Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Eidgenössisches Departement des Innern EDI
Bundesamt für Statistik BFS

Plausi++: Automatische Plausibilitätsprüfung der Qualität und Zuverlässigkeit von Administrativ- und Umfragedaten

Abschlussbericht zum Dateninnovationsprojekt

EXPERIMENTAL STATISTICS

Neuchâtel, 2024

Herausgeber:	Bundesamt für Statistik (BFS)	**Layoutkonzept:**	Sektion PUB
Auskunft:	persfinhs@bfs.admin.ch	**Grafiken:**	Sektion BILD-P
Redaktion:	Mehmet Aksözen, BFS	**Download:**	www.statistik.ch
Themenbereich:	15 Bildung und Wissenschaft	**Copyright:**	BFS, Neuchâtel 2024 Wiedergabe unter Angabe der Quelle für nichtkommerzielle Nutzung gestattet
Originaltext:	Deutsch		
Übersetzung:	Sektion BILD-P		

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

Inhalt

1 Management Summary 2
1.1 Ausgangslage 2
1.2 Zielsetzung 2
1.3 Zielerreichung 3
1.4 Projektorganisation 3
2 Detailliertere Zielbeschreibungen 4
2.1 Ziel 1: Test des Feedback-Mechanismus 4
2.2 Ziel 2: Einbezug der Datenlieferanten 6
2.3 Ziel 3: Verfolgung des Change Managements 8
2.4 Ziel 4: Modellaktualisierung 10
3 Lessons learned 12
4 Referenzen 13

1 Management Summary

1.1 Ausgangslage

Dieses Dokument baut auf dem Bericht «Plausi++: Automatische Plausibilitätsprüfung der Qualität und Zuverlässigkeit von Administrativ- und Umfragedaten» [1] auf und soll als Ergänzung und Abschluss des Pilotprojekts dienen.

Die Idee für das Pilotprojekts war es, durch einen Machine Learning Algorithmus die bislang manuelle Plausibilisierung mit einem automatisierten Teil zu ergänzen und potenzielle bislang nicht identifizierte Fehler (Problemfälle) zu finden.

Die genutzten Daten basieren auf den Daten der Personalstatistik der universitären Hochschulen, ergänzt um Daten aus den Statistiken der Studierenden und Examen sowie mathematischen Kennzahlen (siehe [1], S.6).

Das Modell gibt für jeden Datensatz eine vorhergesagte Personalkategorie zurück. Wenn diese vorhergesagte Personalkategorie nicht der gelieferten entsprach (Problemfall), wurde nach Kandidaten von Variablen gesucht, deren Werte zu diesem Unterschied beigetragen haben könnten.

1.2 Zielsetzung

Im Ausblick des vorherigen Berichts [1], S. 6 werden mehrere Punkte für die Weiterführung des Projekts aufgelistet. Diese entsprechen der Zielsetzung der hier vorgestellten Arbeiten und betreffen wie folgt:

1. Test des Feedback-Mechanismus: Der in einem ersten Entwurf erstellte Feedback-Mechanismus gibt einen Hinweis zu Problemfällen (Kandidaten von Variablen, die fehlerhaft sein können), welche dann von den Datenlieferanten überprüft werden.
2. Einbezug von Datenlieferanten: Die Datenlieferanten erhalten das Feedback zu möglichen Kandidaten und überprüfen diese. Die Datenlieferanten erläutern, weshalb die Daten korrekt sind, oder liefern korrigierte Daten. Zudem können sich organisatorische Unterschiede zwischen den Datenlieferanten in den Daten durch systematische Abweichungen abbilden, die wiederum Gruppen von Problemfällen ergeben können.
3. Change Management an den universitären Hochschulen berücksichtigen: Neben den Problemfällen können Revisionen oder andere Änderungen an den Hochschulen dazu führen, dass sich die Datenstrukturen ändern (z.B. Wahl einer anderen Personalkategorie als bislang wegen interner Anpassungen), so dass das Modell vermehrt Problemfälle finden kann, die auf diese Änderungen zurückführbar sind.
4. Monitoring des Modells: Wenn Unterschiede in den Daten zwischen zwei Jahren während der Erhebung gemessen werden, wird der Algorithmus mit den verfügbaren Daten neu trainiert. Ansonsten erfolgt ein neues Training mit dem Algorithmus nach dem Abschluss der Erhebung, welcher mit dem alten Modell verglichen wird.

1.3 Zielerreichung

Ziel 1: Den Datenlieferanten wurden vier Gruppen von möglichen Problemfällen zugestellt, die etwa 28% aller Fälle ausmachen. Um den Aufwand bei den Datenlieferanten möglichst gering zu halten wurde der Feedback-Mechanismus zusätzlich vor dem Versand überprüft. So wurden Problemfälle gemeldet bei denen die Wahrscheinlichkeit, einen Fehler zu finden, besonders hoch erschien. Zusätzlich wurden zu diesen Fällen weitere Angaben hinzugezogen.

Ziel 2: Die Datenlieferanten konnten alle ausgewählten Problemfälle (potenzielle Fehler) als korrekt zurückmelden, auch solche die zum Teil selten vorkommen. Aus weiteren Rückmeldungen der Datenlieferanten zu einer weiteren Gruppe von möglichen Problemfällen (32% aller Fälle mit Bezug zur Verwaltung der Hochschule) stellte sich heraus, dass es mehrere strukturelle Gründe für Abweichungen geben kann:

- Die Erfassung der Daten erfolgt unterschiedlich (Berechnung oder Vollerhebung).
- Die Verwaltung der Institutionen ist unterschiedlich gestaltet, was sich auf die Personalkategorien unterschiedlich auswirkt.
- Das Leistungsangebot der Datenlieferanten hat einen unterschiedlichen Fokus (mehr Lehre) gegenüber der Mehrheit (mehr Forschung und Entwicklung).

Ziel 3: Mit der Berechnung eines Population Stability Index wurde ein Modul ergänzt, um die Verteilung zwischen den Jahren zu überprüfen und die Einsatzfähigkeit des vorherigen Algorithmus zu beurteilen. Die Verteilung der Daten pro Personalkategorie, pro Hochschule und bei weiteren gelieferten Daten unterschied sich zwischen den Jahren nicht merklich.

Ziel 4: Nach jeder Erhebung wurde der Algorithmus mit den aktuellsten Daten neu trainiert. Die sehr hohe Genauigkeit dieser jährlich neu trainierten Modelle blieb stabil. Die Genauigkeit der Modelle bei der Vorhersage der Personalkategorie änderte sich auch kaum, wenn Modelle auf Daten anderer Jahre angewandt wurden.

Die Datenqualität der Personalstatistik ist sehr hoch. Die Unterschiede an den Hochschulen sind erhebungstechnischer und struktureller Natur und wurden zum Abschluss des Projekts mit den Hochschulen besprochen. Bei Bedarf seitens der Hochschulen kann der Algorithmus wieder genutzt werden.

1.4 Projektorganisation

Die Weiterführung des Projekts fiel in eine Phase, in der sowohl die Datenlieferanten als auch die Projektleitung wenig Ressourcen für einen Testdurchlauf hatten (v.a. die Corona-Pandemie). Um den Aufwand für die Institutionen auf ein geringes Mass zu reduzieren, hat die Projektleitung etwas mehr Zeit und Ressourcen investiert und eine Auswahl an potenziellen Problemfällen mit hoher Wahrscheinlichkeit für eine Korrektur zusammengestellt. Auf Grund dieser etwas schwieriger zu planenden Situation hat die Projektleitung den Einbezug weiterer Mitarbeitenden aus den Sektionen BILD-P oder BILD-S auf die notwendige Datenaufbereitung seitens der IT der Sektion BILD-P auf das Minimale reduziert.

2 Detailliertere Zielbeschreibungen

2.1 Ziel 1: Test des Feedback-Mechanismus

Bei 8% der Datensätze entsprach die vorhergesagte Personalkategorie nicht der gelieferten Personalkategorie. Die Tabelle 1 zeigt die Sensitivität und Spezifität der einzelnen Personalkategorien (D = Direktions- und administrativ-technisches Personal, A = Assistierende und Doktorierende, U = übrige Dozierende, P = Professorinnen und Professoren, W = wissenschaftliche Mitarbeitende). Die Personalkategorie D wurde zu 90.5% (Sensitivität) korrekt zugeordnet. Für die Spezifität hat sie einen Wert von 98.7%, d.h. bei 1.3% (100% - 98.7%) der Fälle in denen die Personalkategorie eine andere als die Personalkategorie D war, wurde fälschlicherweise Personalkategorie D vorhergesagt.

	A	D	P	U	W
Sensitivität	0.944	0.905	0.951	0.933	0.887
Spezifität	0.960	0.987	0.996	0.985	0.968

Tabelle 1 Sensitivität und Spezifität der Vorhersagen der Personalkategorien durch das Modell

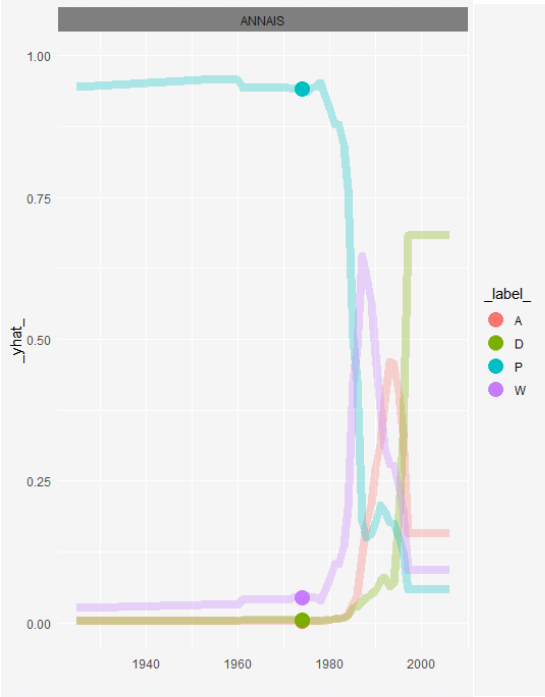
Bei 32% der Problemfälle wurde statt der Personalkategorie D eine andere Personalkategorie vorhergesagt. Zudem gab es vier Kombinationen von gelieferter und vorhergesagter Personalkategorie, die zusammen 28% der Problemfälle ausmachten. Wie in Tabelle 2 zu sehen kamen die übrigen Kombinationen von gelieferter und vorhergesagter Personalkategorie auf 40%.

gelieferte Personalkategorie	vorhergesagte Personalkategorie	Anzahl der Problemfälle	Anteil an allen Problemfällen
D	A, P, U, W	2667	32%
A	W	1240	28%
U	W	489	
W	U	357	
U	P	207	
übrige 12 Kombinationen		3302	40%
Total		8262	100%

Tabelle 2 Anzahl Problemfälle für die Kombinationen der gelieferten und vorhergesagten Personalkategorien

Die Ermittlung von Kandidaten, die als Erklärung für die Problemfälle dienen können, erfolgt durch eine Untersuchung der Daten nach dem ceteribus paribus Prinzip, wenn die vorhergesagte Personalkategorie nicht der gelieferten entsprach. In diesen Fällen wurden die Werte aller anderen Variablen ausser der Personalkategorie einzeln verändert, um zu schauen, ob bei einer anderen Ausprägung der jeweiligen Variablen die vorhergesagte und gelieferte Personalkategorie übereinstimmen (siehe [1], S.4, 16, 18 für mehr Informationen).

Dieses Prinzip lässt sich besonders gut anhand der Variablen «Alter» darstellen (G 1).



G 1 Wahrscheinlichkeit der Personalkategorien in Abhängigkeit des Alters für eine Person, aus [1], S.17

Auf der y-Achse (vertikal) ist die Wahrscheinlichkeit für die Personalkategorien eingetragen und horizontal auf der x-Achse das Geburtsjahr. Bis zum Geburtsjahr 1978 ist die Wahrscheinlichkeit, dass diese Mitarbeitende eine Professorin ist über 90% (P, blaue Linie). Bei einem jüngeren Geburtsjahr geht die Wahrscheinlichkeit zurück und ab dem Geburtsjahr 1985 ist die Wahrscheinlichkeit für diese Person eine wissenschaftliche Mitarbeitende (W, violette Linie) zu sein am höchsten. Diese Wahrscheinlichkeit steigt an seinem höchsten Punkt auf 65% und fällt dann wieder (für mehr Informationen, [1], S.16).

Bei der Rücksprache mit den Fachexpertinnen stellte es sich heraus, dass das Alter keine Fehlerquelle darstellt, da dieses über die gelieferte AHVN13 der Datenlieferanten separat durch das BFS überprüft wird. Daher wurden diese Fälle aus dem Feedbackmechanismus rausgenommen.

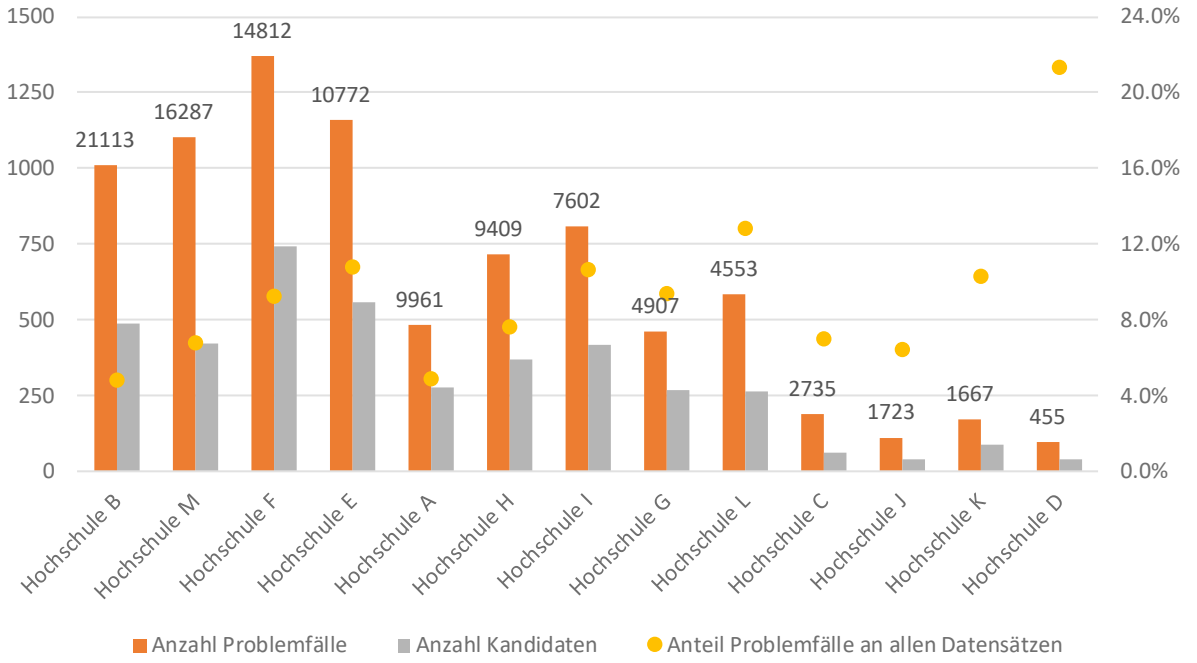
Die Tabelle 3 zeigt, dass für 49% (4055/8262) der nicht übereinstimmenden Personalkategorien der Feedback-Mechanismus potentielle Kandidaten ermitteln konnte. Bei einem anderen Wert dieser Kandidaten hätte die Modellvorhersage in der gelieferten Personalkategorie resultiert. Am häufigsten wurden Kandidaten für Problemfälle gefunden, bei denen das Modell fälschlicherweise die Personalkategorie W oder A (1357 oder 1344 Fälle) vorhersagte. Das machte jeweils 33% aller anders vorhergesagten Personalkategorien aus. Innerhalb dieser Personalkategorien bedeutete das, dass bei 49% der fälschlicherweise der Personalkategorie W bzw. A zugewiesenen Fälle ein Kandidat zur Verfügung stand, nach deren potentieller Korrektur die gelieferte Personalkategorie erwartet werden könnte.

Vorhergesagte Personalkategorie	**A**	**D**	**P**	**U**	**W**	**Total**
Problemfälle	2761	1017	381	1326	2777	8262
Kandidaten	1344	423	185	746	1357	4055
Anteil der Kandidaten an den Problemfällen pro Personalkategorie	49%	42%	49%	56%	49%	49%
Verteilung der Kandidaten auf die Personalkategorien	33%	10%	5%	18%	33%	100%

Tabelle 3 Durch das Modell ermittelte Problemfälle mit den vorhergesagten Personalkategorien und möglichen Erklärungskandidaten um auf die gelieferten Personalkategorien zu kommen. Mit der Verteilung der Kandidaten auf die Personalkategorien, z.B. 10% aller Kandidaten hätten bei einem entsprechenden Wert dazu führen können, dass statt Personalkategorie D die gelieferte Personalkategorie durch das Modell vorhergesagt wird.

2.2 Ziel 2: Einbezug der Datenlieferanten

Die Datenlieferanten unterscheiden sich stark in Bezug auf die Anzahl der Datensätze, die Anzahl der Problemfälle und die Anzahl der Kandidaten (G 2). Die Anzahl der Datensätze entspricht nicht der Anzahl der Mitarbeitenden (MA). MA haben häufig Aufgaben in mehreren Leistungsarten (Lehre, Forschung & Entwicklung, Dienstleistungen und Weiterbildung) mit unterschiedlicher Finanzierung, sodass pro MA mehrere Datensätze geliefert werden. Die Hochschule B mit den meisten Datensätzen (21'113) hat mit 1010 die viertmeisten Problemfälle.



G 2 Die Reihenfolge der Hochschulen von links nach rechts erfolgt nach Anzahl der Datensätze (Zahl über den orangenen Balken). Die Anzahl der Problemfälle und der Kandidaten entsprechen der linken Achse und der Anteil der Problemfälle an allen Datensätzen ist auf der rechten Achse zu sehen.

In vorherigen Rückmeldungen der Hochschulen wurde darauf hingewiesen, dass für die Personalkategorie D die Tätigkeitsanteile pauschal berechnet werden und dass für die anderen Personalkategorien unterschiedliche Verfahren für die Ermittlung der Tätigkeitsanteile eingesetzt werden. Bei den Datenlieferanten wurde in Erfahrung gebracht, ob sie die Unterschiede in der Ermittlung und die Besonderheit der Personalkategorie D in Tabelle 2 erklären können.

Erfassung der Tätigkeitsanteile

Sieben Hochschulen haben für das Projekt ihre Erfassungsmethode für die Tätigkeitsanteile erläutert.

Die Hochschulen B, M, A, H und J haben im Vergleich zu den anderen Hochschulen einen geringeren Anteil an Problemfällen über alle Datensätze hinweg. Auf Rückfrage haben die ersten vier Hochschulen angegeben, dass sie die Erfassung ihrer Tätigkeitsanteile mit Pflichtenheften, Schlüsseln oder höher aggregierten Einheiten (z.B. pro Institut) durchführen. Das heisst die Ermittlung der Tätigkeitsanteile basieren vermehrt auf Berechnungen statt auf Erhebungen pro Person. Die Hochschule J lässt eine Erhebung pro MA durchführen.

Die Hochschulen E und G haben einen höheren Anteil an Problemfällen über alle Datensätze hinweg. Sie haben angegeben, dass sie die Erhebungen für die Personalkategorien A, P, U und W pro MA durchführen. Dabei könne es vorkommen, dass die erbrachten Leistungen von MA nicht den vertraglich ausgemachten gemäss der Personalkategorie entsprechen müssen. In diesen Fällen vermuteten die Ansprechpartner daher, dass die Hochschulen, die ihre Erhebung für die Tätigkeitsanteile pro MA machen, durch das Modell eher Problemfälle zugewiesen bekommen, auch wenn die Tätigkeitsanteile korrekt erhoben wurden.

Hier erscheint die Anzahl der Datensätze pro MA eine gute Kennzahl. Die Hochschulen mit einem geringeren Anteil an Problemfällen über alle Datensätze hinweg haben mehr Datensätze pro MA. Es scheint, dass die Leistungen viel detaillierter anhand des Pflichtenhefts u. ä. mit jeweils der passenden Personalkategorie erfasst und den Personen zugeordnet wird. Bei den Hochschulen mit Erhebungen pro MA werden die Aufgaben stärker pro MA gebündelt geliefert.

Unterschiede zwischen den Personalkategorien

Bei der Personalkategorie D erläuterten die Hochschulen, dass die Tätigkeitsanteile anhand der Pflichtenhefte oder der Arbeitsverträge berechnet werden. Bei den Problemfällen hat sich herausgestellt, dass es eine Rolle spielt, ob eine Hochschule eher zentral verwaltet wird oder nicht. Daher kann es vorkommen, dass das Modell bei der Personalkategorie D eine andere Personalkategorie erwartet. Als Kandidat gibt es dann an, dass diese MA zur Personalkategorie D wechseln würden, wenn ihnen für die Variable Fachbereich der Fachbereich «Zentralverwaltung» zugewiesen worden wäre.

Wegen dieser Besonderheit der Personalkategorie D erfolgten die Fragen an die Datenlieferanten zu den Problemfällen und Kandidaten der Personalkategorien A, P, U und W. Um den Aufwand für die Datenlieferanten gering zu halten, wurden die Daten umfangreich vorbereitet, so dass die Datenlieferanten möglichst viele Informationen erhalten haben. Wenn die Wahrscheinlichkeiten für die gelieferte und vorhergesagte Personalkategorie nahe beieinander lagen, wurden diese nicht ausgewählt. Es wurden vor allem Datensätze gewählt, bei denen die Wahrscheinlichkeit für die gelieferte Personalkategorie durch das Modell als besonders tief und für eine andere Personalkategorie als besonders hoch eingeschätzt wurde. Diese Datensätze mit Problemfällen wurden ergänzt um weitere korrekt vorhergesagten Datensätze der jeweiligen Person, um ein Gesamtbild der Anstellung dieser oder dieses MA an der jeweiligen Hochschule zu erhalten.

Die Hochschulen konnten auf alle Fragen Erklärungen abgeben:

- Das Modell hat Personalkategorie U statt der gelieferten Personalkategorie W vorhergesagt, da für diese MA mehr Tätigkeitsanteile in der Lehre gemeldet wurden als das Modell erwartet hat.
  Erklärung seitens der Hochschulen: Unterschiede bei den Tätigkeitsanteilen bei der Zuteilung zu den Personalkategorien W und U können vorkommen. In der Personalkategorie W angestellte MA können zum Beispiel wegen Vakanz einer Professur mehr Lehre im Laufe des Jahres anbieten. Zudem können organisatorische Unterschiede eine Rolle spielen. Die Hochschulen L und D mit den höchsten Anteilen an Problemfällen über alle Datensätze hinweg haben einen spezielleren Fokus (mehr Lehre). Dadurch sind fast die Hälfte der Personalkategorie W einer anderen Personalkategorie zugewiesen worden.
- Das Modell hat für MA statt der Personalkategorie U die Personalkategorie W erwartet, da deren Anstellungsstatus «fest angestellt» waren.
  Erklärung seitens der Hochschulen: Dieser Anstellungsstatus komme zwar seltener vor als andere Anstellungsstatus, aber er komme vor.
- Statt der Personalkategorie A sagte das Modell wegen des Tätigkeitsanteils in der Forschung und Entwicklung die Personalkategorie W vorher.
  Erklärung seitens der Hochschulen: Diese MA hatten mehrere Verträge hintereinander. Mit dem Abschluss des Doktorats in der Personalkategorie A wurden MA als Postdoc in der Personalkategorie W angestellt.
- MA in der Personalkategorie U wurden durch das Modell die Personalkategorie P zugewiesen.
  Erklärung seitens der Hochschulen: Manche MA haben mehrere Verträge seriell oder parallel. Das heisst sie sind eventuell in einem Fachbereich in der Personalkategorie P angestellt, aber haben auch Verträge in der Personalkategorie U in anderen Fachbereichen.

2.3 Ziel 3: Verfolgung des Change Managements

Wenn die Hochschulen grössere Änderungen in ihrem Personal zwischen den Jahren haben, kann es sein, dass das bisherige Modell diesen Änderungen nicht genug Rechnung trägt. Das kann dazu führen, dass die Genauigkeit (Accuracy) des Modells abnimmt und dass die Anzahl der Problemfälle steigt und somit die Datenqualität scheinbar sinkt.

Mit dem neuen Modul zum Population Stability Index (PSI) wird die Nutzbarkeit des Modells von einem Jahr auf das andere seitens der Daten überprüft. Der PSI ist eine statistische Messgrösse, die den Unterschied zwischen zwei Verteilungen angibt. In diesem Fall zeigt er an, wenn eine Verschiebung der Verteilung zwischen des bisherigen Datensatzes und eines neueren Datensatzes vorliegt. Für eine detailliertere Beschreibung und weitere Literaturhinweise sei auf die Dissertation von Bilal Yurdakul «Statistical Properties of Population Stability Index» [2] verwiesen.

Der PSI wird berechnet, indem eine multinominale Klassifizierung der kardinalgeordneten Personalkategorie in sogenannte *bins* erfolgt. Wenn die Variable nicht kardinalgeordnet wäre, werden diese *bins* zunächst definiert. Hier wird die Definition aus [2] verwendet, wobei N die Population des Vorjahres ist und M die Population des neuen Jahres.

In [2], S.3 wird der PSI definiert als:

$$PSI = \sum_{i=1}^{B} \left(\frac{n_i}{N} - \frac{m_i}{M}\right) \times \left(\ln\frac{n_i}{N} - \ln\frac{m_i}{M}\right)$$

wobei n_i and m_i die Anzahl im i-ten *bin* ist.

Als Daumenregel wird angenommen, dass bei einem PSI < 0.1 das existierende Modell weiterhin genutzt werden kann, dass bei einem PSI bis 0.25 die Gründe für die Erhöhung des PSI untersucht werden sollten und dass bei einem PSI über 0.25 ein neues Modell entwickelt werden sollte ([2], S.1). In seiner Arbeit weisst Yurdakul auf die Grenzen dieses verbreitet angewandten Ansatzes hin ([2], S.41). Er schlägt eine Präzisierung der Kennzahl für sich stark unterscheidende N und M (Anzahl N und M bis zu 1000) sowie eine Abhängigkeit von der Anzahl *bins* (insbesondere bei 10 oder mehr *bins*) vor ([2], S.40).

Für die Berechnung des PSI wurde der R-Code aus «psi: Population Stability Index (PSI)» [3], ebenfalls von Bilal Yurdakul, verwendet.

In unserem Fall ist die Frage, ob die Verteilung der Personalkategorien sich zwischen zwei Jahren verändert hat. Es wurden jährlich die Verteilungen der Personalkategorien zwischen zwei Jahren verglichen. Dabei wurden die gesamten Datensätze zweier Jahre vollständig verwendet. Zudem wurden auch zur Sicherheit die Verteilungsunterschiede pro Hochschule, pro Personalkategorie und bei weiteren gelieferten Variablen untersucht. In unserem Fall, mit einer hoher Anzahl N und M in einer ähnlichen Grössenordnung sowie fünf Personalkategorien (*bins*), ist die Daumenregel nutzbar.

Das Ergebnis des Vergleichs zweier Datensätze gibt der R-Code aus [3] wie in Tabelle 4 aus.

```
$res
           psi    cv.zscore        psi**     cv.chisq   ci
1 0.0001030045 0.0001589636 0.0001023611 0.0001743115 0.95

$tbl
  bin n.base   pct.base n.target pct.target        psi.b
1   A  36109 0.34066380    37799 0.33784702 2.338740e-05
2   D  27993 0.26409487    29680 0.26527949 5.301826e-06
3   P   5156 0.04864335     5277 0.04716576 4.557853e-05
4   U  17730 0.16727046    18850 0.16848108 8.730239e-06
5   W  19008 0.17932752    20276 0.18122665 2.000653e-05
```

Tabelle 4 PSI beim Vergleich zweier Datensätze

Der PSI für den Populationsvergleich im oberen Teil des Ergebnisses zeigt einen Wert unter 0.1. Zu den fünf Personalkategorien (*bins*) werden unter anderem die Anzahl pro Datensatz (base = alter Datensatz, target = neuer Datensatz) sowie die PSI-Werte angegeben. Die PSI-Werte pro Hochschule und bei den weiteren gelieferten Variablen waren ebenfalls jeweils kleiner als 0.1. Das heisst gemäss diesen PSI-Werten kann das bisherige Modell weiterverwendet werden.

2.4 Ziel 4: Modellaktualisierung

Wenn der PSI während der Datenerhebung durch das BFS eine Verschiebung in der Verteilung einer Hochschule angeben würde, müsste das Modell aktualisiert werden. Bislang erfolgte die Berechnung des PSI erst nach Abschluss der gesamten Erhebung. Daher wurde das Modell mit dem neuen Datensatz des aktuellsten Jahres neu erstellt.

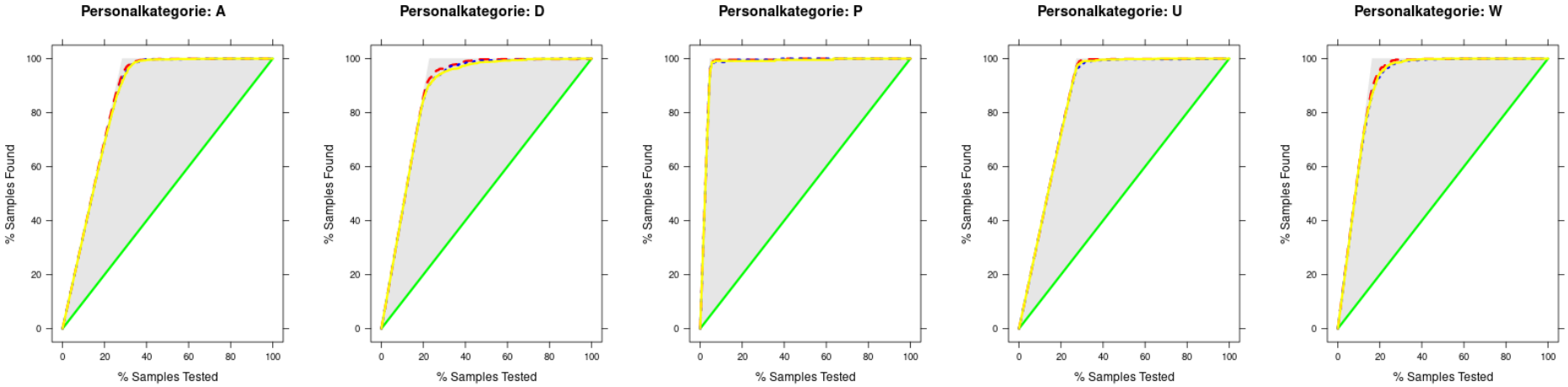
Zur Sicherheit wurde das jeweils neue Modell mit dem alten verglichen:
1. Die Accuracy der Modelle mit den ihnen zugrundeliegenden Daten wurde mit mehreren Hyperparametern für mehrere Datensätze beziehungsweise Jahre untersucht.
2. Die älteren Modelle wurden mit neueren Datensätzen auf ihre Performanz hin betrachtet, anhand von Gains-Kurven.
3. Die älteren Modelle wurden mit neueren Datensätzen auf ihre Sensitivität und Spezifizität hin betrachtet, anhand von ROC-Kurven.

Zu 1. Wie bereits in [1], S.15, erwähnt, wurde die Accuracy der jährlich neu erstellten Modelle (Algorithmus: «Gradient Boosting Machines») mit mehreren Hyperparametern getestet. Die Accuracy dieser neueren Modelle hat sich im Vergleich zu [1] nicht geändert: 0.94 (0.935 - 0.945 mit einem 95% Vertrauensintervall).

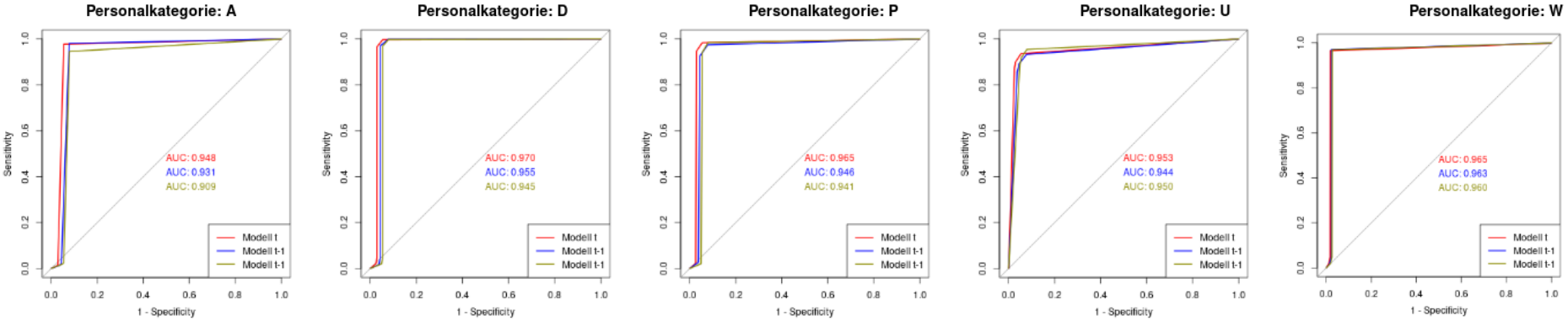
Zu 2. Für den Performanzvergleich wurden die Modelle der Vorjahre auf den jeweils aktuellen Datensatz angewandt. Gains-Kurven können die Performanz mehrerer Modelle abbilden (siehe [1], S.14 und für weitere Erläuterungen [4]). G 3 zeigt die Performanz dreier Modelle beim Einsatz auf den Datensatz aus dem Jahr t und eine Baseline. Hier wurde das Modell t-2 mit den Daten des Vorvorjahrs (in Gelb), Modell t-1 mit den Daten des Vorjahrs (in Blau) und Model t mit den aktuellen Daten (in Rot) erstellt. Die Baseline ist in Grün. Die Überlappung der drei Gains-Kurven zeigt, dass die Performanz der Modelle sich wenig unterscheiden. Wie schon in [1] erwähnt, zeigt die Steilheit der Kurven bei welchen Personalkategorien die Modelle besonders gut (z.B. Personalkategorie P) funktionieren.

Zu 3. G 4 zeigt Receiver Operating Characteristic (ROC) Kurven für die gleichen drei Modelle pro Personalkategorie (siehe [5] für das Package pROC). Dabei werden die Sensitivität und die Spezifizität in Bezug gesetzt. Bei einem perfekten Modell wären die Sensitivität und die Spezifität gleich 1 (damit 1 – Spezifität = 0). Dann wäre die Fläche unter der Kurve (Area Under Curve, AUC) gleich 1. Die älteren Modelle haben kleinere AUC-Werte. Das heisst sie haben geringere Sensitivität und/oder Spezifität, wenn auch die Modelle alle hohe AUC-Werte haben.

Der Vergleich der Modelle dreier Jahre zeigt, dass das neuste Modell am besten abschneidet, aber die anderen immer noch als sehr gute Modelle gesehen werden können. Von einem Jahr auf das andere kann ein Modell erneut verwendet werden. Das wäre wichtig, wenn ein vorjähriges Modell während einer aktuellen Erhebung genutzt würde. In dem Fall gäbe es noch keinen vollständigen Datensatz, so dass die Einsatzfähigkeit des vorjährigen Modells eingeschätzt werden müsste. Mit den obigen Ergebnissen wäre diese Einsatzfähigkeit gegeben gewesen.



G 3 Gains-Kurven der Modelle aus den Jahren t (in Rot), t-1 (in Blau) und t-2 (in Gelb) beim Einsatz auf den Datensatz des Jahres t für die fünf Personalkategorien, Baseline in Grün.



G 4 ROC-Kurven der Modelle aus den Jahren t (in Rot), t-1 (in Blau) und t-2 (in Gelb) beim Einsatz auf den Datensatz des Jahres t für die fünf Personalkategorien.

3 Lessons learned

Der Ansatz Fehler in den Daten zu finden, hat dazu geführt, dass angebotsseitige (Fokus zweier Hochschulen auf die Lehre), strukturelle (mehr oder weniger zentrale Verwaltung) und erfassungstechnische (Berechnungen oder Erhebungen für die Tätigkeitsanteile, teilweise abhängig von der Personalkategorie) Unterschiede gefunden wurden. Diese beeinflussen die Modellbildung soweit stark genug, dass viele vom Modell ermittelten Problemfälle keine realen Problemfälle darstellen. Die Rückmeldungen zu den wahrscheinlichsten Problemfällen hat gezeigt, dass diese seitens der Hochschulen erklärt werden können. Wie im Bericht zuvor erwähnt kann zwar nicht ausgeschlossen werden, dass Fehler Teil des Modells sind, siehe auch [1], S.6. Dass das Modell bislang keine bestätigten Fehler in den Datensätzen gefunden hat, spricht jedoch für eine hohe Datenqualität.

Die Weiterentwicklung des Feedback-Mechanismus mit neueren Ansätzen wie sie in den letzten Jahren hinzugekommen sind ([1], S.6), erschien nach den informativen und verständlichen Rückmeldungen der Hochschulen nicht als prioritär.

Für die Modellierung wurden die Daten mit Studierenden- und Examensdaten verknüpft. Diese wurden bei Problemfällen wegen des Persönlichkeits- und Datenschutzes nicht mitgeschickt. Aber die ergänzten Daten könnten eine Rolle spielen, zumal es für über 50% der Problemfälle keine Kandidaten aus den Erhebungsdaten gab: Zum Beispiel, wenn MA neben der Hochschulanstellung studieren und nicht möchten, dass dies an der Hochschule bekannt ist. Das Modell hätte diese Information und könnte wegen des Studiums eine andere Personalkategorie zuweisen. Zur Sicherheit wurden den Hochschulen nur Variablen aus der Erhebung der Hochschulpersonalstatistik als Kandidaten zurückgemeldet.

Das Coaching und das training-on-the-job verliefen gut. Die BFS-Mitarbeitenden brauchen bei einer derartigen Begleitung gute bis sehr gute Programmier- und Fachkenntnisse, um die Vorschläge aus dem Coaching umsetzen zu können.

Die gleichbleibend hohe Qualität der Modelle über mehrere Jahre hinweg und die hohe Populationsstabilität bedeutet, bezogen auf die Datensätze der Hochschulen, dass die gelieferten Daten der Hochschulen sich von einem Jahr auf das andere nicht in ihrer Qualität unterscheiden, obwohl es sicherlich zu Personalwechseln an den Hochschulen kommt.

Die Datenqualität der Personalstatistik ist sehr hoch, daher werden die Plausibilitätsprüfungen mittels Machine Learning nicht standardmässig in die Kontrollarbeiten implementiert. Die Unterschiede an den Hochschulen sind erhebungstechnischer und struktureller Natur. Diese wurden zum Abschluss des Projekts mit den Hochschulen besprochen. Bei Bedarf seitens der Hochschulen, zum Beispiel bei der Umstellung der Datenerfassung an den Hochschulen, kann der Algorithmus wieder genutzt werden.

Bei der Anwendung des Ansatzes dieses Projekts auf andere Projekte wäre es wichtig in einem ersten Schritt grössere der Inputdaten inhärente strukturelle Unterschiede ausfindig zu machen um diese allenfalls in die Entwicklung der Anwendung des Algorithmus auf einzelne Datensätze zu integrieren. Der Algorithmus ist auch unter unterschiedlichen Rahmenbedingungen anwendbar, skalierbar und mit Anpassungen wiederverwendbar.

4 Referenzen

[1] Ruiz, C. et al., 2019: «Plausi++: Automatische Plausibilitätsprüfung der Qualität und Zuverlässigkeit von Administrativ- und Umfragedaten»,
Webseite: https://www.bfs.admin.ch/bfs/de/home.assetdetail.9847917.html (letzter Zugriff 25.03.2024)

[2] Yurdakul, B., 2018: «Statistical Properties of Population Stability Index», Dissertation,
Webseite: https://scholarworks.wmich.edu/cgi/viewcontent.cgi?article=4249&context=dissertations (letzter Zugriff 25.03.2024)

[3] Yurdakul, B., 2018: «psi: Population Stability Index (PSI)»,
Webseite: https://rdrr.io/cran/PDtoolkit/man/psi.html (letzter Zugriff 25.03.2024)

[4] RPubs by RStudio, 2020: Introduction to Cumulative Gains, Kappa Measures and Their Applications, Webseite: https://rpubs.com/tshute/577248 (letzter Zugriff 25.03.2024)

[5] Robin, X., 2023: pROC: Display and Analyze ROC Curves (uned.ac.cr),
Webseite: https://mirror.uned.ac.cr/cran/web/packages/pROC/pROC.pdf (letzter Zugriff 25.03.2024)

Herzlichen Dank an Prof. Dr. Diego Kuonen für seine Unterstützung und Vorschläge!