Segmented Individual Claim Reserve Estimation using BART Refinement

Adhitya Ronnie Effendie

Actuarial Science Study Program, Department of Mathematics, Universitas Gadjah Mada, Yogyakarta

E-mail: adhityaronnie@ugm.ac.id

Amran Rahim

Department of Mathematics, Faculty of Mathematics and Natural Sciences University of Hasanuddin, Makassar

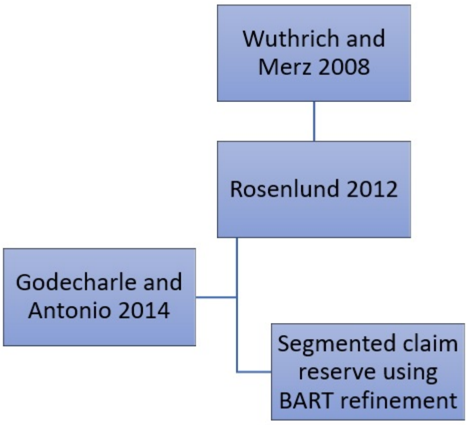
E-mail: amran@science.unhas.ac.id

Abstract. We want to extend an individual claim reserving method proposed by (Rosenlund 2012) and also (Godecharle and Antonio 2014) by using segmented calculation. This method is an individual method of claims reserve estimation which involves detailed condition on claim characteristics in the calculation process. Data is divided into several segments according to combination of background variables. We then apply RDC method to find estimated IBNR and RBNS reserves for each segment. Bayesian Additive Regression Tree (BART) is used to refine the estimated reserves. This is because the estimation become unstable due to a lot of combination factors for each segment.

Keywords: Individual claim reserving method, IBNR, Bayesian Additive Regression Tree

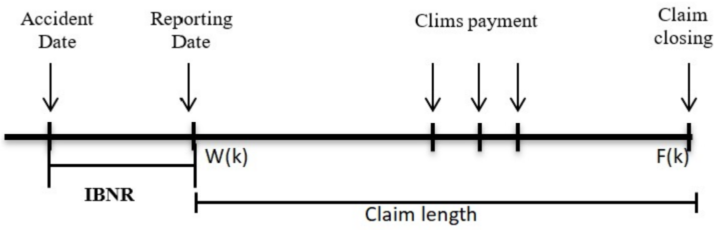
1. Introduction

Starting point of this paper is the Reserve by Detailed Conditioning (RDC) method, as introduced by (Rosenlund, 2012). RDC - in its original specification - is a deterministic reserving method, designed for individual claims in discrete time. A remarkable and innovative aspect of the method is its ability to condition on claim characteristics, which are used for identification or clustering of similar claims. Conditional on a specific set of claim characteristics, a best estimate for the reserve attached to an open claim is obtained from the observed, historical development of claims from the same cluster, hence with similar characteristics (Effendie, A.R., Pebriawan, R. 2017).



2. The Claim Process

In general, the loss reserve is the total outstanding payments of all incurred claims, whether reported or not. In other words, it is an aggregation of the outstanding payments for every single claim. The claim process reflects the dynamics of the development of a single claim and is discussed in (Wüthrich and Merz,2008).



Following (Rosenlund 2012) we determine reserves by conditioning on claims characteristics. These characteristics summarize information registered during the development of a claim. They allow for the identification of similar claims. In this work we consider the claim length, the last observed cumulative payment and the reporting delay as claim characteristics.

i	**Claim ID**	**Development period** j					
		1	2	3	$\cdots$	$n-1$	n
1	$c_{1,1}$	$Y(c_{1,1},1)$	$Y(c_{1,1},2)$	$Y(c_{1,1},3)$	$\cdots$	$Y(c_{1,1},n-1)$	$Y(c_{1,1},n)$
	$c_{1,2}$	$Y(c_{1,2},1)$	$Y(c_{1,2},2)$	$Y(c_{1,2},3)$	$\cdots$	$Y(c_{1,2},n-1)$	$Y(c_{1,2},n)$
	$\vdots$	$\vdots$					
2	$c_{2,1}$	$Y(c_{2,1},1)$	$Y(c_{2,1},2)$	$Y(c_{2,1},3)$	$\cdots$	$Y(c_{2,1},n-1)$	
	$c_{2,2}$	$Y(c_{2,2},1)$	$Y(c_{2,2},2)$	$Y(c_{2,2},3)$	$\cdots$	$Y(c_{2,2},n-1)$	
	$\vdots$	$\vdots$					
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		
n	$c_{n,1}$	$Y(c_{n,1},1)$					
	$c_{n,2}$	$Y(c_{n,1},1)$					
	$\vdots$	$\vdots$					

2.1. Claim Length

The claim length is the duration from claim reported up to claim finalized. We denote the length of claim k $(k = 1, 2, \cdots, N)$ by $L(k)$ and define it as follow:

$$L(k) = F(k) - W(k) + 1. \quad (1)$$

Conditional claim length probability:

$$P(L=\lambda|L>t)=\Big[\prod_{k=t+1}^{\lambda-1}P(L>k|L\geq k)\Big]P(L=\lambda|L\geq\lambda) \tag{2}$$

for $0\leq t\leq n-1$ and $t+1\leq\lambda\leq n$.

2.2. Mean payment
Define the sum of amounts paid up to and including period t from reporting as

$$H(t)=\sum_{h=1}^{t}Y(h+W-1), t=0,1,\cdots,n \tag{3}$$

where h is counted from reporting with the reporting period W having $h=1$. We want to predict the expected remaining payment sum from the known sum. Consider this expression:

$$E[H(L)-H(t)|L>t,H(t),W] \tag{4}$$

For $t=n-i-W+2$ an estimate of this expression gives the RBNS (Reported But Not Settled) reserve of a reported open claim. For $t=0$ we obtain the IBNR (Incurred But Not Reported) reserve per claim.

2.3. Rosenlund's Estimator of claim reserve
Define the underlying reserve for a claim as

$$R(q,w,t)=E[H(L)-H(t)|L>t,Q_t=q,W\wedge w_0=w] \tag{5}$$

$$\hat{R}(q,w,t)=\sum_{\lambda=t+1}^{n}\sum_{h=t+1}^{\lambda}\hat{p}_{\lambda}(q,w,t)\hat{\mu}_{\lambda h}(q,w,t) \tag{6}$$

where

$$p_{\lambda}(q,w,t)=P(L=\lambda|L>t,Q_t=q,W\wedge w_0=w) \tag{7}$$

is the probability of claim length and

$$\mu_{\lambda h}(q,w,t)=E[Y(h+W-1)|L=\lambda,Q_t=q,W\wedge w_0=w] \tag{8}$$

is the expected of claim payment for $0\leq t\leq n-1$, $t+1\leq\lambda\leq n$ and $t+1\leq h\leq\lambda n$.

3. Data

The dataset that we used in this research contains information on 58,573 individual claims of BPJS Kesehatan (Indonesian Social Health Insurance) during 2015. 51,978 claims are closed and remaining 6,595 claims are open. All claims come from Special Capital Region of Jakarta, reported from 4 major public hospitals with different specialties (Heart center, Cancer center, Children and Mother hospital and general hospital) There are four background variables available from the data:

- **Prov (f1)**: Provider: 1. Cipto Mangunkusumo general hospital, 2. Dharmais National Cancer Center, 3. Harapan Kita National Heart Center and 4. Harapan Kita Children and Mother hospital.

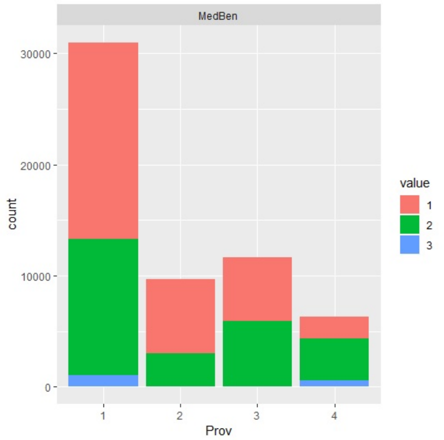


Figure 1. Figure caption for first of two sided figures.

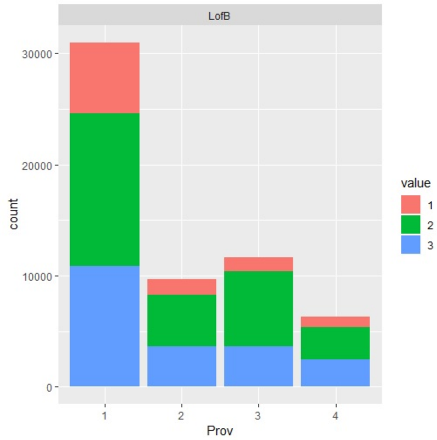


Figure 2. Figure caption for second of two sided figures.

- **LofB (f2)**: Line of Business: 1. PBI (Funded by Government) 2. PPU (Employee obligation) 3. PBPU (Other sources)
- **MedBen (f3)**: Type of Medical Benefit: 1. Procedure, 2. Non-procedure, 3. Maternity
- **Resint (f4)**: Resource intensity: 1. Low, 2. Medium, 3. High, 4. Outpatient

Claim currency was converted to Canadian dollar with currency rate on September 1, 2018. The mean, median and standard deviation of severity was \$2,245.9, \$722.3 and \$7,585.59, respectively.

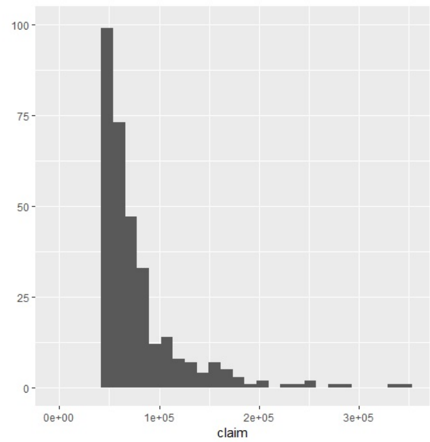


Figure 3. Figure caption for first of two sided figures.

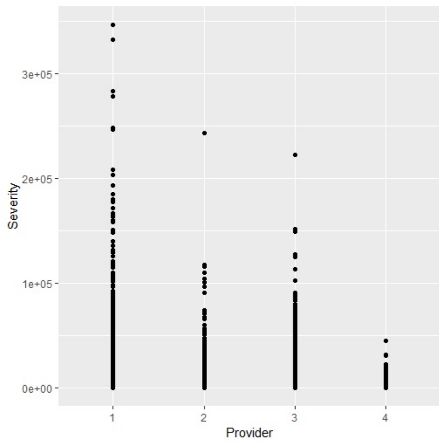


Figure 4. Figure caption for second of two sided figures.

3.1. Aggregate Incremental Run-Off Triangle

We will show the following aggregate incremental Run-off triangle of the data:

3446688	6179760	2909263	386201	210011	74709	67606	125005	212358	185341	185983	132102
5012170	5372674	494351	327627	256527	111248	181139	228289	211127	240085	185512	
5200872	5700927	500055	368327	563105	218751	502796	340458	352503	351399		
5399711	5431463	323241	601192	321764	344862	410149	424878	284449			
2077345	5323001	3656416	486279	109214	210426	367260	378070				
1592788	5341758	3749137	107095	564520	549810	639113					
3010580	4498306	1487821	531795	618776	401084						
1439465	4230602	3052373	978017	780673							
2830825	4363460	2163495	876193								
3726501	4793857	1067356									
3259496	4044510										
3863873											

Here we give full (cumulative) triangle, computed by classical (Chain Ladder) method:

3446688	9626448	12535711	12921912	13131923	13206632	13274238	13399243	13611601	13796942	13982925	14115027	14323848
5012170	10384844	10879195	11206822	11463349	11574597	11755736	11984025	12195152	12435237	12620749	12739982	12928460
5200872	10901799	11401854	11770181	12333286	12552037	13054833	13395291	13747794	14099193	14298863	14433950	14647489
5399711	10831174	11154415	11755607	12077371	12422233	12832382	13257260	13541709	13807659	14003201	14135494	14344618
2077345	7400346	11056762	11543041	11652255	11862681	12229941	12608011	12864949	13117608	13303377	13429060	13627732
1592788	6934546	10683683	10790778	11355298	11905108	12544221	12841541	13103239	13360578	13549788	13677798	13880150
3010580	7508886	8996707	9528502	10147278	10548362	10859413	11116801	11343350	11566126	11729923	11840740	12015915
1439465	5670067	8722440	9700457	10481130	10724899	11041157	11302852	11533192	11759697	11926235	12038907	12217013
2830825	7194285	9357780	10233973	10626802	10873960	11194612	11459945	11693487	11923139	12091992	12206230	12386811
3726501	8520358	9587714	10059341	10445467	10688408	11003589	11264393	11493950	11719684	11885656	11997944	12175444
3259496	7304006	8971874	9413208	9774532	10001868	10296804	10540856	10755668	10966903	11122214	11227289	11393389
3863873	9637297	11837971	12420291	12897041	13197000	13586154	13908171	14191605	14470319	14675244	14813887	15033047
LDF												
2,494	1,228	1,049	1,038	1,023	1,029	1,024	1,020	1,020	1,014	1,009	1,015	

The outstanding reserve is CAD 27,425,948

4. RDC Method

Expand the triangle into individual run-off triangle:

Row	ID	accident	report	final	Status	f1	f2	f3	f4	dev1	dev2	dev3	dev4	dev5	dev6	dev7	dev8	dev9	dev10	dev11	dev12
1	6452	1	12	NA	O	1	2	1	2	0	0	0	0	0	0	0	0	0	0	0	12480777
2	6453	1	12	NA	O	1	3	1	1	0	0	0	0	0	0	0	0	0	0	0	3961750
3	6454	1	12	NA	O	1	3	1	1	0	0	0	0	0	0	0	0	0	0	0	3855043
4	6455	1	12	NA	O	1	3	1	1	0	0	0	0	0	0	0	0	0	0	0	3190220
5	6456	1	12	NA	O	1	3	1	1	0	0	0	0	0	0	0	0	0	0	0	1723164
6	6457	1	12	NA	O	1	2	1	1	0	0	0	0	0	0	0	0	0	0	0	1723164
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
58568	159370	12	12	NA	O	1	2	1	4	1191458	0	0	0	0	0	0	0	0	0	0	0
58569	159371	12	12	NA	O	1	2	1	4	1191458	0	0	0	0	0	0	0	0	0	0	0
58570	159372	12	12	NA	O	1	1	1	4	1191458	0	0	0	0	0	0	0	0	0	0	0
58571	159373	12	12	NA	O	1	3	1	4	1191458	0	0	0	0	0	0	0	0	0	0	0
58572	159374	12	12	NA	O	1	3	1	4	1191458	0	0	0	0	0	0	0	0	0	0	0
58573	159375	12	12	NA	O	1	1	1	4	1191458	0	0	0	0	0	0	0	0	0	0	0

4.1. Claim characteristics

Basic statistics of payment delay (developments):

	dev1	dev2	dev3	dev4	dev5	dev6	dev7	dev8	dev9	dev10	dev11	dev12
Min.	0	0	0	0	0	0	0	0	0	0	0	0
Median	32,55	168,4	0	0	0	0	0	0	0	0	0	0
Mean	697,6	943,8	331,27	79,61	58,47	32,62	58,47	25,55	18,11	13,26	6,342	2,255
Max.	173592,86	173592,9	141894,71	37085,8	37085,8	22759,33	22759,33	13195,65	11614,38	10042,24	10042,24	10042,24

Some options: $w_0 = 3, 7, 12$ and $q_0 = 10, 15, 37$ (Sturgess)

4.2. RDC results

w0	q0	IBNR	RBNS	Total
3	10	18.553.258	6.749.386	25.302.644
3	15	18.553.258	6.986.395	25.539.653
3	37	18.553.258	6.814.946	25.368.204
7	10	17.913.593	6.989.734	24.903.327
7	15	17.913.593	7.255.622	25.169.215
7	37	17.913.593	7.074.267	24.987.860
12	10	17.478.848	6.895.006	24.373.854
12	15	17.478.848	7.150.855	24.629.703
12	37	17.478.848	6.965.673	24.444.521

4.3. Comments on RDC results
In general, RDC method has lower claim reserve estimation compare to Chain Ladder method. This is agree with some opinion that Chain Ladder method is over estimate. At the same maximum claim reported period, w_0, IBNR result is not change. IBNR result going down as w_0 increase. Within the same w_0, RBNS initially increase but at some quantiles it reaches its asymptotic value. As w_0 increases, RBNS result is also increase. As this model doesn't count the effects of several background variables (rating factors) and the estimate is much lower than standard method (Chain Ladder), we may think that the result of standard RDC method is under estimate. Need a new method that count the effect of rating factors and adjust the estimate value from its base factor

5. RDC Segmented Calculation
The basic idea of RDC segmented calculation method is, we calculate RDC claim estimation for every segment (i.e. every combination of background variable). In this case we will have response variable for every combination of background variable (excluding the zero combinations). We can calculate claim estimation directly from this result, but will give "raw estimate". We need a "smoothing" method, to smooth the result from segmented calculation.

5.1. Bayesian Additive Regression Tree (BART) overview
BART is a Bayesian approach to nonparametric function estimation using regression trees. Regression trees rely on recursive binary partitioning of predictor space into a set of hyper-rectangles in order to approximate some unknown function f. Predictor space has dimension of the number of variables. Tree-based regression models have an ability to flexibly interactions and nonlinearities. Models composed of sums of regression trees have an even greater ability than single trees to capture interactions and non-linearities as well as additive in f. We choose package **bartMachine** from R library

5.2. RDC-BART Segmented Calculation
We divide the method into the following stages:

- **RDC segmentation** Group data into segments. Each segment represents unique combination of background variables. In this case we have $4x3x3x4 = 144$ combination of background variables.Then choose appropriate w_0 and q_0 and apply RDC method to get IBNR and RBNS estimate for each segment.
- **bartMachine** setup: clean data that was obtained from previous stage from NAs, build response vector (we take log of the reserve as continuous response) and predictor matrix (set of four categorical variables), setting Java heap (up to 5GB of RAM) and setting number of core used (we use 4 cores).

5.3. RDC-BART Segmented Calculation
BART model building: Setting hyperparameters (in our case) : $m = 50, \alpha = 0.95, \beta = 2, k = 2, q = 0.9, \nu = 3$.Set probabilities of the GROW/ PRUNE/ CHANGE steps to 28% / 28% / 44%. Set the number of burn-in Gibbs samples to 250 and number of post-burn-in samples to 1,000. Set the covariates to be equally important *a priori*

5.4. RDC-BART Segmented Calculation
R output

```
#for IBNR
> bart_machine1
bartMachine v1.2.3 for regression
training data n = 709 and p = 26
```

```
built in 1.4 secs on 4 cores, 50 trees, 250 burn-in and 1000 post. samples
sigsq est for y beforehand: 1.999
avg sigsq estimate after burn-in: 0.3064
in-sample statistics:
 L1 = 254.45
 L2 = 166.99
 rmse = 0.49
 Pseudo-Rsq = 0.9291
p-val for shapiro-wilk test of normality of residuals: 0
p-val for zero-mean noise: 0.98395
#for RBNS
> bart_machine2
bartMachine v1.2.3 for regression
training data n = 497 and p = 26
built in 0.8 secs on 4 cores, 50 trees, 250 burn-in and 1000 post. samples
sigsq est for y beforehand: 1.658
avg sigsq estimate after burn-in: 0.70499
in-sample statistics:
 L1 = 299.09
 L2 = 297.92
 rmse = 0.77
 Pseudo-Rsq = 0.8235
p-val for shapiro-wilk test of normality of residuals: 0
p-val for zero-mean noise: 0.97742
```

5.5. *RDC-BART Segmented Calculation Steps*

- Find the best model
- Predict the reserve (IBNR and RBNS) based on the winning model, (R2)
- Choose a rating factor as a base factor. Here we choose **Prov** as it is most directly connected to the expected loss rather than other available rating factors.
- Calculate the reserve with data from segmented base factor only (R1). Here we use standard RDC method
- Calculate final reserve estimation as $R3 = R2 \times \frac{\sum_{u=1}^{n} R1(u)}{\sum_{u=1}^{n} R2(u)}$ with n is the number of total period (12 in this case).

5.6. *Result*

	month1	month2	month3	month4	month5	month6	month7	month8	month9	month10	month11	month12	Total
IBNR	3367945	3749727	3193851	3049135	1905846	1567745	927956,3	1181085	67154,22	57619,69	31471,36	186,3662	19099721,9
RBNS	877371,8	1447403	1028357	1367744	326937	520898,1	295373,2	821208,1	25089,48	44074,01	25188,86	331,9048	6779976,75
Total	4245316,8	5197130	4222208	4416879	2232783	2088643	1223330	2002293	92243,7	101693,7	56660,22	518,271	25879698,7

So this method gives total reserve CAD 25,879,699 with $w_0 = 3$ and $q_0 = 10$

6. Summary

RDC-BART Segmented Method gives estimate slightly greater than standard RDC but still lower than Chain Ladder method. RDC-BART Segmented Method depends on several factors, including q_0 and w_0, hyper-parameters, number of trees, BART winning model. RDC-BART Segmented Method also depends on base factor we choose. Some improvements could be made including adding inference, RMSE, MSEP, change the base to categorical level and better algorithm to shorten execution time.

7. References

[1] Dorman L I 1975 *Variations of Galactic Cosmic Rays* (Moscow: Moscow State University Press) p 103

[2] Effendie, A.R., Pebriawan, R. 2017, *Estimation of IBNR and RBNS reserve by detailed conditioning method* (Far East Journal of Mathematical Sciences) vol 101(12), pp. 2785-2801

[3] Godecharle and Antonio 2014, *Reserving by conditioning on markers of individual claims: a case-study using historical simulation* (KU Leuven Faculty of Economics and Business)

[4] Rosenlund 2012, *Bootstrapping individual claim histories* (ASTIN Bulletin) vol 42 pp 291-324

[5] Wüthrich and Merz 2008, *Stochastic claims reserving methods in insurance* (John Wiley & Sons)