Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Eidgenössisches Departement des Innern EDI
Bundesamt für Statistik BFS

Verläufe im System der Sozialen Sicherheit

Identifikation von Erwerbs- und Sozialleistungsverlaufsmuster durch unsupervised machine learning bei neuen Arbeitslosen

EXPERIMENTAL STATISTICS

Neuchâtel, 2019

Herausgeber:	Bundesamt für Statistik (BFS)	**Layoutkonzept:**	Sektion DIAM
Inhalt:	Luzius von Gunten, Nora Meister, Philippe Meyer, Thomas Ruch	**Download:**	www.statistik.ch
Themenbereich:	00 Statistische Grundlagen	**Copyright:**	BFS, Neuchâtel 2019
Originaltext:	Deutsch		Wiedergabe unter Angabe der Quelle für nichtkommerzielle Nutzung gestattet
Übersetzung:	Sprachdienste BFS		

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

Inhaltsverzeichnis

Executive Summary

Arbeitslosigkeit kann für die betroffenen Personen sehr unterschiedlich verlaufen. Die so entstehenden Verlaufsbiographien sind geprägt durch (wiederholte) Sozialleistungsbezüge aus dem System der Sozialen Sicherheit (Arbeitslosen-, Invalidenversicherung, Sozialhilfe), Erwerbsarbeit oder auch Rückzug aus dem Erwerbsleben. Im vorliegenden Projekt sind Angaben zu den individuellen Verläufen unter Anwendung induktiver statistischer Methoden analysiert worden, um typische Verlaufsmuster zu identifizieren. Dazu wurde der Datensatz SHIVALV+IK verwendet. Dieser enthält monatsgenaue Informationen zu individuellen Sozialleistungsbezügen aus der Sozialhilfe (SH), der Invalidenversicherung (IV) und der Arbeitslosenversicherung (ALV), sowie zur Erwerbstätigkeit (IK).
Als Grundgesamtheit diente eine Kohorte von Personen zwischen 18 und 64 Jahren, die im Jahr 2010 erstmals Taggelder der Arbeitslosenversicherung (ALV) bezogen hatten. Alle Informationen zum Bezug von Sozialleistungen sowie zur Erwerbstätigkeit während den folgenden 48 Monaten sind in die Analyse miteinbezogen worden.
Mit Sequence Clustering werden ähnliche Verläufe auf der Basis eines Algorithmus gruppiert und so typische Verlaufsmuster identifiziert, ohne diese im Vornherein zu kennen (unsupervised machine learning). Mittels diesem induktiven Verfahren haben sich 10 Cluster typischer Verlaufsmuster herausgebildet. Jedes dieser Cluster wurde anschliessend mit Verlaufsindikatoren und grafischen Darstellungen («State Distribution Plots», «Sankey Plots») vertieft analysiert und inhaltlich interpretiert.
Es zeichnen sich mehrere Cluster ab, in denen sich die Personen nach einer Phase der Arbeitslosigkeit wieder in den Arbeitsmarkt integrieren (Abb. 1; Cluster 3,7 und 6). Untereinander unterscheiden sich diese Cluster nach der Dauer der Arbeitslosigkeit. Deutlich wird auch die Ausdifferenzierung eines Clusters mit Zwischenverdienst in dem die durchschnittliche Dauer des Bezugs von ALV gesamthaft 22 Monate beträgt (Abb. 1; Cluster 9). Daneben entstehen Cluster, die klare Tendenzen entweder zum dauerhaften Bezug von IV-Renten oder Leistungen der Sozialhilfe zeigen (Abb. 1; Cluster 2 und 8, 10). In einem davon (Cluster 10) wird das Einkommen aus Erwerbstätigkeit durch Leistungen der Sozialhilfe ergänzt. Die durchschnittlichen Dauern von IV-Renten respektive des Bezugs von Leistungen der Sozialhilfe betragen in allen drei Clustern über 30 Monate. Ein anderes Cluster vereint diejenigen Personen, die während des Beobachtungszeitraumes dauerhaft aus den untersuchten Systemen hinausfallen (Abb. 1; Cluster 1). Unter anderem durch Teilnahme an einer Bildungsmassnahme wegen Aufnahme einer Ausbildung, Rückzug aus dem Arbeitsmarkt oder Ausreise. Daneben entstehen zwei kleinere Cluster, die heterogene Verläufe vereinen (Abb. 1; Cluster 5 und 4).
Das Projekt hat somit 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. Im weiteren Verlauf des Projektes werden Sensitivitätsanalysen für unterschiedliche Kohorten sowie einzelner Datenaufbereitungsschritte durchgeführt. Anschliessend wird geprüft, wie ein zeitlich stabiles Clustermodell auf neue Kohorten übertragen werden kann. In einem zweiten Projektteil soll mithilfe von Prädiktionsmodellen für jede Person in der Kohorte die Wahrscheinlichkeit für die Zugehörigkeit zu jedem der zehn Verlaufscluster geschätzt werden (supervised machine learning). Die Modelle werden auf denjenigen Kohorten entwickelt (trainiert), bei welchen als abhängige Variable die Clusterzugehörigkeit für jede Person bekannt ist. Diese Modelle können anschliessend auf Kohorten angewendet werden, deren Erstbezug von ALV weniger weit in der Vergangenheit liegt. Auf diese Art können – je nach Robustheit der Modelle – die voraussichtlichen Verläufe frühzeitig erkannt werden.

1 Zweck

Das vorliegende Dokument dient der zusammenfassenden Dokumentation der Ereignisse und Ergebnisse des Projekts „Machine Learning Soziale Sicherheit - ML_SoSi“ und hält die wichtigsten Schlüsse, welche aus der Projektarbeit gezogen werden können, für künftige Revisionen oder Folgeprojekte fest.

2 Zusammenfassung

Arbeitslosigkeit kann für die betroffenen Personen in Bezug auf die Inanspruchnahme staatlicher Angebote und/oder der Wiedereingliederung in den Arbeitsmarkt sehr unterschiedlich verlaufen. Die so entstehenden Verlaufsbiographien sind geprägt durch (wiederholte) Sozialleistungsbezüge aus dem System der Sozialen Sicherheit (Arbeitslosen-, Invalidenversicherung, Sozialhilfe), Erwerbsarbeit oder auch Rückzug aus dem Erwerbsleben. Mit dem vorliegenden Projekt wurde gezeigt, wie solche Verläufe unter Anwendung induktiver statistischer Methoden analysiert werden können.

Dazu wurde der Datensatz SHIVALV+IK verwendet. Dieser enthält monatsgenaue Informationen zu individuellen Sozialleistungsbezügen aus der Sozialhilfe (SH), der Invalidenversicherung (IV) und der Arbeitslosenversicherung (ALV), sowie zur Erwerbstätigkeit (IK). Mit sequence clustering ist eine Gruppierung der Verläufe vorgenommen worden. Es wurden 10 Cluster ausdifferenziert und anhand von Verlaufsindikatoren und grafischen Methoden wie «state distribution plots» und «sankey plots» wurde eine sowohl statistisch als auch inhaltlich valide Interpretation sozialer Verlaufsmuster (Cluster) erreicht. Damit wurde aufgezeigt, dass «Sequence Clustering» für die Analyse und Darstellung individueller Verlaufsdaten im Bereich der Sozialen Sicherheit einen erheblichen Mehrwert erbringt.

Voraussetzung für die Anwendung dieses Verfahrens sind umfangreiche konzeptionelle Überlegungen zum Datenmodell und der Analysestrategie. Für deren Umsetzung sind vertiefte Kenntnisse einschlägiger Methoden und Programmiertechniken notwendig. Das Projektteam konnte sich diese einerseits in den vom BFS organsierten Kursen aneignen und sie «on-the-job», d.h. während der Projektbearbeitung vertiefen.

Für ein produktives Umfeld konnten wichtige Erkenntnisse erzielt werden: Neben dem Nachweis des Mehrwerts induktiver statistischer Methoden und von Visualisierungsmethoden wurde ein Datenmodell (sowie die Schritte für dessen Erstellung) entwickelt, das eine effiziente Analyse der Daten ermöglicht. Zudem konnte die Vorteile der Kohortensichtweise, einer standardisierten Beobachtungsdauer von Verläufen und die Vorteile von darauf aufbauenden Verlaufsindikatoren aufgezeigt werden.

Die zur Verfügung gestellte Infrastruktur (Programm «R»; Cloud «Atlantica») kam leistungsmässig an ihre Grenzen, was eine konzeptionell komplexere und womöglich weniger effiziente Modellierung zur Folge hatte.

Die angewandte agile Arbeitsweise mit Intensivworkshops hat sich im gegebenen methodischen Setting sowohl in Bezug auf die Zielerreichung und den Teamgeist positiv ausgewirkt. Voraussetzung ist jedoch eine kompetente Führung. Zentral ist auch eine laufende, nachvollziehbare Dokumentation.

3 Ausgangslage und Zielsetzungen

Das Pilotprojekt «Machine Learning SoSi» im Rahmen der «Data Innovation Strategie» des BFS konzentriert sich inhaltlich auf die Integrationserfolge von Sozialleistungsbezügerinnen. Denn die nachhaltige (Re-)Integration aller Bevölkerungsgruppen in den Arbeitsmarkt stellt die **zentrale sozialpolitische Herausforderung** der nächsten Jahre dar. Angesichts der Tendenzen sozialer Entgrenzung von Armut und der Zunahme diskontinuierlicher Bildungs- und Erwerbsbiographien einerseits, und der

Fragmentierung staatlicher Unterstützungsangebote zur Existenzsicherung und (Wieder-)Eingliederung in den Arbeitsmarkt im System der Sozialen Sicherheit andererseits, kommt einer **verlaufsorientierten Perspektive** der Arbeitsmarktintegration eine immer grössere Bedeutung zu. Nur so lassen sich soziale Verlaufsmuster und das Zusammenspiel staatlicher Unterstützungsangebote in ihrer Komplexität untersuchen.

Arbeitslosigkeit ist für viele Personen Teil ihrer Erwerbsbiografie geworden und steht oft am Anfang von sozialen Abstiegsprozessen. Entsprechend hoch ist die Bedeutung der Arbeitslosenversicherung für die soziale Absicherung der Bevölkerung: rund 11% der Erwerbsbevölkerung ist in einer Dreijahresperiode auf Taggelder der Arbeitslosenversicherung (ALV) angewiesen[1]; sie gilt als Eintrittssystem in das System der Sozialen Sicherheit. Kohorten von neu arbeitslosgewordenen Personen, die Taggelder der ALV beziehen, bilden deshalb die Grundgesamtheiten im vorliegenden Projekt (siehe Abschnitt 4.2). Die Wege aus der ALV zurück in die Erwerbsarbeit sind zu einem grossen Teil nicht geradlinig, sondern können durch weitere Sozialleistungsbezüge, erneute Arbeitslosigkeit oder (zeitweiligen) Rückzug aus dem Arbeitsmarkt gekennzeichnet sein. Aus diesem Grund werden neben den Verläufen der Arbeitslosentaggelder und der Erwerbsverläufe auch die Bezugsverläufe der Invalidenversicherung (IV) und der Sozialhilfe (SH) berücksichtigt. Zusammen mit der ALV bezwecken die IV und die SH die Existenzsicherung bei Verlust des Erwerbseinkommens und in Notlagen und bilden den Grossteil des Schweizerischen Systems der Sozialen Sicherheit für die Erwerbsbevölkerung. Die entsprechenden Verläufe sind im SHIVALV+IK-Datensatz abgebildet (siehe Abschnitt 4.1).

Eine zentrale Aufgabe im vorliegenden Projekt ist es, einen Analyseansatz zu entwickeln, um steuerungsrelevante Informationen nutzbar zu machen. Zu diesem Zweck wird ein zweistufiger Analyseansatz gewählt: in einem ersten Schritt werden Verlaufsmuster aus den Daten extrahiert. Damit gewinnt man grundlegende Erkenntnisse über typische Verläufe ab Beginn der Arbeitslosigkeit. In einem zweiten Schritt werden statistische Modelle zur Vorhersage solcher Verlaufsmuster erstellt. Auf dieser Basis lassen sich sozialpolitische Frühwarnindikatoren entwickeln. Zu diesen Zwecken werden insbesondere induktive statistische Verfahren aus dem Bereich des Machine Learning angewendet. Damit lassen sich die Identifikation von Verlaufsmustern und deren Vorhersagen möglichst nahe an den Daten realisieren, ohne diese durch deduktive, theoretische Festlegungen zu beeinflussen (der zweite Schritt konnte noch nicht umgesetzt werden). Zudem handelt es sich hier als Pilotprojekt im Rahmen der Data Innovation Strategie des Bundesamts für Statistik, die einen Fokus auf alternative Verfahren, wie das Machine Learning setzt. Details zum Vorgehen und dessen Begründung finden sich in Abschnitt 4.2.

Die leitenden Fragestellungen zum Projekt lauten: Wie können komplexe soziale Verlaufsmuster im Bereich der Sozialen Sicherheit zielgerichtet analysiert und sinnvoll beschrieben werden? Wie exakt können mit bestehenden Daten die individuellen Verlaufsmuster vorhergesagt werden? Inwieweit können mit induktiven statistischen Methoden für Politik und Verwaltung steuerungsrelevante Informationen, Indikatoren und Analysen erarbeitet werden?

Im Vordergrund stehen demnach folgende Zielsetzungen:

- Darstellung der typischen Erwerbs- und Bezugsverläufe von Personen die Leistungen aus dem System der sozialen Sicherung beziehen, unter Anwendung der Methode des sequence clustering.
- Definition zentraler, aussagekräftiger Indikatoren und Visualisierungen um die Komplexität individueller Verläufe synthetisiert darzustellen.
- Datengetriebenes sequence clustering, um ähnliche Verlaufsmuster zusammenzufassen (unsupervised machine learning).

[1] 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.

- Identifizierung von zentralen Aspekten für den Transfer der Projekterkenntnisse in ein produktives Umfeld.

Das letztere Ziel ist insofern von Bedeutung, da das BFS in Zukunft voraussichtlich standardmässig Kennzahlen zu Beständen, Ab- und Übergängen aus dem System SHIVALV (Sozialhilfe, Invalidenversicherung, Arbeitslosenversicherung) unter dem Titel «Soziale Verlaufsmuster» veröffentlichen wird. Zudem legt das Modernisierungsprojekt der Sozialhilfestatistik einen Fokus auf Verlaufsanalysen. Erkenntnisse aus dem Projekt hinsichtlich der Datenaufbereitung und Visualisierungen bieten Synergien für diese Arbeiten.

Die detaillierten Projektziele und eine Einschätzung der Zielerreichung findet sich im Anhang, Abschnitt 9.5.

4 Vorgehen

4.1 Daten

Für das Pilotprojekt wird der Datensatz SHIVALV 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. Somit enthält die im Projekt verwendete Datenbasis SHIVALV+IK pro Person in der Grundgesamtheit, pro Monat im Beobachtungszeitraum und pro betrachtetes System (SH, IV, ALV, IK/Erwerb) die Information, ob ein Leistungsbezug bzw. Erwerbsarbeit vorliegt oder nicht. Aus den vier Grundzuständen respektive Stati ergeben sich 16 mögliche Zustands- oder Statuskombinationen (SH+IV; SH+ALV; SH+IK; ALV+IV, usw.; vgl. Tabelle 1 auf S. 11) für jeden Monat. Zusätzlich stehen pro Jahr und Person soziodemografische und -professionelle Variablen zur Verfügung.

Der Datensatz wird vom Bundesamt für Sozialversicherungen (BSV) aufbereitet und enthält die zeitlichen Bezugs- und Erwerbsverläufe aller Personen zw. 18 und 65 Jahren, die in einem betrachteten Jahr eine Sozialleistung erhalten haben und/oder erwerbstätig waren (Vollerhebung). Für dieses Projekt wurden die SHIVALV+IK-Daten für den Zeitraum zwischen Jahren 2010-2016 ausgewertet.

Die Zweckmässigkeit von SHIVALV+IK für die Analyse von Verläufen im System der Sozialen Sicherheit konnte in unterschiedlichen Studien aufgezeigt werden (siehe z.B. Fluder et. al. 2009[1], Fritschi et al. 2013[2], Fluder et al. 2017[3]).

Die Aufbereitung und Bereinigung der Daten konzentriert sich auf mögliche fehlende Werte, unbeabsichtigte Überschneidungen oder Lücken auf der Zeitachse sowie auf die Datenstruktur. Die Struktur der Daten ist so zu wählen, dass am Ende ein Basisdatensatz vorhanden ist, der Ausgangspunkt für möglichst alle durchzuführenden Analyseschritte darstellt. Eine vertiefte Darstellung der Datenaufbereitung für das Projekt findet sich im Anhang, Abschnitt 9.2.

[2] 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.

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

4.2 Analysestrategie

Das Vorgehen umfasst vier Schritte: (1) Abgrenzung und Beschreibung der Grundgesamtheiten (Kohorten), (2) sequence clustering der Verläufe, (3) Deskription der Cluster anhand von Verlaufsindikatoren, (4) Prädiktion der Clusterzugehörigkeit neuer Kohorten.

Machine Learning

Bei deduktiven Verfahren werden Annahmen über die Welt (Theorien, Hypothesen) anhand geeigneter Daten in statistischen Modellen überprüft (Falsifikation). Induktive statistische Verfahren funktionieren in umgekehrter Weise: Daten sind ein Abbild der Welt und ohne Vorkenntnisse darüber werden durch geeignete Algorithmen die darin enthaltenen Strukturen und Zusammenhänge in Form von statistischen Modellen möglichst genau abgebildet. Daraus lassen sich Erkenntnisse über die Welt erzielen (Hypothesen, Theorien). Induktive Verfahren der Statistik können dem Machine Learning zugeordnet werden.

Man kann zwei Formen des Machine Learnings unterscheiden, supervised und unsupervised machine learning. Bei Letzteren handelt es sich um Verfahren zur Mustererkennung in Daten, ohne dass die resultierenden Muster im Vornherein bekannt sind; z.B. können Supermarktkunden anhand ihrer präferierten Produkte, Ausgaben pro Einkauf und Einkaufsfrequenz in Kundensegmente unterteilt werden.

Beim supervised machine learning ist das Resultat (abhängige Variable) im Vornherein bekannt und die Verfahren passen anhand aller verfügbarer Daten (unabhängige Verfahren) Modelle an, um das Resultat möglichst präzise vorherzusagen; im Unterschied zur klassischen Statistik wird hier die Modellbildung (z.B. Variablenselektion, Interaktionen, Gewichte) einem geeigneten Algorithmus überlassen. Er profitiert von möglichst vielen Daten. Die Modelle können dann auf neue Datenpunkte angewendet werden, bei welchen die abhängige Variable noch nicht bekannt ist.

Grundgesamtheit, Kohorte und Beobachtungsdauer (1)

Die Frage, wie sich die Verläufe von arbeitslos gewordenen Personen durch das System der Sozialen Sicherheit hindurch gestalten, steht im Zentrum des Projekts. Die Grundgesamtheit für die Analyse umfasst daher eine Kohorte aller neu arbeitslos gewordene Personen in einem bestimmten Jahr. Um Fehlerquellen für die Interpretation der Resultate auszuschalten, wird mit folgenden Kriterien eine homogene Definition der Kohorte angestrebt:

- Für Personen in der Kohorte wurde im Referenzjahr eine formelle Anmeldung als arbeitslose Person und somit eine Rahmenfrist[4] eröffnet,
- sie weisen einen Taggeldbezug nach Eröffnung der Rahmenfrist auf,
- in den 24 Monaten vor dem ersten Taggeldbezug haben sie nicht bereits Taggelder bezogen (siehe auch Anhang, Abschnitt 9.2) und
- sie erreichen im Beobachtungszeitraum das Rentenalter nicht.

Mit der Kohortensichtweise lässt sich die Komplexität von Verlaufsdaten stark reduzieren, indem die Bedeutung der Linkszensierung reduziert wird.

Um individuelle Verläufe zu analysieren, ist es wichtig, dass alle Personen in der Grundgesamtheit über eine gleich lange Dauer beobachtet werden; so stellt man sicher, dass alle Kohortenmitglieder

[4] Die Rahmenfrist bezeichnet den Zeitraum von 2 Jahren vor und nach der Anmeldung. Ab dem Referenzdatum der Rahmenfrist können maximal während 2 Jahren ALV-Taggelder bezogen werden. Um überhaupt einen Anspruch zu begründen, müssen im Zeitraum von 2 Jahren vor dem Referenzdatum genügend Beiträge einbezahlt worden sein; ausgenommen sind Personen, die aus bestimmten Gründen von der Beitragspflicht befreit waren.

die gleiche zeitliche Grundwahrscheinlichkeit für nachfolgende Ereignisse (z.B. Sozialhilfebezug) aufweisen. Der Startpunkt des Beobachtungszeitraums für die Kohorte ist der erste Taggeldbezug und die Beobachtungsdauer ist 48 Monate; in anderen Studien wurde aufgezeigt, dass 48 Monate aufgrund der zur Verfügung stehenden Zeitreihen eine optimale Beobachtungsdauer ist[3].

Im Projekt werden Kohorten neuer Arbeitslosen für die Jahre 2010 und 2011 gebildet und zur Überprüfung der Plausibilität anhand soziodemografischer Merkmale mit der Bevölkerung verglichen.

Sequence Clustering (2)

Um typische Muster von sozialen Verläufen in der Sozialhilfe der Invaliden- und Arbeitslosenversicherung und in der Erwerbsarbeit von neuen Arbeitslosen zu untersuchen, wird ein rein induktiver statistischer Analyseansatz gewählt. Analyseansätze, in denen Verläufe theoretisch kategorisiert und als Verlaufstypologie ausgewiesen wurden (deduktives Vorgehen), wurden mehrfach erfolgreich erprobt [3,5]. Bei einem deduktiven Vorgehen läuft man jedoch Gefahr, neue und unerwartete Verlaufsmuster nicht zu erkennen und die Relevanz von Verlaufsmuster rein theoretisch festzulegen. Induktive Methoden stützen sich hingegen allein auf die in den Daten abgebildeten Strukturen und gruppieren Verläufe nach ihrer Ähnlichkeit. Damit wird eine wertneutrale Beschreibung der Verlaufsmuster unterstützt und Raum gelassen für die Beschreibung unerwarteter Verlaufsmuster. Dies bedingt jedoch, dass die Nachverfolgung von gewissen Verlaufsmustern erschwert werden können. Methoden des sequence clustering können dem unsupervised machine learning zugeordnet werden.

Der Clusteralgorithmus wurde im Projekt aufgrund der begrenzten Leistungsfähigkeit der Infrastruktur in folgende zwei Schritte aufgeteilt: In einem ersten Schritt werden die Verläufe der Sozialleistungsbezüge und die Erwerbsverläufe der untersuchten Kohorten mit dem effizienten k-means-Algorithmus in 3000 Gruppen eingeteilt. Für jede dieser Gruppen wird ein repräsentativer bzw. der mittlere Verlauf (vom Typ eines Medians) bestimmt. Auf der Basis der 3000 mittleren Verläufe wird im zweiten Schritt das rechenintensive pairwise-optimal-matching-Verfahren angewendet, um die definitiven Verlaufscluster zu bestimmen.

Deskription der Cluster anhand von Verlaufsindikatoren (3)

Die inhaltliche Interpretation und Beschreibung einer gewählten Clusterlösung ist eine zentrale Aufgabe bei der Clusteranalyse. Zu diesem Zweck liegt ein Fokus des Projekts auf dem Austesten verschiedener Methoden der Datenvisualisierung für Verlaufsdaten, insbesondere state distribution plots. Dabei handelt es sich um die relative Verteilung aller 16 möglichen Verlaufszustände pro Monat, betrachtet über die 48 monatige Beobachtungsdauer.

Zudem werden Verlaufsindikatoren entwickelt, welche Teilaspekte der Verläufe abbilden und zur inhaltlichen Qualifizierung der Cluster dienen. Z.B. die Anzahl Bezugsperioden von Arbeitslosentaggelder oder die Anzahl Monate mit Sozialhilfeunterstützung.

Durch die deskriptive Analyse können die Cluster beschrieben, inhaltlich eingeordnet und gelabelt werden.

Prädiktion der Clusterzugehörigkeit (4)

Dieser Analyseschritt wurde im Pilotprojekt bisher nicht umgesetzt und wird in den Resultaten nicht ausgewiesen. Der folgende Abschnitt gibt jedoch einen Überblick über das Vorhaben.

Als letzter Analyseschritt sollen mithilfe von Prädiktionsmodellen die Wahrscheinlichkeit für die Zugehörigkeit zu jedem der zehn Verlaufscluster geschätzt werden. Die Modelle werden auf den Kohorten 2011 und 2010 entwickelt (trainiert), bei welchen die Clusterzugehörigkeit für jede Person in der Kohorte bekannt ist (abhängige Variable). Dieses Modell kann anschliessend auf neue Kohorten von Arbeitslosen angewendet werden, um neue Entwicklungen in den Verlaufsmustern frühzeitig erkennen

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

zu können. So liesse sich z.B. der Anteil der Personen in der Kohorte 2018 schätzen, die einem Verlaufscluster angehören, welches vornehmlich durch Sozialhilfebezug charakterisiert ist; diese Information kann als sozialpolitischer Frühwarnindikator dienen.

Für diese Schätzungen wird ein «supervised machine learning»-Ansatz angewendet: Für die Kohorte 2010 (und auch 2011) kennt man die Clusterzugehörigkeit (abhängige Variable). Die Algorithmen passen ein Modell auf alle verfügbaren unabhängigen Variablen an, sodass eine möglichst hohe Prädiktionsgüte erreicht wird. Als unabhängige Variablen dienen alle Informationen, die zum Start des Beobachtungszeitraums (erster Taggeldbezug) zur Verfügung stehen. Zudem können auch Informationen verwendet werden, die diesem Zeitpunkt vorausgehen («retrospektive» Verläufe) oder die für diesen Zeitpunkt aus weiteren Datenquellen gewonnen werden können (Verknüpfungen). Im Gegensatz zur klassischen deduktiven statistischen Modellierung, wird die Modellspezifikation (Formen des Modells) dem Algorithmus überlassen.

Für die Modellbildung müssen voraussichtlich mithilfe systematischen Experimentierens eine Reihe von Entscheidungen getroffen werden, u.a. bezüglich der Wahl des Algorithmus und Optimierungsverfahrens, der Kriterien zur Einschätzung der Modellgüte, dem Umgang mit unterschiedlichen Clustergrössen, der Gewichtung der Fehlerraten.

4.3 Agile Arbeitsweise

Um dem innovativen Charakter der Arbeiten sicherzustellen und die geplanten Ressourcen möglichst effizient einzusetzen, wurde eine agile Arbeitsweise gewählt: in zweitägigen Intensivworkshops arbeitete das ganze Team gemeinsam in einem Raum an den Projektaufgaben. Durch die räumliche Nähe sind die Kommunikationswege kurz, auf offene Fragen kann rasch eingegangen werden und die Iterationen zwischen Statistik und Interpretation können rasch durchlaufen werden. Zudem eignet sich diese Arbeitsweise gut für sektionsübergreifende Zusammenarbeit. Voraussetzung bildet eine intensive Vor- und Nachbereitung der Workshops. Aufgebaut wurden die Workshops alle nach demselben Basisablauf:

- Zieldefinition und Verantwortungsaufteilung im Plenum zu Beginn
- ca. dreimal pro Workshop Präsentation der Zwischenresultate im Plenum für das Alignment der Aktivitäten
- Einschätzung der Zielerreichung in Prozent am Schluss
- und Festlegung der Pendenzen für die Zwischenzeit bzw. den nächsten Workshop.

5 Ergebnisse Pilotprojekt

5.1 Beschreibung der Kohorten

Erste Resultate liefert die Beschreibung der Kohorten nach verschiedenen Merkmalen. Die Kohorte 2010 besteht aus 126'359 Personen. Davon sind knapp über 60% Schweizer, etwas über die Hälfte (51%) Männer und das Durchschnittsalter beträgt 35.8 Jahre. Im Vergleich zur ständigen Wohnbevölkerung 2010 sind Ausländer also übervertreten (39% in der Kohorte, 23% in der gesamten Schweiz) und das Durchschnittsalter ist tiefer als das Durchschnittalter der Erwerbsbevölkerung, das im Jahr 2010 bei 40.7 Jahren lag. Aufgrund der Verteilung des Arbeitslosigkeitsrisikos in der Bevölkerung entsprechen diese Resultate den Erwartungen.

Die Unterschiede zwischen den Kohorten 2010 und 2011 sind minim. Die Kohorte 2011 ist kleiner (111'650 gegenüber den 126'359 Personen der Kohorte 2010), weist aber hinsichtlich der Zusammensetzung nur unwesentliche Unterschiede auf.

5.2 Wahl der Anzahl Cluster

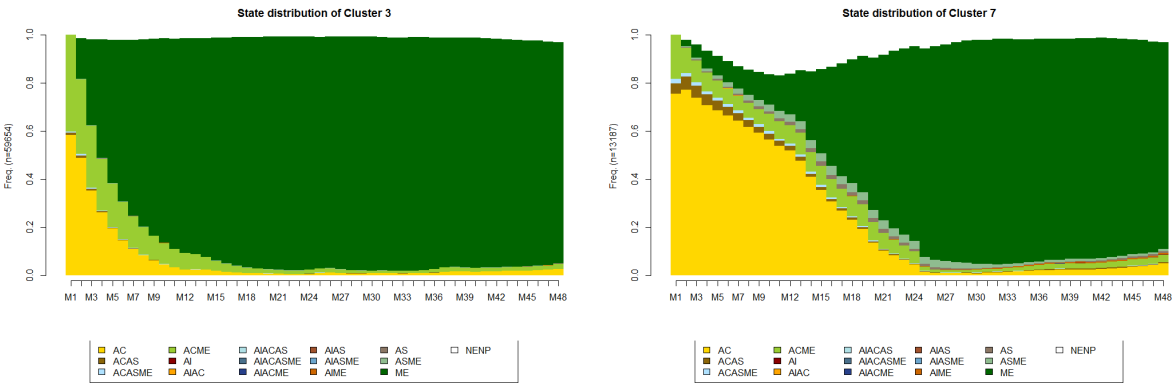
Das Projektteam kam zum Schluss, dass aufgrund statistischer und fachlicher Kriterien die Wahl von **zehn Cluster** am besten geeignet sind, um die typischen Verläufe im System der Sozialen Sicherheit zu beschreiben. Dazu wurden einerseits statistische Kriterien ausgewertet, wie die interne Heterogenität der Cluster (within sum of squares) z.B. anhand der Elbow-Methode. Andererseits wurden die inhaltliche Plausibilität der Clusterlösung anhand der «State-Distributions-Plots» pro Cluster beurteilt. Eine vertiefte Darstellung findet sich im Anhang, Abschnitt 9.4.

5.3 Interpretation der definitiven Clusterlösung

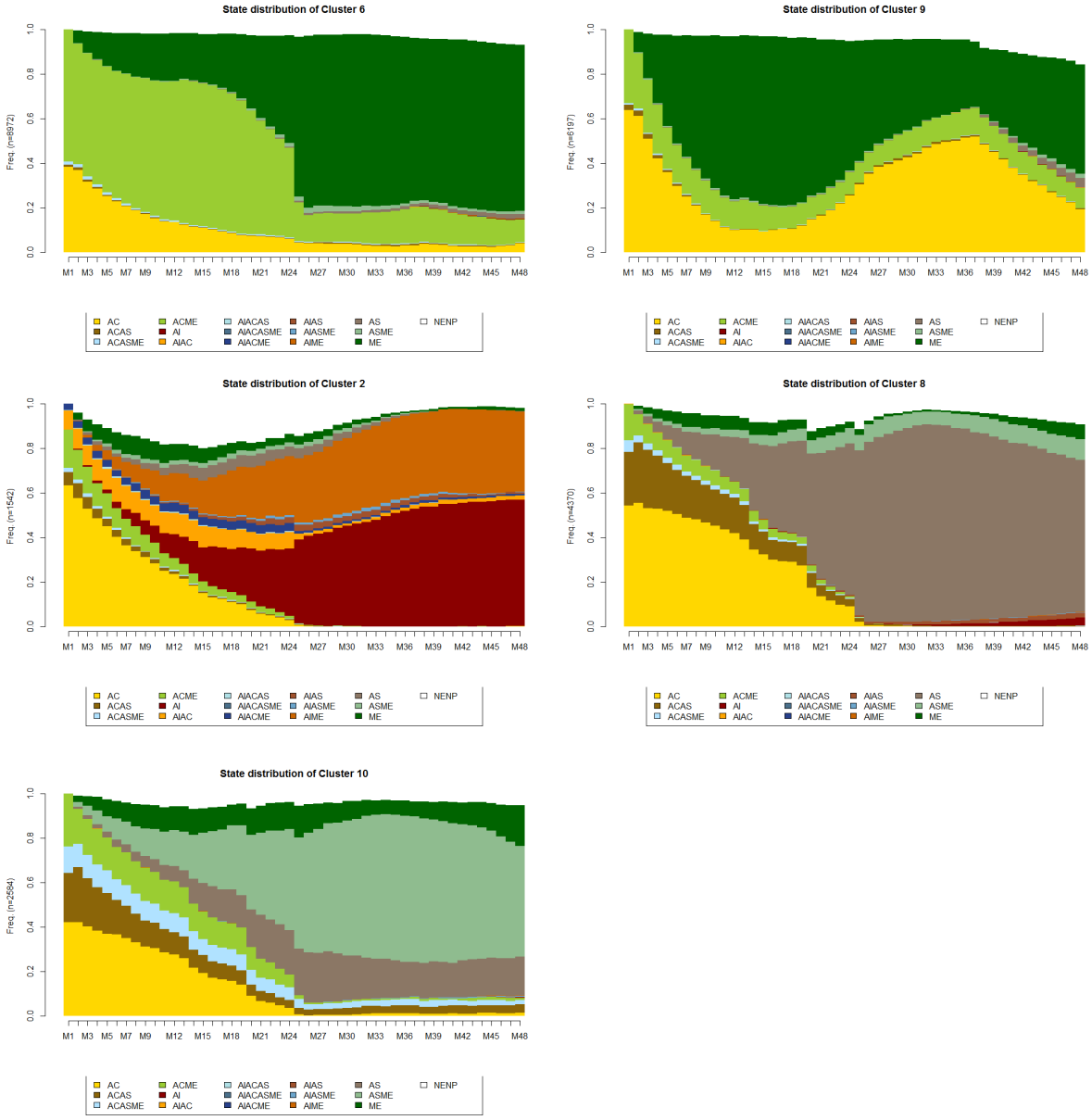
Anhand der **«State-Distributions-Plots»** für eine Lösung mit zehn Clustern lassen sich in erster Sichtung sieben Cluster relativ gut interpretieren. Nachfolgende Interpretationen der Cluster der Kohorte 2010 beziehen sich nicht auf die einzelnen Verläufe, sondern auf die Beobachtung der Verteilung der Zustände der Clusterelemente über die Beobachtungsdauer. Cluster (C) 3, 7, 6 und 9 beziehen sich fast ausschliesslich auf den Bezug von ALV und Erwerbsarbeit (knapp 70% der ganzen Kohorte). C3 beschreibt Verläufe mit einer Integration in den Arbeitsmarkt nach kurzer Arbeitslosigkeit und stellt mit 47% aller Personen in der Kohorte den «Normalverlauf» dar. C7 vereint ca. 10% aller Personen und zeigt ein ähnliches Bild, jedoch dauert dort die Arbeitslosigkeit deutlich länger. C6 vereint Verläufe, die vor allem durch Teilzeitarbeitslosigkeit geprägt sind (ca. 7% aller Personen). C9 beschreibt Verläufe bei denen sich Bezugsperioden von ALV und Zwischenverdienst abwechseln (ca. 5%) aller Personen.

Bei Cluster 2, 8 und 10 kommen die beiden anderen Sicherungssysteme, die Invalidenversicherung und die Sozialhilfe ins Spiel (knapp 7% der Kohorte). C2 vereint alle Verläufe, die letztendlich in die Invalidenversicherung führen (ca. 1% der Kohorte). Zudem ist hier die Rolle der Sozialhilfe ersichtlich, die oft bei der Überbrückung der Zeit zwischen IV-Anmeldung und IV-Entscheid einspringt. C8 beschreibt ein Verlaufsmuster, das durch den alleinigen Bezug von Sozialhilfe gekennzeichnet ist (3.5% der Kohorte). Wie in anderen Cluster auch (C7, C6), sind hier deutlich Aussteuerungseffekte[6] nach 24 Monaten zu sehen. Das vorherrschende Muster in C10 (2% der Kohorte) ist der Bezug von Sozialhilfe bei gleichzeitiger Erwerbsarbeit. Neben den Aussteuerungseffekten ist auch die relativ hohe Heterogenität dieser Verläufe gut abgebildet.

Abbildung 1: state distribution plots für sieben gut interpretierbare Cluster, Kohorte 2010



[6] in der Regel erlischt der Anspruch auf ALV nach 400 Taggelder oder 24 Monaten



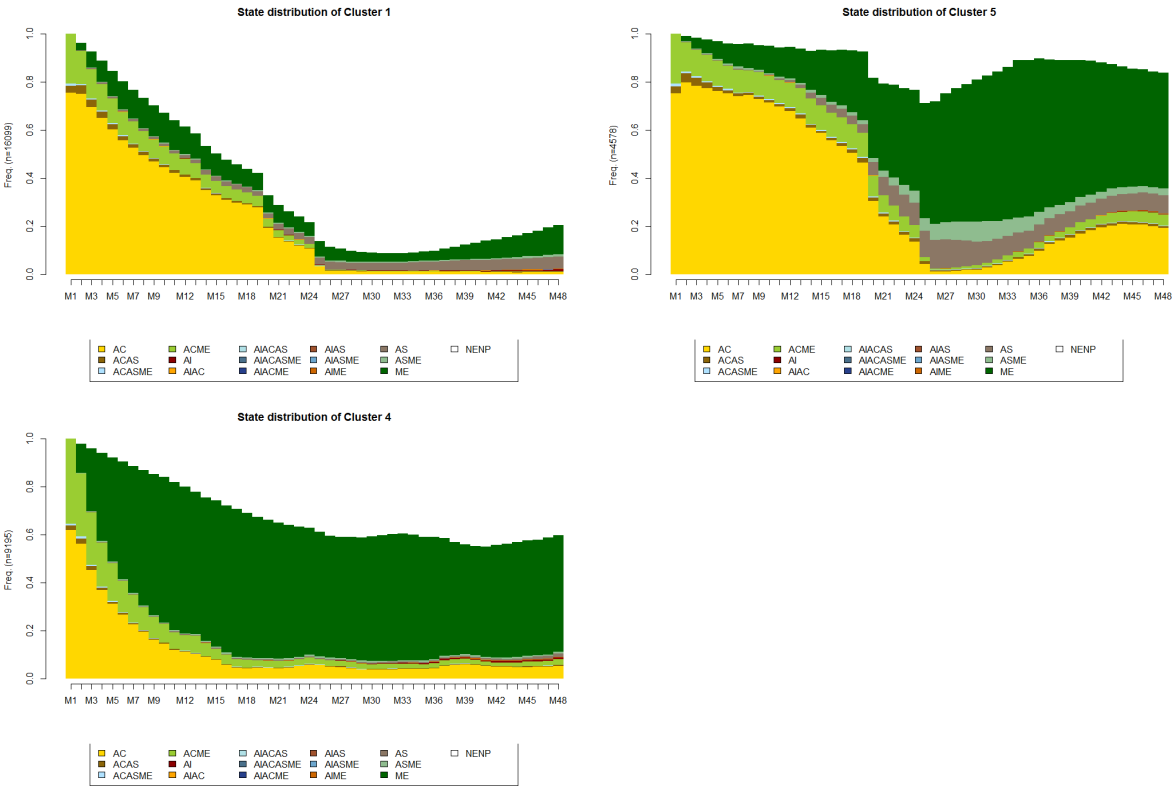
Quelle: SHIVALV-IK 2010-2015, eigene Berechnungen

Tabelle 1: Legende Status/-kombinationen

AC	=	Arbeitslosenversicherung (ALV)
ACME	=	Arbeitslosenversicherung (ALV) + Erwerbstätigkeit
AIACAS	=	Invalidenversicherung (IV) + Arbeitslosenversicherung (ALV) + Sozialhilfe (SH)
AIAS	=	Invalidenversicherung (IV) + Sozialhilfe (SH)
AS	=	Sozialhilfe (SH)
ACAS	=	Arbeitslosenversicherung (ALV) + Sozialhilfe (SH)
AI	=	Invalidenversicherung (IV)
AIACASME	=	Invalidenversicherung (IV) + Arbeitslosenversicherung (ALV) + Sozialhilfe (SH) + Erwerbstätigkeit
AIASME	=	Invalidenversicherung (IV) + Sozialhilfe (SH) + Erwerbstätigkeit
ASME	=	Sozialhilfe (SH) + Erwerbstätigkeit
ACASME	=	Arbeitslosenversicherung (ALV) + Sozialhilfe (SH) + Erwerbstätigkeit
AIAC	=	Invalidenversicherung (IV) + Arbeitslosenversicherung (ALV)
AIACME	=	Invalidenversicherung (IV) + Arbeitslosenversicherung (ALV) + Erwerbstätigkeit
AIME	=	Invalidenversicherung (IV) + Erwerbstätigkeit
ME	=	Erwebstätigkeit
NENP	=	keine Erwerbstätigkeit + keine Sozialleistung

Drei Cluster (C1, C4 und C5) sind inhaltlich weniger gut interpretierbar. Die Ausdifferenzierung dieser Cluster findet relativ früh statt, so dass sie nicht durch die Wahl einer geringeren Anzahl an Cluster verhindert werden können. C1 beschreiben Verläufe von Personen, die nach einer relativ langen Bezugsperiode von ALV-Taggelder nicht mehr in den untersuchten Systemen auftauchen. Entweder haben sie sich aus der Erwerbsarbeit zurückgezogen, weil sie sich ihren Lebensbedarf anderweitig sichern können (z.B. durch eine vorgezogene Altersrente oder indem ein anderes Familienmitglied genügend Ressourcen aufbringt) oder befinden sich in Ausbildung. Auch möglich ist, dass sie ausgereist oder verstorben sind. Mit beinahe 13% der Kohortenmitglieder handelt es sich um ein relativ grosses Cluster. C5 beschreibt ebenfalls Verläufe mit langen Bezugsdauern von Arbeitslosentaggelder. Nach dieser Phase treten heterogene Verlaufsmuster mit Erwerbsarbeit, Sozialhilfe, Rückzug aus dem Arbeitsmarkt/Ausreise und erneuter Arbeitslosigkeit auf. Die Interpretationsschwierigkeiten können anhand der Verlaufsindikatoren (siehe unten) behoben werden. Mit 3.6% der Kohorte handelt es sich um eine kleine Gruppe. C4 hingegen vereint Verläufe mit kurzer Bezugsdauer von Arbeitslosentaggeldern und anschliessend mutmasslich heterogene Verläufe mit Erwerbsarbeit und Rückzug von der Erwerbsarbeit/Ausreise. C4 ist mit 7.4% der Kohortenmitglieder eine mittelgrosse Gruppe.

Abbildung 2: state distribution plots für drei schwierig interpretierbare Cluster, Kohorte 2010

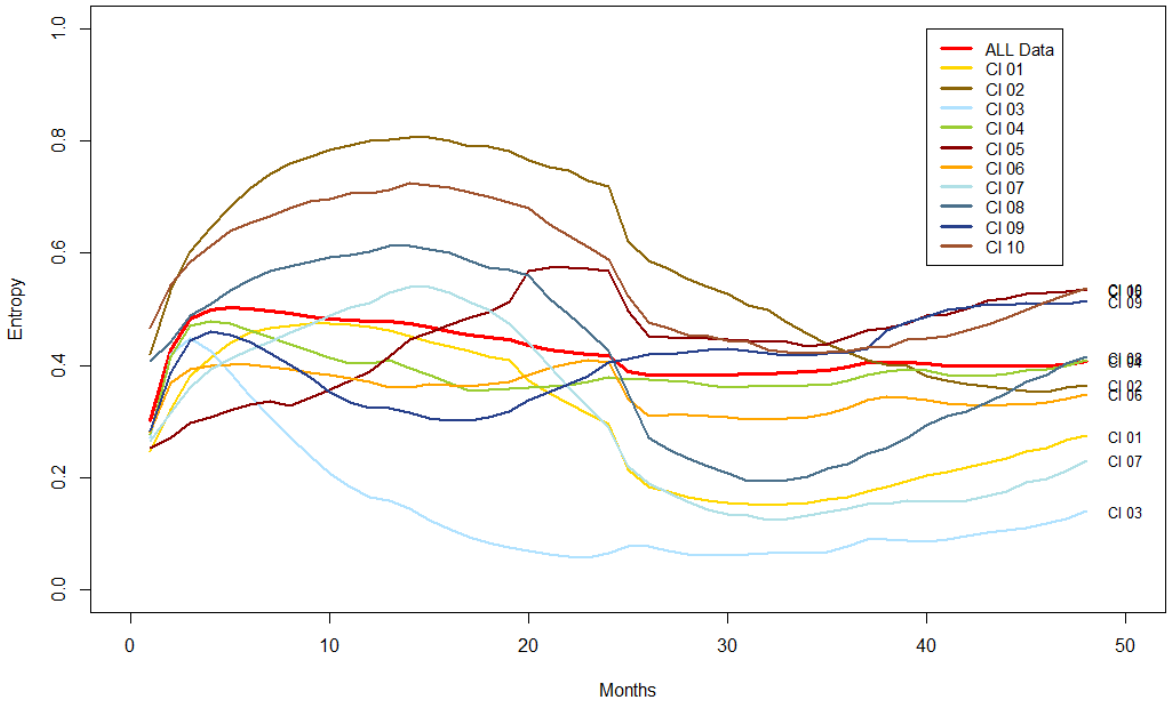


Quelle: SHIVALV-IK 2010-2015, eigene Berechnungen

Die **Heterogenität eines Verlaufsclusters** kann man nicht nur für das Cluster insgesamt, sondern auch zu jedem Zeitpunkt innerhalb des Clusters berechnen. Dazu eignet sich die «transversal entropy», die auf den Häufigkeitsverteilungen der Zustände zu einem bestimmten Verlaufszeitpunkt (Monat) beruht. Je höher die Entropie desto höher die Heterogenität. Abbildung 3 zeigt die «Entropieverläufe» für die ganze Kohorte und für jedes Cluster einzeln aus.

Ausgehend von einer homogenen Ausgangslage – alle sind neu arbeitslos – entwickelt sich die Heterogenität in den einzelnen Clustern sehr unterschiedlich (z.B. C2, C3 und C9). Bei Clustern mit einem ähnlichen «Verlaufsmuster» der Entropie, ist deren Level jedoch sehr stark unterschiedlich ausgeprägt (z.B. C10, C8 und C7). Schlussendlich zeigen sich in allen Clustern klare Abweichungen zum «durchschnittlichen Entropieverlauf» für die ganze Kohorte. Dies zeigt auf, dass die Cluster unterschiedliche «Geschichten» erzählen und damit eine gute inhaltliche Separierung gelungen ist.

Abbildung 3: Transversal Entropy für die Kohorte und für alle Cluster nach Monat, Kohorte 2010



Quelle: SHIVALV-IK 2010-2015, eigene Berechnungen

Die «State Distribution Plots» zeigen die Häufigkeitsverteilung der vorkommenden Verlaufszustände zu einem bestimmten Zeitpunkt und geben, wie oben erwähnt, die individuellen Verläufe in einem bestimmten Cluster nur ungenau wieder. Um die inhaltliche Validität der Cluster weiter zu prüfen werden deshalb **Verlaufsindikatoren** gebildet und deren Mittelwerte pro Cluster ausgewiesen. Zum Beispiel die Anzahl Monate mit Arbeitslosenversicherungstaggelder (ALV) pro Person im Cluster, aufsummiert über und gemittelt durch alle Personen im Cluster.

In Tabelle 2 werden zehn Indikatoren präsentiert, mit welchen die meisten der zehn Cluster gut charakterisiert und voneinander separiert werden können. Eindeutig lassen sich die Cluster C2, C8 und C10 unterscheiden:

- **C2** ist klar dadurch definiert, dass alle Personen im Cluster in der Beobachtungsperiode eine IV-Rente beziehen (I9). In der 48-monatigen Beobachtungsdauer beträgt die durchschnittlich IV-Bezugsdauer 31.4 Monate, das ist rund zwei Drittel der Beobachtungszeit.
- In **C8** beziehen alle Personen Sozialhilfe (I6) und sind gleichzeitig nicht erwerbstätig (I8, I5). Der Sozialhilfebezug dauert durchschnittlich 33.0 Monate.
- **C10** unterscheidet sich in diesem Punkt wesentlich von C8. In diesem Cluster sind die Personen während des Sozialhilfebezuges (I6; im Mittel 32.0 Monate) während durchschnittlich 21.8 Monaten erwerbstätig (I8).

Auch die Cluster C3, C7, C6, C9 und C1 sind anhand der Indikatoren gut charakterisierbar:

- Die Personen in **C6** zeichnet sich als einzige durch eine vergleichsweise sehr hohe Dauer von ALV[7] bei gleichzeitiger Erwerbsarbeit aus(I3), die im Mittel 17.3 Monate dauert. Dabei handelt es sich vermutlich um Zwischenverdienste im Rahmen der ALV. Insgesamt arbeiten die Personen in diesem Cluster fast über die ganze Beobachtungsdauer (I5). Mit 22.2 Monaten ist die mittlere Bezugsdauer von ALV insgesamt auch sehr hoch (I1), verteilt auf mehrere Bezugsperioden (I4).
- **C9** weist mit 3.02 eindeutig die höchste mittlere Anzahl ALE-Bezugsperioden auf (I4) und ist damit vor allem durch das Muster mehrfacher Arbeitslosigkeit gekennzeichnet die durch zwischenzeitliche Erwerbstätigkeit (Zwischenverdienste) unterbrochen wird. Entsprechend lang ist auch die durchschnittliche ALE-Bezugsdauer von 21.2 Monaten (I1). Gegenüber C6 ist Dauer von Teilzeitarbeitslosigkeit (I3) kürzer und die Erwerbsintegration (I5) schwächer.
- **C1** umfasst Personen, die sich nach einer kurzen Phase der Arbeitslosigkeit (I1, I2) aus dem Arbeitsmarkt zurückziehen, aus der Schweiz ausreisen oder versterben (I11). Das ist das einzige Cluster mit einer sehr hohen Verweildauer ausserhalb der untersuchten Systeme. Die Heterogenität dieses Clusters führt zu Interpretationsschwierigkeiten.
- **C3** scheint anhand des «State Distribution Plots» eindeutig separierbar zu sein. Mithilfe der Verlaufsindikatoren sind die Unterschiede zu den anderen Clustern (insbesondere C7) jedoch feiner. Die erste Bezugsperiode von ALE ist vergleichsweise lang, wobei die mittlere ALE-Bezugsdauer (I1) tief ist. Gleichzeitig sind diese Personen durchschnittlich während 44.4 Monaten erwerbstätig (I5) und Sozialhilfebezüge kommen kaum vor (4% der Personen).
- **C7** zählt wie C3 und C6 zu den Cluster mit der höchsten mittleren Dauer von Erwerbstätigkeit. Die Dauer und Anzahl Bezugsperioden von ALE ist deutlich länger als in C3. Teilzeitarbeitslosigkeit (I3) ist in C7 im Durchschnitt kürzer als in C6. Der Anteil an Sozialhilfebezüger ist ähnlich hoch wie in C6, aber die Anzahl von ALE-Bezugsperioden ist deutlich geringer (I5).

Immer noch schwierig zu interpretieren sind die Cluster C5 und C4:

- **C5** weist eine hohe mittlere ALE-Bezugsdauer (I1) und im Vergleich eine erhöhte mittlere Anzahl ALE-Bezugsperioden (I4). Ebenso findet man eine mittlere Erwerbsdauer (I5) und ein relativ hoher Anteil Kohortenmitglieder mit Sozialhilfebezügen (I6), die im Mittel jedoch nur kurz andauern (I7). Insbesondere die mehrfachen ALE-Bezüge und der hohe Anteil an Personen mit Sozialhilfe lassen eher auf komplexe Verläufe schliessen. Die Entropie für dieses Cluster (siehe Abbildung 3) steigt denn auch von einer homogenen Ausgangslage zum heterogensten Schlusszustand an.
- **C4** ist einerseits durch eine relativ kurze ALE-Bezugsdauer (I1) gekennzeichnet, andererseits durch eine mittlere Erwerbsdauer (I5) und einer vergleichsweise hohen Dauer ohne Leistungsbezug und Erwerbsarbeit (I11). In Zusammenhang mit den «State Distribution Plots» liegt der Schluss nahe, dass es sich um Personen handelt, die nach der Arbeitslosigkeit leicht in den Weg in die Erwerbsarbeit finden, sich aber nach einer Weile aus dem beobachteten System zurückziehen. Mit den vorgeschlagenen Indikatoren lassen sich die Abfolge dieser Prozesse und der Rückzug aus dem System jedoch nicht abschliessend inhaltlich qualifizieren.

Eine wichtige allgemeine Beobachtung zeigt sich in Bezug auf die Sozialhilfe. Sozialhilfebezüge kommen in allen Clustern vor, insbesondere auch in C2, das durch IV-Bezüge gekennzeichnet ist. Dies zeigt die Bedeutung und die subsidiäre Wirkung der Sozialhilfe bei der Existenzsicherung auf, wenn alle anderen Quellen der Wohlfahrt versiegen oder der Leistungsanspruch in Abklärung ist.

[7] In einigen Grafiken und Tabellen taucht die Abkürzung ALE auf, sie ist synonym zu ALV (Arbeitslosenversicherungstaggelder) zu verstehen.

In der Tabelle 3 sind Vorschläge aufgeführt, wie die Cluster auf der Basis obenstehender Analyse bezeichnet («Labelling») werden können. Die aktuellen Labels orientieren sich an einer institutionellen Logik. Im weiteren Verlauf der Arbeiten könnten auch andere Arten der Unterscheidungen verfolgt werden. Zum Beispiel, indem der Grad der Verfestigung der Bezugskarriere über alle Unterstützungssysteme hinweg ins Zentrum gerückt wird. Unterscheidungen können nach Verfestigung (langfristiger Bezug von Leistungen), Konsolidierung (wiederholter Bezug von Leistungen, mit Untergruppen nach Aktivierung strukturiert) oder Optimierung (erfolgreicher dauerhafter Ausstieg) der Karrieren vorgenommen werden.

Tabelle 2: Verlaufsindikatoren nach Cluster (Mittelwerte, Anteil), Kohorte 2010

Indikator		C3	C7	C6	C9	C2	C8	C10	C1	C5	C4
I1	Anzahl Monate mit ALV	5.75	14.21	22.20	21.20	12.05	14.12	15.54	12.21	20.32	8.43
I2	Dauer der ersten ALV-Bezugsperiode (Monate)	4.14	10.53	13.11	6.08	9.34	11.33	10.57	9.84	13.27	5.10
I3	Anzahl Monate ALV und Erwerbsarbeit kombiniert (Teilzeitarbeitslosigkeit)	2.80	2.55	17.29	6.01	2.77	1.73	5.95	1.81	3.14	2.79
I4	Anzahl Bezugsperioden ALV	1.51	1.74	2.51	3.02	1.46	1.45	1.87	1.41	2.03	1.80
I5	Anzahl Monate mit ME	44.42	32.99	41.39	29.87	17.09	6.19	29.83	5.63	22.50	27.13
I6	Anteil Personen mit mindestens einer SH-Bezugsperiode	4%	18%	14%	17%	27%	100%	100%	17%	38%	12%
I7	Anzahl Monate mit SH	0.17	2.16	1.33	1.08	3.49	32.98	31.91	1.86	4.91	0.74
I8	Anzahl Monate SH und Erwerbsarbeit kombiniert	0.11	1.20	0.78	0.41	1.07	3.11	21.82	0.27	1.69	0.33
I9	Anteil Personen mit mindestens einer IV-Bezugsperiode	0%	1%	1%	0%	100%	6%	1%	1%	1%	1%
I10	Anzahl Monate mit IV	0.00	0.11	0.05	0.03	31.39	0.82	0.11	0.11	0.04	0.15
I11	Anzahl Monate ohne Erwerbsarbeit und ohne Sozialleistungsbezug	0.63	3.01	1.34	2.57	4.72	2.77	1.98	30.65	5.69	14.92
Clusterlabels		ALV Kurzzeit	ALV Langzeit	Zwischen-verdienst	ALV Mehrfach	IV	SH	«Working Poor»	Leavers	Komplex	Unklar

Quelle: SHIVALV-IK 2010-2015, eigene Berechnungen

Anmerkung: ALE = Arbeitslosenentschädigung/Arbeitslosentaggelder, SH = Leistungen der Sozialhilfe, IV = Leistungen der Invalidenversicherung, ME = marché d'emploi/Erwerbsarbeit

5.4 Vergleich der Clusterlösung zwischen Kohorte 2010 und 2011

Der Clusteralgorithmus führt insgesamt zu stabilen Resultaten. Wird das sequence clustering mit der nachfolgenden Kohorte von Erstbezügern von ALV durchgeführt, ergeben sich sehr ähnliche Cluster. Lediglich die schwierig zu beschreibenden Cluster (in Kohorte 2010 die C4 und C5 siehe Abschnitt 5.3). Auch für die Kohorte 2011 lässt sich die Anzahl von 10 Cluster anhand der Elbow-Methode gut rechtfertigen (siehe Anhang, Tabelle A 1,

Abbildung A 2 und Abbildung A 3).

Um Auswirkungen durch Regulationen im Bereich des Systems der sozialen Sicherheit, oder wirtschaftliche und gesellschaftliche Auswirkungen verfolgen zu können, ist es unabdingbar, die Cluster einer Kohorte auf andere Kohorten auf individueller Ebene übertragen zu können. Mit anderen Worten geht es darum, sicherzustellen, dass gleiche Verläufe, den gleichen Cluster zugeordnet werden, unabhängig von der Kohorte. Dies könnte noch einen Einfluss auf die Anzahl Cluster haben, welche mit einer genügenden Sicherheit über die Jahre gebildet werden können, was Voraussetzung für die Stabilität von Längsschnittanalysen ist. Die Arbeiten in diesem Bereich konnten im Rahmen dieses Pilotprojekts nicht abschliessend durchgeführt werden.

Es ist anzunehmen, dass sich aufgrund des wirtschaftlichen und gesellschaftlichen Wandels sowie der Entwicklung der Regulationen im Bereich der Sozialgesetzgebung neue Verlaufsmuster für spätere Kohorten entstehen und bestehende gewissermassen verschwinden.

5.5 Fazit

In einem Umfeld, in dem Verlaufsdaten zunehmend an Bedeutung gewinnen, ist das sequence clustering ein vielversprechendes Verfahren, das es erlaubt die Komplexität von Verlaufsdaten deutlich zu verringern, und inhaltlich valide und analytisch unerwartete Resultate zu erzeugen (Erkennung deduktiv nicht antizipierter Muster). Neben der Festlegung der Anzahl Cluster ist anhand grafischer Methoden und anhand von Verlaufsindikatoren eine inhaltliche Interpretation der Cluster gelungen. Die Cluster von sozialen Verlaufsmuster bilden einerseits bekannte Umstände ab, zeigen aber auch unerwartete Wege aus der Arbeitslosigkeit in die Erwerbsarbeit auf.

Offene Punkte betreffen die bessere Interpretation unklarer Verlaufsmuster (C4 und C5). Allenfalls können durch Verknüpfung mit weiteren Datenquellen Rentenbezüge, Ausbildungsteilnahmen, Ausreisen aus der Schweiz oder Todesfälle besser identifiziert werden.

Mit der Analyse einer zweiten Kohorte konnte mittels State-Distributions-Plots auch die Stabilität des Algorithmus aufgezeigt werden. Ein offener Punkt betrifft jedoch noch die Übertragung der Cluster einer Kohorte auf andere.

6 Lessons learned

6.1 Sequence Clustering

Beruhend auf der Kohortensichtweise konnte im ML_SoSi-Projekt gezeigt werden, dass das angewendete «sequence clustering»-Verfahren zu statistisch nachvollziehbaren und inhaltlich validen Resultaten führt, um soziale Verlaufsmuster im System der Sozialen Sicherheit und der Erwerbsarbeit zu beschreiben (siehe Abschnitt 5.3). Es wurde eine beschränkte Anzahl Cluster bestimmt, deren inhaltliche Interpretation sich als gut machbar erwies. Das heisst, diese Methode (als Teil von «unsupervised machine learning») eignet sich gut, die Komplexität der Verlaufsdaten so zu reduzieren, dass sie analysierbar werden.

Wie unter 5.4 erwähnt, wird es in Zukunft eine Herausforderung sein, mit dem Wandel der Verlaufsmuster von «späteren» resp. neuen Kohorten aufgrund regulatorischer und wirtschaftlicher Veränderungen umzugehen. Als mögliches Vorgehen wurde die Bildung von Cetroiden (statistisch berechneter zentraler Verlauf pro Cluster) angedacht. Verläufe aus neuen Kohorten werden jeweils mit diesen zentralen Verläufen pro Cluster verglichen und jenem Cluster zugeordnet, bei welchem die höchste Übereinstimmung gefunden wurden. Falls die Zuordnung nicht eindeutig ist, oder die Homogenität stark beeinträchtigt, könnte dies ein Hinweis auf neue Verlaufsmuster sein. Das initiale Clustermodell muss somit regelmässig darauf überprüft werden, ob es die gesellschaftlichen Realitäten immer noch adäquat abbildet.

6.2 Datenvisualisierung

Im ML_SoSi-Projekt konnte gezeigt werden, dass mehrere vielversprechende Visualisierungsmethoden für individuelle Verlaufsdaten zur Verfügung stehen. In Zusammenhang mit dem sequence clustering sind dies insbesondere «state distribution plots» und «transversal entropy plots» (vgl. Abschnitt 5.3). Besonders Erstere bieten einen intuitiven Zugang zur Clusteranalyse und sind in Präsentationen vor Publikum auf grosse Resonanz gestossen.

Ebenfalls vielversprechend sind Sankey-Plots. Ein rudimentäres Beispiel dazu ist in Abschnitt 10.1, Abbildung 4) zu sehen. Sie visualisieren Personenströme zwischen verschiedenen Zeitpunkten. Im erwähnten Beispiel sind die Zeitpunkte jeweils ein Beobachtungsjahr. An den Strömen ist zu sehen, wie viele Personen von dem einen in den anderen Zustand wandern. Obschon eine vielversprechende Darstellungsform, besteht bei vielen verschiedenen Zuständen die Gefahr der Unübersichtlichkeit. Dies kann behoben werden, indem ein Betrachtungszeitpunkt über mehrere Monate synthetisiert wird (der Zustand für Person i zum Zeitpunkt T ergibt sich aus dem häufigsten Zustand in den Monaten t-2, t-1, t, t+1, t+2 → Modalwert).

6.3 Kohortensichtweise, Standardisierung des Beobachtungszeitraums und Verlaufsindikatoren

Der Ansatz, die Grundgesamtheit als homogene Kohorten mit einem gemeinsamen Initialereignis (erstmaliger Bezug von ALE) zu definieren und die Beobachtungsdauer zu standardisieren, hat sich bewährt. Erstens ist die Abgrenzung so relativ einfach durchzuführen und das Konzept kann gut vermittelt werden.

Der Kohortenansatz wird auch in den Längsschnittstatistiken im Bildungsbereich (LABB) mehrheitlich angewendet.

Verlaufsindikatoren können polyvalent zur Beschreibung von Cluster, ganzen Kohorten oder Subgruppen davon verwendet werden. Einzelne Indikatoren, die im ML_SoSi entwickelt wurden, können direkt für die Weiterentwicklung der SHIVALV-Indikatoren im BFS bzw. für ein produktives Umfeld übernommen werden. Es besteht zudem Potential, um weitere Kennzahlen zu entwickeln.

Der Nachteil der Kohortenperspektive ist, dass darauf beruhende Kennzahlen oftmals einen Zeitraum abdecken, der vergleichsweise weit in der Vergangenheit liegt. Ein Lösungsansatz dazu wäre das Abstützen auf Prädiktionen.

6.4 Datenmodell

Im Rahmen des Projekts wurde auf der Basis der Ursprungsdaten des BSV ein Datenmodell entwickelt, das für die Analyse und ein produktives Umfeld zielführend ist. Es besteht aus einem Datensatz im «long»-Format, welches die Verlaufsinformationen pro System (SH, IV, ALV, IK/Erwerb) pro Person und pro Monat enthält. Zusätzlich enthält es mehrere Datensätze mit soziodemografisch und -professionellen Informationen. Diese liegen auf Jahresbasis vor, sind nach Quellsystem (SH, IV, ALV) getrennt und gelten entsprechend nur für die Personen, die Leistungen diesen Systemen erhalten. Für Geschlecht, Zivilstand, Nationalität und Alter existiert ein harmonisiert Datensatz über alle Systeme (siehe auch Anhang, Abschnitt 9.2). Das Datenmodell ist in der Lage den Informationsgehalt von SHIVALV+IK vollständig zu repräsentieren. Bei Bedarf können weitere Datenformate («wide», «spell») abgeleitet werden, die für die Beantwortung spezifischer Bedürfnisse Vorteile bieten. Konvertierungscodes dazu können externen Nutzern direkt zur Verfügung gestellt oder über eine Plattform wie Github öffentlich zugänglich gemacht werden.

Das Datenformat ist mit einfachen Codes direkt für die Analyse bereit; viele vordefinierte Funktionen können das Datenformat lesen.

Da in Zukunft das BFS für die Produktion und Abgabe der SHIVALV-Daten zuständig ist, können die

Erfahrungen mit dem Datenmodell dort in die Produktion umgesetzt und als Standard definiert werden.

6.5 Mehrwert für die Sozialhilfeempfängerstatistik (SHS)

ML_SoSi hat auf methodischer Ebene eine Basis gelegt, die für den SHS-Kontext direkt übernommen werden kann. Dies ist insbesondere für eine modernisierte SHS von Bedeutung, da dort Verlaufsindikatoren eine stärkere Rolle spielen sollen:

- Datenaufbereitung und Struktur: Das SHIVALV+IK-Datenmodell ist direkt übertragbar auf die SHS. Es wird ein Konzeptpapier ausgearbeitet, das die exakt gleiche Datenstruktur für Standardverlaufsdatensätze der SHS vorsieht. Im Unterschied zu SHIVALV werden jedoch keine Einschränkungen für das Alter vorgeschlagen. Zudem stellt sich die Frage, ob neben dem Bezugsverlauf weitere Informationen als Verläufe modelliert werden müssen (z.B. Wohnortswechsel, Dossierwechsel) und ob Verläufe der WSH, AsylStat und FlüStat in einem einzigen Standardverlaufsdatensatz geführt werden können.
- Kohorte und Beobachtungszeitraum: Die Kohortendefinition kann analog auf Sozialhilfebeziehende in der SHS übernommen werden, ebenso die Abgrenzung von Personen mit früheren Sozialhilfebezügen. Dies gilt auch für die Definition des Startzeitpunkts für die Beobachtung (Monat mit erster Auszahlung) und für den standardisierten Beobachtungszeitraum.
- Verlaufsindikatoren: Im Projekt schon definierte Verlaufsindikatoren wie Anzahl Monate Sozialhilfebezug oder Anzahl Sozialhilfebezugsperioden im Beobachtungszeitraum können direkt in die Produktion übernommen werden. Zudem schlagen wir vor, die Typologie der Sozialhilfeverläufe aus der Studie «Verläufe in der Sozialhilfe»[8] zu reproduzieren.
- Sequence Clustering: Mit der Anwendung dieses induktiven Ansatzes ausschliesslich auf den Sozialhilfeverläufen kann die zuvor erwähnte Typologie datengetrieben «validiert» bzw. können neue Verlaufsmuster identifiziert werden.

6.6 Infrastruktur

Die Statistiksoftware R hat sich als Programmiersprache sehr bewährt, insbesondere das Package TraMineR der Universität Genf.

Für die Modellbildung (unsupervised und supervised learning) sind die persönlichen Rechner zu wenig performant. Aufgrund der Restriktionen auf den Servern der Cloud Atlantica, die in der Pilotphase zur Verfügung standen, tauchten auch dort Performanceprobleme auf. Diese Problematik wird in einem separaten Bericht über alle Pilotprojekt erörtert.

6.7 Agile Arbeitsweise und Zusammenarbeit über Sektionen

Die agile Arbeitsweise mit Intensivworkshops hat sich im gegebenen methodischen Setting sowohl in Bezug auf die Zielerreichung und den Teamgeist positiv ausgewirkt und wird von uns für ähnliche Settings empfohlen. Voraussetzung ist jedoch eine kompetente Führung.

Die Zusammenarbeit über die Sektionsgrenze hinweg (SHS/SOZAN) war eine grosse Bereicherung und hat gut funktioniert, insbesondere auch die Co-Leitung des Projekts. Die Planung der Ressourcen und Aktivitäten haben dadurch jedoch an Komplexität zugenommen.

Das Arbeitsprozessmodell CRISP-DM und die in Confluence zur Verfügung gestellten Dokumentationsvorlagen dazu haben sich mehrheitlich bewährt. Nachfolgeprojekten wird ein pragmatischer Umgang mit dem Modell empfohlen.

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

6.8 Data Science Kultur

Data Science verknüpft u.a. Kompetenzen aus Analyse, Computerwissenschaften und Data Management und ein Mindset, dass diese Kompetenzen gewinnbringend zu verbinden vermag. Im Pilotprojekt wurde offensichtlich, dass vor allem die IT-seitigen Kompetenzen wie kohärente Programmierung und die Organisation der Daten eine Herausforderung für das Team darstellten. Insbesondere der Anspruch, die Codes unterschiedlicher Personen optimal aufeinander abzustimmen und mit Parametrisierungen eine effiziente Reproduktion der Resultate zu gewährleisten, konnte nicht vollständig eingelöst werden. Für ähnliche Projekte lohnt es sich, Aspekte wie funktionenorientierte Programmierung, R-Package-Programmierung und GIThub stärker zu gewichten.

6.9 Arbeitsaufwände für Kompetenzaufbau und Datenaufbereitung

Die Alltagsarbeiten der Teammitglieder unterscheidet sich deutlich von der Arbeit in diesem Projekt. Das heisst, dass wesentliches Know-how aufgebaut werden musste, was trotz der Unterstützung durch die angebotenen Weiterbildungskurse relativ viel Ressourcen in Anspruch nahm. Themen wie die Statistiksoftware «R», methodische Elemente rund um induktives Arbeiten sowie die konzeptionellen Gesichtspunkte für den Aufbau der Datenbasis benötigen viel Zeit für die Ausbildung «on-the-job». Nachfolgeprojekte sollten diesen Umständen unbedingt Rechnung tragen.

Wie in der Forschungsarbeit üblich, wurden im Projekt ca. 80% für die Datenaufbereitung und 20% für die Analyse eingesetzt. Je nachdem wie nah ein künftiges produktives Umfeld an den Inhalten und Procederes dieses Projektes angesiedelt sind, dürfte sich der Aufwand für die Datenaufbereitung wesentlich reduzieren. Das heisst, es kommen Effekte durch die Skalen- und Standardisierungseffekte realisieren.

7 Weiteres Vorgehen

Unter Berücksichtigung der angespannten Ressourcensituation wird der Fokus der Arbeiten ab September 2019 auf folgende Themenbereiche gerichtet:

a. Sensitivitätsanalysen: Untersucht werden soll dabei der Einfluss von Datenaufbereitungsschritten wie in Kapitel 4 und ausführlicher im Anhang (Kapitel 9.2.) dargestellt sind. Des Weiteren soll der Einfluss unterschiedlicher Beobachtungsdauern (aktuell 48 Monate) identifiziert werden. Ziel ist der Erkenntnisgewinn, ob und in welchem Ausmass entsprechende Datenbereinigungen notwendig und sinnvoll sind.
b. Übertragung der Cluster: Anschliessend wird die Frage angegangen werden, wie sich die Cluster trotz Veränderungen der Kontextfaktoren (z.B. wirtschaftliche Entwicklung, Regulationen) auf andere Kohorten übertragen lassen. Ziel ist hierbei sicherzustellen, dass gleiche Verläufe, den gleichen Cluster zugeordnet werden, unabhängig von der Kohorte (siehe dazu auch Kapitel 5.4 und 5.5).
c. Prädiktion: Als weiterer Schritt sollen mithilfe von Prädiktionsmodellen die Wahrscheinlichkeit für die Zugehörigkeit der beobachteten Individuen zu jedem der zehn Verlaufscluster geschätzt werden. Für diese Schätzungen wird ein «supervised machine learning»-Ansatz angewendet: Mit den Daten der Kohorten 2011 und 2010 wird ein Modell entwickelt (trainiert), bei welchen als abhängige Variable die Clusterzugehörigkeit für jede Person in der Kohorte bekannt ist. Im Rahmen der Modellbildung müssen voraussichtlich mithilfe systematischen Experimentierens eine Reihe von Entscheidungen getroffen werden, u.a. bezüglich der Wahl des Algorithmus und Optimierungsverfahrens, der Kriterien zur Einschätzung der Modellgüte, dem Umgang mit unterschiedlichen Clustergrössen, der Gewichtung der Fehlerraten.

8 Projektorganisation

Projektteam

Wer	Vornehmliche Projektrolle
Luzius von Gunten, SHS	Co-Projektleitung: Konzeption, Analysestrategie, Workshopleitung AG NDS
Thomas Ruch, SOZAN	Co-Projektleitung: Konzeption, Koordination, Dokumentation AG NDS
Nora Meister, SOZAN	Datenmodell, Indikatorenberechnung, sequence clustering
Sheila Planta, SHS	Datenkorrekturen
Gerhard Gillmann, SHS	Datenkorrekturen, Indikatorenberechnung, Sankey-Plots
Joaquim Golay, SHS	Datenmodell, Sensitivitätsanalyse
Philippe Meyer, SHS	Imputationen, sequence clustering, Auswertung und Visualisierung

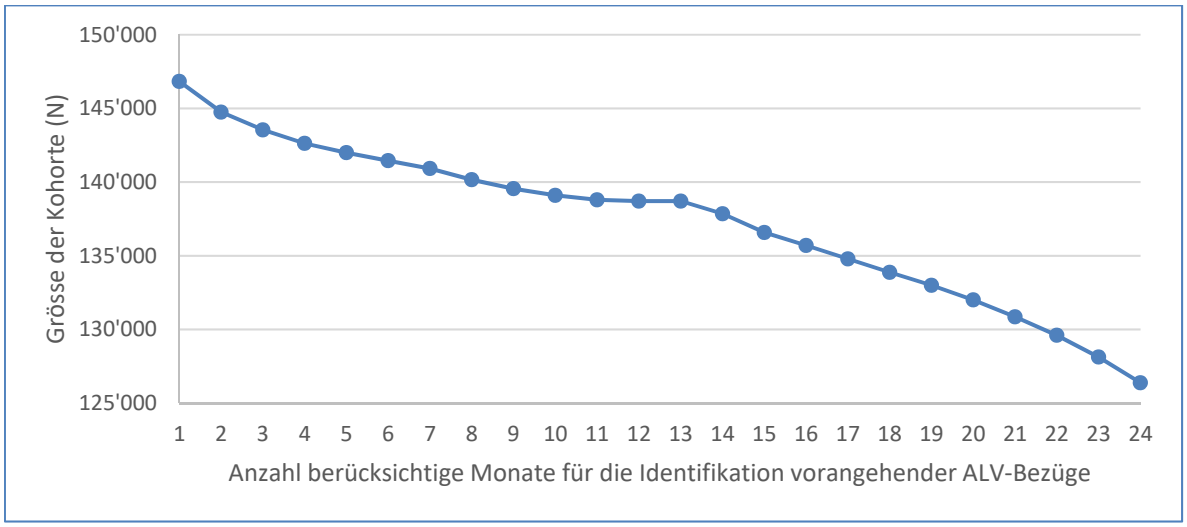
Unterstützung METH

Wer	Vornehmliche Projektrolle
Daniel Kilchmann	Methodische Beratung, kritische Begleitung
Michael Leuenberger	Methodische Beratung, Validierung Clustermodelle, entropy plots
Athanassia Chalimourda	Methodische Beratung, Validierung Clustermodelle, elbow plots
Kaspar Stucki	Methodische Beratung, Analysestrategie sequence clustering

9 Anhang

9.1 Tabellen und Grafiken

Abbildung A 1: Taille de la cohorte ALV 2010, selon nombre de mois de vérification rétrospective



Quelle: SHIVALV 2010, AVAM/ASAL 2008/2009, eigene Berechnungen

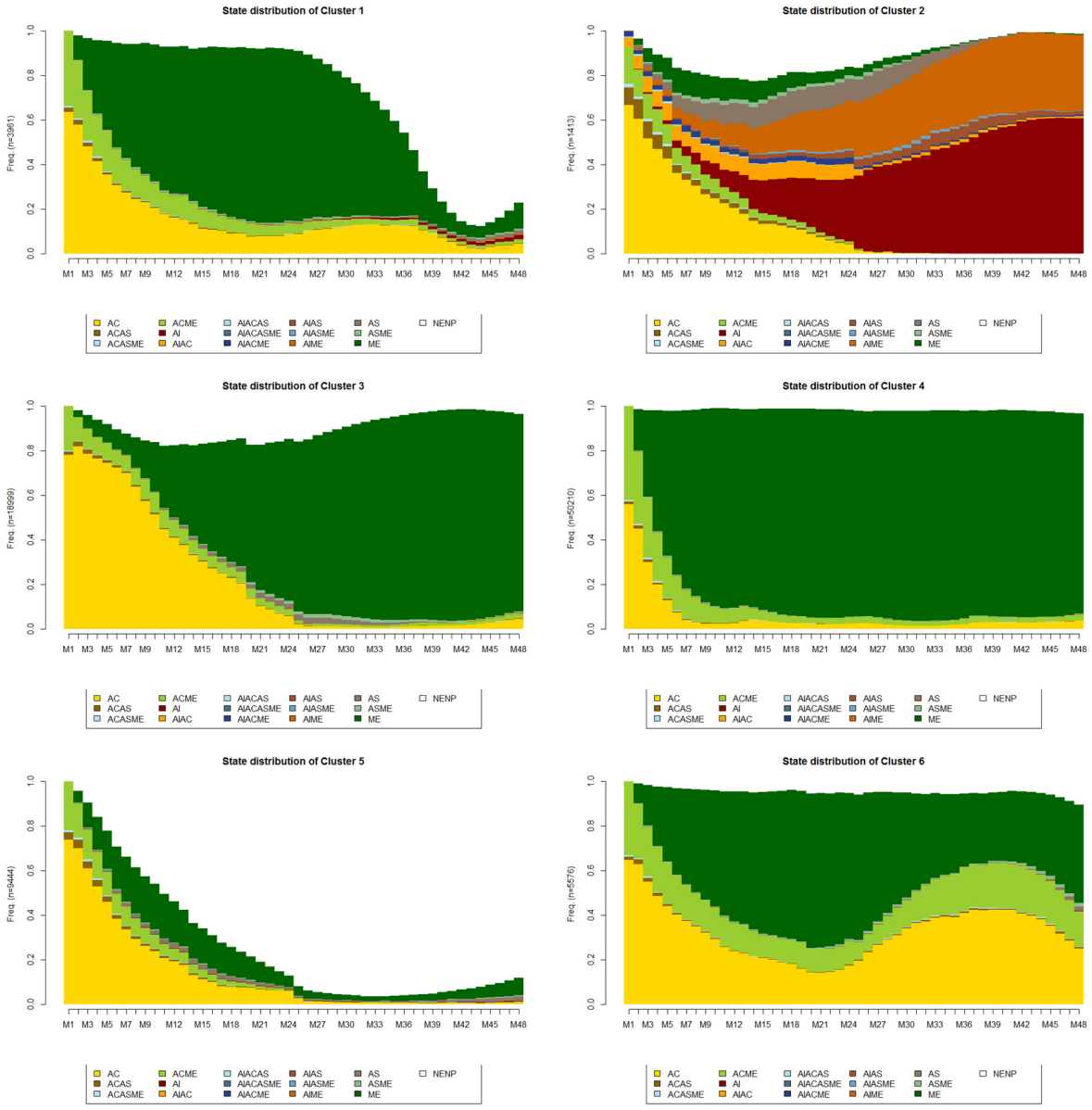
Tabelle A 1: Korrespondenztabelle für Zehn-Clustermodell Kohorte 2010 und 2011

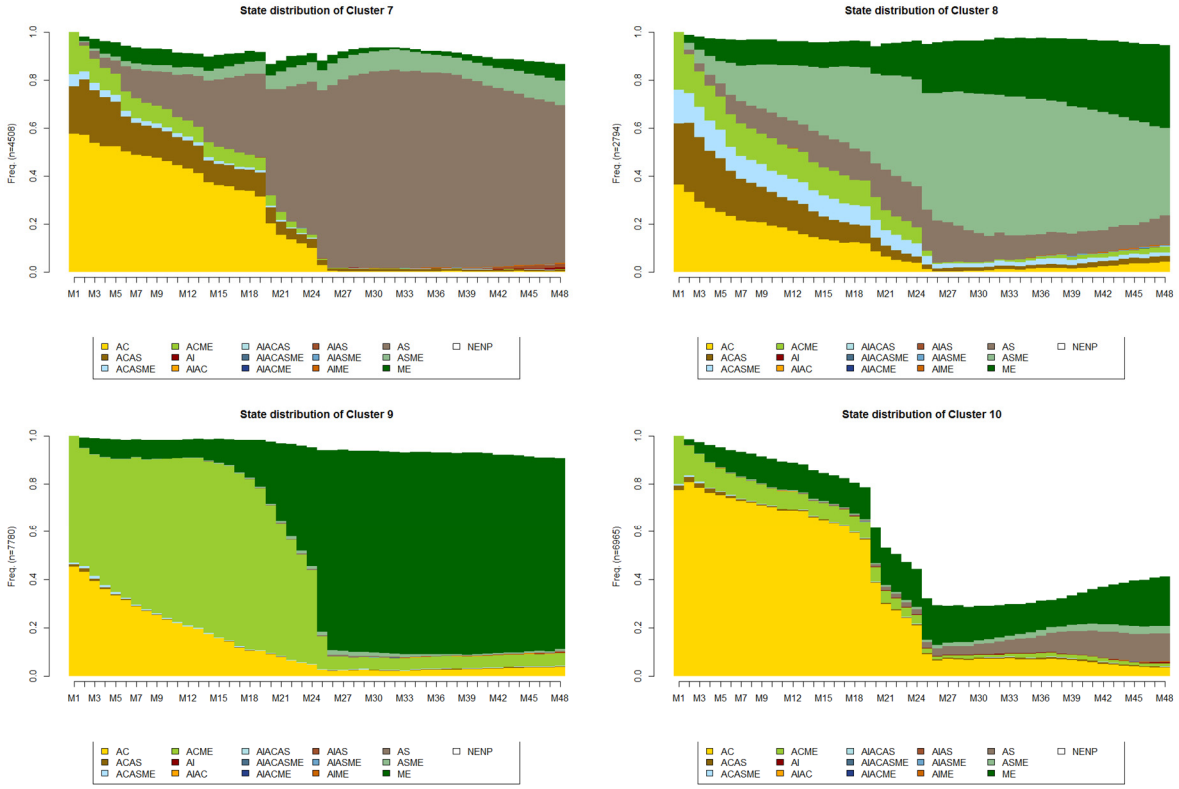
Cluster K2010	Cluster K2011

DIS Pillotprojekt ML_SoSi_GS

1	5
2	2
3	4
4	*1*
5	*10*
6	9
7	3
8	7
9	6
10	8

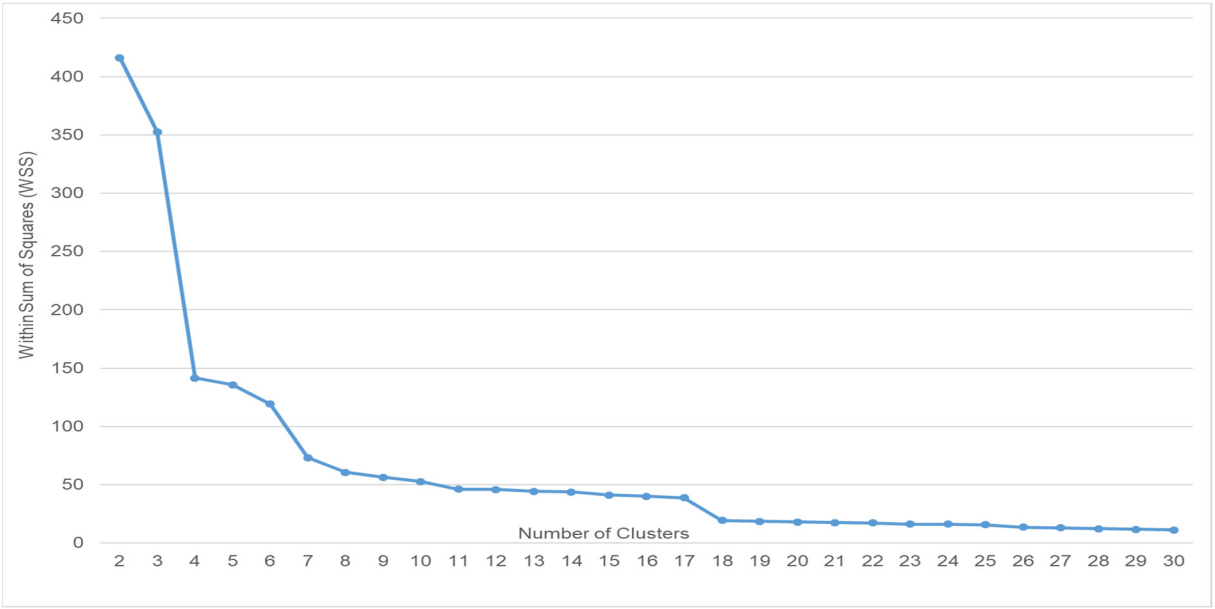
Abbildung A 2: state distribution plots für Zehn-Clustermodell, Kohorte 2011





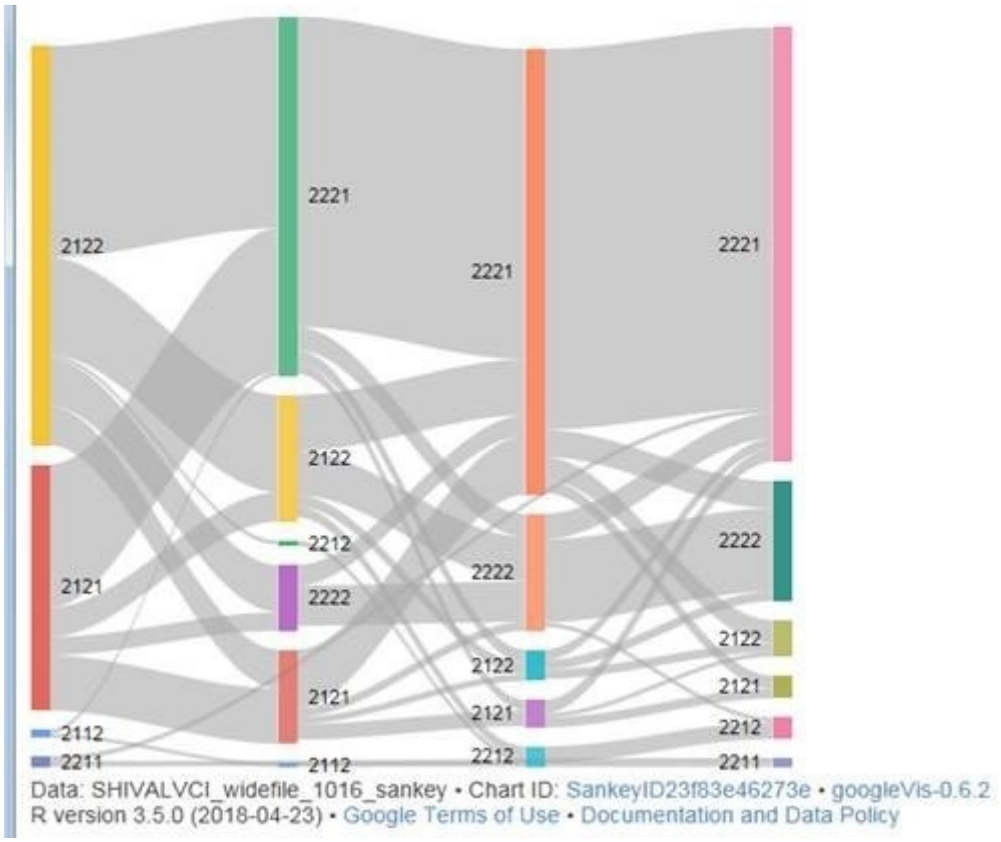
Quelle: SHIVALV-IK 2011-2016, eigene Berechnungen

Abbildung A 3: Elbow-Methode zur Bestimmung der optimalen Anzahl Cluster, Kohorte 2011



Quelle: SHIVALV-IK 2011-2016, eigene Berechnungen

Abbildung A 4: sankey-plot (explorativ, ohne Beschriftungen)



9.2 Datenmodell und Datenaufbereitung

Datenmodell

Für die Analysen im Projekt sind mehrere Datensätze notwendig. Konkret werden ein LONG-File, eine WIDE-File und mindestens vier SPELL-Files (eines pro System) generiert[9]. Die folgenden Kapitelverweise beziehen sich auf die Dokumentation des R-Packages TraMineR: http://mephisto.unige.ch/pub/TraMineR/doc/TraMineR-Users-Guide.pdf

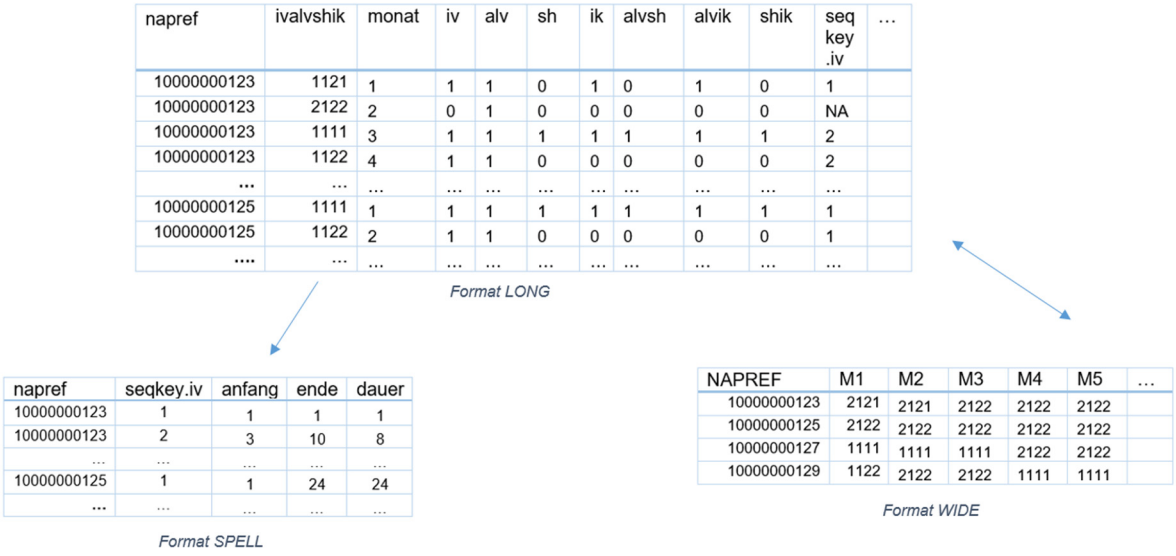
- Datenstruktur LONG: In der Grundform von LONG jeder mögliche Status/Zustand einer Person als eigene Spalte in binärer Form aufgeführt (Status aktuell vorhanden = Ja(1)/Nein(0); z.B. ALV-Leistungen = 0/1, Erwerbstätigkeit = 0/1). Jede Spalte bezeichnet ein System (ALV, IV, SH, Erwerbstätigkeit) oder eine Systemkombination (z.B. ALV/SH, ALV/Erwerbstätigkeit, usw.). Pro Monat wird eine Zeile geschrieben, so dass für jede Person mehrere Statusinformationen (Zeilen) registriert werden können. Jede Statusänderung begründet eine neue Sequenz die durchnummeriert werden («seqkey»). Im Minimum existiert eine einzige Sequenz, wenn nämlich eine Person während der gesamten Beobachtungszeit im selben Status verbleibt (z.B. immer nur ALV-Leistungsbezug).
- Datenstruktur WIDE: Pro Person (Zeile) und Zeitpunkt (Spalte) ist die Kombination aller aktuellen Stati in einem Stringcode aufgeführt. Jede Position innerhalb des Strings ist einem System zugeordnet. Bei vier möglichen Systemen (ALV, IV, SH, Erwerbstätigkeit) besteht der String aus immer genau vier Positionen. Jede Position repräsentiert immer dasselbe System, so entspricht zum Beispiel die 2. Stelle im String dem Leistungsbezug in der Sozialhilfe. Ein

[9] Je nach Erkenntnisinteresse werden zusätzlich pro Statuskombinationen weitere Files generiert

Wert 1 an einer bestimmten Position bedeutet, dass aus dem entsprechenden System Leistungen bezogen werden oder die Person erwerbstätig ist. . Der Wert 2 bedeutet kein Bezug von Leistungen oder Erwerbstätigkeit. Die Datenstruktur entspricht dem "State Sequence"-Format (STS).

- Datenstruktur SPELL: Pro System (ALV, IV, SH, Erwerbstätigkeit) und Systemkombinationen (z.B. ALV/SH, ALV/Erwerbstätigkeit, usw.) werden für jede Person Startmonat, Endmonat und Dauer des Verbleibs im System berechnet (Periode=spell). Eine Person kann somit mehrere Perioden aufweisen wobei jede Periode in einer separaten Zeile abgelegt wird. Pro Person sind demnach mehrere Zeilen möglich. Die Perioden pro Person sind nummeriert (seqkey="Sequenzschlüssel).

Abbildung A 5: Schematische Darstellung der Datenstrukturen LONG, WIDE und SPELL



Quelle: eigene Darstellung

Je nach Analysezweck werden nur einzelne Leistungsbezugsverläufe oder Kobinationen davon weiterverarbeitet. Wenn nur die Datensätz WIDE und SPELL zur Verfügung stehen, ist diese Weiterverarbeitung schwierig, da Abfragen auf den so strukturierten Datensets schnell aufwändig und komplex werden. Im LONG sind entsprechende Abfragen sehr einfach zu bewerkstelligen, da ohne grossen Aufwand flexibel neue Spalten mit kombinierten/reduzierten Informationen hinzugefügt werden können.

Je nach Analyse- oder Datenaufbereitungssituation haben SPELL und WIDE ihre Berechtigung: Mit SPELL lassen sich z.B. sehr einfach Indikatoren berechnen wie "Dauer der letzten Periode mit kombiniertem SH-IV-Bezug im Beobachtungszeitraum" oder "Anzahl ALV-Bezugsperioden, die >3 Monate dauern" oder "Dauer zwischen Ende des ersten ALV-Bezugs und Beginn der ersten Erwerbsperiode". WIDE eignet sich sehr gut für Analysen wie die "State Distribution Plots" und für die Nutzung von TraMineR-Funktionen.

Die drei Datenstrukturen lassen sich einfach in die jeweils andere Form überführen. Da LONG die flexibelste und am einfachsten zu bearbeitende Datenstruktur ist, verwenden wir LONG als Master. die beiden anderen Datenstrukturen werden daraus nach Bedarf abgeleitet.

Zusätzlich enthält das Datenmodell mehrere Datensätze mit soziodemografisch und -professionellen Informationen; diese liegen auf Jahresbasis vor, sind nach Quellsystem (SH, IV, ALV) getrennt und gelten entsprechend nur für die Personen, die Leistungen dieser Systeme erhalten. Für Geschlecht,

Zivilstand, Nationalität und Alter existiert ein harmonisierter Datensatz über alle Systeme.

Korrektur der Verlaufsdaten: Ein-Monats-Bezüge, Ein-Monats-Unterbrechungen und Ein-Monats-Überlappungen

Aus rein administrativen Gründen (z.B. Auszahlung von Sozialleistungen zu Beginn oder am Ende des Monats) können kurze Bezugslücken oder Überlappungen zwischen den verschiedenen Systemen in den Bezugsverläufen vorkommen. Zudem stellt sich die Frage ob einmonatige Bezugs- bzw. Erwerbsperioden eine zentrale Rolle für die Existenzsicherung spielen und deshalb analytisch relevant sind. In verschiedenen externen Forschungsprojekten wurden entsprechende Datenkorrekturen umgesetzt. Im Rahmen einer Sensitivätsanalyse soll hier getestet werden, wie sich Korrekturen auf den Verlaufsdaten auf die Verlaufscluster und Verlaufsindikatoren auswirken. Während eine Version der Verlaufsdaten inklusive Korrekturen umgesetzt werden konnte, ist die Sensitivitätsanalyse noch ausstehend.

Folgende Korrekturen wurden in konsekutiven Arbeitsschritten realisiert (Umsetzung der sogenannten «2-Monatsregel»):

1. Korrektur einmonatige Bezugs-/Erwerbslücken: diese werden auf den Einzelsystemen ALV, SH, IV, IK aufgefüllt (0 → 1)
2. Korrektur einmonatige Bezugs-/Erwerbsperioden: diese werden auf den Einzelsystemen ALV, SH, IV gelöscht (1 → 0)
3. Korrektur von einmonatigen Überlappungen zw. den Systemen: Relevant sind folgende Überlappungen:
    a. ALV-SH: SH wird auf 0 gesetzt, ALV bleibt 1
    b. ALV-IK: IK wird auf 0 gesetzt, ALV bleibt 1
    c. IK-SH: SH wird auf 0 gesetzt, IK bleibt 1

    In der IV kommen Überlappungen nur sehr selten vor, weshalb sie dort nicht korrigiert werden.

Imputation fehlender Werte in den deskriptiven Variablen

Eine Analyse der fehlenden Werte für die soziodemografischen Variablen ergab nur bei zwei Variablen eine problematisch hohe Missingquote. Für die Kohorte 2010 sind das ungefähr 15% aller Werte der Variable «sbn2000» (zuletzt ausgeübter Beruf) sowie 7% der Fälle für die Variable «ausbild», welche die Bildungsabschlüsse der Kohortenmitglieder beschreibt. Die Werte für die Kohorte 2011 sind mit 13% für «sbn2000» sowie 10% für «ausbild» sehr ähnlich. Andere Variablen, etwa der Zivilstand, haben Missingquoten von weit unter einem Prozent. Der Grenzwert, ab dem eine Imputation in Betracht gezogen wurde, wurde bei 5% festgesetzt.

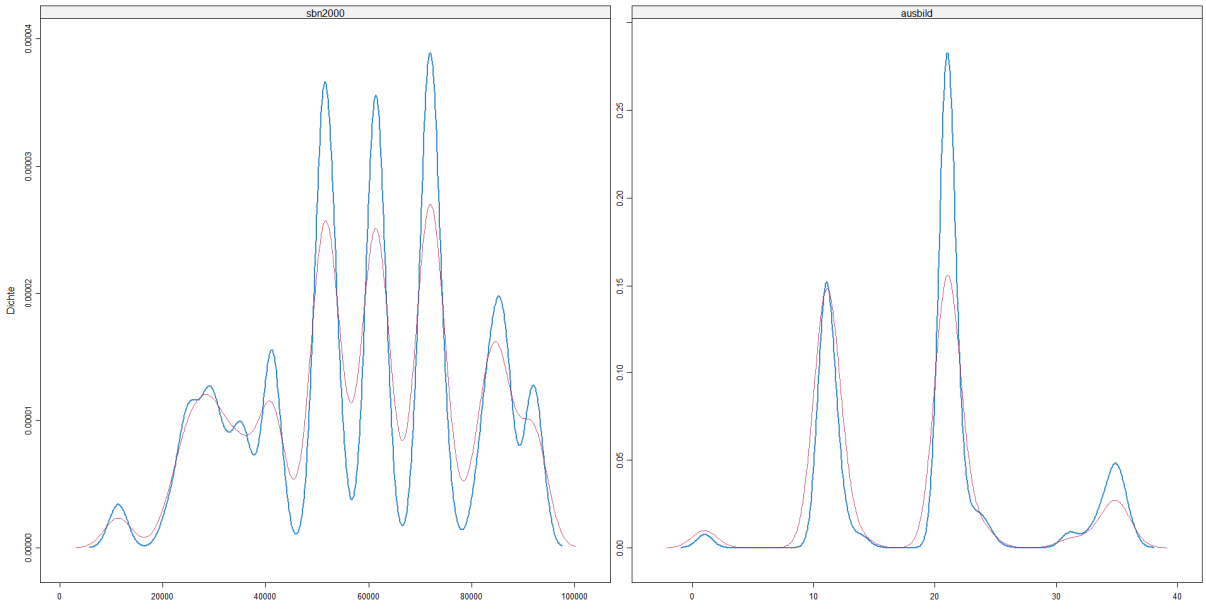
In einem weiteren Schritt wurden die fehlenden Werte auf ihre Zufälligkeit überprüft, also ob die Werte Missing Completly At Random (MCAR), Missing At Random (MAR) oder Not Missing At Random (NMAR) sind. Wenn die fehlenden Werte rein zufällig verteilt sind (MCAR), kann von einer Imputation abgesehen werden. Falls die Missings jedoch systematisch unterschiedlich verteilt sind (MAR), sind die Voraussetzungen für eine Imputation gegeben. NMAR kann nicht erkannt werden, da dies heisst, dass die Information, um den Antwortmechanismus zu modellieren, nicht vorhanden ist. Somit kann nicht ausgeschlossen werden, dass der MAR Mechanismus eigentlich NMAR ist. Die Auswertung sowohl für die Kohorte 2010 als auch die Kohorte 2011 ergab jeweils systematische Verzerrungen in den Daten aufgrund der fehlenden Werte in den beiden Variablen «sbn2000» und «ausbild».

Die Imputation selbst wurde in r mithilfe des Pakets ‘mice’ durchgeführt. Die verwendete Methode ist ‘cart’, kurz für classification and regression trees. Dabei wird mit einem machine-learning-Ansatz für die zu imputierende Variable ein Modell mit allen vorhanden Variablen im Datensatz erstellt, um die fehlenden Werte so genau wie möglich zu ersetzen.

Die Grafik unten (ein sogenannter Dichteplot) zeigt die vorhandenen Variablenwerte (in blau) sowie

die imputierten Variablenwerte (in rot) für die Kohorte 2011. Wie zu sehen ist, sind die beiden Dichteverteilungen sehr nah beieinander.

Abbildung A 6: Dichteverteilungen für vorhandene und imputierte Werte



Quelle: AVAM/ASAL 2010, eigene Berechnungen

9.3 Sensitivitätsanalyse: Abgrenzung der Kohorte

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 ein bis 24 Monate vor dem Beginn der Arbeitslosigkeit 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 der Arbeitslosigkeit im Kohortenjahr ALV-Taggelbezüge aufweisen, aus der Kohorte ausgeschlossen. 24 Monate entspricht einer Rahmenfrist der ALV.

9.4 Wahl der Anzahl Cluster

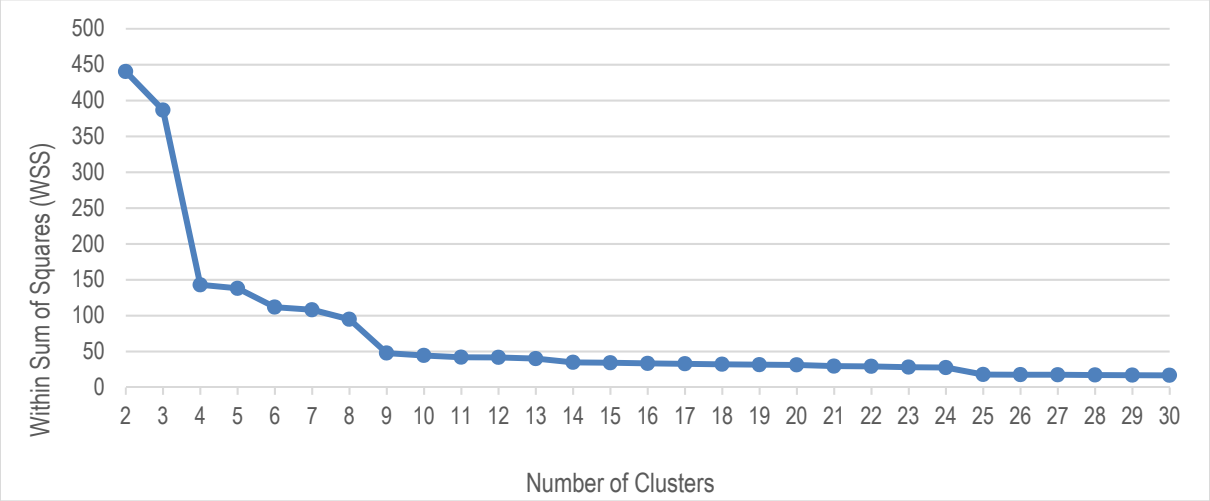
Das Projektteam kommt zum Schluss, dass aufgrund von statistischen und fachlichen Kriterien die Wahl von zehn Cluster am besten geeignet sind, um die typischen Verläufe im System der Sozialen Sicherheit zu beschreiben. Im Anschluss werden die Kriterien kurz beschrieben und deren Inhalt erläutert.

Elbow-Methode

Anhand der Elbow-Methode lassen sich in der Clusteranalyse aus statistischer Sicht die Anzahl Cluster ermitteln. Dabei werden im vorliegenden Projekt für jede Clusteranzahl die quadrierten Distanzen zwischen den Clusterelementen pro Cluster berechnet und über alle Cluster aufsummiert. Diese within-sum-of-squares (WSS) ist somit ein Mass für die Dissimilarität innerhalb der Cluster (je grösser WSS, desto heterogener die Cluster). Ab einer bestimmten Anzahl Cluster sinkt der marginale Gewinn einer noch grösseren Anzahl Cluster. Das zeigt sich durch einen Knick in der Grafik («Elbow»). In diesem Knick kann die statistisch effizienteste Anzahl Cluster abgelesen werden. Die Elbow-Methode

suggeriert, dass für die Kohorte 2010 aus statistischer Sicht mindesten 9 Cluster sinnvoll sind (sieh Grafik unten). Für die definitive Festlegung der Anzahl müssen jedoch zusätzlich inhaltliche Kriterien hinzugezogen werden.

Abbildung A 7: Elbow-Methode zur Bestimmung der Anzahl Cluster, Kohorte 2010



Quelle: SHIVALV-IK 2010-2015, eigene Berechnungen

«State distributions»

Für die Wahl der Anzahl Cluster werden in einem weiteren Schritt die «state distributions» pro Monat betrachtet. Dabei handelt es sich um die relative Verteilung der möglichen Verlaufszustände pro Monat über die 48 Monaten betrachtet. Die möglichen Grundzustände sind marché de l'emploi (ME), assurance chômage (AC), aide social (AS) und assurance invalidité (AI); Daraus ergeben sich 16 möglichen Kombinationen der vier Grundzustände inkl. die Abwesenheit aller vier Grundzustände (ni emploi ni prestation (NENP)). Die «state distribution plots» zeigen keine individuellen Verläufe, geben aber einen sehr guten Überblick über die in einem Cluster gruppierten Informationen. Bei dieser Analyse konnte festgestellt werden, dass bei einer Wahl von sechs Cluster oder weniger jene Verläufe, die vor allem durch die Invalidenversicherung bzw. durch die Sozialhilfe geprägt sind, nicht voneinander separiert werden. Die relativ hohe Anzahl von zehn Cluster stellt hingegen sicher, dass inhaltliche wichtige Differenzierung von erwerbstätigen Sozialhilfebeziehenden, Teilzeitarbeitslose, Kurz- und Langzeitarbeitslose sowie Mehrfacharbeitslose zuverlässig ausdifferenziert werden. Erwerbstätige Sozialhilfebeziehende hätten bei neun Clustern (Elbow) nicht identifiziert werden können. Eine hohe Heterogenität der Verläufe ist angesichts der hohen Zahl an Kombinationsmöglichkeiten der Grundzustände über 48 Monat, der sehr unterschiedlichen Risiken, die die drei sozialen Sicherungssysteme abdecken, und der Vielfalt an Erwerbsbiografien nicht verwunderlich. Entsprechend ist auch aus dieser Sicht eine relativ hohe Anzahl von 10 Clustern gerechtfertigt, um diese Heterogenität gebührend abbilden zu können.

Heterogenität innerhalb der Clusters

Als weiteres Kriterium kann man die Entwicklung der Heterogenität innerhalb der Clusters anschauen. Als Mass dafür kann die Summe der mittleren quadrierten Abweichungen eines jeden Verlaufs zu allen anderen verwendet werden (Sum of Squares). Dieses Mass ist jedoch abhängig von der Anzahl Fälle in einem Cluster, sodass grosse Cluster wie C3 und C1 generell höhere Werte aufweisen. Die Wahl von zehn Cluster anstatt von acht oder neun rechtfertigt sich auch in dieser Hinsicht, da bei zehn Cluster die Heterogenität innerhalb des Clusters mit den höchsten Sum of Squares nochmals deutlich gesenkt werden konnte. Die Clustergrössen mit einigen Heterogenitätsmassen für die Lösung mit zehn Cluster finden sich in untenstehender Tabelle.

Tabelle A 2: Clusterheterogenität, Kohorte 2010

Clusters	*Size*	*Sum of Squares*	*MaxDist*	*MeanDist*
1	16099	8.4079	89.4105	35.4176
2	1542	0.3149	95.6484	59.4487
3	59654	21.8373	41.6771	17.2327
4	9195	3.7701	74.6986	36.7017
5	4578	0.8455	77.0469	37.3845
6	8972	2.6192	79.2725	34.6271
7	13187	3.9101	77.2882	27.9983
8	4370	0.8428	95.7805	38.0395
9	6197	1.4309	81.0088	34.1617
10	2584	0.5427	81.2137	49.3994
Total	126378	1410.1878	96.0000	50.9605

Quelle: SHIVALV-IK 2010-2015, eigene Berechnungen

9.5 Zielerreichung

Ziele	Grad der Zielerreichung	Bemerkung
methodisch/technisch:		
Insgesamt wird die Entwicklung eines kohärenten Analyseansatzes für die SHIVALV+IK-Verlaufsdaten angestrebt.	70%	Der letzte Schritt zur Prädiktion, gewisse Sensitivitätsanalysen, Sankey-Plots und die Analyse weiterer Kohorten fehlen noch.
Definition eines Datenmodells und der dazu notwendigen Datenaufbereitungsschritte	90%	Das Datenmodell entspricht dem state of the art, die Analyse der Notwendigkeit von Datenkorrekturen ist noch nicht fertiggestellt.
Definition zentraler, aussagekräftiger Indikatoren und Visualisierungen um die Komplexität individueller Verläufe synthetisiert darzustellen	90%	Verlaufsindikatoren können noch weiterentwickelt werden und bestimmte Erkenntnisinteressen besser abbilden.
Datengetriebenes sequence clustering, um ähnliche Verlaufsmuster zusammenzufassen (unsupervised machine learning)	100%	Ein erfolgreicher Ansatz wurde entwickelt.
Entwicklung von Modellen zur Vorhersage der Zugehörigkeit zu den Clustern unter Anwendung von Methoden des supervised machine learning	10%	Erste Analysestrategien wurden festgelegt. Noch keine Umsetzung.

Ziele inhaltlich:		
Darstellung der typischen Erwerbs- und Bezugsverläufe von Personen die Leistungen aus dem System der sozialen Sicherung beziehen.	100%	Siehe sequence clustering.
Identifikation der Anzahl und Verlaufsbiographien von Personen, die a) sich rasche wieder in den Arbeitsmarkt integrieren b) überdurchschnittlilche Dequalilfiktations- und Ausgrenzungtendenzen aufweisen und c) die "nicht-systeminduzierte" Verlaufsmuster aufweisen ("Pendler", "Escaper", «Langzeitbeziehende», etc.[10] .).	100%	Siehe sequence clustering
Abschätzung der zukünftigen Entwicklungen typischer Verlaufsmuster aufgrund aktuell verfügbarer Informationen.	0%	Prädiktionen noch nicht umgesetzt.

[10] s.a.: Buhr, Petra (1995): Dynamik von Armut: Dauer und biographische Bedeutung von Sozialhilfebezug. Opladen: Westdeutscher Verlag.