Vol.2021-IS-155 No.7
2021/3/8

植物工場の果菜類を対象とした収量予測モデル生成プロセスの開発

外舘有希[†1]　大場みち子[†2]　高森満[†3]

概要：トマトやキュウリ等の果菜類は，植物工場の主要な栽培品目であるため，収量予測に基づいた適切な生産・販売計画が求められる．しかし，現在，植物工場の果菜類の収量予測は，気象や人為的な影響を受けることから困難とされている．そこで，本研究の目的は，太陽光型植物工場の果菜類の収量に関わる要因を分析し，収量予測モデルを開発することである．目的を達成するために，(1)モデル生成用データの選定，(2)データの前処理，(3)データの可視化，(4)特徴量の設計，(5)予測手法の選定，(6)モデルの適正化の 6 つのサブプロセスから成る収量予測モデル生成プロセスを開発した．予測モデルの構築には，施設内外の環境データと過去の収量の実績データを利用し，重回帰分析，一般化加法モデル，多変量適応的回帰スプラインを用いた．本稿では，提案手法の有効性を確かめるために，開発した予測モデル生成プロセスをミニきゅうりとミニトマトに適用して実験した結果について報告する．

キーワード：収量予測，植物工場，果菜類，統計モデリング, tsfresh パッケージ

Development of Yield Prediction Model Generation Process for Fruit Vegetables in Plant Factories

YUKI TODATE[†1]　MICHIKO OBA[†2]　MITSURU TAKAMORI[†3]

Abstract: Fruit type vegetables such as tomatoes and cucumbers are major crops grown in plant factories, so appropriate production and sales plans based on yield prediction are required. However, yield prediction of fruit type vegetables in plant factories is currently difficult because of weather and human influences. Therefore, the purpose of this research is to analyze the factors related to the yield of fruit type vegetables in sunlight-based plant factories and to develop a yield prediction model. To achieve the purpose, we developed a yield prediction model generation process consisting of six sub-processes: (1) data selection for model generation, (2) data preprocessing, (3) data visualization, (4) feature design, (5) selection of prediction methods, and (6) model optimization. We used environmental data inside and outside the facility and actual data of past yields as features, and also used Multiple linear regression, generalized additive model, and multivariate adaptive regression spline to construct the prediction model. In this paper, we report the experimental results of applying the developed prediction model generation process to mini cucumbers and mini tomatoes to verify the effectiveness of the proposed method.

Keywords: Yield Prediction, Plant Factory, Fruit Type Vegetables, Statistical Modeling, tsfresh Package

1. はじめに

農林水産省は，野菜・果樹・花きといった園芸作物を食糧の支出金額に占める割合が最も高く，消費生活上重要な品目であると位置付けている[1]．それゆえ，消費者ニーズに応え安定供給するためには，施設内で植物の光や養分等の生育環境を制御して栽培を行う施設園芸による周年供給が必須であるとしている[1]．そのような背景の中，施設園芸の最先端の形態である植物工場が近年大きな関心と期待を集めている．

植物工場とは，「高度な環境制御と生育予測を行うことにより，野菜等の植物の周年・計画生産が可能な施設園芸農業の一形態」である[2]． 植物工場は，従来のビニールハウスや加温設備等の単一の環境のみを制御可能な温室と比較して，より高度な環境制御が可能であるとされている．現在，国内の施設面積に占める植物工場の割合は 2.8%[1]とまだ少なく，植物工場は発展途上の段階にあるとされている[3][4].

植物工場で栽培する主要品目に，果菜類と呼ばれるトマ

†1　公立はこだて未来大学システム情報科学研究科
Graduate School of Systems Information Science, Future University Hakodate.
†2　公立はこだて未来大学
Future University Hakodate
†3　株式会社アプレ
Apure Inc.

トやキュウリ等の果実を食用とする野菜が挙げられる．果菜類は生育に強い光を必要とすることから，太陽光を主に利用して栽培する太陽光型植物工場で主に栽培されている．果菜類は，植物工場の主要な栽培品目であるため，収量予測に基づいた適切な生産・販売計画が求められる．しかし，果菜類の収量予測は困難である．

植物工場の果菜類の収量予測を困難にしている要因には，主に 3 つある．1 つ目の要因に，不安定な気象変動の影響が挙げられる．果菜類は太陽光型植物工場で栽培されているため，太陽光や気温等の気象に左右される．特に，梅雨や冬場等の不安定な天候が多い時期はその影響を受ける．2 つ目の要因に，果菜類の生物学的な特徴が挙げられる．果菜類は開花や結実という生育期間を持つため，葉菜類等の他の種類の作物と比較し，環境の影響を受ける期間が長いという特徴がある．1 つの株から複数の実を収穫可能であることも予測を難しくさせている一因である．最後に，3 つ目の要因に，人為的な要因が挙げられる．植物工場の収量は，出荷・調製された量であり，栽培されている野菜の収穫可能量ではない.例えば，出荷計画に応じて意図的に収穫しないといいったことが挙げられる．植物工場の果菜類を対象とした研究には，施設内の最適な環境制御の方法を調査した研究[5]や作物の生育を解析した研究[6]がある．しかし，人為的な要因も関係することから，実際の施設の環境の条件下で栽培された作物に，これらを直接適用して収量を予測することは困難である．

以上のことから，本研究の目的は，太陽光型植物工場の果菜類の収量に関わる要因を分析し，収量予測モデルを開発することである．研究目的を達成するための目標として，果菜類の収量を予測するための予測モデル生成プロセスを開発することを研究目標とする．予測モデルの生成プロセスを明確にすることにより，予測モデルの開発効率性や正確性の向上を目指し，複数の果菜類や他の施設においても予測モデルを適用可能にすることも目指す．

2. 関連研究とその課題

2.1 農業分野における予測モデル生成プロセスの開発

予測モデル生成プロセスを農業分野において活用した研究として，耕作放棄量予測モデル生成プロセスを開発した研究がある[7]．この研究では，地域ごとに耕作放棄の特性が異なることから全てを説明する汎用的なモデルの構築が困難という問題に対し，様々な地域に適応可能な予測モデル生成プロセスを開発した．予測モデル生成プロセスとして，(1)モデル生成用データの選定，(2)応答変数および説明変数の設計，(3)モデル構造の選定，(4)モデルチューニングとバリデーション，の 4 つのサブプロセスを提案し，各サブプロセスの詳細を述べている．実験では，予測モデル生成プロセスを耕作放棄があった異なる地域に適用し，中高耕作放棄ケースにおける提案手法の有効性を明らかにした．

この研究の課題として，予測モデルの開発に必要なサブプロセスが網羅的に定義されていない点が挙げられる．例えば，予測モデルを開発する上で一般的に最も重要とされている前処理に関するサブプロセスが含まれていない．予測モデルの開発に必要なサブプロセスを全て定義しなければ他の地域で再現・適用することが難しいと考える．

2.2 特徴量の抽出方法

特徴量の抽出の過程に関する関連研究の課題について述べる．関連研究では基本統計量のみを用いており，特徴量のバリエーションが少ない．そのため，収量に関係する特徴量を見つけ出す作業に時間を要し，モデルの精度を最大まで高められないという課題があると考える．今回の対象データである収量実績のデータは，時系列データに分類される．時系列データは，一つ一つのデータセットが密接に関係していることが特徴であるため，その関係性を表せるような特徴量の生成を行い，目的変数との関係性を見つけ出すことが重要である．そのため，時系列データの特徴量を漏れなく抽出可能な特徴量を生成する方法を検討する必要がある．その際に，効率的かつ正確に特徴量を抽出できるようすることもまた重要である．時系列データの特性から，時系列データでは無数に特徴量を抽出できるため，特徴量の検討や抽出に多くの時間や労力を要するからである．

2.3 予測手法の選定

機械学習は高精度な収量の予測を実現している[8][9]．その一方で，機械学習手法を収量予測において活用するにはいくつかの障壁がある．その一つに，機械学習の有用性は非常に大量の学習データを準備できるか否かに大きく依存するといった点が挙げられる[10]．この点は，多くの場面でビジネス応用する際に課題となり得る．今回実験対象としている植物工場においても，機器の導入による環境変動や栽培品種の変動が著しいため，膨大な数のデータセットを用意することが難しい.それゆえ，比較的少ないデータ数でも利用可能な予測モデル予測手法が求められる.その他では，解釈の困難性に関する問題がある．機械学習は，処理プロセスをブラックボックス化してしまうため，どうしてそのような推定結果になったのかが陽に提示できないという解釈の困難性が度々指摘されている．農業従事者にとって納得感のない予測モデルは受け入れ難いものであるため，農業従事者が予測モデルの理論を理解可能なモデル開発が重要である．

3. 提案手法

3.1 研究課題と解決アプローチ

以上の関連研究での課題より，本研究における研究課題は，次に示す通りである．

課題 1: 予測モデルを構築するまでの全プロセスが不明確
課題 2: 時系列データから効率的かつ正確に有効な特徴量を抽出する方法の選定

課題3: 少ないサンプル数でも予測精度の高いモデルの構築が可能，かつモデル構造が解釈可能な予測手法の選定

上記で述べた研究課題に対する解決アプローチは，次に示す通りである．アプローチの詳細は，3.2 節で述べる．

アプローチ 1: 予測モデルを構築するまでの全プロセスを網羅的かつ明確に定義

アプローチ 2: 特徴量抽出と特徴量選択アルゴリズムを含む時系列分析用パッケージ tsfresh の利用

アプローチ 3: 予測手法として統計モデリングの利用

3.2 予測モデル生成プロセスの構築方法

2.1 節で述べた関連研究[7]における課題や一般的な予測モデル構築の手順を参考にして，6 つのサブプロセスを定義する．そのサブプロセスとは，(1) モデル生成用データの選定，(2) データの前処理，(3) データの可視化，(4) 特徴量の設計,(5) 予測手法の選定， (6) モデルの適正化である．本研究では，これらのプロセスを簡易な演算手順として定式化することで，そのプロセスに従えば様々な施設固有の状況に応じた収量に関係する要因の特定と収量を精度良く予測ができるモデルの生成を可能とすることを目指す．以下で，(1) から(6) の各サブプロセスについて具体的に説明する．

3.2.1 モデル生成用データの選定

予測モデル生成に用いるデータとして，過去の収量の実績データ，施設内の環境データ， 屋外の気象データの 3 種類のデータを用いることにする．主な理由は 2 つある． 1 つ目は，2.2 節「関連研究」でも述べたように，温室の果菜類を対象として収量予測をした研究において，これらの 3 種類のデータが特徴量として有効であるとの報告があったからである．温室と植物工場は施設内の環境制御の仕方が異なるが，「施設内」で作物を栽培している大きな特徴が共通している．それゆえ，植物工場の果菜類にも有効である可能性が高いと考えるため，それらの属性を利用する． 2 つ目は，植物工場において継続的に取得・記録されているデータであるからである．関連研究の中には，植生指標である NDVI や作物の茎径の測定データ等の作物の外観に関する データを用いたものがある[11][12][13]．しかし，これらの方法では外観データをセンシングするための機器を新たに設置する必要があるため，データ収集と導入コストが高い．その点， 施設内外の環境データと収量の実績データは，一般的な植物工場では取得・記録していることが一般的であるため，データ収集が容易である．以上の理由から，施設内外の環境データと収量の実績データを用いる．

3.2.2 データの前処理

実施する前処理の種類は，主に 3 つである．

1 つ目は，欠損値に関わる処理である．センシング機器の故障等による欠損値に対しては，欠損値があった日の前後 3 日間の平均値で補完する．

2 つ目は，データスケールの統一に関わる処理である．非正規分布のデータにも対応できるようにするため，ロバストｚスコアによる標準化処理をする．ロバストｚスコアには，中央値は分布の代表値として分布形に影響されないという利点や，四分位はバラツキの統計量として分布両端の外れ値に影響されないという利点があるため選択する．

3 つ目は，予測に利用する環境データの期間に関わる処理である．その前処理の方法とは，予測日より 7 日前からの 1 日の平均値の累積値で作成するというものである．この期間は，ミニきゅうり，ミニトマト共に，開花から収穫までの果実成長期間に相当する．この期間の環境要因が収量に大きな影響を及ぼすとされているため[9]，この期間のデータを用いることとした．加えて，1 日の中の日中の時間帯のみで属性を作成する．日中の時間帯は年間を通して変化するため，実験対象の施設がある地域の日の出入りの時刻の 1 年の平均値を利用する．

3.2.3 データの可視化

目的変数との関係性や特徴量間の関係性の分析，データ内の異常値や外れ値を見つけることを目的にデータの可視化を行う．実施するデータの可視化項目と概要について，以下で説明する．

(1) 週別収量の傾向

週別収量の傾向の可視化では，時系列データに現れるデータの傾向パターンについて分析する．時系列データに現れるデータの傾向パターンとして，傾向変動，循環変動，季節変動，不規則変動の主に 4 つある．

(2) 施設内環境データの傾向

温度や二酸化炭素濃度等の施設内の環境データの 1 日の平均，最高値，最低値，最高値と最低値の差といった複数の統計量を算出し，グラフで可視化することによって，各環境データの日単位，年単位での推移とデータの分布を分析する．

(3) 目的変数との単相関

この分析には散布図とヒートマップ図を用いる．ヒートマップとは，行列型の数値データの強弱を色別に可視化してデータを可視化する方法である．属性間の相関関係をより明確に把握するために，散布図とあわせて相関行列のヒートマップ図を利用する．

3.2.4 特徴量の設計

本項では，3.1 節で述べた課題解決アプローチ 2 の「特徴量抽出と特徴量選択アルゴリズムを含む時系列分析用パッケージ tsfresh の利用」についての詳細を述べる．

特徴量の設計には，特徴量の抽出方法と特徴量の選択方法に関しての 2 つのステップが含まれる．特徴量の抽出では，3.1 節の解決アプローチでも述べた時系列データの特徴を抽出した様々な特徴量を生成する．時系列データ特有の様々な特徴量を抽出するため，tsfresh(Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests)[14]という Python パッケージを用いる．tsfresh は，特徴量抽出と特徴

量選択アルゴリズムを含む時系列分析用の Python パッケージである．図 1 に tsfresh で提供されている機能の概念図を示す．tsfresh は 63 の特徴化手法を提供している．例えば，ピーク数，中央値，標準偏差，フーリエ変換，自己相関係数を用いた特徴量，時間反転対称性特徴量がある．tsfresh は時系列データ特有の特徴量を網羅的に生成することが可能であり，抽出・選択処理の並列化により実行時間を大幅に短縮することが可能なため利用する．また，tsfsresh は多くの論文で活用の報告がされており，これまでに病気の予測，機械の障害発生の検出，交通量の予測などの多様な活用事例があることから，今回のケースにおいても十分に適用可能であると考えた．

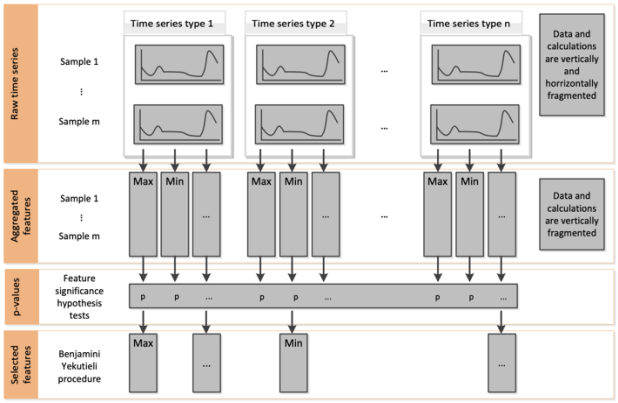

図 1: 特徴量抽出から選択までのデータ処理の流れ
(文献[15]から引用)

次に，特徴量選択について説明する．tsfresh によって多くの特徴量が抽出されるため，効果的な特徴量選択をする必要がある．本研究では，特徴量選択を 4 段階に分けて実施する．

1 段階目は，tsfresh で提供している select_features モジュールを用いた方法である．tsfresh では特徴量の抽出だけでなく，特徴量を選択するモジュールも提供している．select_features は，統計的な有意差がありそうな特徴量のみになるように統計的仮説検定を用いて特徴量を選択するモジュールである．特徴量の有意性検定法には，教師付き機械学習問題のタイプ（分類/回帰）と特徴量のタイプ（カテゴリカル/連続）に応じて，Exact Fisher test of independence, Kolmogorov-Smirnov test, Kolmogorov-Smirnov test, Kendal rank test の 4 つの方法が用いられる[15]．これらの方法による検証の結果は，p 値のベクトルであり，ラベルやターゲットを予測するための各特徴の重要性を定量化する．p 値は保持する特徴量を決定するために，Benjamini-Yekutieli 手順に基づいて評価される．

2 段階目は，相互情報量(Mutual Infomartion)を用いた方法である．相互特徴量は，ある特徴量 X と別の特徴量 Y の間の同時分布 P(X,Y)と個々の分布の積 P(X)P(Y) がどれだけ似ているのかを算出する．もし互いに独立であれば Mutual Infomartion は 0 になる．相互特徴量により，特徴量の数を「学習データのサンプル数/10」となるように選択する．なぜなら，学習データの数より特徴量の数が多すぎることは，ノイズのパターン学習や学習スピードの観点から適切でないからである．

3 段階目は，一般的な統計的変数選択手法である赤池情報量基準(AIC)に基づく stepwise 法である．AIC は，次の項で説明する重回帰分析と一般化加法モデルの予測モデルにおいて，最も良い特徴量の組み合わせを調べるために用いる．もう一つ予測手法として利用する多変量適応的回帰スプラインは，アルゴリズム内で特徴量選択を自動してくれるため，この特徴量選択方法を適用するのは重回帰分析と一般化加法モデルの 2 つの手法である

最後に，4 段階目は，分散拡大要因(VIF: Variance Inflation Factor)を用いた方法である．VIF は多重共線性が発生していないかを確認するために用いられる．多重共線性がある場合は該当している変数を削除することで，特徴量を選択する．

3.2.5 予測手法の選定

3.1 で述べた「少ないデータ量かつモデル構造が解釈可能な手法の利用」のアプローチの詳細について述べる．利用する予測手法の選定基準は，i. 予測モデルに対する特徴量の回帰構造の探索と評価が可能, ii.少ないデータ数でも予測精度の高いモデルの構築が可能という 2 点とした．i と ii を担保できる以下の 3 つを予測手法の候補として選定する．

第 1 の候補は，重回帰分析(MLR: Multiple Linear Regression)である．MLR は 2 つ以上の特徴量を使用して連続値の目的変数を予測するための一般的な統計手法である．MLR は, 露地栽培の収量予測において多く用いられており，露地栽培の果菜類の収量予測においても多く用いられている．MLR によって，一定の予測精度を得ていることから，本研究でも MLR を選択する．

第 2 の候補は，一般化加法モデル(GAM: Generalized Additive Model)である．GAM は，Generalized Linear Model の各特徴量に重みをつけるだけでなく，関数とする事で複雑な現象も表現することができる非線形にも対応したモデルである．GAM は，目的変数と特徴量の関係性が分かりやすいという線形モデルの利点を保ったまま，機械学習のモデルと同レベルでの精度の予測が可能であるため，選定する．

最後に，第 3 の候補として，多変量適応的回帰スプライン(MARS: Multivariate Adaptive Regression Splines)を選定した． MARS では，CART(CART: Classification And Regression Trees)のステップ関数近似を打ち切りベキ乗基底関数に変更し，局所的に線形モデルをあてはめることで予測精度を高めたものである．GLM と比較して特徴量間の交互作用を樹木構造でティッピングポイントも含めて明示的に表現できるところに強みがある[7]ため，選定する．

3.2.6 モデルの適正化

予測モデルの評価は，予測精度，解釈性，納得性によって多面的な評価を実施する．

まず，1 つ目の評価指標である精度についてである．精度の評価には，平均絶対誤差率(MAPE: Mean Absolute Percentage Error)を利用する．MAPE は，誤差の絶対値を実測値で割って誤差率とし，これを平均化したものである．本研究では複数の作物を対象としているため，異なる作物の精度を比較できることが必要である．そのため，他の作物と精度の比較がしやすい MAPE を評価指標として用いることとした．データセットの分割は，75%を学習データに，25%をテストデータに分割して leave-one-out 法によって評価をする．

次に，2 つ目の評価指標である解釈性について説明する．解釈性には，何に対する解釈性かで 2 つの評価観点があるとされている[16]．一つは，予測結果に対して特徴量の何が影響してその結果になったのかを解釈できるかである．もう一つは，予測モデルが特徴量の何を重視しているかを解釈できるかである．これらの 2 つの評価観点は，解釈しやすい予測手法を利用することで条件を満たすと考えている．本研究では，結果を解釈しやすい統計モデリングの手法を予測手法として用いるため，この評価観点については既に満たすと考える．

最後に，3 つ目の評価指標である納得性について説明する．納得性は，主に現場の人が予測結果の根拠に納得できるかを評価するためのものである．納得性を評価するために，関連研究[17]を参考に，過去の実績値と予測値，寄与率などをレポート形式にしてまとめる．作成した分析結果のレポートを用いて，植物工場の従業員に開発した予測モデルについて説明を行い，予測モデルに対しての納得感についてヒアリングすることで，納得性の評価をする．

4. 実験

4.1 対象施設

本研究の対象とする実験施設は，北海道函館市にある太陽光型植物工場(以下，植物工場 A と表記)である．植物工場 A は水耕栽培による野菜の生産・直販を行い，果菜類 7 種，葉菜類 17 種の計 24 種という多品目の野菜を栽培している．温度，湿度，CO2 濃度，養液濃度等のハウス内の環境データのセンシングに加え，気温，湿度，光量等の気象データのセンシングも行なっている．植物工場 A での果菜類の収量予測の需要については，マネジメント業務を担う社員と実際に生産業務を担う社員へのヒアリングから確認した．

4.2 対象作物

ミニきゅうりの(ラリーノホワイト)とミニトマト（アイコ）を予測対象の作物とする．この 2 つの作物は日本の施設園芸作物の栽培割合の上位 3 つに入る主要品目である．植物工場 A においても主要栽培作物であるため，栽培面積が大きく，ほぼ毎日出荷をしている．以上のことから収量予測モデルを開発する需要が高いと考えたため，実験対象の作物とした．

4.3 問題設定

植物工場の従業員へのヒアリングにより，目的変数として，収量を予測する日(以下，予測日)の次の日から 1 週間分の収量(以下，週別収量)を設定した．現在植物工場 A では果菜類の収量予測が全くできていないため，なるべく早く活用段階に持っていくことが求められる．これらの事情と予測モデルが実用可能とされる一般的な精度基準を踏まえて，予測精度の目標値として，MAPE 20.0%以内を目指す．

4.4 モデル生成用データの選定と前処理

モデル構築とその評価に用いるデータは，ミニきゅうりは 2018 年 4 月～2019 年 5 月の全 59 週分，ミニトマトは，2018 年 2 月～2019 年 6 月の期間の全 66 週分のデータである．属性の選定は，3.2.1 項「モデル生成用データ選定」で説明した属性の選択基準に基づいて行った．また，3.2.2 項「前処理」で説明した前処理をした．その結果，表 1 の属性を作成した．

表 1: 作成した属性

表記	属性の名前
WY(Weekly yield)	週別収量[kg]
PY01(Past weekly yield a week ago)	1週間前の収量[kg]
PY02(Past weekly yield two week ago)	2週間前の収量[kg]
PY03(Past weekly yield three week ago)	3週間前の収量[kg]
PY04(Past weekly yield forth week ago)	4週間前の収量[kg]
AT(Air temperature)	外の気温(℃)
AH(Air humidity)	外の湿度(%)
SR(Solar radiation)	日射量(kWh m^{-2})
FT(Temperature in the facility)	施設内の温度(℃)
FCD(Carbon dioxide concentration in the facility)	二酸化炭素濃度(ppm)
FS (Satiety in the facility)	飽差(g/m^3)

4.5 データの可視化

4.5.1 収量の推移の分析

図 2 は，予測対象の週別収量の推移をグラフにしたものである．週別収量の平均値は，ミニきゅうりが 135kg，ミニトマトが 24.5kg であった．ミニトマトと比較して，ミニきゅうりは収量の変動幅が大きいことがわかった．時系列データの傾向分析では，ミニきゅうりの収量は季節変動があることがわかった．図 2 のミニきゅうりの収量が大きく増加している時期は夏場の時期に相当することから，光量が増大する夏場にかけて収量が最大になっていることが分かった．一方，ミニトマトの収量は循環変動とわずかに季節変動の傾向があることが分かった．図 2 より収量が大きく増減する周期がいくつか発生していることから，期間を空けて新たな実が成るという果菜類の生物学的特徴が見られたと考えられる．また，ミニトマトにおいて季節変動の傾向があまり出なかった要因を調査するため，植物工場の従業員にヒアリングしたところ，対象データの期間はミニトマト以外の他の作物の栽培に力を入れていていたため，ミニトマトの栽培管理が疎かになっていたことが分かった．具体的には，廃棄すべき株をいつまでも残し栽培していた

ことや，害虫被害への対応が遅れてしまったことが挙げられる．以上のことより，栽培管理を適切に行えていると，ミニきゅうりのように，夏場の時期が最も収量が高くなるような季節変動の傾向になると推測される．

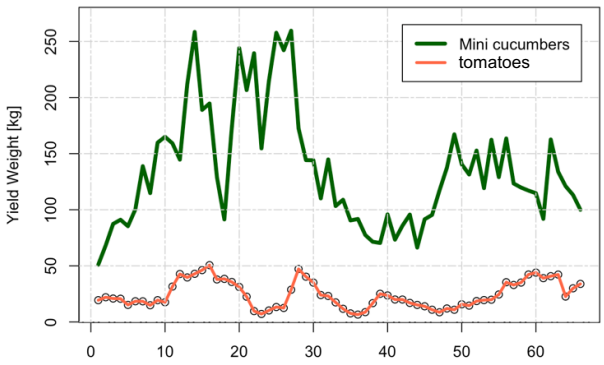

図 2：両作物の収量の推移

4.5.2 属性間の関係性の可視化

目的変数と特徴量の関係性と特徴量間の関係性の分析をした結果について説明する． 図 3，図 4 に相関係数のヒートマップ図を示す．単相関係数の算出には，Spearman の順位相関係数を用いた．

データの分析の結果について説明する．まずミニきゅうりにおいて収量との相関関係が最も強く見られたのは， 1 週間前の収量(PY01)の単相関係数 0.79，続いて，二酸化炭素濃度(FCD)が -0.71, 気温(AT)が 0.69 という結果となった．一方で，ミニトマトにおいて目的変数との相関関係が最も強く見られたのは，1 週間前の収量(PY01)の 0.83，続いて，二酸化炭素濃度(FCD)が -0.34，気温(AT)が 0.33 という結果となった．

ミニきゅうりとミニトマトの両方の作物で，概ね同じ属性が収量との相関関係が強くなるという結果となった．これらの相関関係が強く見られた属性をモデルに加味することによって，モデルの精度向上につながる可能性が高い．各属性の相関係数の値の観点では，各作物で大きな差異が見られた．それゆえ，作物によって環境要因の影響度が異なると推察される．

図 3：ミニきゅうりの相関行列のヒートマップ図

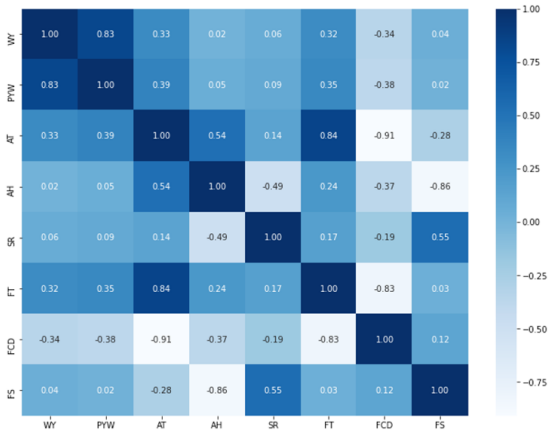

図 4：ミニトマトの相関行列のヒートマップ図

4.6 特徴量の設計

まず，表 1 の属性を tsfresh の extract_features モジュールによって，時系列データの特徴を反映した特徴量を抽出した． その結果，ミニきゅうりでは，13,243 個の特徴量を抽出し，ミニトマトでは 5,278 個の特徴量を抽出した．次に，同じく tsfresh の select_features モジュールを用いた統計的仮説検定により，ミニきゅうりは 265 個，ミニトマトは 69 個まで特徴量を絞り込みした．次に，相互特徴量により，ミニきゅうりとミニトマトでそれぞれ 5 個の特徴量を選択した．最後に，VIF の値を算出したところ，全ての特徴量が 5 未満であったため，多重共線性の疑いは低いことから 5 つ全ての特徴量を利用することとした．

特徴量選択の結果，得られた特徴量を表 2 と表 3 に示す．特徴量名は以下の 3 つの要素から構成される．(1) 特徴量を抽出する時系列属性，(2)その特徴量を抽出するために使用された特徴量計算機の名前，(3) 特徴量計算機を構成するパラメータのキーと値のペア:

[kind] _ [calculator] _ [parameterA] _ [valueA] _ [parameterB] _ [valueB]

表 2: ミニきゅうりの特徴量選択の結果

表記	特徴量名	説明
C_TS_SRMAX	solarRadiation_max_period_7	予測日 7 日間前からの期間の光量の最大値
C_TS_CDC	carbDioxConcent_average_period_7	予測日 7 日間前からの期間の CO^2濃度の累積値
C_TS_ATMIN	airTemp__min_period_7	予測日 7 日間前からの期間の気温の最小値
C_TS_PYW	pastYield01__cwt_coefficients__coeff_0__w_20__widths_(2, 5, 10, 20)	1 週間前の収量のRicker ウェーブレットの連続ウェーブレット変換値
C_TS_AT	airTemp_average_period_7	予測日 7 日間前からの期間の気温の累積値

表 3: ミニトマトの特徴量選択の結果

表記	特徴量名	説明
T_TS_PYW	pastYieldWeight_cwt_coefficients_widths_(2, 5, 10, 20)_coeff_0_w_2	予測日 1 週間前の収量のRickerウェーブレットの連続ウェーブレット変換値
T_TS_AT	airTemp_average_period_7	予測日 7 日間前からの期間の気温の1 日の平均の累積値
T_TS_SR	solarRadiation_abs_energy	日射量の2 乗値の合計値
T_TS_FT	facilityTemp_fft_coefficient_coeff_0_attr_angle	施設温度に関する 1 次元離散高速フーリエ変換のフーリエ係数値
T_TS_CDC	carbonDioxideConcentration_fft_coefficient_coeff_0_attr_angle	CO^2濃度の1 次元離散高速フーリエ変換のフーリエ係数値

選択された特徴量について説明する．まず，時系列データ特有の特徴が反映された特徴量について説明する．表 2,3 の 1 次元離散フーリエ変換やリッカーウェーブレットの連続ウェーブレット変換を用いた C_TS_CDC，C_TS_SR, T_TS_PYW, T_TS_HT, T_ TS_CDC がそれにあたる．これらの特徴量は周期的な変化を表す特徴量である．この周期性は，週別収量の傾向の分析時にも季節変動や循環変動として両作物において確認されていたことから，特徴量の選択時に選ばれたことでも納得がいく．

4.7 予測手法の選定とモデルの適正化

学習データ 66 件のうち，50 件(約 75%)を学習データに，16 件(約 25%)をテストデータに分割した．各予測手法を学習データにおいて，最適な特徴量の組み合わせとハイパーパラメータの探索をした．表 4, 表 5 に, その結果を示す．対象作物や予測手法によって，選択された特徴量が異なった．特徴量には，相関分析の際に収量との相関関係が高いとされた属性が多く選ばれる傾向となった．表 6 には，チューニングした特徴量とハイパーパラメータ を用いて，テストデータにおいて予測した精度の結果を示す．両作物とも予測精度が最も高かったのは GAM を用いた予測モデルであった． tsfresh の有効性を評価するため，tsfresh を利用して特徴量抽出をした場合と tsfresh を利用せずに表 1 の特徴量をそのまま使った場合を比較した．比較する予測手法には，各作物で最も高い精度を示した GAM を用いた. 表 7 にその結果を示す. 結果として，tsfresh 利用によって全体的に予測精度が向上したことを確認した．

表 4: ハイパーパラメータのチューニング結果

対象作物	MLR	GAM	MARS
ミニきゅうり	なし	select=FALSE method=REML	degree=1 nprune=2
ミニトマト	なし	Select = TRUE Method = REML	degree=1 nprune=12

表 5: 特徴量のチューニング結果

対象作物	MLR	GAM	MARS
ミニきゅうり	C_TS_AT	C_ TS_AT C_ TS_PYW	C_ TS_AT
ミニトマト	T_ TS_PYW T_TS_SR T_ TS_CDC	T_ TS_PYW	T_ TS_PYW T_ TS_SR

表 6: 各予測手法による予測精度の結果

作物	平均絶対誤差率 (MAPE[%])		
	MLR	GAM	MARS
ミニきゅうり	19.2	**17.0**	17.7
ミニトマト	42.0	**19.8**	42.6

表 7 ： tsfresh の利用有無による予測精度の比較

tsfresh利用有無	ミニきゅうり		ミニトマト	
	Train	Test	Train	Test
tsfresh有	**13.2**	**17.0**	25.7	**19.8**
tsfresh無	16.9	22.9	19.7	19.9

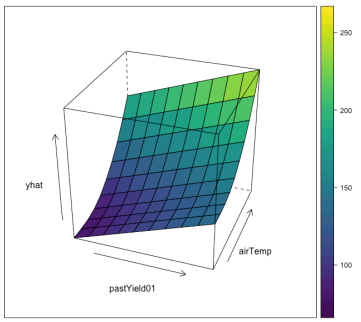

図 5：週別収量と特徴量との関係性を表した 3D モデル

4.8 考察

まず，精度の良し悪しの観点で考察を述べる．ミニきゅうりの精度はMAPEで17.0%，ミニトマトは19.8%となった．この精度は，4.3節「問題設定」で述べた，精度の目標値であるMAPE = 20.0%を満たす結果となった．

次に，予測モデルに組み込まれた特徴量について考察を述べる. まず, ミニきゅうりでは, 最も精度が良かった GAM の予測モデルに, 予測日一週間前の収量に関する特徴量（C_ TS _PYW）と気温に関する特徴量（C_ TS _AT）の 2 つの特徴量が含まれる結果となった．図 5 より，気温と一週間前の収量が高いほど，週別収量も大きくなる傾向があることが分かった．一方，ミニトマトでは一週間前の収量(T_ TS_PYW)に関する特徴量のみが選ばれる結果となった．この結果は，4.5.1 項「収量の推移の分析」の際に分かった，ミニトマトの収量の変動幅の小ささや周期変動の傾向があ

るという分析結果とつながる．

最後に，植物工場の従業員に，精度と納得性の観点で予測モデルに対する意見をヒアリングした結果について説明する．まず，精度に関して述べる．植物工場の従業員より，今回利用したデータを用いた週別収量の予測精度は，今回の実験結果の精度で最善の精度であるとコメント頂いた．しかし，今回開発した予測モデルでは収量の増減が正確に捉えきれていないため，実際に活用するのは難しいとのご意見を頂いた．納得性に関しては，予測モデルに対して現場の感覚と一致するとの意見で納得感を示す回答であった．

5. おわりに

5.1 まとめ

本研究の目的は，太陽光型植物工場の果菜類の収量に関わる要因を分析し，収量予測モデルを開発することである．目的を達成するために，解決アプローチを(1) 予測モデルを構築するまでの全プロセスを網羅的かつ明確に定義，(2)特徴量抽出と特徴量選択アルゴリズムを含む時系列分析用パッケージ tsfresh の利用，(3)モデル構造が解釈可能な統計モデリングの利用として，これらのアプローチを含めた，(1)モデル生成用データの選定，(2)データの前処理，(3)データの可視化，(4)特徴量の設計，(5)予測手法の選定，(6)モデルの適正化の 6 つのサブプロセスから成る収量予測モデル生成プロセスを開発した．提案手法の有効性を確かめるために，開発した予測モデル生成プロセスをミニきゅうりとミニトマトに適用して実験した．予測モデルの構築には，施設内外の環境データと過去の収量の実績データを利用し，重回帰分析，一般化加法モデル，多変量適応的回帰スプラインを用いた．実験の結果，両作物において最も精度が高かったのは一般化加法モデルを用いた予測モデルで，両作物とも目標値である平均絶対誤差率 20%以内を達成した．加えて，予測モデルの納得性を評価するために実施した植物工場の従業員へのヒアリングによって，今回開発した予測モデルは，現場の感覚と一致する納得感のある予測モデルであることが明らかとなった．以上より，予測精度，解釈性，納得性の 3 つの評価観点で提案手法の有効性を確認した．

5.2 今後の課題

今後の課題は以下の通りである．

(1) 新たな特徴量の検討

予測モデルの精度向上には，生産管理方法や出荷調整方法等の人為的要因に関する特徴量を予測モデルに加味する必要があると考える．

(2) 目的変数の変更

本研究では，目的変数として，植物工場 A で栽培している対象作物の「全ての株からとれる週別収量」を設定していた．そのため，ミニきゅうりのように，株数が多く，収量の変動幅が非常に大きい作物における収量予測がより困難であった．この問題に対し，「一部の株を対象にした週別収量」を設定するなどして，目的変数を細分化して考える方法が有効であると考えている．

参考文献

[1] 農林水産省：施設園芸をめぐる情勢（ オンライン），入手 先⟨https://www.maff.go.jp/j/seisan/ryutu/engei/sisetsu/attach/pdf/index-20.pdf⟩（参照 2020-12-24）．

[2] 農林水産省：植物工場の説明 一般財団法人・社会開発研究センター（オ ンライン），入手先⟨http://www．maff．go．jp/j/heya/sodan/1308/01.html⟩（参照 2020-12-24）．

[3] 藤本真狩：日本における植物工場の現状と今後の展望，精密工学会誌, Vol. 81, No. 9, pp. 811–814 (2015).

[4] 土屋和：植物工場をめぐる現状と課題，一般社団法人日本施設園芸協会（オンライン），入手先⟨https://www.alic.go.jp/content/000127162.pdf⟩（参照 2020-12-24）．

[5] Cunha, J. & Moura Oliveira, Paulo. :Optimal Management Of Greenhouse Environments, Proceedings of EFITA 2003 Conference (2003)

[6] R. Xu, J. Dai, W. Luo, X. Yin, Y. Li, X. Tai, L. Han, Y. Chen, L. Lin, G. Li, C. Zou, W. Du, M. Diao: A photothermal model of leaf area index for greenhouse crops, Agricultural and Forest Meteorology,Vol. 150, No. 4, pp.541-552, (2010).

[7] 町村尚，松井孝典：機械学習アルゴリズムによる耕作放棄の要因分析および予測モデル の開発，土木学会論文集 G（環境），Vol. 70, No. 6, pp. 131–139 (2014).

[8] Alhnaity, B., Pearson, S., Leontidis, G. and Kollias, S. D.: Using Deep Learning to Predict Plant Growth and Yield in Greenhouse Environments(2019).

[9] González-Sanchez, A., Frausto-Solis, J. and Ojeda, W.: Predictive ability of machine learning methods for massive crop yield prediction, SPANISH JOURNAL OF AGRICULTURAL RESEARCH, (2014).

[10] Marcus, G.: Deep Learning: A Critical Appraisal., CoRR, Vol. abs/1801.00631 (2018).

[11] Lin, W. and Hill, B.: Neural network modelling to predict weekly yields of sweet peppers in a commercial greenhouse, Canadian Journal of Plant Science, Vol. 88, pp. 531–536 (2008).

[12] Stas, M., Van Orshoven, J., Dong, Q., Heremans, S. and Zhang, B.: A comparison of machine learning algorithms for regional wheat yield prediction using NDVI time series of SPOT-VGT, IEEE, pp. 258–262 (2016).

[13] Pantazi, X., Moshou, D., Alexandridis, T., Whetton, R. and Mouazen, A.: Wheat yield prediction using machine learning and advanced sensing techniques, Computers and Electronics in Agriculture, Vol.121, pp.57– 65 (2016).

[14] Christ, M., Braun, N., Neuffer, J. and Kempa-Liehr A.W. (2018). Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh -- A Python package). Neurocomputing, Vol. 307, pp. 72-77(2018).

[15] M. Christ, A.W. Kempa-Liehr, M. Feindt, Distributed and parallel time series feature extraction for industrial big data applications. Asian Machine Learning Conference (ACML) 2016, Workshop on Learning on Big Data (WLBD), Hamilton (New Zealand)(2016).

[16] 本橋洋介：機械学習を用いた業務システムの機能と評価に関する考察，技術報告.

[17] 梅津圭介，本橋洋介：分析成果の業務活用とモデルの解釈性についての一考察，人工知能学会全国大会論文集，Vol. JSAI2016, pp. 3K34–3K3）, DOI: 10.11517/pjsai.JSAI2016.0_3K34 (2016).