```
Stichtag,Geschlecht,Altersgruppen (unter 6; 65 oder älter),,,,,,,,,
,,Insgesamt,unter 6,6 bis unter 15,15 bis unter 18,18 bis unter 25,25 bis unter 30,30 bis
unter 40,40 bis unter 50,50 bis unter 65,ab 65
31.12.2006,männlich,6118977,344474,604411,222077,512433,377722,893802,1064334,1109712,990012
,weiblich,6373681,327284,573127,210589,503504,378591,868923,1021520,1118752,1371391
,Insgesamt,12492658,671758,1177538,432666,1015937,756313,1762725,2085854,2228464,2361403
31.12.2010,männlich,6158439,330333,562446,205868,537235,386600,782873,1084206,1215578,1053300
,weiblich,6380257,314057,533717,194213,515805,382400,774349,1044177,1226852,1394687
,Insgesamt,12538696,644390,1096163,400081,1053040,769000,1557222,2128383,2442430,2447987
31.12.2015,männlich,6352172,350563,528115,209020,549997,443322,826411,938870,1378744,1127130
,weiblich,6491342,332761,501078,189234,502978,412664,802903,919640,1385878,1444206
,Insgesamt,12843514,683324,1029193,398254,1052975,855986,1629314,1858510,2764622,2571336
31.12.2020,männlich,6512595,394325,538965,182735,527821,435653,901337,823568,1500896,1207295
,weiblich,6627588,375841,510302,174114,478672,401570,855105,820260,1494956,1516768
,Insgesamt,13140183,770166,1049267,356849,1006493,837223,1756442,1643828,2995852,2724063
© Bayerisches Landesamt für Statistik; Fürth 2022 | Stand: 09.12.2022 / 09:48:53,,,,,,,,,,
```

Aggregation Detection in CSV Files

Outline

- Problem definition
- Aggregation detection approach
  - Aggregations in CSV files
  - AGGRECOL workflow
- Evaluation
- Discussion

Aggregations

- Sum: $A = \sum_{i=1}^{n} B_i$
- Average: $A = \frac{1}{n} \sum_{i=1}^{n} B_i$
- Difference: $A = B - C$
- Division: $A = B/C$
- Rel. Change: $A = (C - B)/B$

Aggregations

- Number format
- Error-Level
- Coverage:
  Proportion of columns with same aggregation

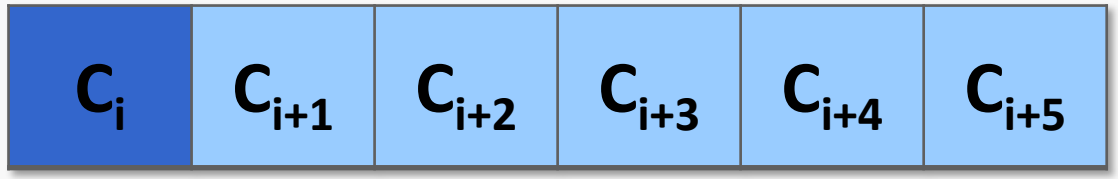
A=B/C	B	C
10.5771	20345	1923.5

AggreCol

Workflow in three stages

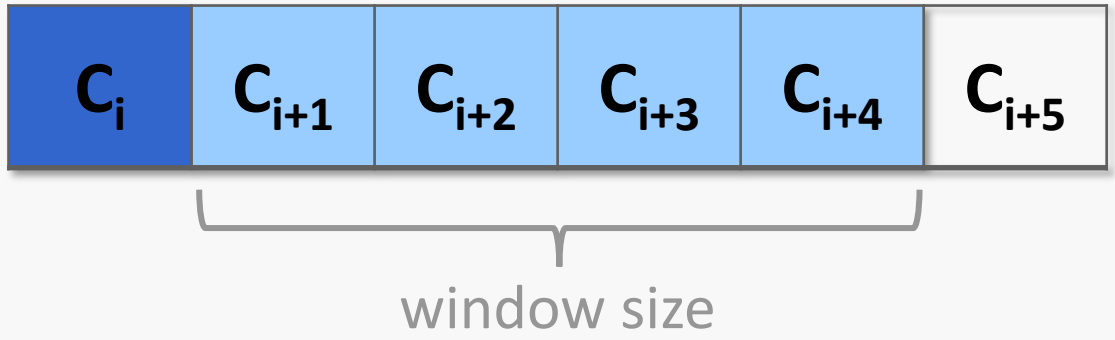
AggreCol: Stage 1

- Individual aggregation detection
  - commutative aggregations: adjacency list strategy



AggreCol: Stage 1

- Individual aggregation detection
  - commutative aggregations: adjacency list strategy
  - non-commutative aggregations: sliding window strategy



AggreCol: Stage 1

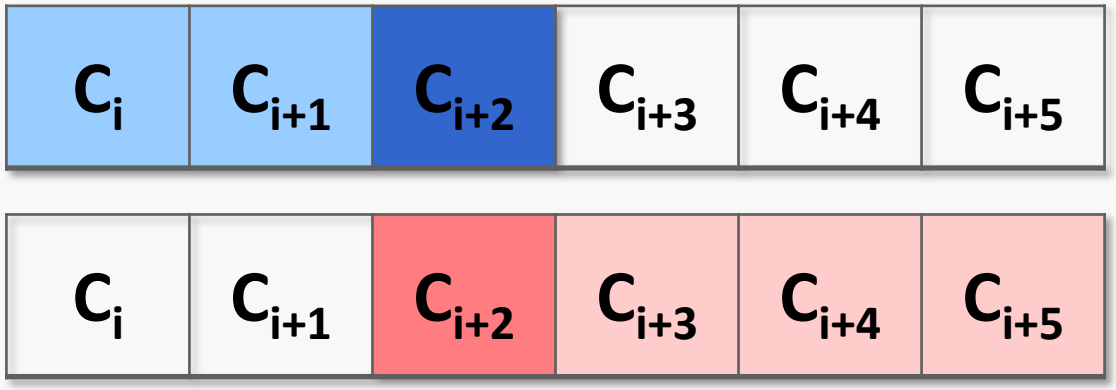
- Prioritization of groups
  - Number of group members
  - Average error

Type: Sum,
Aggregator: c_{j5},
Range elems: {c_{j6}, c_{j7}, c_{j8}}

AggreCol: Stage 1

- Prioritization of groups
  - Number of group members
  - Average error
- Pruning rules
  - Directional disagreement

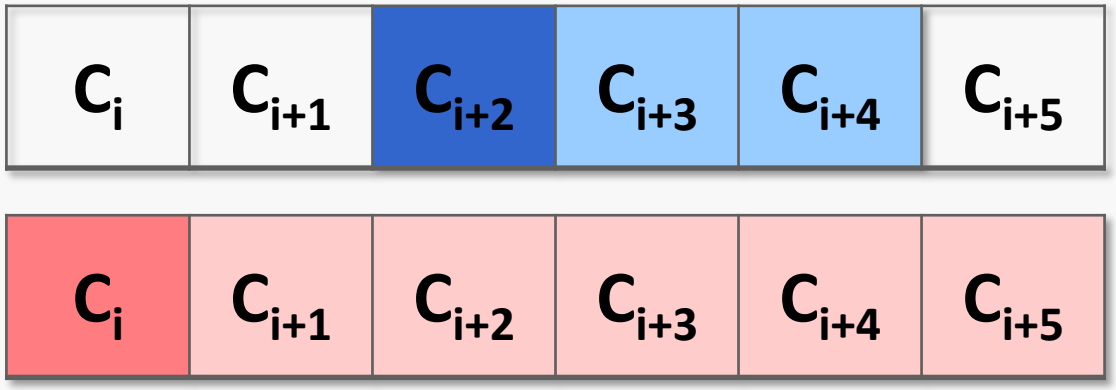
Type: Sum,
Aggregator: c_{j5},
Range elems: $\{c_{j6}, c_{j7}, c_{j8}\}$



AggreCol: Stage 1

- Prioritization of groups
  - Number of group members
  - Average error
- Pruning rules
  - Directional disagreement
  - Complete inclusion

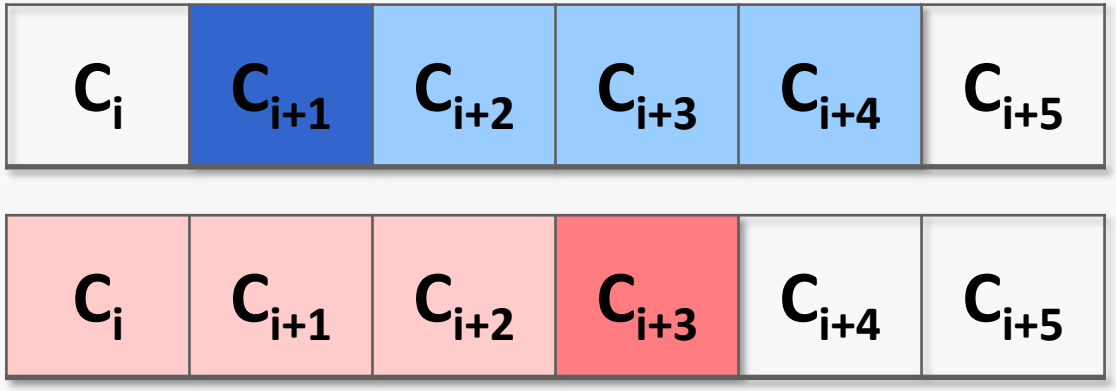
Type: Sum,
Aggregator: c_{j5},
Range elems: $\{c_{j6}, c_{j7}, c_{j8}\}$



AggreCol: Stage 1

- Prioritization of groups
  - Number of group members
  - Average error
- Pruning rules
  - Directional disagreement
  - Complete inclusion
  - Mutual inclusion

Type: Sum,
Aggregator: c_{j5},
Range elems: $\{c_{j6}, c_{j7}, c_{j8}\}$



AggreCol: Stage 2

Collective aggregation detection

- Prioritization of groups
- Pruning rules
  - Complete inclusion
  - Mutual inclusion
  - Same aggregator with overlapping ranges

AggreCol: Stage 3

Supplemental aggregation detection

- Loop of:
  - Modifying files
  - Individual aggregation detection
- Pruning as in stage 1

Outline

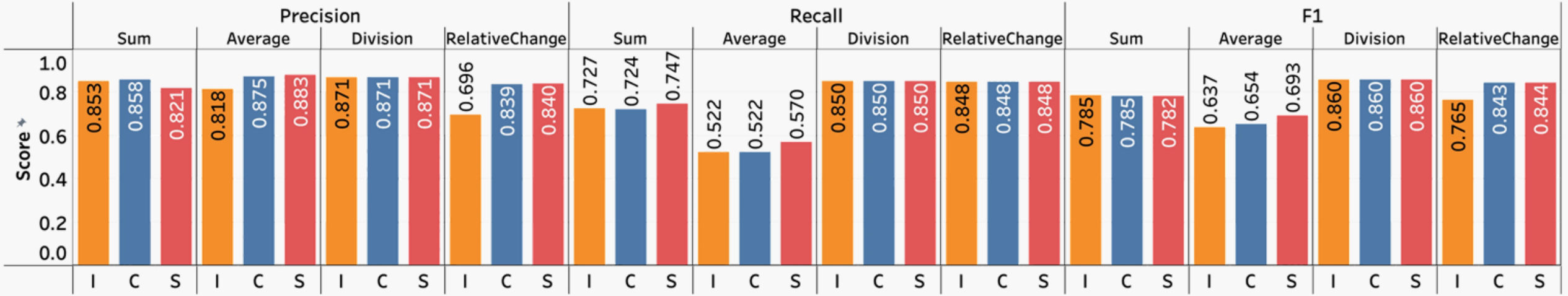
- Problem definition
- Aggregation detection approach
- Evaluation
  - Methods
  - Results
- Discussion

Evaluation Methods

- Metrics:
  - Precision: correctly detected / all detected
  - Recall: correctly detected / actual aggregations
  - F_1-score: harmonic mean of precision and recall
- Validation dataset and unseen dataset
- Empirical tests:
  - Consideration of the individual stages
  - Comparison of the results of both datasets

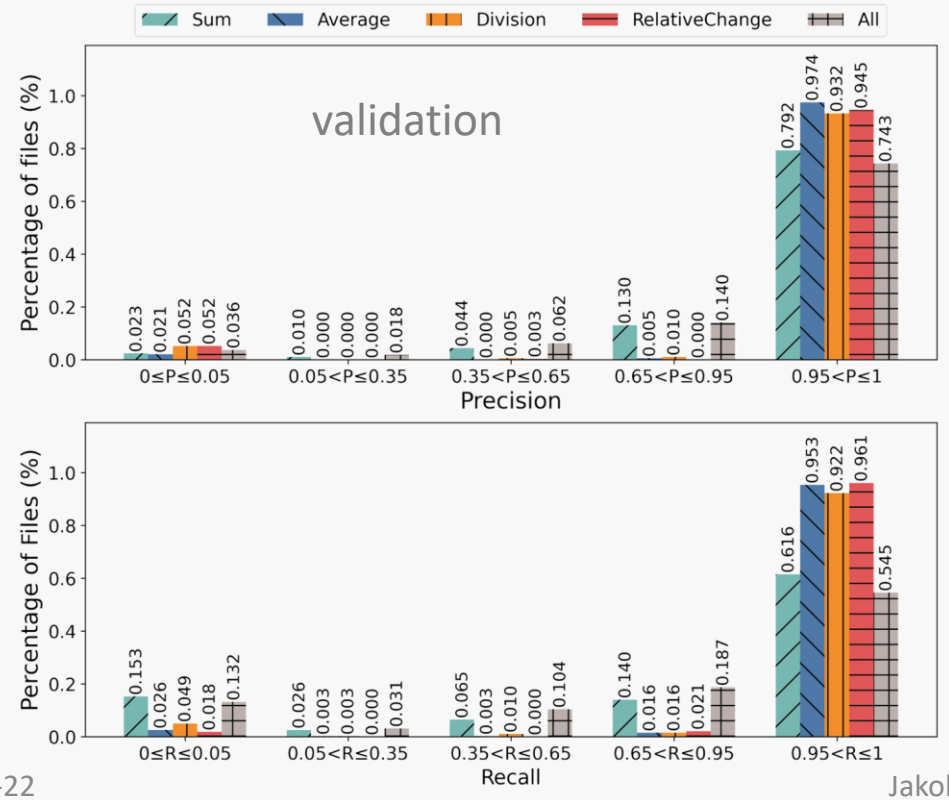
Results

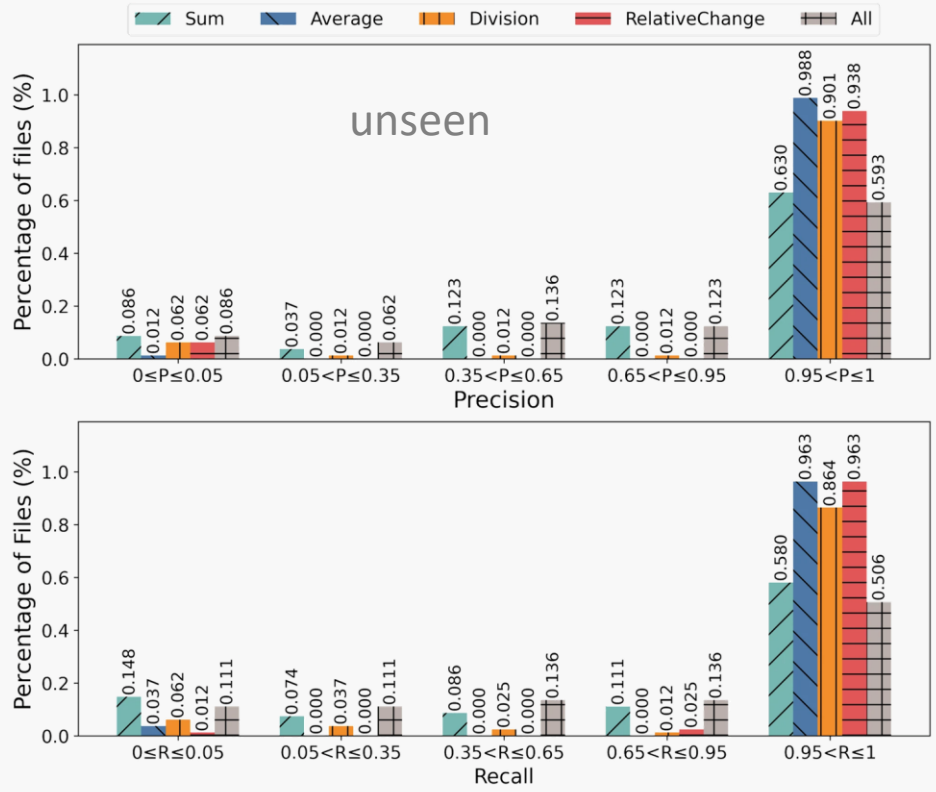
- Consideration of the individual stages



Results

- Consideration of the individual stages
- Comparison of the results of both datasets (file-level)





Outline

- Problem definition
- Aggregation detection approach
- Evaluation
- Discussion

Discussion

- Conclusion
  - First proposed solution for this problem
  - In terms of metrics, promising results for a large part of the files (exception: sum)
- Criticism/Proposals
  - File-level evaluation reasonable, but biased scores
  - Error level per file
  - Extension by keyword matching
- Example

Sources

- Paper:
  Lan Jiang, Gerardo Vitagliano, Mazhar Hameed, and Felix Naumann. 2022. Aggregation Detection in CSV Files. In Proceedings of the 25th International Conference on Extending Database Technology, EDBT 2022, Edinburgh, UK, March 29 - April 1, 2022. OpenProceedings.org, 2:207–2:219. https://doi.org/10.48786/edbt.2022.10
- Source Code:
  Lan Jiang, Gerardo Vitagliano, Mazhar Hameed, and Felix Naumann. 2021. AggreCol Version 0.1. https://github.com/lanchiang/AggreCol. [Online; accessed 24-November-2022]
- Example data:
  Bayerisches Landesamt für Statistik, 2022
  https://www.statistikdaten.bayern.de/genesis//online?operation=table&code=12411-004z&bypass=true&levelindex=0&levelid=1670575612358#abreadcrumb