Q7-20

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]農業研究機構

背景・目的

農業の技術継承の問題とは ？

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）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。

内閣府「研究開発と Society 5.0 との橋渡しプログラム（ BRIDGE）」

農業労働力の減少を補うための AI 技術の確立を目指している。
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**）で補う。

農業 LLM 構築に必要となる「技術継承の文書」とは？

各都道府県の農業普及センターや公設試験研究機関が公開している。
例えば、**農林業基準技術についての文章** がある。

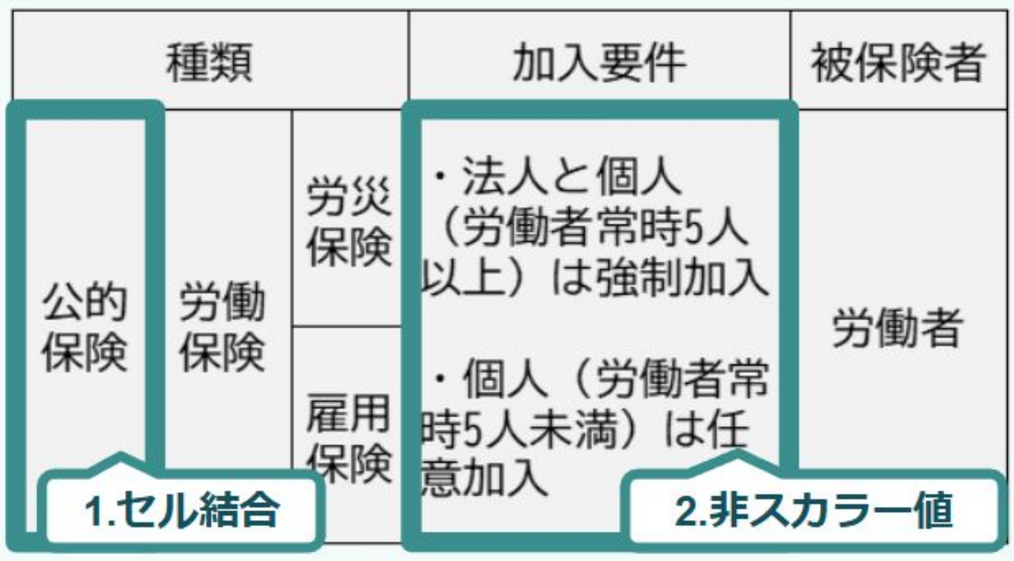
しかしながら、**機械判読可能な形式ではない** 。

また、技術継承の文章には、**機械判読が困難な「表」が数多く含まれている** 。

本研究の目的

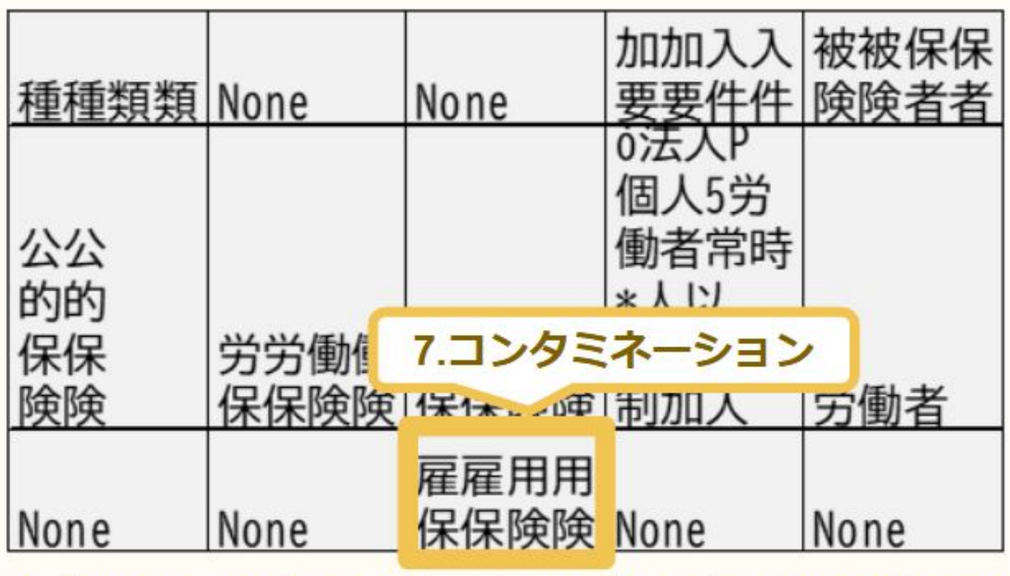
農業基準技術文書に含まれる「表」を機械判読可能な形式に変換する。
第一段階として、PDF形式の表をCSV形式に変換する際に生じる課題を明らかにする。

PDF分析
視覚的に認識できる対象データ自体の問題



タイプ1: セル結合 タイプ2: 非スカラー値 タイプ3: 図的利用 タイプ4: 罫線なし表
タイプ5: 横枠なし表 タイプ6: 表ではない タイプ7: ページを横断する タイプ8: ｷｬﾌﾟﾁｬ表

CSV分析
CSV変換時に生じる**変換ツール由来**の問題



タイプ1: 情報の不足 タイプ2: ｵﾌﾞｼﾞｪｸﾄ矩形誤認 タイプ3: ﾃｷｽﾄ矩形誤認 タイプ4: 非ｽｶﾗｰ値
タイプ5: 合成誤認 タイプ6: 重複誤認 **タイプ7: ｺﾝﾀﾐﾈｰｼｮﾝ (混入)** タイプ8: セル外文字列混入

タイプ1. セル結合

作業の種類	栽培技術	
	技術内容	作業時期
品種の選定	①種子注文	
播種	①耕起	10～11

タイプ2. 非スカラー値

	種類・規模	数 量
建物・施設	加温ハウス（連棟標準型AP）	3
	簡易ハウス	1
	作業収納舎66㎡	1
	貯蔵庫33㎡	1
	園内道	10
	重油タンク	3
	防油堤	3
	計	

タイプ3. 図的利用

	1	2	3	4	5	6
ハウスびわ（一般出荷型）	↑摘果	↑袋かけ	■◆	■◆	↑せん定・施肥	↑芽かき
ハウスびわ「はるたより」	↑摘果	↑袋かけ		■◆■	◆↑せん定・施肥	↑芽かき
簡易ハウスびわ「なつたより」（無加温）		↑ ↑摘果	∪袋かけ	■	◆■◆↑せん定・	施肥

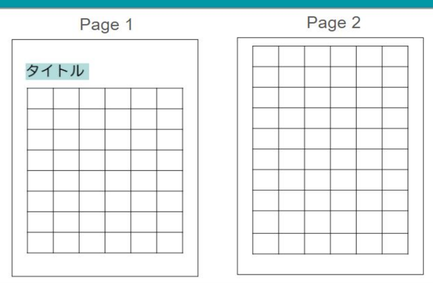
タイプ4. 罫線なし表

水　田	灰色低地土	70~90%
	黒ボク土	60~90%
普通畑	森林褐色、灰色低地土	70~90%
	黒ボク土	60~90%
樹園地	茶園以外	50~80%
	茶園	25~50%

タイプ5. 横枠なし表

色	黄～黄褐色(2)、褐色(5)、黒褐色～黒色(10)
形状	現物の形状をとどめる(2)、かなりくずれる(5)、ほとんど認めない(10)
臭気	糞尿臭強い(2)、糞尿臭弱い(5)、たい肥臭(10)
水分	握ると指の間からしたたる・・・70%以上(2) 強く握ると手のひらにかなりつく・・・60%前後(5) 強く握っても手のひらにあまりつかない・・・50%前後(10)
最高温度	50℃以下(2)、50～60℃(10)、60～70℃(15)、70℃以上(20)

タイプ6. 表ではない表

販売金額＋副産物価格等	農業経営費	物財費		
		雇用労働費		
		支払地代		
		支払利子		
		販売経費		
	農業所得	家族労働費		
		自作地代	農業純収益	
		自己資本利子		家族労働報酬
		農企業利潤		

タイプ7. ページをまたぐ表