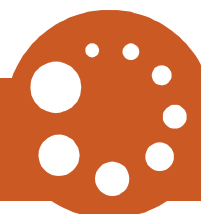




Plausi++: Automatic data validation to check the quality and reliability of administrative data and survey data

Management Summary

EXPERIMENTAL STATISTICS



Neuchâtel, 2024

Published by: Federal Statistical Office (FSO)
Information: persfinhs@bfs.admin.ch
Editors: Mehmet Aksözen, FSO
Topic: 15 Education and science
Original text: German
Translation: Section BILD-P

Layout concept: Section PUB
Downloads: www.statistics.ch
Copyright: FSO, Neuchâtel 2024
Reproduction with mention of source authorised
(except for commercial purposes)

1 Background

This document builds on the report "Plausi++: Automatic data validation to check the quality and reliability of administrative data and survey data" [1] and is intended to serve as a supplement and conclusion to the pilot project.

The idea for the pilot project was to use a machine learning algorithm to add an automated element to the previously manual plausibility check and to find potential previously unidentified errors (problematic cases).

The data used is based on data from the staff statistics in universities and institutes of technology (UIT), supplemented by data from the student and diploma statistics as well as mathematical key figures (see [1], p.6).

The model returns a predicted category of staff for each data set. If this predicted category of staff did not correspond to the provided one (problematic case), a search was made for candidate variables whose values could have contributed to this difference.

2 Objectives

In the previous report [1], p. 6, several points are listed for the continuation of the project. These correspond to the objectives of the work presented here as follows:

1. Testing the feedback mechanism: The feedback mechanism created in a first draft gives an indication of problematic cases (candidates of variables that may be incorrect), which are then checked by the data providers.
2. Inclusion of data suppliers: The data providers receive feedback on possible candidates and check them. The data providers explain why the data is correct or provide corrected data. In addition, organisational differences between the data suppliers can be reflected in the data through systematic deviations, which in turn can result in groups of problematic cases.
3. Taking into account changes at the UIT: In addition to the problematic cases, revisions or other changes at the UIT can lead to changes in the data structures (e. g. selection of a different category of staff than previously due to internal adjustments), so that the model can find more problematic cases that can be traced back to these changes.
4. Monitoring of the model: If differences in the data between two years are measured during the survey, the algorithm is retrained with the available data. Otherwise, the algorithm is retrained after the survey is completed and compared with the old model.

3 Achievement of objectives

Objective 1: Four groups of possible problematic cases were sent to the data providers, which account for around 28% of all cases. In order to minimise the workload for the data providers, the feedback mechanism was also checked prior to dispatch. Problematic cases were reported where the probability of finding an error appeared to be particularly high. In addition, further information was added to these cases.

Objective 2: The data suppliers were able to report all selected problematic cases (potential errors) as correct, even those that occur infrequently. Further feedback from the data suppliers on another group of possible problematic cases (32% of all cases relating to the UIT's administration) revealed that there may be several structural reasons for deviations.

- Data is collected in different ways (calculation or complete survey).
- The administration of the institutions is organised differently, which has a different impact on the staff categories.
- The range of services offered by the data providers has a different focus (more teaching) compared to the majority (more research and development).

Objective 3: With the calculation of a Population Stability Index, a module was added to check the distribution between the years and to assess the usability of the previous algorithm. The distribution of data per personnel category, per UIT and for other data provided did not differ noticeably between the years.

Objective 4: After each survey, the algorithm was retrained with the latest data. The very high accuracy of these annually retrained models remained stable. The accuracy of the models in predicting the personnel category also hardly changed when models were applied to data from other years.

The data quality of the personnel statistics is very high. The differences at the UIT are of a technical and structural nature and were discussed with the UIT at the end of the project. If required by the UIT, the algorithm can be used again.

4 Project organisation

The continuation of the project coincided with a phase in which both the data providers and the project management had few resources for a test run (especially during the coronavirus pandemic). In order to minimise the effort for the institutions, the project management invested a little more time and resources and compiled a selection of possible problematic cases with a high probability of correction. Due to this situation, the project management minimised the involvement of other employees from the Educational Processes section or the Education System section. The necessary data preparation was carried out by the IT group of the Educational Processes section.

5 References

- [1] Ruiz, C. et al., 2019: «Plausi++: Automatische Plausibilitätsprüfung der Qualität und Zuverlässigkeit von Administrativ- und Umfragedaten»,
website: <https://www.bfs.admin.ch/bfs/de/home.assetdetail.9847917.html> (last access 25.03.2024)
- [2] Yurdakul, B., 2018: «Statistical Properties of Population Stability Index», Dissertation,
website: <https://scholarworks.wmich.edu/cgi/viewcontent.cgi?article=4249&context=dissertations> (last access 25.03.2024)