# Representation Learning and Clustering for Multi-Domain Sequence Analysis: An Application to Dutch Register Data

Maximilian Reichert[0009−0003−8352−7288]

Erasmus School of Social and Behavioural Sciences, Erasmus University Rotterdam, Rotterdam 3062 PA, The Netherlands
reichert@essb.eur.nl

**Abstract.** Life-course theory emphasises that lives unfold across interdependent domains. Yet, standard tools for time series clustering are difficult to scale to large sample sizes observed through many concurrent life domains. To address this, I introduce a representation-learning workflow for multi-domain life-course analysis that scales to large samples and many indicators observed over time. Its central component is a bottlenecked Long Short-Term Memory (LSTM) autoencoder that compresses each person's multi-domain sequence into a global embedding used for clustering. Substantive interpretation is recovered by applying social sequence analysis for visualisation and sequence description. I demonstrate the workflow using administrative register data for the complete 2010 birth cohort of children in the Netherlands ($N$ = 165,204), observed yearly from age 1 to 15 across 18 contextual indicators, such as parental income, wealth, or union status. I cluster the learned global embeddings of multi-domain sequences into five childhood context types, which reveal a clear social gradient: advantaged clusters are marked by stable nuclear-family arrangements, owner-occupation, higher parental income and wealth, and greater childcare use, whereas disadvantaged clusters show more family and residential complexity, lower resources, and lower levels of childcare participation. This suggests that representation learning and clustering can recover substantive variation in childhood contexts. In future applications, this workflow may be expanded to link descriptive, explanatory, and predictive approaches in the study of life-course stratification.

**Keywords:** Multi-Domain Sequence Analysis; LSTM; Recurrent Neural Networks; Clustering; Life-Course Analysis

## 1 Introduction

Life-course theory conceptualises individual lives as unfolding across interdependent domains, where developments in family, work, health, housing, care, and economic resources may condition or reinforce one another over time [5]. This multi-domain perspective creates a methodological challenge. Methodological approaches should be capable of capturing how several life domains unfold jointly

within persons over time, while also identifying recurring patterns of longitudinal dynamics across a population. Person-centred approaches, such as sequence analysis, are central to this task because they identify subgroups of individuals characterised by similar trajectories in one or multiple indicators observed over time. Clustering these trajectories usually results in a typology that can be used descriptively, to inform theory about the prevalence of typical and atypical life-course configurations, or analytically, either to explain within-population heterogeneity given some baseline variable or to explain later outcomes given the uncovered heterogeneity [19]. In life-course research, sequence analysis is the most prominent approach in this person-centred tradition [2,19], but other approaches, such as group-based multi-trajectory modelling [21,9] or Mixture Hidden Markov Models are increasingly used as well [14]. Yet, the multi-domain perspective called for by life-course theory remains difficult to operationalise at scale with existing person-centred approaches [17].

The difficulty is both computational and substantive. Existing person-centred approaches, and especially sequence-analysis, struggle when both the number of observational units and the number of concurrent sequences per unit of observation increases [32,17]. In such settings, the challenge is to apply trajectory clustering while retaining interpretable descriptions of how domains combine within trajectories. As administrative data become more readily available and provide insight into increasingly complex life-course dynamics, the need to cluster and describe interdependencies among life-course sequences at population scale becomes more pressing.

At the same time, a growing literature treats individual life-courses as event sequences and uses machine learning to predict later outcomes [28,35,38,29]. These approaches are attractive because they scale to large longitudinal data and can learn compact representations of trajectories observed across multiple life domains [22,38]. Their main focus, however, is usually predictive performance rather than theory-driven description of patterns of substantive life-course dynamics within a population [27,30,29,34]. A key emerging finding is that the predictive performance of such models appears to be socially stratified, pointing to substantial within-population heterogeneity [20,37]. These learned representations could be useful for describing heterogeneity, but they are rarely connected back to the interpretive tools of social sequence analysis. Understanding and describing such heterogeneity at population scale remains a gap in the bridge between life-course theory and applications in life-course prediction.

To address this gap, this paper proposes a representation-learning workflow for representing, clustering, and interpreting multi-domain trajectories in population-scale administrative data. The workflow transforms multi-domain sequences into structured token sequences and applies an LSTM autoencoder [15] to encode each complete trajectory into a global embedding used for clustering. To retain substantive interpretation, the resulting cluster labels are linked back to the original domain-specific sequences and described using state-distribution plots and sequence-complexity indicators. The contribution is a workflow that

connects vector-based sequence representation with the person-centred and descriptive logic of social sequence analysis [24,17].

I demonstrate this workflow using Dutch register data [4] for the complete 2010 birth cohort of children in the Netherlands ($N = 165{,}204$), observed yearly from age 1 to age 15 across 18 contextual indicators, including family structure, housing, income, wealth, childcare use, and parental health. The empirical illustration shows that the learned embeddings can be clustered into types that differ systematically in family stability, housing tenure, parental resources, childcare participation, and sequence complexity. The remainder of the paper situates the proposed representation-based approach within existing person-centred life-course research, outlines the workflow, presents the empirical illustration, and concludes with limitations and extensions.

## 2 Background

The methodological problem addressed in this paper is how to construct interpretable person-centred typologies when both the number of individuals and the number of observed life domains are large. Administrative registers are especially relevant in this context because they cover entire populations, long observation windows, and often provide data on several concurrent life-course domains per individual. Clustering such data requires representations that preserve timing, ordering, duration, and cross-domain interdependence, while remaining computationally feasible and substantively interpretable [5,12,7,17].

A large share of life-course research relies on person-centred, as opposed to variable-centred analytical strategies. The aim of the person-centred approach is to identify subgroups (clusters) of individuals characterised by similar patterns of longitudinal dynamics in one or multiple observed indicators. This clustering of life-course trajectories usually results in a typology that serves different aims. First, it is used descriptively and to inform theory about the prevalence of specific types of trajectories [19]. Often, these trajectories are described based on the distribution of states at different points in time to identify typical and atypical life-course configurations. Second, the subsequent aim is either to explain within-population heterogeneity given some baseline variable, using clusters as outcomes, or to explain some outcome given the uncovered heterogeneity, using clusters as predictors. Thus far, these approaches have been used on a wide variety of questions and data, e.g., employment [18,23], income [36], health [9], or childcare use trajectories [11].

The most prominent approach is sequence analysis [2,19], which was adapted to the social sciences in close conversation with life-course theory development [1,2,3]. It treats trajectories as ordered sequences of states and provides tools for calculating sequence dissimilarities [24]. These dissimilarities can then be used for clustering and typology construction [33,31]. In standard applications, sequence clustering typically depends on a researcher-specified dissimilarity matrix. This requires choices about substitution, insertion, and deletion costs. These choices

are theoretically consequential because they determine which forms of similarity the clustering solution can recover.

These decisions become more difficult in multi-domain sequence analysis, where several concurrent life domains are analysed jointly [25]. Multi-domain sequence analysis usually proceeds either by combining domains into joint states or by aggregating domain-specific distances. The former preserves simultaneity but leads to rapidly expanding and sparse state spaces. The latter avoids joint-state explosion but requires assumptions about how similarities across domains should be weighted. Both strategies become increasingly difficult to operationalise and execute computationally when large populations are observed across many concurrent domains simultaneously [25]. For example, simply storing a $160{,}000 \times 160{,}000$ distance matrix requires roughly 204.8 GB of memory. The challenge is therefore both computational and interpretive: large dissimilarity matrices become difficult to compute, store, and cluster at full-population scale, while combined-state representations can become too sparse to describe meaningfully [25,32].

Other person-centred approaches also provide ways to study several longitudinal outcomes jointly. Group-based multi-trajectory modelling [21], for example, allows multiple continuous and repeated outcomes in different domains to define latent groups through sets of trajectories [21,36]. These models are useful for deriving interpretable typologies and relating group membership to covariates, but they usually require a limited number of outcomes and relatively restrictive assumptions about functional form and outcome distributions. Mixture Markov Models and Mixture Hidden Markov Models [14] instead approach categorical sequences from a probabilistic and transition-focused perspective, modelling state changes as Markov processes that may differ across latent subpopulations or latent states [8]. These models can represent latent dynamics, transition structures, and missingness. However, when the number of domains, states, latent classes, and therefore possible transitions increases, they can become difficult to interpret, especially when the aim is exploratory description of within-population heterogeneity.

One way to address these limits is to replace pairwise sequence dissimilarities, the basis for clustering in sequence analysis, with vector representations (also called features or embeddings). Khalifa et al. review several approaches, agnostic of discipline [17]. But also within sequence analysis, the idea of representing sequences as vectors [10] is not new. For example, Elzinga's combinatorial vector representation maps token sequences into Euclidean space by enumerating subsequences and common precedence relations, so that similarity reflects shared ordered structure. In this approach, the vector representation is calculated algorithmically from the original sequence.

Recent work on life-course prediction, natural language processing, and time-series clustering suggests a second approach, namely learning embeddings from sequence data using sequence models, such as recurrent neural networks or transformers [6,17]. LSTMs, for example, are recurrent neural networks designed to process ordered data by updating a hidden state over time while retaining infor-

mation from earlier events that is relevant for later reconstruction or prediction [15]. For life-course data, this hidden state can represent between-person differences in timing, order, duration, and co-occurrence of states. The hidden state observed at the end of a processed sequence provides a fixed-length representation of the complete observed trajectory, which can then be used for downstream tasks such as clustering, visualisation, or prediction.

The workflow proposed here uses this idea for exploratory multi-domain sequence analysis. It learns one global embedding per individual. Individuals are clustered in the learned representation space, and the resulting groups are interpreted by describing the original domain-specific sequences. This combination is intended to preserve the person-centred and interpretive strengths of social sequence analysis while avoiding the main computational bottleneck of large multi-domain dissimilarity matrices.

## 3 Framework

The workflow consists of five steps: first, transform multi-domain sequences into discrete tokens; second, arrange tokens into structured yearly sequences; third, compress each full trajectory into a global embedding using an LSTM autoencoder; fourth, cluster individuals based on these embeddings; and fifth, interpret the clusters by returning to the original domain-specific sequences using visualisations and sequence-complexity indicators. Figure 1 summarises the workflow from balanced panel input, through tokenisation and representation learning, to clustering, trajectory-type description, and external validation. The empirical application in Section 4 demonstrates the clustering and interpretation of trajectory types, but stops short of the explanatory or predictive analysis stage, which would involve linking the clusters to baseline covariates or outcomes.

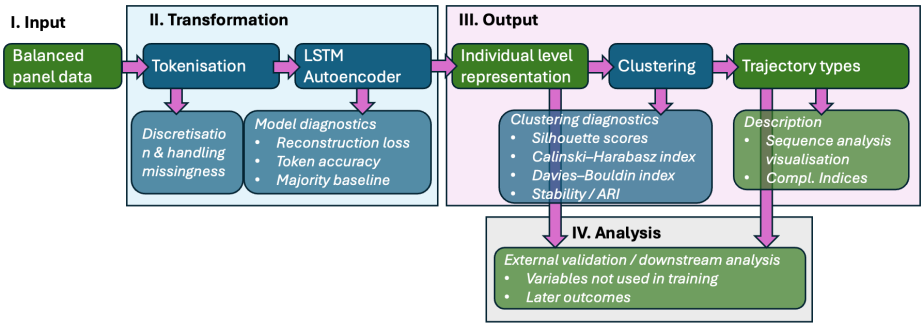


**Fig. 1.** Workflow for representation learning, clustering, and interpretation of multi-domain trajectories.

### 3.1 Data transformation

The first step transforms a balanced panel into an integer-encoded token matrix for the LSTM autoencoder. The procedure starts from a long-format administrative data file in which each row corresponds to one observation of one person at one point in time $t$. Preprocessing requires substantive and technical decisions about how raw variable values are translated into sequence states. This includes discretising continuous or count-based variables and explicitly handling missingness before tokenisation. Continuous variables may be binned into quantiles or manually discretised into sociologically meaningful categories based on domain expertise.

After discretisation and missing-value handling, every person-$t$-variable cell is converted into a token of the form `variable=value`. These tokens (which would be the combined alphabet in sequence analysis [25]) are mapped to one unique integer each, the token IDs. For each individual, the final model input is an ordered sequence of all token IDs within each $t$ over all observed values of $t$. Individuals are then randomly split at the person level into training, validation, and test sets.

### 3.2 LSTM autoencoder

The integer-encoded person-$t$ token sequences are compressed using a bottlenecked LSTM autoencoder. LSTMs are computationally lighter than transformer architectures for the modest sequence lengths considered here, preserve temporal ordering, and yield hidden states at each $t$ that can be inspected over time [15], but other sequence models, such as transformers [37], may be explored in future applications.

Each token ID is first mapped to a learned token embedding. Within each $t$, the $V$ token embeddings are concatenated and projected through a linear layer with nonlinear activation to obtain a fixed-length vector representation for each $t$. This produces one vector representation for each individual at each $t$. The encoder, a unidirectional LSTM, reads these vectors in temporal order. For each individual and at each $t$, it produces a contextual hidden state $h_{it}$, summarising the observed sequence up to and including $t$. I refer to the final hidden state $h_{iT}$ as the global embedding, which represents the complete observed multi-domain trajectory and serves as input to the downstream clustering step.

A central design feature is the global bottleneck. The decoder receives only the final encoder state $h_{iT}$, repeated across the time dimension, and attempts to reconstruct the original sequence of tokens. The same global embedding initialises the decoder hidden state, while the decoder cell state is initialised at zero. This prevents the decoder from reconstructing time point $t$ directly from the corresponding encoder state $h_{it}$. Successful reconstruction therefore requires the encoder to compress information about the complete multi-domain trajectory into the global embedding.

The decoder produces token logits for every time point and variable position. The reconstruction objective is a token-level cross-entropy loss between observed

and predicted tokens across individuals, time points, and variable positions. The objective also includes a temporal smoothing penalty on consecutive encoder states, discouraging unnecessary representational drift while retaining changes that improve reconstruction. The model therefore outputs both a global embedding for each complete person-sequence $h_{iT}$ and time-specific embeddings $h_{it}$ that could be used to trace latent change over time. Diagnostics include reconstruction loss, token-level reconstruction accuracy, and comparison to a majority baseline.

### 3.3 Clustering

Each individual is represented by one global embedding summarising the complete observed multi-domain trajectory. This embedding is the input to the clustering stage. In the empirical illustration below, clustering is performed on the global embedding $h_{iT}$, the final hidden state of the upper encoder layer. Depending on the empirical structure of the embedding space, suitable algorithms include $k$-means, hierarchical clustering, Gaussian mixture models, or density-based approaches such as HDBSCAN, potentially combined with UMAP. The clustering step can be evaluated using diagnostics such as silhouette scores, the Calinski–Harabasz index, the Davies–Bouldin index, and stability checks based on the Adjusted Rand Index [31].

### 3.4 Interpretation and visualisation

The learned embedding space can be inspected by projecting the high-dimensional global embeddings onto two dimensions using methods such as UMAP or t-SNE [39]. These projections may provide a useful visual check of whether the encoder has learned a structured latent space, for example through gradients, dense regions, or separated groups. As an external validation check, the embedding projection and cluster membership can be compared with baseline variables that were not included in model training. Systematic variation in such held-out variables across the embedding space or clusters provides an exploratory indication that the global embeddings capture substantively meaningful structure beyond the reconstruction task.

The embedding-derived clusters may be interpreted by linking cluster labels back to the original domain-specific sequences. This allows the resulting typology to be described with standard sequence analysis tools, i.e., sequence distribution plots or sequence complexity indexes [13,24].

In sum, this workflow supports exploratory, person-centred typology construction. The neural model compresses the multi-domain trajectories, clustering is performed using the algorithm of choice [17], and substantive interpretation is restored by returning to the original sequences and describing them with tools from social sequence analysis [19].

## 4 Empirical Illustration

### 4.1 Data

The empirical illustration uses a variety of Dutch registers provided by Statistics Netherlands [4] to construct an annual balanced child-year panel for the complete 2010 birth cohort ($N$ = 165,204). Each child is observed annually from age 1 through age 15, yielding 15 observations per child and 2,478,060 child-year records; children who leave the Netherlands during the observation period are excluded. For each child-year, I construct 18 indicators describing the child's context, including residential mobility, housing tenure, household composition, family structure, income, wealth, childcare use, and parental health. Figure 2 shows a selection. These indicators combine information linked through the child's registered household with information linked through biological relations to parents, so that parental resources can be retained even when co-residence changes.

The resulting panel is transformed into the tokenised input used by the model as outlined in Section 3.1. Each person-year-variable cell is converted into a token, and all observed tokens are mapped to token IDs. Within each year, tokens are arranged in a fixed variable order. This integer-encoded sequence is the input to the LSTM autoencoder.

### 4.2 Model and clustering

I train the bottlenecked LSTM autoencoder on the integer-encoded child-year sequences and extract one global embedding per child from the final encoder state. The model uses a token embedding dimension of 32. For each child-year, the 18 token embeddings are concatenated and projected through a linear layer with ReLU activation to a 64-dimensional yearly representation. The encoder is a two-layer unidirectional LSTM with hidden size 64 and dropout of 0.1. The final hidden state $h_{iT}$ of the upper encoder layer is used as the 64-dimensional global embedding. The decoder receives only this global embedding, repeated across the time dimension, and reconstructs the original tokens for every year and variable position. It is initialised with the global embedding as hidden state and a zero cell state, enforcing the bottleneck between encoder and decoder.

The model is trained with the Adam optimiser using a learning rate of 0.001, a batch size of 512, and a 70/15/15 train-validation-test split. The reconstruction objective is token-level cross-entropy loss, augmented by a temporal smoothing penalty on consecutive encoder states with weight $\lambda = 0.0005$. Training was stable over 50 epochs. The validation loss decreased to 0.2742, and the held-out test loss remained nearly identical at 0.2747, suggesting good generalisation and no evidence of overfitting. Training took approximately 1.5 hours on standard servers provided by Statistics Netherlands (CPU-based training, dual-core processor, 32 GB RAM). After training, I extract one global embedding for each child.

For clustering, I robust-scale the 64-dimensional global embeddings and estimate $k$-means and bisecting $k$-means solutions for several values of $k$. The

empirical illustration uses the bisecting $k$-means solution with $k = 5$, selected on the basis of clustering diagnostics and substantive interpretability. Bisecting $k$-means yields higher silhouette scores than standard $k$-means for all values of $k$ considered. The $k = 5$ solution is also a local optimum according to the Calinski–Harabasz and Davies–Bouldin indices in both clustering algorithms. Substantive interpretation is based on linking the cluster labels back to the original domain-specific sequences shown in Figure 2.

### 4.3 Results

The five clusters each represent a sizeable share of the cohort and capture combinations of, among others, family, housing, economic, care, and health conditions that recur across childhood (see Figure 2 for sequence distribution plots for select indicators, and Table 1 for cluster sizes and sequence complexity measures).

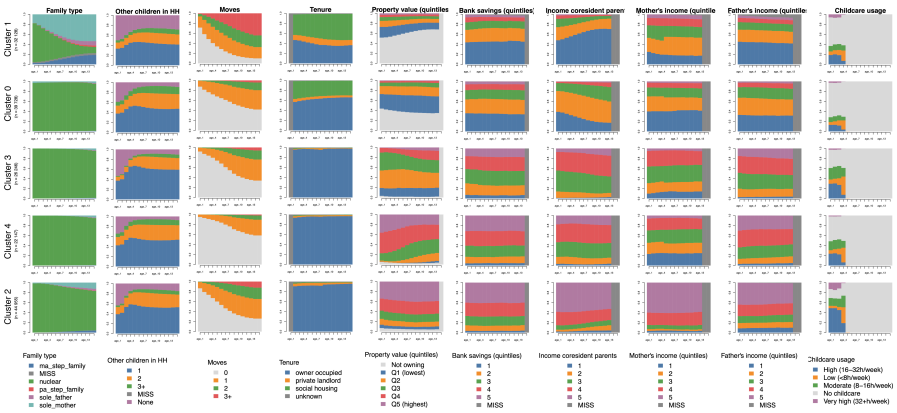


**Fig. 2.** Cluster-specific multi-domain sequence distributions from age 1 to age 15. Rows correspond to the clusters identified from the global embeddings; absolute cluster size is reported on the left, and relative shares are shown in Table 1. Clusters are ordered substantively from less to more advantaged contexts; columns show the age-specific distribution of states for selected indicators. Stacked areas represent the proportion of children in each categorical state at each age.

The descriptive plots in Figure 2 relate the clustering solution back to the original multi-domain trajectories. The sequence distribution plots [13] suggest that the learned representations capture a clear social gradient in childhood contexts. At one end are children growing up in comparatively stable and resource-rich contexts (Clusters 4, 2): predominantly nuclear families, high rates of owner occupation, high property values, higher bank savings, higher parental incomes, relatively little residential mobility, and high rates of childcare use. At the other end are children exposed to more complex and less resource-rich contexts (Clusters 1, 0), including higher shares of sole-mother or step-family arrangements,

**Table 1.** Cluster size and mean sequence complexity by domain. Clusters are ordered substantively from less to more advantaged contexts.

| Cluster | *N* | Share (%) | Family Complexity | Housing Complexity | Income Complexity | Wealth Complexity | Childcare Complexity | Health Complexity |
|---|---|---|---|---|---|---|---|---|
| 1 | 32,128 | 19.4 | 0.144 | 0.277 | 0.360 | 0.289 | 0.150 | 0.562 |
| 0 | 39,728 | 24.0 | 0.058 | 0.166 | 0.353 | 0.284 | 0.094 | 0.567 |
| 3 | 26,246 | 15.9 | 0.051 | 0.164 | 0.366 | 0.378 | 0.187 | 0.562 |
| 4 | 22,147 | 13.4 | 0.048 | 0.136 | 0.372 | 0.349 | 0.145 | 0.549 |
| 2 | 44,955 | 27.2 | 0.070 | 0.185 | 0.323 | 0.357 | 0.250 | 0.560 |

overall fewer co-resident children, lower income and wealth, higher exposure to rental or social housing, lower childcare use, and more frequent moves. One intermediate cluster combines stable family arrangements with more moderate economic resources or distinctive childcare-use patterns (Cluster 3). Cluster labels do not imply an ordinal ordering. A supplementary external validation check using a held-out baseline variable not included in model training also supported this interpretation, as the variable varied systematically across the embedding space and differentiated the resulting clusters.

Table 1 summarises cluster size and mean sequence complexity across sequences grouped by substantive life-course domain. Sequence complexity is a standardised longitudinal measure reflecting the number of transitions and the diversity of the state space within individual sequences. While mean sequence complexity varies only modestly across most domains, the table shows that the social gradient identified in the sequence plots is also reflected in higher instability, especially in the family and housing complexity of Cluster 1.

## 5 Discussion and outlook

### 5.1 Contribution

This paper contributes a representation-learning workflow for trajectory clustering and interpretation. Methodologically, it shows how multi-domain sequences can be compressed into embeddings, clustered, and interpreted by returning to the original domain-specific sequences. This preserves the person-centred logic of sequence analysis while making it feasible to analyse many concurrent life domains for an entire cohort in register-based settings.

I show that childhood contexts are stratified as multidimensional configurations of inequalities in income, family structure, housing, or childcare. This extends previous research that operationalises childhood contexts as unidimensional or multidimensional count measures or indexes (for a review see [16]). Empirically, the illustration shows that these domains combine into distinct and interpretable childhood context types.

The added value of the method is that this conclusion depends on observing many domains jointly and longitudinally. A single-domain analysis could show clusters with lower income, different family forms, or housing disadvantage, but it would not show how these dimensions combine into multi-domain trajectory types. Likewise, conventional multi-domain sequence analysis would computationally struggle to recover this structure across 18 concurrent indicators for an

entire population cohort. The workflow therefore enables a methodological conclusion that is difficult to reach with conventional multi-domain sequence analysis at this scale: representation learning can recover interpretable, person-centred configurations of life-course trajectories across many concurrent domains. Explaining how and to what extent baseline variables are associated with these configurations, and whether they might be predictive of later life outcomes, is left for future substantive research.

### 5.2 Limitations

Several limitations should be noted. First, the workflow is exploratory and the resulting clusters in the empirical example should be interpreted as a data-driven typology rather than as fixed latent classes. The embedding space and clustering solution may depend on several components of the workflow, including preprocessing decisions, model architecture, embedding dimension, smoothing strength, dimensionality reduction, clustering algorithm, and random initialisation. A full benchmarking exercise across alternative architectures and clustering pipelines is beyond the scope of this conference paper.

Second, the empirical illustration is limited in that it serves as a proof of concept, not a fully fledged sociological analysis. Several additional diagnostics laid out in Figure 1, such as broader model comparisons, extended stability checks, and more extensive external validation, would be useful to gauge the robustness of the results. For example, the illustration does not include a full comparison against conventional multi-domain sequence analysis pipelines, such as optimal matching or other dissimilarity-based clustering strategies. Such comparisons are computationally demanding at the full population scale and require additional decisions about multi-domain costs and domain weighting [32], but would be feasible on tractable subsamples or simulated datasets.

Third, the autoencoder is evaluated primarily through reconstruction diagnostics and downstream interpretability. Although reconstruction accuracy indicates that the global embeddings preserve substantial information about the original multi-domain sequences, it does not by itself guarantee that the resulting clusters are substantively optimal. Future applications should therefore combine reconstruction metrics, cluster stability diagnostics, external validation variables, and substantive domain knowledge.

### 5.3 Outlook

The proposed framework opens up several avenues for methodological development and substantive research. First, the workflow directly extends the scope of multi-domain sequence analysis in register-based contexts. It provides a scalable alternative to the computationally demanding task of calculating large dissimilarity matrices across numerous domains and individuals. Future work can combine the strengths of both approaches by using representation learning to handle initial sequence compression and clustering, and then applying the descriptive

sequence-analysis toolbox such as indicators of timing, order, duration, transition rates, and longitudinal complexity to examine resulting groups in detail [26,32].

Second, the framework can be extended to develop more fine-grained, model-specific interpretability tools in contexts of life-course prediction. For example, variable-importance measures could be calculated and averaged by cluster, variable, and year. These measures would identify which variables at which point in time contribute most to the final hidden state $h_{iT}$ which is used for clustering. This could reveal, for example, whether a cluster is primarily defined by early family structure, subsequent housing moves, or late-childhood economic resources.

Third, learned sequence embeddings can help diagnose socially stratified predictability in life-course models [20,37] by showing whether predictive performance differs across clusters. To this end, relative variable importance within each $t$ and domain could be calculated not relative to the final hidden state but to an outcome that the researcher is interested in predicting. Alternatively, the LSTM may be trained in a supervised manner on a specific outcome. This would still produce a hidden state that may be clustered, and would shed light on whether pathways with a specific outcome are meaningfully differentiated.

Finally, future applications can leverage the model's yearly hidden states. By extracting and tracing the yearly encoder states $h_{it}$ over time, researchers can map individual trajectories as continuous paths through the learned life-course space. Distances between consecutive time-specific embeddings could then be used as dynamic indicators of instability, convergence, or divergence. This would enable an exploratory, person-centred analysis of life-course tempo, drawing on established methods for tracing semantic shifts in temporal language models [6]. Taken together, these directions define a broader research agenda in which learned sequence representations are used to compress and cluster life-course data in register-based contexts, and to connect description, explanation, and prediction in the study of social stratification over time.

**Acknowledgments.** I am thankful for the continued feedback and support of Julia Leesch, Linda Vecgaile, and Tom Emery, and for the hospitality of Nicole Hiekel at the Max Planck Institute for Demographic Research. This work is supported by European Research Council grant Childcare Strategies (Grant agreement ID:101042662). Funded by the European Union. Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. The data used in this study are from Statistics Netherlands Microdata [4]. The analyses were performed in CBS environment 9424. Code will be made available via the ODISSEI Code Library upon publication.

**Disclosure of Interests.** The author has no competing interests to declare that are relevant to the content of this article.

## References


1. Abbott, A., Forrest, J.: Optimal Matching Methods for Historical Sequences. The Journal of Interdisciplinary History **16**(3), 471–494 (1986). https://doi.org/10.2307/204500
2. Abbott, A., Tsay, A.: Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect. Sociological Methods & Research **29**(1), 3–33 (Aug 2000). https://doi.org/10.1177/0049124100029001001
3. Abbott, A.D.: Processual Sociology. The University of Chicago Press, Chicago ; London (2016)
4. Bakker, B.F., Van Rooijen, J., Van Toor, L.: The system of social statistical datasets of Statistics Netherlands: An integral approach to the production of register-based social statistics. Statistical Journal of the IAOS **30**(4), 411–424 (2014)
5. Bernardi, L., Huinink, J., Settersten, R.A.: The life course cube: A tool for studying lives. Advances in Life Course Research **41**, 100258 (Sep 2019). https://doi.org/10.1016/j.alcr.2018.11.004
6. Chen, Y., Perozzi, B., Al-Rfou, R., Skiena, S.: The Expressive Power of Word Embeddings (May 2013). https://doi.org/10.48550/arXiv.1301.3226
7. Cullati, S., Kliegel, M., Widmer, E.: Development of reserves over the life course and onset of vulnerability in later life. Nature Human Behaviour **2**(8), 551–558 (Aug 2018). https://doi.org/10.1038/s41562-018-0395-3
8. Eberlein, L., Pavlopoulos, D., Garnier-Villarreal, M.: Starting flexible, always flexible? The relation of early temporary employment and young workers employment trajectories in the Netherlands. Research in Social Stratification and Mobility **89**, 100861 (Feb 2024). https://doi.org/10.1016/j.rssm.2023.100861
9. Elsenburg, L.K., Rieckmann, A., Bengtsson, J., Jensen, A.K., Rod, N.H.: Application of life course trajectory methods to public health data: A comparison of sequence analysis and group-based multi-trajectory modeling for modelling childhood adversity trajectories. Social Science & Medicine **340**, 116449 (Jan 2024). https://doi.org/10.1016/j.socscimed.2023.116449
10. Elzinga, C.H.: Combinatorial Representations of Token Sequences. Journal of Classification **22**(1), 87–118 (Jun 2005). https://doi.org/10.1007/s00357-005-0007-6
11. Emery, T.: Inequalities in early childcare strategies: Evidence from Dutch administrative data. Advances in Life Course Research **67**, 100727 (Mar 2026). https://doi.org/10.1016/j.alcr.2026.100727
12. Fasang, A.E., Gruijters, R.J., Van Winkle, Z.: The life course boat: A theoretical framework for analyzing variation in family lives across time, place, and social location. Journal of Marriage and Family p. jomf.13012 (Jun 2024). https://doi.org/10.1111/jomf.13012
13. Gabadinho, A., Ritschard, G., Müller, N.S., Studer, M.: Analyzing and Visualizing State Sequences in R with TraMineR. Journal of Statistical Software **40**, 1–37 (Apr 2011). https://doi.org/10.18637/jss.v040.i04
14. Helske, S., Helske, J.: Mixture Hidden Markov Models for Sequence Data : The seqHMM Package in R. Journal of Statistical Software **88**(0) (2019). https://doi.org/10.18637/jss.v088.i03
15. Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Computation **9**(8), 1735–1780 (Nov 1997). https://doi.org/10.1162/neco.1997.9.8.1735
16. Kazakova, Y., Leturcq, M., Panico, L.: Multidimensional Child Deprivation: Constructing Longitudinal Indicators for the Early Childhood Period. Child

Indicators Research **17**(6), 2629–2669 (Dec 2024). https://doi.org/10.1007/s12187-024-10178-7
17. Khalifa, O., Balendran, A., Tran, V.T., Petit, F.: Clustering methods for categorical time series and sequences : A scoping review. BMC Medical Research Methodology **26**(1), 123 (May 2026). https://doi.org/10.1186/s12874-026-02857-6
18. Killewald, A., Zhuo, X.: U.S. Mothers' Long-Term Employment Patterns. Demography **56**(1), 285–320 (Feb 2019). https://doi.org/10.1007/s13524-018-0745-9
19. Liao, T.F., Bolano, D., Brzinsky-Fay, C., Cornwell, B., Fasang, A.E., Helske, S., Piccarreta, R., Raab, M., Ritschard, G., Struffolino, E., Studer, M.: Sequence analysis: Its past, present, and future. Social Science Research **107**, 102772 (Sep 2022). https://doi.org/10.1016/j.ssresearch.2022.102772
20. Lundberg, I., Brown-Weinstock, R., Clampet-Lundquist, S., Pachman, S., Nelson, T.J., Yang, V., Edin, K., Salganik, M.J.: The origins of unpredictability in life outcome prediction tasks. Proceedings of the National Academy of Sciences **121**(24), e2322973121 (Jun 2024). https://doi.org/10.1073/pnas.2322973121
21. Nagin, D.S., Jones, B.L., Passos, V.L., Tremblay, R.E.: Group-based multi-trajectory modeling. Statistical Methods in Medical Research **27**(7), 2015–2023 (Jul 2018). https://doi.org/10.1177/0962280216673085
22. Rahal, C., Verhagen, M., Kirk, D.: The rise of machine learning in the academic social sciences. AI & SOCIETY **39**(2), 799–801 (Apr 2024). https://doi.org/10.1007/s00146-022-01540-w
23. Reichert, M., Emery, T., Bartova, A.: The critical juncture of childbirth: Turbulence in the employment trajectories of Mothers in Europe (Dec 2024). https://doi.org/10.31235/osf.io/y4f3j
24. Ritschard, G.: Measuring the Nature of Individual Sequences. Sociological Methods & Research **52**(4), 2016–2049 (Nov 2023). https://doi.org/10.1177/00491241211036156
25. Ritschard, G., Liao, T.F., Struffolino, E.: Strategies for Multidomain Sequence Analysis in Social Research. Sociological Methodology **53**(2), 288–322 (Aug 2023). https://doi.org/10.1177/00811750231163833
26. Rowold, C., Struffolino, E., Fasang, A.E.: Life-Course-Sensitive Analysis of Group Inequalities: Combining Sequence Analysis With the Kitagawa–Oaxaca–Blinder Decomposition. Sociological Methods & Research **54**(2), 646–705 (May 2025). https://doi.org/10.1177/00491241231224226
27. Salganik, M.J., Lundberg, I., Kindel, A.T., Ahearn, C.E., Al-Ghoneim, K., Almaatouq, A., Altschul, D.M., Brand, J.E., Carnegie, N.B., Compton, R.J., Datta, D., Davidson, T., Filippova, A., Gilroy, C., Goode, B.J., Jahani, E., Kashyap, R., Kirchner, A., McKay, S., Morgan, A.C., Pentland, A., Polimis, K., Raes, L., Rigobon, D.E., Roberts, C.V., Stanescu, D.M., Suhara, Y., Usmani, A., Wang, E.H., Adem, M., Alhajri, A., AlShebli, B., Amin, R., Amos, R.B., Argyle, L.P., Baer-Bositis, L., Büchi, M., Chung, B.R., Eggert, W., Faletto, G., Fan, Z., Freese, J., Gadgil, T., Gagné, J., Gao, Y., Halpern-Manners, A., Hashim, S.P., Hausen, S., He, G., Higuera, K., Hogan, B., Horwitz, I.M., Hummel, L.M., Jain, N., Jin, K., Jurgens, D., Kaminski, P., Karapetyan, A., Kim, E.H., Leizman, B., Liu, N., Möser, M., Mack, A.E., Mahajan, M., Mandell, N., Marahrens, H., Mercado-Garcia, D., Mocz, V., Mueller-Gastell, K., Musse, A., Niu, Q., Nowak, W., Omidvar, H., Or, A., Ouyang, K., Pinto, K.M., Porter, E., Porter, K.E., Qian, C., Rauf, T., Sargsyan, A., Schaffner, T., Schnabel, L., Schonfeld, B., Sender, B., Tang, J.D., Tsurkov, E., van Loon, A., Varol, O., Wang, X., Wang, Z., Wang, J., Wang, F., Weissman, S., Whitaker, K., Wolters, M.K., Woon, W.L., Wu, J., Wu, C., Yang, K., Yin,

J., Zhao, B., Zhu, C., Brooks-Gunn, J., Engelhardt, B.E., Hardt, M., Knox, D., Levy, K., Narayanan, A., Stewart, B.M., Watts, D.J., McLanahan, S.: Measuring the predictability of life outcomes with a scientific mass collaboration. Proceedings of the National Academy of Sciences **117**(15), 8398–8403 (Apr 2020). https://doi.org/10.1073/pnas.1915006117

28. Savcisens, G., Eliassi-Rad, T., Hansen, L.K., Mortensen, L.H., Lilleholt, L., Rogers, A., Zettler, I., Lehmann, S.: Using sequences of life-events to predict human lives. Nature Computational Science **4**(1), 43–56 (Jan 2024). https://doi.org/10.1038/s43588-023-00573-5
29. Shmatko, A., Jung, A.W., Gaurav, K., Brunak, S., Mortensen, L.H., Birney, E., Fitzgerald, T., Gerstung, M.: Learning the natural history of human disease with generative transformers. Nature pp. 1–9 (Sep 2025). https://doi.org/10.1038/s41586-025-09529-3
30. Sivak, E., Pankowska, P., Mendrik, A., Emery, T., Garcia-Bernardo, J., Hocuk, S., Karpinska, K., Maineri, A., Mulder, J., Nissim, M., Stulp, G.: Combining the Strengths of Dutch Survey and Register Data in a Data Challenge to Predict Fertility (PreFer) (Mar 2024). https://doi.org/10.48550/arXiv.2402.00705
31. Studer, M.: WeightedCluster Library Manual: A practical guide to creating typologies of trajectories in the social sciences with R (2013). https://doi.org/10.12682/LIVES.2296-1658.2013.24
32. Studer, M.: Sequence Analysis for Large Databases (2024). https://doi.org/10.12682/LIVES.2296-1658.2024.104
33. Studer, M., Ritschard, G., Gabadinho, A., Müller, N.S.: Discrepancy Analysis of State Sequences. Sociological Methods & Research **40**(3), 471–510 (Aug 2011). https://doi.org/10.1177/0049124111415372
34. Vafa, K., Athey, S., Blei, D.M.: Estimating wage disparities using foundation models. Proceedings of the National Academy of Sciences **122**(22), e2427298122 (Jun 2025). https://doi.org/10.1073/pnas.2427298122
35. Vafa, K., Palikot, E., Du, T., Kanodia, A., Athey, S., Blei, D.M.: CAREER: A Foundation Model for Labor Sequence Data (Feb 2024). https://doi.org/10.48550/arXiv.2202.08370
36. Vargha, L.: Fathers' and mothers' joint longitudinal employment patterns around first birth in Germany, 1990-2020 (Oct 2025). https://doi.org/10.31235/osf.io/8ekth_v2
37. Vecgaile, L., Spata, A., Vecchietti, L., Zagheni, E.: Predicting Individual Life Trajectories: Addressing Uncertainty in Social Employment Transitions (Jun 2025). https://doi.org/10.31235/osf.io/7ut9m_v1
38. Verhagen, M.D., Stroebl, B., Liu, T., Liu, L.T., Salganik, M.J.: The Book of Life approach: Enabling richness and scale for life course research (Jul 2025). https://doi.org/10.48550/arXiv.2507.03027
39. Wang, Y., Huang, H., Rudin, C., Shaposhnik, Y.: Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMap, and PaCMAP for Data Visualization. Journal of Machine Learning Research **22**(201), 1–73 (2021)