1 Visualization of different structure format

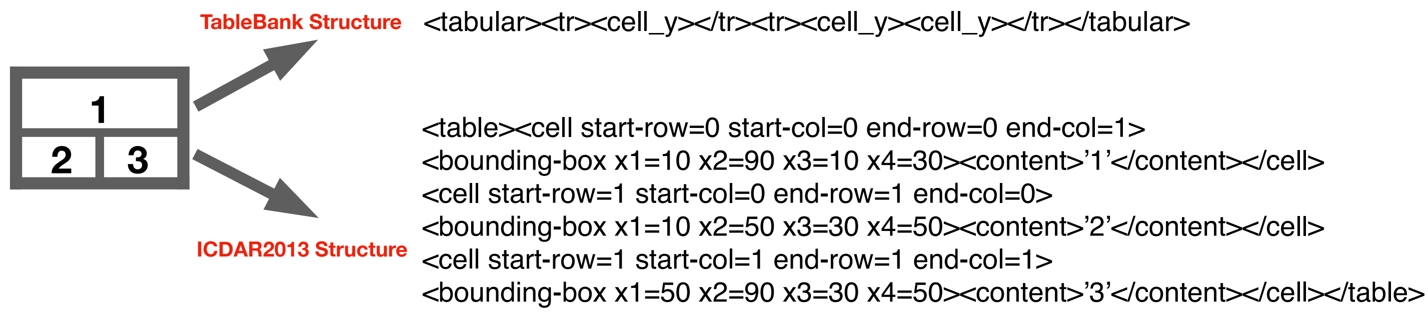


Figure 1: TableBank versus ICDAR2013 structure annotations

2 Experimental Details

2.1 GTE-Table Network

We make a few changes to the original RetinaNet model in GTE-Table. We add anchors with aspect ratio $\{0.1, 0.25\}$ to each feature map for wide tables. The input image size is $900 * 643$.

2.2 GTE-Cell Network

The GTE-Cell Network is composed of a line classifier network at the top of the hierarchy and two object detection models that specialize on different styles of tables. The graphic line classifier network is a ResNet50 model with a binary classifier on top. This network is first pretrained with the attributes derived from SD-Tables dataset and then fine-tuned on the ICDAR train dataset. The ground truth data is derived from the presence of nearby vertical graphical lines (as detected by a PDF parser) for each cell. We make the following changes to the original RetinaNet model in GTE-Cell for cell object detection. Since the scale of cell is generally small, we use pyramid levels $P3$ and $P5$. We find that skipping $P4$ allows us to add additional anchors while keeping a similar level of computational efficiency. We add anchors with aspect ratio $\{0.1, 0.25\}$ to each feature map to better detect very wide cells. For denser scale objects, at each level we use anchors of sizes $\{0.5, 0.7, 1, 1.2, 1.6\}$ of the set of aspect ratio anchors. We add additional smaller scale anchors because the majority of cells are much smaller than the anchors generated from $P3$. The input image size is $965 * 1350$.

2.3 Hyper-parameter Selection

For joint training, our hyper-parameters are selected from characteristics of the ICDAR training data. On average, the height of a character is 10 pixels. We wanted to check the text density of tables just inside and just outside of the table; we chose 5 pixels (or half a character height) for this purpose. As a result, we chose $\mu_1 = 5$ and $\mu_2 = 5$. We also chose $\alpha = 1/8$ (the density threshold) as we calculated the cell density of tables in the training set and found that the value at the lower end of the density scale (5th percentile) was around 1/8. We did not select the minimum (which was around 0.1) in case there are outliers in the training set. Finally, $\gamma_1 = 1/10$ in Eq.**??** gives less penalty to false negative bounding boxes to better reflect the proportion between false positive and false negative bounding boxes (as we found that an equal penalty caused the iterative training to become unstable very quickly).

For inference time, we found there may be overlapping tables that can be quite different in shape while having similar confidence levels. Thus, we choose a set of parameters (μ_5, μ_6, γ_2, ϵ, δ) to prioritize tables with the most tabular characteristics. In particular, we prioritize tables not having any cells within 2 lines of text outside the table ($\mu_5 = -20$ pixels) that are not contained already by other non-overlapping tables, while having many cells just inside the table, up to 0.25 of area (i.e., $\mu_6 = \{0.25 * (x2 - x1), 0.25 * (y2 - y1)\}$ pixels).

3 Cluster-based Algorithm for Generating Cell Structure

4 Additional cell detection examples

See Figures 2 and 3.

Algorithm 1 Cell Boundary to Structure Cluster Algorithm

procedure Preprocess Cell Bounding Boxes
    for b in *cellboxes* **do**
        if not INTERSECT(b, *textboxes*) **then**
            DELETE b
        if INTERSECT(b, *textboxes*) **then**
            b.bounding_box = MAX(b.bounding_box, *textbox*.bounding_box)
        if INTERSECT(b, *cellboxes*) **then**
            b.bounding_box = MAX(b.bounding_box, *cellbox*.bounding_box)
procedure Assign Cell Row and Column Location
    while not INTERSECT(b, *cellboxes*) **do**
        $b.x1 \leftarrow b.x1 - 5$
        $b.x2 \leftarrow b.x2 + 5$
    for b in *cellboxes* **do**
        $num_{col} \leftarrow$ MAX(CNT_INTERSEC($b.midx$, *cellboxes*), num_{col})
        $num_{row} \leftarrow$ MAX(CNT_INTERSEC($b.midy$, *cellboxes*), num_{row})
    $alignment_x$, $alignment_y \leftarrow$ GET_XY_ALIGNMENT(*cellboxes*)
    for b in *cellboxes* **do**
        $b.align_x \leftarrow$ ALIGN_DATA($b.x1$, $b.midx$, $b.x2$, $alignment_x$)
        $b.align_y \leftarrow$ ALIGN_DATA($b.y1$, $b.midy$, $b.y2$, $alignment_y$)
    $col_{posx} \leftarrow$ KMeans($cell_boxes.align_x$, num_{col})
    $row_{pos}x \leftarrow$ KMeans($cell_boxes.align_y$, num_{row})
    for b in *cellboxes* **do**
        $b.col \leftarrow$ ALIGN_TO_COL($b.align_x$, col_{posx}, $alignment_x$)
        $b.row \leftarrow$ ALIGN_TO_ROW($b.align_y$, col_{posy}, $alignment_y$)
procedure Assign Text Lines to Table
    for b in *textboxes* **do**
        if INTERSECT(b, *cellboxes*) **then**
            $b.col \leftarrow cellbox.col$
            $b.row \leftarrow cellbox.row$
        else
            $b.col \leftarrow$ ALIGN_TO_COL($b.align_x$, col_{posx}, $alignment_x$)
            $b.row \leftarrow$ ALIGN_TO_ROW($b.align_y$, col_{posy}, $alignment_y$)
procedure Split Cell Text Lines When Neighbor is Empty
    for r in num_{row} **do**
        for c in num_{col} **do**
            if IS_EMPTY(r, c) **then**
                $neighbor_{text} \leftarrow$ GET_CELLS($r-1$, c) + GET_CELLS($r+1$, c)
                for b in $neighbor_{text}$ **do**
                    $b.col \leftarrow$ ALIGN_TO_COL($b.align_x$, col_{posx}, $alignment_x$)
                    $b.row \leftarrow$ ALIGN_TO_ROW($b.align_y$, col_{posy}, $alignment_y$)

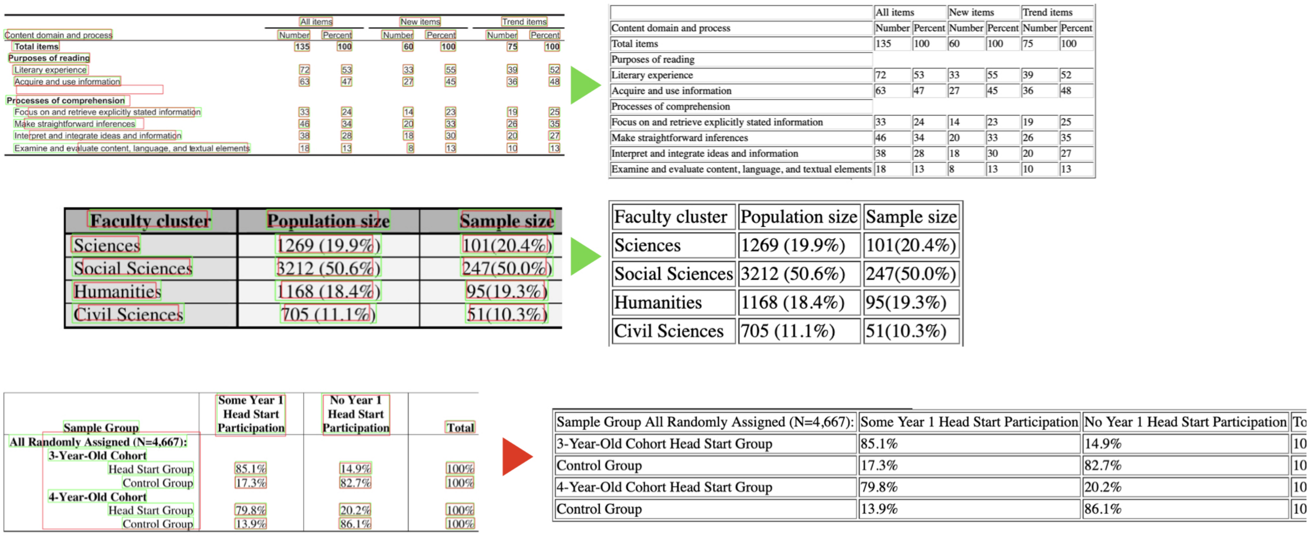


Figure 2: Additional cell boundary to structure examples

5 Detailed ICDAR13 Results

See Tables 1 and 2.

Table 1: ICDAR 2013 table detection results with additional comparisons

Category	Method	Input type	Recall	Precision	F1	Cpt	Pu
Commercial Softwares	*FineReader*	PDF	99.71	97.29	98.48	142	148
	OmniPage	PDF	96.44	95.69	96.06	141	130
	Nitro	PDF	93.23	93,97	93.60	124	144
	Acrobat	PDF	87.38	93.65	90.40	110	141
Non Deep Learning	*ICST-Table*[1]	PDF	26.97	74.96	39.67	28	41
	TableSeer[6]	PDF	33.35	88.36	48.64	0	29
	Nurminen[2]	PDF	90.77	92.10	91.43	114	151
	TABFIND[9]	PDF	98.31	92.92	95.54	149	137
	pdf2table[11]	PDF	85.30	63.99	73.13	100	94
	TEXUS[7]	PDF	90.23	88.32	89.26	114	138
Deep Learning	*Hao*[3]	Image	97.24	92.15	94.63	/	/
	DeepDeSRT[8]	Image	96.15	97.40	96.77	/	/
	TableBank[5]	Image	/	/	96.25	/	/
Ours	GTE	Image	99.77	98.97	**99.31**	146	146

Table 2: Cell Structure results on ICDAR2013 with additional comparisons

Category	Method	GT Border?	Rec.	Prec.	F1
Commercial Softwares	*FineReader*	N	88.35	87.10	87.72
	OmniPage	N	83.80	84.60	84.20
	Nitro	N	67.93	84.59	75.35
	Acrobat	N	72.62	81.59	76.85
Academic Systems	*Nurminen*[2]	N	80.78	86.93	83.74
	TEXUS[7]	N	84.23	81.02	82.59
	KYTHE[2]	N	48.11	57.40	52.20
	pdf2table[11]	N	59.51	57.52	58.50
	TABFIND[9]	N	70.52	68.74	69.62
Ours	GTE	N	92.72	94.41	**93.50**
Academic Systems	*Tensmeyer*[10]	Y	94.64	95.89	95.26
	Nurminen[2]	Y	94.09	95.12	94.60
	Khan[4]	Y	90.12	96.92	93.39
	TABFIND[9]	Y	64.01	61.44	62.70
Ours	GTE	Y	95.77	96.76	**96.24**

6 ICDAR19 evaluation metric ambiguities

See Figure 4.

References

[1] J. Fang, L. Gao, K. Bai, R. Qiu, X. Tao, and Z. Tang. A table detection method for multipage pdf documents via visual seperators and tabular structures. *2011 International Conference on Document Analysis and Recognition*, pages 779–783, 2011.

[2] M. C. Göbel, T. Hassan, E. Oro, and G. Orsi. Icdar 2013 table competition. *2013 12th International Conference on Document Analysis and Recognition*, pages 1449–1453, 2013.

[3] L. Hao, L. Gao, X. Yi, and Z. Tang. A table detection method for pdf documents based on convolutional neural networks. *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, pages 287–292, 2016.

[4] S. A. Khan, S. M. D. Khalid, M. A. Shahzad, and F. Shafait. Table structure extraction with bi-directional gated recurrent unit networks. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1366–1371. IEEE, 2019.

[5] M. Li, L. Cui, S. Huang, F. Wei, M. Zhou, and Z. Li. Tablebank: Table benchmark for image-based table detection and recognition. *ArXiv*, abs/1903.01949, 2019.

[6] Y. Liu, K. Bai, P. Mitra, and C. L. Giles. Tableseer: automatic table metadata extraction and searching in digital libraries. In *JCDL*, 2007.

[7] R. Rastan, H.-Y. Paik, and J. Shepherd. Texus: A task-based approach for table extraction and understanding. In *DocEng*, 2015.

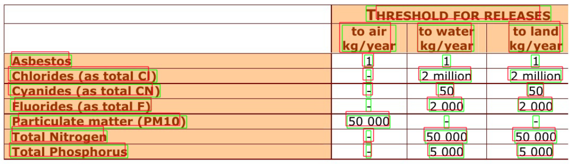
[8] S. Schreiber, S. Agne, I. Wolf, A. Dengel, and S. Ahmed. Deepdesrt: Deep learning for detection and structure recognition of tables in document images. *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 01:1162–1167, 2017.

[9] A. C. e. Silva. Parts that add up to a whole: a framework for the analysis of tables. *PH.D. dissertation, The University of Edinburgh*, 2010.

[10] C. Tensmeyer, V. I. Morariu, B. Price, S. Cohen, and T. Martinez. Deep splitting and merging for table structure decomposition. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 114–121. IEEE, 2019.

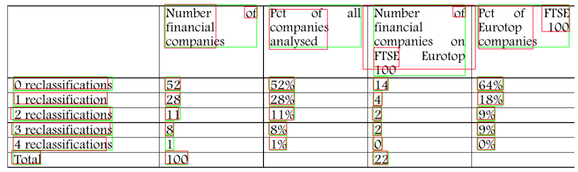
[11] B. Yildiz, K. Kaiser, and S. Miksch. pdf2table: A method to extract table information from pdf files. In *IICAI*, 2005.

Figure 3: Example Cell detection errors



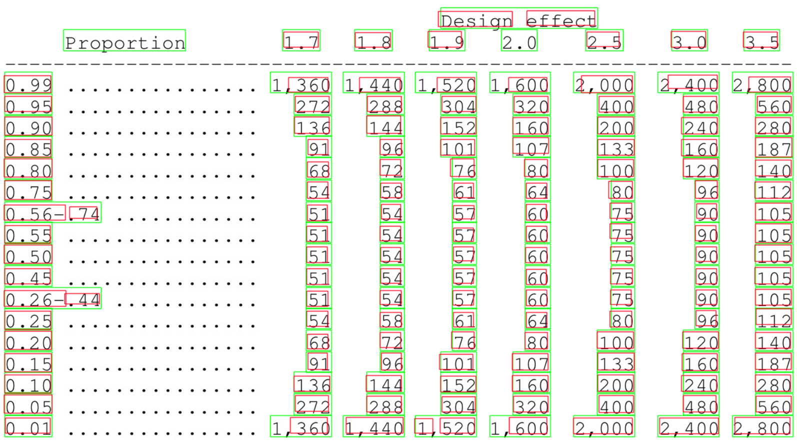
	THRESHOLD FOR RELEASES		
	to air kg/year	to water kg/year	to land kg/year
Asbestos	1	1	1
Chlorides (as total Cl)	-	2 million	2 million
Cyanides (as total CN)	-	50	50
Fluorides (as total F)	-	2 000	2 000
Particulate matter (PM10)	50 000	-	-
Total Nitrogen	-	50 000	50 000
Total Phosphorus	-	5 000	5 000

(a) Correct cell detection



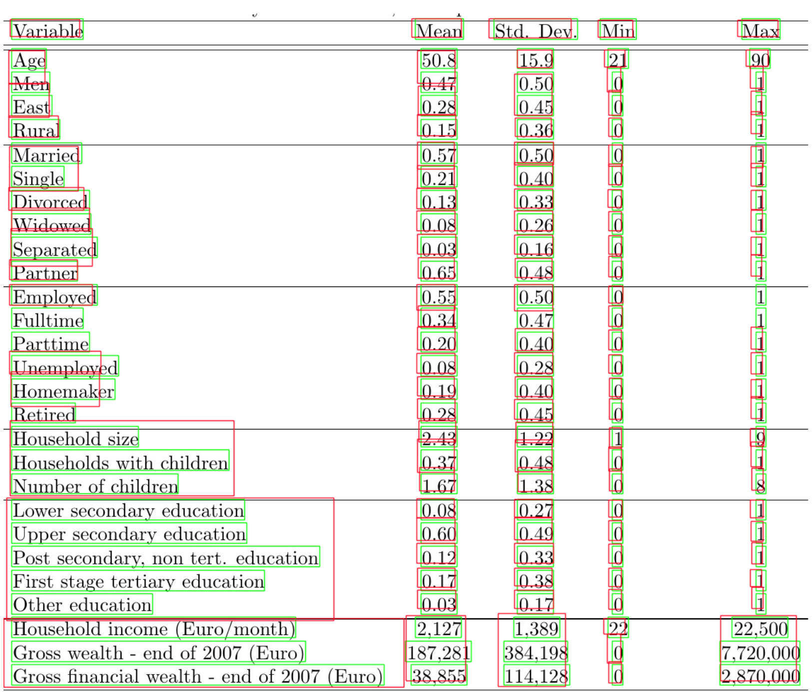
	Number of financial companies	Pct of all companies analysed	Number of financial companies on FTSE Eurotop 100	Pct of Eurotop FTSE 100 companies
0 reclassifications	52	52%	14	64%
1 reclassification	28	28%	4	18%
2 reclassifications	11	11%	2	9%
3 reclassifications	8	8%	2	9%
4 reclassifications	1	1%	0	0%
Total	100		22	

(b) Oversplit cell detection



			Design effect				
Proportion	1.7	1.8	1.9	2.0	2.5	3.0	3.5
0.99	1,360	1,440	1,520	1,600	2,000	2,400	2,800
0.95	272	288	304	320	400	480	560
0.90	136	144	152	160	200	240	280
0.85	91	96	101	107	133	160	187
0.80	68	72	76	80	100	120	140
0.75	54	58	61	64	80	96	112
0.56-.74	51	54	57	60	75	90	105
0.55	51	54	57	60	75	90	105
0.50	51	54	57	60	75	90	105
0.45	51	54	57	60	75	90	105
0.26-.44	51	54	57	60	75	90	105
0.25	54	58	61	64	80	96	112
0.20	68	72	76	80	100	120	140
0.15	91	96	101	107	133	160	187
0.10	136	144	152	160	200	240	280
0.05	272	288	304	320	400	480	560
0.01	1,360	1,440	1,520	1,600	2,000	2,400	2,800

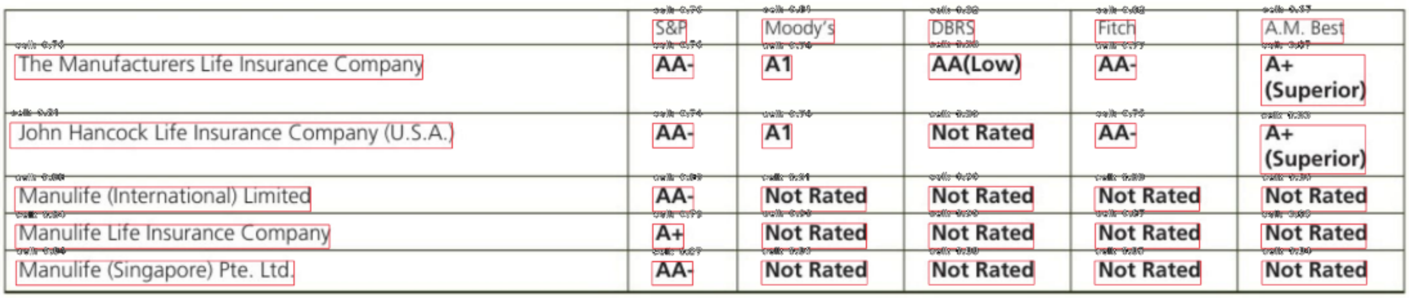
(c) Missing Cell detection



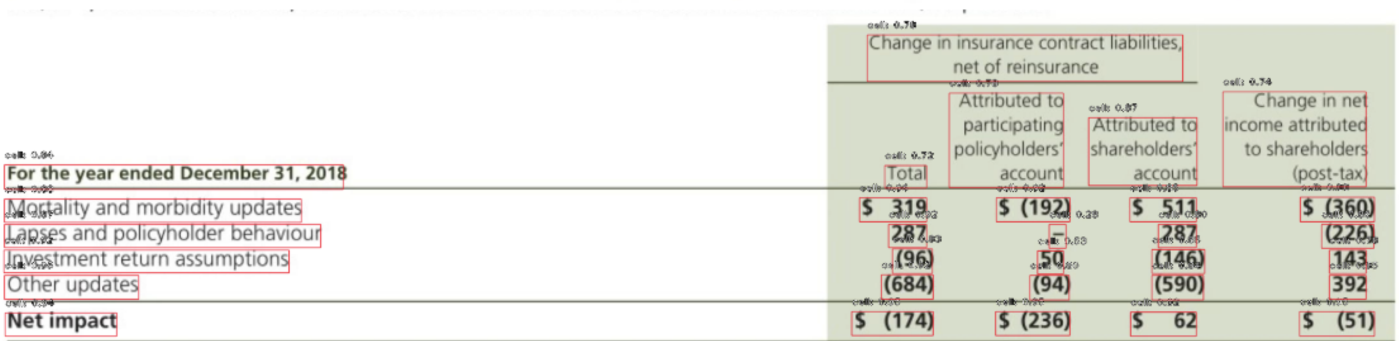
Variable	Mean	Std. Dev.	Min	Max
Age	50.8	15.9	21	90
Men	0.47	0.50	0	1
East	0.28	0.45	0	1
Rural	0.15	0.36	0	1
Married	0.57	0.50	0	1
Single	0.21	0.40	0	1
Divorced	0.13	0.33	0	1
Widowed	0.08	0.26	0	1
Separated	0.03	0.16	0	1
Partner	0.65	0.48	0	1
Employed	0.55	0.50	0	1
Fulltime	0.34	0.47	0	1
Parttime	0.20	0.40	0	1
Unemployed	0.08	0.28	0	1
Homemaker	0.19	0.40	0	1
Retired	0.28	0.45	0	1
Household size	2.43	1.22	1	9
Households with children	0.37	0.48	0	1
Number of children	1.67	1.38	0	8
Lower secondary education	0.08	0.27	0	1
Upper secondary education	0.60	0.49	0	1
Post secondary, non tert. education	0.12	0.33	0	1
First stage tertiary education	0.17	0.38	0	1
Other education	0.03	0.17	0	1
Household income (Euro/month)	2,127	1,389	22	22,500
Gross wealth - end of 2007 (Euro)	187,281	384,198	0	7,720,000
Gross financial wealth - end of 2007 (Euro)	38,855	114,128	0	2,870,000

(d) Overmerged cell detection

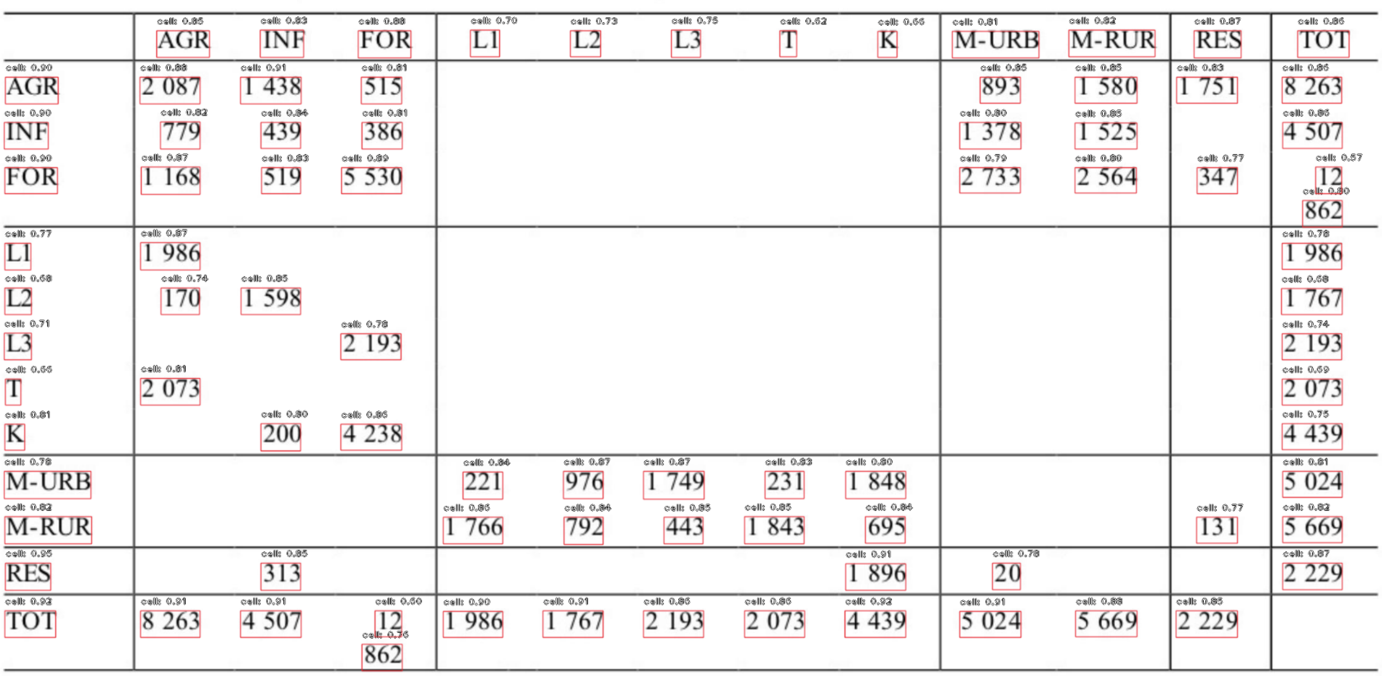
Figure 4: The detected cell bounding boxes in the following images seem to be correct by eye and include all characters in the ground truth cell but has zero matches at IOU=0.9.



	S&P	Moody's	DBRS	Fitch	A.M. Best
The Manufacturers Life Insurance Company	AA-	A1	AA(Low)	AA-	A+ (Superior)
John Hancock Life Insurance Company (U.S.A.)	AA-	A1	Not Rated	AA-	A+ (Superior)
Manulife (International) Limited	AA-	Not Rated	Not Rated	Not Rated	Not Rated
Manulife Life Insurance Company	A+	Not Rated	Not Rated	Not Rated	Not Rated
Manulife (Singapore) Pte. Ltd.	AA-	Not Rated	Not Rated	Not Rated	Not Rated



	Change in insurance contract liabilities, net of reinsurance			
For the year ended December 31, 2018	Total	Attributed to participating policyholders' account	Attributed to shareholders' account	Change in net income attributed to shareholders (post-tax)
Mortality and morbidity updates	$ 319	$ (192)	$ 511	$ (360)
Lapses and policyholder behaviour	287	–	287	(226)
Investment return assumptions	(96)	50	(146)	143
Other updates	(684)	(94)	(590)	392
Net impact	$ (174)	$ (236)	$ 62	$ (51)



	AGR	INF	FOR	L1	L2	L3	T	K	M-URB	M-RUR	RES	TOT
AGR	2 087	1 438	515						893	1 580	1 751	8 263
INF	779	439	386						1 378	1 525		4 507
FOR	1 168	519	5 530						2 733	2 564	347	12 862
L1	1 986											1 986
L2	170	1 598										1 767
L3			2 193									2 193
T	2 073											2 073
K		200	4 238									4 439
M-URB				221	976	1 749	231	1 848				5 024
M-RUR				1 766	792	443	1 843	695			131	5 669
RES		313						1 896	20			2 229
TOT	8 263	4 507	12 862	1 986	1 767	2 193	2 073	4 439	5 024	5 669	2 229	