Competitive Premium Pricing and Cost Savings for Insurance Policy Holders: leveraging Big Data

Zvezdov, Ivelin

AIR Worldwide, VERISK Analytics Corp.

7 February 2017

Competitive Premium Pricing and Cost Savings for Insurance Policy Holders: leveraging Big Data

Ivelin M. Zvezdov, M.Phil. Senior Product Manager,
AIR Worldwide, VERISK Analytics Corp., 131 Dartmouth Street, Boston 02116, MA, USA
izvezdov@air-worldwide.com, ivelin.zvezdov@gmail.com

Sebastian Rath, PhD, Principal Insurance Risk Officer,
NN Group, Rotterdam, The Netherlands, *Sebastian.Rath@gmail.com*

Abstract

Examining the intersection of research on the effects of (re)insurance risk diversification and availability of big insurance data components for competitive underwriting and premium pricing is the purpose for this paper. We study the combination of physical diversification by geography and insured natural peril with the complexity of aggregate structured insurance products, and furthermore how big historical and modeled data components impact product underwriting decisions. Under such market conditions, the availability of big data components facilitates accurate measurement of inter-dependencies among risks, and the definition of optimal and competitive insurance premium at the level of the firm and the policy holders. We extend the discourse to a notional micro-economy and examine the impact of diversification and insurance big data components on the potential for developing strategies for sustainable and economical insurance policy underwriting. We review concepts of parallel and distributed algorithmic computing for big data clustering, mapping and resource reducing algorithms.

Key-words

Effects of insurance risk diversification on premium definition; contribution of big data components to measuring inter-dependencies; rational for sustainable and economic underwriting practices and cost savings

Introduction

This working paper will examine how big data and fast compute platforms solve some complex premium pricing and portfolio structuring and accumulation problems in context of flood insurance markets. Our second objective is to measure the effects of geo-spatial insurance risk diversification through modeling of interdependencies and show that such measures have impact on single risk premium definition and its market cost. The single product case studies examine the pricing of insurance umbrella coverage. They are selected to address scenarios relevant to current (re)insurance market conditions under intense premium competition. Then we extend the discourse to a micro-economy of multiple policy holders and aim to generalize some finding on economies of scale and diversification. The outcomes of all case studies and theoretical analysis depend on the availability of big insurance data components for modeling and pricing workflows. The quality, usability and computational cost of such data components determine their direct impact on the underwriting and pricing process and on definition of the single risk cost of insurance.

1.0 Pricing Aggregate Umbrella Policies

Insurers are competing actively for insured's premiums and looking for economies of scale to offset and balance premium competition and thus develop more sustainable long term underwriting strategies. While writing competitive premium policies and setting up flexible contract structures, insurers are mindful of risk concentration, and the lower bounds of fair technical pricing. Structuring of aggregate umbrella policies lends itself to underwriting practices of larger scales in market share and diversification. Only large insurers have the economies of scale to offer such products to their clients.

Premium pricing of umbrella and global policies relies on both market conditions and mathematical modeling argument. On the market and operational side the insurer relies on lower cost of umbrella products due to efficiencies of scale in brokerage, claims management, administration, and even in the computational scale-up of the modeling and pricing internal functions of its actuarial departments. In our study we will first focus on the statistical modeling argument, and then we will define big data components, which allow for solving such policy structuring and pricing problem.

We first set up the case study on a smaller scale in context of two risks - with insured limits for flood of 90 and 110 million respectively. These risks are priced for combined river-rain and storm surge flood coverage, first with both single limits separately and independently and then in an aggregate umbrella insurance product with a combined limit of 200 million:

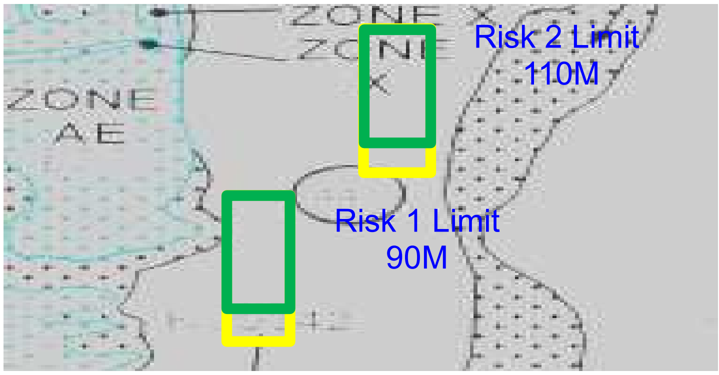
$$Umbrella(200M) = Limit\ 1\ (90M) + Limit\ 2\ (110M) \tag{1.0}$$

Table 1: Policy Set-Up and Limit Coverage

Policy Set Up	Limit Coverage
Policy 1, π (S)	90M
Policy 2, π (Q)	110M
Umbrella, π (S + Q)	200M

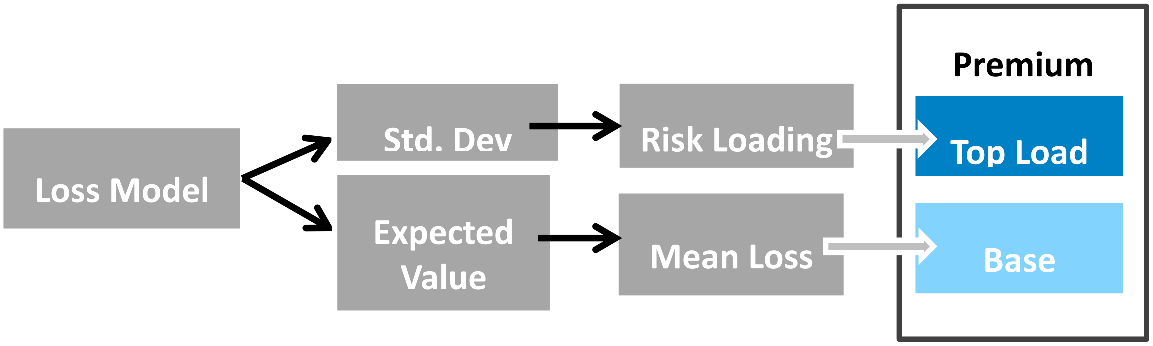
The two risks are owned by a single insured, and are located in a historical flood zone, in geo-spatial proximity to each other, of less than 1 kilometer.

Figure 1: Geospatial location of insured risk with less than one kilometer of proximity



For premium pricing we assume a traditional approach dependent on modeled expected values of insured loss and standard deviation of loss.

Figure 2: Basic Insurance Premium Components and Construction



To set the statistical mechanics of the case study for both risks, we have a modeled flood insurance loss data samples Q_t *and* S_t respectively for both risks, from a stochastic simulation - T. Modeled insured losses have an expected value $E[.]$ and a standard deviation $\sigma[.]$, which define a standard policy premium of $\pi(.)$

When both policies' premiums are priced independently, by the standard deviation pricing principle we have:

$$\pi(S_t) = E[S_t] + \sigma[S_t]$$

$$\pi(Q_t) = E[Q_t] + \sigma[Q_t] \qquad (1.1)$$

With non-negative loadings, it follows that:

$$\pi(S_t) \geqq E[S_t]$$
$$\pi(Q_t) \geqq E[Q_t] \qquad (2.0)$$

Since both risks are owned by the same insured we aggregate the two standard premium equations, using traditional statistical accumulation principles for expected values and standard deviations of loss.

$$\pi(Q_t) + \pi(S_t) = E[S_t] + \sigma[S_t] + E[Q_t] + \sigma[Q_t]$$

$$\pi(Q_t) + \pi(S_t) = E[S_t + Q_t] + \sigma[S_t] + \sigma[Q_t] \tag{3.0}$$

The theoretical joint insured loss distribution function $f_{S,Q}(S_t, Q_t)$ of the two risks will have an expected value of insured loss:

$$E[S_t + Q_t] = E[S_t] + E[Q_t] \tag{4.0}$$

And a joint theoretical standard deviation of insured loss:

$$\sigma[S_t + Q_t] = \sqrt{\sigma^2[S_t] + \sigma^2[Q_t] + 2\rho\sigma[S_t] * \sigma[Q_t]} \tag{4.1}$$

We use further these aggregation principles to express the sum of two single risks premiums - $\pi(Q_t), \pi(S_t)$, as well as to derive a combined premium $\pi(Q_t + S_t)$ for an umbrella coverage product insuring both risks with equivalency in limits as in (1.0). An expectation for full equivalency in premium definition produces the following equality:

$$\pi(Q_t + S_t) = E[S_t + Q_t] + \sqrt{\sigma^2[S_t] + \sigma^2[Q_t] + 2\rho\sigma[S_t] * \sigma[Q_t]} = \pi(Q_t) + \pi(S_t) \tag{4.2}$$

The expression introduces a correlation factor ρ between modeled insured losses of the two policies. In our case study this correlation factor specifically expresses dependencies between historical and modeled losses for the same insured peril due to geo-spatial distances. Such correlation factors are derived by algorithms which measure dependencies of historical and modeled losses on their sensitivities to geo-spatial distances among risks. In this article we will not delve into the definition of such geo-spatial correlation algorithms. Three general cases of dependence relationships among flood risks due to their geographical situation and distances are examined in our article: *full independence, full dependence and partial dependence*.

2.0 Sub-Additivity, Dependence and Diversification

Scenario 2.1: Two Boundary Cases of Fully Dependent and Fully Independent Risks

In the first boundary case, where we study *full dependence between risks*, expressed with a unit correlation factor, we have from first statistical principles that the theoretical sum of the standard deviations of loss of the fully dependent risks is equivalent to the standard deviation of the joint loss distribution of the two risks combined, as defined in equation (4.1).

$$\sigma[S_t + Q_t] = \sqrt{\sigma^2[S_t] + \sigma^2[Q_t] + 2\sigma[S_t] * \sigma[Q_t]} = \sigma[S_t] + \sigma[Q_t] \qquad (4.3)$$

For expected values of loss, we already have a known theoretical relationship between single risks' expected insurance loss and umbrella product expected loss in equation (4.0). The logic of summations and equalities for the two components in standard premium definition in (4.0) and (4.3) leads to deriving a relationship of proven full additivity in premiums between the single policies and the aggregate umbrella product, as described in equation (4.2), and shortened as:

$$\pi(Q_t + S_t) \ = \ \pi(Q_t) + \pi(S_t) \qquad (4.4)$$

Some underwriting conclusions are evident from this analysis. When structuring a combined umbrella product for fully dependent risks, in very close to identical geographical space, same insured peril and line-of-business - the price of the aggregated umbrella product should approach the sum of single risk premiums priced independently. The absence of diversification in geography and insured catastrophe peril prevents any significant opportunities for cost savings or competitiveness in premium pricing. The summation of riskiness form single policies to aggregate forms of products is linear and co-monotonic. Economies of market share scale do not play a role in highly clustered and concentrated pools of risks, where diversification is not achievable, and inter-risk dependencies are close to perfect. In such scenarios the impact of big data components to underwriting and pricing practices is not as prominent, because formulation of standard premiums for single risks and aggregated products could be achieved by theoretical formulations.

In our second boundary case of *full and perfect independence*, when two or more risks with two separate insurance limits are priced independently and separately, the summation of their premiums is still required for portfolio accumulations by line-of-business and geographic and administrative region. This premium accumulation task or 'roll-up' of fully independent risks is accomplished by practitioners accordingly with the linear principles of equation (3.0). However, if we are to structure an aggregate umbrella cover for these same single risks with an aggregated premium of $\pi(Q_t + S_t)$, the effect of statistical independence expressed with a zero correlation factor will reduce equation (3.0) to equation (5.0).

$$\pi(Q_t + S_t) \;=\; E[S_t + Q_t] + \sqrt{\sigma^2[S_t] + \sigma^2[Q_t]} \tag{5.0}$$

Full independence among risks more strongly than any other cases supports the premium sub-additivity principle, which is stated I (6.0).

$$\pi(Q_t + S_t) \;\leqq\; \pi(Q_t) + \pi(S_t) \tag{6.0}$$

An expanded expression of the subadditivity principle is easily derived from the linear summation of premiums in (3.0) and the expression of the combined single insurance product premium in (5.0).

Some policy and premium underwriting guidelines can be derived from this regime of full statistical independence. Under conditions of full independence, when two risks are priced independently and separately the sum of their premiums will always be larger than the premium of an aggregate umbrella product covering these same two risks. The physical and geographic characteristics of full statistical independence for modeled insurance loss are large geo-spatial distances and independent insured catastrophe perils and business lines. In practice this is generally defined as insurance risk portfolio diversification by geography, line, and peril. In insurance product terms, we proved that diversification by geography, peril and line-of-business, which are the physical prerequisites for statistical independence, allow to structure and price an aggregate umbrella product with a premium less than the sum of the independently priced premiums of the underlying insurance risks.

In this case, unlike with the case of full dependence, big data components have a computing and accuracy function to play in the underwriting and price definition process. Once the subadditivity of the aggregate umbrella product premium as in (6.0) is established, this premium is then back-allocated to the single component risks covered by the insurance product. This is done in order to measure the relative riskiness of the assets under the aggregate insurance coverage and each risk individual contribution to the formation of the aggregate premium. The back-allocation procedure is described further in the article in the context of a notional micro economy case.

Scenario 2.2: Less Than Fully Dependent Risks

In our case study we have geo-spatial proximity of the two insured risks in a known flood zone with measured and available averaged historical flood intensities, which leads to a measurable statistical dependence of modeled insurance loss. We express this dependence with a computed correlation factor in the interval $[0 < \rho' < 1.0]$.

Partial dependence with a correlation factor $0 < \rho' < 1.0$ has immediate impact on the theoretical standard deviation of combined modeled loss, which is a basic quantity in the formulation of risk and loading factors for premium definition.

$$\sigma[S_t + Q_t] = \sqrt{\sigma^2[S_t] + \sigma^2[Q_t] + 2\rho^{'}\sigma[S_t]\sigma[Q_t]} \leqq \sigma[S_t] + \sigma[Q_t]$$

This leads to redefining the equality in (4.3) to an expression of inequality between the premium of the aggregate umbrella product and the independent sum of the single risk premiums, as in the case of complete independence.

$$\pi(Q_t + S_t) \; = \; E[S_t + Q_t] + \sqrt{\sigma^2[S_t] + \sigma^2[Q_t] + 2\rho'\sigma[S_t] * \sigma[Q_t]} \leqq \pi(Q_t) + \pi(S_t) \quad (7.0)$$

The principle of premium sub-additivity (6.0), as in the case of full independence, again comes into force. The expression of this principle is not as strong with partial dependence as with full statistical independence, but we can clearly observe a theoretical ranking of aggregate umbrella premiums $\pi(Q_t + S_t)$ in the three cases reviewed so far.

$$\pi^{Full\ Independence} \leqq \pi^{Partial\ Dependence} \leqq \pi^{Full\ Dependence} \qquad (7.1)$$

This theoretical ranking is further confirmed in the next section with computed numerical results.

Less than full dependencies, i.e. partial dependencies among risks, could still be viewed as a statistical modeling argument for diversification in market share geography, line-of-business, and insured peril. Partial but effective diversification still offers an opportunity for competitive premium pricing. In insurance product and portfolio terms our study proves that partial or imperfect diversification by geography affects the sensitivity of premium accumulation, and allows for cost savings in premium for aggregate umbrella products vs. the summation of multiple single risk policy premiums.

3.0 Numerical Results of Single Risk and Aggregate Premium Pricing Cases

In our flood risk premium study, we modeled and priced three scenarios, using classical formulas for a single risk premium in equation (1.0) and for umbrella policies in equation (7.0). In our first scenario we price each risk separately and independently with insured limits of 90M and 110M. In the second and third scenarios, we price an umbrella product with a limit of 200M, in three sub-cases with $\{1.0, 0.3\ and\ 0.0\}$ correlation factors, respectively to represent full dependence, partial dependence and full independence of modeled insured loss. We use stochastic modeled insurance flood losses computed with high geo-spatial granularity of 30 meters.

Table 2: Numerical results of premium pricing under three dependence structures

Insured Limit(s)	Policy & Premium Set Up	Premium	Dependence & Additivity
90M	Policy 1: π(1)	512K	
110M	Policy 2: π(2)	725K	
200M	Premium Sum: π(1) +π(2)	1.24M	Full Dependence & Additivity
200M	Umbrella: π(1+2): 100% correlation	1.24M	Full Dependence & Additivity
200M	Umbrella: π(1+2): 30% correlation	1.02M	Partial Dependence & Sub-Additivity
200M	Umbrella: π(1+2): 0% correlation	0.9M	Full Independence & Sub-Additivity

The numerical results of our experiment fully support the conclusions and guidelines which we earlier derived from theoretical statistical relationships. For fully dependent risks in close proximity, the sum of single risk premiums approaches the price of an umbrella product, which is priced with 1.0 (100%) correlation factor. This is the stochastic relationship of full premium additivity. For partially dependent risks, the price of a combined product, modeled and priced with a 0.3 (30%) correlation factor, could be less than the sum of single risk premiums. For fully independent risks, priced with a 0 (0.0%) correlation factor the price of the combined insurance cover will further decrease to the price of an umbrella on partially dependent risks (30% correlation). Partial dependence and full independence support the stochastic ordering principle of premium sub-additivity. The premium ranking relationship in (7.1) is strongly confirmed by these numerical pricing results.

Less than full dependence among risks, which is a very likely and practical measurement in real insurance umbrella coverage products, could still be viewed as the statistical modeling argument for diversification in market share geography. Partial and incomplete dependence theoretically and numerically supports the argument that partial but effective diversification offers an opportunity for competitive premium pricing.

4.0 Theoretical Expansion to a Single Firm Micro-Economy Case

We expand the discourse to a simple theoretical micro-economy, and examine if the same principles derived for the aggregate umbrella insurance product still hold on the larger scale of an insurance firm. In a notional economy with $\{1 \ldots to \ldots N\}$ insurance risks $r_{1,N}$ and policy holders respectively, we have only one insurance firm, which at time T, does not have an information data set θ_T about dependencies among per-risk losses. Each premium is estimated by the traditional standard deviation principle in (1.1). For the same time period T the insurance firm collects a total premium $\pi_T[total]$ equal to the linear sum of all $\{1 \ldots to \ldots N\}$ policy premiums $\pi_T[r_N]$ in the notional economy.

$$\pi_T[r_1] + \ldots + \pi_T[r_N] = \sum_{i=1}^{N} \pi_T = \pi_T[total] \qquad (8.0)$$

There is full additivity in portfolio premiums, and because of unavailability of data on inter-risk dependencies for modeling, the insurance firm cannot take advantage of competitive premium cost savings due to market share scale and geographical distribution and diversification of the risks in its book of business. For coherence we assume that all insurance risks and policies belong to the same line of business and cover the same insured natural peril - flood, so that the only insurance risks diversification possible is due to insurance risk independence derived from geo-spatial distances. A full premium additivity equation similar to an aggregate umbrella product premium (3.0), extended for the case of the total premium of the insurance firm in our micro-economy, is composed in (9.0)

$$\pi_T[total] = \pi_T[r_1] + \ldots + \pi_T[r_N] = E[r_1 + \cdots + r_N] + \sigma[r_1] + \cdots + \sigma[r_N] \qquad (9.0)$$

In the next time period $T+1$ the insurance firm acquires a data set θ_{T+1} which allows it to model geo-spatial dependencies among risks and to identify fully dependent, partially dependent and fully independent risks. The dependence structure is expressed and summarized in a $[NxN]$ correlation matrix - $\rho_{i,N}$. Traditionally, full independence between any two risks is modeled with a zero correlation factor, and partial dependence is modeled by a correlation factor less than one. With this new information we can extend the insurance product expression (7.0) to the total accumulated premium $\pi_{T+1}[total]$ of the insurance firm at time $T+1$

$$\sum_{i=1}^{N} \pi_{T+1} = E[r_1 + \cdots + r_N] + \sqrt{\sum_{1,N} \sigma^2[r_i] + \sum_{1,N} 2\rho_{i,N}\sigma[r_i]\sigma[r_N]} \qquad (10.0)$$

The impacts of full independence and partial dependence, which are inevitably present in a full insurance book of business, guarantee that the sub-additivity principle for premium accumulation comes into effect. In our case study sub-additivity has two related expressions. Between the two time periods the acquisition of the dependence data set θ_T which is used for modeling and definition of the correlation structure $\rho_{i,N}$ provides that a temporal sub-additivity or inequality between the total premiums of the insurance firm can be justified in (10.1).

$$\sum_{i=1}^{N} \pi_{T+1} \leq \sum_{i=1}^{N} \pi_T \qquad (10.1)$$

It is undesirable for any insurance firm to seek lowering its total cumulative premium intentionally because of reliance on diversification. However an underwriting guidelines' implication could be that after the total firm premium is accumulated with a model taking account of inter-risk dependencies, then this total monetary amount can be back-allocated to individual risks and policies and thus provide a sustainable competitive edge in pricing. The business function of diversification and taking advantage of its consequent premium cost savings is achieved through two statistical operations: accumulating pure flood premium with a correlation structure, and then back-allocating the total firms' premium down to single contributing risk granularity. A backwardation relationship for the back-allocated single risk and single policy premium $\pi'_{T+1}[r_N]$ can be derived with a standard deviations' proportional ratio. This per-risk back-allocation ratio is constructed from the single risk standard deviation of expected loss $\sigma_{T+1}[r_N]$ and the total linear sum of all per-risk standard deviations $\sum_{i=1}^{N} \sigma_{T+1}[r_N]$ in the insurance firm's book of business

$$\pi'_{T+1}[r_N] \ = \sum_{i=1}^{N} \pi'_{T+1}\,[r_N] * \left[\frac{\sigma_{T+1}[r_N]}{\sum_{i=1}^{N} \sigma_{T+1}[r_N]}\right] \qquad (11.0)$$

From the temporal sub-additivity inequality between total firm premiums in (10.1) and the back-allocation process for total premium $\sum_{i=1}^{N} \pi'_{T+1}\,[r_N]$ down to single risk premium in (11.0), it is evident that there are economies of scale and cost in insurance policy underwriting between the two time periods for any arbitrary single risk r_N. These cost savings are expressed in (12.0).

$$\pi'_{T+1}[r_N] \leq \pi_T[r_N] \qquad (12.0)$$

In our case study of a micro economy and one notional insurance firms' portfolio of one insured peril, namely flood, these economies of premium cost are driven by geo-spatial diversification among the insured risks. We support this theoretical discourse with a numerical study.

4.1 Notional Flood Insurance Portfolio Case Study

We construct two notional business units each containing ten risks, and respectively ten insurance policies. The risks in both units are geo-spatially clustered in high intensity flood zones – Jersey City in New Jersey – 'Unit NJ' and Baton Rouge in Louisiana – 'Unit BR'. For each business unit we perform two numerical computations for premium accumulation under two dependence regimes. Each unit's accumulated *fully dependent* premium is computed by equation (9.0). Each unit's accumulated *partially dependent* premium, modeled with a constant correlation factor of 0.6 (60%), between any two risks, for both units is computed by equation (10.0). The total insurance firm's premium under both cases of full dependencies and partial dependence is simply a linear sum – 'business unit premiums' roll up to the book total.

Table 3: Results for accumulated premium for two business units and the portfolio total

Total Insurance Firm Premium		
	Fully Dependent Premium	**Partially Dependent Premium**
Unit NJ	37.8M	32.5M
Unit BR	27.1M	23.9M
Total Book	64.9M	56.4M

In all of our case studies we have focused continuously on the impact of measuring geo-spatial dependencies and their interpretation and usability in risk and premium diversification. For the actuarial task of premium accumulation across business units, we assume that the insurance firm will simply roll - up unit total premiums, and will not look for competitive pricing as a result of diversification across business units. This practice is justified by underwriting and pricing guidelines being managed somewhat autonomously by geo-admin business unit, and premium and financial reporting being done in the same manner.

In our numerical case study we prove that the theoretical inequality (10.1), which defines temporal subadditivity of premium with and without dependence modeled impact is maintained. Total business unit premium computed without modeled correlation data and under assumption of full dependence $\sum_{i=1}^{N} \pi_T$ always exceeds the unit's premium under partial dependence $\sum_{i=1}^{N} \pi_{T+1}$ computed with acquired and modeled correlation factors.

$$\sum_{i=1}^{N} \pi_{T+1}(Unit\ NJ) \leq \sum_{i=1}^{N} \pi_T(Unit\ NJ)$$

$$\sum_{i=1}^{N} \pi_{T+1}(Unit\ BR) \leq \sum_{i=1}^{N} \pi_T(Unit\ BR)$$

This justifies performing back-allocation in both business units, using procedure (11.0), of the total premium $\sum_{i=1}^{N} \pi_{T+1}$ computed under partial dependence. In this way competitive cost savings can be distributed down to single risk premium. In table 4, we show the results of this back-allocation procedure for all single risks in both business units:

Table 4: Single Risk Premiums by Unit under two Correlation Factors

Single Risk Premiums					
Unit NJ Risks	Fully Dependent Premiums	Partially Dependent Premiums	Unit BR Risks	Fully Dependent Premiums	Partially Dependent Premiums
risk 1	1,373,677	1,314,438	risk 11	496,449	323,495
risk 2	790,016	750,127	risk 12	7,225,247	6,601,950
risk 3	1,225,628	1,160,409	risk 13	7,225,247	6,601,950
risk 4	3,837,894	3,391,682	risk 14	147,973	97,815
risk 5	3,837,894	3,391,682	risk 15	267,605	169,304
risk 6	9,533,304	8,560,567	risk 16	812,826	579,865
risk 7	7,897,792	6,278,738	risk 17	232,896	148,851
risk 8	7,871,039	6,253,646	risk 18	10,155,420	9,082,536
risk 9	181,688	174,465	risk 19	113,118	80,000
risk 10	1,241,295	1,203,113	risk 20	378,275	242,799
Total Unit	37,790,226	32,478,869		27,055,056	23,928,565

For each single risk we observe that the per-risk premium inequality (12.0) is maintained by the numerical results. Partial dependence, which can be viewed as the statistical – modeling expression of imperfect insurance risk diversification proves that it could lead to opportunities for competitive premium pricing and premium cost savings for the insured on a per-risk and per-policy cost savings.

4.2 Premium Mapping and Quantile Pricing

The pure technical insurance premium can be expressed as a value-at-risk VaR or tail-value-at-risk TVaR metric computed at exceedance probability α from the full insurance loss distribution S_n of each insured risk r_n, such that

$$VaR_\alpha(S_n) = \inf\{s|P(S_n > s)1 - \alpha\} \quad (13.0)$$

$$TVaR_{\alpha}(S_n) = \frac{1}{1-\alpha}\int_{\alpha}^{1} VaR(S_n)dt \tag{13.1}$$

For our micro-economy case study, we map each risk premium absolute value – partially dependent and back-allocated from *table 4* to a VaR and TVaR value from the full risk insurance loss distribution.

Table 5: Back-allocated dependent single risk premiums mapped to VaR and TVaR

Single Risk Premiums mapped to VaR and TVaR							
NJ Risks	**VAR α**	**TVAR α**	**Premiums**	**BR Risks**	**VAR α**	**TVAR α**	**Premiums**
risk 1	0.0037	0.0511	1,314,438	risk 11	0.0910	0.2969	323,495
risk 2	0.0054	0.0489	750,127	risk 12	0.0050	0.0600	6,601,950
risk 3	0.0121	0.0545	1,160,409	risk 13	0.0050	0.0600	6,601,950
risk 4	0.0236	0.1045	3,391,682	risk 14	0.0884	0.2927	97,815
risk 5	0.0235	0.1045	3,391,682	risk 15	0.0987	0.3198	169,304
risk 6	0.0202	0.0405	8,560,567	risk 16	0.0692	0.2294	579,865
risk 7	0.0622	0.1712	6,278,738	risk 17	0.0904	0.3148	148,851
risk 8	0.0622	0.1722	6,253,646	risk 18	0.0078	0.0687	9,082,536
risk 9	0.0117	0.0432	174,465	risk 19	0.0718	0.2359	80,000
risk 10	0.0032	0.0454	1,203,113	risk 20	0.0901	0.3106	242,799
Total Line	0.0205	0.0738	32,478,869		0.0069	0.1675	23,928,565

It is evident that in a quantile premium pricing practice, where the policy premium is derived purely from a VaR or TVaR value following

$$\pi_T[r_N] = \frac{1}{1-\alpha}\int_{\alpha}^{1} VaR(S_n)dt \tag{13.2}$$

For the quantile premium to approach the traditional premium computed from expected value and standard deviation as in expression (1.1), the exceedance probability α in the premium pricing formula (13.2) needs to vary significantly by each insured risk. This may create an issue for practitioners when such probability tolerance is defined by risk in underwriting guidelines, and will not stay constant for the whole book of business or unit. Furthermore we proved that to measure dependencies and diversification for an insurance book of business (see expression 12.0) single policies' premiums need to be derived through back-allocation from a total accumulated dependent line – unit premium, through a probabilistic technique, as we do in expression (11.0), using a standard deviation ratio. Still the exceedance probability of an insurance premium mapped as a VaR and TVaR metric is practical and very useful in capital reserving tasks. It identifies scenarios with a probability weight α where policy loss in a single scenario – VaR, or on average – TVaR could exceed the policy premium.

$$\pi_T[r_N] \leq VaR_\alpha(S_n) \leq TVaR_\alpha(S_n)$$

The data scale - size dimension of the big data component to support such task at a portfolio and business unit level is the availability of the full per-risk insurance loss simulations. The frequency dimension of the big data component is contained in updating and preserving full insurance loss simulations for every task, as the practitioner varies underwriting parameters such as load factors or exceedance probability thresholds.

5.0 Functions and Algorithms for Insurance Data Components

5.1 Definition of Insurance Big Data Components

Large insurance data component facilitate and practically enable the actuarial and statistical tasks of measuring dependencies, modeled loss accumulations and back-allocation of total business unit premium to single risk policies. For this study our definition of big insurance data components covers historical and modeled data at high geospatial granularity, structured in up to one million simulation geo-spatial maps. For modeling of a single (re)insurance product for a single or few insured risks, a single map can contain a few hundred historical and modeled physical measure data points, such as water depth in the case of flood insurance. For a large book of business or a portfolio simulation, one map may contain millions of such data points. Time complexity is another feature of big data. Global but structured and distributed data sets are updates asynchronously and oftentimes without a schedule, depending on scientific and business requirements and computational resources. Thus such big data components have a critical and indispensable role in defining competitive premium cost savings for the insureds, which otherwise may not be found sustainable by the policy underwriters and the insurance firm.

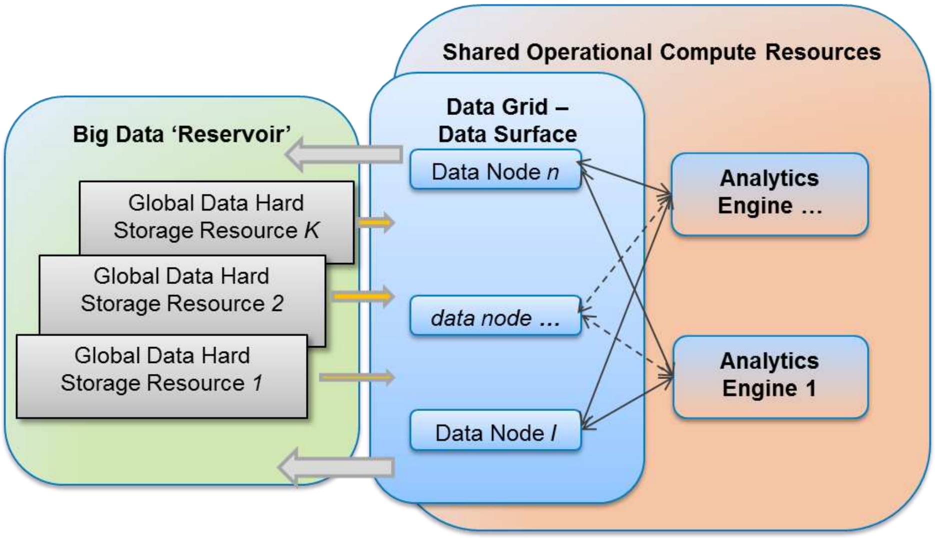
5.2 Intersections of Exposure, Physical and Modeled Simulated data sets

Fast compute and big data platforms are designed to provide various complex, and demanding on computational resources geospatial modeling and analysis tasks. One such fundamental task is the projection of an exposure map of insured risks and computing of its intersection with multiple simulated stochastic flood intensity scenarios and geo-physical properties maps containing attributes such as coastal and river banks elevations and distances to water bodies. Such big data algorithms will typically perform as a first step a spatial cashing and indexing of all latitude and longitude geo-coded units and grid-cells with all-and-any attributes relevant to the required intersection definition of insured risk exposure and modeled stochastic flood intensity. Geo-spatial interpolation is also employed to compute and adjust peril intensities to distances and geo-physical elevations of the insured risks. In a second step a distance based computation between indexes with insured risk attributes and those with modeled intensity attributes derives the intersection of the scenario simulation and the insured risks map, so that further data operations and analytics are performed only on this smaller data sub-set.

5.3 Reduction and Optimization through Mapping and Parallelism

One relevant definition of Big Data to our own study is datasets that are too large and too complex to be processed by traditional database technologies and algorithms. In principle moving data between processes and algorithms or between platforms is the most computationally expensive task in solving big geo-spatial scale problems. Two such tasks in todays' insurance firms' workflow are modeling and measuring inter-risk dependencies and diversification within an insurance portfolio. The cost and expense of big geo-spatial solutions is magnified by the size of required data sets typically being distributed across multiple hard physical computational environments as a result of their large scale and structure. The fundamental solution is to achieve distributed optimization, which is constructed by a sequence of algorithms. As a first step a mapping and splitting algorithm will divide large data sets into sub-sets and perform statistical and modeling computations on the smaller sub-sets. In our computational case study for flood insurance the smaller data chunks represent insurance risks and policies in geo-physically dependent zones, such as river basins and coastal segments. The smaller data sets are processed as smaller sub-problems in parallel by assigned and managed sufficiently appropriate computational resources. In our case study, following these principles, we solve smaller scale and chunked data sets computations for flood intensity and then for modeling and estimating of fully simulated and probabilistic insurance loss. Once the cost effective sub-set operations are complete on the smaller sub-sets, a second algorithm will collect and map together the results of the first stage compute for consequent next tier and higher level operations and data analytics.

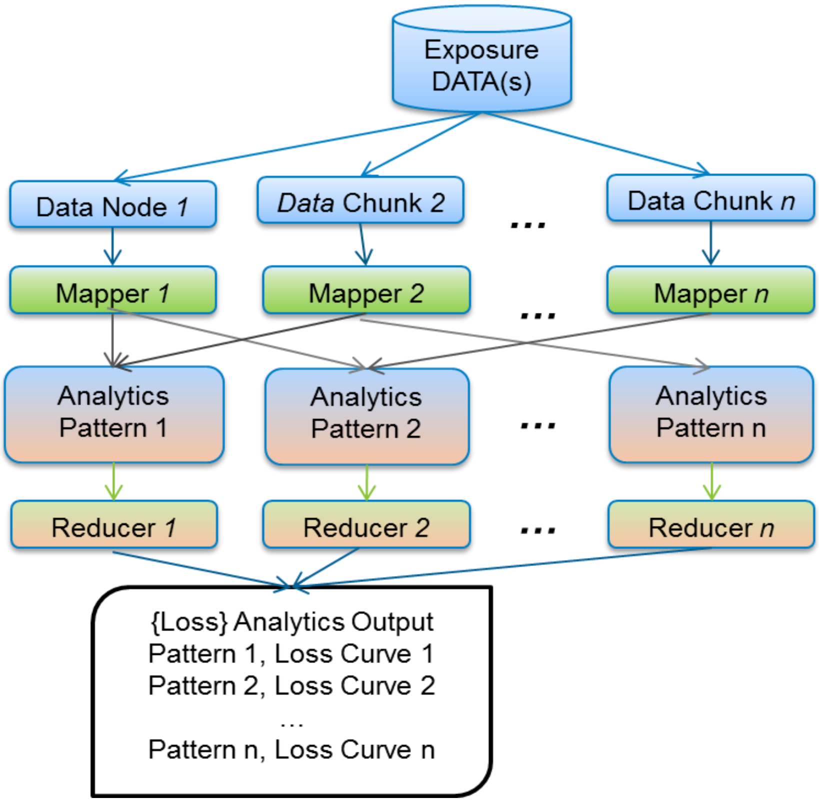
Figure 1: Distributed Computational Resources, Storage and Data Grid Framework


Shared Operational Compute Resources
Big Data 'Reservoir'
Data Grid – Data Surface
Data Node *n*
Global Data Hard Storage Resource *K*
Global Data Hard Storage Resource 2
Global Data Hard Storage Resource *1*
data node …
Data Node *l*
Analytics Engine …
Analytics Engine 1

For single insurance products, business units and portfolios an ordered accumulation of risks is achieved via mapping and controlling the order by scale of the strength or lack thereof inter-risk dependencies. Data sets and algorithmic tasks with identical characteristics could be grouped together and resources for their processing significantly reduced by avoiding replication or repetition of computational tasks, which have already been mapped and now can be reused. Post-analytics and post-processed data could also be distributed on different physical storage capacities by a secondary scheduling algorithm, which intelligently allocates chunks of modeled and post-processed data to available storage resources. This family of techniques is generally known as MapReduce.

.

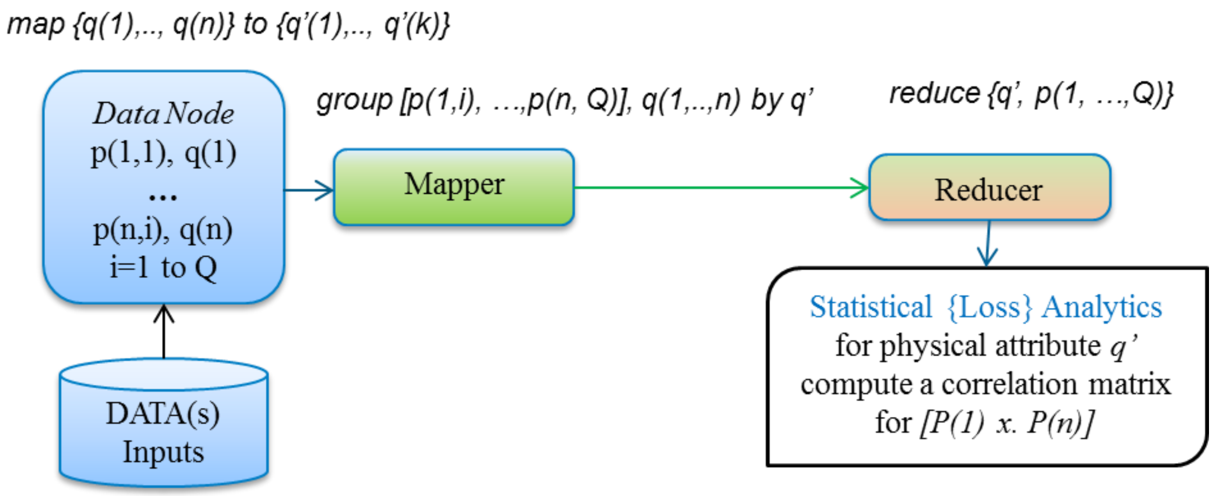
Figure 2: Conceptual View of MapReduce Algorithm in Loss Estimation Analytics



In our flood risk modeling case studies, one application of this family of optimization algorithms is found very appropriately in computing intersections and reducing dimensionality of big geo-spatial data and simulation problems. In more formal terms, we need to build an intersection and dimensionality reduction optimization algorithm, for e set of insured flood risks $\{1, \ldots, n\}$ with geo-spatial co/ordinates $\{q(1), \ldots, q(n)\}$, subject to a flood intensity simulation, which in practice measures flooding water depth: $\{p(1,1), \ldots, p(n, Q)\}$ for a simulation of size: $i=1$ *to* Q. In the mapping phase of the algorithm, we build geo-spatial polygons q', which cluster nearby insured risks from the whole dataset: $\{q(1), \ldots, q(n)\}$ by some distance measure $\{d\}$. In the second grouping step of the algorithm, for the polygons $\{q'(1), \ldots, q'(k)\}$, who now cumulatively cover the entire geo-spatial distribution of insured risks, we create a sub-set of the simulation $[Q]$.

Thus for one and any geo-polygon *q'(1)*, which contains *m* insured risks, we have reduced the required simulated data for analytics operations from *{p(1,1),…,p(n, Q)}* to m**p(1, …,Q)*.

Figure 3: Mathematics workflow in Mapping and Reduction Algorithms


map {q(1),.., q(n)} to {q'(1),.., q'(k)}
Data Node
p(1,1), q(1)
…
p(n,i), q(n)
i=1 to Q
DATA(s)
Inputs
group [p(1,i), …,p(n, Q)], q(1,..,n) by q'
Mapper
reduce {q', p(1, …,Q)}
Reducer
Statistical {Loss} Analytics
for physical attribute q'
compute a correlation matrix
for [P(1) x. P(n)]

With this optimization approach computing insured losses and correlation matrices within each polygon and subsequently for the entire geo-spatial distribution of risks becomes a much more manageable and sustainable proposition.

5.4 Scheduling and Synchronization by Service Chaining

Distributed and service chaining algorithms process geo-spatial analysis tasks on data components simultaneously and automatically. For logically independent processes, such as computing intensities or losses on uncorrelated scenarios of a simulation, service chaining algorithms will divide and manage the tasks among separate computing resources. Dependencies and correlations among such data chunks may not exist because of large geo-spatial distances, as we saw in some of the modeling and pricing scenarios in our cases studies. Hence they do not have to be modeled explicitly and performance improvements are gained immediately. For such scenarios both input data and computational tasks can be broken down to pieces and sub-tasks respectively. For logically inter-dependent tasks, such as accumulations of inter-dependent quantities such as losses in geographic proximity, chaining algorithms automatically order the commencement and completion of dependent sub-tasks. In our modeled scenarios, the simulated loss distributions of risks in immediate proximity are accumulated first, where dependencies are expected to be the strongest. A second tier of accumulations for risks with partial dependence due to longer distances, and full independence measures is scheduled for once the first tier of accumulations of highly dependent risks is complete. Service chaining methodologies work in collaboration with auto-scaling memory algorithms, which provide or remove computational memory resources, depending on the intensity of modeling and statistical tasks. Challenges still are significant in processing shared data structures. An insurance risk management example, which we are currently developing for our a next working paper, would be pricing a complex multi-tiered product, comprised of many geo-spatially dependent risks, and then back-allocating a risk metric, such as tail-value-at-risk TVaR down to single risk granularity. On the statistical level this back-allocation and risk management task involves a process called de-convolution or also component convolution. A computational and optimization challenge is present when highly dependent and logically connected statistical operations are performed with chunks of data distributed across different hard data storage resources. Solutions are being developed for multi-threaded implementations of map-reduce algorithms, which address such computationally intensive tasks. In such procedures the mapping is done by task definition and not directly onto the raw and static data.

Some Conclusions and Further Work

With advances in computational methodologies for natural catastrophe and insurance portfolio modeling, practitioners are producing increasingly larger data sets of modeled physical, loss and risk metrics. Simultaneously single product and portfolio optimization techniques are used in insurance premium underwriting, which take advantage of metrics in diversification and inter-risk dependencies. Such optimization techniques significantly increase the frequency of production of insurance underwriting data, and require new types of algorithms, which can process multiple large, distributed and frequently updated sets. Such algorithms have been developed theoretically and now they are entering from a proof of concept phase in the academic environments to implementations in production in the modeling and computational systems of insurance firms.

Both traditional statistical modeling methodologies such as premium pricing, and new advances in definition of inter-risk variance-covariance and correlation matrices and policy and portfolio accumulation principles, require significant data management and computational resources to account for the effects of dependencies and diversification. Accounting for these effects allows the insurance firm to support cost savings in premium value for the insurance policy holders.

With many of the reviewed advances at present, there are still open areas for research in statistical modeling, single product pricing and portfolio accumulation and their supporting optimal big insurance data structures and algorithms. Algorithmic communication and synchronization cost between global but distributed structured and dependent data is expensive. Optimizing and reducing computational processing cost for data analytics is a top priority for both scientists and practitioners. Optimal partitioning and clustering of data, and particularly so of geospatial images, is one other active area of research.

References

Ayma, V. A. et al (2015) Classification of Algorithms for Big Data Analysis, A Map Reduce Approach, Remote Sensing and Spatial Information Sciences Conference 2015

Fedak, Jilles (2013) MapReduce Runtime Environments, INRIA, University Of Lyon, France

Goovaerts, Mark, Laeven Roger (2011) Premium Calculation and Insurance Pricing

Hurlimann, Werner (2006) On a Robust Parameter Free Pricing Principle: Fair Value and Risk Adjusted Principle

Jin, B. X, et al (2015) Building Spatiotemporal Cloud Platform for Supporting GIS Application, International Workshop on Spatiotemporal Computing 2015

Isaac, Luke P. (2014) Basics of Map Reduce Algorithm Explained with a Simple Example, Geek Stuff, May 2014

Nandakumar, A. N. et al (2014) A Survey of Data Mining Algorithms on Apache Hadoop, International Journal of Emerging Technology and Advanced Engineering, Volume 4, Issue 1, January 2014

Nivranshu, Hans, (2015) Big Data Clustering Using Genetic Algorithm On Hadoop MapReduce, International Journal of Emerging Technology and Advanced Engineering, Volume 4, Issue 04, April 2015

Rau Chaplin, Andrew (2015) Scaling up to Big Data: Algorithmic Engineering + HPC, Statistical and Computational Analytics for Big data Conference 2015