Online Supplemental Materials

Appendix A: Summary of Key Terms in Alphabetical Order

Term	Definition
AAL	Average annual loss, the annualized expected cost of all stochastic losses covered by an insurance policy
Actuarial fairness	the principle that each entity being insured should pay for the risk it introduces to the insurance pool, also referred to as chance solidarity
Catastrophe	a natural disaster that occurs with a relatively small probability but causes significant and often sudden damages. Examples include flood, hurricane, earthquake, etc.
Catastrophe insurance	a type of insurance that covers losses caused by certain types of catastrophes, e.g., hurricane windstorm insurance, flood insurance, etc.
Exceedance probability	the probability for an insurer to fall into insolvency as the total amount of covered loss exceeds the total capital prepared by the insurer (after taking into account reinsurance)
Expense load	the annualized cost of expenses incurred by the insurer (beyond AAL and capital requirement) as associated with an insurance policy
Insurance claim	a request from a policy holder to the insurer for the compensation of losses covered by the insurance policy
Insured loss	the amount of damage sustained by a policy holder and compensated by the insurer in consequence of events covered by the insurance policy
Insured peril	a cause of loss (e.g., hurricane wind damage, flood damage) that is covered by an insurance policy
Insurer	an entity that offers insurance, also referred to as insurance company, insurance provider, underwriter
Insurance policy	a contract between the insurer and a policy holder to specify the potential claims (e.g., perils, limits) covered by the insurer in exchange of the premium paid by the policy holder
Insurance pool	the set of all policies currently issued by an insurer for covering similar claims (e.g., those of the same peril type)
Insurance premium	the amount of money charged to a policy holder for covering the potential claims specified in the insurance policy
Loss distribution	the probability distribution of the annualized insured loss associated with an insurance policy
Policy holder	an entity that pays an insurer a premium in exchange for the coverage of certain potential claims.
Ratemaking	the process of determining the premium charged by the insurer for an insurance policy
Risk load for a policy	the portion of the annualized cost of capital requirement (beyond the total AAL) that is allocated to an insurance policy
Risk-load location	the process of allocating the total risk load for a set of insurance policies to the portion assigned to each policy in the set
Solidarity principle	the principle that each entity being insured should pay the same premium (or a premium that commensurates with the entity's income level), also referred to as subsidizing solidarity
Total loss distribution	the probability distribution of the annualized sum of insured losses associated with a set of insurance policies
Total risk load	the annualized cost of capital requirement (beyond the total AAL) for a set of catastrophe insurance policies

Appendix B: Mathematical Theorems and Proofs

B.1. Proof of Theorem 1

Theorem 1. *For given Ω and $\mathcal{L}$, the only $r(s|\Omega,\mathcal{L})$ that satisfies all three axioms is Shapley Value*

$$r_{\mathrm{SV}}(s|\Omega,\mathcal{L}) = \sum_{S\subseteq\Omega\backslash\{s\}} \frac{(n-1)!|S|!}{n!}(\mathcal{L}(S\cup\{s\})-\mathcal{L}(S)), \tag{8}$$

where $(\cdot)!$ represents the factorial.

Proof of Theorem 1. To prove the theorem, we need to show that Equation 8 is both a sufficient and a necessary condition for $r(s|\Omega,\mathcal{L})$ to satisfy the three axioms. We draw upon the proof of Proposition 2.1 by Ghorbani and Zou (2019) to construct our proof. Note that the only difference between the three axioms there and the three axioms in our paper is our second axiom, i.e., higher risk, higher pay. Our second axiom requires that, if two policies s_1 and s_2 satisfy $\mathcal{L}(S\cup\{s_1\})\geq\mathcal{L}(S\cup\{s_2\})$ for all $S\subseteq\Omega\backslash\{s_1,s_2\}$, then there must be $r(s_1|\Omega,\mathcal{L})\geq r(s_2|\Omega,\mathcal{L})$. In contrast, Ghorbani and Zou (2019) only requires an *equality axiom*, i.e., there must be $r(s_1|\Omega,\mathcal{L}) = r(s_2|\Omega,\mathcal{L})$ if s_1 and s_2 satisfy $\mathcal{L}(S\cup\{s_1\}) = \mathcal{L}(S\cup\{s_2\})$ for all $S\subseteq\Omega\backslash\{s_1,s_2\}$. Clearly, our second axiom is stronger as it implies that the equality axiom must be true. Thus, the necessity proof for Equation 8 follows directly from that of Proposition 2.1 by Ghorbani and Zou (2019).

For the sufficiency proof, we only need to show that the Shapley value design in Equation 8 satisfies our second axiom, as the other two follow directly from Ghorbani and Zou (2019). Equation 8 yields

$$\begin{aligned}
r_{\mathrm{SV}}(s_1|\Omega,\mathcal{L}) - r_{\mathrm{SV}}(s_2|\Omega,\mathcal{L}) &= \sum_{S\subseteq\Omega\backslash\{s_1\}} \frac{(n-1)!|S|!}{n!}(\mathcal{L}(S\cup\{s_1\})-\mathcal{L}(S)) - \sum_{S\subseteq\Omega\backslash\{s_2\}} \frac{(n-1)!|S|!}{n!}(\mathcal{L}(S\cup\{s_2\})-\mathcal{L}(S)) && (9)\\
&= \sum_{S\subseteq\Omega\backslash\{s_1,s_2\}} \frac{(n-1)!|S|!}{n!}(\mathcal{L}(S\cup\{s_1\})-\mathcal{L}(S)) + \\
&\quad \sum_{S\subseteq\Omega\backslash\{s_1,s_2\}} \frac{(n-1)!(|S|+1)!}{n!}(\mathcal{L}(S\cup\{s_2\}\cup\{s_1\})-\mathcal{L}(S\cup\{s_2\})) - \\
&\quad \sum_{S\subseteq\Omega\backslash\{s_1,s_2\}} \frac{(n-1)!|S|!}{n!}(\mathcal{L}(S\cup\{s_2\})-\mathcal{L}(S)) - \\
&\quad \sum_{S\subseteq\Omega\backslash\{s_1,s_2\}} \frac{(n-1)!(|S|+1)!}{n!}(\mathcal{L}(S\cup\{s_2\}\cup\{s_1\})-\mathcal{L}(S\cup\{s_1\})) && (10)\\
&\geq 0. && (11)
\end{aligned}$$

Note that the deduction from (10) to (11) is because $\mathcal{L}(S\cup\{s_1\})\geq\mathcal{L}(S\cup\{s_2\})$ for all $S\subseteq\Omega\backslash\{s_1,s_2\}$. This completes the sufficiency proof.

B.2. Proof of Theorem 2

Theorem 2. *Consider two policies s and s' such that the loss distribution of s is independent of all other policies in Ω, while the covariance between s' and any other policy s_j is positive and proportional to c. When n is sufficiently large, we have*

$$\frac{r_{\mathrm{S}}(s)}{r_{\mathrm{S}}(s')} < \frac{r_{\mathrm{SV}}(s)}{r_{\mathrm{SV}}(s')} \sim \frac{r_{\mathrm{SD}}(s)}{r_{\mathrm{SD}}(s')} < \frac{r_{\mathrm{V}}(s)}{r_{\mathrm{V}}(s')} \sim \frac{r_{\mathrm{MV}}(s)}{r_{\mathrm{MV}}(s')} \sim \frac{r_{\mathrm{MS}}(s)}{r_{\mathrm{MS}}(s')} \tag{12}$$

when $c = 0$ and $\sigma(s) > \sigma(s')$, and

$$\frac{r_{\mathrm{SV}}(s)}{r_{\mathrm{SV}}(s')} \sim \frac{r_{\mathrm{MV}}(s)}{r_{\mathrm{MV}}(s')} \sim \frac{r_{\mathrm{MS}}(s)}{r_{\mathrm{MS}}(s')} < \frac{r_{\mathrm{SD}}(s)}{r_{\mathrm{SD}}(s')} < \frac{r_{\mathrm{V}}(s)}{r_{\mathrm{V}}(s')} < \frac{r_{\mathrm{S}}(s)}{r_{\mathrm{S}}(s')} \tag{13}$$

when $c = \sigma^2$ and $\sigma(s) = \sigma(s')$, where $\sim$ represents asymptotic equality, and r_{S}, r_{V}, r_{SD}, r_{MV}, r_{MS}, and r_{SV} represent the risk-load allocation by the solidarity, variance, standard deviation, marginal variance, marginal surplus, and Shapley-value based methods, respectively.

Proof of Theorem 2. Consider the comparison between solidarity (i.e., r_{S}) and the rest of the allocation methods first. Obviously, except solidarity, which assigns equal risk load to all policies, all other methods assign a higher risk load to s (than s') in the first case and s' in the second case. This proves the position of r_{S} in (12) and (13).

To compare the rest of the allocation methods over a policy s (or s'), one subtle issue that needs to be addressed is the dependency of the marginal variance and marginal surplus methods (i.e., r_{MV} and r_{MS}) on the order in which the policies are added to the pool. To arrive at a stable estimate, many existing implementations (e.g., Mango 1997) opt for a "leave-one-out" approach. That is, one first compute the marginal variance/surplus for each policy s_i when setting the existing pool as $\Omega\backslash\{s_i\}$, before normalizing the computed results to a sum of $\mathcal{L}(\Omega)$. We follow this approach when computing r_{MV} and r_{MS}.

To simplify the notations, we designate the policy under consideration, say s, as the n-th policy (i.e., s_n) in the pool. With this notation, the following formulae follow directly from the definitions of the existing allocation methods.

$$r_{\mathrm{V}} \propto \lambda_n^2 \tag{14}$$

$$r_{\mathrm{SD}} \propto \lambda_n \tag{15}$$

$$r_{\mathrm{MV}} \propto \sigma^2\lambda_n^2 + 2\sum_{i=1}^{n-1} c\lambda_i\lambda_n = \sigma^2\lambda_n^2 + 2(n-1)c\lambda\lambda_n \tag{16}$$

$$r_{\mathrm{MS}} \propto \sqrt{\sum_{i=1}^{n}\sigma^2\lambda_i^2 + 2\sum_{i<j\le n} c\lambda_i\lambda_j} - \sqrt{\sum_{i=1}^{n-1}\sigma^2\lambda_i^2 + 2\sum_{i<j<n} c\lambda_i\lambda_j} \tag{17}$$

[Case 1] ($c = 0$). In this case, there is clearly $r_{\mathrm{SD}} \propto \lambda_n$, $r_{\mathrm{MV}} \propto \lambda_n^2$, and $r_{\mathrm{V}} \propto \lambda_n^2$. For r_{MS}, we can rewrite (17) as

$$\lim_{n\to\infty} r_{\mathrm{MS}} \propto \sqrt{\sum_{i=1}^{n}\sigma^2\lambda_i^2} - \sqrt{\sum_{i=1}^{n-1}\sigma^2\lambda_i^2} \approx \frac{\lambda_n^2}{2\sum_{i=1}^{n-1}\lambda_i^2} \propto \lambda_n^2 \tag{18}$$

In other words, among r_{SD}, r_{V}, r_{MV}, and r_{MS}, we have

$$\frac{r_{\mathrm{SD}}(s)}{r_{\mathrm{SD}}(s')} < \frac{r_{\mathrm{V}}(s)}{r_{\mathrm{V}}(s')} \sim \frac{r_{\mathrm{MV}}(s)}{r_{\mathrm{MV}}(s')} \sim \frac{r_{\mathrm{MS}}(s)}{r_{\mathrm{MS}}(s')}. \tag{19}$$

For r_{SV}, since our goal is to derive its asymptotic limit, we assign $\sigma = 1$ and $\lambda_i = 1$ for $i \in [1, n-1]$ to simplify the notations without loss of generality. This yields

$$r_{\mathrm{SV}} = \frac{1}{n}\left(\lambda_n + \sqrt{\lambda_n^2+1} - 1 + \sqrt{\lambda_n^2+4} - 2 + \cdots + \sqrt{\lambda_n^2+(n-1)^2} - (n-1)\right) \tag{20}$$

$$\approx \frac{\lambda_n}{n} + \left(\frac{1}{\sqrt{2}} + \frac{1}{\sqrt{5}} + \cdots + \frac{1}{\sqrt{1+(n-1)^2}}\right)\frac{\lambda_n - 1}{n} + \left(\frac{1}{2^{3/2}} + \frac{1}{5^{3/2}} + \cdots + \frac{1}{(1+(n-1)^2)^{3/2}}\right)\frac{(\lambda_n-1)^2}{2n} \tag{21}$$

$$\approx \frac{1}{n}\left(\lambda_n + (\lambda_n - 1)\log n + \frac{\zeta(3)}{2}(\lambda_n - 1)^2\right) \tag{22}$$

$$\propto \lambda_n, \tag{23}$$

where $\zeta(\cdot)$ is the Riemann zeta function. Note that the transition from (20) to (21) follows Taylor expansion of $f(\lambda_n) = \sqrt{\lambda_n^2 + c^2} - c$ at the point 1. The transition from (21) to (22) follows the approximation and limit of the sum of two well-known power series,

$$\sum_{r=1}^{n} \frac{1}{r} \approx \log n, \text{ and} \tag{24}$$

$$\lim_{n\to\infty} \sum_{r=1}^{n} \frac{1}{r^3} = \zeta(3) \approx 1.202, \tag{25}$$

where $\zeta(3)$ is commonly known as Apéry's constant. Combining (19) and (23), we have

$$\frac{r_{\mathrm{SD}}(s)}{r_{\mathrm{SD}}(s')} \sim \frac{r_{\mathrm{SV}}(s)}{r_{\mathrm{SV}}(s')} < \frac{r_{\mathrm{V}}(s)}{r_{\mathrm{V}}(s')} \sim \frac{r_{\mathrm{MV}}(s)}{r_{\mathrm{MV}}(s')} \sim \frac{r_{\mathrm{MS}}(s)}{r_{\mathrm{MS}}(s')}. \tag{26}$$

[Case 2] ($c = \sigma^2$). A key observation for this case is that, due to the perfect risk correlation between any pair in $\Omega\backslash\{s\}$, for all s_i where $i \in [1, n-1]$ (i.e., all policies but s), the marginal surplus for s_i when being added to any subset of $\Omega\backslash\{s_i, s\}$ is σ. Once again, to simplify the notations, consider the case where $\sigma = 1$ and $\lambda_i = 1$ for all but s. We have

$$\frac{r_{\mathrm{V}}(s)}{r_{\mathrm{V}}(s')} = \lambda_n^2 \tag{27}$$

$$\frac{r_{\mathrm{SD}}(s)}{r_{\mathrm{SD}}(s')} = \lambda_n \tag{28}$$

$$\lim_{n\to\infty} \frac{r_{\mathrm{MV}}(s)}{r_{\mathrm{MV}}(s')} = \lim_{n\to\infty} \frac{\lambda_n^2}{1 + 2(n-1)} = 0 \tag{29}$$

$$\lim_{n\to\infty} \frac{r_{\mathrm{MS}}(s)}{r_{\mathrm{MS}}(s')} = \lim_{n\to\infty} \frac{\lambda_n^2}{2(n-1)} = 0 \tag{30}$$

$$\lim_{n\to\infty} \frac{r_{\mathrm{SV}}(s)}{r_{\mathrm{SV}}(s')} = \lim_{n\to\infty} \frac{\lambda_n}{n} = 0. \tag{31}$$

Here (27)-(29) follow directly from the definitions in (14)-(16), respectively. The earlier observation on the constant marginal surplus, combined with (17), yields (30). Similarly, recall from earlier discussions that r_{SV} is a weighted average of marginal surpluses. Thus, a constant marginal surplus implies $r_{\mathrm{SV}} = r_{\mathrm{MS}}$ for s_i with $i \in [1, n-1]$. This, combined with (22), yields (31). Combining (27)-(31), we have

$$\frac{r_{\mathrm{SV}}(s)}{r_{\mathrm{SV}}(s')} \sim \frac{r_{\mathrm{MV}}(s)}{r_{\mathrm{MV}}(s')} \sim \frac{r_{\mathrm{MS}}(s)}{r_{\mathrm{MS}}(s')} < \frac{r_{\mathrm{SD}}(s)}{r_{\mathrm{SD}}(s')} < \frac{r_{\mathrm{V}}(s)}{r_{\mathrm{V}}(s')}. \tag{32}$$

The proof for the general case (where $\sigma \neq 1$ and/or $\lambda_i \neq 1$) follows directly in analogy.

B.3. Proof of Theorem 3

THEOREM 3. *Algorithm FAST-SV returns an (ϵ, δ)-approximation of $r_{\mathrm{SV}}(s|\Omega, \mathcal{L})$, i.e., with an estimation error ω satisfying $\Pr\{|\omega| < \epsilon\} > 1 - \delta$, after making $\mathcal{O}(1)$ calls to $\mathcal{L}$ and incurring a computational cost of $\mathcal{O}((n/\epsilon^2)\log(n/\delta))$.*

Proof of Theorem 3. According to Theorem 8 in Jia et al. (2019b), which was proven using Hoeffding's inequality, Algorithm FAST-SV returns an (ϵ, δ)-approximation of $r_{\mathrm{SV}}(s|\Omega, \mathcal{L})$ if

$$m \geq \frac{2r^2 n}{\epsilon^2} \log \frac{2n}{\delta}, \tag{33}$$

where r is the range of Shapley value, which is obviously a constant in our case as $r_{\mathrm{SV}}(s|\Omega, \mathcal{L})$ is, by definition, bounded from the above by $\sigma(s)$. Note from Algorithm FAST-SV that it makes $c+1$, i.e., $\mathcal{O}(1)$, calls to $\mathcal{L}(\cdot)$ and has a computational complexity of $\mathcal{O}(m)$ (i.e., from the loop between Lines 5 and 7). Taking Inequality 33 into it, we have a computational cost of $\mathcal{O}((n/\epsilon^2)\log(n/\delta))$ for Algorithm FAST-SV.

Appendix C: Summary Statistics

Table 7 lists the summary statistics for the amount of loss. To highlight the geographic variation of losses, we grouped the claims into ten categories according to the first digit of the ZIP code in which the loss occurred. According to the design of ZIP code, this first digit represents its geographic region. For example, the regions most frequently suffer flood damage from hurricanes are 3* (i.e., the southeast region covering Florida, Alabama, etc.) and 7* (i.e., the southwest region covering Texas, Louisiana, etc.). Indicatively, the two groups have the largest number of claims (**n**) and the highest average loss ($\bar{\mathbf{x}}$) in the table. Table 7 also shows that, for all groups, the standard deviation (**s**) far exceeds the average loss ($\bar{\mathbf{x}}$), which is in turn higher than the median loss ($\widetilde{\mathbf{x}}$). In other words, the loss distribution exhibits both a wide dispersion and a positive skew (i.e., a heavy-tailed distribution), two features typical of catastrophic losses (Embrechts et al. 2013).

Table 7 Summary Statistics for NFIP Claim Dataset

Var	ZIP	n	Min	$\mathbf{q_1}$	$\widetilde{\mathbf{x}}$	$\bar{\mathbf{x}}$	$\mathbf{q_3}$	Max	s
Loss	0*	241387	0.01	3245.50	10014.38	31093.53	32165.04	4022518.64	61275.82
	1*	203996	0.01	2933.91	10212.47	33587.16	37658.36	9467720.05	73864.06
	2*	187612	0.01	3118.10	8733.00	23348.97	23976.09	2054595.38	45666.26
	3*	316492	0.01	3117.69	10540.17	35368.06	35292.54	10000000.00	82350.71
	4*	66825	6.76	2165.82	6585.40	17365.26	18518.92	1000000.00	36707.33
	5*	42993	0.01	1671.29	6132.01	21797.53	20000.00	1250000.00	52971.74
	6*	89982	0.01	2176.18	6563.61	18802.33	18878.06	1007405.42	42434.89
	7*	702087	0.01	5000.00	22435.35	54298.85	77095.46	9023557.85	77406.53
	8*	10297	0.01	1938.57	6500.15	21974.97	19563.70	1000000.00	50737.02
	9*	53138	0.01	2496.30	7896.72	21562.16	23052.36	1400000.00	43822.25
	all	1914809	0.01	3369.52	11963.59	38236.44	44000.00	10000000.00	70803.31

Note. **Var** = Variable. **ZIP** = first digit of the property ZIP code. **n** = number of records. **Min** = minimum. $\mathbf{q_1}$ = first quartile. $\widetilde{\mathbf{x}}$ = median. $\bar{\mathbf{x}}$ = mean. $\mathbf{q_3}$ = third quartile. **Max** = maximum. **s** = standard deviation. Loss = Total payout for a claim in 2020 dollars.

Table 8 lists the summary statistics for the joined dataset of NFIP policies and Census sociodemographic variables. As can be seen from the table, the sociodemographic variables vary widely by region. For example, the percentage of African Americans (**race**) differ by nearly an order of magnitude between the mountain states (2% for ZIP = 8*) and the mid-Atlantic region (18% for ZIP = 2*). Similarly, there is an uneven distribution of policies across regions. The hurricane-prone southeast (3*) and southwest (7*) regions, for example, account for 64% of all policies in the pool.

Table 8 Summary Statistics for NFIP Policy Dataset

ZIP	n	age	race	income
0*	22405	46.16	0.04	104706.18
1*	18847	43.27	0.09	99296.09
2*	38604	42.10	0.18	71449.21
3*	169194	48.83	0.11	69967.86
4*	46075	40.05	0.05	59737.44
5*	6409	40.81	0.02	90085.74
6*	4162	37.58	0.07	76958.88
7*	130341	36.78	0.18	77066.81
8*	8728	43.65	0.02	100128.38
9*	19815	42.30	0.03	81396.19
all	464580	42.15	0.08	83079.28

Note. **age**, **race**, **income** are per-policy average.