Identifying outlier insurance firms

Giorgio Miraglia, 03/03/2022

Introduction and report overview

This report identifies outlier insurance firms based on three main criteria: firm size, being an outlier 'from the norm', and whether the business profile of the firm is changing. The criteria are used separately from one another, each forming a section of this report. In each section the report identifies which firms are most likely to be outlier firms and therefore require increased supervisory resources from the PRA. For the most part, identification of the outlier firms is based on univariate analysis, however the main results from the 'outlier from the norm' section, which most comprehensively seeks to identify outliers, is confirmed via ML methods: isolation forests and DBSCAN clustering. There are what seem to be significant data quality issues, which are addressed in the first section.

In general, the findings of the report are:

- Most outliers stay the same from one reporting period to the next, and this finding is supported by ML models described in the Annex.
- There are a small number of very large firms.
- Most 'features' (variables) in the insurance market are not evenly or normally distributed. There are many firms far from the centre of distributions for most variables.
- Further analysis should be done on the outlier firms this report identifies to ascertain the nature and validity of their outlierness.

The main caveat of the analysis is that the results are heavily dependent on how erroneous outliers are identified. These are outlier observations that I believe to be instances of misreporting. Further work on this report would first and foremost require a more robust approach to spotting these erroneous outliers.

Removing erroneous outliers

Some observations for some firms have values that are either larger than the rest of the population by several orders of magnitude or are intuitively impossible values for the data field to take. As a first step in this analysis, I try to label these observations. Using boxplots (in the annex), such values are quickly identified. For example, it must be wrong that a firm has an SRC coverage ratio of 1 billion. These types of outliers, which massively skew the averages and standard deviations of the variables are removed. This is important because my method for detecting outliers depends on the relationship between the mean and standard deviation.

The second way in which I remove anomalous observations is by excluding observations that intuitively should not exist. For example, net combined ratio should not take values less than zero, so those

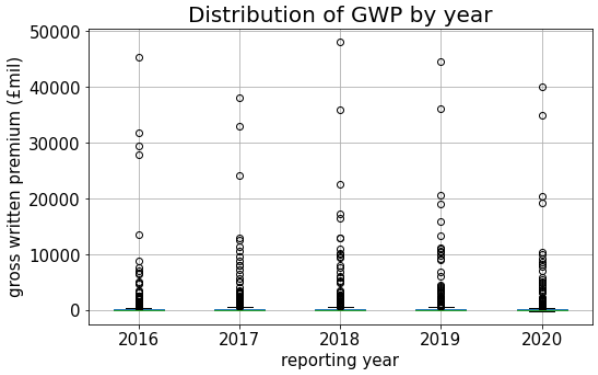
observations are excluded. I use the general rule that the 'ratio' variables with earned premiums in the denominator cannot take values greater than ten and must take values greater than zero.

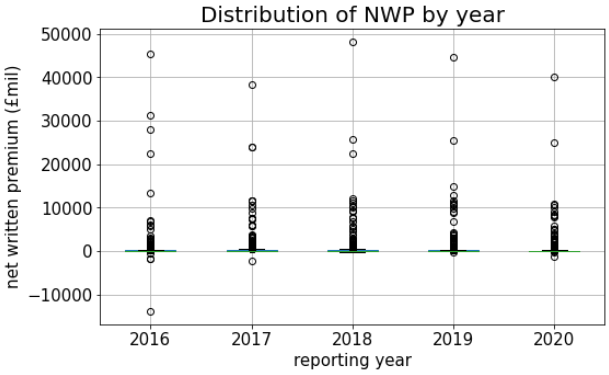
In all, I label 8.5% of observations as being misreported, and exclude these from the rest of the analysis. There is scope to remove misreported figures more surgically by using domain knowledge, understanding the reporting requirements on firms, and looking for further patterns in the suspicious observations.

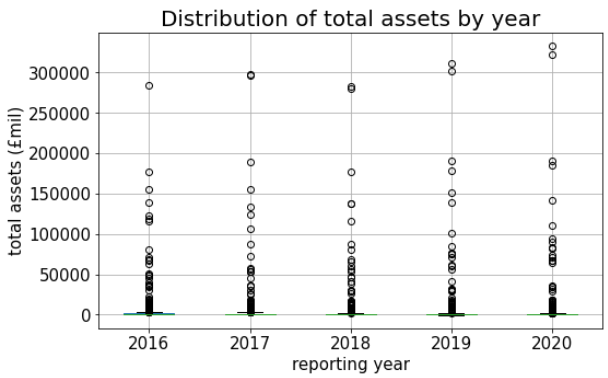
Identifying outliers by size

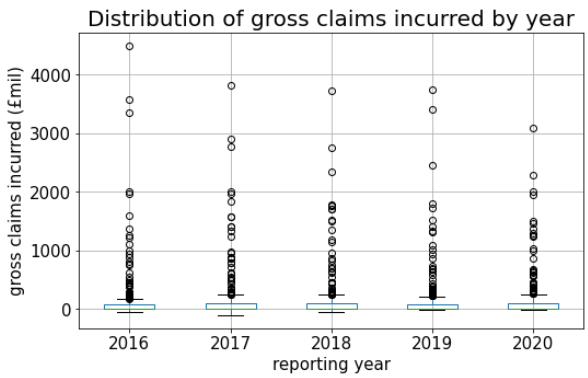
To identify outlier firms in *size*, I focus on four variables that I think best measure the size of an insurance firm. These are gross written premium (GWP), net written premium (NWP), total assets and gross claims incurred. The first two show how much business the firm is generating, total assets measures the firm's size according to the balance sheet, and gross claims incurred can be considered a measure of size on the cost side of the market. To find outliers in size, I focus my attention to average values of these variables for firms across the five reporting periods. I do this because it seems likely that from a supervisory perspective the largest firms over time are of most importance, and because it seems unlikely that the largest firms should be different from one period to the next.

Below are the plotted boxplots for these variables by year, which mostly show observations beyond the interquartile range. Some firms appearing much larger than the rest in these boxplots is consistent with the concentrated nature of insurance markets.









The table below gives an idea of the distributions of the 'size' variables, their skewness and the likely presence of outliers, as shown by high kurtosis. The large difference between mean and median, and the closeness of the means to the 90th percentile confirm the presence of a low number of very large firms.

var name	mean	10% quantile	median	90% quantile	kurtosis	skewness
GWP (£mil)	934	0	24	1721	61	7
NWP (£mil)	751	0	12	1087	78	8
Total assets (£mil)	7344	4	183	11564	53	7
Gross claims incurred (£mil)	143	0	8	296	27	5

In the next table I give the outlier threshold, above which a firm is classified as an 'outlier', and the percentile on the distribution that this threshold corresponds to. The outlier threshold is defined as three times the standard deviation above the mean. I recurringly use this (or a similar) definition to identify outliers in the report.

Variable	Outlier threshold	Percentile of threshold
GWP (£mil)	12000	97.8%
NWP (£mil)	10517	98.4%
Total Assets (£mil)	97700	98.1%
Gross claims incurred (£mil)	1260	97.2%

The final table in this section shows which firms surpass the outlier thresholds for each 'size' variable. The bolded firms are those that appear in more than one 'size' variable, and so likely will require the most attention from supervisors. Only firm 105 is an outlier across all 'size' variable. This firm is of interest also later in the report.

GWP		NWP		Total assets		Gross claims incurred	
Outliers	Value	Outliers	Value	Outliers	Value	Outliers	Value
4	**33900**	**4**	**24500**	7	93700	17	2760
26	**16000**	**26**	**16200**	10	145600	22	1350
34	**16000**			**34**	**151000**	52	1630
105	**14900**	**105**	**11600**	**105**	**185000**	**105**	**2300**
210	**39800**	**210**	**39800**	**210**	**302000**	112	3200
247	**16200**	**247**	**16000**			216	2180
311	**13700**			**311**	**268000**	234	1490
						283	1670
						286	1500

It is also of interest that most outlier firms for gross claims incurred are not outliers for written premiums or total assets. This may warrant further investigation. If the firms incurring the most claims are not also the largest, perhaps these are at risk of failing and require supervisory attention.

Identifying outliers from the norm

In this section I identify outlier firms in a more systematic way, within and across reporting periods. Using the same outlier definition as the previous section (greater than 3 x standard deviations above or below the mean), I calculate whether a firm is an outlier for each variable within each reporting period separately and add up the number of variables for which a firm is an outlier. I call this the 'outlier score'.

The full tables showing the breakup by variables for each firm are in the annex. In the below table I show the ten firms with the largest outlier scores in each reporting period. Some interesting findings arise:

- Mostly the same firms have the greatest outlier scores in each period. In every time period there are one or two new firms entering the top ten. This makes it easier to prioritise supervisory resources.
- There appear to be 'groups' of outliers. What I mean by this is that some firms are more-so outliers for 'size' type variables, some firms are more-so outliers for 'cost' type variables, and some for both. There is interesting structure here to be further explored.
- Using DBSCAN and isolation forest methods on the same data yields largely similar outliers in each reporting period. This adds robustness to the results. There is more on this in the Annex and the Jupyter notebook.

2016		2017		2018		2019		2020	
Firm	Score	Firm	Score	Firm	Score	Firm	Score	Firm	Score
105	10	105	10	105	9	105	8	105	8
34	5	34	9	311	7	4	7	34	6
101	5	4	6	4	6	101	7	311	6
311	5	311	5	34	6	311	7	**146**	6
4	5	101	5	101	5	34	6	7	5
210	4	**88**	4	210	4	**190**	4	210	4
247	4	247	4	52	3	210	4	88	4
10	3	210	4	**25**	3	88	4	17	3
17	3	**283**	3	17	3	**52**	3	52	3
216	3	17	3	216	3	17	**3**	**22**	3

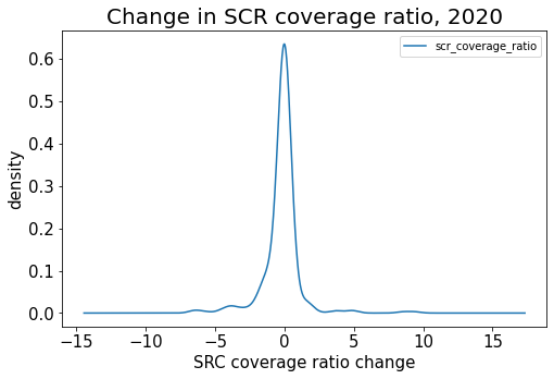
*Firms entering the top ten for the first time are bolded

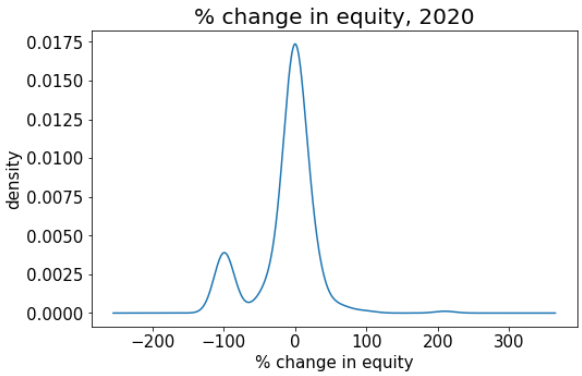
Identifying outliers in changing business profile

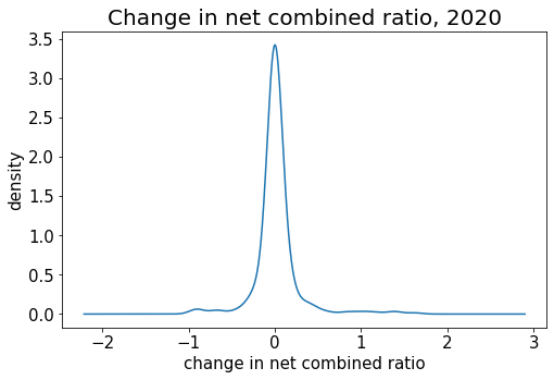
Like in section one, to identify outliers in changing business profile I focus on four variables. I think these have the most potential for harm. They are:

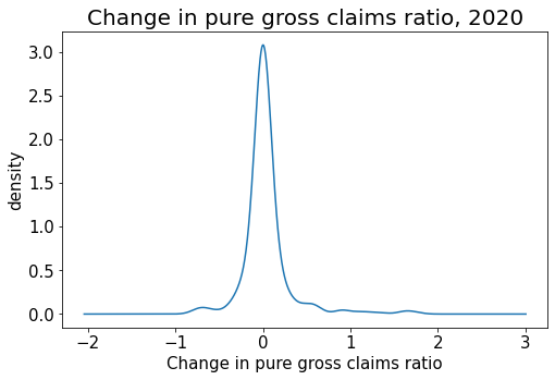
- *Change in SRC coverage ratio*: which in the case of rapid deterioration warrants the PRA's attention.
- *% change in equity*: if there are large decreases in equity this can indicate problems in the firm's balance sheet and potential future insolvency (in severe cases).
- *Change in net combined ratio*: A large increase in net combined ratio means decreasing profitability which can be a prudential risk.
- *Change in pure gross claims ratio*: similar to changes in net combined ratio, if there is a sharp increase in claims as a ratio of earned premiums that can be a significant source of cost pressure on the insurance firm.

Furthermore, I decide to focus only on the above variables for the 2020 period, which is the most recent in the data. Outliers in changing business profiles in 2020 should be more important than outliers from previous years.









In the above charts, I show the density plots for the variables of interest. Along with the next table, we see that the variables are centred around zero. However, the mean change in equity and SRC coverage ratios is negative, and the mean change in net combined ratio and pure gross claims ratios is positive. These are 'bad' trends and may reflect wider exogenous shocks caused by the pandemic. For example, the claims incurred over pandemic insurance. Another point of interest is how we treat the percentage change in equity variable. The minimum value is -100%, which is a significant deterioration of the balance sheet if true. Given this bounded nature of the variable, I choose -100% to be the outlier threshold for the % change in equity variable. The other thresholds are:

- SRC coverage ratio: **less than** 3 x below the mean
- Net combined ratio: **more than** 3 x above the mean
- Pure gross claims ratio: **more than** 3 x above the mean

The thresholds are in the directions that are potentially harmful. It seems likely that supervisors would not be concerned if an insurance firm has a sharp increase in their SRC coverage ratio.

var name	mean	10% quantile	median	90% quantile	kurtosis	skewness
SCR coverage ratio (Δ)	-1	-2	0	0	37	-5
Equity (%Δ)	-10	-100	-2	21	41	4
Net combined ratio (Δ)	0	0	0	0	40	3
Pure gross claims ratio (Δ)	0	0	0	0	40	2

Variable	Outlier threshold	Percentile of threshold
SCR coverage ratio (Δ)	13.2	2.3%
Equity (%Δ)	-100	5.0%
Net combined ratio (Δ)	1.9	98.3%
Pure gross claims ratio (Δ)	1.3	98.3%

SCR coverage ratio (Δ)		Net combined ratio (Δ)		Pure gross claims ratio (Δ)	
Outliers	Value	Outliers	Value	Outliers	Value
18	-16.1	39	5.8	21	1.6
53	-20.4	146	4.4	146	3.9
103	-31.0	203	2.3	29	1.3
163	-18.0	214	2.0	214	2.2
177	-30.0				
319	-39.2				
323	-35.5				

The outlier table for the % change in equity variable is in the Annex. None of the net combined ratio change or pure gross claims ratio change outliers have also had a 100% decrease in equity. This is reassuring. However, three outliers in SRC coverage ratio change have seen a 100% decrease in equity, and are bolded in the above table. These firms, which have had a deterioration of their balance sheets and a sharp decrease in their SRC coverage ratio should be prioritised by supervisors.

Annex

In the first part of the annex I briefly go over the ML techniques I use to support my earlier findings. I use isolation forests and DBSCAN, separately on data from each reporting period to find the top 10-15 outliers according to these methods, and check whether they line up with previous sections of the report.

Specifically, I'm interested in seeing whether my 'total outlier score' will find similar outliers to these two ML methods. Although they work quite differently, all three have largely overlapping results. This is reassuring and gives robustness to the findings.

The isolation forest is straightforwardly fitted on the data for each reporting period separately. The model provides a 'score' which allows me to rank the outliers as I've done with my 'outlier score' method. I specify the *contamination* parameter as being 0.05, which means we expect 5% of observations to be outliers. I chose this amount because I think it is reasonable that supervisors would want to target a small but not insignificant number of firms.

DBSCAN is a little different as it is effectively a clustering algorithm based on density. I iterate through multiple possible values for the 'neighbourhood size' (eps) parameter until the model is unable to categorise as close to 15 firms as possible. These are the outliers. The second parameter, min_samples, is kept constant at five. This means an insurance firm must be clustered with at least four other insurance firms for it not to be considered an outlier. I think this is a reasonable assumption given the size of the dataset.

Below are the tables showing the outlier firms for each reporting period and for each ML technique. The bolded firms are those that are also classified as a top ten outlier in the 'outlier from norm' section. As you can see, there is significant overlap between these two.

Isolation Forest Outliers (ranked)

2016	2017	2018	2019	2020
105	**4**	**4**	**4**	**105**
4	**105**	**105**	**210**	**210**
210	**34**	**210**	**105**	**311**
216	**210**	**311**	**311**	**146**
26	**311**	**34**	**88**	**34**
34	**88**	30	**34**	**88**
47	**247**	247	**101**	199
311	10	**101**	**190**	**7**
10	**101**	7	7	26
101	7	73	199	247

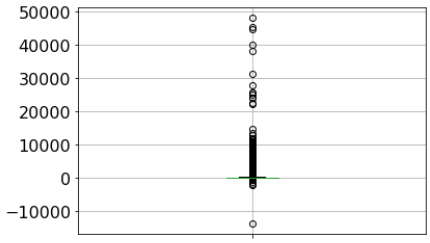
*Bolded numbers are outliers also detected in the 'outliers from norm' section.

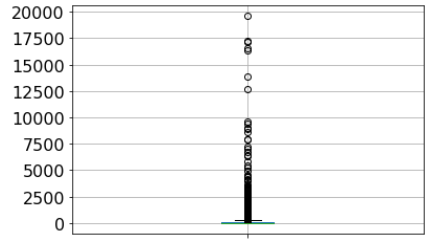
DBSCAN Outliers

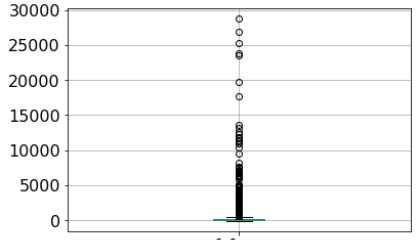
2016	2017	2018	2019	2020
1	**4**	**4**	**4**	10
4	7	7	7	**17**
10	10	10	10	26
17	27	30	**17**	**34**
26	**34**	**34**	26	39
31	88	**101**	**34**	73
34	**101**	**105**	71	**88**
101	**105**	112	73	**105**
105	112	**210**	**88**	112
210	**210**	**216**	**101**	**146**
216	**247**	247	**105**	199
247	**311**	**311**	112	**210**
311	319	316	129	234
319			177	247
			190	311
			199	
			210	
			311	

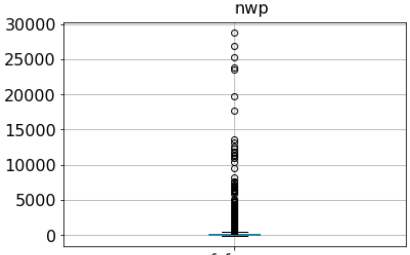
*Bolded numbers are outliers also detected in the 'outliers from norm' section.

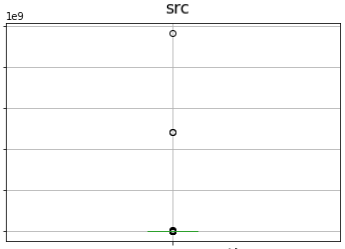
Boxplots for identifying instances of misreporting

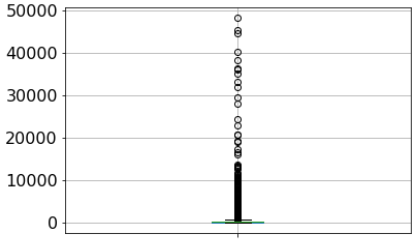


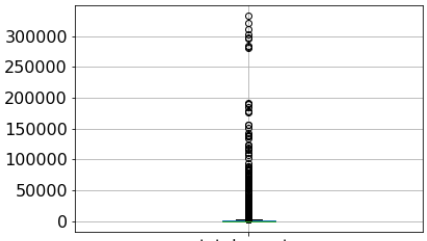


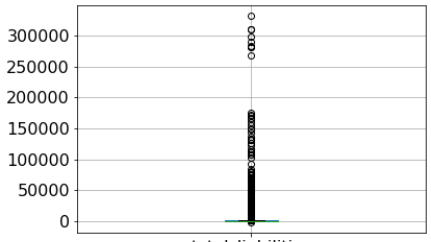


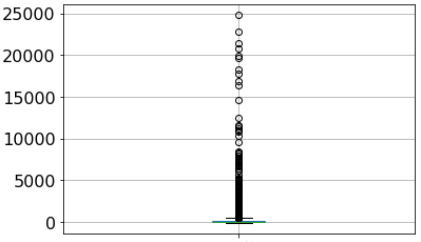


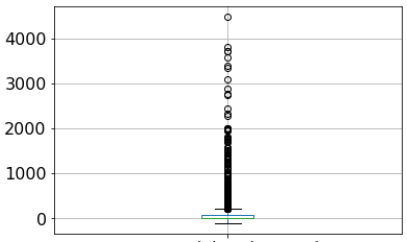


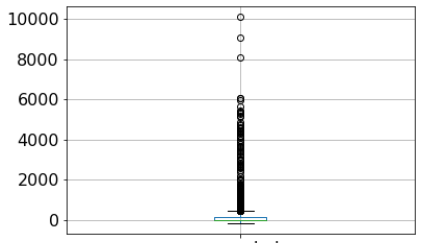


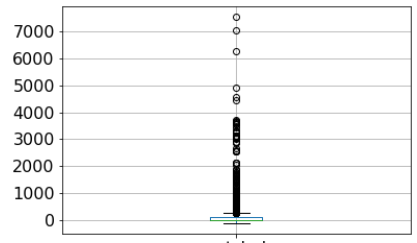


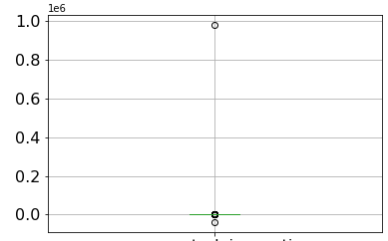


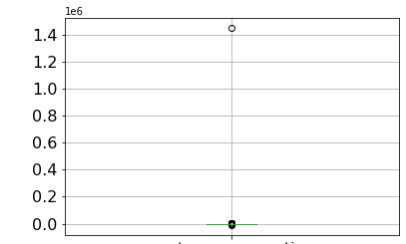


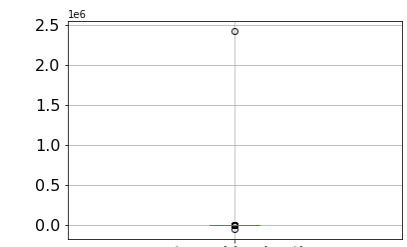


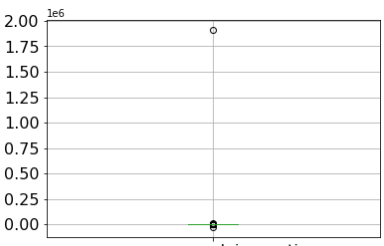


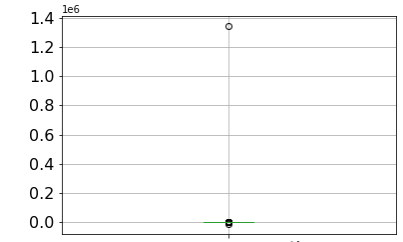


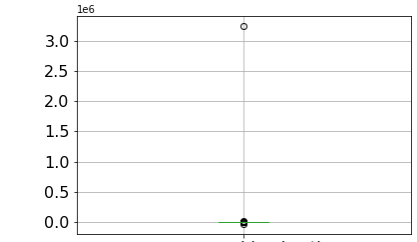












% Equity Change Outliers, 2020

Change in equity outliers (-100%)
17
18
44
54
64
84
101
106
121
124
163
164
167
171
181
193
238
239
242
243
269
272
277
289
290
306
318
319
325

Outlier from norm section tables:

2016																		
firm	**Total outlier score**	**NWP**	**SRC**	**EOF for SRC**	**SCR coverage ratio**	**GWP**	**Total assets**	**Total liabilities**	**equity**	**Gross claims incurred**	**Gross BEL**	**Net BEL**	**Pure net claims ratio**	**Net expense ratio**	**Net combined ratio**	**Pure gross claims ratio**	**Gross expense ratio**	**Gross combined ratio**
105	10	1	1	1		1	1	1	1	1	1	1						
34	5		1	1			1	1	1									
101	5		1	1			1	1	1									
311	5		1	1			1	1	1									
4	5	1	1	1		1			1									
210	4	1				1	1	1										
247	4	1				1	1	1										
10	3						1	1	1									
17	3									1	1	1						
216	3									1	1	1						

2017																		
firm	**Total outlier score**	**NWP**	**SRC**	**EOF for SRC**	**SCR coverage ratio**	**GWP**	**Total assets**	**Total liabilities**	**equity**	**Gross claims incurred**	**Gross BEL**	**Net BEL**	**Pure net claims ratio**	**Net expense ratio**	**Net combined ratio**	**Pure gross claims ratio**	**Gross expense ratio**	**Gross combined ratio**
105	10	1	1	1		1	1	1	1	1	1	1						
34	9		1	1			1	1	1					1	1		1	1
4	6	1	1	1		1			1				1					
311	5		1	1			1	1	1									
101	5		1	1			1	1	1									
88	4													1	1		1	1
247	4	1				1	1	1										
210	4	1				1	1	1										
283	3									1	1	1						
17	3									1	1	1						

2018																		
firm	Total outlier score	NWP	SRC	EOF for SRC	SCR coverage ratio	GWP	Total assets	Total liabilities	equity	Gross claims incurred	Gross BEL	Net BEL	Pure net claims ratio	Net expense ratio	Net combined ratio	Pure gross claims ratio	Gross expense ratio	Gross combined ratio
105	9		1	1		1	1	1	1	1	1	1						
311	7	1	1	1		1	1	1	1									
4	6	1	1	1		1			1				1					
34	6		1	1		1	1	1	1									
101	5		1	1			1	1	1									
210	4	1				1	1	1										
52	3									1	1	1						
25	3									1	1	1						
17	3									1	1	1						
216	3									1	1	1						

2019																		
firm	Total outlier score	NWP	SRC	EOF for SRC	SCR coverage ratio	GWP	Total assets	Total liabilities	equity	Gross claims incurred	Gross BEL	Net BEL	Pure net claims ratio	Net expense ratio	Net combined ratio	Pure gross claims ratio	Gross expense ratio	Gross combined ratio
105	8		1	1			1	1	1	1	1	1						
4	7	1	1	1		1			1				1		1			
101	7	1	1	1		1	1	1	1									
311	7	1	1	1		1	1	1	1									
34	6		1	1		1	1	1	1									
190	4												1		1	1		1
210	4	1				1	1	1										
88	4													1	1		1	1
52	3									1	1	1						
17	3									1	1	1						

2020																		
firm	**Total outlier score**	**NWP**	**SRC**	**EOF for SRC**	**SCR coverage ratio**	**GWP**	**Total assets**	**Total liabilities**	**equity**	**Gross claims incurred**	**Gross BEL**	**Net BEL**	**Pure net claims ratio**	**Net expense ratio**	**Net combined ratio**	**Pure gross claims ratio**	**Gross expense ratio**	**Gross combined ratio**
105	8		1	1			1	1	1	1	1	1						
34	6		1	1		1	1	1	1									
311	6		1	1		1	1	1	1									
146	6												1	1	1	1	1	1
7	5		1	1			1	1	1									
210	4	1				1	1	1										
88	4													1	1		1	1
17	3									1	1	1						
52	3									1	1	1						
22	3									1	1	1						