農林業基準技術に含まれる表を対象とした PDF から CSV へ変換する際の課題分析

杉山陽菜乃 [1] 阿部瑞稀 [1] 中村彩乃 [1] 前多陸玖 [1] 坂口遥哉 [1] 佐藤栄作 [1] 木村泰知 [1]
小林暁雄 [2] 大友将宏 [2] 石原潤一 [2] 桂樹哲雄 [2] 川村隆浩 [2]
[1] 小樽商科大学 [2] 農業・食品産業技術総合研究機構
{g1202219307,g1202200907,g1202228307,g2021344,g2021163}@edu.otaru-uc.ac.jp
g1202453006@edu.otaru-uc.ac.jp kimura@res.otaru-uc.ac.jp
{akio.kobayashi,masahiro.otomo,ishiharaj612,t.katsuragi,takahiro.kawamura}@naro.go.jp

概要

内閣府「研究開発と Society 5.0 との橋渡しプログラム（BRIDGE）」の農林水産省課題「AI 農業社会実装プロジェクト」では，国内農業知識を学習させた LLM の開発を進めている．これを実現するには，国内に散在する農業に関するデータを収集し，機械可読な形式へ変換する必要がある．本研究では，多くの都道府県から公開されている農業基準技術に関する文書を対象とし，そこに含まれている，人手で作成された経営類型や技術体系や作業別労働時間などの表から適切な形式でのデータの抽出を目的とする．本稿では，その基礎的な分析として，PDF 形式の表を CSV 形式に変換する際に生じる課題を明らかにした．

1 はじめに

農業の技術継承は，高齢化が進む日本において非常に重要な問題となっている．国内就農者の平均年齢は 68.7 歳（2023 年）[1)] と高齢であり，今後 20 年間で現在の 3/4 ほどの就農者が離農するという予測が立てられている [2)]．

これに対し，内閣府「研究開発と Society 5.0 との橋渡しプログラム（BRIDGE）」[3)] における農林水産省実施施策「AI 農業社会実装プロジェクト」[4)] では，農業従事者の減少による農業労働力の減少を補うための AI 技術の確立を目的としており，普及指導員による営農指導を農業ドメイン知識を学習した LLM（農業 LLM）によって補うことなどを想定している．

このプロジェクトでは，各都道府県の農業普及センターや公設試験研究機関（公設試）が公開している，普及指導員などの専門家によって作成された農業に関する技術文書を収集し，農業 LLM の構築を進めている．これらの収集された技術文書の多くは，機械可読な形式での構築は行われておらず，農業 LLM に使用するにあたり前処理として機械可読な形式への変換が必要になる．特に，各公設試がまとめている，設置都道府県における作物別の基準技術文書（農業基準技術文書）は殆どが異なる都道府県間で共通の表形式の文書となっている．これらの表データを適切に機械可読な形式とすることで，農業 LLM を用いる RAG システムなどで活用できると考えられる．

このような背景から，本研究では，農業基準技術文書に含まれる表を機械可読なデータベースに変換することを目的とする．本稿では，そのための基礎的な分析として，PDF 形式の表を CSV 形式に変換する際に生じる課題を明らかにし，自動化の可能性を検討する．これにより，地域コミュニティを基盤とした技術継承モデルの構築を目指す．

本研究の貢献は，以下の 3 点である．

- **農林業基準技術の PDF に含まれる表の整理**：長崎県の 119 件の PDF ファイル（総ページ数 723 ページ）に含まれる表を整理した
- **視覚的に確認可能な PDF 上の課題分析**：セル結合や非スカラ値などの変換時に生じるする具体的な課題を 8 つのタイプへ分類した
- **CSV 自動変換の際に確認される課題分析**：既存の変換ツールを使用することで，初めて気づく

1） https://www.maff.go.jp/j/tokei/sihyo/data/08.html
2） https://www.affrc.maff.go.jp/docs/press/attach/pdf/240604-2.pdf
3） https://www8.cao.go.jp/cstp/bridge/index.html
4） https://www.affrc.maff.go.jp/docs/bridge/attach/pdf/2023bridge-7.pdf

ことができる課題を分析した

2 関連研究

機械判読を目的とした表の分類の関連研究には，有価証券報告書を対象とした分類がある [1, 2]．この研究では，有価証券報告書に含まれる機械判読が困難な表のセル分類を目的として行った．その中で，有価証券報告書内の機械判読が困難な表を「小見出し行」「複数の Header（セル結合を含む）」「空白セル」「非スカラー値」「特殊な形」の 5 つのタイプに分類した．それらを含んだデータセットに対して一般的な機械学習手法を用いたアプローチを行うことで，有価証券報告書における表を対象としたセル分類の困難性を実証した．

Ye ら（2023）は LLM を用いた表形式データの推論の向上を目的とし，表形式データを小さな集合に分割し，複雑な質問をより単純な部分質問に分割することで，より正確な検索が可能であることを示した [3]．その中で，LLM が表構造データの推論を不得意とする理由について，「LLM のトークン制限を超過する巨大な表形式データ」「比較，集計，演算などを必要とする複雑な質問」「テキストと表形式データの複雑な相互理解」の 3 点にあると指摘した．また，従来の表形式推論の手法が，列やヘッダーなどの表の構造的な情報を主に利用し，セルのテキスト情報が重視されていなかった点を指摘し，LLM によりセルテキストの意味を抑えた表構造理解が可能であることを示した．

3 対象データ

本研究では，令和 5 年度版の「長崎県農林業基準技術」を対象のデータとして分析を行う．「長崎県農林業基準技術」は，長崎県の公式ホームページ上で PDF ファイルとして公開されており[5)]，長崎県の農林業技術及び経営の基本指標として 5 年ごとに策定されている．同サイトから Python のライブラリである Beautiful Soup を用いてスクレイピングを行いデータを収集した．また PDF から CSV への変換に際しては，Python のライブラリである pdfplumber を用いて行った．

集計した PDF ファイル数はで 119 件，総ページ数は 723 ページ，pdfplumber が表として認識したものの総数は 1,197 件であった．pdfplumber が表として認識したもののの中で，表及び表として取得したいものは 1,007 件であった．件数は「(pdfplumber によって検知した表数+人手による検知数)-(図的利用+表ではない+ページを横断する表+空白誤認+オブジェクト矩形誤認+テキスト矩形誤認+合成誤認+重複誤認)」という計算式で算出した．

表のフォーマットや記述内容は，PDF ファイルごとに概ね共通したものがみられる．PDF ファイルは大きく「経営類型・単一品目」「品目別施肥基準」「長崎県農林業技術の策定に伴う基本的な考え方」「家畜部門・経営類型」「農耕地土壌の概要と改良基準，土壌管理」の 5 つのファイルと 13 のラベルに区分される．表 1 は PDF ファイル・表ラベルごとの統計情報である．

表 1 PDF ファイル・表ラベルごとの統計情報

PDF ファイル・表ラベル	合計
経営類型・単一品目のファイル	
技術体系の特徴	103
資本装備と減価償却費	104
技術体系	315
品目の作付体系	97
作業別・月旬別労働時間	214
総労働時間	102
家畜部門・経営類型のファイル	
畜舎及び付帯施設算出基礎	12
養分必要量	10
常時飼育頭数	8
品目別施肥基準	
施肥量	17
施肥時期及び割合	12
農耕地土壌概要と改良基準のファイル	
土壌改良基準	9
長崎県農林業技術の策定のファイル	
部門別の経営類型と単一品目 (作型)	19
その他	112
合計	1,007

4 PDF を CSV に変換する際の課題

PDF の表構造は視覚的には整っているが，機械判読可能な形式へ変換する際には，多くの課題が存在する．CSV 形式への変換作業では，表の認識やデータ抽出が困難になる場合がある．

そこで，人間が視覚的に判断できる PDF 上の課題，および，CSV へ自動変換する際に生じる課題の

5) https://www.pref.nagasaki.jp/bunrui/shigoto-sangyo/nogyo/nouringyoukijyungijyutu/

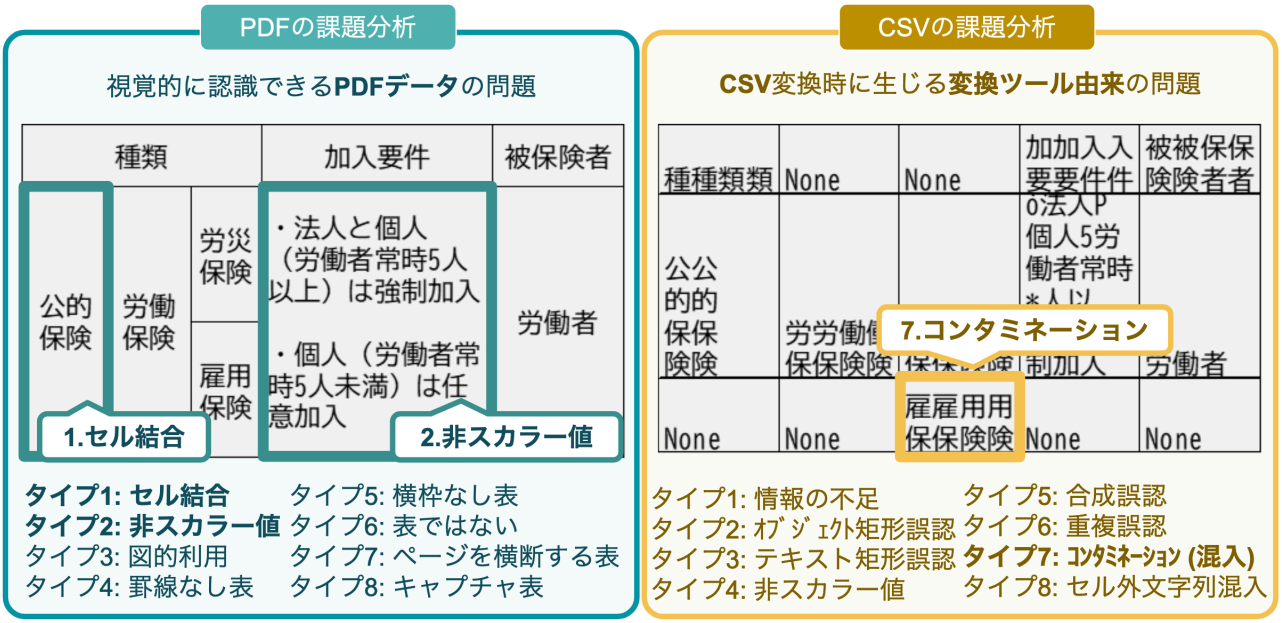

図 1　「視覚的に確認可能な PDF 上の課題分析」および「CSV 自動変換の際に確認される課題分析」のイメージ

２つの観点から，農林業基準技術の PDF 文書に含まれる「表」を対象として，PDF を CSV に変換する際の課題について述べる．図 1 に２つのデータ分析のイメージを示す．

4.1　視覚的に確認可能な PDF 上の課題

PDF に含まれる表を CSV へ変換する際には，PDF 形式特有の課題が存在する．本節では，CSV 変換前に認識可能な課題，すなわち，人間が視覚的に判断できる PDF 上の課題について，予備調査に基づき 8 つのタイプに整理した結果について述べる．

タイプ 1: セル結合　複数のセルが一つのセルに結合されており，データの分割が困難な表
タイプ 2: 非スカラ値　一つのセルに複数の値が含まれており，値の抽出が困難な表
タイプ 3: 図的利用　表がデザイン的に加工され，図の一部として利用されている表
タイプ 4: 罫線なし表　罫線がなく，セルの区切りが不明確なため，値の分割が困難な表
タイプ 5: 横枠なし表　表全体を囲む枠線がなく，データ領域を検出するのが難しい表
タイプ 6: 表ではない　表の形状をしているが，データとして扱えない擬似的な表
タイプ 7: ページを横断する表　複数ページにまたがるため，ヘッダー情報が欠落する表
タイプ 8: キャプチャ表　表に見えるが，画像として保存されており，データの取得が困難な表

4.2　CSV 自動変換の際に確認される課題

PDF に含まれる表を CSV へ変換する際には，変換ツールを使用することで，初めて気づくことができるデータ抽出の課題が存在する．これらの課題は，PDF 内の記述方法や情報の保存形式に依存しており，自動化を妨げる要因となっている．本節では，pdfplumber [6] を用いて PDF から CSV に変換した際に確認された課題について，予備調査に基づき 8 つのタイプに整理した内容について述べる．

タイプ 1: 情報の不足　CSV への変換時にセル内のデータが欠落し，必要な情報を取得できない
タイプ 2: オブジェクト矩形誤認　表ではない矩形 (四角形の領域) を表として誤認する
タイプ 3: テキスト矩形誤認　文字列や数字を囲む四角い装飾が表として認識される
タイプ 4: 非スカラ値　セル分割されている値が CSV 変換時に一つのセルに集約される
タイプ 5: 合成誤認　二つ以上の独立した表が，一つの表として認識される
タイプ 6: 重複誤認　同じ表が複数回 CSV に変換され，重複したデータとして扱われる
タイプ 7: コンタミネーション　PDF 上では視認できない不要な情報が CSV 変換時に混入する
タイプ 8: セル外文字列混入　表タイトルなどのセル外の文字列を CSV データとして認識する

6)　https://github.com/jsvine/pdfplumber

5　PDF および CSV の課題分析

本分析の目的は，農林業基準技術に含まれる PDF ファイルを CSV 形式に変換する際に生じる課題を明らかにすることである．本分析では，3 で述べた，長崎県の農林業基準技術の文書に含まれる 1,197 件の表を対象として，4 で定義した PDF から CSV へ変換する際の課題ごとに，人手で分類することで，自動化に向けた基礎情報を提供する．

5.1　分析の手順

PDF ファイルを CSV に変換し, そのデータから表に見られる諸特徴の抽出・ラベリングを行った. 次に, このラベリング結果を基に, 分析対象の項目を整理するための体系的なチェックリストを作成した. その後，119 件の PDF ファイルを手作業で 3 人の人手で確認し, 作成したチェックリストを活用して表の分類を可視化した.

5.2　分析の結果

5.2.1　PDF データ分析の結果

表 2 は，PDF 文書上で確認された表の特徴ごとの数と割合を示している．最も多かった特徴は「セル結合」であり，全体の 70.6%を占めている．「非スカラ値」の表も 60.5%と多く，これらが PDF 文書内で一般的な特徴であることがわかる．一方，「図的利用」（8.4%）や「ページを横断する表」（1.9%），「罫線なし表」（0.6%）などは比較的少数である．

表 2　PDF 上で確認される特徴的な表の割合

表のタイプ	タイプ名	表の数	割合
1	セル結合	830	70.6%
2	非スカラ値	712	60.5%
3	図的利用	99	8.4%
4	罫線なし表	7	0.6%
5	横枠なし表	11	0.9%
6	表ではない	32	2.7%
7	ページを横断する表	22	1.9%
8	キャプチャ表	12	1.2%
-	その他	64	5.4%

5.2.2　CSV データ分析の結果

表 3 は，CSV データ上で確認された表の特徴的な誤りや問題点について，そのタイプごとの数と割合を示している．「情報の不足」が最も多く確認され，全体の 10.9%を占めている．「コンタミネーション」（7.1%）や「オブジェクト矩形誤認」（3.9%）も比較的多く，これらの誤認識がデータの正確性を損なう要因となっている．その他，「非スカラ値誤認」（3.1%），「セル外文字列混入」（1.1%），「合成誤認」（1.0%），「重複誤認」（0.4%）は比較的少数だが，これらの対処も必要である．これらの結果から，CSV データの解析では，誤認識やデータ不足といった問題に対応する必要があることが明らかになった．

表 3　CSV 上で確認される特徴的な表の割合

表のタイプ	タイプ名	表の数	割合
1	情報の不足	110	10.9%
2	オブジェクト矩形誤認	40	3.9%
3	テキスト矩形誤認	15	1.5%
4	非スカラ値誤認	31	3.1%
5	合成誤認	10	1.0%
6	重複誤認	4	0.4%
7	コンタミネーション	71	7.1%
8	セル外文字列混入	11	1.1%
-	その他	104	10.3%

6　おわりに

本稿では，最終目標に到達するための基礎的な分析として，農業基準技術文書に含まれる表をデータベースに変換する際の課題と，その自動化の可能性について述べた．

謝辞

本研究は，JSPS 科研費 JP21H03769 および，内閣府「研究開発と Society 5.0 との橋渡しプログラム（BRIDGE）」における農林水産省実施施策「AI 農業社会実装プロジェクトの助成を受けて実施された．また，協力頂いた合同会社 Nita Limo に感謝します．

参考文献

[1] 前多陸玖, 奥山和樹, 佐藤栄作, 木村泰知. 有価証券報告書を対象とした機械判読が困難な表のセル分類に向けて. 人工知能学会全国大会論文集, Vol. JSAI2024, pp. 3M5OS12b03–3M5OS12b03, 2024.

[2] 奥山和樹, 木村泰知. 有価証券報告書を対象とした機械判読が困難な表構造の分析. 言語処理学会第 30 回年次大会 (NLP2024), pp. P3–20–, 3 2024.

[3] Yunhu Ye, Binyuan Hui, Min Yang, Binhua Li, Fei Huang, and Yongbin Li. Large language models are versatile decomposers: Decompose evidence and questions for table-based reasoning, 2023.

付録