IBNR Project

Amedeo Zito

For the cours ACT-7005
Travail actuariel pratique en entreprise

Presented to professor

Ilie Radu Mitric

the 4^{th} of may 2020

Faculty of Science and Engineering
Actuarial Science
Laval University
Winter 2020

Summary

bla

Acknowledgement

Radu Mitric has my full gratitude for allowing me to do such a project within my master degree program.

In addition, I want to thank my entire team at Intact, Ying He, Constance Filiozat, Stephane Caron, Simon Valois and Jean-Denis Grenier, without it would not have been possible to develop the model and continue making it better.

Contents

List of Figures

List of Tables

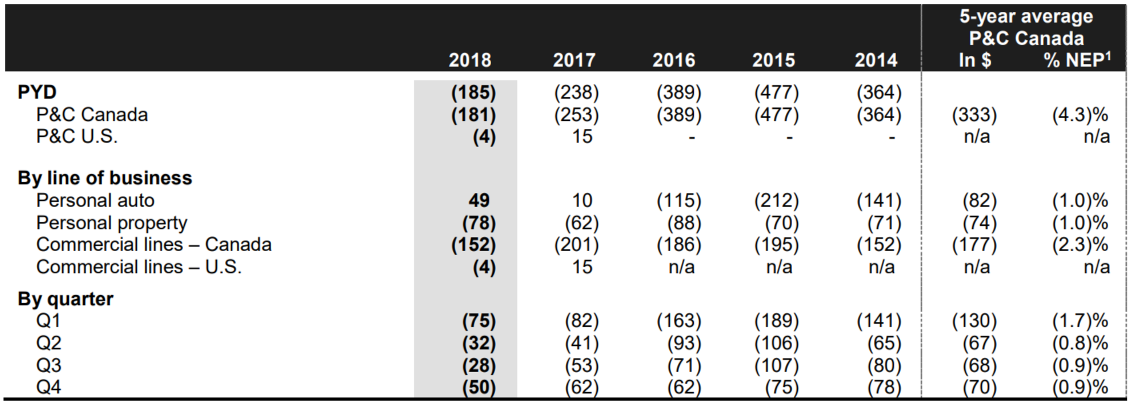
1 Introduction

This winter I was working for the Data Lab of Intact Insurance. Specifically, I joined the Claims squad. The squad develops programs and models which are used internally for process optimization and cost reduction.

....

2 Project objectives

In this section, we will briefly discuss the project and its objectives.



	2018	2017	2016	2015	2014	5-year average P&C Canada	
						In $	% NEP[1]
PYD	**(185)**	(238)	(389)	(477)	(364)		
P&C Canada	**(181)**	(253)	(389)	(477)	(364)	(333)	(4.3)%
P&C U.S.	**(4)**	15	-	-	-	n/a	n/a
By line of business							
Personal auto	**49**	10	(115)	(212)	(141)	(82)	(1.0)%
Personal property	**(78)**	(62)	(88)	(70)	(71)	(74)	(1.0)%
Commercial lines – Canada	**(152)**	(201)	(186)	(195)	(152)	(177)	(2.3)%
Commercial lines – U.S.	**(4)**	15	n/a	n/a	n/a	n/a	n/a
By quarter							
Q1	**(75)**	(82)	(163)	(189)	(141)	(130)	(1.7)%
Q2	**(32)**	(41)	(93)	(106)	(65)	(67)	(0.8)%
Q3	**(28)**	(53)	(71)	(107)	(80)	(68)	(0.9)%
Q4	**(50)**	(62)	(62)	(75)	(78)	(70)	(0.9)%

Figure 1: Unfavourable (favourable) prior year development, [Intact, 2018]

The IBNR project arises from the results of [Intact, 2018] annual report, see figure 1. The prior year development (PYD) of the Personal auto line is at 49 million of which 20 million are auto physical damage. PYD represents the change in total prior year claims liabilities during a specific period, in this case 2018. An increase in claims liabilities is referred to as an unfavorable prior year development. This means that the actuarial department underestimated the claim losses by 49 million. Even if percentage wise this is not very significant, it still is a large amount for a line of business, which should not be fluctuating as much. Such unfavourable development is not desirable and therefore Intact's higher management launched an investigation regarding the origin of this development. They decided that my team should investigate the issue and develop a new model which should exist in parallel with the model of the actuarial department. This project started in summer 2019. I was involved in this project by autumn 2019.
The idea is to have a second model with allows the actuarial department to asses if their model works correctly. If both models converge, they can have more confidence in their booked numbers. If the discrepancies are to large, it will trigger further investigation. It is important to note that the booked PYD is shown in the Intact annual reports and is often used by investor to determine Intact's performance.

In addition, the actuarial department wants a model which is interpretable and comprehensive. At this stage, a black box model is not a solution, since it does not allow an exact understanding of the results. The model itself uses historical claims data in order to predict the incurred but not reported (IBNR) personal auto claims for a specific month. The actuarial department uses an advanced chain-ladder approach. We were asked to find a different method which we will discus in more detail in section 4. Consequently, the main objectives of this project are:

- Develop a model which outperforms the current model used for booking the PYD.
- The model should be interpretable and not a black box
- The model should be dynamic and able to capture recent data changes

Before diving into the model itself, we have to fully understand the data the model will use for the predictions. Thus, in the next section, we will analyse the data we use for our model.

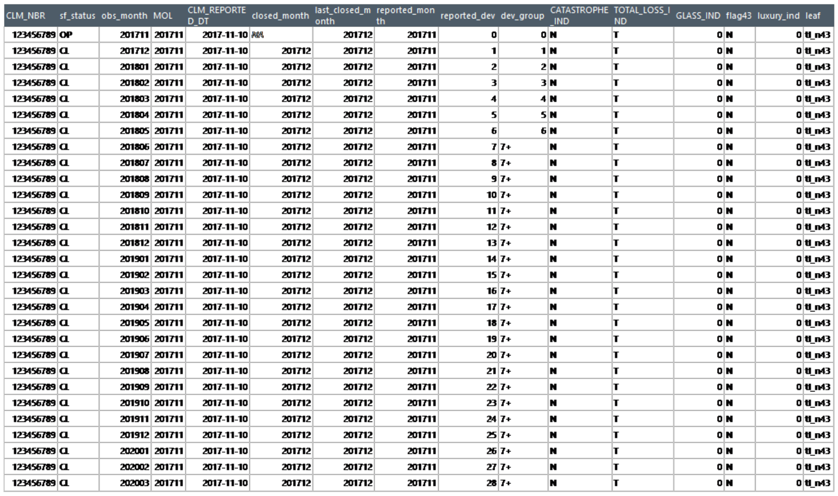
3 Data Analysis

In this section, we will deep dive into the data used in our model. Frist, we discuss the available datasets and the relevant columns. Then, we will show some charts with interesting patterns.

We have over one million lines of Canada wide monthly claims data, starting in january 2016 until today. We can not use data earlier than 2016, since prior 2016 claims were registered in an older system. This significantly changes the underlying claim distribution and makes prior 2016 data non-representative of future data.
The data is divided into databases for each region and line of business. The regions we will cover are Quebec ("QC"), Ontario ("ON") and Alberta ("AB"). The two line of business we cover are physical damage ("PHYSDAM") and liability ("LIPD"). The former consists of collisions and comprehensive coverage (theft, vandalism etc.), while the latter includes all damage caused by the insured to a third party. Note that in Quebec due to regulatory differences there is not separation between the two line of businesses. In Quebec the insurance company covers the loss only for its own insured independent of the responsibility and accountability. We name the single line of business in Quebec "PDPD". We will adjust our model hyper-parameters to each of the regions and line of businesses.

3.1 Data sample

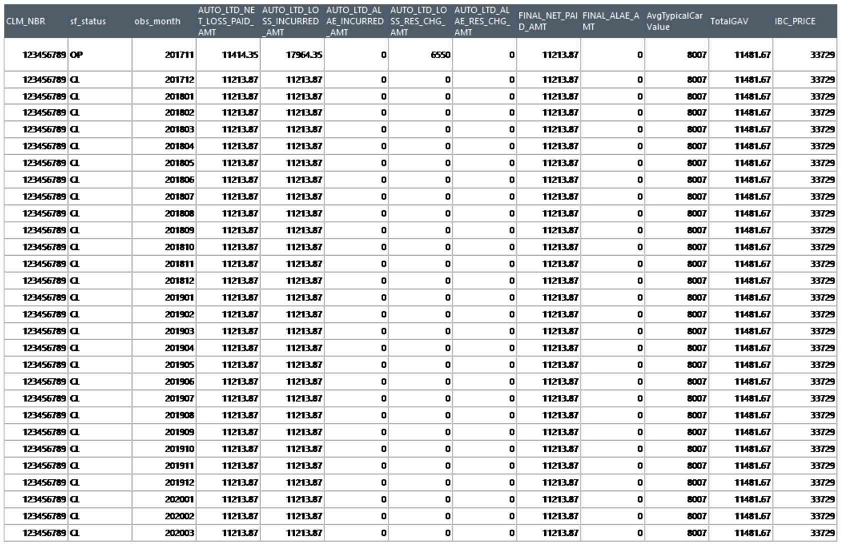
The dataset has over 120 columns, so we have to first determine what variables are relevant for us. The figures x and y show and extract for a fictive claim number. The claim number is unique for each claim. Each line represents a month of observation (`obs_month`) and is the snapshot of that claim in that specific month. With exception of the variables `last_closed_month`, `FINAL_NET_PAID_AMT` and `FINAL_ALAE_AMT`, all the information shown is the information we would have for that month of observation, while the three mentioned variables is information we know today (after the observation month). `sf_status` is the variable that indicated if the claim is open or closed at the observation month. On figure 2, we also have the month of loss (`MOL`), the reported date (`CLM_REPORTED_DT`) and the month of closure (`closed_month`). `last_closed_month` is the month at which the claim closed for the last time, since claims can reopen, this is important information we don't have when we predict the IBNR. `reported_dev` is the number of months since the claim has been reported, i.e. the age of the claim. `CATASTROPHE_IND`, `TOTAL_LOSS_IND`, `GLASS_IND`, `flag_43` and `luxury_ind` are variables we use to classify our data into `leaf`. We will discuss this in further detail in section 4.



CLM_NBR	sf_status	obs_month	MOL	CLM_REPORTED_DT	closed_month	last_closed_month	reported_month	reported_dev	dev_group	CATASTROPHE_IND	TOTAL_LOSS_IND	GLASS_IND	flag43	luxury_ind	leaf
123456789	OP	201711	201711	2017-11-10	NA	201712	201711	0	0	N	T	0	N	0	tl_n43
123456789	CL	201712	201711	2017-11-10	201712	201712	201711	1	1	N	T	0	N	0	tl_n43
123456789	CL	201801	201711	2017-11-10	201712	201712	201711	2	2	N	T	0	N	0	tl_n43
123456789	CL	201802	201711	2017-11-10	201712	201712	201711	3	3	N	T	0	N	0	tl_n43
123456789	CL	201803	201711	2017-11-10	201712	201712	201711	4	4	N	T	0	N	0	tl_n43
123456789	CL	201804	201711	2017-11-10	201712	201712	201711	5	5	N	T	0	N	0	tl_n43
123456789	CL	201805	201711	2017-11-10	201712	201712	201711	6	6	N	T	0	N	0	tl_n43
123456789	CL	201806	201711	2017-11-10	201712	201712	201711	7	7+	N	T	0	N	0	tl_n43
123456789	CL	201807	201711	2017-11-10	201712	201712	201711	8	7+	N	T	0	N	0	tl_n43
123456789	CL	201808	201711	2017-11-10	201712	201712	201711	9	7+	N	T	0	N	0	tl_n43
123456789	CL	201809	201711	2017-11-10	201712	201712	201711	10	7+	N	T	0	N	0	tl_n43
123456789	CL	201810	201711	2017-11-10	201712	201712	201711	11	7+	N	T	0	N	0	tl_n43
123456789	CL	201811	201711	2017-11-10	201712	201712	201711	12	7+	N	T	0	N	0	tl_n43
123456789	CL	201812	201711	2017-11-10	201712	201712	201711	13	7+	N	T	0	N	0	tl_n43
123456789	CL	201901	201711	2017-11-10	201712	201712	201711	14	7+	N	T	0	N	0	tl_n43
123456789	CL	201902	201711	2017-11-10	201712	201712	201711	15	7+	N	T	0	N	0	tl_n43
123456789	CL	201903	201711	2017-11-10	201712	201712	201711	16	7+	N	T	0	N	0	tl_n43
123456789	CL	201904	201711	2017-11-10	201712	201712	201711	17	7+	N	T	0	N	0	tl_n43
123456789	CL	201905	201711	2017-11-10	201712	201712	201711	18	7+	N	T	0	N	0	tl_n43
123456789	CL	201906	201711	2017-11-10	201712	201712	201711	19	7+	N	T	0	N	0	tl_n43
123456789	CL	201907	201711	2017-11-10	201712	201712	201711	20	7+	N	T	0	N	0	tl_n43
123456789	CL	201908	201711	2017-11-10	201712	201712	201711	21	7+	N	T	0	N	0	tl_n43
123456789	CL	201909	201711	2017-11-10	201712	201712	201711	22	7+	N	T	0	N	0	tl_n43
123456789	CL	201910	201711	2017-11-10	201712	201712	201711	23	7+	N	T	0	N	0	tl_n43
123456789	CL	201911	201711	2017-11-10	201712	201712	201711	24	7+	N	T	0	N	0	tl_n43
123456789	CL	201912	201711	2017-11-10	201712	201712	201711	25	7+	N	T	0	N	0	tl_n43
123456789	CL	202001	201711	2017-11-10	201712	201712	201711	26	7+	N	T	0	N	0	tl_n43
123456789	CL	202002	201711	2017-11-10	201712	201712	201711	27	7+	N	T	0	N	0	tl_n43
123456789	CL	202003	201711	2017-11-10	201712	201712	201711	28	7+	N	T	0	N	0	tl_n43

Figure 2: Sample from database

In figure 3, we have for each month of observation the amounts paid for the loss and alea. `AUTO_LTD_NET_LOSS_PAID_AMT` is the paid amount known at the observation month. Any type of recovery can decrease the paid amount. `AUTO_LTD_LOSS_RES_CHG_AMT` and `AUTO_LTD_LOSS_RES_CHG_AMT` are the case reserve amounts at a given observation month. `AUTO_LTD_LOSS_INCURRED_AMT` and `AUTO_LTD_ALAE_INCURRED_AMT`, represent the incurred at the given observation month. The variables `FINAL_NET_PAID_AMT` and `FINAL_ALAE_AMT` are the final amounts we know today, they are also called the ultimate amount for that claim. `AvgTypicalValue` (`ACV`) is an estimate of the market value of the accident vehicle. `TotalGAV` (`GAV`) is the gross appraisal value, which is the garage cost estimate to repair the vehicle. `IBC_PRICE` is the initial purchasing price of the vehicle.



CLM_NBR	sf_status	obs_month	AUTO_LTD_NET_LOSS_PAID_AMT	AUTO_LTD_LOSS_INCURRED_AMT	AUTO_LTD_ALAE_INCURRED_AMT	AUTO_LTD_LOSS_RES_CHG_AMT	AUTO_LTD_ALAE_RES_CHG_AMT	FINAL_NET_PAID_AMT	FINAL_ALAE_AMT	AvgTypicalCarValue	TotalGAV	IBC_PRICE
123456789	OP	201711	11414.35	17964.35	0	6550	0	11213.87	0	8007	11481.67	33729
123456789	CL	201712	11213.87	11213.87	0	0	0	11213.87	0	8007	11481.67	33729
123456789	CL	201801	11213.87	11213.87	0	0	0	11213.87	0	8007	11481.67	33729
123456789	CL	201802	11213.87	11213.87	0	0	0	11213.87	0	8007	11481.67	33729
123456789	CL	201803	11213.87	11213.87	0	0	0	11213.87	0	8007	11481.67	33729
123456789	CL	201804	11213.87	11213.87	0	0	0	11213.87	0	8007	11481.67	33729
123456789	CL	201805	11213.87	11213.87	0	0	0	11213.87	0	8007	11481.67	33729
123456789	CL	201806	11213.87	11213.87	0	0	0	11213.87	0	8007	11481.67	33729
123456789	CL	201807	11213.87	11213.87	0	0	0	11213.87	0	8007	11481.67	33729
123456789	CL	201808	11213.87	11213.87	0	0	0	11213.87	0	8007	11481.67	33729
123456789	CL	201809	11213.87	11213.87	0	0	0	11213.87	0	8007	11481.67	33729
123456789	CL	201810	11213.87	11213.87	0	0	0	11213.87	0	8007	11481.67	33729
123456789	CL	201811	11213.87	11213.87	0	0	0	11213.87	0	8007	11481.67	33729
123456789	CL	201812	11213.87	11213.87	0	0	0	11213.87	0	8007	11481.67	33729
123456789	CL	201901	11213.87	11213.87	0	0	0	11213.87	0	8007	11481.67	33729
123456789	CL	201902	11213.87	11213.87	0	0	0	11213.87	0	8007	11481.67	33729
123456789	CL	201903	11213.87	11213.87	0	0	0	11213.87	0	8007	11481.67	33729
123456789	CL	201904	11213.87	11213.87	0	0	0	11213.87	0	8007	11481.67	33729
123456789	CL	201905	11213.87	11213.87	0	0	0	11213.87	0	8007	11481.67	33729
123456789	CL	201906	11213.87	11213.87	0	0	0	11213.87	0	8007	11481.67	33729
123456789	CL	201907	11213.87	11213.87	0	0	0	11213.87	0	8007	11481.67	33729
123456789	CL	201908	11213.87	11213.87	0	0	0	11213.87	0	8007	11481.67	33729
123456789	CL	201909	11213.87	11213.87	0	0	0	11213.87	0	8007	11481.67	33729
123456789	CL	201910	11213.87	11213.87	0	0	0	11213.87	0	8007	11481.67	33729
123456789	CL	201911	11213.87	11213.87	0	0	0	11213.87	0	8007	11481.67	33729
123456789	CL	201912	11213.87	11213.87	0	0	0	11213.87	0	8007	11481.67	33729
123456789	CL	202001	11213.87	11213.87	0	0	0	11213.87	0	8007	11481.67	33729
123456789	CL	202002	11213.87	11213.87	0	0	0	11213.87	0	8007	11481.67	33729
123456789	CL	202003	11213.87	11213.87	0	0	0	11213.87	0	8007	11481.67	33729

Figure 3: Severity sample from database

Already with this single extract we can get a small understanding of our data. Specifically, we notice that the final paid amount, i.e. the ultimate, is close to ACV and GAV amount. This might indicate that the ACV and GAV are good predictors for our model. Consequently, we want to identify the dependence structure between the ultimate amount and the ACV or GAV.

3.2 Dependence structure

[Embrechts et al., 2001] show how copulas are used for modelling dependence between random variables. Even thou we do not plan to model de dependence structure itself, we will use the copula to visualize the dependence. Since we are interested in more than only linear dependence, we will use Kendall's tau as dependence measure. The definition of Kendall's tau for a random vector pair (X, Y) is given as

$$\tau(X,Y) = \Pr((X-\widetilde{X})(Y-\widetilde{Y}) > 0) - \Pr((X-\widetilde{X})(Y-\widetilde{Y}) < 0)$$

, where $(\widetilde{X}, \widetilde{Y})$ is an independent copy of (X, Y).

It is the probability of concordance minus the probability of discordance. Concordance measures how X and Y move in the same direction relative to their independent copy. Discordance measures how X and Y move in opposite direction relative to their independent copy. It also can be interpreted as the correlation coefficient between the quantiles of X and Y, which have a relationship defined by a copula. Kendall's tau has a value between -1 and 1. -1 indicates perfect negative dependence, also called countermonotonic, while 1 indicates perfect positive dependence, comonotonic. If Kendall's tau is close to 0, the pairs are likely independent.

A copula is a cumulative distribution function of a multivariate uniformed distribution. The copula of two independent uniform distribution $U_1 \sim U(0,1)$ and $U_2 \sim U(0,1)$, is defined as

$$C(u_1, u_2) = u_1 \times u_2$$

.

A copula can be visualized by plotting pairs of quantiles of the uniform distributions. For the bivariate independent copula, the pairs are even distributed on the graph. Kendall's tau should be close to 0 since it measures the correlation coefficient of these pairs.

Before we can plot the ultimate and the ACV/GAV, we need to find their empirical quantile values between 0 and 1. We rank the values according to their relative size and divide each rank by the total number of observations. In addition, we will only use a sample of 10000 pairs, because we would have to many data points on the graph. We also use one single pair per claim number. Furthermore, we group the data into age since reported date categories.

Starting with Quebec, figures 4 to 5 shows the copulas for each age grouping. On the x-axis we plot the quantiles of the ultimate amount (paid loss + alae) and on the y-axis we can see the quantiles of theGAVor ACV. We use theGAVfor claims with repairable vehicles and the ACV for claims with vehicles that are total loss. On figure XXX we can observe the relationship between the ultimate and GAV. The relationship seems strong especially for younger claims (¿ 0.7). Note that Kendall's tau decreases the older the claims become, indicating that older claims become more complex and incurred additional payments or recoveries which do not depend on the damage estimation. Figure XXX demonstrates a weaker dependence between the ACV and the ultimate. However, the dependence is still positive and not negligible. Further, it shows that the ACV might not always be a good estimation of the actual market value. The weaker dependence is therefore only caused by additional fees but also by intrinsic estimation error of the actual market value of the vehicle.

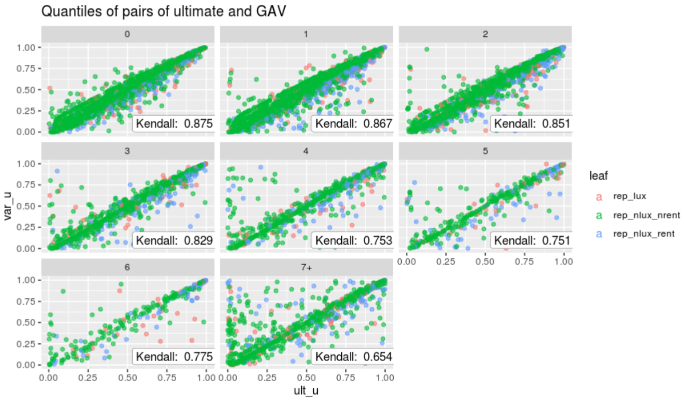


Figure 4: Quantiles pairs for Quebec repairable claims. x-axis is the ultimate quantiles and y-axis the GAV quantiles

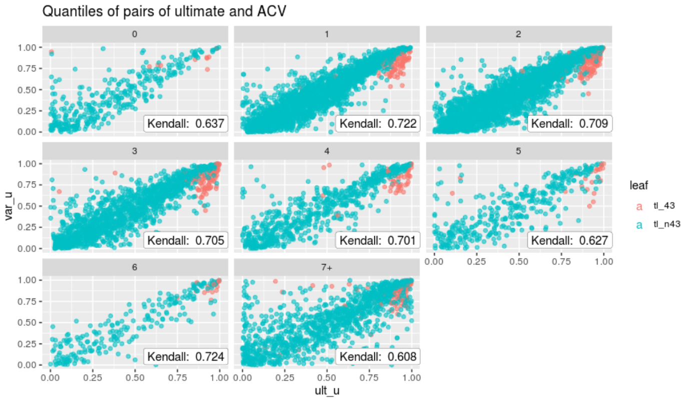


Figure 5: Quantiles pairs for Quebec total loss claims. x-axis is the ultimate quantiles and y-axis the ACV quantiles

For Ontario figure6 to 7 should similar pattern than for Quebec. However, since we have two different line of businesses PHYSDAM and LIPD, it is interesting to observe the difference in pattern. LIPD tends to be more on the lower half of the diagonal. Third party liabilities seem to incur higher losses than the GAV would suggest and that it incurs more additional fees.

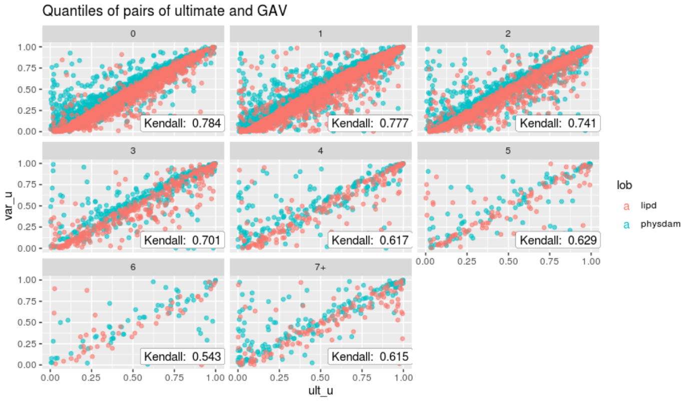


Figure 6: Quantiles pairs for Ontario repairable claims. x-axis is the ultimate quantiles and y-axis the GAV quantiles

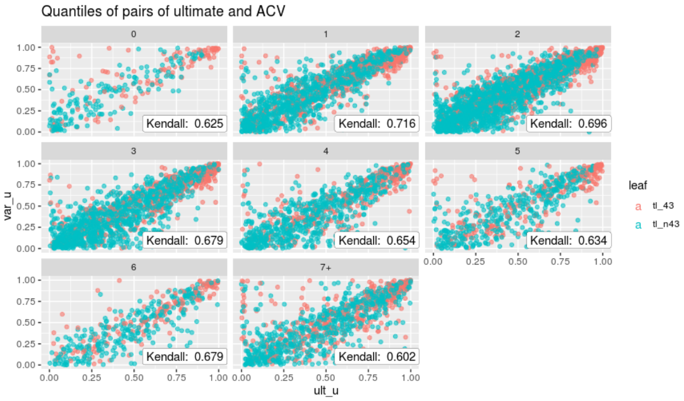


Figure 7: Quantiles pairs for Ontario total loss claims. x-axis is the ultimate quantiles and y-axis the `ACV` quantiles

Alberta shown in figure 8 to 9 has an interesting patterns. Again, we see weaker dependence for older claims. Albeit, there is a descriptive force which seem to strongly impact the dependence structure and leads to more claims with very low ultimate compared to the `GAV` or `ACV`. We identified this disruptive force as subrogation and recoveries. Subrogation is a slow process at which the insurance company can recover paid losses, if the insured was not responsible for the accident. This mean that if Intact paid the entire loss, they might be able to recover a part or the entire loss with a lawsuit.

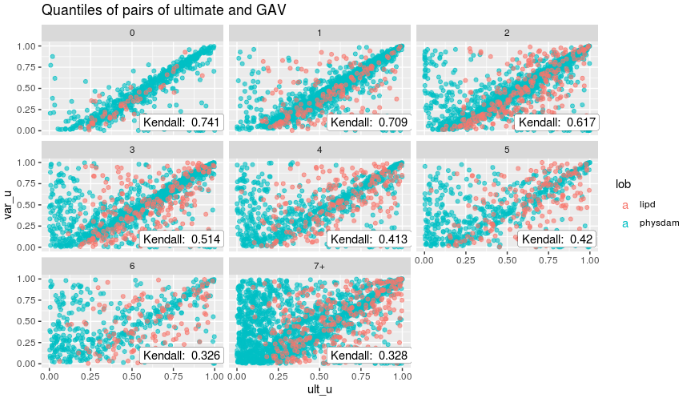


Figure 8: Quantiles pairs for Alberta repairable claims. x-axis is the ultimate quantiles and y-axis the `GAV` quantiles

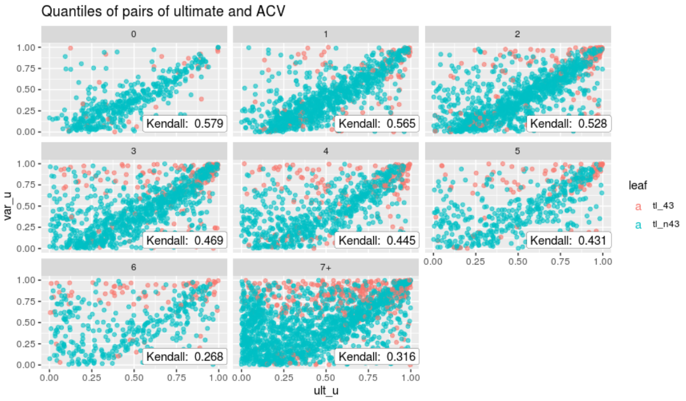


Figure 9: Quantiles pairs for Alberta total loss claims. x-axis is the ultimate quantiles and y-axis the ACV quantiles

The data indicates that it can take more than a year until the subrogation process is finished. Consequently, a proportion of claims in Alberta need much longer to fully develop to the ultimate which might even fall to 0 or negative. This can be problematic, since we are often unable to predict if the claim falls into the subrogation category or not. We can not filter out these claims, because for a given observation month, we do not know which claims a affected. We can partly mitigate this issue by aggregating the data and using averages. We can expect higher volatility in our model for Alberta. All of these copulas also indicate a slightly stronger dependence for large/extreme values.

3.3 Trend analysis

In our model we do not want to estimate the ultimate on a claim by claim basis, so we will aggregate the data. When aggregation data is important to verify the trends in data. If we use the GAV and the ACV as a predictor for the ultimate, we should verify that mean growth rates are similar. For each observation month we will calculate the mean ultimate and the mean GAV/ACV of open claims. Then, we will compare their monthly growth rates. We could also do the same per month of loss, however, we want to analyse the underlying distribution of what we are trying to predict. Our model will predict the ultimate based on observation month data. Each observation month will contain a proportion of claims with different month of loss. While using aggregation per observation month, we have to be aware of possible fluctuations related to different number of claims and different mixtures of month of loss.
When looking at figure 10, we can observe that in Quebec the average ultimate grows faster than the average GAV/ ACV. Using theGAVand ACV as predictor might tend to underestimate the ultimate if we use past averages.

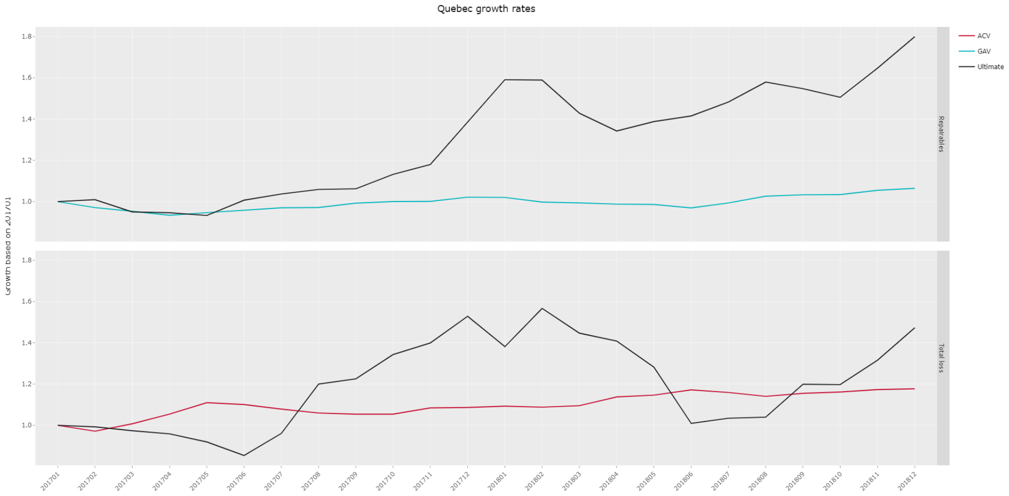


Figure 10: Quebec ultimate and `GAV/ACV` growth relative to January 2017 (= 1)

In Ontario figure 11, the opposite seems to happen for total loss claims but not for repairable. Thus, we might tend to overestimate total loss claims ultimate.

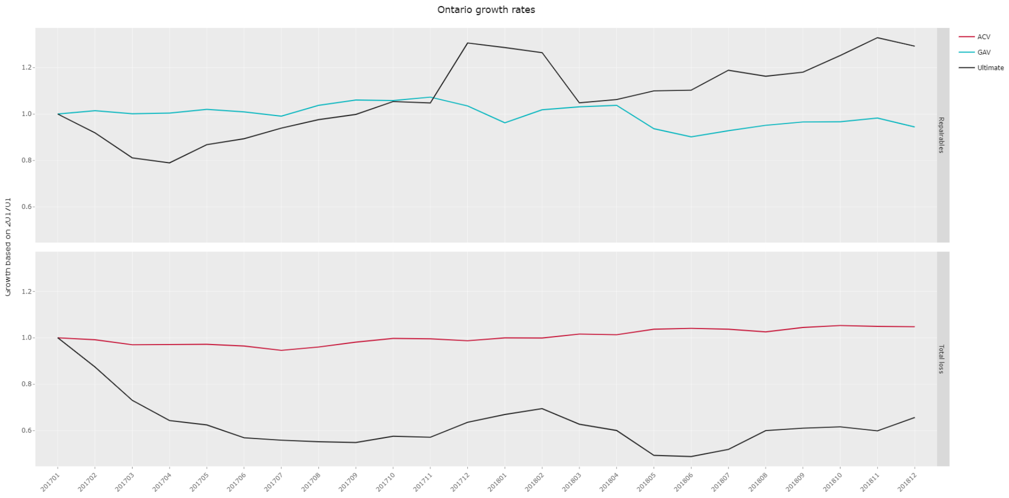


Figure 11: Ontario ultimate and `GAV/ACV` growth relative to January 2017 (= 1)

On figure 12, Alberta has a similar but reversed pattern. Total loss claims ultimate growth fluctuates around the 1 value. While for repairables, growth rates for the ultimate are lower than for the `GAV`.

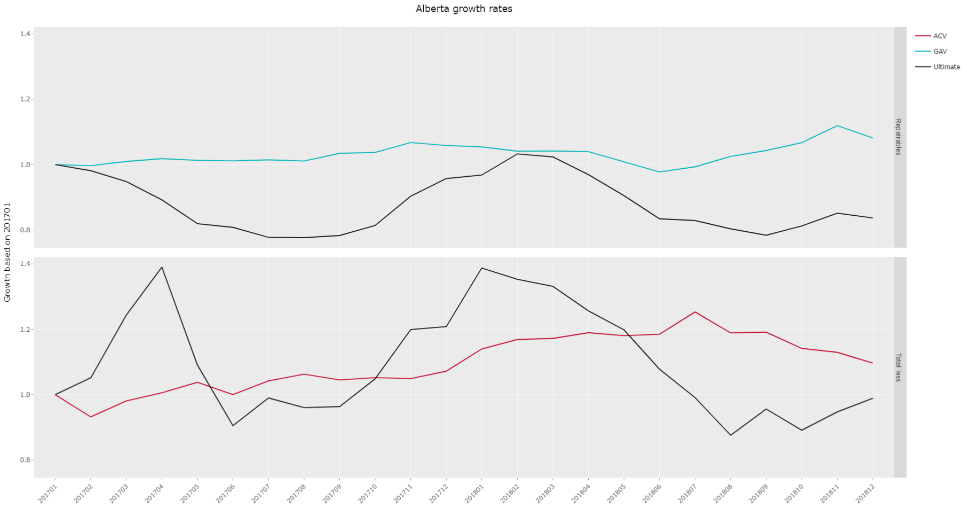


Figure 12: Alberta ultimate and GAV/ACV growth relative to January 2017 (= 1)

CONFIDENTIEL??? Furthermore, the GAV and ACV might not capture the additional costs related to the claim. The ultimate allocated loss adjustment expense (ALAE) is often unrelated to the GAV or ACV. A larger proportion of ALAE can cause greater estimation error. Figure 13 shows the ALAE to loss ratio for Quebec. The average is around 0.0125. In Ontario, seen in figure 14, the ALAE to loss ratio is similar to the Quebec, although after December 2017 there is clearly a spike which might cause prediction errors. Alberta in figure 15 show again an different pattern than the other 2 regions. While most ratios are lower than in Quebec and Ontario, le non-luxury non-rental repairable vehicles show proportionally larger ratios.

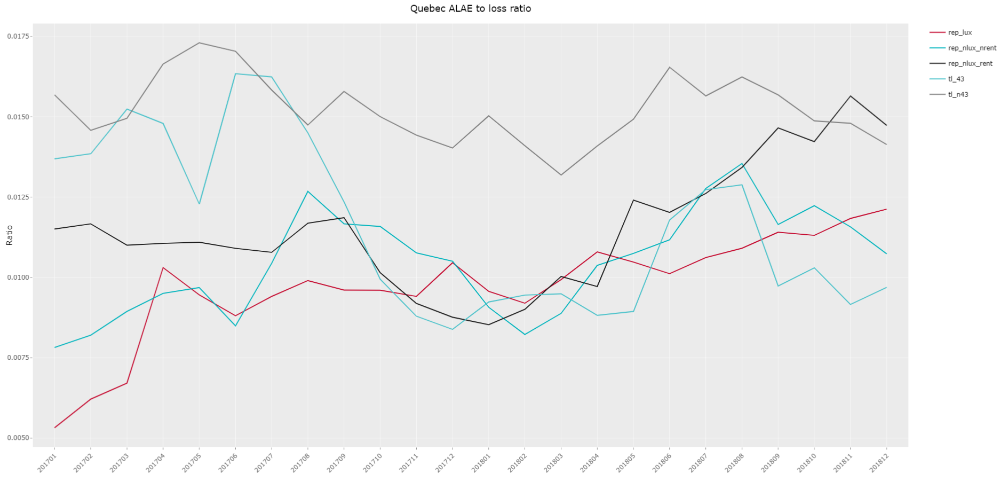


Figure 13: Quebec ALAE to loss ratio per `leaf`

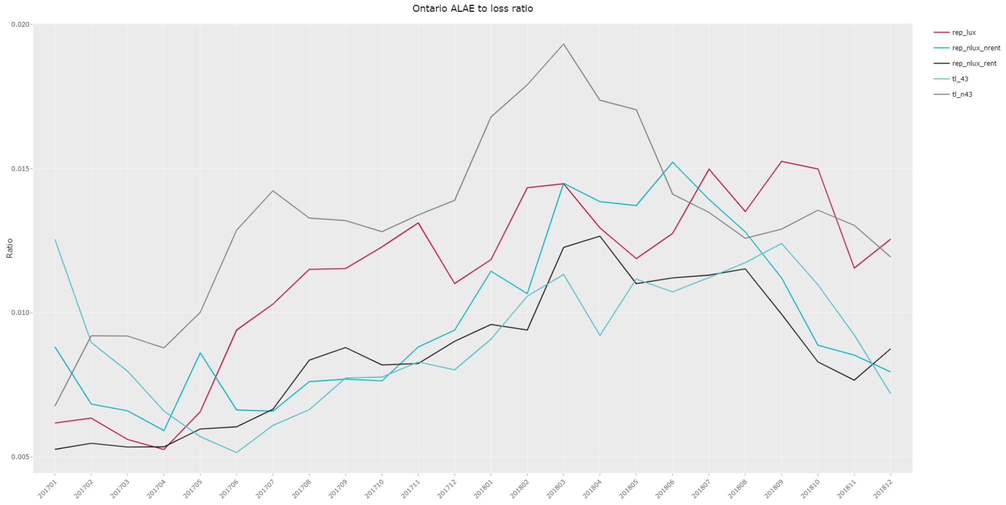


Figure 14: Ontario ALAE to loss ratio per `leaf`

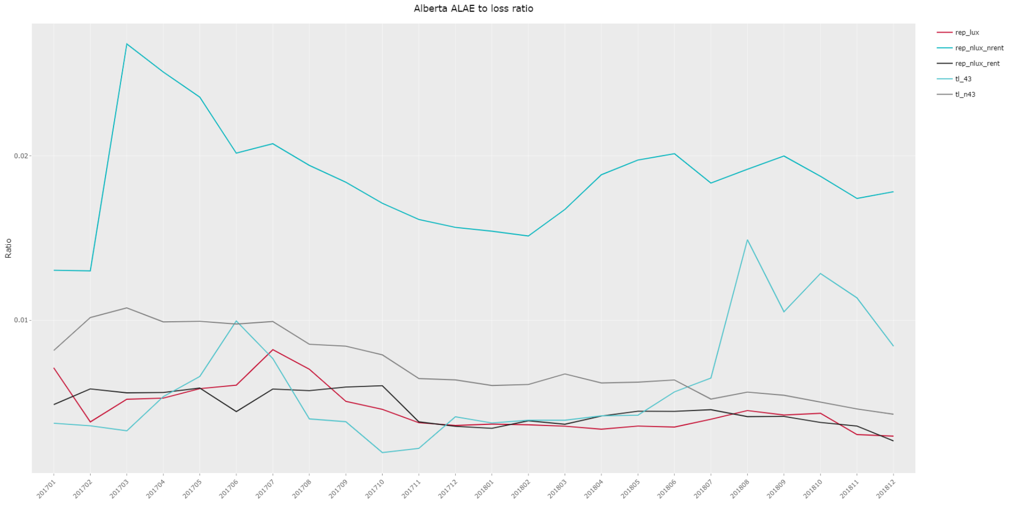


Figure 15: Alberta ALAE to loss ratio per `leaf`

In order to better understand the impact of recovery on our data, figures 16 to 18 shows the recovery to ultimate ratio for all 3 regions. Quebec and Ontario both have a ratio below 0.17, while Alberta has ratios between 1 and 0.3. The discrepancy is very significant and has to be considered in our model.

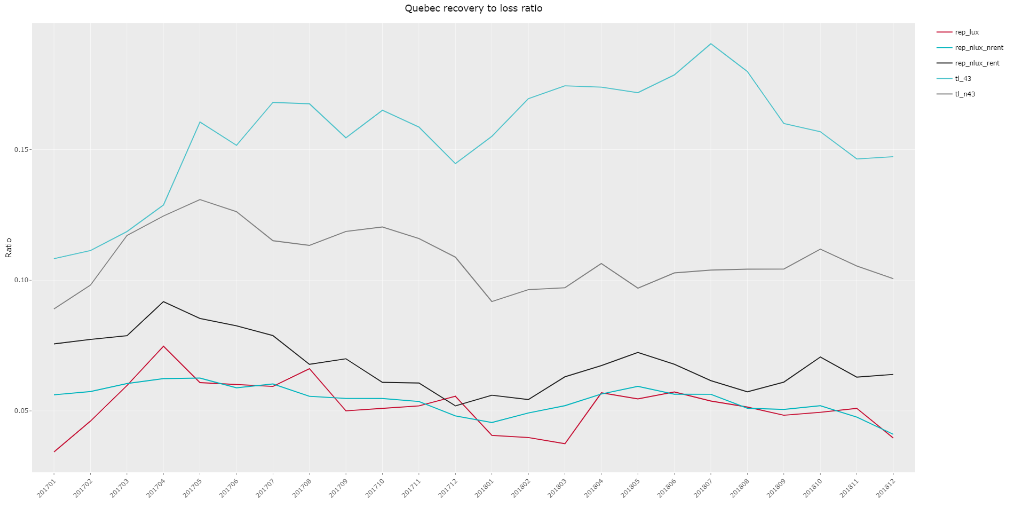


Figure 16: Quebec recovery to loss ratio per `leaf`

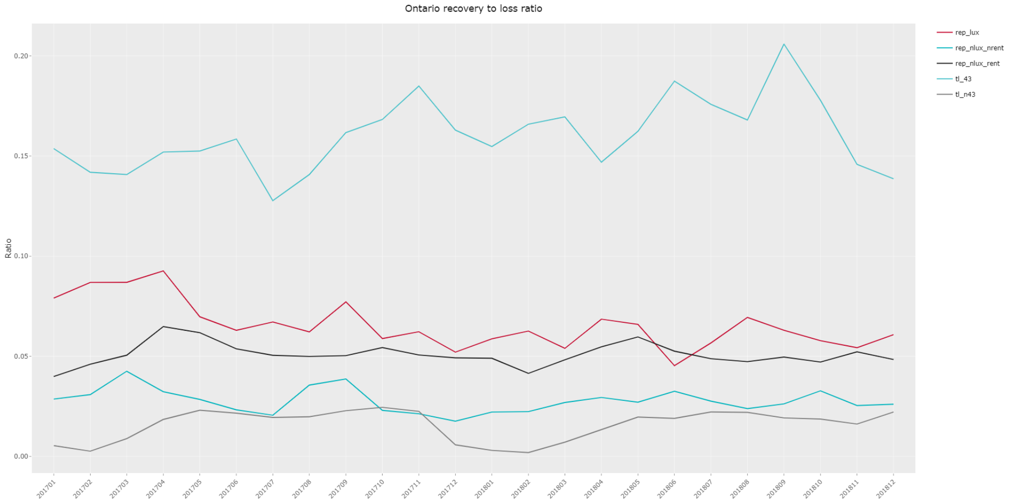


Figure 17: Ontario recovery to loss ratio per `leaf`

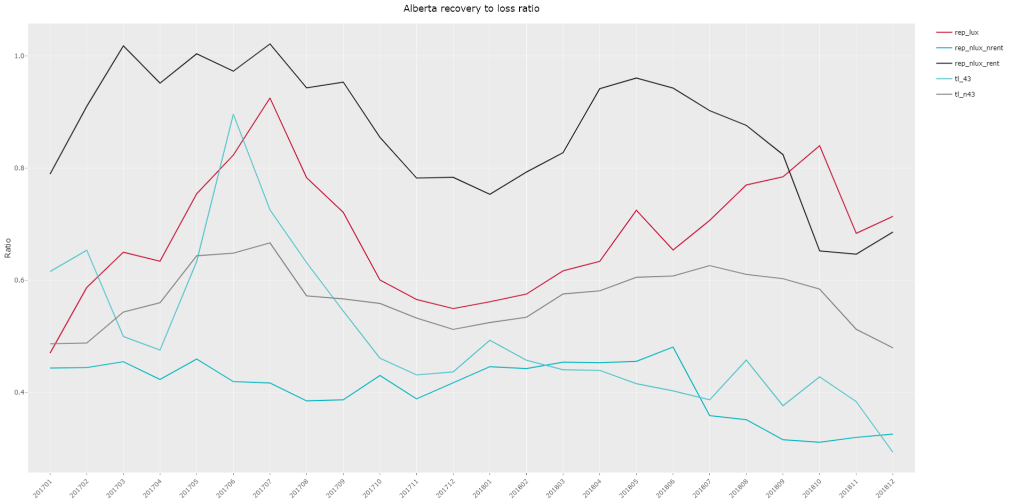


Figure 18: Alberta recovery to loss ratio per `leaf`

Now that we have a better understanding of our data, we will discuss the model structure and methodology.

4 Model methodologie

4.1 Incurred but not reported

The task of the actuarial model is to predict the IBNR, the incurred but not reported claims. The IBNR can be divided into 3 distinct elements, which we defined as pure IBNR, IBNER and unpure IBNR. Pure IBNR are claims which are not reported at the observation date, meaning the insurer has no information on them. The insurer only knows that a claim happened. IBNER, incurred but not enough reported, are claims which have been reported and the insurer the information on the claims in their database. Unpure IBNR consist of claims which might reopen at any given time. This mean that a claim which closed in 2017 might reopen in 2018 or 2019. Unpure IBNR is a small proportion of the total IBNR, but still should be considered in the model. Figure 19 gives a visual representation of these categories.

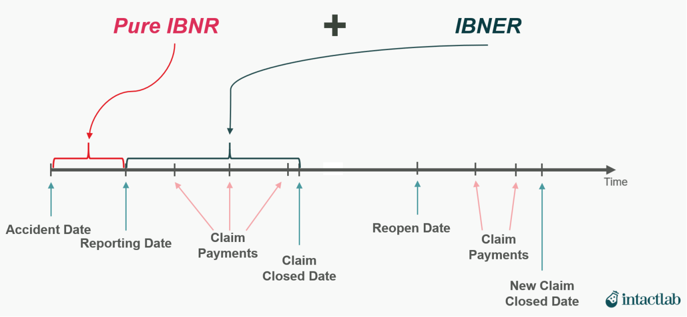


Figure 19: Timeline of a claim

4.2 Hierarchical approach

The actuarial department uses a modified chain-ladder method for their model. We will try a more hierarchical approach, where we cluster our data in more homogeneous groups. First, we develop a model for each of the three IBNR types. Our team focuses on the IBNER part, while the pure and unpure IBNR models are still chain-ladder based and were developed by the actuarial department. For the IBNER model, we grouped the data according the following claims characteristics: total loss, total loss without (43) replacement cost endorsement (`n43`), luxury repairable vehicles (`rep_lux`), non luxury non rental repairable vehicles (`rep_nlux_nrent`) and non luxury rental repairable vehicles (`rep_nlux_rent`). We suppose that the frequency and severity distributions are very similar within these groups. Figure 20 gives an overview of the hierarchical approach.

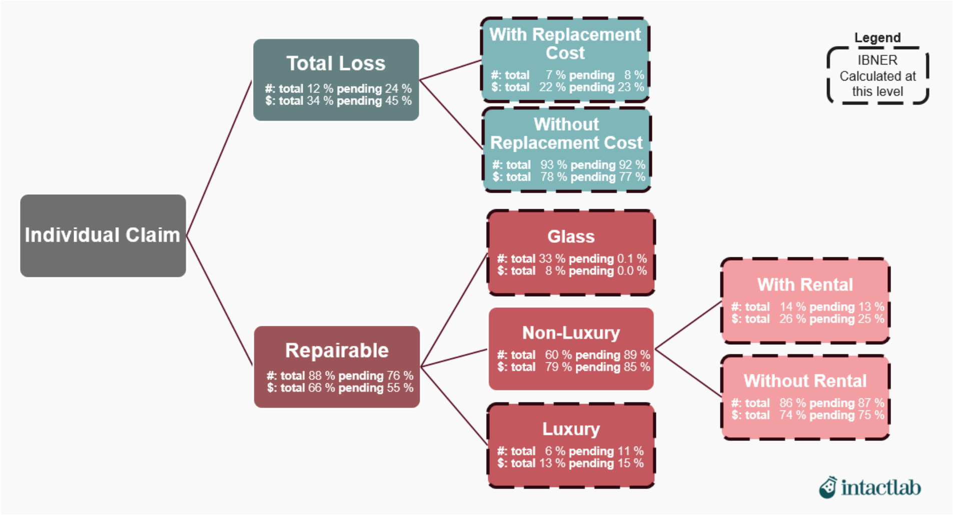

Individual Claim
Total Loss
#: total 12 % pending 24 %
$: total 34 % pending 45 %
With Replacement Cost
#: total 7 % pending 8 %
$: total 22 % pending 23 %
Without Replacement Cost
#: total 93 % pending 92 %
$: total 78 % pending 77 %
Repairable
#: total 88 % pending 76 %
$: total 66 % pending 55 %
Glass
#: total 33 % pending 0.1 %
$: total 8 % pending 0.0 %
Non-Luxury
#: total 60 % pending 89 %
$: total 79 % pending 85 %
Luxury
#: total 6 % pending 11 %
$: total 13 % pending 15 %
With Rental
#: total 14 % pending 13 %
$: total 26 % pending 25 %
Without Rental
#: total 86 % pending 87 %
$: total 74 % pending 75 %
Legend
IBNER Calculated at this level
intactlab

Figure 20: Hierarchical model structure

4.3 Key formulas

As mentioned in section ???, the `ACV` and `GAV` have strong predictive strength. We will use a basic formula to link the `GAV`/`ACV` with the ultimate and use historical observation on the pending (open) claims.

Definition 1. *We define $\hat{L}_i$ as ultimate loss prediction for claims pending/open in period i and X_i as the predictor, in our case `GAV` or `ACV` used to predict period i. Their relationship is defined as*

$$\hat{L}_i = \hat{\Theta}_i \times X_i$$

Where $\hat{\Theta}_i$ is the factor for time period i.

We need to calculate the factor $\hat{\Theta}_i$ with the available historical data.

Definition 2. *We define $\widetilde{L}_{j,i}$ as total incurred for claims in time period j as of i. X_j is the predictor in time period j. Thus, the factor is defined as*

$$\hat{\theta}_i = \frac{\widetilde{L}_{j,i}}{X_j}$$

Note the difference between $\hat{L}_i$ and $\widetilde{L}_{j,i}$. The former is the ultimate we want to predict, so we do not know its value in observation month i. The latter is the total incurred for claims in period j we know as of i. For illustration in the sample data of figures 2 and 3, we want to calculate the factor $\hat{\theta}_{201804}$. We suppose we want to use open claim with CLM_NBR = 123456789 to calculate this factor, then

$$\begin{aligned}\widetilde{L}_{201711,201804} &= \texttt{AUTO\_LTD\_NET\_LOSS\_PAID\_AMT} + \texttt{AUTO\_LTD\_ALAE\_INCURRED\_AMT} \\ &\quad + \texttt{AUTO\_LTD\_LOSS\_RES\_CHG\_AMT} \\ &= 11213.87\end{aligned}$$

as of 201804 (obs_month = 201804) and X_j = AvgTypicalCarValue = 8007. As the example illustrated, the ultimate and the predictor are historical values which should be fully developed. It is necessary to have a least 5 to 12 months of development, so that the factors are stable enough. The difference in time between the moment we want to predict the pending and the historical data, $i - j$ is defined as the lag. How many historical observation month j we use for the calculation is defined as period length. Figure 21 shows a 5 months lag and 3 months period length. We want to predict the ultimate of the December 2018, $i = 201812$ pending (open) claims. We go back 5 months and use the historical claim data from open claims in May, June and July 2018, $j = 201805, 201806, 201807$.

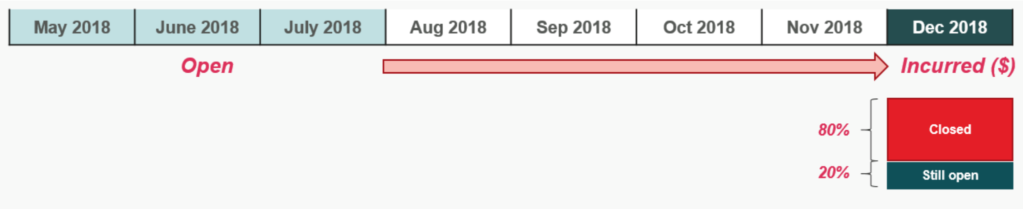
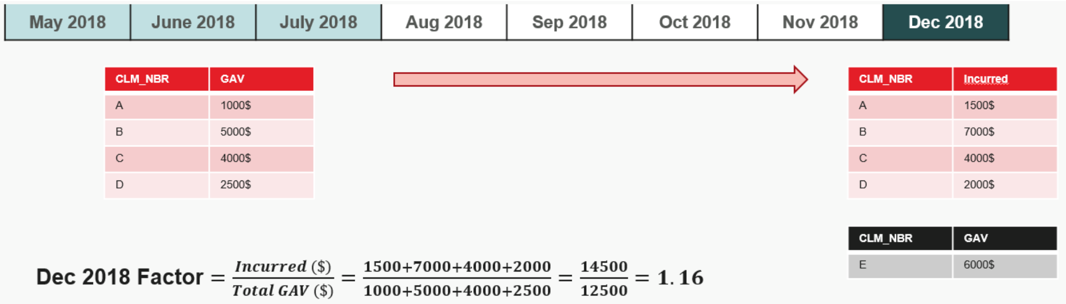


Figure 21: Lag and period length visualization

This approach is similar to a lagged moving average model. We use the incurred as of December 2018 of claims open in May, June and July. Thus, the incurred had a minimum of 5 months to develop. Then, we divide this incurred by the aggregated GAV or ACV in May, June and July. We only keep the most recent data line for each claim, in other words we don't have any duplicated per claim number. Figure 22 gives a numerical example.



CLM_NBR	GAV
A	1000$
B	5000$
C	4000$
D	2500$

CLM_NBR	Incurred
A	1500$
B	7000$
C	4000$
D	2000$

CLM_NBR	GAV
E	6000$

$$\text{Dec 2018 Factor} = \frac{\textit{Incurred}\ (\$)}{\textit{Total GAV}\ (\$)} = \frac{1500+7000+4000+2000}{1000+5000+4000+2500} = \frac{14500}{12500} = 1.16$$

Figure 22: Factor calculation example

Note that in figure 21 the incurred is subdivided into open and closed claims, since it is possible that claims remain open even after 5 months. Therefore, we will calculate a claim number weighted average of closed and open factors. We get as final factor

Definition 3. *$\hat{\Theta}_i$ is the final factor used for the prediction. $\hat{\theta}_{i,open}$ is the factor for claims that are still open during i and $\hat{\theta}_{i,closed}$ is the factor for claims that are closed during i. Furthermore, n_{open} is the number of open claims in period i and n_{closed} is the number of closed claims in period i. Thus, we have*

$$\hat{\Theta}_i = n_{open}\hat{\theta}_{i,open} + n_{closed}\hat{\theta}_{i,closed}$$

We multiply this factor by the aggregated `GAV` or `ACV` of pending claims (December 2018 in our example) to get the ultimate amount.

Definition 4. *We define the $IBNER$ in period i as*

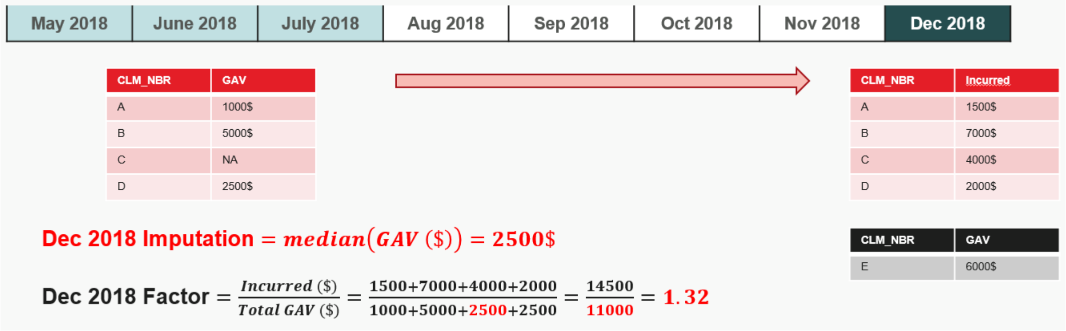
$$IBNER_i = \hat{L}_i - I_i$$

, where I_i is the incurred payments and reserve for claims pending in period i.

Note that once claims are fully developed $L_i = I_i$, where L_i is the real observed ultimate loss for claims open in period i. The prediction results for the model will be discussed in section XX.

4.4 Imputation methodology

It is important to note that no data is perfect. Our dataset contains a non negligible amount of missing `GAV` and `ACV` values. About 17% of the claims have missing `GAV` or `ACV`, some regions and line of business are worse than others. Figure 23 demonstrates how missing values are imputed. If we have missing values, we calculate the median of the existing values in the time window. We replace the missing values by the median and execute the factor calculations without any further readjustment. This method is not perfect and will be revised. For instance, in our example, we would overestimate the factor since we impute with a lower value.



CLM_NBR	GAV
A	1000$
B	5000$
C	NA
D	2500$

CLM_NBR	Incurred
A	1500$
B	7000$
C	4000$
D	2000$

CLM_NBR	GAV
E	6000$

Dec 2018 Imputation $= median(GAV\ (\$)) = 2500\$$

Dec 2018 Factor $= \frac{Incurred\ (\$)}{Total\ GAV\ (\$)} = \frac{1500+7000+4000+2000}{1000+5000+2500+2500} = \frac{14500}{11000} = 1.32$

Figure 23: Factor calculation example with missing values

4.5 Adjusting the lag and period length

As we have noticed in the analysis of the data, Alberta clearly has different patterns than the other regions. Claims in Quebec and Ontario settle considerably faster (after 4-5 months) and recovery is less impactful. In Alberta having a lag of 5 is clearly insufficient, thus we will use a lag of 10. This gives the claims at least 10months to fully develop. In addition, a lag and period length that includes the 12th month, is beneficial if we have seasonal effects.

4.6 Second model iteration: Historical pending but now closed claims

After having used the first iteration of the model for January 2020. We noticed systematic error, constant over- or underestimations. This might be related to the open claims (after the lag) with are used to calculate the factors. Since open claims still have a case reserve the incurred used for the factor might be inflated, thus explaining overestimation. Indeed, the factor for open claims is always larger than the factor for closed. If we overestimate constantly, we attribute to much weight to still open claims. In the second iteration of the model we discard the weighted average and we only use the factor of closed claims. In short, based on definition 3, we have

$$\hat{\Theta}_i = \hat{\theta}_{i,closed}$$

. This also has the benefit that closed claims usually should not develop further and thus have less volatility in the calculation of the factors. The results will be discussed in section 5.

4.7 Third model iteration: Historical closed claims

The final iteration of the model we will discuss in this report is based on the idea of the second iteration. We only want to use the factor of closed claims. However, in the second iteration we use the factor of closed claims of the lagged 3 month period, while in this iteration we want to increase the number of claims used in the calculations and add more recent claims. Consequently, instead of using the pending claims in a historical time window, we calculate factors for all claims that closed in the time window. This simply implies changing the filter on `sf_status`. If we have use the previous example, we want to predict the pending claims in December 2018, we use all claims that closed between May and July for our factor calculations. Since the claims are closed it is unlikely that we have still open claims in December.

5 Results

In this section we will analyse the results of all 3 model iterations. For the first two iterations we will show the prediction error on the ultimate for each `leaf` and province, while for the last iteration will will give the average monthly error. It should be notes that we only show the ultimate predictions for claims pending in 2017 and 2018, although we use the incurred as of March 2020. Pending claims in December 2018 had more than 14 months to develop to the ultimate.

5.1 First model iteration: Historical pending claims

Quebec: For our first model, the average monthly prediction error is 957757.84, based on a total volume of about 100 million. Figure 24 show that the error has a light seasonal pattern. In winter we tend to slightly underestimate, while in most months we overestimate the ultimate loss. The seasonality is difficult to confirm, since we only have 2 years of fully credible data. The error does not seem to be all random and thus should be further investigated.

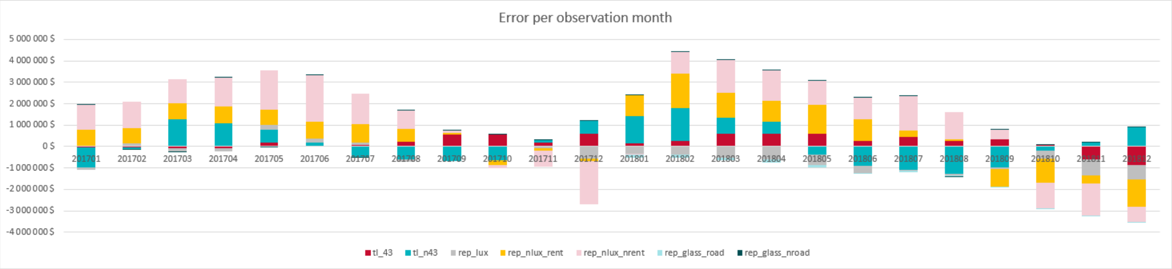


Figure 24: Quebec "Historical pending claims" model, prediction error by observation month

Figure 25 indicates that the largest proportion of error comes form `tl_n43` and `rep_nlux_nrent`.These two `leaf` are also the largest in terms of volume.

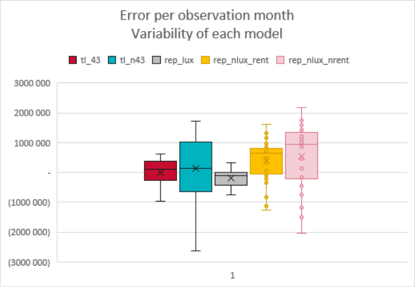


Figure 25: Quebec "Historical pending claims" model, monthly prediction error boxplot

Ontario: In Ontario, our model has the tendency to underestimate the ultimate amount. The average error per observation month shown in figure 26 is at -4632281.66. Specifically the winter proves difficult for our model. We are currently investigating this issue, but it might be related to the lag we use and/or the trends observed in the data.

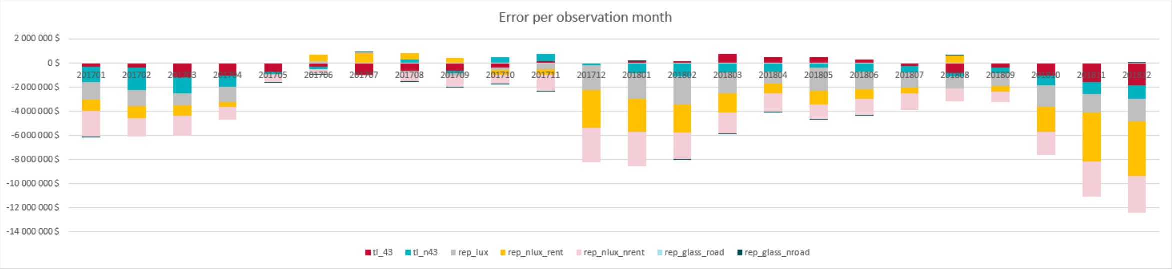


Figure 26: Ontario ”Historical pending claims” model, prediction error by observation month

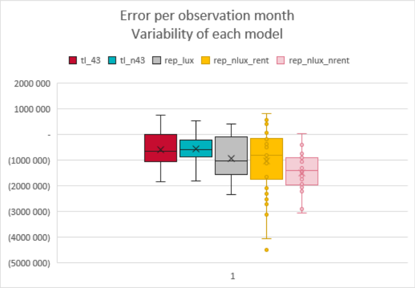


Figure 27: Ontario ”Historical pending claims” model, monthly prediction error boxplot

Alberta: The average monthly prediction error amounts to 3547625.79 on a total volume of about 80 million. There is no clear seasonal pattern, however the model constantly overestimates. The cause of the overestimation is the subrogation, which even after 12 months is not completed. We found about 10% of the claims in Alberta still receive recovery payments after 12 months. Figure 29 shows that `rep_nlux_rent` is the biggest source of error for this model.

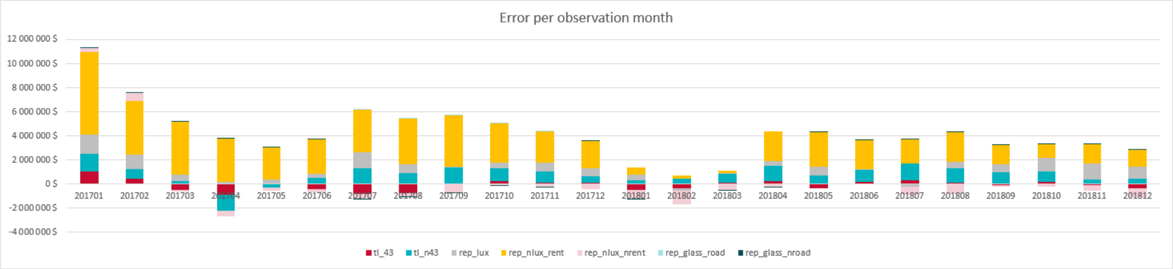


Figure 28: Alberta ”Historical pending claims” model, prediction error by observation month

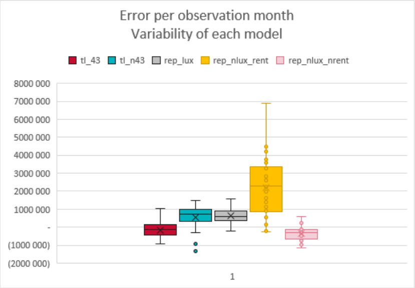


Figure 29: Alberta "Historical pending claims" model, monthly prediction error boxplot

5.2 Second model iteration: Historical pending but now closed claims

Excluding open claims in the factor calculations has a strong impact on the model. The average monthly error are -1379997.20, -5134385.50 and -431987.83, for Quebec, Ontario and Alberta respectively. This model has a strong tendency to underestimate. Albeit, for Alberta, when looking at figure 34, the results are considerably better than the first iteration model. Adding open claims seems to multiply the error by a factor of 10. This indicated that once a claim is closed, the subrogation process seems to also be finished or partly finished. For the other two provinces, it that adding open claims captures something we are not yet able to fully explain. We suspect that the issue is related to the imputation method and observed trends.

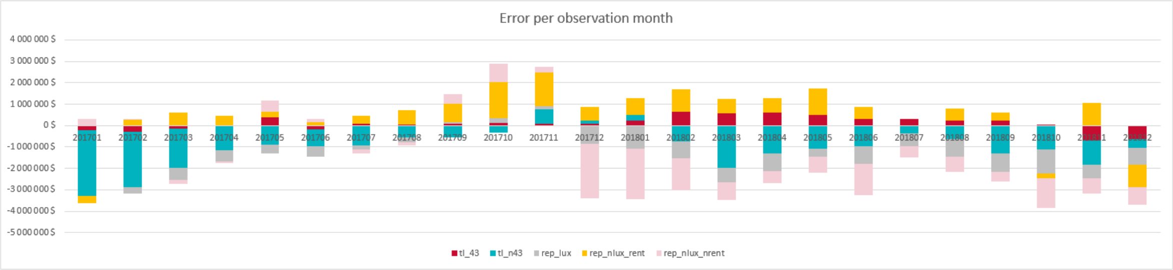


Figure 30: Quebec "Historical pending but now closed claims" model, prediction error by observation month

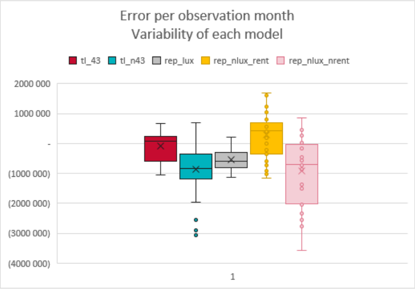


Figure 31: Quebec ”Historical pending but now closed claims” model, monthly prediction error boxplot

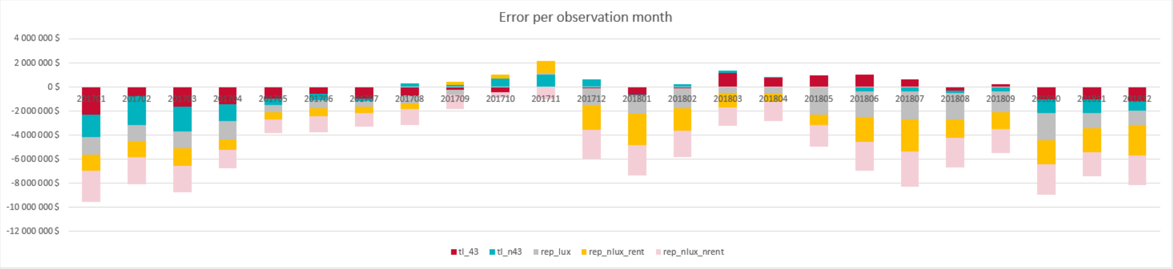


Figure 32: Ontario ”Historical pending but now closed claims” model, prediction error by observation month

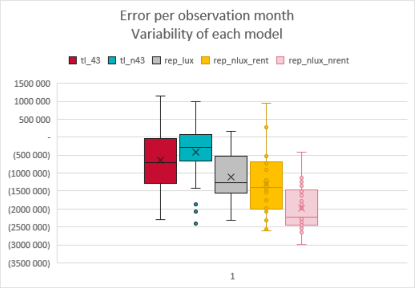


Figure 33: Ontario ”Historical pending but now closed claims” model, monthly prediction error boxplot

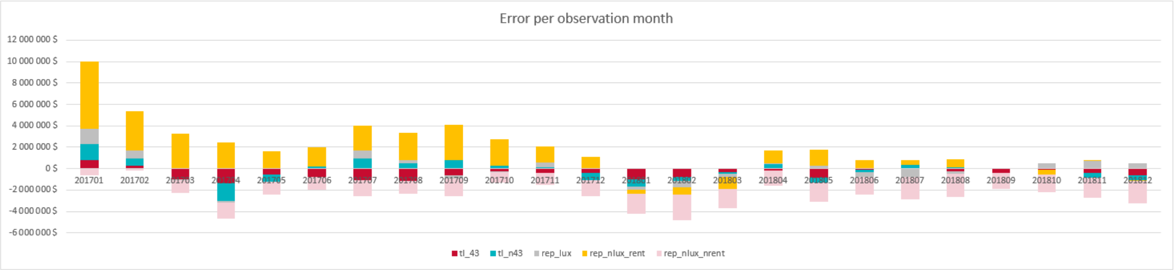


Figure 34: Alberta "Historical pending but now closed claims" model, prediction error by observation month

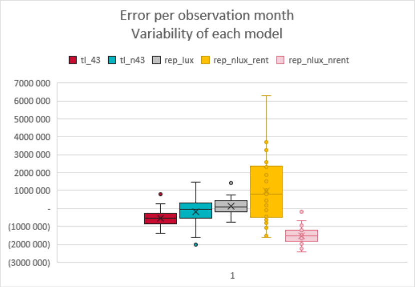


Figure 35: Alberta "Historical pending but now closed claims" model, monthly prediction error boxplot

5.3 Third model iteration: Historical closed claims

Lastly, the third iteration model seems not to function as well as expected. Quebec has an average monthly error of -7196717.05, Ontario of -13153105.83 and Alberta of 4148104.04. One issue with this model is related to the claims in the window. While the two previous model use a small window of pending claims, this model uses all claims that closed in the time window. Thus, we are using claims that can be very old relative to the observation month. Therefore the age of a claim might have a considerable impact on this model.

6 Conclusion

We were able to develop a functional and practical model for December 2019 and thus delivered 2019 year end predictions to compare with the corporative actuarial department booked numbers. The results were positive. However, we were not fully satisfied with the model and wanted to increase accuracy and consistency. ... areas of improvement ... Advanced AI model??

References

[Embrechts et al., 2001] Embrechts, P., Lindskog, F., and McNeil, A. (2001). Modelling dependence with copulas. *Rapport technique, Département de mathématiques, Institut Fédéral de Technologie de Zurich, Zurich*, 14.

[Intact, 2018] Intact (2018). 2018 annual report.