研　究　報　告　書

「膨大なレガシー栽培データを蘇生する（データさきがけ）」

研究タイプ：通常型
研究期間：　2015 年 10 月～2019 年 3 月
研　究　者：　小野木　章雄

1．　研究のねらい

現在人口増加や気候変動により世界的に食料問題が深刻化しており、食料増産や生産の効率化及び安定化が危急の課題となっている。この課題解決のために、気象、圃場環境、栽培管理及びゲノムに関するビッグデータを用いたデータセントリックな品種改良・栽培管理法に大きな期待が寄せられている。特に全ゲノム情報を用いたゲノミックセレクション法（genomic selection、注）は新たな品種改良手法として導入が期待される。しかしこれら技術の実現には大きな障壁がある。それは品種改良や栽培管理の直接のターゲットとなる作物の性質、つまり収量や形態、収穫期、耐病性などに関する栽培データが不足していることである。栽培データが他のデータに比べ不足する要因は、そのデータ量が年（時間）により制約を受けるためである。作物の栽培は基本的に年に 1 回しか行うことができず、さらに栽培結果は同じ場所であっても年により変動するため、ある品種がその場所で平均的にどのような性質を示すかを知るためには多年に渡る栽培が必要となる。

一方で、近代的な品種改良は数十年にわたり全国の公設試験場などで様々な作物について行われており、その過程で膨大な栽培データが蓄積されている。例えばダイズ品種改良の栽培試験は一部試験場では戦前から開始され、戦後始まった国主導の事業では全国数か所にある主要育成地とほぼ全国の地方公設試験場で行われてきた。しかしこの過程で得られた膨大な栽培データは各育成地や試験地に死蔵されており、レガシーデータとなっていた。

そこで本課題では、こういった過去のダイズ品種改良の過程で収集され、その後死蔵されている栽培データを「蘇生」し、データセントリックな品種改良・栽培管理法を迅速に構築することを目標とした。ここでいう「蘇生」とは栽培データの収集及び整理と、それを活用するために必要な情報、つまりゲノムや気象データ、との関連付け、さらに活用に向けた統計学的手法や枠組みを提供することを指す。この研究により、作物の増産や生産の高効率化が迅速となり、日本及び世界の食糧問題解決に向けた大きな貢献となることが期待される。

（注）ゲノムデータと作物の性質を統計学的或いは機械学習的手法で関連付け、ゲノムデータのみから作物の性質を予測し、栽培することなく優良品種を選抜する手法。品種改良の高速化や高効率化への貢献が期待される。

2．　研究成果

（1）概要

栽培データは主要ダイス育成地、すなわち農業・食品産業技術総合研究機構（農研機構）の4センター（東北農業研究センター、次世代作物開発研究センター、西日本農業研究センター、九州沖縄農業研究センター）、北海道立総合研究機構の2試験場（中央農業試験場、十

勝農業試験場）、及び長野県野菜花き試験場を中心に収集した。アメダスにおける気象観測が 1961 年から開始されているため収集の対象を 1961 年以降とした。品種改良における栽培試験では、これら育成地から全国の試験地へ種子が配布され試験されるため、実際にデータに含まれる試験地は 460 箇所以上になる。収集した栽培データは主に以下の3種類のデータから成る。（1）ある品種をある年ある試験地で栽培した結果示した性質を記録したデータ（以下試験結果と呼ぶ）、（2）施肥や播種期など栽培条件を記録したデータ（以下耕種概要）、及び（3）試験時の生育状況を記録したデータ（以下生育概要）である。これら栽培データはデジタル化・統合を経て、農研機構内のサーバーにデータベースとして格納した。

さらに収集した栽培データを気象データと関連付けるために、過去約50年にわたる全国の公設試験場の沿革を調査して試験場所を特定し、最寄りのアメダス観測所との関連付けを行った。またゲノムデータについては育成地やジーンバンクから 2,208 品種（系統）について種子提供を受け、DNA 抽出の後ゲノムデータを取得した。

全国規模かつ多年にわたるヒストリカルな栽培データと、気象及びゲノムデータをどのように活用するかという問題は必ずしも明らかではなく、国際的にも同様の取り組みは乏しい。そのため本課題ではまず気象及びゲノムデータの利用について、双方を同時に利用し作物の生育を高精度に予測する統計学的手法の提案とパッケージの開発を行った。またヒストリカル栽培データは不均衡な多変量データであるが、有用遺伝子の検出手法とゲノムから性質を予測する統計学的手法について不均衡多変量データにおける性能の検討を行った。さらに後者の知見に基づきゲノミックセレクション法で用いる統計学的予測モデルの構築とそのためのフローを開発し、ダイズだけでなく他の作物に対しても適用できるような枠組みを構築した。

（2）詳細

【データ収集と整理】

栽培データはどの育成地においても概ね 2000 年前後までは試験成績書と呼ばれる紙媒体で、その後はエクセルファイルなどのデジタル媒体で保存されていた。そのためまず専門業者委託による紙媒体のデジタル化を行った。続いてデータのフォーマットが育成地により様々であり時代による変遷もあること、さらに例えば収量であれば、「アールあたりの子実重 kg」や「子実重 kg/a」など、性質の名称も一貫性がなかったために、フォーマット及び性質名の統一を行った。性質名については性質名に加え、測定時期、測定単位をアンダーバーで結合する簡易的なオントロジーを用いて統一した。以上の作業は栽培データを構成する3つのデータ（試験結果、耕種概要、生育概要）について行い、それぞれをテーブルとして構成した後、これらテーブル間での関連付けを行った。現在確定しているテーブルサイズは 83,886 行×606 列（試験結果）、10,376 行×95 列（耕種概要）、5987 行×13 列（生育概要）となる。以上のデータを格納するために農研機構高度解析センターにデータベース「作物育種データを管理するDBとWebシステム」を構築し、現在ダイズ品種改良関係者に順次公開している。以上の取り組み及び成果は研究成果（3）の1）及び2）で報告した。

【気象及びゲノムデータとの関連付け】

公設試験場は移転することがあり、同じ名前でも過去と現在で所在地が異なることも珍しくな

い。また所在地が同じでも組織改編により名称が変更することも多い。所在地の明確化はその後の統計解析や気象データとの関連付けにおいて重要であることから、沖縄県を除いた関係する 46 都道府県の公設試験場についてその沿革を調査し、試験場の所在地の変遷を可能な限り特定した。これにより各試験場について最寄りのアメダスから気象データを得ることや、公開されている気象及び土壌データベースから情報を得ることも可能となった。ゲノム情報については 2,208 品種（系統）の種子を発芽させ本葉から DNA を抽出した。ゲノムデータは全ゲノムシークエンス法、アレイ（Thermo Fisher Scientific 社）法、または RAD-seq 法により得た。これらの手法は得られるゲノムデータの量に差があり、全ゲノムシークエンス法では全塩基配列（約 1.1G 塩基）、アレイ法では約 150K 塩基、RAD-seq 法では約 1K 塩基分のデータが得られる。これらの手法には品種の重要度に合わせてそれぞれ 9、 575、及び 2,208 品種（全品種）を供した。これらゲノムデータは試験結果と関連付けることが可能であり、収集した栽培データを用いたゲノミックセレクション法や有用遺伝子の検出などに利用することが可能となった。

【データ活用手法に関する研究】
ヒストリカル栽培データと気象及びゲノムデータを今後の農業や品種改良にどのように役立てればよいかという問題は未だ道筋がついておらず、国際的に見ても取り組みは乏しい。ここでは特に品種改良への利用で重要となる以下の４点について研究を行った。

1）気象及びゲノムデータ双方を考慮した作物生育予測手法
安定した収量及び品質を得るには、その土地の気候に作物が適応していることが必須であり、例えば開花や成熟がその気候の下での望ましい時期に起こる必要がある。また管理面から播種や収穫時期が限定されていることも多く、任意の気候の下でどのような生育をするかデザイン可能であると品種改良が効率化される。本課題担当者は以前に植物生理学的生育モデルとゲノム情報を組み合わせることで、高精度に開花予測が可能なことをイネにおいて例証したが（文献参照）、同様のモデルをイネの開花だけではなくダイズを含めた様々な作物の生育において構築可能な汎用 R パッケージ「GenomeBasedModel」を開発した（研究成果（3）の3））。このパッケージではユーザーが任意の生育モデルとゲノム情報を引数として与えると、その生育モデルのパラメータとゲノム情報を自動的に関連付け、任意の気候の下での生育をベイズ統計の枠組みで予測する。同パッケージは現在 Github（https://github.com/Onogi/GenomeBasedModel）で公開中である。

2）多変量混合効果モデルに基づくゲノムワイド関連解析の有効性について
ゲノムワイド関連解析は広大なゲノム領域から標的の性質を調節する遺伝子を検出する解析を指すが、混合効果モデルに基づく検定は現在その解析の中心的手法である。一方ヒストリカル栽培データは性質や試験地などの面で多変量データであるため、通常用いられる単変量モデルではなく、多変量モデルを用いるという選択肢も考えられる。しかしながらゲノムワイド関連解析において単変量と多変量モデルの優劣は明らかではなかった。そのためシミュレーションを用いて様々な条件下で比較を行い、3つのパラメータ、つまり変量間の相関、遺伝子効果の相対的な大きさ、データ欠損率、によりその優劣が変化することを見出した（図1及

び研究成果（3）の4））。

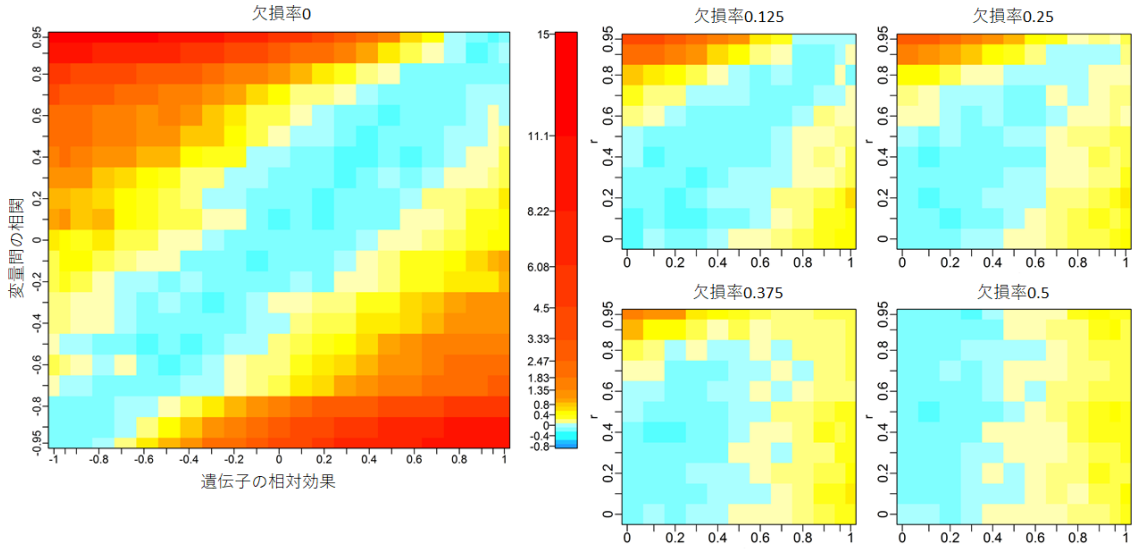

図1　多変量（変量数は2）と単変量モデルでの検定で得られた－log10P 値の差。赤は多変量で青は単変量で検出力が高かったことを示す。縦軸が変量間の相関、横軸が遺伝子効果の相対的な大きさ。欠損率が 0 の場合は変量間相関と遺伝子相対効果は－1 から 1 の値を、欠損率が 0 以上の場合は 0 から 1 の値の結果を図示。

3）多変量混合効果モデルに基づくゲノミックセレクション法の有効性について

混合効果モデルはゲノムから性質を予測するゲノミックセレクション法においても中心的な手法であるが、2）のゲノムワイド関連解析同様、多変量と単変量モデルの予測能力の優劣については十分に明らかにされていなかった。そのためシミュレーションにより両者の比較を詳細に行い、遺伝的要因により生じる相関（遺伝相関）と、それ以外の要因により生じる相関（残差相関）の差が変量間で大きいほど多変量混合効果モデルが優れていることを見出した（図2）。

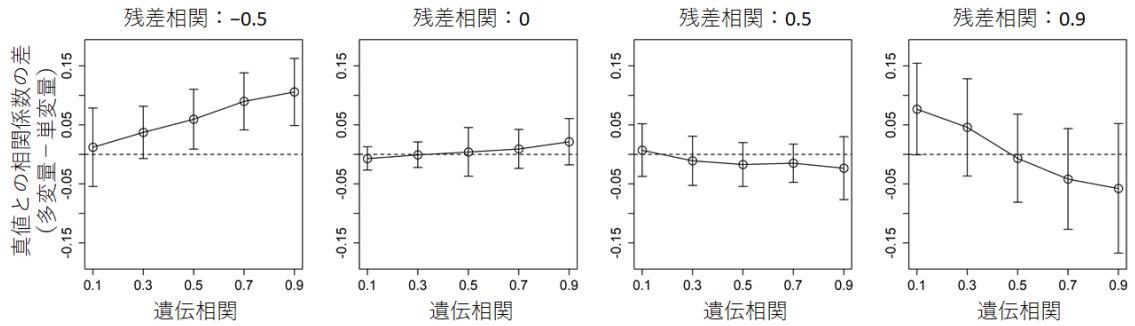

図2　多変量（変量数は2）と単変量による予測の差。予測能力は真値とのピアソン相関係数で測定し、多変量から単変量による相関係数を引いた値を図示した。100 回のシミュレーションの平均値（バーは標準偏差）。

4）ヒストリカル栽培データからのゲノミックセレクション法のための統計学的予測モデル構築

前述の3）からヒストリカル栽培データを多変量としてモデリングすることは必ずしも良い予測

をもたらさないことが明らかとなった。そのため各試験地の各性質を単変量として扱うモデルを基本とし、3）で明らかとなった条件に合致するものだけ統合し多変量として扱う方法が適していることが示唆された。試験地によっては様々な栽培条件（施肥、栽植密度、播種期など）が試みられているため、栽培条件と遺伝の相互作用を考慮しながら主要 15 地域の 18 試験地において予測モデルの構築を行った（図3及び研究成果（3）の5））。図3の例から、同じ品種であっても地域及び播種日により期待される収量が異なることがわかる。これらモデルを品種候補の選抜や交配計画策定に用いることにより今後各試験地での品種改良が高効率化できると考えられる。さらに一連の予測モデル構築を他の作物に対しても適用可能なようにモデル構築フローを開発した（研究成果（3）の5））。

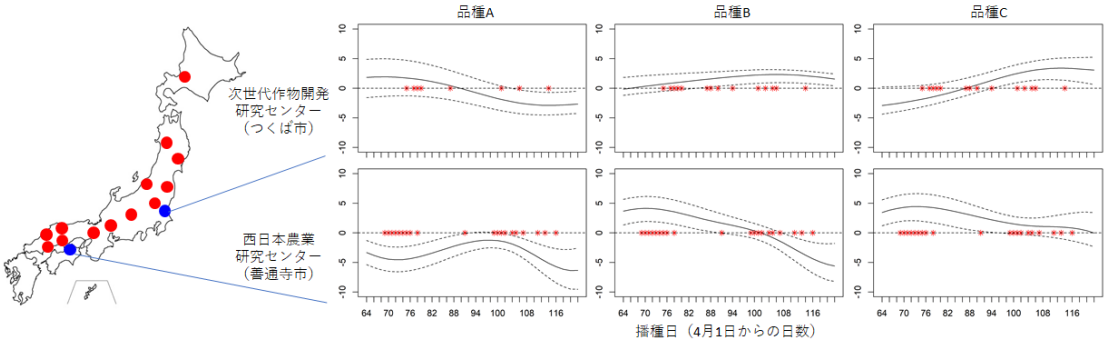

図3　ゲノミックセレクション法のモデルを構築した 15 地域と解析例。例は農研機構次世代作物開発研究センター及び西日本農業研究センターの結果であり、代表的な3品種（A から C）について収量（子実重）の結果を示した。遺伝と播種日の相互作用をスプライン関数で表現しており、縦軸は期待される収量の偏差、横軸は播種日（4 月 1 日からの日数）を示す。図中の赤いアスタリスクは栽培データが存在する点、実線は事後平均、破線は 50% highest density interval を示す。

以上1）から4）の成果は、今回収集したダイズのヒストリカル栽培データのみならず、今後収集が進むと考えられる他作物の栽培データについても適用可能であり、今後のヒストリカル栽培データを利用した品種改良においての重要な貢献であると考える。

文献）Onogi et al. (2016) Theor. Appl. Genet., 129: 805–817

3．　今後の展開

【データ収集及びデータベースの拡張】

本課題ではヒストリカル栽培データを格納するデータベースを構築したが、品種改良のための栽培試験は毎年行われているため、データは今後も蓄積していくことになる。そのためそれらデータを再び死蔵データにしないために、定期的にデータベースへアップロードしていく仕組みが必須となる。そこで短期的（今後1～2年間）には新たなデータをデータベース管理者である本課題担当者が集積し一括アップロードする仕組みを、より長期的（今後2～3年以降）にはダイズ品種改良担当者らが直接栽培データをアップロードする仕組みを構築する予定である。また当該データベースは検索・表示・ダウンロードが主な機能となるが、それをさらに拡張し、図

表によるサマリーなどの出力や簡単な統計処理、または気象や土壌データベースとの連携なども可能なように拡張していく予定である（今後3年以内を予定）。

【ゲノムデータとヒストリカル栽培データを用いた品種改良手法の実装】
ゲノミックセレクション法の予測モデルをより広域的に多くの試験地において構築し、国内全体をカバーできるように拡大させる予定である。また多変量化を含めモデルのブラッシュアップを行う。いずれも今後１～２年程度の短期的な目標と考えている。またさらにモデルの予測能力を実証するために、実際に栽培試験を行う。これから３～５年程度かけモデルのブラッシュアップと実証試験を繰り返し、ゲノムとデータ解析に基づくよりデータセントリックな作物品種改良手法を実装していきたい。

４．　自己評価

本課題の柱であるレガシー化していた栽培データの蘇生については、主たるダイズ育成地からこの分野では世界的にも稀な規模のデータを収集し、データベース構築まで行えた点で目的を果たしたと考える。また合わせてゲノムデータの取得も行ったことにより、栽培データの利用価値が格段に向上し、ゲノミックセレクション法に代表されるデータセントリックな品種改良手法近の社会実装へ大きな貢献ができたと言える。一方で情報・数理系分野など農業分野以外の研究者との連携は希薄であったこと、論文発表が遅れたことは反省すべき点であり今後の課題としたい。

５．　主な研究成果リスト

（1）論文（原著論文）発表

該当成果なし

（2）特許出願

該当成果なし

（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）

1）小野木（2017）育種と統計とデータさきがけ，オペレーションズ・リサーチ，62: 233–238

2）小野木（2017）ダイズ育種におけるヒストリカルデータ，日本育種学会第 132 回講演会ワークショップ「ダイズにおける育種の新しい流れ」，岩手大学

3）小野木（2018）GenomeBasedModel：任意の生物学的統計モデルのパラメータとそのパラメータに対するゲノムワイド SNP の効果を同時推定するための R パッケージ，日本育種学会第 134 回講演会口頭発表，岡山大学

4）Onogi (2019) Comparison of F-tests for Univariate and Multivariate Mixed-Effect Models in Genome-Wide Association Mapping. Front. Genet. 10:30. doi: 10.3389/fgene.2019.00030

5）小野木（2019）全国規模のヒストリカルデータをいかにゲノミックセレクションへ利用するか：ダイズにおける研究例，日本育種学会第 135 回講演会口頭発表，千葉大学