# Difference on Evaluation Scores Considering Image Descriptions

uRos2023

12-14 Dec. 2023, Bucharest

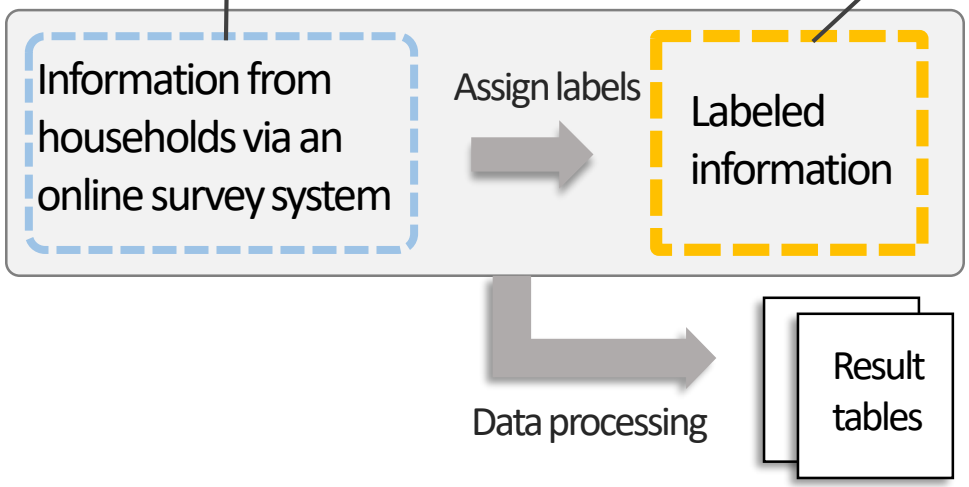
Yukako Toko (National Statistics Center)

Mika Sato-Ilic (University of Tsukuba)

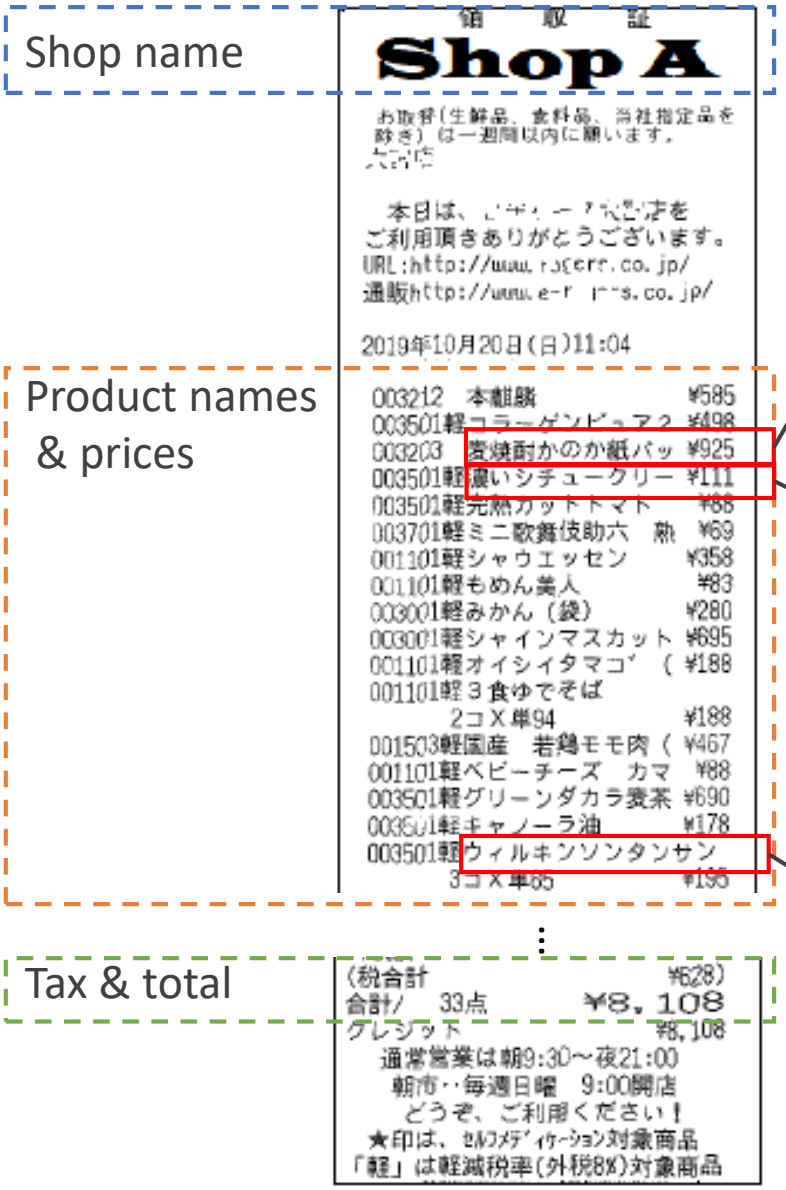
# Coding task for the Family Income and Expenditure Survey

Assigning corresponding category labels (or classes):
An essential activity for data processing in official statistics

| Purchased items and their uses | Quantities | Unit | Cash disbursements (Yen) | Label | Category |
|---|---|---|---|---|---|
| Sandals (for ladies) | 1 | pair | 3,564 | 672 | Shoes (for ladies) |
| Toilet rolls | | | 646 | 532 | Toilet rolls |
| Chicken | 279 | g | 538 | 222 | Chicken meat |
| Chocolate assortment box | | | 149 | 352 | Chocolate |
| Hot chilli sauce | 300 | ml | 214 | 328 | Sauce |



# Image of Japanese shopping receipt




Ex: **Contains multiple types of characters**

芋焼酎かのか紙パッ

: Kanji : Hiragana : Katakana

Ex: **Incomplete product name**

濃い シチュー クリー

rich taste stew crea...

Complete product name in English:
Rich taste cream stew cubes

Ex: **No space between words**

ウィルキンソンタンサン

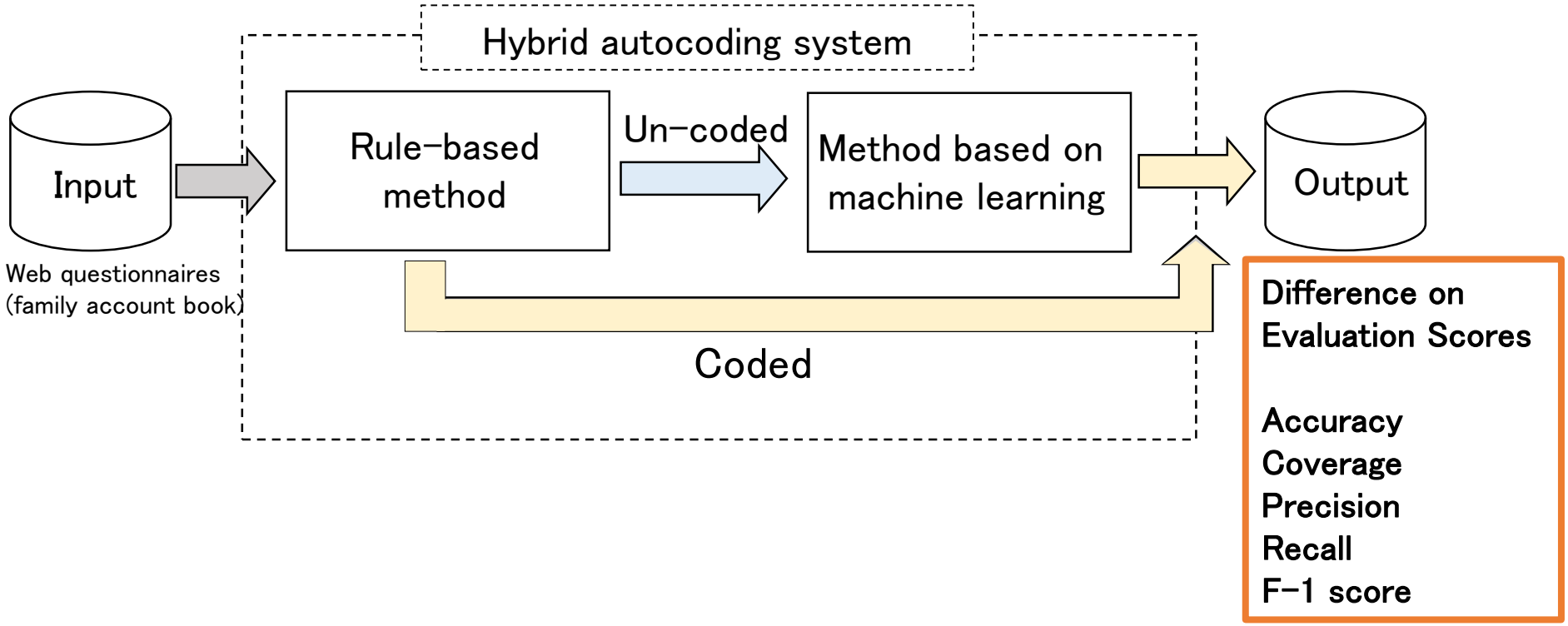
wilkinsonsparklingwater

Product name with spaces in English:
Wilkinson sparkling water

# Hybrid autocoding system (HAS)

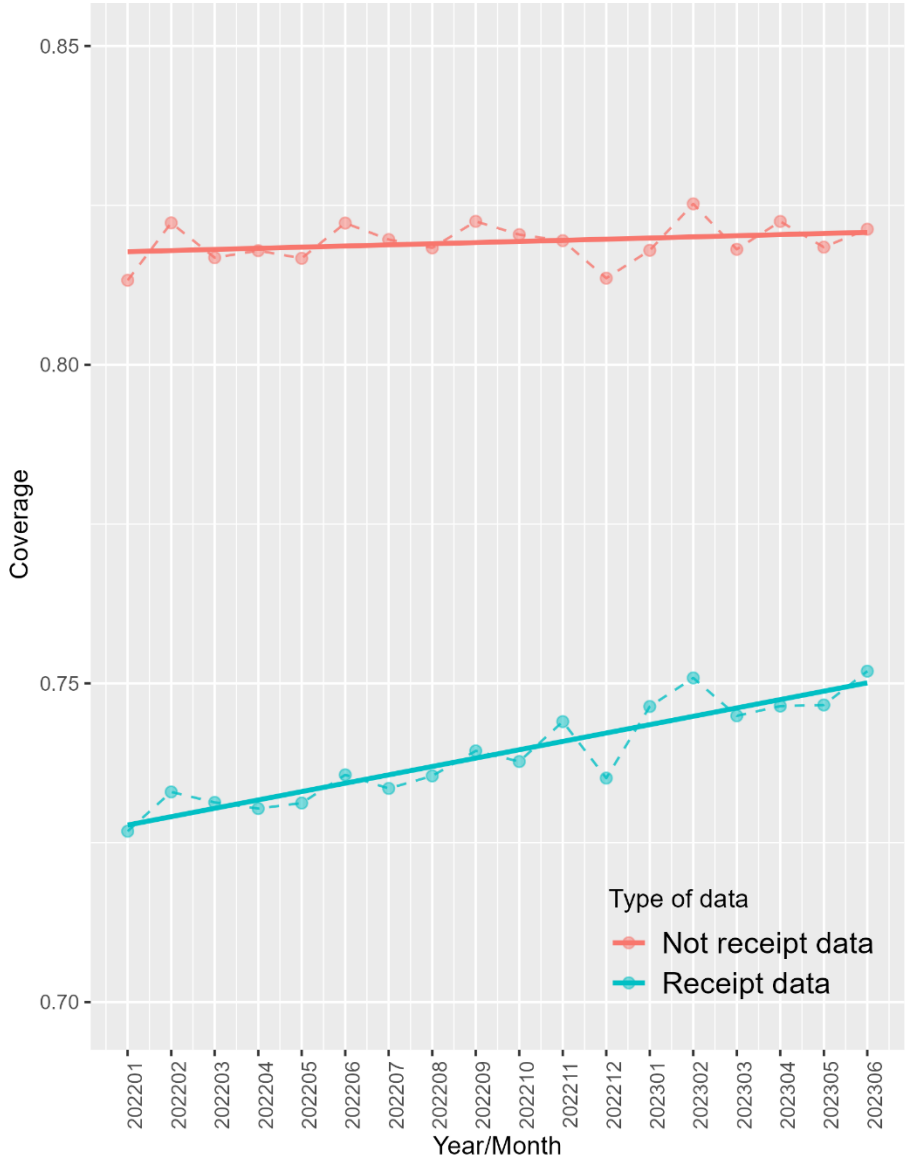
✓ Combined method of rule based classification method & classifier based on machine learning technique

✓ Implemented in data processing for the Family Income and Expenditure Survey since Jan. 2022



# Result of coverage of HAS


Number of target data



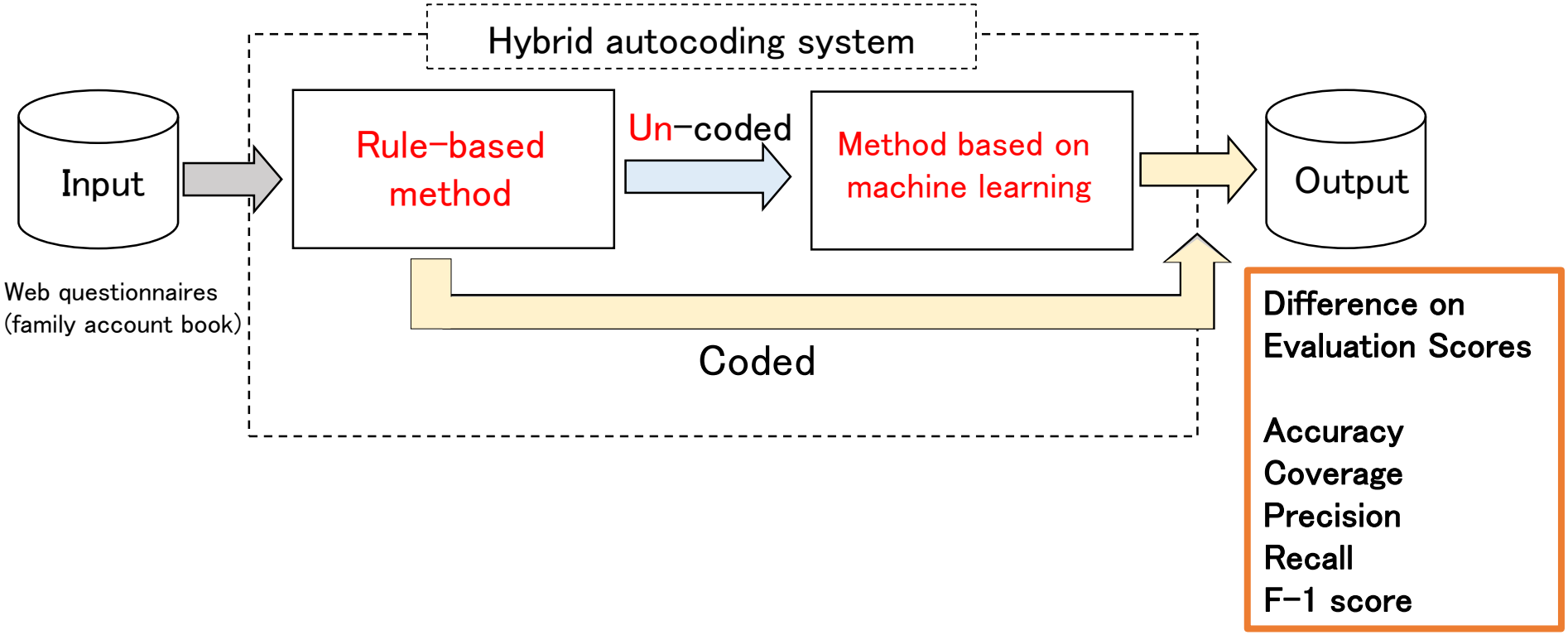

Result of coverage

->Difficulty in coding receipts data
Various kinds of product name

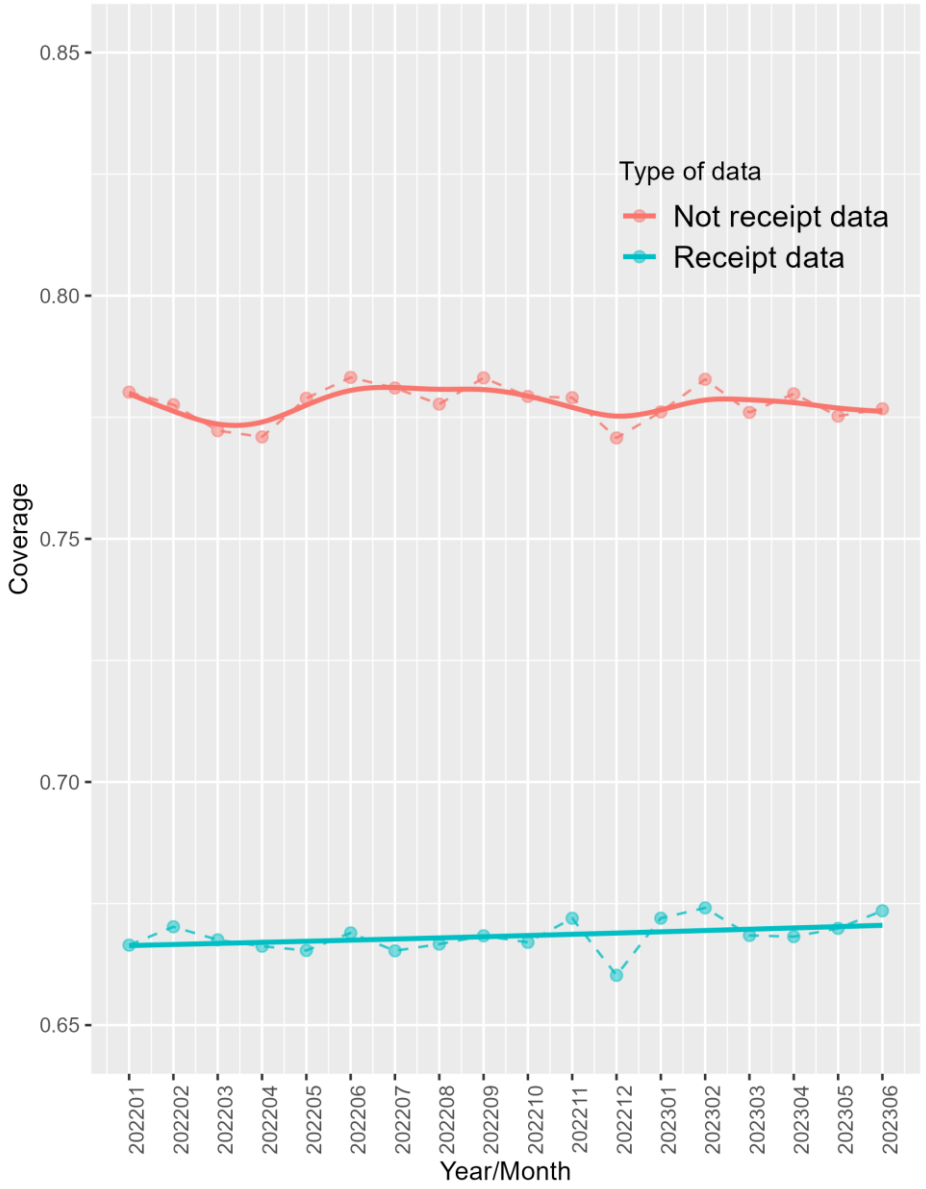
# Hybrid autocoding system (HAS)

✓ Combined method of rule based classification method & classifier based on machine learning technique

✓ Implemented in data processing for the Family Income and Expenditure Survey since Jan. 2022



# Results of coverage of HAS




Result of coverage of rule-based method


Result of coverage of method based on machine learning

# Numerical Examples

## Family Income and Expenditure Survey

- ✓ Sampling survey monthly conducted by Statistics Bureau of Japan
- ✓ Dataset contains purchased items name or receipt items name in short text descriptions including descriptions obtained from shopping receipt images, and corresponding labels
- ✓ Approx. **520** different category labels are available

Training data

Data from Jan. 2018 to the previous month's data of the evaluation data
ex) We use Jan. 2018 to Aug. 2022 data if the evaluation data is Sep. 2022 data

Number of training data : approx. 30 million

Evaluation data

Data from Sep. 2022 to Jun. 2023

Number of evaluation data : approx. 990,000 per month

# Method based on machine learning

$\bar{p}_{jk}$: Reliability score of $j$-th object to a class $k$

$$\bar{p}_{jk} = T\left(\tilde{\tilde{p}}_{jk}, \sum_{m=1}^{\tilde{K}_j} \tilde{\tilde{p}}_{jm}^2\right), \qquad j = 1, \dots, J, \qquad k = 1, \dots, \tilde{K}_j.$$

$$\bar{p}_{jk} = T\left(\tilde{\tilde{p}}_{jk}, 1 + \sum_{m=1}^{K_j} \tilde{\tilde{p}}_{jm} \log_K \tilde{\tilde{p}}_{jm}\right), \qquad j = 1, \dots, J, \qquad k = 1, \dots, \tilde{K}_j.$$

***Explanation of the uncertainty of the training data.***
***Utilization on the deference of measurement of uncertainty.***

**Probability measure**

Relative frequency of object $\boldsymbol{j}$ to class $\boldsymbol{k}$

**Fuzzy measure**

Transformation from $\tilde{\tilde{p}}_{jk}$ to classification status of object $\boldsymbol{j}$

Classification status of object $\boldsymbol{j}$ over the $\tilde{K}_j$ classes

$$\bar{\bar{p}}_{jk} = g(n_j)\bar{p}_{jk}$$

The classifier arranges $\{p_{j1}, \dots, p_{jK}\}$ in descending order and creates $\{\tilde{p}_{j1}, \dots, \tilde{p}_{jK}\}$, such as $\tilde{p}_{j1} \geq \cdots \geq \tilde{p}_{jK}, j = 1, \dots, J$.
After that, $\{\tilde{\tilde{p}}_{j1}, \dots, \tilde{\tilde{p}}_{j\tilde{K}_j}\}, \tilde{K}_j \leq K$ are created.

$T$ : *T*-norms (Menger, K., 1942)

$p_{jk}$: Relative frequency of object $\boldsymbol{j}$ to class $\boldsymbol{k}$

$$p_{jk} = \frac{n_{jk}}{n_j}, \qquad n_j = \sum_{k=1}^{K} n_{jk}, \qquad j = 1, \dots, J, \qquad k = 1, \dots, K$$

$n_{jk}$ : Number of text descriptions in a class $\boldsymbol{k}$ with $\boldsymbol{j}$-th object in the training dataset

$g(n_j)$ : Weight for control size of object $\boldsymbol{j}$ $\qquad g(n_j) = n_j / \sqrt{1 + {n_j}^2}, \qquad g(n_j) = \tanh n_j$

# T-norm (Menger, 1942, Schweizer and Skla, 2005)

- Boundary conditions

$$0 \le T(a,b) \le 1, \qquad T(a,0) = T(0,b) = 0, \qquad T(a,1) = T(1,a) = a$$

- Monotonicity

$$a \le c\ , b \le d\ \rightarrow T(a,b) \le T(c,d)$$

- Symmetry

$$T(a,b) = T(b,a)$$

- Associativity

$$T(T(a,b),c) = T\big(a,T(b,c)\big)$$

where $\forall a,b,c,d \in [0,1]$

- Algebraic product

$$T(a,b) = ab$$

- Hamacher product

$$T(a,b) = \frac{ab}{p + (1-p)(a+b-ab)}, \qquad p \ge 0$$

- Minimum

$$T(a,b) = min\{a,b\}$$

- Einstein product

$$T(a,b) = \frac{ab}{1 + (1-a)(1-b)}$$

# Evaluation measures for classification

$$accuracy = \frac{TP + TN}{N}$$

$$macro\ precision = \frac{1}{K}\sum_{l=1}^{K}\frac{TP_l}{TP_l + FP_l}$$

$$macro\ recall = \frac{1}{K}\sum_{l=1}^{K}\frac{TP_l}{TP_l + FN_l}$$

$$macro\ f1score = \frac{1}{K}\sum_{l=1}^{K}\left(2 * \frac{precision_l * recall_l}{precision_l + recall_l}\right)$$

$K$: number of classes $N$: number of text descriptions

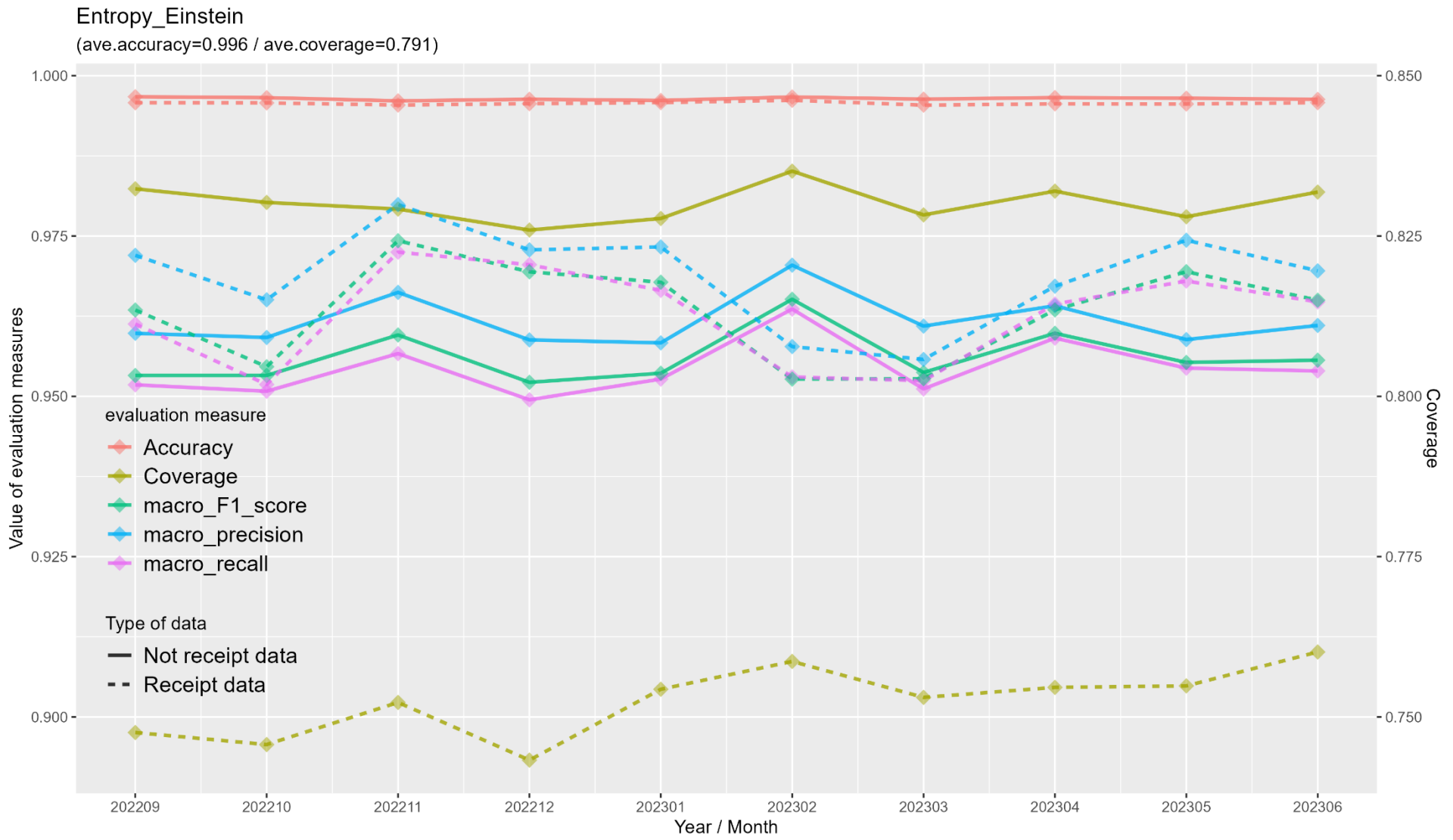
$TP$: number of true positive text descriptions

$TN$: number of true negative text descriptions

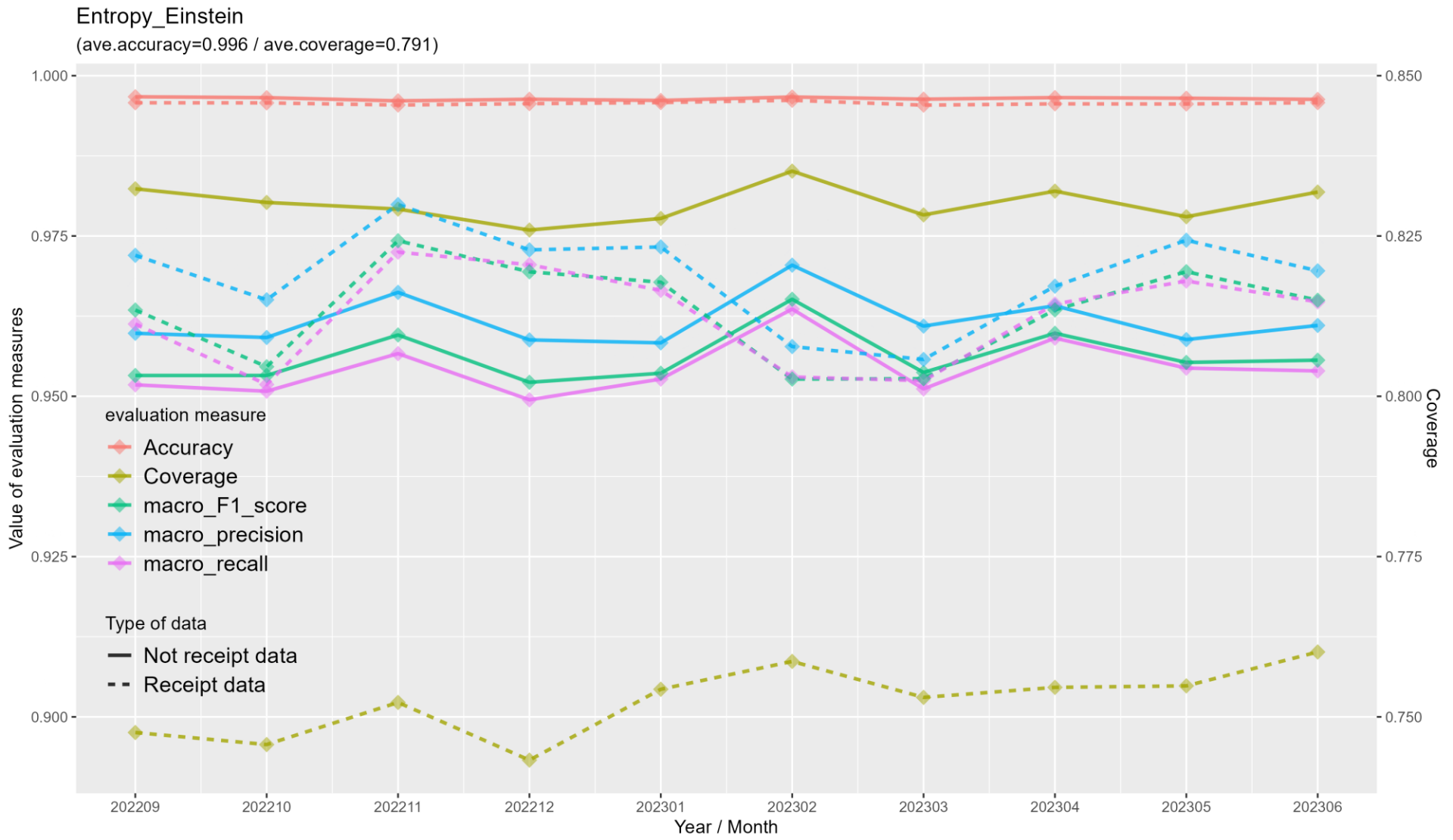
$FP$: number of false positive text descriptions

$FN$: number of false negative text descriptions

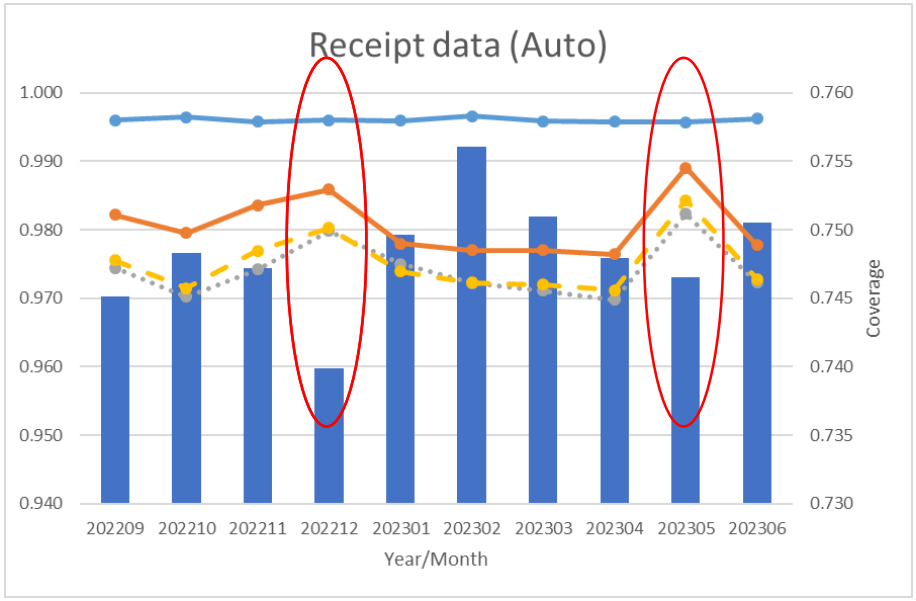
# Numerical Examples

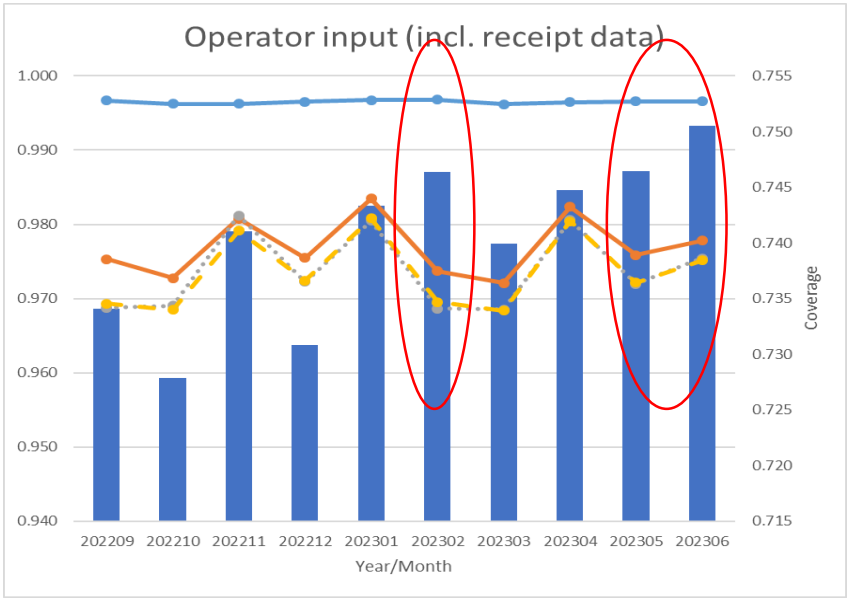


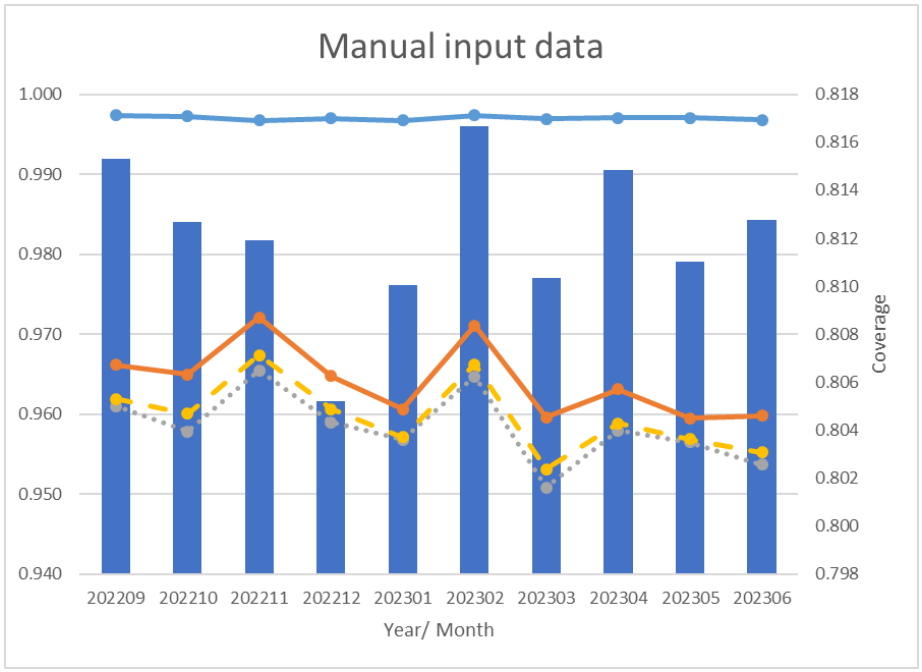
# Numerical Examples

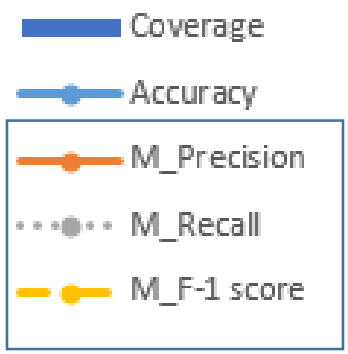


# Results of evaluation measures on kinds of input data




Special feature of inclusion of false part

# Conclusions

1. Evaluation of different measures for HAS under several kinds of data

- Previously used evaluation measures are only accuracy and coverage

- Various features based on different evaluation measures (false part based evaluation measures(Precision, Recall, f-1 score)) are captured
  - Comparison of receipt and manually inputted data
    -> Machine learning based method is covered for increase of receipt data
  - Automatically recognized receipt data
    -> Although scores of coverage is lower, scores of false part based evaluation measures are higher
  - Operator based receipt data -> although scores of coverage is higher, scores of false part based evaluation measures are lower
  - Manually inputted data -> simultaneous change between coverage and scores of false part based evaluation measures

2. Increase of amounts of data depends on increase of receipt data

Increase of autocoding of receipt data is treated by the machine learning method

Coverage of receipt data by using the machine learning method is almost continuously higher than one of not receipt data

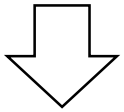
For all of evaluation measures, evaluation scores of receipt data are higher than evaluation sores of manually inputted data.

The above fact shows stability of machine learning method for the receipt data

# Conclusions

## Further study

Investigation of details of evaluation scores highlights difference of features among methods of autocoding



Development of new method or adoptable local application of methods for autocoding system

# References


1. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York (1981)
2. Bezdek, J.C., Keller J., Krisnapuram, R., Pal, N.R.: Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. Kluwer Academic Publishers (1999)
3. Menger, K.: Statistical metrics, Proceedings of the National Academy of Sciences of the United States of America, 28, pp. 535-537 (1942)
4. Schweizer, S., Sklar, A.: Probabilistic metric spaces. Dover Publications, New York (2005)
5. Toko, Y., Iijima, S., Sato-Ilic, M.: Overlapping Classification for Autocoding System, Journal of Romanian Statistical Review, 4, pp. 58-73 (2018)
6. Toko, Y., Iijima, S., Sato-Ilic, M.: Generalization for Improvement of the Reliability Score for Autocoding, Journal of Romanian Statistical Review, 3, pp. 47-59 (2019)
7. Toko, Y., Sato-Ilic, M.: Improvement of the training dataset for supervised multiclass classification, Czarnowski, I., Howlett, R.J., Jain, L. C. (Eds.), Smart Innovation, Systems and Technologies, 193, pp. 291-302. Springer, Singapore (2020)
8. Toko, Y., Sato-Ilic, M.: Efficient Autocoding Method in High Dimensional Space, Romanian Statistical Review, Vol. 1, pp. 3-16 (2021)
9. Toko, Y., Sato-Ilic, M.: Autocoding based Multi-Class Support Vector Machine by Fuzzy c-Means, Romanian Statistical Review, Vol. 1, pp.27-39 (2022)
10. Toko, Y., Sato-Ilic, M., Sasajima, T.: Improvement of Model Construction based on Reliability Scores of Objects for Autocoding, Romanian Statistical Review, Vol. 1, pp.34-48 (2023)
11. Statistics Bureau of Japan: “Outline of the Family Income and Expenditure Survey”. Available at: https://www.stat.go.jp/english/data/kakei/1560.html, last accessed 2023/11/21