

帳票における修正内容の推定方法について

前田一穂¹ 遠藤進¹ 大塚浩¹

概要： RPAにより定型的な業務の効率化が進んでいるが、RPAでは自動化が難しいものとして人の判断を要する業務が残っている。本稿では、人の判断を要する業務の一つである突合作業を取り上げる。まず、人がどのように突合作業で判断をしているかを、自治体での住民税賦課業務を題材に考察する。さらに、人が突合作業を効率的に行うため用いているノウハウの情報を学習データに埋め込むことで、過去の修正履歴をもとに、未修正の帳票の修正内容を機械学習により推定する方法を提案する。修正者の作業前に、推定した修正内容をシステムが提示することで、修正者にとって修正内容の見当がつけやすくなり、突合作業の効率化が期待される。

キーワード： 帳票、修正、突合作業、RPA、機械学習、税賦課業務

A Method of Estimating How to Correct Forms

KAZUHO MAEDA^{†1} SUSUMU ENDO^{†1}
HIROSHI OTSUKA^{†1}

Abstract: Though RPA is improving the efficiency of routine work, the work that requires human judgment is still difficult to automate with RPA. In this paper, we will focus on the matching work, which is one of the work that requires human judgment. First, we will consider how people make decisions in the matching work, using the assessment of resident taxes in local governments as the subject. Furthermore, by embedding the information of the know-how that people use to efficiently perform the matching work in the learning data, we propose a method of estimating how to correct the uncorrected forms by machine learning from the past correction history. By presenting the estimated correction contents before the work of the corrector, it becomes easier for the corrector to guess the correction contents, and it is expected that the efficiency of the matching work will be improved.

Keywords: forms, correction, matching work, RPA, machine learning, assessment of taxes

1. 背景

労働環境改善意識の高まりや、少子高齢化による労働力不足により、労働力の確保が喫緊の社会課題となりつつある。そのための有効な解決手段の一つが AI や ICT であり [1]、ソフトウェアロボットにより業務プロセスを自動化する RPA(Robotic Process Automation)の導入が進んでいる。その代表的な技術が、紙の書類を自動的に機械で読み取る OCR であり [2]、近年では不定形な書類も高精度な電子化が可能になっている [3]。

高精度な電子化が可能になった書類の一つが帳票である [4]。帳票は、企業等が経営活動のために作成した書類の総称であり、仕訳帳、出納帳、請求書、領収書等が含まれる。

しかし、RPAでも自動化が難しい業務として非定型業務、特に人の判断を要する業務が残っている [1,5]。特に、判断のルールが明文化されておらず、あるいは明文化に多大な労力を要するため、過去の経験（暗黙的な知識）に基づいて実施されている業務については自動化が困難である。

本稿では、自動化が困難な非定型業務として帳票の突合作業を取り上げる。突合作業は、複数の帳票に記載された

データ項目を見比べることで、それぞれの帳票の正しさを確認したり、データの誤りを修正したりする作業である。突合作業を伴う業務を以下に例示する（図 1）。

- 自治体における税賦課業務
- 企業における会計監査業務 [6]
- 医療現場におけるレセプトの突合業務 [7]



図 1 帳票の突合作業を伴う業務の例

Figure 1 Examples of form matching work

本稿では、突合作業の際に人がどのように判断をしているかを、自治体での住民税の賦課業務を題材に考察する。さらに、考察から得られた知見を用い、人が突合作業を効率的に行うため用いているノウハウの情報を学習データに埋め込むことで、過去の修正履歴をもとに未修正の帳票の修正内容を推定する方法を提案する。

¹ 富士通株式会社
Fujitsu Limited

修正者の作業前に、推定した修正内容をシステムが提示することで、修正者にとって修正内容の見当がつけやすくなり、突合作業の効率化が期待される。

2. 自治体での住民税賦課業務

突合作業の例として、2章では自治体での住民税賦課業務の詳細を述べる。

住民税賦課業務において、自治体の職員は、住民や事業所から提出された様々な帳票種別(表1)を突き合わせて住民税額に誤りがないか確認している。

表1 住民税賦課業務の帳票種別

Table 1 form types for assessment of resident taxes

帳票種別	作成者	主な情報
確定申告書	納税者	収入額、控除額、扶養人数
住民税申告書	納税者	収入額、控除額、扶養人数
給与支払報告書(給報)	勤め先	給与収入額、控除額、扶養人数
公的年金等支払報告書(年報)	年金支払い元	年金額、扶養人数

図3 給与支払報告書の様式[9]

Figure 3 A format of salary payment reports

図2 確定申告書の様式[b] (一部) [8]

Figure 2 A format of tax returns (a part)

図4 公的年金等支払報告書の様式[9]

Figure 4 A format of pension payment reports

表1に示した確定申告書、給報、年報の様式を図2、図3、図4に示す。赤枠で示したように、扶養人数に関する情報は、これらの帳票のいずれにも記載されている。このような、異なる帳票種別に記載されている共通の情報が、互いに食い違っているのが、帳票の不整合の一例である。

不整合が発生した場合、職員は、帳票を見直し、紙ベースのオリジナル帳票や関連する帳票を確認しながら、必要なら帳票を修正して住民税額を変更する(図5)。

ある自治体では、不整合の発生件数は住民数の概ね1割程度あり、住民税徴収開始に間に合うよう、これだけの帳票の確認作業を2か月程度で終えなければならない。このため、短期間に膨大な作業量が必要となり、職員への負荷が大きく、課題となっている。また、個々の職員が経験を積みベテランになるに従い作業の効率性と正確性が向上し

ていくが、せっかく蓄積したノウハウの継承が難しく、職場ローテーションによりベテラン職員が異動するたび、職場全体の効率性が低下することも課題である。



図 5 住民税賦課業務のイメージ図

Figure 5 An image of assessment of resident taxes

3. 帳票の突合作業の考察

3章では、帳票の突合作業について考察する。

帳票は、実際に発生している事象の情報を、経営活動で扱いやすいよう格納したものとみなせる。例えば、請求書や領収書には、実際に納入した物品の情報の中から経営活動で扱う情報（品名や金額など）のみを抽出して記載している。不要と思われる色や重さは記載されない。このように、各々の帳票は、現実の事象の情報的一部分を、一定のルールに従って記載したものとみなせる。

ここで、同一の事象を別々に記載した2つの帳票が存在し、一方を誤って記載した場合を考える。誤りの原因として以下が考えられる。

- **単純ミス**：帳票作成者の記入ミス。帳票に記入するときに誤った数字を記入した場合など。
- **記載ルール誤解**：帳票作成者による記載ルールの誤認識。何を記載するかや、どの項目に記入するかを誤解した場合など。
- **記載時期ずれ**：記載の時期のずれ。一方の帳票に古い情報が残っている場合など。

突合作業で、この2つの帳票を比較すれば、現実の事象と帳票を比較しなくとも、いずれかの記載に誤りがあったことがわかる。

すなわち、突合作業は、現実の事象から帳票へ正しく記載されたかを、同じ事象を記載した複数の帳票を突き合わせることで確認する作業とみなせる（図 6）

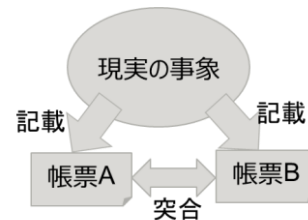


図 6 突合作業のイメージ

Figure 6 An image of form matching work

4. 帳票の修正作業の考察

3章で突合作業について考察し、現実の事象と比較しなくとも、突合作業で帳票の誤りを見つけられることは説明できた。しかし、職員による住民税賦課業務を観察すると、どう修正すべきかまで帳票だけから判断できているケースがしばしばあった。4章では、現実の事象と比較しなくとも、なぜ人が帳票を修正できるのかを、住民税賦課業務を題材に考察する。

ある自治体で住民税賦課業務についてヒアリングしたところ、帳票の不整合が見つかった場合、職員はまず不整合を起こした帳票を見比べて帳票の状況を確認していた。さらに、想定されうる不整合の中から、この状況を説明可能なものを探索し、納得できたときは帳票のみから修正を実施していた。

ここで、職員が想定している不整合から一部を抜粋する。

- **単純ミス（記入ミス）**：別の帳票から数値を書き写すときに書き間違えた
- **単純ミス（帳票忘却）**：副業の給与を申告し忘れた
- **記載ルール誤解（記入欄誤り）**：小規模企業共済等掛金控除を誤って社会保険料控除の内数として記入した（源泉徴収票では小規模企業共済等掛金控除を社会保険料控除の内数として記載されているが、確定申告書では社会保険料控除の外数として記入するため間違えやすい）
- **記載ルール誤解（非記入）**：4人目の扶養者の情報を確定申告書に記入しなかった（扶養者の記入欄が3人分しかないため[c]、4人目を書くのをあきらめたケース。一人分の欄に複数人分記入するのが正しい）
- **記載時期ずれ（古い情報）**：年報の扶養者数が古いまま何年も更新されていないため、現在の扶養者数と異なった（年報の扶養者数は他の帳票種別に比べ更新されにくい）
- **記載時期ずれ（変更）**：年末調整の時点では子を父の扶養としていたが、確定申告で母の扶養に変更したため、給報と確定申告書で扶養者数が異なった

c 令和元年用まで。令和2年用以降は扶養者の欄が増えている

さらに、住民税賦課業務の場合、不整合の原因に加え、不整合が発生する条件や、どの帳票のどの項目をどの値に修正すべきかもノウハウとして各職員が持っており、修正のときに活用して効率的に作業を実施していることも分かった。以下では、職員が持っていたこれらのノウハウを「不整合パターン」と呼ぶ。(表 2)

表 2 不整合パターンの定義

Table 2 Definitions of inconsistent patterns

不整合原因	記載が誤っている原因
不整合条件	不整合が発生しているとみなせる条件
修正箇所	どの帳票と項目が誤っているか
正しい値	どの値に修正すべきか

自治体で形式知として活用されているフローチャートや職員様へのヒアリングをもとに抽出した住民税賦課業務における不整合パターンの例を表 3 に示す[d]。

表 3 から、不整合パターンには以下の傾向が見て取れる。

- **不整合条件**: 異なる帳票種別間で同一項目の値が異なることを条件とするケースが多い。
- **修正箇所と正しい値**: 誤っている帳票の項目を他の帳票種別の同一項目の値に変更するケースが多い。どの帳票種別が誤っていて、どの帳別種別が正しいかはケースバイケース

これらの傾向を持つ不整合パターンの活用では、帳票種別間で値を比較する必要がある。このとき問題となるのが、一つの帳票種別に複数の帳票が存在する場合である。住民税賦課業務でも、転職や副業で増えた勤め先の数だけ給報が存在するが、確定申告書と比較する前に、給報で一つの値にまとめておく必要がある。このように、帳票種別間での比較のため一つにまとめた値を、本稿では代表値とよぶ。

この代表値の計算方法は、帳票種別や項目により様々である。例えば住民税賦課業務の場合、同じ給報でも、扶養者数は甲給報[e]の値を採用すべきだが、給与収入は全ての給報の値の総和を採用すべきであるように、項目が異なれば適切な計算方法が異なっていた。職員は、項目ごとの代表値の計算方法をノウハウとして持っており、複数の帳票がある場合でも、効率的に不整合パターンを活用できていた(図 7)。

このように、住民税賦課業務において、職員は代表値計算方法や不整合パターンといったノウハウを活用して突合作業を進めていた。特にベテラン職員は、経験を通じ、より多くのノウハウを獲得しており、このため正確かつ効率的に修正作業を進められるものと考えられる。

d 例え不整合パターンに合致しても、他の帳票等を調査した上で異なる最終判断を下すこともある

e 給与所得者が扶養控除等申告書を提出した給報。本業の給報の可能性が高い。

表 3 住民税賦課業務での不整合パターン例

Table 3 examples of mistake patterns in assessment of resident taxes

単純ミス (記入ミス)	
不整合原因	確定申告書に給与収入を記入するときに書き間違えた
不整合条件	給報が一つのみで、確定申告書と給報で給与収入の数字が異なり、互いの数字が似ている(168万と186万など)
修正箇所	確定申告書の給与収入が誤り
正しい値	給報の給与収入
単純ミス (帳票忘却)	
不整合原因	副業の給与収入を確定申告書に記入し忘れた
不整合条件	甲給報[e]と確定申告書の給与収入が等しく、乙給報[f]が存在する
修正箇所	確定申告書の給与収入
正しい値	甲給報と乙給報の給与収入の合計
記載ルール誤解 (記入欄誤り)	
不整合原因	小規模企業共済等掛金控除を誤って社会保険料控除の内数として記入した
不整合条件	給報の社会保険料控除+小規模企業共済等掛金控除=確定申告書の社会保険料控除
修正箇所	確定申告書の社会保険料控除
正しい値	給報の社会保険料控除
記載ルール誤解 (非記入)	
不整合原因	4人目の扶養者を確定申告書に記入しない
不整合条件	紙で確定申告書を提出しており、確定申告書の扶養者数が3人で、給報の扶養者数が4人以上
修正箇所	確定申告書の扶養者数
正しい値	給報の扶養者数
記載時期ずれ (古い情報)	
不整合原因	年報の扶養者数が古いまま何年も更新されていないため、確定申告書の扶養者数と異なる
不整合条件	確定申告書と年報の扶養者数が異なる
修正箇所	年報の扶養者数
正しい値	確定申告書の扶養者数
記載時期ずれ (変更)	
不整合原因	年末調整の時点では子を父の扶養としていたが、確定申告で母の扶養に変更した
不整合条件	確定申告書と給報の扶養者数が異なり、本人と配偶者の扶養者数の合計が確定申告書と給報で等しい
修正箇所	給報の扶養者数
正しい値	確定申告書の扶養者数

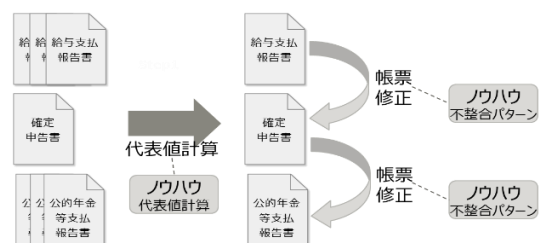


図 7 住民税賦課業務におけるノウハウ利用イメージ

Figure 7 Image of using know-how in assessment of resident taxes

f 給与所得者が扶養控除等申告書を提出していない給報。本業以外の給報の可能性が高い。

5. 対象とする帳票

5章では、本稿で想定する帳票について述べる。

まず、一般の文書と比較した帳票の特徴として、以下があげられる。

- **項目と値の組**: 一つの帳票は項目と値の組の集合で表現できる。
- **帳票種別**: 帳票には種別が存在する。帳票種別により、帳票を構成する項目が異なる。
- **値の型**: 項目ごとにその値の型が決まっている。型は数値や定義済のコードが多く、自由記述は少ない。

これらの特徴を住民税賦課業務で用いる帳票を例に説明する。

- **項目と値の組**: 給与収入が〇〇円、社会保険料控除額が××円のように、帳票は項目と値の組み合わせの集合で表現される。
- **帳票種別**: 確定申告書や給報のように、帳票には種別が存在する。確定申告書なら収入額や控除額、給報なら給与額のように、帳票種別により帳票を構成する項目が異なる
- **値の型**: 給与額や控除人数なら数値、性別ならコードのように、項目により、その値の型が決まっている。帳票のほとんどの型は、金額・人数等の数値か、性別のような定義済のコードであり、テキストが入る項目は氏名、住所、摘要等、まれである。

問題の単純化のため、本稿で扱う帳票には、さらに以下の限定を追加する。

- **型の制約**: 値の型は単一の数値、日付、定義済のコードのいずれかに限定する

他の型の項目は、除去する、あるいは表4で示した方法で前処理することで、上記制約を満たす型にあらかじめ変換しておくことを想定している。

表4 前処理の例

Table 4 examples of preprocessing

前処理前の型	前処理の例
テキスト	クラスタリングしてコード化
住所	町名単位にまとめてコード化
複数の数値の集合 (一つの帳票に同一の項目名の欄が複数ある場合)	総和をとって単一の数値に変換

6. 対象とする問題

本稿で提案する推定方法の目的は、修正済の帳票を学習データとして、未修正の帳票に対し修正すべき箇所と正しい

い値を推定することである。

そのための準備として、全ての帳票を以下の3テーブルに変換しておく。

- 修正済の帳票を、修正前テーブルと修正後テーブルの2テーブルに変換する。ある帳票のある項目が修正されている場合、修正前テーブルと修正後テーブルの間で該当するセルの値が異なる。
- 未修正の帳票を、未修正テーブルに変換する。

各テーブルは以下の性質を持つものとする。

- 1帳票当たり1行
- 全ての帳票は、共通の列としてユーザID、帳票ID、帳票種別を持つ。これらの列には欠損が存在しない。
- その他の列は、項目あたり1列で、全ての帳票で共通。仮に帳票種別ごとに項目名が異なっても、意味が同等なら共通の列にまとめる。型も全ての帳票で共通であり、単一の数値、日付、定義済コードのいずれか。欠損を許容する(その帳票種別に存在しない項目には欠損が入る)

5章で述べた本稿で想定する帳票は、この3テーブルに容易に変換できる。

表5に修正前テーブルと修正後テーブルの例を示す。例えば、「帳票ID0001」の「確申」の「給与収入」のセルの数値が修正前後のテーブルで異なることが、ここが修正されたことを表している(赤字セル)。また、年報には給与収入の欄が存在しないため、テーブル上では欠損となっている。

表5 修正済帳票を変換したテーブル
(上:修正前テーブル、下:修正後テーブル)

Table 5 Tables converted from corrected forms
(upper: before correction, lower: after correction)

ユーザID	帳票ID	帳票種別	給与収入	扶養者数	...
001	0001	確申	300万	2	...
001	0002	給報	150万	2	...
001	0003	給報	200万	2	...
002	0004	年報		1	...
...

ユーザID	帳票ID	帳票種別	給与収入	扶養者数	...
001	0001	確申	350万	2	...
001	0002	給報	150万	2	...
001	0003	給報	200万	2	...
002	0004	年報		1	...
...

表6に未修正の帳票を変換した未修正テーブルの例を示す。列は修正前テーブル、修正後テーブルと共通となる。

表 6 未修正帳票を変換した未修正テーブル

Table 6 A table converted from uncorrected forms

ユーザ ID	帳票 ID	帳票種別	給与収入	扶養者数	...
003	0005	確申	100 万	3	...
003	0006	給報	100 万	2	...
003	0007	給報	50 万	2	...
004	0008	年報		1	...
...

7 章以降で提案する推定方法は、修正前テーブルと修正後テーブルを学習データとして、未修正テーブルの修正すべきセルと正しい値を推定する。

7. 提案方法の概要

本稿で提案する推定方法は、4 章で考察した、ノウハウを使って人が帳票を修正する方法をベースに構築する。以下の 3 ステップで推定する (図 8)。

Step1: 帳票種別の代表値の計算

複数の帳票の値を代表値にまとめる計算方法を、複数用意しておく。代表値計算方法ごとに計算結果を生成し、それらを別々の帳票種別の代表値とみなして、Step2 以降の推定処理を実施する

Step2: 修正箇所を推定:

以下のデータを機械学習で学習して修正箇所を推定する。

- 説明変数: 典型的な不整合条件 (異なる帳票種別間の代表値の差分)
- 目的変数: 帳票種別 × 項目ごとの修正要否

Step3: 正しい値の推定:

以下のデータを機械学習で学習して正しい値を推定する。

- 説明変数: Step2 と同じ
- 目的変数: 修正箇所ごとに、どの帳票種別の代表値に修正されたか

この推定方法を採用した意図は、4 章で議論した、修正作業の効率化のためのノウハウである「不整合パターン(不整合条件と修正箇所と正しい値)」と「代表値計算方法」を機械学習に学習させることにある。これらを学習データ(説明変数と目的変数)に埋め込むことで、代表値計算方法と不整合条件と修正箇所を一括して Step2 で学習し、代表値計算方法と不整合条件と正しい値を一括して Step3 で学習することを想定している。[g]

g 修正箇所として項目 × 帳票種別を、正しい値として代表値を推定する。このため、修正箇所の帳票種別に帳票が複数ある場合、どの帳票をどう修

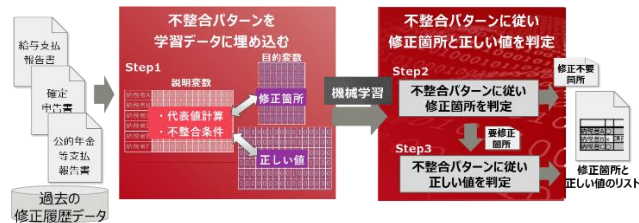


図 8 提案手法のイメージ

Figure 8 An image of proposed method

8. 提案方法の詳細

8 章では提案方法の詳細を述べる。

提案方法では、説明変数および目的変数の生成方法は全ユーザで共通なので、以下は、1 ユーザに限定して説明する。説明で使用する記号の定義を表 7 に列挙しておく。

表 7 記号の定義

(上:インデックス、中:変数、下:関数)

Table 7 Definitions of symbols
(upper: index, middle: variables, lower: functions)

記号	意味	インデックス範囲
m	項目のインデックス	$1 \leq m \leq L$
i, l	帳票種別のインデックス	$1 \leq i, l \leq I$
k	帳票のインデックス	$1 \leq k \leq K(i)$
a, b	代表値計算関数のインデックス	$1 \leq a, b \leq A$

記号	意味
v_{mik}	項目 m の帳票種別 i の帳票 k の修正前の値
\bar{v}_{mik}	項目 m の帳票種別 i の帳票 k の修正後の値
r_{mia}	項目 m の帳票種別 i の代表値 a
d_{mialb}	2 つの代表値 r_{mia} と r_{mlb} の差分
e_{mialb}	説明変数の値 (修正箇所推定と正しい値推定で共通)
o_{mi}^p	修正箇所推定の目的変数の値 (項目 m の帳票種別 i が修正されれば 1)
$o_{m\tilde{i}a}$	正しい値推定の目的変数の値 (項目 m の帳票種別 i が代表値 r_{mia} に修正されれば 1、それ以外の値に修正されれば 0)

記号	意味
f_a	代表値計算関数 a
g_m	項目 m の代表値差分の変換関数

帳票種別の代表値は式(1)で計算できる。

$$r_{mia} = f_a(\{v_{mi1}, \dots, v_{miK(i)}\}) \quad (1)$$

説明変数は式(2)(3)で計算できる。

$$d_{mialb} = r_{mia} - r_{mlb} \quad (2)$$

$$e_{mialb} = g_m(d_{mialb}) \quad (3)$$

修正箇所推定の目的変数は式(4)で計算できる。

$$o_{mi}^p = \begin{cases} 1 & \text{if } \exists_k v_{mik} \neq \bar{v}_{mik} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

正しい値推定の目的変数は式(5)で計算できる。

正すべきかを、推定結果を参考に人が判断する必要がある

$$o_{mila}^v = \begin{cases} na & \text{if } o_{mi}^p = 0 \text{ or } v_{mi1} = r_{mla} \\ 1 & \text{else if } \bar{v}_{mi1} = r_{mla} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

次に、表 8 のサンプルデータを用いて、提案手法によって、どのような学習用データが生成されるかを具体的にみていく。表 8 において修正内容を赤字で示しており、この修正は以下のノウハウに基づいている。

- ① 給報における給与収入の代表値は総和である
- ② 給報における扶養者数の代表値は最大値である
- ③ 確申と給報で給与収入が異なる場合は、単純ミスとみなして、確申を給報に合わせる
- ④ 確申と給報で扶養人数が異なる場合は、記載時期ずれ(変更)とみなして、給報を確申に合わせる

表 8 修正済帳票を変換したテーブルのサンプル
(上:修正前、下:修正後)

Table 8 Samples of tables converted from corrected forms
(upper: before correction, lower: after correction)

ユーザ ID	帳票 ID	帳票種別	給与収入	扶養者数
001	0001	確申	300	2
001	0002	給報	150	2
001	0003	給報	200	2
002	0004	確申	250	1
002	0005	給報	150	1
002	0006	給報	100	1
003	0007	確申	250	2
003	0008	給報	250	3

ユーザ ID	帳票 ID	帳票種別	給与収入	扶養者数
001	0001	確申	350	2
001	0002	給報	150	2
001	0003	給報	200	2
002	0004	確申	250	1
002	0005	給報	150	1
002	0006	給報	100	1
003	0007	確申	250	2
003	0008	給報	250	2

表 8 から学習用データを作るにあたり、関数を以下のように定義する

- 代表値計算関数 f_a : 総和をとる sum と最大値をとる max の 2 つ
- 代表値差分の変換関数 g_m : 符号関数

表 8 から式(1)~(3)を用いて説明変数を計算した結果を表 9 に、式(4)(5)を用いて目的変数を計算した結果を表 10 と表 11 に示す。

表 9 説明変数 e_{mialb} の計算結果
(上: r_{mia} 、中: d_{mialb} 、下: e_{mialb})

Table 9 Tables of explanatory variables e_{mialb}
(upper: r_{mia} , middle: d_{mialb} , lower: e_{mialb})

ユーザ ID	帳票種別		給与収入	扶養者数
	i	a		
001	確申	max[h]	300	2
001	給報	sum	350	4
001	給報	max	200	2
002	確申	max[h]	250	1
002	給報	sum	250	2
002	給報	max	150	1
003	確申	max[h]	250	2
003	給報	sum	250	3
003	給報	max	250	3

ユーザ ID	m	給与収入		扶養者数			
	i	確申	給報	確申	給報		
	a	max	sum	max	sum		
	l	給報		給報			
	b	sum	max	sum	max		
001		-50	100	150	-2	0	2
002		0	100	100	-1	0	1
003		0	0	0	-1	-1	0

ユーザ ID	m	給与収入		扶養者数			
	i	確申	給報	確申	給報		
	a	max	sum	max	sum		
	l	給報		給報			
	b	sum	max	sum	max		
001		-1	1	1	-1	0	1
002		0	1	1	-1	0	1
003		0	0	0	-1	-1	0

組 A 組 B

表 10 修正箇所推定の目的変数 o_{mi}^p の計算結果

Table 10 A table of objective variables o_{mi}^p

ユーザ ID	m	給与収入		扶養者数	
	i	確申	給報	確申	給報
001		1	0	0	0
002		0	0	0	0
003		0	0	0	1

組 A 組 B

表 11 正しい値推定の目的変数 o_{mila}^v の計算結果

Table 11 A table of objective variables o_{mila}^v

ユーザ ID	m	給与収入						扶養者数					
	i	確申			給報			確申			給報		
	l	確申	給報	確申	給報	確申	給報	確申	給報	確申	給報	確申	給報
	a	m a x	s u m	m a x	m a x	s u m	m a x	m a x	s u m	m a x	m a x	s u m	m a x
	b												
001			1	0									
002													
003											1		

h 確申は max と sum の代表値が等しいので、sum は省略し max のみ記載している

説明変数(表9)と修正箇所推定の目的変数(表10)の間に高相関な列が2組存在する(表内の組Aと組B)。また、正しい値推定の目的変数(表11)にも、それぞれの組に対応する範囲が存在する(赤枠と緑枠)。高相関な列の組は機械学習で学習されやすいことから、この学習データを学習した機械学習の挙動は表12のように想定される。この挙動がノウハウ①～④に合致したことは、提案方法で生成した学習データに、ノウハウの情報が埋め込まれていることを示している。

表12 想定される機械学習の挙動

Table 12 Expected machine learning behavior

	不整合条件	修正箇所	正しい値	合致ノウハウ
組A	確申の給与収入と給報の給与収入の総和が異なれば	確申の給与収入を修正する	給報の総和に修正する	①③
組B	確申の扶養者数と給報の扶養者数の最大が異なれば	給報の扶養者数を修正する	確申に修正する	②④

最後に、これまでに作成した説明変数と目的変数を学習する機械学習の手法について議論する。さきほど示したように、提案手法は、説明変数の列が直接ノウハウを表現するよう設計されている。このことから、単独列分岐の組み合わせをベースにしたモデルが適していると考えており、決定木や、そのアンサンブル法[10]、もしくはWide Learning法[11]等が候補として挙げられる。

9. まとめ

本稿では、突合作業の際に人がどのように判断をしているかを、自治体での住民税賦課業務を題材に考察し、不整合パターンや代表値の計算方法といったノウハウを用いて、人が突合作業を効率的に行っている知見を得た。この知見を活用し、このノウハウの情報を学習データに埋め込むことで、過去の修正履歴をもとに未修正の帳票の修正内容を推定する方法を提案した。また、サンプルデータを用い、提案方法で生成した学習データに、このノウハウが埋め込まれていることを確認した。

修正者の作業前に、提案手法が推定した修正内容をシステムが提示することで、修正者にとって修正内容の見当がつけやすくなり、突合作業の効率化が期待される。

参考文献

[1] 北原聖子：“AI等の技術が労働市場に与える影響に関する内外の研究動向について”，ESRI Research Note No.43, 2018
 [2] 柏木恵：“AI-OCR・RPAの活用の現状と課題”，月刊『税』2019年9月号, 2019
 [3] D. Baviskar, S. Ahirrao, V. Potdar and K. Kotecha：“Efficient Automated Processing of the Unstructured Documents Using Artificial Intelligence: A Systematic Literature Review and Future

Directions”，IEEE Access, vol. 9, 72894-72936, 2021
 [4] B. Nguyen, T. Tran, T. Dang, D. Nguyen and H. Le：“HYCEDIS: Hybrid Confidence Engine for Deep Document Intelligence System”，Document Intelligence (DI) Workshop at KDD 2021, 2021
 [5] D.H. Autor, F. Levy and R.J. Mumane：“The Skill Content of Recent Technological Change: An Empirical Exploration”，Quarterly Journal of Economic，118(4), 1279-1333, 2003.
 [6] 河本高文, 二木厚吉, 吉岡信和：“部門ごとの伝票突合せを反映した業務プロセスの信頼性のアセスメント手法”，情報処理学会論文誌, 59(9), 1699-1708, 2018
 [7] “突合点検・縦覧点検”，<https://www.ssk.or.jp/shinryohoshu/tenken/>, (参照 2022-02-03).
 [8] “税の情報・手続・用紙”，<https://www.nta.go.jp/taxes/index.htm/>, (参照 2022-02-01).
 [9] “地方税分野の主な申告手続等における様式【税目別】”，https://www.soumu.go.jp/kojinbango_card/mynumber_tax_zeimokubetu.html, (参照 2022-02-01)
 [10] L. Breiman：“Random Forests”，Machine Learning 45, 5-32, 2001
 [11] 大堀 耕太郎 他：“知識発見によって信頼をつなぐ Wide Learning 技術”，雑誌 FUJITSU 70, 4, 2019