Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Eidgenössisches Departement des Innern EDI
Bundesamt für Statistik BFS

ML_SoSi – Machine Learning Soziale Sicherheit

Pilotprojekt im Rahmen der Dateninnovationsstrategie des BFS

Schlussbericht

EXPERIMENTAL STATISTICS

Neuchâtel, 2023

Herausgeber:	Bundesamt für Statistik (BFS)	**Layoutkonzept:**	Sektion PUB
Redaktion:	Luzius von Gunten, Frank Schubert, Brandon Qorri Gonzalez, Athanassia Chalimourda	**Download:**	www.statistik.ch
Themenbereich:	00 Statistische Grundlagen	**Copyright:**	BFS, Neuchâtel 2023 Wiedergabe unter Angabe der Quelle für nichtkommerzielle Nutzung gestattet
Originaltext:	Deutsch		

ESPACE DE L'EUROPE 10 CH-2010 NEUCHÂTEL WWW.STATISTIK.CH

Inhaltsverzeichnis

Dank ... **3**

1 Management Summary ... **4**

2 Ausgangslage, Fragestellung und Zielsetzung ... **6**

3 Vorgehen ... **8**

3.1 Übersicht Analysedesign ... 8

3.2 Datengrundlage ... 8

3.3 Kohortendesign und standardisierter Beobachtungszeitraum ... 9

3.4 Sequence Clustering ... 10

3.4.1 Methodische und technische Vorüberlegungen ... 10

3.4.2 Umgesetztes Vorgehen ... 11

3.5 Visualisierung und Verlaufsindikatoren ... 13

3.6 Vergleich von Clusterlösungen über die Zeit ... 15

3.6.1 Motivation und Lösungsweg ... 15

3.6.2 Entwicklung Prädiktionsmodell ... 15

3.6.3 Prädiktion ... 17

3.7 Aktualisierung der initialen Clusterlösung ... 17

3.7.1 Deskriptive Evaluation der Ähnlichkeit zwischen Referenz und Prädiktion ... 18

3.7.2 Post-hoc Evaluation der Zuverlässigkeit der Prädiktion ... 20

3.7.3 Analyse der Korrespondenz zwischen Prädiktion und neuer Clusterlösung in einer bestimmten Kohorte ... 20

4 Resultate ... **22**

4.1 Übersicht zu den Grundgesamtheiten ... 22

4.2 Initiale Clusterlösung ... 23

4.2.1 Lösung mit zehn Clustern ... 23

4.2.2 Inhaltliche Interpretation ... 24

4.2.3 Vergleich von Clusterlösungen über die Kohorten ... 30

4.3 Übertragung der initialen Clusterlösung auf neue Kohorten mittels Prädiktion ... 31

4.4 Analysen zur Aktualisierungsnotwendigkeit der initialen Clusterlösung ... 37

4.4.1 Deskriptive Evaluation der Ähnlichkeit zwischen Referenz und Prädiktion ... 38

4.4.2 Post-hoc Evaluation der Zuverlässigkeit der Prädiktion ... 41

4.4.3 Analyse der Korrespondenz zwischen Prädiktion und neuer Clusterlösung in einer bestimmten Kohorte ... 43

4.4.4 Schlussfolgerungen ... 49

5 Key Learnings und Empfehlungen ... **51**

6 Transfer in die Produktion ... **55**

6.1 Auswirkungen von ML_SoSi auf die statistische Standardproduktion ... 55

6.2 Weiteres Vorgehen beim Transfer in die Produktion ... 56

6.3 Generischer, induktiver Analyseansatz für individuelle Verlaufsdaten in der Produktion 56

7 Anhang ... **59**

8.1 Tabellen im Haupttext ... 125

8.2 Abbildungen im Haupttext ... 125

8.3 Tabellen im Anhang ... 126

8.4 Abbildungen im Anhang ... 127

Dank

Neben den Autorinnen und Autoren des Schlussberichts haben folgende Personen am Pilotprojekt mitgewirkt oder dieses mit Beratung und wertvollen Rückmeldungen unterstützt. Für diese Unterstützung gebührt ihnen grosser Dank.

Philippe Meyer, Daniel Kilchmann, Michael Leuenberger, Gerhard Gillmann, Diego Kuonen, Christian Ruiz, Nora Meister, Thomas Ruch, Sheila Planta, Joaquim Golay, Kaspar Stucki.

1 Management Summary

Arbeitslosigkeit kann für die betroffenen Personen sehr unterschiedlich verlaufen. Die so entstehenden Verlaufsbiographien sind unter anderem geprägt durch (wiederholte) Sozialleistungsbezüge aus dem System der Sozialen Sicherheit (Arbeitslosen-, Invalidenversicherung, Sozialhilfe), Wiedereintritt in die Erwerbsarbeit oder auch Rückzug aus dem Erwerbsleben und Migration. Im Pilotprojekt «ML_SoSi» werden Angaben zu den individuellen Verläufen unter Anwendung induktiver statistischer Methoden analysiert und typische Verlaufsmuster identifiziert. Neben den so erzielten Resultaten ist es das Ziel des Projektes, einen datengetriebenen Ansatz zur Analyse individueller Verläufe in Längsschnittdaten für die öffentliche Statistik zu entwickeln.

Als Datenbasis dient ein anonymisierter verknüpfter Datensatz, der monatsgenaue Informationen zu individuellen Sozialleistungsbezügen aus der Sozialhilfe (SH), der Invalidenversicherung (IV) und der Arbeitslosenversicherung (ALV), sowie zur Erwerbstätigkeit (IK) enthält. Im Rahmen dieses Berichts wird für diesen Datensatz die Abkürzung „SHIVALV+IK“ verwendet. Zur Grundgesamtheit zählen Personen zwischen 18 und 65 Jahren, die im Zeitraum 2010–2015 neu Taggelder der Arbeitslosenversicherung (ALV) beziehen. Die Analyse wird auf der Basis von Jahreskohorten umgesetzt. In die Analyse einbezogen werden Informationen zum Bezug von Sozialversicherungs- und Sozialhilfeleistungen sowie zur Erwerbstätigkeit während den folgenden 48 Monaten bzw. 4 Jahren.

In der methodischen Umsetzung werden zunächst mit der Kohorte 2010 in einem zweistufigen Sequenzclusteringverfahren (unsupervised machine learning) typische Verlaufsmuster identifiziert und mittels grafischer Darstellungen («State Distribution Plots») sowie Verlaufsindikatoren analysiert und inhaltlich interpretiert. Anschliessend wird diese initiale Clusterlösung auf die Kohorten der Folgejahre 2011–2015 mittels «supervised machine learning» übertragen (Prädiktion). Bei jeder Übertragung wird dabei die Validität des Modells geprüft, wobei verschiedene Kriterien evaluiert werden.

Mit diesem Vorgehen liegt der Fokus auf der Erkennung und Analyse aggregierten typischen Verlaufsmustern und deren Übertragung auf weitere Kohorten. Die Nutzbarmachung individueller Vorhersagen zu jedweden Zwecken wird ausgeschlossen.

Insgesamt zehn Cluster zur Beschreibung der typischen Verlaufsmuster von neuen Arbeitslosentaggeldbeziehender wurden identifiziert (siehe Tabelle 4), die mehrheitlich auch im Kohortenvergleich 2010-2015 stabil bleiben (8 von 10 der typischen Verlaufsmuster). Inhaltlich zeichnen sich mehrere Cluster ab, in denen sich die Personen nach einer Phase des Bezugs von Arbeitslosentaggeld wieder in den Arbeitsmarkt integrieren. Dabei unterscheiden sich die Cluster nach der Dauer des Taggeldbezugs (Cluster 1 und 2), nach Vorhandensein einer Zwischenverdienstphase (Cluster 3) sowie durch mehrfache ALV-Bezugsperioden mit zwischenzeitlicher Erwerbstätigkeit (Cluster 4). Daneben entstehen Cluster, die klare Tendenzen entweder zum dauerhaften Bezug von IV-Renten oder von Leistungen der Sozialhilfe zeigen (Cluster 5, 6, 7, 8 und 9). Hierzu gehören zwei Cluster mit Neubezug dieser Leistungen (Cluster 5, 9) und zwei Cluster, mit jeweils ausgeprägten Phasen mit ergänzendem Erwerbseinkommen (Cluster 6, 7) sowie ein Cluster, mit Personen, die bereits vor dem Bezug von Arbeitslosentaggeld dauerhaft oder wiederholt auf Sozialhilfe angewiesen waren. Ein letztes Cluster schliesslich vereint diejenigen Personen, die während des Beobachtungszeitraumes dauerhaft nicht mehr von den untersuchten Systemen (SH, IV, ALV, IK/Erwerb) erfasst werden (Cluster 10). Das Projekt hat gezeigt, dass das «Sequence Clustering» ein vielversprechendes Verfahren ist, um inhaltlich valide und analytisch relevante Resultate zu erzeugen. Es erlaubt eine deutliche Verringerung der Komplexität der Verlaufsdaten und erweitert damit die Analysemöglichkeiten durch die Erkennung von Muster, die deduktiv nicht antizipiert werden konnten.

Um diese Informationen für das Publikum der öffentlichen Statistik, unter anderem für die politische Steuerung, noch relevanter zu machen, sind Zeitreihendaten von grosser Bedeutung. Die initiale Clusterlösung kann jedoch nicht ohne weiteres in einer neuen Kohorte reproduziert werden. Im vorliegenden Projekt wurde diese Schwierigkeit gelöst indem die initiale Lösung mittels Prädiktion auf neue Kohorten übertragen wurden. Dieses Vorgehen funktioniert gut und die Kriterien, nach welchen entschieden wurde, ab welchem Zeitpunkt die Übertragung nicht mehr valide ist, hat sich in diesem Fall bewährt. Die Erkenntnisse haben einen konkreten Mehrwert für die statistische Standardproduktion,

sowohl bezüglich der neu entwickelten Längsschnittindikatoren und deren Visualisierung als auch für die Bildung von beschreibenden, quantitativen Verlaufsprofilen[1].

Erkenntnisse, Möglichkeiten und Beschränkungen bei der Anwendung von datengetriebenen Methoden in der öffentlichen Statistik werden im Bericht vertieft diskutiert. Auf der Basis von key learnings werden Empfehlungen für ähnlich gelagerte Projekte im BFS dargestellt. Die Schlussfolgerungen münden in einem generischen, induktiven Analyseansatz für individuelle Verlaufsdaten in der Statistik-Produktion im BFS.

[1] Siehe Publikation «Verläufe im System der sozialen Sicherheit 2021»: https://www.bfs.admin.ch/bfs/de/home/statistiken/soziale-sicherheit/analysen-verlaeufe-system/analysen.assetdetail.25385873.html

2 Ausgangslage, Fragestellung und Zielsetzung

Das Ziel des Pilotprojekts «ML_SoSi» (Machine Learning Soziale Sicherheit) im Rahmen der Dateninnovationsstrategie des BFS ist es, typische Bezugs- und Reintegrationsverläufe im System der sozialen Sicherheit mit datengetriebenen Methoden zu identifizieren und zu beschreiben und für die Produktion öffentlicher Statistik nutzbar zu machen. Es verfolgt damit schwerpunktmässig methodische aber auch inhaltliche Ziele.

Der gesellschaftliche und wirtschaftliche Wandel der letzten Jahrzehnte hat sich und wirkt sich weiterhin tiefgreifend auf die Lebensverläufe der Bevölkerung aus. Neben der traditionellen Normalbiografie sind vielfältige davon abweichende Lebensentwürfe und Biografien zur Regel geworden.[2] Diese stehen einem System der Sozialen Sicherheit in der Schweiz gegenüber, das sich einerseits weiterhin an kontinuierlichen Erwerbsbiografien orientiert und in dem andererseits die sozialpolitischen Zuständigkeiten stark durch den Föderalismus gekennzeichnet ist.[3] Die Zuständigkeiten sind geprägt durch bundesgesetzlich geregelte Sozialversicherungen (hier: Arbeitslosenversicherung und Invalidenversicherung) mit unterschiedlichen Zielsetzungen und den Sozialversicherungen nachgelagerte Bedarfsleistungen, die in kantonaler und teilweise kommunaler Kompetenz liegen (hier: wirtschaftliche Sozialhilfe). In dieser Situation, in der eine Vielfalt neuer Lebensentwürfe auf ein stark ausdifferenziertes System der Sozialen Sicherheit trifft, ist die Analyse des Zusammenspiels der einzelnen Sozialwerke und des Leistungsbezugs durch Risikogruppen von zunehmender Komplexität gekennzeichnet.

Deshalb kommt einer verlaufsorientierten Perspektive eine grosse Bedeutung zu. Dabei kann ein induktiver, datengetriebener Ansatz einerseits dazu dienen, bestehende Resultate aus der Forschung zu verifizieren, welche vornehmlich auf deduktiver Typenbildungen von Verläufen im System der Sozialen Sicherheit beruhen, und andererseits können Veränderungen in den Verlaufsmustern sichtbar bzw. neu entstehende typische Verläufe identifiziert werden.

An der grundlegenden Ausrichtung des Projekts hat sich seit der Publikation der letzten Zwischenergebnisse[4] nichts Wesentliches geändert. Grundgesamtheit bleiben die Personen, die in einem bestimmten Jahr neu Arbeitslosentaggeld beziehen (Kohorten). Typische Verlaufsmuster werden durch ein Sequenzclusteringansatz gebildet (unsupervised machine learning) und mit Verlaufsindikatoren und Visualisierungen beschrieben. Nicht mehr weiterverfolgt wurde das Ziel, die Zugehörigkeit zu einem bestimmten Verlaufsmuster für die Mitglieder einer neuen Kohorte zu schätzen, für welche die Daten zu den Bezugsverläufen noch nicht oder noch nicht vollständig vorliegen (forecasting). Die vorliegenden Resultate bestätigen die Stabilität der typischen Verläufe über die Zeit nicht vollständig. Zudem zeigen sie, dass die Verlaufsinformationen für die Prädiktion von zentraler Bedeutung sind; Eine Prädiktion der Clusterzugehörigkeit alleine auf den soziodemografischen Angaben zum Zeitpunkt des Systemeintritts ist nicht erfolgsversprechend. Hingegen wurde ein Ansatz entwickelt, um eine bestimmte Clusterlösung zwischen unterschiedlichen Kohorten, für welche die Verlaufsdaten vollständig vorliegen, stabil über die Zeit vergleichen zu können (supervised machine learning).

Es werden demnach folgende Hauptfragestellungen bearbeitet:

- Wie sehen typischen Erwerbs- und Bezugsverläufe von Personen aus, die Leistungen aus dem System der sozialen Sicherung beziehen?
- Wie können komplexe Verlaufsmuster im Bereich der Sozialen Sicherheit mit Hilfe induktiver statistischer Methoden (unsupervised machine learning) analysiert und sinnvoll beschrieben werden?
- Inwieweit können mit induktiven statistischen Methoden für Politik und Verwaltung steuerungsrelevante Indikatoren und Analysen erarbeitet werden?

[2] Hardering, Friedericke (2015). «Prekarität und Prekarisierung. Jüngere Tendenzen der Debatte über die neue soziale Unsicherheit». In: König Helmut, Schmidt, Julia, Sicking, Manfreg (Hg.), «Die Zukunft der Arbeit in Europa. Chancen und Risiken neuer Beschäftigungsverhältnisse». transcript: Bielefeld.

[3] Von Gunten, Luzius, Fluder, Robert, Zürcher, Pascale et al. (2015). «Existenzsicherung im Alter. Risikofaktoren und Ursachen für EL-Bezüge bei AHV-Neurentnern und -Neurentnerinnen». Berner Fachhochschule: Bern.

[4] https://www.experimental.bfs.admin.ch/expstat/de/home/projekte/ml-sosi.html

Entsprechend werden im Projekt die Ziele verfolgt, einen Analyseansatz für Verlaufsdaten in der öffentlichen Statistik zur Identifikation, Beschreibung und Darstellung von typischen Verlaufsmuster zu entwickeln und die Herausforderungen eines Transfers in die Statistikproduktion bezüglich dieses Ansatzes zu identifizieren sowie Lösungsansätze dazu zu formulieren.

3 Vorgehen

3.1 Übersicht Analysedesign

Ausgehend von einem verknüpften Datensatz basierend auf Daten aus Sozialversicherungsregistern und der Sozialhilfestatistik (siehe Abschnitt 3.2) werden Jahreskohorten neuer Beziehenden von Arbeitslosentaggeld bestimmt. Für jede Person in der Kohorte wird ab dem ersten Taggeldbezugsmonat ein standardisierter 48-monatiger Beobachtungszeitraum festgelegt (Abschnitt 3.3). Für jeden Monat ist die Information verfügbar, ob eine Leistung der Arbeitslosen- (ALV), Invaliditätsversicherung (IV) und/oder der Sozialhilfe (SH) bezogen wurde bzw. ob ein Erwerbseinkommen erzielt wurde.

Mit diesen 48-monatigen individuellen Sequenzen werden für die Kohorte 2010 mit einem zweistufigen Clusteringverfahren zehn typische Verlaufsmuster identifiziert (unsupervised machine learning, Abschnitt 3.4). Um die Datenmenge zu verarbeiten, werden in der ersten Stufe die ca. 126'000 Verläufe mit dem k-means-Algorithmus in 3000 Cluster eingeteilt. In der zweiten Stufe wird aus jedem Cluster der «repräsentativste» Verlauf ausgewählt, welcher wiederum mit Hilfe des «hierarchical clustering»-Algorithmus und auf Basis der «edit»-Distanz zu zehn typischen Verlaufsmustern (cluster) zusammengefasst werden. Die Zugehörigkeit zu einem dieser zehn Verlaufsmuster wird über die Repräsentanten der 3000 Cluster aus der ersten Stufe auf die Grundgesamtheit übertragen.

Die gefundene Lösung mit 10 Verlaufsmustern wird mit Verlaufsindikatoren und Visualisierungen (state distribution plots) interpretiert und validiert, sodass für jedes der zehn Muster ein Label vergeben werden kann (Abschnitt 3.5).

Um die Häufigkeitsverteilung der zehn Verlaufsmuster über die Zeit (bzw. verschiedene Kohorten) analysieren zu können, wird die initiale Clusterlösung auf Basis der Kohorte 2010 mittels «supervised machine learning» auf die Kohorten der Folgejahre übertragen (Prädiktion). Da sich sowohl die strukturellen Merkmale der Kohortenmitglieder als auch die gesellschaftlichen und institutionellen Rahmenbedingungen mit der Zeit ändern, muss bei jeder Übertragung der initialen Clusterlösung auf eine neue Kohorte dessen Validität geprüft werden. Dazu kommen verschiedene Kriterien zum Einsatz: deskriptiver Vergleich der initialen und übertragenen Cluster (Kompaktheit, Clustergrössen, visuell, Verlaufsindikatoren), Zuverlässigkeit der Clusterübertragung (mithilfe der probabilistischen Basis der majority vote im «random forest»-Algorithmus), Übereinstimmung der übertragenen Clusterlösung mit einer neuen Clusterlösung auf derselben Kohorte mittels externer Masse, die auf der Konfusionsmatrix basieren.

3.2 Datengrundlage

Für das Pilotprojekt wird ein Datensatz genutzt, der Informationen zu individuellen Sozialleistungsbezügen aus der Sozialhilfe (SH), der Invalidenversicherung (IV) und der Arbeitslosenversicherung (ALV) enthält. Ergänzt werden diese Angaben mit Informationen zum Erwerbsverlauf, indem die Erwerbsperioden aus den individuellen Konten (IK) der Zentralen Ausgleichskasse hinzugefügt werden. Im Rahmen dieses Berichts wird für diesen Datensatz die Abkürzung „SHIVALV+IK“ verwendet. Es handelt sich dabei weitgehend um Sekundärdaten im Sinne der Dateninnovationsstrategie[5] (S.10), die sich aus den Administrativdaten der grossen Sozialversicherungen und teilweise der Sozialdienste speisen. Die Daten wurden anonymisiert.

Die im Projekt verwendete Datenbasis wurde aus den Sekundärdatenquellen so aufbereitet, dass sie pro Person pro Monat im verfügbaren Zeitraum und pro betrachtetes System (SH, IV, ALV, IK/Erwerb) die Information enthält, ob ein Leistungsbezug bzw. Erwerbsarbeit vorliegt oder nicht. Aus den vier Grundzuständen ergeben sich für jeden Monat 16 mögliche Statuskombinationen (bspw. ALV, SH+ALV, IV+IK, SH+IV+IK); Eine vollständige Liste der möglichen Status findet sich im Anhang (Abschnitt 7.1).

Die Grundgesamtheit der SHIVALV+IK-Daten setzt sich aus allen Personen zwischen 18 und 65 Jahren zusammen, die in einem betrachteten Jahr mindestens eine Leistung aus einem der drei Sicherungssystemen erhalten haben und/oder erwerbstätig waren (Vollerhebung). Der Datensatz wird

[5] https://www.bfs.admin.ch/bfs/de/home/aktuell/neue-veroeffentlichungen.assetdetail.3862237.html

seit 2020 im BFS von der Sektion Sozialhilfestatistik aufbereitet, zuvor war das Bundesamt für Sozialversicherungen zuständig.

Die Zweckmässigkeit der verwendeten Datenbasis für die Analyse von Verläufen im System der sozialen Sicherheit konnte in unterschiedlichen Studien aufgezeigt werden (siehe z.B. Fluder et. al. 2009[6], Fritschi et al. 2013[7], Fluder et al. 2017[8]).

3.3 Kohortendesign und standardisierter Beobachtungszeitraum

Die jährlichen Grundgesamtheiten im vorliegenden Projekt bilden Kohorten von Personen, die in einem spezifischen Jahr neu Arbeitslosentaggeld beziehen. Die Kohorten wurden folgendermassen definiert:

- Personen, die im Referenzjahr offiziell als arbeitslose Person gemeldet wurden und für die demnach eine Rahmenfrist[9] besteht.
- Sie weisen einen Taggeldbezug nach Eröffnung der Rahmenfrist und im selben Jahr des Beginns der Rahmenfrist auf
- In den 24 Monaten vor dem ersten Taggeldbezug haben sie keine Arbeitslosentaggelder bezogen
- sie erreichen im 48-monatigen Beobachtungszeitraum (siehe unten) das Rentenalter nicht.

Damit werden «neue» Arbeitslosentaggeldbeziehende nicht in einem gesetzlichen Sinne (neue Rahmenfrist), sondern in einem konzeptuellen Sinn definiert (zuvor keine Taggelder). Personen, welche strukturell Arbeitslosentaggelder beziehen, das heisst immer wieder für einzelne Perioden ohne 24-monatigen Unterbruch von der Arbeitslosenversicherung abhängig sind, werden in der Analyse nicht berücksichtigt. Nicht berücksichtigt sind auch Personen, die arbeitslos gemeldet sind, jedoch keine Taggelder beziehen. Damit wird die Operationalisierung einer homogenen Grundgesamtheit mit ähnlichen Voraussetzungen bezüglich des Bezugs von Arbeitslosentaggeldern erreicht. Dies verringert die Komplexität der repräsentierten Verläufe.

Um die Notwendigkeit des Ausschlusses von Personen mit vorangehenden Bezügen von Arbeitslosentaggelder einzuschätzen, wurde eine Sensitivitätsanalyse gemacht. Es wurde analysiert, wie stark sich die Grösse der Kohorte verändert, wenn man Personen ausschliesst, die bis 24 Monate vor dem Beginn des Taggeldbezugs schon einmal Taggelder der Arbeitslosenversicherung bezogen haben. Mit zunehmender Anzahl retrospektiv berücksichtigter Monate nimmt die Anzahl Personen in der Kohorte linear ab. Es findet sich also kein Optimum, um die für die Kohortenbildung retrospektiv berücksichtigten Monate auf ein Minimum zu begrenzen. Es wurden deshalb alle Personen, die innerhalb von 24 Monaten vor dem Beginn des Taggeldbezugs im Kohortenjahr ALV-Taggeldbezüge aufweisen, aus der Kohorte ausgeschlossen. 24 Monate entsprechen einer Rahmenfrist der ALV.

Im Vergleich zum Zwischenbericht wurden an der Implementation der Kohortenbestimmung im Programmcode kleine Korrekturen vorgenommen. Damit haben sich die Grundgesamtheiten leicht verändert. Die Interpretierbarkeit der resultierenden Verlaufsmuster hat sich dadurch verbessert, die Anzahl der Cluster musste nicht angepasst werden.

Für jedes Mitglied einer Kohorte wurde ein standardisierter Beobachtungszeitraum festgelegt: Ab dem Monat mit dem ersten Taggeldbezug (= Monat 1) werden die folgenden 47 Monate für die Analyse berücksichtigt. Dies bedeutet, dass der Monat 1 nicht für alle Kohortenmitglieder im gleichen Monat liegt,

[6] Fluder, Robert, Thomas Graf, Rosmarie Ruder und Renate Salzgeber (2009). «Quantifizierung der Übergänge zwischen Systemen der Sozialen Sicherheit (IV, ALV und Sozialhilfe)». Bundesamt für Sozialversicherungen: Bern.

[7] Fritschi, Tobias, Oliver Hümbelin, Christoph Schaller, Robert Fluder, Bernhard Anrig, Urs Sauter, Kilian Koch, Livia Bannwart, Luca Bösch (2013). «Data Mining mit Administrativdaten der Sozialen Sicherheit». Berner Fachhochschule: Bern.

[8] Fluder, Robert, Renate Salzgeber, Tobias Fritschi, Luzius von Gunten und Larissa Luchsinger (2017). «Berufliche Integration von arbeitslosen Personen». Berner Fachhochschule: Bern

[9] Die Rahmenfrist bezeichnet den Zeitraum von zwei Jahren vor und nach der Anmeldung. Ab dem Referenzdatum der Rahmenfrist können maximal während zwei Jahren ALV-Taggelder bezogen werden. Um überhaupt einen Anspruch zu begründen, müssen im Zeitraum von zwei Jahren vor dem Referenzdatum genügend Beiträge einbezahlt worden sein.

jedoch stets im gleichen Jahr. Pro Kohortenmitglied und für jeden einzelnen der insgesamt 48 Monate im Beobachtungszeitraum ist der Status bekannt (16 verschiedene mögliche Zustände, siehe Abschnitt 0). Durch diese Standardisierung wird für jedes Kohortenmitglied die gleiche Anzahl Monate analysiert und der Beginn des Beobachtungszeitraums ist mit einer einheitlichen Definition festgelegt. Dadurch wird sichergestellt, dass alle Mitglieder ähnliche Voraussetzungen für das Eintreffen nachfolgender Ereignisse (z.B. Sozialhilfebezug) aufweisen. Andere Studien haben aufgezeigt, dass 48 Monate eine optimale Beobachtungsdauer ist[10].

3.4 Sequence Clustering

Im Gegensatz zu bestehenden Forschungsansätzen, die in einem deduktiv-deterministischen Verfahren Verlaufstypen im System der Sozialen Sicherheit bestimmt haben (siehe z.B. Fussnote 8), werden im vorliegenden Projekt typische Verlaufsmuster mit einem Sequenzclusteringverfahren identifiziert (induktives, datengetriebenes Vorgehen oder unsupervised machine learning).

Die öffentliche Statistik hat unter anderem zum Ziel, statistische Informationen für die politische Steuerung zur Verfügung zu stellen. Zu diesem Zweck sind Zeitreihen von zentraler Bedeutung, damit die Entwicklung der interessierenden Phänomene verfolgt werden können. Für das vorliegende Projekt bedeutet dies, dass die typischen Verlaufsmuster im System der Sozialen Sicherheit nicht nur für einen Beobachtungszeitraum aufgezeigt werden müssen, sondern für mehrere Zeiträume, sodass Entwicklungen über die Zeit sichtbar werden.

Die Basis, um zeitliche Entwicklungen aufzeigen zu können, ist ein Set von typischen Verlaufsmustern für einen bestimmten Beobachtungszeitraum, die **initiale Clusterlösung**. Diese Clusterlösung wird im Anschluss auf neue Kohorten übertragen, damit ein sinnvoller Zeitvergleich dieser Muster möglich wird (siehe Abschnitt 3.6). Im vorliegenden Abschnitt wird das Vorgehen für die Etablierung der initialen Clusterlösung beschrieben. Die initiale Clusterlösung wird auf der Kohorte neu ALV-Taggelder beziehender Personen 2010 entwickelt, der frühesten verfügbaren Kohorte.

3.4.1 Methodische und technische Vorüberlegungen

Im vorliegenden Projekt wurde ein zweistufiges Clusterverfahren verwendet, welches zwei Clustering-Ansätze verbindet. Die erste Stufe beruht auf dem k-means Algorithmus, der voraussetzt, dass die Anzahl Cluster im Vornherein bekannt ist. In dieser Stufe werden mit dem k-means Algorithmus 3000 Verlaufscluster gebildet. Der jeweils «repräsentativste» Verlauf aus jedem Cluster wird ausgewählt und in die zweite Stufe übergeben. Die zweite Stufe beruht auf dem «hierarchical clustering»-Algorithmus und erlaubt eine freie Auswahl der definitiven Anzahl Cluster anhand statistischer und fachlicher Kriterien. Die erste Stufe hat im Projekt also die Funktion, die Datenmenge für das Clustering in der zweiten Stufe zu verringern. Mit der zweiten Stufe werden die relevanten Verlaufsmuster identifiziert.

Das direkte «hierarchical clustering» ohne vorgängige Datenreduktion wäre methodisch die geeignetste Vorgehensweise zur Identifikation typischer Verlaufsmuster, da es auf der für Vergleiche von Sequenzen idealen «edit»-Distanz beruht, ein einfach verständliches agglomeratives Verfahren für die Bildung der Cluster beinhaltet und eine etabliert induktive Vorgehensweise ist. Problematisch ist jedoch, dass mit zunehmender Kohortengrösse die Speichergrösse der zugrundeliegenden Distanzmatrix quadratisch und die Rechenzeit kubisch zunimmt[11]. Die Distanzmatrix für die Kohorte mit 126'000 Mitglieder wies annähernd 8 Mrd. Zellen (N*(N-1)/2) auf und verlangte schätzungsweise über 50 Gb Arbeitsspeicher alleine für die Distanzmatrix.

Für ein direktes «hierarchical clustering» konnte bis zuletzt keine technische Lösung gefunden werden (Performance). Auch mit dem Transfer von der BIT-Infrastruktur auf die Data Science Plattform «Renku» des Swiss Data Science Centers konnten die angestrebten Berechnungen nicht durchgeführt

[10] Salzgeber, Renate, Tobias Fritschi, Luzius von Gunten, Oliver Hümbelin und Kilian Koch (2016). «Verläufe in der Sozialhilfe (2006-2011)». Bundesamt für Statistik: Neuchâtel.

[11] http://danifold.net/fastcluster.html?section=1

werden. Die Datenreduktion in der ersten Stufe des zweistufigen Vorgehens war also aufgrund der begrenzten Leistungsfähigkeit der zur Verfügung stehenden Recheninfrastruktur notwendig.

3.4.2 Umgesetztes Vorgehen

Die Umsetzung des zweistufigen Clusteralgorithmus umfasst mehrere Schritte

- Codierung der relevanten Informationen
- K-means-Clustering und Wahl der Repräsentanten zur Datenreduktion
- Hierarchisches Clustering der Repräsentanten
- Übertragung der definitiven Clusterlösung auf die Grundgesamtheit

Umkodierung der relevanten Informationen

Bei den Verlaufsdaten handelt es sich grundsätzlich um kategoriale Informationen (16 verschiedene Zustände, siehe Anhang 7.1). Der K-means-Algorithmus verlangt jedoch nach einer numerischen Repräsentation der Daten, da ihm die euklidische Distanz zugrunde liegt. Zu diesem Zweck werden die monatlichen Statusinformationen in numerische Vektoren umgewandelt:

- Die vierstelligen Stringcodes (siehe Anhang 7.1) werden in numerische Vektoren mit 4 Elementen transformiert. Die vier Elemente stehen weiterhin für die vier betrachteten Systeme.
- Die vierelementigen Vektoren werden so normiert, dass die Summe der Elemente jeweils =1 ergibt
- Liegt kein Bezug oder keine Erwerbsarbeit vor, haben die entsprechenden Vektorelemente den Wert =0

Tabelle 1: Beispiele für Transformation der Zustandscodierung

Zustandsinformation	**Ursprüngliche Codierung** (String)	**Transformierte Codierung** (Vektor mit vier Elementen)
Arbeitslosentaggeld, Erwerbsarbeit	"2121"	(0, 0.5, 0, 0.5)
IV-Rente	"1222"	(1, 0, 0, 0)
IV-Rente, Arbeitslosentaggeld, Sozialhilfe	"1112"	(0.333, 0.333, 0.333, 0)
Weder Sozialleistung noch Erwerbsarbeit	"2222"	(0, 0, 0, 0)

Anmerkung: In der ursprünglichen bedeutet die Codierung «1» einen Bezug/ein Erwerbseinkommen und «2» bedeutet keinen Bezug/kein Erwerbseinkommen

Mit dieser Codierung der Daten werden die Anzahl Spalten der ursprünglichen Datenmatrix von 49 (1x ID + 48x Monatszustände) auf 193 erweitert (1x ID + 48x Monatszustände x 4 Elemente der numerischen Vektoren). Die insgesamt 192 Vektorelemente, welche den 48-monatigen Verlauf im euklidischen Raum repräsentieren, werden direkt dem K-means-Algorithmus übergeben. Weitere Transformationen wurden getestet, die zu analogen Resultaten geführt haben.

K-means-Clustering und Wahl der Repräsentanten zur Datenreduktion

Die initiale Clusterlösung wird auf der Basis der Kohorte 2010 ermittelt. Entsprechend wird das K-means-Clustering[12] auf die rund 126'000 Kohortenmitglieder angewendet, deren Verläufe in diesem Schritt in 3000 Cluster gruppiert werden (K=3000). Die K Schwerpunkte als Startpunkte werden zufällig aus den rund 126'000 Datenpunkten (Zeilen) des Ursprungsdatensatzes ausgewählt. Die Homogenität in den Clustern wird durch den Algorithmus in max. 20 Iterationen optimiert, wobei die Distanzen zwischen den Schwerpunkten und den Datenpunkten (Verläufen) mit der euklidischen Distanz bestimmt und alle 192 Features berücksichtigt werden, die die 48-monatigen Verläufe der Kohortenmitglieder abbilden. Es wird die «k-means»-Funktion aus dem R-base-package «stats» verwendet.

[12] G. James, D. Witten, T. Hastie, R. Tibshirani (2013): An Introduction to Statistical Learning. Springer.

Für das Hierarchische Clustering in der zweiten Stufe wird nun die gesamte Datenmenge (126'000 Verläufe bzw. Kohortenmitglieder) mithilfe der 3000 Clustern reduziert. Zu diesem Zweck wird für jedes der 3000 Cluster der repräsentativste Verlauf bestimmt, der im zweiten Clusteringschritt (hierarchisches Cluster) weiterverwendet wird. Die Repräsentanten werden identifiziert, indem pro Cluster jener Verlauf ermittelt wird, bei dem die Summe aller Distanzen zu den anderen Verläufen im selben Cluster am geringsten ist.

Als Distanzmass wird die edit-Distanz[13] verwendet. Sie findet einen breiten und weit anerkannten Einsatz, um die Ähnlichkeit von Strings zu ermitteln (z.B. in Übersetzungsalgorithmen). Indem die Verläufe, also die Abfolge von jeweils 16 möglichen Zuständen über 48 Monate, als Strings interpretiert werden, ist die edit-Distanz optimal geeignet, um die Ähnlichkeit der Verläufe zu ermitteln. Sie berechnet sich aus den Anzahl Operationen (Insert, Delete, Substitute), welche notwendig sind, um einen Verlauf exakt in einen anderen Verlauf überzuführen. Die Kosten für der Insert/Delete-Operationen werden gleich gewichtet (=1), die Kosten der Substitute-Operationen werden anhand der empirischen Transformationswahrscheinlichkeiten gewichtet.

Hierarchisches Clustering der Repräsentanten und Übertragung der definitiven Clusterlösung auf die Grundgesamtheit

Die initiale Clusterlösung, also die Identifikation typischer Verlaufsmuster, welche für die Fragestellung relevant sind, wird anhand der Verlaufsdaten der 3000 Repräsentanten aus dem vorangehenden Schritt ermittelt. Für diesen Zweck wird ein agglomeratives hierarchisches Clustering auf der Basis der edit-Distanz durchgeführt.

Die Distanzmatrix für die 3000 Repräsentanten wird mit der edit-Distanz berechnet. Auch hier werden die Kosten für Insert/Delete-Operationen gleich (=1) und jene der Substitute-Operationen anhand der empirischen Transformationswahrscheinlichkeiten gewichtet.

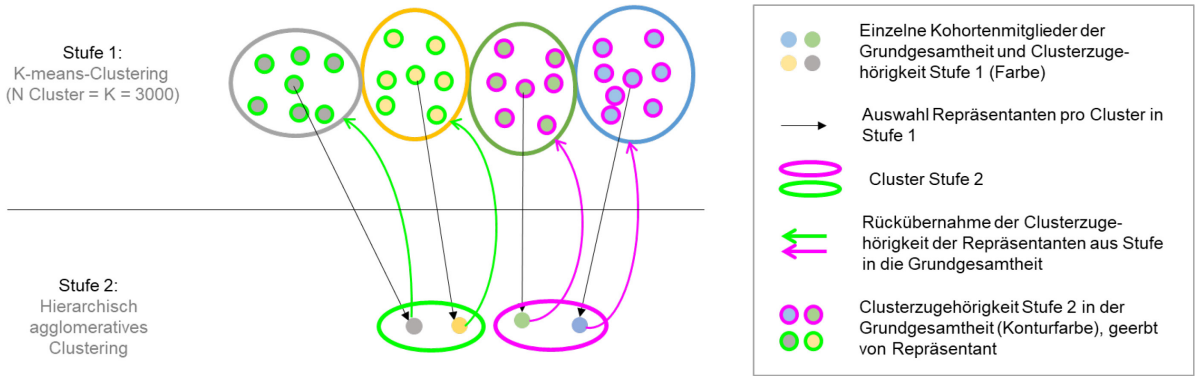
Das klassische agglomerativ-hierarchische Clustering[14] nimmt als Input die Distanzmatrix. Als «agglomeration»-Verfahren wurde mit «ward's minimum variance criterion» eine Varianz-Methode angewendet (in Abgrenzung zu Linkage-Methoden). Dabei wird für jedes Cluster die Summe der quadrierten Distanzen der einzelnen Verläufe vom jeweiligen Cluster-Zentroiden berechnet und über alle Verläufe im selben Cluster aufsummiert. Im nächsten Schritt werden die beiden Cluster vereinigt, bei deren Fusion die geringste Erhöhung der Gesamtsumme der quadrierten Distanzen auftritt. Mit dem Ward-Verfahren wird die grösstmögliche Homogenität der Cluster als Kriterium bei der Clusterbildung verwendet, was der Fragestellung des Projekts entspricht (typische Verlaufsmuster). Der Algorithmus iteriert, bis alle Datenpunkte zu einem grossen Cluster zusammengefügt wurden. Für die initiale Clusterlösung wird angenommen, dass sie mindestens ein und maximal 40 Cluster umfasst. Das ergibt 40 Clusterlösungen mit unterschiedlichen Anzahl Clustern als Kandidaten für die initiale Clusterlösung.

Da diese 40 Clusterlösungen nur für die 3000 Repräsentanten der ersten Clusteringstufe berechnet wurde, müssen die Clusterzugehörigkeiten für jede Lösung auf die gesamte Kohorte 2010 (126'000 Mitglieder) rückübertragen werden. Umgesetzt wird dies, indem alle Clusterangehörige aus der ersten Stufe (K-means) die Clusterzugehörigkeit ihres Repräsentanten aus der zweiten Stufe (hierarchisches Clustering) erben (siehe Abbildung 1).

[13] https://en.wikipedia.org/wiki/Edit_distance

[14] K. Backhaus, B. Erichson, W. Plinke, R. Weiber (2016): Multivariate Analysemethoden. Springer

Abbildung 1: Übersicht zweistufiges Clustering



Quelle: BFS

Wahl der Anzahl definitiver Cluster

Welche der 40 Clusterlösung (Anzahl Cluster) für die Beschreibung typischer Verlaufsmuster im System der Sozialen Sicherheit zielführend ist, wird im Anschluss anhand statistischer und fachlicher Kriterien ermittelt. Basis für diese Evaluation ist die gesamte Kohorte 2010 mit ihren geerbten Clusterzugehörigen aus Stufe 2.

- Statistisches Kriterium: Die Summe der quadrierten Distanzen zwischen allen Verläufen innerhalb eines Clusters

$$S^2 = \frac{1}{2 * n * (n-1)} \sum_{i,j=1}^{n} d(V_i, V_j)^2$$

  wobei $d(V_i, V_j)$ der edit-Distanz zwischen zwei Verläufen und n der Clustergrösse entsprechen. Über alle Cluster aufsummiert kann diese «innerhalb-der-Cluster»-Varianz als Funktion der Anzahl Cluster (Abwandlung des «elbow plots») aufgezeigt werden. Anhand der Knicke in der Kurve, siehe Abbildung 5, erkennt man bei welcher Anzahl Cluster die Gesamt-Homogenität stark zunimmt. Bei Abflachen der Kurve ist der Homogenitätsgewinn nicht mehr gross und zusätzliche Cluster führen somit zu keinem nennenswerten Gewinn.
- Fachliche Kriterien: Lösungen mit verschiedenen Anzahl Clustern werden mithilfe von state distribution plots visualisiert und mithilfe von geeigneten Verlaufsindikatoren beschrieben (siehe Abschnitt 3.5). Dies ermöglicht eine inhaltliche Interpretation der Cluster, also die Einschätzung, ob die in den Clustern abgebildeten typischen Verlaufsmuster aus fachlicher Sicht sinnvoll verstanden werden können.

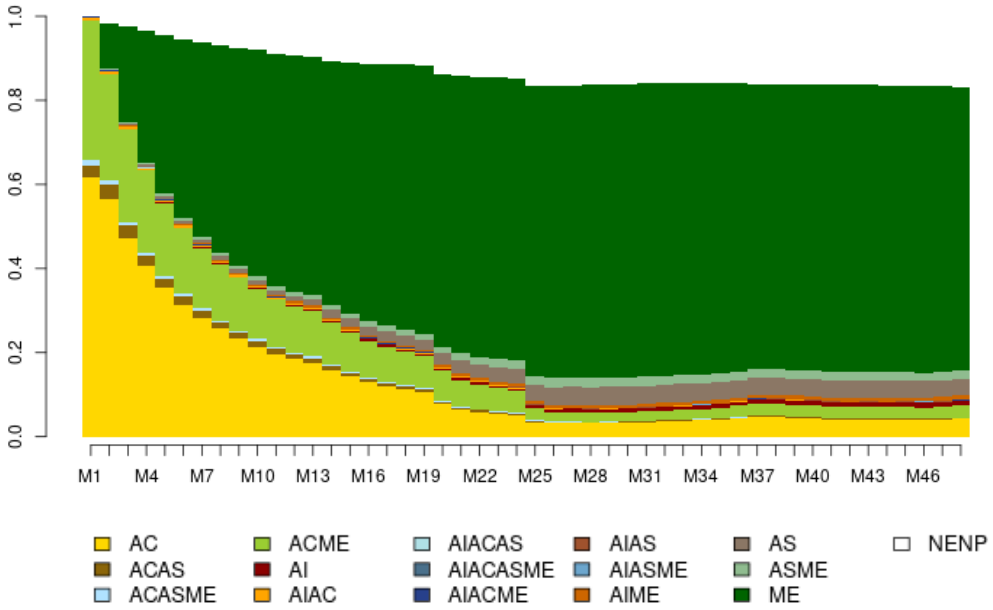
Die anhand der beschriebenen Kriterien gewählte Clusterlösung wird als **initiale Clusterlösung** bezeichnet.

3.5 Visualisierung und Verlaufsindikatoren

Die typischen Verlaufsmuster in der initialen Clusterlösung müssen inhaltlich beschrieben werden, um die Fragestellungen des vorliegenden Projekts zu bearbeiten. Zu diesem Zweck werden Visualisierungen der Verlaufsmuster sowie Verlaufsindikatoren herangezogen.

Für die visuelle Darstellung von Verlaufsdaten eignen sich «state distribution plots». Zur Abbildung typischer Verlaufsmuster wird für die gesamte Kohorte bzw. für ein bestimmtes Cluster von Verläufen die relative Häufigkeitsverteilung der vorgefundenen Zustände pro Monat visualisiert.

Abbildung 2: State distribution plot, gesamte Kohorte 2010



Quelle: BFS - SHIVALV-IK 2010-2014

Abbildung 2 zeigt einen state distribution plot für die gesamte Kohorte 2010. Es ist ersichtlich, dass in Monat 1 rund 60% der Kohorte neuer Arbeitslosentaggeldbeziehender 2010 einzig Arbeitslosenentschädigungen bezieht (AC), etwas weniger als 40% beziehen Arbeitslosenentschädigungen und erzielen gleichzeitig ein Erwerbseinkommen z.B. im Rahmen eines Zwischenverdienstes (ACME). In Monat 48 erzielen rund 70% einzig ein Erwerbseinkommen (ME), etwas weniger als 17% beziehen weder Sozialleistungen noch sind sie erwerbstätig (NENP) und ca. 15% beziehen eine Sozialleistung oder haben andere Kombinationen (Für eine Erklärung der Abkürzungen in Grafiklegende siehe Anhang 7.1). Diese Darstellung von Verlaufsdaten ist sehr gut geeignet um Strukturen in Verlaufsdaten visuell darzustellen; wie sich zeigen wird, können damit Unterschiede in den Verlaufsmustern der Cluster sehr gut abgebildet werden.

Als Alternative wurden Sankeyplots getestet. Sie visualisieren zusätzlich die Zustandswechsel zwischen den einzelnen Zeitpunkten, bilden so die Dynamik innerhalb der Verläufe besser ab und schöpfen das Potential individueller Verlaufsdaten besser aus. Sankeyplots sind jedoch nur geeignet, um Verläufe mit wenigen Beobachtungszeitpunkten darzustellen. Mit zu vielen Beobachtungszeitpunkten ist die Komplexität sehr gross und die Grafik kann kaum mehr zielführend interpretiert werden. Als Lösung kann die Anzahl Beobachtungspunkte reduziert werden (Selektion, Synthese), was jedoch mit Informationsverlust einhergeht. In diesem trade-off bieten state distribution plots die grösseren Vorteile.

Verlaufsindikatoren synthetisieren bestimmte Informationen der individuellen Verläufe der Kohortenmitglieder. So kann z.B. die mittlere Anzahl Monate mit Arbeitslosenentschädigung innerhalb der 48 Monate Beobachtungszeit für die ganze Kohorte oder für die Mitglieder eines Clusters ermittelt werden. Andere Beispiele sind die mittlere Anzahl Bezugssequenzen von Arbeitslosenentschädigung oder der Anteil Kohorten- bzw. Clustermitglieder, der in den 48 Monaten mindestens einmal Sozialhilfe bezogen hat. Die konkret verwendeten Verlaufsindikatoren hängen von der initialen Clusterlösung ab und werden bei den Resultaten in Abschnitt 4.2.2, Tabelle 5 präsentiert.

In der fachlichen Evaluation zur Festlegung der Anzahl Cluster in der initialen Clusterlösung wurden sowohl state distribution plots als auch Verlaufsindikatoren verwendet.

3.6 Vergleich von Clusterlösungen über die Zeit

3.6.1 Motivation und Lösungsweg

Gegenüber deduktiv-deterministischen Verfahren für die Ermittlung von Verlaufstypologien (z.B. Zuordnung von Verläufen zu Verlaufstypen mittels if-then-Bedingungen) bieten Sequenzclusteringansätze (induktive, datengetriebene Verfahren oder unsupervised machine learning) den Vorteil, dass auch unvorhergesehene oder neue Verlaufsmuster sichtbar werden, die in einem deduktiven Verfahren allenfalls nicht antizipiert werden. Der Nachteil ist jedoch, dass die Erstellung von Zeitreihen mit einem Clusteringansatz mit zusätzlichen Schwierigkeiten verbunden ist. Bei deduktiven Verlaufstypologien können dieselben Regeln für die Bestimmung der Typologie in der Ursprungskohorte auf neue Kohorten angewendet werden, was die Vergleichbarkeit der Resultate über die Zeit sicherstellt. Bei Clusteringansätzen sind die Resultate des Algorithmus auf neuen Kohortendaten jedoch kaum vorhersehbar und es ist wahrscheinlich, dass die Gruppierung von ähnlichen Verläufen andere Muster ergibt als bei der Ursprungskohorte, da der Algorithmus mehr oder weniger stark auf Änderungen in den Grunddaten reagiert. Die initiale Clusterlösung für die Kohorte 2010 ist damit nur bedingt mit einer Clusterlösung desselben Algorithmus für die Kohorte 2011 vergleichbar. Entsprechend können Entwicklungen über die Zeit (z.B. Veränderung der Clustergrössen), was für die öffentliche Statistik besonders interessant ist, nicht oder nur ungenügend analysiert werden.

Die Gewährleistung von mehr Stabilität kann dadurch erreicht werden, die initiale Clusterlösung mittels supervised machine learning auf eine neue Grundgesamtheit zu übertragen. Zu diesem Zweck wird auf der Basis der Grunddaten der Kohorte 2010 ein statistisches Prädiktionsmodell trainiert, welches erlaubt für jedes Kohortenmitglied die Clusterzugehörigkeit vorherzusagen. Das Prädiktionsmodell enthält damit die implizit aus den Daten hergeleiteten «Regeln» für die Clusterzuordnung in der Initiallösung. Dieses Modell wird danach auf eine neue Kohorte angewendet, wodurch jeder neue Verlauf anhand der «Regeln» im Prädiktionsmodell jenem Cluster aus der initialen Clusterlösung zugeordnet wird, dass ihm am ähnlichsten ist. Da diese Zuordnung mit einer bestimmten Wahrscheinlichkeit verbunden ist, handelt es sich um ein probabilistisches Verfahren (im Gegensatz zu den theoriegeleiteten, expliziten if-then-Regeln bei deduktiv-deterministischen Verlaufstypologien). Auf aggregierter Ebene erhält man damit ein Abbild der initialen Clusterlösung der Kohorte 2010 in einer neuen Kohorte. Unterschiede zwischen der initialen Clusterlösung und dem Abbild sind auf die Güte des Prädiktionsmodells und Unterschiede in den Grunddaten zurückzuführen.

3.6.2 Entwicklung Prädiktionsmodell

Ziel des Prädiktionsmodells ist es für die Kohorte 2010 anhand eines Sets an Inputvariablen die Clusterzugehörigkeit (Output) in der initialen Clusterlösung vorherzusagen und das Modell mit der besten Prognosegüte zu identifizieren. Um ein gutes Prädiktionsmodell zu trainieren, ist systematisches Experimentieren notwendig; es gibt im Vornherein keine Hinweise darauf, welches Vorgehen das Beste ist (no free lunch). Aus diesem Grund wird das gesuchte Modell über vier Achsen optimiert, den Prädiktonsalgorithmus (5), die Hyperparameter (grid-search), die Features (2 sets) und das Datenformat (3, falls möglich). Die ganze Modellentwicklung wurde mit dem Framework des R-Packages «caret» umgesetzt:

- **Prädiktionsalgorithmen**: Tuning über fünf Algorithmen mit unterschiedlichen Ansätzen, alle angewendeten Algorithmen wurden über die im «caret»-Package verfügbaren Funktionen und Parameter umgesetzt:
    - Random Forest (RF)[15]: Decision-tree-based Ensemble-Classifier mit Bagging-Ansatz und randomisierter Auswahl der features, umgesetzt mit der «ranger»-Funktion (gbm package)
    - Gradient Boosting Machines (GBM)[16]: Decision-tree-based Ensemble-Classifier mit Boosting- und Bagging-Ansatz, umgesetzt mit der «gbm»-Funktion in R (gbm package)

[15] https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

[16] https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/

- Support Vector Machines[17] (SVM Poly): Large Margin Classifier, der Hyperebenen sucht, welche die Cluster im N-dimensionalen Vektorraum (N = Anzahl Prädiktoren im Modell [features]) optimal trennen, es wird nur ein polynomialer Kernel angewendet, umgesetzt mit der «ksvm»-Funktion (kernlab package)
- K-Nearest-Neighbors (KNN): Instance-based Classifier, der die Clusterzugehörigkeit eines neuen Datenpunkts anhand der Zugehörigkeiten der k am nächsten liegenden Datenpunkten ableitet, umgesetzt mit der «knn»-Funktion (class package)
- Neural Networks mit Model Averaging (avNNet)[18,19]: feed-forward neural networks mit einem single hidden layer für multinomiale log-lineare Modelle, mehrere Modelle werden mit unterschiedlichen Intitialgewichten trainiert und dann gemittelt, umgesetzt mit der «avNNet» -Funktion

- **Hyperparameter**: Für jeden Algorithmus können unterschiedliche Ausgangsparameter gesetzt werden (z.B. bei RF die Anzahl decision trees im Modell und bei KNN die Anzahl berücksichtigter Nachbarn). Das Tuning sucht in der Trainingsphase pro Algorithmus jenes Set an Hyperparametern, welches die beste Prädiktionsgüte liefert, und wurde mittels grid search umgesetzt. Bei einigen Hyperparametern wurden in einer vorgelagerten Testphase geeignete Parameterräume festgelegt. Eine Übersicht der optimierten Parameter findet sich in Anhang 7.18.
- **Features**: Die Features, welche die Clusterzugehörigkeit vorhersagen sollen, wurden in zwei Sets aufgeteilt. Das Tuning in der Trainingsphase wird mit beiden Sets durchgeführt:
  - «Withoutsociodem»: Die 48 Monatsvariablen M1 bis M48 mit den darin enthaltenen Statusinformationen (siehe u.a. Anhang 7.1)
  - «Withsociodem»: Die 48 Monatsvariablen M1 bis M48 und zusätzlich folgende sozio-demografische Information (diese Informationen liegen für das Kohortenstartjahr zur Verfügung, in der Regel zum Zeitpunkt des Erstbezugs eines Arbeitslosentaggelds): Geschlecht, Alter in Jahren, Zivilstand, Nationalität, Wohnkanton.
- **Datenformat**: Je nach Algorithmus werden die Daten in unterschiedlichen Formaten verlangt. Aus diesem Grund wurden die Inputdaten in unterschiedlichen Repräsentationen aufbereitet. Jedes Modell in der Trainingsphase wird mit allen erlaubten Datenrepräsentationen berechnet:
  - «Factor»: In diesem Datenformat werden alle Variablen als «Factor» (kategorielle Daten) verstanden. Auch das Alter in Altersjahren, als einzige echte numerische Variable, wird als Factor interpretiert. Verwendete Methoden, die kategorielle Daten erlauben, sind Random Forest und GBM.
  - «Mixed»: Alle kategoriellen Variablen werden als «Factor» verstanden, ausser die Variable Alter, welche genuin als numerische Variable behandelt wird. Verwendete Methoden, die gemischte Datenformate akzeptieren, sind Random Forest und GBM.
  - «One-Hot» : Unter «one hot encoding» (oder Dummy-Codierung) versteht man ein vollständig numerisches Datenformat für Modelle, die keine kategoriellen Variablen verarbeiten können. Eine kategorielle Variable X mit beispielsweise fünf Ausprägungen (X=1,2,3,4,5) wird dabei in fünf binäre Variablen aufgesplittet. X wird also zu X1, X2, X3, X4 und X5, wobei Xi den Wert 1 annimmt, falls X den Wert i aufweist; ansonsten hat Xi den Wert 0. Alle verwendeten Methoden können das «One-hot»-Datenformat verarbeiten. Speziell zu erwähnen sind folgende Variablen
    - Die Monatsvariablen mit den Statusinformationen M1 bis M48 werden jeweils in vier binäre Variablen aufgeteilt, die die vier betrachteten Bereiche repräsentieren (M1 wird z.B. zu M1_IV, M1_ALV, M1_SH, M1_IK)
    - Die Variable kanton wird zu kanton_1, kanton_2… kanton_26
    - Das Alter ist bereits in einem rein numerischen Datenformat und wird nicht «one hot»-codiert.

[17] https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[18] https://rdrr.io/cran/caret/man/avNNet.html

[19] http://www.stats.ox.ac.uk/~ripley/PRbook/

In der Trainingsphase wird für die meisten Kombinationen der Parameter über die vier Achsen (Algorithmen, Hyperparameter, Features, Datenformat) ein Modell gerechnet und evaluiert. Gewisse Kombinationen von Parametern haben sich in der vorgelagerten Testphase als nicht zielführend erwiesen und wurden deshalb weggelassen.

Die Konfiguration der Trainingsphase ist wie folgt:

- Randomisierte Aufteilung der Grunddaten (Kohorte 2010) in 80% der Daten für das Training und 20% der Daten für die finale Evaluation (Testdatenset).
- Berechnung eines Modells für die meisten Kombinationen der Parameter über die vier Achsen «Algorithmen», «Hyperparameter», «Features» und «Datenformat»[20] auf den Trainingsdaten.
- Anwendung von 5-fold cross validation ohne Wiederholung für jedes Modell.
- Das Modell mit der besten cross validation Prädiktionsgüte (im Sinne der mean balanced accuracy, siehe weiter unten) wird als finales Modell ausgewählt.
- Die definitive Prädiktionsgüte wird anhand des Testdatensets ermittelt.

Zuletzt braucht es eine geeignete Metrik, um die Prädiktionsgüte eines Modells zu evaluieren. Eine einfach zu vermittelnde Metrik ist die Accuracy. Sie weist ganz einfach aus wie oft die vorhergesagten Werte im Mittel mit den tatsächlichen Werten übereinstimmen. Bei Prädiktionsproblemen, in welchen die Häufigkeiten der vorherzusagenden Kategorien sehr unterschiedlich sind («imbalanced data»), kann jedoch ein sehr schlechtes Modell eine sehr gute Accuracy erreichen. Wenn z.B. ein vorherzusagendes Cluster1 95% aller Datenpunkte im Trainingsdatensatz umfasst und ein Modell in jedem Fall nur die Zugehörigkeit zum Cluster1 vorhersagt, würde dennoch eine Accuracy von 95% erzielt. Aus diesem Grund wird in diesem Fall eine differenziertere Version der Accuracy, die balanced bzw. mean balanced accuracy angewendet, welche die unterschiedlichen Clustergrössen berücksichtigt. Die mean balanced accuracy wird berechnet als der Mittelwert der Anteile korrekt vorhergesagten Clusterzugehörigkeit pro Cluster. Die Balanced Accuracy ist im vorliegenden Projekt sinnvoll, da die Cluster sehr unterschiedlich gross sind (imbalance).

3.6.3 Prädiktion

Das Prädiktionsmodell, welches in der Trainingsphase die beste Performance gezeigt hat, wird auf eine neue Kohorte angewendet, um dort für jedes Kohortenmitglied eine Clusterzugehörigkeit auf Basis der initialen Clusterlösung zu erhalten. Es wird dabei davon ausgegangen, dass die Validität des Modells für die Zuordnung eines Verlaufs zu einem Cluster gegeben ist. Diese Annahme gilt es zu überprüfen, siehe dazu nachfolgenden Abschnitt.

3.7 Aktualisierung der initialen Clusterlösung

Die Übertragung einer initialen Clusterlösung (Basis Kohorte 2010) auf eine neue Kohorte mittels supervised machine learning (Prädiktion) ermöglicht die Etablierung von stabilen Zeitreihen für clusteringbasierte Verlaufstypologien (siehe vorangehender Abschnitt). Eine neue Schwierigkeit ergibt sich jedoch aus dem Umstand, dass sich die Verläufe im System der Sozialen Sicherheit aufgrund von individuellen Faktoren (z.B. Zusammensetzung der Migrationsbevölkerung, Scheidungshäufigkeit, Familienzusammensetzung) und Kontextfaktoren (z.B. Gesetzesänderungen, wirtschaftliche Schocks) mit der Zeit ändern können und so die initiale Clusterlösung an Relevanz bzw. Validität verliert. Änderungen dieser Faktoren können zu neuen relevanten Clusterlösungen führen. Es braucht daher ein Vorgehen, um entscheiden zu können, wann die Validität der initialen Clusterlösung nicht mehr genügend ist und diese aktualisiert werden muss.

Im Folgenden wird ein Vorgehen beschrieben, das sich auf verschiedene Analysen und Kennzahlen abstützt, um eine informierte Entscheidung zu ermöglichen. Ein Vorgehen, das eine Aktualisierung der initialen Clusterlösung bei Überschreitung von fixen Schwellenwerten vorschreibt, ist nicht zielführend,

[20] Gewisse Kombinationen von Parametern haben sich in der vorgelagerten Testphase als nicht zielführend erwiesen und wurden deshalb weggelassen

da die Schwellenwerte mehr oder weniger willkürlich festgelegt werden müssen. Zudem wären für die Entwicklung entsprechender Schwellenwerte Daten aus einem grösseren Zeitraum notwendig.

Um entscheiden zu können, wann eine initiale Clusterlösung aktualisiert werden muss, sind Vergleiche zwischen drei Clusterlösungen notwendig (zeitunabhängige, generische Bezeichnungen):

- **Referenz**: aktuell gültige initiale Clusterlösung, im Projekt auf Basis der Kohorte 2010
- **Prädiktion**: aktuell gültige initiale Clusterlösung, die mittels Prädiktion auf eine neue Kohorte übertragen wurde, im Projekt Kohorten 2011 bis 2015
- **Neue Clusterlösung**: neue Clusterlösung, welche mithilfe desselben zweistufigen Algorithmus wie bei der initialen Clusterlösung auf einer neuen Kohorte berechnet wurde, im Projekt Kohorten 2011 bis 2015

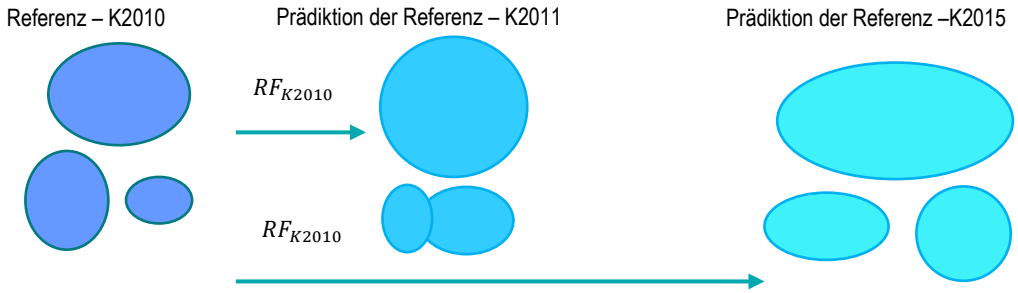
3.7.1 Deskriptive Evaluation der Ähnlichkeit zwischen Referenz und Prädiktion

Die Güte der Prädiktion wurde in der Trainingsphase auf der Kohorte 2010 mittels cross validation score und test score ermittelt. In diesem Schritt werden nun die Daten und die vorhergesagten Clusterzugehörigkeiten in einer nachfolgenden Kohorte mit der Ursprungskohorte 2010 deskriptiv verglichen. Es handelt sich also um einen Vergleich zwischen Referenz und ihren Prädiktionen in späteren Kohorten. Die Vergleiche umfassen folgende Aspekte:

- Visueller Vergleich mithilfe der state distribution plots
- Vergleich der Ausprägungen der Verlaufsindikatoren
- Vergleich der Clustergrössen
- Vergleich der Varianzanteile und weiterer interner Massen der Clustervalidierung

Der letzte Vergleich wird folgend vertieft erläutert: Die Evolution der initialen Clusterlösung (Referenz) in der Zeit, wird durch seine Prädiktionen auf späteren Kohorten ausgedrückt. Diese projizieren die Referenz-Kategorisierung in der Zeit (Abbildung 3). Um diese Projektionen zu evaluieren, werden sowohl bei der initialen Clusterlösung als auch bei seinen Prädiktionen die gleichen internen Masse berechnet. Diese werden als intern bezeichnet, da sie eine Clusterlösung als solches und nicht im Vergleich mit anderen Clusterlösungen charakterisieren (externe Masse werden eingesetzt, um verschiedene Clusterlösungen miteinander zu vergleichen, siehe auch Abschnitt 3.7.3). Sie quantifizieren Eigenschaften wie zum Beispiel Anzahl der Verläufe bzw. Personen pro Cluster, mittlere und maximale Distanzen zwischen den Verläufen innerhalb der Cluster, Varianzanteile innerhalb der Cluster in Bezug zur Gesamtvarianz der Grundgesamtheit und Trennschärfe zwischen den Clustern. Bleiben die internen Masse in der Zeit ungefähr stabil, ist dies ein Indiz dafür, dass die Prädiktionen einige Eigenschaften der initialen Clusterlösung in der Zeit gut konservieren und es noch keinen Grund gibt, die initiale Clusterlösung zu erneuern. Weichen die interne Masse mit der Zeit deutlich von jenen der initialen Clusterlösung ab, ist dies ein Indiz dafür, die initiale Clusterlösung zu aktualisieren.

Abbildung 3: Schematische Darstellung der Übertragung der Initialen Clusterlösung (Referenz) auf zukünftige Kohorten



Anmerkung: Beispiel der Kohorten K2011 und K2015. RF_{K2010} bezeichnet das Prädiktionsmodell (hier random forest) das auf den Daten der Kohorte K2010 trainiert wurde.

Es wird erwartet, dass die Cluster der Prädiktionen mit der Zeit an Kompaktheit verlieren. Dies würde sich durch eine Erhöhung der Varianz innerhalb der Cluster im Vergleich zur Gesamtvarianz ausdrücken. Der Anteil der Varianz innerhalb der Cluster der Clusterlösung zur Gesamtvarianz als auch ebendieser Anteil pro Cluster sind eng mit den «within-sum-of-squares» (wss) verbunden. Letzteres wird über alle Cluster als Summe der quadratischen Distanzen der Verläufe vom «mittleren Verlauf» jedes Clusters berechnet. Da es sich bei den Verläufen einer Kohorte um Sequenzen handelt und ein «mittlerer Verlauf» nur indirekt ermittelt werden kann, benutzen wir für die Varianz innerhalb der Cluster folgenden äquivalenten Ausdruck:

$$\frac{1}{n-1}\sum_{j=1}^{k}\sum_{\boldsymbol{x}\in C_j}\|\boldsymbol{x}-\overline{\boldsymbol{x}}\|^2 = \frac{1}{2n(n-1)}\sum_{j=1}^{k}\sum_{\boldsymbol{x},\boldsymbol{y}\in C_j}\|\boldsymbol{x}-\boldsymbol{y}\|^2$$

C_j ist das Cluster j einer Clusterlösung mit k Clustern, $j = 1, \dots, k$.

Mit zunehmender Clusteranzahl nimmt diese Summe kontinuierlich ab. Im Trivialfall, ist jeder Verlauf in seinem eigenen Cluster und die Varianz innerhalb der Cluster ist gleich null. Wir haben die Anzahl von Clustern der Referenz mithilfe von fachlichen Überlegungen über ihre Zusammensetzung und der Varianz innerhalb der Cluster getroffen (siehe Abschnitt 4.2).

Ein weiteres internes Mass, das sowohl für die Referenz als auch für ihre Prädiktionen in der Zeit berechnet wird, ist der mittlerer Silhouette-Koeffizient. Der Silhouette-Koeffizient s_i, quantifiziert wie gut ein Verlauf i, seinem Cluster zugeordnet ist, indem er die mittleren Distanzen des Verlaufs i von den Verläufen seines Clusters und des nächstgelegenen Clusters berücksichtigt. Es sei a_i die mittlere Distanz des Verlaufs i von allen anderen Verläufen seines Clusters und b_i seine mittlere Distanz zu den Verläufen des nächstliegenden Clusters. Letzteres ist das Cluster mit der kleinsten mittleren Distanz zu i. Der Silhouette Koeffizient s_i des Verlaufs i wird dann wie folgt definiert:

$$s_i = \frac{b_i - a_i}{\max(b_i, a_i)}$$

Verläufe die gut klassifiziert sind, haben eine kleine mittlere Distanz a_i zu den Verläufen des eigenen Clusters und eine grosse mittlere Distanz b_i zu den Verläufen des nächstliegenden Clusters. Somit haben gut klassifizierte Verläufe einen Wert für s_i nahe Eins. Ist s_i nahe Null, könnte dieser Verlauf auch benachbarten Clustern zugeordnet werden. Ein negativer Koeffizient bedeutet, dass der Verlauf sich im falschen Cluster befindet. Der Silhouette Koeffizient s_i nimmt Werte zwischen -1 und 1. Für die ganze Clusterlösung wird der Mittelwert der Silhouette-Koeffizienten s_i aller Verlaufe i der Kohorte berechnet. Diese Grösse wird der Einfachheit halber auch als Silhouette Koeffizient bezeichnet, präziser wäre mittlere Silhouette-Breite (in den Abbildungen, «average Silhouette width»).

Wie bereits erwähnt, eine deutliche Abweichung der internen Massen von jenen der Referenz deutet darauf hin, dass die Referenz aktualisiert werden sollte. Konkret, für die hier beschriebene internen Masse: eine deutliche Zunahme der Varianzanteile und der Distanzen innerhalb der Prädiktionscluster, deutet auf eine Abnahme der Kompaktheit der Referenz-Kategorisierung hin, und eine deutliche Abnahme der Silhouette Koeffizienten deutet auf Verlust der Trennbarkeit der Cluster hin. Solche Entwicklungen der messbaren Grössen Kompaktheit und Trennbarkeit bieten Argumente für die Aktualisierung der Clusterlösung.

Die Berechnung von den oben erwähnten internen Massen ist mit Schwierigkeiten verbunden, da die gesamte Distanzmatrix aller Kohortenmitglieder dafür berechnet werden muss. Dies ist jedoch aufgrund technischer Einschränkungen aktuell nicht möglich (siehe Abschnitt 3.4.1). Aus diesem Grund werden die entsprechenden Kennzahlen auf einfachen Zufallsstichproben der Referenz und der Prädiktion berechnet.

3.7.2 Post-hoc Evaluation der Zuverlässigkeit der Prädiktion

Die Güte der Prädiktion ergibt sich in einem ersten Schritt aus der Analyse der balanced accuracy während der Modellentwicklung in der cross validation und beim finalen Test auf den Testdaten. Nachdem das Prädiktionsmodell auf eine neue Kohorte angewendet wurde, ergeben sich jedoch neue Evaluationsmöglichkeiten. In unserem Fall berechnen die meisten Prädiktionsmodelle für einen neuen Datenpunkt (d.h. einen neuen Verlauf aus einer neuen Kohorte) nicht nur eine Vorhersage der Clusterzugehörigkeit, sie liefern auch Werte für die «Zuordnungsgüte», annähernd eine Wahrscheinlichkeit mit welcher das Prädiktionsmodell sich bei einem bestimmten Datenpunkt für eine Kategorie der Referenz entscheidet. Der neue Verlauf wird jenem Cluster zugeordnet, für welches die höchste Zuordnungsgüte ausgewiesen wird.

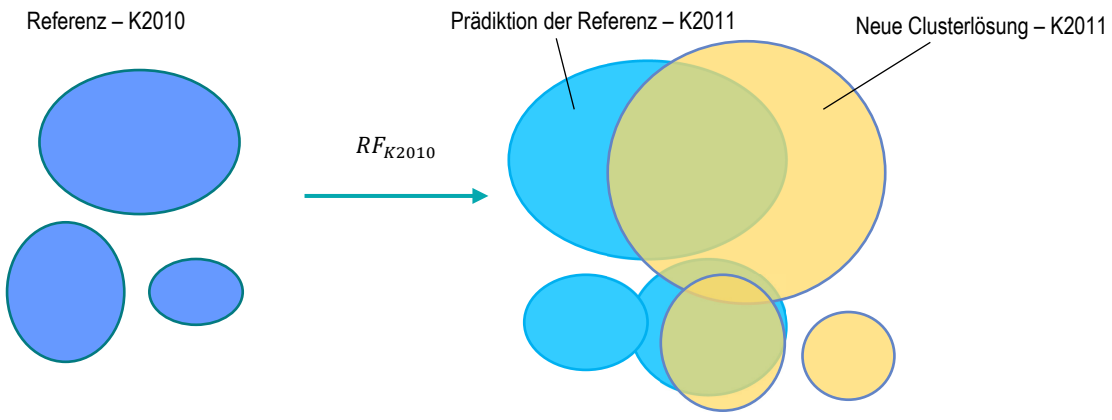
Die Zuordnungsgüte, mit welcher die Clusterzugehörigkeit für einen neuen Verlauf vorhergesagt wird, kann als «Verlässlichkeit der Prädiktion» interpretiert werden. Verringert sich die Zuordnungsgüte in den neuen Kohorten im Vergleich zur Referenz, ist dies ein Hinweis dafür, dass die initiale Clusterlösung nicht mehr zu den neuen Grundgesamtheiten passt. Mögliche Gründe dafür wären zum Beispiel Veränderungen in den Verlaufsmustern oder in der strukturellen Zusammensetzung neuer Kohorten. Die Verteilung dieser Zuordnungsgüte, kann für die Referenz und deren Prädiktionen für die Kohorten K2011 bis K2015 verglichen werden, um eine allfällige Abnahme der Verlässlichkeit der Prädiktion festzustellen.

Wie in Tabelle 12 ersichtlich, weisen diese Verteilungen starke Asymmetrien auf. Daher sehen wir von einem quantitativen Vergleich mithilfe von Mittelwerten ab. Sinnvoller ist, neben den visuellen Vergleichen der Verteilungen der Zuordnungsgüte, die Entwicklung deren Dezile in der Zeit zu beobachten. Nehmen die Anzahl Verläufe der höchsten Dezile mit der Zeit ab, werden immer mehr Verläufe mit zunehmend kleinerer Verlässlichkeit einer Kategorie der Referenz zugewiesen.

3.7.3 Analyse der Korrespondenz zwischen Prädiktion und neuer Clusterlösung in einer bestimmten Kohorte

Bisher wurde auf die Abweichung bzw. Übereinstimmung zwischen Referenz und Prädiktion(en) fokussiert. Um die Notwendigkeit der Aktualisierung der initialen Clusterlösung zu evaluieren, ist es zusätzlich sinnvoll, zu vergleichen, wie gut die Prädiktion der initialen Clusterlösung in einer bestimmten Kohorte mit einer neuen Clusterlösung übereinstimmt, die für diese neuen Kohorte *mit demselben Algorithmus und identischer Parametrierung* berechnet wurde wie die initiale Lösung. Die neue sowie die initiale Clusterlösung ergeben dieselbe Anzahl Cluster. Man kann jedoch davon ausgehen, dass bei Veränderungen in den Verläufen zugrundeliegenden erklärenden Variablen (features), die neue Clusterlösung inhaltlich nicht identisch mit der initialen Lösung sein wird, und dass neue typische Verlaufsmuster (Cluster) auftauchen. Dies wird schematisch in Abbildung 4 dargestellt.

Abbildung 4: Schematische Darstellung des Vergleichs zwischen Prädiktion der Referenz auf der Kohorte 2011 (Hellblau) und einer neuen Clusterlösung für dieselbe Kohorte (Gelb).



Da die beiden Clusterlösungen (initiale und neue Lösung) für dieselbe Grundgesamtheit bzw. Kohorten vorliegen, auf demselben Algorithmus beruhen und die gleiche Anzahl Cluster generieren, kann der Grad ihrer Übereinstimmung mit der (Balanced) Accuracy bzw. mit dem Cohen's-Kappa-Koeffizienten ermittelt werden (externe Masse). Die Accuracy wird im supervised machine learning oft als Kennzahl zur Modellperformance in der Trainingsphase verwendet, wo sie das Verhältnis zwischen den korrekt vorhergesagten Werten in Bezug zum ganzen Trainingssample beschreibt. Im vorliegenden Fall wird sie als Kennzahl verwendet, um zu evaluieren, wie gut eine neue Clusterlösung in der Kohorte XY die Prädiktion der initialen Clusterlösung (K2010) in der Kohorte XY zu reproduzieren vermag, wobei beide Lösungen auf dem exakt gleichen Algorithmus beruhen.

Die «Accuracy» ist in diesem Fall das Verhältnis der Verläufe, die sowohl in der Prädiktion als auch in der neuen Clusterlösung im *inhaltlich* äquivalenten Cluster kategorisiert werden, gegenüber allen Verläufen in der Kohorte. Die «Balanced Accuracy» mittelt diese Anteile ungewichtet über alle Cluster. Dadurch spielt die Clustergrösse keine Rolle, unterschiedlich grosse Cluster tragen gleichermassen zum Wert der «Balanced Accuracy» bei. Bei Cohen´s Kappa, wird von der Accuracy die erwartete Übereinstimmung abgezogen. Letztere ist die Übereinstimmung die sich daraus ergibt, wenn die Verläufe zufällig proportional zur Grösse der Cluster der Referenz zugeordnet werden. Man beachte, dass bei einem grossen Cluster, wie ALV-Kurzzeit, das 52% der Verläufe der Kohorte 2010 gruppiert, viele Verläufe rein zufällig korrekt zugeordnet werden würden. Beim Cohen´s Kappa, geht es nun darum, die accuracy um eben diesen Anteil der zufälligen Übereinstimmung zu korrigieren.

Inwiefern ein Cluster in der Prädiktion und in der neuen Clusterlösung *inhaltlich* übereinstimmend ist, lässt sich jedoch nicht ohne Weiteres feststellen. Dazu bedarf es einer qualitativen Analyse. Zu diesem Zweck, werden visuelle Vergleiche mithilfe der state distribution plots sowie Analysen auf Basis der Konfusionsmatrix der beiden Clusterlösungen umgesetzt (absolute, relative Verteilungen, Jaccard-Matrix). Auf dieser Basis können dann die Cluster aus der neuen Clusterlösung auf Basis qualitativer Einschätzungen einem bekannten Cluster aus der initialen Clusterlösung zugeordnet bzw. neue typische Verlaufsmuster (Cluster) identifiziert werden. Mithilfe der Konfusionsmatrix werden offensichtlich inhaltlich übereinstimmende Cluster in zwei Clusterlösungen identifiziert. Wenn eine inhaltliche Übereistimmung nicht ganz eindeutig ist, kann folgende Grösse, die an der Jaccard-Metrik angelehnt ist, Abhilfe schaffen: $|\alpha \cap \beta|/|\alpha \cup \beta|$, wo α ein Cluster der Prädiktion ist und β ist ein Cluster eines neuen Modells. Die daraus resultierende Kennzahl beschreibt den Anteil der Verläufe, die zwei Cluster α und β aus unterschiedlichen Clusterlösungen gemeinsam haben, gemessen an der Summe aller Verläufen die α und β zugeordnet sind.

4 Resultate

Die Resultate sind in vier Abschnitten strukturiert. Im ersten Abschnitt werden die Grundgesamtheiten (Kohorten) anhand soziodemografischer Merkmale charakterisiert. Im zweiten Abschnitt werden die Resultate zur Herleitung der initialen Clusterlösung mittels Sequenzclustering präsentiert. Drittens finden sich in Abschnitt 4.3 die Resultate bezüglich der Übertragung der initialen Clusterlösung auf weitere Kohorten mittels Prädiktion und in Abschnitt 4.4 zeigen wir viertens die Resultate zur Frage, wann die initiale Clusterlösung aktualisiert werden muss.

4.1 Übersicht zu den Grundgesamtheiten

In Tabelle 2 sind die Grössen der Kohorten sowie deren Verteilungen über die wichtigsten soziodemografischen Merkmale dargestellt. Bei der Bestimmung der Kohorten wurden im Vergleich zu den letzten Zwischenergebnissen (siehe Fussnote 4, S. 6) geringe Korrekturen vorgenommen, wobei einzelne Personen ohne Arbeitslosentaggeld im ersten Beobachtungsmonat oder mit Taggeldbezug in den zwei Jahren vor dem Kohorteneintritt aus den Kohorten gelöscht wurden.

Tabelle 2: Kohorten neu Arbeitslosentaggeld beziehender Personen 2010 bis 2015 nach soziodemografischen Merkmalen

Kohortenjahr ➔	2010		2011		2012		2013		2014		2015	
	N	%	N	%	N	%	N	%	N	%	N	%
Total	123'786	100.0	107'733	100.0	120'452	100.0	127'919	100.0	127'875	100.0	138'043	100.0
Geschlecht												
Männer	63'974	51.7	54'209	50.3	63'297	52.6	68'239	53.4	68'120	53.3	74'282	53.8
Frauen	59'812	48.3	53'524	49.7	57'155	47.5	59'680	46.7	59'755	46.7	63'761	46.2
Nationsklasse												
Schweizer/innen	75'817	61.3	64'313	59.7	70'027	58.1	73'897	57.8	73'167	57.2	77'900	56.4
Ausländer/innen	47'969	38.8	43'420	40.3	50'425	41.9	54'022	42.2	54'708	42.8	60'143	43.6
Altersklasse												
18-24	24'035	19.4	19'966	18.5	22'203	18.4	22'171	17.3	21'743	17.0	23'129	16.8
25-39	53'169	43.0	46'003	42.7	51'529	42.8	55'489	43.4	55'545	43.4	60'198	43.6
40-54	38'663	31.2	34'765	32.3	38'754	32.2	41'502	32.4	41'512	32.5	44'652	32.4
55-65	7'919	6.4	6'999	6.5	7'966	6.6	8'757	6.9	9'075	7.1	10'064	7.3
Zivilstand												
Ledig	59'723	48.3	50'768	47.1	57'792	48.0	61'688	48.2	62'029	48.5	67'510	48.9
Verheiratet	49'625	40.1	43'929	40.8	48'807	40.5	51'535	40.3	51'398	40.2	55'185	40.0
Verwitwet	833	0.7	730	0.7	746	0.6	811	0.6	758	0.6	814	0.6
Geschieden	12'667	10.2	11'525	10.7	12'423	10.3	13'154	10.3	12'991	10.2	13'884	10.1
Getrennt	920	0.7	763	0.7	669	0.6	712	0.6	696	0.5	644	0.5
Ohne Angabe	18	0.01	18	0.0	15	0.0	19	0.0	3	0	6	0

Quelle: BFS - SHIVALV-IK 2010-2015

Zu den Grundgesamtheiten können folgende Beobachtungen festgehalten werden:

- Die Grösse der Kohorten bleibt zwischen 2010 und 2015 relativ stabil nahe bei 125'000 Personen mit Ausnahme der Kohorten 2011 (~110'000) und 2015 (~138'000)
- Die Verteilung nach Geschlecht ist relativ ausgeglichen mit einem leicht höheren Anteil für die Männer. Der Unterschied hat über den Betrachteten Zeitraum tendenziell zugenommen (51.7% Männer 2010 gegenüber 53.8% 2015)
- Der Anteil der Schweizerinnen und Schweizer hat über die betrachteten Kohorten kontinuierlich abgenommen (61.3% im Jahr 2010 gegenüber 56.4% im Jahr 2015)
- Die 25-39-Jährigen sind mit etwas mehr als 40% die grösste Altersgruppe in allen Kohorten, gefolgt von den 40-54-Jährigen (ca. 30%). 55-65-Jährige bilden mit 7.3% und weniger eine Minderheit. Die Häufigkeitsverteilung über die Altersklassen in den Kohorten ist zwischen

2010 und 2016 sehr stabil geblieben. Die grössten Bewegungen findet sich bei 18-24-Jährigen, die 2010 19.4% und 2015 16.8% der Kohorten ausmachten. Effekte der Reform der Arbeitslosenversicherung im Jahr 2011 auf die Altersverteilung in den Kohorten sind keine sichtbar, da vor allem die Bezugsdauer von Arbeitslosentaggeldern altersabhängig angepasst wurde jedoch nicht die Eintrittsschwelle in den Leistungsbezug. Die Reform wirkt sich daher weniger auf die Altersstruktur der Kohorten, sondern eher auf die eigentlichen Verläufe aus.

- Die grosse Mehrheit der Kohortenmitglieder sind ledig (~48%) oder verheiratet (~40%). Rund 10% sind geschieden und weniger als 2% getrennt oder verwitwet. Dieser Anteil bleibt über den betrachteten Zeitraum in den Kohorten stabil.

Im state distribution plot auf Seite 14 (Abbildung 2) sind die aggregierten Verläufe der gesamten Kohorte 2010 abgebildet, welche die Basis für die initiale Clusterlösung ist (die state distribution plots für die Gesamtheit der Kohorten 2011-2015 finden sich in Anhang 7.3 bis 7.17). Zu Beginn der 48-monatigen Beobachtungsdauer, im Monat 1 (M1) beziehen alle Kohortenmitglieder Arbeitslosentaggelder. Etwas weniger als 40% sind zugleich erwerbstätig oder in einem Zwischenverdienst engagiert (hellgrüne Fläche), 60% beziehen ausschliesslich ALV-Taggelder. Den meisten gelingt der Wiedereinstieg in die Erwerbsarbeit: Rund 70% sind im Monat 25 erwerbstätig ohne Sozialleistungsbezug (dunkelgrüne Fläche); Dieser Anteil bleibt bis zum Ende der Beobachtungsperiode in etwa stabil. Parallel zur Abnahme des Anteils Taggeld beziehender Kohortenmitglieder nimmt der Anteil der Personen, welche Leistungsbezüge aus der IV und der Sozialhilfe aufweisen, zu Beginn kontinuierlich zu. Diese Anteile bleiben ab ca. dem 25. Monaten ebenfalls stabil. Zusammen mit Personen, die weiterhin oder erneut auf ALV-Taggelder angewiesen sind, betrifft dies ab Monat 25 rund 15% der Kohorte. Im gleichen Zeitraum bewegt sich der Anteil jener Personen in der Kohorte, die sich entweder aus dem Erwerbsleben und den beobachteten Sozialleistungen zurückziehen oder aus der Schweiz ausreisen (oder versterben), ebenfalls bei 15% (weisse Fläche).

State distribution plots zeichnen für jeden Beobachtungszeitpunkt die relative Häufigkeitsverteilung der aktuellen Status ab. Informationen zu den Verläufen auf individueller Ebene gehen hierbei verloren. Die Darstellungsweise lässt dennoch vermuten, dass ab dem 25. Beobachtungsmonat kaum mehr Dynamik in den individuellen Verläufen vorkommt, womit die 48-monatige Beobachtungsdauer hinterfragt werden kann. Einige der typischen Verlaufsmuster, die mit dem Sequenzclustering identifiziert werden, zeigen jedoch auch nach dem 25. Beobachtungsmonat erhebliche Dynamiken (siehe folgende Abschnitte).

4.2 Initiale Clusterlösung

Für die Bildung der initialen Clusterlösung wird das im Abschnitt 3.4 beschriebene Vorgehen umgesetzt. Sie wurde auf der Basis der Kohorte 2010 berechnet. Dies stellt sicher, dass für die Erstellung von Zeitreihen eine maximale Anzahl nachfolgender Kohorten zur Verfügung steht.

4.2.1 Lösung mit zehn Clustern

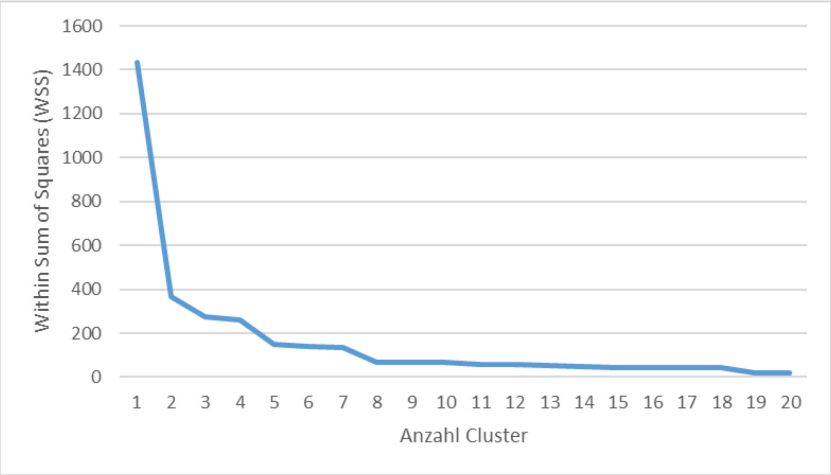
Die Einschätzung aus dem Zwischenbericht (siehe Fussnote 4), dass zehn Cluster zur Beschreibung der typischen Verlaufsmuster neu Arbeitslosentaggeld beziehender Personen im System der Sozialen Sicherheit zielführend ist, hat sich auch mit den neuen, leicht korrigierten Kohorten nicht verändert. Bei der Auswahl der Anzahl Cluster für die initiale Clusterlösung wird aus statistischer Sicht auf die Heterogenität innerhalb der Cluster geachtet (siehe Abschnitt 3.4.2).

In Abbildung 5 wird dargestellt, wie sich die Within-Sum-of-Squares (die Summe der quadrierten Distanzen zwischen allen Verläufen innerhalb eines Clusters über alle Cluster aufsummiert) als Funktion der Anzahl Cluster in einer Clusterlösung entwickelt. Dabei zeigt sich, dass die Heterogenität in Clusterlösungen mit bis zu fünf Clustern jeweils stark abnimmt. Die letzte deutliche Abnahme zeigt sich zwischen einer Lösung mit sieben und acht Clustern. Dort differenziert sich das wichtige Cluster «ALV Langzeit» aus (siehe Tabelle 4), so dass aus fachlicher Sicht eine Lösung mit mehr als sieben Clustern klar zu bevorzugen ist.

Eine Clusterlösung mit mehr als 8 Clustern ist anhand der statistischen Kriterien nicht zwingend angezeigt. Die Heterogenität nimmt ab dem 8. Cluster nicht mehr wesentlich ab. Aus fachlicher Sicht ist jedoch eine Lösung mit 10 Clustern interessant, da die letzten beiden Ausdifferenzierungen relevante Erkenntnisse über die Verlaufsstruktur im System der Sozialen Sicherheit ermöglichen:

- Mit der Ausdifferenzierung eines neunten Clusters teilt sich ein Cluster mit Verläufen, die hauptsächlich durch IV-Bezüge gekennzeichnet sind, in zwei Verlaufsmuster mit IV-Verläufen auf, einmal mit und einmal ohne Erwerbsarbeit. Diese Unterscheidung ist fachlich relevant.
- Mit einer Lösung mit zehn Clustern teilt sich ein Sozialhilfecluster, welches von Sozialhilfebezug ohne gleichzeitige Erwerbstätigkeit gekennzeichnet ist, in zwei neue Sozialhilfecluster auf. Die Beschreibungen finden sich in Tabelle 4 unter «Sozialhilfe wiederholt» und «Sozialhilfe neu». Diese Unterscheidung ist von Bedeutung, da ersteres Verlaufsmuster langwierige Sozialhilfeverläufe ev. mit mehreren Bezugsperioden abbildet während letzteres auf tatsächliche Neueintritte in die Sozialhilfe verweist. Der grösste Teil der Neueintretenden wird keine Verstetigung des Sozialhilfebezugs erfahren.

Abbildung 5: Annäherung eines Elbow-Plots, Kohorte K2010



Anmerkung: Die Annäherung ist aus den Distanzen zwischen den 3000 Repräsentanten des ersten Clustering-Schrittes berechnet, gewichtet durch die entsprechenden Clustergrössen.
Quelle: BFS - SHIVALV-IK 2010-2014

4.2.2 Inhaltliche Interpretation

Die fachliche Interpretation wird anhand von state distribution plots und Verlaufsindikatoren durchgeführt (siehe Abschnitt 3.5). Insgesamt differenzieren sich die Cluster grob entlang folgender Achsen:

- Bezug von Arbeitslosentaggelder alleine oder Bezug von weiteren Sozialleistungen (Sozialhilfe, IV)
- Dauer des Bezugs von Arbeitslosentaggeldern bzw. bis zum Austritt aus der Arbeitslosenversicherung
- Gleichzeitigkeit von Bezug einer Sozialleistung und Erwerbsarbeit

Für die fachliche Interpretation der Cluster werden Verlaufsindikatoren berechnet, die verschiedene Aspekte dieser Achsen abbilden. Sie sind in Tabelle 5 aufgelistet. Die typischen Verlaufsmuster in den Clustern sowie die fachliche Interpretation ist Tabelle 4 aufgezeigt. Um die Interpretation der Grafiken zu vereinfachen, seien hier die Farblegende und wichtigsten Ausprägungen erneut aufgeführt (siehe auch Anhang 7.1):

Tabelle 3: Legende state distribution plots und wichtigste Ausprägungen

Legende:

AC	ACME	AIACAS	AIAS	AS	NENP
ACAS	AI	AIACASME	AIASME	ASME	
ACASME	AIAC	AIACME	AIME	ME	

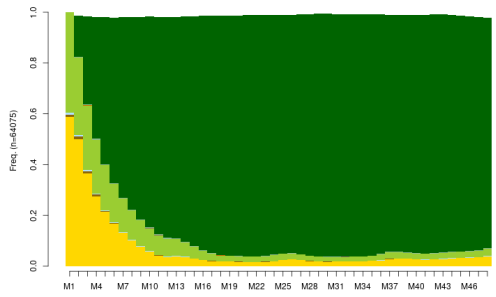
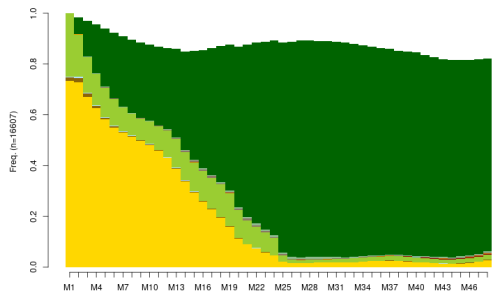
Fünf Grundstatus als Basis der 16 möglichen Status:

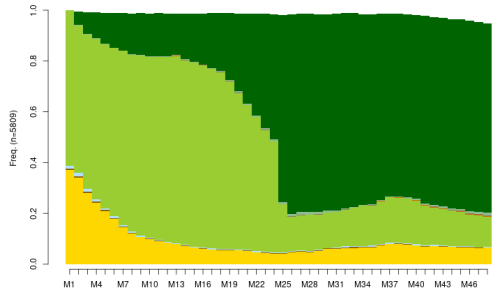
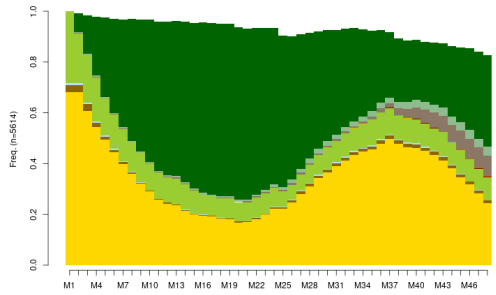
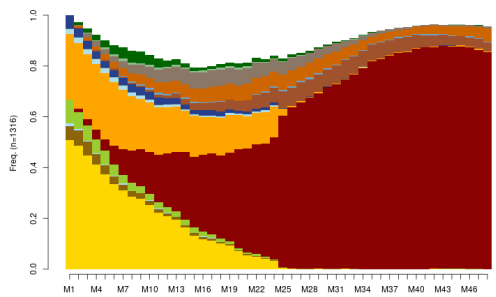
- ME: Erwerbsarbeit (**M**arché de l'**e**mploi)
- AC: Arbeitslosentaggeld (**A**ssurance **c**hômage)
- AS: Sozialhilfe (**A**ide **s**ociale)
- AI: IV-Rente (**A**ssurance **i**nvalidité)
- NENP: Weder Sozialleistung noch Erwerbsarbeit (**N**i en **e**mploi **n**i **p**restations)

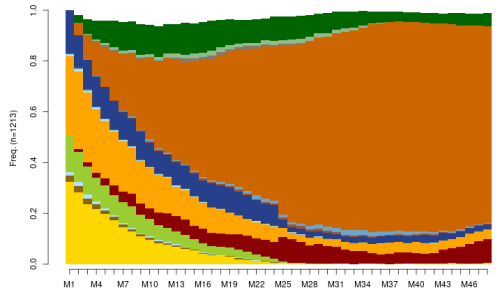
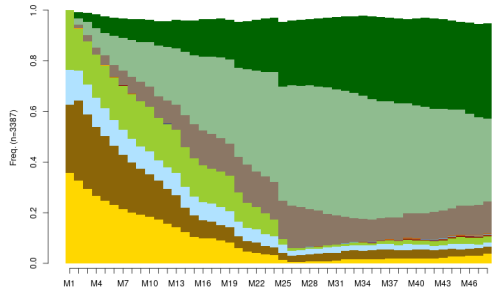
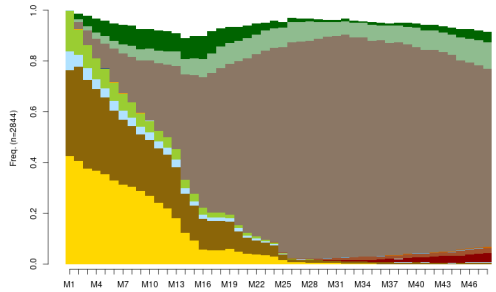
Wichtigste Ausprägungen:

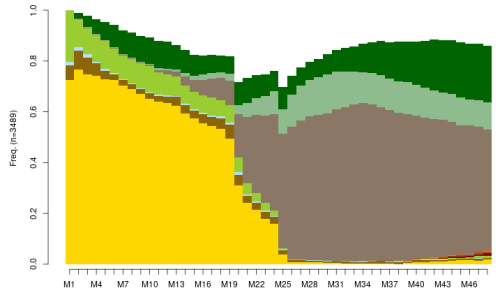
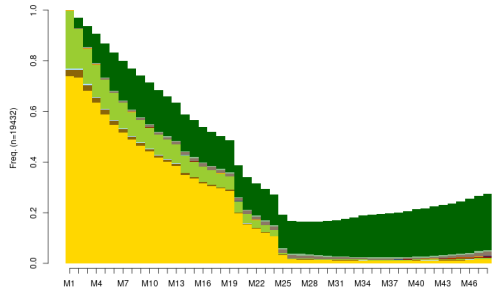
- Gelb (AC): Arbeitslosentaggeld
- Dunkelgrün (ME): Erwerbsarbeit
- Hellgrün (ACME): Arbeitslosentaggeld und Erwerbsarbeit kombiniert
- Dunkelrot (AI): IV-Rente
- Dunkelorange (AIME): IV-Rente und Erwerbsarbeit kombiniert
- Grau-Braun (AS): Sozialhilfe
- Türkis-Grün (ASME): Sozialhilfe kombiniert mit Erwerbsarbeit
- Braun (ACAS): Arbeitslosentaggeld kombiniert mit Sozialhilfe
- Weiss (NENP): Kein Leistungsbezug, keine Erwerbsarbeit

Tabelle 4: Typische Verlaufsmuster und fachliche Interpretation

Nr.	State distribution plots, Kohorte 2010	Label und Interpretation
1		**«ALV Kurzzeit»** - Verläufe mit Integration in den Arbeitsmarkt nach kurzem Bezug Arbeitslosentaggeldern - Im Mittel dauert die erste ALV-Bezugsperiode 4 Monate; nach rund 4 Monaten sind rund 50% der Personen wieder voll im Arbeitsmarkt integriert - Mit 44 Monaten ist die mittlere Dauer mit Erwerbsarbeit in der Beobachtungsperiode hoch. N=64'075 / Anteil an Kohorte=52% Zentrale Verlaufsindikatoren: VI2, VI5
2		**«ALV Langzeit»** - Verläufe mit Integration in den Arbeitsmarkt nach langem ALV-Taggeldbezug - Im Mittel dauert die erste ALV-Bezugsperiode mit 10 Monaten überdurchschnittlich lange; nach ca. 14 Monaten sind rund 50% der Personen wieder voll im Arbeitsmarkt integriert - Mit 32 Monaten liegt die mittlere Dauer mit Erwerbsarbeit in der Beobachtungsperiode im Schnitt der gesamten Kohorte. N=16'607 / Anteil an Kohorte=13% Zentrale Verlaufsindikatoren: VI2, VI5

3		**«Zwischenverdienst»** - Verläufe mit Integration in den Arbeitsmarkt nach langer Arbeitslosigkeit mit ALV-subventioniertem Zwischenverdienst (Taggeld) oder Teilzeitarbeitslosigkeit (bei Verlust einer von mehreren Stellen) - Im Vergleich führt diese Situation zu sehr hohen ALV-Bezugsdauern von durchschnittlich 24 Monaten über die gesamte Beobachtungsperiode - Der gleichzeitige Bezug von Taggeldern und Erwerbsarbeit dauert in der Beobachtungsperiode im Schnitt 19 Monate - Steiler Abfall der ALV-Bezüge ab Monat 20 und v.a. im Monat 24, da Anspruch auf ALV in der Regel nach 24 Monaten erlischt - Ein Teil der Personen bezieht danach erneut ALV-Taggelder, oft in Kombination mit einem Erwerb. N=5'809 / Anteil an Kohorte=5% Zentrale Verlaufsindikatoren: VI1, VI3, VI5
4		**«ALV Mehrfach»** - Verläufe mit mehreren ALV-Bezugsperioden und zwischenzeitlicher Erwerbsarbeit - Im Schnitt weisen die Personen in diesem Cluster 2.8 ALV-Bezugsperioden auf. - Nach der anfänglichen Arbeitslosigkeit steigt der Anteil ALV-Taggeldbeziehenden ab Monat 21 wieder an und erreicht im Monat 37 mit rund 60% (inkl. hellgrün) den zweiten Höchststand. - Die mehrfachen Bezugsperioden führen zu einer überdurchschnittlichen mittleren ALV-Bezugsdauer von 22 und einer unterdurchschnittlichen Gesamtdauer der Erwerbsperioden von 26 Monaten - Ein Viertel der Kohortenmitglieder sind in der Beobachtungsperiode mind. einmal auf eine Zahlung der Sozialhilfe angewiesen. N=5'614 / Anteil an Kohorte=5% Zentrale Verlaufsindikatoren: VI1, VI4, VI5
5		**«IV-Rente»** - Verläufe, die nach anfänglichem Bezug von Arbeitslosentaggeld zu einer IV-Rente ohne Erwerbsarbeit führen - Der ALV-Taggeldbezug vor dem IV-Entscheid dauert durchschnittlich 10 Monate lang, nach Ende der Rahmenfrist von 24 Monaten bezieht niemand mehr ALV-Taggelder - Alle Clustermitglieder beziehen (zumindest vorübergehend) eine IV-Rente, die Bezugsdauer in der Beobachtungsperiode liegt bei durchschnittlich 34 Monaten - Rund 30% der Clustermitglieder sind auf Überbrückungsleistungen der Sozialhilfe von durchschnittlich 5 Monaten angewiesen - Das Cluster umfasst auch Personen, die bereits vor Beginn der Arbeitslosigkeit eine IV-Rente bezogen (hellorange). Ihr Anteil nimmt über die Zeit nur wenig ab. Dies zeigt die Schwierigkeiten von invaliden Personen bei der beruflichen Reintegration auf. N=1'316 / Anteil an Kohorte=1% Zentrale Verlaufsindikatoren: VI9, VI10

6		**«IV-Rente und Erwerb»** - Verläufe, die nach anfänglicher Arbeitslosigkeit zu einer IV-Teilrente mit Erwerbsarbeit führen - Die Dauer der anfänglichen Arbeitslosigkeit dauert in diesem Cluster im Schnitt weniger lange als beim Cluster «IV-Rente». Der Übergang in die IV ist immer wieder mit Erwerbsarbeit kombiniert (hellgrün, dunkelgrün). - Alle Clustermitglieder beziehen (zumindest vorübergehend) eine IV-Rente und die Dauer von IV-Rente und Erwerbsarbeit kombiniert liegt im Schnitt bei 31 Monaten. - In diesem Cluster leistet die Sozialhilfe weniger Überbrückungshilfen als im Cluster «IV-Rente» (bei 16% der Personen). Die Erwerbseinkommen helfen hier stärker den Lebensunterhalt zu stemmen. - Auch dieses Cluster umfasst Personen, die bereits vor Beginn der Arbeitslosigkeit eine IV-Rente bezogen (hellorange, dunkelblau). N=1'213 / Anteil an Kohorte=1% Zentrale Verlaufsindikatoren: VI9, VI11
7		**«Sozialhilfe und Erwerb»** - Verläufe, die nach anfänglichem ALV-Taggeld zu Sozialhilfebezug kombiniert mit Erwerbsarbeit führen. - Alle Clustermitglieder beziehen in der Beobachtungsperiode mind. einmal Sozialhilfe - Im Schnitt beziehen sie in der Beobachtungsperiode 28 Monate Sozialhilfe, in 18 Monaten davon sind sie gleichzeitig erwerbstätig. - Dabei weisen sie im Schnitt 2.4 separate Erwerbsperioden aus und arbeiten im Schnitt 33 Monate, teilweise reicht das Einkommen aus, um nicht auf Sozialhilfeleistungen angewiesen zu sein (zunehmend im Verlauf der Beobachtungsperiode → dunkelgrün). - Der Anteil der Sozialhilfebeziehenden bleibt über die ganze Beobachtungsperiode in etwa konstant, was darauf hindeutet, dass die Situation «Sozialhilfe mit Erwerbstätigkeit» dauerhafte oder sich wiederholende Zustände sind. - Entsprechend beziehen rund 40% der Clustermitglieder bereits zu Beginn des ALV-Taggeldbezugs Sozialhilfe (vermutlich, weil sie bereits zuvor darauf angewiesen waren). N=3'387 / Anteil an Kohorte=3% Zentrale Verlaufsindikatoren: VI6, VI8
8		**«Sozialhilfe wiederholt»** - Verläufe, die nach anfänglichem ALV-Taggeldbezug zu Sozialhilfebezug ohne weiteren Leistungsbezug bzw. Erwerbsarbeit führen. - Die anfängliche Arbeitslosigkeit dauert mit 9 Monaten überdurchschnittlich lange. - Alle Clustermitglieder beziehen in der Beobachtungsperiode mind. einmal Sozialhilfe und verbleiben dort im Schnitt 37 Monate. - Auch in diesem Cluster beziehen rund 40% der Mitglieder bereits zu Beginn der Arbeitslosigkeit Sozialhilfe, mit grosser Wahrscheinlichkeit, weil sie bereits zuvor darauf angewiesen waren. N=2'844 / Anteil an Kohorte=2% Zentrale Verlaufsindikatoren: VI2, VI6, VI7

9		**«Sozialhilfe neu»** - Verläufe, die nach anfänglichem Arbeitslosentaggeld in die Sozialhilfe mit marginaler Erwerbsarbeit führen. - Der anfängliche ALV-Taggeldbezug dauert im Schnitt 13 Monate. Der Anteil Taggeldbeziehende fällt mit der Aussteuerung ab Monat 19 stark ab. - Gleichzeitig steigt der Anteil Personen mit Sozialhilfebezug stark an. Alle Clustermitglieder beziehen mind. einmal Sozialhilfe in der Beobachtungsperiode, der im Schnitt 20 Monate dauert. - Ein Teil der Clustermitglieder zieht sich vorübergehend aus dem Arbeitsmarkt und den Sozialwerken zurück (→weiss), insbesondere im Übergang zwischen ALV und Sozialhilfe. - Ein kleiner Teil der Clustermitglieder ist während der Sozialhilfephase erwerbstätig, oft gleichzeitig zum Sozialhilfebezug. - Im Unterschied zum Cluster 8 verweisen diese Verläufe auf Personen, die neu auf Sozialhilfe angewiesen sind, da sie den Wiedereinstieg in den Arbeitsmarkt aus der Arbeitslosigkeit heraus nicht geschafft haben. N=3’489 / Anteil an Kohorte=3% Zentrale Verlaufsindikatoren: VI2, VI6, VI7
10		**«Leavers»** - Verläufe, die nach anfänglicher Arbeitslosigkeit zu einem «Systemaustritt» führen, das heisst sie beziehen weder eine der drei Sozialleistungen noch sind sie erwerbstätig - Die anfängliche Arbeitslosigkeit dauert im Schnitt 10 Monate und liegt damit über dem Durchschnitt. - Der Anteil der Personen, welche «aus dem beobachteten System austreten» nimmt bis Monat 25 kontinuierlich zu; Dieser Zustand dauert im Schnitt 28 Monate an. - Gründe für den «Systemaustritt» können die folgenden sein: Rückzug aus dem Erwerbsleben (z.B. wegen familiären Pflichten oder fortgeschrittenem Alter), Ausbildungsphasen ohne Erwerbsarbeit, Ausreisen aus der Schweiz, Todesfälle. - Gegen Ende nimmt der Anteil der Clustermitglieder, die im Arbeitsmarkt wieder aktiv sind, leicht zu. N=19’432 / Anteil an Kohorte=16% Zentrale Verlaufsindikatoren: VI12

Anmerkungen: ALV=Arbeitslosenversicherung/-taggeld, SH=Sozialhilfe, IV=Invalidenversicherung
Quelle: BFS - SHIVALV-IK 2010-2014

Tabelle 5: Verlaufsindikatoren nach Cluster, Kohorte 2010 (Mittelwerte oder Anteile pro Cluster)

ID	Verlaufsindikator ↓ Cluster →	ALV Kurz-zeit 1	ALV Lang-zeit 2	Zwischen-verdienst 3	ALV mehr-fach 4	IV-Rente 5	IV-Rente und Erwerb 6	Sozialhilfe und Erwerb 7	Sozialhilfe wiederholt 8	Sozialhilfe neu 9	Leavers 10	Kohorte insgesamt
VI1	Anzahl Monate mit ALV	7	13	**24**	**22**	12	12	15	11	17	12	11
VI2	Dauer der ersten ALV-Bezugsperiode (Monate)	**4**	**10**	14	7	10	8	10	**9**	**13**	10	7
VI3	Anzahl Monate ALV und Erwerbsarbeit kombiniert	3	3	**19**	5	1	5	6	2	2	2	4
VI4	Anzahl Bezugsperioden ALV	1.6	1.6	2.7	**2.8**	1.3	1.6	1.9	1.5	1.6	1.4	1.7
VI5	Anzahl Monate mit Erwerbsarbeit	**44**	**32**	**43**	**26**	4	36	33	7	11	10	34
VI6	Anteil Personen mit mindestens einer SH-Bezugsperiode	5%	10%	10%	26%	30%	16%	**100%**	**100%**	**100%**	12%	16%
VI7	Anzahl Monate mit SH	0	0	1	2	5	2	28	**37**	**20**	1	3
VI8	Anzahl Monate SH und Erwerbsarbeit kombiniert	0	0	0	1	0	1	**18**	4	4	0	1
VI9	Anteil Personen mit mindestens einer IV-Bezugsperiode	0%	1%	0%	0%	**100%**	**100%**	1%	7%	2%	1%	3%
VI10	Anzahl Monate mit IV	0	0	0	0	**34**	38	0	1	0	0	1
VI11	Anzahl Monate IV und Erwerbsarbeit kombiniert	0	0	0	0	3	**31**	0	0	0	0	0
VI12	Anzahl Monate ohne Erwerbsarbeit und Sozialleistungen	1	6	1	3	5	1	2	3	7	**28**	6
Anzahl Personen		*64'075*	*16'607*	*5'809*	*5'614*	*1'316*	*1'213*	*3'387*	*2'844*	*3'489*	*19'432*	*123'786*
Anteil an Kohorte		*52%*	*13%*	*5%*	*5%*	*1%*	*1%*	*3%*	*2%*	*3%*	*16%*	*100%*

Anmerkungen: Die rote Schrift hebt Ausprägungskombinationen von Indikatoren hervor, die für ein bestimmtes Cluster charakteristisch sind.
Quelle: BFS - SHIVALV-IK 2010-2014

Anhand der Verlaufsmuster in den Clustern lässt sich insgesamt sagen, dass rund 70% der Personen in Kohorte 2010 nach anfänglichem Bezug von Arbeitslosentaggeldern ohne grössere Verwerfungen in der Erwerbsbiografie den Wiedereinstieg in den Erwerbsprozess finden; dies sind Personen aus den Clustern «ALV kurz» (52%), «ALV lang» (13%) und «Zwischenverdienst» (5%). Weitere 16% der Kohortenmitglieder ziehen sich aus unterschiedlichen Gründen aus dem Erwerbsleben zurück oder reisen aus, ohne dass weitere Unterstützung aus der ALV notwendig ist oder auf die IV oder die SH rekurriert werden muss, respektive kann («Leavers»). Bei weiteren 2% wird im Laufe des initialen Taggeldbezugs eine (teilweise) Erwerbsunfähigkeit aus gesundheitlichen Gründen festgestellt, so dass sie Ansprüche auf finanzielle Leistungen der IV geltend machen können («IV Rente» und «IV Rente und Erwerb»).

Bei 5% der Personen in der Kohorte zeigt das Verlaufsmuster eine Geschichte von wiederholtem Bezug von Arbeitslosentaggeldern; Die Anzeichen einer Prekaritätsspirale verdeutlichen sich durch den Umstand, dass rund ein Viertel von Ihnen kurzfristig auf Sozialhilfeunterstützung angewiesen ist («ALV mehrfach»). Sozialhilfebezug kommt in unterschiedlicher Häufigkeit in allen Verlaufsmuster vor, so dass rund 16% der Kohortenmitglieder mindestens einmal (d.h. mindestens während eines Monats) eine finanzielle Leistung der Sozialhilfe bezogen haben. Aber nur bei 8% der Kohortenmitglieder ist das Verlaufsmuster eindeutig durch die Sozialhilfe geprägt («Sozialhilfe wiederholt», «Sozialhilfe neu» und «Sozialhilfe und Erwerb»). Neben der umfassenden wirtschaftlichen Hilfe in diesen Fällen, zeigt sich hier auch die wichtige Überbrückungsfunktion der Sozialhilfe.

Im Vergleich zum Zwischenbericht (siehe Fussnote 4) zeigen sich in der hier präsentierten Clusterlösung typische Verlaufsmuster, welche alle verhältnismässig klar interpretiert werden und mit bestehendem Wissen zum System ALV-IV-SH und zu sozialen Risiken in Verbindung gebracht werden kann. Bei der Umsetzung der Kohortendefinition im Zwischenbericht waren die Korrekturen, welche den vorliegenden Daten zugrunde liegen, noch nicht umgesetzt. Es zeigt sich, dass 8 der 10 Cluster in beiden Berichten die identischen typischen Verlaufsmuster aufzeigen. Im Zwischenbericht zeigten sich aber auch zwei Cluster (mit den interimistischen Labels «komplex» und «unklar»), die inhaltlich nicht vollständig interpretiert werden konnten. Mit der korrigierten Datenbasis sind diese Verlaufsmuster in der initialen Clusterlösung jedoch nicht mehr präsent. An ihrer Stelle hat sich dafür das Cluster mit IV-Renten in zwei Cluster mit IV-Renten aufgeteilt (einmal mit und einmal ohne Erwerbstätigkeit) und im letzten verbleibenden Cluster («Sozialhilfe neu») ist ein neues, interpretierbares Verlaufsmuster sichtbar geworden. Die Korrekturen bei der Umsetzung der Kohortendefinition haben in der initialen Clusterlösungen auf Basis der Kohorte 2010 zu besseren Resultaten geführt.

4.2.3 Vergleich von Clusterlösungen über die Kohorten

Um eine erste Einschätzung zu erhalten, ob die initiale Clusterlösung ein Artefakt für eine bestimmte Kohorte darstellt oder ob dieselben typischen Verlaufsmuster auch in weiteren Kohorten auftritt, werden für alle weiteren Kohorten (2011 bis 2015) Clusterlösungen mit 10 Clustern berechnet und zwar mit identischen Algorithmen wie bei der initialen Clusterlösung der Kohorte 2010. In der Folge werden die state distribution plots verglichen (siehe Anhänge 7.3 bis 7.17). Daraus lassen sich folgende Erkenntnisse erzielen:

- Acht typische Verlaufsmuster können in allen Kohorten auf Basis visueller Vergleiche wiedergefunden werden. Es handelt sich dabei einerseits um «ALV kurz» (1), «ALV lang» (2), «Zwischenverdienst» (3), «ALV wiederholt» (4), «Sozialhilfe und Erwerb» (7) und «Leavers» (8). Andererseits können auch die beiden Verlaufsmuster «IV-Rente» (5) und «IV-Rente und Erwerb» dazu gezählt werden. Einzig in Kohorte 4 finden sich diese beiden Muster nicht getrennt, sondern gemeinsam in einem einzigen Cluster («Neu4»); Die Grundmuster sind aber auch dort wiederzuerkennen.

- Bei den verbleibenden Clustern aus der initialen Clusterlösungen, «Sozialhilfe wiederholt» und «Sozialhilfe neu», zeigt sich, dass diese beiden Verlaufsmuster in den anderen Kohorten nicht ausdifferenziert werden. In den Kohorten 2011 bis 2015 sind nur zwei Sozialhilfecluster zu erkennen, wobei «Sozialhilfe und Erwerb» mit der initialen Clusterlösung übereinstimmt. Das zweite Sozialhilfecluster in diesen Kohorten («Neu1») vereint Eigenschaften aus den beiden Clustern «Sozialhilfe wiederholt» und «Sozialhilfe neu» der initialen Lösung. Dies zeigt sich visuell anhand der state distribution plots: sowohl ein klarer Anteil an Clustermitglieder mit ALV-Taggeldern und Sozialhilfe ab dem ersten Monat ist ersichtlich, als auch der markante Abfall des Anteils Arbeitslosentaggeld beziehender Personen ab Monat 20 (siehe Anhang 7.14).
- Anstelle eines dritten Sozialhilfeclusters differenzieren sich in den Kohorten 2011 bis 2015 zwei weitere typische Verlaufsmuster aus. Das eine beschreibt Verläufe, bei welchen nach dem anfänglichen Bezug von Arbeitslosentaggelder ein zwischenzeitlicher Rückzug aus dem Erwerbsleben und den Sozialleistungen oder eine temporäre Ausreise aus der Schweiz stattfindet bevor gegen Ende des Beobachtungszeitraums beinahe alle Clustermitglieder ohne weiteren Leistungsbezug erwerbstätig sind («Neu2», Kohorten 2011, 2013, 2014; siehe Anhang 7.15). Das andere beschreibt Verläufe, die nach anfänglichem Taggeldbezug zurück in die Erwerbsarbeit führen, gegen Ende des Beobachtungszeitraums hingegen mit einem Rückzug aus dem Erwerbsleben und den Sozialleistungen bzw. mit einer Ausreise aus der Schweiz einhergehen («Neu3», Kohorten 2012, 2014, 2015; siehe Anhang 7.16).

Zusammenfassend lässt sich festhalten, dass 8 von 10 typischen Verlaufsmustern in allen Kohorten deskriptiv wiedererkannt werden können. Lässt man die Kombination von «Sozialhilfe wiederholt» und «Sozialhilfe neu» in einem neuen Cluster («Neu1») in den Kohorten 2011 bis 2015 ebenfalls als bekanntes Verlaufsmuster gelten, so zeigt sich, dass sich dort jeweils nur ein Cluster klar von der initialen Lösung auf Basis der Kohorte 2010 unterscheidet. In diesem einen abweichende Cluster zeigen sich je nach Kohorte zwei neue Verlaufsmuster («Neu2» und «Neu3»).

Aus dieser ersten Analyse über mehrere Kohorten lassen sich noch keine Schlussfolgerungen darüber ziehen, wann die initiale Clusterlösung aktualisiert werden muss (siehe dazu Abschnitt 4.4). Sie zeichnen jedoch ein erstes deskriptives Bild zur Stabilität des Algorithmus über unterschiedliche Grundgesamtheiten.

4.3 Übertragung der initialen Clusterlösung auf neue Kohorten mittels Prädiktion

Wie in Abschnitt 3.6.2 diskutiert, wurden eine Vielzahl an Prädiktionsmodellen trainiert, um jenes mit der besten Performance zu wählen. Fünf verschiedene Algorithmen (random forest, gbm, svmPoly, knn et avNNet) wurden getestet unter Variation ihrer Hyperparameter und unter Einbezug verschiedener Variablensets («with sociodem» oder «without sociodem») bzw. Datenformate («one-hot encoded», «factor» oder «mixed»).

Aus technischen Gründen wurden nicht alle Parameterkombinationen dieser vier Achsen (Algorithmus, Hyperparameter, Variablensets, Datenformate) umgesetzt. Zudem musste für die Modelle mit den Algorithmen svmPoly und avNNet die Datenmenge reduziert werden (Verwendung von nur 30% der Ursprungsdaten mittels einfach Zufallsstichprobe vor der train/test-Split) und für einige Modelle musste im Vergleich zu den anderen Modellen der Raum der getesteten Hyperparameter eingeschränkt werden («knn», «gbm mixed» und «gbm factor without sociodem»).

Die Modelle wurden auf 80% der Daten trainiert (train-split), die restlichen 20% wurden für die finale Performanceevaluation des gewählten Modells verwendet (test-split). Im Modelltraining wurde eine fünffache cross-validation (eine Repetition) umgesetzt. Die Abbildung 6 zeigt die 20 performantesten Modelle aus der Trainingszeit anhand der mean balanced accuracy auf.

Der Random Forest Algorithmus scheint im Vergleich insgesamt die besten Resultate zu erzielen, gefolgt vom GBM. SVM, KNN und avNNet sind ebenfalls in der Liste der Top 20 vertreten, es sind jedoch keine klaren Vorteile gegenüber dem besten Modell ersichtlich. Grundsätzlich kann man auch feststel-

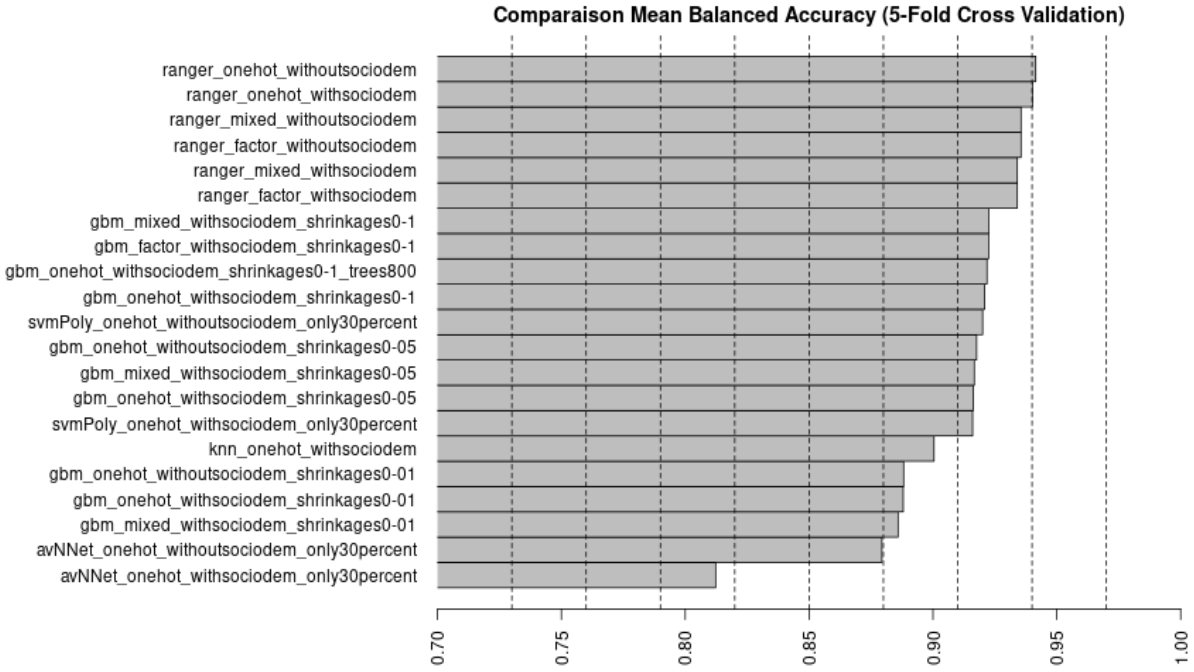
len, dass die Performance eher besser ist, wenn die soziodemografischen Variablen nicht berücksichtig werden. Sie scheinen für die Vorhersage einer Clusterzugehörigkeit im Vergleich zu den Informationen zu den Verläufen (48 Monatsvariablen) kaum eine Erklärungskraft zu haben. Da sie bei der Clustergenerierung ebenfalls nicht als Inputinformation berücksichtigt wurden, gibt es a priori auch keinen starken Grund dies anzunehmen.

Das beste Modell erzielt eine mean balanced accuracy von 94.15% und beruht auf dem Algorithmus Random Forest (ranger) mit einem «one hot»-Encoding und ohne soziodemografischen Variablen. Die Hyperparameter im besten Modell entsprechen folgenden Werten:

- mtry=17
- min.node.size=10
- splitrule="gini"[21].

Die mean balanced accuracy auf dem Testdaten liegt bei 94.06% und damit sehr nahe bei der Trainingsperformance.

Abbildung 6: Performancevergleich der Prädiktionsmodelle (Training) anhand der Mean Balanced Accuracy.



Anmerkung: Ranger=Random Forest
Quelle: BFS - SHIVALV-IK 2010-2014

Eine Analyse der Performance pro Cluster ergibt sehr ähnliche Resultate. In Tabelle 6 ist ersichtlich, dass die höchste Balanced Accuracy im Cluster «IV-Rente» (97.59%) und die tiefste im Cluster «Zwischenverdienst» (88.43%) erzielt wird. Ausser im Cluster «Zwischenverdienst» erreiche alle Cluster einen Wert von über 90%. Aufgrund der teilweise geringen Clustergrössen sind die anderen Kolonnen in der Tabelle mit Vorsicht zu interpretieren.

[21]«Mtry» repräsentiert die Anzahl der Variablen, die bei jedem Knoten eines Entscheidungsbaums berücksichtigt werden. «Min.node.size» bezeichnet die minimale Anzahl Beobachtungen, die für die Endknoten eines Entscheidungsbaums vorgeschrieben werden. Je kleiner die Anzahl desto komplexer der Entscheidungsbaum. « Splitrule » beschreibt die verwendete Regel, mit der an einem Knoten die Daten separiert werden. Hier besagt die Regel, dass jene Separierung gewählt werden soll, die die Gini-Impurity minimiert.

Tabelle 6: Prädiktionsperformance des finalen Modells nach Cluster

	Sensitivity (True Positive Rate)	Specificity (True Negative Rate)	Precision	Balanced Accuracy
ALV Kurzzeit	98.25%	95.97%	96.30%	97.11%
ALV Langzeit	88.22%	98.32%	88.95%	93.27%
Zwischenverdienst	77.24%	99.62%	91.12%	88.43%
ALV Mehrfach	82.28%	99.06%	80.56%	90.67%
IV-Rente	95.20%	99.97%	96.99%	97.59%
IV-Rente und Erwerb	92.15%	99.98%	97.81%	96.06%
Sozialhilfe und Erwerb	80.68%	99.71%	88.03%	90.19%
Sozialhilfe wiederholt	90.96%	99.84%	93.18%	95.40%
Sozialhilfe neu	89.21%	99.56%	86.00%	94.39%
Leavers	95.79%	99.28%	96.14%	97.54%

Quelle: BFS - SHIVALV-IK 2010-2014

Um eine Einschätzung des Variance-Bias-Tradoffs zu erhalten, wurden mit dem finalen Modell Lernkurven berechnet.[22] Dabei wurden das Modell mit einer zunehmenden Anzahl Beobachtungen mehrfach neu trainiert. Die dafür verwendeten Stichproben unterschiedlicher Grösse sind immer eine Teilpopulation der ursprünglichen 80% der Trainingsdaten; aus 20% der Testdaten wurden keine Daten für das Training verwendet. Hingegen wurde die Performance immer an der konstanten Population der 20% der Testdaten berechnet.

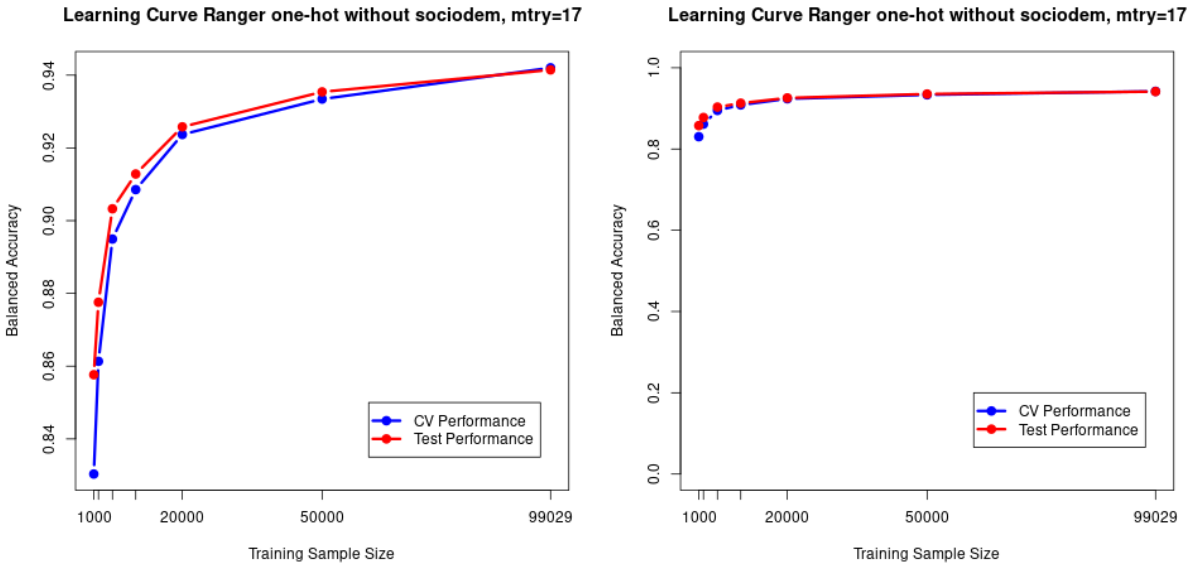
Aus der Abbildung 7 lassen sich verschiede Erkenntnisse ableiten:

- Die Modellperfomance sowohl im Training als auch in der Testphase ist hoch und erreicht mit einer grossen Stichprobe Werte nahe der maximal zu erwartender Performance (siehe weiter oben). Daraus leiten wir ab, dass das Modell genügend gut spezifiziert ist und kein under-fitting vorliegt (low bias).
- Die Lernkurven für die Training- und Testperfomance liegen ab einer Stichprobengrösse von 10'000 Beobachtungen sehr nahe beieinander und reduzieren sich mit zunehmender Stichprobengrösse weiter. Es liegt daher kaum over-fitting vor; Die Qualität der Prädiktionen im finalen Modell besteht damit nicht nur für die Trainingsdaten, sondern können auch auf neuen Daten reproduziert werden (low variance).
- Mit dem vollen Trainingssample erreicht das finale Modell sowohl im Training als auch in der Testphase eine Performance von knapp über 94% (siehe oben). Wenn das Trainingssample halbiert würde (N=50'000), wäre immer noch eine Performance von 93% erreicht worden, und selbst bei einem Trainingssample von N=10'000 läge die Performance noch über 90%. Ab ca. einer Trainingssamplegrösse von N=20'000 nimmt der Nutzen zusätzlicher Datenpunkte für das Training nur noch wenig zu. Dies sind Hinweise darauf, dass mit einer Reduktion des Trainingssamples gewisse technische Limiten behoben werden könnten, ohne auf viel Performance verzichten zu müssen (siehe auch Abschnitt 5 «Key Learnings und Empfehlungen»).

Die «low bias, low variance»-Situation erlaubt es, das finale Modell mit Zuversicht auf neue Kohorten anzuwenden, um so die initiale Clusterlösung auf eine neue Population zu übertragen.

[22] Siehe auch «Bias-Variance-Trade-Off»: https://www.rootstrap.com/blog/gentle-introduction-to-the-bias-variance-trade-off-in-machine-learning

Abbildung 7: Learning curves des finalen Modells (links: Zoom y-Achse, rechts: vollständige y-Achse)

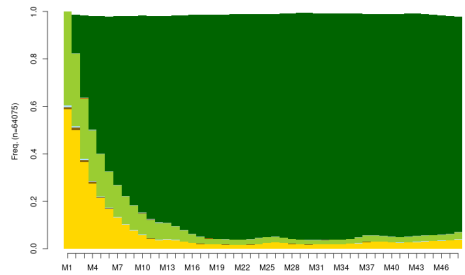
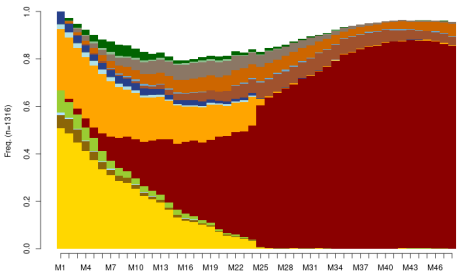
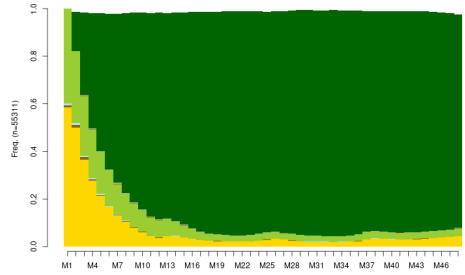
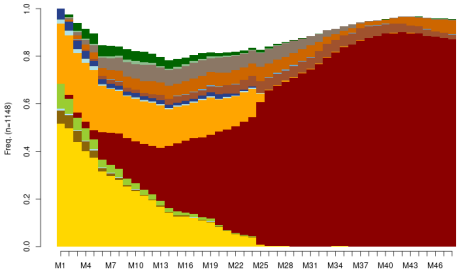
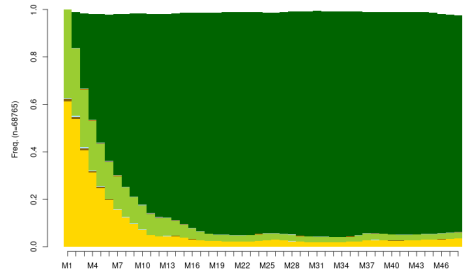
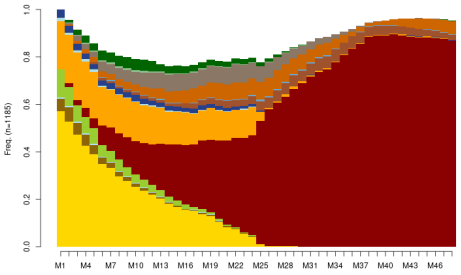


Anmerkung: Die X-Achse repräsentiert die Stichprobengrösse für das Trainingssample. Die Y-Achse repräsentiert die Modellperformance in Form der Mean Balanced Accuracy. Die blaue Kurve beschreibt Trainingsperformance (inkl. Cross-Validation). Obwohl es sich immer um eine Teilpopulation der Gesamtheit aller Trainingsdaten handelt, ändert sich die Reihenfolge der Beobachtungen in jeder Stichprobe. Dies kann das Training bei der Cross-Validation und somit die Trainings- und Testperformance des Modells erheblich beeinflussen und erklärt auch, weshalb beim Berechnen der Lernkurven mit dem vollen Testsample (N=99'029) leicht andere Performance werden erreicht werden als bei der Wahl des finalen Modells (siehe oben). Die rote Kurve beschreibt die Testperformance, welche immer auf Basis der konstanten Testdaten berechnet wird (20% test split).
Quelle: BFS - SHIVALV-IK 2010-2014

Das finale Prädiktionsmodell dient der Klassifizierung der Verläufe von Personen aus neuen Kohorten (2011 bis 2015) in die zehn Cluster der initialen Clusterlösung. Wenn auf Basis dieser Prädiktionen die state distribution plots pro Cluster erstellt werden, können diese in einer visuellen Analyse mit den entsprechenden state distribution plots aus der initialen Clusterlösung verglichen werden.

Die Tabelle 7 zeigt exemplarisch die State Distribution Plots zwei Cluster aus dem initialen Clustermodell der Kohorte 2010, auf dem das Prädiktionsmodell trainiert wurde, sowie die State Distribution Plots derselben Cluster, die mit dem Prädiktionsmodell in den Kohorten 2011 und 2015 vorhergesagt wurden. Man stellt trotz kleinen Unterschieden eine sehr hohe visuelle Übereinstimmung der Grafiken fest. Das Prädiktionsmodell war daher in der Lage die individuellen Verläufe den Clustern so zuzuteilen, dass die initialen visuellen Verlaufsmuster aus der Kohorte 2010 reproduziert werden konnten. Gleichzeitig lässt auch sagen, dass die Kohorten 2010, 2011 und 2015 ähnlich zusammengesetzt waren. Entsprechende Resultate erhält man auch die acht weiteren Cluster als auch für die weiteren Jahre (siehe Anhang 7.4 bis 7.17).

Tabelle 7: State distribution plots für die Cluster «ALV-Kurzzeit» und «IV-Rente» für Kohorte 2010, 2011 und 2015

Cluster	ALV Kurzzeit	IV-Rente
2010 (initiale Clusterlösung)		
2011 (Prädiktion)		
2015 (Prädiktion)		

Quelle: BFS - SHIVALV-IK 2010-2019

Neben der Clusterübereinstimmung in initialer Lösung und Prädiktion interessiert auch die Entwicklung der Clustergrössen mit jeder weiteren Prädiktion (siehe Tabelle 8). Obwohl die Kohortengrösse sich über die Zeit verändert hat, ist die relative Häufigkeitsverteilung der Cluster innerhalb einer Kohorte konstant geblieben.

- Das Cluster «ALV Kurzzeit» umfasst in allen Jahren rund 51% der Kohortenmitglieder mit einer leichten Reduktion von 52% auf 50% über die Zeit
- Das Cluster «ALV Langzeit» hat zwischen 2010 und 2015 leicht zugenommen mit einem Anteil von rund 13% zu Beginn und 16% am Schluss der betrachteten Zeitperiode.
- Das Cluster «Leavers», welches jeweils rund 16% der Kohortenmitglieder repräsentiert, wuchs auf 17% im 2015 mit einem Höchststand von 18% im Jahr 2014.
- Noch weniger Bewegungen gab es in den restlichen Clustern.

Diese Beobachtungen sind möglicherweise auf einen Konjunkturwandel zwischen 2010 und 2015 zurückzuführen, der zu mehr Arbeitslosigkeit geführt hat. So gab es 2015 etwas weniger Kurzzeitarbeitslosigkeit und etwas mehr Langzeitarbeitslosigkeit.

Tabelle 8: Clustergrössen nach Kohorte

Cluster	K2010 (Original) N	%	K2011 (Prädiktion) N	%	K2012 (Prädiktion) N	%	K2013 (Prädiktion) N	%	K2014 (Prädiktion) N	%	K2015 (Prädiktion) N	%
Total	123'786	100%	107'733	100%	120'452	100%	127'919	100%	127'875	100%	138'043	100%
ALV Kurzzeit	64'075	52%	55'311	51%	61'736	51%	65'058	51%	64'063	50%	68'765	50%
ALV Langzeit	16'607	13%	14'586	14%	16'889	14%	18'042	14%	18'473	14%	21'584	16%
Zwischenverdienst	5'809	5%	4'411	4%	5'045	4%	5'348	4%	5'439	4%	5'965	4%
ALV Mehrfach	5'614	5%	5'087	5%	5'948	5%	6'559	5%	5'891	5%	5'843	4%
IV-Rente	1'316	1%	1'148	1%	1'087	1%	1'065	1%	1'092	1%	1'185	1%
IV-Rente und Erwerb	1'213	1%	920	1%	944	1%	950	1%	927	1%	978	1%
Sozialhilfe und Erwerb	3'387	3%	2'802	3%	2'903	2%	2'774	2%	2'915	2%	3'022	2%
Sozialhilfe wiederholt	2'844	2%	2'283	2%	2'302	2%	2'376	2%	2'480	2%	2'388	2%
Sozialhilfe neu	3'489	3%	3'431	3%	3'775	3%	4'205	3%	4'200	3%	4'255	3%
Leavers	19'432	16%	17'754	16%	19'823	16%	21'542	17%	22'395	18%	24'058	17%

Quelle: BFS - SHIVALV-IK 2010-2019

Mit Verlaufsindikatoren können Unterschiede zwischen der initialen Clusterlösung einerseits und deren Prädiktion in späteren Kohorten andererseits mit einem quantitativeren Ansatz analysiert werden. Tabelle 9 und Tabelle 10 zeigen die Verlaufsindikatoren für die Kohorten 2010 und 2015 auf (Die Verlaufsindikatoren für alle Kohorten finden sich im Anhang). Folgende Unterschiede zwischen den beiden Zeitpunkten können beobachtet werden:

- Für die Gesamtkohorte beobachtet man eine Zunahme der Bezugsdauer von Arbeitslosentaggelder (Mittelwert steigt von 10.6 Monaten in der Kohorte 2010 auf 11.3 Monate im Jahr 2015). Im selben Zeitraum nimmt auch die Dauer der ersten ALV-Bezugsperiode von 7.1 auf 7.7 Monate zu. Die mittlere Anzahl Bezugsperioden hingegen bleibt hingegen konstant.
- Dieselbe Entwicklung zeigt sich ebenfalls in den Clustern «ALV Kurzzeit», «ALV Langzeit» und «ALV Mehrfach» sowie in den beiden Clustern «Sozialhilfe neu» und «Leavers».
- Im Cluster «IV-Rente» ist hingegen eher ein leichter Rückgang in der Bezugsdauer von Arbeitslosentaggeldern festzustellen, ohne dass jedoch die Bezugsdauer von IV-Renten zunimmt. Dasselbe gilt für das Cluster «IV-Rente und Erwerb»
- Im Cluster «Sozialhilfe und Erwerb» scheint sich die mittlere Erwerbsdauer erhöht zu haben, Einerseits hat sich die Zeit in einem Zwischenverdienst (ALV-Taggeld und Erwerbsarbeit kombiniert) leicht von 6.5 auf 6.8 Monaten erhöht, andererseits hat die Anzahl Monate, in welchen sowohl Sozialhilfe bezogen wurde als auch Erwerbstätigkeit vorliegt, von 18 auf 20 Monate zugenommen. Umgekehrt verhält es sich im Cluster «Sozialhilfe wiederholt», bei dem ein Rückgang bei beiden Indikatoren sowie bei der Anzahl Monate mit Arbeitslosentaggeld festzustellen ist.
- Neben diesen leichten Verschiebungen sind die Verlaufsindikatoren im betrachteten Zeitraum relativ stabil geblieben. Auch in dieser Hinsicht liefert das Prädiktionsmodell valide Resultate. Zudem ist die beobachtete Stabilität wohl auch ein Hinweis darauf, dass die Verläufe zwischen den Kohorten keine grundlegenden Veränderungen erfahren haben.

Tabelle 9: Verlaufsindikatoren Kohorte 2010, initiale Clusterlösung

Cluster ➔ / Verlaufsindikatoren ⬇	ALV Kurzzeit	ALV Langzeit	Zwischenverdienst	ALV Mehrfach	IV-Rente	IV-Rente und Erwerb	Sozialhilfe und Erwerb	Sozialhilfe wiederholt	Sozialhilfe neu	Leavers	Kohorte
Anzahl Monate mit ALV	6.5	12.8	24.0	22.3	12.0	12.0	14.9	11.3	16.9	12.4	10.6
Dauer der ersten ALV-Bezugsperiode (Monate)	4.3	9.7	13.7	7.4	9.9	8.4	9.9	8.8	13.2	9.9	7.1
Anzahl Monate ALV und Erwerbsarbeit kombiniert	3.0	3.1	19.4	5.0	1.4	4.7	6.5	1.8	2.2	2.1	3.8
Anzahl Bezugsperioden ALV	1.6	1.6	2.7	2.8	1.3	1.6	1.9	1.5	1.6	1.4	1.7
Anzahl Monate mit Erwerbsarbeit	44	32	43	26	4	36	33	7	11	10	34
Anteil Personen mit mindestens einer SH-Bezugsperiode	5%	10%	10%	26%	30%	16%	100%	100%	100%	12%	16%
Anzahl Monate mit SH	0	0	1	2	5	2	28	37	20	1	3
Anzahl Monate SH und Erwerbsarbeit kombiniert	0	0	0	1	0	1	18	4	4	0	1
Anteil Personen mit mindestens einer IV-Bezugsperiode	0%	1%	0%	0%	100%	100%	1%	7%	2%	1%	3%
Anzahl Monate mit IV	0	0	0	0	34	38	0	1	0	0	1
Anzahl Monate IV und Erwerbsarbeit kombiniert	0	0	0	0	3	31	0	0	0	0	0
Anzahl Monate ohne Erwerbsarbeit und Sozialleistungen	1	6	1	3	5	1	2	3	7	28	6

Quelle: BFS - SHIVALV-IK 2010-2014

Tabelle 10: Verlaufsindikatoren Kohorte 2015, Prädiktion

Cluster ➔ / Verlaufsindikatoren ⬇	ALV Kurzzeit	ALV Langzeit	Zwischenverdienst	ALV Mehrfach	IV-Rente	IV-Rente und Erwerb	Sozialhilfe und Erwerb	Sozialhilfe wiederholt	Sozialhilfe neu	Leavers	Kohorte
Anzahl Monate mit ALV	7.0	14.1	24.2	23.1	11.6	11.8	14.9	10.3	17.3	13.2	11.3
Dauer der ersten ALV-Bezugsperiode (Monate)	4.5	10.8	15.0	8.2	9.5	8.3	10.2	8.2	13.7	10.6	7.7
Anzahl Monate ALV und Erwerbsarbeit kombiniert	3.2	3.2	20.6	4.9	1.4	4.1	6.8	1.4	2.0	2.1	3.8
Anzahl Bezugsperioden ALV	1.7	1.7	2.5	2.8	1.3	1.7	1.9	1.4	1.6	1.5	1.7
Anzahl Monate mit Erwerbsarbeit	43	32	44	25	4	36	34	6	11	9	33
Anteil Personen mit mindestens einer SH-Bezugsperiode	4%	9%	8%	20%	29%	14%	100%	100%	100%	10%	14%
Anzahl Monate mit SH	0	0	0	2	4	2	29	38	19	1	2
Anzahl Monate SH und Erwerbsarbeit kombiniert	0	0	0	1	0	1	20	3	3	0	1
Anteil Personen mit mindestens einer IV-Bezugsperiode	0%	0%	1%	1%	100%	100%	1%	6%	2%	1%	2%
Anzahl Monate mit IV	0	0	0	0	31	38	0	1	0	0	1
Anzahl Monate IV und Erwerbsarbeit kombiniert	0	0	0	0	3	30	0	0	0	0	0
Anzahl Monate ohne Erwerbsarbeit und Sozialleistungen	1	5	0	4	7	2	1	3	6	27	6

Quelle: BFS - SHIVALV-IK 2015-2019

4.4 Analysen zur Aktualisierungsnotwendigkeit der initialen Clusterlösung

Die initiale Clusterlösung wurde, wie schon in vorherigen Abschnitten beschrieben, mit einem Prädiktionsmodell auf die Kohorten K2011 bis K2015 übertragen. Um die Frage zu beantworten wann die initiale Clusterlösung aktualisiert werden soll, wird diese Lösung, ihre Prädiktionen in weiteren Kohorten und neue Clusterlösungen für diese Kohorten untersucht werden (siehe auch Abschnitt 3.7).

4.4.1 Deskriptive Evaluation der Ähnlichkeit zwischen Referenz und Prädiktion

Visueller Vergleich mithilfe der state distribution plots

Wie im Absatz 4.3 bereits analysiert, zeigen die state distribution plots der Referenz im Vergleich zu den Prädiktionen keine nennenswerten Unterschiede, siehe Tabelle 7 für einen Vergleich zwischen Cluster *ALV Kurzzeit* und *IV Rente* der Referenz und deren Prädiktionen auf die Kohorten K2011 und K2015 (und auch Anhänge 7.4 bis 7.13).

Vergleich der Ausprägungen der Verlaufsindikatoren

Wie im Absatz 4.3 bereits analysiert, verändern sich die Verlaufsindikatoren in den Prädiktionen im Vergleich zur Referenz nur sehr leicht. Insbesondere hat die Dauer des Bezugs von Arbeitslosentaggelder in mehreren Clustern zugenommen, das gilt auch für die ganze Kohorte. Ausnahme sind die beiden Cluster mit IV-Rente, wo die Dauer des Taggeldbezugs eher abgenommen hat. Zudem haben die Unterschiede der Dauer der Erwerbstätigkeit im Cluster «Sozialhilfe und Erwerb» tendenziell zugenommen, während sie bei «Sozialhilfe wiederholt» eher abnimmt. Grundsätzlich sind die Ausprägungen der Verlaufsindikatoren für die einzelnen Cluster in den Prädiktionen relativ stabil geblieben und die Ausprägungsmuster sind in den Prädiktionen weiterhin deutlich erkennbar.

Vergleich der Clustergrössen

Wie im Absatz 4.3 bereits analysiert, bleibt die Häufigkeitsverteilung der Cluster im Vergleich zwischen der Referenz und den Prädiktionen weitgehend stabil. Sehr kleine Verschiebungen (maximal +/- 3 Prozentpunkte) sind bei den Clustern «ALV Kurzzeit» (grösstes Cluster), «ALV Langzeit» und «Leavers» zu finden. Die Stabilität der relativen Clustergrössen ist trotz sich klar verändernder Kohortengrösse zu beobachten.

Vergleich der Varianzanteile und weiterer interner Masse

In diesem Absatz werden die initiale Clusterlösung und ihre Prädiktionen in den späteren Kohorten mit den internen Massen untersucht, die im Abschnitt 3.7.1 beschrieben wurden. Es werden die Anzahl der Verläufe bzw. Personen pro Cluster, die mittleren und maximalen Distanzen innerhalb jedes Clusters, der Varianzanteil innerhalb der Cluster im Vergleich zur Gesamtvarianz und der mittlerer Silhouette Koeffizient jedes Clusters und der gesamten Clusterlösung berechnet. Die Kompaktheit der Cluster wird durch die Varianzanteile und die Distanzen ausgedrückt, die Trennbarkeit durch den mittleren Silhouette Koeffizienten. Veränderungen der messbaren Grössen Kompaktheit und Trennbarkeit bieten Argumente in Bezug auf die Aktualisierung der Clusterlösung.

Distanzen und Varianzberechnungen für die Prädiktionen basieren auf den Distanzmatrizen der Kohorten. Aufgrund limitierter Rechenkapazitäten können sie nur auf Stichproben berechnet werden, die maximal 20% der ursprünglichen Grösse der Kohorte erfassen (grössere Stichproben können vom Arbeitsspeicher nicht mehr verarbeitet werden). Da die Kohorte K2015 grösser als die vorherigen ist, werden die interne Masse aus rechentechnischen Gründen auf eine Stichprobe der Grösse 18.5% der Kohorte berechnet. Der Konsistenz und der Vergleichbarkeit halber, werden die internen Masse für die initiale Clusterlösung ebenfalls auf einer einfachen Zufallsstichprobe von 20% der Kohorte K2010 gerechnet. Basierend auf dem Stichprobenfehler gibt es eine erwartete Fluktuation bei den Distanzen und den Varianzanteilen (nicht ausgewiesen).

Tabelle 11 fasst einige internen Masse für die Referenz (initiale Clusterlösung) und deren Prädiktion in den Kohorten 2011 und 2015 zusammen. In den Zeilen «Stichprobe N» und «Stichprobe %» werden die Anzahl der Verläufe bzw. Personen pro Cluster und die anteilige Clustergrösse angegeben. Anschliessend werden die mittleren und maximalen Distanzen innerhalb jedes Clusters berechnet. Der «Varianzanteil %» bezeichnet den Anteil der Varianz innerhalb der Cluster zur Gesamtvarianz. In der letzten Zeile wird der mittlerer Silhouette Koeffizient angegeben.

Tabelle 11: Interne Masse für die initiale Clusterlösung (Kohorte K2010) sowie deren Prädiktion in den Kohorten 2011 und 2015

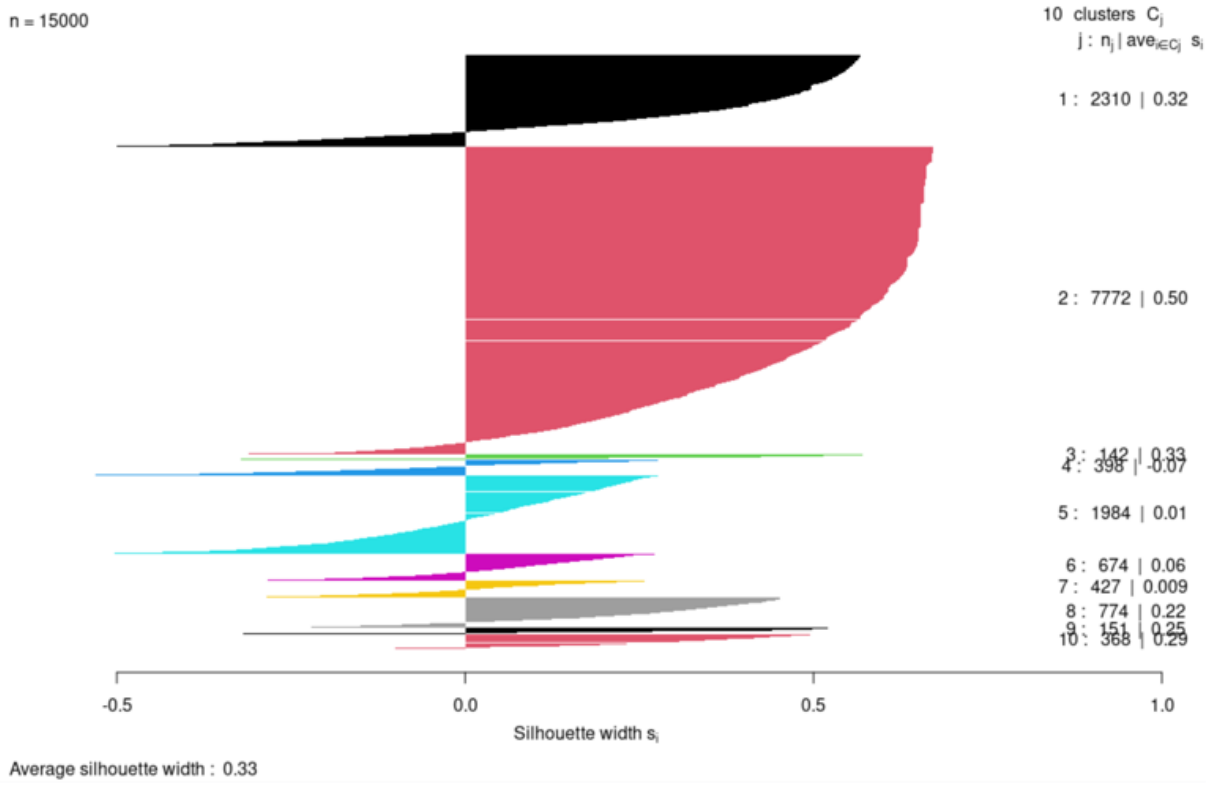
	Kohorte	Clusterlösung insgesamt	1 ALV Kurzzeit	2 ALV Langzeit	3 Zwischenverdienst	4 ALV Mehrfach	5 IV-Rente	6 IV-Rente und Erwerb	7 Sozialhilfe und Erwerb	8 Sozialhilfe wiederholt	9 Sozialhilfe neu	10 Leavers
Stichprobe N	Referenz K2010	24'757	3'804	12'828	247	665	3'329	1'119	693	1'218	264	590
	Prädiktion K2011	21'547	3'530	11'045	184	555	2'964	1'027	643	894	236	469
	Prädiktion K2015	25'538	4'587	12'643	159	549	4'010	1'046	795	1'097	219	433
Stichprobe %	Referenz K2010	100.0	15.4	51.8	1.0	2.7	13.5	4.5	2.8	4.9	1.1	2.4
	Prädiktion K2011	100.0	16.4	51.3	0.9	2.6	13.8	4.8	3.0	4.2	1.1	2.2
	Prädiktion K2015	100.0	18.0	49.5	0.6	2.2	15.7	4.1	3.1	4.3	0.9	1.7
Mittlere Distanz	Referenz K2010	46.1	36.2	15.8	46.6	57.0	29.8	38.2	44.8	29.8	51.5	37.3
	Prädiktion K2011	36.6	36.7	16.5	44.4	55.3	29.7	38.5	43.6	26.0	47.6	36.7
	Prädiktion K2015	46.0	36.0	16.5	46.4	55.1	28.2	36.7	42.5	25.5	47.4	36.6
Maximale Distanz	Referenz K2010	96.0	94.4	77.7	94.4	94.9	80.2	84.1	89.9	91.1	95.7	80.8
	Prädiktion K2011	94.5	94.5	80.8	95.4	94.6	82.8	90.7	92.8	87.7	94.7	82.5
	Prädiktion K2015	96.0	94.0	65.9	92.8	94.8	96.0	92.5	91.8	89.5	95.4	85.7
Varianzanteil %	Referenz K2010	5.55	1.28	3.22	0.01	0.09	0.65	0.11	0.06	0.10	0.01	0.03
	Prädiktion K2011	5.89	1.46	3.42	0.01	0.07	0.66	0.12	0.06	0.05	0.01	0.03
	Prädiktion K2015	6.11	1.76	3.24	0.00	0.05	0.81	0.09	0.07	0.06	0.01	0.02
Silhouette-Koeffizient	Referenz K2010	0.33	0.32	0.50	0.33	-0.07	0.01	0.06	0.01	0.22	0.25	0.29
	Prädiktion K2011	0.34	0.31	0.48	0.34	-0.01	0.05	0.08	0.06	0.36	0.25	0.29
	Prädiktion K2015	0.32	0.30	0.47	0.30	-0.01	0.09	0.07	0.05	0.31	0.22	0.26

Anmerkung: Interne Masse sind berechnet auf einer einfachen Zufallsstichprobe von 20% (bzw. 18.5% bei K2015)
Quelle: BFS - SHIVALV-IK 2010-2019

Am Anfang des Absatzes haben wir die Hypothese aufgestellt, dass die Varianzanteile der Cluster in den Prädiktionen mit der Zeit grösser werden. Beim Vergleich in Tabelle 11 stellen wir fest, dass dies tatsächlich generell leicht der Fall ist. Der Anstieg ist aber nicht ausgeprägt genug, um eine Aktualisierung der initialen Clusterlösung zu rechtfertigen. Zudem könnten nicht quantifizierte Effekte der Stichprobengrösse und der analysierten Stichprobe eine Rolle spielen.

Die mittleren Distanzen innerhalb der Cluster der Stichproben sind für die Referenz und für die Prädiktionen in K2011 und K2015 ähnlich. Das gleiche gilt für die maximalen Distanzen innerhalb der Cluster. Die Silhouette-Koeffizienten, für die einfachen Stichproben berechnet, sowohl für die Clusterlösung insgesamt aber auch pro Cluster, sind sehr stabil. Je näher die Koeffizienten bei 1 sind, desto besser die Trennbarkeit der Cluster. Der Silhouette-Plot gibt eine noch detailliertere Sicht, wie gut die Verläufe den Clustern zugeordnet werden, da der Silhouette Koeffizient jedes einzelnen Verlaufs graphisch dargestellt wird. Abbildung 8 zeigt den Silhouette-Plot für eine einfache Zufallsstichprobe von 15'000 Verläufen der initialen Clusterlösung der Kohorte K2010. Dies ist die maximale Anzahl Elemente die rechentechnisch verarbeitet werden konnte. Im Silhouette Plot werden die Koeffizienten aller Verläufen der Clusterlösung in absteigender Reihenfolge, pro Cluster gruppiert und farblich kenngezeichnet dargestellt. Die Breite der Gruppe der Koeffizienten pro Cluster spiegelt die Clustergrösse wider. Die Ergebnisse der Silhouette-Plots waren über verschiedene Stichprobengrössen hinweg stabil.

Abbildung 8: Silhouette-Plot der initialen Clusterlösung (K2010).



Quelle: BFS - SHIVALV-IK 2010-2014

Lesehilfe: Die Silhouette Koeffizienten für die einzelne Verläufe werden pro Cluster gruppiert und farblich kenngezeichnet dargestellt. Die Koeffizienten nehmen Werte zwischen -1 und 1. Je besser ein Verlauf klassifiziert ist, desto näher zu Eins ist sein Silhouette Koeffizient. Ein Koeffizient nahe Null, deutet auf einen Verlauf hin der dem nächstliegenden Cluster angehören könnte. Ein negativer Silhouette Koeffizient bedeutet, dass der Verlauf dem nächstliegenden Cluster zugeordnet werden sollte.

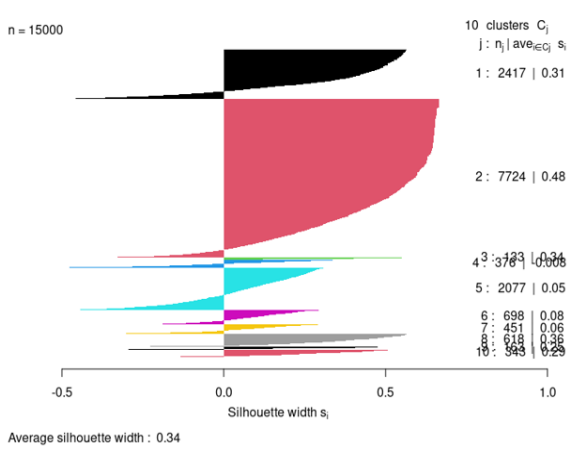
Beim Silhouette-Plot für die Kohorte K2010 ist auffällig, dass die Cluster nicht gut voneinander getrennt werden können. Es scheint, dass in allen Clustern Verläufe vorkommen, die eher zu anderen Clustern gehören. Das ist auch der Fall für eine initiale Clusterlösung mit weniger Clustern (siehe Anhang 7.21.1). Es kommt deswegen die Vermutung auf, dass Verläufe mit zueinander grosser Distanz vom Algorithmus eher zufällig einem Cluster zugewiesen werden. Von diesen Verläufen, scheint es viele zu geben, siehe Abbildung 8. Ein Distanzmass für Verläufe, das fachlich besser motiviert ist und das zwischen Zuständen und deren zeitlichen Abständen besser differenziert, würde eventuell Abhilfe schaffen.

Diese Tatsache wiederholt sich für die Prädiktionen der Referenz, wie die Silhouette-Plots für die Kohorten K2011 und K2015 veranschaulichen (siehe Abbildung 9). Es ist offensichtlich, dass die Prädiktionen der Referenz sehr ähnlich sind. Die mittleren Silhouette-Koeffizienten der Prädiktionen sind ebenfalls sehr ähnlich zur Referenz, siehe auch Tabelle 11. Obwohl die Evolution der mittleren Silhouette-Koeffizienten sehr stabil ist, ist bemerkenswert, dass einige Cluster (sowohl Referenz als auch Prädiktionen) Silhouette-Koeffizienten nahe Null haben. Dies deutet darauf hin, dass viele Verläufe in diesen Clustern auch anderen Clustern angehören können (siehe Definition des Silhouette-Koeffizienten im Abschnitt 3.7.1).

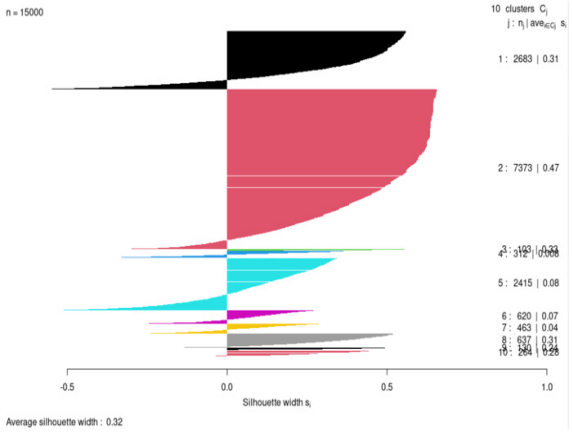
Im Anhang 7.21.1 finden sich Abbildungen, die eine detailliertere Vorstellung über die Verteilung der Distanzen in der gesamten Kohorte K2010 und für einige ausgewählte Cluster geben. 25% der Distanzen innerhalb der Kohorte 2010 sind zwischen 70 und der maximalen Distanz von 96. Die entsprechenden Verläufe scheinen sich auf mehrere Cluster zu verteilen. Das grösste Cluster der initialen Cluster, ist das mit den kleinsten Distanzen.

Abbildung 9: Silhouette-Plots für die Prädiktionen der Referenz auf die Kohorten K2011 und K2015.

Kohorte 2011



Kohorte 2015

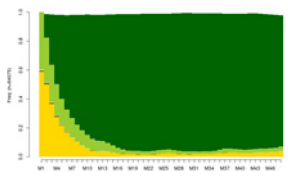
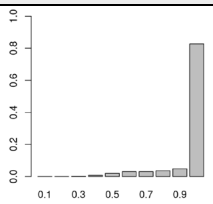
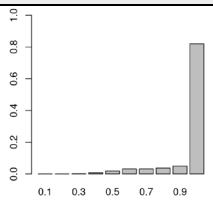
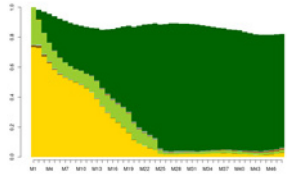
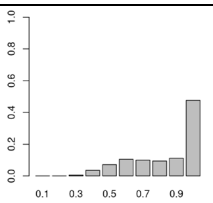
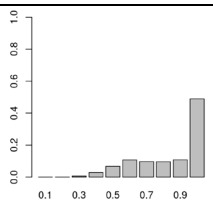
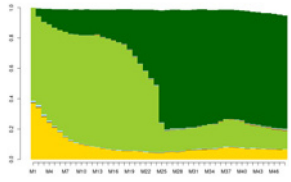
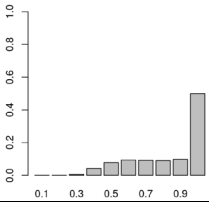
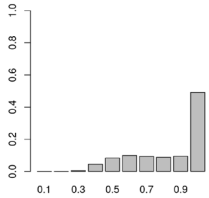


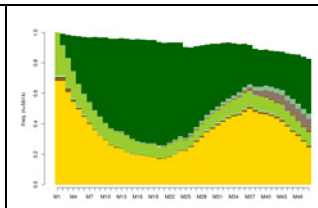
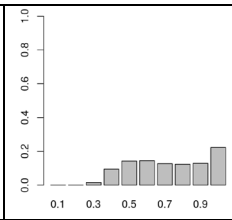
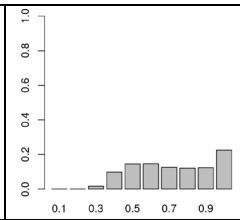
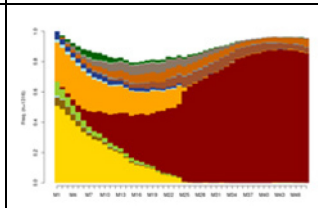
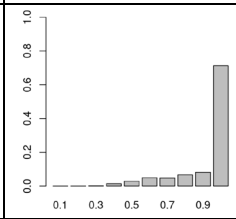
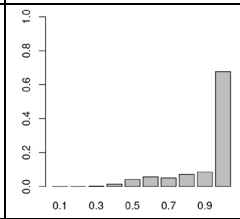
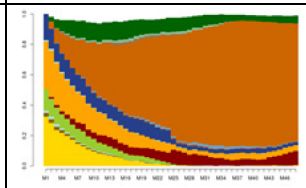
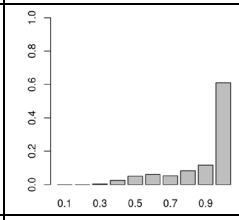
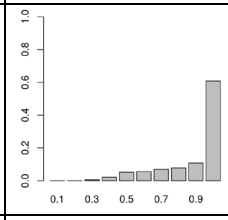
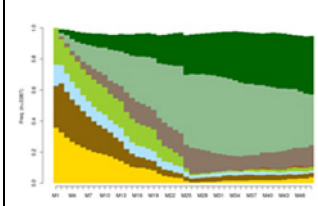
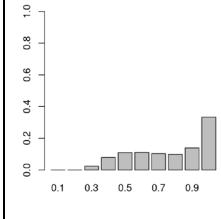
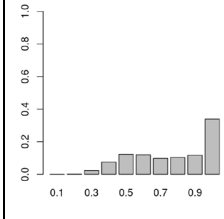
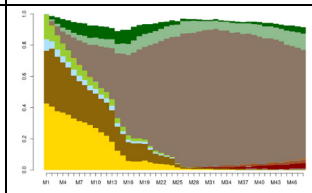
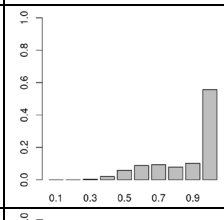
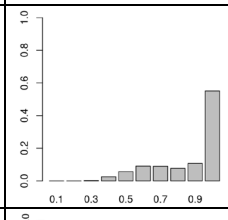
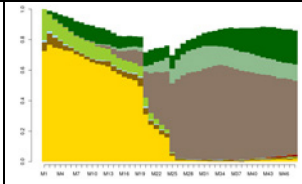
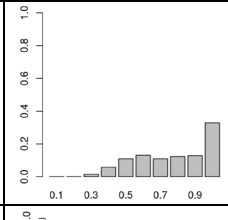
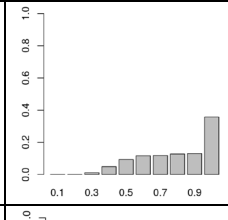
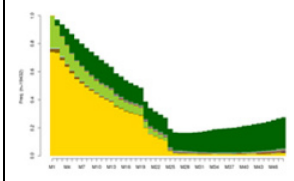
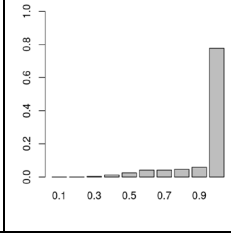
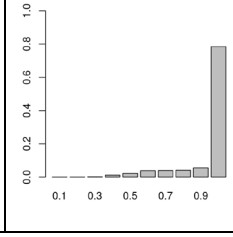
Quelle: BFS - SHIVALV-IK 2011-2019

4.4.2 Post-hoc Evaluation der Zuverlässigkeit der Prädiktion

In Tabelle 12 werden die state distribution plots der Referenz und die Verteilungen der Zuordnungsgüte dargestellt, mit welchen die Verläufe der Kohorten K2011 und K2015 den Clustern der Referenz bei der Prädiktion zugewiesen werden. Es werden die relativen Häufigkeiten (Y-Achse) der Zuordnungsgüte (X-Achse) für jedes Referenzcluster dargestellt. Über 80% der Verläufe der Kohorten 2011 und 2015, die dem Cluster «ALV Kurzzeit» zugeordnet werden, weisen eine Zuordnungsgüte von über 90% auf. Der Anteil der Verläufe, die mit einer sehr hohen Zuordnungsgüte von über 90% dem Cluster zugewiesen werden, ist sowohl in der Kohorte 2011 und 2015 z.B. in den Clustern «ALV-Langzeit», «ALV-Mehrfach» und «Sozialhilfe neu» deutlich geringer. Der visuelle Vergleich der Verteilungen zeigt auch auf, wie ähnlich diese für die Kohorten K2011 und K2015 sind.

Tabelle 12: relative Häufigkeitsverteilung (y-Achse) der Zuordnungsgüte (x-Achse) für die Prädiktionen in den Kohorten 2011 und 2015.

Clusterbezeichnung (Label)	**Referenz** State distribution plot	**Prädiktion Kohorte 2011** Verteilung der Zuordnungsgüte	**Prädiktion Kohorte 2015** Verteilung der Zuordnungsgüte
ALV-Kurzzeit (1)			
ALV-Langzeit (2)			
Zwischenverdienst (3)			

ALV Mehrfach (4)			
IV-Rente (5)			
IV-Rente und Erwerb (6)			
Sozialhilfe und Erwerb (7)			
Sozialhilfe Wiederholt (8)			
Sozialhilfe neu (9)			
Leavers (10)			

Quelle: BFS - SHIVALV-IK 2010-2019

In Tabelle 13 werden die Dezilwerte der Zuordnungsgüte für die Prädiktionen in den Kohorten K2011 und K2015 abgebildet. Es ist bemerkenswert wie stabil diese Werte sind. Tabelle 12 beschreibt auf qualitative, visuelle Weise und Tabelle 13 auf deskriptive, quantitative Weise die gleiche Tatsache: die Verläufe in den Kohorten 2011 und 2015 werden den Clustern der Referenz mit sehr ähnlicher Zuordnungsgüte zugewiesen.

Tabelle 13: Dezilwerte der Zuordnungsgüte für die Prädiktionen der Clusterzugehörigkeit in den Kohorten K2011 und K2015

	Kohorte	min	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
ALV Kurzzeit	**K2011**	0.21	0.74	0.94	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	K2015	0.20	0.72	0.93	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ALV Langzeit	**K2011**	0.22	0.49	0.58	0.68	0.79	0.88	0.95	0.99	1.00	1.00	1.00
	K2015	0.21	0.50	0.59	0.69	0.80	0.89	0.96	0.99	1.00	1.00	1.00
Zwischenverdienst	**K2011**	0.22	0.47	0.58	0.69	0.80	0.90	0.97	1.00	1.00	1.00	1.00
	K2015	0.19	0.46	0.57	0.67	0.78	0.89	0.96	1.00	1.00	1.00	1.00
ALV Mehrfach	**K2011**	0.21	0.39	0.46	0.53	0.60	0.68	0.76	0.84	0.92	0.98	1.00
	K2015	0.22	0.39	0.46	0.53	0.60	0.67	0.75	0.84	0.92	0.98	1.00
IV-Rente	**K2011**	0.26	0.62	0.79	0.92	0.98	1.00	1.00	1.00	1.00	1.00	1.00
	K2015	0.25	0.57	0.75	0.87	0.96	1.00	1.00	1.00	1.00	1.00	1.00
IV-Rente und Erwerb	**K2011**	0.24	0.56	0.72	0.82	0.90	0.96	0.99	1.00	1.00	1.00	1.00
	K2015	0.21	0.54	0.69	0.82	0.91	0.96	0.99	1.00	1.00	1.00	1.00
Sozialhilfe und Erwerb	**K2011**	0.19	0.40	0.49	0.58	0.68	0.77	0.85	0.92	0.97	1.00	1.00
	K2015	0.19	0.40	0.48	0.56	0.66	0.76	0.85	0.92	0.97	1.00	1.00
Sozialhilfe wiederholt	**K2011**	0.20	0.52	0.63	0.75	0.87	0.94	0.99	1.00	1.00	1.00	1.00
	K2015	0.25	0.52	0.64	0.75	0.86	0.94	0.99	1.00	1.00	1.00	1.00
Sozialhilfe neu	**K2011**	0.19	0.43	0.51	0.59	0.68	0.77	0.85	0.92	0.98	1.00	1.00
	K2015	0.22	0.45	0.54	0.63	0.71	0.79	0.87	0.94	0.98	1.00	1.00
Leavers	**K2011**	0.18	0.65	0.87	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	K2015	0.20	0.66	0.88	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Quelle: BFS - SHIVALV-IK 2011-2019

4.4.3 Analyse der Korrespondenz zwischen Prädiktion und neuer Clusterlösung in einer bestimmten Kohorte

Eine weitere Perspektive in der Frage, wann die Referenz aktualisiert werden soll, wird durch Vergleiche ihrer Prädiktion in einer neuen Kohorte mit einer neuen Clusterlösung (selber Algorithmus und Parametrisierung) für eben diese Kohorte aufgezeigt. Grosse Abweichungen einer neuen Clusterlösung von der Prädiktion der Referenz, weisen darauf hin, dass die initiale Clusterlösung aktualisiert werden sollte. Im Idealfall stimmen die Cluster einer Prädiktion inhaltlich eins zu eins mit den Clustern einer neuen Clusterlösung für die Kohorte in Frage überein. Es ist jedoch zu erwarten, dass dies in der Regel nicht der Fall ist. Tauchen neue typische Verlaufsmuster auf (d.h. neue Cluster), stellt sich die Frage, wie stark die Entsprechung der neuen Clusterlösung mit der Prädiktion ist.

In diesem Abschnitt werden zuerst Analysen zur Bestimmung der inhaltlichen Übereinstimmung zwischen den einzelnen Clustern in der Prädiktion und in der neuen Clusterlösung präsentiert, mit dem Ziel, eine Zuordnung der neuen Cluster zu jener der Prädiktion vorzuschlagen. Im Anschluss können auf dieser Basis externe Masse (Accuracy, Cohen's Kappa) für die Übereinstimmung zwischen Prädiktion und neuer Clusterlösung berechnet werden.

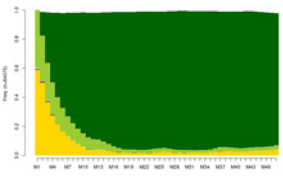
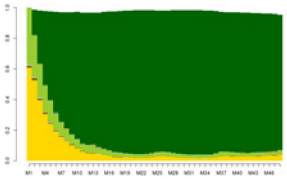
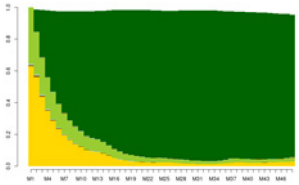
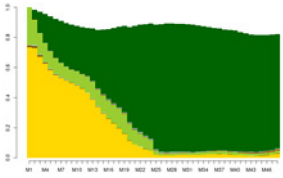
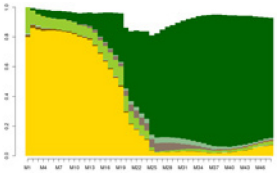
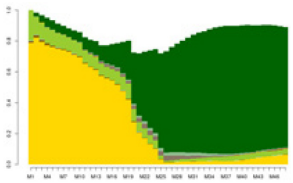
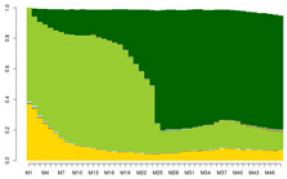
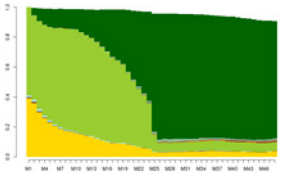
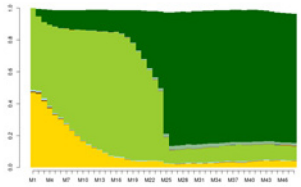
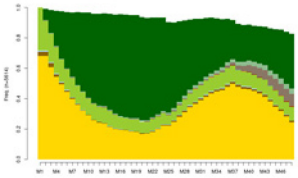
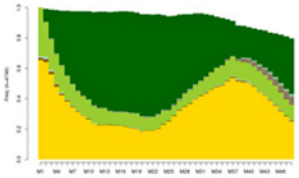
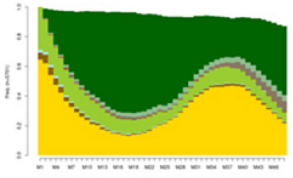
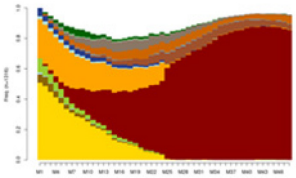
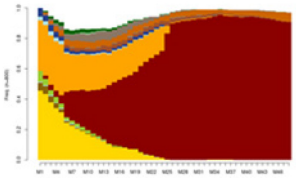
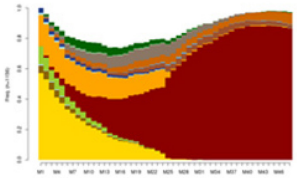
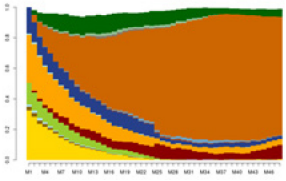
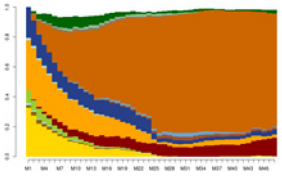
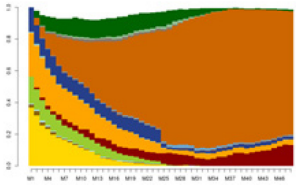
Visueller Vergleich mithilfe der state distribution plots

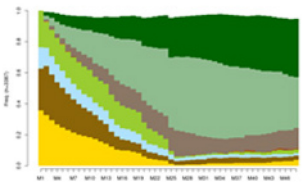
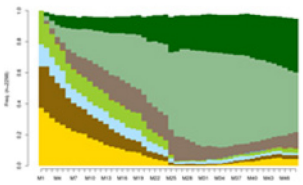
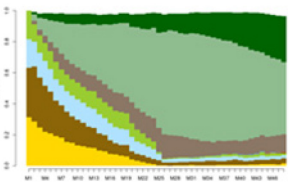
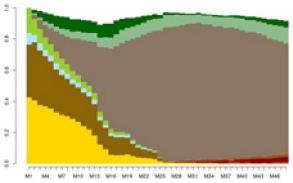
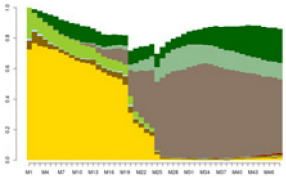
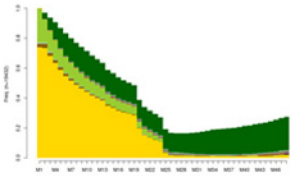
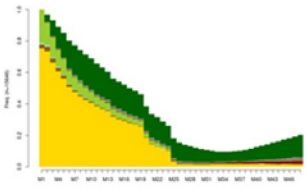
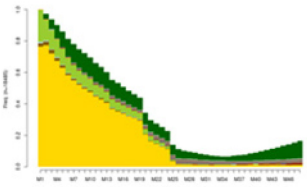
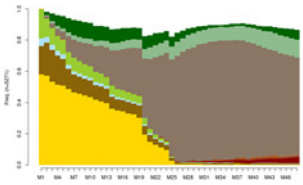
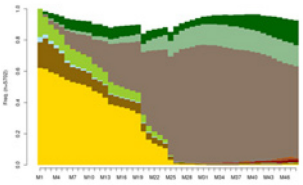
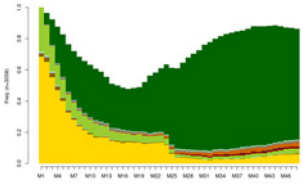
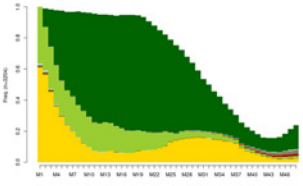
Aufgrund der rein visuellen Analysen, können acht typische Verlaufsmuster aus der Prädiktion[23] auch in den Kohorten K2011 und K2015 identifiziert werden (siehe Tabelle 14; vergleiche auch Abschnitt 4.2.3). Es handelt sich um *ALV kurz* (2), *ALV lang* (5), *Zwischenverdienst* (8), *ALV mehrfach* (6), *Sozialhilfe und Erwerb* (4), *IV-Rente* (9), *IV-Rente und Erwerb* (3), und *Leavers* (1). Die verbleibenden

[23] Als Vereinfachung wird in Tabelle 14 die Referenz und nicht die jeweilige Prädiktion dargestellt. Die hohe Übereinstimmung zwischen Referenz und Prädiktion legitimiert dieses Vorgehen

zwei Clustern aus der Prädiktion *Sozialhilfe wiederholt* (10) und *Sozialhilfe neu* (7) werden in den anderen Kohorten nicht ausdifferenziert. In den Kohorten K2011 bis K2015 sind nur zwei anstatt drei Sozialhilfe-Cluster zu erkennen.

Tabelle 14: State distribution plots für die Referenz und neue Clusterlösungen für die Kohorten K2011 und K2015.

Clusterbezeichnung	**Referenz* (Kohorte K2010)**	**Neue Clusterlösung Kohorte K2011**	**Neue Clusterlösung Kohorte K2015**
ALV-Kurzzeit (1)		Cluster B 	Cluster R 
ALV-Langzeit (2)		Cluster C 	Cluster Q 
Zwischen-Verdienst (3)		Cluster G 	Cluster U 
ALV mehrfach (4)		Cluster F 	Cluster S 
IV-Rente (5)		Cluster E 	Cluster Z 
IV-Rente und Erwerb (6)		Cluster J 	Cluster W 

Sozialhilfe und Erwerb (7)		Cluster D 	Cluster X 
Sozialhilfe Wiederholt (8)		Keine Entsprechung	Keine Entsprechung
Sozialhilfe neu (9)		Keine Entsprechung	Keine Entsprechung
Leavers (10)		Cluster H 	Cluster T 
Neu 1		Cluster I 	Cluster V 
Neu 2		Cluster A 	
Neu 3			Cluster Y 

Anmerkung: Als Vereinfachung wurden hier die Referenz und nicht die jeweiligen Prädiktionen dargestellt. Die hohe Übereinstimmung legitimiert dieses Vorgehen. Die Zuordnung der Cluster der neuen Lösung zu jenen der Referenz beruht hier auf einer rein visuellen Interpretation der state distribution plots.

Quelle: BFS - SHIVALV-IK 2010-2019

Konfusionsmatrix der initialen und neuen Clusterlösungen und Jaccard-Matrix

Neben dem visuellen Vergleich bieten Analysen auf Basis der Konfusionsmatrix vertiefte Analysemöglichkeiten (siehe Tabelle 15). In den Zeilen sind die Cluster der Prädiktion, in den Spalten die Cluster eines neuen Models für diese Kohorte angeordnet. Handelt es sich bei der neuen Clusterlösung um eine identische Lösung wie bei der Prädiktion, müssten alle Verläufe eines Clusters in der neuen Lösung genau einem Cluster in der Prädiktion zugeordnet sein. Dies ist offensichtlich nicht der Fall.

Tabelle 15: Konfusionsmatrix für die Cluster der Prädiktion und einer neuen Clusterlösung, Kohorte K2011

Kohorte 2011			neue Clusterlösung										Total
			A	B	C	D	E	F	G	H	I	J	
Prädiktion initiale Clusterlösung	ALV-Kurzzeit	1	253	50'554	499	55	0	827	3'121	5	2	3	55'319
	ALV-Langzeit	2	1'693	5'064	4'949	56	0	272	1'756	860	6	7	14'663
	Zwischenverdienst	3	6	248	30	7	0	210	3'936	6	0	1	4'444
	ALV mehrfach	4	194	461	409	68	0	3'342	330	152	81	4	5'041
	IV-Rente	5	24	0	0	0	783	0	1	175	102	63	1'148
	IV-Rente und Erwerb	6	62	16	4	1	15	0	30	1	4	780	913
	Sozialhilfe und Erwerb	7	19	259	145	1'864	0	3	190	25	269	10	2'784
	Sozialhilfe wiederholt	8	0	1	8	72	0	0	3	25	2'214	1	2'324
	Sozialhilfe neu	9	15	15	604	146	1	8	37	116	2'421	0	3'363
	Leavers	10	792	62	1'672	29	1	84	632	14'281	172	9	17'734
Total			3'058	56'680	8'320	2'298	800	4'746	10'036	15'646	5'271	878	107'733

Quelle: BFS - SHIVALV-IK 2011-2015

Man kann ebenfalls die relativen Häufigkeitsverteilungen der beiden Clusterlösungen betrachten (für die Kohorte 2011 siehe Tabellen im Anhang 7.21.2, Tabelle A 34 und Tabelle A 35):

- Zeilenprozente zeigen auf, wie sich Verläufe eines Clusters aus der Prädiktion auf die Cluster der neuen Lösung verteilen. Insbesondere das initiale Cluster «ALV – Langzeit» verteilt sich auf mehrere Cluster der neuen Lösung
- Spaltenprozente geben Auskunft, wie sich Verläufe eines Clusters aus der neuen Lösung auf die Cluster der Prädiktion verteilen. Vor allem die neuen Cluster G und I speisen sich aus mehreren Clustern der initialen Lösung.

Eine genauere Einschätzung dieser Übereinstimmungen bzw. Differenzen erlaubt eine Jaccard-Metrik: Beispielsweise berechnet man für ein Cluster α der Prädiktion und ein Cluster β eines neuen Modells das Verhältnis der Überschneidungsmenge in Bezug zur Vereinigungsmenge ($|\alpha \cap \beta|/|\alpha \cup \beta|$). Daraus resultiert als Kennzahl der Anteil der Verläufe, die zwei Cluster aus unterschiedlichen Clusterlösungen gemeinsam haben, gemessen an der Summe aller Verläufen die α und β zugeordnet sind. Diese Grösse ist dem Jaccard-Index angelehnt, der für den Vergleich von ganzen Clusterlösungen verwendet wird. Wenn diese Grösse paarweise für alle Cluster der zwei Lösungen berechnet wird, ergibt sich folgende «Jaccard-Matrix» am Beispiel der Kohorte K2011:

Tabelle 16: Jaccard Matrix für die Cluster der Prädiktion und einer neuen Clusterlösung, Kohorte 2011.

Kohorte 2011			neue Clusterlösung									
			A	B	C	D	E	F	G	H	I	J
Prädiktion initiale Clusterlösung	ALV-Kurzzeit	1	0.4%	82.3%	0.8%	0.1%	0.0%	1.4%	5.0%	0.0%	0.0%	0.0%
	ALV-Langzeit	2	10.6%	7.6%	27.4%	0.3%	0.0%	1.4%	7.7%	2.9%	0.0%	0.0%
	Zwischenverdienst	3	0.1%	0.4%	0.2%	0.1%	0.0%	2.3%	37.3%	0.0%	0.0%	0.0%
	ALV mehrfach	4	2.5%	0.8%	3.2%	0.9%	0.0%	51.9%	2.2%	0.7%	0.8%	0.1%
	IV-Rente	5	0.6%	0.0%	0.0%	0.0%	67.2%	0.0%	0.0%	1.1%	1.6%	3.2%
	IV-Rente und Erwerb	6	1.6%	0.0%	0.0%	0.0%	0.9%	0.0%	0.3%	0.0%	0.1%	77.2%
	Sozialhilfe und Erwerb	7	0.3%	0.4%	1.3%	57.9%	0.0%	0.0%	1.5%	0.1%	3.5%	0.3%
	Sozialhilfe wiederholt	8	0.0%	0.0%	0.1%	1.6%	0.0%	0.0%	0.0%	0.1%	41.1%	0.0%
	Sozialhilfe neu	9	0.2%	0.0%	5.5%	2.6%	0.0%	0.1%	0.3%	0.6%	39.0%	0.0%
	Leavers	10	4.0%	0.1%	6.9%	0.1%	0.0%	0.4%	2.3%	74.8%	0.8%	0.0%

Quelle: BFS - SHIVALV-IK 2011-2015

Aus Tabelle 16 entnehmen wir, dass in der Kohorte 2011 folgende Clusterpaare der Prädiktion und der neuen Clusterlösung mehr als 50% ihrer Verläufe gemeinsam haben. Auch in der visuellen Analyse zeigt sich in diesen Fällen eine gute Übereinstimmung:

- *ALV Kurzzeit* (1) und B (82.3%)
- *ALV Mehrfach* (4) und F (51.9%)
- *IV - Rente* (5) und E (67.2%)
- *IV-Rente und Erwerb* (6) und J (77.2%)
- *Sozialhilfe und Erwerb* (7) und D (57.9%)
- *Leavers* (10) und H (74.8%)

Eine wichtige Erkenntnis ist, dass das grösste Cluster der Referenz, ALV-Kurzzeit, Zeit-invariant zu sein scheint. Für dieses Cluster, das in jeder Kohorte mindestens 40% aller Verläufe beinhaltet, gibt es jeweils eine grosse Übereinstimmung zwischen Prädiktion und neuem Modell. Im Sinne des Jaccard-Index beträgt die Überreinstimmung für die späteren Kohorten: K2011: 82%, K2012: 83%, K2013: 78%, K2014: 86%, K2015: 82%.

Ausgehend von der Prädiktion gibt es einige Cluster, die sich im neuen Modell nicht mehr klar wiederfinden. Dies trifft auf die Cluster ALV *Langzeit* (2), *Zwischenverdienst* (3), *Sozialhilfe wiederholt* (8) und *Sozialhilfe neu* (9) zu. Diesbezüglich können drei Beobachtungen festgehalten werden:

- Die Cluster *Sozialhilfe wiederholt* (8) und *Sozialhilfe neu* (9) der Prädiktion fusionieren im neuen Modell im Cluster I. Dies ist auch gut in den relativen Verteilungen der Konfusionsmatrix (siehe Anhang 7.21.2) und in der visuellen Analyse (siehe Tabelle 14) nachzuvollziehen.
- In der neuen Lösung taucht zudem ein neues Verlaufsmuster bzw. Cluster auf (A). Im neuen Modell werden im Vergleich zur Prädiktion jene Verläufe neu aufgeteilt, die alleine durch Arbeitslosentaggeldbezug oder Leavers geprägt sind. Dies betrifft insbesondere das Cluster *ALV-Langzeit* (2) aus der Prädiktion: Es verteilt sich in der neuen Lösung auf das neue Cluster A sowie auf die Cluster B, C und G. Zudem zeigt sich, dass das Cluster G in der neuen Lösung visuell zwar eine hohe Übereinstimmung mit dem Cluster *Zwischenverdienst* (3) aufweist, es enthält jedoch auch viele Verläufe, die in der Prädiktion den Clustern *ALV-Kurzzeit* (1) und *ALV-Langzeit* (2) zugeordnet sind.
- In der Regel, sind die typischen Verlaufsmuster aus der Prädiktion visuell in der neuen Lösung wiederzufinden (siehe visuelle Analyse in Tabelle 14); die entsprechenden Cluster beruhen aber teilweise auf einer anderen Selektion von Verläufen als in der Prädiktion.

Für die Kohorte 2015 sind ähnliche Beobachtungen zu machen: Die Cluster *ALV Kurzzeit* (1), *IV - Rente* (5), *IV-Rente und Erwerb* (6) und *Leavers* (10) haben in der neuen Clusterlösung ebenfalls eine klare Übereinstimmung mit einem neuen Cluster (sowohl auf Basis der Jaccard-Matrix als

auch visuell; siehe Anhang 7.21.2, Tabelle A 39 und Tabelle 14). Im Vergleich zur Kohorte 2011 gilt das zusätzlich auch für das Cluster *Zwischenverdienst* (3). Die Cluster *Sozialhilfe wiederholt* (8) und *Sozialhilfe neu* (9) fusionieren in der neuen Lösung wie in der Kohorte 2011 in einem neuen Sozialhilfecluster (V), welches zudem Verläufe aus dem Cluster *Sozialhilfe und Erwerb* (7) aufnimmt. In der neuen Lösung wird für das zuletzt genannte Cluster gemäss Jaccard-Matrix keine klare Übereinstimmung mehr gefunden (visuell bleibt das Verlaufsmuster jedoch eindeutig bestehen). Wie in der Kohorten 2011 taucht ein neues Verlaufsmuster auf (Y), das sich insbesondere aus den Clustern *ALV-Langzeit* (2) und *ALV mehrfach* (4) der Prädiktion speist. In diesem Zusammenhang ist auch hier zu beobachten, dass die Verläufe, die vor allem durch Arbeitslosentaggeldbezug und Leavers geprägt sind, neu aufgeteilt werden. Für die Cluster *ALV-Langzeit* (2) und *ALV mehrfach* (4) finden sich in der neuen Clusterlösung daher keine klaren Entsprechungen (kein Jaccard-Index >=50%). Visuell bleiben die typischen Verlaufsmuster aus der Prädiktion jedoch weitgehend bestehen.

Zuordnung inhaltlich übereinstimmender Cluster zwischen initialer und neuer Clusterlösung

Um externe Masse der Übereinstimmung zweier Clusterlösungen zu berechnen (Accuracy, Balanced Accuracy, Cohen's Kappa), gilt die Voraussetzung, dass jedem Cluster aus der einen Lösung exakt ein Cluster aus der anderen Lösung zugeordnet wird; nur so kann der Grad der Übereinstimmungen bestimmt werden. Hier bedeutet dies, dass jedem Cluster aus der Prädiktion jenes Cluster aus dem neuen Modell zugeordnet wird, bei welchem die Übereinstimmung anhand des Jaccard-Indexes am grössten ist.

Falls in der neuen Lösung keine eindeutig neuen Verlaufsmuster auftauchen, bereitet dieses Vorgehen keine Schwierigkeiten. Anderenfalls kann es dazu kommen, dass Cluster einander zugeordnet werden, obwohl der Jaccard-Index eine sehr geringe oder keine Überschneidung nahelegt. Dies führt in der Konsequenz zu niedrigeren Übereinstimmungswerten, was in diesem Fall jedoch zielführend ist.

Auf der Basis der vorangehenden Analysen zeigen sich folgende Zuordnungsergebnisse, die durch die visuellen Analysen gut unterstützt werden (siehe Tabelle 14):

Tabelle 17 Zuordnungsergebnisse für die Cluster in Prädiktion und neuer Clusterlösung für die K2011 und K2015, sortiert nach Jaccard-Index K2011

Cluster der Prädiktion (Basis: initiale Clusterlösung K2010)		**K2011**		**K2015**	
		Zuordnung Cluster neue Lösung zu Cluster Prädiktion	Jaccard-Index	Zuordnung Cluster neue Lösung zu Cluster Prädiktion	Jaccard-Index
ALV Kurzzeit	**1**	B	82.3%	R	82.8%
IV-Rente und Erwerb	**6**	J	77.2%	W	83.4%
Leavers	**10**	H	74.8%	T	69.6%
IV-Rente	**5**	E	67.2%	Z	79.3%
Sozialhilfe und Erwerb	**7**	D	57.9%	X	44.8%
ALV Mehrfach	**4**	F	51.9%	S	45.6%
Sozialhilfe wiederholt	**8**	I	41.1%	Y	0.0%
Zwischenverdienst	**3**	G	37.3%	U	54.8%
ALV Langzeit	**2**	C	27.4%	Q	31.8%
Sozialhilfe neu	**9**	A	0.2%	V	38.5%

Quelle: BFS - SHIVALV-IK 2011-2019

In der Kohorte 2011 fusionieren die Cluster *Sozialhilfe wiederholt* (8) und *Sozialhilfe neu* (9) in der Prädiktion zum Cluster I in der neuen Lösung. Da der Jaccard-Index von *Sozialhilfe wiederholt* (8) mit dem neuen Sozialhilfe-Cluster I leicht höher ist als für *Sozialhilfe neu* (9) (41.1% vs 39%) wird das neue Cluster I ersterem zugeordnet. Hingegen werden die Cluster *Sozialhilfe neu* (9) und das neuen Cluster A einander zugeordnet, obwohl keine gemeinsamen Verläufe bestehen (0.2%); es handelt sich um die letzten beiden verbliebenen Cluster ohne Pendant in der jeweils anderen Clusterlösung.

In der Kohorte 2015 kann dieselbe Fusion wie in Kohorte 2011 beobachtet, nur fällt die Zuordnung umgekehrt aus: Da der Jaccard-Index von *Sozialhilfe neu* (9) mit dem neuen Sozialhilfe-Cluster V

leicht höher ist als für *Sozialhilfe wiederholt* (8) (38.5% vs 36.8%) wird das neue Cluster V ersterem zugeordnet. Hingegen werden die Cluster *Sozialhilfe wiederholt* (8) und das neuen Cluster Y einander zugeordnet, obwohl keine gemeinsamen Verläufe bestehen (0.0%); es handelt sich auch hier um die letzten beiden verbliebenen Cluster ohne Pendant in der jeweils anderen Clusterlösung.

Auf Basis dieser Zuordnungen kann nun die Konfusionsmatrix neu geordnet und die externen Masse zur Übereinstimmung von Prädiktion und neuer Clusterlösung berechnet werden.

Tabelle 18: Externe Masse für den Vergleich zwischen Prädiktion und einer neuen Clusterlösung für die Kohorten K2011 und K2015.

Externe Masse	Kohorte 2011	Kohorte 2015
Accuracy	0.77	0.78
Balanced Accuracy	0.68	0.65
Cohen's Kappa	0.66	0.67

Quelle: BFS - SHIVALV-IK 2011-2019

Grundsätzlich kann nur eine mittlere Übereinstimmung zwischen der initialen und der neuen Clusterlösung festgestellt werden (zur Interpretation siehe nachfolgender Abschnitt). Eine sehr hohe Übereinstimmung wäre bei Werten nahe 1 gegeben. Die Accuracy ist deutlich höher, da sie die unterschiedlichen Clustergrössen nicht berücksichtigt.

Betrachtet man die Accuracy separat pro Cluster der Referenz (auch «True Positive Rate», siehe Anhang 7.21.2,Tabelle A 40 und Tabelle A 41), zeigen sich für die Kohorte 2011 und 2015 für dieselben Cluster Übereinstimmungswerte ähnlicher Grössenordnung.

Von Bedeutung ist vor allem, dass die externen Masse über die Zeit stabil bleiben und nicht abnehmen. Das heisst auch, dass einerseits die auf den jeweiligen Kohorten neu berechneten Clusterlösungen auch über die Zeit ähnliche Resultate zeitigen (visuell kann das z.B. in Tabelle 14 nachvollzogen werden), und dass andererseits die Distanz dieser neuen Modelle zur jeweiligen Prädiktion der Referenzclusterlösung konstant bleiben.

4.4.4 Schlussfolgerungen

In den vorangegangenen Abschnitten beschäftigte uns die Frage: Wann muss die initiale Clusterlösung aktualisiert werden? Dieser Absatz fasst die Erkenntnisse der Abschnitte 4.4.1 bis 4.4.3 zusammen.

Die deskriptive **Evaluation der Ähnlichkeit zwischen Referenz und Prädiktion** in Abschnitt 4.4.1 zeigt, dass Visualisierungen, Verlaufsindikatoren, Häufigkeitsverteilungen und interne Masse der Cluster über die Kohorten stabil bleiben. Die Stabilität der Varianzanteile widerspricht unserer anfänglichen Hypothese, dass die Prädiktionen mit der Zeit an Aussagekraft und die Cluster an Kompaktheit verlieren. Gemäss der deskriptiven Evaluation mithilfe von internen Massen gibt es keine Hinweise, dass die Referenz aktualisiert werden muss.

In Abschnitt 4.4.2 wurden bei der Prädiktion der Clusterzugehörigkeit in Bezug auf die initiale Clusterlösung (Referenz) für Verläufe aus den Kohorten 2011 und 2015 die **Zuordnungsgüte evaluiert**. Auffällig ist, dass diese Verteilungen ähnlich wie die interne Masse im Abschnitt 4.4.1) sehr stabil in der Zeit zu sein scheinen. Verläufe der Kohorten K2011 und K2015 werden zu den Clustern der initialen Clusterlösung ähnlich stark oder schwach zugeordnet. Die Annahme, dass die Zuordnungsgüte aufgrund von Veränderungen in den Grundgesamtheiten abnimmt, bestätigt sich in den betrachteten Kohorten nicht. Auch diese Ergebnisse legen keine Notwendigkeit für eine Aktualisierung der initialen Clusterlösung nahe.

Der **Vergleich der Prädiktion der initialen Clusterlösung mit einer neuen Clusterlösung** (selber Algorithmus) in einer bestimmten Kohorte, siehe Abschnitt 4.4.3, bietet eine andere Perspektive für die Evaluation, ob die Referenz aktualisiert werden soll. Die Kennzahlen zur Übereinstimmung dieser beiden Clusterlösungen (externe Masse; Accuracy, Balanced Accuracy, Cohen's Kappa) wurden

für unterschiedliche Kohorten berechnet und es zeigt sich, dass die externen Masse über die Zeit nicht abnehmen, sondern stabil bleiben. Die neuen Clusterlösungen führen nicht zu grundlegend anderen Verlaufsmuster (siehe auch nächster Absatz). Auch das weist nicht darauf hin, dass die initiale Clusterlösung aktualisiert werden muss.

Es zeigt sich jedoch, dass die Prädiktionen und die neuen Clusterlösungen oft nur begrenzt übereinstimmen (in der Regel findet man für vier von zehn Verlaufsmustern in der Referenz keine genügend gute Übereinstimmung in den neuen Clusterlösungen). Detailanalysen zeigen einerseits, dass oft zwei Cluster aus der Referenz in den neuen Clusterlösungen späterer Kohorten in einem Cluster vereinigt sind; diese kombinierten Cluster sind daher nicht eigentlich neue Verlaufsmuster, sondern beruhen auf bekannten Mustern (siehe Anhang 7.14 und 7.17). Andererseits treten neue Verlaufsmuster (Cluster) auf, die in der Referenz ganz fehlen. Über die Kohorten 2011 bis 2015 konnten insgesamt zwei solcher Verlaufsmuster identifiziert werden (siehe Anhang 7.15 und 7.16), die je nach Kohorte auftreten. Nach den Ergebnissen des Abschnitts C, lässt sich die initiale Clusterlösung (bzw. deren Prädiktion) bei erneutem Laufenlassen des Algorithmus in späteren Kohorten nicht eindeutig reproduzieren.

Grundsätzlich zeigen die Analysen in den drei vorangehenden Abschnitten keine Notwendigkeit auf, die Referenz nach fünf neuen Kohorten zu aktualisieren.

Hingegen liegt der Schluss nahe, dass die beobachteten Auffälligkeiten (Silhouette plots, Vergleich initiale und neue Clusterlösungen) mit einer nicht optimalen Wahl der initialen Clusterlösung (Referenz) zu tun haben. Es ist auffällig, dass die Abgrenzung der Cluster statistisch gesehen Mängel aufweist. Bei der Wahl der initialen Clusterlösung sollten daher verstärkt statistische Merkmale, wie z.B. Silhouette-Koeffizienten berücksichtigt werden (z.B. bei der Wahl der Anzahl Cluster). Dies hilft auch, das Clusterproblem aus statistischer Sicht besser zu verstehen. So gibt es Clustering-Probleme, denen eine suboptimale Abgrenzung zwischen den Clustern inhärent ist; hier können z.B. «fuzzy» Clusteralgorithmen Alternativen bieten. Auch andere Algorithmen (z.B. mit Bootstrappingansatz) oder ein angepasstes Distanzmass könnten zu einer besseren Trennbarkeit zwischen den Clustern führen. Nicht zuletzt könnte ein Pooling von Kohorten als Basis für die Entwicklung der initialen Lösung zu einer besseren Stabilität über die Zeit führen, was hingegen erweiterte Rechenressourcen voraussetzt und die Entdeckung von systembedingten Änderungen in den Verläufen erschweren könnte.

5 Key Learnings und Empfehlungen

Im Folgenden sollen im Sinne eines «institutionellen Lernens» zentrale Erkenntnisse aus dem ML_SoSi-Projekt in generalisierter Form präsentiert werden, um einen Beitrag für das lösungsorientierte Arbeiten in ähnlichen Projekten im BFS zu leisten.

Infrastruktur

1) **Performante Infrastruktur ist notwendig:** Viele statistische Methoden aus dem Bereich der Datenwissenschaften funktionieren über iterative Optimierung (z.B. gradient descent, cross validation, tuning via grid search) und einige statistischen Programme (z.B. R) erfordern die Haltung grosser Datenmengen im Arbeitsspeicher (z.B. Distanzmatrizen beim Clustering). Daneben profitieren viele Algorithmen von grossen Datenmengen und auch die verfügbaren Daten im BFS wachsen stetig. Einige Anwendungen verlangen daher nach einer Recheninfrastruktur, die deutlich über die derzeit verfügbaren Lösungen im BFS hinausgeht. Für die Pilotprojekte, die im Rahmen der Dateninnovationsstrategie des BFS umgesetzt wurden, stand eine besonders performante Analyseumgebung zur Verfügung. Dennoch konnte damit im Projekt ML_SoSi nicht alle Vorgehensschritte wie vorgesehen umgesetzt werden (z.B. direktes hierarchischen Clustering ohne k-means-Zwischenschritt. Diese Erfahrungen mündeten in entsprechende Anforderungen für die Weiterentwicklung der Datenanalyseinfrastruktur des BFS.
2) **Technische Hindernisse algorithmisch lösen:** Eine performante Infrastruktur ist notwendig, sie ist jedoch auch teuer und garantiert nicht in jedem Fall bessere Resultate. Aus diesem Grund sollten sich Projektteams mit ähnlicher Ausrichtung wie ML_SoSi bei grossen Datenmengen nicht alleine auf die Skalierung der Infrastruktur verlassen. Das ML_SoSi-Projekt hat gezeigt, dass die folgenden Herangehensweisen mutmasslich zielführende Alternativen sind, um die Anforderungen an die technische Infrastruktur zu minimieren:
    - Anwendung eines Bootstrapping-Ansatzes zur Ermittlung der initialen Clusterlösung bei grossen Datenmengen: Die pro Iteration zu verarbeitender Datenmenge wird durch Stichprobenziehung minimiert, was z.B. ein direktes hierarchisches Clustering ermöglicht. Durch Wiederholung der Stichprobenziehung und Clusterbildung können weitere wertvolle Informationen für die Evaluation der initialen Clusterlösung gewonnen werden (siehe auch key learning Nr. 5).
    - Frühzeitig die notwendige Datenmenge für Prädiktionen ermitteln: Die learning curves als ein Mittel der Evaluation der Prädiktionsgüte (siehe Abschnitt 4.3) zeigen auf, wie klein die Datenmenge in der Trainingsphase sein darf, um dennoch ein Prädiktionsmodell mit guter Performance zu erhalten. Dies passt auch gut zu einem Bootstrapping-Ansatz. Aus diesem Grund empfehlen wir, dass in ähnlichen Projekten der gesamte Prozess vom clustering bis zur prediction früh pilotiert wird, um die optimale Datenmenge für das Modelltraining mithilfe von learning curves zu ermitteln.
3) **Flexible Arbeitsumgebung:** Die Anforderungen an statistische Software und Methoden sind insbesondere im Bereich der Datenwissenschaften projektabhängig und können je nach Problemstellung sehr spezifisch sein. Zudem bringen Projektmitarbeitende unterschiedliche Kompetenzen und Präferenzen bezüglich Programmiersprachen mit. Entsprechend sind flexible Arbeitsumgebungen, in welchen die präferierte Programmiersprache gewählt und Methodenpackages nach Bedarf installiert werden können, von grossem Mehrwert (Jupyter Notebooks erlauben z.B. die Programmierung in Python, R und weiteren Sprachen in einer einheitlichen Umgebung). Ein zusätzlicher Mehrwert entsteht, wenn die Arbeitsumgebung die Zusammenarbeit im Projektteam unterstützt, z.B. durch Codeversionierung und -sharing (Stichwort: GIT). Im Projekt wurden positive Erfahrungen mit der Datascience-Plattform Renku gesammelt, die diese Funktionalitäten vereint.

Statistische Methoden

4) **Einfache Lösungen bevorzugen:** Der angewendete Zweistufen-Algorithmus für das Clustering war weniger dem Untersuchungsgegenstand als den begrenzten Rechenkapazitäten geschuldet (Siehe Abschnitt 3.4). Ein klareres Vorgehen aus einer Hand könnte (frei nach «Ockhams Rasiermesser») zur Validität und Transparenz von Resultaten beitragen (mutmasslich könnte auch hier ein Bootstrappingansatz Vorteile aufweisen). Die Bevorzugung von einfachen Lösungen ist nicht auf Problemstellungen von unsupervised machine learning (Strukturerkennung) begrenzt, sondern gilt auch für Probleme des supervised machine learning (Prädiktion). Aufgrund der fachlichen Einschätzung der initialen Clusterlösung sowie der Übereinstimmungen mit anderen Forschungsresultaten (siehe Fussnoten 6 bis 8) ist es hingegen zu bezweifeln, dass ein alternativer Ansatz grosse Abweichungen in den typischen Verlaufsmustern zu Tage gebracht hätte.

5) **Statistische Kriterien bei Evaluation von Clusterlösungen hoch gewichten:** Im ML_SoSi-Projekt wurde aus statistischer Sicht vor allem die Clusterhomogenität anhand der within-sum-of-squares begutachtet. Aus fachlicher Sicht wurde zudem eine Lösung mit 10 Clustern bevorzugt, was aus statistischer Sicht nicht zwingend angezeigt war. Während der Entwicklung der Kriterien für die Aktualisierung der initialen Clusterlösung zeigte sich dann anhand der Silhouette-plots, dass in der initialen Clusterlösung die Abgrenzung der Cluster nicht ideal war. Silhouette-plots können daher zielführend bereits bei der Evaluation der initialen Clusterlösung beigezogen werden (sie sind jedoch nicht geeignet für grosse Datenmengen, weshalb eine Stichprobenziehung notwendig werden könnte). Aus einem Bootstrapping-Ansatz in der Clusteringphase könnten zudem Kennzahlen zur Clusterstabilität, zur statistisch optimalen Clusteranzahl und zur Modellwahl gewonnen werden. Eine stärkere Berücksichtigung statistischer Kriterien bei der Wahl der initialen Clusterlösung führt tendenziell zu einer besseren Clusterabgrenzung (Vorbehalte siehe auch key learning 7). Dies wirkt sich auch bei der Übertragung der initialen Clusterlösung auf neue Kohorten und bei der Einschätzung, wann die initiale Clusterlösung aktualisiert werden muss, vorteilig aus, da weniger Verläufe an der Grenze zwischen verschiedenen Clustern zu befürchten sind.

6) **Distanzmass hinterfragen und dem Erkenntnisinteresse anpassen:** Wie zuvor ausgeführt (key learning 5), zeigt sich in der initialen Clusterlösung, dass die Abgrenzung zwischen den Clustern nicht ideal ist. Dieses Resultat bleibt bestehen, auch wenn die Anzahl Cluster schrittweise reduziert wird. Dies kann in Zusammenhang mit dem Distanzmass stehen. Um eine Clusterlösung zu verbessern, lohnt es sich daher, das Distanzmass zu hinterfragen. Während die edit-Distanz (entspricht Levensthein-Distanz) ideal ist, um Sequenzen miteinander zu vergleichen, sind die Damerau-Levensthein-Distanz und die Hamming-Distanz etwas robustere Varianten. Weiter könnten die Substitutionskosten anhand von Fachwissen und abhängig von den Projektzielen gewichtet werden, anstatt anhand der empirischen Transformationswahrscheinlichkeiten.

7) **Clusteringproblem genau analysieren:** Die nicht optimale Abgrenzung der Cluster in der initialen Lösung kann auch Ursachen unabhängig vom Distanzmass haben. Es gibt Clusteringprobleme, bei welchen grundsätzlich nicht jeder Datenpunkt eindeutig einem Cluster zugeordnet werden kann. So können einzelne Verläufe im System der Sozialen Sicherheit Charakteristiken verschiedener typischer Verlaufsmuster in sich vereinen, insbesondere wenn die Verläufe einen längeren Zeitraum abdecken. Bei solchen Clusteringproblemen können Methoden des Fuzzy-Clustering angewendet werden, bei welchen Datenpunkte für jedes Cluster einer spezifischen Clusterlösung einen Zugehörigkeitsgrad erhalten (für jedes Cluster ein Wert zwischen 0 und 1). Datenpunkte, die für kein Cluster einen hohen Zugehörigkeitsgrad ausweisen, können z.B. aus der Analyse ausgeschlossen werden, um so die Abgrenzung zwischen den Clustern zu verbessern.

8) **Standards bei Training von Prädiktionsmodellen einhalten:** Das Projektteam hat mit folgenden Prinzipien sehr gute Erfahrungen bei der Erstellung des Prädiktionsmodells gemacht:

   - Genügend grosse Datenmenge, ev. Daten poolen.

- Training-Test-Datensplitt: 80% der Daten für das Training, 20% der Daten für die Schlussevaluation der Performance.
- Der Problemstellung angepasste Performancemetrik wählen (data imbalance beachten).
- Mindestens 5-fold cross validation, um den Generalisierungsfehler zu minimieren.
- Systematisches Experimentieren (no free lunch): Um das beste Prädiktionsmodell zu erhalten, sehr unterschiedliche statistische Methoden jeweils mit unterschiedlicher Parametrierung testen (Stichwort: grid search).
- Scheinbar unwichtige Aspekte wie Datenformatierung (z.B. factor vs one-hot) mitberücksichtigen.

9) **Zeitreihen einer initialen Clusterlösung erstellen mittels Prädiktion:** Es ist Clusteringverfahren inhärent, dass sie als induktive Methoden der Mustererkennung (unsupervised machine learning) trotz gleicher Parametrierung mit hoher Wahrscheinlichkeit nur bedingt vergleichbaren Resultate auf unterschiedlichen Grundgesamtheiten liefern. Um dennoch Zeitreihenanalysen mit einer konsolidierten Clusterlösung zu ermöglichen, wurde im Projekt ein Ansatz entworfen, mit dem die initiale Clusterlösung auf eine neue Grundgesamtheit (Kohorte) mittels Prädiktion (supervised machine learning) übertragen wird. Dieser Ansatz funktioniert mit hoher Präzision; Voraussetzung ist jedoch eine gut konsolidierte Clusterlösung (siehe key learnings 4) bis 7) und ein sehr performantes Prädiktionsmodell (siehe key learning 8), was wiederum von der Problemstellung und der zur Verfügung stehenden Datenmenge abhängig ist. Auch die verwendeten Evaluationskriterien, wann die initiale Clusterlösung aktualisiert werden muss, haben sich bewährt. Im Projekt wird geschlussfolgert, dass dieser Ansatz für die Darstellung und Analyse von Entwicklungen über die Zeit mithilfe von Clusterlösungen zielführend ist. Wir empfehlen bei ähnlichen Problemstellungen, diesen Ansatz an erster Stelle zu testen.

Resultate

10) **Erkenntnisinteresse hinterfragen:** Als einziges Projekt in der ersten Welle von Pilotprojekten im Rahmen der Dateninnovationsstrategie des BFS hatte ML_SoSi zum Ziel, mithilfe komplementärer, induktiver Analysemethoden einen Ansatz zu entwickeln, wie neue statistische Kennzahlen für das Zielpublikum des BFS entwickelt werden können (bei den anderen Pilotprojekten stand vor allem die Effizienz- und Qualitätssteigerung in der Statistikproduktion im Vordergrund). Es stellt sich heute die Frage, ob ein ausschliesslich induktives Verfahren zu diesem Zweck sinnvoll ist. Die Anspruchsgruppen der öffentlichen Statistik haben oft konkrete Informationsbedürfnisse, die anhand von BFS-Publikationen bedient werden (z.B. Bezifferung der Übertrittsquote aus der Arbeitslosenversicherung in die Sozialhilfe oder Anstieg an Sozialhilfefällen während der Covid-19-Pandemie). Entsprechend werden anhand der Informationsbedürfnisse hypothesengeleitet Indikatoren und Indikatorensysteme entwickelt, die Antworten auf spezifischen Fragen liefern (deduktives Vorgehen). Induktive Methoden sind hingegen nicht in der Lage zielgerichtet Resultate zu spezifischen Fragen zu produzieren; sie bilden vielmehr die in den Daten vorhandenen Strukturen ab, ungeachtet dem Erkenntnisinteresse, welches an die Datenbasis herangetragen wird. Wir empfehlen bei ähnlich gelagerten Projekten deshalb zu Beginn, das leitende Erkenntnisinteresse genau zu hinterfragen. Wir sind dabei überzeugt, dass induktive Methoden in vielen Fällen einen hohen Mehrwert generieren können und sehen vor allem zwei Szenarien (siehe auch Abschnitt 6.3):

- *Rein induktiver Ansatz zur Produktion von neuen statistischen Kennzahlen* (analog ML_SoSi): Die klassische Abfolge von Strukturerkennung (z.B. mithilfe von Clustering und unsupervised machine learning) und Prädiktion von entdeckten Mustern (supervised machine learning) ist sinnvoll bei gesellschaftlichen Phänomenen, die mit (ev. neuen) grossen Datenbeständen abgebildet werden können und über die wenig bekannt ist oder unklare Informationsbedürfnisse bestehen.
- *Kombination von deduktiven und induktiven Verfahren zur Produktion von neuen Statistischen Kennzahlen*: In diesem Szenario werden aus den Informationsbedürfnissen und Fragestellungen der Anspruchsgruppen Indikatoren(-systeme) abgeleitet, welche Fakten für die Diskussion derselben beisteuern können. Induktive Methoden werden komplementär dazu eingesetzt, um zu überprüfen, ob in den Daten weitere relevante Strukturen

und Entwicklungen abgebildet sind, die bisher (noch) nicht in den Fokus der Anspruchsgruppen gelangt sind. Induktive Methoden können deshalb helfen zu prüfen, ob in der öffentlichen Diskussion die richtigen Fragen gestellt werden bzw. ob relevante Entwicklungen übersehen werden.

6 Transfer in die Produktion

6.1 Auswirkungen von ML_SoSi auf die statistische Standardproduktion

Die erfolgversprechende erste Phase des Projekts liess Erwartungen entstehen, dass das induktive Clusterverfahren möglicherweise wie erprobt in die statistische Standardproduktion überführt werden könnte. Aus inhaltlichen und aus technischen Gründen geschieht dies jedoch nicht direkt. Vielmehr wurde im Sinne des «induktiv-deduktiven Verfahrenszyklus» viele Erkenntnisse aus dem induktiven Clustering in deduktive Verlaufsprofile übersetzt (regelbasiert) und können so in einfacherer Weise für den Datenkonsumenten in Wert gesetzt werden. Diese Profile bilden zentrale Elemente der im Clustering herausgearbeiteten Eigenheiten in den Verläufen ab und beschreiben diese mit Längsschnittindikatoren.

Im Rahmen des Projekts ML_SoSi wurde zudem ein hohes Verständnis für die komplexe Datenstruktur aufgebaut. Durch eine neue Datenorganisation im Long-Format konnten die parallelen und nicht-kontinuierlichen Verläufe in den drei Leistungssystemen leichter für die Analyse zugänglich gemacht. Diese wurden für die weitere SHIVALV-Produktion und Diffusion übernommen. Auch im definitorischen Bereich für die Beschreibung der Daten mittels Längsschnittindikatoren wurden umfangreiche Tests und Entscheide im Projekt notwendig. Dies beinhaltete unter anderem sowohl Anpassungen aufgrund von kurzen Überschneidungen und Unterbrüchen, die grossmehrheitlich auf Ursachen im Bereich administrativer Datenerfassung und -verarbeitung beruhen, als auch grundlegende definitorische Arbeiten im Bereich von Neueintritten, Austritten und Rückkehrern sowie Doppelbezug. In der Folge konnten die Projektergebnisse, die mit einer relativ selektiven Eintrittskohorte[24] in einem Sozialsystem (ALV) mit induktiven Methoden in ML_SoSi erarbeitet wurden, direkt in die Entwicklung der Längsschnittindikatoren einfliessen (siehe Publikationsverweis in Fussnote 1).

Eine wesentliche Erkenntnis aus den Clusteranalysen in ML_SoSi ist, dass zusätzliche Registerdaten für das Verständnis der Verläufe in den Sozialsystemen notwendig sind. Dies beinhaltet neben dem im Projekt bereits berücksichtigten, aber in SHIVALV noch nicht integrierten, Arbeitsmarktstatus insbesondere auch die Beschreibung der im Projekt unter «Leavers» zusammengefassten Zustände. Für die verlaufsstatistischen Analysen im Bereich der sozialen Sicherheit werden daher auch zusätzliche Datenquellen integriert, die hier weitere Informationen liefern. Vor allem über die Statistik der Bevölkerung (STATPOP) sollen Veränderungen der ständigen Wohnbevölkerung mit Todesfällen sowie Ein- und Auswanderungen und auch Übertritte in das Rentenalter integriert und zum Verständnis der Verläufe in den Sozialsystemen nutzbar gemacht werden.

Auch im Bereich der Darstellung der Verläufe im System der sozialen Sicherheit wurde das Pilotprojekt ML_SoSi für die Erprobung der Verständlichkeit und Kommunizierbarkeit von verschiedenen Längsschnittmassen genutzt. Einerseits umfasste dies graphische Verfahren. Dabei zeigte sich, dass für das Verständnis von Sankeyplots die Komplexität der Verläufe im Sozialsystem stark reduziert werden müsste. Diese Vereinfachung ist aufgrund der Vielfalt der Übergänge zwischen den Systemen der Sozialen Sicherheit nicht ohne weiteres möglich, sodass diese Darstellungsform für die Basisindikatoren und deren Visualisierung eher keine Verwendung mehr finden wird. Als sehr geeignet erwiesen sich hingegen State Distribution Plots, die auch in der Publikation als Standard Eingang finden. Das Projekt zeigte dabei dennoch auch Herausforderungen auf, wie beispielsweise die klare Abgrenzung von State Distribution Plots und Grafiken mit individuellen Verläufen, um Fehlinterpretationen zu vermeiden.

Aus dem Projekt konnten schliesslich Erkenntnisse zur Weiterentwicklung der statistischen Analyseinfrastruktur im BFS gewonnen werden. Eine der direkten Folgen für die Produktion war die gestiegene Bedeutung von GIT als Tool für kollaborative Skripterstellung im BFS. Auch bezüglich technischer Anforderungen an die Analyseinfrastruktur konnten Spezifikationen aus dem Projekt abgeleitet werden.

[24] Neu-ALV-Taggeldbezüger ohne Taggeldbezug in den vorangegangenen 24 Monaten

6.2 Weiteres Vorgehen beim Transfer in die Produktion

Der Transfer des Pilotprojekts ML_SoSi in die Produktion wird stufenweise und in enger Abstimmung mit den Stakeholdern und Statistik-Konsumenten umgesetzt, wobei in der Publikation vom Juni 2023 bereits mehrere Elemente berücksichtigt wurden (siehe Publikationsverweis in Fussnote 1).

1) In einer ersten Phase wurde die **Kohortenbetrachtungsweise** in der Weiterentwicklung der Statistik SHIVALV integriert. Mit dem einhergehenden Perspektivwechsel wurde die bisherige im SHIVALV-Monitoring angewandte querschnittliche Kalenderjahrperspektive ergänzt. Die hierfür nötigen definitorischen Arbeiten bauen auf den zentralen Erkenntnissen in ML_SoSi auf. Dies beinhaltet insbesondere die Abgrenzung von «Neu-Bezügern», wobei die Definition am Beispiel der ALV im Projekt ML_SoSi aufgrund der gewonnenen Ergebnisse weiterentwickelt wurde.
2) Die in ML_SoSi für die Beschreibung der Cluster verwendeten **Längsschnittindikatoren** wurden als Basis für die Entwicklung von beschreibenden Längsschnittindikatoren in SHIVALV verwendet. Dies beinhaltet insbesondere Verbleibs-, Rückkehr und Austrittsindikatoren zu relevanten Kohorten.
3) Auf Basis der Ergebnisse des Clusterings in ML_SoSi wurden **Profile** von Verläufen gebildet, deren quantitative Anteile im Zeitverlauf beobachtet werden können. Die Präsentation nach Aggregationsniveau erfolgt dabei mit steigender Komplexität. Unterschieden werden drei Grundprofile (Einfach, Rückkehr, Mehrfachbezug), die dann nach Leistungssystem, Dauer und Häufigkeit ausdifferenziert werden.
4) Die **Darstellung** der längsschnittlichen Strukturen und Indikatoren erfolgt einerseits mit Überlebenskurven. Andererseits wird angestrebt, die in ML_SoSi eingesetzten state-distribution-plots für die beschreibenden Analysen einzusetzen.
5) Der Transfer der induktiven Methoden, insbesondere des **Clusterings**, in der SHIVALV-Produktion wird zunächst nicht weiterverfolgt. Nach Einführung der vorgenannten Schritte wird der Informationsbedarf und der Mehrwert der komplexen induktiven Vorgehensweise nochmals mit den Stakeholdern evaluiert. Voraussetzung für eine Übernahme in die Produktion ist auch eine performante Analyseinfrastruktur.

6.3 Generischer, induktiver Analyseansatz für individuelle Verlaufsdaten in der Produktion

Ziel des Projektes war es, unter anderem einen Analyseansatz zu entwerfen, mit welchem komplementäre induktive Verfahren für die Analyse von individuellen Verlaufsdaten im Rahmen der öffentlichen Statistik in Wert gesetzt werden können. Darin lebt die Hoffnung, wichtige Erfahrung für ähnlich gelagerte Projekte generisch aufzubereiten und zum institutionellen Lernen beizutragen.

Der im Folgenden skizzierte generische Analyseansatz eignet sich für Datenanalysen, in denen typische Verlaufsmuster in individuellen Verlaufsdaten identifiziert und analysiert werden sollen. Er ist dann sinnvoll, wenn entweder keine klaren Informationsbedürfnisse oder Fragestellungen zu einem gesellschaftlichen Phänomen bestehen bzw. wenn kein konsolidierter Wissensbestand dazu existiert und/oder wenn die zugrundeliegenden Daten aus Sekundärquellen stammen, die nicht spezifisch für den Untersuchungsgegenstand designt wurden. Wir gehen dabei davon aus, dass die Daten in ihrem Rohzustand bereits vorliegen. Wir empfehlen eine agile Organisations- und Arbeitsweise mit entsprechenden.

Der generische Analyseansatz zeigt anhand zweier Szenarien auf, wie die ausgetesteten induktiven Verfahren in die Statistikproduktion eingebettet werden können. Einerseits wird ein rein induktives Vorgehen und andererseits eine konkrete Anwendung des «induktiv-deduktiven Verfahrenszyklus» (siehe Dateninnovationsstrategie, S. 11, Fussnote 5) dargestellt. Die Begründung für die Aufsplittung auf zwei Szenarien findet sich in key learning Nr. 10). Die weiterführenden methodischen Details zu den einzelnen Schritten finden sich u.a. in Abschnitt 3 bzw. in den key learnings (Abschnitt 5).

Tabelle 19: Generischer Analyseansatz für die Integration induktiver Verfahren bei der Analyse von Verlaufsdaten in der Statistikproduktion

Nr.	Vorbereitung	Key learning	GSBPM
	Erkenntnisinteresse klären: - Haben Anspruchsgruppen klare Informationsbedürfnisse und Fragestellungen? - Besteht ein fundiertes empirisches Wissen über Zusammenhänge bezüglich der interessierenden Phänomene? - Wurde die Datenbasis designt, um spezifische Fragestellungen zu untersuchen (Primäranalyse)? Oder handelt es sich um Sekundärdaten ohne eindeutige Anbindung an analytische Erkenntnisinteressen und mit dem Potential, sie so aufzubereiten, damit die gewünschten Phänomene abgebildet werden können (Sekundäranalyse)?	10)	Specify needs
	Szenario wählen und anpassen		Design
	Anforderungen an Infrastruktur evaluieren (Rechenleistung, Software) und zur Verfügung stellen	1) und 3)	Build
	Sekundärdaten so **aufbereiten**, damit sie das interessierende gesellschaftliche Phänomen gut abbilden.		Design, Process
	Szenario 1: Identifikation von typischen Verlaufsmustern mit Sequenzclustering und Erstellung von Zeitreihen mittels Prädiktion		
	Clustering: Bootstrapping oder nicht? Anhand der Datenmenge entscheiden ob, ein direktes hierarchisches Clustering oder ein Bootstrap-Ansatz sinnvoll ist (letzteres bei grossen Datenmengen)	2) und 4)	Design
	Pipeline von Clustering bis Prädiktion **früh und provisorisch pilotieren**, um mit learning curves die notwendige Datenmenge für eine gute Prädiktionsgüte zu ermitteln.	2)	Design
	Initiale Clusterlösung konsolidieren, dabei insbesondere edit-Distanz hinterfragen und Agglomerationsverfahren festlegen. Bei der Evaluation der Clusterlösungen und insbesondere bei der Wahl der Anzahl Cluster statistische Kriterien hochgewichten. State distribution plots und Verlaufsindikatoren für die fachliche Interpretation nutzen, bei wenigen Beobachtungszeitpunkten Sankeyplots testen.	5), 6), 7) und 8)	Analyse
	Falls die Datenmenge zu gross für die direkte Verarbeitung in der Pipeline ist und falls weniger Daten für die Analysen erforderlich sind, als in der Grundgesamtheit vorhanden: **Stichprobenziehung** aus der vom Clusteralgorithmus gelabelten Grundgesamtheit, so dass optimale Samplegrösse erreicht wird. Andernfalls Infrastruktur skalieren.	2) und 1)	Collect
	Prädiktionsmodelle trainieren: Dazu auf die Problemstellung angepasste Evaluationsmetrik wählen und best practice anwenden. Modell mit der besten Performance evaluieren. Modelle mit ähnlich guter Performance ebenfalls erwägen («Ockhams Rasiermesser»). In diesem Punkt ist lückenlose Dokumentation besonders wichtig, da Transparenz bei supervised machine learning sonst schwierig herzustellen ist.	4) und 8)	Analyse
	Zeitreihen mittels Prädiktion erstellen: Anwenden des Prädiktionsmodells auf weitere Grundgesamtheiten/Kohorten. Deskriptive Analyse der Prädiktionsresultate und Vergleich mit den Resultaten, wenn der Clusteringalgorithmus direkt auf die neuen Kohorten angewendet wird (state distribution plots und Verlaufsindikatoren).	9)	Process
	Kriterien für die **Aktualisierung der initialen Clusterlösung** anhand bereits vorhandener Zeitreihen entwickeln und Schwellenwerte festlegen.	9)	Analyse
	Szenario 2: Fragestellung- bzw. hypothesengeleitete Herleitung von Indikatoren, Verlaufstypologien und Erklärungsmodellen (Deduktion) und Ergänzung		

	der Resultate mit Sequenzclustering (Induktion) als Anwendung des «induktiv-deduktiven Verfahrenszyklus».		
	Entwicklung von Indikatoren und Erklärungsmodellen, welche die Informationsbedürfnisse der Anspruchsgruppen möglichst direkt abbilden. **Iterative Verbesserung** dieser Resultate im Austausch mit den Anspruchsgruppen.		Specify needs, Design
	Umfassende deskriptive und inferenzstatistische Analyse der Fragestellungen		Analyse
	Entscheiden mit welchen **Informationen** (Subsamples, Features) und **Clusteringverfahren** die in Frage stehenden gesellschaftlichen Phänomene am besten abgebildet werden können. Anhand der **Datenmenge** entscheiden ob, Clustering auf der gesamten Grundgesamtheit möglich ist oder ob ein Verfahren mit Datenreduktion angezeigt ist (Bootstrapping, Sampling, spezieller Clusteralgorithmus)	2), 4) und 5)	Design
	Clusterlösungen anhand der Achsen «Anzahl Cluster», «Clusteralgorithmus und Parametrierung» und «Distanzmass» **variieren**. Evaluation der unterschiedlichen Clusterlösungen anhand statistischer Kriterien.	5), 6) und 7)	Analyse
	Deskription und fachliche Interpretation der Cluster mithilfe von Verlaufsindikatoren und state distribution plots oder Sankeyplots. Relevanz erklären. Ziel in diesem Verfahrensschritt ist es nicht eine konsolidierte Clusterlösung zu erhalten, sondern Phänomene und Entwicklungen zu entdecken, die nicht im Fokus der Informationsbedürfnisse der Anspruchsgruppen stehen, aber trotzdem relevant sind und in der weiteren Analyse Beachtung finden sollten.		Analyse

Im Projekt ML_SoSi konnten verschiede Aspekte dieses generischen Analyseansatzes getestet werden. Die gemachten Erfahrungen halfen, diesen weiter zu konkretisieren. Im Nachhinein bevorzugen wir für die Statistikproduktion das Szenario 2, welches derzeit auch umgesetzt wird. Bezüglich des SHIVALV-Datensatzes bestehen spezifische Informationsbedürfnisse für Analysen, die auch im Rahmen von Begleitgruppen- und Arbeitssitzungen mit zentralen Stakeholdern aufgenommen werden und in die Weiterentwicklung der Indikatoren einfliessen. Die Entwicklung der nun konstruierten Indikatoren hat dabei umfangreich von den Erkenntnissen aus dem Projekt ML_SoSi profitiert.

7 Anhang

7.1 Codierung der Statusinformationen und Liste der Statuskombinationen

Codierung der Statusinformationen:

Jeder Verlauf ist eine Abfolge von monatlichen Statusinformationen, die in einem vierstelligen numerischen Code festgehalten sind (z.B. 2211). Die verschiedenen Stellen des Codes bilden die betrachteten Systeme ab; Die numerischen Werte bilden ab, ob eine Leistung bezogen wurde bzw. ein Erwerbseinkommen erziel wurde oder nicht.

Die vier Stellen im Code:

- erste Stelle im vierstelligen Code = Invalidenversicherung, IV
- zweite Stelle im vierstelligen Code = Arbeitslosentaggeld, ALV-
- dritte Stelle im vierstelligen Code = Sozialhilfe, SH
- vierte Stelle im vierstelligen Code = Erwerbsarbeit

Die verwendeten Werte im Code:

- 1 = eine Leistung aus dem entsprechenden System wurde bezogen / Erwerbsarbeit liegt vor
- 2= keine Leistung wurde bezogen / keine Erwerbsarbeit liegt vor

Code «2211» für einen Status einer beliebigen Person in einem beliebigen Monat bedeutetet also folgendes:

IV	**ALV**	**SH**	**Erwerb**
2=kein Bezug	2=kein Bezug	1=Bezug	1=Arbeit

Hauptstatus:

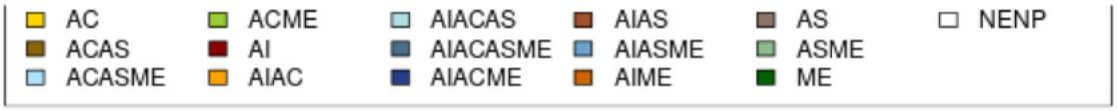
- ME (2221) : Marché de l'emploi
- AC (2122) : Assurance chômage
- AS (2212) : Aide sociale
- AI (1222) : Assurance invalidité
- NENP (2222) : Ni en emploi ni prestations

Kombinationen:

- AIACAS (1112) : Assurance invalidité, assurance chômage, aide sociale
- AIACME (1121) : Assurance invalidité, assurance chômage, marché de l'emploi
- AIAC (1122) : Assurance invalidité, assurance chômage
- AIASME (1211) : Assurance invalidité, aide sociale, marché de l'emploi
- AIAS (1212) : Assurance invalidité, aide sociale
- AIME (1221) : Assurance invalidité, marché de l'emploi
- ACAS (2112) : Assurance chômage, aide sociale
- ASME (2211) : Aide sociale, marché de l'emploi
- AIACASME (1111) : Assurance invalidité, assurance chômage, aide sociale, marché de l'emploi
- ACME (2121) : Assurance chômage, marché de l'emploi
- ACASME (2111) : Assurance chômage, aide sociale, marché de l'emploi

7.2 Legende

Legende für die state distribution plots

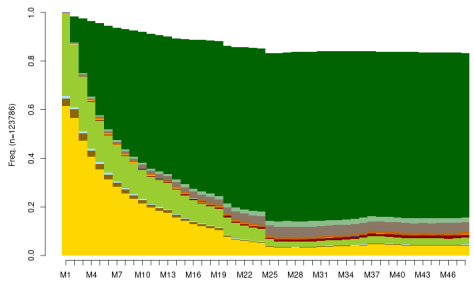
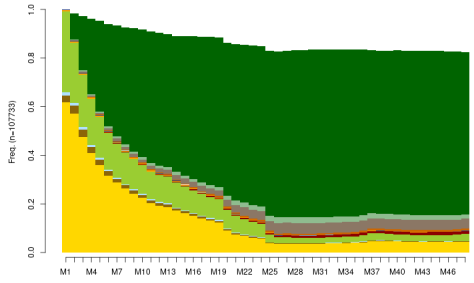
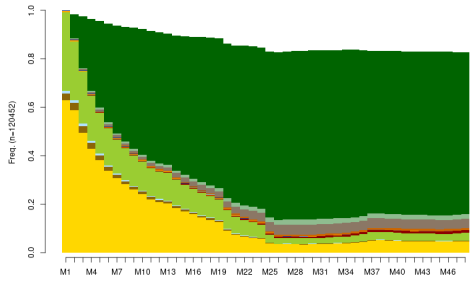
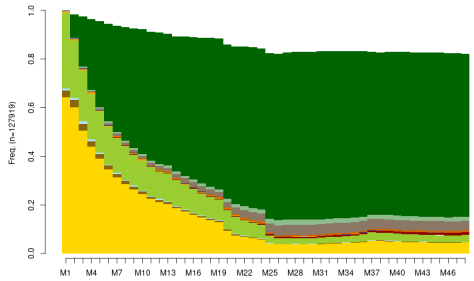


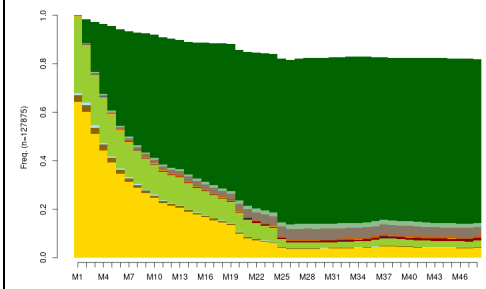
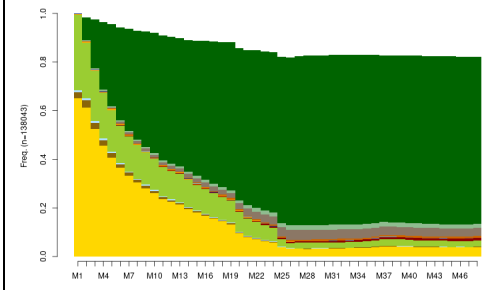
Legende für die Verlaufsindikatoren:

- VI1: Anzahl Monate mit ALV
- VI2: Dauer der ersten ALV-Bezugsperiode (Monate)
- VI3: Anzahl Monate ALV und Erwerbsarbeit kombiniert
- VI4: Anzahl Bezugsperioden ALV
- VI5: Anzahl Monate mit Erwerbsarbeit
- VI6: Anteil Personen mit mindestens einer SH-Bezugsperiode
- VI7: Anzahl Monate mit SH
- VI8: Anzahl Monate SH und Erwerbsarbeit kombiniert
- VI9: Anteil Personen mit mindestens einer IV-Bezugsperiode
- VI10: Anzahl Monate mit IV
- VI11: Anzahl Monate IV und Erwerbsarbeit kombiniert
- VI12: Anzahl Monate ohne Erwerbsarbeit und Sozialleistungen

7.3 Kohorten insgesamt, SDPs und VIs, K2010-K2015

Tabelle A 1: Kohorten insgesamt, SDPs und VIs, K2010-K2015

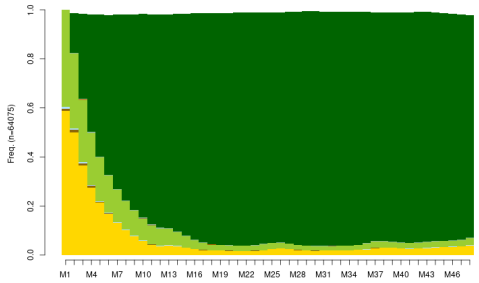
Kohorte	State Distribution Plot	Verlaufsindikatoren
2010		VI1 10.6 VI2 7.1 VI3 3.8 VI4 1.7 VI5 34 VI6 16% VI7 3 VI8 1 VI9 3% VI10 1 VI11 0 VI12 6
2011		VI1 10.8 VI2 7.2 VI3 3.8 VI4 1.7 VI5 33 VI6 16% VI7 3 VI8 1 VI9 2% VI10 1 VI11 0 VI12 6
2012		VI1 11.1 VI2 7.4 VI3 3.9 VI4 1.7 VI5 33 VI6 15% VI7 3 VI8 1 VI9 2% VI10 1 VI11 0 VI12 6
2013		VI1 11.2 VI2 7.5 VI3 3.8 VI4 1.7 VI5 33 VI6 15% VI7 2 VI8 1 VI9 2% VI10 1 VI11 0 VI12 6

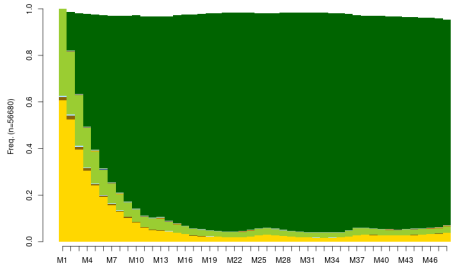
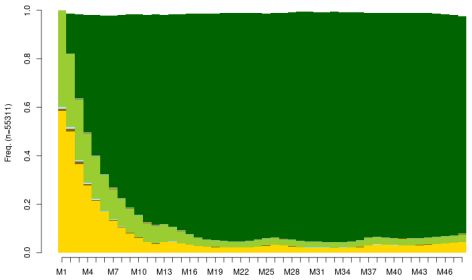
2014		VI1 11.2 VI2 7.5 VI3 3.8 VI4 1.7 VI5 33 VI6 15% VI7 2 VI8 1 VI9 2% VI10 1 VI11 0 VI12 6
2015		VI1 11.3 VI2 7.7 VI3 3.8 VI4 1.7 VI5 33 VI6 14% VI7 2 VI8 1 VI9 2% VI10 1 VI11 0 VI12 6

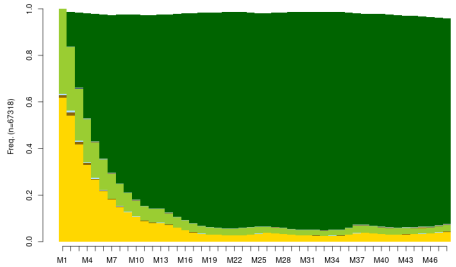
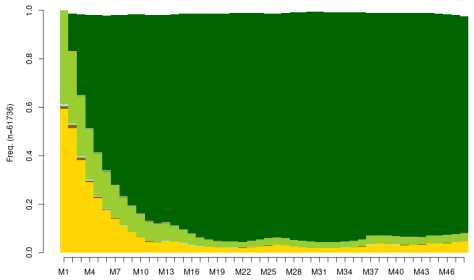
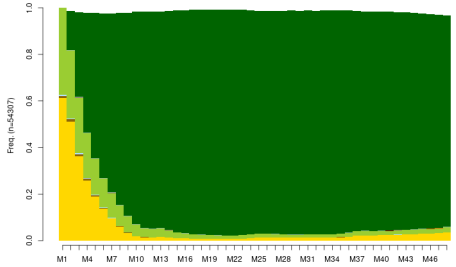
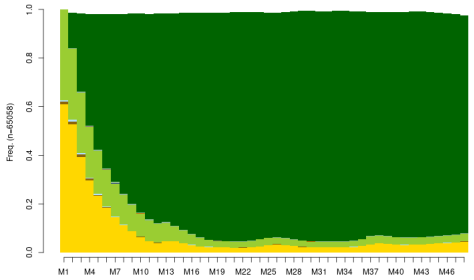
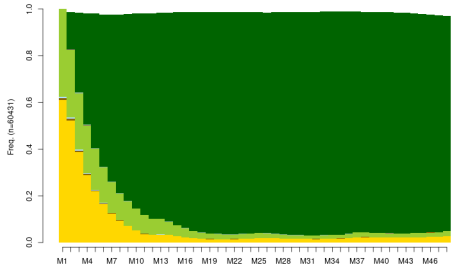
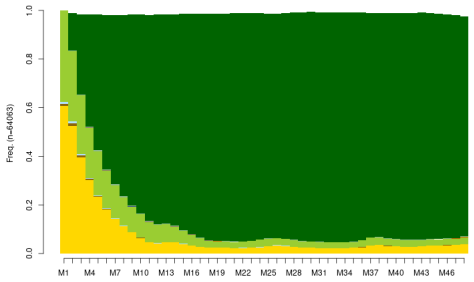
Quelle: BFS - SHIVALV-IK 2010-2019

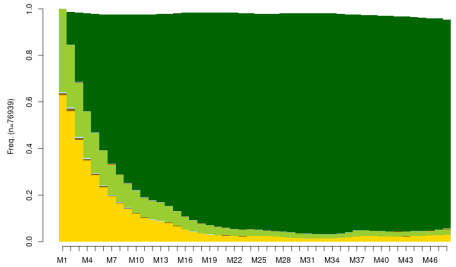
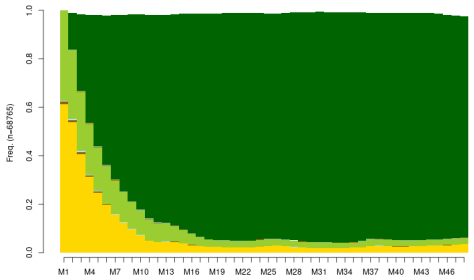
7.4 Cluster 1 - ALV Kurzzeit, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015

Tabelle A 2: Cluster 1 - ALV Kurzzeit, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015

Ko-horte	Referenz			
	State Distribution Plot	Verlaufsindikatoren		
2010		VI1	6.5	
		VI2	4.3	
		VI3	3.0	
		VI4	1.6	
		VI5	44	
		VI6	5%	
		VI7	0	
		VI8	0	
		VI9	0%	
		VI10	0	
		VI11	0	
		VI12	1	

	Neue Clusterlösung				Prädiktion			
	State Distribution Plot	Verlaufsindikatoren			State Distribution Plot	Verlaufsindikatoren		
2011		VI1	6.5			VI1	6.8	
		VI2	4.2			VI2	4.2	
		VI3	2.6			VI3	3.1	
		VI4	1.6			VI4	1.7	
		VI5	43			VI5	44	
		VI6	6%			VI6	5%	
		VI7	0			VI7	0	
		VI8	0			VI8	0	
		VI9	0%			VI9	0%	
		VI10	0			VI10	0	
		VI11	0			VI11	0	
		VI12	1			VI12	1	

2012		VI1 7.3 VI2 4.6 VI3 2.9 VI4 1.7 VI5 42 VI6 5% VI7 0 VI8 0 VI9 0% VI10 0 VI11 0 VI12 1		VI1 7.0 VI2 4.4 VI3 3.2 VI4 1.7 VI5 44 VI6 5% VI7 0 VI8 0 VI9 0% VI10 0 VI11 0 VI12 1
2013		VI1 5.4 VI2 3.8 VI3 2.4 VI4 1.5 VI5 44 VI6 4% VI7 0 VI8 0 VI9 0% VI10 0 VI11 0 VI12 1		VI1 7.1 VI2 4.4 VI3 3.2 VI4 1.7 VI5 44 VI6 5% VI7 0 VI8 0 VI9 0% VI10 0 VI11 0 VI12 1
2014		VI1 6.2 VI2 4.2 VI3 2.9 VI4 1.6 VI5 44 VI6 4% VI7 0 VI8 0 VI9 0% VI10 0 VI11 0 VI12 1		VI1 7.0 VI2 4.4 VI3 3.2 VI4 1.7 VI5 44 VI6 4% VI7 0 VI8 0 VI9 0% VI10 0 VI11 0 VI12 1

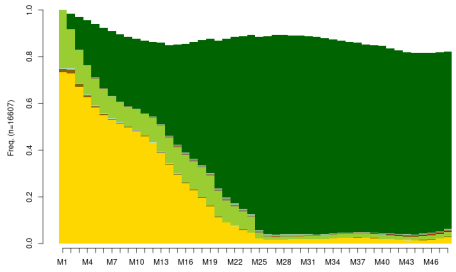
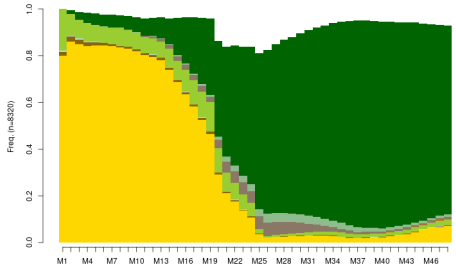
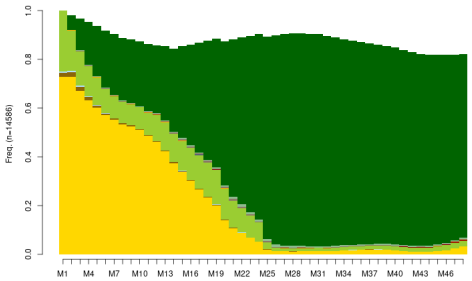
2015		VI1 7.4 VI2 5.0 VI3 3.0 VI4 1.7 VI5 42 VI6 4% VI7 0 VI8 0 VI9 0% VI10 0 VI11 0 VI12 1		VI1 7.0 VI2 4.5 VI3 3.2 VI4 1.7 VI5 43 VI6 4% VI7 0 VI8 0 VI9 0% VI10 0 VI11 0 VI12 1

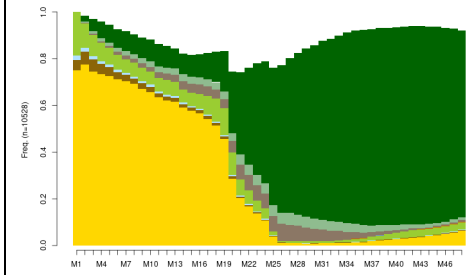
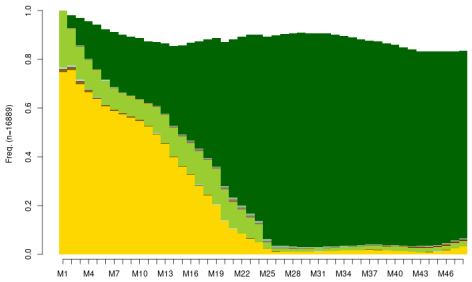
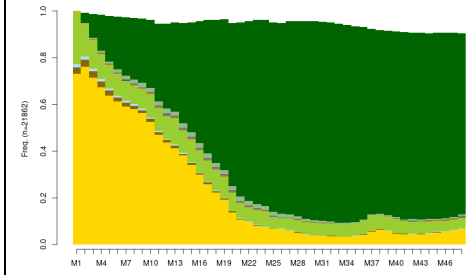
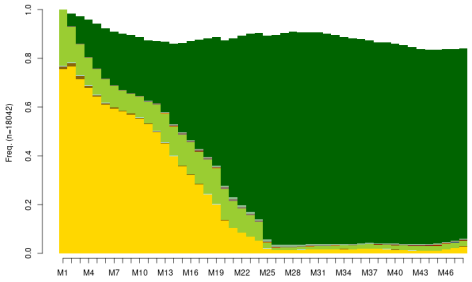
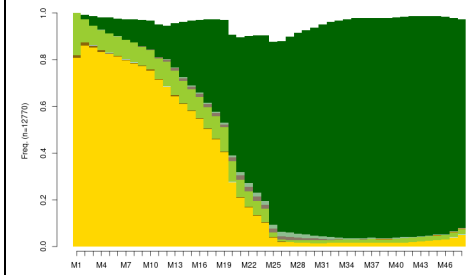
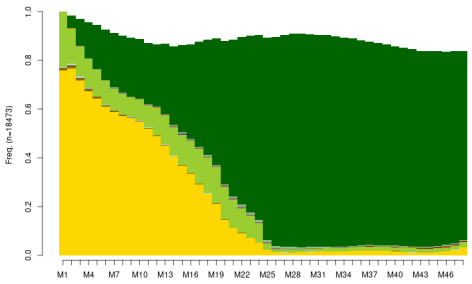
Anmerkung: Die Zuordnung des Clusters der neuen Lösung zum Clustern der Prädiktion bezieht sich auf eine visuelle Interpretation der state distribution plots und bezieht die Resultate der Jaccard-Matrix nicht mit ein.

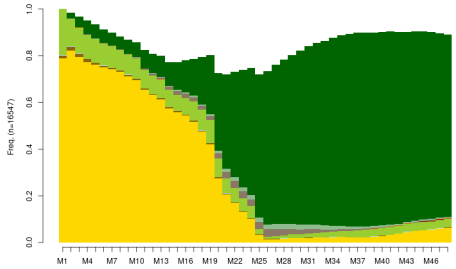
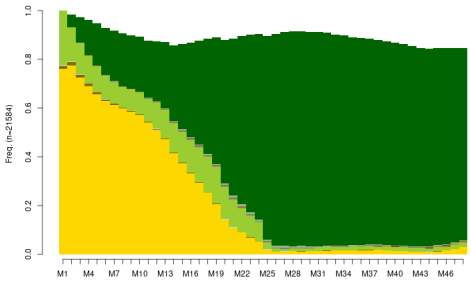
Quelle: BFS - SHIVALV-IK 2010-2019

7.5 Cluster 2 - ALV Langzeit, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015

Tabelle A 3: Cluster 2 - ALV Langzeit, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015

Ko-horte	Referenz		Neue Clusterlösung		Prädiktion	
	State Distribution Plot	Verlaufsindikatoren	State Distribution Plot	Verlaufsindikatoren	State Distribution Plot	Verlaufsindikatoren
2010		VI1 12.8 VI2 9.7 VI3 3.1 VI4 1.6 VI5 32 VI6 10% VI7 0 VI8 0 VI9 1% VI10 0 VI11 0 VI12 6				
2011				VI1 18.8 VI2 14.4 VI3 2.4 VI4 1.7 VI5 27 VI6 21% VI7 2 VI8 1 VI9 0% VI10 0 VI11 0 VI12 3		VI1 13.4 VI2 10.1 VI3 3.3 VI4 1.7 VI5 32 VI6 10% VI7 1 VI8 0 VI9 1% VI10 0 VI11 0 VI12 6

2012		VI1 16.8 VI2 12.9 VI3 2.4 VI4 1.7 VI5 26 VI6 27% VI7 4 VI8 2 VI9 0% VI10 0 VI11 0 VI12 6		VI1 13.8 VI2 10.6 VI3 3.2 VI4 1.6 VI5 32 VI6 10% VI7 1 VI8 0 VI9 0% VI10 0 VI11 0 VI12 5
2013		VI1 15.2 VI2 9.7 VI3 3.8 VI4 2.1 VI5 34 VI6 12% VI7 1 VI8 1 VI9 0% VI10 0 VI11 0 VI12 3		VI1 13.9 VI2 10.6 VI3 3.2 VI4 1.7 VI5 32 VI6 9% VI7 0 VI8 0 VI9 1% VI10 0 VI11 0 VI12 5
2014		VI1 17.3 VI2 13.1 VI3 2.5 VI4 1.7 VI5 31 VI6 11% VI7 1 VI8 0 VI9 0% VI10 0 VI11 0 VI12 2		VI1 14.0 VI2 10.6 VI3 3.2 VI4 1.7 VI5 32 VI6 9% VI7 0 VI8 0 VI9 0% VI10 0 VI11 0 VI12 5

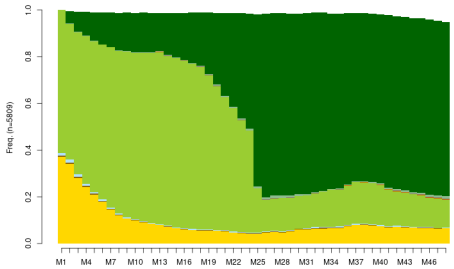
2015					
		VI1	17.1		VI1	14.1
		VI2	12.8		VI2	10.8
		VI3	2.8		VI3	3.2
		VI4	1.7		VI4	1.7
		VI5	26		VI5	32
		VI6	16%		VI6	9%
		VI7	1		VI7	0
		VI8	0		VI8	0
		VI9	0%		VI9	0%
		VI10	0		VI10	0
		VI11	0		VI11	0
		VI12	7		VI12	5

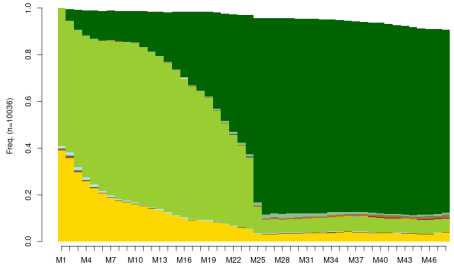
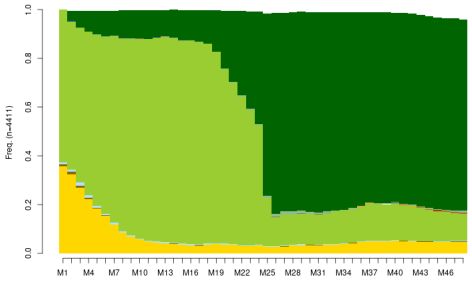
Anmerkung: Die Zuordnung des Clusters der neuen Lösung zum Clustern der Prädiktion bezieht sich auf eine visuelle Interpretation der state distribution plots und bezieht die Resultate der Jaccard-Matrix nicht mit ein.

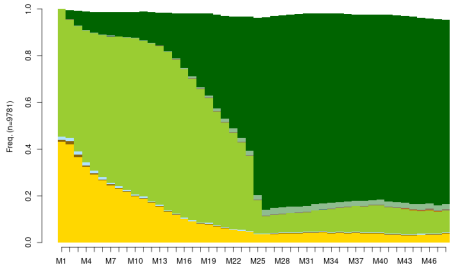
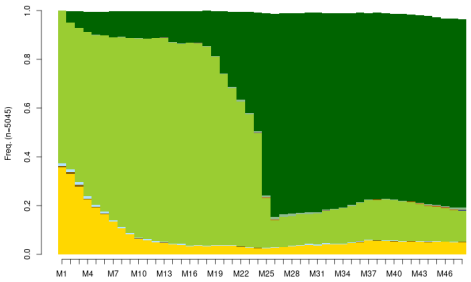
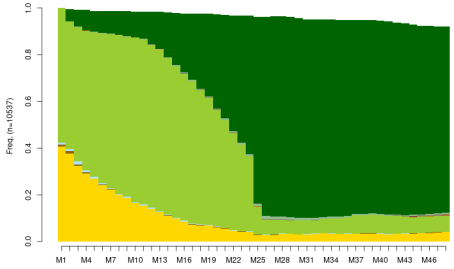
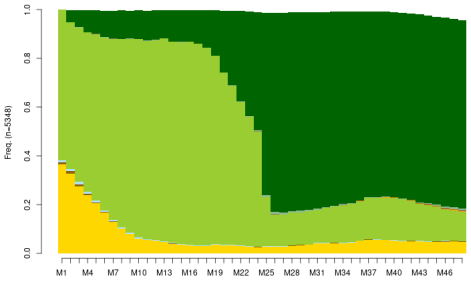
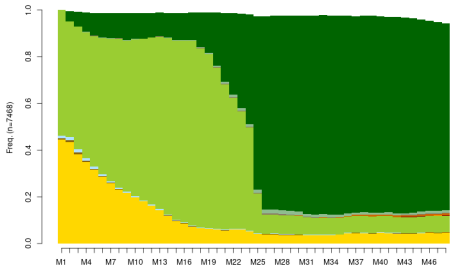
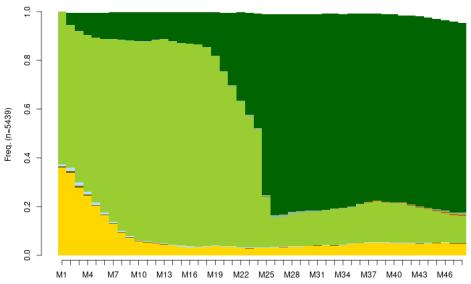
Quelle: BFS - SHIVALV-IK 2010-2019

7.6 Cluster 3 - Zwischenverdienst, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015

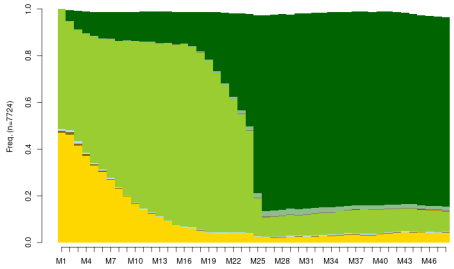
Tabelle A 4: Cluster 3 - Zwischenverdienst, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015

Ko-horte	Referenz			
	State Distribution Plot	Verlaufsindikatoren		
2010		VI1	24.0	
		VI2	13.7	
		VI3	19.4	
		VI4	2.7	
		VI5	43	
		VI6	10%	
		VI7	1	
		VI8	0	
		VI9	0%	
		VI10	0	
		VI11	0	
		VI12	1	

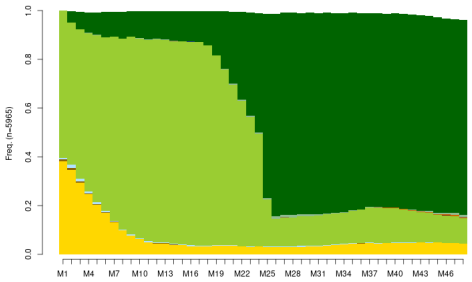
	Neue Clusterlösung		Prädiktion	
	State Distribution Plot	Verlaufsindikatoren	State Distribution Plot	Verlaufsindikatoren
2011		VI1 20.2 VI2 13.1 VI3 15.4 VI4 2.2 VI5 41 VI6 11% VI7 1 VI8 1 VI9 1% VI10 0 VI11 0 VI12 2		VI1 24.5 VI2 15.1 VI3 21.0 VI4 2.4 VI5 44 VI6 9% VI7 0 VI8 0 VI9 0% VI10 0 VI11 0 VI12 1

2012		VI1 21.7 VI2 14.0 VI3 16.3 VI4 2.3 VI5 41 VI6 13% VI7 1 VI8 1 VI9 0% VI10 0 VI11 0 VI12 1		VI1 24.7 VI2 15.2 VI3 21.1 VI4 2.4 VI5 44 VI6 9% VI7 0 VI8 0 VI9 0% VI10 0 VI11 0 VI12 0
2013		VI1 20.6 VI2 13.7 VI3 15.9 VI4 2.2 VI5 41 VI6 9% VI7 1 VI8 0 VI9 0% VI10 0 VI11 0 VI12 2		VI1 24.6 VI2 15.0 VI3 21.1 VI4 2.5 VI5 44 VI6 8% VI7 0 VI8 0 VI9 0% VI10 0 VI11 0 VI12 0
2014		VI1 22.8 VI2 15.2 VI3 17.4 VI4 2.3 VI5 41 VI6 10% VI7 1 VI8 1 VI9 1% VI10 0 VI11 0 VI12 1		VI1 24.6 VI2 15.0 VI3 21.0 VI4 2.5 VI5 44 VI6 8% VI7 0 VI8 0 VI9 0% VI10 0 VI11 0 VI12 0

2015



VI1	22.8
VI2	14.7
VI3	17.7
VI4	2.3
VI5	42
VI6	10%
VI7	1
VI8	1
VI9	0%
VI10	0
VI11	0
VI12	1



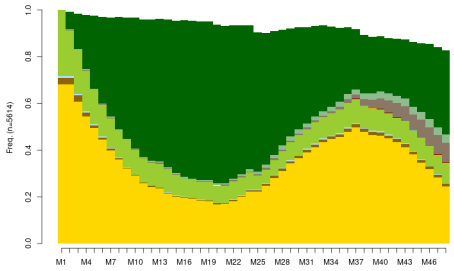
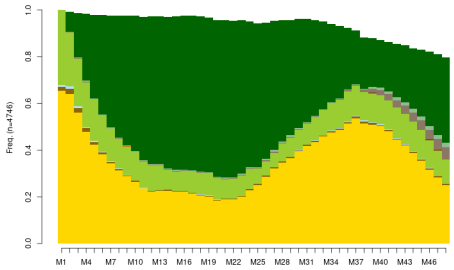
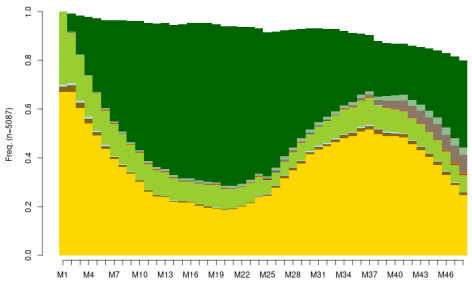
VI1	24.2
VI2	15.0
VI3	20.6
VI4	2.5
VI5	44
VI6	8%
VI7	0
VI8	0
VI9	1%
VI10	0
VI11	0
VI12	0

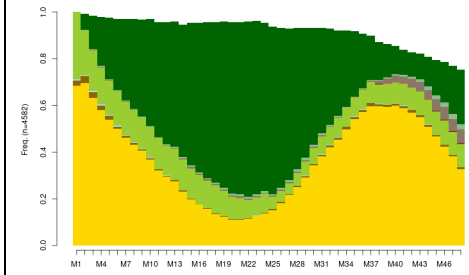
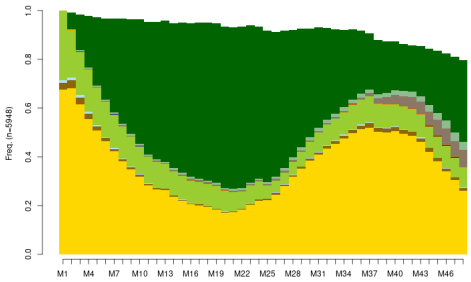
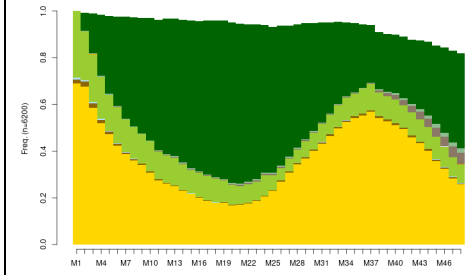
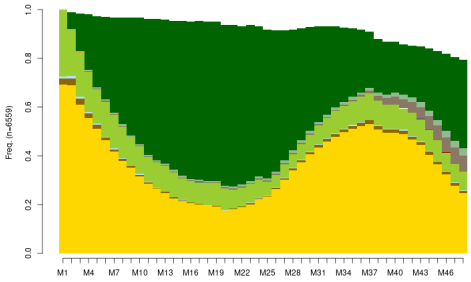
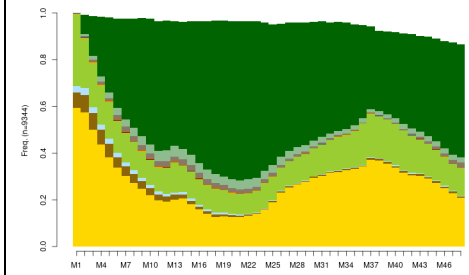
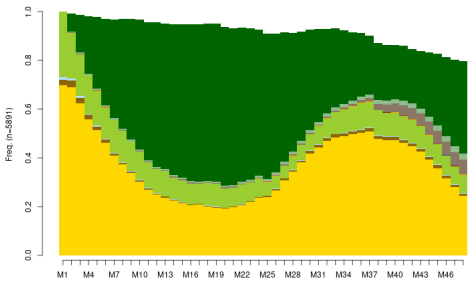
Anmerkung: Die Zuordnung des Clusters der neuen Lösung zum Clustern der Prädiktion bezieht sich auf eine visuelle Interpretation der state distribution plots und bezieht die Resultate der Jaccard-Matrix nicht mit ein.

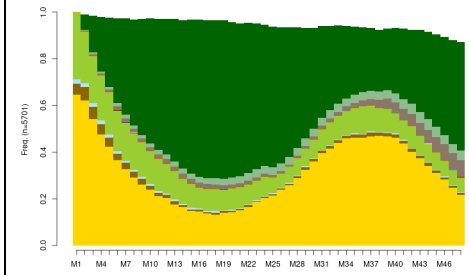
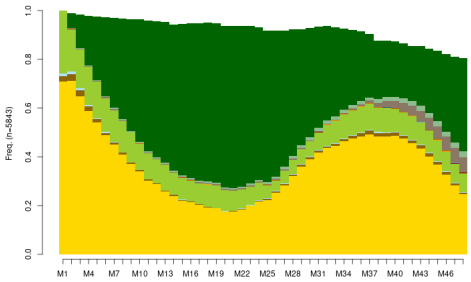
Quelle: BFS - SHIVALV-IK 2010-2019

7.7 Cluster 4 - ALV Mehrfach, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015

Tabelle A 5: Cluster 4 - ALV Mehrfach, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015

<table>
<tr><th rowspan="2">Ko-horte</th><th colspan="4">Referenz</th></tr>
<tr><th colspan="2">State Distribution Plot</th><th colspan="2">Verlaufsindikatoren</th></tr>
<tr><td>2010</td><td colspan="2">  </td><td colspan="2"><table>
<tr><td>VI1</td><td>22.3</td></tr>
<tr><td>VI2</td><td>7.4</td></tr>
<tr><td>VI3</td><td>5.0</td></tr>
<tr><td>VI4</td><td>2.8</td></tr>
<tr><td>VI5</td><td>26</td></tr>
<tr><td>VI6</td><td>26%</td></tr>
<tr><td>VI7</td><td>2</td></tr>
<tr><td>VI8</td><td>1</td></tr>
<tr><td>VI9</td><td>0%</td></tr>
<tr><td>VI10</td><td>0</td></tr>
<tr><td>VI11</td><td>0</td></tr>
<tr><td>VI12</td><td>3</td></tr>
</table></td></tr>
<tr><td></td><td colspan="3">Neue Clusterlösung</td><td colspan="3">Prädiktion</td></tr>
<tr><td></td><td colspan="2">State Distribution Plot</td><td colspan="2">Verlaufsindikatoren</td><td colspan="2">State Distribution Plot</td><td>Verlaufsindikatoren</td></tr>
<tr><td>2011</td><td colspan="2">  </td><td colspan="2"><table>
<tr><td>VI1</td><td>23.1</td></tr>
<tr><td>VI2</td><td>7.1</td></tr>
<tr><td>VI3</td><td>5.7</td></tr>
<tr><td>VI4</td><td>3.0</td></tr>
<tr><td>VI5</td><td>27</td></tr>
<tr><td>VI6</td><td>18%</td></tr>
<tr><td>VI7</td><td>1</td></tr>
<tr><td>VI8</td><td>0</td></tr>
<tr><td>VI9</td><td>0%</td></tr>
<tr><td>VI10</td><td>0</td></tr>
<tr><td>VI11</td><td>0</td></tr>
<tr><td>VI12</td><td>3</td></tr>
</table></td><td colspan="2">  </td><td><table>
<tr><td>VI1</td><td>23.1</td></tr>
<tr><td>VI2</td><td>7.5</td></tr>
<tr><td>VI3</td><td>5.0</td></tr>
<tr><td>VI4</td><td>2.9</td></tr>
<tr><td>VI5</td><td>25</td></tr>
<tr><td>VI6</td><td>24%</td></tr>
<tr><td>VI7</td><td>2</td></tr>
<tr><td>VI8</td><td>1</td></tr>
<tr><td>VI9</td><td>1%</td></tr>
<tr><td>VI10</td><td>0</td></tr>
<tr><td>VI11</td><td>0</td></tr>
<tr><td>VI12</td><td>4</td></tr>
</table></td></tr>
</table>

2012		VI1 23.9 VI2 8.4 VI3 5.2 VI4 2.9 VI5 25 VI6 21% VI7 1 VI8 0 VI9 0% VI10 0 VI11 0 VI12 4		VI1 23.3 VI2 7.9 VI3 5.0 VI4 2.9 VI5 25 VI6 24% VI7 2 VI8 1 VI9 1% VI10 0 VI11 0 VI12 4
2013		VI1 23.2 VI2 7.7 VI3 5.2 VI4 2.9 VI5 27 VI6 17% VI7 1 VI8 0 VI9 0% VI10 0 VI11 0 VI12 3		VI1 23.3 VI2 7.8 VI3 4.9 VI4 2.9 VI5 25 VI6 22% VI7 2 VI8 1 VI9 0% VI10 0 VI11 0 VI12 4
2014		VI1 20.8 VI2 6.7 VI3 6.9 VI4 3.1 VI5 31 VI6 25% VI7 3 VI8 2 VI9 1% VI10 0 VI11 0 VI12 2		VI1 23.0 VI2 7.7 VI3 4.9 VI4 2.8 VI5 25 VI6 21% VI7 2 VI8 1 VI9 0% VI10 0 VI11 0 VI12 4

2015		VI1	21.6		VI1	23.1
		VI2	7.6		VI2	8.2
		VI3	5.6		VI3	4.9
		VI4	2.8		VI4	2.8
		VI5	28		VI5	25
		VI6	28%		VI6	20%
		VI7	4		VI7	2
		VI8	2		VI8	1
		VI9	1%		VI9	1%
		VI10	0		VI10	0
		VI11	0		VI11	0
		VI12	3		VI12	4

Anmerkung: Die Zuordnung des Clusters der neuen Lösung zum Clustern der Prädiktion bezieht sich auf eine visuelle Interpretation der state distribution plots und bezieht die Resultate der Jaccard-Matrix nicht mit ein.

Quelle: BFS - SHIVALV-IK 2010-2019

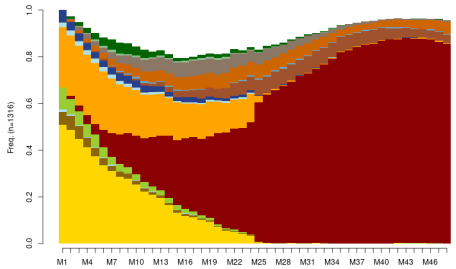
7.8 Cluster 5 - IV-Rente, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015

Tabelle A 6: Cluster 5 - IV-Rente, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015

Kohorte 2010

Referenz

State Distribution Plot



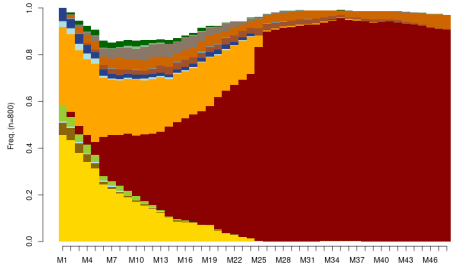
Verlaufsindikatoren

VI1	12.0
VI2	9.9
VI3	1.4
VI4	1.3
VI5	4
VI6	30%
VI7	5
VI8	0
VI9	100%
VI10	34
VI11	3
VI12	5

2011

Neue Clusterlösung

State Distribution Plot

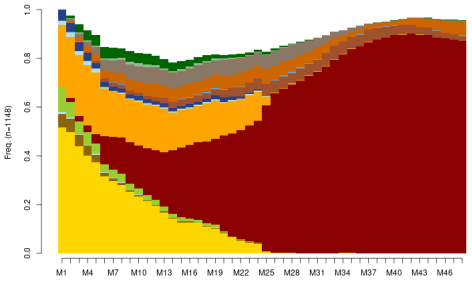


Verlaufsindikatoren

VI1	11.1
VI2	9.5
VI3	0.9
VI4	1.2
VI5	3
VI6	23%
VI7	2
VI8	0
VI9	100%
VI10	39
VI11	2
VI12	3

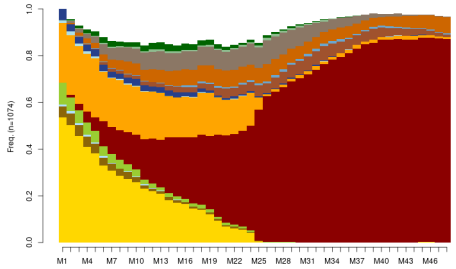
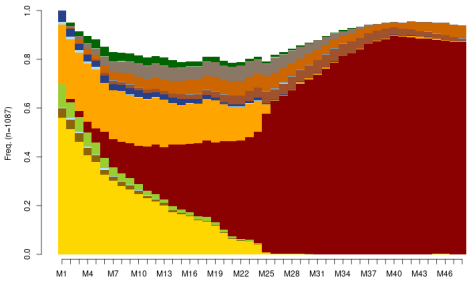
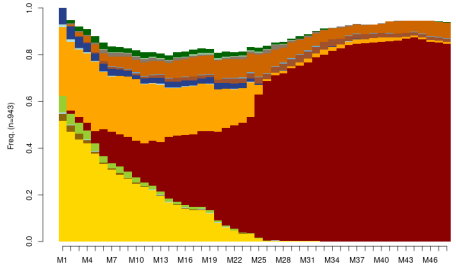
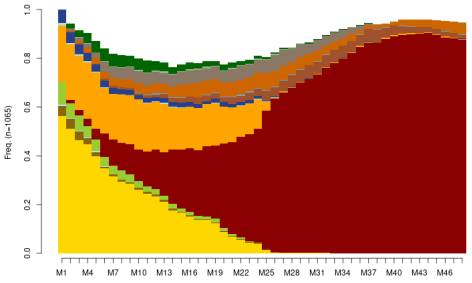
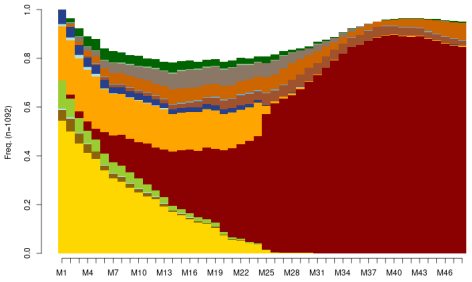
Prädiktion

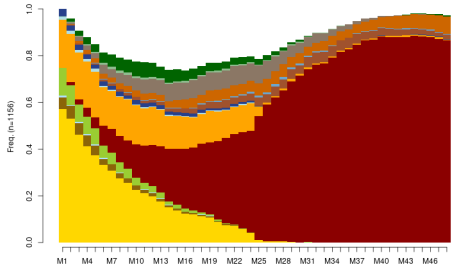
State Distribution Plot



Verlaufsindikatoren

VI1	11.1
VI2	9.4
VI3	1.1
VI4	1.3
VI5	4
VI6	30%
VI7	4
VI8	0
VI9	100%
VI10	33
VI11	2
VI12	6

2012		VI1 12.6 VI2 10.2 VI3 1.3 VI4 1.3 VI5 5 VI6 34% VI7 5 VI8 1 VI9 100% VI10 34 VI11 3 VI12 4		VI1 11.8 VI2 9.8 VI3 1.2 VI4 1.3 VI5 4 VI6 28% VI7 3 VI8 0 VI9 100% VI10 33 VI11 3 VI12 6
2013		VI1 12.8 VI2 10.4 VI3 1.3 VI4 1.4 VI5 5 VI6 17% VI7 2 VI8 0 VI9 100% VI10 35 VI11 3 VI12 6		VI1 11.9 VI2 9.9 VI3 1.2 VI4 1.3 VI5 4 VI6 27% VI7 3 VI8 0 VI9 100% VI10 32 VI11 2 VI12 6
2014	In der neuen Clusterlösung für diese Kohorte findet sich kein entsprechendes Verlaufsmuster. Dafür sind neue Verlaufsmuster aufgetreten (siehe Cluster Neu 1, Neu 2, Neu 3 und Neu 4)			VI1 11.6 VI2 9.6 VI3 1.5 VI4 1.3 VI5 5 VI6 26% VI7 4 VI8 0 VI9 100% VI10 32 VI11 3 VI12 6

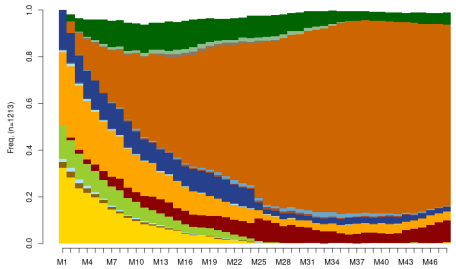
2015					
		VI1	11.2	VI1	11.6
		VI2	8.9	VI2	9.5
		VI3	1.3	VI3	1.4
		VI4	1.4	VI4	1.3
		VI5	5	VI5	4
		VI6	29%	VI6	29%
		VI7	4	VI7	4
		VI8	0	VI8	0
		VI9	100%	VI9	100%
		VI10	31	VI10	31
		VI11	2	VI11	3
		VI12	7	VI12	7

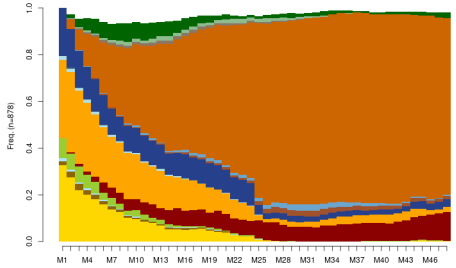
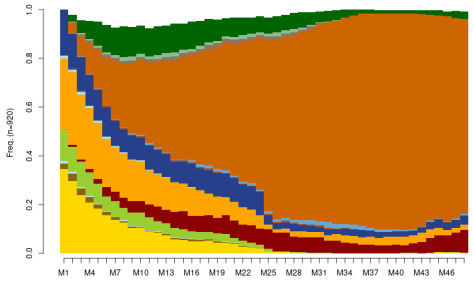
Anmerkung: Die Zuordnung des Clusters der neuen Lösung zum Clustern der Prädiktion bezieht sich auf eine visuelle Interpretation der state distribution plots und bezieht die Resultate der Jaccard-Matrix nicht mit ein.

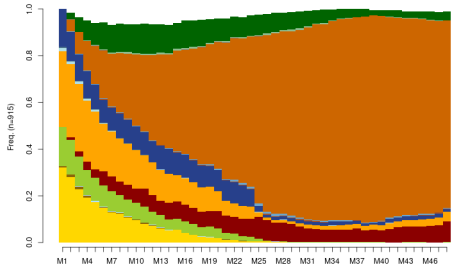
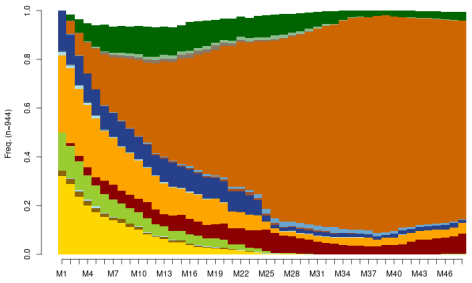
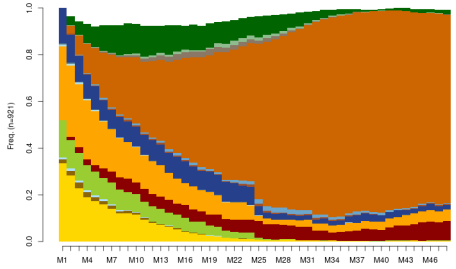
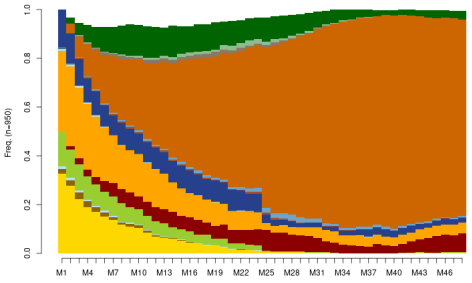
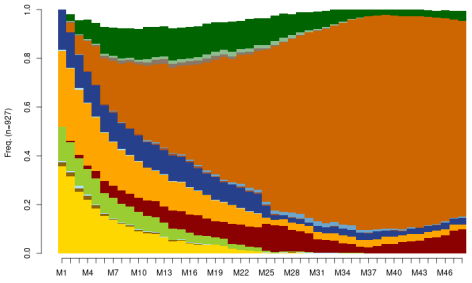
Quelle: BFS - SHIVALV-IK 2010-2019

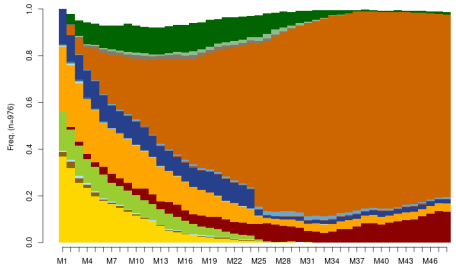
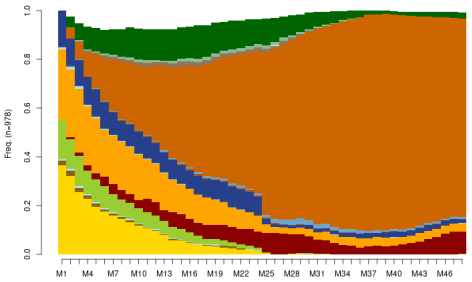
7.9 Cluster 6 - IV-Rente und Erwerb, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015

Tabelle A 7: Cluster 6 - IV-Rente und Erwerb, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015

Ko-horte	Referenz			
	State Distribution Plot		Verlaufsindikatoren	
2010			VI1	12.0
			VI2	8.4
			VI3	4.7
			VI4	1.6
			VI5	36
			VI6	16%
			VI7	2
			VI8	1
			VI9	100%
			VI10	38
			VI11	31
			VI12	1

	Neue Clusterlösung		Prädiktion	
	State Distribution Plot	Verlaufsindikatoren	State Distribution Plot	Verlaufsindikatoren
2011		VI1 11.9 VI2 8.3 VI3 4.1 VI4 1.7 VI5 35 VI6 16% VI7 2 VI8 1 VI9 100% VI10 41 VI11 32 VI12 2		VI1 11.8 VI2 8.0 VI3 4.6 VI4 1.7 VI5 36 VI6 15% VI7 2 VI8 1 VI9 100% VI10 39 VI11 32 VI12 1

2012		VI1 11.2 VI2 8.3 VI3 4.5 VI4 1.6 VI5 37 VI6 10% VI7 1 VI8 0 VI9 100% VI10 39 VI11 31 VI12 1		VI1 11.3 VI2 8.1 VI3 4.2 VI4 1.6 VI5 37 VI6 14% VI7 2 VI8 1 VI9 100% VI10 39 VI11 31 VI12 1
2013		VI1 11.8 VI2 7.8 VI3 4.6 VI4 1.8 VI5 36 VI6 16% VI7 2 VI8 1 VI9 100% VI10 38 VI11 31 VI12 2		VI1 11.8 VI2 7.8 VI3 4.3 VI4 1.8 VI5 36 VI6 15% VI7 2 VI8 1 VI9 100% VI10 38 VI11 31 VI12 2
2014	In der neuen Clusterlösung für diese Kohorte findet sich kein entsprechendes Verlaufsmuster. Dafür sind neue Verlaufsmuster aufgetreten (siehe Cluster Neu 1, Neu 2, Neu 3 und Neu 4)			VI1 11.7 VI2 8.0 VI3 4.7 VI4 1.7 VI5 36 VI6 13% VI7 1 VI8 1 VI9 100% VI10 38 VI11 30 VI12 2

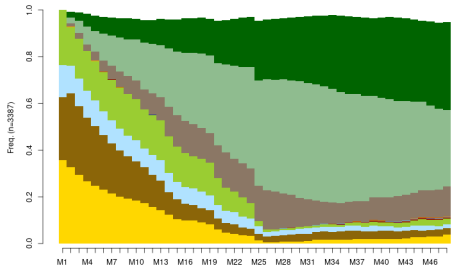
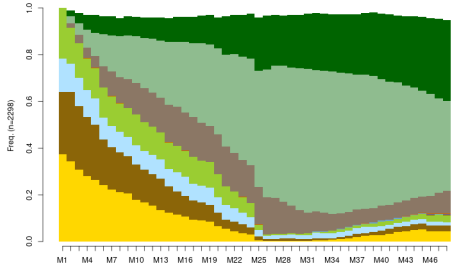
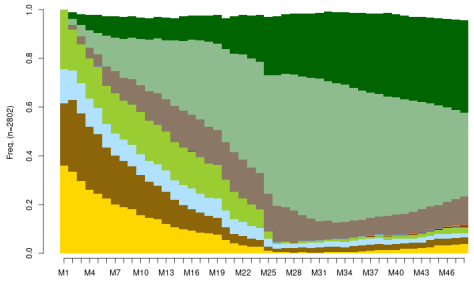
2015		VI1	12.0		VI1	11.8
		VI2	8.6		VI2	8.3
		VI3	4.6		VI3	4.1
		VI4	1.6		VI4	1.7
		VI5	36		VI5	36
		VI6	15%		VI6	14%
		VI7	2		VI7	2
		VI8	1		VI8	1
		VI9	100%		VI9	100%
		VI10	38		VI10	38
		VI11	31		VI11	30
		VI12	2		VI12	2

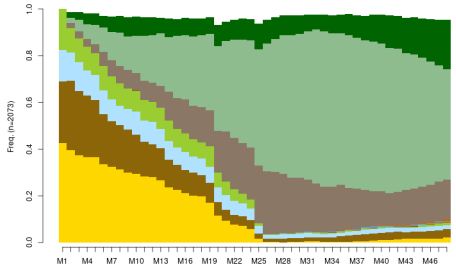
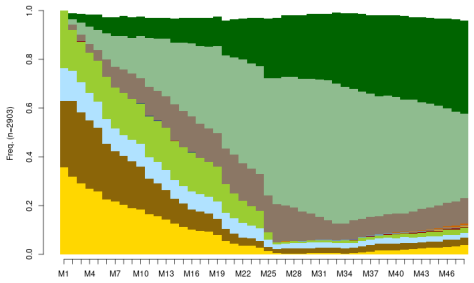
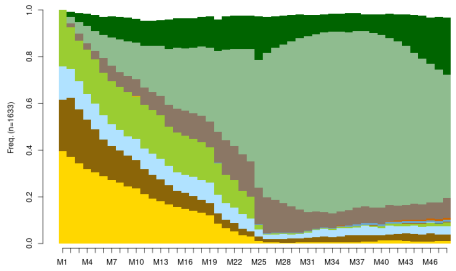
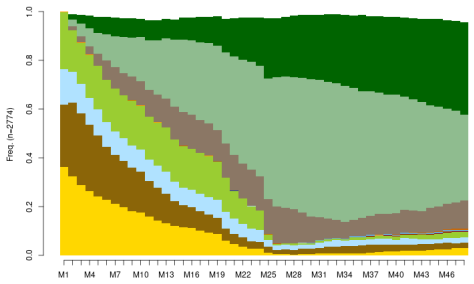
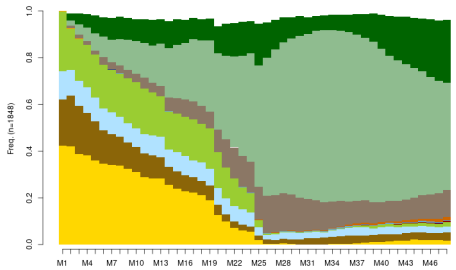
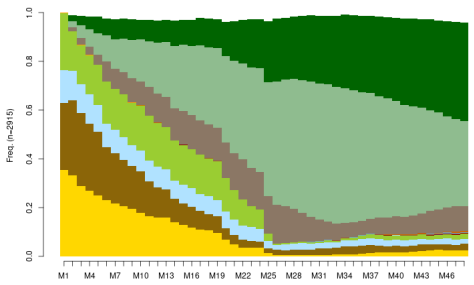
Anmerkung: Die Zuordnung des Clusters der neuen Lösung zum Clustern der Prädiktion bzw. eine fehlende Zuordnung bezieht sich auf eine visuelle Interpretation der state distribution plots und bezieht die Resultate der Jaccard-Matrix nicht mit ein.

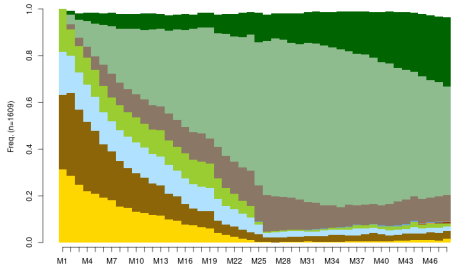
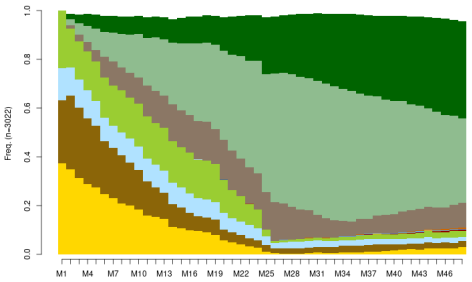
Quelle: BFS - SHIVALV-IK 2010-2019

7.10 Cluster 7 - Sozialhilfe und Erwerb, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015

Tabelle A 8: Cluster 7 - Sozialhilfe und Erwerb, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015

Ko-horte	Referenz			
	State Distribution Plot		Verlaufsindikatoren	
2010			VI1	14.9
			VI2	9.9
			VI3	6.5
			VI4	1.9
			VI5	33
			VI6	100%
			VI7	28
			VI8	18
			VI9	1%
			VI10	0
			VI11	0
			VI12	2
	Neue Clusterlösung		**Prädiktion**	
	State Distribution Plot	Verlaufsindikatoren	State Distribution Plot	Verlaufsindikatoren
2011		VI1 13.8 VI2 9.3 VI3 5.7 VI4 1.9 VI5 33 VI6 100% VI7 30 VI8 21 VI9 1% VI10 0 VI11 0 VI12 2		VI1 14.4 VI2 9.7 VI3 6.5 VI4 1.9 VI5 34 VI6 100% VI7 29 VI8 20 VI9 1% VI10 0 VI11 0 VI12 1

2012		VI1 15.0 VI2 10.6 VI3 4.7 VI4 1.8 VI5 29 VI6 100% VI7 34 VI8 22 VI9 1% VI10 0 VI11 0 VI12 2		VI1 14.7 VI2 10.1 VI3 6.5 VI4 1.9 VI5 34 VI6 100% VI7 29 VI8 20 VI9 2% VI10 0 VI11 0 VI12 1
2013		VI1 15.2 VI2 10.5 VI3 7.1 VI4 1.9 VI5 34 VI6 100% VI7 32 VI8 25 VI9 2% VI10 0 VI11 0 VI12 1		VI1 14.9 VI2 10.1 VI3 6.8 VI4 1.9 VI5 34 VI6 100% VI7 29 VI8 20 VI9 2% VI10 0 VI11 0 VI12 1
2014		VI1 16.6 VI2 11.2 VI3 7.0 VI4 2.0 VI5 33 VI6 100% VI7 30 VI8 23 VI9 2% VI10 0 VI11 0 VI12 1		VI1 15.0 VI2 10.0 VI3 6.7 VI4 1.9 VI5 34 VI6 100% VI7 28 VI8 20 VI9 2% VI10 0 VI11 0 VI12 1

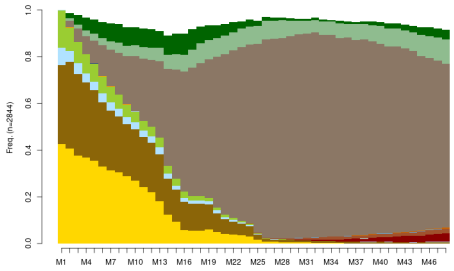
2015		VI1	13.7		VI1	14.9
		VI2	9.3		VI2	10.2
		VI3	6.2		VI3	6.8
		VI4	1.9		VI4	1.9
		VI5	35		VI5	34
		VI6	100%		VI6	100%
		VI7	35		VI7	29
		VI8	26		VI8	20
		VI9	1%		VI9	1%
		VI10	0		VI10	0
		VI11	0		VI11	0
		VI12	1		VI12	1

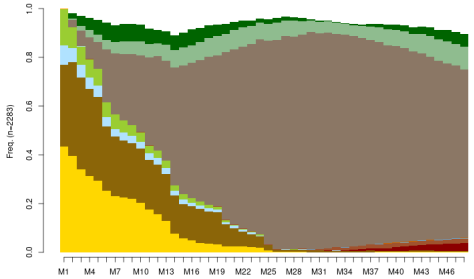
Anmerkung: Die Zuordnung des Clusters der neuen Lösung zum Clustern der Prädiktion bezieht sich auf eine visuelle Interpretation der state distribution plots und bezieht die Resultate der Jaccard-Matrix nicht mit ein.

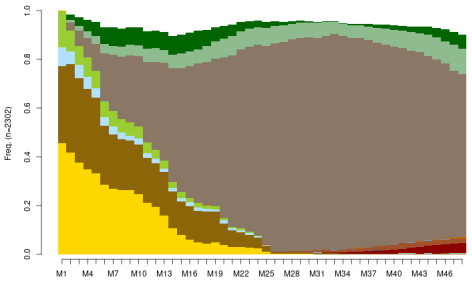
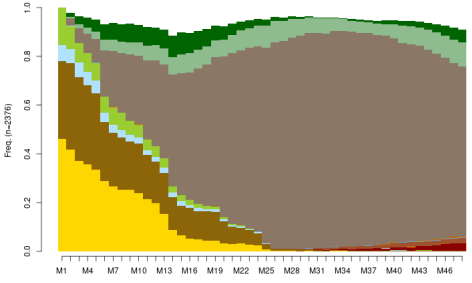
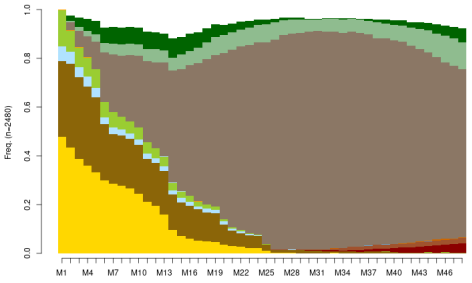
Quelle: BFS - SHIVALV-IK 2010-2019

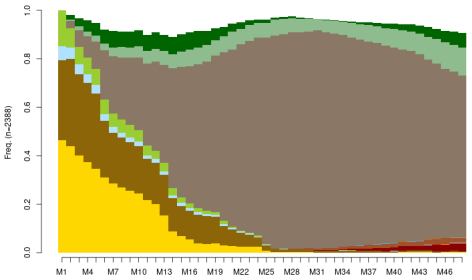
7.11 Cluster 8 - Sozialhilfe wiederholt, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015

Tabelle A 9: Cluster 8 - Sozialhilfe wiederholt, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015

Ko-horte	Referenz	
	State Distribution Plot	Verlaufsindikatoren
2010		VI1 11.3 VI2 8.8 VI3 1.8 VI4 1.5 VI5 7 VI6 100% VI7 37 VI8 4 VI9 7% VI10 1 VI11 0 VI12 3

	Neue Clusterlösung		Prädiktion	
	State Distribution Plot	Verlaufsindikatoren	State Distribution Plot	Verlaufsindikatoren
2011	In der neuen Clusterlösung für diese Kohorte findet sich kein entsprechendes Verlaufsmuster. Dafür sind neue Verlaufsmuster aufgetreten (siehe Cluster Neu 1, Neu 2, Neu 3 und Neu 4)			VI1 10.3 VI2 8.0 VI3 1.6 VI4 1.4 VI5 6 VI6 100% VI7 38 VI8 4 VI9 6% VI10 1 VI11 0 VI12 3

2012	In der neuen Clusterlösung für diese Kohorte findet sich kein entsprechendes Verlaufsmuster. Dafür sind neue Verlaufsmuster aufgetreten (siehe Cluster Neu 1, Neu 2, Neu 3 und Neu 4)		VI1 10.5 VI2 8.2 VI3 1.6 VI4 1.4 VI5 6 VI6 100% VI7 38 VI8 4 VI9 7% VI10 1 VI11 0 VI12 3
2013	In der neuen Clusterlösung für diese Kohorte findet sich kein entsprechendes Verlaufsmuster. Dafür sind neue Verlaufsmuster aufgetreten (siehe Cluster Neu 1, Neu 2, Neu 3 und Neu 4)		VI1 10.4 VI2 8.3 VI3 1.7 VI4 1.4 VI5 7 VI6 100% VI7 38 VI8 4 VI9 6% VI10 1 VI11 0 VI12 3
2014	In der neuen Clusterlösung für diese Kohorte findet sich kein entsprechendes Verlaufsmuster. Dafür sind neue Verlaufsmuster aufgetreten (siehe Cluster Neu 1, Neu 2, Neu 3 und Neu 4)		VI1 10.5 VI2 8.2 VI3 1.6 VI4 1.4 VI5 6 VI6 100% VI7 38 VI8 3 VI9 7% VI10 1 VI11 0 VI12 3

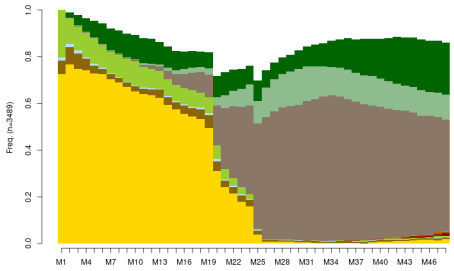
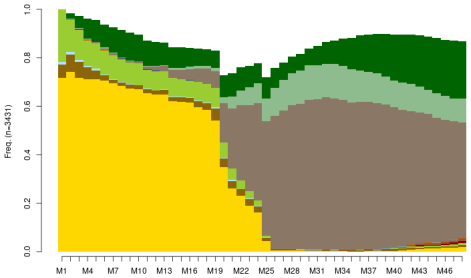
2015	In der neuen Clusterlösung für diese Kohorte findet sich kein entsprechendes Verlaufsmuster. Dafür sind neue Verlaufsmuster aufgetreten (siehe Cluster Neu 1, Neu 2, Neu 3 und Neu 4)		VI1 10.3 VI2 8.2 VI3 1.4 VI4 1.4 VI5 6 VI6 100% VI7 38 VI8 3 VI9 6% VI10 1 VI11 0 VI12 3

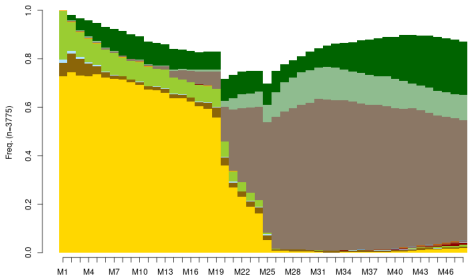
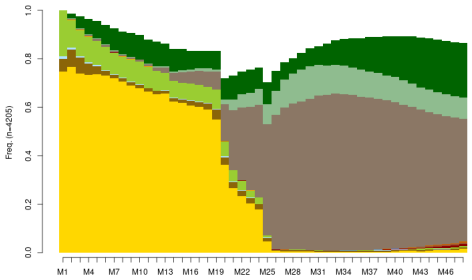
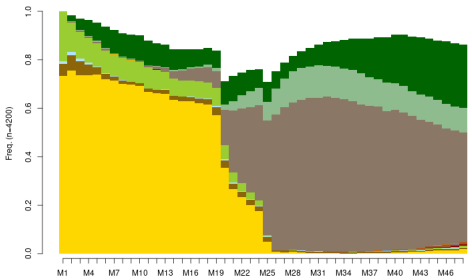
Anmerkung: Die fehlende Zuordnung eines Clusters der neuen Lösung zum Clustern der Prädiktion bezieht sich auf eine visuelle Interpretation der state distribution plots und zieht die Resultate der Jaccard-Matrix nicht mit ein.

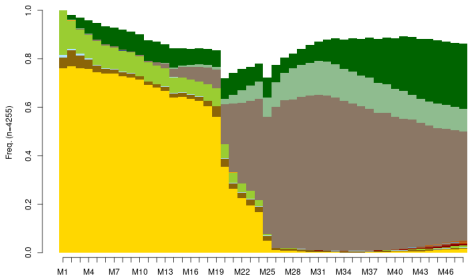
Quelle: BFS - SHIVALV-IK 2010-2019

7.12 Cluster 9 - Sozialhilfe neu, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015

Tabelle A 10: Cluster 9 - Sozialhilfe neu, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015

<table>
<tr><th rowspan="2">Ko-horte</th><th colspan="4">Referenz</th></tr>
<tr><th colspan="2">State Distribution Plot</th><th colspan="2">Verlaufsindikatoren</th></tr>
<tr><td>2010</td><td colspan="2">  </td><td colspan="2"><table>
<tr><td>VI1</td><td>16.9</td></tr>
<tr><td>VI2</td><td>13.2</td></tr>
<tr><td>VI3</td><td>2.2</td></tr>
<tr><td>VI4</td><td>1.6</td></tr>
<tr><td>VI5</td><td>11</td></tr>
<tr><td>VI6</td><td>100%</td></tr>
<tr><td>VI7</td><td>20</td></tr>
<tr><td>VI8</td><td>4</td></tr>
<tr><td>VI9</td><td>2%</td></tr>
<tr><td>VI10</td><td>0</td></tr>
<tr><td>VI11</td><td>0</td></tr>
<tr><td>VI12</td><td>7</td></tr>
</table></td></tr>
<tr><th rowspan="2"></th><th colspan="2">Neue Clusterlösung</th><th colspan="2">Prädiktion</th></tr>
<tr><th>State Distribution Plot</th><th>Verlaufsindikatoren</th><th>State Distribution Plot</th><th>Verlaufsindikatoren</th></tr>
<tr><td>2011</td><td>In der neuen Clusterlösung für diese Kohorte findet sich kein entsprechendes Verlaufsmuster. Dafür sind neue Verlaufs-muster aufgetreten (siehe Cluster Neu 1, Neu 2, Neu 3 und Neu 4)</td><td></td><td>  </td><td><table>
<tr><td>VI1</td><td>17.0</td></tr>
<tr><td>VI2</td><td>13.1</td></tr>
<tr><td>VI3</td><td>2.2</td></tr>
<tr><td>VI4</td><td>1.6</td></tr>
<tr><td>VI5</td><td>11</td></tr>
<tr><td>VI6</td><td>100%</td></tr>
<tr><td>VI7</td><td>20</td></tr>
<tr><td>VI8</td><td>3</td></tr>
<tr><td>VI9</td><td>2%</td></tr>
<tr><td>VI10</td><td>0</td></tr>
<tr><td>VI11</td><td>0</td></tr>
<tr><td>VI12</td><td>7</td></tr>
</table></td></tr>
</table>

2012	In der neuen Clusterlösung für diese Kohorte findet sich kein entsprechendes Verlaufsmuster. Dafür sind neue Verlaufsmuster aufgetreten (siehe Cluster Neu 1, Neu 2, Neu 3 und Neu 4)		VI1 17.1 VI2 13.4 VI3 2.0 VI4 1.6 VI5 10 VI6 100% VI7 20 VI8 3 VI9 2% VI10 0 VI11 0 VI12 7
2013	In der neuen Clusterlösung für diese Kohorte findet sich kein entsprechendes Verlaufsmuster. Dafür sind neue Verlaufsmuster aufgetreten (siehe Cluster Neu 1, Neu 2, Neu 3 und Neu 4)		VI1 17.1 VI2 13.3 VI3 2.1 VI4 1.6 VI5 10 VI6 100% VI7 20 VI8 3 VI9 2% VI10 0 VI11 0 VI12 7
2014	In der neuen Clusterlösung für diese Kohorte findet sich kein entsprechendes Verlaufsmuster. Dafür sind neue Verlaufsmuster aufgetreten (siehe Cluster Neu 1, Neu 2, Neu 3 und Neu 4)		VI1 17.2 VI2 13.4 VI3 2.1 VI4 1.6 VI5 11 VI6 100% VI7 19 VI8 3 VI9 2% VI10 0 VI11 0 VI12 7

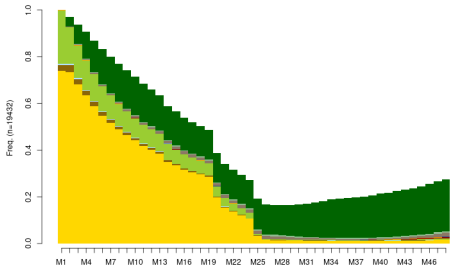
2015	In der neuen Clusterlösung für diese Kohorte findet sich kein entsprechendes Verlaufsmuster. Dafür sind neue Verlaufsmuster aufgetreten (siehe Cluster Neu 1, Neu 2, Neu 3 und Neu 4)		VI1 17.3 VI2 13.7 VI3 2.0 VI4 1.6 VI5 11 VI6 100% VI7 19 VI8 3 VI9 2% VI10 0 VI11 0 VI12 6

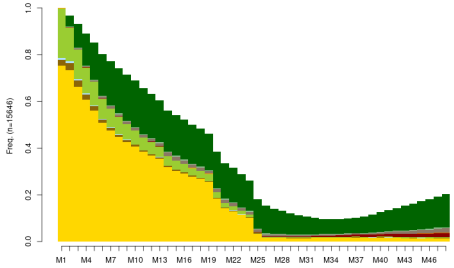
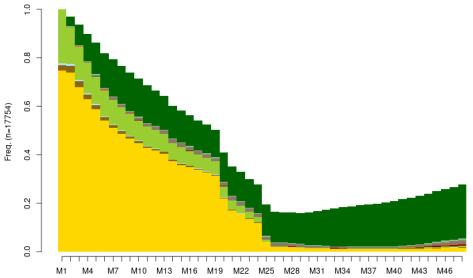
Anmerkung: Die fehlende Zuordnung eines Clusters der neuen Lösung zum Clustern der Prädiktion bezieht sich auf eine visuelle Interpretation der state distribution plots und bezieht die Resultate der Jaccard-Matrix nicht mit ein.

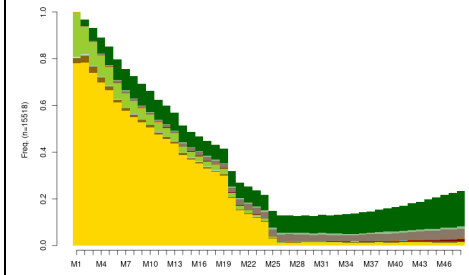
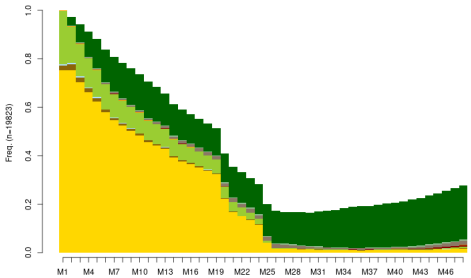
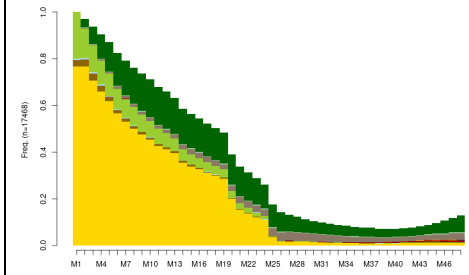
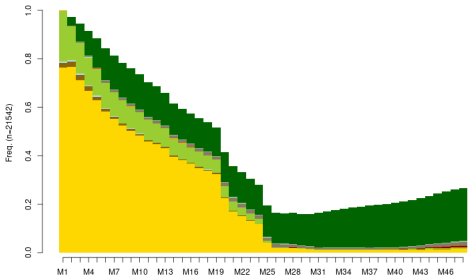
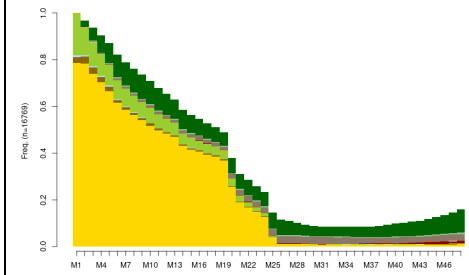
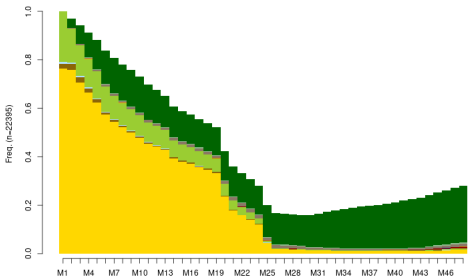
Quelle: BFS - SHIVALV-IK 2010-2019

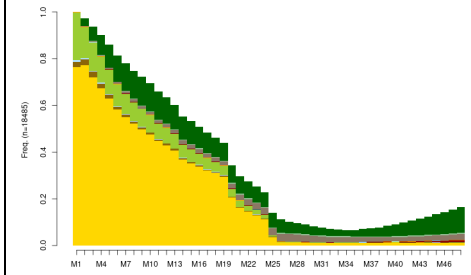
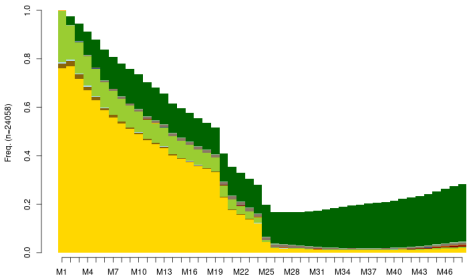
7.13 Cluster 10 - Leavers, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015

Tabelle A 11: Cluster 10 - Leavers, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015

Ko-horte	Referenz			
	State Distribution Plot	Verlaufsindikatoren		
2010		VI1	12.4	
		VI2	9.9	
		VI3	2.1	
		VI4	1.4	
		VI5	10	
		VI6	12%	
		VI7	1	
		VI8	0	
		VI9	1%	
		VI10	0	
		VI11	0	
		VI12	28	

	Neue Clusterlösung		Prädiktion	
	State Distribution Plot	Verlaufsindikatoren	State Distribution Plot	Verlaufsindikatoren
2011		VI1 11.4 VI2 8.9 VI3 1.6 VI4 1.4 VI5 8 VI6 12% VI7 1 VI8 0 VI9 2% VI10 0 VI11 0 VI12 30		VI1 12.7 VI2 10.0 VI3 2.1 VI4 1.5 VI5 9 VI6 13% VI7 1 VI8 0 VI9 1% VI10 0 VI11 0 VI12 28

2012		VI1 12.5 VI2 10.2 VI3 1.3 VI4 1.4 VI5 6 VI6 17% VI7 2 VI8 0 VI9 2% VI10 0 VI11 0 VI12 30		VI1 13.0 VI2 10.4 VI3 2.1 VI4 1.4 VI5 9 VI6 12% VI7 1 VI8 0 VI9 1% VI10 0 VI11 0 VI12 27
2013		VI1 12.1 VI2 9.9 VI3 1.7 VI4 1.4 VI5 6 VI6 15% VI7 2 VI8 0 VI9 1% VI10 0 VI11 0 VI12 30		VI1 13.1 VI2 10.5 VI3 2.1 VI4 1.4 VI5 9 VI6 11% VI7 1 VI8 0 VI9 1% VI10 0 VI11 0 VI12 27
2014		VI1 13.3 VI2 10.9 VI3 1.5 VI4 1.4 VI5 5 VI6 15% VI7 1 VI8 0 VI9 2% VI10 0 VI11 0 VI12 30		VI1 13.1 VI2 10.4 VI3 2.1 VI4 1.5 VI5 9 VI6 12% VI7 1 VI8 0 VI9 1% VI10 0 VI11 0 VI12 27

2015					
		VI1	12.6		VI1	13.2
		VI2	10.2		VI2	10.6
		VI3	1.8		VI3	2.1
		VI4	1.4		VI4	1.5
		VI5	5		VI5	9
		VI6	14%		VI6	10%
		VI7	1		VI7	1
		VI8	0		VI8	0
		VI9	1%		VI9	1%
		VI10	0		VI10	0
		VI11	0		VI11	0
		VI12	31		VI12	27

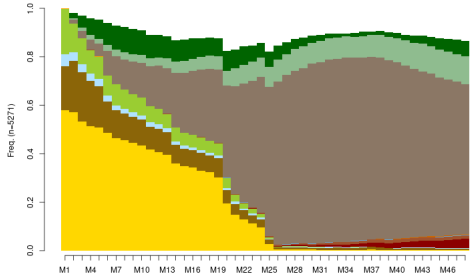
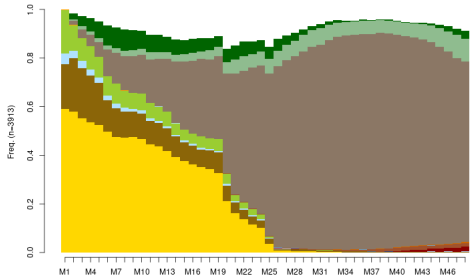
Anmerkung: Die Zuordnung des Clusters der neuen Lösung zum Clustern der Prädiktion bezieht sich auf eine visuelle Interpretation der state distribution plots und bezieht die Resultate der Jaccard-Matrix nicht mit ein.

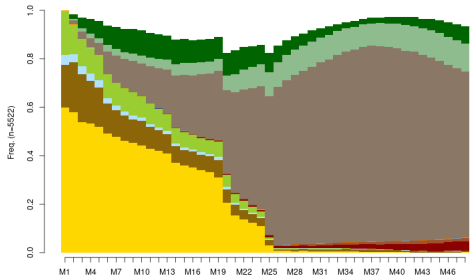
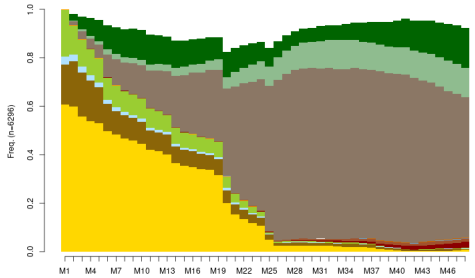
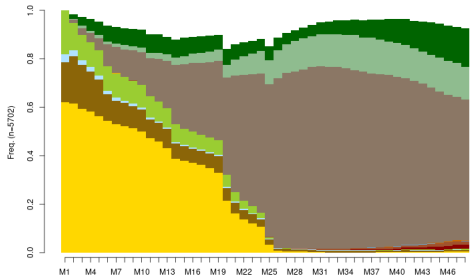
Quelle: BFS - SHIVALV-IK 2010-2019

7.14 Cluster Neu 1, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015

Tabelle A 12: Cluster Neu 1, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015

Kohorte	Referenz	
	State Distribution Plot	Verlaufsindikatoren
2010	Für dieses neue Cluster gibt es keine Entsprechung in der Referenz	

Kohorte	Neue Clusterlösung		Prädiktion	
	State Distribution Plot	Verlaufsindikatoren	State Distribution Plot	Verlaufsindikatoren
2011		VI1 13.8 VI2 10.6 VI3 2.1 VI4 1.5 VI5 8 VI6 100% VI7 29 VI8 4 VI9 6% VI10 1 VI11 0 VI12 5	Für dieses neue Cluster gibt es keine Entsprechung in der Referenz bzw. in deren Prädiktion	
2012		VI1 13.9 VI2 11.0 VI3 1.9 VI4 1.5 VI5 6 VI6 100% VI7 31 VI8 3 VI9 4% VI10 0 VI11 0 VI12 4	Für dieses neue Cluster gibt es keine Entsprechung in der Referenz bzw. in deren Prädiktion	

2013		VI1	14.1	Für dieses neue Cluster gibt es keine Entsprechung in der Referenz bzw. in deren Prädiktion
		VI2	10.9	
		VI3	2.3	
		VI4	1.5	
		VI5	9	
		VI6	100%	
		VI7	29	
		VI8	4	
		VI9	6%	
		VI10	1	
		VI11	0	
		VI12	4	
2014		VI1	14.2	Für dieses neue Cluster gibt es keine Entsprechung in der Referenz bzw. in deren Prädiktion
		VI2	10.7	
		VI3	2.2	
		VI4	1.6	
		VI5	10	
		VI6	100%	
		VI7	28	
		VI8	4	
		VI9	5%	
		VI10	1	
		VI11	0	
		VI12	4	
2015		VI1	14.5	Für dieses neue Cluster gibt es keine Entsprechung in der Referenz bzw. in deren Prädiktion
		VI2	11.4	
		VI3	2.2	
		VI4	1.5	
		VI5	10	
		VI6	100%	
		VI7	29	
		VI8	4	
		VI9	4%	
		VI10	0	
		VI11	0	
		VI12	4	

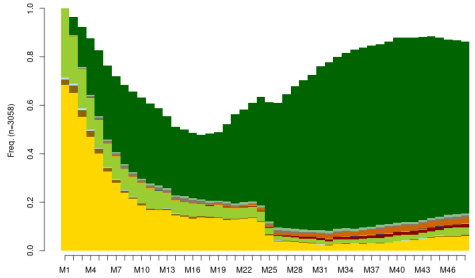
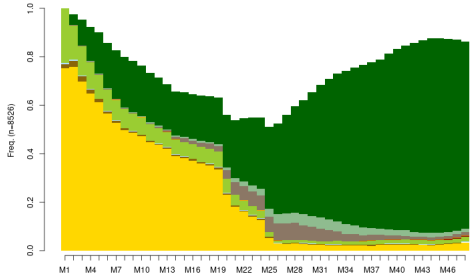
Anmerkung: Die fehlende Zuordnung des Clusters der neuen Lösung zu einem Cluster der Prädiktion bezieht sich auf eine visuelle Interpretation der state distribution plots und bezieht die Resultate der Jaccard-Matrix nicht mit ein.

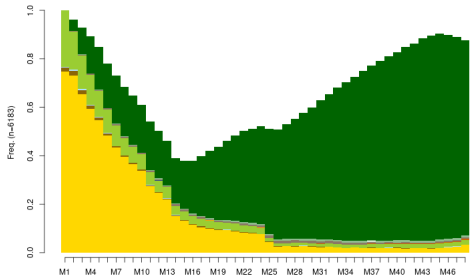
Quelle: BFS - SHIVALV-IK 2010-2019

7.15 Cluster Neu 2, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015

Tabelle A 13: Cluster Neu 2, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015

Kohorte	Referenz	
	State Distribution Plot	Verlaufsindikatoren
2010	Für dieses neue Cluster gibt es keine Entsprechung in der Referenz	

Kohorte	Neue Clusterlösung		Prädiktion	
	State Distribution Plot	Verlaufsindikatoren	State Distribution Plot	Verlaufsindikatoren
2011		VI1 9.9 VI2 5.6 VI3 2.7 VI4 1.9 VI5 27 VI6 16% VI7 1 VI8 0 VI9 4% VI10 1 VI11 0 VI12 13	Für dieses neue Cluster gibt es keine Entsprechung in der Referenz bzw. in deren Prädiktion	
2012	Dieses neue Cluster taucht in der Kohorte 2012 nicht auf		-	
2013		VI1 13.5 VI2 9.7 VI3 2.3 VI4 1.7 VI5 23 VI6 24% VI7 2 VI8 1 VI9 1% VI10 0 VI11 0 VI12 12	Für dieses neue Cluster gibt es keine Entsprechung in der Referenz bzw. in deren Prädiktion	

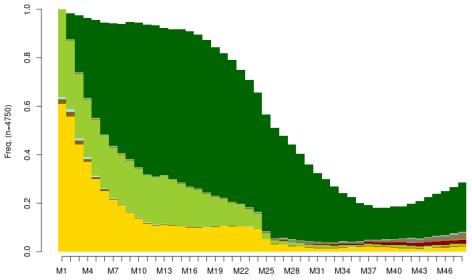
<table>
<tr>
<td>2014</td>
<td> 
</td>
<td>
<table>
<tr><td>VI1</td><td>9.9</td></tr>
<tr><td>VI2</td><td>6.9</td></tr>
<tr><td>VI3</td><td>2.0</td></tr>
<tr><td>VI4</td><td>1.7</td></tr>
<tr><td>VI5</td><td>24</td></tr>
<tr><td>VI6</td><td>12%</td></tr>
<tr><td>VI7</td><td>1</td></tr>
<tr><td>VI8</td><td>0</td></tr>
<tr><td>VI9</td><td>1%</td></tr>
<tr><td>VI10</td><td>0</td></tr>
<tr><td>VI11</td><td>0</td></tr>
<tr><td>VI12</td><td>16</td></tr>
</table>
</td>
<td>Für dieses neue Cluster gibt es keine Entsprechung in der Referenz bzw. in deren Prädiktion</td>
</tr>
<tr>
<td>2015</td>
<td colspan="2">Dieses neue Cluster taucht in der Kohorte 2015 nicht auf</td>
<td>-</td>
</tr>
</table>

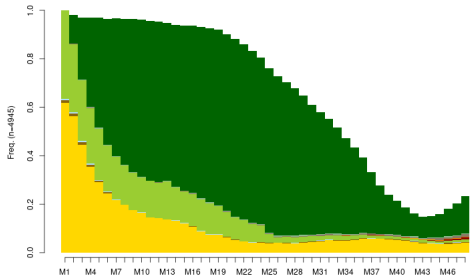
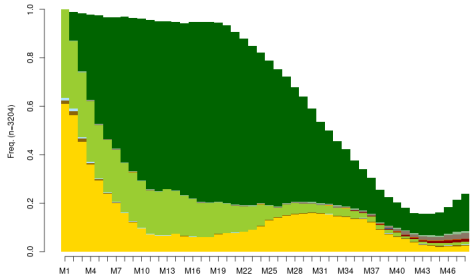
Anmerkung: Die fehlende Zuordnung des Clusters der neuen Lösung zu einem Cluster der Prädiktion bezieht sich auf eine visuelle Interpretation der state distribution plots und bezieht die Resultate der Jaccard-Matrix nicht mit ein.

Quelle: BFS - SHIVALV-IK 2010-2019

7.16 Cluster Neu 3, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015

Tabelle A 14: Cluster Neu 3, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015

<table>
<tr><th rowspan="2">Kohorte</th><th colspan="2">Referenz</th></tr>
<tr><td>State Distribution Plot</td><td>Verlaufsindikatoren</td></tr>
<tr><td>2010</td><td colspan="2">Für dieses neue Cluster gibt es keine Entsprechung in der Referenz</td></tr>
<tr><th rowspan="2">Kohorte</th><th colspan="2">Neue Clusterlösung</th><th colspan="2">Prädiktion</th></tr>
<tr><td>State Distribution Plot</td><td>Verlaufsindikatoren</td><td>State Distribution Plot</td><td>Verlaufsindikatoren</td></tr>
<tr><td>2011</td><td colspan="2">Dieses neue Cluster taucht in der Kohorte 2011 nicht auf</td><td colspan="2">-</td></tr>
<tr><td>2012</td><td> 
</td><td>
<table>
<tr><td>VI1</td><td>10.0</td></tr>
<tr><td>VI2</td><td>6.7</td></tr>
<tr><td>VI3</td><td>4.7</td></tr>
<tr><td>VI4</td><td>1.7</td></tr>
<tr><td>VI5</td><td>23</td></tr>
<tr><td>VI6</td><td>12%</td></tr>
<tr><td>VI7</td><td>1</td></tr>
<tr><td>VI8</td><td>0</td></tr>
<tr><td>VI9</td><td>2%</td></tr>
<tr><td>VI10</td><td>0</td></tr>
<tr><td>VI11</td><td>0</td></tr>
<tr><td>VI12</td><td>20</td></tr>
</table>
</td><td colspan="2">Für dieses neue Cluster gibt es keine Entsprechung in der Referenz bzw. in deren Prädiktion</td></tr>
<tr><td>2013</td><td colspan="2">Dieses neue Cluster taucht in der Kohorte 2013 nicht auf</td><td colspan="2">-</td></tr>
</table>

2014		VI1	10.0	Für dieses neue Cluster gibt es keine Entsprechung in der Referenz bzw. in deren Prädiktion
		VI2	6.3	
		VI3	4.2	
		VI4	1.8	
		VI5	26	
		VI6	9%	
		VI7	0	
		VI8	0	
		VI9	1%	
		VI10	0	
		VI11	0	
		VI12	16	
2015		VI1	11.6	Für dieses neue Cluster gibt es keine Entsprechung in der Referenz bzw. in deren Prädiktion
		VI2	6.3	
		VI3	5.0	
		VI4	2.0	
		VI5	25	
		VI6	11%	
		VI7	1	
		VI8	0	
		VI9	1%	
		VI10	0	
		VI11	0	
		VI12	16	

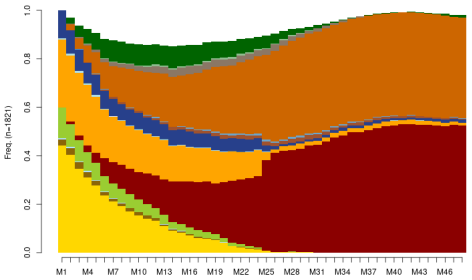
Anmerkung: Die fehlende Zuordnung des Clusters der neuen Lösung zu einem Cluster der Prädiktion bezieht sich auf eine visuelle Interpretation der state distribution plots und bezieht die Resultate der Jaccard-Matrix nicht mit ein.

Quelle: BFS - SHIVALV-IK 2010-2019

7.17 Cluster Neu 4, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015

Tabelle A 15: Cluster Neu 4, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015

Kohorte	Referenz	
	State Distribution Plot	Verlaufsindikatoren
2010	Für dieses neue Cluster gibt es keine Entsprechung in der Referenz	

Kohorte	Neue Clusterlösung		Prädiktion	
	State Distribution Plot	Verlaufsindikatoren	State Distribution Plot	Verlaufsindikatoren
2010	Dieses neue Cluster taucht in der Kohorte 2010 nicht auf		-	
2011	Dieses neue Cluster taucht in der Kohorte 2011 nicht auf		-	
2012	Dieses neue Cluster taucht in der Kohorte 2012 nicht auf		-	
2013	Dieses neue Cluster taucht in der Kohorte 2013 nicht auf		-	
2014		VI1 11.6 VI2 8.8 VI3 3.0 VI4 1.5 VI5 20 VI6 18% VI7 2 VI8 0 VI9 100% VI10 37 VI11 16 VI12 4	Für dieses neue Cluster gibt es keine Entsprechung in der Referenz bzw. in deren Prädiktion	
2015	Dieses neue Cluster taucht in der Kohorte 2015 nicht auf		-	

Anmerkung: Die fehlende Zuordnung des Clusters der neuen Lösung zu einem Cluster der Prädiktion bezieht sich auf eine visuelle Interpretation der state distribution plots und bezieht die Resultate der Jaccard-Matrix nicht mit ein.

Quelle: BFS - SHIVALV-IK 2010-2019

7.18 Übersicht Training Prädiktionsmodelle

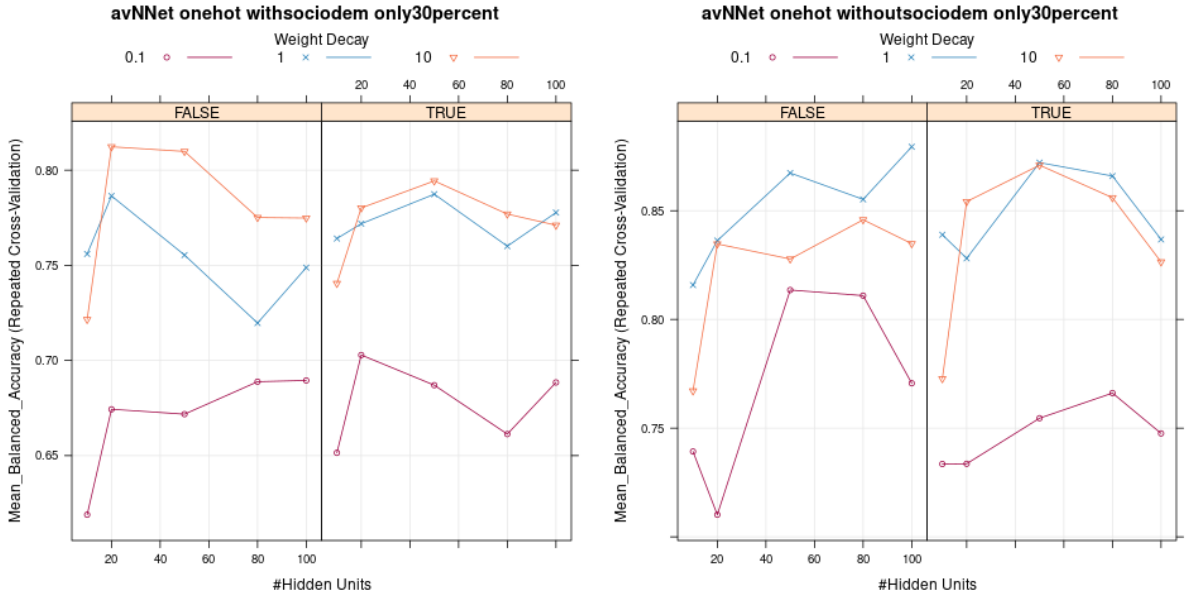
Tabelle A 16: Übersicht Training Prädiktionsmodelle

Algorithme	Format	Variables	Remarque et Tuning hyperpa-ramètres
RF	Factor	Withsociodem	Tuning: - num.trees=500 - mtry=[1,53] - min.node.size=1 - splitrule=«gini»
RF	Factor	Withoutsociodem	Tuning: - num.trees=500 - mtry=[1,53] - min.node.size=1 - splitrule=«gini»
RF	Mixed	Withsociodem	Tuning: - num.trees=500 - mtry=[1,53] - min.node.size=1 - splitrule=«gini»
RF	Mixed	Withoutsociodem	Tuning: - num.trees=500 - mtry=[1,53] - min.node.size=1 - splitrule=«gini»
RF	One-hot	Withsociodem	Tuning: - num.trees=500 - mtry=[1,53] - min.node.size=1 - splitrule=«gini»
RF	One-hot	Withoutsociodem	Tuning: - num.trees=500 - mtry=[1,53] - min.node.size=1 - splitrule=«gini»
GBM	Factor	Withsociodem	Tuning: - shrinkage=0.1 - n.trees={100,200,500,700} - interaction.depth=2 - n.minobsinnode=10
GBM	Mixed	Withsociodem	Tuning: - shrinkage={0.01,0.05,0.1} - n.trees={100,200,500,700} - interaction.depth=2 - n.minobsinnode=10
GBM	One-hot	Withsociodem	Tuning: - shrinkage={0.01,0.05,0.1} - n.trees={100,200,500,700} - interaction.depth=2 - n.minobsinnode=10
GBM	One-hot	Withoutsociodem	Tuning: - shrinkage={0.01,0.05} - n.trees={100,200,500,700} - interaction.depth=2

			- n.minobsinnode=10
SVM Poly	One-hot	Withsociodem	Aufgrund von Perfomanceeinschränkungen konnten nur 30% der Daten des Trainingssamples verwendet werden. Tuning: - degree={1,2,3} - C={0.001, 0.01, 0.1, 1, 10, 100, 1000} - scale=1
SVM Poly	One-hot	Withoutsociodem	Aufgrund von Perfomanceeinschränkungen konnten nur 30% der Daten des Trainingssamples verwendet werden. Tuning: - degree={1,2,3} - C={0.001, 0.01, 0.1, 1, 10, 100, 1000} - scale=1
KNN	One-hot	Withsociodem	Tuning: - K={1,2,3,4,5,8,10}
avNNet	One-hot	Withsociodem	Aufgrund von Perfomanceeinschränkungen konnten nur 30% der Daten des Trainingssamples verwendet werden. Tuning: - size={10,20,50,80,100} - decay={0.1,1,10} - bag={true,false}
avNNet	One-hot	Withoutsociodem	Aufgrund von Perfomanceeinschränkungen konnten nur 30% der Daten des Trainingssamples verwendet werden. Tuning: - size={10,20,50,80,100} - decay={0.1,1,10} - bag={true,false}

7.19 Trainingsergebnisse Prädiktionsmodelle

Abbildung A 1: avNNet - Ergebnisse Training Prädiktionsmodell

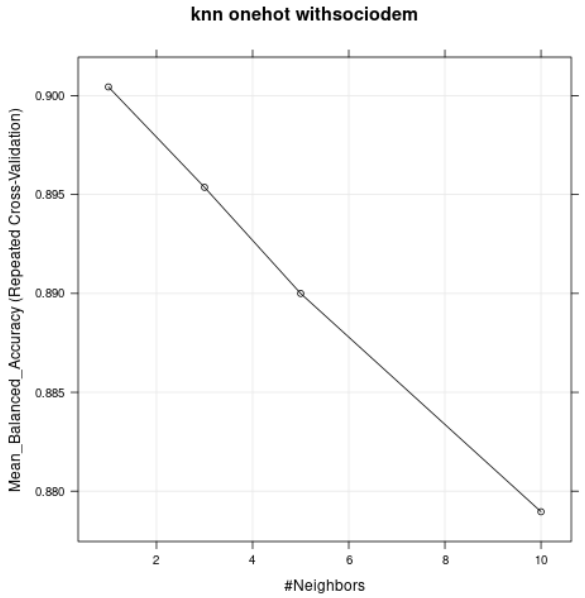


Quelle: BFS - SHIVALV-IK 2010-2014

Abbildung A 2: GBM - Ergebnisse Training Prädiktionsmodell

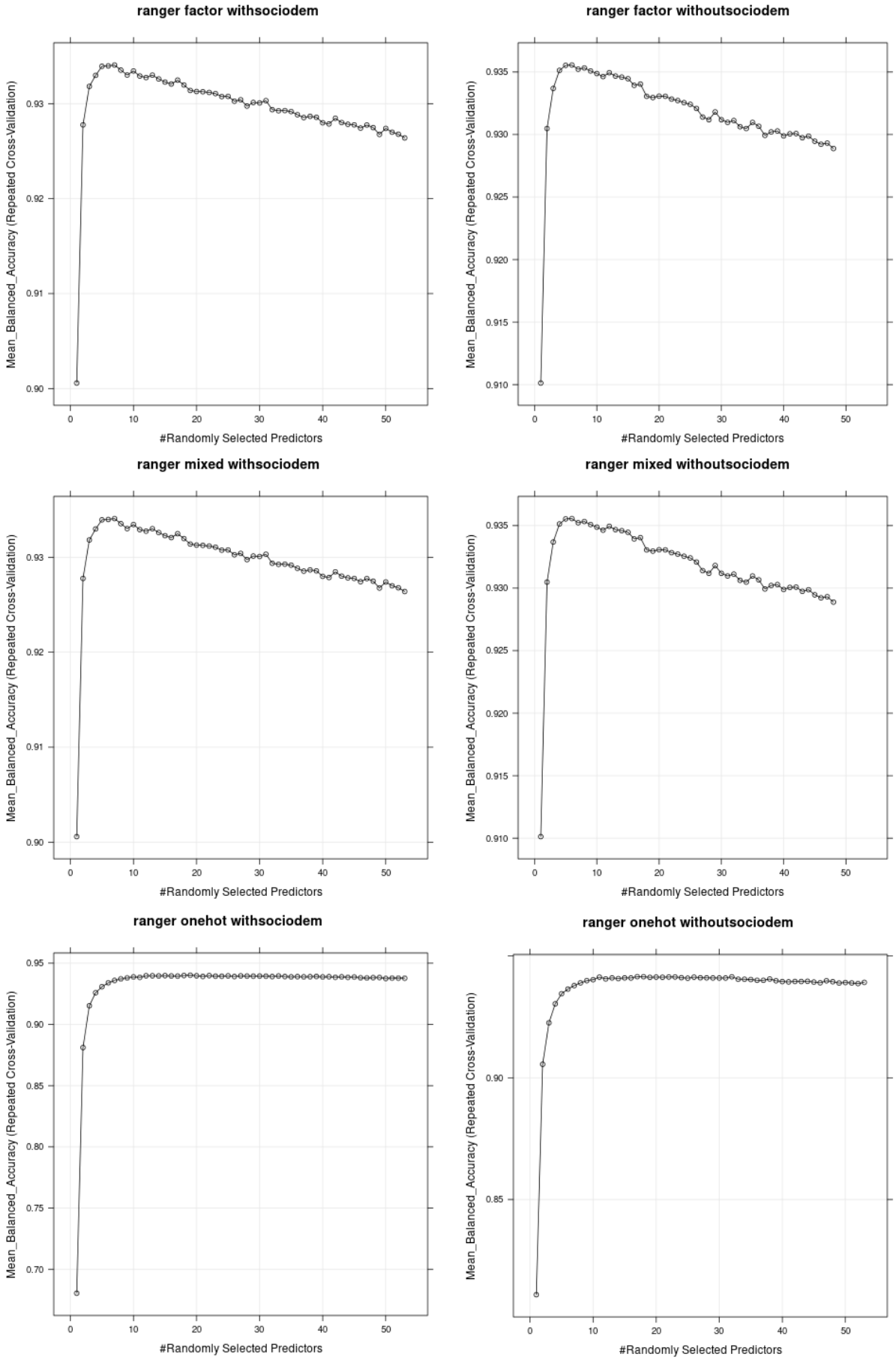
Quelle: BFS - SHIVALV-IK 2010-2014

Abbildung A 3: knn - Ergebnisse Training Prädiktionsmodell



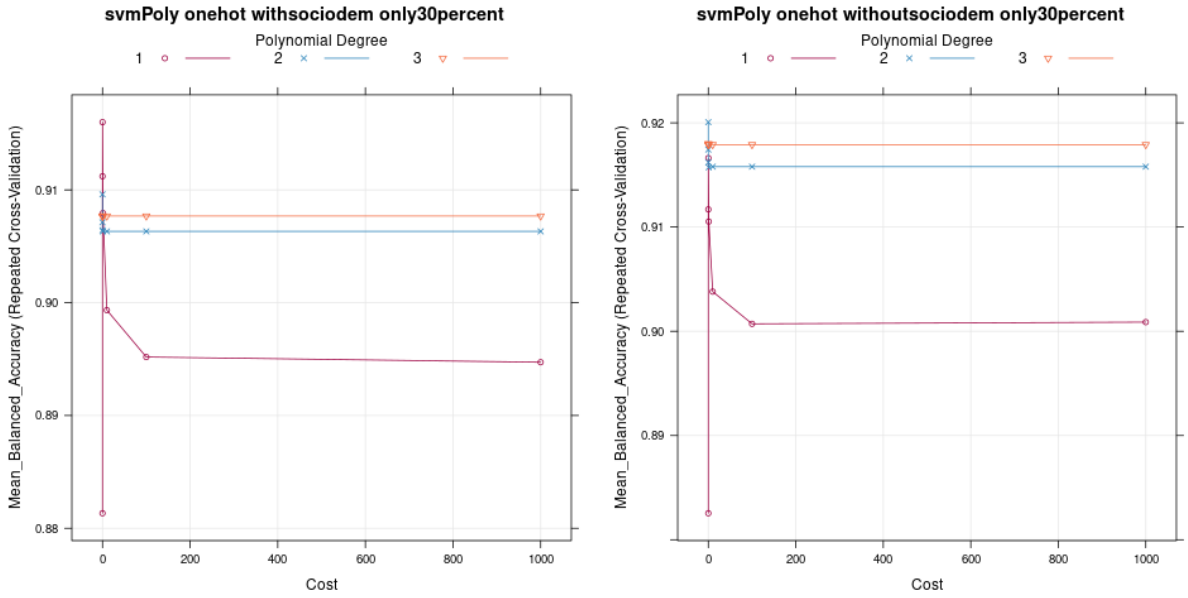
Quelle: BFS - SHIVALV-IK 2010-2014

Abbildung A 4: ranger - Ergebnisse Training Prädiktionsmodell



Quelle: BFS - SHIVALV-IK 2010-2014

Abbildung A 5: svmPoly - Ergebnisse Training Prädiktionsmodell



Quelle: BFS - SHIVALV-IK 2010-2014

7.20 Verlaufsindikatoren

Tabelle A 17: Referenz K2010- Verlaufsindikatoren

Verlaufsindikatoren 2010 *Referenz*	**ALV Kurzzeit**	**ALV Lang-zeit**	**Zwischen-verdienst**	**ALV Mehr-fach**	**IV-Rente**	**IV-Rente und Erwerb**	**Sozialhilfe und Erwerb**	**Sozialhilfe wiederholt**	**Sozialhilfe neu**	**Leavers**	**Kohorte ins-gesamt**
Anzahl Monate mit ALV	6.5	12.8	24.0	22.3	12.0	12.0	14.9	11.3	16.9	12.4	10.6
Dauer der ersten ALV-Bezugsperiode (Monate)	4.3	9.7	13.7	7.4	9.9	8.4	9.9	8.8	13.2	9.9	7.1
Anzahl Monate ALV und Erwerbsarbeit kombiniert	3.0	3.1	19.4	5.0	1.4	4.7	6.5	1.8	2.2	2.1	3.8
Anzahl Bezugsperioden ALV	1.6	1.6	2.7	2.8	1.3	1.6	1.9	1.5	1.6	1.4	1.7
Anzahl Monate mit Erwerbsarbeit	44	32	43	26	4	36	33	7	11	10	34
Anteil Personen mit mindestens einer SH-Bezugs-periode	5%	10%	10%	26%	30%	16%	100%	100%	100%	12%	16%
Anzahl Monate mit SH	0	0	1	2	5	2	28	37	20	1	3
Anzahl Monate SH und Erwerbsarbeit kombiniert	0	0	0	1	0	1	18	4	4	0	1
Anteil Personen mit mindestens einer IV-Bezugs-periode	0%	1%	0%	0%	100%	100%	1%	7%	2%	1%	3%
Anzahl Monate mit IV	0	0	0	0	34	38	0	1	0	0	1
Anzahl Monate IV und Erwerbsarbeit kombiniert	0	0	0	0	3	31	0	0	0	0	0
Anzahl Monate ohne Erwerbsarbeit und Sozial-leistungen	1	6	1	3	5	1	2	3	7	28	6

Quelle: BFS - SHIVALV-IK 2010-2014

Tabelle A 18: Neue Clusterlösung K2011- Verlaufsindikatoren

Verlaufsindikatoren 2011 ***Neue Clusterlösung***	**ALV Kurzzeit**	**ALV Langzeit**	**Zwischenverdienst**	**ALV Mehrfach**	**IV-Rente**	**IV-Rente und Erwerb**	**Sozialhilfe und Erwerb**	**Leavers**	**Neu1**	**Neu2**	**Kohorte insgesamt**
Anzahl Monate mit ALV	6.5	18.8	20.2	23.1	11.1	11.9	13.8	11.4	13.8	9.9	10.8
Dauer der ersten ALV-Bezugsperiode (Monate)	4.2	14.4	13.1	7.1	9.5	8.3	9.3	8.9	10.6	5.6	7.2
Anzahl Monate ALV und Erwerbsarbeit kombiniert	2.6	2.4	15.4	5.7	0.9	4.1	5.7	1.6	2.1	2.7	3.8
Anzahl Bezugsperioden ALV	1.6	1.7	2.2	3.0	1.2	1.7	1.9	1.4	1.5	1.9	1.7
Anzahl Monate mit Erwerbsarbeit	43	27	41	27	3	35	33	8	8	27	33
Anteil Personen mit mindestens einer SH-Bezugsperiode	6%	21%	11%	18%	23%	16%	100%	12%	100%	16%	16%
Anzahl Monate mit SH	0	2	1	1	2	2	30	1	29	1	3
Anzahl Monate SH und Erwerbsarbeit kombiniert	0	1	1	0	0	1	21	0	4	0	1
Anteil Personen mit mindestens einer IV-Bezugsperiode	0%	0%	1%	0%	100%	100%	1%	2%	6%	4%	2%
Anzahl Monate mit IV	0	0	0	0	39	41	0	0	1	1	1
Anzahl Monate IV und Erwerbsarbeit kombiniert	0	0	0	0	2	32	0	0	0	0	0
Anzahl Monate ohne Erwerbsarbeit und Sozialleistungen	1	3	2	3	3	2	2	30	5	13	6

Anmerkung: Die Zuordnung der Cluster der neuen Lösung zu den Clustern der Referenz sowie die Identifikation neuer Cluster (Neu1 etc.) beziehen sich auf eine visuelle Interpretation der state distribution plots und bezieht die Resultate der Jaccard-Matrix nicht mit ein.
Quelle: BFS - SHIVALV-IK 2011-2015

Tabelle A 19: Prädiktion K2011- Verlaufsindikatoren

Verlaufsindikatoren 2011 ***Prädiktion***	**ALV Kurzzeit**	**ALV Langzeit**	**Zwischenverdienst**	**ALV Mehrfach**	**IV-Rente**	**IV-Rente und Erwerb**	**Sozialhilfe und Erwerb**	**Sozialhilfe wiederholt**	**Sozialhilfe neu**	**Leavers**	**Kohorte insgesamt**
Anzahl Monate mit ALV	6.8	13.4	24.5	23.1	11.1	11.8	14.4	10.3	17.0	12.7	10.8
Dauer der ersten ALV-Bezugsperiode (Monate)	4.2	10.1	15.1	7.5	9.4	8.0	9.7	8.0	13.1	10.0	7.2
Anzahl Monate ALV und Erwerbsarbeit kombiniert	3.1	3.3	21.0	5.0	1.1	4.6	6.5	1.6	2.2	2.1	3.8
Anzahl Bezugsperioden ALV	1.7	1.7	2.4	2.9	1.3	1.7	1.9	1.4	1.6	1.5	1.7
Anzahl Monate mit Erwerbsarbeit	44	32	44	25	4	36	34	6	11	9	33
Anteil Personen mit mindestens einer SH-Bezugsperiode	5%	10%	9%	24%	30%	15%	100%	100%	100%	13%	16%
Anzahl Monate mit SH	0	1	0	2	4	2	29	38	20	1	3
Anzahl Monate SH und Erwerbsarbeit kombiniert	0	0	0	1	0	1	20	4	3	0	1
Anteil Personen mit mindestens einer IV-Bezugsperiode	0%	1%	0%	1%	100%	100%	1%	6%	2%	1%	2%
Anzahl Monate mit IV	0	0	0	0	33	39	0	1	0	0	1
Anzahl Monate IV und Erwerbsarbeit kombiniert	0	0	0	0	2	32	0	0	0	0	0
Anzahl Monate ohne Erwerbsarbeit und Sozialleistungen	1	6	1	4	6	1	1	3	7	28	6

Quelle: BFS - SHIVALV-IK 2011-2015

Tabelle A 20: Neue Clusterlösung K2012- Verlaufsindikatoren

Verlaufsindikatoren 2012 ***Neue Clusterlösung***	**ALV Kurz-zeit**	**ALV Lang-zeit**	**Zwischen-verdienst**	**ALV Mehr-fach**	**IV-Rente**	**IV-Rente und Er-werb**	**Sozialhilfe und Er-werb**	**Leavers**	**Neu1**	**Neu3**	**Kohorte insgesamt**
Anzahl Monate mit ALV	7.3	16.8	21.7	23.9	12.6	11.2	15.0	12.5	13.9	10.0	11.1
Dauer der ersten ALV-Bezugsperiode (Monate)	4.6	12.9	14.0	8.4	10.2	8.3	10.6	10.2	11.0	6.7	7.4
Anzahl Monate ALV und Erwerbsarbeit kombiniert	2.9	2.4	16.3	5.2	1.3	4.5	4.7	1.3	1.9	4.7	3.9
Anzahl Bezugsperioden ALV	1.7	1.7	2.3	2.9	1.3	1.6	1.8	1.4	1.5	1.7	1.7
Anzahl Monate mit Erwerbsarbeit	42	26	41	25	5	37	29	6	6	23	33
Anteil Personen mit mindestens einer SH-Bezugs-periode	5%	27%	13%	21%	34%	10%	100%	17%	100%	12%	15%
Anzahl Monate mit SH	0	4	1	1	5	1	34	2	31	1	3
Anzahl Monate SH und Erwerbsarbeit kombiniert	0	2	1	0	1	0	22	0	3	0	1
Anteil Personen mit mindestens einer IV-Bezugspe-riode	0%	0%	0%	0%	100%	100%	1%	2%	4%	2%	2%
Anzahl Monate mit IV	0	0	0	0	34	39	0	0	0	0	1
Anzahl Monate IV und Erwerbsarbeit kombiniert	0	0	0	0	3	31	0	0	0	0	0
Anzahl Monate ohne Erwerbsarbeit und Sozialleis-tungen	1	6	1	4	4	1	2	30	4	20	6

Anmerkung: Die Zuordnung der Cluster der neuen Lösung zu den Clustern der Referenz sowie die Identifikation neuer Cluster (Neu1 etc.) beziehen sich auf eine visuelle Interpretation der state distribution plots und bezieht die Resultate der Jaccard-Matrix nicht mit ein.
Quelle: BFS - SHIVALV-IK 2012-2016

Tabelle A 21: Prädiktion K2011- Verlaufsindikatoren

Verlaufsindikatoren 2012 ***Prädiktion***	**ALV Kurzzeit**	**ALV Langzeit**	**Zwischenver-dienst**	**ALV Mehrfach**	**IV-Rente**	**IV-Rente und Erwerb**	**Sozialhilfe und Erwerb**	**Sozialhilfe wie-derholt**	**Sozialhilfe neu**	**Leavers**	**Kohorte insge-samt**
Anzahl Monate mit ALV	7.0	13.8	24.7	23.3	11.8	11.3	14.7	10.5	17.1	13.0	11.1
Dauer der ersten ALV-Bezugsperiode (Monate)	4.4	10.6	15.2	7.9	9.8	8.1	10.1	8.2	13.4	10.4	7.4
Anzahl Monate ALV und Erwerbsarbeit kombi-niert	3.2	3.2	21.1	5.0	1.2	4.2	6.5	1.6	2.0	2.1	3.9
Anzahl Bezugsperioden ALV	1.7	1.6	2.4	2.9	1.3	1.6	1.9	1.4	1.6	1.4	1.7
Anzahl Monate mit Erwerbsarbeit	44	32	44	25	4	37	34	6	10	9	33
Anteil Personen mit mindestens einer SH-Be-zugsperiode	5%	10%	9%	24%	28%	14%	100%	100%	100%	12%	15%
Anzahl Monate mit SH	0	1	0	2	3	2	29	38	20	1	3
Anzahl Monate SH und Erwerbsarbeit kombi-niert	0	0	0	1	0	1	20	4	3	0	1
Anteil Personen mit mindestens einer IV-Be-zugsperiode	0%	0%	0%	1%	100%	100%	2%	7%	2%	1%	2%
Anzahl Monate mit IV	0	0	0	0	33	39	0	1	0	0	1
Anzahl Monate IV und Erwerbsarbeit kombiniert	0	0	0	0	3	31	0	0	0	0	0
Anzahl Monate ohne Erwerbsarbeit und Sozial-leistungen	1	5	0	4	6	1	1	3	7	27	6

Quelle: BFS - SHIVALV-IK 2012-2016

Tabelle A 22: Neue Clusterlösung K2013- Verlaufsindikatoren

Verlaufsindikatoren 2013 ***Neue Clusterlösung***	**ALV Kurzzeit**	**ALV Langzeit**	**Zwischenverdienst**	**ALV Mehrfach**	**IV-Rente**	**IV-Rente und Erwerb**	**Sozialhilfe und Erwerb**	**Neu1**	**Neu2**	**Leavers**	**Kohorte insgesamt**
Anzahl Monate mit ALV	5.4	15.2	20.6	23.2	12.8	11.8	15.2	14.1	13.5	12.1	11.2
Dauer der ersten ALV-Bezugsperiode (Monate)	3.8	9.7	13.7	7.7	10.4	7.8	10.5	10.9	9.7	9.9	7.5
Anzahl Monate ALV und Erwerbsarbeit kombiniert	2.4	3.8	15.9	5.2	1.3	4.6	7.1	2.3	2.3	1.7	3.8
Anzahl Bezugsperioden ALV	1.5	2.1	2.2	2.9	1.4	1.8	1.9	1.5	1.7	1.4	1.7
Anzahl Monate mit Erwerbsarbeit	44	34	41	27	5	36	34	9	23	6	33
Anteil Personen mit mindestens einer SH-Bezugsperiode	4%	12%	9%	17%	17%	16%	100%	100%	24%	15%	15%
Anzahl Monate mit SH	0	1	1	1	2	2	32	29	2	2	2
Anzahl Monate SH und Erwerbsarbeit kombiniert	0	1	0	0	0	1	25	4	1	0	1
Anteil Personen mit mindestens einer IV-Bezugsperiode	0%	0%	0%	0%	100%	100%	2%	6%	1%	1%	2%
Anzahl Monate mit IV	0	0	0	0	35	38	0	1	0	0	1
Anzahl Monate IV und Erwerbsarbeit kombiniert	0	0	0	0	3	31	0	0	0	0	0
Anzahl Monate ohne Erwerbsarbeit und Sozialleistungen	1	3	2	3	6	2	1	4	12	30	6

Anmerkung: Die Zuordnung der Cluster der neuen Lösung zu den Clustern der Referenz sowie die Identifikation neuer Cluster (Neu1 etc.) beziehen sich auf eine visuelle Interpretation der state distribution plots und bezieht die Resultate der Jaccard-Matrix nicht mit ein.
Quelle: BFS - SHIVALV-IK 2013-2017

Tabelle A 23: Prädiktion K2013- Verlaufsindikatoren

Verlaufsindikatoren 2013 ***Prädiktion***	**ALV Kurzzeit**	**ALV Langzeit**	**Zwischenverdienst**	**ALV Mehrfach**	**IV-Rente**	**IV-Rente und Erwerb**	**Sozialhilfe und Erwerb**	**Sozialhilfe wiederholt**	**Sozialhilfe neu**	**Leavers**	**Kohorte insgesamt**
Anzahl Monate mit ALV	7.1	13.9	24.6	23.3	11.9	11.8	14.9	10.4	17.1	13.1	11.2
Dauer der ersten ALV-Bezugsperiode (Monate)	4.4	10.6	15.0	7.8	9.9	7.8	10.1	8.3	13.3	10.5	7.5
Anzahl Monate ALV und Erwerbsarbeit kombiniert	3.2	3.2	21.1	4.9	1.2	4.3	6.8	1.7	2.1	2.1	3.8
Anzahl Bezugsperioden ALV	1.7	1.7	2.5	2.9	1.3	1.8	1.9	1.4	1.6	1.4	1.7
Anzahl Monate mit Erwerbsarbeit	44	32	44	25	4	36	34	7	10	9	33
Anteil Personen mit mindestens einer SH-Bezugsperiode	5%	9%	8%	22%	27%	15%	100%	100%	100%	11%	15%
Anzahl Monate mit SH	0	0	0	2	3	2	29	38	20	1	2
Anzahl Monate SH und Erwerbsarbeit kombiniert	0	0	0	1	0	1	20	4	3	0	1
Anteil Personen mit mindestens einer IV-Bezugsperiode	0%	1%	0%	0%	100%	100%	2%	6%	2%	1%	2%
Anzahl Monate mit IV	0	0	0	0	32	38	0	1	0	0	1
Anzahl Monate IV und Erwerbsarbeit kombiniert	0	0	0	0	2	31	0	0	0	0	0
Anzahl Monate ohne Erwerbsarbeit und Sozialleistungen	1	5	0	4	6	2	1	3	7	27	6

Quelle: BFS - SHIVALV-IK 2013-2017

Tabelle A 24: Neue Clusterlösung K2014- Verlaufsindikatoren

Verlaufsindikatoren 2014 *Neue Clusterlösung*	ALV Kurzzeit	ALV Langzeit	Zwischenverdienst	ALV Mehrfach	Sozialhilfe und Erwerb	Leavers	Neu1	Neu2	Neu3	Neu4	Kohorte insgesamt
Anzahl Monate mit ALV	6.2	17.3	22.8	20.8	16.6	13.3	14.2	9.9	10.0	11.6	11.2
Dauer der ersten ALV-Bezugsperiode (Monate)	4.2	13.1	15.2	6.7	11.2	10.9	10.7	6.9	6.3	8.8	7.5
Anzahl Monate ALV und Erwerbsarbeit kombiniert	2.9	2.5	17.4	6.9	7.0	1.5	2.2	2.0	4.2	3.0	3.8
Anzahl Bezugsperioden ALV	1.6	1.7	2.3	3.1	2.0	1.4	1.6	1.7	1.8	1.5	1.7
Anzahl Monate mit Erwerbsarbeit	44	31	41	31	33	5	10	24	26	20	33
Anteil Personen mit mindestens einer SH-Bezugsperiode	4%	11%	10%	25%	100%	15%	100%	12%	9%	18%	15%
Anzahl Monate mit SH	0	1	1	3	30	1	28	1	0	2	2
Anzahl Monate SH und Erwerbsarbeit kombiniert	0	0	1	2	23	0	4	0	0	0	1
Anteil Personen mit mindestens einer IV-Bezugsperiode	0%	0%	1%	1%	2%	2%	5%	1%	1%	100%	2%
Anzahl Monate mit IV	0	0	0	0	0	0	1	0	0	37	1
Anzahl Monate IV und Erwerbsarbeit kombiniert	0	0	0	0	0	0	0	0	0	16	0
Anzahl Monate ohne Erwerbsarbeit und Sozialleistungen	1	2	1	2	1	30	4	16	16	4	6

Anmerkung: Die Zuordnung der Cluster der neuen Lösung zu den Clustern der Referenz sowie die Identifikation neuer Cluster (Neu1 etc.) beziehen sich auf eine visuelle Interpretation der state distribution plots und bezieht die Resultate der Jaccard-Matrix nicht mit ein.
Quelle: BFS - SHIVALV-IK 2014-2018

Tabelle A 25: Prädiktion K2014- Verlaufsindikatoren

Verlaufsindikatoren 2014 *Prädiktion*	ALV Kurzzeit	ALV Langzeit	Zwischenverdienst	ALV Mehrfach	IV-Rente	IV-Rente und Erwerb	Sozialhilfe und Erwerb	Sozialhilfe wiederholt	Sozialhilfe neu	Leavers	Kohorte insgesamt
Anzahl Monate mit ALV	7.0	14.0	24.6	23.0	11.6	11.7	15.0	10.5	17.2	13.1	11.2
Dauer der ersten ALV-Bezugsperiode (Monate)	4.4	10.6	15.0	7.7	9.6	8.0	10.0	8.2	13.4	10.4	7.5
Anzahl Monate ALV und Erwerbsarbeit kombiniert	3.2	3.2	21.0	4.9	1.5	4.7	6.7	1.6	2.1	2.1	3.8
Anzahl Bezugsperioden ALV	1.7	1.7	2.5	2.8	1.3	1.7	1.9	1.4	1.6	1.5	1.7
Anzahl Monate mit Erwerbsarbeit	44	32	44	25	5	36	34	6	11	9	33
Anteil Personen mit mindestens einer SH-Bezugsperiode	4%	9%	8%	21%	26%	13%	100%	100%	100%	12%	15%
Anzahl Monate mit SH	0	0	0	2	4	1	28	38	19	1	2
Anzahl Monate SH und Erwerbsarbeit kombiniert	0	0	0	1	0	1	20	3	3	0	1
Anteil Personen mit mindestens einer IV-Bezugsperiode	0%	0%	0%	0%	100%	100%	2%	7%	2%	1%	2%
Anzahl Monate mit IV	0	0	0	0	32	38	0	1	0	0	1
Anzahl Monate IV und Erwerbsarbeit kombiniert	0	0	0	0	3	30	0	0	0	0	0
Anzahl Monate ohne Erwerbsarbeit und Sozialleistungen	1	5	0	4	6	2	1	3	7	27	6

Quelle: BFS - SHIVALV-IK 2014-2018

Tabelle A 26: Neue Clusterlösung K2015- Verlaufsindikatoren

Verlaufsindikatoren 2015 ***Neue Clusterlösung***	**ALV Kurzzeit**	**ALV Langzeit**	**Zwischenverdienst**	**ALV Mehrfach**	**IV-Rente**	**IV-Rente und Erwerb**	**Sozialhilfe und Erwerb**	**Leavers**	**Neu1**	**Neu3**	**Kohorte insgesamt**
Anzahl Monate mit ALV	7.4	17.1	22.8	21.6	11.2	12.0	13.7	12.6	14.5	11.6	11.3
Dauer der ersten ALV-Bezugsperiode (Monate)	5.0	12.8	14.7	7.6	8.9	8.6	9.3	10.2	11.4	6.3	7.7
Anzahl Monate ALV und Erwerbsarbeit kombiniert	3.0	2.8	17.7	5.6	1.3	4.6	6.2	1.8	2.2	5.0	3.8
Anzahl Bezugsperioden ALV	1.7	1.7	2.3	2.8	1.4	1.6	1.9	1.4	1.5	2.0	1.7
Anzahl Monate mit Erwerbsarbeit	42	26	42	28	5	36	35	5	10	25	33
Anteil Personen mit mindestens einer SH-Bezugsperiode	4%	16%	10%	28%	29%	15%	100%	14%	100%	11%	14%
Anzahl Monate mit SH	0	1	1	4	4	2	35	1	29	1	2
Anzahl Monate SH und Erwerbsarbeit kombiniert	0	0	1	2	0	1	26	0	4	0	1
Anteil Personen mit mindestens einer IV-Bezugsperiode	0%	0%	0%	1%	100%	100%	1%	1%	4%	1%	2%
Anzahl Monate mit IV	0	0	0	0	31	38	0	0	0	0	1
Anzahl Monate IV und Erwerbsarbeit kombiniert	0	0	0	0	2	31	0	0	0	0	0
Anzahl Monate ohne Erwerbsarbeit und Sozialleistungen	1	7	1	3	7	2	1	31	4	16	6

Anmerkung: Die Zuordnung der Cluster der neuen Lösung zu den Clustern der Referenz sowie die Identifikation neuer Cluster (Neu1 etc.) beziehen sich auf eine visuelle Interpretation der state distribution plots und bezieht die Resultate der Jaccard-Matrix nicht mit ein.
Quelle: BFS - SHIVALV-IK 2015-2019

Tabelle A 27: Prädiktion K2011- Verlaufsindikatoren

Verlaufsindikatoren 2015 ***Prädiktion***	**ALV Kurzzeit**	**ALV Langzeit**	**Zwischenverdienst**	**ALV Mehrfach**	**IV-Rente**	**IV-Rente und Erwerb**	**Sozialhilfe und Erwerb**	**Sozialhilfe wiederholt**	**Sozialhilfe neu**	**Leavers**	**Kohorte insgesamt**
Anzahl Monate mit ALV	7.0	14.1	24.2	23.1	11.6	11.8	14.9	10.3	17.3	13.2	11.3
Dauer der ersten ALV-Bezugsperiode (Monate)	4.5	10.8	15.0	8.2	9.5	8.3	10.2	8.2	13.7	10.6	7.7
Anzahl Monate ALV und Erwerbsarbeit kombiniert	3.2	3.2	20.6	4.9	1.4	4.1	6.8	1.4	2.0	2.1	3.8
Anzahl Bezugsperioden ALV	1.7	1.7	2.5	2.8	1.3	1.7	1.9	1.4	1.6	1.5	1.7
Anzahl Monate mit Erwerbsarbeit	43	32	44	25	4	36	34	6	11	9	33
Anteil Personen mit mindestens einer SH-Bezugsperiode	4%	9%	8%	20%	29%	14%	100%	100%	100%	10%	14%
Anzahl Monate mit SH	0	0	0	2	4	2	29	38	19	1	2
Anzahl Monate SH und Erwerbsarbeit kombiniert	0	0	0	1	0	1	20	3	3	0	1
Anteil Personen mit mindestens einer IV-Bezugsperiode	0%	0%	1%	1%	100%	100%	1%	6%	2%	1%	2%
Anzahl Monate mit IV	0	0	0	0	31	38	0	1	0	0	1
Anzahl Monate IV und Erwerbsarbeit kombiniert	0	0	0	0	3	30	0	0	0	0	0
Anzahl Monate ohne Erwerbsarbeit und Sozialleistungen	1	5	0	4	7	2	1	3	6	27	6

Quelle: BFS - SHIVALV-IK 2015-2019

7.21 Anhang zum Abschnitt «Analysen zur Aktualisierungsnotwendigkeit der initialen Clusterlösung» (4.44)

7.21.1 Ergänzungen zu Abschnitt 4.4.1

Interne Masse der initialen Clusterlösung auf Basis der Repräsentanten des ersten Clustering-Schritts

Folgende Tabelle gibt die internen Masse für die initiale Clusterlösung auf Basis der 3000 Repräsentanten aus dem ersten Clustering-Schritt wieder (siehe auch Abschnitt 3.4.2). Sie dient als Vergleich zu den internen Massen die auf einfachen Stichproben berechnet wurden (siehe Tabelle 13). Eine Approximation wie bei Tabelle A 28 mithilfe von Repräsentanten ist aber nur bei neu berechneten Clusterlösungen mittels des hier angewendeten Zwei-Schritt-Clusterings möglich. Die Varianzanteile die mithilfe dieser Approximation berechnet wurden sind auch für das Elbow-plot benutzt worden. Es ist zu beachten, dass sie etwas kleiner ausfallen als diese die mithilfe der Approximation mit den Stichproben berechnet wurden.

Tabelle A 28: Interne Masse für die Referenz auf Basis der 3000 «Repräsentanten» des ersten Clustering-Schritts

Cluster	Gesamt	1	2	3	4	5	6	7	8	9	10
Anzahl	123786	19432	64075	1213	3387	16607	5614	3489	5809	1316	2844
Anz.(%)	100.0	15.70	51.76	0.98	2.74	13.42	4.54	2.82	4.69	1.06	2.30
mean(d)	50.84	38.58	20.59	45.97	54.81	29.84	34.82	42.97	34.45	48.31	36.97
max(d)	96	88.33	58.31	71.64	94.29	64.74	75.22	82.44	77.31	93.49	73.24
var (%)	4.57	1.25	2.48	0.01	0.08	0.52	0.08	0.05	0.07	0.01	0.02
S. Koef.	0.33	0.32	0.50	0.33	--0.07	0.01	0.06	0.009	0.22	0.25	0.29

Quelle: BFS - SHIVALV-IK 2010-2014

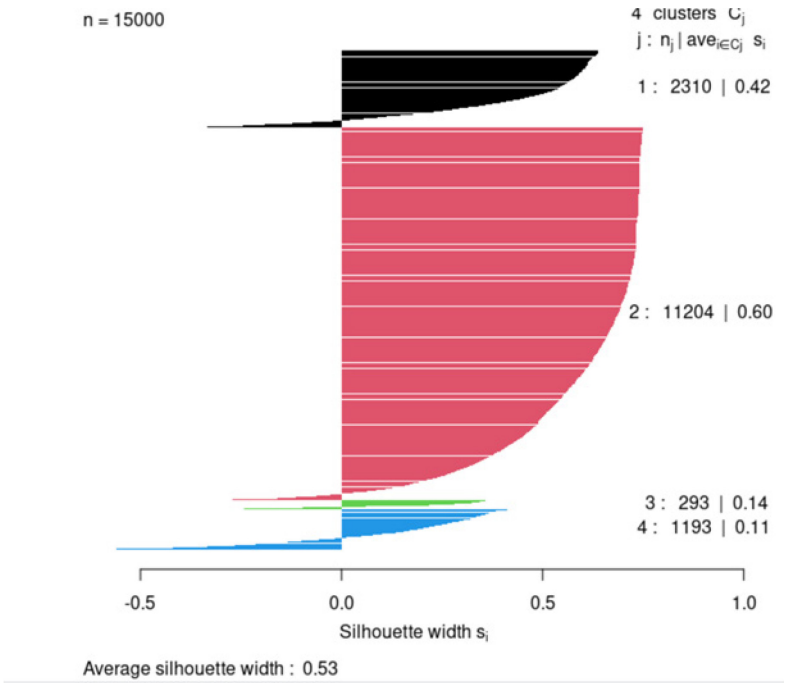
Interne Masse für eine initiales Clusterlösung mit vier Clustern

Tabelle A 29: Interne Masse für eine initiale Clusterlösung mit vier Clustern, Kohorte 2010

Cluster	Gesamt	1	2	3	4
Stichpr.	24757	3804	18494	511	1948
St. (%)	100.0	15.37	74.70	2.06	7.87
mean(d)	46.1	36.20	26.56	65.73	58.21
max(d)	96.0	94.40	94.14	95.98	95.75
var (%)	100.0	1.284	18.118	0.072	0.797
S. Koef.	0.53	0.42	0.60	0.14	0.11

Quelle: BFS - SHIVALV-IK 2010-2014

Abbildung A 6: Silhouette-Plot für eine initiale Clusterlösung mit vier Clustern, Kohorte 2010



Quelle: BFS - SHIVALV-IK 2010-2014

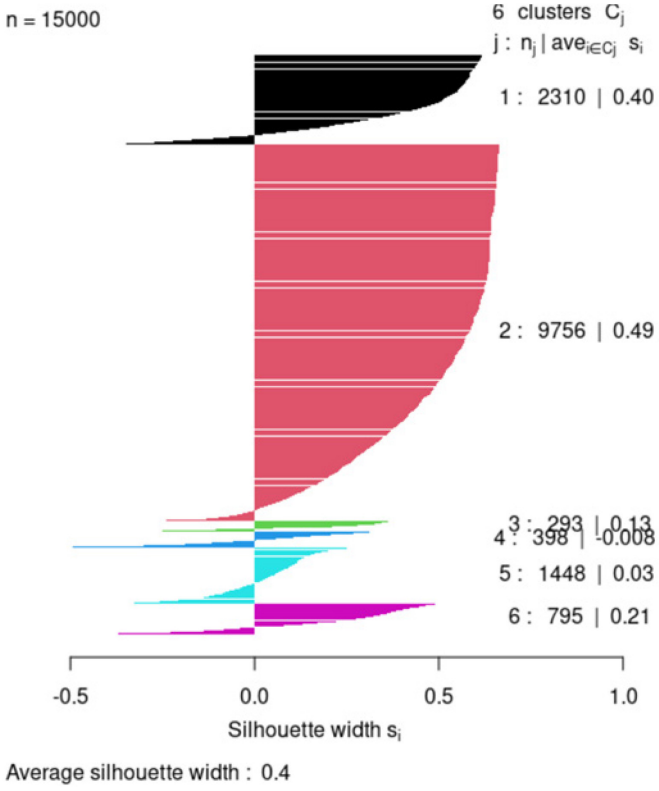
Interne Masse für eine initiale Clusterlösung mit sechs Clustern

Tabelle A 30: Interne Masse für eine initiale Clusterlösung mit sechs Clustern, Kohorte 2010

Cluster	Gesamt	1	2	3	4	5	6
Stichpr.	24757	3804	16157	511	665	2337	1283
St. (%)	100.0	15.37	65.26	2.06	2.69	9.44	5.18
mean(d)	46.1	36.20	21.59	65.73	56.96	40.55	47.40
max(d)	96.0	94.40	80.17	95.98	94.89	92.23	95.25
var (%)	100.0	1.284	9.199	0.072	0.087	0.568	0.232
S. Koef.	0.40	0.40	0.49	0.13	-0.008	0.03	0.21

Quelle: BFS - SHIVALV-IK 2010-2014

Abbildung A 7: Silhouette-Plot für eine initiale Clusterlösung mit sechs Clustern, Kohorte 2010



Quelle: BFS - SHIVALV-IK 2010-2014

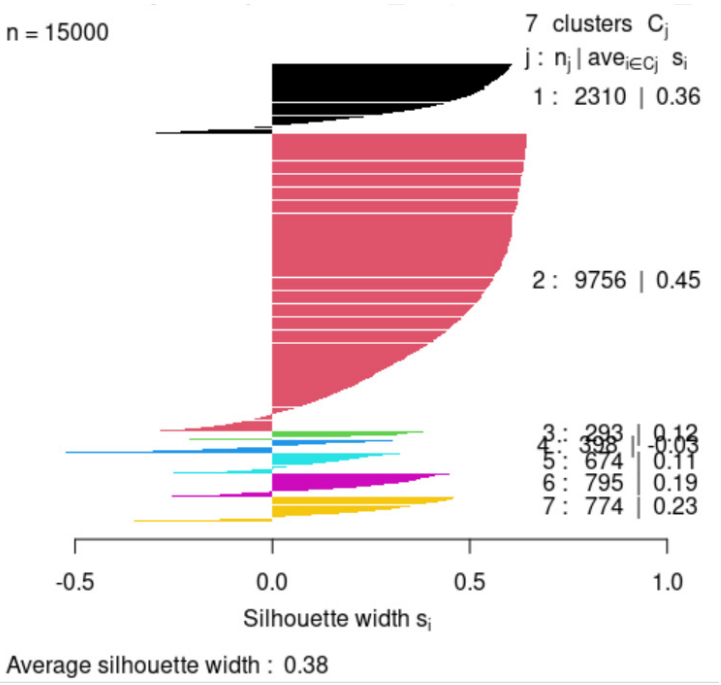
Interne Masse für eine initiale Clusterlösung mit sieben Clustern

Tabelle A 31: Interne Masse für eine initiale Clusterlösung mit sieben Clustern, Kohorte 2010

Cluster	Gesamt	1	2	3	4	5	6	7
Stichpr.	24757	3804	16157	511	665	1119	1283	1218
St. (%)	100.0	15.370	65.26	2.06	2.69	4.52	5.18	4.92
mean(d)	46.1	36.200	21.59	65.73	56.96	38.17	47.40	29.78
max(d)	96.0	94.400	80.17	95.98	94.89	84.10	95.25	91.14
var (%)	100.0	1.284	9.199	0.072	0.087	0.111	0.232	0.095
S. Koef.	0.38	0.36	0.45	0.12	-0.03	0.11	0.19	0.23

Quelle: BFS - SHIVALV-IK 2010-2014

Abbildung A 8: Silhouette-Plot für eine initiale Clusterlösung mit sieben Clustern, Kohorte 2010



Quelle: BFS - SHIVALV-IK 2010-2014

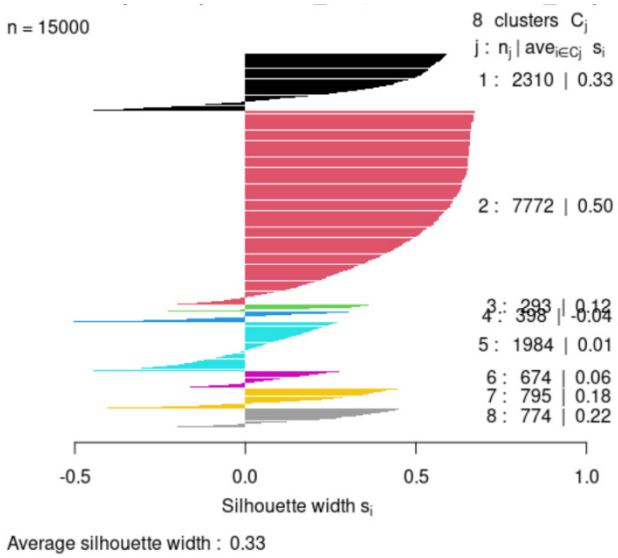
Interne Masse für eine initiale Clusterlösung mit acht Clustern

Tabelle A 32: Interne Masse für eine initiale Clusterlösung mit acht Clustern, Kohorte 2010

Cluster	Gesamt	1	2	3	4	5	6	7	8
Stichpr.	24757	3804	12828	511	665	3329	1119	1283	1218
St. (%)	100.0	15.37	51.82	2.06	2.69	13.45	4.52	5.18	4.92
mean(d)	46.1	36.20	15.78	65.73	56.96	29.79	38.17	47.40	29.78
max(d)	96.0	94.40	77.73	95.980	94.89	80.17	84.10	95.25	91.14
var (%)	100.0	1.284	3.223	0.072	0.087	0.636	0.111	0.232	0.095
S. Koef.	0.33	0.33	0.50	0.12	-0.04	0.01	0.06	0.18	0.22

Quelle: BFS - SHIVALV-IK 2010-2014

Abbildung A 9: Silhouette-Plot für eine initiale Clusterlösung mit acht Clustern, Kohorte 2010



Quelle: BFS - SHIVALV-IK 2010-2014

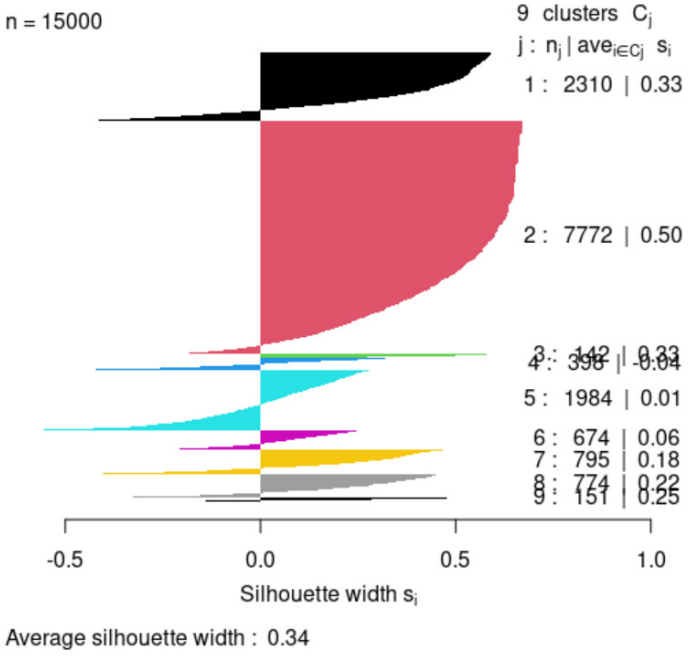
Interne Masse für eine initiale Clusterlösung mit neun Clustern

Tabelle A 33: Interne Masse für eine initiale Clusterlösung mit neun Clustern, Kohorte 2010

Cluster	Gesamt	1	2	3	4	5	6	7	8	9
Stichpr.	24757	3804	12828	247	665	3329	1119	1283	1218	264
St. (%)	100.0	15.37	51.82	1.00	2.69	13.45	4.52	5.18	4.92	1.07
mean(d)	46.1	36.20	15.78	46.60	56.96	29.79	38.17	47.40	29.78	51.53
max(d)	96.0	94.40	77.73	94.43	94.89	80.17	84.10	95.25	91.14	95.66
var (%)	100.0	1.284	3.223	0.009	0.087	0.636	0.111	0.232	0.095	0.012
S. Koef.	0.34	0.33	0.50	0.33	-0.04	0.01	0.06	0.18	0.22	0.25

Quelle: BFS - SHIVALV-IK 2010-2014

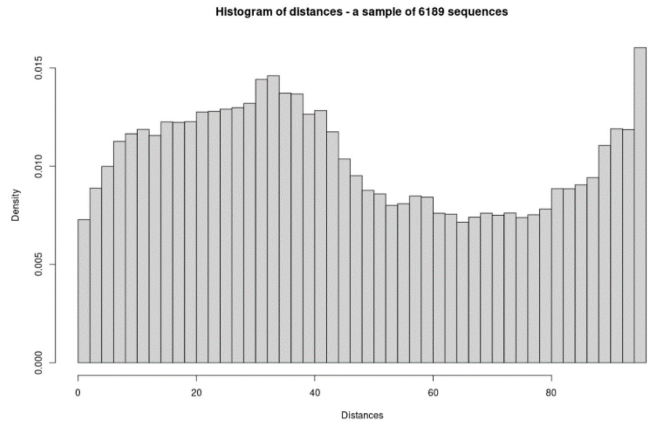
Abbildung A 10: Silhouette-Plot für eine initiale Clusterlösung mit neun Clustern, Kohorte 2010



Quelle: BFS - SHIVALV-IK 2010-2014

Distanzverteilung Kohorte 2010 insgesamt und in ausgewählten Clustern der initialen Clusterlösung

Abbildung A 11: Verteilungsdichte und Quantile der Distanzen zwischen den Verläufen, K2010



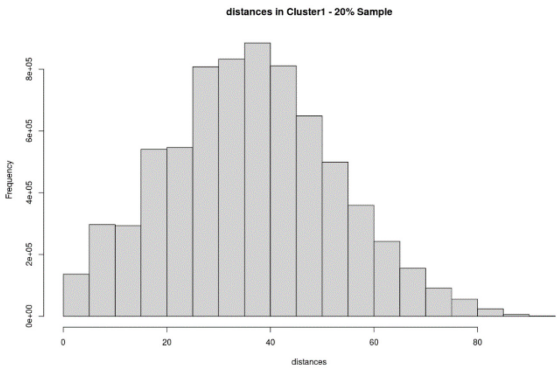
Quantile der Distanz-Verteilung

0%	25%	50%	75%	100%
0	22.85	40.96	69.97	96

Anmerkung: Verteilung wurde berechnet auf Basis einer einfachen Zufallsstichprobe von 5% der Kohorte 2010 (6189 Verläufe)
Quelle: BFS - SHIVALV-IK 2010-2014

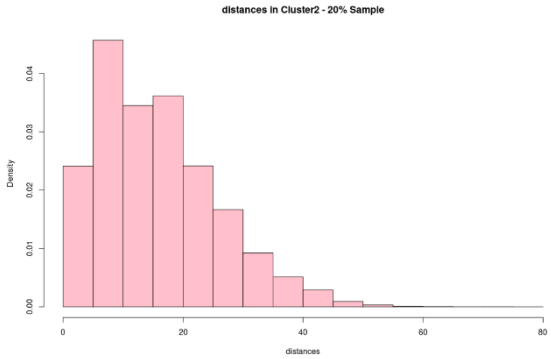
Abbildung A 12: Verteilungsdichten und Quantile der Distanzen zwischen den Verläufen für eine Auswahl von Clustern der initialen Clusterlösung, K2010

Cluster 1



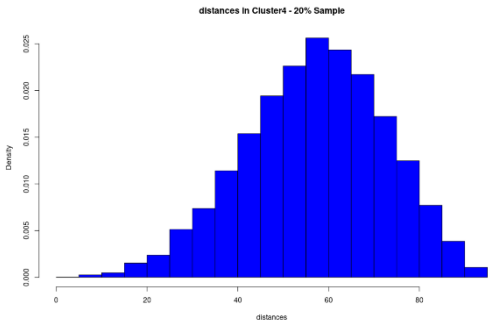
0%	25%	50%	75%	100%
0	24.95	35.80	46.81	94.4

Cluster 2



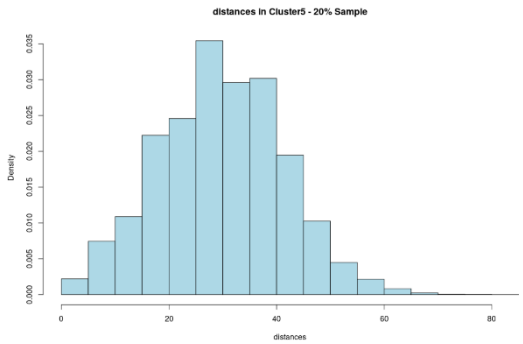
0%	25%	50%	75%	100%
0	7.74	14.06	21.72	77.73

Cluster 4



0%	25%	50%	75%	100%
1.97	46.79	57.8	68.16	94.9

Cluster 5



0%	25%	50%	75%	100%
0	21.63	29.65	37.62	80.17

Quelle: BFS - SHIVALV-IK 2010-2014

7.21.2 Ergänzungen zu Abschnitt 4.4.3

Tabelle A 34: Relative Häufigkeitsverteilung (Zeilenprozent) der Konfusionsmatrix für die Cluster der Prädiktion und einer neuen Clusterlösung, Kohorte K2011

Kohorte 2011 Relative Verteilung Zeile			**neue Clusterlösung**										**Total**
			A	B	C	D	E	F	G	H	I	J	
Prädiktion initiale Clusterlösung	ALV-Kurzzeit	1	0.5%	91.4%	0.9%	0.1%	0.0%	1.5%	5.6%	0.0%	0.0%	0.0%	100%
	ALV-Langzeit	2	11.5%	34.5%	33.8%	0.4%	0.0%	1.9%	12.0%	5.9%	0.0%	0.0%	100%
	Zwischenverdienst	3	0.1%	5.6%	0.7%	0.2%	0.0%	4.7%	88.6%	0.1%	0.0%	0.0%	100%
	ALV mehrfach	4	3.8%	9.1%	8.1%	1.3%	0.0%	66.3%	6.5%	3.0%	1.6%	0.1%	100%
	IV-Rente	5	2.1%	0.0%	0.0%	0.0%	68.2%	0.0%	0.1%	15.2%	8.9%	5.5%	100%
	IV-Rente und Erwerb	6	6.8%	1.8%	0.4%	0.1%	1.6%	0.0%	3.3%	0.1%	0.4%	85.4%	100%
	Sozialhilfe und Erwerb	7	0.7%	9.3%	5.2%	67.0%	0.0%	0.1%	6.8%	0.9%	9.7%	0.4%	100%
	Sozialhilfe wiederholt	8	0.0%	0.0%	0.3%	3.1%	0.0%	0.0%	0.1%	1.1%	95.3%	0.0%	100%
	Sozialhilfe neu	9	0.4%	0.4%	18.0%	4.3%	0.0%	0.2%	1.1%	3.4%	72.0%	0.0%	100%
	Leavers	10	4.5%	0.3%	9.4%	0.2%	0.0%	0.5%	3.6%	80.5%	1.0%	0.1%	100%

Quelle: BFS - SHIVALV-IK 2011-2015

Tabelle A 35: Relative Häufigkeitsverteilung (Spaltenprozent) der Konfusionsmatrix für die Cluster der Prädiktion und einer neuen Clusterlösung, Kohorte K2011

Kohorte 2011 Relative Verteilung Spalte			**neue Clusterlösung**									
			A	B	C	D	E	F	G	H	I	J
Prädiktion initiale Clusterlösung	ALV-Kurzzeit	1	8.3%	89.2%	6.0%	2.4%	0.0%	17.4%	31.1%	0.0%	0.0%	0.3%
	ALV-Langzeit	2	55.4%	8.9%	59.5%	2.4%	0.0%	5.7%	17.5%	5.5%	0.1%	0.8%
	Zwischenverdienst	3	0.2%	0.4%	0.4%	0.3%	0.0%	4.4%	39.2%	0.0%	0.0%	0.1%
	ALV mehrfach	4	6.3%	0.8%	4.9%	3.0%	0.0%	70.4%	3.3%	1.0%	1.5%	0.5%
	IV-Rente	5	0.8%	0.0%	0.0%	0.0%	97.9%	0.0%	0.0%	1.1%	1.9%	7.2%
	IV-Rente und Erwerb	6	2.0%	0.0%	0.0%	0.0%	1.9%	0.0%	0.3%	0.0%	0.1%	88.8%
	Sozialhilfe und Erwerb	7	0.6%	0.5%	1.7%	81.1%	0.0%	0.1%	1.9%	0.2%	5.1%	1.1%
	Sozialhilfe wiederholt	8	0.0%	0.0%	0.1%	3.1%	0.0%	0.0%	0.0%	0.2%	42.0%	0.1%
	Sozialhilfe neu	9	0.5%	0.0%	7.3%	6.4%	0.1%	0.2%	0.4%	0.7%	45.9%	0.0%
	Leavers	10	25.9%	0.1%	20.1%	1.3%	0.1%	1.8%	6.3%	91.3%	3.3%	1.0%
Total			100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

Quelle: BFS - SHIVALV-IK 2011-2015

Tabelle A 36: Konfusionsmatrix für die Cluster der Prädiktion und einer neuen Clusterlösung, Kohorte K2015

Kohorte 2015 absolute Werte			neue Clusterlösung Q	R	S	T	U	V	W	X	Y	Z	**Total**
Prädiktion initiale Clusterlösung	ALV-Kurzzeit	1	747	66'008	1'111	1	830	3	2	20	64	0	68'786
	ALV-Langzeit	2	9'223	9'551	143	96	1'432	29	4	7	1'183	16	21'684
	Zwischenverdienst	3	232	600	281	10	4'876	1	8	1	48	0	6'057
	ALV mehrfach	4	808	444	3'577	59	253	16	4	9	551	2	5'723
	IV-Rente	5	0	0	0	69	0	12	60	0	0	1'029	1'170
	IV-Rente und Erwerb	6	6	18	11	6	1	0	887	1	0	44	974
	Sozialhilfe und Erwerb	7	131	189	340	41	232	613	2	1'419	5	4	2'976
	Sozialhilfe wiederholt	8	0	0	39	82	0	2'184	4	91	0	20	2'420
	Sozialhilfe neu	9	618	12	75	678	16	2'765	2	55	12	6	4'239
	Leavers	10	4'782	117	124	17'443	84	79	3	6	1'341	35	24'014
Total			16'547	76'939	5'701	18'485	7'724	5'702	976	1'609	3'204	1'156	138'043

Quelle: BFS - SHIVALV-IK 2015-2019

Tabelle A 37: Relative Häufigkeitsverteilung (Zeilenprozent) der Konfusionsmatrix für die Cluster der Prädiktion und einer neuen Clusterlösung, Kohorte K2015

Kohorte 2015 Relative Verteilung Zeile			neue Clusterlösung Q	R	S	T	U	V	W	X	Y	Z	**Total**
Prädiktion initiale Clusterlösung	ALV-Kurzzeit	1	1.1%	96.0%	1.6%	0.0%	1.2%	0.0%	0.0%	0.0%	0.1%	0.0%	100%
	ALV-Langzeit	2	42.5%	44.0%	0.7%	0.4%	6.6%	0.1%	0.0%	0.0%	5.5%	0.1%	100%
	Zwischenverdienst	3	3.8%	9.9%	4.6%	0.2%	80.5%	0.0%	0.1%	0.0%	0.8%	0.0%	100%
	ALV mehrfach	4	14.1%	7.8%	62.5%	1.0%	4.4%	0.3%	0.1%	0.2%	9.6%	0.0%	100%
	IV-Rente	5	0.0%	0.0%	0.0%	5.9%	0.0%	1.0%	5.1%	0.0%	0.0%	87.9%	100%
	IV-Rente und Erwerb	6	0.6%	1.8%	1.1%	0.6%	0.1%	0.0%	91.1%	0.1%	0.0%	4.5%	100%
	Sozialhilfe und Erwerb	7	4.4%	6.4%	11.4%	1.4%	7.8%	20.6%	0.1%	47.7%	0.2%	0.1%	100%
	Sozialhilfe wiederholt	8	0.0%	0.0%	1.6%	3.4%	0.0%	90.2%	0.2%	3.8%	0.0%	0.8%	100%
	Sozialhilfe neu	9	14.6%	0.3%	1.8%	16.0%	0.4%	65.2%	0.0%	1.3%	0.3%	0.1%	100%
	Leavers	10	19.9%	0.5%	0.5%	72.6%	0.3%	0.3%	0.0%	0.0%	5.6%	0.1%	100%

Quelle: BFS - SHIVALV-IK 2015-2019

Tabelle A 38: Relative Häufigkeitsverteilung (Spaltenprozent) der Konfusionsmatrix für die Cluster der Prädiktion und einer neuen Clusterlösung, Kohorte K2015

Kohorte 2015 Relative Verteilung Spalte			neue Clusterlösung Q	R	S	T	U	V	W	X	Y	Z
Prädiktion initiale Clusterlösung	ALV-Kurzzeit	1	4.5%	85.8%	19.5%	0.0%	10.7%	0.1%	0.2%	1.2%	2.0%	0.0%
	ALV-Langzeit	2	55.7%	12.4%	2.5%	0.5%	18.5%	0.5%	0.4%	0.4%	36.9%	1.4%
	Zwischenverdienst	3	1.4%	0.8%	4.9%	0.1%	63.1%	0.0%	0.8%	0.1%	1.5%	0.0%
	ALV mehrfach	4	4.9%	0.6%	62.7%	0.3%	3.3%	0.3%	0.4%	0.6%	17.2%	0.2%
	IV-Rente	5	0.0%	0.0%	0.0%	0.4%	0.0%	0.2%	6.1%	0.0%	0.0%	89.0%
	IV-Rente und Erwerb	6	0.0%	0.0%	0.2%	0.0%	0.0%	0.0%	90.9%	0.1%	0.0%	3.8%
	Sozialhilfe und Erwerb	7	0.8%	0.2%	6.0%	0.2%	3.0%	10.8%	0.2%	88.2%	0.2%	0.3%
	Sozialhilfe wiederholt	8	0.0%	0.0%	0.7%	0.4%	0.0%	38.3%	0.4%	5.7%	0.0%	1.7%
	Sozialhilfe neu	9	3.7%	0.0%	1.3%	3.7%	0.2%	48.5%	0.2%	3.4%	0.4%	0.5%
	Leavers	10	28.9%	0.2%	2.2%	94.4%	1.1%	1.4%	0.3%	0.4%	41.9%	3.0%
Total			100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

Quelle: BFS - SHIVALV-IK 2015-2019

Tabelle A 39: Jaccard Matrix für die Cluster der Prädiktion und einer neuen Clusterlösung, Kohorte 2015

Kohorte 2015 Jaccard-Matrix			**neue Clusterlösung**									
			Q	R	S	T	U	V	W	X	Y	Z
Prädiktion initiale Clusterlösung	ALV-Kurzzeit	1	0.9%	82.8%	1.5%	0.0%	1.1%	0.0%	0.0%	0.0%	0.1%	0.0%
	ALV-Langzeit	2	31.8%	10.7%	0.5%	0.2%	5.1%	0.1%	0.0%	0.0%	5.0%	0.1%
	Zwischenverdienst	3	1.0%	0.7%	2.4%	0.0%	54.8%	0.0%	0.1%	0.0%	0.5%	0.0%
	ALV mehrfach	4	3.8%	0.5%	45.6%	0.2%	1.9%	0.1%	0.1%	0.1%	6.6%	0.0%
	IV-Rente	5	0.0%	0.0%	0.0%	0.4%	0.0%	0.2%	2.9%	0.0%	0.0%	79.3%
	IV-Rente und Erwerb	6	0.0%	0.0%	0.2%	0.0%	0.0%	0.0%	83.4%	0.0%	0.0%	2.1%
	Sozialhilfe und Erwerb	7	0.7%	0.2%	4.1%	0.2%	2.2%	7.6%	0.1%	44.8%	0.1%	0.1%
	Sozialhilfe wiederholt	8	0.0%	0.0%	0.5%	0.4%	0.0%	36.8%	0.1%	2.3%	0.0%	0.6%
	Sozialhilfe neu	9	3.1%	0.0%	0.8%	3.1%	0.1%	38.5%	0.0%	0.9%	0.2%	0.1%
	Leavers	10	13.4%	0.1%	0.4%	69.6%	0.3%	0.3%	0.0%	0.0%	5.2%	0.1%

Quelle: BFS - SHIVALV-IK 2015-2019

Tabelle A 40: Berechnung der externen Masse zur Übereinstimmung von Prädiktion und neuer Clusterlösung, K2011

		neue Clusterlösung										Total	True Positive TP per Cluster	False Negative (FN) per Cluster	TP-Rate per Cluster (TP/TP+FN)
ursprüngliche Clusterbenennung -->		B	C	G	F	E	J	D	I	A	H				
Zuordnung gemässe Jaccard -->		1	2	3	4	5	6	7	8	9	10				
Prädiktion initiale Clusterlösung	ALV - Kurzzeit 1	50'554	499	3'121	827	0	3	55	2	253	5	55'319	50'554	4'765	0.91
	ALV - Langzeit 2	5'064	4'949	1'756	272	0	7	56	6	1'693	860	14'663	4'949	9'714	0.34
	Zwischenverdienst 3	248	30	3'936	210	0	1	7	0	6	6	4'444	3'936	508	0.89
	ALV mehrfach 4	461	409	330	3'342	0	4	68	81	194	152	5'041	3'342	1'699	0.66
	IV-Rente 5	0	0	1	0	783	63	0	102	24	175	1'148	783	365	0.68
	IV – Rente und Erwerb 6	16	4	30	0	15	780	1	4	62	1	913	780	133	0.85
	Sozialhilfe und Erwerb 7	259	145	190	3	0	10	1'864	269	19	25	2'784	1'864	920	0.67
	Sozialhilfe wiederholt 8	1	8	3	0	0	1	72	2'214	0	25	2'324	2'214	110	0.95
	Sozialhilfe neu 9	15	604	37	8	1	0	146	2'421	15	116	3'363	15	3'348	0.00
	Leavers 10	62	1'672	632	84	1	9	29	172	792	14'281	17'734	14'281	3'453	0.81
Total		56'680	8'320	10'036	4'746	800	878	2'298	5'271	3'058	15'646	107'733			

Accuracy	0.77
Balanced Accuracy	0.68
Cohen's Kappa	0.66

Quelle: BFS - SHIVALV-IK 2011-2015

Tabelle A 41: Berechnung der externen Masse zur Übereinstimmung von Prädiktion und neuer Clusterlösung, K2015

		neue Clusterlösung										Total	True Positive TP per Cluster	False Negative (FN) per Cluster	TP-Rate per Cluster (TP/TP+FN)
ursprüngliche Clusterbenennung -->		R	Q	U	S	Z	W	X	Y	V	T				
Zuordnung gemässe Jaccard -->		1	2	3	4	5	6	7	8	9	10				
Prädiktion initiale Clusterlösung	ALV - Kurzzeit 1	66'008	747	830	1'111	0	2	20	64	3	1	68'786	66'008	2'778	0.96
	ALV - Langzeit 2	9'551	9'223	1'432	143	16	4	7	1'183	29	96	21'684	9'223	12'461	0.43
	Zwischenverdienst 3	600	232	4'876	281	0	8	1	48	1	10	6'057	4'876	1'181	0.81
	ALV mehrfach 4	444	808	253	3'577	2	4	9	551	16	59	5'723	3'577	2'146	0.63
	IV-Rente 5	0	0	0	0	1'029	60	0	0	12	69	1'170	1'029	141	0.88
	IV – Rente und Erwerb 6	18	6	1	11	44	887	1	0	0	6	974	887	87	0.91
	Sozialhilfe und Erwerb 7	189	131	232	340	4	2	1'419	5	613	41	2'976	1'419	1'557	0.48
	Sozialhilfe wiederholt 8	0	0	0	39	20	4	91	0	2'184	82	2'420	0	2'420	0.00
	Sozialhilfe neu 9	12	618	16	75	6	2	55	12	2'765	678	4'239	2'765	1'474	0.65
	Leavers 10	117	4'782	84	124	35	3	6	1'341	79	17'443	24'014	17'443	6'571	0.73
Total		76'939	16'547	7'724	5'701	1'156	976	1'609	3'204	5'702	18'485	138'043			

Accuracy	0.78
Balanced Accuracy	0.65
Cohen's Kappa	0.67

Quelle: BFS - SHIVALV-IK 2015-2019

8 Tabellen und Abbildungsverzeichnis

8.1 Tabellen im Haupttext

Tabelle 1: Beispiele für Transformation der Zustandscodierung 11
Tabelle 2: Kohorten neu Arbeitslosentaggeld beziehender Personen 2010 bis 2015 nach soziodemografischen Merkmalen 22
Tabelle 3: Legende state distribution plots und wichtigste Ausprägungen 25
Tabelle 4: Typische Verlaufsmuster und fachliche Interpretation 25
Tabelle 5: Verlaufsindikatoren nach Cluster, Kohorte 2010 (Mittelwerte oder Anteile pro Cluster)..... 29
Tabelle 6: Prädiktionsperformance des finalen Modells nach Cluster 33
Tabelle 7: State distribution plots für die Cluster «ALV-Kurzzeit» und «IV-Rente» für Kohorte 2010, 2011 und 2015 35
Tabelle 8: Clustergrössen nach Kohorte 36
Tabelle 9: Verlaufsindikatoren Kohorte 2010, initiale Clusterlösung..... 37
Tabelle 10: Verlaufsindikatoren Kohorte 2015, Prädiktion 37
Tabelle 11: Interne Masse für die initiale Clusterlösung (Kohorte K2010) sowie deren Prädiktion in den Kohorten 2011 und 2015 39
Tabelle 12: relative Häufigkeitsverteilung (y-Achse) der Zuordnungsgüte (x-Achse) für die Prädiktionen in den Kohorten 2011 und 2015..... 41
Tabelle 13: Dezilwerte der Zuordnungsgüte für die Prädiktionen der Clusterzugehörigkeit in den Kohorten K2011 und K2015..... 43
Tabelle 14: State distribution plots für die Referenz und neue Clusterlösungen für die Kohorten K2011 und K2015. 44
Tabelle 15: Konfusionsmatrix für die Cluster der Prädiktion und einer neuen Clusterlösung, Kohorte K2011 46
Tabelle 16: Jaccard Matrix für die Cluster der Prädiktion und einer neuen Clusterlösung, Kohorte 2011. 47
Tabelle 17 Zuordnungsergebnisse für die Cluster in Prädiktion und neuer Clusterlösung für die K2011 und K2015, sortiert nach Jaccard-Index K2011..... 48
Tabelle 18: Externe Masse für den Vergleich zwischen Prädiktion und einer neuen Clusterlösung für die Kohorten K2011 und K2015. 49
Tabelle 19: Generischer Analyseansatz für die Integration induktiver Verfahren bei der Analyse von Verlaufsdaten in der Statistikproduktion 57

8.2 Abbildungen im Haupttext

Abbildung 1: Übersicht zweistufiges Clustering 13
Abbildung 2: State distribution plot, gesamte Kohorte 2010 14
Abbildung 3: Schematische Darstellung der Übertragung der Initialen Clusterlösung (Referenz) auf zukünftige Kohorten 18
Abbildung 4: Schematische Darstellung des Vergleichs zwischen Prädiktion der Referenz auf der Kohorte 2011 (Hellblau) und einer neuen Clusterlösung für dieselbe Kohorte (Gelb). 20
Abbildung 5: Annäherung eines Elbow-Plots, Kohorte K2010 24
Abbildung 6: Performancevergleich der Prädiktionsmodelle (Training) anhand der Mean Balanced Accuracy. 32
Abbildung 7: Learning curves des finalen Modells (links: Zoom y-Achse, rechts: vollständige y-Achse) 34

Abbildung 8: Silhouette-Plot der initialen Clusterlösung (K2010). 40
Abbildung 9: Silhouette-Plots für die Prädiktionen der Referenz auf die Kohorten K2011 und K2015. 41

8.3 Tabellen im Anhang

Tabelle A 1: Kohorten insgesamt, SDPs und VIs, K2010-K2015 61
Tabelle A 2: Cluster 1 - ALV Kurzzeit, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015..... 63
Tabelle A 3: Cluster 2 - ALV Langzeit, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015..... 66
Tabelle A 4: Cluster 3 - Zwischenverdienst, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015 69
Tabelle A 5: Cluster 4 - ALV Mehrfach, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015..... 72
Tabelle A 6: Cluster 5 - IV-Rente, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015..... 75
Tabelle A 7: Cluster 6 - IV-Rente und Erwerb, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015 78
Tabelle A 8: Cluster 7 - Sozialhilfe und Erwerb, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015 81
Tabelle A 9: Cluster 8 - Sozialhilfe wiederholt, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015 84
Tabelle A 10: Cluster 9 - Sozialhilfe neu, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015 87
Tabelle A 11: Cluster 10 - Leavers, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015..... 90
Tabelle A 12: Cluster Neu 1, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015 93
Tabelle A 13: Cluster Neu 2, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015 95
Tabelle A 14: Cluster Neu 3, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015 97
Tabelle A 15: Cluster Neu 4, SDPs und VIs für Referenz, Prädiktion und neue Clusterlösung, K2010-2015 99
Tabelle A 16: Übersicht Training Prädiktionsmodelle 101
Tabelle A 17: Referenz K2010- Verlaufsindikatoren..... 107
Tabelle A 18: Neue Clusterlösung K2011- Verlaufsindikatoren..... 108
Tabelle A 19: Prädiktion K2011- Verlaufsindikatoren 108
Tabelle A 20: Neue Clusterlösung K2012- Verlaufsindikatoren..... 109
Tabelle A 21: Prädiktion K2011- Verlaufsindikatoren 109
Tabelle A 22: Neue Clusterlösung K2013- Verlaufsindikatoren..... 110
Tabelle A 23: Prädiktion K2013- Verlaufsindikatoren 110
Tabelle A 24: Neue Clusterlösung K2014- Verlaufsindikatoren..... 111
Tabelle A 25: Prädiktion K2014- Verlaufsindikatoren 111
Tabelle A 26: Neue Clusterlösung K2015- Verlaufsindikatoren..... 112
Tabelle A 27: Prädiktion K2011- Verlaufsindikatoren 112
Tabelle A 28: Interne Masse für die Referenz auf Basis der 3000 «Repräsentanten» des ersten Clustering-Schritts..... 113
Tabelle A 29: Interne Masse für eine initiale Clusterlösung mit vier Clustern, Kohorte 2010.....114

Tabelle A 30: Interne Masse für eine initiale Clusterlösung mit sechs Clustern, Kohorte 2010 115
Tabelle A 31: Interne Masse für eine initiale Clusterlösung mit sieben Clustern, Kohorte 2010 116
Tabelle A 32: Interne Masse für eine initiale Clusterlösung mit acht Clustern, Kohorte 2010 117
Tabelle A 33: Interne Masse für eine initiale Clusterlösung mit neun Clustern, Kohorte 2010 118
Tabelle A 34: Relative Häufigkeitsverteilung (Zeilenprozent) der Konfusionsmatrix für die Cluster der Prädiktion und einer neuen Clusterlösung, Kohorte K2011 120
Tabelle A 35: Relative Häufigkeitsverteilung (Spaltenprozent) der Konfusionsmatrix für die Cluster der Prädiktion und einer neuen Clusterlösung, Kohorte K2011 120
Tabelle A 36: Konfusionsmatrix für die Cluster der Prädiktion und einer neuen Clusterlösung, Kohorte K2015 121
Tabelle A 37: Relative Häufigkeitsverteilung (Zeilenprozent) der Konfusionsmatrix für die Cluster der Prädiktion und einer neuen Clusterlösung, Kohorte K2015 121
Tabelle A 38: Relative Häufigkeitsverteilung (Spaltenprozent) der Konfusionsmatrix für die Cluster der Prädiktion und einer neuen Clusterlösung, Kohorte K2015 121
Tabelle A 39: Jaccard Matrix für die Cluster der Prädiktion und einer neuen Clusterlösung, Kohorte 2015 122
Tabelle A 40: Berechnung der externen Masse zur Übereinstimmung von Prädiktion und neuer Clusterlösung, K2011 123
Tabelle A 41: Berechnung der externen Masse zur Übereinstimmung von Prädiktion und neuer Clusterlösung, K2015 124

8.4 Abbildungen im Anhang

Abbildung A 1: avNNet - Ergebnisse Training Prädiktionsmodell 103
Abbildung A 2: GBM - Ergebnisse Training Prädiktionsmodell 103
Abbildung A 3: knn - Ergebnisse Training Prädiktionsmodell 104
Abbildung A 4: ranger - Ergebnisse Training Prädiktionsmodell 105
Abbildung A 5: svmPoly - Ergebnisse Training Prädiktionsmodell 106
Abbildung A 6: Silhouette-Plot für eine initiale Clusterlösung mit vier Clustern, Kohorte 2010 114
Abbildung A 7: Silhouette-Plot für eine initiale Clusterlösung mit sechs Clustern, Kohorte 2010 115
Abbildung A 8: Silhouette-Plot für eine initiale Clusterlösung mit sieben Clustern, Kohorte 2010 116
Abbildung A 9: Silhouette-Plot für eine initiale Clusterlösung mit acht Clustern, Kohorte 2010 117
Abbildung A 10: Silhouette-Plot für eine initiale Clusterlösung mit neun Clustern, Kohorte 2010 118
Abbildung A 11: Verteilungsdichte und Quantile der Distanzen zwischen den Verläufen, K2010 119
Abbildung A 12: Verteilungsdichten und Quantile der Distanzen zwischen den Verläufen für eine Auswahl von Clustern der initialen Clusterlösung, K2010 119