k^{\text{gross}}(T_R), \quad (6.11)$$

where $T_R = 1/\text{OEP}_*(R)$ is the return period corresponding to R on the gross OEP curve. As with $\omega_k(200)$, the quantities $\omega_k^{\text{net}}(200)$ measure proportions of expected annual exceedance frequencies, not proportions of losses.

The net omega decomposition equals the gross omega evaluated at the return period T_R associated with the retention. The treaty does not alter the underlying risk structure of any cluster: it shifts the regulatory evaluation point from SCR_*^g down to R on the existing gross omega curve of Figure 5.7.

For the GMM-tied portfolio at $R = 1,000 \text{ M€}$, the OEP at the retention is $\text{OEP}_*(R) = 0.01539$, corresponding to a return period $T_R \approx 65$ years on the gross OEP. At that return period, the net exceedance frequency shares are:

- $\omega_0^{\text{net}}(200) = 0.2152$ (Central Scheldt)
- $\omega_1^{\text{net}}(200) = 0.7725$ (Meuse)
- $\omega_2^{\text{net}}(200) = 0.0124$ (Low-risk Scheldt)

All three clusters now contribute positively, in contrast to the gross decomposition where $\omega_1(200) = 1.000$. The apparent rebalancing away from full Meuse dominance does not reflect a change in the underlying risk of either basin: it reflects that at the 65-year level all three basins have non-negligible exceedance frequency in the gross model, and the treaty has moved the regulatory measurement point into that range. Figure 5.7 confirms this reading, since these values correspond to reading $\omega_k(T)$ at $T \approx 65$ on the gross curve.

This shift carries a practical implication for the insurer. Before the treaty, Scheldt-zone exposure had negligible capital impact because $\nu_0(\text{SCR}_*^g) \approx 0$. After the treaty, the net capital condition is evaluated at R , where $\nu_0(R) > 0$: new Central Scheldt underwriting now accounts for 21.5% of the net SCR condition, and even the Low-risk Scheldt cluster contributes 1.2%. The insurer must therefore update their decisions accordingly, using $\omega_k^{\text{net}}(200)$ rather than $\omega_k(200)$ to assess the capital impact of new exposures on the net portfolio.

6.3 Solvency II Use Test

The Solvency II use test requires that the internal model be embedded in the insurer's business decisions rather than used solely for regulatory reporting [14]. The cluster decomposition provides a direct link between the internal model and underwriting decisions, illustrated by the following simplified example.

Using the GMM-tied internal model, an insurer knows that $\omega_1(200) = 1.000$ thus the entire portfolio SCR is driven by Meuse events. The underwriter is presented with two new flood policies of equal sum insured, one located in Liège (Meuse basin, Cluster 1) and one in Antwerp (Scheldt basin, Cluster 0).

Adding the Liège policy increases $\nu_1(\text{SCR}_*^g)$ and therefore raises SCR_*^g : the new exposure contributes directly to the binding SCR condition. Adding the Antwerp policy increases $\nu_0(\text{SCR}_*^g)$, but since $\nu_0(\text{SCR}_*^g) \approx 0$, the effect on SCR_*^g is negligible.

The internal model therefore delivers a differentiated underwriting signal. The Meuse-zone exposure consumes capital at the full portfolio rate, while Scheldt-zone exposure has negligible capital impact at the 200-year horizon. This decision rule uses the same $\omega_k(200)$ decomposition that produces the SCR, satisfying the internal model use test requirement in order to drive business decisions [14].

After the unlimited treaty at $R = 1,000 \text{ M€}$, the underwriter monitors $\omega_k^{\text{net}}(200)$ over time. If new Meuse underwriting gradually pushes ω_1^{net} back toward one, the internal model signals that the treaty is becoming insufficient relative to the growing Meuse exposure and a revision of the attachment point is necessary.

6.4 Climate Change and Future Capital Needs

Climate change is expected to alter hydrological and precipitation patterns across the globe, with increasing evidence of more frequent and more intense extreme weather events [31]. Belgium is also projected to experience significant climatic changes over the course of the twenty-first century. As mentioned by Belgian Climate Portal [4], projections indicate a strengthening of precipitation seasonality by the end of the century, with summer precipitation potentially decreasing by up to 25% and winter precipitation increasing by up to 22%. Furthermore, the impacts of climate change are not expected to be spatially homogeneous across Belgium, as regional climatic and hydrological characteristics may lead to differing local responses to changing precipitation regimes. Regulatory expectations for the integration of climate change risk scenarios into the Own Risk and Solvency Assessment have been formalised by European Insurance and Occupational Pensions Authority [15], requiring insurers to account for both physical risks linked to evolving hazard frequency and severity, and transition risks associated with the shift toward a low-carbon economy.

Without geographic granularity, climate scenarios can only be applied as a uniform national loading on the full event catalogue. The cluster decomposition enables basin-

specific differential loadings: Meuse event rates λ_i for $i \in \mathcal{C}_1$ can be stressed independently of Scheldt rates, reflecting distinct physical mechanisms and available climate projections for each catchment. The resulting stressed $\omega_k(200)$ then quantifies which basin drives the additional capital need under each scenario.

This granularity has direct consequences for prevention policy and communication. Since $\omega_1(200) = 1.000$ in the current portfolio, the entire regulatory capital is exposed to Meuse hazard intensification. Investments in flood prevention and adaptation measures in the Meuse catchment directly reduce $\nu_1(\text{SCR}_*^g)$ and therefore SCR_*^g , while equivalent investments in the Scheldt basin have smaller capital impact at the regulatory threshold. The cluster framework therefore provides a quantitative basis for orienting prevention budgets and communicating the capital relevance of adaptation measures toward the basin where first-order regulatory impact can be expected.

6.5 Conclusions

This chapter shows how the cluster decomposition of the Belgian flood event distribution translates into concrete Solvency II applications. The main findings are the following.

The portfolio gross SCR of 4,017 M€ is driven entirely by Cluster 1 (Meuse) at the 200-year horizon ($\omega_1(200) = 1.000$), where $\omega_k(200)$ denotes the proportion of expected annual exceedances of SCR_*^g attributable to cluster k , not a proportion of losses. The 30.5% diversification benefit reflects the hydrological independence of the basins, made visible and quantifiable at the regulatory level through the cluster structure.

The unlimited XL treaty at $R = 1,000$ M€ reduces the portfolio SCR to the retention exactly, with capital relief of 3,017 M€ attributable entirely to the truncation of Cluster 1's exceedance frequency above R . The ceded AAL of Clusters 0 and 2, while positive due to super-200-year events, inflates the pure premium without contributing to capital relief. The shift in the net ω_k^{net} values from full Meuse dominance to a more balanced decomposition reflects the movement of the regulatory evaluation on a return period of approximately 65 years, not any movement in the underlying risk. This shift does carry a practical consequence: Scheldt exposure now contributes to the net SCR condition and the underwriting guidelines must be updated accordingly.

The use test application shows that $\omega_k(200)$ provides a direct underwriting signal: Meuse-zone policies consume capital at the full portfolio rate while Scheldt-zone policies have negligible capital impact. This differentiation is only available through the internal model and is consistent with the use test requirements of EIOPA [14].

Finally, the cluster structure provides the geographic resolution needed to apply basin-specific climate stress scenarios, directly connecting prevention investments to their quantitative capital impact under the ORSA framework [15].

General Conclusion

The central question of this thesis was whether the output of a probabilistic catastrophe model could be used to construct a segmentation of Belgian flood risk that is both hydrologically grounded and actuarially meaningful, without relying on any prior geographic convention.

Chapter 1 established the catastrophe modelling context and introduced the two datasets on which the entire analysis rests: the Event Loss Table and the Damage Distribution Table. Chapter 2 provided the physical argument for why a data-driven segmentation should be attempted at all, showing that the Scheldt and Meuse basins operate through structurally distinct flood-generating mechanisms and can flood independently. Chapter 3 confirmed that this physical structure is recoverable from the data itself: the spatial footprint of flood events, encoded in the DDT as provincial loss proportions, already mirrors the basin structure before any clustering is applied. Chapters 2 and 3 together constitute the geographical and empirical motivation that makes the clustering exercise meaningful rather than arbitrary.

Chapter 4 developed the theoretical tools needed to interpret the results actuarially. The Occurrence Exceedance Probability framework and the additivity of the aggregate exceedance frequency $\nu(l)$ provided a natural bridge between the clustering and the capital applications of Chapter 6: by decomposing portfolio tail risk into cluster-level contributions through the shares $\omega_k(T)$, the partition becomes directly readable in terms of regulatory capital, reinsurance structuring, and underwriting signals.

Chapter 5, the methodological core of the thesis, introduced and compared five clustering families. A particularly notable contribution is the use of energy distance-based evaluation through the DISCO F-ratio and normalised H-matrix, which assess cluster quality in terms of distributional differences rather than geometric compactness. These indices proved especially informative in a setting where clusters differ in shape and tail behaviour rather than in location alone, and their joint use alongside geometry-based indices such as the silhouette score provided a more complete picture of partition quality than either family alone. The results show that all algorithms consistently recover the Meuse–Scheldt divide, and that the risk hierarchy of the Belgian flood portfolio is robust across methods: Meuse events dominate the tail beyond the 150-year return period, the central Scheldt cluster carries intermediate tail risk, and the remaining events are actuarially negligible at high return periods.

Chapter 6 demonstrated that the cluster decomposition translates directly into concrete Solvency II and reinsurance applications. It is worth noting that the analytical framework of that chapter is entirely independent of the clustering methodology: any partition of the event catalogue, whether produced by one of the algorithms of Chapter 5 or even by an administrative zoning, plugs directly into the same exceedance frequency decomposition. In that sense, Chapter 6 could stand as an independent contribution, offering a general methodology for capital allocation, Cat XL treaty analysis, use test compliance, and climate stress testing for any partitioned flood portfolio.

The main finding is that the answer to the initial question is affirmative and robust. The Belgian flood risk landscape is sufficiently structured that the spatial footprint of events, rather than their administrative label, already encodes the information needed for meaningful risk zonation. A natural next step for validation of the methodology would be to test it with data from other countries, other catastrophe model outputs, or other natural catastrophe perils, and to explore additional distributional quantifications for clustering evaluation beyond those considered here.

Bibliography

- [1] J. Aitchison. A new approach to null correlations of proportions. *Mathematical Geology*, 13:175–189, 1981. doi: 10.1007/BF01031393.
- [2] J. Aitchison and J. J. Egozcue. Compositional data analysis: Where are we and where should we be heading? *Mathematical Geology*, 37:829–850, 2005. doi: 10.1007/s11004-005-7383-7.
- [3] David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In *ACM-SIAM Symposium on Discrete Algorithms*, 2007. URL <https://api.semanticscholar.org/CorpusID:1782131>.
- [4] Belgian Climate Portal. Précipitations, 2025. URL <https://climat.be/changements-climatiques/changements-observe/precipitations>. Accessed: 2026-05-29.
- [5] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.
- [6] Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27, 1974.
- [7] Gilles Celeux and Gérard Govaert. Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5):781–793, 1995. ISSN 0031-3203. doi: [https://doi.org/10.1016/0031-3203\(94\)00125-6](https://doi.org/10.1016/0031-3203(94)00125-6). URL <https://www.sciencedirect.com/science/article/pii/0031320394001256>.
- [8] ClimateChangePost. Flash floods and urban flooding – Belgium, 2020. URL <https://www.climatechangepost.com/belgium/flash-floods-and-urban-flooding/>. Accessed: May 2026.
- [9] Consorcio de Compensación de Seguros. Consorcio de compensación de seguros – inicio. Official website, 2024. URL <https://www.conorseguros.es/inicio>. Public entity under the Spanish Ministry of Economy, Trade and Business.
- [10] CRESTA. Cresta zones. <https://www.cresta.org>. Accessed: May 2026.

- [11] Crisis Centre Belgium. Surface water flooding. Belgian National Risk Assessment. URL <https://crisiscenter.be/en/risks-belgium/natural-risks/surface-water-flooding>. Accessed: May 2026.
- [12] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39(1):1–38, 1977.
- [13] European Commission, Directorate-General for Environment, and Royal HaskoningDHV. *Current practice in flood risk management in the European Union – September 2021*. Publications Office of the European Union, 2021. doi: doi/10.2779/235272.
- [14] European Insurance and Occupational Pensions Authority. Guidelines on the use of internal models. Technical Report EIOPA-BoS-14/180 EN, European Insurance and Occupational Pensions Authority, 2015. URL https://www.eiopa.europa.eu/publications/guidelines-use-internal-models_en.
- [15] European Insurance and Occupational Pensions Authority. Public statement on the monitoring exercise on the use of climate change scenarios in the ORSA. Technical Report EIOPA-BoS-25-094, European Insurance and Occupational Pensions Authority, 2025. URL <https://www.eiopa.europa.eu>.
- [16] European Parliament and Council. Directive 2009/138/EC on the taking-up and pursuit of the business of insurance and reinsurance (Solvency II). Official Journal of the European Union, L 335, 2009. URL <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:02009L0138-20190113>.
- [17] Ronald A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, London, 1925.
- [18] Gouvernement wallon. Bassins et sous-bassins hydrographiques en Wallonie. Connaître la Wallonie. URL <https://connaitrelawallonie.wallonie.be/fr/histoire/atlas/bassins-et-sous-bassins-hydrographiques-en-wallonie>. Accessed: May 2026.
- [19] Michael Greenacre. Compositional data analysis. *Annual Review of Statistics and Its Application*, 8:271–299, 2021. doi: 10.1146/annurev-statistics-042720-124436.
- [20] Patricia Grossi and Howard Kunreuther, editors. *Catastrophe Modeling: A New Approach to Managing Risk*. Huebner International Series on Risk, Insurance and Economic Security. Springer, New York, NY, 2005. ISBN 978-0-387-23082-5. doi: 10.1007/b100669.
- [21] J. A. Hartigan and M. A. Wong. Algorithm AS 136: A k-means clustering algorithm. *Applied Statistics*, 28(1):100–108, 1979.
- [22] Natalia Humphreys. Exceedance probability in catastrophe modeling. *CAS E-Forum*, Winter 2021:1–61, 2021.
- [23] IMMERSE Project. The scheldt estuary. Interreg VB North Sea Region Programme. URL <https://northsearegion.eu/immerse/project-estuaries/the-scheldt-estuary/index.html>. Accessed: 29 May 2026.

- [24] Jacquet, Philippe, Raymond, Mireille, Martin, Olivier, and Rouas, Guy. Modélisation des crues de la meuse. *La Houille Blanche*, (6):94–103, 2003. doi: 10.1051/lhb/2003118. URL <https://doi.org/10.1051/lhb/2003118>.
- [25] Matthew Jones, Kirsten Mitchell-Wallace, Matthew Foote, and John Hillier. *Fundamentals*, chapter 1, pages 1–46. John Wiley & Sons, Ltd, 2017. ISBN 9781118906057. doi: 10.1002/9781118906057.ch1. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118906057.ch1>.
- [26] Michel Journée, Edouard Goudenhoofdt, Stéphane Vannitsem, and Laurent Delobbe. Quantitative rainfall analysis of the 2021 mid-July flood event in Belgium. *Hydrology and Earth System Sciences*, 27:3169–3189, 2023. doi: 10.5194/hess-27-3169-2023.
- [27] Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York, 1990. doi: 10.1002/9780470316801.
- [28] Stuart A. Klugman, Harry H. Panjer, and Gordon E. Willmot. *Loss Models: From Data to Decisions*. John Wiley & Sons, Hoboken, NJ, 3 edition, 2008. ISBN 9781118315323.
- [29] Godfrey N. Lance and William T. Williams. A general theory of classificatory sorting strategies. I. Hierarchical systems. *The Computer Journal*, 9(4):373–380, 1967. doi: 10.1093/comjnl/9.4.373.
- [30] Songzi Li and Maria L. Rizzo. K-groups: A generalization of k-means clustering. 2017. doi: 10.48550/arXiv.1711.04359.
- [31] V. Masson-Delmotte, P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, Ö. Yelekçi, R. Yu, and B. Zhou, editors. *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, 2021. doi: 10.1017/9781009157896.
- [32] Geoffrey McLachlan and David Peel. *Finite Mixture Models*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., September 2000. ISBN 9780471006268. doi: 10.1002/0471721182.
- [33] Dietmar Pfeifer. Insurance risk management for catastrophic events. Lecture notes, Carl von Ossietzky Universität Oldenburg / AON, 2005.
- [34] Maria L. Rizzo and Gábor J. Székely. DISCO analysis: A nonparametric extension of analysis of variance. *The Annals of Applied Statistics*, 4(2):1034–1055, 2010. doi: 10.1214/09-AOAS245.
- [35] Maria L. Rizzo and Gábor J. Székely. Energy distance. *WIREs Computational Statistics*, 8(1):27–38, 2016. doi: 10.1002/wics.1375.
- [36] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

- [37] C. C. Sampson, T. J. Fewtrell, F. O’Loughlin, F. Pappenberger, P. B. Bates, J. E. Freer, and H. L. Cloke. The impact of uncertain precipitation data on insurance loss estimates using a flood catastrophe model. *Hydrology and Earth System Sciences*, 18(6): 2305–2324, 2014. doi: 10.5194/hess-18-2305-2014. URL <https://hess.copernicus.org/articles/18/2305/2014/>.
- [38] Erich Schubert and Peter J. Rousseeuw. Faster k-medoids clustering: Improving the PAM, CLARA, and CLARANS algorithms. In *Similarity Search and Applications (SISAP)*, volume 11807 of *Lecture Notes in Computer Science*, pages 171–187. Springer, 2019. doi: 10.1007/978-3-030-32047-8_16.
- [39] SEGEFA. Belgique – rivières et bassins hydrographiques. Département de Géographie, Université de Liège, 2008. URL http://www.enseignement.be/index.php?page=23827&do_id=3161. Réalisée à la demande du Service général du Pilotage du Système éducatif. Republié sur dailyscience.be. Accessed: May 2026.
- [40] Gábor J. Székely and Maria L. Rizzo. Hierarchical clustering via joint between-within distances: Extending Ward’s minimum variance method. *Journal of Classification*, 22(2):151–183, 2005. doi: 10.1007/s00357-005-0012-9.
- [41] Task Committee on Hydrology Handbook of Management Group D of the American Society of Civil Engineers. *Hydrology Handbook*. Number 28 in ASCE Manuals and Reports on Engineering Practice. American Society of Civil Engineers, New York, 2nd edition, 1996. ISBN 0-7844-0138-1.
- [42] Vlaamse overheid. River basin management plan Scheldt and Meuse 2022–2027: Non-technical summary. Integraal Waterbeleid, 2022. URL <https://sgbp.integraalwaterbeleid.be/beheerplan/nts-engels-rbmp3-web.pdf>. Adopted by the Flemish Government on 1 July 2022.
- [43] Jr. Ward, Joe H. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963. doi: 10.1080/01621459.1963.10500845.
- [44] Eric W. Weisstein. Beta distribution. MathWorld – A Wolfram Resource, . URL <https://mathworld.wolfram.com/BetaDistribution.html>. Accessed: May 2026.
- [45] Eric W. Weisstein. Regularized beta function. MathWorld – A Wolfram Resource, . URL <https://mathworld.wolfram.com/RegularizedBetaFunction.html>. Accessed: May 2026.
- [46] Wikipedia. Yser. Wikipedia, The Free Encyclopedia, 2025. URL <https://en.wikipedia.org/wiki/Yser>. Accessed: May 2026.
- [47] Wikipédia. Yser. Wikipédia, l’encyclopédie libre, 2025. URL <https://fr.wikipedia.org/wiki/Yser>. Accessed: May 2026.

