Dlugosz, Stephan

**Working Paper**

# Clustering life trajectories: A new divisive hierarchical clustering algorithm for discrete-valued discrete time series

ZEW Discussion Papers, No. 11-015

**Provided in Cooperation with:**
ZEW - Leibniz Centre for European Economic Research

*Suggested Citation:* Dlugosz, Stephan (2011) : Clustering life trajectories: A new divisive hierarchical clustering algorithm for discrete-valued discrete time series, ZEW Discussion Papers, No. 11-015, Zentrum für Europäische Wirtschaftsforschung (ZEW), Mannheim, https://nbn-resolving.de/urn:nbn:de:bsz:180-madoc-31434

This Version is available at:
https://hdl.handle.net/10419/44458

Discussion Paper No. 11-015

# Clustering Life Trajectories

## A New Divisive Hierarchical Clustering Algorithm for Discrete-valued Discrete Time Series

Stephan Dlugosz

ZEW
Zentrum für Europäische Wirtschaftsforschung GmbH
Centre for European Economic Research

Discussion Paper No. 11-015

# Clustering Life Trajectories

## A New Divisive Hierarchical Clustering Algorithm for Discrete-valued Discrete Time Series

Stephan Dlugosz

**Non Technical Summary**

The goal of clustering is to partition the objects (here: persons) into groups which ideally should be homogenous and well-separated (i.e. low within-group and high between-groups heterogeneity). In the context of data preparation, cluster analysis reveals underlying structures in the data to provide subgroups that are easier to handle in subsequent steps of the data analysis.

Most clustering techniques use a distance or dissimilarity matrix. Finding a good—i.e. interpretable—distance measure for a particular clustering task is hard. Therefore, a more direct approach to clustering might perform better: Homogeneity and heterogeneity describe the two extreme points of a measure of dispersion.

In this paper, a new clustering procedure for discrete-time discrete-valued life course trajectories is introduced that does not depend on a dissimilarity measure but on dispersions. The applied measure of dispersion has to deal with nominal data appropriately. Moreover, a discrete measure of association is needed to cope with the dependency structure of the time series. Both measures are discussed, a model for clustering discrete time series is introduced and the applicability of the new algorithm is demonstrated on a quite large data set from the German pension insurance.

This paper offers a technical foundation for accounting for the heterogenous histories of the participants of observational studies with greater precision, without immediately being confronted with the problems of dimensionality. This is particularly useful for policy evaluation.

### Das Wichtigste in Kürze

Die Cluster-Analyse ist eine Technik, um Objekte (hier: Personen) in Gruppen einzuteilen, welche idealerweise homogen in sich und heterogen untereinander sind. Im Rahmen der Datenaufbereitung erreicht man so eine Sytematisierung der Beobachtungen einer ansonsten schwer beherrschbaren Grundgesamtheit, auf welche in den nachfolgenden Analyseschritten Bezug genommen werden kann.

Die meisten Clustertechniken basieren auf der Definition von Distanzmaßen, aber die Festlegung eines guten—d.h. interpretierbaren—Maßes ist schwierig. Aus diesem Grund können direktere Methoden, welche auf Streuungsmaßen basieren, häufig leichter interpretierbare Ergebnisse erzielen, insbesondere wenn die Variablen kategoriell sind: Homogenität und Heterogenität können dann als die zwei Extrempunkte eines Streuungsmaßes verstanden werden.

In diesem Papier wird eine neue Clustertechnik für diskrete Zeitreihen mit kategoriellen Werten zum Clustern von Lebensläufen eingeführt, welches anstatt auf Distanzmaßen auf Streuungsmaßen basiert und auch nominale Werte berücksichtigen kann. Zusätzlich werden kategoriale Assoziationsmaße definiert um die temporale Abhängigkeitstruktur der Zeitreihe zu berücksichtigen. Die Maßdefinitionen werden diskutiert, ein Clustermodell eingeführt und die Anwendbarkeit des neuen Algorithmus anhand eines recht großen Datensatz der Deutschen Rentenversicherung demonstriert.

Dieses Papier liefert die technische Grundlage, um die heterogene Vergangenheit von Personen in Beobachtungsstudien (beispielweise im Rahmen von Politikevaluationen) präziser berücksichtigen zu können, ohne sofort mit einem Dimensionalitätsproblem konfrontiert zu werden.

# Clustering life trajectories

## A new divisive hierarchical clustering algorithm for discrete-valued discrete time series*

Stephan Dlugosz†

A new algorithm for clustering life course trajectories is presented and tested with large register data. Life courses are represented as sequences on a monthly timescale for the working-life with an age span from 16–65. A meaningful clustering result for this kind of data provides interesting subgroups with similar life course trajectories. The high sampling rate allows precise discrimination of the different subgroups, but it produces a lot of highly correlated data for phases with low variability. The main challenge is to select the variables (points in time) that carry most of the relevant information. The new algorithm deals with this problem by simultaneously clustering and identifying critical junctures for each of the relevant subgroups. The developed divisive algorithm is able to handle large amounts of data with multiple dimensions within reasonable time. This is demonstrated on data from the Federal German pension insurance.

**Keywords:** clustering, measures of association, discrete data, time series
**JEL C33, C38, J00**

* This work was supported by the German Research Foundation through the *Statistical Modelling of Errors in Administrative Labour Market Data* grant.

† ZEW Center for European Economic Research, L7 1, 68161 Mannheim, Germany, E-mail: dlugosz@zew.de

# 1 Clustering life trajectories

Assume a data set with different persons and regularly repeated measurements of their status. For example, samples from the population of some region and their labor market statuses, 'working' or 'not working'. For many reasons, it is interesting to group these persons not simply according to time-constant variables—like sex, year of birth or primary education—, but to use all the time-dependent information to identify 'similar' life trajectories, such as health status health status or more detailed socio-economic status descriptions. These groups might also present similar attitudes, socialisation and they especially share similar histories. This can give you clues for future labour market performance, health development etc. In particular, this is important for observational studies, where treatment and control groups are balanced according to their observable characteristics [10]. The grouping reveals some of the latent (potentially) unbalanced characteristics.

The goal of clustering is to partition the objects (here: persons) into groups which ideally should be homogenous and well-separated (i.e. low within-group and high between-groups heterogeneity). Most clustering techniques use a distance or dissimilarity matrix. Finding a good—i.e. interpretable—distance measure for a particular clustering task is hard. Therefore, a more direct approach to clustering might perform better: Homogeneity and heterogeneity describe the two extreme points of a measure of dispersion. Unfortunately, measures of dispersion are only available for single variables. However, if the data is discrete (either in its nature or by grouping), it can be reduced to a single variate (with all current combinations as different values) and dispersion-based clustering is applicable.

Assume a data structure consisting of repeated measurements of a nominal quantity. This is a discrete time-series, often known as panel data in econometrics. Life course trajectories are highly dependent on their history, they are quite restricted in their length, and show low in-series variability. The resulting dependency structure is quite special as it nearly includes the whole history. Therefore, neatly reducing the dimensionality is crucial for obtaining reliable clustering results.

Current methods for time-series clustering either require parametric model assumptions in order to estimate and compare parameters for different subgroups (e.g. Markov chains), or rely on quite arbitrary assumptions for defining the 'distance' of two time series [6]. Although the time series in question covers the entire life span of a person, status transitions are quite rare, which complicates estimation of a transition matrix. Moreover, transition rates are very likely to change over one's lifetime. These render a

homogenous Markov model unsuitable. We also lack the necessary number of transitions for more heterogenous models.

There are two specifically designed approaches for life trajectory clustering: The first one is optimal matching—although is was originally not designed for that purpose [5]. This distance measure has recently been used for clustering survey data on life trajectories [9]. Optimal matching calculates the transformation costs that are needed to match two sequences. Possible transformations are insertion, deletion, replacement or reordering of subsequences of the trajectories. The optimal-matching algorithm searches for the sequence of transformations with the lowest total costs. These costs are used as a measure for dissimilarity. The costs for each type of transformation are predefined by the analyst. At first sight, it seems that these costs can be directly interpreted; closer examination reveals that finding plausible costs is often problematic. In addition, the solution strongly depends on these transformation costs.

The other one depends on a more careful modelling of distances [4]. For each element of the sequence, distances between the persons are calculated as usual. The combined distance matrix of the trajectories is obtained by calculating the weighted sum of the different distance matrices. The weights should be different for the different points in time in order to reflect the ordering structure of time. The authors proposed a compound interest model. The rationale behind this approach is the discounting of future developments: differences today or in the near future are more relevant. The argumentation may be switched to the history instead of the future or even include both directions. Again, the analyst has to define the (time-dependent) interest rate, which highly influences the resulting partition. Moreover, the analysis is focused on a particular point in time. If you are interested in the global partition, you have to find a way to "average" the results for the different reference points.

These approaches basically use dissimilarities as the basis for their clustering algorithms. In this paper, a new clustering procedure for discrete-time discrete-valued life course trajectories is introduced that does not depend on a dissimilarity measure but on dispersions. The applied measure of dispersion has to deal with nominal data appropriately. Moreover, a discrete measure of association is needed to cope with the dependency structure of the time series.

These measures are discussed in Section 2 together with some basic assumptions about the data generating method, including the notion of 'critical junctures'. The algorithm itself is introduced in Section 3 and some interesting properties are derived. The applicability of the introduced algorithm is demonstrated on a quite large data set from the German pension insurance consisting of about 15 600 individuals with about 400 repeated

measurements in Section 4.

# 2 Model

The intended application provides a panel data set with high dimensions, both in time (variables) and individuals. The data is nominal by nature. Thus, the clustering procedure has to deal with high dimensionality and a large set of objects.

The huge set of variables forming a large clustering space is characterised by a time-dependent correlation structure. Preprocessing methods like factor analysis or feature selection could help to reduce the number of variates. Such a preprocessing step would reduce the dimensionality on a global scale. Here, local structures are of special interest, especially to identify critical junctures in someone's life. Therefore, an integrated dimensionality reduction and clustering algorithm is needed.

Observational data is large, but not very precise. It is often contaminated with erroneous measurements and lacks details. However, the large number of observations allows for reliable results. We chose a divisive hierarchical clustering algorithm because this group of algorithms is well-suited for large sample sizes.

The next subsection introduces the basic assumptions of the model. We refer to the term 'model' in a strictly nonparametric sense in this paper. Subsection 2.2 introduces measures of dispersion and association needed to describe a clustering algorithm for discrete, nominal data.

## 2.1 Clustering model

Clustering models allow us to derive some general properties of the solution of a clustering algorithm.

Assume, persons are chosen from a finite number $O$ of classes. These classes differ in their probabilities that a member of the class is in a specific status at time $t$. These groups of persons are mixed and we observe their status over time. This can be modelled by a multinomial mixture with unknown mixing parameter $(p_o)_{o \in O}$:

- $X_t \sim \mathcal{M}\big(1, p(c; X_1, \ldots, X_{t-1})\big)$
- $\mathcal{X}_o = (X_1, \ldots, X_T)$
- $\mathcal{X} = \sum_o p_o \mathcal{X}_o$

Life trajectories are highly dependent on their own history where everybody starts at the same status: schooling. Afterwards, some people decide to continue education,

some start to work and others become housewives or husbands. The life trajectory develops according to the decisions, which are dependent on former decisions. This leads to recurrent splitting of the groups; spawning a hierarchical structure,

$$\exists E \subset \{1,\dots,T\}, E \text{ connected} : D(X_E|c_{super}) = A(X_E|c_{sub}) \text{ for all } c_{sub} \subset c_{super} \ .$$

This assumption of a hierarchical structure is slightly more general than described in the former paragraph. It also allows for reunifications. Therefore, the resulting cluster tree is not the corresponding tree of individual's decisions.

## 2.2 Discrete measures of dispersion and association

**Definition 1** (measure of dispersion, [7]). *Let $\mathcal{P}$ denote the class of all finite stochastic vectors, i.e. $\mathcal{P}$ is the union of the sets $\mathcal{P}_k$ comprising all probability vectors of length $k \geq 2$. $D : \mathcal{P} \to [0, \infty[$ is a measure of dispersion, iff*

*DPI* $D\big(p_{\sigma(1)},\dots,p_{\sigma(K)}\big) = D(p_1,\dots,p_K)$ *for all permutations* $\sigma$
*DMD* $D(p) = 0$ *iff* $p$ *is an unit vector*
*DMA* $p <_m q \to D(p) \geq D(q)$
*DSC* $D(p_1,\dots,p_{k-1},r,s,p_{k+1},\dots,p_k) \geq D(p_1,\dots,p_{k-1},p_k,p_{k+1},\dots,p_K)$
*DMP* $D\big((1-r)p+rq\big) \geq (1-r)D(p) + rD(q)$ *for* $0 < r < 1$
*DEC* $D(p_1,\dots,p_k,0) \geq D(p_1,\dots,p_K)$.

For convenience, the class $D$ is restricted to functions

$$D_g(p) = \sum_k g(p_k)$$

with $g$ continuous, concave on $[0,1]$ with $g(0) = g(1) = 0$ and $0 < g(t)$ for $0 < t < 1$. This class includes popular measures of dispersion like the Shannon entropy and the Gini index $D_G(p) = 1 - \sum_k p_k^2$ [7]. $D_g(p)$ can efficiently be estimated from a multinomial i.i.d. sample $X_1,\dots,X_n$, $\hat{p}_n = N^{-1}\sum_i X_i$. The estimator $D_g(\hat{p}_n)$ is the strongly consistent ML-estimator of $D_g(p)$:

**Proposition 1** ([7], Prop. 3a). *Let $p$ be an interior point of $\mathcal{P}$ and $\Sigma$ denote the asymptotic covariance matrix of $\lim_{n\to\infty} \mathcal{L}\big(\sqrt{n}(\hat{p}_n - p)\big)$.*

*a) If $p$ not uniform, then*

$$\mathcal{L}_p\Big(\sqrt{n}\big(D_g(\hat{p}_n) - D_g(p)\big)\Big) \to \mathcal{N}(0, \Gamma_g) \ ,$$

*where* $\Gamma_g = (\dots, g'(p_k), \dots)\Sigma(\dots, g'(p_k), \dots)^t$.

*b) If p uniform, i.e.* $p = u_K$*, then*

$$\mathcal{L}_p\Big(n\big(D_g(\hat{p}_n) - D_g(p)\big)\Big) \to \mathcal{L}\left(\sum_k \lambda_k Y_k^2\right) \ ,$$

*where* $\{Y_k\}_k$ *is a sample of gaussian variates and* $\lambda_k$ *denotes the k-th eigenvalue of* $\Sigma^{1/2} diag(\dots, g''(p_k), \dots)\Sigma^{1/2}$.

The measure of predictive association can be based upon measures of dispersion[1]:

**Definition 2** (measure of predictive association, [2])**.** *Let D be a measure of dispersion and* $L_x$ *an appropriate measure of location. Furthermore, let* $p = (p_{ij})_{i,j}$ *be a two-way-table*[2]*. Then is* $P : \mathcal{P} \to [0, 1]$ *with*

$$P(Y, X) = 1 - \frac{L_x\big(D(Y|X = x)\big)}{D(Y)}$$

*called a measure of predictive association.*

The choice of $L_x$ is quite arbitrary and usually you will take one of the standard functionals. Choosing $D$ is more delicate as it depends on the scaling [2]. The most prominent example is Pearson's $R^2$, but there are also Kendall's $\tau_b$ and others.

Definition 2 is quite general. For the purpose of this paper, we are only interested in nominal measures and we focus on the room of two-way-tables. Again, we are interested in the smaller class based upon a measure of dispersion $D_g$: $P_{D_g}$. For convenience, we choose the average as measure of location $L$.

Then, distributional aspects can be established using the $\Delta$-method:

**Proposition 2.** *Let* $p = (p_{ij})_{i,j}$ *be an interior point of a two-way-table,* $\Sigma$ *as in Proposition 1.*

*a) If* $(p_{ij})_{i,j}$ *not uniform, then*

$$\mathcal{L}_p\Big(\sqrt{n}\big(P_{D_g}(\hat{p}_n) - P_{D_g}(p)\big)\Big) \to \mathcal{N}(0, \Gamma_g) \ ,$$

[1] This is obviously related to the general proportional reduction-in-error principle.

[2] i.e. $p \in \mathcal{P}$, if indexed properly, e.g. row-wise.

*where* $\Gamma_g = d_{ij}\Sigma d_{ij}^t$ *with*

$$d_{ij} = \left( \frac{\sum_{i,j} g(p_{ij}) \cdot g'(p_{.j}) \cdot p_{ij} + g'(p_{ij}) \cdot \sum_j g(p_{.j})}{\left[\sum_j g(p_{.j})\right]^2} \right)_{ij} .$$

*b) If* $(p_{ij})_{i,j}$ *uniform, then*

$$\mathcal{L}_p\Big(n\big(P_{D_g}(\hat{p}_n) - P_{D_g}(p)\big)\Big) \to \mathcal{L}\left(\sum_k \lambda_k Y_k^2\right) ,$$

*where* $\{Y_k\}_k$ *is a sample of gaussian variates and* $\lambda_k$ *denotes the* $k$*-th eigenvalue of* $\Sigma^{1/2} diag(\dots, P''_{D_g}(p_k), \dots)\Sigma^{1/2}$.

*Proof.*

a) directly follows from the $\Delta$ method.

b) follows from [11], Satz 5.127, p. 134. □

The common way to reduce dimensionality is to find a (smaller) subspace under the constraint of minimal loss in 'information', which is often measured in correlation or association between variables. Variable-combining methods like factor analysis, but also variable selection methods like the lasso are based upon this principle.

**Definition 3** (measure of association)**.** $A : \mathcal{P}_{1,2} \to [0,1]$ *is a measure of association, iff*

*API* $A\big(p_{\sigma_1(1),\sigma_2(1)}, \dots, p_{\sigma_1(K_1)\sigma_2(K_2)}\big) = A(p_{1,1}, \dots, p_{K_1,K_2})$ *for all permutations* $\sigma_1, \sigma_2$
*ASY* $A(X,Y) = A(Y,X)$
*ANO* $A = 0$ *if* $X \perp Y$ *and* $A = 1$ *if* $X = f(Y)$ *with bijective* $f$
*AMP* $A\big((1-r)p + rq\big) \le (1-r)A(p) + rA(q)$ *for* $0 < r < 1$.

Axioms API, ASY, ANO are very common in literature (see [11, 8] for instance). Axiom AMP is quite special but very useful for the application we have in mind. There is no direct ordering axiom like DMA for measures of dispersion and AMP partly fills this structural gap. This axiom is not very restrictive as popular measures of association like Pearson's correlation coefficient obeys this axiom (if added to the usual axioms for measures of association for continuous variates). Other examples are the Cramér's $V = \chi(1+\chi^2)^{-1/2}$, the corrected contingency coefficient $C_{corr} = \chi\big(\min(d_1, d_2) - 1\big)^{-1/2}$, where $\chi^2$ represents the $\chi^2$-coefficient on probabilities.

The measure of association which is based on the predictive measure of association, $A_P = \min\big(P_D(Y,X), P_D(X,Y)\big)$, has an intuitive interpretation as the 'minimal information redundance' that two variables $X$ and $Y$ share.

**Proposition 3.** *The minimal information redundance $A_P$ is a measure of association.*

*Proof.* Axioms API, ASY and ANO are trivially true. AMP follows from DMP.

**Proposition 4.** *Let $p = (p_{ij})_{i,j}$ be an interior point of a two-way-table, $\Sigma$ as in Proposition 1.*

*a) If $P_{D_g}(p) \neq P_{D_g}(p^t)$, then*

$$\mathcal{L}_p\Big(\sqrt{n}\big(P_{D_g}(\hat{p}_n) - P_{D_g}(p)\big)\Big) \to \mathcal{N}(0, \Gamma_g) \ ,$$

*where $\Gamma_g = d_{ij}\Sigma d_{ij}^t$ with*

$$d_{ij} = \begin{cases} d_{ij}^{(1)} = \left( \frac{\sum_{i,j} g(p_{ij}) \cdot g'(p_{.j}) \cdot p_{ij} + g'(p_{ij}) \cdot \sum_j g(p_{.j})}{\left[\sum_j g(p_{.j})\right]^2} \right)_{ij} & P_{D_g}(p) < P_{D_g}(p^t) \\ d_{ij}^{(2)} = \left( \frac{\sum_{i,j} g(p_{ij}) \cdot g'(p_{i.}) \cdot p_{ij} + g'(p_{ij}) \cdot \sum_j g(p_{i.})}{\left[\sum_j g(p_{i.})\right]^2} \right)_{ij} & P_{D_g}(p) > P_{D_g}(p^t) \end{cases} .$$

*b) If $p$ symmetric but not uniform, then*

$$\mathcal{L}_p\Big(\sqrt{n}\big(A_{P_{D_g}}(\hat{p}_n) - A_{P_{D_g}}(p)\big)\Big) \to \mathcal{N}(0, \Gamma_g) \ ,$$

*where $\Gamma_g = d_{ij}\Sigma d_{ij}^t$ with $d_{ij} = d_{ij}^{(1)}$.*

*c) If $p$ uniform, then*

$$\mathcal{L}_p\Big(n\big(A_{P_{D_g}}(\hat{p}_n) - A_{P_{D_g}}(p)\big)\Big) \to \mathcal{L}\left(\sum_k \lambda_k Y_k^2\right) \ ,$$

*where $\{Y_k\}_k$ is a sample of gaussian variates and $\lambda_k$ denotes the $k$-th eigenvalue of $\Sigma^{1/2} diag(\dots, P''_{D_g}(p_k), \dots)\Sigma^{1/2}$.*

*d) If $p$ asymmetric and $P_{D_g}(p) = P_{D_g}(p^t)$, then*

$$\mathcal{L}_p\Big(\sqrt{n}\big(P_{D_g}(\hat{p}_n) - P_{D_g}(p)\big)\Big) \to \mathcal{N}(0, \Gamma'_g) \ ,$$

*where $\Gamma'_g \leq d_{ij}\Sigma d_{ij}^t$ with $d_{ij} = \max_v(d_{ij}^{(1)}, d_{ij}^{(2)})$ and $\max_v$ operates component-wise.*

*Proof.*

a) Proposition 3, part a).

b),c) symmetry implies differentiability of $A_{P_{D_g}}$ on all interior points of $p$. Proposition 3 for the results.

d) one-sided differentiability only allows for bounds on the convergence, Proposition 3, part a) □

# 3 Algorithm

A clustering algorithm for trajectory data should be stable in the asymptotic sense, i.e. for $n \to \infty$. Furthermore, it should deal with the time-series structure of the data. Both requirements are met by using measures of association and dispersion. The data model implies a hierarchical cluster structure that the clustering algorithm should find.

## 3.1 Clustering algorithm

The algorithm works as follows:

1 dimension reduction: cluster points in time

- 1.1 calculate association between neighbouring points in time
- 1.2 combine points in time with maximum association ($a$), association between groups of dates is the minimum of the pairwise associations (complete linkage)
- 1.3 iterate steps 1.1 and 1.2 until each point in time provides additional information ($0 \leq a < \alpha \leq 1$)

2 calculate average (mode) values for groups of dates

3 split data set according to variable with maximal total dispersion reduction (CART-like: $\min_{t,k} p_1 D_1 + p_2 D_2$)

4 iterate steps 1 to 3 until groups are homogeneous ($0 \leq d_c < \beta \leq 1$)

At each split, perform a dimensionality reduction for the current subset of the data, find the best split and iterate. The splitting procedure can be compared to the CART-strategy [1, 7].

Steps 1 and 2 perform the dimensionality reduction. First, we look for similar variables and connected clusters of them. Here, when saying connected, we mean variables that represent neighbouring points in time. This clustering procedure is agglomerative[3] and

[3] Note that the hierarchical nature of this variable clustering procedure is as restrictive as any other clustering heuristic because of the order structure.

we use the notion of complete linkage for the calculation of association between a single variable and a cluster of variables. Again, we ensure that the predictive information of the variable on the other variables in its cluster is sufficiently large. As a result, the time line is divided into a set of periods of various lengths. This procedure obeys the ordering of the variables that form our time-series. Now, we have to select a representative value for each period (cluster of points in time) and we choose the mode. We get a set of new variables, where each of the variables represents a set of points in time. This set is used in the next steps for a particular split.

In step 3, the algorithm looks for the best split, i.e. the variable on which the split is based that provides the maximal reduction in the dispersion measured over all (newly defined) variables of the subgroups. The basic idea to minimise the dispersion in this way has been transferred from CART [1] to cluster-trees in [7].

These steps are iterated for each newly defined subset until the homogeneity criterion is met or all objects have become their own leafs in the cluster-tree. The exact choice of the measure of dispersion and the measure of association is arbitrary to some extent. We will use the Gini for dispersion and the minimised predictive association for association. Other choices will provide slightly different results.

There are two tuning parameters that have to be fixed. The first parameter ($\alpha$) controls the clustering of variables in the dimensionality reduction step. It ensures that the clusters contain only those variables that are sufficiently similar. The choice of the parameter $\alpha$ highly influences the structure of the resulting tree and, thus, the parameter has to be chosen carefully. The second parameter ($\beta$) controls the size of the cluster tree. It represents the homogeneity criterion that the clusters have to meet.

## 3.2 Properties

The resulting partition should be stable, i.e. increasing the number of observations should (after a certain point) not influence the partition found so far. Our data can grow in two directions: The number of persons and the sampling frequency can both increase. While it is classic that the partition stays stable for an increasing number of people, it is less intuitive for increasing frequency. We assumed that the time series is discrete in time by nature. This implies that the proportion of identical measurements increases with the sampling frequency. This behaviour should not affect the resulting partition.

**Theorem 1** (cluster stability)**.** *The algorithm generates a stable partition of the sample into $k \in \mathbb{N}$ clusters:*

*a) Let $C_k^n(i)$ denote the cluster assignment[4] of observation $i$ for $k$ clusters found in a sample of size $n$. Then holds*

$$P(C_k^n(i) = C_k^{n-1}(i)) \to_p 1 \ \ \forall i \ .$$

*b) Let $C_k^f(i)$ denote the cluster assignment of observation $i$ for $k$ clusters found with sampling frequencies $f \geq f'$. Then holds*

$$\exists_{f*} \forall_{f' > f*} P(C_k^f(i) = C_k^{f'}(i)) \to_p 1 \ \ \forall i \ .$$

*Proof.* Direct results from propositions 1, 2 and 4. □

Note that the whole tree is stable in the sense of theorem 1 and not only the leaves.

The algorithm is designed for large data sets. Let $n$ denote the number of persons and $T$ the length of the time series. Then it is true:

**Theorem 2.** *The algorithm belongs to the computational complexity class $O(n^2 \times T^3)$.*

*Proof.* Calculation of two-way contingency tables: $O(n \times T^2)$ for the measures of association. Cluster tree induction: $O(n \times T)$ for the measures of dispersion. Combined, we have $O((n \times T^2) \times (n \times T))$. □

We can expect that the algorithm performs much better in the average case, especially because of the reduced $T$ after the dimensionality reduction step. Additionally, reusing calculations from former splits reduces the calculation time significantly for certain configurations of measures of dispersion and association.

# 4 Example

## 4.1 Data

We use the Versichertenkontenstichprobe (VSKT, sample of insured persons and their insurance accounts) of the German Federal Pension Insurance [3]. It is an administrative merged dataset with detailed monthly information on 60 821 individuals aged 15 to 67 of 2005. This sample is stratified to cover the whole range of covariates and keep the sample quite small at the same time. The VSKT is a monthly panel that starts in January of

[4] The identification of the clusters is straightforward. For example, you could use depth-first or breadth-first search.

the year in which the person becomes 16, and ends in December of the year in which the person reaches 67. In total, there are up to 600 observations per individual.

We restrict the sample to the birth cohort 1940–1950. This includes 15 566 life course trajectories, which correspond to 9 269 888 individuals. 8 979 observed persons are females which correspond to 4 798 073 individuals. Many of the observations at the beginning and the end of the life courses are missing. Thus, we use only the months 49 to 432, which is about age 18 to 50. The dataset includes a time series of the social-economic status that is measured on a scale with thirteen different values. These thirteen values are combined to four classes:

0 employed (full-time or part-time)

1 unemployed (with or without compensation)

2 absence from the labour market due to education, family responsibilities (children, old-age care), military service

3 absence from the labour market due to long-term sickness, pension; and unknown status

The results become more detailed if you opt for a less crude grouping of the thirteen statuses. We, however, decided to use this more simplistic setup to demonstrate the potential of the new algorithm.

## 4.2 Results

We use the Gini as measure for dispersion and the minimal predictive association for measuring associations. The tuning parameters are fixed as $\alpha = 0.5$ and $\beta = 0.1$. Additionally, we stopped the splitting process, if a node contains less than 1 000 observations. Figure 1 shows the resulting tree. We labelled the nodes (clusters) by consulting table 1 and analysing figure 2.

Table 1 shows the distribution of sex and education among the clusters. Extraordinary values are highlighted. These provides a first clue to the characteristics of the clusters. Figure 2 shows employment rates for each of the clusters. Together, these two pieces of information allow for a good interpretation of the various groups.

The most obvious groups are the two female dominated groups, early- and three-phases-housewives. The first is characterised by a low lifetime labour market attachment whilst females in the latter group interrupt their careers to raise children. Both groups represent a quite small fraction of the population. Nevertheless, they represent two extremes that

Figure 1: Cluster tree

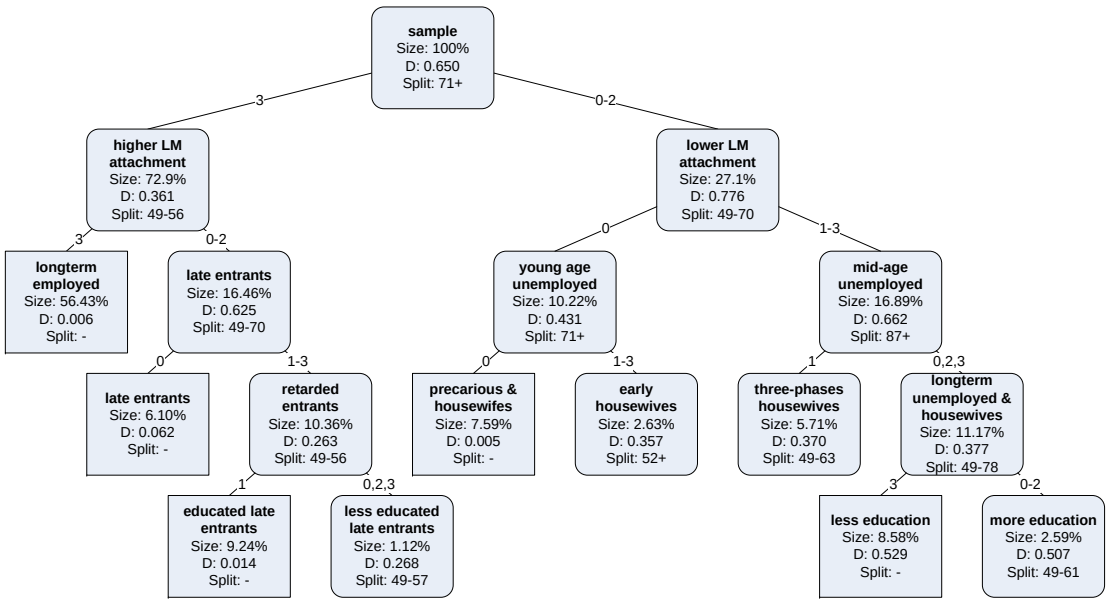
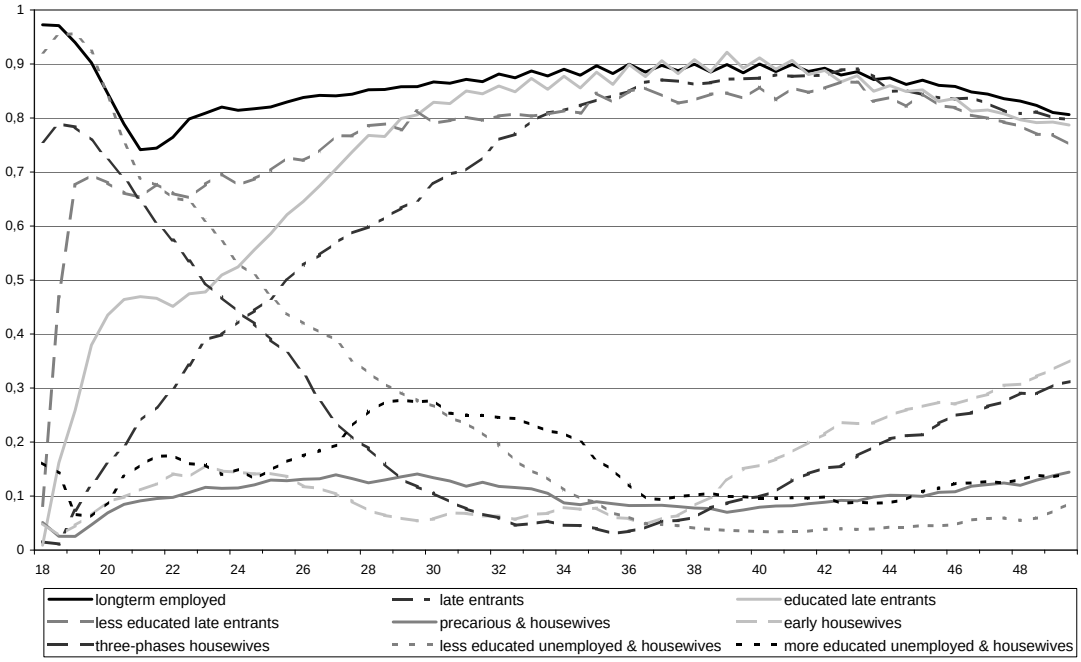



Table 1: Cluster sizes, sexual composition and distribution of educational degrees

| | long-term unempl. | late entrants | educated late entrants | less edu. late entrants | pre-carious & housew. | early housew. | three-phases housew. | less edu. unemp. & housew. | more edu. unemp. & housew. |
|---|---|---|---|---|---|---|---|---|---|
| size | 56.4% | 6.1% | 9.2% | 1.1% | 7.6% | 2.6% | 5.7% | 8.6% | 2.6% |
| female | 41.0% | 67.7% | 53.2% | 58.9% | 61.5% | **99.0%** | **99.1%** | 52.7% | 57.1% |
| | | | | | | | | | |
| missing/na. | 54.5% | 58.7% | 41.3% | 60.2% | **85.1%** | 75.3% | 73.1% | **86.9%** | 75.3% |
| school | 5.4% | **10.3%** | 3.4% | **12.4%** | 3.2% | 6.3% | 7.3% | 3.0% | 2.9% |
| voc. training | **35.6%** | 25.0% | 25.6% | 25.5% | 9.8% | 17.4% | 17.8% | 9.0% | 10.2% |
| high school | 0.1% | 0.0% | **1.2%** | 0.0% | 0.1% | 0.0% | 0.1% | 0.2% | 0.5% |
| h.s. & voc. tr. | 1.0% | 0.9% | **5.1%** | 0.9% | 0.5% | 1.0% | 0.1% | 0.3% | 1.4% |
| techn. college | 2.3% | 2.7% | **4.1%** | 1.0% | 0.7% | 0.0% | 0.3% | 0.6% | 0.4% |
| university | 1.3% | 2.5% | **19.2%** | 0.1% | 0.6% | 0.0% | 1.3% | 0.0% | **9.4%** |

Figure 2: Employment rates for the nine clusters over time




are clearly distinguishable from other life trajectories patterns. There are, of course, housewives in the other leaves of the right branch of the cluster tree.

The critical junctures are found by looking at the split variables denoted in the cluster tree. Whether a person will have a high labour market attachment is decided at life trajectory month 71, which corresponds to the age of about 20. In contrast, Figure 2 suggests that the high labour market attachment becomes visible in the data only after the age of about 24. The other critical junctures also lie around age 20. The latest juncture is at the age of 21, where persons that show low overall employment rates split. Only a few are coming back to the labour market after their 38th birthday. To conclude, the life trajectory till the 20th or 21st birthday of a person strongly influences the remaining trajectory.

## References

[1] Breiman L, Friedman J H, Olshen R A, Stone C J (1984) Classification and regression trees. Chapman & Hall, London

[2] Dlugosz S, Müller-Funk U (2008) Predictive classification and regression trees. In:

Fink A, Lausen B, Seidel W, Ultsch A (eds) Advances in data analysis, data handling and business intelligence. Springer Berlin Heidelberg New York, pp 127-134

[3] Himmelreicher R K, Stegmann M (2008) New possibilities for socio-economic research through longitudinal data from the research data centre of the German federal pension insurance (FDZ-RV). J Appl Soc Sci Studies (Schmollers Jahrbuch) 128: 647-660

[4] Košmelj K, Batagelj V (1990) Cross-sectional approach for clustering time varying data J Classification 7: 99-109. DOI 10.1007/BF01889706

[5] Kruskal J B (1983) An overview of sequence comparison. In: Sankoff D, Kruskal J B (eds) Time warps, string edits, and macromolecules: practice of sequence comparison. Addison-Wesley Reading Mass., pp 1-44

[6] Liao T W (2005) Clustering of time series data—a survey. Pattern Recognition 38: 1857–1874. DOI 10.1016/j.csda.2005.04.012

[7] Müller-Funk U (2008) Measures of dispersion and cluster-trees for categorical data. In: Preisach C, Burkhardt H, Schmidt-Thieme L, and Decker R (eds) Data analysis, machine learning and applications. Springer Berlin Heidelberg New York, pp 163-170

[8] Nelsen R B (2006) An introduction to copulas. Springer, Berlin Heidelberg New York

[9] Piccarreta R, Billari F C (2007) Clustering work and family trajectories by using a divisive algorithm. J Royal Statistical Society A 170: 1061–1078. DOI 10.1111/j.1467-985X.2007.00495

[10] Rosenbaum P R, Rubin D (1983) The central role of the propensity score in observational studies for causal effekts. Biometrika 70: 41-55.

[11] Witting H, Müller-Funk U (1995) Mathematische Statistik II. Teubner Stuttgart