研　究　報　告　書

「レガシーデータに基づくイネの品質と生産性に関わる因果関係の解析と機械学習を用いたオンサイト生育診断技術の開発」

研究タイプ：通常型
研究期間：　2016 年 10 月～2020 年 3 月
研　究　者：　西内　俊策

1．　研究のねらい

農業において重要なことは、生産性と品質の両立である。しかし、実際の農業現場では生産性と品質の向上を両立することは難しく、生産性か品質かのいずれかを優先する必要があるという認識が存在している。稲作に限らず、農業分野ではこのような生産性と品質のトレードオフは良く見られる状況であり、その中での利益の最大化を目指した具体的な提案が求められている。また、環境変動下で持続可能で安定した作物生産の為には、栽培地域毎に動的な栽培暦を用意することに加えて、確率モデルを元に気象変動に対するリスク評価を行っておくなど、農家の判断の助けとなる情報を用意し、提供する必要がある。本研究の目的は、農学的知見と情報分野の技術を組み合わせ、生産性と品質の因果関係を記述することで、農業上の意思決定を支援する技術の確立に貢献することである。

栽培地や気象条件、品種に合わせた栽培管理の最適化に資する知見を得るためには、信頼のおける栽培記録を蓄積することと、集めたデータに基づき、一般化した予測モデルの開発とその評価が必要である。日本では、それぞれの地域に適合した品種の選定などを目的として、水稲の試験栽培が各都道府県の農業試験場などで行われており、水稲の出穂予測モデルの開発などに利用されてきたが、栽培条件の詳細が不明であるなどの問題から、生産性や品質に掛かる解析に活用することは十分に行われてこなかった。

そこで本研究では、全国で蓄積されたイネ栽培レコードから、メッシュ気象情報を用いて各試験実施時のイネの生育状況を分類し、各年度の気候が収量と品質にどのように影響したかを確率的グラフィカルモデルにより解析し、それらの因果関係の強弱を予測する。それにより、気象条件や施肥体系が、どのように高収量や高品質に結びつくか、一般化された情報を得ることを目指す。このことは、生育状況に合わせた施肥、水管理等の栽培管理の最適化に資する知見となることが期待される。また、本研究では、他作物に比べて研究の進んだ水稲を対象として解析を行うことで、既存の知見が栽培情報の解析から抽出可能か、また新知見が得られるかを検証する。この解析を通して、他作物における栽培記録の価値について間接的に評価を行うことも目標とする。

本提案課題においては、

2．　研究成果

（1）概要

本研究では、全国の農業試験場で蓄積されたイネ栽培のレコードと、愛知県のより詳細な栽培記録のデジタル化を通して米の品質と収量を目的変数とする水稲の確率的グラフィカルモデルを記述し、米の品質と収量に影響を与える可能性の高い環境要因について推定を行

った。

研究テーマ A「イネ栽培レコードのデジタル化とデータセット整備」では、実験ノートや野帳のスキャン作業から、データ入力、データクリーニング、他のデータベースとの結合までを行い、約25000枚の画像データから約11000件の奨励品種決定試験相当のデータをデジタル化した。これは、国立研究開発法人農業・食品産業技術総合研究機構の管理する水稲奨励品種決定基本調査成績データベースのデータ数の 5%に相当し、愛知県のデータ量を 3 倍にすることが出来た。

研究テーマ B「収量と品質に関する予測モデルの開発」では得られたデータセットを用い、水稲の収量や品質について推定モデルの開発を行った。差分データセットを生成することにより、推定精度を向上することが出来、15 品種で相関係数 0.7 以上で収量推定が可能であった。

また、研究テーマ C「グラフィカルモデルを利用した確率モデルの設計」では、グラフィカルモデルを利用した確率モデルから、水稲の品種毎に見られる気象との因果関係について解析を行い、収量の増加に影響すると知られている元肥や追肥の量、栽植密度、追肥回数といった栽培管理が、必ずしも収量の増加に等しく寄与しない、ということが明らかになった。

これらの解析から、レガシーデータのデジタル化とその評価からこれまで試験栽培等で示されてきた既存の知見と同等の結果を情報解析により求められることが明らかとなった。

（2）詳細

研究テーマ A「イネ栽培レコードのデジタル化とデータセット整備」

本研究では 1980 年以降、全国の農業総合試験場での水稲の栽培記録である奨励品種決定試験データと、愛知県、岐阜県の農業総合試験場での試験栽培記録について、そのデジタル化とクリーニングを行い、解析に用いるデータセットの整備を行った。奨励品種決定試験のデータとして、国立研究開発法人農業・食品産業技術総合研究機構の管理する水稲奨励品種決定基本調査成績データベースを用いた。このデータベース中には、全国で行われた水稲奨励品種決定試験のデータが記載されており、30 年間で全国 20 万レコードを超える品種名、田植え日、出穂日、登熟日という基礎的なデータに加え、施肥体系、収量、品質といったデータが含まれている。しかし、日付表記の揺れやデータの抜けなどが散見された為、それらを探し出し削除や修正を行った。また、各県で記録され集められたデータであるため、品質の評価軸が試験場毎に基準が異なる等比較が難しい状態であったため、データの正規化、階級値の再割り当てを行い比較可能な形にした。

愛知県、岐阜県の農業総合試験場の記録は、研究利用という枠内で複写、デジタル化等の許可を得て入手した。年度毎に発刊された試験場成績報告書に加え、圃場での記録媒体となる野帳、当初の予定である試験計画書とそれに書き込まれた変更や備考について、試験毎に確認を行った。それにより、水稲奨励品種決定試験基本調査成績データベースに記載された項目に加えて、データベースに残っていない施肥体系や中干し実施の有無など、可能な限り項目の収集とクリーニングを行った。また、栽培地の緯度経度を元に、国立研究開発法人農業・食品産業技術総合研究機構の全国 1km メッシュ農業気象データと国立研究開発法人農業・食品産業技術総合研究機構農業　環境変動研究センターの日本土壌インベントリー

を外部データベースとして紐付け、各栽培データに生育期間中の気象情報と栽培地の土壌分類を加えられるデータセットとして纏めた。しかし、事前の想定以上の資料の散逸状況や難読性に直面し、1978 年以前の栽培情報のデジタル化は断念することになった。

総スキャン枚数は約 25000 枚で 15000 件相当の栽培データを追加し、データクリーニング後、約 11000 件の奨励品種決定試験に準ずるデータのデジタル化を完了した。水稲奨励品種決定試験基本調査成績データベースに記載のある愛知県の試験情報は約 4700 件であり、愛知県での栽培情報を約 3 倍に増やすことが出来た。

研究テーマ B「収量と品質に関する予測モデルの開発」

データセットを用い、水稲の収量と品質を推定するモデルの開発を行った。一件の栽培情報につき、作業内容を示す説明変数に加えてサンプリングの対象や集計方法の異なる気象データを 420 変数用意し、主成分分析や PLS 回帰、XGBoost 等いくつかの機械学習を試した。データセット中に 200 レコード以上含まれている 84 品種について、気象データと生育値の差分を説明変数とした主成分回帰により、収量の推定を行った。80%を教師データ、残り 20%をテストデータとして、テストデータを用いた推定値と実測値の相関係数を求めたところ、相関係数は平均 0.6 程度であった(図 1)。上位 15 品種については相関係数は 0.7 を上回っており、気象と栽培管理が収量との間で強い関連があることが示唆されたが、一方で気象や栽培管理との関連が弱い品種も見出されており、より高度な生理モデルの必要性が示唆された。

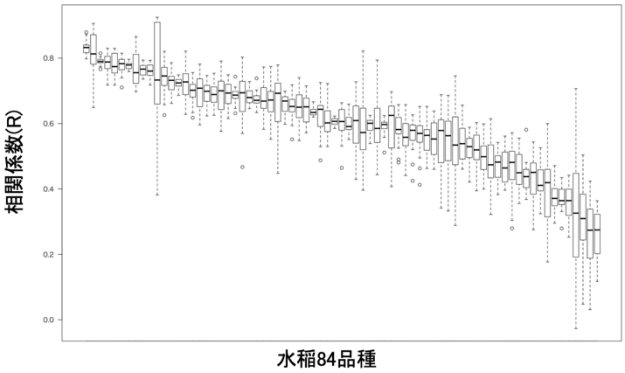

図1. 時間的、空間的に近接した栽培データによる収量推定
水稲84品種の栽培データを用い、時間的、空間的な影響を加味して収量推定モデルを作出し、実測収量と推定収量間の相関係数(R)を示した。

研究テーマ C「グラフィカルモデルを利用した確率モデルの設計」

水稲の表現型には、「環境」と「栽培管理」の影響を受けるものと受けないものがある。さらにそれは品種によって異なると仮定した。品種毎に表現型を説明するグラフィカルモデルを作成し、有意に出現頻度の高い因果関係を抽出することで、一般化した因果情報を記述出来ると考えた。そのために、研究テーマ A で得られたデータセットを用い、水稲の品種毎に品質と収量を目的変数とするグラフィカルモデルの作成を行った。1 例として収量性に優れる品種に共通する因果関係を示したネットワークを示す(図 2)。

モデル作成にはベイジアンネットワークを用い、説明変数として採用した気象情報や形質値については、必要に応じて標準化と離散化を行った。得られたネットワーク構造から、高頻度で観察される因果関係を抽出し、それらの類似性から水稲品種のクラスタリングを行った。それにより、品種の遺伝型に依らず気象や栽培管理に対する応答性の類似した品種クラスタを纏め、そのクラスタ毎に収量に影響を与える条件を抽出した。

これまでの水稲研究から、収量の増加に影響すると知られている元肥や追肥の量、栽植密度、追肥回数といった栽培管理が、必ずしも収量の増加に等しく寄与しない、ということが明ら

かになった。そして、栽植密度が収量増に強く貢献する品種が見出されるなど、品種毎に適した栽培管理方法があることを示す結果を得た。また、多収品種と収量性の低い品種を比較した場合に、多収品種で共通して見られる栽培管理への応答性が明らかとなり、育種時の選抜基準として注目すべき形質があることを示す結果を得た。

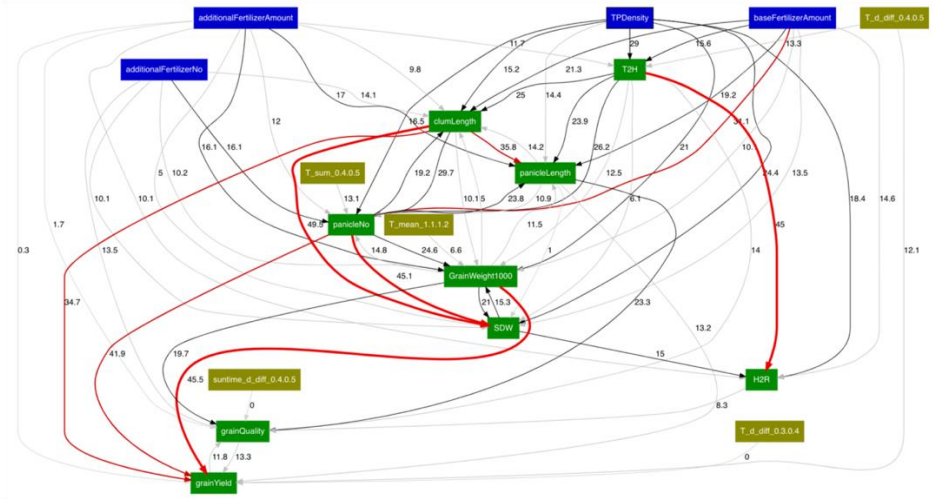

図2. 多収品種におけるベイジアンネットワーク図
収量性に優れる10品種の栽培データを用いて、収量に繋がる因果関係をネットワーク構造として記述した。線の太さと色が結合の強さを示している。

本研究で、レガシーデータのデジタル化とその評価からこれまで試験栽培等で示されてきた既存の知見と同等の結果を情報解析により求められることが明らかとなった。一方で、登穂日数に日長の変動が影響する、といった、過去に棄却された仮説に類似する知見も出てきたことから、情報解析により見出された知見の新規性については改めてデータを選抜して検証する必要があることが分かった。

3. 今後の展開

本課題は、過去の栽培情報を元に、グラフィカルモデルを利用した確率モデルを作成し、収量と品質に関わる気象リスク評価を行うという考え方を稲作に適応したものである。今後は作成したモデルの実証評価を重ね、より普遍的に利用できる技術を開発する。

これまでに多くの作物で試験栽培が行われており、デジタル化されたデータに加えて各試験機関や研究機関に紙媒体で保存されているものを含めると相当量の栽培記録が日本に存在している。全国区のデータからは汎用的な知見が、特定の地域のデータからは、より地域の実情に近い知見が解析により得られると考えられる。それら蓄積されたレガシーデータは、今後、知的資産として再評価され、積極的に利用されることによって ICT を活用した農業現場での栽培管理指針に大きく貢献することが期待される。

また、過去の栽培記録に裏打ちされたモデルは、イネに限らず、農業現場における意思決定において有用な判断基準となると考えられ、他の作物でも同様のアプローチが適応されることが期待される。そのために必要な具体的な栽培データの記述や気象情報の扱いが今後研究されると考えられる。

4.　自己評価

研究目的の達成状況としては、基盤となるデータセットの整備は、愛知県、岐阜県の栽培記録についてデジタル化を進め、解析に利用可能なデータセットとして検証出来たことから当初の目的は達成できた。しかし、1985 年以前の栽培情報のデジタル化が出来なかったことや、確率モデルの設計とその評価で得られた知見が既存の個別研究を上回るものではなく、取り組む余地を残した結果となった。解析手法についても今後検討が必要だと考えている。

栽培記録のデジタル化については、当初想定した以上にトラブルに見舞われたものの、十分な予算があったため、解析に値する栽培記録について整備することができた。研究実施体制は計画通りに研究に必須となるデータの入力や確認を担当する研究補助員を確保した。研究費は人件費を始め、デジタル化や解析機器など、本課題遂行の為に適切に執行した。

機械学習や深層学習といったモデル化技術において高品質なデータを多数揃えることが重要であり、一年一作が基本の農業分野では、データ準備が課題となることは明らかである。本研究で取り組んだレガシーデータのデジタル化と整備は、レガシーデータの発掘と再評価が今後重要なデータ確保手段になり得るかどうかの試金石であった。本研究で、レガシーデータのデジタル化とその評価からこれまで試験栽培等で示されてきた既存の知見と同等の結果を情報解析により求められたことは、今後の作物生産の最適化を実現するための重要な情報源として、栽培情報を蓄積する意義を示したという点で価値があったと考えている。

5.　主な研究成果リスト

（1）論文（原著論文）発表

該当無し

（2）特許出願

研究期間累積件数：0 件

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

【国内学会】

・西内俊策. 気象とイネの生産性にみられる因果関係の解析. 日本育種学会第 131 回講演会. 2017 年 3 月.

・近藤拓也, 西内俊策. イネの出穂予測精度向上を目指した農業情報の利用. 人工知能学会第 31 回全国大会. 2017 年 5 月.

・西内俊策. イネのレガシーデータ解析って何が分かるのか？. 第 5 回農学中手の会. 2019 年 12 月.

・西内俊策, 松井秀俊. 栽培記録から見出された水稲の環境応答変動性の解析. 日本育種学会第 137 回講演会. 2020 年 3 月.

他 10 件