European Actuarial Journal (2023) 13:755–774
https://doi.org/10.1007/s13385-022-00338-w

ORIGINAL RESEARCH PAPER

A resimulation framework for event loss tables based on clustering

Benedikt Funke[1] · **Harmen Roering**[2]

Received: 24 May 2022 / Revised: 16 September 2022 / Accepted: 20 November 2022 /
Published online: 26 December 2022

Abstract

Catastrophe loss modeling has enormous relevance for various insurance companies due to the huge loss potential. In practice, geophysical-meteorological models are widely used to model these risks. These models are based on the simulation of meteorological and physical parameters that cause natural events and evaluate the corresponding effects on the insured exposure of a certain company. Due to their complexity, these models are often operated by external providers—at least seen from the perspective of a variety of insurance companies. The outputs of these models can be made available, for example, in the form of event loss tables, which contain different statistical characteristics of the simulated events and their caused losses relative to the exposure. The integration of these outputs into the internal risk model framework is fundamental for a consistent treatment of risks within the companies. The main subject of this work is the formulation of a performant resimulation algorithm of given event loss tables, which can be used for this integration task. The newly stated algorithm is based on cluster analysis techniques and represents a time-efficient way to perform sensitivities and scenario analyses.

Keywords Resimulation · Event loss tables · Natural catastrophe models · Clustering

✉ Benedikt Funke
benedikt.funke@th-koeln.de

Harmen Roering
harmen.roering@oliverwyman.com

[1] Institute for Insurance Studies, TH Köln-University of Applied Sciences, Gustav-Heinemann-Ufer 54, Cologne 50968, Germany

[2] Oliver Wyman BV, Strawinskylaan 4101, 1077 ZX Amsterdam, The Netherlands

1 Introduction

This article deals with a numerically efficient way to resimulate event loss tables (ELTs), which contain simulated losses due to natural catastrophe events and for which the resimulation forms the basis to incorporate them into internal risk models of insurance companies. Specifically, a natural catastrophe is an accumulation event, namely a single extreme event that causes an enormous number of different small or moderate losses that have—once summed up—a huge loss potential. Within the insurance industry, the individual losses affect several different lines of business and are often together in the billions of euros. For an extensive description of this topic, we recommend [8] which provides a wide-ranging and excellent overview of the subject of natural catastrophe modeling in general.

Basically, there are two different approaches for modeling these losses within insurance companies, e.g. see [3, 4]. One class of models are the so-called mathematical-statistical models, which are built on historical loss data (both internal and external). These models are based on the idea of describing the loss caused by natural catastrophes using a collective model. In the course of this, probability distributions are fitted to observed loss frequencies and loss severities using historical and indexed data. A key drawback of this method is its dependence on a suitable data base. Catastrophe risk, by definition, includes events that have caused enormous losses. Often, a multi-year time series of data, usually a maximum of 50 years, is available to be used for estimation. However, the events that are in the tail, such as those with returning periods 200, 1000 or even 10,000 years, cannot be estimated adequately in this way without extrapolation. Even with a hypothetical longer time series, the necessary indexing, that takes changes in the geographical distribution, the number of risks, the sum insured or portfolio shifts into account, is a major challenge, so in general the historical losses would not be useful for current modeling.

In view of this, the second class of models referred to as geophysical-meteorological or exposure-based models is more often used on the market and is the main focus of this article. The outputs of these models are usually supplied to insurance companies by external providers. In order to derive these outputs, granular portfolio data is made available to the service provider that subsequently simulated the losses relative to this portfolio resulting from different synthetic and historic natural events. The resulting outputs in the form of ELTs can be integrated into the company's own risk models by means of resimulation. In practice, problems arise in the resimulation of ELTs, such as that of a long runtime or an increased simulation error. In view of this, the main purpose of our article is to present a clustering algorithm that searches for similarities in these outputs and acts then as a much more performant resimulation tool and thus allows a faster evaluation of sensitivities or scenario analyses, e.g. in the context of the Own Risk and Solvency Assessment (ORSA) process under Solvency II.

The rest of our work is organized in such a way that we first describe the general procedure of resimulation of vendor outputs in Sect. 2. After that, we will present our enhanced clustering approach for more performant resimulation in Sect. 3. In Sect. 4 we apply our algorithm to a set of ELTs representing different perils and compare the outputs to the traditional resimulation approach in order to demonstrate the practical

relevance of our contribution. We conclude with a brief discussion and an outlook on the handling of multi-portfolio ELTs.

2 Resimulating event loss tables: the traditional approach

2.1 Structure of an event loss table

In this subsection we want to introduce the structure of an ELT, which is needed for our purposes. In general, the actual structure of the outputs of exposure-based models depends on the choice of the external provider. In particular, we will address outputs that contain parameters that can be used to mathematically model the listed loss events. The rows of an ELT represent stochastically independent scenarios, each of which can be understood as a separate independent collective model.

One example ELT can be found in Table 1. For each row of this table we can read parameters for the average number of events per year, as well as information about the expected value and the variability of the row-wise scenario losses.

Specifically, each event is uniquely coded by an EVENTID. The RATE-column indicates the probability of occurrence of the specific event within a year. The PERSPVALUE-column ("Perspective Value") contains the expected value of the loss severity distribution when the event occurs. The columns STDDEVI and STDDEVC contain information about the standard deviation of the loss level at event occurrence. A distinction is made between the risk-specific standard deviation STDDEVI and the standard deviation across affected risks STDDEVC. The total standard deviation per row is calculated as the sum of these two variables.

An ELT contains information about two different sources of uncertainty. The uncertainty regarding the occurrence of the event is coded via the RATE-column and is also referred to as primary uncertainty. On the other hand, the uncertainty in the amount of loss caused is parameterized by the two standard deviation parameters STDDEVI and STDDEVC and is accordingly referred to as secondary uncertainty.

The EXPVALUE-column ("Exposed Value") indicates the insured values affected by the respective event (e.g. the total sum insured or the sum of the PMLs of the insured risks) and thus provides an upper limit for the maximum event loss.

Table 1 Example of an event loss table

EVENTID	RATE	PERSPVALUE	STDDEVI	STDDEVC	EXPVALUE
4180731	0.000432996	223,870	30,170	164,421	29,901,935,239
4180732	0.000321448	5,236,225	85,976	3,665,470	1,922,520,000
4180734	0.0034441	15,665	5258	12,238	1,820,097,738
4180737	0.0000202	428,075	18,632	356,513	34,006,033,640
4180738	0.000329058	72,138	15,865	50,591	2,813,630,546
4180739	0.001621555	3517	2353	2632	203,059,682

In general, external models are independent modeling platforms. Independent here means that they are independent of the company-specific internal risk modelling framework. Hence, the provided outputs of these models must therefore first be suitably integrated into the simulation engine of the internal model of the company. The required embedding of these results is done by resimulating the catastrophic losses per peril and by relying on different distributional assumptions on the event numbers and severities. The losses resimulated in this way can be embedded in the runs of the internal model across different risks, portfolios and local business units. The choice of a unique seed is an important factor here, as this is the only way to ensure reproducibility and guarantee consistency in the way randomness is incorporated.

In the following subsection, we will explain how the resimulation of catastrophe losses from ELTs can be implemented in practice and which modeling assumptions underlie it in our case.

2.2 Resimulation from ELTs

We assume below that each row i, $1 \leq i \leq n$, of an ELT represents a collective model S_i

$$S_i := \sum_{j=1}^{N_i} X_{ij},$$

whereby N_i is a random variable that models the claim numbers and is assumed to have a Poisson distribution with parameter λ_i. This parameter coincides with the corresponding RATE in row i. Moreover, the severity of the j-th claim of collective model i is modelled by a random variable X_{ij} with mean and standard deviation that are denoted by

$$E[X_{ij}] := \mu_i \text{ and } \sigma[X_{ij}] := \sigma_i.$$

We assume that X_{ij}, $1 \leq j \leq N_i$, are independent and identically distributed random variables. Moreover, μ_i equals the perspective value PERSPVALUE_i and σ_i is the sum of the independent and the correlated standard deviations ($\sigma_i = \text{STDDEVI}_i + \text{STDDEVC}_i$).

We assume that different collective models S_i, $1 \leq i \leq n$, are also all stochastically independent. In addition, for each row i, we assume that the claim number N_i and the claim severities X_{ij} are stochastically independent, too. Even if this assumption might be violated in practice—for a practical example of this dependency, we refer to [2] which treats this dependence for flood scenarios—it is a common assumption when dealing with collective models to assume independence here. In particular, the amount of the losses often increases with the number of claims in practice. The reason for this is due to the fact that claims settlement is often more of a lump-sum type when there are a large number of claims within an event.

Another key assumption concerns the distribution of the degree of loss

$$z_{ij} := X_{ij}/E_i,$$

whereby E_i is the exposed value EXPVALUE$_i$ of collective model i. Using this definition we note that the mean and the standard deviation of the degree of loss can be represented according to

$$E[z_{ij}] = \mu_i/E_i := \bar{\mu_i} \text{ and } \sigma[z_{ij}] = \sigma_i/E_i := \bar{\sigma_i}.$$

Due to the fact that $z_{ij} \in [0, 1]$, it is a common practice to choose the Beta distribution with parameters α_i and β_i for z_{ij} (see [3, 4]).

For the resimulation of the total annual loss, we now proceed as follows. First, for simulation path (M) and each row i the number of losses N_i is simulated by a draw of a corresponding Poisson distributed random number $\tilde{n}_i^{(M)}$. This random number indicates how many events of the collective model i occurred in path (M). For each event that occurred, we now simulate a Beta-distributed random number $\tilde{z}_{ij}^{(M)}$ under the use of the parameters α_i and β_i. Since these parameters are unknown, we estimate them canonically by using method of moments-type estimators as follows:

$$\alpha_i = \left[\frac{\bar{\mu_i} \cdot (1 - \bar{\mu_i})}{\bar{\sigma_i}^2} - 1\right] \cdot \bar{\mu_i} \text{ and } \beta_i = \left[\frac{\bar{\mu_i} \cdot (1 - \bar{\mu_i})}{\bar{\sigma_i}^2} - 1\right] \cdot (1 - \bar{\mu_i}).$$

Subsequently, the loss degrees generated in this way are multiplied by the respective exposure values E_i to yield the simulated losses $\tilde{x}_{ij}^{(M)} = \tilde{z}_{ij}^{(M)} \cdot E_i$ for collective model i and simulation path (M).

Finally, in order to calculate the Aggregate Exceedance Probability (AEP) curve, the simulated losses are summed up across all collective models and result in the total annual loss for this single simulation path (M)

$$S^{(M)} := \sum_{i=1}^{n} \sum_{j=1}^{\tilde{n}_i^{(M)}} \tilde{x}_{ij}^{(M)}.$$

The Occurrence Exceedance Probability (OEP) curve in simulation path (M) is analogously determined by using the maximum simulated event over all rows $1 \leq i \leq n$.

In the next section, we will now present how the problems that occur within the context of resimulating ELTs, such as that of a long runtime as well as that of an increased simulation error, can be tackled by a clustering-based approach. Moreover, this section contains our main result, which builds up on the facts introduced so far.

3 Clustering-based resimulation

3.1 Introduction to a clustering-based approach

A possible solution for the runtime problems and indirectly for the increased simulation error is to perform the resimulation of an ELT through a clustered approach. One basic assumption here is formed by the independence of the rows of the ELT. In particular, we assume that the number of events across different scenarios $i_1, \ldots, i_k$ are stochastically independent. This results in the fact that for distinct scenarios $i_1, \ldots, i_k$

- the sum of the number of events follows again a Poisson distribution with an accumulated frequency parameter $\sum_{j=1}^{k} \lambda_{i_j}$ due to the convolution property of the Poisson distribution and
- the conditional distribution $\left(N_{i_1}, \ldots, N_{i_k}\right) \mid \left(\sum_{j=1}^{k} N_{i_j} = n\right)$ fulfills

$$\left(N_{i_1}, \ldots, N_{i_k}\right) \Bigg| \left(\sum_{j=1}^{k} N_{i_j} = n\right) \sim Multinom\left(k, \pi_k\right),$$

whereby

$$\pi_k := \left(\pi_{i_1}, \ldots, \pi_{i_k}\right)$$

and

$$\pi_{i_l} := \frac{\lambda_{i_l}}{\sum_{j=1}^{k} \lambda_{i_j}},\ l = 1, \ldots, k.$$

The latter relation represents the fact that by simulating from the aggregated scenarios, we do not lose any information about the occurrence of the individual scenarios.

Now because the events occur independent from one another and because the sum of events from different rows follows a Poisson distribution with an accumulated frequency, we can state that (re-)simulating scenarios of rows $i_1, \ldots, i_k$ can be done with one single collective model as long as the severity distributions are equal.

To keep things in perspective let us now specifically look at two distinct rows i and i' such that the degrees of loss and the exposed values fulfill

$$z_{ij}, z_{i'j} \sim \mathcal{B}(\alpha_i, \beta_i) \overset{\mathcal{D}}{=} \mathcal{B}(\alpha_{i'}, \beta_{i'}) \text{ and } E_i = E_{i'},$$

then resimulation in path (M) of losses of both events could be done by summing up the corresponding frequencies (λ_i and $\lambda_{i'}$) and using the same severity distribution for both as in

$$S_i^{(M)} + S_{i'}^{(M)} := \sum_{j=1}^{\tilde{n}_i^{(M)}} \tilde{x}_{ij}^{(M)} + \sum_{j=1}^{\tilde{n}_{i'}^{(M)}} \tilde{x}_{i'j}^{(M)} = \sum_{j=1}^{\tilde{n}_i^{(M)} + \tilde{n}_{i'}^{(M)}} \tilde{x}_{ij}^{(M)}.$$

Therefore, finding severity distributions among the rows in the ELT that are (approximately) equal allows for a reduction in the amount of collective models used in the resimulation. This would reduce the runtime as the total amount of simulations needed for obtaining stable results of the OEP- and AEP-curve decrease. Note that through the second Poisson property given above, it is still possible to simulate the losses per individual row after using the clustered resimulation framework.

3.2 Clustering the ELTs

The proposed solution to resimulate by using a clustering framework requires an algorithm that combines scenarios that have nearly equal severity distributions. How exactly "nearly equal" is defined in our case is explained in more detail below.

Finding such nearly identical pairs or groups of severity distributions in the ELT is often performed by an unsupervised learning technique. However, most conventional unsupervised learning methods are focused on finding a (preferably) small number of clusters to represent the data. In our case, we are not interested in finding the smallest number of clusters to comprehensibly represent the data. We rather want to find a number of clusters that minimizes the runtime while preserving the original data characteristics. More specifically, we require the within-cluster-distances to be small to ensure the clustered framework being a good representation of the original resimulation framework. In order to reach this aim, a careful treatment of outliers is necessary.

The first challenge when considering a clustering algorithm is to define the distance between severity distributions. To define distances in our case, we propose to directly use the parameters of the severity distribution (α_i, β_i and E_i). However, we note that these parameters—in particular the different exposed values E_i—can become quite large. At the same time, it is expected that the shape parameters α_i and β_i of the Beta distributions would not become that large. Therefore, we encounter a scale issue when using distance measures such as the Euclidean norm. To prevent this, we will standardize all parameters in order to make them comparable.

To create our clusters, we propose an algorithm somewhat related to DBSCAN ("Density-Based Spatial Clustering of Applications with Noise") introduced by [5] and also similar to the contribution in [1]. The actual DBSCAN algorithm is widely used in the context of cluster analysis. Although there are very different fields of application, it has not been used in its original form, nor have variants of this algorithm been applied in the context of actuarial problems to the best of our knowledge.

In general, the underlying idea is to examine neighborhoods $N_\varepsilon(D_i)$ around each single data object D_i. In our case, the data objects are the points in the three-dimensional space, whose coordinates are determined by the scaled parameters of the Beta distribution and the scaled exposed value of the individual scenarios i. In accordance to that we define the ε-neighborhood

$$N_\varepsilon(D_i) := \left\{ y \in \bigcup_{j=1}^{n} D_j \,\middle|\, 0 \leq ||y - D_i||_2 \leq \varepsilon \right\},$$

where $\varepsilon > 0$, $||\cdot||_2$ denotes the three-dimensional Euclidean norm and $D_i := (\alpha_{i,\text{scaled}}, \beta_{i,\text{scaled}}, E_{i,\text{scaled}})'$ consists of the scaled parameters of scenario i. We note that we restrict the neighborhood $N_\varepsilon(D_i)$ only to three-dimensional data points that belong to the scenarios of the underlying ELT, as these points act as our objectives.

Our algorithm clusters the data objects by first determining the cardinality of $N_\varepsilon(D_i)$ for each different data object D_i and then arrange $N_\varepsilon(D_i)$ in descending order with respect to this quantity. The neighborhood $N_{\varepsilon,\max}(D_{i'})$, into which the most points fall, acts as the first cluster. Hence, all data objects that belong to this neighborhood including the center $D_{i'}$ fall into this first cluster. The clustered severity distribution is then given by the average of the original parameters within $N_{\varepsilon,\max}(D_{i'})$ while the frequency parameter is given by the sum of all frequencies of the data objects within this neighborhood. These parameters act as representatives of the first cluster. We then remove all data objects that are clustered and repeat this process until no cluster can be created anymore.

As this algorithm requires the computation of distances between the data objects, we speed up the algorithm by splitting the data into disjoint subsets that together contain all the data. Using a proper splitting strategy prevents redundant distance calculations of data objects that are far apart in our parameter space. One possible splitting strategy is to divide the parameter space based on the quantiles of each univariate parameter distribution.

As an illustrative example, we will consider our splitting strategy for a data set that contains at least five different data points per dimension. If we split the α, β and E univariate parameter spaces based on the corresponding parameter quantiles in five disjoint parts for each dimension, then this particular subdivision leads to a total of $5^3 = 125$ subsets. The union of these disjoint subsets covers all data objects, even though some of these sets may be empty. The previous statement can be understood by the fact that all α-values in a data set must fall between the minimum and maximum α of the data set—the same reasoning applies to the β and the exposure values.

In particular, the first subset in this example is given by

$$T_1 = \left\{ y \in \bigcup_{j=1}^{n} D_j | \alpha_{min} \leq \alpha_y < \alpha_{20\%}, \beta_{min} \leq \beta_y < \beta_{20\%}, E_{min} \leq E_y < E_{20\%} \right\},$$

where

$$y := (\alpha_y, \beta_y, E_y)' := (\alpha_{y,\text{scaled}}, \beta_{y,\text{scaled}}, E_{y,\text{scaled}})'$$

and $\alpha_{x\%}, \beta_{x\%}, E_{x\%}$ denote the $x\%-$quantiles of the three scaled parameters.

The second subset T_2 would be obtained by considering the next disjoint part for one of the parameters, as in

$$T_2 = \left\{ y \in \bigcup_{j=1}^{n} D_j \middle| \alpha_{20\%} \leq \alpha_y < \alpha_{40\%}, \beta_{min} \leq \beta_y < \beta_{20\%}, E_{min} \leq E_y < E_{20\%} \right\}.$$

The, say, last of the sets constructed in this way is then given by

$$T_{125} = \left\{ y \in \bigcup_{j=1}^{n} D_j \,\middle|\, \alpha_{80\%} \leq \alpha_y < \alpha_{100\%} + 1, \beta_{80\%} \leq \beta_y < \beta_{100\%} + 1,\ E_{80\%} \leq E_y < E_{100\%} + 1 \right\}.$$

The indexing of the sets does not play a role and has only illustrative reasons. We point out here that whenever a subset imposes a condition on the maximum of the values of α, β or E, we formally increase the upper bound by 1 in order to be able to assign all points accordingly.

Algorithm 1 shows the pseudo-code of our clustering resimulation framework.

Algorithm 1 ELT clustering algorithm

```
C ← ∅
D ← {D_1, D_2, ..., D_n}
Define T ← {T_1, T_2, ..., T_t : for each h and g, h ≠ g, T_h ∩ T_g = ∅ and Σ_{j=1}^{t} T_j = D} based on the parameter values
for j = 1 : t do
    D^(j) = D ∩ T_j
    while |D^(j)| ≠ 0 do
        ∀ D_i ∈ D^(j), compute |N_ε(D_i^(j))|
        Find k = argmax_{D_i ∈ D^(j)} |N_ε(D_i^(j))|
        Label as cluster N_ε(D_k^(j))
        C ← C + N_ε(D_k^(j))
        D^(j) ← D^(j) − C
    end while
end for
```

Our algorithm requires two tuning parameters, ε and the choice of disjoint sets which is given by a set

$$T = \left\{ T_1, T_2, ..., T_t : \text{for each } h \text{ and } g,\ h \neq g, T_h \cap T_g = \emptyset, \sum_{j=1}^{t} T_j = D \right\}$$

where $D = \bigcup_{j=1}^{n} D_j$. The tuning parameter ε controls to what extent different severity distributions are grouped together within a cluster. Therefore, an ε-value of (nearly) 0 would represent the original resimulation framework without clustering whereas a sufficiently large ε would result in a resimulation framework based on a single large event. To find the optimal value of ε, a similar approach as in a K-means algorithm is applied by using an elbow plot that displays the sum of within-cluster-distances as a function of the number of clustered data points. The use of this plot to select a suitable ε-value is also taken in the pioneering work [5] in which DBSCAN was first

introduced. The procedure proposed in [5] is to choose the value for ε at which this graph has an "elbow", i.e. a kink (see Fig. 4).

An ε of 0 should return a dissimilarity of 0 while a sufficiently large ε would return the sum of all distances. We suggest that all ε-values on the more conservative side (hence, less dissimilarity within the clusters) are eligible for the resimulation and therefore, we consider the knot in the elbow plot as the maximum suitable number for ε.

For the choice of disjoint sets in T, we consider the trade-off between the decrease in the cluster efficiency and the increase in the efficiency of the clustering algorithm.

3.3 Testing the clustered resimulation output

In order to evaluate the performance of our proposed clustered resimulation framework, we must obtain significantly similar outputs of both (clustered and unclustered) resimulation frameworks. To test this, we simulated 10,000 simulations paths for a 100 times with both the original and the clustered resimulation framework. We then test for each run with regards to the original resimulation framework $R^{(i)}_{\text{original}}$, whether this simulated run $R^{(i)}_{\text{original}}$ significantly differs in its distribution compared to the previous simulated run $R^{(i-1)}_{\text{original}}$ using a paired Kolmogorov-Smirnov (KS) test under a 10%-significance level. In comparison, we use the same test statistic and significance level to test whether the distribution of the run $R^{(i)}_{\text{clustered}}$ significantly differs from the distribution of $R^{(i-1)}_{\text{original}}$. The idea is to test to what extent the distribution of $R^{(i)}_{\text{clustered}}$ mimics that of $R^{(i)}_{\text{original}}$, since this is the main goal of a proper clustering algorithm in our case. After 100 resimulation runs, we compare the number of rejections of both frameworks.

4 A comparison between both frameworks

In this section, we will now compare the results of the two resimulation frameworks. For this purpose, we use a data set consisting of four different and fictitious ELTs and covering the perils windstorm, flood and earthquake. For the windstorm hazard, two different ELTs are available.

4.1 Discussion on the characteristics of the data set used

The most important descriptive information of the data set we used can be found in Table 2. In this table, S denotes the total loss of the complete ELT for each individual peril. Each ELT in turn represents a collective model, so that the mean and the variance across all scenarios can be calculated using Wald's equations as well as Bienayme's identity due to the independence of the individual scenarios and the independence between the claim number and severity distributions in each scenario i.

Table 2 Descriptive statistics of the present ELTs including secondary uncertainty

Peril	Scenarios	Sum of frequencies	$E[S]$	$Std(S)$	$CoV(S)$ (%)
Windstorm 1	21,010	41.66	37,797,65	64,889,212	171.7
Windstorm 2	17,747	38.07	5,647,010	9,192,134	162.8
Flood	7045	2.19	10,952,003	90,486,752	826.2
Earthquake	4803	0.05	442,074	12,785,084	2892.1

Table 3 Descriptive statistics of the present ELTs excluding secondary uncertainty

Peril	$Std(S)$	$CoV(S)$ (%)
Windstorm 1	61,342,690	162.3
Windstorm 2	8,386,852	148.5
Flood	84,912,569	775.3
Earthquake	11,019,135	2492.6

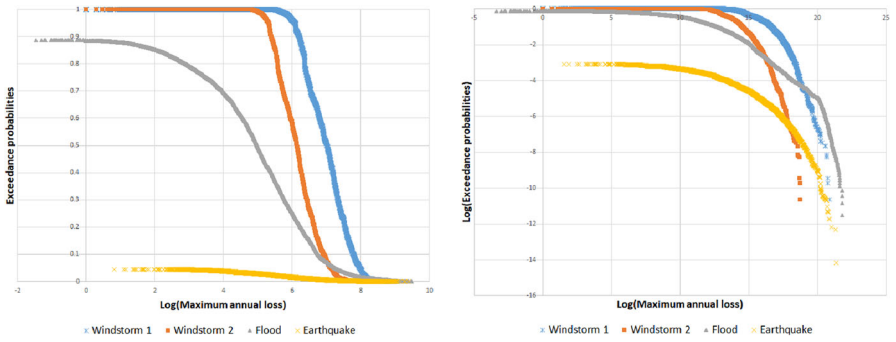


Fig. 1 Logarithmized annual maximum losses and the associated exceedance probabilities without taking the secondary uncertainty into account

We remark that the number of scenarios in Table 2 varies widely and that the distribution for the earthquake hazard has an enormously high coefficient of variation. The number of expected events also varies greatly, so that for the earthquake hazard, one expects only a single occurrence of a corresponding event every 200 years from the present ELT scenario catalogue. How to deal in general with estimating the loss frequency of such extremely rare risks in practice, is presented in the current contribution [6].

The values for the standard deviation of the total loss shown in Table 2 also take into account the secondary uncertainty. For the sake of completeness, it should be mentioned that the coefficients of variation of the data sets change without taking into account the secondary uncertainty as listed in Table 3. The only source of uncertainty here is the probability of occurrence as measured by the rates. The loss severities are deterministic and defined by the expected values of the loss severity distributions (PERSPVALUE$_i$).

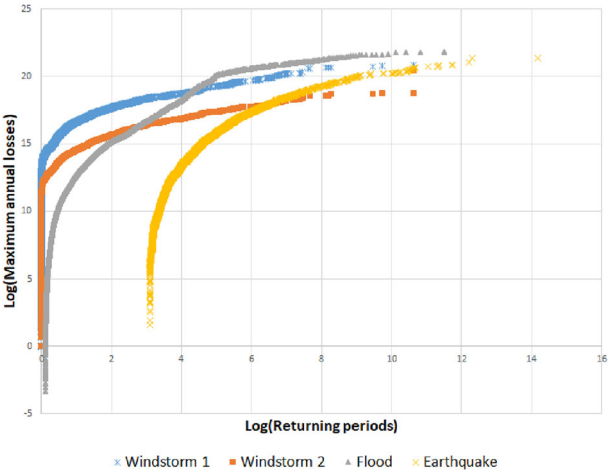


Fig. 2 OEP-curves of the four perils studied without considering the secondary uncertainty

In Fig. 1, the logarithmized annual maximum losses are displayed against the corresponding exceedance probabilities. For this representation as well as for the representation of the corresponding OEP-curves in Fig. 2, we have not considered the secondary uncertainty.

The reason for not considering the secondary uncertainty is based on the fact that in this case a closed analytical representation of the OEP-curve is possible, which significantly affects and simplifies the representation, see [7].

Figure 1 clearly shows the heaviness of the tails of the peril distributions. The two windstorm distributions are the ones with the lightest tail as expected and in contrast, the distribution of earthquake events is extremely heavy-tailed. The probability that no earthquake event is realized in this case is about 95%. For better comparability, the logarithmized exceedance probabilities can be taken from the right graph of Fig. 1.

Figure 2 shows the OEP-curves of the four ELT distributions considered. For both axes a logarithmic scaling was chosen to achieve comparability. The impressions of Fig. 1 are reinforced by Fig. 2. We conclude that the earthquake distribution has the largest return periods within the tail, but the highest total loss with even a smaller return period is shown by the flood ELT. The comparatively lighter tails of the two windstorm ELTs can also be identified.

4.2 Hyperparameter tuning

Our first step is to select the hyperparameters ε and T. For T, we created $6^3 = 216$ sets by dividing each of the three univariate parameter spaces in six disjoint subspaces based on their quantile values. The choice of T was based on the computational efficiency.

To illustrate the effect that the number of clusters falls when ε grows, we have plotted in Fig. 3 the relative shrinkage-proportions of the number of scenarios against the respective ε-values. Here, the monotonic behavior can be found for all hazards. We

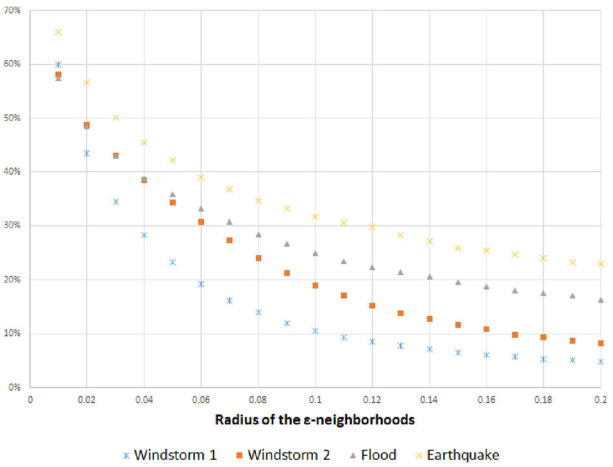


Fig. 3 Shrinkage-proportions of the number of scenarios compared to the unclustered ELT

further note that due to the extreme skewness of the earthquake distribution, the relative proportions of the scenario shrinkages are the largest. This is not very surprising, since we want to keep the characteristics of the original distribution and many of the extreme scenarios are classified as outliers.

In order to select ε, we have to identify the knots in the corresponding elbow plots. In Fig. 4, we show the elbow plots corresponding to the four different ELT data sets. In an elbow plot, the number of clusters are plotted against the corresponding summed discrepancy measures. The summed discrepancy for a given cluster is defined as the sum of the Euclidean distances of the clustered scenarios to the representative of this cluster. These distances are then summed over all clusters and result in the discrepancy measures shown for ε-values between 0 and 0.20.

As can be seen from the plots, the ELT data sets windstorm 2 and flood seem to have a knot around the ε-value of 0.01 and around the ε-value 0.10. In contrast, the ELT data sets windstorm 1 and earthquake seem to have a single knot around the ε-value 0.05. To test our methodology, we will pick the values of 0.05 and 0.10 for ε, as these might be considered as knot values.

4.3 Comparing the simulated losses

Most important for the clustered resimulation framework is that the output is comparable to the original resimulation framework. To compare the outputs of both resimulation frameworks, we first present the descriptive statistics of the clustered data for ε-values of 0.05 and 0.1. The corresponding figures for the base case can be found in Table 2. Subsequently, we use the scheme presented in Sect. 3.3 and a Q-Q plot analysis based on the average quantile estimates and their confidence intervals after 100 resimulations with a resimulation path of 10,000.

Table 4 shows the main descriptive results of the total losses of the entire ELTs of the four considered perils. The enormous reduction in the number of scenarios is striking.

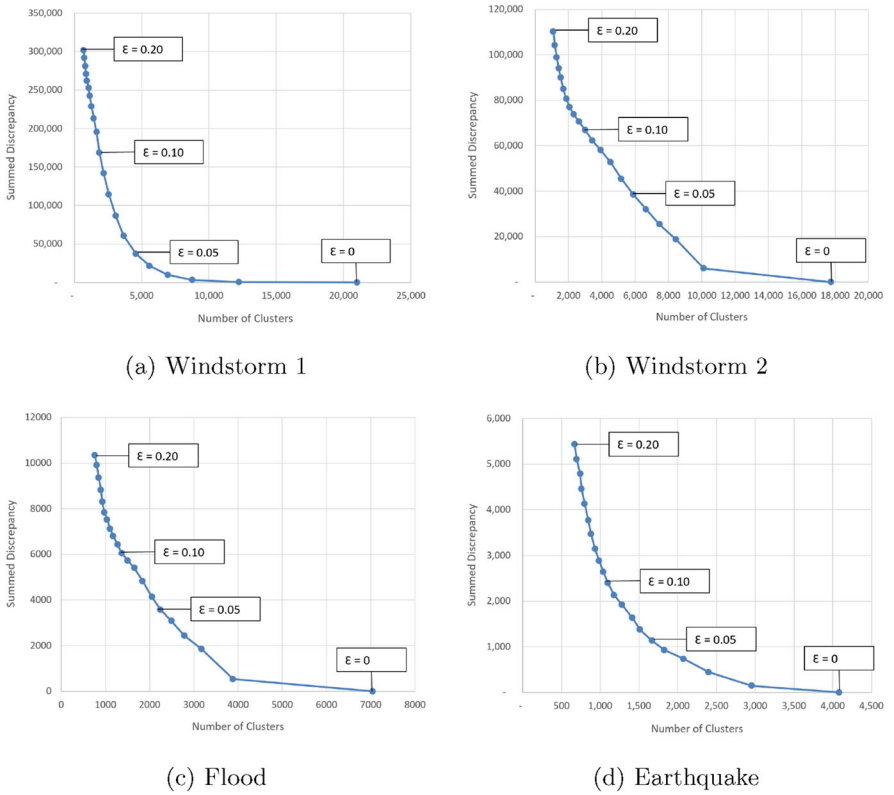


Fig. 4 Elbow plots of the amount of ELT clusters against a dissimilarity level for ε-values ranging from 0 up to 0.20

The sum of the frequencies is not reduced by our clustering algorithm, since per cluster the frequencies of the individual clustered scenarios are summed up. It can be derived that the differences in the expected total loss and the standard deviation of the total loss are very small, which indicates that the characteristics of the original distribution are approximately maintained. It is noticeable that the coefficient of variation is reduced by our clustering algorithm, which is in accordance with the reduction of the number of scenarios and the smoothing of the distribution parameters.

Table 5 shows the results of our testing framework for both ε-values of 0.05 and 0.10. For an ε-value of 0.05, we can see that the rejection rate is for both frameworks around 10% of the amount of resimulations with the exception of the earthquake ELT. However, the amount of rejections for both resimulation frameworks is the same for the earthquake ELT such that the high rejection rate is more due to the variation in the simulation than due to the clustering framework.

In the case of an ε-value of 0.10, we can see that the rejection rates for the clustered ELT starts to behave differently compared to what is expected. For the windstorm 1 data set, the rejection rate is much lower than that of the original data set. Moreover, the ELT data sets windstorm 2 and flood have notably different rejection rates, too.

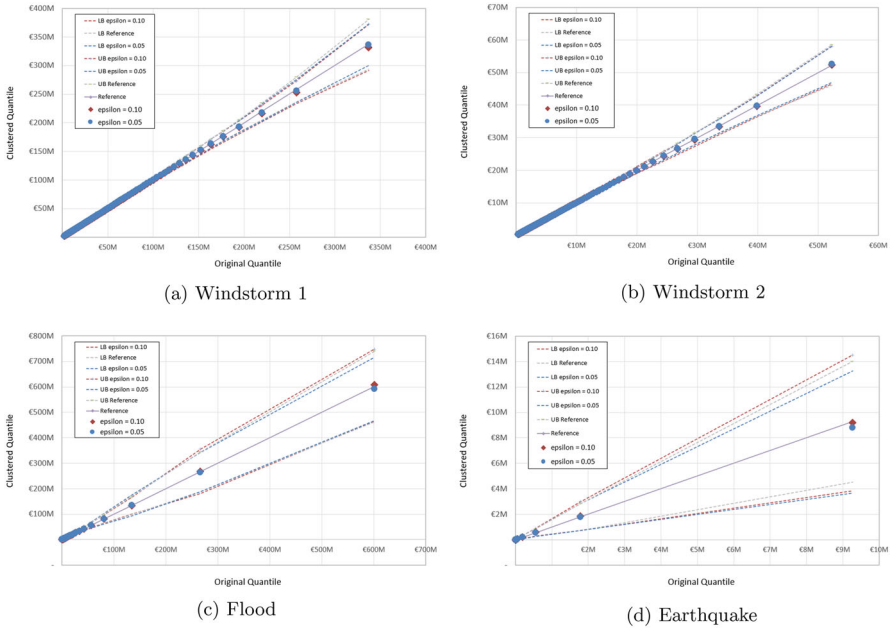
Table 4 Descriptive key figures of the total loss S for $\varepsilon = 0.05$ and $\varepsilon = 0.10$ including the secondary uncertainty

	Perils	Windstorm 1	Windstorm 2	Flood	Earthquake
$\varepsilon = 0.05$	No. of scenarios				
	Clustered data set	4888	6102	2523	2028
	Difference to the base case in %	− 77.73	− 65.62	− 64.19	− 57.78
	E(S)				
	Clustered data set	37,734,045	5,668,573	10,932,948	442,273
	Difference to the base case in %	− 0.17	0.38	− 0.17	0.04
	Std(S)				
	Clustered data set	64,702,593	9,210,767	90,213,234	12,769,220
	Difference to the base case in %	− 0.29	0.20	− 0.30	−0.12
	CoV(S)				
	Clustered data set	171.5	162.5	825.2	2887.2
	Difference to the base case in p.p.	− 0.2	− 0.3	− 1.1	− 4.9
$\varepsilon = 0.10$	No. of scenarios				
	Clustered data set	2211	3377	1759	1521
	Difference to the base case in %	− 89.48	− 80.97	−75.03	− 68.33
	E(S)				
	Clustered data set	37,785,796	5,695,088	10,961,460	441,611
	Difference to the base case in %	− 0.03	0.85	0.09	− 0.10
	Std(S)				
	Clustered data set	64,707,513	9,198,838	90,444,750	12,712,466
	Difference to the base case in %	− 0.28	0.07	− 0.05	− 0.57
	CoV(S)				
	Clustered data set	171.2	161.5	825.1	2878.7
	Difference to the base case in p.p.	− 0.4	− 1.3	− 1.1	− 13.4

For the earthquake data set, there is no remarkable difference in the rejection rates but this might also be caused by the variance in the resimulation in general. Note that it is interesting to see that the ELT data sets windstorm 2 and flood perform the worst as those were the data sets we assumed to have two knots. These results support the presumption that in the case of multiple elbows, a smaller, i.e. more conservative, ε-value is preferable.

Table 5 Comparison of the rejection rates for both frameworks using a paired KS test under a 10%-significance level

	Perils	Windstorm 1	Windstorm 2	Flood	Earthquake
$\varepsilon = 0.05$	Proportion rejected				
	Original data set	0.12	0.12	0.11	0.96
	Clustered data set	0.08	0.10	0.12	0.96
$\varepsilon = 0.10$	Proportion rejected				
	Original data set	0.10	0.11	0.09	0.97
	Clustered data set	0.05	0.41	0.17	0.95



(a) Windstorm 1 (b) Windstorm 2

(c) Flood (d) Earthquake

Fig. 5 Q–Q plots displaying the averaged quantiles of 100 resimluations and corresponding 95% upper and lower bounds of the quantile estimates of the original framework and the clustered approach using $\varepsilon = 0.05, 0.1$

To compare the outputs more directly, we use Q-Q plots to see whether the original resimulation framework and the clustered resimulation framework produce similar quantiles in general. We do this by taking the average over 100 resimulations of the quantile estimates. The quantiles we select to plot are corresponding to the 0%-quantile up to the 99.5%-quantile with a 0.5% step between two consecutive quantiles. Figure 5 shows the resulting Q-Q plots, which contain additionally the 95% confidence intervals of the simulated quantiles of the original resimulation framework ("UB reference" and "LB Reference") and of our clustering-based approach ("UB epsilon" and "LB epsilon") for two different values $\varepsilon = 0.05$ and 0.1, respectively.

As can be seen from the plots, both the clustered resimulation with an ε-value of 0.05 and 0.10 seem to match the quantile estimates of the original resimulation quite well. However, we do see that the clustered resimulation with an ε of 0.05 for the earthquake data does not result in a proper averaged estimate for the 99.5%-quantile. This again might be caused by the simulation variance and due to the heavy-tailed distribution of the earthquake ELT.

4.4 Comparing the runtime

In order to benefit from the clustered algorithm, a sufficient decrease must be obtained in the runtime. In addition, the time it takes to cluster the ELTs must also be taken into account as this adds to the runtime. In Fig. 6a the runtime is shown for the original resimulation framework together with the clustered resimulation framework using an ε-value of 0.05 for 10,000 resimulations. We can deduce from the figure that the runtime of the clustered framework together with the time it takes to cluster is roughly half of the runtime of the original framework. When using a more aggressive clustering approach using an ε-value of 0.10, the runtime further decreases to roughly a third or a fourth of the original resimulation runtime as shown in Fig. 6b.

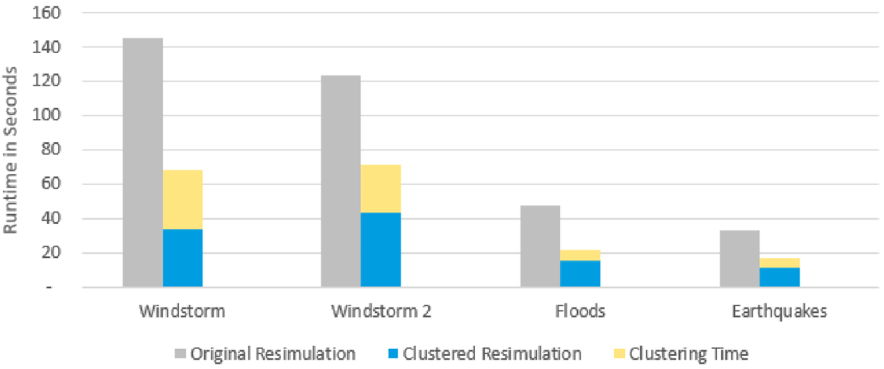
Note that the time it takes to cluster is constant with respect to the simulation path while the frameworks are roughly linearly increasing in runtime with respect to the simulation path. This means that if we would increase the simulation paths to, say, 100,000, we would see a further improvement of the relative runtime.

4.5 Further comparison with the same runtime

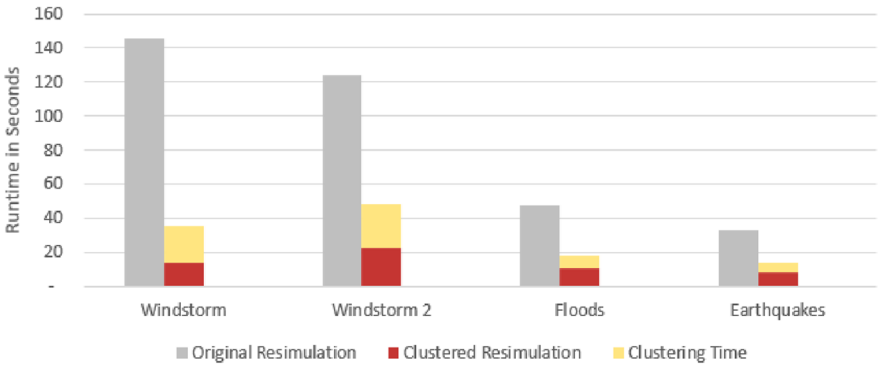
In this section we want to compare both simulation techniques on the basis of an additional aspect. We have seen that a more aggressive, i.e. larger, choice of ε works well on average when comparing the resulting quantiles and that this aggressive approach reduces the runtime more than the conservative choice. This in turn opens up time for additional simulation paths. To compare the benefit of the additional simulation paths to the variance of simulated quantiles, we reran both frameworks with the same runtime, whereby the time constraint is imposed by the original framework.

Table 6 shows the gains that are obtained from the additional resimulation paths for the 99.5%-quantile estimate. As can be seen, for both hyperparameters the average bias in the output seems to be small. Only the earthquake peril witnesses biases over 1%. In contrast, the standard deviation is reduced up to 73% for the aggressive approach and up to 55% for the more conservative approach which emphasizes the gains in variance by our approach.

In general, a more aggressive choice of ε leads to an increased bias of the clustered output due to the larger number of scenarios clustered together. Together with the results in Table 6 this reflects some kind of bias-variance trade-off that should be mitigated as far as possible by an appropriate choice of ε.



(a) Runtime clustered resimulation with an ε-value of 0.05.



(b) Runtime clustered resimulation with an ε-value of 0.10.

Fig. 6 Runtime of a single resimulation of 10,000 samples with the original framework and the clustered framework using an ε-value of 0.05 and 0.10

5 Conclusion

In this work, we presented a performant resimulation algorithm for ELTs based on a clustering-based approach. Empirical evidence so far in this work suggests that with the correct tuning, the clustered approach produces outputs that are statistically not significantly different from the original resimulation approach.

We have seen that the runtime can be a fraction of the original approach even when additionally taking the time to cluster into account. Thus, our clustered approach could help to solve the runtime versus variance issue.

Looking ahead, we plan to improve this methodology by deriving a more mathematical way to optimize the hyperparameters. We expect that the runtime gains versus the decrease in accuracy could be mathematically described with respect to the parameters. This in turn should end up in a data-driven way of optimizing the hyperparameters.

Table 6 Descriptive statistics of the 99.5%-quantile of the total loss when running the resimulation framework a 100 times with $\varepsilon = 0.05$ and 0.10, whereby the time budget is the same as for the original framework

	Perils	Windstorm 1	Windstorm 2	Flood	Earthquake
$\varepsilon = 0.05$	$\bar{q}_{99.5\%}$				
	Clustered data set	337,261,496	52,499,090	596,785,420	8,946,989
	Difference to the base case in %	0.01	0.37	− 0.77	− 3.48
	Std($q_{99.5\%}$)				
	Clustered data set	9,711,722	1,634,446	32,632,702	1,375,803
	Difference to the base case in %	− 55.33	− 47.04	− 53.00	− 41.98
$\varepsilon = 0.10$	$\bar{q}_{99.5\%}$				
	Clustered data set	335,705,451	52,520,277	600,980,558	9,393,852
	Difference to the base case in %	− 0.45	0.41	− 0.07	1.34
	Std($q_{99.5\%}$)				
	Clustered data set	5,945,100	1,247,679	31,472,377	1,241,614
	Difference to the base case in %	− 72.66	− 59.57	− 54.67	− 47.64

An interesting and practically relevant extension of our approach could be to generalize the proposed algorithm to the case of multi-portfolio ELTs. In this case, single events trigger different portfolios so that different business units that belong to one company experience losses that are caused by the same event and that have to be shared across this company. In this situation, the scenarios within the ELT possess not a single but multiple severity distributions. It is well known in the literature, that the increase of dimensionality is quite cumbersome for clustering algorithms. The adaption of our algorithm is under investigation and is the subject of a future contribution.

Acknowledgements The authors thank the Handling Editor and an anonymous referee for their helpful remarks and suggestions.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

1. Chen X (2015) A new clustering algorithm based on near neighbor influence. Expert Syst Appl 42(21):7746–7758
2. Deutsche Gesellschaft für Versicherungs- und Finanzmathematik: Interne Risikomodelle in der Schaden-/Unfallversicherung (2008)
3. Diers D (2007) Interne Unternehmensmodelle in der Schaden- und Unfallversicherung : Entwicklung Eines Stochastischen Internen Modells Für die Wert- und Risikoorientierte Unternehmenssteuerung und Für die Anwendung Im Rahmen Von Solvency II. ifa-Verlag, Ulm
4. Diers D (2009) Stochastic modelling of catastrophe risks in internal models. German Risk Insur Rev (GRIR) 5(1):1–28
5. Ester M, Kriegel H-P, Sander J, Xu X et al (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd 96:226–231
6. Fackler M (2022) Premium rating without losses. Eur Actuar J 12:275–316
7. Homer D, Li M (2017) Notes on using property catastrophe model results. Casual Actuar Soc E-Forum 2:29–44
8. Mitchell-Wallace K, Jones H, Foote M (2017) Natural catastrophe risk management and modelling: a practitioner's guide. Wiley-Blackwell, New Jersey