

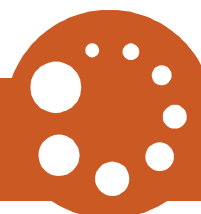


ML_SoSi _ Machine learning and social security

Pilot project within the framework of the FSO's
data innovation strategy

Management Summary

EXPERIMENTAL STATISTICS



Neuchâtel, 2023

Published by: Federal Statistical Office (FSO)
Editors: Luzius von Gunten, Frank Schubert,
Brandon Qorri Gonzalez, Athanassia Chalimourda
Topic: 00 Statistical Basis and Overviews
Original text: German
Translation: FSO language services

Layout concept: Section PUB
Downloads: www.statistics.ch
Copyright: FSO, Neuchâtel 2023
Reproduction with mention of source authorised
(except for commercial purposes)

Management Summary

The course of unemployment can vary greatly for the persons concerned. Individual unemployment trajectories are characterised - amongst other things - by (repeated) receipt of social benefits from the social security system (unemployment and invalidity insurance, social assistance), workforce re-entry or withdrawal, and migration. In the "ML-SoSi" pilot project, information on individual trajectories is analysed using inductive statistical methods to identify typical trajectory patterns. In addition to the findings obtained, the aim of the project is to develop a data-driven approach to the analysis of individual trajectories in longitudinal data for official statistics.

The data base is an anonymised linked data set containing monthly information on individual receipts of social benefits from the social assistance (SH), the invalidity insurance (IV) and the unemployment insurance (ALV), as well as on employment (social security accounts/IK). In this report, the abbreviation "SHIVALV+IK"¹ is used for this data set. The statistical population comprises persons aged between 18 and 65 who were new recipients of daily allowances from the unemployment insurance (UI) from 2010-2015. Analysis is conducted on the basis of annual cohorts. Included in the analysis is information on the receipt of social insurance and social assistance benefits as well as on employment during the subsequent 48 months (4 years).

In implementation of the methods, typical trajectory patterns are first identified with the 2010 cohort in a two-stage sequence clustering procedure (unsupervised machine learning). These patterns are then analysed in graph form (state distribution plots) by means of trajectory indicators and their content interpreted. This initial cluster solution is then transferred to the cohorts of the following years 2011-2015 using supervised machine learning (prediction). The validity of the model is checked with each transfer by assessing various criteria.

This approach focuses on the recognition and analysis of aggregated typical trajectory patterns and their transfer to further cohorts. The use of individual predictions for any purpose whatsoever is excluded.

In total, ten clusters were identified to describe the typical trajectory patterns of new daily allowance recipients (Table 1), most of which also remain stable in the 2010-2015 cohort comparison (8 out of 10 of the typical trajectory patterns). There are several clusters whose content shows persons who, after a phase of receiving daily allowances, re-enter the labour force. The clusters can be differentiated by length of daily allowance receipt (clusters 1 and 2), by presence of an interim earnings phase (cluster 3) and by multiple periods of daily allowance receipt with interim employment (cluster 4). Furthermore, clusters emerge showing a clear tendency towards either permanent receipt of invalidity pensions or social assistance benefits (Table 2; clusters 5, 6, 7, 8 and 9). Among these are two clusters with new receipt of these benefits (clusters 5 and 9) and two clusters with distinct phases of supplementary income from employment (clusters 6 and 7) as well as one cluster with persons who were already permanently or repeatedly dependent on social assistance prior to receiving daily allowances. One final cluster (cluster 10) incorporates persons no longer recorded on an ongoing basis by the systems investigated (social assistance, invalidity and unemployment insurance, social security accounts /employment) during the observation period. The project has shown sequence clustering to be a promising algorithm for producing results whose content is valid and which are analytically relevant. It enables a considerable reduction in the complexity of the trajectory data, thus increasing the possibilities for analysis by recognising patterns that could not be anticipated deductively.

Time series data are of great importance to make this information even more relevant for the official statistics audience, including for political steering. However, the initial cluster solution cannot simply

¹ SHIVALV+IK: Sozialhilfe (SH), Invalidenversicherung (IV), Arbeitslosenversicherung (ALV), Individuelle Konten der AHV (IK)

be reproduced in a new cohort. In the present project, this difficulty was overcome by transferring the initial solution to new cohorts by means of prediction. This approach works well and the criteria used to decide at what point the transfer is no longer valid have proved to be successful in this case. The findings have a concrete added value for standard statistical production, both in terms of the newly developed longitudinal indicators and their visualisation, as well as for the formation of descriptive, quantitative trajectory profiles².

Findings, opportunities and limitations in the application of data-driven methods in official statistics are discussed in depth in the report. Based on key learnings, recommendations for similar projects in the FSO are presented. The conclusions lead to a generic, inductive analysis approach for individual trajectory data in statistics production at the FSO.

² See publication "Verläufe im System der sozialen Sicherheit 2021" (only available in German and French):
<https://www.bfs.admin.ch/bfs/de/home/statistiken/soziale-sicherheit/analysen-verlaeufe-system/analysen.assetdetail.25385873.html>