RSMB Limited
77 Kingsway, London, WC2B 6SR
Tel: +44 (0) 20 4570 6886
Email: contact@rsmb.co.uk
Web: www.rsmb.solutions

RSMB

Comparing Constrained vs Unconstrained Data Fusion

A practical guide to probabilistic data integration.

White Paper

Noel O'Sullivan, Chief Statistician, © RSMB Ltd. 2025

Introduction

In today's advertising and marketing landscape, data is everything. Integrating diverse datasets into something coherent and actionable is a major challenge, and this is where data fusion comes in. By combining information from multiple sources, data fusion allows advertisers and marketers to get a complete picture of their audiences. This paper explores two key approaches to data fusion, unconstrained and constrained, examining their strengths, limitations, and applications in the advertising and marketing industries.

What is Data Fusion?

At its core, data fusion is the process of merging information from different datasets to create a unified, richer dataset. For example, combining a survey on media consumption with another on purchasing habits can provide a more comprehensive analysis by linking them through shared characteristics such as demographics or geographic location. This enables advertisers and marketers to analyse previously unconnected variables, like the correlation between TV viewership and buying behaviour.

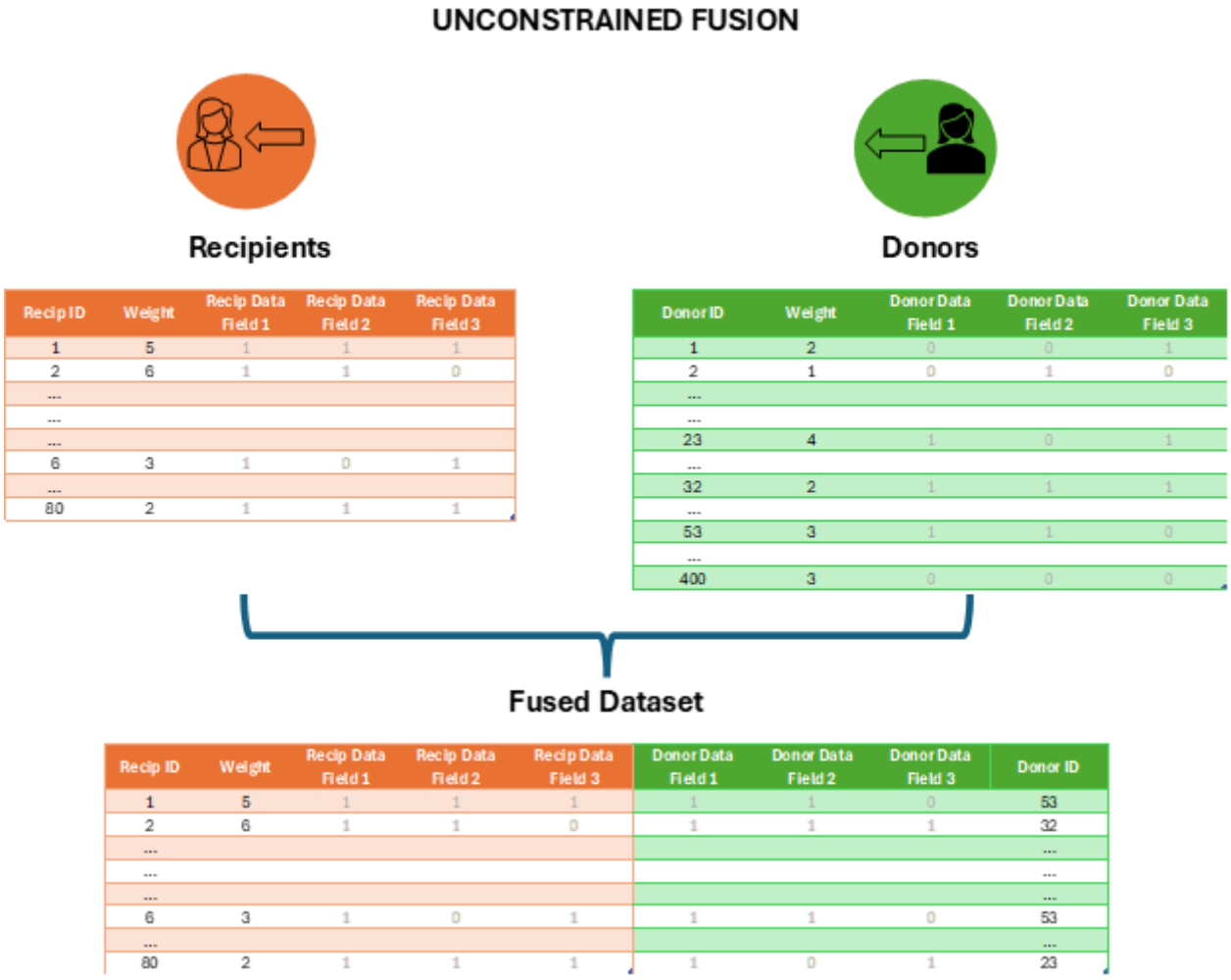
What is Data Fusion?

Directly matching records using unique identifiers is often impractical in the advertising and marketing world because many datasets don't share common identifiers and their structures may differ significantly. For example, a TV viewing dataset might capture household-level behaviours, while a retail dataset records individual purchase histories. Deterministic methods struggle to integrate such datasets.

Data fusion relies on probabilistic techniques to link datasets. In advertising, this is particularly valuable for tasks like cross-platform audience measurement—combining TV panel data with digital analytics to map out audience interactions across channels. Another example is integrating campaign exposure data with consumer surveys to understand how advertising impacts purchase intent. These capabilities make data fusion an essential tool when deterministic methods (i.e. direct matching by some sort of identifier like an email address) fall short

Unconstrained Data Fusion

Unconstrained fusion takes a flexible approach, designating one dataset as the "recipient" and the other as the "donor." The fusion algorithm matches donor records to recipient records based on similarity in shared variables, such as demographics. The recipient dataset's structure and composition remain intact, while the donor data is adjusted to fit. The diagram below illustrates how unconstrained fusion works.



Recipients

Recip ID	Weight	Recip Data Field 1	Recip Data Field 2	Recip Data Field 3
1	5	1	1	1
2	6	1	1	0

6	3	1	0	1

80	2	1	1	1

Donors

Donor ID	Weight	Donor Data Field 1	Donor Data Field 2	Donor Data Field 3
1	2	0	0	1
2	1	0	1	0

23	4	1	0	1

32	2	1	1	1

53	3	1	1	0

400	3	0	0	0

Fused Dataset

Recip ID	Weight	Recip Data Field 1	Recip Data Field 2	Recip Data Field 3	Donor Data Field 1	Donor Data Field 2	Donor Data Field 3	Donor ID
1	5	1	1	1	1	1	0	53
2	6	1	1	0	1	1	1	32
---								---
---								---
---								---
6	3	1	0	1	1	1	0	53
---								---
80	2	1	1	1	1	0	1	23

There are a couple of points to note in particular from the diagram:

- The recipient data is unchanged (e.g. 80 records, same weight).
- Donors can be used more than once if a good match e.g. Donor ID 53 has been used twice.
- There are only 80 matches so many of the 400 donors will not be used.
- The above is when the donor sample is bigger than the recipient sample; in some instances the opposite will be true. In these cases it is likely that most if not all of the recipients will be used but inevitably some will be used more than once.

As unconstrained fusion allows donor records to be reused or excluded, it delivers higher precision when exploring granular behaviours or niche audience segments. Unconstrained fusion enables the selection of an optimal record for the recipient, regardless of varying sample sizes. Even if the donor is utilised multiple times, it remains a highly suitable (and potentially the best) match for the respondent. This approach theoretically supports the creation of a fused product with the most accurate cross-correlations, which is the ultimate objective of the fusion process.

However, excluding some donor records might introduce biases, skewing results. If the donor dataset represents highly fragmented consumer behaviours, underutilisation could result in missed insights. Despite these challenges, the flexibility and computational efficiency of unconstrained fusion make it ideal for projects with tight timelines or budgets.

Key Characteristics

Optimisation-Focused: The matches prioritise closeness of common variables, such as demographics or media behaviours, rather than equal distribution of weights. This ensures higher precision in reflecting real-world correlations.

Recipient Priority: The recipient dataset retains its structure, including weights and sample composition. Donor data is adjusted to fit the recipient's framework.

Advantages

- Flexibility in Matching: By allowing donors to be reused or excluded, the algorithm achieves higher granularity and precision.
- Retention of Recipient Data Integrity: Ensures that critical metrics in the recipient dataset remain unchanged, which is beneficial in media or audience research where the recipient dataset often defines currency (e.g., television ratings or survey panels).
- Cost and Efficiency: Requires fewer computational resources than constrained fusion, making it suitable for projects with tight timelines or budgets.
- Calibration: For donated continuous variables (e.g., consumption metrics), there is potential to calibrate any distortions from the original dataset.

Limitations

- Risk of Sample Bias: Excluding some donor records can lead to an imbalance, potentially skewing results if donor data represents key population segments.
- Loss of Donor Metrics: Important metrics in the donor dataset, like weight distributions or coverage ratios, may be diluted or lost altogether.
- Currency Mismatch: Outputs may lack the balance needed for applications that rely on precise representation from both datasets.

Best Use Cases

- Exploratory Research: When the goal is to generate hypotheses or insights without strict representational requirements.
- Media and Consumer Behaviour Analysis: Ideal for integrating audience behaviour datasets with rich demographic information.

Constrained Data Fusion

Constrained fusion is more structured, ensuring that every record from both datasets is used exactly once. This method uses mathematical optimisation to fragment and align records based on shared characteristics and preserving the weight of both datasets simultaneously. The result is an equitable dataset where neither source dominates. This fragmentation and alignment means that:

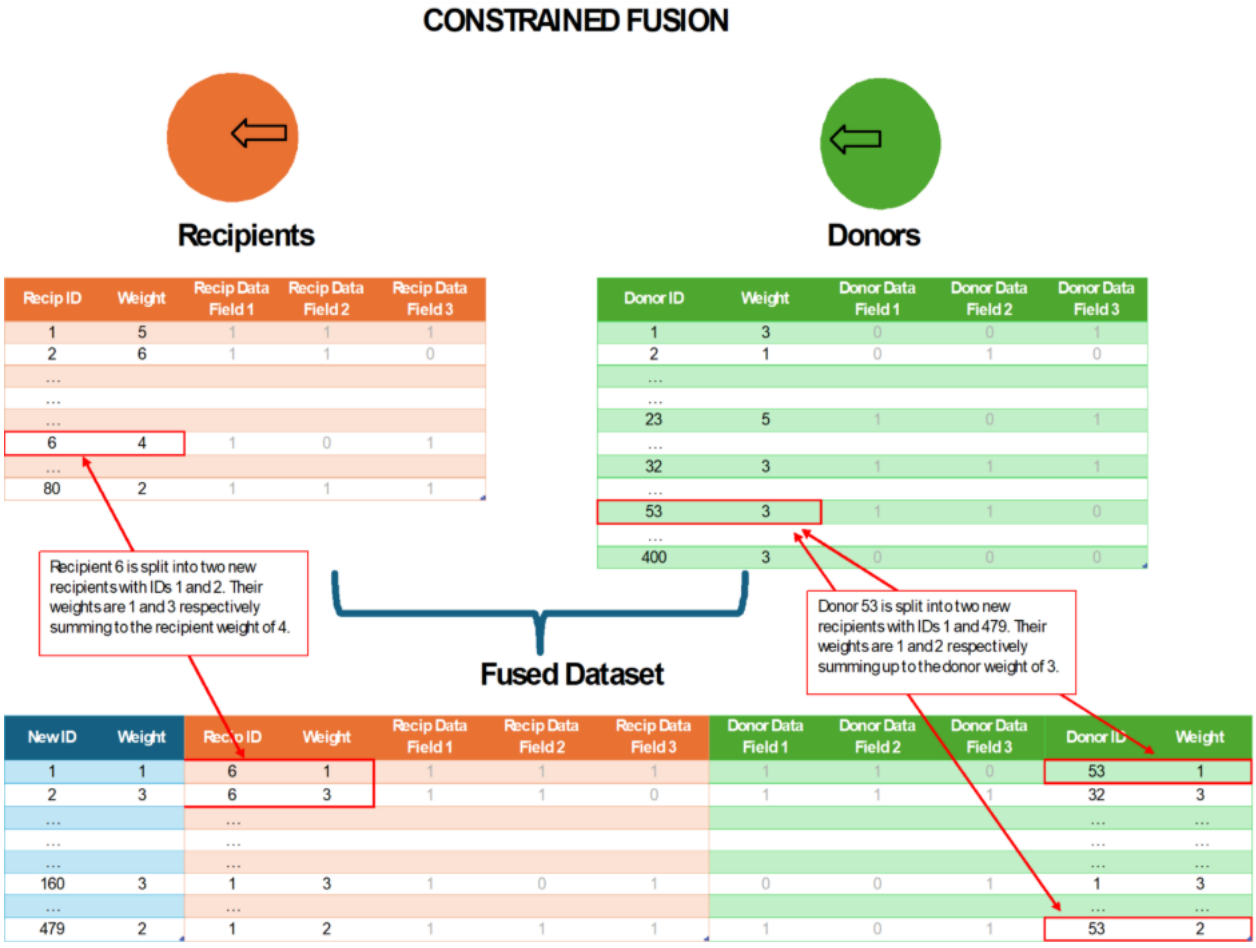
- The total number of records in the fused dataset will be the sum of the combined number of records minus 1.
- The sum of weights for any fragmented donor or recipient will always sum to their original weight. The sum of fragmented weights will always total the sum of weights as no fragments are discarded.

In marketing applications, constrained fusion is particularly valuable for standardised reporting. For instance, when combining data from a digital ad server with a consumer survey, constrained fusion ensures that all data points contribute equally, preserving the integrity of both datasets. This approach is also ideal for maintaining categorical data, such as demographic groupings or geographic regions, where balance and accuracy are paramount.

However, the rigid constraints of this method can reduce precision. Matches may be forced, creating artificial correlations that don't exist in reality. Additionally, constrained fusion is

computationally demanding, which can be a barrier for projects with limited resources. Despite these drawbacks, its ability to preserve key metrics makes it indispensable for projects requiring representational fairness, such as joint audience measurement for TV and digital platforms.

The diagram below illustrates how constrained fusion works.



There are a couple of points to note in particular from the diagram:

- The number of respondents in the fused file is 479 which equates to the sum of the 80 Recipient and 400 Donor respondents minus one.
- The weights of the fragmented respondents sum back to their original weight, so for example Recipient 6 had an original weight of 4; they have been fragmented into two respondents in the fused dataset, one with a weight of 1 and the other with a weight of 3. Similarly, Donor 53 had an original weight of 3; they have been fragmented into two respondents in the fused dataset, one with a weight of 1 and the other with a weight of 2.

Key Characteristics

Equal Representation: Every record from the donor and recipient datasets is incorporated into the fusion process. Adjustments are made to sample weights, splitting records into smaller "fragments" to align proportions.

Balance Across Datasets: Ensures that neither dataset dominates the fusion output, making it suitable for scenarios requiring fairness and accuracy across sources.

Advantages

- Preservation of Metrics: Retains original distributions and weights from both datasets, ensuring key statistics are not distorted.
- Suitability for Complex Data: Ideal for datasets with disproportionate structures or varying levels of granularity.
- Enhanced Credibility: Outputs are more balanced and suitable for stakeholders who require equal emphasis on all integrated datasets.

Limitations

- Reduced Precision in Matching: The strict constraints may force matches that are less optimal, leading to slightly less accurate cross-dataset correlations.
- Higher Computational Demand: The need to balance weights for all records significantly increases processing complexity and time requirements.
- Potential Distortions: While preserving metrics, the algorithm may create artificial relationships between records to meet constraints.

Best Use Cases

- Standardised Reporting: When precise representational balance is required for regulatory or commercial reporting (e.g., joint audience ratings in media).
- Categorical Data: For preserving categorical data where unconstrained is unable to calibrate.
- Currency Maintenance: Essential in scenarios where multiple datasets define overlapping measurement standards.

Detailed Comparison: Unconstrained vs. Constrained Fusion

Unconstrained and constrained fusion each serve different purposes, and their choice depends on the goals of the project. Unconstrained fusion prioritises flexibility and precision, making it ideal for exploratory analyses and granular audience insights. In contrast, constrained fusion focuses on balance and fairness, ensuring that all records contribute equally to the integrated dataset.

For example, a marketer conducting exploratory research into cross-platform consumer behaviours might choose unconstrained fusion to optimise matches for granularity. Conversely, a company preparing a joint TV and digital advertising report would likely opt for constrained fusion to ensure representational balance across datasets.

The table below compares the two approaches in more detail.

Comparison Table

Aspect	Unconstrained Fusion	Constrained Fusion
Primary Goal	Maximising granularity and precision in matching	Balancing and preserving metrics across datasets
Usage of Records	Selective (records may be excluded or reused)	Complete (all records are used exactly once)
Flexibility	High, as donor records can be optimised for matching	Low, as constraints limit matching flexibility
Weight Preservation	Prioritises recipient dataset's structure	Balances weights of both datasets
Suitability for Large Datasets	Moderate; computational demand is lower	High; requires significant computational resources
Granularity	Better for identifying subtle correlations	Lower, as constraints can dilute precision
Applications	Media research, exploratory studies	Regulatory reporting, multi-source integrations

Evaluating Fusion Quality

To ensure a successful data fusion, robust evaluation methods are essential. One critical step is assessing the alignment of records based on fusion hooks, such as demographics or behaviours. Fusion diagnostics can help identify inconsistencies or biases, particularly in unconstrained fusion, where some donor records may be excluded.

Currency preservation tests are another important tool. For example, when fusing TV ratings with digital ad exposure data, it's crucial to verify that audience demographics remain consistent after the integration. Regression-to-the-Mean (RTM) analysis provides additional insight, measuring how fusion impacts correlation strength compared to single-source datasets. Low RTM values indicate effective fusion.

Split sample tests offer a practical way to simulate real-world outcomes. By dividing single-source datasets into subsets and performing fusion, marketers can validate the reliability of the results before applying them to separate datasets.

Practical Recommendations

Choosing the right fusion method depends on the project's objectives:

- Unconstrained fusion is ideal for targeted and strategic analyses, offering the flexibility to uncover niche audience behaviours.
- Constrained fusion is better suited to standardised reporting and regulatory requirements, where equitable representation and preservation is critical.

To ensure reliable results:

- Invest in preparing high-quality datasets, cleaning and/or modelling data where necessary.
- Ensure you are using robust fusion hooks.
- Use advanced evaluation techniques like fusion diagnostics, currency preservation tests, and regression to the mean analysis.

Conclusion

Data fusion is a powerful tool for integrating diverse datasets in the advertising and marketing industries. Whether conducting exploratory research or preparing a report, choosing the right approach—unconstrained or constrained—is key to achieving your objectives. Unconstrained fusion offers flexibility and precision, making it ideal for granular analyses, while constrained fusion provides balance and fairness, ensuring all data sources contribute equally. However it is important to note that the complexity of constrained fusion limits its availability to few practitioners, who might struggle to enhance the process with necessary additions like suitable statistical distance measurements or discriminant analysis.

By understanding the strengths and limitations of each method, marketers can make informed decisions that align their data integration strategies with their analytical goals. Ultimately, effective data fusion enhances the value of integrated datasets, enabling deeper insights and more impactful marketing decisions.

RSMB Fusion

RSMB has over 30 years of experience conducting high-quality fusions using both constrained and unconstrained methods. Our fusion platform is now available in the cloud as a SaaS product, via an API, and as a native Snowflake app, making it easier for organisations to run high-quality fusions. We also manage the whole end-to-end process in-house for companies that prefer to tap into our expertise. There is more information about RSMB Fusion at https://www.rsmb.solutions/fusion-home.