Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Bundesamt für Statistik BFS
Office fédéral de la statistique OFS
Ufficio federale di statistica UST
Federal Statistical Office FSO

Integrating a System for Automatic Classification of Economic Activities into Statistical Production: Challenges and Solutions

Athanassia Chalimourda, Lorenz Helbling

Data Science, AI and statistical methods
Interoperability and Registers

use of R in official statistics, 12-14 December 2023

Introduction

The classification of the economic activities according to the *Nomenclature générale des activités économiques* (**NOGA**), the Swiss **NACE**, is performed manually by coding experts.

NOGAuto is an assistance system for automatic classification and interaction with the coding expert written in **R**.

- Automatic Classification of Economic Activities
- Performance Measures and Decision Making
- Interaction with Experts and Expert-Systems
- R – Packages
- Conclusions

NOGAuto, Shiny App, Methods (Business Registers Data):
Lorenz Helbling, Mathias Constantin,
Cindia Duc Sfez, Daniele Marx
Methodological Support (Statistical Methods):
Athanassia Chalimourda, Daniel Assoulin

Automatic Classification of Economic Activities

Involves Natural Language Processing and automatic classification of economic activities descriptions in French, German and Italian, currently performed manually

- Evaluation of the automatic classification
- How can an innovative system under continuous development be integrated into statistical production that needs reliability and stability?

Automatic Classification

The *Nomenclature générale des activités économiques* (**NOGA**) has hierarchical categories – Example:

*The operation of a drugstore and the marketing of all drugstore, herbal, dietetic and cosmetic products, medicines and health products (NOGA – Code: **477501**)*

- Sector (3 classes): **3** – *Services*
- Section (21): **G** – *Trade; Maintenance and repair in motor vehicles*
- Two digits (Division, 88): **47** – *Retail trade (excluding trade in motor vehicles)*
- Four digits (615): **4775** – *Retail trade in cosmetic and body care products*
- Six Digits (794): **477501** – *Drugstore*

The NOGAuto classification

- An activity description is turned into a vector (text2vec)
- Supervised machine leaning with a gradient boosting machine (GBM) which associates a NOGA-Code to an activity description
- The predicted code is assigned to a description with a prediction probability approximated by the number of GBM-trees that voted for that code

Performance Measures and Decision Making

How can performance measures be used in order to:

- assess the overall quality of NOGAuto
- guide the automation process

Measures for overall performance as well as measures for performance by class are employed.

Global Performance Measures

We use global performance measures based on the confusion matrix which compares the actual with the predicted classes.

Example: The confusion matrix for the 21 Classes of NOGA-Section

Predicted

Actual

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	sum
A	364	0	29	0	0	6	21	0	0	0	0	0	0	4	0	1	2	1	2	0	0	430
B	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19
C	2	0	1010	0	0	7	64	2	1	4	2	2	4	6	0	2	2	4	7	0	0	1119
D	1	0	1	56	0	1	0	1	0	2	0	0	0	0	0	0	0	0	0	0	0	62
E	0	0	2	0	83	1	3	1	0	0	0	0	0	2	0	0	0	0	0	0	0	92
F	1	0	8	0	0	290	6	1	0	1	0	2	0	5	0	0	0	0	0	0	0	314
G	2	0	57	1	0	1	1377	2	0	9	0	2	3	6	0	1	1	0	6	0	1	1469
H	0	1	6	0	0	1	5	238	1	2	0	0	0	4	0	0	0	0	2	0	0	260
I	0	0	1	0	0	0	4	0	132	0	0	2	1	4	0	0	3	2	0	0	0	149
J	0	0	2	0	0	0	5	0	0	289	1	0	2	4	0	0	0	3	1	0	0	307
K	0	0	1	0	0	0	2	1	0	0	212	2	2	6	0	0	1	0	0	0	0	227
L	0	0	1	0	0	1	1	0	1	0	0	82	0	3	0	0	0	1	0	0	0	90
M	0	0	10	0	0	0	6	1	1	5	3	0	321	0	0	3	4	2	2	0	0	358
N	0	0	5	0	0	9	13	6	0	2	1	1	7	442	0	2	0	2	0	0	0	490
O	0	0	0	0	0	0	0	1	0	0	0	0	0	0	21	1	4	0	1	0	0	28
P	0	0	1	0	0	0	0	1	0	1	0	0	1	3	7	178	7	3	0	0	0	202
Q	0	0	4	0	0	0	1	0	1	0	0	0	1	2	1	3	240	1	3	0	0	257
R	0	0	2	0	0	0	5	0	3	6	0	0	3	2	4	5	1	179	2	0	0	212
S	0	0	18	0	0	2	14	2	0	0	1	0	2	6	0	1	3	1	246	0	0	296
T	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	3	0	4
U	1	0	0	0	0	1	4	0	1	1	0	0	2	0	0	1	1	1	0	0	10	23
sum	371	20	1158	57	83	320	1531	257	142	322	220	93	349	499	33	198	269	200	272	3	11	6408

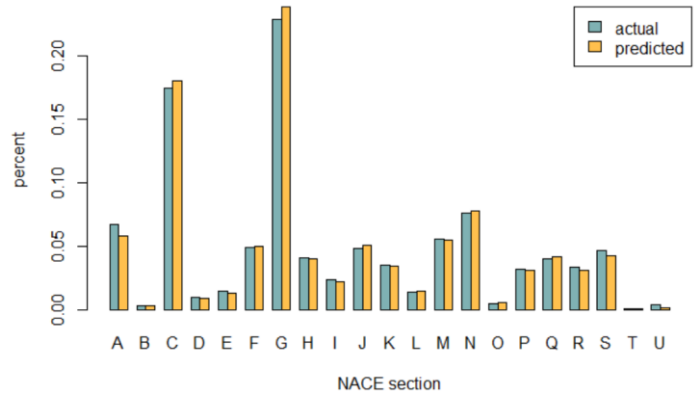
- True positive (TP) the activity is correctly predicted to class G
- False positive (FP) the activity is falsely predicted to class G
- False negative (FN) the activity is falsely *not* predicted to class G

Global Performance Measures

- Accuracy: Overall percentage of elements for which the predicted and the actual class are the same
- Balanced Accuracy: Mean value of the agreement percentages per class (with respect to the actual classes)
- Cohen‘s Kappa: The accuracy is corrected for the class agreement expected by chance
- Comparison of the distributions of the classes‘ percentages

	NOGA-Section	NOGA-Division
Accuracy	0.90	0.88
Balanced Accuracy	0.87	0.86
Cohen’s Kappa	0.89	0.87

Distributions of the percentages of the classes in a small sample (n_test = 6408) of activity descriptions in French



Performance Measures per Class

Positive predicted value (ppv, precision):

#TP / (#TP + #FP). High precision implies low number of descriptions that are falsely predicted to a certain class.

True positive rate (tpr, recall):

#TP / (#TP + #FN). High recall implies low number of false negatives. It means that a class is well captured.

NOGA-section	# Activities	ppv	tpr
A: Agriculture	430	0.98	0.85
C: Manufacturing	1119	0.87	0.90
G: Trade	1469	0.90	0.94

The true positive rate in the subset of activities with prediction probabilities $\geq 80\%$

NOGA-section	# Activities	tpr
A: Agriculture	375	0.94
C: Manufacturing	934	0.98
G: Trade	1323	0.98

The true positive rate in the subset of activities with prediction probabilities $< 80\%$

NOGA-section	# Activities	tpr
A: Agriculture	55	0.18
C: Manufacturing	185	0.51
G: Trade	146	0.58

NOGAuto interacts with Experts and Expert-Systems

Maximize automation while controlling quality (error propagation, predictive and distributional accuracy)

Assist coding experts, leaving challenging classifications to them

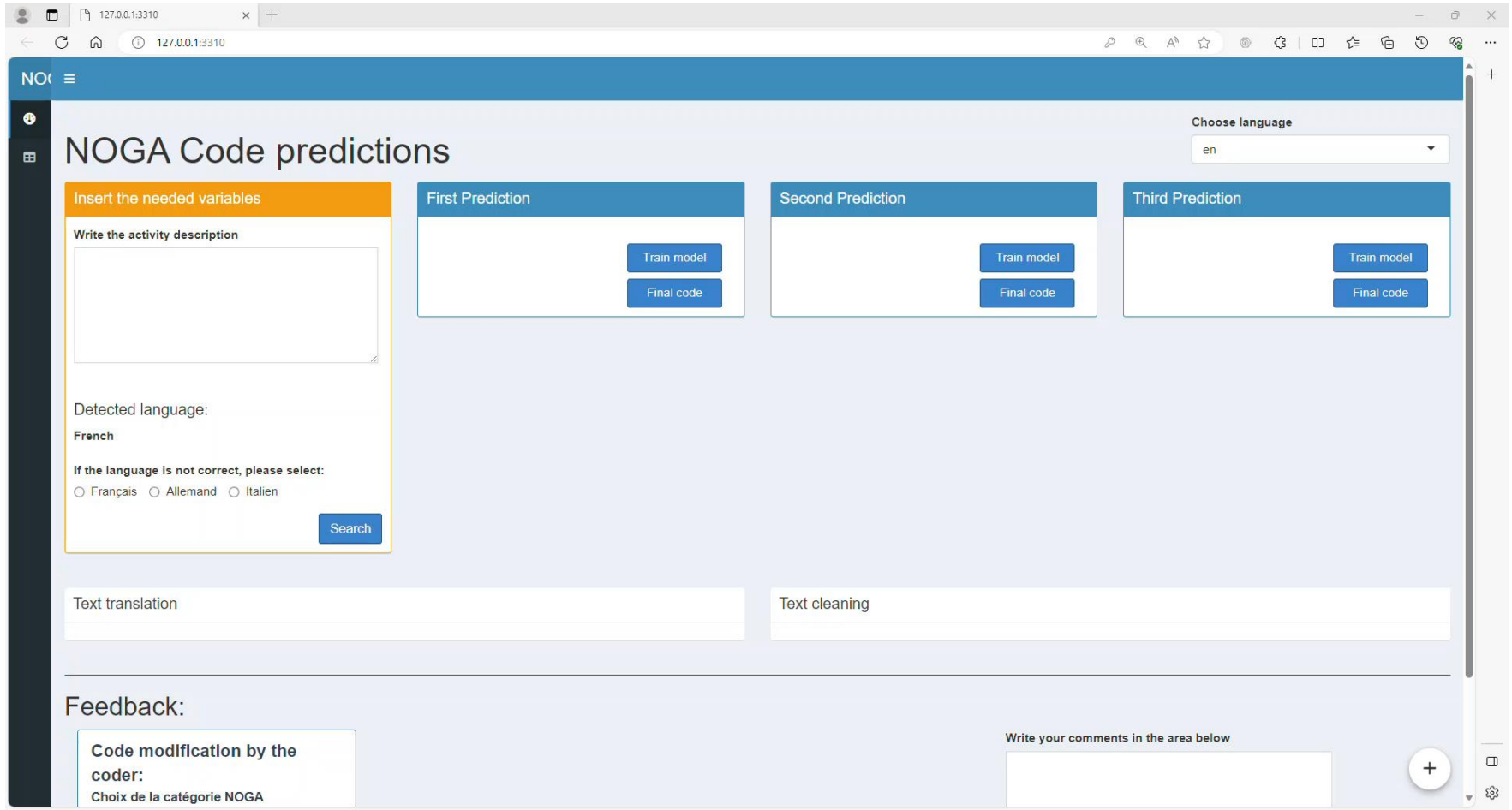
Select units for manual control of the NACE - Code

Example of Interaction with Experts

- Proceed automatically for descriptions with high prediction probability, which belong to classes with high true positive rate.
- When the requirements on true positive rate and prediction probability are not fulfilled, NOGAuto stops and makes code-suggestions based on the prediction probabilities.
- The expert can accept a NOGAuto suggestion or decide for a completely different code. The expert's choice is recorded for evaluation purposes.
- Challenging activity descriptions are left to the expert´s judgement.

Interaction with Expert-Systems

Example (German): ***«Die Gesellschaft erbringt sämtliche Dienstleistungen im Bereich Grafik und Illustration. Ausserdem unterstützt sie Unternehmen, Institutionen und Einzelpersonen in Kommunikationsfragen.»***



R – Packages

Some of the R-Packages used:

- Dplyr
- Tokenizers
- Tidyverse
- Text2vec
- Xgboost
- Caret
- Shiny
- Flexdashboard

Conclusions

- Integrating an innovative assistance system in statistical production should allow for progress while simultaneously assuring quality and stability.
- Global performance measures assure overall quality.
- Performance measures should account for class imbalance, since for economic activities all classes are important independently of their size.
- Combination of performance measures and prediction probabilities can be used to delimitate where NOGAuto performs best, leaving the remaining units to the expert.
- NOGAuto can be combined with other expert systems.

Future work includes

- Selection of units for manual NACE – Code control
- Improvement and adjustment of thresholds for the expert's intervention
- Maintenance of the data set over time for training, testing and adjustment

Thank you for your attention!