IMPERIAL

Imperial College London
Department of Mathematics

Beyond Single-View Learning: Multi-View Clustering for Social Data

MAX LABARRE

CID: 02428325

Supervised by Dr Marina Evangelou

August 31, 2025

Submitted in partial fulfilment of the requirements for the
MSc in Machine Learning and Data Science at Imperial College London

The work contained in this thesis is my own work unless otherwise stated.

Visual Studio Code with GitHub Copilot (including Agent Mode) was used to assist in writing and debugging code produced for this project. This tool provides real-time suggestions and completions based on my prompts and existing code.

All use of such tools was directed by me, and I take full responsibility for the conceptual design, implementation decisions, analysis, and interpretation of results presented in this work.

Signed: Max Labarre Date: August 31, 2025

Acknowledgements

I would like to express my sincere gratitude to Dr Marina Evangelou, my supervisor, for her invaluable guidance, encouragement, and expertise throughout the course of this project and I am also grateful to Ella Orme (3rd year PhD student under her supervision) whose insightful advice and active participation in our group discussions greatly enriched the work. Their combined expertise in multi-view learning, particularly in unsupervised settings, has been an invaluable resource.

I also wish to thank my project working group: Rafal Chen, Adam Muhtar, Margaret Snape, and Pavlos Tranakidis for their contributions to our collective literature reviews. Several aspects of this report are a direct outcome of the collaborative discussions and shared research efforts within the group.

My thanks extend to the Machine Learning and Data Science programme staff and lecturers at Imperial College London, whose dedication and support have been instrumental during the past two years. In particular, I am grateful to Dr Zak Varty, my personal tutor, for his patience, constructive guidance, and encouragement throughout the programme.

I would also like to acknowledge James McCrae, Ben Ford, and Will Haste at the Office for National Statistics (ONS) for publishing their work on *Clustering local authorities against subnational indicators in England* and making their code repository publicly available, enabling us to replicate and critically review their approach. Although we have not been in direct contact, their openness in sharing both methodology and code has been instrumental in shaping the comparative aspects of this research.

More broadly, I extend my appreciation to the ONS for publishing the datasets that underpin this analysis.

Without the generosity, openness, and contributions of all those acknowledged above, this research would not have been possible.

Abstract

Clustering local authorities using socio-economic indicators has become a cornerstone of policy decision-making in the UK. Yet, current practices, most notably those employed by the Office for National Statistics (ONS), rely heavily on single-view learning or naïve early integration techniques that overlook structural complexity, missingness, and inter-view consensus or disagreement.

This project addresses these limitations by critically evaluating the ONS methodology and introducing a comprehensive multi-view learning (MVL) framework tailored to noisy, incomplete social data. Using the same dataset as the ONS, the study replicates their pipeline, uncovers methodological weaknesses, and demonstrates significant gaps in robustness, interpretability, and replicability.

The core contribution of this work lies in implementing and benchmarking a diverse portfolio of Multi-View Clustering (MVC) strategies spanning early, intermediate, and late integrations. These include similarity-based, matrix factorisation, and neural network approaches designed to handle missing data without imputation. Notably, Multi-View Similarity Fusion and Masked Autoencoders outperform baseline models while maintaining transparency.

The findings highlight the limitations of traditional clustering methods and show that MVL delivers more robust and interpretable segmentations, directly addressing long-standing challenges in Social Sciences around missingness and transparency. Given the strategic importance of the UK's Levelling Up agenda, we hope these advances offer an alternative for informing fair and effective data-driven social policies.

Keywords: Multi-View Learning, Multi-View Clustering, Similarity Fusion, Matrix Factorisation, Masked Autoencoder, Spatial View, Missing Data Handling, Social Science Analytics, Local Authority Segmentation, Policy Decision-Making, ONS Clustering Methodology, Levelling Up Agenda, Interpretability in Clustering, Robustness and Stability.

Attached files: Everything necessary to reproduce this project can be found at: https://github.com/MaxLabarre/imperial_Multi_View_Clustering.git

Contents

Chapter 1. Foundations and Scope 1
1.1 Context and Motivation . . . 1
1.2 Problem Statement . . . 1
1.3 Objectives and Research Questions . . . 2
1.3.1 Objectives of the project . . . 2
1.3.2 Research Questions . . . 2
1.4 Contributions . . . 3
1.5 Structure of the Report . . . 3

Chapter 2. Clustering for Social Research 4
2.1 Literature Review . . . 4
2.1.1 Unsupervised Machine Learning in Social Sciences . . . 4
2.1.2 Disciplinary Applications and Perspectives . . . 5
2.1.3 Common Methods: Clustering and Dimensionality Reduction . . . 5
2.2 Single-View Learning (SVL) in Practice: ONS Clustering Case Study . . . 7
2.2.1 The ONS Dataset . . . 7
2.2.2 The ONS Objectives . . . 9
2.2.3 The ONS Single-View Clustering Methodology . . . 10
2.3 Critical Review of the ONS analysis . . . 12
2.3.1 Evaluation and Results . . . 12
2.3.2 Limitations and Missed Opportunities . . . 12
2.3.3 Toward Multi-View Learning . . . 14

Chapter 3. Multi-View Clustering (MVC) for Social Research 15
3.1 Literature Review . . . 15
3.1.1 Multi-View Data Integration (or Fusion) Strategies . . . 15
3.1.2 Common Methods in Multi-View Clustering . . . 16
3.1.3 Motivations for Multi-View Clustering . . . 17
3.2 Application of Multi-View Clustering (MVC) to ONS Dataset . . . 18
3.2.1 Experimental Framework and Methodology . . . 18
3.2.2 Evaluation and Diagnostics Framework . . . 29

Chapter 4. Evaluating MVC: Results, Diagnostics, and Benefits 31
4.1 Experimental Results . . . 31
4.1.1 Comparative Analysis of Methods . . . 31

4.1.2 Further exploration . . . 32
4.2 Evaluating Multi-View Similarity Fusion (MVSF) . . . 32
4.2.1 Insights . . . 32
4.2.2 Diagnostics . . . 34
4.2.3 Interpretability . . . 35
4.3 Summary of MVC Benefits and Findings . . . 36

Chapter 5. Discussions & Conclusions **37**
5.1 Discussions . . . 37
5.1.1 Summary of Key Insights . . . 37
5.1.2 Limitations of Current Work . . . 38
5.1.3 Ethical and Policy Considerations . . . 38
5.2 Conclusion & Future Work . . . 39
5.2.1 Recap of Research Contributions . . . 39
5.2.2 Implications for Local Authority Segmentation . . . 39
5.2.3 Future Research Directions . . . 40

Supplementary Matter **1**
A Term Definitions and Glossary . . . 2
B Reproducibility Statement . . . 4
C Exploratory Data Analysis Results and Visualisations . . . 9
D Detail of Experimental Methods . . . 14
E Experimental Results, Diagnostics and Visualisations . . . 27
F Supplementary Experiments and Ablations . . . 40

This report was produced under a strict page limit for the main body of 40 pages.

Chapter 1. Foundations and Scope

1.1. Context and Motivation

The application of Machine Learning (ML) techniques to social data has grown rapidly in recent years, driven by the increasing availability of large volumes of rich data. These datasets offer unprecedented opportunities to understand and support social intervention and public policy. However, they also present unique challenges: social data often reflects complex, overlapping societal processes and is affected by structural biases, data missingness or sparsity, and variability across time and space.

Segmentation, the task of grouping similar entities (e.g., individuals, regions, or communities) based on observed features, is a very common use-case in Social Sciences that allows researchers and decision-makers get insights from unstructured data.

Despite this, segmentation exercises in practice often rely on simplistic methodologies. A common approach is to apply standard clustering algorithms on a large dataset, often a concatenation of multiple datasets, where features from different domains are merged into a single vector space. This naive integration strategy may ignore the unique structure and relevance of each data view, for example, by treating them all equally while not accounting for varying missingness among them. One such example is the clustering methodology employed by the UK's Office for National Statistics (2023a), which applied single-view clustering to segment Local Authorities in England.

These practices often lead to poor and unreliable results. Sensitivity to random initialisation, preprocessing decisions, and missing data are rarely addressed systematically. The resulting segmentation may be difficult to reproduce or interpret and risk oversimplifying complex social phenomena. This highlights a critical methodological gap: the need for approaches that can account for multi-view structure, handle missingness, and generate more interpretable outcomes.

1.2. Problem Statement

Segmentation of local authorities is a critical tool for understanding regional disparities and informing place-based policy, particularly in the context of the UK government's Levelling Up agenda (UK Government, 2022), which aims to reduce regional inequality and promote equitable growth across England. However, current analytical approaches, such as those adopted by the Office for National Statistics (2023b), typically rely on single-view clustering or naive feature concatenation across multiple views (e.g. economy, education, health). These methods do not fully exploit the multi-faceted nature of social data and fail to account for view-specific structure, redundancy, and potential disagreement across views. Social data is often noisy, incomplete, and biased, with each view contributing distinct but partial information. Ignoring

these characteristics can result in segmentations that are difficult to interpret and reproduce. There is a need for methodologies that can integrate diverse views while preserving their individual contributions to reveal meaningful groupings.

Multi-View Learning (MVL) refers to machine learning from data that provide multiple "views" or feature sets for the same underlying entities. These views could be different modalities (e.g. images, text, audio), different sensors, or heterogeneous data sources describing the same samples. Integrating information from multiple views can leverage the complementary nature of the data to improve learning performance and insights.

1.3. Objectives and Research Questions

1.3.1. Objectives of the project

This research addresses the central problem of how MVL methods can improve upon traditional single-view or naive integration approaches for unsupervised segmentation in social datasets, using local authority-level data in England as a case study.

We critically evaluate existing single-view and naive integration approaches for social segmentation. We then develop and benchmark MVL methods that handle noisy, incomplete data while assessing robustness and interpretability.

1.3.2. Research Questions

1. Baseline Comparisons: How do current single-view and naive integration clustering methods perform on real-world social datasets in terms of performance against common metrics, stability, and interpretability?
2. Multi-View Effectiveness: Can MVL approaches provide more stable and meaningful segmentations of local authorities compared to traditional methods?
3. Handling Data Quality Issues: How do different methods handle noise and missingness, and how do these factors affect the quality of the resulting clusters?
4. Enhancing Interpretability: To what extent can MVL approaches enhance interpretability by enabling the assessment of individual view contributions to the clustering outcomes?

1.4. Contributions

This research makes several contributions to the study of unsupervised learning and clustering in social data contexts:

- Critical Evaluation of Single-View Learning (SVL) Approaches: We benchmark and stress-test commonly used SVL methods, specifically those adopted by the ONS, for clustering local authorities, revealing their limitations in robustness, interpretability, and sensitivity to change in modelling parameters.
- Introduction of a MVL Framework: We propose a framework tailored to social data. This includes the application of integration strategies that preserve both shared and view-specific structure, handle missing data, and provide insights into the contributions of views and features to the segmentation results.
- First Application to Policy-Related Segmentation: To our knowledge, this is the first application of MVL methods to the segmentation of English local authorities using multi-view indicators, offering a more robust, nuanced alternative to the independently segmented or naively concatenated models currently used in official statistics and policy analysis.

1.5. Structure of the Report

Following this introductory chapter, this report is structured with the following chapters: Chapter 2 reviews existing Single-View clustering approaches, including a real-world case study of the Office for National Statistics segmentation of local authorities in England. It establishes a baseline for comparison with Multi-View methods; Chapter 3 is a review of existing Multi-View clustering approaches and their motivations. We then present a suite of MVC methods and the advantages we hope to gain from their application. Chapter 4 presents and discusses results gained from the application of MVC methods presented in chapter 3. It investigates how these applications affect clustering quality, interpretability, and alignment with policy-relevant geography. Finally, Chapter 5 concludes our report by synthesising the findings, reflecting on methodological implications, and outlining future directions for research and application in policy-oriented segmentation and MVL.

Each chapter contributes to advance a more rigorous, robust, and interpretable framework for unsupervised segmentation of complex social data.

Chapter 2. Clustering for Social Research

2.1. Literature Review

2.1.1. Unsupervised Machine Learning in Social Sciences

Unsupervised Machine Learning (UML), a set of algorithms designed to uncover hidden patterns or structures in data without the use of predefined labels or outcomes (Hastie et al., 2009), has always played a pivotal role in Social Sciences, where researchers often work with large, complex, and imperfect datasets. UML methods are well-suited to tackle several persistent challenges in the analysis of social data:

- Lack of ground truth labels: Social phenomena like political ideology, cultural norms, social class, or mental health profiles often have no objective or agreed-upon classification. UML enables researchers to uncover latent structures without requiring labeled data (Waggoner, 2020).
- High dimensionality and multivariate data: Surveys, administrative records, and social media data often contain dozens or hundreds of variables per individual or case. Dimensionality reduction techniques help simplify these datasets, making them more interpretable while retaining meaningful variance (Wasserman and Faust, 1994).
- High variability and heterogeneity: Social populations are inherently diverse, with overlapping or fuzzy boundaries. Clustering methods allow analysts to identify subgroups or profiles in data without assuming linear relationships or fixed categories (Zakharov, 2016).
- Missingness and sparsity: The collection of social data often presents challenges, respondents may skip questions in surveys or systems may produce errors and incomplete information. UML techniques, especially techniques that tolerate missing values allow partial data to still contribute meaningfully to pattern discovery (Sinha et al., 2021).
- Imperfect measurements and biases: Social data are often influenced by self-reporting biases, measurement error, and cultural framing. UML is less reliant on strict model assumptions and can be used to explore data structure without overfitting to noise or requiring fully reliable input (Lundberg et al., 2022).

Overall, UML provides a flexible, data-driven approach that is compatible with the exploratory, theory-building orientation of much social science research. Rather than specifying hypotheses in advance, researchers can use UML to discover patterns that may warrant further investigation, validation, or theorisation.

2.1.2. Disciplinary Applications and Perspectives

- Political Science: While traditionally reliant on supervised and deductive approaches, political science is increasingly leveraging UML to explore latent ideologies, regime types, and opinion clusters (Waggoner, 2020) or to studying relational structures in sociopolitical systems (Wasserman and Faust, 1994).
- Sociology: Sociologists have long used unsupervised methods to derive typologies of individuals, neighborhoods, and cultures. Factor analysis and clustering are standard tools for mapping social stratification, cultural values, and lifestyles. For example, dimensionality reduction can uncover global cultural patterns (Inglehart and Welzel, 2005) and latent representations are frequently used in survey analysis to identify latent social classes or belief systems (Sinha et al., 2021).
- Economics: Economists often have to deal with high-dimensional data to which they need to apply reduction to construct indices (e.g., economic freedom, financial development), while clustering techniques are used to segment markets or consumers and help classify macroeconomic regimes when clear labels are unavailable (Rezankova, 2014).
- Anthropology and Archaeology: Clustering and dimensionalty reduction are used in anthropology to identify cultural patterns across societies and designate taxonomies of artifacts, skeletal remains, or linguistic features (Lundberg et al., 2022).
- Psychology: Psychology has a long tradition of using unsupervised methods, particularly clustering and exploratory factor analysis, to classify individuals based on cognitive, behavioral, or affective traits and ultimately develop conceptual typologies of psychologies (Zakharov, 2016).

2.1.3. Common Methods: Clustering and Dimensionality Reduction

Clustering Techniques.

- K-means: A centroid-based clustering algorithm that partitions observations into k clusters by minimising within-cluster variance; introduced by MacQueen (1967). It is widely used in psychology and sociology to uncover latent profiles (Zakharov, 2016).
- Hierarchical clustering: Builds a nested tree (dendrogram) of clusters via agglomerative or divisive strategies; originally formalized by Johnson (1967). It is commonly used in sociology for typology construction and cross-national comparisons (Rezankova, 2014).
- Gaussian Mixture Models (GMM) A probabilistic model assuming data are generated from a mixture of Gaussian distributions; introduced in its modern Expectation-Maximization (EM) form by Dempster et al. (1977). GMM has been used in psychology and education to model latent ability groups (Sinha et al., 2021).

- Density-Based Spatial Clustering of Applications with Noise (DBSCAN): A density-based algorithm that forms clusters based on the density of points and can discover arbitrarily shaped clusters; introduced by Ester et al. (1996). It has been used in spatial analysis and housing data (Waggoner, 2020).

- Spectral clustering: A graph-based method that partitions data by analyzing the eigenstructure of a similarity or Laplacian matrix derived from a graph representation of the data. Its early theoretical formulation was introduced by Donath and Hoffman (1972), it underpins many community detection algorithms used in social network analysis (Wasserman and Faust, 1994).

Dimensionality Reduction.

- Principal Component Analysis (PCA): A linear technique that projects data onto orthogonal components explaining maximal variance; first introduced by Pearson (1901). It is extensively used in political science and sociology for index construction and attitudinal scaling (Inglehart and Welzel, 2005).

- t-distributed Stochastic Neighbor Embedding (t-SNE): A nonlinear technique for visualizing high-dimensional data by preserving local structure in 2D or 3D space; proposed by Van der Maaten and Hinton (2008). It has been used in cultural analytics and visualizing clusters of survey responses (Lundberg et al., 2022).

- Uniform Manifold Approximation and Projection (UMAP): A nonlinear dimensionality reduction method that preserves both local and global data structure; introduced by McInnes et al. (2018). Applications include clustering patterns in political behavior, mobility data, and ethnographic coding (Waggoner, 2020).

- Topic Modeling (e.g., Latent Dirichlet Allocation): A generative probabilistic model for uncovering topics in large text corpora; developed by Blei et al. (2003). LDA has transformed political science, media studies, and policy analysis by enabling scalable content analysis (Grimmer and Stewart, 2013).

- Non-negative Matrix Factorization (NMF): A parts-based factorization approach producing additive, interpretable components; first formalized by Lee and Seung (1999). NMF has been applied in recommender systems, cultural analysis, and user behavior segmentation on social platforms (Zhang and Peng, 2022).

Unsupervised machine learning, particularly clustering and dimensionality reduction, offers powerful tools for exploratory, inductive research across the social sciences. These methods help reveal latent typologies, social structures, and dimensions underlying complex, high-dimensional datasets. Their adaptability makes them attractive across disciplines and their capacity to handle unlabeled data aligns with the data realities of many social phenomena.

However, the interpretive burden of UML remains high. Researchers must be vigilant about overfitting, misclassification, and the sociological implications of the categories they derive.

Methodological rigor, theoretical grounding, transparency and validation strategies are essential to ensure that the clusters or dimensions discovered are meaningful and actionable.

Methodological Considerations.

- Interpretability: UML outputs must be interpreted in light of theory. The same dataset can produce very different clusterings depending on preprocessing, distance metric, or algorithm used (Lundberg et al., 2022).
- Cluster validation: Without ground truth, researchers use performance metrics such as silhouette scores, Bayesian information criterion (BIC), bootstrapping, and external validity checks to ensure robustness. But these do not guarantee that the identified clusters reflect meaningful, generalisable, or view-relevant structures in the underlying data.(Sinha et al., 2021; Zakharov, 2016).
- Preprocessing: Variable scaling, handling of missingness, and encoding of categorical data significantly affect outcomes. Mixed data types require careful processing and algorithmic parameters, careful selection (Rezankova, 2014).
- Scalability and complexity: High-dimensional or large-scale data may require deep embeddings, mini-batch clustering, or dimensionality reduction as a preprocessing step (Zhang and Peng, 2022).
- Bias and fairness: Unsupervised methods can reinforce existing social biases if not carefully interpreted. Ethical interpretation is essential, especially when clustering individuals or communities (Lundberg et al., 2022).

2.2. Single-View Learning (SVL) in Practice: ONS Clustering Case Study

To illustrate the limitations of Single-View Clustering, the following section provides a critical review of a real-world explample of single-view clustering approach applied to social data. We examine the assumptions, preprocessing decisions, and validation strategy employed. This review sets the stage for a methodological rethinking grounded in multi-view learning frameworks.

2.2.1. The ONS Dataset

This work uses the same dataset as the Office for National Statistics in their Clustering local authorities against subnational indicators analysis (Office for National Statistics, 2023b). Specifically, it comprises a curated selection of indicators (economic, connectivity, education, etc.) aggregated at the local authority level in England between 2019 and 2021. The datasets (views) include view-specific metrics derived from administrative and survey sources such as

the Annual Population Survey (APS), Public Health England, and other subnational statistical releases. In their white paper (UK Government, 2022), the UK government outlines 12 Leveling Up Missions, several of which align very closely with the views used by the ONS in its analysis as shown in Table 2.1.

Table 2.1.: Linkage between ONS Views and the Government's Leveling Up Initiative

Leveling Up	Description	View	# Local Authorities	# Social Indicators
Mission 1	Pay, employment and productivity	Economic	309	4
Mission 3	Local public transport connectivity	Connectivity	340	5
Mission 5	Primary school attainment	Educational Attainment	309	9
Mission 6	Skills training	Skills	309	4
Mission 7	Health and life expectancy	Health	309	8
Mission 8	Well-being	Wellbeing	309	4

The full dataset is available on the website for the Office for National Statistics (2022).

The issue of missing data. However, Table 2.2 highlights the substantial fragmentation and missingness of the dataset comprising the 309 common Local Authorities (LAs) in England.

- 205 Local Authorities appear in all six views, forming the joint intersection across them. Only 81 of which have no missing data across all views.
- 104 appear in only a subset of views.
- 24 were explicitly excluded by the ONS due to boundary changes or data reliability concerns, common due to boundary changes, non-reporting, or insufficient sample sizes (e.g., Buckinghamshire, West and North Northamptonshire).

Table 2.2.: Missingness and row retention by view after preprocessing

View	% Rows Missing Data	% Rows Excluded (ONS)	% Rows Retained
Economic	0.00%	0.65%	99.35%
Connectivity	13.24%	2.35%	86.76%
Educational attainment	**59.87%**	1.94%	**39.48%**
Skills	0.00%	0.97%	99.35%
Health	**60.94%**	5.63%	**26.88%**
Wellbeing	1.29%	0.65%	98.71%

Notes. Missingness reflects the proportion of rows containing at least one `NA` across the view. Retention reflects the percentage of local authorities retained after preprocessing (e.g., excluding suppressed or incomplete entries). **Notable values highlighted in bold**.

The Health and Educational Attainment views exhibit the highest attrition after preprocessing, retaining just 86 and 122 rows respectively, which means their cleaned versions only cover 39% and 27% of LAs in England. The Connectivity view also contains a significant number of missing rows. The ONS acknowledges those patterns are not random: data availability is systematically lower in certain geographies (e.g., rural or newly merged authorities), so resulting clusters may reflect reporting artifacts rather than meaningful socio-economic structures. This challenges direct interpretability of single-view clusters and their comparability across views.

The risks of imputation. The ONS indicates in the latest update of its methodology that data imputation was applied during preprocessing. They imputed lower tier data by setting it to be the same as the upper tier data for all missing lower tier local authorities within an upper tier local authority (Office for National Statistics, 2024a). We imagine this decision was driven by their choice of methodology which requires complete data (detailed later in subsection 2.2.3) but imputing data when over 70% of Local Authorities have missing data in some views, or when two out of six views have missing data in 60% of their rows, it poses serious risks. Such risks include introducing bias, inflating correlations, and smoothing away meaningful variation, especially if missingness is not at random.

2.2.2. The ONS Objectives

The ONS segmentation objectives appear to be two-fold:

1. **View-specific models** aiming to uncover distinct patterns specific to each thematic view (e.g., economy, connectivity, health), allowing policymakers to identify strengths and weaknesses in isolation.

2. A **Headline model**, by contrast, seeks to provide a holistic, cross-view segmentation that synthesises patterns across views to support high-level strategic planning and place-based policy interventions.

While this dual approach is conceptually valid, balancing granularity with integrative insight, the execution raises concerns. A critical issue with the Headline model is the arbitrary selection of "one headline metric from each theme where available [by consulting] with cross-government topic experts on which metric was most appropriate to include" (Office for National Statistics, 2023b), rather than leveraging the full breadth of available data across all views. This risks introducing subjective bias (even if coming from experts): what counts as "headline" or representative is not methodologically justified and may exclude features that carry key signals for clustering. Consequently, the Headline segmentation may reflect the designers' preconceptions more than the true multivariate structure of differences between Local Authorities.

The resulting clusters were labeled with holistic descriptions. For the Headline model, the labels are from both the quantitative results gained from the unified view and qualitative adjustments informed by the single views' results. Thus, the Headline model is essentially a hybrid of single-view and early integration clustering over concatenated features, not a true fusion of separate views. Table 2.3 shows a subset of the ONS clustering labels across views (Office for National Statistics, 2023c).

Each of the five thematic models (Economic, Connectivity, etc.) was based on a single view of indicators. As touched on earlier, the Headline model is aimed to integrate across the five thematic views. But more on that later (see subsection 2.2.3).

Table 2.3.: The ONS qualitative cluster labels for a subset of Local Authorities across views

Local Authority	Headline	Economic	Connectivity	Educational Attainment	Skills	Health	Well-being
Hartlepool	High connectivity, low health and well-being	Below median on all economic metrics	Slightly below median on all connectivity metrics	Higher KS2, lower GCSE performance	Above median apprenticeships and further education, below median Level 3+ qualifications	Far below median on all health metrics	Slightly above median on all well-being metrics
Warrington	High health and productivity, low well-being	Above median employment rate, below median productivity	Far above median on all connectivity metrics	Far higher against most education metrics	Slightly below median on all skills metrics	Broadly median on all health metrics	Better than median anxiety, worse than median on all other metrics
North Lincolnshire	High health and well-being, moderate educational performance	Above median employment rate, below median productivity	Below median on all connectivity metrics	Higher KS2, lower GCSE performance	Above median apprenticeships and further education, below median Level 3+ qualifications	Far below median on all health metrics	Far above median on all well-being metrics

Notes. KS2: Key Stage 2 scores. GCSE: General Certificate of Secondary Education. "Post-16" includes apprenticeships and further education. "Level 3+" refers to qualifications equivalent to A-levels or higher.

2.2.3. The ONS Single-View Clustering Methodology

Pipeline. The ONS applies a single-view clustering approach to segment Local Authorities (LAs). Their pipeline includes preprocessing of variables which are then clustered independently after dimensionality reduction. We replicate this pipeline, simplified in Figure 2.1, in our implementation (`ONS_implementation.py`), which is further detailed in Appendix D1.

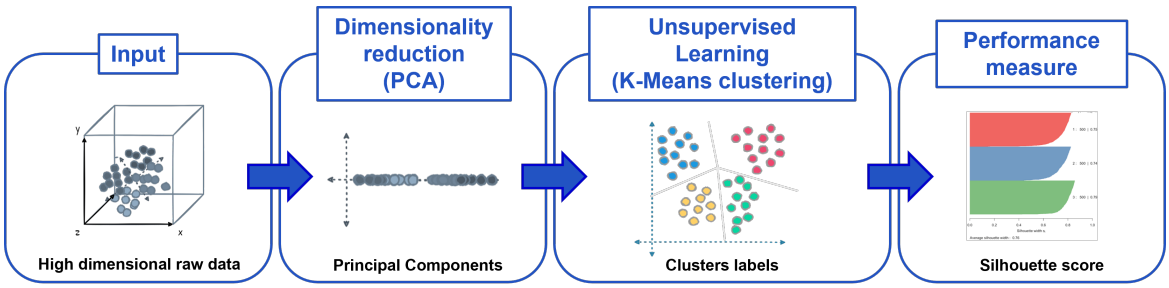

Figure 2.1.: Simplified Conceptualisation of the ONS Clustering Pipeline

The goals of this pipeline are:

- Compress high-dimensional data into a smaller space while preserving structure.
- Discover natural groupings in the reduced space.
- Evaluate how well-separated and cohesive the clusters are.

Let the data matrix (view) be $\mathbf{X} \in \mathbb{R}^{n \times d}$ where n is the number of samples and d is the number of features. Assuming $\mathbf{X}$ has been centered (zero mean) and optionally standardized.

PCA (Principal Component Analysis). It reduces dimensionality by projecting onto directions of maximum variance (Jolliffe and Cadima, 2016).

We compute covariance matrix $\Sigma = \frac{1}{n}\mathbf{X}^T\mathbf{X}$

We solve eigenvalue problem: $\Sigma\mathbf{W} = \mathbf{W}\Lambda$ where $\mathbf{W}$ contains eigenvectors and Λ the eigenvalues.

We keep the top p components $\mathbf{Z} = \mathbf{X}\mathbf{W}_{[:,1:p]} \in \mathbb{R}^{n\times p}$. This retains directions with most variance $\text{argmax}_{\|\mathbf{w}\|=1}\text{Var}(\mathbf{X}\mathbf{w})$ to explain at least 25% of the variance.

K-Means Clustering. Partitions the projected data $\mathbf{Z}$ into k clusters by minimizing within-cluster variance $\min_{\{C_i\}_{i=1}^k} \sum_{i=1}^k \sum_{\mathbf{z}_j \in C_i} \|\mathbf{z}_j - \boldsymbol{\mu}_i\|^2$ where C_i: set of points in cluster i and $\boldsymbol{\mu}_i$: centroid of cluster i

The algorithm alternates assigning points to nearest centroid and recomputing centroids as means of assigned points (MacQueen, 1967).

Silhouette Score. Evaluates clustering quality using cohesion vs separation:

For each sample i, we have $a(i)$: mean intra-cluster distance and $b(i)$: mean nearest-cluster distance.

Then $s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \in [-1, 1]$

The overall silhouette score is the mean over all points (Rousseeuw, 1987).

The same process is applied to the concatenated subset of features selected from the various views for their Headline model.

Motivation. We suspect the ONS deliberate choice of PCA, K-means clustering, and Silhouette Score comes from their explainability, advantageous in domains of high public scrutiny. PCA reveals linear relationships between variables and condenses them into interpretable components, while K-means forms intuitive clusters based on proximity to centroids, an idea that is straightforward to communicate to non-technical stakeholders. This contrasts with more complex non-linear or distance-based methods, which can obscure interpretation and complicate explanations. Silhouette Score was probably preferred over alternatives like the indices from Caliński and Harabasz (1974) or Davies and Bouldin (1979) because it provides a balance of within-cluster cohesion and between-cluster separation, without relying on abstract ratios or assumptions about cluster shape.

2.3. Critical Review of the ONS analysis

2.3.1. Evaluation and Results

Replication. We found challenges in our attempt to replicate the ONS' local authority clustering methodology across the six thematic views as well as the composite headline model, using the publicly available methodological documentation and indicator datasets. To evaluate the fidelity of our reproduction, we compared the claimed silhouette scores obtained against the best we could get from the replicated methodology. We also compared the resulting cluster assignments against the official ONS cluster labels using two label-invariant measures: Adjusted Rand Index (Hubert and Arabie, 1985) and Normalised Mutual Information (Vinh et al., 2010). These metrics assess structural agreement between clustering, accounting for arbitrary label permutations. Table 2.4 shows a comparison between our results and those claimed by the ONS:

Table 2.4.: Comparison of ONS claimed and replicated Clustering results across views

View	n	K (ONS)	Silh. (ONS)	Silh. (Rep)	Silh. (Grid)	Δ Silh.	ARI	NMI
Economic	307	4	**0.62**	0.37	0.44	-0.25	**0.781**	**0.759**
Connectivity	304	4	0.53	0.50	0.51	-0.1	0.530	0.585
Educational attainment	303	8	0.21	**0.38**	0.43	0.17	0.413	0.597
Skills	307	4	**0.43**	0.37	0.41	-0.06	0.560	0.567
Health	291	4	0.28	**0.39**	0.42	0.11	0.546	0.565
Wellbeing	307	4	0.25	**0.32**	0.43	0.08	0.578	0.611
Headline model	309	4	**0.48**	0.38	—	-0.1	**0.620**	**0.702**

Notes. n: number of local authorities in preprocesssed view. **K (ONS)**: number of clusters claimed by ONS. **Silh.**: silhouette score. **Rep**: replicated with fixed k. **Grid**: silhouette score after grid search over k. **Δ Silh.**: difference between claimed and replicated. **ARI**: Adjusted Rand Index. **NMI**: Normalised Mutual Information. Missing grid search values are marked as "—". **Notable values highlighted in bold**.

A detailed description of the analysis and results can be found in Appendix E1. These results reveal a significant misalignment between the ONS's claimed performance and what we were able to replicate. Where our silhouette scores match the ONS-claimed values (e.g., Connectivity), ARI/NMI remain modest, suggesting label misalignment despite similar cluster cohesion. Ironically, when similarity with ONS labels is high (e.g., Economic or Headline model), our silhouette scores are nowhere near those claimed by the ONS. There is also a big question mark around some ONS labels we have perused, for example St Albans (E07000240), a Local Authority with all five connectivity variables (incl. three imputed) higher than their respective medians, is somehow described as being part of the "Slightly below median on all connectivity metrics" cluster. In conclusion, our replication efforts have called into question the transparency adn replicability of the ONS clustering pipeline or the data used.

2.3.2. Limitations and Missed Opportunities

Regardless of whether we can reproduce the ONS results, it is important to note that the silhouette coefficients they report are very modest to say the least, with maybe the exception of the Economic view achieving a silhouette score of 0.62, which can be considered as evidence of

good structural separation and little overlap (Rousseeuw, 1987). Most other views are closer to 0.50 or substantially lower, indicating only moderate cluster separation. Of particular interest to our study, the ONS Headline model attains a silhouette score of just 0.48, which falls into a range generally interpreted as reflecting weak structure, where clusters may overlap and separation is less distinct. Such values, while not uncommon in high-dimensional and noisy social datasets, highlight that interpretability and robustness may be as critical as raw numerical scores when evaluating segmentation quality. From our replication, the silhouette scores are even lower, painting a grimmer picture of cluster separation and overall model quality.

The ONS identifies several limitations in their clustering analysis (Office for National Statistics, 2023b), which are common challenges in Social Sciences (already outlined in subsection 2.1.3):

1. Sensitivity to Initialisation: As with many iterative algorithms, K-means is sensitive to its random starting points.
2. Model Instability: The clustering outcomes can shift significantly with small changes to the data or preprocessing steps. We suspect, this is the source of our challenges in replication.
3. Data Missingness: K-means requires complete data. As such, local authorities with missing data (often due to recent boundary changes) had to be excluded from clustering. This is a practical compromise to maintain algorithmic requirements.
4. Limited Interpretability: While the clustering reveals similarities between local areas, the method does not directly indicate which variables drive these groupings.

These reflect the trade-offs of clustering within the constraints of official statistics, balancing technical rigour with clarity for public and policy audiences.

Mitigations. For 1 & 2, the ONS' choice of combining PCA with K-means happened to be an astute one. In an experiment detailed in Appendix E2, we tested three dimensionality reduction techniques across four clustering algorithms, from the most common methods used in Social Sciences (see subsection 2.1.3). We first conducted a grid search over different values of components and clusters to identify optimal configurations based on silhouette scores. We then evaluated the stability of each method by running 100 iterations with varying random seeds, capturing both the distribution of silhouette scores and the consistency of cluster assignments (again, using ARI and NMI). Across all metrics, PCA combined with K-means consistently outperformed other combinations (aside from maybe Agglomerative Clustering), demonstrating both high clustering quality and stability, reinforcing its suitability for transparent, interpretable analysis of social indicators.

For 3, as outlined in subsection 2.2.1, missingness is handled through imputation which carries significant risks. This will be directly addressed in later sections of this report.

For 4, aside from acknowledging the limitation and presenting the clusters as descriptive (Table 2.3) to align with the intended use of the output, the ONS does not address the issue of interpretability. This will be directly addressed in later sections of this report.

2.3.3. Toward Multi-View Learning

Aside from the common methodological limitations, a particularly underexamined and pressing issue, highlighted in the ONS case study, is the dominance of SVL approaches in Social Sciences.

Although general MVL surveys (Yan et al. (2021), Fang et al. (2023) or Yu et al. (2024)) chart widespread adoption of MVL across domains like Computer Vision, Natural Language Processing, and Bioinformatics, there is no evidence from the Social Research literature that broader MVL paradigms have been broadly applied to Social data.

Cruickshank and Carley (2020) emphasises that prior to their study (one of the first real-world application of Multi-View Clustering on social-data), most social media analyses employed Single-View Clustering despite the inherently Multi-View or Multi-Modal nature of most social datasets, where individuals, institutions, or places are characterised by multiple sources of information (e.g., demographic traits, behavioral logs, attitudinal surveys, social networks). Analyses are still typically conducted on a single flattened table, such as that of the ONS.

What we presented so far illustrates both the known methodological considerations (subsection 2.1.3) and common practical limitations (subsection 2.3.2) of single-view clustering when applied to complex social datasets which can be summarised by:

- Structural Oversimplification: By treating each view equally or by selecting a subset of representative features, the approach disregards the interdependencies and complementary signals across views.
- Data Incompleteness: The scale of view-specific missingness of data challenges the integrity of single-view models and questions the validity of imputation.
- Subjectivity in Integration: The Headline model relies on expert-chosen indicators rather than principled fusion of full view information, introducing the risk of bias, and relying on the assumption that each selected view or indicator is representative of broader socio-economic patterns.
- Interpretability: Although the method produces descriptive clusters, it lacks mechanisms to assess the contribution of individual views or indicators to the segmentation output.

These challenges are not unique to the ONS case, they are emblematic of broader limitations in traditional clustering approaches when applied to real-world social data. Such data is inherently multi-view: indicators emerge from distinct domains, collected through separate instruments, each capturing partial and sometimes conflicting perspectives on underlying societal structures. This makes social data particularly well-suited to multi-view learning.

Chapter 3. Multi-View Clustering (MVC) for Social Research

3.1. Literature Review

3.1.1. Multi-View Data Integration (or Fusion) Strategies

In Multi-View, a crucial question is how to fuse or integrate the views. Broadly, integration strategies are categorised by when the fusion occurs in the Machine Learning pipeline. The three commonly discussed approaches are:

- **Early Integration (or Feature-Level Fusion)** combines all feature sets before learning, typically via concatenation. It treats MVL as standard SVL on a unified feature space. Probably the most common because of its simplicity and how easy it is to implement, it enables standard algorithms to leverage all data at once and can capture cross-view interactions if model is expressive enough (Xu et al., 2013). However, it is sensitive to dimensionality and feature imbalance if views differ in scale or size, it is sensitive to missing data and does not model inter-view dependencies explicitly as it treats all views or features equally unless they are specifically weighted (Gaw et al., 2021; Li et al., 2016).

- **Intermediate Integration (or Latent/Hybrid Fusion)** combines views during learning, typically via a shared latent space. It balances view-specific and shared representations and captures cross-view dependencies effectively (Li et al., 2016). Intermediate fusion also excels when views are complementary but heterogeneous or of different modalities (e.g., text and images) (Rappoport and Shamir, 2018). However, it has higher complexity than the other strategies and needs careful parameter tuning (Gao et al., 2020). Most modern models (deep multimodal networks, variational fusion methods, etc.) use this strategy (Baltrusaitis et al., 2018).

- **Late Integration (or Decision-Level Fusion)** trains separate models per view and combines decisions (e.g., majority voting or averaging) similar to ensemble learning. This strategy is less common and excels when data views are too different to merge early. It simplifies preprocessing, views remain independent and it is robust to missing views. However, it cannot model inter-view interactions at feature level and the fusion step may be overly simplistic. It is generally less efficient at leveraging view synergy but is useful in systems with asynchronous or missing views (Liu et al., 2018). It is often used in healthcare and recommender systems where expert models exist for each view. However, it is usually outperformed by intermediate fusion when data volume allows joint training (Gaw et al., 2021; Zhao et al., 2024).

Figure 3.1 shows how those strategies differ conceptually, compared to when no integration is employed (Single-View Learning). The choice of strategy is typically driven by the trade-offs researchers and practitioners are willing to make in their work: Early integration is fast but naive, Late integration is modular but limited, Intermediate integration is expressive but complex. Recent literature trends favor intermediate and hybrid fusion models, especially deep models with shared bottlenecks, attention-based weighting, and modality-specific encoders (Li and Tang, 2024; Zhao et al., 2024).

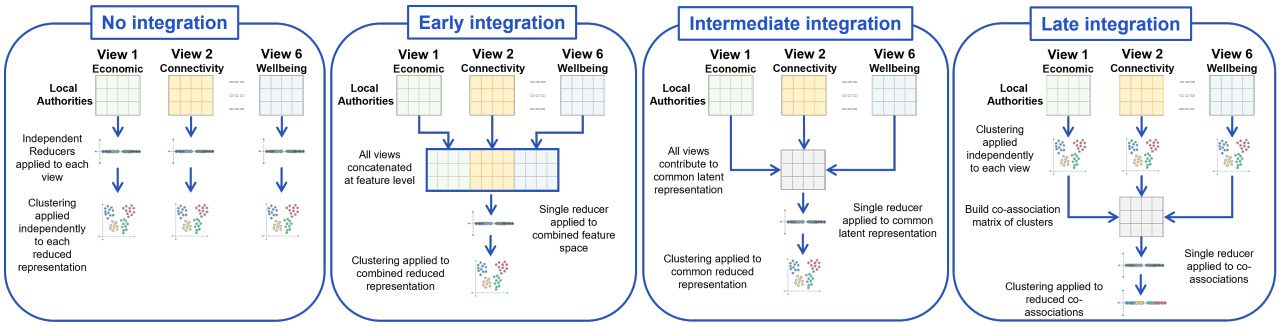

Figure 3.1.: Simplified Conceptualisations of different Data Integration Strategies applied to the Clustering pipeline described in Figure 2.1

3.1.2. Common Methods in Multi-View Clustering

Multi-view clustering seeks to exploit complementary information from multiple views to improve cluster quality. Many common techniques are covered in the literature such as:

- Co-training which iteratively transfers knowledge between views to enhance clustering (Kumar and Daumé III, 2011).
- Similarity fusion which merges view-specific similarity matrices into one unified matrix (Wang et al., 2014b).
- Manifold learning which learns low-dimensional embeddings capturing shared structure (Rodosthenous et al., 2024).
- Ensemble methods which aggregates clustering results from each view (Liu et al., 2013).
- Deep learning for non-linear joint representations (Chowdhury et al., 2025).

Many of the techniques we have enountered in the literature share underlying methodological principles, which we group into four broad families.

Similarity-based Methods. Where the idea is to represent each view as a similarity or distance matrix (e.g. Gaussian kernel, cosine similarity), then combine them. A subset of those methods are referred to as Similarity Fusion methods which combine per-view similarity graphs into a consensus graph using iterative diffusion processes Wang et al. (2014a). An example of this is Multi-view spectral clustering which builds a consensus graph Laplacian from fused similarities Kumar et al. (2011). These methods are often seen in bioinformatics, text, and image clustering, but are limited by scalability and sensitivity to the similarity metric used.

Matrix Factorisation-based Methods. In those methods each view's data is factorised, typically via Non-Negative Matrix Factorisation (NMF) or tri-factorization (NMTF) to enforce alignment across views. For example, Co-regularized NMF Liu et al. (2013) minimizes view-specific reconstruction loss plus a cross-view divergence penalty. Multi-view tri-factorization such as Restricted NMTF learns cluster indicator matrices that are shared or constrained across views Orme et al. (2025). These methods are interpretable, effective for sparse data but have strict linear assumptions and are sensitive to initialisation.

Bayesian Methods. The key idea with these methods is to model data using probabilistic models with shared latent variables across views which enable uncertainty quantification and the integration of prior knowledge. Bayesian methods are especially useful in biological datasets with missing modalities or sparse samples (Shapiro and Battle, 2024). They are flexible and robust to missing data but they are computationally expensive and rely on expert knowledge for the choice of prior(s).

Autoencoder-based Methods. Deep neural networks are used to learn non-linear latent embeddings per view. Some of the latest developments include Multi-view Variational Embedding Clustering (MVAEC) Zhao et al. (2017) which integrates representation learning with probabilistic clustering. They are powerful for nonlinear and high-dimensional data but are prone to training instability and interpretability is more challenging.

3.1.3. Motivations for Multi-View Clustering

Multi-View Clustering (MVC) offers a framework to address the limitations described in subsection 2.3.3. Rather than flattening or discarding structural richness, MVC approaches aim to integrate information across views while preserving view-specific characteristics. This mitigates against oversimplifications, as reducing to a single-view or flattened model can sacrifice transparency and potentially weakens the robustness of outcomes (Yu et al., 2024). It also allows better utilisation of incomplete data, as MVC can incorporate entities that are only partially observed by leveraging similarity or redundancy across views (Wen et al., 2023). Furthermore, view-aware integration methods can highlight the importance of individual views and provide insight into the cohesion and separation of clusters across them (Chao et al., 2021). By avoiding reliance on any single view, MVC can improve robustness and stability, reducing sensitivity

to noise, outliers, and view-specific biases (Yang and Wang, 2018). Finally, certain MVC approaches enhance interpretability by enabling cluster assignments to be traced back to specific views and features, thus increasing transparency (Chao et al., 2021).

In light of these advantages, our research continues with the premise that segmentation of local authorities, like many social science tasks, requires more than a collection of parallel single-view models or a naïve concatenation of indicators. It calls for an integrated, yet view-sensitive, modelling approach that can accommodate data incompleteness, promote robustness, and provide actionable insights to policymakers.

3.2. Application of Multi-View Clustering (MVC) to ONS Dataset

3.2.1. Experimental Framework and Methodology

To ensure methodological fidelity and some comparability with the ONS baseline (Figure 2.1), we fixed core elements of our experimental design to match the official ONS pipeline (subsection 2.2.3, Appendix D1). Specifically, all experiments applied PCA-based dimensionality reduction (retaining 2–6 components and $\geq$25% variance), followed by K-means clustering ($k \in [4, 15]$) evaluated via the silhouette score. This provides a controlled baseline while allowing us to directly assess how multi-view learning methods handle missingness and maintain interpretability. This enables us to test whether MVL can extract valuable information from all entities, even if partially observed, without exclusions, while mitigating the missingness–reliability trade-off, unlike the ONS approach.

To ensure breadth, we implement a methodologically diverse portfolio spanning three of the most common families of methods (described in subsection 3.1.2): Similarity-based, Matrix Factorisation-based, and Autoencoder-based approaches. Within each family, we select representative methods covering different integration points in the pipeline (early, intermediate, or late integration, see subsection 3.1.1). All methods share the same pre-processing, dimensionality reduction, clustering, and evaluation framework, with variations only in the methodological family and integration stage used, and further detailed in Appendix C2).

The methods employed represent simple, baseline formulations within each methodological family, chosen for their conceptual clarity and ease of implementation. We acknowledge that these are not representative of the architectural innovations or algorithmic optimisations that characterise the current state-of-the-art. Methodological breadth was prioritised over depth to ensure comparability across integration strategies, maintain transparency, and focus on assessing the relative behaviour of approach families rather than achieving maximal clustering performance. At the same time, we are breaking new ground in applying these methods to this specific problem context, with limited prior literature to guide the adaptation of more advanced methods, making our methodology necessarily more exploratory and application-driven rather than seeking incremental refinements on established best-in-class techniques.

Similarity-based Methods. They construct similarity graphs per view (even with missing values) and integrating them across views using graph-based operations that rely on distance measures.

- **Masked Cosine Similarity (MCS)** (Yin and Sun, 2022) is an early integration / similarity-based, missing-value aware. We compute a single sample–sample similarity directly from (preprocessed and standardised) features, using only the coordinates jointly observed for each pair of entities.

  Let $x_i \in \mathbb{R}^p$ denote the concatenated feature vector for entity i across all views (after alignment and scaling). For a pair (i, j) define the index set of commonly observed features $M_{ij} = \{\, m : \; x_{im} \text{ and } x_{jm} \text{ are observed} \,\}$. The masked cosine similarity is

  $$S_{ij} \;=\; \frac{\sum\limits_{m \in M_{ij}} x_{im}\, x_{jm}}{\sqrt{\sum\limits_{m \in M_{ij}} x_{im}^2}\;\sqrt{\sum\limits_{m \in M_{ij}} x_{jm}^2}},$$

  with $S_{ii} = 1$. When $M_{ij} = \varnothing$ we set $S_{ij} = 0$ (pair ignored) to avoid division by zero.

  To reduce degree/self-similarity effects we optionally form a row-stochastic matrix after zeroing the diagonal: $\tilde{S} \;=\; D^{-1}(S - \mathrm{diag}(S))\,, \qquad D_{ii} = \sum_j (S_{ij} - \delta_{ij} S_{ii})\,.$

  We treat the similarity as a kernel and embed it via Kernel PCA before clustering. We then symmetrise and perform eigenvalue clipping for numerical robustness, then apply Kernel PCA with a precomputed kernel, followed by K-means over a grid of dimensions and cluster counts $K = \frac{1}{2}\left(\tilde{S} + \tilde{S}^\top\right), \qquad Z = \mathrm{PCA}(K) \in \mathbb{R}^{n \times d}.$

  And K-means is applied to the resulting embeddings Z.

- **Multi-View Similarity Fusion (MVSF)**, a simplified version of Similarity Network Fusion (Wang et al., 2014b), an Intermediate integration / Similarity-based approach. Per-view similarities $S^{(v)} \in [0,1]^{n \times n}$ are built using 3.2.1 from each cleaned view via masked-cosine similarity on the union of entities; optional row-normalisation can be applied before fusion. Before PCA we symmetrise and, if needed, clip tiny negative eigenvalues to ensure a valid kernel. The fused similarity is the average:

  $$S_{\text{fused}} = \frac{1}{V} \sum_{v=1}^{V} S^{(v)}.$$

  We then symmetrise: $K = \frac{1}{2}\left(S_{\text{fused}} + S_{\text{fused}}^\top\right), \quad H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top, \quad K_c = HKH.$

  A d-dimensional representation is obtained via kernel PCA on K_c, and K-means is applied to the resulting embeddings.

The main difference with SNF from Wang et al. (2014b) is that SNF iteratively diffuses information across KNN graphs using several hyperparameters (K, T, μ), preserving view-specific neighbourhoods while blending information. MVSF deliberately removes the diffusion and KNN steps and uses the arithmetic mean; it is therefore parameter-free at the fusion stage and serves as a strong, transparent baseline. The method is detailed in Figure 3.2.

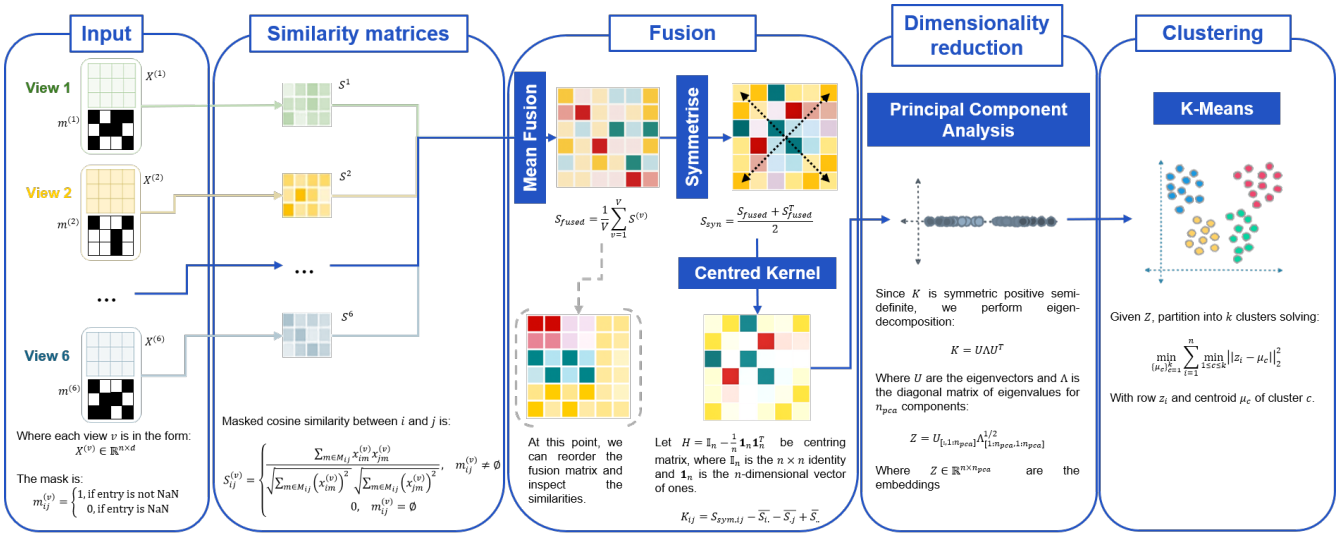

Figure 3.2.: Conceptual Diagram of Multi-View Similarity Fusion (MVSF) method

- **Co-association Similarity Ensemble (CSE)** (Fred and Jain, 2005) is a late-integration / similarity-based consensus method. Each view is clustered independently; the resulting partitions are aggregated via a co-association similarity, embedded with PCA, and re-clustered to obtain a single consensus labeling.

  Given V views on the same n entities, view v yields a partition $\{c_i^{(v)}\}_{i=1}^n$ from k_v-means after view-specific preprocessing. Let $\mathcal{V}_{ij} = \{\, v : \text{entities } i \text{ and } j \text{ both appear in view } v \,\}$ and $m_{ij} = |\mathcal{V}_{ij}|$.

  We define the co-association matrix $C \in [0,1]^{n\times n}$ by

$$C_{ij} = \begin{cases} 1, & i = j, \\ \dfrac{1}{m_{ij}} \displaystyle\sum_{v\in\mathcal{V}_{ij}} \mathbb{1}\left[c_i^{(v)} = c_j^{(v)}\right], & m_{ij} > 0,\ i \neq j, \\ 0, & m_{ij} = 0,\ i \neq j, \end{cases} \tag{1}$$

  where $\mathbb{1}[\cdot]$ is the indicator function. When every entity appears in every view ($m_{ij} = V$), reduces to $C_{ij} = \frac{1}{V}\sum_{v=1}^{V} \mathbb{1}[c_i^{(v)} = c_j^{(v)}]$.

  We treat C as a similarity kernel and apply PCA in the feature space induced by C. We first symmetrise and centre: $K = \frac{1}{2}\left(C + C^\top\right), \qquad H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top, \qquad K_c = HKH.$

Let $K_c = U\Lambda U^\top$ with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$. The d-dimensional embedding used for consensus clustering is $Z = U_{1:d}\,\Lambda_{1:d}^{1/2} \in \mathbb{R}^{n\times d}$.

We apply K-means to Z to obtain the final consensus labels. In our experiments, $d \in \{2,\ldots,5\}$ and $k \in \{4,\ldots,15\}$ are selected by the best average silhouette score (with fixed random seeds for reproducibility).

CSE is parameter-free at the fusion stage and naturally handles partial coverage via m_{ij} in (1). This is directly inspired from the evidence–accumulation clustering of Fred and Jain (2005); our choice of PCA on the co-association kernel provides a simplified and transparent consensus function before K-means. The method is detailed in Figure 3.3.

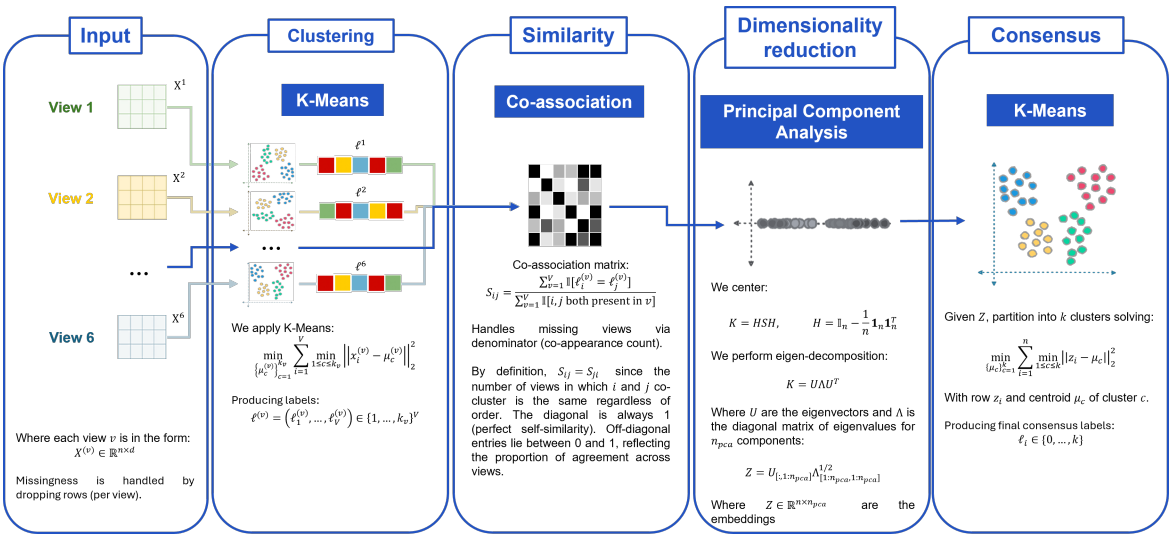

Figure 3.3.: Conceptual Diagram of Co-Association Similarity Ensemble (CSE) method

Matrix Factorisation-based Methods. These methods infer latent low-rank structure directly from observed entries, avoiding imputation and allowing selective handling of unassigned entries.

- **Masked Non-negative Matrix Factorisation (MNMF)** is an early integration / factorisation-based method inspired by Lee and Seung (1999). We factorise a single, union-aligned, nonnegative feature matrix $X \in \mathbb{R}_+^{n\times p}$ built from all views, under an elementwise missingness mask $M \in \{0,1\}^{n\times p}$, and then cluster the resulting low-dimensional representation via PCA $\rightarrow$ K-means.

  Given target rank r, MNMF seeks nonnegative factors $G \in \mathbb{R}_+^{n\times r}$ (row factors; sample embedding) and $H \in \mathbb{R}_+^{p\times r}$ (column loadings) by minimising the masked Frobenius loss

$$\min_{G,H\geq 0} \left\| M \odot \left(X - GH^\top\right) \right\|_F^2, \tag{2}$$

  where $\odot$ denotes the Hadamard product and $M_{ij} = 1$ if X_{ij} is observed (else 0). In our implementation X is zero-filled on missing entries and the mask M ensures that unobserved entries do not contribute to the loss.

With nonnegative fixed seed initialisation, we iterate masked Lee–Seung-style updates (with a small $\varepsilon > 0$ for numerical stability):

$$G \leftarrow G \odot \frac{(M \odot X)H}{\big(M \odot (GH^\top)\big)H + \varepsilon},$$
$$H \leftarrow H \odot \frac{(M \odot X)^\top G}{\big(M \odot (GH^\top)\big)^\top G + \varepsilon}.$$

We check the masked objective in (2) periodically and stop early when the relative improvement falls below a tolerance; maximum iterations and tolerance are user-set (defaults: `max_iter=200`, `tol=`10^{-4}).

We take the row factors G as the sample embedding and apply linear PCA to obtain a d-dimensional representation $Z = \mathrm{PCA}_d(G) \in \mathbb{R}^{n \times d}$. We then cluster Z with K-means. A grid search selects hyperparameters by silhouette score:

$$r \in \{2, \ldots, 5\}, \quad d \in \{2, \ldots, 5\} \cap \{1, \ldots, r\}, \quad k \in \{4, \ldots, 15\}.$$

For each (r, d, k) we fit MNMF $\rightarrow$ PCA(d) $\rightarrow$ K-means(k) and record the average silhouette on Z; the best configuration is reported.

Inputs are expected nonnegative (consistent with NMF); missing entries are handled by (X, M) as above. PCA dimensionality cannot exceed the MNMF rank r so requests beyond r are skipped in our implementation.

- **Per-view NMF with Concatenation (PVNMF)** is an intermediate integration / factorisation-based method. Each view is factorised independently via NMF to obtain a per-view, nonnegative embeddings; these embeddings are column-concatenated into a single representation that is embedded with PCA and clustered by K-means (Liu et al., 2013).

  For view v with nonnegative matrix $X^{(v)} \in \mathbb{R}_+^{n_v \times p_v}$ we solve

$$\min_{G^{(v)},\, F^{(v)} \geq 0} \left\| X^{(v)} - G^{(v)} {F^{(v)}}^\top \right\|_F^2,$$

  obtaining $G^{(v)} \in \mathbb{R}_+^{n_v \times r_v}$ and $F^{(v)} \in \mathbb{R}_+^{p_v \times r_v}$, where r_v is the per-view rank (capped by $\min(n_v, p_v)$ in code and initialised with `nndsvda`). Rows with any missing values are dropped per view and views with negative entries are skipped to respect NMF's nonnegativity constraint.

  Align on the union index of entities and column-concatenate the row factors:

$$G_{\text{concat}} = \left[\, G^{(1)} \;\middle|\; \cdots \;\middle|\; G^{(V)} \,\right] \in \mathbb{R}_+^{n \times R}, \qquad R = \sum_{v=1}^{V} r_v.$$

Normalise each row to unit length (with $\varepsilon > 0$ for stability):

$$\widehat{G}_{i\cdot} = \frac{G_{\text{concat},\, i\cdot}}{\sqrt{\sum_{k=1}^{R} g_{ik}^2 + \varepsilon}}, \qquad i = 1, \ldots, n.$$

Let $\mathbf{1} \in \mathbb{R}^n$ be the all-ones vector and compute the column mean $\mu = \frac{1}{n}\widehat{G}^\top \mathbf{1} \in \mathbb{R}^R$. Center rows $Q = \widehat{G} - \mathbf{1}\mu^\top$. Form the Gram matrix and eigendecompose: $K = QQ^\top \in \mathbb{R}^{n\times n}$, $K = U\Lambda U^\top$ with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$. The d-dimensional PCA scores are $Z = U_{1:d}\,\Lambda_{1:d}^{1/2} \in \mathbb{R}^{n\times d}$. K-means is applied to Z. PVNMF is depicted in Figure 3.4.

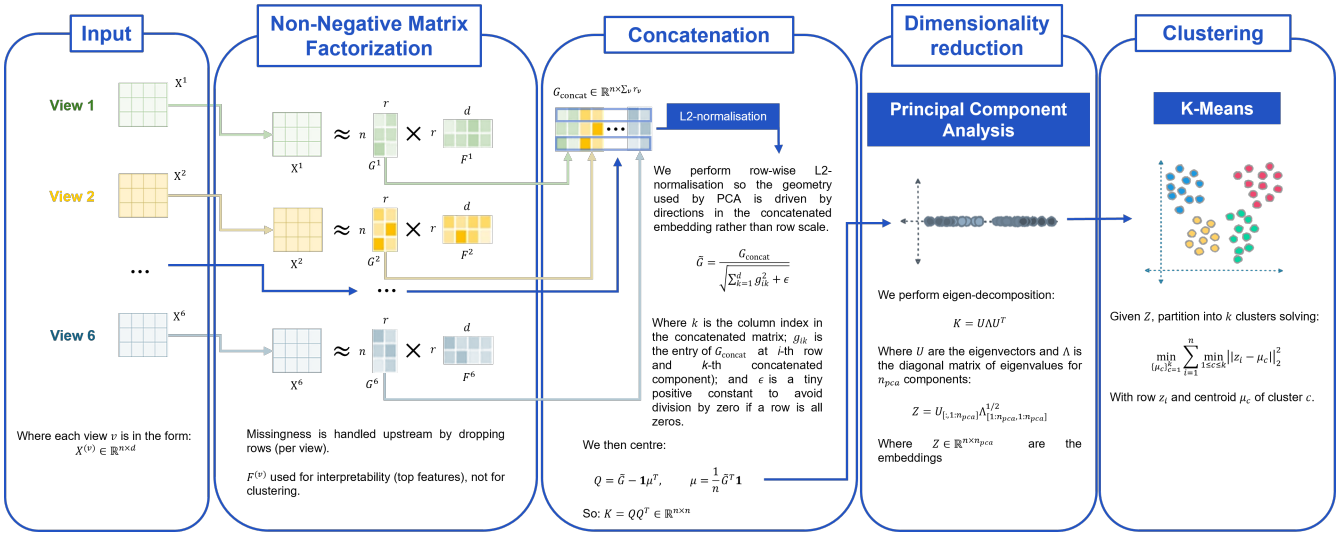

Figure 3.4.: Conceptual Diagram of Per-View Non-Negative Matrix Factorization (PVNMF) method

- **Co-Association Factorization Ensemble (CFE)**, a late integration / factorisation-based consensus. Each view is factorised by Nonnegative Matrix Tri-Factorisation (NMTF) to obtain a low-noise, parts-based partition of the entities; we then aggregate these partitions through a co-association similarity and recluster the induced embedding.

  For view v with nonnegative matrix $X^{(v)} \in \mathbb{R}_+^{n_v \times p_v}$, we fit an NMTF model of ranks (k_v, ℓ_v),

$$\min_{G^{(v)},\, S^{(v)},\, H^{(v)} \geq 0} \left\| X^{(v)} - G^{(v)} S^{(v)} H^{(v)\top} \right\|_F^2, \quad G^{(v)} \in \mathbb{R}_+^{n_v \times k_v},\ S^{(v)} \in \mathbb{R}_+^{k_v \times \ell_v},\ H^{(v)} \in \mathbb{R}_+^{p_v \times \ell_v},$$

  optionally with row/column-orthogonality regularisation on $G^{(v)}$ and $H^{(v)}$ as in **?**. We then derive a hard partition for the entities in view v by clustering the rows of $G^{(v)}$ with K-means into k_v groups (fixed seeds).

  Our code reuses the exact co-association → centring → PCA → K-means path shown in the CSE method (Figure 3.3). The only change is where we compute NMTF factors $(G^{(v)}, S^{(v)}, H^{(v)})$ and cluster the rows of $G^{(v)}$ to obtain labels before building C.

NMTF models $X^{(v)} \approx G^{(v)} S^{(v)} H^{(v)\top}$ and thus discovers simultaneous clusters of rows and columns; $G^{(v)}$ emphasises sample groups that share coherent feature blocks via $S^{(v)}$, typically yielding crisper per-view entity partitions than two-factor NMF when features are redundant or view-specific (Ding et al., 2006). The middle map $S^{(v)}$ absorbs cross-cluster interactions and rescales between feature groups, which we found stabilises the ensuing co-association (less spurious agreement). On our data, the NMTF-based CFE produced high-quality consensus partitions (silhouette often > 0.80), while a late-integration variant built from per-view NMF partitions and the same co-association pipeline plateaued around ~ 0.60, and a per-view feature CFE baseline that clusters in the original feature spaces and then co-associates never exceeded ~ 0.51. A diagram for CFE is shown in Figure 3.4.

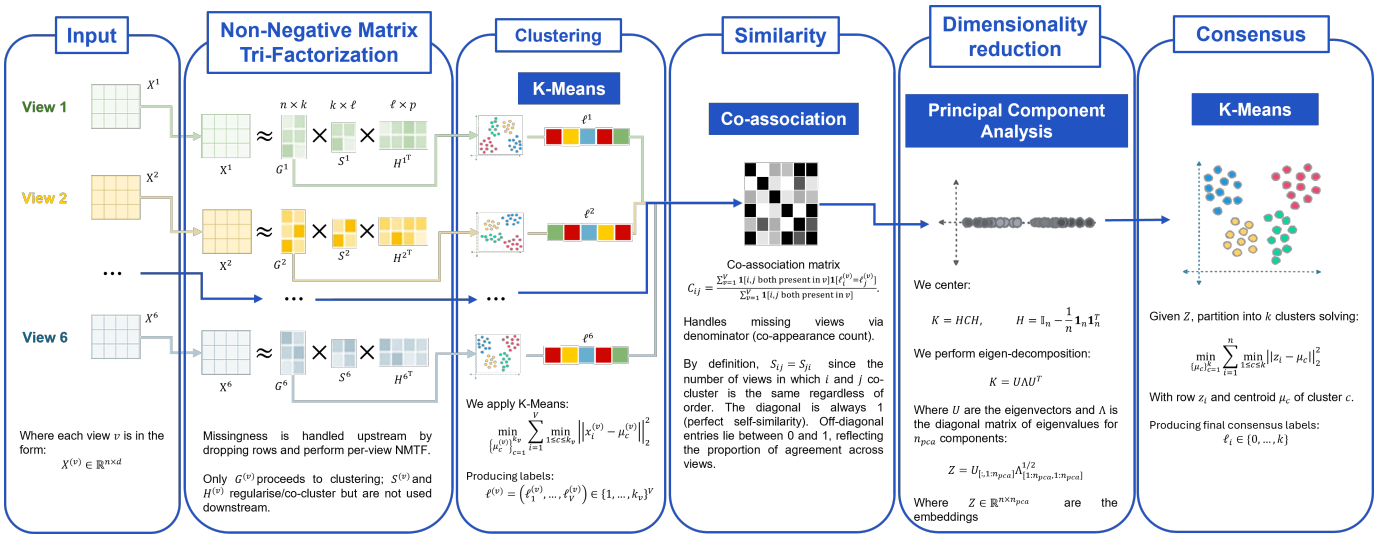

Figure 3.5.: Conceptual Diagram of Co-Association Factorization Ensemble (CFE) method

Autoencoder-based Methods. Learn non-linear latent representations from partial views using embeddings, with masking to explicitly disregard missing values.

- **Masked Autoencoder (MA)**, an early integration / autoencoder-based method. We train a single autoencoder on the feature matrix and reconstruct only the observed entries via a masked loss, then apply PCA and K-means on the learned latent codes (cf. Vincent et al. 2008).

  After aligning entities across all views, let $X \in \mathbb{R}^{n \times p}$ be the real-valued feature matrix and $M \in \{0,1\}^{n \times p}$ the observation mask, $M_{ij} = 1$ if X_{ij} is observed and 0 otherwise. We apply NaN-aware robust scaling feature-wise: for feature j, with median m_j and IQR $s_j > 0$ computed on observed values only,

$$X_{ij}^{\mathrm{sc}} = \frac{\tilde{X}_{ij} - m_j}{s_j}, \qquad \tilde{X}_{ij} = \begin{cases} X_{ij}, & M_{ij} = 1, \\ m_j, & M_{ij} = 0, \end{cases}$$

(IQR = 0 falls back to $s_j = 1$). Missing entries remain excluded by M in the loss.

The encoder is a two-layer MLP with ReLU nonlinearity, $f_\theta : \mathbb{R}^p \to \mathbb{R}^r$ and the decoder $g_\phi : \mathbb{R}^r \to \mathbb{R}^p$ mirrors this size:

$$z_i \;=\; f_\theta(x_i^{\mathrm{sc}}), \qquad \hat{x}_i^{\mathrm{sc}} \;=\; g_\phi(z_i),$$

where x_i^{sc} is the i-th row of X^{sc}.

We minimise

$$\mathcal{L}(\theta, \phi) = \frac{1}{n} \sum_{i=1}^{n} \frac{\left\| \left(\hat{x}_i^{\mathrm{sc}} - x_i^{\mathrm{sc}}\right) \odot m_i \right\|_2^2}{m_i^\top \mathbf{1} + \varepsilon}, \tag{3}$$

where m_i is the i-th row of M and $\varepsilon > 0$ prevents division by zero. This per-row normalisation avoids biasing the loss toward rows with many observed values. We optimise with Adam (lr 10^{-3}) for a fixed number of epochs (default 100), with deterministic seeds for PyTorch/NumPy for reproducibility.

After training, we obtain latent codes $H = [h_1, \ldots, h_n]^\top \in \mathbb{R}^{n \times r}$ and compute a linear PCA embedding $Z = \mathrm{PCA}_d(H) \in \mathbb{R}^{n \times d}$. We then run K-means on Z to produce labels. We perform a grid search over latent dimension $L \in \{5, 8, 10\}$, PCA dimension $C \in \{2, \ldots, 5\}$, and clusters $k \in \{4, \ldots, 15\}$; for each pair (d, k) we retain the latent L that maximises the average silhouette score (deduplicated grid).

We return the selected labels, PCA embedding, assignments table, and the deduplicated grid of results reporting the best latent per (C, k).

Conceptually, the masked objective in (3) relates to denoising autoencoders (Vincent et al., 2008): training focuses on reconstructing only the reliable (observed) coordinates, which yields a robust latent representation under missingness, without explicitly corrupting inputs.

- **Multibranch Autoencoder with Shared Bottleneck (MBASB)** is an intermediate integration / autoencoder-based method. Each view has its own encoder; their latent outputs are fused by an elementwise mean to form a shared bottleneck H, which a single decoder uses to reconstruct the concatenated feature vector. We then apply PCA to H and cluster with K-means (Ngiam et al., 2011).

  Let $X^{(v)} \in \mathbb{R}^{n \times p_v}$ be the cleaned/aligned per-view matrices with masks $M^{(v)} \in \{0, 1\}^{n \times p_v}$. After NaN-aware robust scaling, concatenate along features:

$$X = \left[X_{\mathrm{sc}}^{(1)} \,\middle|\, \cdots \,\middle|\, X_{\mathrm{sc}}^{(V)}\right] \in \mathbb{R}^{n \times p}, \quad M = \left[M^{(1)} \,\middle|\, \cdots \,\middle|\, M^{(V)}\right] \in \{0, 1\}^{n \times p}, \quad p = \textstyle\sum_v p_v.$$

  Each encoder $f_\theta^{(v)} : \mathbb{R}^{p_v} \to \mathbb{R}^r$ is a two-layer perceptron. For row i,

$$h_i^{(v)} = f_\theta^{(v)}\big(x_{i,\mathrm{sc}}^{(v)}\big) \in \mathbb{R}^r, \qquad H_i = \frac{1}{V} \sum_{v=1}^{V} h_i^{(v)} \;\; \text{(shared bottleneck)}.$$

A single decoder $g_\phi : \mathbb{R}^r \to \mathbb{R}^p$ reconstructs all features:

$$\hat{x}_i = g_\phi(H_i) \in \mathbb{R}^p.$$

We minimise the same masked per-row reconstruction loss defined in Eq. (3) using $X, \hat{X}, M$. Backpropagating through the mean gives

$$\frac{\partial \mathcal{L}}{\partial h_i^{(v)}} = \frac{1}{V} \frac{\partial \mathcal{L}}{\partial H_i},$$

so each encoder receives an equal gradient share.

After training, we follow exactly the same process as in MA above. Our design shown in Figure 3.6 follows the multimodal shared-representation idea of Ngiam et al. (2011), but uses a single decoder to all features with a masked per-row loss to handle missingness—a practical choice for our incomplete, heterogeneous views. If independent per-view decoders were desired, the loss would split across views with their respective masks; our implementation does not do this.

- **Autoencoders Ensemble with Consensus Clustering (AECC)** is a late integration / autoeconder-based approach. We train a separate masked autoencoder on each view, cluster each view's PCA embedding, and combine the resulting partitions via a co-association consensus (as in Strehl and Ghosh 2002). This parallels our CSE pipeline but replaces the per-view "feature → K-means" step with "AE → PCA → K-means".

  For each view v, with scaled data $X^{(v)} \in \mathbb{R}^{n \times p_v}$ and mask $M^{(v)}$, we fit a masked autoencoder (two-layer perceptron encoder/decoder; same architecture as in MA) by minimising the same masked per-row reconstruction loss $\mathcal{L}_{\text{masked}}$ in Eq. (3), applied to $(\hat{X}^{(v)}, X^{(v)}, M^{(v)})$. After training we extract the bottleneck matrix

  $$H^{(v)} \in \mathbb{R}^{n \times r},$$

  then compute a linear PCA embedding $Z^{(v)} = \text{PCA}_d\big(H^{(v)}\big) \in \mathbb{R}^{n \times d}$, and obtain per-view labels $c^{(v)} = \text{KMeans}_k\big(Z^{(v)}\big)$. Rows with no observed features in view v are marked as missing (label -1) and excluded from agreements in the next step.

  From $\{c^{(v)}\}_{v=1}^{V}$ we build an $n \times n$ co-association matrix C (a label-derived similarity) exactly as in our CSE:

  $$C_{ij} = \frac{1}{m_{ij}} \sum_{v \in \mathcal{V}_{ij}} \mathbb{1}\Big[c_i^{(v)} = c_j^{(v)}\Big], \qquad \mathcal{V}_{ij} = \{\, v : \; c_i^{(v)} \neq -1, \; c_j^{(v)} \neq -1 \,\}, \quad m_{ij} = |\mathcal{V}_{ij}|,$$

  with $C_{ii} = 1$ and $C_{ij} = 0$ when $m_{ij} = 0$

  We apply PCA directly to C to obtain the consensus embedding $Z = \text{PCA}_d(C)$ and then run K-means with the same k (our code ties the per-view and consensus d, k grids).

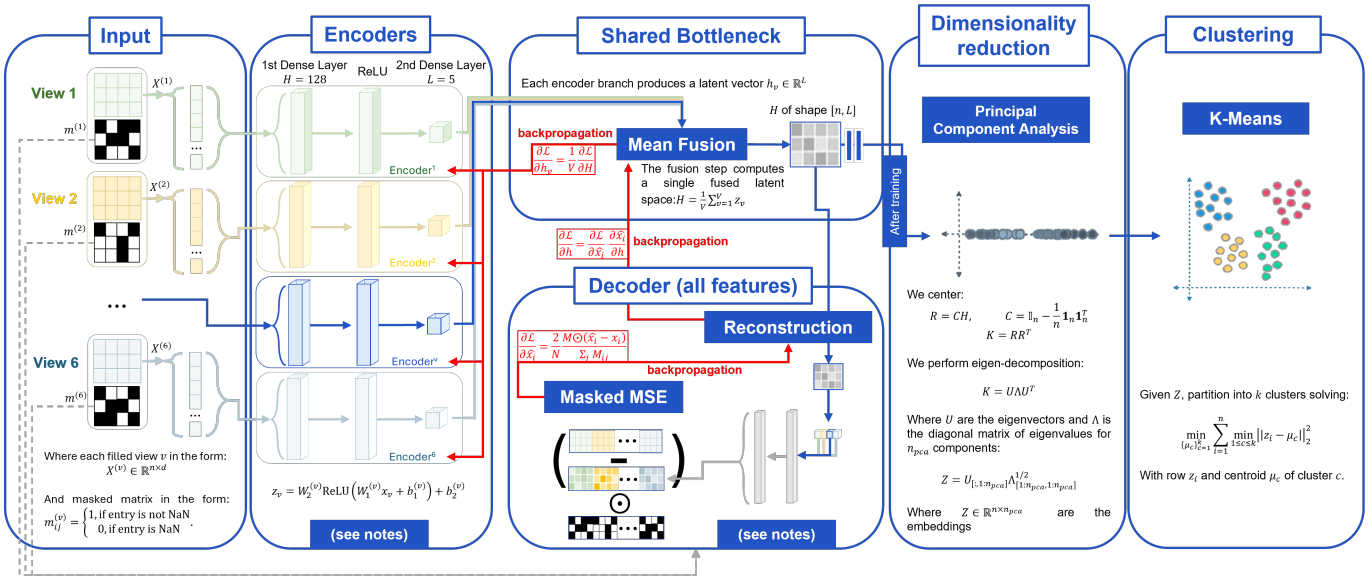

Figure 3.6.: Conceptual Diagram of Multibranch Autoencoder with Shared Bottleneck (MBASB) method

---

Notes.

$$x^{(v)} \in \mathbb{R}^{d_v}, \quad h^{(v)} = \mathrm{ReLU}\left(W_1^{(v)} x^{(v)} + b_1^{(v)}\right), \quad z^{(v)} = W_2^{(v)} h^{(v)} + b_2^{(v)} \in \mathbb{R}^r.$$

- $W_1^{(v)} \in \mathbb{R}^{q \times d_v}$, $b_1^{(v)} \in \mathbb{R}^q$ (encoder hidden width q).
- $W_2^{(v)} \in \mathbb{R}^{r \times q}$, $b_2^{(v)} \in \mathbb{R}^r$ (latent dim r).
- $\mathrm{ReLU}(t) = \max(0, t)$.

Fusion (shared bottleneck):

$$H = \frac{1}{V} \sum_{v=1}^{V} z^{(v)} \in \mathbb{R}^r \quad \text{(per sample; stacked over samples gives } H \in \mathbb{R}^{n \times r}\text{)}.$$

Single decoder (all features):

$$\hat{h} = \mathrm{ReLU}(W_{d1} H + b_{d1}), \qquad \hat{x} = W_{d2} \hat{h} + b_{d2} \in \mathbb{R}^p,$$

with $W_{d1} \in \mathbb{R}^{q \times r}$, $b_{d1} \in \mathbb{R}^q$, $W_{d2} \in \mathbb{R}^{p \times q}$, $b_{d2} \in \mathbb{R}^p$.
Reconstruction loss: use the same masked per-row loss as MA, Eq. (3), applied to the concatenated $(X, \hat{X}, M)$.
Backprop through mean fusion: $\frac{\partial \mathcal{L}}{\partial z^{(v)}} = \frac{1}{V} \frac{\partial \mathcal{L}}{\partial H}$.

The best (d, k) is selected by average silhouette; for each (d, k) we keep the latent r that attains the highest score (deduplicated grid).

AECC is a label-derived similarity ensemble—identical in spirit to CSE's co-association consensus, but where each base partition comes from an AE (MA-style) rather than direct K-means on features. This method is detailed in Figure 3.7.

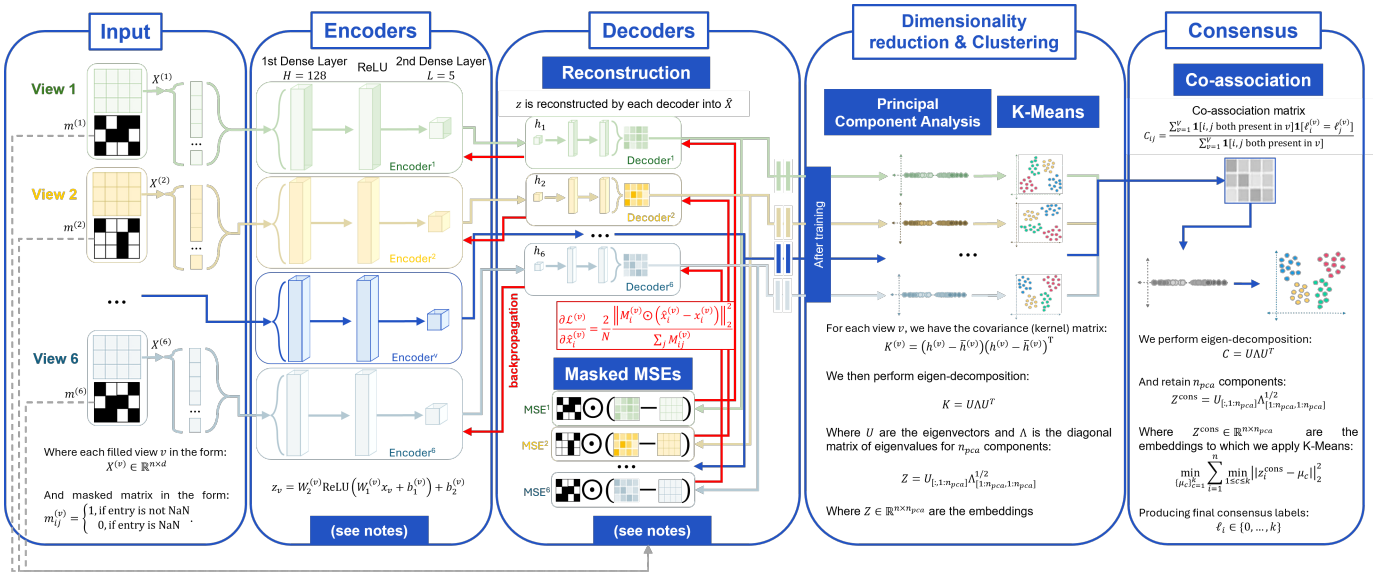

Figure 3.7.: Conceptual Diagram of Autoencoder Ensemble with Consensus Clustering (AECC) method

Addressing Interpretability. We will of course employ summary statistics to inspect cluster characteristics, but multi-view learning unlocks a richer suite of tools tailored to enhance interpretability. The integration of multiple partial views via common latent representation allows for a deeper investigation into both the formation and coherence of clusters. For example, by inspecting a fused similarity matrix or latent embeddings, we gain structured insight into how different views influence clustering outcomes. Table 3.1 summarises the techniques employed to ensure interpretability across structural, view-level, and local authority (LA)-level dimensions.

Together, these techniques provide a multi-scale interpretability framework. Structural measures clarify the overall shape and reliability of the clustering solution. View-level diagnostics reveal the cohesion or dominance exerted by views. Meanwhile, LA-level tools such as graph-based attribution uncover the local dynamics driving pairwise consensus. This layered approach ensures not only transparency in how clusters arise, but also in understanding the role of each contributing view, feature and even entity (LA) to a shared representation, balancing both global structure and local nuance.

The methods described inTable 3.1 are further detailed in Appendix E4.

Table 3.1.: Interpretability techniques used in the multi-view clustering pipeline

Technique	What it explains	Coverage
Common Latent Representation + hierarchical reordering	Provides a global overview of pairwise LA similarity and reveals structural clusters via hierarchical ordering.	Structural
View ablation for silhouette	Measures the impact of each view on overall cluster structure by re-running clustering without one view at a time.	View-level
Feature ablation for silhouette	Measures the impact of each feature (within each view) on overall cluster structure by re-running clustering without one feature at a time.	Feature-level
Per-View contributions to shared representation	Captures how much each view contributes to the final latent space, globally.	View-level
Per-LA contributions to shared representation	Captures how much each view influences the range of contributions by LAs to the final latent space, locally.	LA-level
Graph-based attribution to shared representation	Visualises LA-to-LA attribution with edges and highlights which view dominates each connection, informing global and local contributions to the latent space.	Structural & LA-level

Notes. These techniques were chosen to ensure interpretability of both global and local structures emerging from data fusion and clustering. Each contributes toward understanding not just which clusters exist, but why they form, and how different data views shape the results.

3.2.2. Evaluation and Diagnostics Framework

Risks of the experimental design and its evaluation. The primary risks inherent to our experimental design arise from the absence of ground truth, extensive missingness, and potential noise in the data. Together, these factors make it difficult to determine whether the cluster structures observed reflect genuine patterns or are instead artefacts of these limitations. This is particularly relevant given the reliance on clustering quality metrics such as the Silhouette score.

The Silhouette score measures the extent to which individual samples are well-matched to their assigned cluster compared to other clusters, using distances in the embedding space. When similarity/distance-based methods are used to generate the embeddings and to evaluate the clusters, there is a risk that the score will be inflated. In particular, when missingness patterns are structured in the data, they can themselves drive separation between samples. Many similarity-based approaches address missing values by excluding unobserved features from pairwise calculations, which can inadvertently increase distances between groups with differing levels of completeness. If missingness correlates with substantive characteristics (for example, deprived areas having fewer reported values for education or health), then the embeddings may capture both genuine variation and patterns of data availability.

This does not necessarily invalidate the clusters: if missingness patterns are associated with real-world groupings, then separation driven partly by missingness may still reflect meaningful structure. However, it does highlight the risk that such clusters may not generalise to datasets with more complete or differently distributed missingness.

Evaluation framework. To address these risks and to also support robust interpretation (detailed in subsection 3.2.1), the evaluation framework is built around four complementary diagnostic components: cluster summaries, canonical profiles, a silhouette report, and a fusion

report. These components are designed to be agnostic to the specific clustering or integration method used, and can therefore be applied to both the current approach and future methods.

1. Cluster summaries. They quantify and visualise both the completeness of the data within each cluster and the behaviour of features across clusters. Missingness summaries present the proportion of missing values for each feature in each cluster, enabling identification of clusters that may be defined partly by data availability. Feature behaviour summaries report central tendency (medians) and dispersion (standard deviations) for each feature within clusters. To aid comparison, median values are normalised on a per-feature basis so that relative differences between clusters are emphasised.

2. Canonical profiles. They serve as interpretable baselines for evaluating cluster composition. These are constructed from features that exhibit strong relationships across views or modalities, and include an overall reference profile based on median values, as well as variants defined by external attributes. Comparing derived clusters against these profiles allows identification of those that align with known underlying relationships we identified in the data, as well as those that deviate markedly and may represent noise or methodological artefacts.

3. Silhouette report. It provides both global and per-cluster measures of cohesion and separation. It also supports stratified analysis, for example by grouping samples according to their proportion of missing data and calculating silhouette scores within these strata. Systematically higher scores in strata with high missingness may indicate that the observed silhouette score is inflated by patterns of data unavailability rather than substantive relationships between observations.

4. Fusion report. The fusion report examines the extent to which the integrated representation draws balanced contributions from all views or modalities, and the degree to which cluster structures are consistent across them. Contributions can be summarised at the level of the whole dataset and for individual samples. Cross-view agreement can be assessed using measures such as ARI or NMI (presented earlier), comparing the cluster assignments produced by each view to those from the integrated solution. This ensures that the integrated clusters are not dominated by a single view and that the shared structure across views is preserved.

The methods described in this subsection are further detailed in Appendix E4.

Chapter 4. Evaluating MVC: Results, Diagnostics, and Benefits

4.1. Experimental Results

4.1.1. Comparative Analysis of Methods

Figure 4.1 shows the comparison of each multi-view learning method's silhouette performance by strategy and model, it illustrates the spread and peak of performance across the grid search of PCA components and cluster numbers.

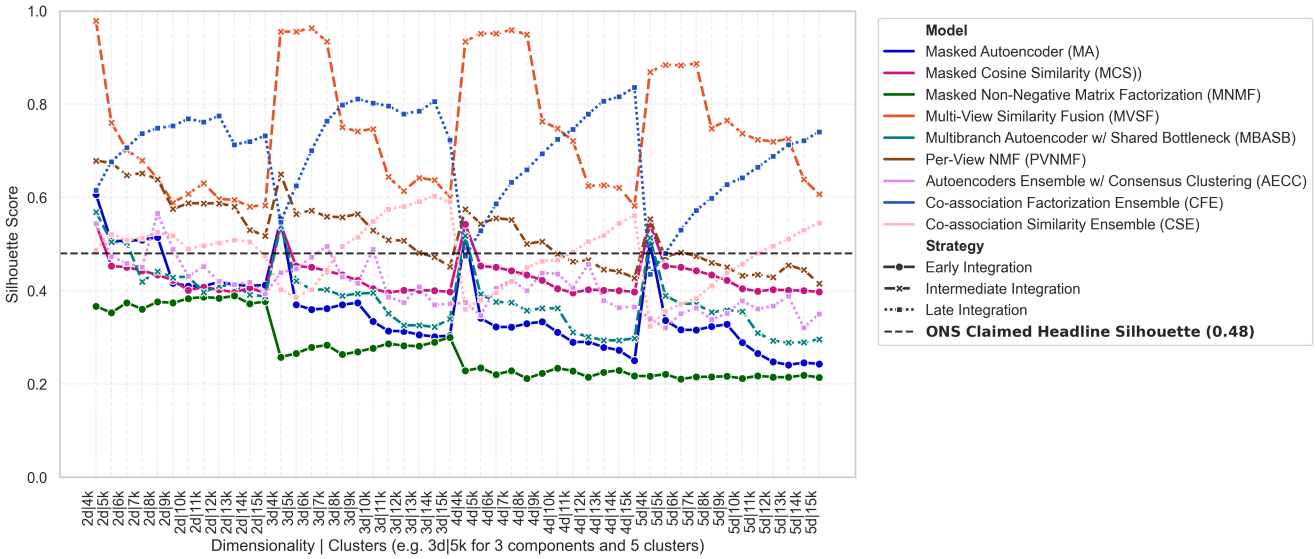

Figure 4.1.: Silouhette scores across the Grid Search by Experimental Model (and Integration Strategy) applied to data with missingness (n=309)

Notes. While we display results only across PCA components (C) and clusters (k), the full grid search also explored additional latent parameters specific to some method:

- Factorisation rank for Matrix Factorisation methods (matching the range of C to avoid breakage)
- Latent Dimension for Autoencoder methods ($\in\{5,8,10\}$)

For each (PCA, k) combination, the best-performing setting of those latent parameters was selected (based on silhouette score) and is reported in both Figure 4.1 and Table 4.1.

Table 4.1 shows that Multi-View Similarity Fusion (Intermediate) in orange, achieved the highest silhouette score by far (0.98), suggesting strong internal cohesion and separation of clusters, despite missing data. However, its performance is highly unstable across parameter combinations, dropping to values around 0.6 in some configurations. This volatility raises important concerns, notably sensitivity to cluster count (k), especially if the fused similarity matrix amplifies

weak signals or noise from dominant views. This is particularly problematic in high-missingness settings, where noisy signals can dominate similarity structure.

This variance is also obvious in our next best model: Co-Association Factorization Ensemble (Late) in blue, with silhouette score at 0.84. But unlike MVSF, it performs best at high values of k. Per-View NMF (Intermediate) is our third best model at 0.68 with the Multibranch Autoencoder at a disappointing 0.57.

All other methods with early or late integrations peaked below 0.6, showing more modest average performance with higher variance, indicating some sensitivity to parameter settings.

Table 4.1.: Multi-View methods by Strategies, Performance and Optimal Parameters

Integration Strategy	**Model**	**Max. Silh**	**Std. Silh**	**Opt. n (PCA)**	**Opt. k**
SVL / Early	ONS PCA + K-Means	0.49	—	—	4
Early	Masked Cosine Similarity	0.54	0.040	2	4
	Masked Non-Negative Matrix Factorization	0.39	0.065	2	13
	Masked Autoencoder	0.61	0.091	2	4
Intermediate	Multi-View Similarity Fusion	**0.98**	**0.135**	2	4
	Per-View NMF w/ Concat	0.68	0.072	2	4
	Multibranch Autoencoder	0.57	0.070	2	4
Late	Co-association Similarity Ensemble	0.60	0.069	3	14
	Co-association Factorization Ensemble	**0.84**	**0.100**	4	15
	Autoencoders Ensemble + Consensus Clustering	0.57	0.055	2	8

Notes. **Max. Silh**: Maximum Silhouette Score. **Std. Silh**: Standard Deviation of Silhouette Scores. **Opt. n (PCA)**: optimal number of PCA components. **Opt. k**: optimal number of clusters determined via grid search. Unknown information is marked as "—". **Notable values highlighted in bold**.

4.1.2. Further exploration

While we evaluated several methods (see Appendix E3), we now detail the results of Multi-View Similarity Fusion (MVSF) because it directly addresses our two key challenges: missingness and interpretability, while delivering the highest silhouette scores, substantially exceeding those reported by the ONS. Despite some sensitivity to hyperparameters, MVSF is methodologically simple, fuses partial observations without imputation, and is transparent: its view-specific similarity matrices expose clustering structure and enable global and local attribution (per LA, feature, and view). This combination of accuracy, robustness to incomplete data, and interpretability makes MVSF the most suitable approach for our setting.

4.2. Evaluating Multi-View Similarity Fusion (MVSF)

4.2.1. Insights

The Multi-View Similarity Fusion approach identified four distinct clusters, including one singleton outlier: Rutland, the smallest local authority in England (Cluster 3). The cluster configuration reflects a clear socio-economic stratification:

- Cluster 1 captures most deprived areas ($n = 85$), closely resembling the low-income canonical profile.
- Cluster 2 represents least deprived areas ($n = 37$), aligning well with the high-income archetype.
- Cluster 0 includes the majority of Local Authorities ($n = 186$ out of $N = 309$) and appears more mixed or average across views despite still producing silhouette scores above 0.8.

This segmentation is illustrated in Figure 4.2 (mean feature profiles), validated against canonical profiles in Figure 4.3, showing strong structural alignment, particularly on features like income, education, and life expectancy. Minor deviations (e.g. employment rate or child obesity) highlight nuanced differences not fully captured by canonical archetypes.

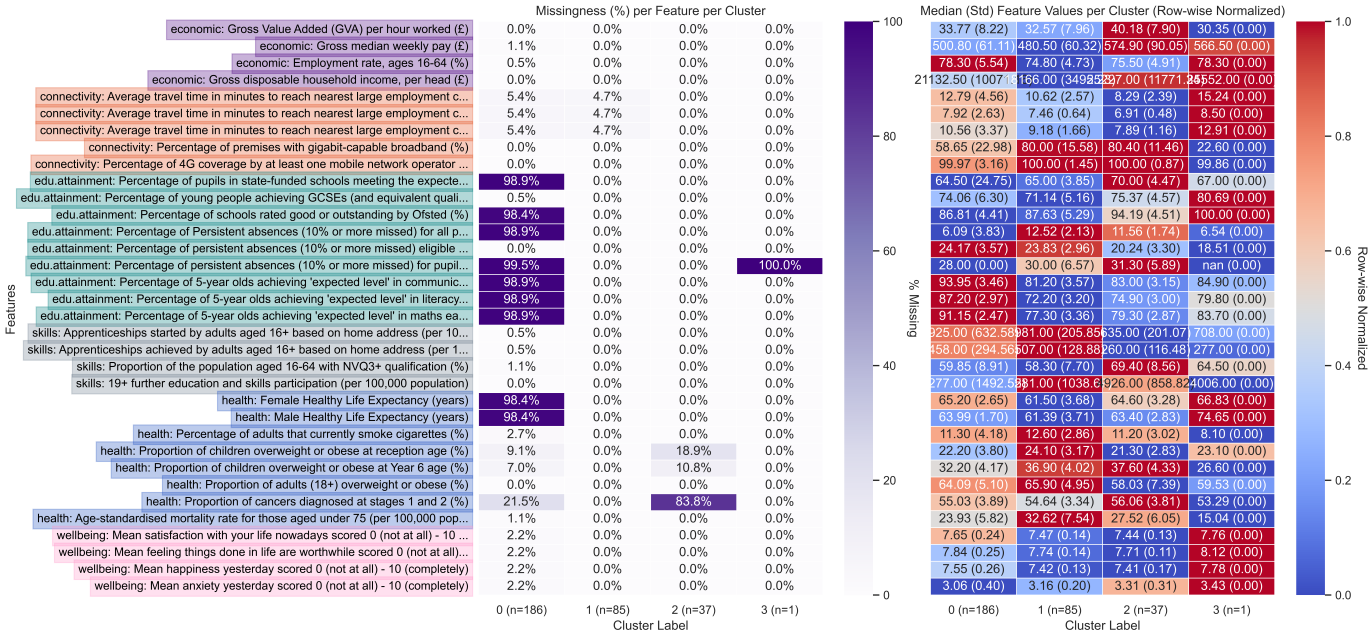

Figure 4.2.: Missingness and Mean Feature values by Cluster (with standard deviation)

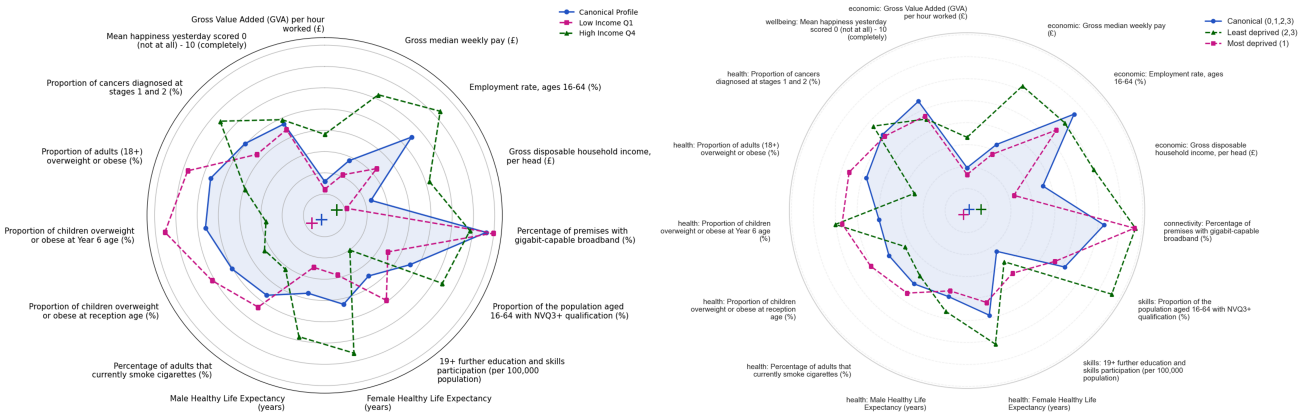

Figure 4.3.: Comparison Canonical Profiles (left) and MVSF clusters (right)

4.2.2. Diagnostics

Missingness by cluster (Figure 4.2) confirms that the largest cluster is dominated by incomplete cases, suggesting the clustering may be based on data missingness, but there is also evidence that it does not rely on missingness patterns alone. Silhouette scores by missingness strata (Figure 4.4) show that cohesion remains relatively consistent across quartiles of missingness, reducing concerns about silhouette inflation from missingness structure. Fusion error by cluster (Figure 4.5) is relatively low and uniform, indicating that the shared representation generalises well across clusters and is not disproportionately influenced by noisy or missing observations.

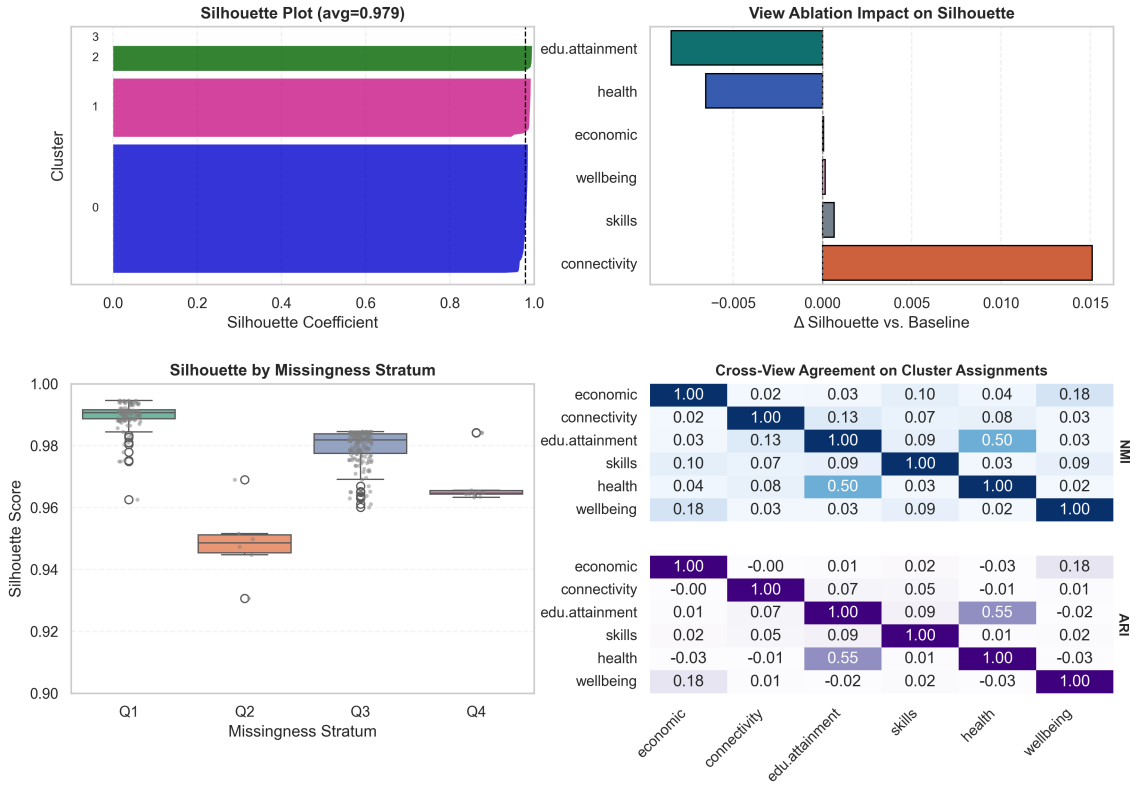

Figure 4.4.: MVSF Silhouette Report: Impact of Views, Missingness and Cross-View Agreement

The view ablation results in Figure 4.4 indicate that removing the educational attainment and health views, precisely the two views with the highest proportion of missing values and the strongest correlation in missingness patterns, introduces a minor drop in silhouette score (around –0.02 combined). According to NMI and ARI on cluster assignments, these views are in significant agreement with the fused clustering, yet their contribution to the fused similarity matrix is minimal compared to other views (see Figure 4.5). This suggests that their influence on clustering quality is not primarily through their similarity weights in the fusion step, but rather through indirect structural effects, potentially reinforcing certain partitions that align with missingness patterns. Thus, while the clustering does not appear to be solely driven by missingness, these views seems to exert disproportionate impact relative to their measured contribution, underscoring the importance of careful interpretation of fusion contributions alongside ablation results.

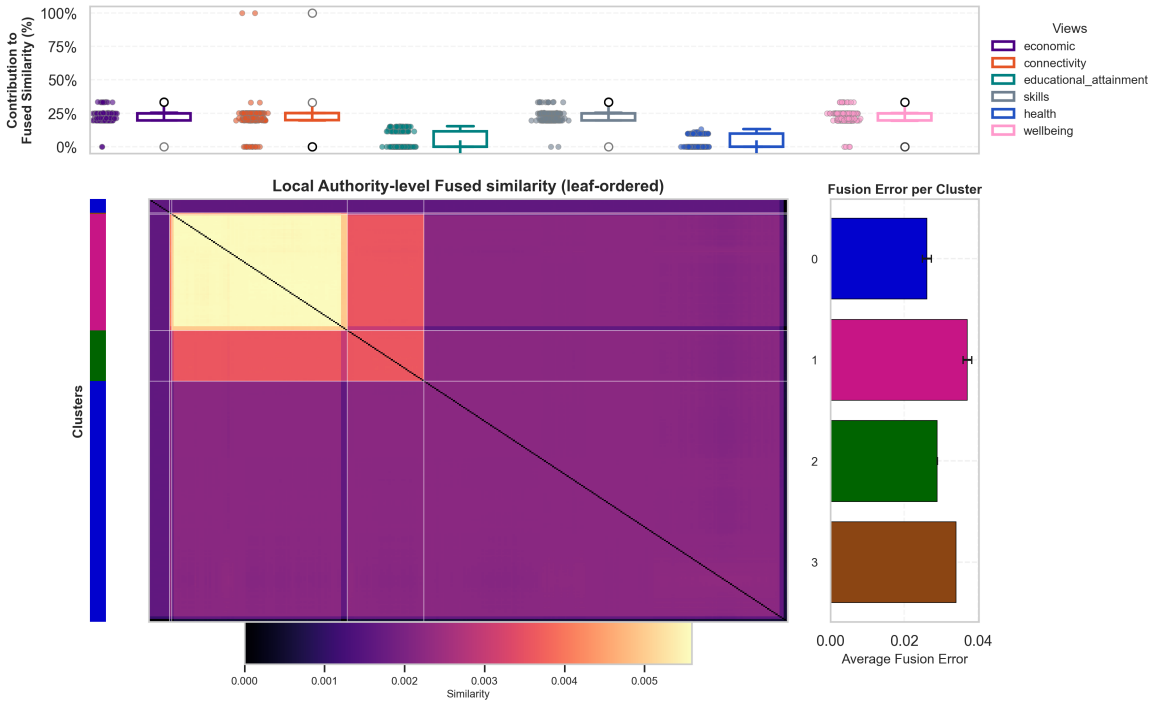

Figure 4.5.: MVSF Fusion Report: Similarity, Contributions and Errors

4.2.3. Interpretability

The MVSF's fused similarity matrix (Figure 4.5), notably the Local Authority-level fusion heatmap in the center, shows that cluster 1 is the most internally cohesive (bright block), cluster 2 has decent cohesion and importantly shows some similarity with cluster 1 (the orange off-diagonal block between 1 and 2). Clusters 0 and 3 are weak/patchy internally (darker blocks), i.e., poor cohesion. The only clear cross-cluster affinity is 1 and 2. Pairs involving 0 or 3 are largely dissimilar, both to each other and to 1 and 2. Cluster 2 has the lowest fusion error which means the views agree most for this group (easiest to reconcile in fusion) while clusters 1 and 3 have the highest fusion error, indicating that even if 1 is cohesive in the fused space, the views disagree about its relationships (fusion is working harder/compromising), and 3 looks both incohesive and hard to explain.

The view contributions to the fused similarity (top of Figure 4.5) show all views but educational attainment and health, each contribute equally to the fusion. However, we have noticed that in the unweighted fusion step, sparse but high-signal views like education and health can exert strong localised influence where data exists, while having no effect where data is missing. Conversely, denser but less-aligned views like connectivity can introduce structure orthogonal to the main socioeconomic gradient, thereby reducing overall cohesion. Also, uniform signal hurts clustering, like widespread 4G coverage, high (90+%) and fairly uniform across virtually all LAs, it produces cross-cluster links that blur boundaries.

In ablation, removing the connectivity view slightly reduces inter-cluster similarity, so silhouette goes up, but when we dive deeper and look at the structure of the similarities in graph form (Figure 4.6), it is a dominant view. This nuance indicate that connectivity may be acting as

a source of noise, with redundant signal already captured by other views (notably economic), measurement or low variance, or the fact that connectivity captures dimensions orthogonal to the socioeconomic axis that underpins the main partition.

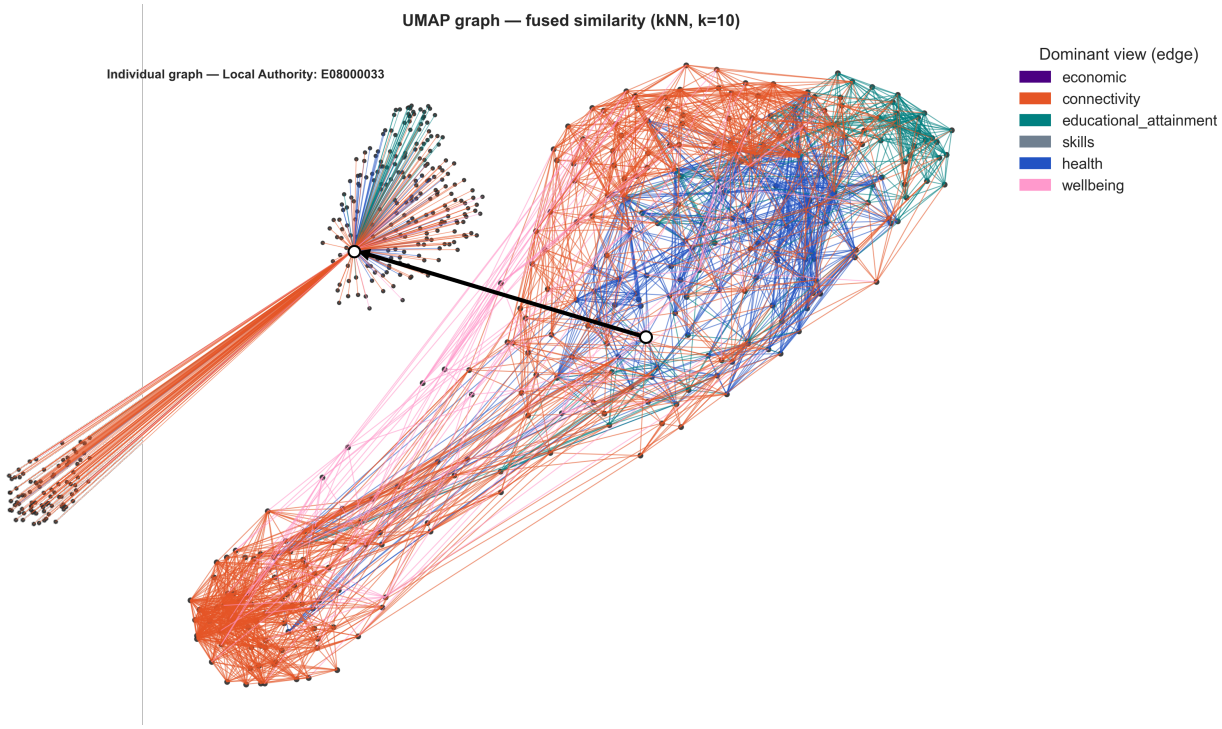

Figure 4.6.: Local Authority-level Similarity Graph (global and individual)

4.3. Summary of MVC Benefits and Findings

From a methodological standpoint, MVSF ensures that pairwise similarities are computed only over shared observed features, avoiding bias from imputation. However, this also means that the similarity basis differs across LA pairs, leading to heterogeneous and potentially underpowered similarity.

Taken together, the findings suggest that missingness does not take away informativeness, in fact, partial views may still capture distinctive structure that strengthens clustering for subsets of the data. However, views that are misaligned or redundant may degrade performance, particularly when fusion treats all views equally regardless of density, noise, or relevance.

In summary, MVSF balances high clustering performance, transparent interpretability, and robustness to missingness. While some evidence suggests missing data influences cluster assignments and the high silhouette score might be inflated by masking, the overall solution is not driven by missingness alone. Its fused similarity representation and per-view diagnostics make it a strong candidate for policy-relevant segmentation of social data, especially when handling incomplete multi-view inputs.

Chapter 5. Discussions & Conclusions

5.1. Discussions

5.1.1. Summary of Key Insights

MVSF's performance. Multi-View Similarity Fusion (MVSF) attains the highest peak silhouette in our grid, on the incomplete (non-imputed) dataset, well above the ONS single-view baseline and other multi-view contenders. It shows a wider spread than alternatives, consistent with fusion potentially amplifying or oversmoothing neighbourhood structure. Targeted diagnostics suggest the result is not merely an artefact of missingness: (i) silhouettes stratified by missingness quartiles are broadly consistent; (ii) removing the two most-missing views (Education, Health) lowers the peak by only ~ 0.02 combined; (iii) fusion error is low and fairly uniform across clusters. The solution yields four interpretable groups (least and most deprived areas, a large mixed cluster, and a singleton), with canonical profiles (income, education, life expectancy) supporting face validity.

Overall performance of MVC methods. Autoencoders (MA, MBASB, AECC) optimise reconstruction under masking, not cluster separation. With $n = 309$ and heterogeneous missingness, tight bottlenecks underfit while wider ones share noise across views; MBASB's equal-weight mean fusion and single decoder can bias learning toward denser/dominant views; AECC hardens each view to labels before fusion (co-association C ignores within-cluster distances and has high variance when co-occurrence m_{ij} is small). Matrix factorisation (MNMF, PVNMF, CFE) minimises Euclidean reconstruction with nonnegativity, favouring parts that explain variance/mass over decision boundaries; early fusion can overweight high-variance blocks (MNMF), while intermediate fusion depends on per-view ranks and concatenation width (PVNMF), which can blur distinctions in subsequent PCA. MVSF works directly at the pairwise level: masked cosine builds view-wise neighbourhoods from overlapping features; mean fusion aggregates these neighbourhoods; kernel PCA operates on the fused similarity. This avoids the reconstruction–clustering objective gap, is scale-invariant per pair, explaining the high peak alongside greater sensitivity to (d, k).

The value of transparency and interpretability. Beyond its performance, MVSF is attractive because it is transparent. Fusion is parameter-free (simple mean); every entry of the fused similarity can be audited back to per-view or even per-LA contributions; most diagnostics we have presented are straightforward to compute and explain. The embedding step is standard (centering, eigendecomposition), and reporting seeds make it deterministic. By contrast, autoencoders and factorisations introduce more knobs (architectures, ranks, regularisers) and optimise objectives that are only indirectly aligned with cluster separation.

MVSF offers the best balance of accuracy and clarity in our setting: it reaches the top silhouette, remains interpretable and auditable, and scales to incomplete multi-view data without feature imputation. Its main caveat is sensitivity to downstream parameters but practical guardrails can be put in place to mitigate (see subsection 5.2.3).

5.1.2. Limitations of Current Work

Our study has several limitations, we highlight the most salient trade-offs we made in Table 5.1.

Table 5.1.: Limitations of Current Work (expressed as trade–offs) summary

Limitation	Description
Comparability over outcome validity.	We selected models by silhouette and face-validity via canonical profiles rather than outcome-based validation, over a modest grid, which risks metric dependence and optimistic selection. Complementary internal criteria (e.g., Calinski–Harabasz, Davies–Bouldin), stability-based selection, and external proxies would give a more robust picture. Preprocessing choices (Winsorisation, standardisation, PCA settings, nonnegativity shifts for NMF) were fixed a priori to mirror ONS and were not exhaustively stress-tested; alternative scalings/weightings could change results. Findings are specific to 2021 indicators and LA boundaries; shifts in definitions, measurement error, or indicator availability limit transportability over time or geography.
Breadth over depth of methods.	We emphasised transparent, comparable baselines and did not include several high-performing modern MVC methods: diffusion/propagation graph fusion (e.g., SNF), learned view weighting/attention or multiple-kernel learning, clustering-aware/contrastive deep multi-view models, and probabilistic/Bayesian multi-view models (e.g., Variational AE variants). These often lift performance by learning view reliability and sharpening boundaries, but add substantial hyperparameters and training complexity; we defer them to future work.
Clarity over sophistication in design.	To keep pipelines auditable, we used equal-weight fusion, different k priors or stability/model-based selection could yield different segmentations. Equal-weight MVSF can amplify consistent but spurious structure and underweight informative sparse views; AE underperformance likely reflects reconstruction–clustering objective mismatch and limited budgets; for matrix factorisation, chosen ranks/concatenation can bias view contributions; we did not add graph/cluster regularisation that could improve separability
Efficiency over robustness to stochasticity and perturbations.	Several components are stochastic (AE initialisation/optimisation, K-means). Results reported with fixed seeds may overstate stability; likewise, segmentations may vary under bootstrap resampling or mild preprocessing changes. Averaging over multiple seeds, reporting dispersion (mean $\pm$sd), and perturbation/stability tests (e.g., ARI/Jaccard under resamples) would strengthen reliability.

5.1.3. Ethical and Policy Considerations

During the course of this work, we only use official publicly available releases under the appropriate Open Government Licence (OGL) and make all implementations publicly available for auditability and reproducibility (see Appendix B). All indicators are at Local Authority (LA) level with no identifiable personal data.

The mechanisms of the methods we presented mitigate imputation bias but do not remove bias if data are not missing at random. Coverage differences across views can shape clusters so monitoring for systemic under/over-representation of specific regions or population groups when collecting data for indicators is paramount. Conversely, cluster labels can stigmatise areas. This

work uses neutral labels, in line with the English Indices of Multiple Deprivation (IMD, also used by the ONS) and avoid pejorative language when describing identified clusters.

We would like to advise readers and users to treat clusters as descriptive evidence to inform deliberation; require domain expert review before any operational decision. We explicitly prohibit decisions about individuals, communities or resource allocation based on our findings or cluster memberships.

5.2. Conclusion & Future Work

5.2.1. Recap of Research Contributions

This work's contributions against each research question are shown in Table 5.2.

Table 5.2.: Synthesis of findings by research question

Research question	Answer
How do current single-view and naive integration clustering methods perform on real-world social datasets in terms of performance against common metrics, stability, and interpretability?	Across nine methods, eight exceeded the ONS claimed silhouette on at least one (d, k) setting, and three did so consistently. Performance, however, was uneven: scores varied noticeably across d_{PCA} and k. Overall, naive baselines can be competitive on peaks but are less transparent than our multi-view approaches.
Can MVL approaches provide more stable and meaningful segmentations of local authorities compared to traditional methods?	Partly. Multi-View Similarity Fusion (MVSF) reached higher peaks than the ONS baseline yet, like other methods, showed sensitivity to d_{PCA} and k, indicating limited stability. With no accepted ground truth for "meaningful" groupings, we cannot claim that MVL is categorically more stable or more meaningful. Its value lies elsewhere: MVL handles heterogeneous missingness without feature imputation, makes per-view contributions auditable, and supports transparent diagnostics.
How do different methods handle noise and missingness, and how do these factors affect the quality of the resulting clusters?	Masked pairwise approaches use only shared, observed coordinates, avoiding ad-hoc imputation; cohesion remains broadly consistent across missingness strata. In MVSF, removing the two most-missing views changes the peak silhouette by only ≈ 0.02, suggesting missingness is influential but not dominant. Reconstruction-based methods can share noise across features/views, and factorisations can overweight high-variance blocks.
To what extent can MVL approaches enhance interpretability by enabling the assessment of individual view contributions to the clustering outcomes?	MVL materially improves transparency. Fused similarity exposes per-LA, per-view contributions and per-cluster fusion errors; co-association matrices quantify agreement frequencies across views; factorisation methods provide parts-based loadings and low-dimensional embeddings that aid explanation even when not top-performing. Autoencoders benefit from post-hoc diagnostics (e.g., reconstruction error by feature/LA). Together, these artefacts support accountable narratives and allow auditors to trace how each view shapes neighbourhoods and final clusters.

5.2.2. Implications for Local Authority Segmentation

Transparency, interpretability, and explainability under high public scrutiny. Local authority segmentation informs public debate and policy, methods must therefore be inspectable. In practice, this means using methods whose steps can be traced (per-view contributions to the fused similarity, cluster profiles anchored to a small set of indicators) and exposing

simple diagnostics that non-specialists can verify. The objective is not only technical validity but explainable evidence that a reasonable person can audit.

Reproducibility as the basis for trust. Reproducibility is a prerequisite for credibility. Our own difficulty in fully replicating the ONS results reflects gaps in disclosed preprocessing and settings; this is not a criticism of the ONS's statistical expertise, but a reminder that even seasoned teams benefit from stronger reproducibility practices.

Segmentation as a framework, not a single method. Public-sector segmentation must be an operational framework that evolves with the data and policy context. Some practical implications include explaining sources of missingness and flagging cases where missingness is likely not at random, or pairing quantitative segments with expert context (e.g., local knowledge, Levelling Up documentation), and record decision points so methodological choices are transparent.

5.2.3. Future Research Directions

As next steps, we prioritise robustness, transparency, and policy usefulness.

- **Robustness beyond a single metric.** We could stress–test preprocessing and selection by running scaling/weighting sweeps, multi-criteria model selection (Silhouette, Calinski–Harabasz, Davies–Bouldin), and complement internal metrics with external outcomes/proxies (e.g., IMD, health/education targets) and year-on-year temporal tests.
- **Adaptive fusion and view weighting (MVSF extensions).** We could replace equal-weight averaging with (i) learned view weights and (ii) robust weights to curb dominance/misalignment and improve resilience to noisy/sparse views. When revisiting autoencoders, couple masked reconstruction with clustering-aware terms to narrow the reconstruction–clustering objective gap.
- **Broader method families.** Add diffusion/propagation fusion (e.g., SNF-style iterative kernels), probabilistic/Bayesian multi-view models for uncertainty quantification, and graph/manifold approaches, comparing them under the same transparency and reporting standards.
- **Handling missingness and coverage.** Audit and model coverage explicitly: for example, by testing sensitivity to synthetic missingness/masking or multiple-imputation comparators to assess MNAR risk.
- **Uncertainty at the LA level.** Publish per-LA confidence summaries (sample silhouette, co-association strength, overlap counts) with flagging rules for borderline or low-coverage cases; provide bootstrap intervals for cluster assignments where feasible.

References

Baltrusaitis, T., Ahuja, C., and Morency, L.-P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 41(2):423–443. Available at https://people.ict.usc.edu/~gratch/CSCI534/Readings/Baltrusaitis-MMML-survey.pdf. Accessed: 2025-07-15.

Bivand, R. S., Pebesma, E., and Gómez-Rubio, V. (2013). *Applied Spatial Data Analysis with R*. Springer, New York, 2 edition.

Blei, D. M., Ng, A., and Jordan, M. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022. Available at https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf. Accessed: 2025-07-29.

Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27. Available at https://www.tandfonline.com/doi/abs/10.1080/03610927408827101. Accessed: 2025-07-05.

Chao, G., Sun, S., and Bi, J. (2021). A survey on multi-view clustering. *IEEE Transactions on Artificial Intelligence*, 2(2):146–168. Available at https://ieeexplore.ieee.org/document/9395530. Accessed: 2025-07-22.

Chowdhury, A. R., Gupta, A., and Das, S. (2025). Deep multi-view clustering: A comprehensive survey of the contemporary techniques. *Information Fusion*, 119. Available at https://www.sciencedirect.com/science/article/abs/pii/S1566253525000855. Accessed: 2025-07-15.

Cruickshank, I. and Carley, K. M. (2020). Characterizing communities of hashtag usage on twitter during the 2020 covid-19 pandemic by multi-view clustering. *Applied Network Science*, 5(66):1–20. Available at https://appliednetsci.springeropen.com/articles/10.1007/s41109-020-00317-8. Accessed: 2025-07-18.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2):224–227. Available at https://dl.acm.org/doi/10.1109/TPAMI.1979.4766909. Accessed: 2025-07-05.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39:1–38. Available at https://www.ece.iastate.edu/~namrata/EE527_Spring08/Dempster77.pdf. Accessed: 2025-07-17.

Ding, C., Li, T., Peng, W., and Park, H. (2006). Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 126 – 135. Available at https://dl.acm.org/doi/10.1145/1150402.1150420. Accessed: 2025-08-02.

Dixon, W. J. (1950). Analysis of extreme values. *The Annals of Mathematical Statistics*, 21(4):488–506. Available at https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-21/issue-4/Analysis-of-Extreme-Values/10.1214/aoms/1177729747.full. Accessed: 2025-07-05.

Donath, W. and Hoffman, A. (1972). Algorithms for partitioning of graphs and computer logic based on eigenvectors of connection matrics. Technical Report 3, IBM Technical Disclosure Bulletin. Available at https://www.kde.cs.uni-kassel.de/mediawiki/images/7/70/Ibm_tdb_15-3_1972_donath-hoffman.pdf. Accessed: 2025-07-24.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, pages 226–231. Available at https://file.biolab.si/papers/1996-DBSCAN-KDD.pdf. Accessed: 2025-06-28.

Fang, U., Li, M., Li, J., Gao, L., Jia, T., and Zhang, Y. (2023). A comprehensive survey on multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12350–12368. Available at https://dl.acm.org/doi/10.1109/TKDE.2023.3270311. Accessed: 2025-07-21.

Fred, A. L. N. and Jain, A. K. (2005). Combining multiple clusterings using evidence accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):835–850. Available at http://ieeexplore.ieee.org/document/1432715. Accessed: 2025-07-15.

Gao, J., Li, P., Chen, Z., and Zhang, J. (2020). A survey on deep learning for multimodal data fusion. *Neural Computation*, 32(5):829–864. Available at https://direct.mit.edu/neco/article/32/5/829/95591/A-Survey-on-Deep-Learning-for-Multimodal-Data. Accessed: 2025-07-25.

Gaw, N., Yousefi, S., and Reisi Gahrooei, M. (2021). Multimodal data fusion for systems improvement: A review. *IISE Transactions*, 54(11):1098–1116. Available at https://www.tandfonline.com/doi/abs/10.1080/24725854.2021.1987593. Accessed: 2025-07-29.

Gondara, L. and Wang, K. (2018). Mida: Multiple imputation using denoising autoencoders. *arXiv preprint*. Available at https://arxiv.org/abs/1705.02737. Accessed: 2025-07-15.

Grimmer, J. and Stewart, B. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297. Available at https://www.cambridge.org/core/journals/political-analysis/article/text-as-data-the-promise-and-pitfalls-of-automatic-content-analysis-methods-for-political-F7AAC8B2909441603FEB25C156448F20. Accessed: 2025-08-02.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics, New York, 2 edition.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1):193–218. Available at https://link.springer.com/article/10.1007/BF01908075. Accessed: 2025-07-05.

Inglehart, R. and Welzel, C. (2005). *Modernization, Cultural Change and Democracy: The Human Development Sequence*. Cambridge University Press. Available at https://www.cambridge.org/core/books/modernization-cultural-change-and-democracy/4321210B04C63808615846DB0E3EEC34. Accessed: 2025-05-01.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323. Available at https://dl.acm.org/doi/10.1145/331499.331504. Accessed: 2025-07-05.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254. Available at https://www.cambridge.org/core/journals/psychometrika/article/abs/hierarchical-clustering-schemes/DFD7B28D0F1DF0348D9C729793CC1ACD. Accessed: 2025-06-24.

Jolliffe, I. and Cadima, J. (2016). *Principal Component Analysis*. Springer, 2nd edition. Available at https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202. Accessed: 2025-07-05.

Kumar, A. and Daumé III, H. (2011). A co-training approach for multi-view spectral clustering. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 393–400. Omnipress. Available at https://dl.acm.org/doi/10.5555/3104482.3104532. Accessed: 2025-07-20.

Kumar, A., Rai, P., and Daume, H. (2011). Co-regularized multi-view spectral clustering. *Proceedings of the 25th International Conference on Neural Information Processing Systems*, pages 1413–1421. Available at https://dl.acm.org/doi/10.5555/2986459.2986617. Accessed: 2025-07-25.

Lee, D. and Seung, S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791. Available at https://www.nature.com/articles/44565. Accessed: 2025-07-29.

Li, S. and Tang, H. (2024). Multimodal alignment and fusion: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (preprint)*. Available at https://arxiv.org/pdf/2411.17040. Accessed: 2025-07-25.

Li, Y., Yang, M., and Zhang, Z. (2016). A survey of multi-view representation learning. *arXiv preprint*. Available at https://arxiv.org/pdf/1610.01206. Accessed: 2025-07-25.

Liu, J., Wang, C., Gao, J., and Han, J. (2013). Multi-view clustering via joint nonnegative matrix factorization. *Proceedings of the 2013 SIAM International Conference on Data Mining*,

pages 252–260. Available at https://epubs.siam.org/doi/10.1137/1.9781611972832.28. Accessed: 2025-07-25.

Liu, X., Zhu, X., Li, M., Wang, L., Tang, C., Yin, J., Shen, D.-g., Wang, H., and Gao, W. (2018). Late fusion for multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(10):2410–2423. Available at https://pmc.ncbi.nlm.nih.gov/articles/PMC6494716/. Accessed: 2025-07-18.

Lundberg, I., Brand, J., and Jeon, N. (2022). Researcher reasoning meets computational capacity: Machine learning for social science. *Social Science Research*, 108:102807. Available at https://www.sciencedirect.com/science/article/abs/pii/S0049089X22001181. Accessed: 2025-05-08.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, 5(1):281–297. Available at https://projecteuclid.org/euclid.bsmsp/1200512992. Accessed: 2025-07-05.

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint*. Available at https://arxiv.org/abs/1802.03426. Accessed: 2025-07-15.

Newman, M. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physiscal Review E*. Available at https://journals.aps.org/pre/abstract/10.1103/PhysRevE.69.026113. Accessed: 2025-07-30.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. (2011). Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*. Available at https://dl.acm.org/doi/10.5555/3104482.3104569. Accessed: 2025-07-23.

Office for National Statistics (2022). Subnational indicators dataset. https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/datasets/subnationalindicatorsdataset/december2022. Accessed: 2025-07-19.

Office for National Statistics (2023a). Clustering local authorities against subnational indicators, england. https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/articles/clusteringlocalauthoritiesagainstsubnationalindicatorsengland/latest. Accessed: 2025-07-19.

Office for National Statistics (2023b). Clustering local authorities against subnational indicators methodology. https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/methodologies/clusteringlocalauthoritiesagainstsubnationalindicatorsmethodology. Accessed: 2025-07-19.

Office for National Statistics (2023c). Clusters labels - clustering local authorities against subnational indicators, england. https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/datasets/clusteringlocalauthoritiesagainstsubnationalindicatorsengland. Accessed: 2025-07-19.

Office for National Statistics (2024a). Clustering similar local authorities in the uk, methodology. https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/methodologies/clusteringsimilarlocalauthoritiesintheukmethodology. Accessed: 2025-07-19.

Office for National Statistics (2024b). Local authority districts (december 2023) boundaries uk bgc. https://www.data.gov.uk/dataset/31c78354-168f-4457-b8f3-80e86adcd0f4/local-authority-districts-december-2023-boundaries-uk-bgc. Accessed: 2025-07-25.

Orme, E., Rodosthenous, T., and Evangelou, M. (2025). Multi-view biclustering via nonnegative matrix tri-factorization. *arXiv preprint*. Available at https://arxiv.org/abs/2502.13698. Accessed: 2025-07-12.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine and Journal of Science*, 2:559–572. Available at https://www.tandfonline.com/doi/abs/10.1080/14786440109462720. Accessed: 2025-08-01.

Rappoport, N. and Shamir, R. (2018). Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Research*, 46(20):10546–10562. Available at https://academic.oup.com/nar/article/46/20/10546/5123392?login=false. Accessed: 2025-07-29.

Rezankova, H. (2014). Cluster analysis of economic data. *Statistika: Statistics and Economy Journal*, 94(1):65–74. Available at https://csu.gov.cz/docs/107508/f9af9161-d47d-7450-23b0-dfaad4eba77f/32019714q1073.pdf?version=1.0. Accessed: 2025-07-01.

Rodosthenous, T., Shahrezaei, V., and Evangelou, M. (2024). Multi-view data visualisation via manifold learning. *PeerJ Computer Science*, 10. Available at https://peerj.com/articles/cs-1993/. Accessed: 2025-07-15.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65. Available at https://www.sciencedirect.com/science/article/pii/0377042787901257. Accessed: 2025-07-05.

Shapiro, B. and Battle, A. (2024). Bayesian multi-view clustering given complex interview structure. *F1000Research*, 11. Available at https://f1000research.com/articles/11-1460/v2. Accessed: 2025-07-15.

Sinha, P., Calfee, C., and Delucchi, K. (2021). Practitioner's guide to latent class analysis: Methodological considerations and common pitfalls. *Critical Care Medicine*, 49(1):e63–e79. Available at https://pubmed.ncbi.nlm.nih.gov/33165028/. Accessed: 2025-05-18.

Snyder, J. P. (1987). *Map Projections–A Working Manual*, volume 1395 of *US Geological Survey Professional Paper*. US Government Printing Office.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101. Available at https://www.jstor.org/stable/pdf/1412159.pdf. Accessed: 2025-07-21.

Strehl, A. and Ghosh, J. (2002). Cluster ensembles – a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617. Available at https://www.jmlr.org/papers/volume3/strehl02a/strehl02a.pdf. Accessed: 2025-07-23.

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46(sup1):234–240. Available at https://www.jstor.org/stable/143141. Accessed: 2025-07-27.

UK Government (2022). Levelling up the united kingdom white paper. https://assets.publishing.service.gov.uk/media/61fd3ca28fa8f5388e9781c6/Levelling_up_the_UK_white_paper.pdf. Accessed: 2025-07-19.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605. Available at https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf. Accessed: 2025-07-29.

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*. Available at https://www.cs.toronto.edu/~larocheh/publications/icml-2008-denoising-autoencoders.pdf. Accessed: 2025-07-23.

Vinh, N. X., Epps, J., and Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11:2837–2854. Available at https://jmlr.csail.mit.edu/papers/volume11/vinh10a/vinh10a.pdf. Accessed: 2025-07-05.

Waggoner, P. (2020). *Unsupervised Machine Learning for Clustering in Political and Social Research*. Cambridge University Press. Available at https://www.cambridge.org/core/books/unsupervised-machine-learning-for-clustering-in-political-and-social-research/BF62D1E8F6DB3237D5CE524FBFCBA33A. Accessed: 2025-05-05.

Wang, B., Mezlini, A., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., and Golden-

berg, A. (2014a). Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, 11:333–337. Available at https://www.nature.com/articles/nmeth.2810. Accessed: 2025-07-12.

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., and Goldenberg, A. (2014b). Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, 11(3):333–337. Available at https://www.nature.com/articles/nmeth.2810. Accessed: 2025-07-22.

Wang, W., Arora, R., Livescu, K., and Bilmes, J. (2015). On deep multi-view representation learning. In Bach, F. and Blei, D., editors, *International Conference on Machine Learning (ICML)*, volume 37 of *Proceedings of Machine Learning Research*, pages 1083–1092. PMLR. Available at https://proceedings.mlr.press/v37/wangb15.html. Accessed: Accessed: 2025-07-20.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press. Available at https://www.cambridge.org/core/books/social-network-analysis/90030086891EB3491D096034684EFFB8. Accessed: 2025-07-01.

Wen, J., Zhang, Z., Fei, L., Zhang, B., Xu, Y., Zhang, Z., and Li, J. (2023). A survey on incomplete multi-view clustering. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 53(2):1136–1149. Available at https://ieeexplore.ieee.org/document/9845473. Accessed: 2025-07-22.

Xu, C., Tao, D., and Xu, C. (2013). A survey on multi-view learning. *arXiv preprint*. Available at https://arxiv.org/abs/1304.5634. Accessed: 2025-07-18.

Yan, X., Hu, S., Mao, Y., Ye, Y., and Yu, H. (2021). Deep multi-view learning methods: A review. *Neurocomputing*, 448:106–129. Available at https://www.sciencedirect.com/science/article/abs/pii/S0925231221004768. Accessed: 2025-07-23.

Yang, Y. and Wang, H. (2018). Multi-view clustering: A survey. *Big Data Mining and Analytics*, 1(2):83–107. Available at https://ieeexplore.ieee.org/document/8336846. Accessed: 2025-07-22.

Yin, J. and Sun, S. (2022). Incomplete multi-view clustering with cosine similarity. *Pattern Recognition*, 123. Available at https://www.sciencedirect.com/science/article/abs/pii/S0031320321005513. Accessed: 2025-07-20.

Yu, Z., Dong, Z., Yu, C., Yang, K., Fan, Z., and Chen, P. (2024). A review on multi-view learning. *Frontiers of Computer Science*, 19(197334):106–129. Available at https://link.springer.com/article/10.1007/s11704-024-40004-w. Accessed: 2025-07-25.

Zakharov, K. (2016). Application of k-means clustering in psychological studies. *The Quantitative Methods for Psychology*, 12(2):87–100. Available at

https://www.researchgate.net/publication/308014650_Application_of_k-means_clustering_in_psychological_studies. Accessed: 2025-05-07.

Zhang, H. and Peng, Y. (2022). Image clustering: An unsupervised approach to categorize visual data in social science research. *Sociological Methods & Research*, 53(3). Available at https://journals.sagepub.com/doi/abs/10.1177/00491241221082603. Accessed: 2025-07-21.

Zhao, F., Zhang, C., and Geng, B. (2024). Deep multimodal data fusion. *ACM Computing Surveys*, (216):1–36. Available at https://dl.acm.org/doi/full/10.1145/3649447. Accessed: 2025-07-25.

Zhao, H., Ding, Z., and Fu, Y. (2017). Multi-view clustering via deep matrix factorization. *Association for the Advancement of Atificial Intelligence*. Available at https://ojs.aaai.org/index.php/AAAI/article/view/10867/10726. Accessed: 2025-07-15.

Zhou, D. and Burges, C. (2007). Spectral clustering and transductive learning with multiple views. In *Proceedings of the 24th International Conference on Machine Learning (ICML)*, pages 1159–1166. DBLP. Available at https://www.researchgate.net/publication/221346148_Spectral_clustering_and_transductive_learning_with_multiple_views. Accessed: 2025-07-18.

Supplementary Matter

A. Term Definitions and Glossary

Table 3.: Term definitions used in methodological formulations

Symbol	Definition
M	Binary mask ($M_{ij} = 1$ if entry observed, 0 otherwise).
$M^{(v)}$	Mask for view v.
$\odot$	Hadamard (element-wise) product.
$n,\ d_v$	Number of samples; number of features in view v.
V	Number of views.
$X^{(v)}$	Data matrix for view v.
$S^{(v)}$	Per-view *similarity* matrix.
S_{fused}	Fused similarity (e.g., mean of $S^{(v)}$).
D	Diagonal matrix of row sums used for row-normalisation.
$K,\ K_c$	(Centred) Gram/kernel matrix used for embedding.
H_c	PCA centering matrix $I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$.
$\mathbf{1}$	All-ones vector of length n.
$U,\ \Lambda$	Eigenvectors and eigenvalues from $K = U\Lambda U^\top$.
Z	PCA embeddings (scores); with superscripts: $Z^{(v)}$ (per-view), Z^{cons} (consensus).
L	Latent dimension (AE bottleneck); $d_{\text{pv}}, d_{\text{cons}}$ PCA dimensions.
$k,\ k_v$	Number of clusters (consensus / per-view).
$\ell_i^{(v)}$	Cluster label of sample i in view v (-1 if missing in that view).
C	Co-association (label-derived similarity) matrix.
$m_{ij},\ \mathcal{V}_{ij}$	Co-occurrence count and set of views where i and j both appear.
$\mathbb{1}[\cdot]$	Indicator function.
Factorisation-specific	
$G^{(v)}, F^{(v)}$	NMF row/column factors; $X^{(v)} \approx G^{(v)} {F^{(v)}}^\top$.
$\mathcal{S}^{(v)}$	NMTF middle map (not to confuse with similarity $S^{(v)}$); $X^{(v)} \approx G^{(v)}\,\mathcal{S}^{(v)}\,{H^{(v)}}^\top$.
$H^{(v)}$	NMTF column factor (features $\times\ \ell_v$).
G_{concat}	Column-concatenated matrix $[\,G^{(1)}\vert\cdots\vert G^{(V)}\,]$.
R	Total columns in G_{concat} ($R = \sum_v r_v$).
Autoencoder-specific	
$f_\theta^{(v)},\ g_\phi$	Per-view encoders; single decoder (MBASB/MA).
$h_i^{(v)}$	Encoder output for sample i in view v.
$H^{(v)}$	Per-view AE bottleneck matrix ($n \times r$) in AECC;
H	Shared bottleneck in MBASB: $H = \frac{1}{V}\sum_v h^{(v)}$ (stacked over samples).
ε	Small constant for numerical stability (e.g., in normalisers).

Table 4.: Glossary of Acronyms and Terms

Term / Acronym	Definition
AE (Autoencoder)	Neural network that learns a compressed latent representation by reconstructing inputs. Used as a building block in MA/MBASB/AECC.
AECC	Autoencoder Ensemble with Consensus Clustering: per-view masked autoencoders; cluster each view; combine labels via co-association to a consensus.
ARI	Adjusted Rand Index: label-invariant clustering agreement metric adjusted for chance.
APS	Annual Population Survey: UK survey source used in ONS indicators.
BIC	Bayesian Information Criterion: model selection criterion used in clustering contexts.
CFE	Co-Association Factorization Ensemble: late-integration consensus built from per-view NMTF partitions.
CSE	Co-association Clustering/Consensus: builds a pairwise co-membership matrix from per-view cluster labels and clusters its PCA embedding.
DBSCAN	Density-Based Spatial Clustering of Applications with Noise: density-based clustering algorithm.
EM	Expectation–Maximization: iterative procedure for maximum-likelihood estimation (e.g., GMM).
GCSE	General Certificate of Secondary Education: UK exam at end of compulsory schooling.
GMM	Gaussian Mixture Model: probabilistic clustering with Gaussian components (typically fit with EM).
Haversine (great-circle)	Formula to compute distances on a sphere from lat/long; used for spatial k-NN distances.
IMD	Index of Multiple Deprivation: measures of relative deprivation across each of the constituent areas of the United Kingdom.
IQR	Interquartile Range: robust dispersion measure used in NaN-aware scaling.
K-means	Centroid-based clustering minimizing within-cluster variance on an embedding.
KNN / k-NN	k-Nearest Neighbours; also used for graph construction in SNF and for spatial distance summaries.
KL divergence (DKL)	Kullback–Leibler divergence: information-theoretic contrast used for feature ranking.
KS2	Key Stage 2: primary-school assessment in England.
LA	Local Authority: UK local administrative unit (England focus in report).
MA	Masked Autoencoder: early-integration AE trained with a per-row masked reconstruction loss; PCA + K-means on latent codes.
MBASB	Multibranch Autoencoder with Shared Bottleneck: per-view encoders, mean-fused bottleneck, single decoder; PCA + K-means.
MNAR	Missing Not At Random: data missing is systematically related to the unobserved data.
MNMF	Masked Nonnegative Matrix Factorization: NMF with an observation mask; PCA + K-means on sample factors.
MVC	Multi-View Clustering: clustering with multiple views/modalities.
MVL	Multi-View Learning: learning framework leveraging multiple views.
MVSF	Multi-View Similarity Fusion: simple (parameter-free) average of per-view similarity matrices; kernel-PCA + K-means.
NMF	Nonnegative Matrix Factorization: parts-based factorization ($X \approx GF^T$) yielding additive components.
NMI	Normalised Mutual Information: label-invariant clustering agreement metric.
NMTF	Nonnegative Matrix Tri-Factorization: $X \approx GSH^T$, yielding simultaneous row/column clusters.
NNDSVDa	Nonnegative Double Singular Value Decomposition (variant 'a'): common initialization for NMF/NMTF.
ONS	Office for National Statistics: UK statistical agency; source of the single-view baseline.
PCA	Principal Component Analysis: linear dimensionality reduction on covariance/kernel matrices.
PSD	Positive Semidefinite: property required for valid kernels; relevant when centring/symmetrising co-association or fused similarities.
PVNMF	Per-View NMF with Concatenation: NMF per view, concatenate row factors, PCA + K-means.
Queen contiguity	Spatial adjacency rule where regions are neighbours if they share a boundary or vertex (libpysal).
ReLU	Rectified Linear Unit activation, $\mathrm{ReLU}(t) = \max(0, t)$, used in AE encoders/decoders.
SNF	Similarity Network Fusion: iterative diffusion-based fusion of per-view KNN graphs (contrasted with MVSF).
SVL	Single-View Learning: learning from one view/domain at a time.
t-SNE	t-distributed Stochastic Neighbor Embedding: nonlinear DR for visualisation (2–3D).
UMAP	Uniform Manifold Approximation and Projection: nonlinear DR preserving local/global structure.
Silhouette score	Cluster validity index comparing within-cluster cohesion to nearest-cluster separation.

B. Reproducibility Statement

Everything necessary to reproduce this project can be found at: https://github.com/MaxLabarre/imperial_Multi_View_Clustering.git

Overview

This project is fully reproducible from the provided notebooks and utility modules. The workflow is:

1. **Datasets_Wrangling_LocalIndicators.ipynb** (builds all processed views: non-imputed, imputed, and a Spatial view).
2. **Multi_Modal_Extension_Spatial_Analysis.ipynb** (the analysis of the Spatial aspects of Local Authorities, detailed in Appendix F1).
3. **ONS_Replication.ipynb** (replicates the ONS single-view pipeline on non-imputed views).
4. **ONS_Replication - IMPUTED.ipynb** (same pipeline on the imputed views).
5. **Dimensionality_Reduction_Clustering_GridSearch.ipynb** (SVL grid search over DR $\times$ clustering).
6. **Multi_View_Learning_Experiment.ipynb** (core multi-view methods).
7. **Multi_View_Learning_Experiment - SPATIAL.ipynb** (same as above, with the Spatial view included).

All stochastic components are seeded; we use a global `random_state = 19042022` throughout and control threading to ensure determinism.

Computing Environment

Core Python packages (typical) `numpy`, `scipy`, `pandas`, `scikit-learn`, `umap-learn`, `matplotlib` (and optionally `seaborn` for EDA); spatial stack for the Spatial view: `geopandas`, `pyproj`, `shapely`, `libpysal`, (`rtree` optional). If you run the optional autoencoder variants, install your chosen DL backend (e.g., Keras/TensorFlow or PyTorch) and enable library-level determinism.

Deterministic threading (recommended) Set single-threaded determinism during numerically sensitive steps:

```
export OMP_NUM_THREADS=1
export MKL_NUM_THREADS=1
export OPENBLAS_NUM_THREADS=1
export NUMEXPR_NUM_THREADS=1
```

Data Dependencies & Layout

Raw inputs

- ONS indicator tables per view (as used by the ONS baseline) available at: https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/datasets/subnationalindi december2022
- ONS cluster labels available at: https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/datasets/clusteringlocal
- ONS imputation mappings available: https://www.data.gov.uk/dataset/ccacfb62-302e-4655-915c-176727e68ff8/local-authority-d
- 2021 Local Authority (LA) boundaries GPKG package available at: https://www.data.gov.uk/dataset/31c78354-168f-4457-b8f3-80e86adcd0f4/local-authority-d

Directory layout (convention)

```
data/
  raw/                    # raw ONS indicators and LA boundaries
  Final/              # written by Datasets_Wrangling_... (parquet/csv)
notebooks/                # the six notebooks listed above
utils/                  # shared modules imported by notebooks
Pictures/                 # images produced by notebooks
Outputs/                  # labels/, metrics/, figures/, logs/
```

Global Determinism

- Global seed: `random_state = 19042022`.
- Apply `random_state` to: K-Means, PCA (where applicable), UMAP, Spectral, GMM init, NMF initialisation (e.g., NNDSVDa), and any train/validation splits.
- For K-Means use a large `n_init` (e.g., `n_init=100`) to eliminate variance from poor initialisations.

- For DL (optional AEs): set framework seeds and deterministic flags; run with a fixed number of threads.

Step 1: Build Processed Views (Wrangling)

Open and run **Datasets_Wrangling_LocalIndicators.ipynb** end-to-end.

- **Load raw indicators** and harmonise LA identifiers; apply ONS exclusions consistently across all views.
- **Winsorisation & Standardisation** (when required by a downstream pipeline): typically winsorise at 1st/99th percentiles, then Z-score standardise per feature.
- **Non-imputed vs Imputed**:
  - Non-imputed: retain original missingness (used for masked similarity/factorisation and ONS non-imputed branch).
  - Imputed: apply the imputation strategy used in the ONS baseline replication (same assumptions across views).
- **Spatial view**:
  - Boundaries: 2021 LA boundaries; compute geometries in British National Grid (EPSG:27700) for area/perimeter; keep WGS84 lat/long for great-circle distances.
  - Contiguity: Queen contiguity adjacency (binary weights) using `libpysal`.
  - Distance features: great-circle (haversine) k-NN mean distances (e.g., $k \in \{1, 3, 5\}$) from LA centroids.
  - Optional geometry summaries: area, perimeter, compactness (e.g., Polsby–Popper-like).

Step 2: ONS Single-View Replication (Non-imputed)

Open **ONS_Replication.ipynb**.

- Input: non-imputed processed views.
- Pipeline per ONS view: **Winsorise** $\rightarrow$ **Z-score** $\rightarrow$ **PCA** $\rightarrow$ **K-Means**.
- PCA selection rule: retain $\geq 25\%$ cumulative variance with a minimum of 2 components (use the same rule throughout for comparability).
- K-Means: use the ONS-specified k (or the study's fixed k per view); set `random_state=19042022` and large `n_init`.

- Outputs to `results/`: labels (CSV), PCA loadings/variance tables, K-Means diagnostics, and figures.

Open **ONS_Replication - IMPUTED.ipynb**. Repeat the identical pipeline on the imputed versions of the views. Save analogous artefacts to `results/`.

Step 3: SVL Grid Search (DR × Clustering)

Open **Dimensionality_Reduction_Clustering_GridSearch.ipynb**.

- Input: non-imputed processed views (per view, or a designated single view for SVL).
- DR candidates: PCA, UMAP, MDS.
- Clustering candidates: K-Means, Agglomerative, Spectral, Gaussian Mixture Model (GMM).
- **Suggested grids (deterministic)**:
  - PCA: $n\_components \in \{2, \ldots, 10\}$.
  - UMAP: $n\_components \in \{2, \ldots, 10\}$; set `random_state`; keep `metric` consistent (e.g., `euclidean` or `cosine`).
  - MDS (metric): $n\_components \in \{2, \ldots, 10\}$.
  - K-Means: $k \in \{4, \ldots, 15\}$, `n_init` large, `random_state`.
  - Agglomerative: $k \in \{4, \ldots, 15\}$; linkage `ward` (euclidean), `average`, `complete`.
  - Spectral: $k \in \{4, \ldots, 15\}$; set `random_state`.
  - GMM: $k \in \{4, \ldots, 15\}$; covariance type `full` or `diag`; set `random_state`.
- **Evaluation**: compute silhouette (primary). If a reference partition is available, also compute ARI/NMI. Repeat over a fixed small set of seeds to quantify stability (include 19042022).
- **Selection**: report the best (DR, clustering, hyperparameters) by average silhouette; record full grid results to CSV.
- Outputs: `results/metrics/` (CSV tables per grid), `results/labels/`, and `results/figures/`.

Step 4: Multi-View Learning (Core)

Open **Multi_View_Learning_Experiment.ipynb**. Unless otherwise stated, all methods end with PCA → K-Means for a fair, common decision layer.

- **Early integration**
  - Masked Cosine Similarity (per view) → optional centering → PCA ($d \in \{2, \ldots, 5\}$) → K-Means ($k \in \{4, \ldots, 15\}$).
  - Masked NMF (MNMF): rank $r \in \{2, \ldots, 5\}$; non-negative inputs or shifted data; init NNDSVDa; max-iter/tol as set in the notebook; PCA with $d \leq r$; K-Means $k \in \{4, \ldots, 15\}$.
- **Intermediate integration**
  - Multi-View Similarity Fusion (MVSF): average per-view (masked) similarities; kernel centering if required; kernel PCA ($d \in \{2, \ldots, 5\}$); K-Means ($k \in \{4, \ldots, 15\}$).
  - PVNMF: NMF per view (NNDSVDa, fixed seed); concatenate row factors; row-wise ℓ_2 normalisation; PCA ($d \in \{2, \ldots, 5\}$); K-Means ($k \in \{4, \ldots, 15\}$).
- **Late integration**
  - Co-association (CSE): cluster each view (fixed k_v or small grid); build pairwise co-membership matrix; symmetrise and (if needed) centre; PCA ($d \in \{2, \ldots, 5\}$); K-Means ($k \in \{4, \ldots, 15\}$).
- **Optional AE variants** (if you ran them): masked loss; fixed library seeds; then the same PCA → K-Means layer and grids as above.
- **Model selection**: choose d and k by average silhouette on the embedding used by K-Means; for MNMF also respect $d \leq r$. Persist all intermediate/final artefacts.

Step 5: Multi-View Learning with Spatial

Open **Multi_View_Learning_Experiment - SPATIAL.ipynb** and repeat the chosen method(s) including `view_spatial` alongside the socio-economic views. Use the same grids and selection rule. Save artefacts under a distinct method tag (e.g., suffix `_spatial`).

Troubleshooting & Notes

- **Numeric drift**: ensure deterministic threads set to 1; pin package versions; avoid mixing MKL/OpenBLAS builds between runs.
- **Memory**: for large grids, prefer writing intermediate CSVs per grid point and aggregating at the end.
- **Spectral/Agglomerative constraints**: `ward` linkage requires Euclidean distances; ensure compatibility with the embedding metric.
- **GMM initialisation**: fix `random_state`; try `covariance_type=diag` if `full` is unstable in low-sample regimes.
- **Reproducible figures**: set a fixed font and save with explicit DPI and tight bounding boxes.

Minimal Run Order (recap)

1. Run **Datasets_Wrangling_LocalIndicators.ipynb** → writes `data/processed/`.
2. Run **ONS_Replication.ipynb** and **ONS_Replication - IMPUTED.ipynb**.
3. Run **Dimensionality_Reduction_Clustering_GridSearch.ipynb**.
4. Run **Multi_View_Learning_Experiment.ipynb** and/or **Multi_View_Learning_Experiment - SPATIAL.ipynb**.

C. Exploratory Data Analysis Results and Visualisations

Appendix C1: Relationships found in the data

Feature-level Correlations. When exploring the relationships in the data, notably through Features' correlations (Figure 1), we find the following linear relationships:

Table 5.: High Feature-Level Correlations and Interpretations

Correlation	**Interpretation**
High Economic performance ↔ High Educational Attainment	Economically stronger areas tend to have better education systems and outcomes.
High Economic performance ↔ Low Apprenticeship and Skills Participation	Academic pathways may be preferred over vocational training in more affluent areas.
High Economic performance ↔ High Qualifications (Level 3+)	Higher education levels are commonly associated with stronger local economies.
High Economic performance ↔ High Life Expectancy	Wealthier regions generally benefit from better healthcare and living conditions.
High Economic performance ↔ Low Smoking and Obesity Rates	Healthier behaviors are more prevalent in socio-economically advantaged areas.
High Economic performance ↔ High Early Cancer Diagnosis Rates	Preventive healthcare and screenings are more accessible in affluent regions.
High Economic performance ↔ Low Mortality Rates	Better healthcare access and living standards reduce premature mortality.
High Connectivity ↔ High Apprenticeship Starts and Achievements	Improved transport infrastructure supports access to vocational education.
High GCSE Attainment ↔ High Life Expectancy	Educational success is a strong predictor of long-term health.
High GCSE Attainment ↔ Low Obesity and Mortality	Education often leads to healthier lifestyle choices and outcomes.
High Persistent Absences ↔ Low Life Expectancy	Absenteeism may reflect broader socioeconomic or health-related challenges.
High Persistent Absences ↔ High Obesity and Mortality	School disengagement often correlates with underlying health and social vulnerabilities.
High Educational Attainment ↔ High Life Expectancy	Education supports better health literacy, income, and lifestyle decisions.
High Educational Attainment ↔ Low Obesity and Mortality	Better-educated populations are more likely to engage in preventive and healthy behaviors.

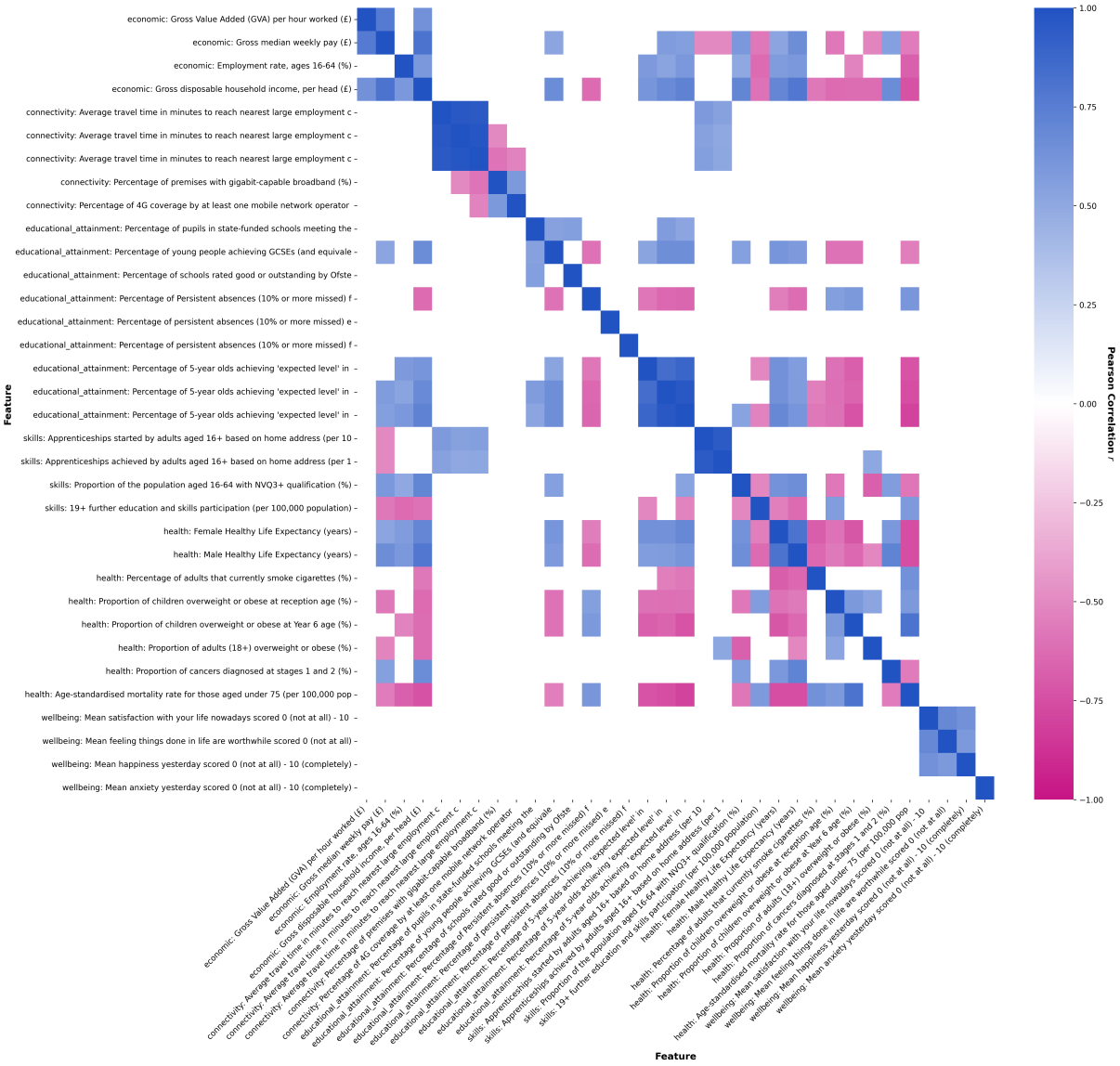

Figure 1.: Features Filtered Correlation Matrix ($|r| > 0.5$)

Canonical Profiles. Based on the strongly correlated features in the correlation matrix (Figure 1), we can create a **"Canonical Profile"**, a profile exhibiting the expected relationships between a set features that exhibit strong pairwise correlations ($|r| > 0.5$) observed across all LAs and views. The plotted lines show their normalised median values across all areas, they summarise the central tendency (median) of mutually correlated variables across the dataset. This Canonical profile is displayed in blue in Figure 2.

Conversely, we can also build an expectation as to what the archetypes on the extremes could look like. To do so, we draw archetypal high- and low-income groups (by top and bottom quartiles). We chose *Gross disposable household income per head* because it is the variable that has the largest number of high correlations across all features ($r > 0.5$ with 18 other features). Those archetypes are displayed in green and red in Figure 2.

This **can be used to assess whether the structure of discovered clusters is consistent with intuitive expectations**: If each cluster's profile resembles the correlation-derived profiles, this means that each cluster expresses a meaningful signal. If their shapes are wildly inconsistent with known strong correlations, or display jagged or erratic shape, they may be spurious.

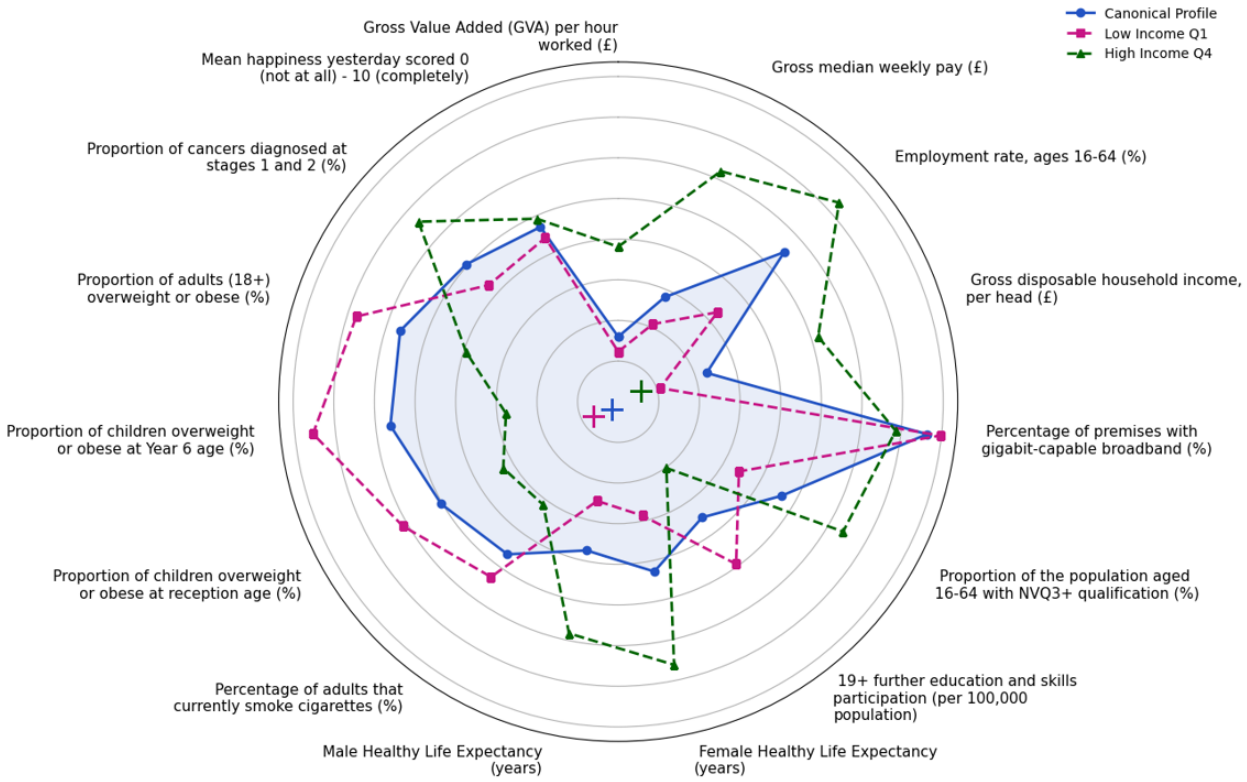

Figure 2.: Canonical Profile based on Correlated Features ($|r| > 0.5$) with High and Low Income Archetypes

Correlations with Missingness. When exploring the correlations between Features with high missingness (more than 10% missing data) with other Features (Figure 3), we find the following linear relationships:

Table 6.: Feature-Level Correlations with Missingness and Interpretations

Missingness Correlation	**Interpretation**
Missing in Edu. Attainment and Health Life Expectancy ↔ High Travel Time to Employment	Remote or poorly connected areas may face data collection challenges and also experience lower service coverage.
Missing in Edu. Attainment and Health Life Expectancy ↔ Low % Premises with Gigabit Broadband	Less digitally connected areas may also lack robust infrastructure for data reporting or monitoring.
Missing in Edu. Attainment and Health Life Expectancy ↔ High Mortality Rate	High-mortality regions may coincide with data gaps in key health and education metrics, potentially masking need.
Missing in Well-being ↔ High Gross Disposable Household Income	May reflect reduced reporting in affluent areas where well-being is assumed rather than systematically measured.
Missing in Diagnosed Cancer Staging (Stage 1 or 2) ↔ High Gross Median Weekly Pay	High-income areas may rely on private systems or alternative data pathways, leading to national-level gaps.
Missing in Diagnosed Cancers (Stage 1 or 2) ↔ Low Proportion of Adults Overweight or Obese	May indicate that healthier populations are underrepresented in public datasets, again possibly due to decentralised health services.

Those insights (Figure 3), combined with what we know of the scale of the fragementation and missingness, especially in certain views can have serious implications:

- **Clusters could reflect data availability patterns**, for example: Remote or underserved areas may form a distinct cluster due to missing education/health values, not due to strong observed similarities. Affluent areas might cluster together if they are underrepresented in public health datasets (e.g., due to reliance on private providers).
- This analysis indicates that **the MCAR (Missing Completely at Random) assumption** might be violated, with some indications that missingness can be systematically related to geography and socioeconomic conditions. Therefore, clusters with many missing entries might align with high-income, well-performing authorities (missing due to alternative reporting pipelines) or low-performing, under-served authorities (missing due to structural data collection gaps).

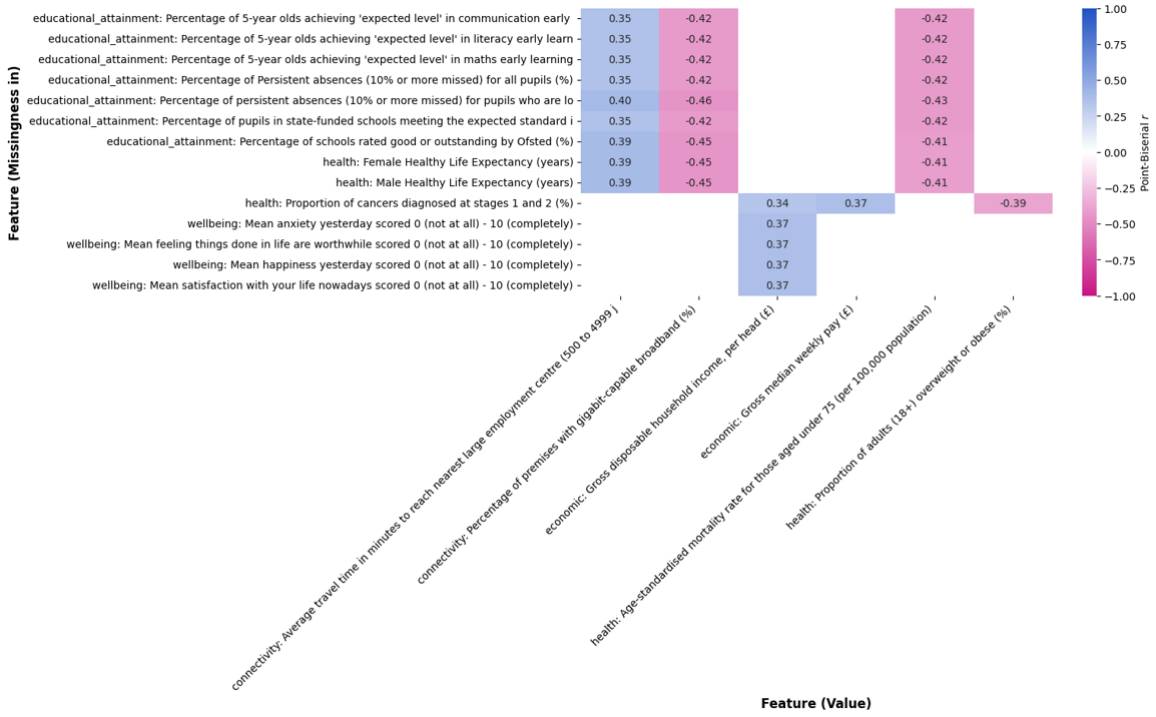

Figure 3.: Correlation between Feature Values and Missingness (only for features missing $> 10\%$, $|r| > 0.3$, $p\text{-}val < 0.05$)

D. Detail of Experimental Methods

Appendix D1: Methods of the ONS pipeline

Here we briefly describe the foundational clustering pipeline in unsupervised learning (as per ONS methodology) and a great basis for comparison.

Its core idea is a simple and widely used clustering pipeline: Raw Data $\rightarrow$ Dimensionality Reduction (PCA) $\rightarrow$ Clustering (KMeans) $\rightarrow$ Evaluation (Silhouette Score).

Its goals are:

- Compress high-dimensional data into a smaller space while preserving structure.
- Discover natural groupings in the reduced space.
- Evaluate how well-separated and cohesive the clusters are.

Let the data matrix be $\mathbf{X} \in \mathbb{R}^{n \times d}$ where n is the number of samples and d is the number of features.

Assume $\mathbf{X}$ has been centered (zero mean) and optionally standardized.

PCA (Principal Component Analysis). It reduces dimensionality by projecting onto directions of maximum variance (Jolliffe and Cadima, 2016).

Compute covariance matrix $\Sigma = \frac{1}{n}\mathbf{X}^T\mathbf{X}$

Solve eigenvalue problem: $\Sigma\mathbf{W} = \mathbf{W}\Lambda$ where $\mathbf{W}$ contains eigenvectors and Λ the eigenvalues.

Keep top p components $\mathbf{Z} = \mathbf{X}\mathbf{W}_{[:,1:p]} \in \mathbb{R}^{n\times p}$

This retains directions with most variance $\text{argmax}_{\|\mathbf{w}\|=1}\text{Var}(\mathbf{X}\mathbf{w})$

K-Means Clustering. Partitions the projected data $\mathbf{Z}$ into k clusters by minimizing within-cluster variance $\min_{\{C_i\}_{i=1}^k} \sum_{i=1}^k \sum_{\mathbf{z}_j \in C_i} \|\mathbf{z}_j - \boldsymbol{\mu}_i\|^2$ where C_i: set of points in cluster i and $\boldsymbol{\mu}_i$: centroid of cluster i

The algorithm alternates assigning points to nearest centroid and recomputing centroids as means of assigned points (MacQueen, 1967).

Silhouette Score. Evaluates clustering quality using cohesion vs separation:

For each sample i, we have $a(i)$: mean intra-cluster distance and $b(i)$: mean nearest-cluster distance.

Then $s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \in [-1, 1]$

The overall silhouette score is the mean over all points (Rousseeuw, 1987).

Pipeline 1: The ONS Single-View Clustering

Require: View-specific dataset $\mathcal{D}$ with Local Authorities as rows and indicators as columns; ONS exclusion list $\mathcal{E}$
Parameters: PCA variance threshold θ (0.25), random seed (19042022), number of optimal clusters k for each view
1: **Impute:** Impute missing data in low-tier Local Authority with higher-tier Local Authority data in $\mathcal{D}$
2: **Filter:** Keep only Local Authority rows from $\mathcal{D}$
3: **Exclude:** Drop Area Codes in $\mathcal{E}$
4: **Winsorize:** Cap each indicator at its 1st and 99th percentiles
5: **Standardize:** Apply Z-score normalization
6: **Reduce Dimensionality:** Apply PCA to retain a minimum of 2 components explaining variance $\geq \theta$
7: **Cluster:** Fit KMeans with k clusters on the PCA components
8: **Evaluate:** Compute silhouette score
return Cluster labels and evaluation metrics

Where **winsorization** (Dixon, 1950) is a technique that limits extreme outliers by capping values at fixed percentiles (typically 1st and 99th), reducing their influence on downstream steps and **standardization** (Jain et al., 1999) transforms each indicator to have zero mean and unit variance (Z-score), ensuring comparability across different scales.

Table 7.: Advantages and Disadvantages of the PCA + KMeans Clustering Pipeline

Advantages	Disadvantages
Dimensionality reduction: PCA reduces noise and captures directions of maximum variance.	**PCA is linear**: Cannot capture nonlinear structure or manifolds in data.
Efficiency: Fast to compute and scalable to high-dimensional data.	**KMeans assumes spherical clusters**: Performs poorly with non-spherical or imbalanced clusters.
Interpretability: PCA axes and KMeans clusters are easy to explain and visualize.	**Sensitive to feature scaling**: Unscaled variables can dominate the PCA and clustering.
Repeatability: Deterministic results with fixed random seed.	**Not robust to outliers**: Both PCA and KMeans are easily influenced by extreme values.
Off-the-shelf availability: Readily available in libraries like `scikit-learn`.	**Requires manual selection of** k: No guarantee of optimal cluster number.

Notes. This pipeline is most suitable for exploratory clustering tasks where linear structure and well-separated clusters are reasonable assumptions.

Appendix C2: Methods used in Experiment

Early integration — Masked Cosine Similarity/Distance (MCS). We concatenate all preprocessed views column-wise to form a single matrix and compute pairwise similarity using a masked cosine that ignores missing coordinates (Yin and Sun, 2022).

Let $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^d$ be two (possibly incomplete) rows. Define the overlap mask

$$\mathbf{m}_{ij} = \neg\big(\mathrm{isnan}(\mathbf{x}_i) \ \vee \ \mathrm{isnan}(\mathbf{x}_j)\big) \in \{0,1\}^d, \quad \mathbf{x}_i^{(m)} = \mathbf{x}_i[\mathbf{m}_{ij}], \quad \mathbf{x}_j^{(m)} = \mathbf{x}_j[\mathbf{m}_{ij}].$$

The **masked cosine similarity** and corresponding distance are

$$\mathrm{sim}_{ij} = \frac{\mathbf{x}_i^{(m)} \cdot \mathbf{x}_j^{(m)}}{\|\mathbf{x}_i^{(m)}\|_2 \, \|\mathbf{x}_j^{(m)}\|_2 + \varepsilon}, \qquad D_{ij} = 1 - \mathrm{sim}_{ij},$$

and if $\sum \mathbf{m}_{ij} = 0$ (no overlap) we set $D_{ij} = 1$ (thus $\mathrm{sim}_{ij} = 0$).

Embedding and clustering. We work with the similarity $S = [\mathrm{sim}_{ij}]$. Let $H_{\mathrm{c}} = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ and $K = H_{\mathrm{c}} \, \frac{1}{2}(S + S^\top) \, H_{\mathrm{c}}$. Compute $K = U \Lambda U^\top$ and obtain PCA scores

$$Z = U_{[:,1:n_{\mathrm{PCA}}]} \, \Lambda^{1/2}_{[1:n_{\mathrm{PCA}}]} \in \mathbb{R}^{n \times n_{\mathrm{PCA}}}.$$

Run K-means on Z for $k \in \{4, \ldots, 15\}$ and select (n_{PCA}, k) by average silhouette.

Optional baseline (feature-space PCA). If a feature-space PCA baseline is desired, impute *the feature matrix* by column means and apply standard PCA on the imputed features; the masked cosine matrix itself is not imputed.

Intermediate integration — Multi-View Similarity Fusion (MVSF). We compute per-view pairwise similarities using a *masked cosine* (handles missingness), fuse them by the mean, then perform kernel PCA on the fused similarity and cluster with K-means, selecting (n_{PCA}, k) by silhouette (intermediate integration at the *similarity* level).

Pipeline 2: Masked Cosine Similarity (Early Integration)

Require: Dictionary of preprocessed views $\{\mathcal{D}^{(v)}\}_{v=1}^{V}$, ONS exclusion list $\mathcal{E}$
Parameters: PCA dimensions $n_{\text{PCA}} \in \{2, \ldots, 5\}$, KMeans clusters $k \in \{4, \ldots, 15\}$, random seed (19042022)
Preprocess Each View: Apply view-specific preprocessing; drop excluded LAs $\in \mathcal{E}$
Align Views: Reindex each view to a common LA set; concatenate features $\rightarrow X_{\text{combined}}$
Compute Similarity: For each pair (i, j), compute masked cosine similarity S_{ij} using overlapping non-missing features (Eq. D)
Form Kernel: Symmetrise and centre S: $K = H_{\text{c}} \frac{1}{2}(S + S^{\top}) H_{\text{c}}$, $H_{\text{c}} = I - \frac{1}{n}\mathbf{1}\mathbf{1}^{\top}$
for $n_{\text{PCA}} \in \{2, \ldots, 5\}$ **do**
  Eigendecompose $K = U\Lambda U^{\top}$; form embedding $Z = U_{[:,1:n_{\text{PCA}}]}\, \Lambda^{1/2}_{[1:n_{\text{PCA}},1:n_{\text{PCA}}]}$
  for $k \in \{4, \ldots, 15\}$ **do**
    Run KMeans with k clusters on Z
    Compute silhouette score
  end for
end for
return Best consensus labels and silhouette score

Table 8.: Advantages and Disadvantages of Masked Cosine Similarity

Advantages	Disadvantages
Handles missing data: Masked cosine similarity naturally avoids imputation bias by ignoring missing elements in pairwise similarity.	**Early integration limits flexibility**: Assumes all features from all views are comparable after standardization, which may not hold.
View alignment: Combines features from multiple views even if each view has different missing areas or coverage.	**Feature imputation may introduce noise**: Mean imputation for PCA may distort distributions, especially with non-random missingness.
Robust to scale: Cosine similarity is magnitude-invariant, reducing outlier influence in clustering.	**No view weighting**: All views are treated equally in concatenation, potentially undervaluing more informative ones.

Notes. This table summarizes the strengths and limitations of an early integration strategy using masked cosine distance, PCA, and KMeans clustering for multi-view data with missing values.

For $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^d$ with missing values, let

$$\mathbf{m}_{ij} = \neg\big(\text{isnan}(\mathbf{x}_i) \ \vee \ \text{isnan}(\mathbf{x}_j)\big), \quad \mathbf{x}_i^{(m)} = \mathbf{x}_i[\mathbf{m}_{ij}], \quad \mathbf{x}_j^{(m)} = \mathbf{x}_j[\mathbf{m}_{ij}].$$

Masked cosine similarity (Yin and Sun, 2022):

$$\text{sim}_{ij} = \frac{\mathbf{x}_i^{(m)} \cdot \mathbf{x}_j^{(m)}}{\|\mathbf{x}_i^{(m)}\|_2 \, \|\mathbf{x}_j^{(m)}\|_2 + \varepsilon}, \qquad \text{if } \sum \mathbf{m}_{ij} = 0 \ \text{ set } \ \text{sim}_{ij} = 0.$$

Let $S^{(v)} \in \mathbb{R}^{n \times n}$ be the per-view similarity matrix (optionally row-normalised with zero diagonal).[1]

$$\textbf{Mean fusion:} \qquad S_{\text{fused}} = \frac{1}{V} \sum_{v=1}^{V} S^{(v)}.$$

Symmetrise & centre (kernel PCA prep) (Kumar and Daumé III, 2011; Wang et al., 2014b; **?**):

$$S_{\text{sym}} = \tfrac{1}{2}\left(S_{\text{fused}} + S_{\text{fused}}^{\top}\right), \qquad H_{\text{c}} = I - \tfrac{1}{n}\mathbf{1}\mathbf{1}^{\top}, \qquad K = H_{\text{c}}\, S_{\text{sym}}\, H_{\text{c}}.$$

[1] Row-stochastic normalisation is optional; it can break symmetry, so we explicitly symmetrise after fusion.

Embedding & clustering. Compute the kernel-PCA scores $Z \in \mathbb{R}^{n \times n_{\mathrm{PCA}}}$ from K and apply K-means on Z; select (n_{PCA}, k) by average silhouette.

Pipeline 3: Similarity Fusion (Intermediate Integration)

Require: Dictionary of preprocessed views $\{\mathcal{D}^{(v)}\}_{v=1}^{V}$, ONS exclusion list $\mathcal{E}$
Parameters: PCA dims $n_{\mathrm{PCA}} \in \{2, \ldots, 5\}$, KMeans clusters $k \in \{4, \ldots, 15\}$, random seed (19042022)
Preprocess Each View: Apply view-specific preprocessing; drop excluded LAs $\in \mathcal{E}$ and metadata columns
Align Views: Reindex each view to a common Local Authority set
for each view v **do**
  Compute pairwise masked-cosine similarities $S^{(v)}$ (set $\mathrm{sim}_{ij} = 0$ if no overlapping features)
  Optional: Zero diagonal and row-normalise $S^{(v)}$ to obtain $\tilde{S}^{(v)}$
end for
Fuse Similarities: $S_{\mathrm{fused}} \leftarrow \frac{1}{V} \sum_{v=1}^{V} \big(\text{use } \tilde{S}^{(v)} \text{ if normalised, else } S^{(v)}\big)$
Symmetrise: $S_{\mathrm{sym}} \leftarrow \frac{1}{2}\big(S_{\mathrm{fused}} + S_{\mathrm{fused}}^{\top}\big)$
Center Kernel: $H_{\mathrm{c}} \leftarrow I - \frac{1}{n}\mathbf{1}\mathbf{1}^{\top}$; $K \leftarrow H_{\mathrm{c}}\, S_{\mathrm{sym}}\, H_{\mathrm{c}}$
Eigendecompose $K = U \Lambda U^{\top}$
for $n_{\mathrm{PCA}} \in \{2, \ldots, 5\}$ **do**
  Form embedding $Z \leftarrow U_{[:,1:n_{\mathrm{PCA}}]}\, \Lambda^{1/2}_{[1:n_{\mathrm{PCA}},1:n_{\mathrm{PCA}}]}$
  for $k \in \{4, \ldots, 15\}$ **do**
    Fit KMeans(k) on Z
    Compute silhouette score
  end for
end for
  return Best labels and silhouette score over the grid

Table 9.: Advantages and disadvantages of Multi-View Similarity Fusion (MVSF)

Advantages	Disadvantages
Handles missing data without feature imputation: masked cosine uses only overlapping coordinates.	**Quadratic cost in** n: computing/storing V similarities is $O(Vn^2)$; eigendecomposition on K is $O(n^3)$.
Intermediate fusion, parameter-free: simple mean across views; reproducible and transparent baseline.	**Equal (unlearned) view weights**: mean fusion may *oversmooth* and underweight informative views.
Scale invariance: cosine reduces sensitivity to per-feature scale/units across heterogeneous views.	**Kernel PCA interpretability**: components are not directly tied to original features.
Per-view diagnostics: enables attribution of contributions and ablations per view.	**Symmetry/PSD caveats**: optional row-normalisation can break symmetry; must symmetrise and centre. PSD not guaranteed in general.
No raw-feature imputation: avoids distortions from early concatenation with mean fills.	**No adaptive fusion**: importance is not learned; robust weighting or SNF-style diffusion may help.

Notes. We compute per-view *similarities* (not distances), fuse by the mean, then *symmetrise and centre* before kernel PCA: $S_{\mathrm{sym}} = \frac{1}{2}(S_{\mathrm{fused}} + S_{\mathrm{fused}}^{\top})$, $H_{\mathrm{c}} = I - \frac{1}{n}\mathbf{1}\mathbf{1}^{\top}$, $K = H_{\mathrm{c}} S_{\mathrm{sym}} H_{\mathrm{c}}$. For large n, consider sparsification/NN graphs or approximate eigensolvers to reduce cost.

Late integration — Co-Association Similarity Ensemble (CSE). Each view is clustered independently with pre-specified k_v, producing labels $c_i^{(v)}$ for entities present in view v. We aggregate these partitions into a *co-association* matrix C that records how often pairs co-cluster, then centre C, embed via PCA, and apply K-means (silhouette selection).

Co-association. Let $\mathcal{I}^{(v)} \subseteq \{1, \ldots, n\}$ be the index set present in view v. For a pair (i, j) define the co-occurrence set $\mathcal{V}_{ij} = \{\, v : \; i, j \in \mathcal{I}^{(v)} \,\}$ and $m_{ij} = |\mathcal{V}_{ij}|$. The co-association matrix $C \in [0, 1]^{n \times n}$ is

$$C_{ij} = \begin{cases} 1, & i = j, \\ \dfrac{1}{m_{ij}} \displaystyle\sum_{v \in \mathcal{V}_{ij}} \mathbb{1}\big[c_i^{(v)} = c_j^{(v)}\big], & m_{ij} > 0,\ i \neq j, \\ 0, & m_{ij} = 0,\ i \neq j. \end{cases}$$

Consensus embedding and clustering. Let $H_c = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ and $C_s = \frac{1}{2}(C + C^\top)$. Form the centred matrix

$$K = H_c\, C_s\, H_c.$$

Compute $K = U\Lambda U^\top$ (clip tiny negative eigenvalues if needed) and take

$$Z^{\text{cons}} = U_{[:,1:d]}\, \Lambda^{1/2}_{[1:d,1:d]} \in \mathbb{R}^{n\times d}.$$

Run K-means on Z^{cons}, selecting (d, k) by average silhouette.

Notes. (i) If every pair appears in all views (full coverage), $m_{ij} = V$ and C is PSD; one may take $K = C$ directly. (ii) This is an evidence–accumulation ensemble; we avoid early feature fusion and operate on label-derived similarity.

Pipeline 4: Co-Association Clustering (Late Integration)

Require: Dictionary of preprocessed views $\{\mathcal{D}^{(v)}\}_{v=1}^{V}$, ONS exclusion list $\mathcal{E}$, per-view k from `ONS_K_VALUES`
Parameters: PCA dims $n_{\text{PCA}} \in \{2,\dots,5\}$, KMeans clusters $k \in \{4,\dots,15\}$, random seed (19042022)
for each view v **do**
  Preprocess $\mathcal{D}^{(v)}$ (winsorize, standardize, exclude $\mathcal{E}$)
  Drop rows with no observed features in view v
  Fit KMeans with $k^{(v)}$ clusters to obtain labels $c^{(v)}$ on index set $\mathcal{I}^{(v)}$
end for
Aggregate (co-association):
Initialize $C \in \mathbb{R}^{n\times n}$ with zeros
for each pair (i, j) **do**
  $\mathcal{V}_{ij} \leftarrow \{\, v :\ i, j \in \mathcal{I}^{(v)} \,\}, \quad m_{ij} \leftarrow |\mathcal{V}_{ij}|$
  if $i = j$ **then**
    $C_{ij} \leftarrow 1$
  else if $m_{ij} > 0$ **then**
    $C_{ij} \leftarrow \dfrac{1}{m_{ij}} \sum_{v\in\mathcal{V}_{ij}} \mathbb{1}[c_i^{(v)} = c_j^{(v)}]$
  else
    $C_{ij} \leftarrow 0$
  end if
end for
Center for PCA: $H_c \leftarrow I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top; \quad K \leftarrow H_c\, \frac{1}{2}(C + C^\top)\, H_c$
Eigendecompose $K = U\Lambda U^\top$
for $n_{\text{PCA}} \in \{2,\dots,5\}$ **do**
  $Z^{\text{cons}} \leftarrow U_{[:,1:n_{\text{PCA}}]}\, \Lambda^{1/2}_{[1:n_{\text{PCA}},1:n_{\text{PCA}}]}$
  for $k \in \{4,\dots,15\}$ **do**
    Fit KMeans(k) on Z^{cons}; compute silhouette
  end for
end for
return Best consensus labels and metrics over the grid

Masked Non-negative Matrix Factorisation (MNMF). We concatenate the (aligned) multi-view data into a single nonnegative matrix $X \in \mathbb{R}_+^{n\times d}$ and learn a low-rank parts-based representation under missingness via a masked NMF objective. We then apply PCA to the sample factors and cluster the PCA scores with K-means (silhouette selection).

Mask and objective. Let $M \in \{0,1\}^{n\times d}$ indicate observed entries ($M_{ij} = 1$ if X_{ij} is observed, else 0). MNMF seeks $G \in \mathbb{R}_+^{n\times r}$, $H \in \mathbb{R}_+^{d\times r}$ with $r \le \min(n, d)$ by

$$\min_{G,H\ge 0} \left\| M \odot \left(X - GH^\top\right) \right\|_F^2.$$

Table 10.: Advantages and disadvantages of Co-association Clustering (CSE)

Advantages	Disadvantages
Preserves view-level structure: each view is clustered separately, maintaining independence.	**Requires k_v per view**: poor choices propagate into the final consensus.
Robust to partial coverage: m_{ij} normalisation avoids penalising pairs missing in some views.	**Coverage imbalance**: pairs with small m_{ij} yield high-variance C_{ij} entries.
Consensus reduces variance: evidence accumulation stabilises noisy base clusterings.	**No learned weighting**: all views (and base clusterings) contribute equally.
Heterogeneous data friendly: no shared feature space is required.	**Information loss from hard labels**: C is binary-agreement based; ignores within-cluster distances.
Transparent fusion: parameter-free at the consensus step; easy to audit.	**PSD/centering caveats**: with m_{ij} normalisation, C need not be PSD; centering and (optional) eigenvalue clipping are required before PCA.
	Complexity: building/storing C is $O(n^2)$ and eigendecomposition is $O(n^3)$.

Notes. Co-association uses $C_{ij} = \frac{1}{m_{ij}} \sum_{v \in \mathcal{V}_{ij}} \mathbb{1}[c_i^{(v)} = c_j^{(v)}]$, $C_{ii} = 1$; centre via $K = H_c \, \frac{1}{2}(C + C^\top) \, H_c$ before PCA. Practical tweaks (not shown in the figure): (i) ignore pairs with m_{ij} below a threshold or shrink C_{ij} toward the global mean; (ii) use view weights or soft co-association (posterior co-membership) if available.

(Only observed entries contribute to the loss; unobserved entries are ignored by M.)

Multiplicative updates with masking. With $\varepsilon > 0$ for numerical stability,

$$G \leftarrow G \odot \frac{(M \odot X)H}{\big(M \odot (GH^\top)\big)H + \varepsilon}, \qquad H \leftarrow H \odot \frac{(M \odot X)^\top G}{\big(M \odot (GH^\top)\big)^\top G + \varepsilon}.$$

We initialise G, H with nonnegative seeds (e.g., NNDSVDa), iterate to a maximum number of steps or until relative improvement falls below a tolerance, and fix the random seed for reproducibility.

Embedding and clustering. Let $Z = \mathrm{PCA}_d(G) \in \mathbb{R}^{n \times d}$ (standard, centred PCA on G); apply K-means to Z. We select (r, d, k) by average silhouette over a grid (e.g., $r \in \{2, \ldots, 5\}$, $d \in \{2, \ldots, 5\} \cap \{1, \ldots, r\}$, $k \in \{4, \ldots, 15\}$).

Notes. (i) Inputs must be nonnegative for NMF; if preprocessing yields negatives, apply a nonnegative transform or shift before MNMF. (ii) PCA dimensionality cannot exceed r. (iii) MNMF avoids feature imputation: missing entries are excluded by the mask in the objective, not filled.

Table 11.: Advantages and disadvantages of Masked NMF (MNMF)

Advantages	Disadvantages
Explicitly handles missing data: objective ignores unobserved entries via a mask.	**Requires nonnegative inputs**: may need shifts/transforms; choice can affect interpretability.
Interpretable latent features: nonnegativity yields sparse, parts-based factors.	**Early fusion may blur view structure**: assumes views are comparable after scaling.
Avoids imputation bias: no filling of missing features before factorisation.	**Sensitive to scaling across views**: concatenation can overweight high-variance views without standardisation.
Simple, tunable pipeline: grid over $(r, d_{\mathrm{PCA}} \le r, k)$ with silhouette selection.	**Non-convex optimisation**: depends on initialisation (NNDSVDa recommended); local minima possible.
Extensible: regularisers (e.g., ℓ_1, graph Laplacian) can be added if desired.	**Scalability**: iterative updates can be slow on large X; memory is $O(nd)$ for M and factors.

Notes. (i) We use Euclidean MNMF with multiplicative updates and ε-stabilisation; for count-like data, KL may be preferable. (ii) Masking assumes missingness is roughly MCAR/MAR; if MNAR, bias can remain. (iii) PCA is applied to G and $d_{\mathrm{PCA}} \le r$.

Pipeline 5: Masked NMF (Early Integration)

Require: Dictionary of preprocessed views $\{\mathcal{D}^{(v)}\}_{v=1}^{V}$, ONS exclusion list $\mathcal{E}$
Parameters: rank grid $r \in$ `rank_range`, PCA dims $n_{\mathrm{PCA}} \in \{2,\dots,5\}$, KMeans $k \in \{4,\dots,15\}$, seed (19042022)
NMF options: init = NNDSVDa (fallback random), max_iter, tol, $\varepsilon > 0$
Preprocess: View-specific transforms; drop rows in $\mathcal{E}$
Align: Reindex all views to union of Area Codes
Concatenate: Column-wise $\to X \in \mathbb{R}^{n\times d}$ (may contain missing)
Mask: $M \in \{0,1\}^{n\times d}$ with $M_{ij} = 1$ if X_{ij} observed; optionally drop rows with $\sum_j M_{ij} = 0$
Ensure nonnegativity: If any $X_{ij} < 0$, apply a nonnegative transform/shift to all columns used for MNMF
for each $r \in$ `rank_range` **do**
  Factorize (masked MNMF): solve $\min_{G,H\geq 0} \| M \odot (X - GH^\top) \|_F^2$
  with multiplicative updates:
  $$G \leftarrow G \odot \frac{(M\odot X)H}{\big(M \odot (GH^\top)\big)H + \varepsilon}, \quad H \leftarrow H \odot \frac{(M\odot X)^\top G}{\big(M \odot (GH^\top)\big)^\top G + \varepsilon}$$
  for each valid $n_{\mathrm{PCA}} \in \{2,\dots,\min(5,r)\}$ **do**
    Embed: $Z \leftarrow \mathrm{PCA}_{n_{\mathrm{PCA}}}(G)$ (standard centred PCA on G)
    for each $k \in \{4,\dots,15\}$ **do**
      Fit KMeans(k) on Z; compute silhouette
    end for
  end for
end for
return Best labels and metrics over (r, n_{PCA}, k) grid (seeded for reproducibility)

Intermediate integration — Per-View NMF (PVNMF). Each view is factorised independently to obtain a nonnegative, low-rank sample embedding; these per-view embeddings are column–concatenated, row-wise ℓ_2–normalised, centred, embedded with PCA, and clustered with K-means (intermediate integration at the latent-factor level).

Per-view factorisation. For view v with nonnegative data $X^{(v)} \in \mathbb{R}_+^{n_v\times d_v}$ and rank r_v,

$$\min_{G^{(v)},\, H^{(v)}\geq 0} \big\| X^{(v)} - G^{(v)} {H^{(v)}}^\top \big\|_F^2, \qquad G^{(v)} \in \mathbb{R}_+^{n_v\times r_v}, \quad H^{(v)} \in \mathbb{R}_+^{d_v\times r_v}.$$

(If some rows are missing within view v, they are omitted for that view; inputs must be nonnegative.)

Alignment, concatenation, and normalisation. Let $\mathcal{I}$ be the union of entity indices across processed views and let $\downarrow_{\mathcal{I}}$ reindex to $\mathcal{I}$ (filling absent rows with 0). Define

$$G_{\mathrm{concat}} = \big[\, G^{(1)}\!\downarrow_{\mathcal{I}} \;\big|\; \cdots \;\big|\; G^{(V)}\!\downarrow_{\mathcal{I}} \,\big] \in \mathbb{R}_+^{n\times R}, \quad R = \sum_{v=1}^{V} r_v.$$

Row-wise ℓ_2 normalisation (as in the figure) uses $d_i = \|G_{\mathrm{concat},\,i\cdot}\|_2 + \varepsilon$ and

$$\widehat{G} \;=\; D^{-1} G_{\mathrm{concat}}, \qquad D = \mathrm{diag}(d_1,\dots,d_n).$$

PCA and clustering. Centre for PCA with $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ and $\widetilde{G} = H\,\widehat{G}$. Let $\widetilde{G}^\top\widetilde{G} = V\Lambda V^\top$; the d-dimensional scores are

$$Z = \widetilde{G}\, V_{1:d} \in \mathbb{R}^{n\times d}.$$

Apply K-means to Z and select (d,k) by average silhouette (with $d \leq \min(5,R)$). *Note:* if you prefer a common rank r across views, set $r_v = r$ so $R = Vr$.

Pipeline 6: Per-View NMF + Latent Space Concatenation

Require: Dictionary of preprocessed views $\{\mathcal{D}^{(v)}\}_{v=1}^{V}$
Parameters: NMF rank r (or per-view r_v), PCA dims $d \in \{2, \ldots, 5\}$, KMeans $k \in \{4, \ldots, 15\}$, seed (19042022)
NMF options: init = NNDSVDa (fallback random), max_iter, tol; enforce nonnegativity
Per-view preprocessing: winsorize/standardize; ensure nonnegativity (shift/transform if needed)
for each view v **do**
  Drop rows with missing features in view v (retain index set $\mathcal{I}^{(v)}$)
  Factorize: fit NMF on $\mathcal{D}^{(v)}$ to obtain $G^{(v)} \in \mathbb{R}_+^{|\mathcal{I}^{(v)}| \times r_v}$, $H^{(v)} \in \mathbb{R}_+^{d_v \times r_v}$
end for
Align & concatenate: let $\mathcal{I} = \bigcup_v \mathcal{I}^{(v)}$; reindex each $G^{(v)}$ to $\mathcal{I}$ (fill absent rows with 0)
  $G_{\text{concat}} \leftarrow [\, G^{(1)}\downarrow_{\mathcal{I}} \mid \cdots \mid G^{(V)}\downarrow_{\mathcal{I}} \,] \in \mathbb{R}_+^{n \times R}$, $R = \sum_v r_v$ (or $R = Vr$)
Row-wise L2 normalisation: $d_i \leftarrow \|G_{\text{concat},\, i\cdot}\|_2 + \varepsilon$; $\widehat{G} \leftarrow D^{-1} G_{\text{concat}}$, $D = \text{diag}(d_1, \ldots, d_n)$
for each valid $d \in \{2, \ldots, \min(5, R)\}$ **do**
  PCA (centred): $Z \leftarrow \text{PCA}_d(\widehat{G})$
  for each $k \in \{4, \ldots, 15\}$ **do**
    Fit KMeans(k) on Z; compute silhouette
  end for
end for
  return Best labels and metrics over (d, k) (and r / $\{r_v\}$ if tuned)

Table 12.: Advantages and disadvantages of Per-View NMF (Intermediate Integration)

Advantages	Disadvantages
Captures per-view structure: each view is factorised independently, preserving local patterns.	**No learned view weighting**: simple concatenation can under/overweight views; larger r_v gain more influence.
Interpretable factors: nonnegativity yields sparse, parts-based latent dimensions per view.	**Nonnegativity required**: views must be shifted/transformed to $\mathbb{R}_+$ before NMF.
Handles heterogeneity: no shared feature space needed; differing d_v and scales are fine.	**Coverage mismatch**: dropping rows per view or zero-filling on alignment can bias the joint space.
Modular and parallelisable: new views can be added/refit independently; per-view ranks r_v are tunable.	**Non-convex optimisation**: sensitive to initialisation; solutions may get stuck in local minima.
Row-wise ℓ_2 normalisation equalises scale: mitigates magnitude imbalance across concatenated blocks.	**Downstream assumptions**: PCA is linear; K-means prefers roughly spherical clusters and may blur view distinctions.

Notes. (i) Report $R = \sum_v r_v$ and enforce $d \leq R$ for PCA. (ii) Fix seeds and use NNDSVDa (fallback random) for reproducibility. (iii) If coverage is highly uneven, consider per-view reweighting or balancing r_v to avoid dominance by a single view. (iv) Per-view ablations and contribution analyses help interpretability.

Early integration — Masked Autoencoder (MA). We concatenate all views (union of features), apply NaN-aware robust scaling, and train a single masked autoencoder to reconstruct only the observed entries. The latent codes are then reduced with PCA and clustered by K-means; we grid-search over latent and PCA dimensions and k, selecting by average silhouette. (Gondara and Wang, 2018)

Let $X \in \mathbb{R}^{n \times d}$ be the (possibly incomplete) feature matrix and $M \in \{0, 1\}^{n \times d}$ the observation mask. With encoder $f_{\text{enc}} : \mathbb{R}^d \to \mathbb{R}^h$ and decoder $f_{\text{dec}} : \mathbb{R}^h \to \mathbb{R}^d$,

$$Z = f_{\text{enc}}(X), \qquad \hat{X} = f_{\text{dec}}(Z).$$

Masked loss (per-row normalisation).

$$\mathcal{L}_{\text{masked-MSE}} = \frac{1}{n} \sum_{i=1}^{n} \frac{\left\| (\hat{x}_i - x_i) \odot m_i \right\|_2^2}{m_i^\top \mathbf{1} + \varepsilon},$$

where m_i is row i of M and $\varepsilon > 0$ avoids division by zero. (Implemented in `masked_mse_loss`.)

Embedding and clustering. Let $Z \in \mathbb{R}^{n\times h}$ be the learned codes. We compute $Z_{\mathrm{PCA}} = \mathrm{PCA}_d(Z)$ and fit $\mathrm{KMeans}_k(Z_{\mathrm{PCA}})$ over a grid of (h, d, k), keeping the best by silhouette.

Pipeline 7: Masked Autoencoder + PCA Clustering (Early Integration)

Require: Combined dataset $X \in \mathbb{R}^{n\times d}$ (union of views; may contain NaNs)
Parameters: latent $r \in$ `latent_dim_range`, PCA dims $d_{\mathrm{pca}} \in \{2, \ldots, 5\}$, $k \in \{4, \ldots, 15\}$, seed (19042022)
Preprocess: For each feature, compute median/IQR on observed entries; robust-scale observed values; set missing entries to the *scaled* median (0 after centring). Build mask $M \in \{0, 1\}^{n\times d}$.
for each latent r **do**
  Train MA: Fit encoder/decoder on (X, M) using the masked per-row loss in Eq. 3 (Adam, fixed seed).
  Encode: $H \leftarrow f_{\mathrm{enc}}(X) \in \mathbb{R}^{n\times r}$ (AE bottleneck codes)
  for each d_{pca} **do**
    PCA: $Z \leftarrow \mathrm{PCA}_{d_{\mathrm{pca}}}(H) \in \mathbb{R}^{n\times d_{\mathrm{pca}}}$ (standard centred PCA)
    for each k **do**
      Fit KMeans(k) on Z; compute silhouette score
    end for
  end for
end for
return Best labels and metrics over (r, d_{pca}, k) (seeded for reproducibility)

Table 13.: Advantages and disadvantages of Masked Autoencoder (Early Integration)

Advantages	Disadvantages
Unified representation: learns a joint latent embedding across all views, capturing global structure.	**No view-specific modelling**: ignores view-wise semantics; subtle inter-view dynamics may be lost.
Handles missing data naturally: masked loss trains only on observed entries; no feature imputation required.	**Sensitive to noisy features**: all inputs are treated equally; low-quality views can dominate without weighting.
Nonlinear capacity: captures cross-view interactions beyond linear methods.	**Reconstruction–clustering mismatch**: objective is reconstruction, not cluster separation; may need regularisation.
End-to-end trainable: joint optimisation of the latent space can improve downstream coherence.	**Interpretability**: latent codes are harder to explain than structured decompositions (e.g., NMF).
Scalable: mini-batch training and GPUs make large d feasible.	**Hyperparameter intensive**: tune latent dim, hidden width, learning rate, epochs, *PCA* dim, and k.

Notes. (i) Use NaN-aware robust scaling before training; missing entries are masked in the loss (Eq. 3). (ii) Fix seeds for AE/PCA/K-means to improve reproducibility. (iii) Rows with very few observed features can yield noisy codes despite per-row normalisation; consider minimum-coverage filters or downstream robustness checks.

Intermediate integration — Multibranch Autoencoder with Shared Bottleneck (MBASB).

This intermediate strategy builds *per-view* encoders that map each view to a common latent space and fuses them by an elementwise mean at a shared bottleneck, which a single decoder uses to reconstruct the concatenated features. It leverages view-specific parameters to preserve within-view structure while enforcing a joint representation at the bottleneck. Related ideas include multi-branch networks (**?**), shared latent space learning (Wang et al., 2015), and masked learning for missing data (Gondara and Wang, 2018).

Let aligned views be $\{X^{(v)}\}_{v=1}^{V}$, $X^{(v)} \in \mathbb{R}^{n\times d_v}$, with a concatenated mask $M \in \{0, 1\}^{n\times\sum_v d_v}$. Each view has an encoder $f_{\theta^{(v)}} : \mathbb{R}^{d_v} \to \mathbb{R}^r$ producing per-sample codes $h_i^{(v)}$, and we fuse by the mean:

$$h_i^{(v)} = f_{\theta^{(v)}}\big(x_i^{(v)}\big) \in \mathbb{R}^r, \qquad H_i = \frac{1}{V}\sum_{v=1}^{V} h_i^{(v)} \ \text{(shared bottleneck)}, \qquad \hat{x}_i = g_\phi(H_i) \in \mathbb{R}^{\sum_v d_v}.$$

Stacking rows gives $H \in \mathbb{R}^{n\times r}$ and $\hat{X} = g_\phi(H) \in \mathbb{R}^{n\times\sum_v d_v}$. Training minimises the *same* masked per-row reconstruction loss as MA (Eq. 3) on $(\hat{X}, X; M)$. After training, we compute PCA scores on the bottleneck,

$$Z = \mathrm{PCA}_d(H) \in \mathbb{R}^{n\times d},$$

and cluster with K-means on Z. We grid-search over latent r, PCA dimension $d \in \{2,\dots,5\}$, and $k \in \{4,\dots,15\}$, selecting the best (d,k) by average silhouette (for each (d,k), retaining the r that attains the highest score).

Pipeline 8: Multibranch Autoencoder with Shared Bottleneck (Intermediate Integration)

Require: Dictionary of *aligned* views $\{\mathcal{D}^{(v)}\}_{v=1}^{V}$ (shared Area Codes)
Parameters: latent $r \in$ `latent_dim_range`, PCA dims $d_{\mathrm{pca}} \in \{2,\dots,5\}$, KMeans $k \in \{4,\dots,15\}$, epochs (100), seed (19042022)
Scale per view: NaN-aware robust scaling (median/IQR on observed); keep per-view masks $M^{(v)}$
Concatenate masks/data: $X \leftarrow [X^{(1)}|\cdots|X^{(V)}]$, $M \leftarrow [M^{(1)}|\cdots|M^{(V)}]$
Model: For each view v, define encoder $f_{\theta^{(v)}} : \mathbb{R}^{d_v} \to \mathbb{R}^r$; fuse by mean
$H_i \leftarrow \frac{1}{V}\sum_{v=1}^{V} f_{\theta^{(v)}}(x_i^{(v)})$; single decoder $g_\phi : \mathbb{R}^r \to \mathbb{R}^{\sum_v d_v}$ with $\hat{x}_i = g_\phi(H_i)$
for each latent r **do**
  Train: minimise masked per-row loss of Eq. 3 on $(\hat{X}, X; M)$ (Adam, fixed seed)
  Encode: $H \leftarrow [H_1;\dots;H_n] \in \mathbb{R}^{n\times r}$ (shared bottleneck codes)
  for each d_{pca} **do**
    PCA: $Z \leftarrow \mathrm{PCA}_{d_{\mathrm{pca}}}(H)$ (standard centred PCA)
    for each k **do**
      Fit KMeans(k) on Z; compute silhouette
    end for
  end for
end for
return Best labels and metrics over (r, d_{pca}, k)

Table 14.: Advantages and disadvantages of Multi-Branch Autoencoder with Shared Bottleneck (MBASB)

Advantages	Disadvantages
View-specific encoding: each view has its own encoder, preserving within-view structure.	**No learned view weighting**: mean fusion treats all branches equally; low-quality views are not down-weighted.
Shared latent space: the bottleneck aligns views into a unified representation before decoding.	**Requires strict sample alignment**: needs identical entity indices across views; partially overlapping views need preprocessing.
Masked loss handles missingness: uses the same per-row masked MSE (Eq. 3); no feature imputation.	**Reconstruction–clustering mismatch**: training optimises reconstruction, not cluster separation; may need regularisation.
Nonlinear capacity: captures cross-view interactions beyond linear fusion.	**Bottleneck sensitivity**: too small r underfits (information loss); too large r risks overfitting/noise sharing.
Modular architecture: per-view encoders allow branch-specific widths/depths and diagnostics.	**Coupled decoding**: a single decoder across all features can bias toward dominant views without care.
Reproducible baseline: simple mean fusion; gradients split evenly ($1/V$) through the bottleneck.	**Scalability/tuning**: more parameters than single-branch AEs; grid over r, d_{pca}, k increases cost.

Notes. (i) After training, PCA is applied to the bottleneck H to obtain Z, then K-means is run on Z. (ii) Consider mitigations for equal-weight fusion (e.g., learned weights/attention, view dropout) and for the objective gap (e.g., regularisation, contrastive or clustering-aware heads). (iii) Fix seeds for AE/PCA/K-means; report the chosen (r, d_{pca}, k).

Late integration — Autoencoder Ensemble with Consensus Clustering (AECC). Each view is encoded and clustered *separately*; the resulting partitions are aggregated via a co-association (evidence-accumulation) matrix, which we embed with PCA and cluster to obtain

a consensus. This mirrors label-ensemble ideas in Fred and Jain (2005); Zhou and Burges (2007).

Per-view AEs, embeddings, and labels. For view v with $X^{(v)} \in \mathbb{R}^{n\times d_v}$ and mask $M^{(v)}$, train a masked autoencoder (same loss as MA, Eq. 3) to obtain bottleneck codes

$$H^{(v)} \in \mathbb{R}^{n\times r}, \qquad Z^{(v)} = \mathrm{PCA}_{d_{\mathrm{pv}}}\big(H^{(v)}\big) \in \mathbb{R}^{n\times d_{\mathrm{pv}}}, \qquad c^{(v)} = \mathrm{KMeans}_{k_{\mathrm{pv}}}\big(Z^{(v)}\big).$$

Co-association (label-derived similarity). Let $\mathcal{I}^{(v)}$ be the index set present in view v. For a pair (i,j), define $\mathcal{V}_{ij} = \{\, v : \; i,j \in \mathcal{I}^{(v)} \,\}$ and $m_{ij} = |\mathcal{V}_{ij}|$. The co-association matrix $C \in [0,1]^{n\times n}$ is

$$C_{ij} = \begin{cases} 1, & i = j, \\ \dfrac{1}{m_{ij}} \sum_{v\in\mathcal{V}_{ij}} \mathbb{1}\big[c_i^{(v)} = c_j^{(v)}\big], & m_{ij} > 0,\ i \neq j, \\ 0, & m_{ij} = 0,\ i \neq j. \end{cases}$$

Special case: with full coverage, $m_{ij} = V$ and this reduces to the uniform average $C_{ij} = \frac{1}{V}\sum_v \mathbb{1}[c_i^{(v)} = c_j^{(v)}]$.

Consensus embedding and clustering (PCA). Let $C_s = \frac{1}{2}(C + C^\top)$ and $H_{\mathrm{c}} = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$. Form $K = H_{\mathrm{c}}\, C_s\, H_{\mathrm{c}}$, compute $K = U\Lambda U^\top$, and take

$$Z^{\mathrm{cons}} = U_{[:,1:d_{\mathrm{cons}}]}\, \Lambda^{1/2}_{[1:d_{\mathrm{cons}},1:d_{\mathrm{cons}}]}.$$

Run $\mathrm{KMeans}_{k_{\mathrm{cons}}}$ on Z^{cons}; select $(d_{\mathrm{pv}}, k_{\mathrm{pv}}, d_{\mathrm{cons}}, k_{\mathrm{cons}})$ by average silhouette.

Table 15.: Advantages and disadvantages of Autoencoder Ensemble with Consensus Clustering (AECC, Late Integration)

Advantages	**Disadvantages**
View-specific modelling: each view is encoded and clustered independently, preserving its semantics.	**No joint alignment**: embeddings are learned per view with no shared supervision or cross-view constraints.
Consensus reduces variance: evidence accumulation dilutes the effect of any one noisy view.	**Coverage imbalance**: pairs with small m_{ij} yield high-variance C_{ij}; may bias the consensus.
Heterogeneity-friendly: views can differ in shape/scale/dimensionality; no shared feature space required.	**Information loss from hard labels**: C captures only co-membership, ignoring within-cluster distances.
Transparent fusion: parameter-light, auditable consensus step (no learned weights).	**PSD/centering caveats**: with m_{ij} normalisation, C need not be PSD; must symmetrise/centre (and possibly clip eigenvalues).
Modular & parallelisable: per-view AEs and clusterings can be trained independently.	**Computational cost & tuning**: build/store C in $O(n^2)$; eigendecomp $O(n^3)$; tune per-view AE/PCA and consensus PCA/KMeans.

Notes. Co-association uses $C_{ij} = \frac{1}{m_{ij}}\sum_{v\in\mathcal{V}_{ij}} \mathbb{1}[c_i^{(v)} = c_j^{(v)}]$, $C_{ii} = 1$; centre via $K = H_{\mathrm{c}}\,\frac{1}{2}(C + C^\top)\,H_{\mathrm{c}}$ before PCA. Hyperparameters involve per-view AE latent r, per-view PCA d_{pv}, consensus PCA d_{cons}, and k. Fix seeds for AE/PCA/KMeans. If desired, learned view weights or soft co-association (posterior co-membership) can mitigate equal-weight fusion and label hardening.

Pipeline 9: Autoencoder Ensemble with Consensus Clustering (Late Integration)

Require: Dictionary of views $\{\mathcal{D}^{(v)}\}_{v=1}^{V}$ (per-view index sets may differ)
Parameters: AE latent r (e.g., $r = 5$), per-view PCA $d_{\text{pv}} \in \{2, \ldots, 5\}$, consensus PCA $d_{\text{cons}} \in \{2, \ldots, 5\}$, KMeans $k \in \{4, \ldots, 15\}$, epochs (100), seed (19042022)
for each view v **do**
  Preprocess per view: robust-scale observed features; build mask $M^{(v)}$; keep index set $\mathcal{I}^{(v)}$
  Train masked AE: minimise Eq. 3 on $(X^{(v)}, M^{(v)})$
  Encode: $H^{(v)} \leftarrow f_{\text{enc}}^{(v)}(X^{(v)}) \in \mathbb{R}^{|\mathcal{I}^{(v)}| \times r}$
  Per-view PCA: $Z^{(v)} \leftarrow \text{PCA}_{d_{\text{pv}}}(H^{(v)})$
  Per-view clustering: $c^{(v)} \leftarrow \text{KMeans}_k(Z^{(v)})$ on indices $\mathcal{I}^{(v)}$
end for
Co-association (label-derived similarity):
Initialise $C \in \mathbb{R}^{n \times n}$ with zeros on the union index $\mathcal{I} = \bigcup_v \mathcal{I}^{(v)}$
for each pair (i, j) on $\mathcal{I}$ **do**
  $\mathcal{V}_{ij} \leftarrow \{\, v : \; i, j \in \mathcal{I}^{(v)} \,\}, \quad m_{ij} \leftarrow |\mathcal{V}_{ij}|$
  if $i = j$ **then**
    $C_{ij} \leftarrow 1$
  else if $m_{ij} > 0$ **then**
    $C_{ij} \leftarrow \dfrac{1}{m_{ij}} \sum_{v \in \mathcal{V}_{ij}} \mathbb{1}[c_i^{(v)} = c_j^{(v)}]$
  else
    $C_{ij} \leftarrow 0$
  end if
end for
Consensus embedding (PCA on centred C):
$H_{\text{c}} \leftarrow I - \frac{1}{n}\mathbf{1}\mathbf{1}^{\top}; \quad K \leftarrow H_{\text{c}} \, \frac{1}{2}(C + C^{\top}) \, H_{\text{c}}$
Eigendecompose $K = U \Lambda U^{\top}$
for each $d_{\text{cons}} \in \{2, \ldots, 5\}$ **do**
  $Z^{\text{cons}} \leftarrow U_{[:, 1:d_{\text{cons}}]} \, \Lambda_{[1:d_{\text{cons}}, 1:d_{\text{cons}}]}^{1/2}$
  for each $k \in \{4, \ldots, 15\}$ **do**
    Fit KMeans_k on Z^{cons}; compute silhouette
  end for
end for
  return Best consensus labels and metrics over $(r, d_{\text{pv}}, d_{\text{cons}}, k)$

E. Experimental Results, Diagnostics and Visualisations

Appendix E1: Replication of ONS single-view methodology

This is a concise summary of what has been done in the `ONS_Replication.ipynb` notebook which replicates the Office for National Statistics (ONS) Local Authority clustering methodology. After loading cleaned indicator data from an Excel workbook (organized by thematic views), it applies the ONS single-view clustering pipeline using the `ONS_implementation.py` module and visualise the results using `plots_and_visualisations.py`. This exercise was necessary for us to gain familiarity with the data and methodology but also to provide a critical review of the ONS' work, with challenges and limitations outlined in subsections 2.2.3 and 2.3.2.

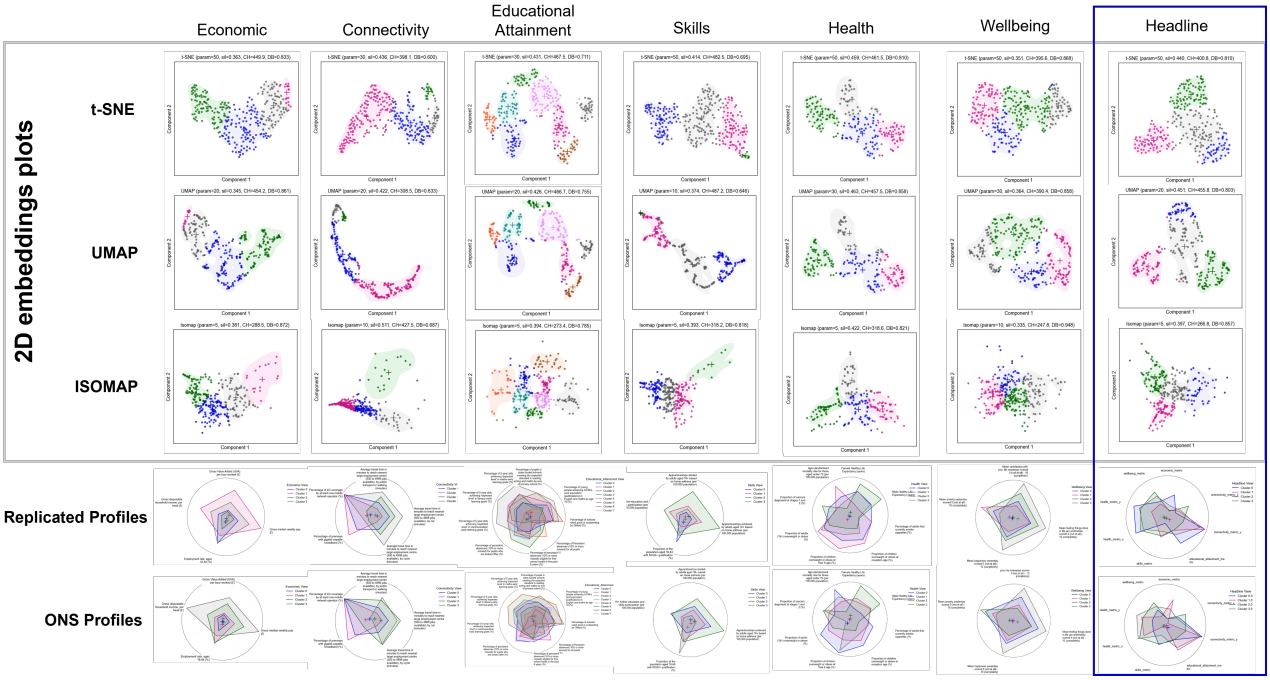

Figure 4.: Clusters Visualisations from ONS Replication

Appendix E2: Results of Dimensionality Reduction + Clustering Grid Search on quality and stability of results

To assess the impact of dimensionality reduction (DR) and clustering algorithm choice, we performed a grid search across three DR techniques: Principal Component Analysis (PCA), Uniform Manifold Approximation and Projection (UMAP), and Multidimensional Scaling (MDS) and four clustering algorithms: K-means, Gaussian Mixture Models (GMM), Agglomerative Clustering, and Spectral Clustering.

- **PCA** preserves global variance structure and consistently led to the highest silhouette scores, especially with K-means.

- **UMAP** preserved local structures well and produced stable clustering with Agglomerative and Spectral methods.
- **MDS** underperformed across all algorithms except for a few high-silhouette outliers with GMM, indicating possible overfitting or noise sensitivity.
- **KMeans** and **Agglomerative Clustering** emerged as the most stable and interpretable across settings.
- **GMM** showed high silhouette variance, making it unreliable without clear underlying Gaussian structure.
- **Spectral Clustering** benefited from UMAP's preservation of local neighborhood graphs but remained less interpretable overall.

Optimal solutions typically involved 2 components and a cluster range of $k = 4$ to 6. Silhouette scores decreased gradually with higher k, as expected due to fragmentation.

Table 16.: Summary of Clustering Algorithm Performance across DR Methods

Algorithm	Best DR	Stability	Max Silhouette	Comment
KMeans	PCA	High	∼0.60	Most stable, interpretable
GMM	UMAP	Very Low	∼0.65 (spike)	Volatile, not robust
Agglomerative	PCA	High	∼0.55	Strong alternative to KMeans
Spectral	UMAP	Medium-High	∼0.50	Best for non-linear structures

We used the same grid search of methods to test their sensitivity to random seed changes:

- **K-means and Agglomerative clustering** produce consistently high silhouette scores with low variance, especially when using PCA or UMAP, indicating robustness to initialisation.
- **GMM** demonstrates very high variability in silhouette scores across all DR methods, particularly PCA, highlighting significant sensitivity to random seed, making it less reliable.
- **Spectral Clustering** exhibits moderate sensitivity, performing best when combined with UMAP, which aligns with its reliance on local structure.
- **MDS** leads to consistently lower performance and greater variability across most algorithms, confirming its poor suitability for these tasks.

Overall, the experiment confirms that K-means (or Agglomerative) + PCA is the most stable and interpretable combination, while GMM is too unstable to recommend without strong prior assumptions.

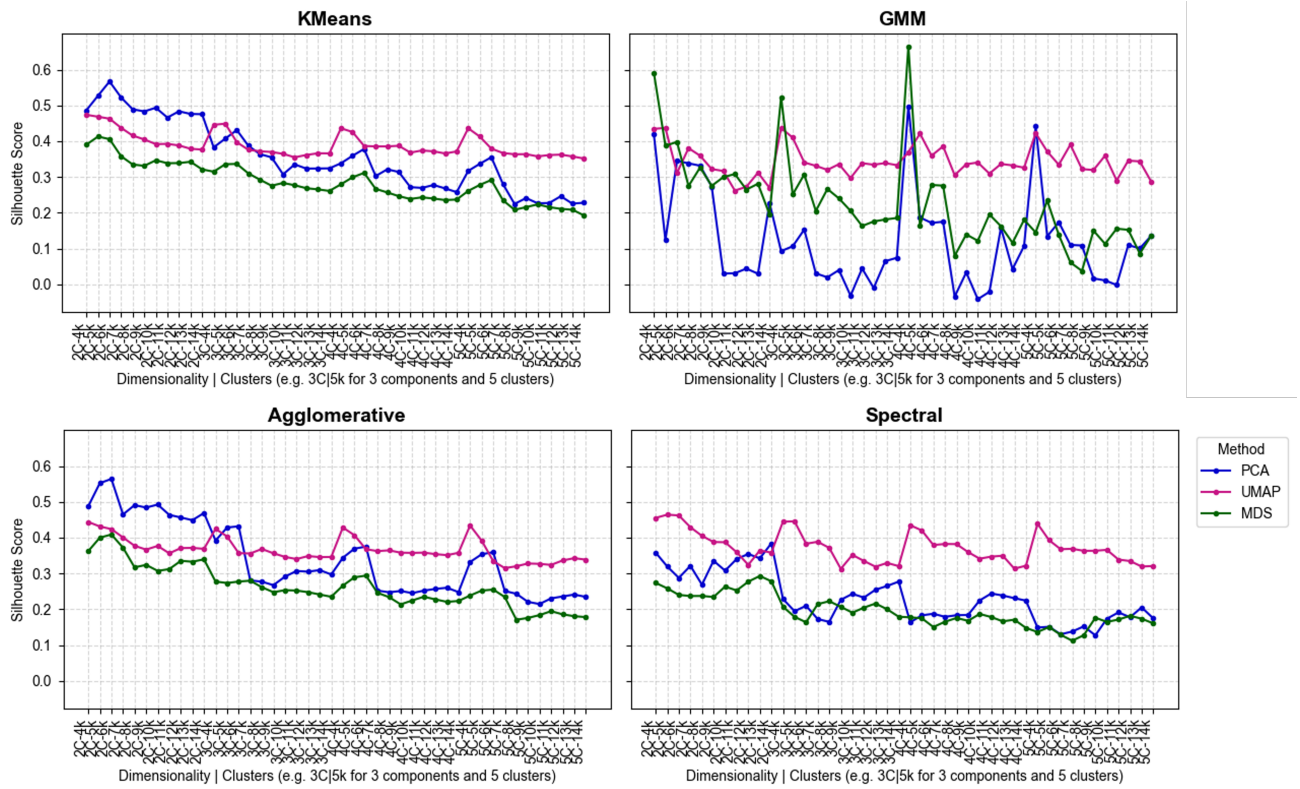

Figure 5.: Silhouette scores across the Grid Search on Concatenated Headline Dataset with no missing values (n=81)

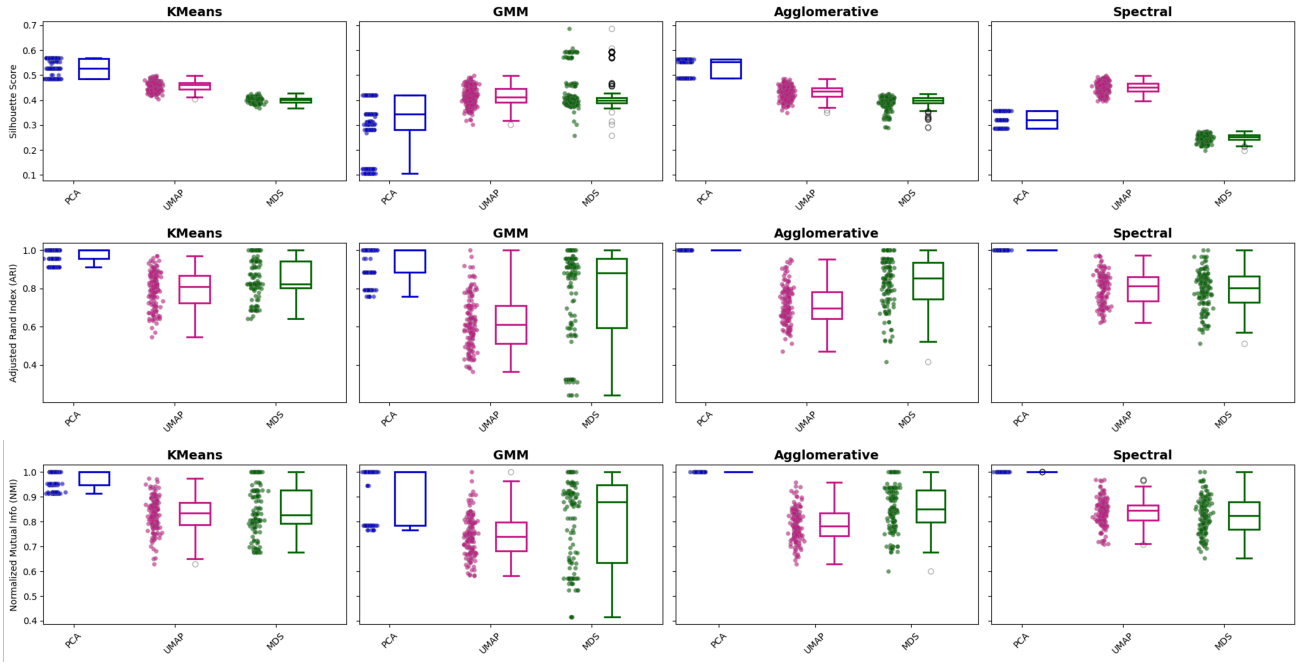

Figure 6.: Stability of Silhouette scores, ARI and NMI across the Grid Search for 100 permutations of random seed on Concatenated Headline Dataset with no missing values (n=81)

Appendix E3: Results of other Multi-View Learning Methods

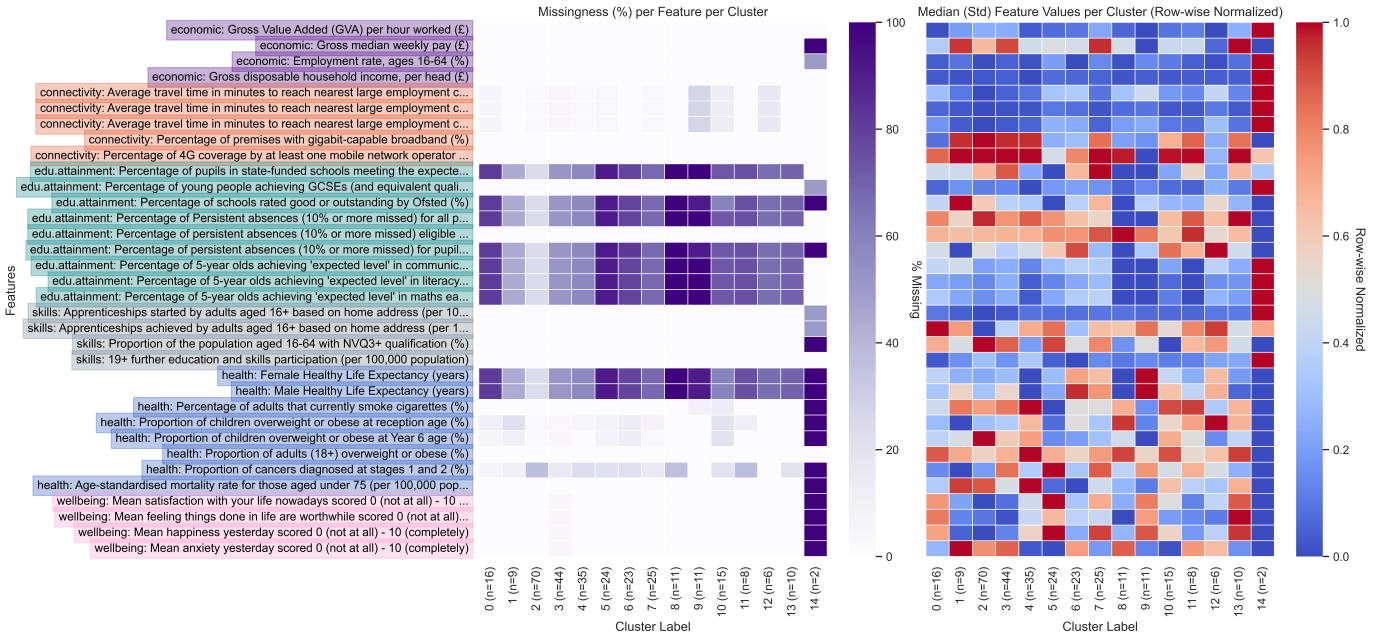

Figure 7.: CFE Clusters Summaries

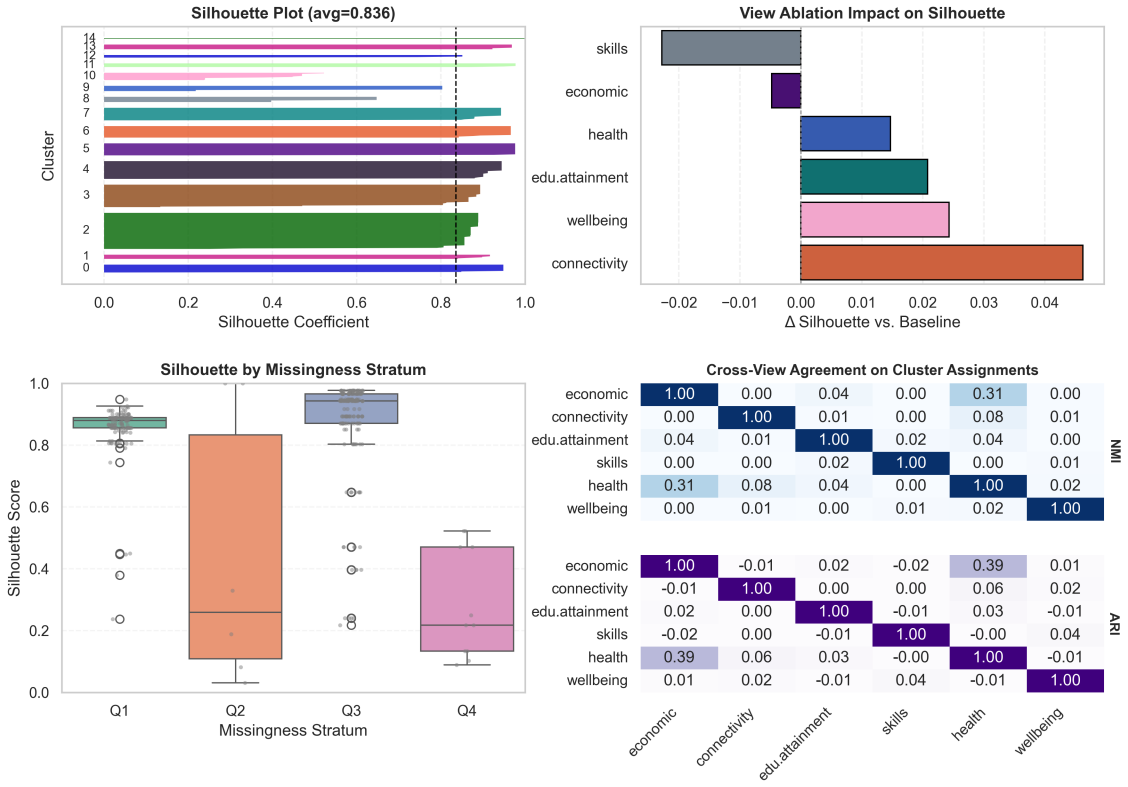

Figure 8.: CFE Silhouette Report

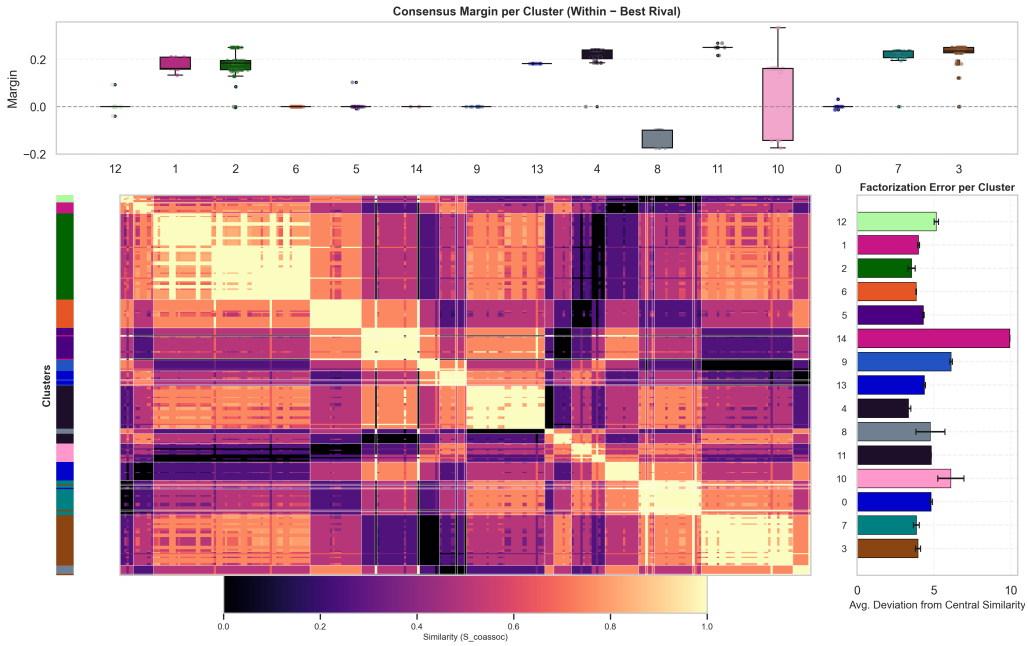

Figure 9.: CFE Factorization Report

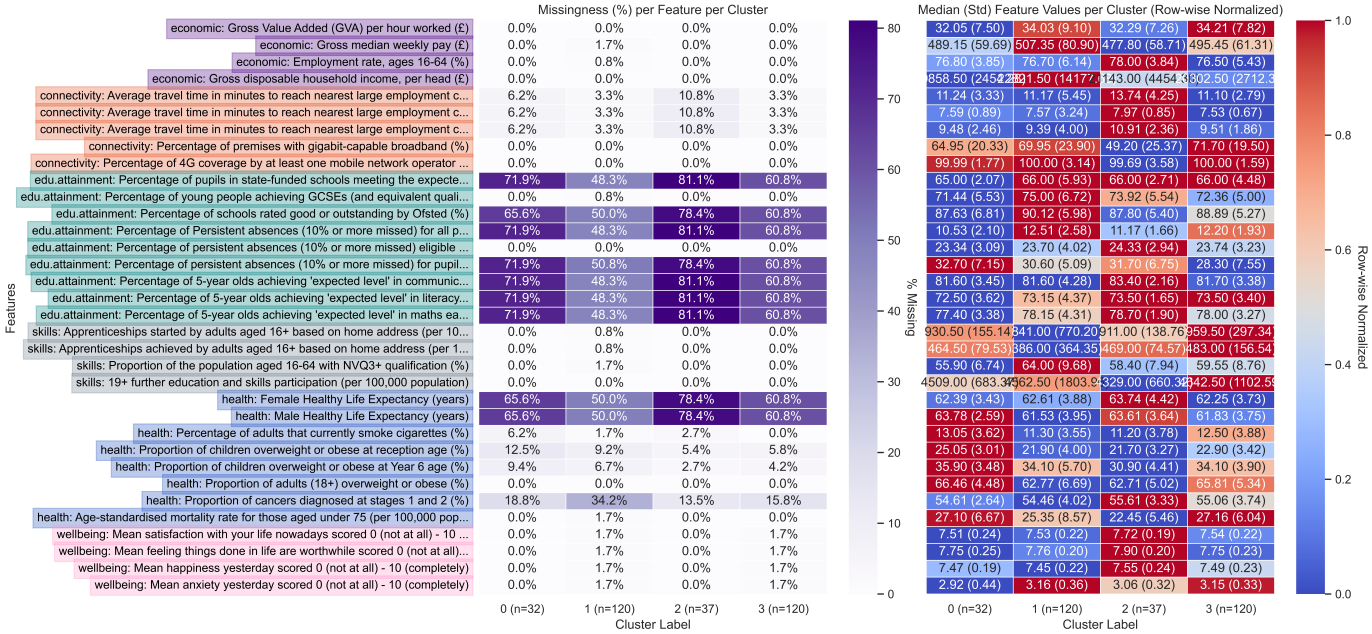

Figure 10.: PVNMF Clusters Summaries

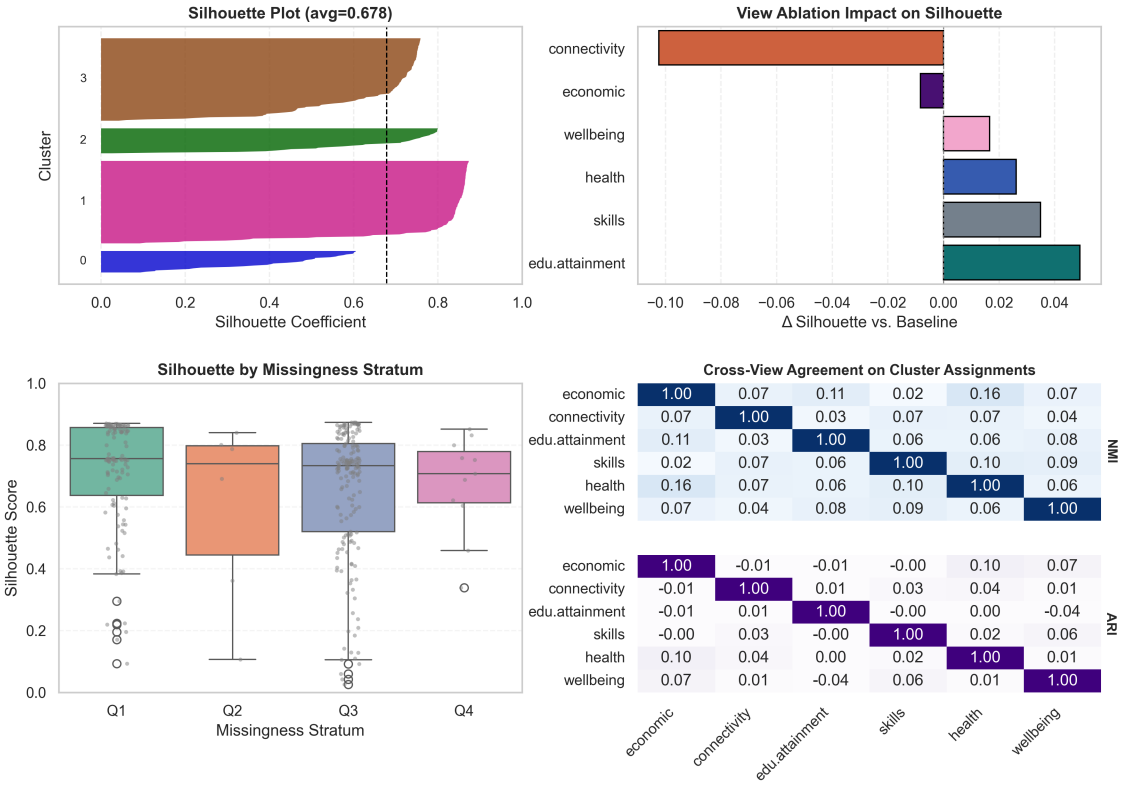

Figure 11.: PVNMF Silhouette Report

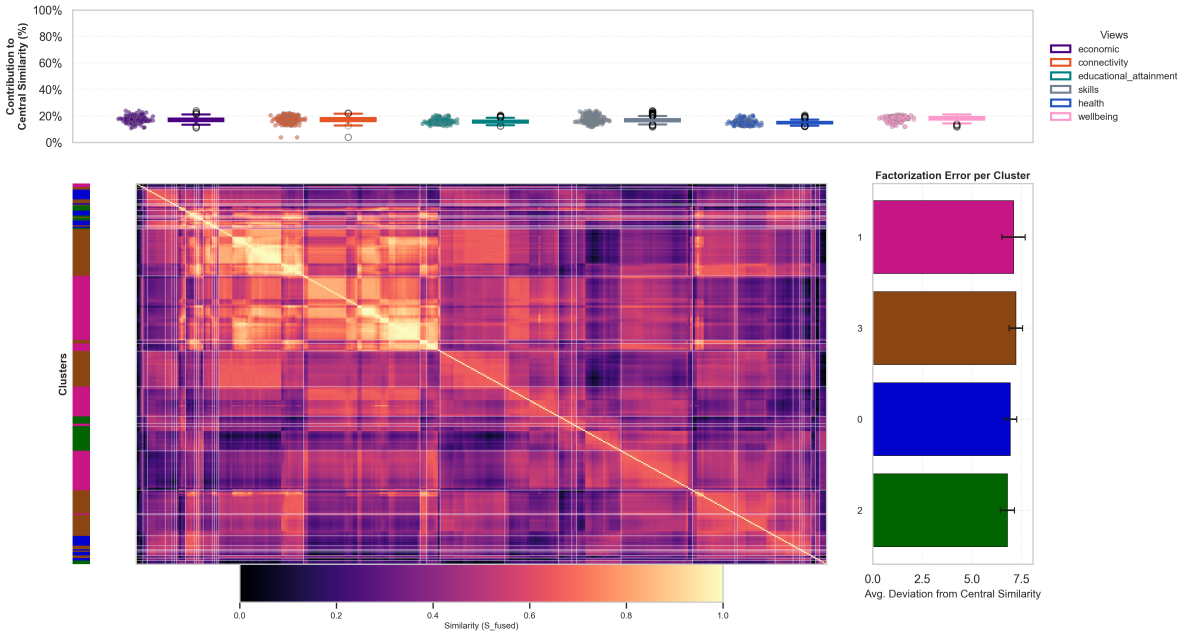

Figure 12.: PVNMF Factorization Report

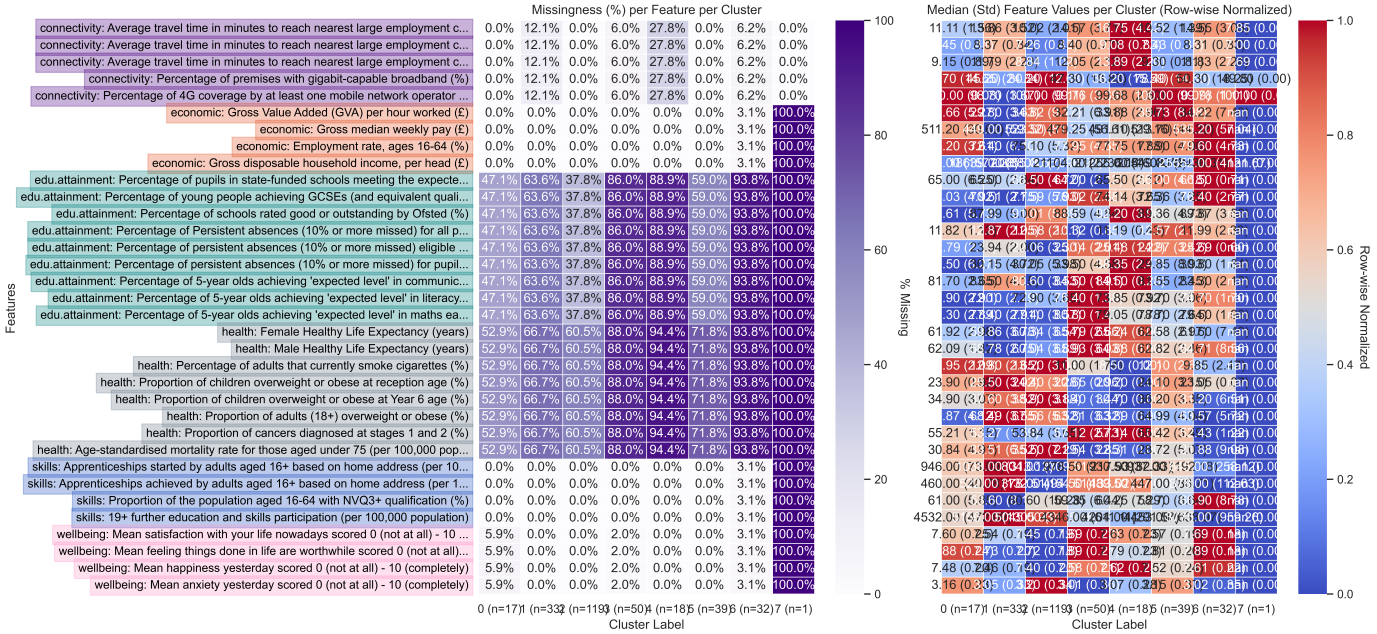

Figure 13.: AECC Clusters Summaries

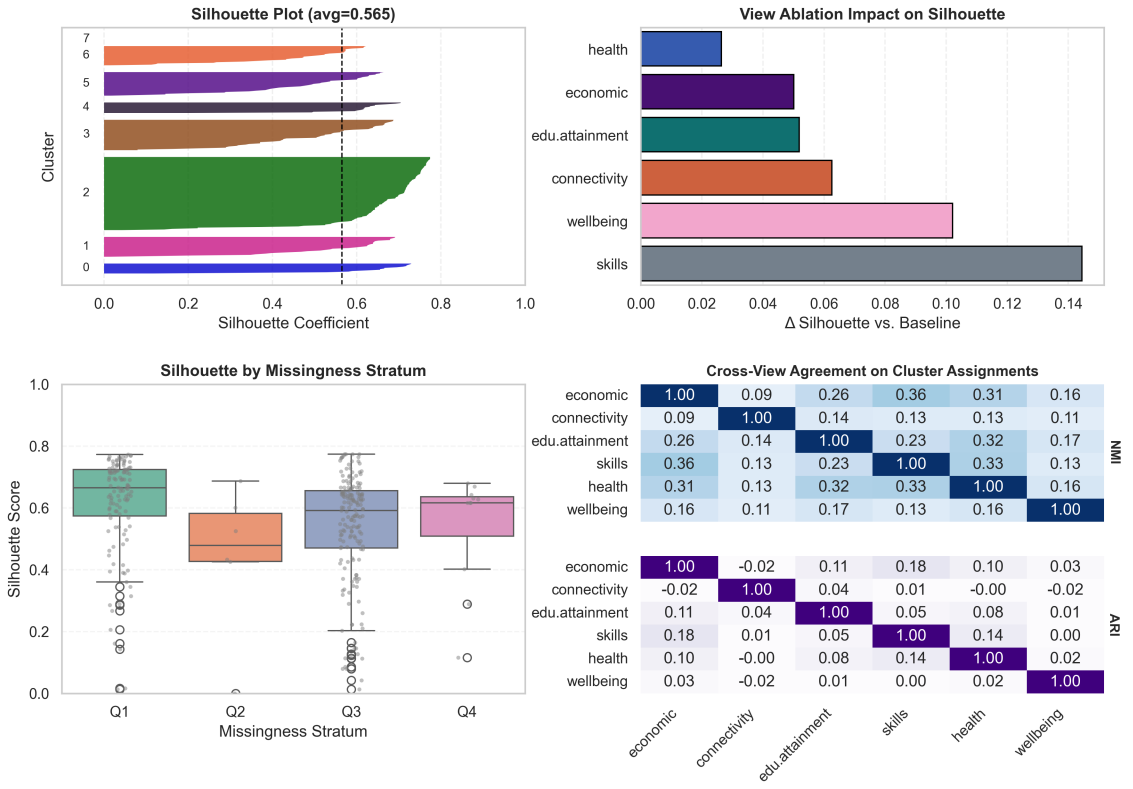

Figure 14.: AECC Silhouette Report

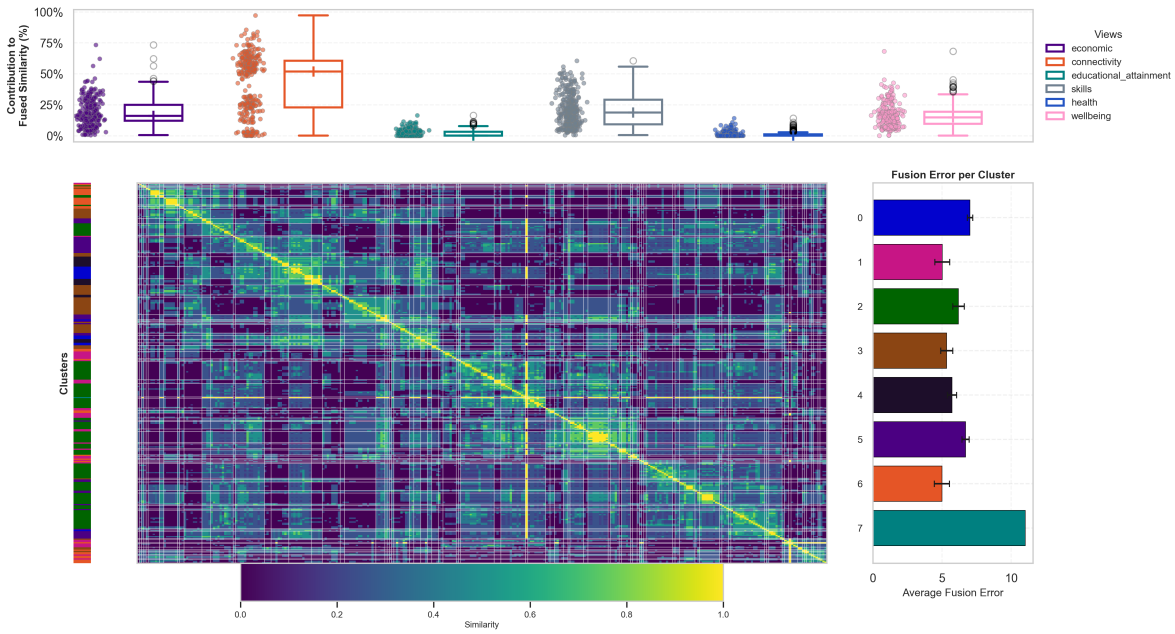

Figure 15.: AECC Reconstruction Report

Appendix E4: Evaluation and Diagnostics Methodology

Cluster Summaries For each view v (e.g., *economic*, *health*), the following steps are applied:

1. Index the table by a unique area identifier (e.g., `Area Code`).
2. Drop meta columns (e.g., `Area`, `Area Level`) if present.
3. Coerce all non-numeric values to `NaN`.
4. Prefix all column names with the view name.

All views are column-concatenated and reindexed to match the ordering of the clustering input. The cluster labels from the model are appended as a `Cluster` column.

Missingness by Feature and Cluster For each feature f and cluster c, the percentage of missing values is computed as:

$$\%\text{Missing}(f, c) = 100 \times \frac{\#\{i \in c \mid x_{if} \text{ is NaN}\}}{|c|}.$$

This yields a feature-by-cluster matrix visualised as a heatmap.

Cluster Medians and Dispersion For each feature f and cluster c, the per-cluster median $\mu_{f,c}$ and standard deviation $\sigma_{f,c}$ are computed. Medians are row-wise min–max normalised to $[0, 1]$ for colour scaling:

$$\tilde{\mu}_{f,c} = \frac{\mu_{f,c} - \min_{c'} \mu_{f,c'}}{\max_{c'} \mu_{f,c'} - \min_{c'} \mu_{f,c'} + \varepsilon}.$$

Annotations display `median (std)` to retain interpretability.

Feature Ranking via Z-Score Contrast

Across clusters, the mean $\bar{\mu}_f$ and standard deviation σ_f for feature f are computed. Cluster-wise z-scores are:

$$z_{f,c} = \frac{\mu_{f,c} - \bar{\mu}_f}{\sigma_f + \varepsilon},$$

and the feature score is $\max_c |z_{f,c}|$.

Feature Ranking via KL Divergence Cluster-specific feature densities $p_{f,c}$ and the overall density q_f are estimated using fixed-width histograms. The Kullback–Leibler divergence is:

$$D_{\text{KL}}(p_{f,c} \,\|\, q_f) = \sum_b p_{f,c}(b) \log \frac{p_{f,c}(b)}{q_f(b)},$$

averaged over clusters.

Visualisation Two heatmaps are shown side-by-side with a shared y-axis (features):

- Left: percentage missingness.
- Right: row-wise normalised cluster medians.

Feature tick backgrounds are shaded according to the view colour palette.

Canonical Profiles Summarise cluster behaviour on a curated subset of "canonical" indicators with known directionality.

For each cluster, the median normalised value is computed per canonical feature. Optionally, clusters may be grouped (e.g., "affluent", "disadvantaged") and medians taken within groups.

Radar charts display one spoke per feature. Profiles are polylines (optional fill) with centroids marked. Direction arrows indicate whether higher is better (+) or lower is better (−).

Silhouette Report For sample i:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}},$$

where $a(i)$ is the mean intra-cluster distance and $b(i)$ the mean nearest-cluster distance.

1. Fuse similarities S across views.
2. Kernel-centre: $K = HSH$ with $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$.
3. PCA to d dimensions; k-means clustering.
4. Compute sample silhouettes and the global average.

1. **Silhouette by cluster:** Sorted samples per cluster with global average line.
2. **View ablation:** Remove one view, recompute silhouette, plot Δ vs. baseline.
3. **Silhouette by missingness:** Bin samples by missingness quantiles, plot distributions.
4. **Cross-view agreement:** Per-view clustering, then compute NMI and ARI between views.

Fusion Report

- Per-view similarities $\{S^{(v)}\}_{v=1}^{V}$ computed via masked cosine similarity.
- Fused similarity:

$$S_{\text{fused}} = \frac{1}{V}\sum_{v=1}^{V} S^{(v)}.$$

- Cluster labels from PCA+k-means on S_{fused}.

Ordering

Row/column order is given by hierarchical clustering on $D = 1 - S_{\text{fused}}$ with optimal leaf ordering.

1. **View contributions:** For sample i and view v,

$$c_i^{(v)} = \frac{1}{n}\sum_{j=1}^{n} \frac{S_{ij}^{(v)}}{S_{\text{fused},ij} + \varepsilon}.$$

   Distribution across i is shown as a boxplot.

2. **Fused similarity heatmap:** Ordered by dendrogram leaves, with cluster boundaries overlaid.

3. **Fusion error per cluster:** For each i,

$$e_i^{(v)} = \left\| S_{i\cdot}^{(v)} - S_{\text{fused},i\cdot} \right\|_2, \quad \bar{e}_i = \frac{1}{V}\sum_{v=1}^{V} e_i^{(v)},$$

   then averaged within clusters.

Multi-view Similarity Graph

Given multiple view–specific similarity matrices over the same set of local authorities (LAs), we build a 2D map of all LAs, draw edges between the most similar pairs, color each edge by the view that most strongly explains that pair, and (optionally) extract an ego–network for a single LA. We also detect communities on the fused, weighted graph and summarise which view dominates inside each community.

We map similarity to a bounded distance via a monotone transform:

$$D_{ij} \;=\; \frac{\max(\bar{S}) - \bar{S}_{ij}}{\max\limits_{p\neq q}\big(\max(\bar{S}) - \bar{S}_{pq}\big) + \varepsilon}, \qquad D_{ii} = 0.$$

Any strictly decreasing transform is acceptable; the above simply rescales to $[0, 1]$.

We embed LAs into $\mathbb{R}^2$ using UMAP McInnes et al. (2018), with a precomputed metric on D:

$$X \in \mathbb{R}^{n\times 2} \;=\; \mathrm{UMAP}\left(D;\; n_{\mathrm{nbrs}},\; \mathrm{min\_dist}\right).$$

We then draw a sparse, undirected, weighted graph $G = (\mathcal{N}, \mathcal{E}, w)$ using one of:

- k–**NN graph** (`mode="knn"`): for each i, connect i to the k largest $\bar{S}_{ij}$, $j \neq i$; keep unique undirected pairs.

$$\mathcal{E} \;=\; \Big\{\{i,j\} :\; j \in \underset{\ell\neq i}{\arg\,\text{top-k}}\;\bar{S}_{i\ell} \text{ or } i \in \underset{\ell\neq j}{\arg\,\text{top-k}}\;\bar{S}_{j\ell}\Big\}.$$

- **Threshold graph** (`mode="threshold"`): include $\{i,j\}$ if $\bar{S}_{ij} \geq \tau$ for a chosen τ.
- **Ego graph** (`mode="ego"`): fix a center c and connect it to its top–k neighbors by $\bar{S}_{c\cdot}$.

Edge weights are $w_{ij} = \bar{S}_{ij}$.

To attribute each edge to the view that best explains it, we compute the dominant view

$$v^*(i,j) \;=\; \underset{v\in\mathcal{V}}{\arg\max}\; S_{ij}^{(v)}.$$

Edges are colored by $v^*(i,j)$; optionally, edge width is made proportional to $S_{ij}^{(v^*(i,j))}$ to reflect the strength of the winning view.

On the fused, weighted graph we detect communities by greedy modularity maximization (Newman and Girvan, 2004). For weighted graphs, modularity is

$$Q \;=\; \frac{1}{2m}\sum_{i,j}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\mathbf{1}\{c_i = c_j\},$$

where $A_{ij} = w_{ij}$, $k_i = \sum_j A_{ij}$ is the (weighted) degree, $2m = \sum_{i,j} A_{ij}$, and c_i is the community label. We summarize each community by the most frequent dominant view on its internal edges.

Nodes are plotted at rows of X; edges $\{i,j\} \in \mathcal{E}$ are rendered as straight segments between X_i and X_j, colored by $v^*(i,j)$. In ego mode, the center is highlighted; neighbor node sizes may scale with the (mean) fused similarity to the center.

Interpretation caveats. Argmax coloring emphasizes frequency of view dominance, not necessarily usefulness for clustering. A view with broadly elevated similarities may dominate many edges yet reduce global separability (e.g., lower silhouette).

Reproducibility Notes

- Random seeds fixed (`random_state=19042022`, `n_init=1000`).
- Missingness computed across the concatenated numeric feature matrix.
- Long view names shortened (e.g., `educational_attainment` → `edu.attainment`).
- High-resolution outputs via `fig.savefig(..., dpi=300, bbox_inches="tight")`.

F. Supplementary Experiments and Ablations

Appendix F1: Supplementary Experiments

Multi-Modal Extension: Spatial Feature View

We attempted to extend the MVL framework by incorporating spatial regularisation, acknowledging the geographic coherence of local authority boundaries, and promoting spatially aware segmentation. However, we were met with disappointing results which we decided to not present in the main body of this report.

Given that Local Authorities (in the context of the Leveling Up agenda) are first and foremost geographical entities, adding geo-spatial data, specifically geographical proximity, to our Multi-View Clustering exercise could be a valuable and justified enhancement. From the ONS' own analysis, we can see that the clusters translate geographically (Figure 16). Our exploratory analysis sought to quantify this relationship and evaluate how much additional explanatory power geography could bring.

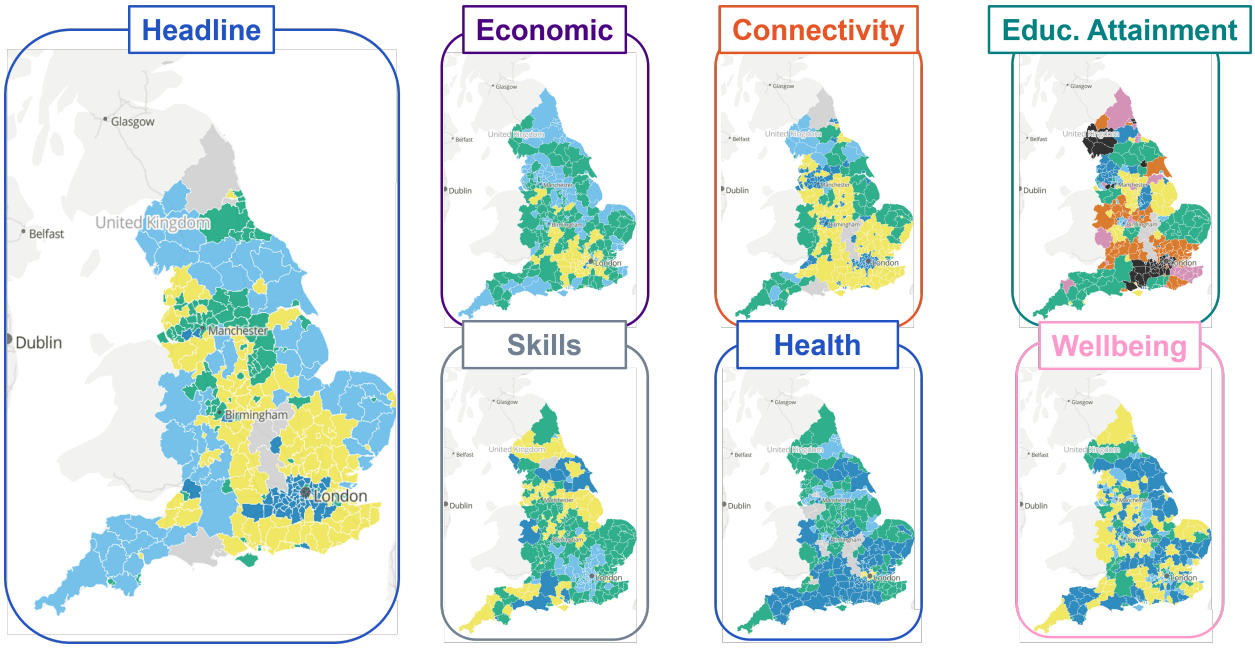

Figure 16.: Geographical visualisation of the ONS clustering results for local authorities in England. *Source:* Office for National Statistics – Subnational indicators explorer (Office for National Statistics, 2023a)

Methodology. We model spatial proximity using a polygon contiguity approach on LA boundaries:

1. Data Acquisition and Preparation: Local Authority boundaries were sourced from the Office for National Statistics (2024b) GPKG dataset and joined to the ONS Headline model cluster assignments. Boundaries were projected to EPSG:4326 for spatial processing.

2. Neighbourhood Definition: Spatial adjacency was defined using Queen contiguity Bivand et al. (2013) via `libpysal`, where two LAs are neighbours if their polygons share either a boundary segment or a vertex:

$$w_{ij} = \begin{cases} 1, & \text{if } \text{geom}_i \cap \text{geom}_j \neq \varnothing \\ 0, & \text{otherwise} \end{cases}$$

   This method captures more relationships than Rook contiguity Bivand et al. (2013), which requires shared edges only, and is suitable for the irregular boundaries of UK local authorities.

3. Adjacency Matrix Construction: The Queen contiguity weights were converted into a binary adjacency matrix $A \in \{0,1\}^{n \times n}$ where $A_{ij} = 1$ if i and j are neighbours. This forms the spatial view in our multi-view framework.

4. Spatial Correlation Analysis: For each LA i and cluster c, we computed the number of neighbours in cluster c, denoted $K_i^{(c)}$ (Spearman, 1904). Using Spearman's rank correlation (ρ), we examined the relationship between $K_i^{(c)}$ and the probability that i belongs to c. Strong positive correlations ($\rho > 0.7$) were found for multiple clusters (Figure 17), indicating that cluster membership is spatially autocorrelated.

5. Neighbour Majority Voting Test: We assigned each LA to the cluster most common among its direct neighbours (breaking ties arbitrarily) and compared these assignments to the ONS labels. This yielded a 27.5% mismatch rate (Figure 18), meaning that while spatial coherence is high, socio-economic features alone do not fully explain spatial patterns.

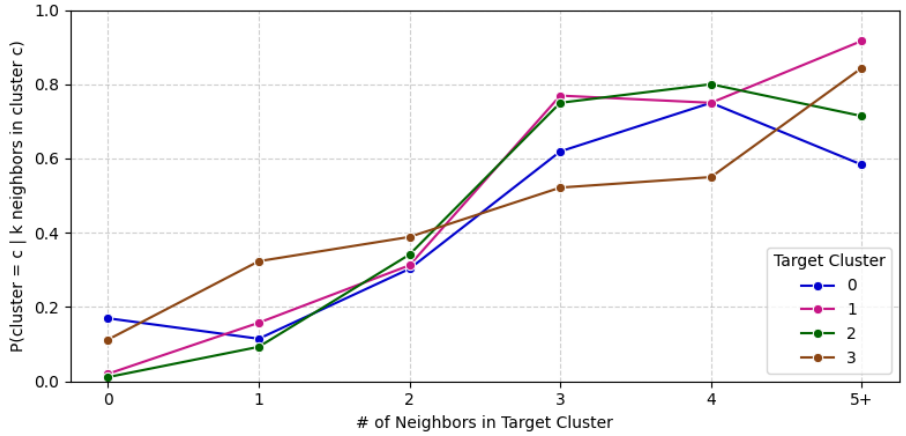

Figure 17.: Probability of being in a Cluster given the number of Neighbors in that Cluster

Interpretation. These results demonstrate that:

- Cluster membership is not spatially random; nearby LAs often share the same cluster, consistent with Tobler's First Law of Geography Tobler (1970).

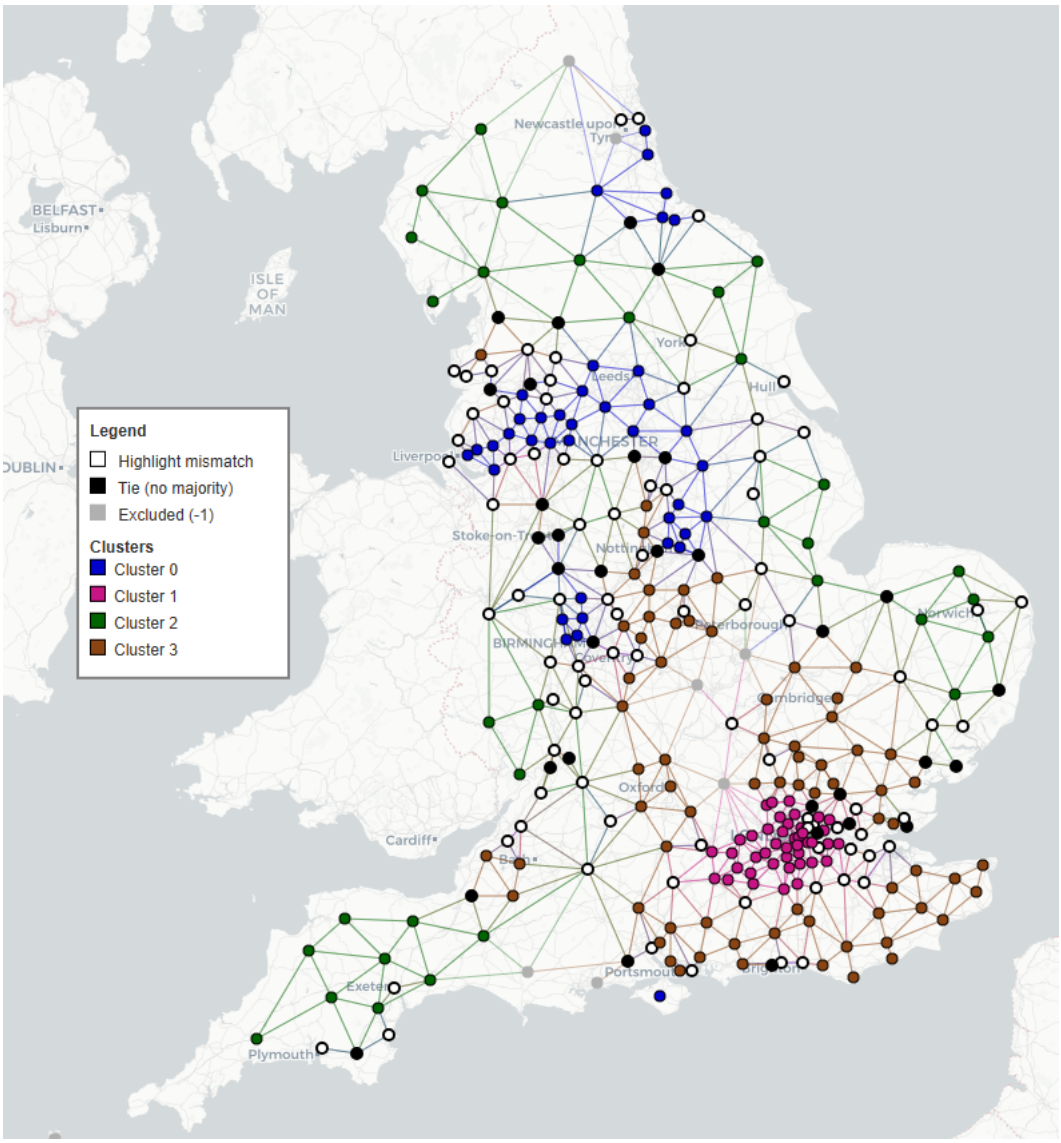

Figure 18.: Graph representation of each Local Authority by its geographical centroid. Cluster assignment is by neighbours' majority voting with ties or mismatch from real cluster label highlighted.

- Spatial adjacency does not perfectly explain ONS clusters, the mismatch rate implies that non-spatial socio-economic features drive differentiation even among neighbours.
- Including spatial proximity in multi-view clustering could encourage spatially coherent clusters, improving interpretability and stakeholder trust; reduce localised overfitting by smoothing noise at LA boundaries; and reveal residual patterns where socio-economic similarity does not align with geography.

Integration Options. Two main strategies could incorporate geography:

1. As an additional view: Treating the adjacency matrix as a similarity/distance view (e.g., masked cosine similarity, spatial kernel) and integrate via similarity fusion or as a dedicated branch in a multi-branch autoencoder.
2. As Spatial Regularisation: Introduce a spatial smoothness term in the loss function (e.g., Laplacian regularisation using the adjacency matrix) to encourage neighbouring LAs to have similar latent representations.

Motivation for a Spatial Feature View. The first integration option is implemented by constructing a spatial feature view that encapsulates morphological, locational, and connectivity characteristics of each LA. This treats spatial information as a distinct modality within the multi-view clustering pipeline, allowing its unique contribution to the shared representation to be explicitly quantified.

Beyond improving spatial coherence in clusters and reducing localised noise, this approach enhances interpretability: the contribution of the spatial view relative to other modalities can be assessed, and further decomposed to identify which spatial features (e.g., compactness, connectivity, remoteness) most strongly influence cluster formation. This supports transparent decision-making and provides stakeholders with concrete explanations for spatial patterns observed in the results.

Methodology. The spatial feature view is constructed as follows:

1. Boundary Data and CRS: Load the ONS 2021 LA boundaries (`.gpkg`) into a `GeoDataFrame` and ensure British National Grid (EPSG:27700) projection if missing.
2. Centroid Coordinates: Extract centroid longitude and latitude.
3. Geometric Properties: In projected coordinates, compute:

$$\begin{aligned}
\text{Area (km}^2\text{)} &= \frac{\text{polygon area (m}^2\text{)}}{10^6}, \\
\text{Perimeter (km)} &= \frac{\text{polygon perimeter (m)}}{10^3}, \\
\text{Compactness} &= \frac{4\pi \cdot \text{Area}}{\text{Perimeter}^2 + \epsilon}.
\end{aligned}$$

4. Topological Connectivity: Using `libpysal`, construct Queen contiguity weights (shared boundary or vertex) to obtain:

   - `deg_queen`: number of neighbours.
   - `is_island`: binary indicator for LAs with no neighbours.

5. Great-Circle k-NN Distances: Let (φ_i, λ_i) denote the latitude and longitude of LA i in radians. The great-circle distance between LAs i and j is computed via the haversine formula (Snyder, 1987):

$$d_{ij} = 2R \arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_j - \varphi_i}{2}\right) + \cos\varphi_i \cos\varphi_j \sin^2\left(\frac{\lambda_j - \lambda_i}{2}\right)}\right),$$

   where $R = 6371.0088$ km is the mean Earth radius. For $k \in \{1, 3, 5\}$, the mean k-NN distance for LA i is:

$$\bar{d}_i^{(k)} = \frac{1}{k} \sum_{m=1}^{k} d_{i,\,\pi_m(i)},$$

   where $\pi_m(i)$ is the m-th nearest neighbour of i. centroid distances and calculate mean distances to the $k = 1, 3, 5$ nearest neighbours.

By constructing a dedicated spatial feature view, we incorporate geographic form, connectivity, and remoteness into the multi-view clustering framework. This not only supports spatially coherent and policy-relevant clusters, but also enables quantification of the spatial modality's contribution to the shared representation, enhancing both methodological rigour and interpretability.

Results. When adding the spatial view to our dataset, encoding geographic features and proximity between local authorities, we found that most multi-view methods experienced a significant drop in silhouette score, and for others, performance stayed essentially unchanged, indicating that the spatial signal was being largely ignored by the clustering. Interestingly, despite observing strong correlations between geography and ONS cluster assignments, our findings suggest that geography may not have been used directly in their clustering either. This would explain why spatial proximity appears aligned to their clusters in practice, yet introduces noise rather than cohesion in our multi-view fusion framework, consistent with their poor silhouette outcomes when a spatial view is included.

To illustrate the change in results, Figure 19 shows the silhouette report of our champion model MVSF. Despite still providing a respectable performance, we can see that the best silhouette score we obtain is ≈ -0.2 that of the ones we find without the Spatial view. This is also confirmed by the view-ablation results in the top-right-hand corner of the report.

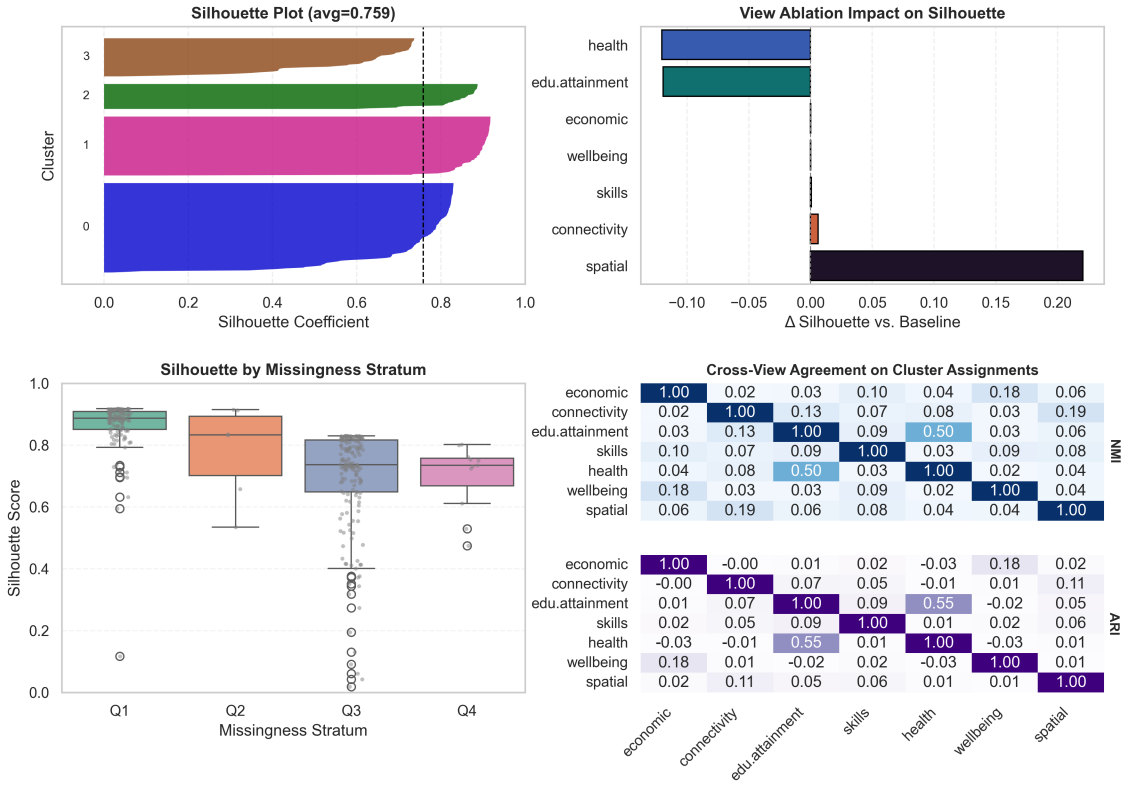

Figure 19.: MVSF Silhouette Report (including Spatial View).

Conclusions. Our attempt to extend the multi-view framework with a spatial feature view highlighted the challenges of incorporating geography into clustering of local authorities. While ONS clusters appear geographically coherent, our results show that explicitly adding spatial features tends to reduce cluster quality, with most methods exhibiting substantial drops in silhouette scores. This suggests that spatial coherence observed in ONS outputs may arise indirectly from socio-economic variables that are themselves spatially correlated, rather than from explicit use of spatial features in their methodology. For our framework, geography introduced noise rather than additional explanatory power, underscoring the difficulty of balancing spatial contiguity with socio-economic differentiation. Nonetheless, these exploratory results offer a valuable insight: spatial regularisation remains a promising but non-trivial extension that may require more sophisticated integration strategies to yield consistent benefits.

Appendix F2: Ablations

Tried DBScan as a clustering technique but it was too sensitive to epsilon which could not be tuned efficiently.

This ablation is shown in the notebook: `Dimensionality_Reduction_Clustering_GridSearch.ipynb`