

気象データの精緻化と 機械学習手法によるてん菜収量予測

国立研究開発法人 農業食品産業技術総合研究機構
北海道農業研究センター 寒地畑作研究領域 スマート畑作グループ
(現・東北農業研究センター 水田輪作研究領域 ICT活用技術グループ) 研究員 金谷 真希

【要約】

北海道の基幹作物であるてん菜は、熟練職員の減少や気候変動の影響により収量予測の難易度が高まり、より客観性の高い収量予測手法へのニーズが高まっている。そこで、筆者は機械学習と高解像度気象データを利用したてん菜収量予測Webアプリを開発した。本アプリは収量予測を行うだけでなく、収量と相関の高い説明変数（予測値に影響を与えている要素）をユーザーに示す機能がある。解像度が高く、解釈が可能な収量予測ツールの開発は、輸送の効率化や経験知の見える化に貢献できる可能性がある。

はじめに

てん菜は、国産の砂糖原料の約8割を占める製糖原料作物であり、北海道の畑輪作体系を支える重要な基幹作物である。てん菜の栽培方法には移植栽培と直播栽培があり、近年は高齢化や労働力不足などの影響から、より省力的な栽培方法である直播栽培への転換が進んでいる。2024年度のてん菜の直播栽培の面積は、作付面積全体の約50%であり、2014年度からの10年間で3倍以上に増加した¹⁾。毎年4～5月頃に植え付けられたてん菜は、10～11月にかけて収穫され、一時保管場所から製糖工場へトラックで輸送される。凍結による品質低下を避けるため、てん菜の輸送はおおむね12月までに終える必要がある。そこで、事前に収量の予測を行い、それに基づいた輸送計画の立案が重要となる。製糖会社・団体では、収穫の約2カ月前に当たる8月下旬に、圃場^{ほじょう}での掘り取り調査や巡回調査の結果を基に収量予測を行ってきた。しかし、熟練職員の減少や近年の気候変動の影響、とりわけ高温多雨傾向や、これに伴う病害の影響によって予測の難易度が高く

なりつつある。そこで、ノウハウの継承の観点からも、気象経過のデータを用いたより客観性の高い予測手法へのニーズが高まっている。

1 高精度気象データの活用が収量予測にもたらす効果

気象データを利用した収量予測に関する先行研究は、気象台やアメダスなどで観測された気象データを用いているものが多い^{2, 3など}。しかし、気象観測露場は、各市町村に存在しているわけではない。さらに、設置場所は市街地や畑地など多様であり、観測データが圃場の気象を正確に反映していない場合があることが指摘されている。気象データの正確性・解像度は、収量予測の精度に強く関与する要素であり、輸送計画を立てる上でも重要なインフラとなる。

本研究では、予測に用いる気象データとして、農研機構1kmメッシュ農業気象データシステム(The Agro-Meteorological Grid Square Data System, NARO) のものを利用した⁴⁾ (<https://amu.rd.naro>).

go.jp/wiki_open/）。このシステムでは、アメダスや気象官署の気象データを使用し、地形などを考慮して補正されたデータが約1キロメートル四方のメッシュに整備されている。日平均気温と降水量は26日先まで、全天日射量は9日先までの予報値が提供される。未来の収量を予測したい場合、気象データを平年値にするか、予報値を含むデータを自ら作成するかする必要はあるが、このシステムを利用することで、実測値と予測値がシームレス（複数のものが一つのものであるかのように利用できる技術や状態）につながった高解像度なデータを簡便に予測に用いることができる。

2 機械学習と高解像度気象データを利用した収量予測モデル

(1) モデルとデータの概要

他の作物と同様に、播種形態などの栽培条件や、掘り取り調査結果などの生育情報、気温や降水量などの気象情報を基に、各要素と収量の相関分析や収量予測を行う試みは、てん菜においても既に^{2、3}行われている^{2、3}。一方で、夏季の高温多湿傾向や高温の長期化をはじめとした気候変動の影響、栽培形態の変化など、てん菜の収量に影響を与える要因やその大きさは年々変化していると考えられる。収量予測において、予測式の説明変数を適切に選択することは重要であるが、変数選択には専門的知識が必要となることが多い。そこで本研究では、統計的手法と機械学習を組み合わせ、データの学習から収量の予測までを半自動的に行う「テンサイ収量予測Webアプリ」を開発した（図1）⁵⁾。また、データの学習の結果、収量と相関が高いとされた変数をアプリに表示することで、予測結果の変動要因を客観化し、経験知の見える化を試みた。

図1 「テンサイ収量予測Webアプリ」

資料：国立研究開発法人 農業食品産業技術総合研究機構

開発したアプリで予測を行うには、予測式を作るために必要な学習データと、そこから作られた予測式の精度検証に用いる検証データが必要となる。こ

れらのデータは、Excel形式のテンプレートを使って作成することで簡単にアプリに取り込むことができる。作成するデータのテンプレートの基本項目と

して、調査地点、栽培方法（移植あるいは直播^{ちよくはん}）、8月20日前後の掘り取り調査結果（1株当たりの根重）、調査地点別収量（トン/ヘクタール）が用意されているほか、任意の項目を追加することができる（モデル式を構築するアルゴリズムは後述する）。これらのデータを予測対象年から過去5年分用意する。

本稿では日本甜菜製糖株式会社、ホクレン農業協同組合連合会、北海道糖業株式会社（五十音順）から提供されたデータを使用し、開発したアプリの実用性を検証した結果を紹介する。これらのデータは、十勝地方の2カ所の工場、およびオホーツク地方の3カ所の工場で収集された。提供されたデータのうち、検証データには2017～22年のデータを、学習データには予測対象の年から過去5年のデータを使用した。例えば、22年のデータを予測する場合は、17～21年のデータを学習データとして使用した。データ項目は、調査地点別収量（1ヘクタール当たりの根重）、8月20日の掘り取り調査結果（1株当たりの根重）、栽培方法（移植あるいは直播）、調査地点（住所あるいは緯度・経度）である。このデータに対応する気象データとして、各調査地点に対応する気温、降水量、全天日射量の農研機構1kmメッシュ農業気象データを利用した。この気象データは、農研機構1kmメッシュ農業気象データの利用登録後、アプリ操作画面の「農研機構1kmメッシュ農業気象データを利用する」にチェックを入れることで、学習データおよび検証データの調査地点情報を基に自動でダウンロードされる。また、自身が用意した気象観測データを利用してアプリを使用することも可能である。

（2）リッジ回帰についての解説—重回帰との違い

このアプリでは、モデル作成の際にリッジ回帰モデルを使用した。リッジ回帰モデルは線形回帰モデ

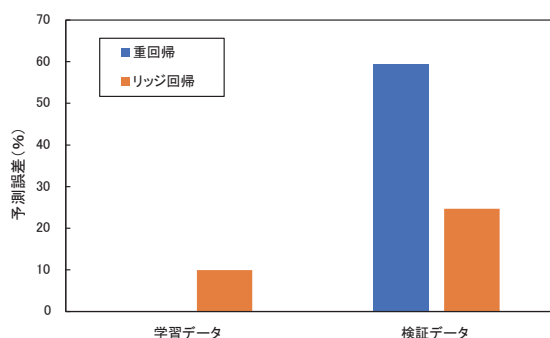
ル（目的変数〈分析の対象となる要素〉と説明変数〈予測値に影響を与えている要素〉の関係を線形に表したモデル）の一つであり、重回帰モデルにL2正則化項（係数が極端に大きくなることを防ぐ数式）を備えることで、過学習を抑制することができる。重回帰モデル、リッジ回帰モデル、過学習について、以下にデータ例を交えながら解説を行う。

複数の説明変数とその影響力を掛けたものを足し合わせることで、目的変数を表現するモデルを重回帰モデルと呼ぶ。重回帰モデルは以下のように表すことができる。

$$\text{目的変数} = \text{切片} + \text{係数1} \times \text{要素1} + \text{係数2} \times \text{要素2} + \dots + \text{係数}p \times \text{要素}p + \text{誤差}$$

重回帰モデルにL2正則化項を加え、1つの係数が極端に大きくなることを防いだものをリッジ回帰モデルと呼ぶ。モデルの係数が極端に大きいとき、そのモデルは学習データに対して過剰に適合し、未知のデータへの当てはまりが悪化する過学習を起こしている可能性がある。過学習は学習データの数が少なく、モデルに影響を与える変数が多い場合や、データのノイズが多い場合に起こりやすい。過学習が起きやすいダミーデータを用いて、一般的な重回帰でモデル化を行った場合と、リッジ回帰でモデル化を行った場合の解析例を図2に示す。

図2 重回帰とリッジ回帰における過学習のしやすさの違い



資料：筆者作成

解析例に用いたダミーデータには、影響力の大きい主要な説明変数の他に、ノイズとなるデータが多数含まれている。このデータを学習データと検証データに分け、重回帰とリッジ回帰でモデル化を行ったのち、予測精度の検証を行った。その結果、重回帰モデルは学習データでの予測誤差が 6.88×10^{-14} （%）、検証データでの予測誤差が59.39%であった。学習データの予測誤差は非常に小さいが、検証データの予測誤差は極端に大きいことから、このモデルは過学習を起こしており、予測対象である検証データに適応できていないことが分かる。一方で、リッジ回帰モデルは学習データの予測誤差が9.95%、検証データの予測誤差が24.69%だった。学習データの予測誤差は重回帰モデルより大きいものの、検証データの予測誤差は重回帰モデルよりも小さかった。また、誤差間の差も極端ではないことから、このモデルは予測対象である検証データに適応できていると言える。

本稿で紹介するモデルは、学習データの期間が予測対象の年から過去5年と限定される点や、モデルに影響を与える変数が多い可能性がある点、データのノイズ除去が現実的ではない点などから、リッジ回帰モデルによる予測式を採用した。

(3) 収量予測のアルゴリズム

この予測モデルでは、学習データの調査地点別の収量と調査地点別収量の平均値の差が目的変数として定義されている。てん菜収量予測Webアプリでは、1) 目的変数と相関の高い気象の要素と時期を特定し、説明変数の候補とする、2) 共線性の高い変数（予測値に不安定な影響を及ぼす要素）を排除する、3) ステップワイズ法に基づいて目的変数に対し相関が高い説明変数を自動で選択する、4) 選択された説明変数とリッジ回帰を使用してモデル式を作る、5) モデル式に検証データを当てはめて予測収量を算出するという流れで予測が行われる。

まず、1)の詳細から説明する。気温・全天日射量・降水量の3要素について、目的変数と最も相関の高い時期を相関分析によって特定する。分析対象は、4月下旬から10月中旬までの連続する旬別平均値である。例えば、4月下旬を起点とした平均気温を計算する際は、4月下旬、4月下旬～5月上旬、4月下旬～5月中旬と期間を徐々に延ばし、4月下旬～10月中旬までを計算した後は、5月上旬を起点とした計算を同様に繰り返すことで、171通りの気象値を気温・降水量・全天日射量の3要素ごとに得る（表1）。

表1 気温・降水量・全天日射量の旬別値の計算の概略図

	5月			6月			...	10月		
	上旬	中旬	下旬	上旬	中旬	下旬		上旬	中旬	下旬
1巡目	*						...			
2巡目	*	*					...			
3巡目	*	*	*				...			
⋮	⋮						...			
18巡目	*	*	*	*	*	*	...	*	*	*
19巡目		*								
20巡目		*	*							
21巡目		*	*	*						
⋮	⋮						...			
35巡目		*	*	*	*	*	...	*	*	*
⋮	⋮									
171巡目										*

資料：国立研究開発法人 農業食品産業技術総合研究機構
 注：収量の平年差と相関の高い気象の期間を特定する際の、旬別値の概略図。5月上旬から10月下旬にかけて、*で示される連続する旬別値と収量の平均差の相関を調査し、最も寄与率の高い期間を特定する。旬別値は期間中の平均値である。気温、全天日射量、降水量について、上記の処理を行う。

このうち、目的変数との相関が最も高いものを説明変数の候補とする。

次に2)で示した共線性について説明する。説明変数の候補として、同じ期間の気温と全天日射量が選択されたとき、変数の性質が似ている（この例の場合、気温が高い日は晴れていることが多い）ために、変数同士の相関が高くなることがある。これを多重共線性と呼ぶ。多重共線性は予測を不安定にし、変数の解釈を難しくする。そのため、このアプリでは変数間の相関を分析し、共線性が高い場合は目的変数との相関がより高い方が説明変数の候補として選択される。

3)のステップでは、選択された気象の変数のほか、学習データの調査項目（栽培条件など）を説明変数の候補として扱い、AIC (Akaike's Information Criterion, 赤池情報量基準)⁶⁾に基づいたステップワイズ法で説明変数の選択が行われる。ステップワイズ法とは、説明変数の候補について全通りの組み合わせでモデル式を構築し、目的変数の予測に最も適した説明変数の組み合わせを探索する手法である。AICは、ステップワイズ法において、モデルの複雑さと当てはまりの良さのバランスを取るための統計的な指標である。

こうして選択された説明変数を使用し、4)のステップとしてリッジ回帰モデルの予測式が作成される。そして5)のステップで、検証データ（予測値を知りたいデータ）がアプリで読み込まれ、作成されたモデル式を使って予測値が計算される。

（４）精度検証に使用した気象データ

気象データについては、農研機構1kmメッシュ農業気象データの「過去データ再現キット」(https://amu.rd.naro.go.jp/wiki_open/doku.php?id=archive)を使用することで過去の予報値を再現し、予測対象の年の8月20日に予測を行ったものとして予測精度の検証を行った。例えば2018年のデー

タを予測する場合は、2018年8月20日までの気象データは実測値、8月21日以降は予報値および平年値を用いた。平均気温と降水量に関しては26日先まで、全天日射量に関しては9日先までを予報値、それ以降の気象値は平年値を用いた。

（５）アプリの実行結果

アプリの利用マニュアルに従ってデータを作成したのち、読み込むデータのファイル名をアプリに入力し、年次や日付を指定後にアプリを実行することで、「予測結果 (t/ha)」と「モデルに選択された説明変数」がアプリの対応する欄に表示される。「予測結果 (t/ha)」で示されるのは調査地点別の予測収量の平均値であるが、工場別の予測値など、用途によって表示内容を変更することができる（詳しくは後述する）。また、「モデルに選択された説明変数」欄には、モデルの説明変数が目的変数と相関の高い順に表示される。

（６）予測精度の計算方法

収量予測の推定精度を評価するため、以下の式で予測誤差を計算した。

$$\text{パーセント誤差} = \frac{\text{予測収量} - \text{実際の収量}}{\text{実際の収量}} \times 100 (\%)$$

また、地域や工場ごとの予測誤差の傾向を確認するため、計算した予測誤差を地域別・工場別に平均し、その地域・工場の予測誤差の平均値とした。

3 てん菜収量予測アプリの予測精度

（１）予測精度について

製糖会社が収集したデータと、農研機構1kmメッシュ農業気象データを使い、アプリを使っただん菜の収量予測を行った。収穫のおよそ2カ月前に収量予測を行った場合の予測精度を確認するため、検

証データは8月20日までに入手可能なデータのみを用いた。例えば、8月21日以降の気象データには予報値と平年値を使用した。その結果、2017～22年の十勝地方における調査地点別収量の予測誤差は6.35～10.20%、平均8.07%だった。また、同様の年次におけるオホーツク地方の予測誤差は2.72～20.50%、平均6.77%だった（表2）。

表2 十勝地方およびオホーツク地方の予測誤差

年次	十勝地方 (%)	オホーツク地方 (%)
2017	6.35	20.50
2018	8.08	5.52
2019	8.54	5.79
2020	7.04	2.72
2021	6.96	10.91
2022	10.20	5.28
全期間平均	8.07	6.77

資料：金谷ら（2024）より引用（引用文献5を参照）

予測誤差が最も大きかったのは2017年のオホーツク地方であり、予測誤差は20.50%だった。特に誤差の大きかった佐呂間町と置戸町は、てん菜褐斑病の市町村別巡回データ⁷⁾から、発病株率がオホーツク地方の平均より高かったことが判明している。てん菜褐斑病は糖度を低下させ、茎葉が枯死すると根重にも影響が及ぶ⁸⁾。一方で、本稿の予測に使用した学習データには、病害に直接関連するデータが含まれていない。これらのことから、データの学習の過程で褐斑病の被害による減収をモデルに反映できなかったために、2017年のオホーツク地方の予測誤差は他の年次より大きかったと考えられた。また、佐呂間町・置戸町と同様に予測誤差が大きかった湧別町は、品種の違いや水はけなどの土壌に関する要因によって予測誤差が大きくなったと考えられた。これらのことから、発病株率など病害の発生状況を表すデータや、飽和透水係数などの土壌

環境に関するデータを学習データに利用することで、収量の予測精度が向上する可能性が示唆された。

（2）説明変数を提示する機能について

今回開発したてん菜収量予測Webアプリは、目的変数と相関が高い順に説明変数を提示する機能がある。提示された説明変数は、地域ごとに異なるものと共通するものがあった。地域ごとに異なっていたのは気象に関する変数であり、十勝地域では7月下旬～9月中旬の降水量が選択されることが多く、オホーツク地域では8月下旬～10月上旬の全天日射量が選択されることが多かった。このとき、降水量と目的変数の相関係数は $-0.62 \sim -0.95$ 、全天日射量と目的変数の相関係数は $0.54 \sim 0.77$ で、どちらも1%水準で有意だった。十勝地方はオホーツク地方よりも夏季の降水量が多い傾向があることや、秋等の好天は根重を増大させると考えられる⁹⁾ことから、提示された説明変数はおおむね過去の研究結果と一致していると言える。また、この二つの変数は、気象値の期間の一部（8月下旬～9月中旬）が重複していたため、両方の変数が説明変数の候補として選択されていても、共線性を排除する過程で一方の説明変数のみが選択された可能性がある。

両地域に共通していた変数は、栽培条件（移植あるいは直播）や8月20日の掘り取り調査の結果であった。栽培条件に関しては、予備解析の結果から、移植栽培と直播栽培を分けて別々に予測を行った場合より、2種類の栽培方法を変数で分けて一緒に予測を行った場合の予測誤差が小さいことが分かっている。また、掘り取り調査の結果は目的変数との相関が高く、掘り取り調査の結果がデータに含まれない場合は予測精度が大幅に低下した。これらの生育調査データが予測精度に大きな影響を与えることから、製糖会社が実施している生育調査は的確かつ信頼性が高く、高精度・高解像度の収量予測を実現する上で非常に重要であると考えられる。

おわりに

本稿では、筆者が開発したてん菜収量予測Webアプリを使用し、製糖会社の収集データを例にその予測精度の検証結果を解説した。このアプリは、データの学習から収量の予測までを半自動的に行うことに加え、収量と相関が高いとされた変数をアプリに表示することで、予測結果の変動要因を客観化することができる。先行研究¹⁰⁾で言及されている通り、解像度の高い収量予測を行うためには、その解像度の高さに比例してデータ収集の労力が大きくなるというジレンマがある。本稿の予測結果から、既存のシステムや収集データ、公的機関のオープンデータを利用することで、過度に労力を増やすことなく高解像度の収量予測を行うことができる可能性が示唆された。また、病害や土壌に関するデータを学習データに加えることで、予測精度が向上する可能性が示されたことから、これらのデータを利用したアプリのバージョンアップと精度の検証を現在実施している。

高解像度気象データを利用した本研究の収量予測手法は、今後さらに予測精度を向上させることで、てん菜の輸送計画の精緻化に活用できる可能性がある。また、収量予測結果として、アプリ画面には調査地点別収量の平均値が表示されるが、輸送計画等を立てる場合は工場別の収量予測値が必要となる。本研究では、工場ごとに調査地点別収量を工場別収量に変換する式を作成し、より実用的なツールとなるよう継続して開発を行っている。予測精度の向上と実用化を見据えた改善を通して、未来の砂糖生産に寄与できる技術を開発していきたい。

謝辞

本稿で紹介した研究成果は、NTT東日本受託研究「小麦および糖業生産システムモデル構築に向けた検証業務」の一環として行われた。加えて、JSPS科研費JP22K05858の助成を受けた。また、データを提供していただいた日本甜菜製糖株式会社、ホクレン農業協同組合連合会、北海道糖業株式会社（五十音順）に感謝申し上げます。

【引用文献】

- 1) 北海道農政部生産振興局農産振興課 (2014-2024)「てん菜生産実績」北海道オープンデータ CC-BY4.0 <<https://creativecommons.org/licenses/by/4.0/deed.ja>>
- 2) 梶山努 (2011)「てんさい」『北海道立総合研究機構農業試験場資料』39、pp. 32-39. 地方独立行政法人北海道立総合研究機構農業研究本部中央農業試験場
- 3) 柳沢朗 (2018)「十勝地方における畑作物の生産性向上と収量変動要因について」『砂糖類・でん粉情報』(2018年4月号)、pp. 46-54. 独立行政法人農畜産業振興機構
- 4) 大野宏之、佐々木華織、大原源二、中園江 (2016)「実況値と数値予報、平年値を組み合わせたメッシュ気温・降水量データの作成」『生物と気象』16、pp. 71-79. 日本農業気象学会
- 5) 金谷真希、森岡涼子、石郷岡康史、辻博之 (2024)「栽培および気象データを用いた機械学習によるテンサイ収量予測モデルの構築」『てん菜研究会報』65、pp. 1-7. 甘味資源振興会
- 6) Akaike H (1973)「Information theory and an extension of the maximum likelihood principle」『Proceedings of the 2nd International Symposium on Information Theory』、pp. 267-281.
- 7) 北海道農政部、北海道病害虫防除所、北海道立総合研究機構農業研究本部 (2017-2022) 農作物有害動植物発生予察事業年報、北海道病害虫防除所・地方独立行政法人北海道立総合研究機構中央農業試験場
- 8) 栢森美如 (2018)「テンサイ褐斑病―近年の薬剤耐性菌事情」『てん菜研究会報』59、pp. 18-24. 甘味資源振興会
- 9) 梶山努 (2011)「てんさい」『北海道立総合研究機構農業試験場資料』39、pp. 32-39. 地方独立行政法人北海道立総合研究機構農業研究本部中央農業試験場
- 10) 村上貴一 (2022)「高解像度気象データと病虫害発生情報に基づくてん菜収量の早期予測技術の開発」『砂糖類・でん粉情報』(2022年6月号)、pp. 57-63. 独立行政法人農畜産業振興機構